

COMMUNICATIONS

CACM.ACM.ORG

OF THE

ACM

12/2016 VOL.59 NO.12



Interactive Visualization of 3D Scanned Mummies at Public Venues

Anticipating Policy and
Social Implications
of Named Data
Networking

3-Sweep

Function at Scale

Can We Trust
Autonomous
Weapons?

Q&A with
Takeo Kanade

acm
creativity
+ cognition
2017

Lifelong Creativity • Learning • Innovation
Singapore, June 27-30
<http://cc.acm.org>

The 11th **ACM Creativity & Cognition (C&C)**
is the first to be run in Asia; bringing together
artists and scientists, designers, educators,
and researchers around the world.

Papers, Pictorials, Artworks, &
Workshops Proposals by
January 6, 2017

Posters, technology demos, &
graduate student symposium by
March 10, 2017



Creativity and Cognition 2017
will present papers addressing:

1. Creativity support environments
2. Studies of technology, people, and creativity
3. Creative works that utilize computing to engage, stimulate, and provoke human experience.



Association for
Computing Machinery

 @acm_cc
 acmcc2017

Inviting Young Scientists



Association for
Computing Machinery

Meet Great Minds in Computer Science and Mathematics

As one of the founding organizations of the Heidelberg Laureate Forum <http://www.heidelberg-laureate-forum.org/>, ACM invites young computer science and mathematics researchers to meet some of the preeminent scientists in their field. These may be the very pioneering researchers who sparked your passion for research in computer science and/or mathematics.

These laureates include recipients of the ACM A.M. Turing Award, the Abel Prize, the Fields Medal, and the Nevanlinna Prize.

The Heidelberg Laureate Forum is **September 24–29, 2017** in Heidelberg, Germany.

This week-long event features presentations, workshops, panel discussions, and social events focusing on scientific inspiration and exchange among laureates and young scientists.

Who can participate?

New and recent Ph.Ds, doctoral candidates, other graduate students pursuing research, and undergraduate students with solid research experience and a commitment to computing research

How to apply:

Online: <https://application.heidelberg-laureate-forum.org/>
Materials to complete applications are listed on the site.

What is the schedule?

Application deadline—**February 14, 2017**.

We reserve the right to close the application website early depending on the volume

Successful applicants will be notified by **end of March/early April 2016**.

More information available on Heidelberg social media



Departments

- 6 **From ACM's Committee on Professional Ethics**
The ACM Code of Ethics: A Call to Action
By Marty J. Wolf
-
- 7 **ACM Code of Ethics and Professional Conduct**
Making a Positive Impact: Updating the ACM Code of Ethics
By Bo Brinkman, Don Gotterbarn, Keith Miller, and Marty J. Wolf
-
- 15 **Cerf's Up**
When Email Isn't Private
By Vinton G. Cerf
-
- 17 **Letters to the Editor**
Reclaim the Lost Promise of the Semantic Web
-
- 18 **BLOG@CACM**
ICER 2016, and Star Trek at 50
 Mark Guzdial reports on promising papers, and Daniel Reed recalls a television show that continues to inspire innovation.
-
- 41 **Calendar**
-
- 130 **Careers**

Last Byte

- 144 **Q&A**
A View to the Future
 Takeo Kanede, recipient of the 2016 Kyoto Prize for Advanced Technology for pioneering contributions to computer vision and robotics, sees respect as an important part of Quality of Life Technology initiatives.
By Leah Hoffmann

News



27

- 21 **Cracking the Code on Biology**
 Genetic engineering promises to revolutionize everything from agriculture and biofuels to medicine and food safety.
By Samuel Greengard
-
- 24 **Containers Push Toward the Mayfly Server**
 The container revolution represents a large-scale shift in thinking about multitasking systems.
By Chris Edwards
-
- 27 **Can We Trust Autonomous Weapons?**
 Nations consider using defense systems that can make their own lethal decisions.
By Keith Kirkpatrick

Viewpoints



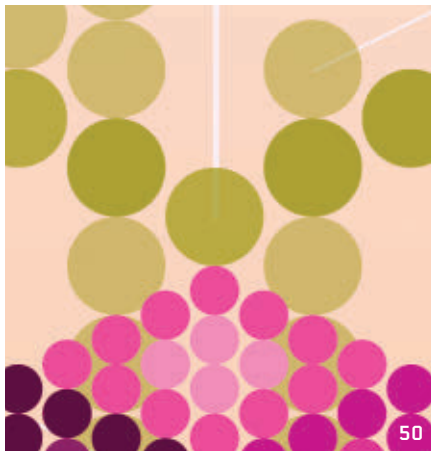
32

- 32 **The Profession of IT Learning to Learn**
 Do you get stuck when it is time to learn something new? Read this.
By Peter J. Denning and Gloria Flores
-
- 37 **Historical Reflections**
Nailing Smoke
 Curation at the bleeding edge of technology.
By David P. Anderson
-
- 40 **Broadening Participation**
Mentoring Female Assistant Professors Enhances Their Success
 A randomized controlled trial validates many of the practices used to retain women in academia.
By Yan Chen
-
- 43 **Viewpoint**
Assessing Abstraction Skills
 What makes a good question?
By Orit Hazzan and Jeff Kramer
-
- 46 **Viewpoint**
(Mis)Managing Parallel Computing Research through EU Project Funding
 Considering the possible negative long-term effects of EU projects on European scientific culture.
By Jesper Larsson Träff



Association for Computing Machinery
 Advancing Computing as a Science & Profession

Practice



50

50 **Functional at Scale**

Applying functional programming principles to distributed computing projects.

By Marius Eriksen

56 **React: Facebook's Functional Turn on Writing JavaScript**

A discussion with Pete Hunt, Paul O'Shannessy, Dave Smith, and Terry Coatta

63 **Industrial-Scale Agile—From Craft to Engineering**

Essence is instrumental in moving software development toward a true engineering discipline.

By Ivar Jacobson, Ian Spence, and Ed Seidewitz



Articles' development led by **acmqueue**
queue.acm.org

About the Cover:

Our cover story tells how museums (in this case, the British Museum) can use visualization techniques and interactive UI to allow visitors to become explorers. Our cover image of the Gebelein Man, a predynastic natural mummy, is a stunning example of how visualization can bring researchers and the public closer together. Cover image courtesy of Interactive Institute Swedish ICT/Interspectral.



Contributed Articles



82

72 **Interactive Visualization of 3D Scanned Mummies at Public Venues**

A full-body virtual autopsy of an ancient Egyptian mummy showed visitors he was likely murdered, stabbed in the back.

By Anders Ynnerman, Thomas Rydell, Daniel Antoine, David Hughes, Anders Persson, and Patric Ljung



Watch the authors discuss their work in this exclusive *Communications* video.
<http://cacm.acm.org/videos/interactive-visualization-of-3d-scanned-mummies-at-public-venues>

82 **Creating Connection with Autonomous Facial Animation**

Biologically based computational modeling promises virtual characters capable of face-to-face human interaction.

By Mark Sagar, Mike Seymour, and Annette Henderson



Watch the authors discuss their work in this exclusive *Communications* video.
<http://cacm.acm.org/videos/creating-connection-with-autonomous-facial-animation>

Review Articles

92 **Anticipating Policy and Social Implications of Named Data Networking**

Considering the social impact of a proposed future Internet architecture.

By Katie Shilton, Jeffrey A. Burke, kc claffy, and Lixia Zhang

102 **Mathematical Foundations for Social Computing**

Social computing benefits from mathematical foundations, but research has barely scratched the surface.

By Yiling Chen, Arpita Ghosh, Michael Kearns, Tim Roughgarden, and Jennifer Wortman Vaughan

Research Highlights

110 **Technical Perspective**
The Chemistry of Software-Defined Batteries
*By Srinivasan Keshav*111 **Software-Defined Batteries**
*By Anirudh Badam, Ranveer Chandra, Jon Dutra, Anthony Ferrese, Steve Hodges, Pan Hu, Julia Meinershagen, Thomas Moscibroda, Bodhi Priyantha, and Evangelia Skiani*120 **Technical Perspective**
3D Image Editing Made Easy
*By Takeo Igarashi*121 **Extracting 3D Objects from Photographs Using 3-Sweep**
By Tao Chen, Zhe Zhu, Shi-Min Hu, Daniel Cohen-Or, and Ariel Shamir



ACM, the world's largest educational and scientific computing society, delivers resources that advance computing as a science and profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources.

Executive Director and CEO
Bobby Schnabel
Deputy Executive Director and COO
Patricia Ryan
Director, Office of Information Systems
Wayne Graves
Director, Office of Financial Services
Darren Ramdin
Director, Office of SIG Services
Donna Cappel
Director, Office of Publications
Bernard Rous
Director, Office of Group Publishing
Scott E. Delman

ACM COUNCIL

President
Vicki L. Hanson
Vice-President
Cherri M. Pancake
Secretary/Treasurer
Elizabeth Churchill
Past President
Alexander L. Wolf
Chair, SGB Board
Patrick Madden
Co-Chairs, Publications Board
Jack Davidson and Joseph Konstan
Members-at-Large
Gabriele Anderst-Kotis; Susan Dumais; Elizabeth D. Mynatt; Pamela Samuelson; Eugene H. Spafford
SGB Council Representatives
Paul Beame; Jenna Neeffe Matthews; Barbara Boucher Owens

BOARD CHAIRS

Education Board
Mehran Sahami and Jane Chu Prey
Practitioners Board
George Neville-Neil

REGIONAL COUNCIL CHAIRS

ACM Europe Council
Dame Professor Wendy Hall
ACM India Council
Srinivas Padmanabhuni
ACM China Council
Jianguang Sun

PUBLICATIONS BOARD

Co-Chairs
Jack Davidson; Joseph Konstan
Board Members
Ronald F. Boisvert; Karin K. Breitman; Terry J. Coatta; Anne Condon; Nikil Dutt; Roch Guerrin; Carol Hutchins; Yannis Ioannidis; Catherine McGeoch; M. Tamer Ozsu; Mary Lou Soffa; Alex Wade; Keith Webster

ACM U.S. Public Policy Office

Renee Dopplick, Director
1828 L Street, N.W., Suite 800
Washington, DC 20036 USA
T (202) 659-9711; F (202) 667-1066

Computer Science Teachers Association
Mark R. Nelson, Executive Director

COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

Communications of the ACM is the leading monthly print and online magazine for the computing and information technology fields. *Communications* is recognized as the most trusted and knowledgeable source of industry information for today's computing professional. *Communications* brings its readership in-depth coverage of emerging areas of computer science, new trends in information technology, and practical applications. Industry leaders use *Communications* as a platform to present and debate various technology implications, public policies, engineering challenges, and market trends. The prestige and unmatched reputation that *Communications of the ACM* enjoys today is built upon a 50-year commitment to high-quality editorial content and a steadfast dedication to advancing the arts, sciences, and applications of information technology.

STAFF

DIRECTOR OF GROUP PUBLISHING
Scott E. Delman
cacm-publisher@cacm.acm.org

Executive Editor
Diane Crawford
Managing Editor
Thomas E. Lambert
Senior Editor
Andrew Rosenbloom
Senior Editor/News
Larry Fisher
Web Editor
David Roman
Rights and Permissions
Deborah Cotton

Art Director
Andrij Borys
Associate Art Director
Margaret Gray
Assistant Art Director
Mia Angelica Balaquiot
Designer
Iwona Usakiewicz
Production Manager
Lynn D'Addesio
Advertising Sales
Juliet Chance

Columnists
David Anderson; Phillip G. Armour;
Michael Cusumano; Peter J. Denning;
Mark Guzdial; Thomas Haigh;
Leah Hoffmann; Mari Sako;
Pamela Samuelson; Marshall Van Alstyne

CONTACT POINTS

Copyright permission
permissions@hq.acm.org
Calendar items
calendar@cacm.acm.org
Change of address
acmhelp@cacm.acm.org
Letters to the Editor
letters@cacm.acm.org

WEBSITE

<http://cacm.acm.org>

AUTHOR GUIDELINES

<http://cacm.acm.org/>

ACM ADVERTISING DEPARTMENT

2 Penn Plaza, Suite 701, New York, NY
10121-0701
T (212) 626-0686
F (212) 869-0481

Advertising Sales
Juliet Chance
acmm mediasales@cacm.acm.org

For display, corporate/brand advertising:
Craig Pitcher
pitcherc@cacm.acm.org T (408) 778-0300

Media Kit acmm mediasales@cacm.acm.org

Association for Computing Machinery (ACM)
2 Penn Plaza, Suite 701
New York, NY 10121-0701 USA
T (212) 869-7440; F (212) 869-0481

EDITORIAL BOARD

EDITOR-IN-CHIEF
Moshe Y. Vardi
eic@cacm.acm.org

NEWS

Co-Chairs
William Pulleyblank and Marc Snir
Board Members
Mei Kobayashi; Michael Mitzenmacher;
Rajeev Rastogi

VIEWPOINTS

Co-Chairs
Tim Finin; Susanne E. Hambrusch;
John Leslie King
Board Members
William Aspray; Stefan Bechtold;
Michael L. Best; Judith Bishop;
Stuart I. Feldman; Peter Freeman;
Mark Guzdial; Rachelle Hollander;
Richard Ladner; Carl Landwehr;
Carlos Jose Pereira de Lucena;
Beng Chin Ooi; Loren Terveen;
Marshall Van Alstyne; Jeannette Wing

PRACTICE

Co-Chair
Stephen Bourne
Board Members
Eric Allman; Peter Bailis; Terry Coatta;
Stuart Feldman; Benjamin Fried;
Pat Hanrahan; Tom Killalea; Tom Limoncelli;
Kate Matsudaira; Marshall Kirk McKusick;
George Neville-Neil; Theo Schlossnagle;
Jim Waldo

The Practice section of the CACM Editorial Board also serves as the Editorial Board of *COMMUNIQUE*.

CONTRIBUTED ARTICLES

Co-Chairs
Andrew Chien and James Larus
Board Members
William Aiello; Robert Austin; Elisa Bertino;
Gilles Brassard; Kim Bruce; Alan Bundy;
Peter Buneman; Peter Druschel; Carlo Ghezzi;
Carl Gutwin; Yannis Ioannidis;
Gal A. Kaminka; James Larus; Igor Markov;
Gail C. Murphy; Bernhard Nebel;
Lionel M. Ni; Kenton O'Hara; Sriram Rajamani;
Marie-Christine Rousset; Avi Rubin;
Krishan Sabnani; Ron Shamir; Yoav Shoham; Larry Snyder; Michael Vitale;
Wolfgang Wahlster; Hannes Werthner;
Reinhard Wilhelm

RESEARCH HIGHLIGHTS

Co-Chairs
Azer Bestavros and Gregory Morrisett
Board Members
Martin Abadi; Amir El Abbadi; Sanjeev Arora;
Michael Backes; Nina Balcan; Dan Boneh;
Andrei Broder; Doug Burger; Stuart K. Card;
Jeff Chase; Jon Crowcroft; Alexei Efros;
Alon Halevy; Norm Jouppi; Andrew B. Kahng;
Sven Koenig; Xavier Leroy; Steve Marschner;
Kobbi Nissim; Guy Steele, Jr.; David Wagner;
Margaret H. Wright; Nikolai Zeldovich;
Andreas Zeller

WEB

Chair
James Landay
Board Members
Marti Hearst; Jason I. Hong;
Jeff Johnson; Wendy E. MacKay

ACM Copyright Notice

Copyright © 2016 by Association for Computing Machinery, Inc. (ACM). Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or fee. Request permission to publish from permissions@hq.acm.org or fax (212) 869-0481.

For other copying of articles that carry a code at the bottom of the first or last page or screen display, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center; www.copyright.com.

Subscriptions

An annual subscription cost is included in ACM member dues of \$99 (\$40 of which is allocated to a subscription to *Communications*); for students, cost is included in \$42 dues (\$20 of which is allocated to a *Communications* subscription). A nonmember annual subscription is \$269.

ACM Media Advertising Policy

Communications of the ACM and other ACM Media publications accept advertising in both print and electronic formats. All advertising in ACM Media publications is at the discretion of ACM and is intended to provide financial support for the various activities and services for ACM members. Current advertising rates can be found by visiting <http://www.acm-media.org> or by contacting ACM Media Sales at (212) 626-0686.

Single Copies

Single copies of *Communications of the ACM* are available for purchase. Please contact acmhelp@cacm.acm.org.

COMMUNICATIONS OF THE ACM

(ISSN 0001-0782) is published monthly by ACM Media, 2 Penn Plaza, Suite 701, New York, NY 10121-0701. Periodicals postage paid at New York, NY 10001, and other mailing offices.

POSTMASTER

Please send address changes to *Communications of the ACM*
2 Penn Plaza, Suite 701
New York, NY 10121-0701 USA

Printed in the U.S.A.



Association for
Computing Machinery





ACM Books



MORGAN & CLAYPOOL
PUBLISHERS

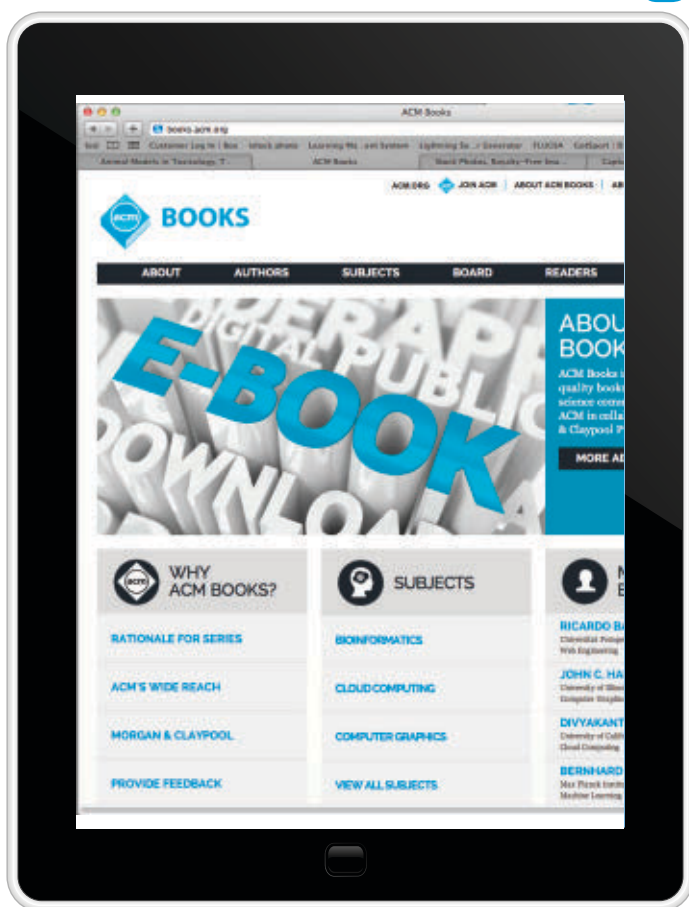
Publish your next book in the ACM Digital Library

ACM Books is a new series of advanced level books for the computer science community, published by ACM in collaboration with Morgan & Claypool Publishers.

I'm pleased that ACM Books is directed by a volunteer organization headed by a dynamic, informed, energetic, visionary Editor-in-Chief (Tamer Özsu), working closely with a forward-looking publisher (Morgan and Claypool).

—Richard Snodgrass, University of Arizona

books.acm.org **ACM Books**



- ◆ will include books from across the entire spectrum of computer science subject matter and will appeal to computing practitioners, researchers, educators, and students.
- ◆ will publish graduate level texts; research monographs/overviews of established and emerging fields; practitioner-level professional books; and books devoted to the history and social impact of computing.
- ◆ will be quickly and attractively published as ebooks and print volumes at affordable prices, and widely distributed in both print and digital formats through booksellers and to libraries and individual ACM members via the ACM Digital Library platform.
- ◆ is led by EIC M. Tamer Özsu, University of Waterloo, and a distinguished editorial board representing most areas of CS.

Proposals and inquiries welcome!

Contact: **M. Tamer Özsu**, Editor in Chief
booksubmissions@acm.org



Association for
Computing Machinery

Advancing Computing as a Science & Profession

The ACM Code of Ethics: A Call to Action

THE ACM COMMITTEE ON Professional Ethics is updating the ACM Code of Ethics and Professional Conduct (the Code).^a It was last changed in 1992. Beginning on page 7 of this issue there is an article that outlines the motivations for the update, describes the update process, includes the first draft of the update, and invites you to take part in this important project. While the draft addresses changes in technology and society, here I reflect on a different view: how those changes motivate a call to action for the profession.

There have been significant changes to the profession of computing in the last 25 years. In 1992, it was still possible for a lone developer to produce a significant piece of software, but that was beginning to change. Numerous free and open source projects were beginning to benefit from collaboration among developers worldwide. Effective software development processes were being shared among experienced developers and being adopted by those new to the profession. Software security and reliability began to change for the better. With the advent of Agile development processes, the circle of collaboration expanded to include the client. The quality of software continued to improve and software itself aligned more with the client's expectations. As the profession welcomed psychologists as collaborators, user interfaces improved. Users became less likely to make mistakes that would lead to harm or even death. Best practices began to mature.

In 1992 there was no doubt that being a computing professional required having a high degree of technical expertise. Today, technical expertise alone is no longer sufficient to be a successful computing professional. Let me explain.

Even those computing professionals with a strong liberal arts background do not have the depth of knowledge and experience in psychology, sociology, philosophy, ethics, and communication theory necessary to fully analyze the impact of some of the complex systems being built. Experts from those fields, and others, who have experience in computing need to be drawn into collaborations as an integral part of the practice of computing in order to advance computing professionalism. It is in this way that the computing profession will offer the best for the common good. Suggested updates to the Code's imperatives emphasize this. The new updated Code imperative 3.7 calls on members to be cognizant of when computing systems are becoming part of the social fabric. This imperative calls attention to the importance of being mindful of the impact computing systems have on others and society in general, as well as being mindful of a system's continuing impact after it has become part of society's infrastructure. Continued stewardship of that system is essential for the responsible computing professional.

The expertise needed to address a broadening range of computing's stakeholders can be found among those studying "computer ethics," understood broadly. As a discipline, computer ethics is even younger than computer science, and in 1992, computer ethics was still in its infancy. Since then, it has experienced growth in both the depth and diversity of scholarship and matured into its own field of study with subdisciplines that intersect with many other disciplines including philosophy, ethics, sociology, communication theory, and design. Some scholars focus on the philosophy of computing. Others focus on computing's impact on society. Still others focus on ethics surrounding the information that is the object of computing. These experts have something to offer the computing pro-

fession, and because of their multidisciplinary focus, they understand the collaborative model.

The ACM Code of Ethics expresses the conscience of the computing profession and is a call to action. It helps professionals address a broader approach to computing. As happened with the sorts of collaborations begun in the 1990s, a renewed widespread, collaborative, and open practice of applying the Code when designing and implementing systems that impact people and societies, will improve our profession. The Code can help us better express the value of the profession, which rests in part on effective ethical analysis and decision making. This in turn, places a special burden on universities to establish a foundation for those abilities in students considering entering the computing profession.

None of this is easy. Some software is of little ethical importance, while other software has the potential to change the nature of what it means to be human. No single computing professional is equipped to think through all of the social and ethical implications of possible systems. It is a complex job and requires special expertise. We can start developing this expertise through collaborations: invite "ethicists" as keynote speakers at technical conferences, send technical experts to computer ethics conferences, hire analysts whose expertise is social change. Updated Code imperative 3.4 is actually a call to action, "The needs of people—including users, other people affected directly and indirectly, customers, and colleagues—should always be a central concern in professional computing." We have benefited from broadened collaboration in the past. More is to be gained by expanding the circle again. 

Marty J. Wolf (mjwolf@acm.org) is the vice chair of ACM's Committee on Professional Ethics and a professor of computer science at Bemidji State University, MN.

^a <https://ethics.acm.org>

Making a Positive Impact: Updating the ACM Code of Ethics

Why Does the Code Need to be Updated?

The ACM website prominently displays: “Advancing Computing as a Science & Profession” and “We see a world where computing helps solve tomorrow’s problems—where we use our knowledge and skill to advance the profession and make a positive impact.”^a These quotes are a high-level description of the goals and purposes of the ACM. The ACM Code of Ethics and Professional Conduct^b (“the Code”) describes what brings us together as a profession. It expresses a social contract we have as professional members of the ACM, a contract describing what we expect of each other and of ourselves as members of the ACM. As members of the ACM we have all consented to the Code; that consent underlies the conscience of our profession and is the foundation for our shared expectations of each other.

The Code provides guidance to ACM members about committing to ethical professional conduct. The Code identifies fundamental considerations for contributing to societal and human well-being. Every ACM member who renews a membership agrees to adhere to this code, a code that was written a quarter of a century ago.

The current version of the Code was approved in 1992. This version of the Code made significant advances over its predecessor. Recognizing that the Code provides guidelines for members of a rapidly developing profession, in 1992 the ACM replaced specific rules that mandated following specific technologies (which might become outdated) with statements of aspirations based on broad ethical principles. In its role of advancing

professionalism and producing a positive impact on society, the ACM also replaced the previous primary function of monitoring member behavior with an emphasis on educating about the principles of ethical behavior in computing and providing guidance in ethical decision making.^c

The 1992 Code was reviewed by ACM membership and received a consensus and a commitment of its members to the ethical principles embodied in it. Sometimes these commitments are expressed as rules and sometimes as ideals, but the essential function was to clarify and formally state the professional’s moral responsibility toward society.

The 1992 Code organized ethical principles into four categories: general moral imperatives, more specific professional responsibilities, organizational leadership imperatives, and compliance. The principles were not canonical pronouncements requiring the use of particular technological solutions. Instead, these ethical markers of professionalism were presented as goals and ideals to which the morally responsible professional practitioner should aspire. The principles were accompanied by guidelines and illustrations showing their application to a developing computing profession.^d

The 1992 Code has been robust and useful in guiding decision-making. Over the years, the Code was used as a guide to instruct students entering the profession, as a decision support tool for computing practitioners, as a standard for the public to judge the professionalism of practitioners,

and as an aid to address legal issues and ethical tensions.^e The ACM Committee on Professional Ethics (COPE) often receives questions about applying the Code; in the last few years, many of those questions were related to artificial intelligence, machine learning, and robotics.

In the 25 years since the drafting of the 1992 Code began, there have been two significant, interconnected, and broad kinds of changes: 1) amazing changes in computing technology, and 2) important changes in how deeply that technology is integrated into social structures and into people’s daily lives. The technical changes are substantial. The number of people impacted and the intensity of that impact have been astonishing.

Twenty-five years ago, a “smart car” had an automatic transmission and, perhaps, antilock brakes. Sending selfies and tweets from your mobile phone were science fiction. The Web was in its infancy.

In 1992 the number of people using and controlling computers seemed limited. Computers were typically in a fixed location, and were just beginning to connect via the Internet. Computers were used to print bills, to control some highly specified processes, and to guide military devices. They managed and recorded financial information, controlled some processes on our automobiles, and controlled microwaves in the air and in our kitchens. It made sense for most scholars in computing to have a narrow focus on the analysis of algorithms and a study of

^e Markoff, John. “Apple’s Engineers, If Defiant, Would Be in Sync with Ethics Code.” *The New York Times Blog*; <http://www.nytimes.com/2016/03/19/technology/apples-engineers-if-defiant-would-be-in-sync-with-ethics-code.html>. Accessed 2016-10-04

Mullin, Joe. Google Puts Its Expert on the Stand to Combat Oracle, Wraps up Its Case. *Ars Technica*; <http://arstechnica.com/tech-policy/2016/05/google-puts-its-expert-on-the-stand-to-combat-oracle-wraps-up-its-case/>. Accessed 2016-10-04

^a ACM Web banner <http://www.acm.org/>. Accessed 2016-10-04

^b Bylaw 15 the ACM Code of Ethics and Professional Conduct <https://www.acm.org/governance/acm-bylaws#bylaw15>

^c Using the New ACM Code of Ethics in Decision Making. Ronald E. Anderson, Deborah G. Johnson, Donald Gotterbarn, and Judith Perrolle. *Comm. ACM* 36, 2 (1993), 98–107; doi:10.1145/151220.151231.

^d Ronald E Anderson. The ACM code of ethics: History, process, and implications. *Social Issues in Computing*. McGraw Hill New York, NY, 1994, 48–71.

the resources needed to execute them.

Now computing is ubiquitous—controlling our transportation and communication, and facilitating many human interactions. Computing today is in our bodies—prosthetics, pace-makers, and insulin pumps. Computing is also integral to the ways in which societies wage war. Computers impact all areas of our lives and many life-preserving functions are relegated to a piece of computer guided machinery.

Many of the newest impacts of computing are invisible. Computers make decisions about who is audited, who gets a heart transplant, and who gets targeted by dangerous devices, be they cars or missiles. The changes in technology and the kinds and number of impacted stakeholders are changing society in fundamental ways.

Social and technical changes connected to computing are interdependent, and they are changing the way computing is accomplished. Given the ease of communications, there is a rise of global software development. Many software developers have less individual control over what a software project will do or how it will operate. The easy development and dissemination of large aggregated datasets is transformational. We are also addressing concepts like accountable algorithms, inconceivable to many of us in 1992.^f

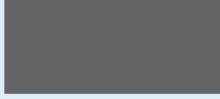
Although the 1992 Code was designed to be flexible, it has not completely kept pace with these changes. ACM's leadership understands that, and is committed to addressing the interaction of the technical and social worlds in the computing profession. The ACM Council charged COPE to develop an update to the Code, so that the members' social contract with the ACM better reflects these changes in computing and society. This update project is called "Code 2018."

Important principles guiding Code 2018 are:

- The Code should continue to document the ethical and professional responsibilities and obligations of computing professionals.



The ACM, like software development in general, is more global than it was during the last century; the updated Code needs to reflect broad, intercultural professional ethics.



- The Code should express the consensus of the computing profession on ethical issues.

- The Code should be used as a guide to decision making.

- The Code should educate both the public and aspiring professionals about the professional obligation of all computing professionals.

How will the Code be updated?

The ACM, like software development in general, is more global than it was during last century; the updated Code needs to reflect broad, intercultural professional ethics. COPE is establishing an international Task Force^g of experts on codes of ethics, current ethical issues of computers in society, public policy, law, and philosophy.^h

The Code 2018 Task Force will do several rounds of draft—review - comment - rewrite—before recommending a draft to the ACM Council for approval and adoption. Each draft will have a wide distribution; the process is designed to include many voices. The goal is to develop a Code by 2018 that will be useful for years to come.

COPE anticipates publishing two updates to this draft version for review and comment (at six-month intervals) before recommending a final version to ACM Council.

Codes, if carefully written and properly promoted, can be powerful instruments when striving for professionalism, and when establishing safeguards for society. But a Code that sits in the unread appendix of a book, or languishes on an unvisited website, has little influence. In addition to the Code 2018 effort, ACM has begun the "Integrity Project," an effort to promote ethical behavior in the computing professions.

New technologies arise so quickly that they may be in widespread use before practitioners can see the social and ethical consequences. In order to help ACM members in particular and the computing community at large to be mindful of professional computing ethics, and how its principles can be applied, the Integrity Project will develop a variety of multimedia materials for use

^f Nicholas Diakopoulos. Accountability in Algorithmic Decision Making. *Commun. ACM* 59, 2 (Feb 2016), 56–62; 10.1145/2844110 <http://cacm.acm.org/magazines/2016/2/197421-accountability-in-algorithmic-decision-making/abstract>. Accessed 2016-10-04

^g For a list of current taskforce members see <http://ethics.acm.org/code-2018>

^h Michael Davis. Eighteen rules for writing a code of professional ethics. *Science and Engineering Ethics* 13, 2 (2007), 171–189.

in schools and industry, including webinars, YouTube videos, and podcasts. The Integrity Project will complement and support the launch of the updated 2018 Code.

Description of the preliminary changes to the Code in the draft 1 version

A preliminary update of the Code (Draft 1) was developed by the Code 2018 Task Force. The initially suggested changes address the increased impact of computing on the infrastructure of society, and the increasing moral responsibility of ACM members. These changes reflect the need for members to have a better understanding of how computing technologies and artifacts impact the social infrastructure. There is a strengthened emphasis on the need for a member's activities to be focused on the common good. Professionalism in computing requires honing one's abilities to anticipate these broader impacts, for accepting responsibility for the consequences of these impacts, and for skillful analysis of trade-offs.

The purpose of this distribution of Draft 1 is to solicit comments from ACM members in particular, and from non-ACM members who are particularly interested. Based on the feedback about Draft 1, the Task Force will develop a revised draft of the code (Draft 2) that will again be presented for review.

Two clauses were added to Draft 1 of the Code. One clause provides a standard to help resolve ethical tension that may exist among Code principles in a particular situation: **"3.4 Ensure that the public good is a central concern during all professional computing work."** This type of clause in an ethics code is often called a "paramountcy clause." It helps a decision maker determine what has priority in her decision. Because it includes this clause, the Code emphasizes the professional's obligations to the public at large. This obligation should be the final arbiter in all professional decisions.

The other new clause reflects the added moral responsibility of computing professionals as computing is further integrated into society: **"3.7 Recognize when computer systems are becoming integrated into the infrastructure of society, and adopt an appropriate standard of care for those systems."**

Recognizing that the resolution of

ethical tensions is not always clear, the update to the Code includes a statement of the importance of professional ethical judgment and of honing ethical analysis skills. The Code does not provide a checklist of what is ethical in all situations, but directs a computer professional to use ethical judgment, acting in a manner that is most consistent with the spirit of the code of ethics in a given set of circumstances.

The suggested modifications to principle 1.4 (**"Be fair and take action not to discriminate unfairly."**) grew from a recognition that *discrimination* can sometimes be used as a synonym for *judgment*, but that the Code intends the term in the more colloquial usage of unjust or *unfair* judgments; there is also an expanded, alphabetical, partial list of unfair bases of discrimination. The intent of this clause is to be open ended, thus enabling it to include consideration of new social concerns in the future.

Imperatives 1.5 and 1.6 in the 1992 code are both focused on intellectual property. Copyright, patent, trade secrets, and licensing are all listed, but there was no mention of Free or Open Source Software. We have suggested combining these two principle into **"1.5 Honor intellectual property rights and give proper credit,"** and having the guidelines address current thinking about this topic.

THE CODE

In order to make your voice heard about the update of the code, please examine Draft 1 (below) and contact the Code 2018 Task Force using the contact information at the bottom of this article.

ACM Code of Ethics and Professional Conduct: Draft 1

Draft 1 was developed by The Code 2018 Task Force. (It is based on the 1992 ACM Code of Ethics and Professional Conduct <http://ethics.acm.org/code-of-ethics/previous-versions/1992-acm-code/>).

Suggested additions to the Code are underlined.ⁱ

Preamble

Commitment to ethical conduct is

ⁱ A complete track changes version of Draft 1 showing all additions and deletions to the 1992 Code is available at <http://ethics.acm.org/code-2018>.

expected of every ACM member. The ACM Code of Ethics and Professional Conduct ("the Code") identifies the elements of such a commitment.

This Code includes 24 imperatives formulated as statements of responsibility. The Code is designed to apply to practicing and aspiring computing professionals. Section 1 outlines fundamental ethical considerations. Section 2 addresses additional, more specific considerations of professional conduct. Section 3 pertains more specifically to individuals who have a leadership role, whether in the workplace or in a volunteer professional capacity. Principles involving compliance with this Code are given in Section 4.

Each imperative is supplemented by guidelines, which provide explanations to assist members in understanding and applying the imperative.

The Code is intended to serve as a basis for ethical decision making in the conduct of professional work. Secondly, it may serve as a basis for judging the merit of a formal complaint pertaining to a violation of professional ethical standards.

The Code as a whole is concerned with how fundamental ethical imperatives apply to one's conduct as a computing professional. The imperatives are expressed in a general form to emphasize that ethical principles which apply to computing professionals are derived from broadly accepted ethical principles.

The Code is not an algorithm for solving ethical dilemmas. Words and phrases in a code of ethics are subject to varying interpretations, and a particular imperative may conflict with other imperatives in specific situations. Questions related to these kinds of conflicts can best be answered by thoughtful consideration of the imperatives and fundamental ethical principles, understanding that the public good is a primary consideration.

1. GENERAL MORAL IMPERATIVES.

As an ACM member I will....

1.1 Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing and its artifacts.

This principle concerning the quality of life of all people affirms an obli-

gation to protect fundamental human rights and to respect the diversity of all cultures. An essential aim of computing professionals is to minimize negative consequences of computing systems, including threats to health, safety, personal security, and privacy. When designing or implementing systems, computing professionals must attempt to ensure that the products of their efforts will be used in socially responsible ways, will meet social needs, and be broadly accessible.

In addition to a safe social environment, human well-being requires a safe natural environment. Therefore, ACM members who design and develop systems must be alert to, and make others aware of, any potential negative impact to the local or global environment.

1.2 Avoid harm to others.

“Harm” means injury or negative consequences, such as undesirable loss of information, loss of property, property damage, or unwanted environmental impacts. This principle prohibits using computing in ways that result in harm to users, the general public, employees, employers, and any other stakeholders. Harmful actions include intentional destruction or modification of files and programs leading to serious loss of resources, or unnecessary expenditure of human resources such as the time and effort required to locate malicious software, purge it from systems, and mitigate its effects.

Well-intended actions, including those that accomplish assigned duties, may lead to harm unexpectedly. In such an event, those responsible are obligated to undo or mitigate the negative consequences as much as possible. Avoiding unintentional harm begins with careful consideration of potential impacts on all those affected by decisions made during design, implementation, use, and removal.

To minimize the possibility of indirectly harming others, computing professionals must minimize errors by following generally accepted best practices for system design, development, and testing. Furthermore, harm can be reduced by assessing the social consequences of systems. If system features are misrepresented to users, coworkers, or supervisors, the individual comput-

ing professional is accountable for any resulting harm.

In the work environment, an ACM member has an additional obligation to report any signs of system risks that might result in serious personal or social harm. If one’s superiors do not act to curtail or mitigate such risks, it may be necessary to “blow the whistle” to help correct the problem or to reduce the risk. However, capricious or misguided reporting of risks can itself be harmful. Before reporting risks, all relevant aspects of the incident must be thoroughly assessed as outlined in imperative 2.5.

1.3 Be honest and trustworthy.

Honesty is an essential component of trust. An ACM member will be fair and not make deliberately false or misleading claims and will provide full disclosure of all pertinent system limitations and potential problems. Fabrication and falsification of data are similarly violations of the Code.

An ACM member has a duty to be honest about his or her own qualifications, and about any limitations in competence to complete a task. ACM members must be forthright about any circumstances that might lead to conflicts of interest or otherwise tend to undermine the independence of their judgment.

Membership in volunteer organizations such as ACM may at times place individuals in situations where their statements or actions could be interpreted as carrying the “weight” of a larger group of professionals. An ACM member will exercise care not to misrepresent ACM or positions and policies of ACM or of any ACM units.

1.4 Be fair and take action not to discriminate unfairly.

The values of equality, tolerance, respect for others, and the principles of equal justice govern this imperative. Unfair discrimination on the basis of age, color, disability, family status, gender identity, military status, national origin, race/ethnicity, religion, sex, sexual orientation, or any other such factor is an explicit violation of ACM policy.

Inequities between different groups of people may result from the use or misuse of information and technology. In a fair society, all individuals have

equal opportunity to participate in, or benefit from, the use of computer resources. However, these ideals do not justify unauthorized use of computer resources, nor do they provide an adequate basis for violation of any other ethical imperatives of this code.

1.5 Honor intellectual property rights and give proper credit.

ACM members are obligated to protect the integrity of intellectual property, unless there is an overriding ethical reason not to do so. Examples of types of violations include (but are not limited to) misrepresentation of authorship, misrepresentation of the origin or ownership of ideas or work, misappropriation of a commons, unauthorized use, unauthorized copying, unauthorized derivative works, and counterfeiting. In normal circumstances, violations of intellectual property laws pertaining to copyrights, patents, trade secrets, non-disclosure agreements, and license agreements are contrary to the Code. Even when not explicitly barred by law, such violations are contrary to the Code.

Fair uses of intellectual property are necessary for the progress of technology in the service of the public good. ACM members should not oppose appropriate fair uses of their intellectual property.

Efforts to help others by contributing time and energy to projects that help society illustrate a positive aspect of this imperative. This includes contributions to projects that are in the public domain, free software, or open source software.

1.6 Respect privacy.

Technology enables the collection and exchange of personal information quickly, inexpensively, and often without the knowledge of the people affected. ACM members should use this personal data for legitimate ends without violating the privacy rights of individuals and organizations. ACM members should therefore implement security measures to maintain the privacy and integrity of personal data. This includes taking precautions to ensure the accuracy of data, as well as protecting it from unauthorized access or accidental disclosure to inappropriate individuals. Computing professionals should establish procedures to allow individuals to

review their personal data and correct inaccuracies.

Only the minimum amount of personal information necessary should be collected in a system. The retention and disposal periods for that information should be clearly defined and enforced, and personal information gathered for a specific purpose should not be used for other purposes without consent of the individual(s).

When data collections are merged, ACM members should take special care for privacy. Individuals may be readily identifiable when several data collections are merged, even though those individuals are not identifiable in any one of those collections in isolation.

1.7 Honor confidentiality.

The ethical obligation for confidentiality holds unless discharged from such obligations by bona fide requirements of law or by other principles of this Code.

User data observed during the normal duties of system operation and maintenance must be treated with strict confidentiality, except in cases where it is evidence for the violation of law, organizational regulations, or this Code. In these cases, the nature or contents of that information must be disclosed only to appropriate authorities.

2. MORE SPECIFIC PROFESSIONAL RESPONSIBILITIES.

As an ACM member with professional responsibilities I will....

2.1 Strive to achieve the highest quality in both the process and products of professional work.

Computing professionals should insist on high quality work from themselves and from colleagues. Professionals must be cognizant of the serious negative consequences that may result from poor quality. High quality includes respecting the dignity of employers, colleagues, clients, users, and anyone effected either directly or indirectly by the work.

2.2 Maintain high standards of professional competence, conduct, and ethical practice.

High-quality computing depends on individuals who take personal and organizational responsibility for acquiring and maintaining professional competence.



Codes, if carefully written and properly promoted, can be powerful instruments when striving for professionalism and when establishing safeguards for society.



Professional competence includes technical knowledge, awareness of the social context in which the work will be deployed, and competence in recognizing and navigating ethical challenges. Upgrading necessary skills should be ongoing and should include independent study, seminars, conferences, and other informal or formal education. The ACM is committed to encouraging and facilitating those activities.

2.3 Know, respect, and apply existing laws pertaining to professional work.

ACM members must obey existing local, state, province, national, and international laws unless there is a compelling ethical justification not to do so. Policies and procedures of the organizations in which one participates must also be obeyed, but compliance must be balanced with the recognition that sometimes existing laws and rules are immoral or inappropriate and, therefore, must be challenged. Violation of a law or regulation may be ethical when that law or rule has inadequate moral basis or when it conflicts with another law judged to be more important. If one decides to violate a law or rule because it is viewed as unethical, or for any other reason, one must fully accept responsibility for one's actions and for the consequences.

2.4 Accept and provide appropriate professional review.

Quality professional work in computing depends on professional reviewing and critiquing. Whenever appropriate, individual members should seek and utilize peer review, and should provide constructive, critical review of the work of others.

2.5 Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks.

ACM members must strive to be perceptive, thorough, and objective when evaluating, recommending, and presenting system descriptions and alternatives. Computing professionals are in a position of special trust, and therefore have a special responsibility to provide objective, credible evaluations to employers, clients, users, and the public. When providing evaluations the professional must also identify any relevant conflicts

of interest, as stated in imperative 1.3.

As noted in the discussion of imperative 1.2 on avoiding harm, any signs of danger from systems must be reported to those who have opportunity and/or responsibility to resolve them. See the guidelines for imperative 1.2 for more details concerning harm, including the reporting of professional violations.

2.6 Accept only those responsibilities for which you are qualified, and honor those commitments.

A computing professional has a responsibility to evaluate every work assignment. Should the evaluation identify reasons that the project should not be attempted, the professional must disclose those reasons to the employer or client. The assignment should not be accepted unless those reasons are mitigated by changes to the nature of the project.

Should the evaluation identify reasons that the professional does not have the expertise to complete the project, the professional must disclose this shortcoming to the employer or client, and request that the project be undertaken by someone with the appropriate qualifications.

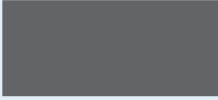
Should the evaluation identify that the project is theoretically impossible to complete by anyone, the professional must disclose this impossibility to the employer or client and request that the project be dropped or modified in order to make the project theoretically possible.

On some occasions, other ethical principles may take greater priority, and a judgment that a specific assignment should not be performed may not be accepted. Only after serious consideration and with full disclosure of risks and concerns to the employer or client, and having clearly identified one's concerns and reasons for that judgment that failed to result in a change to the nature of the project, should one accept the assignment if one is obligated, by contract or by law. The major underlying principle here is the obligation to accept personal accountability for professional work. The computing professional's ethical judgment should be the final guide in deciding whether to proceed. Regardless of the decision, one must accept the responsibility for the consequences.

Computing professionals should en-



New technologies arise so quickly that they may be in widespread use before practitioners can see the social and ethical consequences.



sure that system elements perform as intended. When an ACM member contracts for work with another party, the member has an obligation to keep that party properly informed about progress toward completing that work.

2.7 Improve public understanding of computing, related technologies, and their consequences.

Computing professionals have a responsibility to share technical knowledge with the public by creating awareness and encouraging understanding of computing, including the impacts of computer systems, their limitations, their vulnerabilities, and opportunities they present. This imperative implies an obligation to counter any false views related to computing.

2.8 Access computing and communication resources only when authorized to do so.

Theft or unauthorized destruction of tangible and electronic property is prohibited by imperative 1.2—“Avoid harm to others.” Trespassing and unauthorized use of a computer or communication system is addressed by this imperative. Trespassing includes accessing communication networks and computer systems, or accounts and/or files within those systems, without authorization to do so. Individuals and organizations have the right to restrict access to their systems so long as they do not violate the discrimination principle (see imperative 1.4). No one should access or use another's computer system, software, or data files without permission. One should have appropriate approval before using system resources unless there is an overriding concern for the public good. To support this clause, a computing professional should take appropriate action to secure resources against unauthorized use.

3. ORGANIZATIONAL LEADERSHIP IMPERATIVES.

In this section, “leader” means any member of an organization who has leadership or educational responsibilities. These imperatives generally apply to organizations as well as their leaders. “Organizations” are corporations, government agencies, and other “employers,” as well as volunteer professional organizations.

As an ACM member and an organizational leader, I will....

3.1 Articulate social responsibilities of members of an organizational unit and encourage full acceptance and satisfaction of those responsibilities.

Because organizations have impacts on the public, they must accept responsibilities to society. Organizational procedures and attitudes oriented toward quality, transparency, and toward the welfare of society will reduce harm to members of the public. This serves the public interest and fulfills social responsibility. Therefore, organizational leaders must encourage full participation in meeting social responsibilities and quality performance.

3.2 Manage personnel and resources to design and build systems that enhance the quality of working life.

Organizational leaders are responsible for ensuring that (computer) systems enhance, not degrade, the quality of working life. When implementing a system, organizations must consider the personal and professional development, physical safety, psychological well-being, and human dignity of all workers. Appropriate human-computer ergonomic standards should be considered in system design and in the workplace.

3.3 Establish appropriate rules for authorized uses of an organization's computing and communication resources and of the information they contain.

Organizational leadership has the responsibility to clearly define appropriate and inappropriate uses of organizational computing resources. These rules must be clearly and effectively communicated to those using their computing resources. In addition, the organization must enforce those rules, and take appropriate action when they are violated.

3.4 Ensure that the public good is a central concern during all professional computing work.

The needs of people—including users, other people affected directly and indirectly, customers, and colleagues—should always be a central concern in professional computing. Tasks associated with requirements, design,

development, testing, validation, deployment, maintenance, and disposal should have the public good as an explicit criterion for quality. Computing professionals should keep this focus no matter which methodologies or techniques they use in their practice.

3.5 Articulate, apply, and support policies that protect the dignity of users and others affected by computing systems and related technologies.

Dignity is the principle that all humans are due respect. This includes the general public's right to autonomy in day-to-day decisions.

Designing or implementing systems that deliberately or inadvertently violate, or tend to enable the violation of, the dignity or autonomy of individuals or groups is ethically unacceptable. Computing professionals who are in decision making positions should verify that systems are designed and implemented to protect personal dignity.

3.6 Create opportunities for members of the organization to learn, respect, and be accountable for the principles, limitations, and impacts of computer systems.

Imperative 3.6 complements the imperative on public understanding (imperative 2.7). Educational opportunities are essential to facilitate optimal participation of all organizational members. Opportunities must be available to all computing professionals to help them improve their knowledge and skills in professionalism, the practice of ethics, and computing, including experiences that familiarize them with the consequences and limitations of particular types of systems. Professionals must know the dangers of building systems around oversimplified models, the improbability of anticipating and designing for every possible operating condition, the inevitability of software errors, the ways in which systems impact and are impacted by the contexts in which they are deployed, and other issues related to the complexity of their profession.

3.7 Recognize when computer systems are becoming integrated into the infrastructure of society, and adopt an appropriate standard of care for those systems.

Computing professionals who develop computer systems that have or may become an important part of the infrastructure of society have a responsibility to be good stewards of that commons. Part of that stewardship requires that computing professionals monitor the level of integration into the infrastructure of society. As the level of adoption changes, there are likely to be changes in the ethical responsibilities of the organization. Continual monitoring of how society is using its computer system will allow the organization to remain consistent with their ethical obligation. Where such standards of care do not exist, there may be a duty to develop one.

4. COMPLIANCE WITH THE CODE.

As an ACM member I will....

4.1 Uphold, promote, and respect the principles of this Code.

The future of computing depends on both technical and ethical excellence. ACM members should adhere to the principles expressed in this Code. Each member should encourage and support adherence by all computing practitioners.

4.2 Treat violations of this code as inconsistent with membership in the ACM.

If an ACM member does not follow this code membership in ACM may be terminated.

5. COPE INVITES COMMENTS

The Committee on Professional Ethics is asking you to participate in this project. We invite you to share your views with the Code 2018 Task Force. We are soliciting suggestions about areas of computing that have changed and need to be addressed in an updated Code. For more details, and to provide input, visit <http://ethics.acm.org/code-2018>.

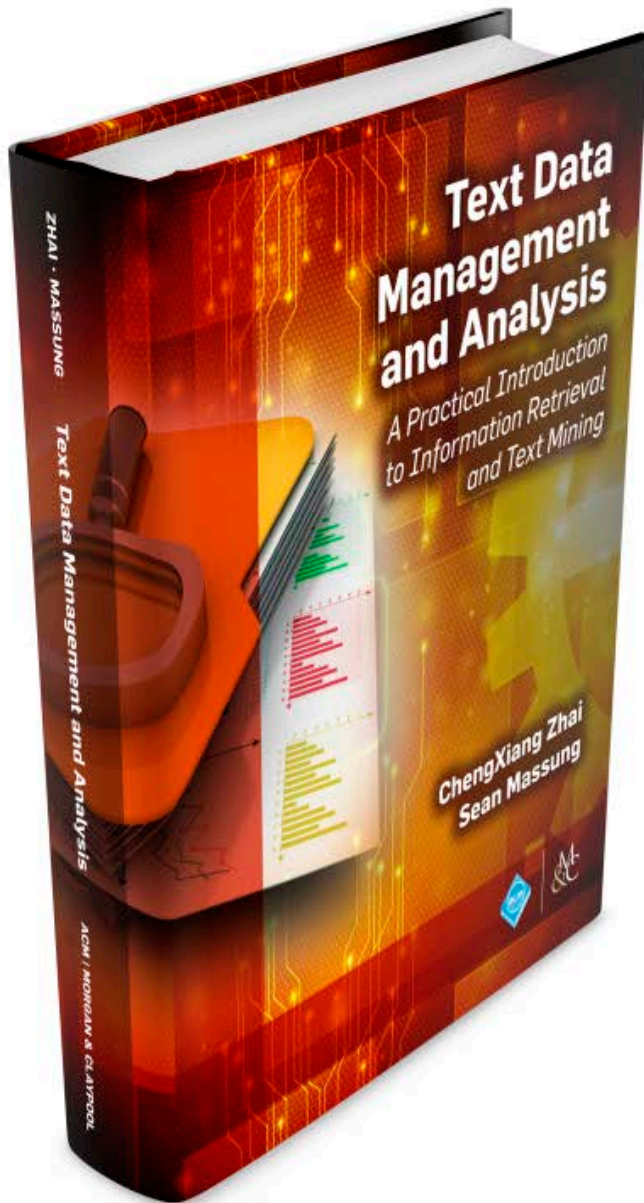
Bo Brinkman (bo.brinkman@miamioh.edu) is an associate professor of computer science and software engineering at Miami University, Oxford, OH.

Don Gotterbarn (gotterbarn@acm.org) is chair of the ACM Committee on Professional Ethics and Professor Emeritus in the Department of Computing at East Tennessee State University, Johnson City.

Keith Miller (millerkei@umsledu) is the Orthwein Endowed Professor for Lifelong Learning in the Sciences College of Education, University of Missouri, St. Louis.

Marty J. Wolf (mjwolf@acm.org) is a professor of computer science at Bemidji State University, Bemidji, MN.

© 2016 ACM 0001-0782/16/12 \$15.00



The most useful and practical knowledge for building a variety of text data applications.

CS STUDENTS (Undergrad & Graduate)
LIBRARY & INFORMATION SCIENTISTS
TEXT DATA PRACTITIONERS

ChengXiang Zhai & Sean Massung (Authors)
University of Illinois at Urbana-Champaign

Text Data Management and Analysis covers the major concepts, techniques, and ideas in **information retrieval** and **text data mining**. It focuses on the practical viewpoint and **includes many hands-on exercises designed with a companion software toolkit** (i.e., MeTA) to help readers learn how to apply techniques of information retrieval and text mining to real-world text data. It also shows readers how to **experiment** with and **improve** some of the algorithms for interesting application tasks. The book can be used as a **text** for computer science undergraduates and graduates, library and information scientists, or as a **reference** for practitioners working on relevant problems in **managing and analyzing text data**.



<http://books.acm.org>
<http://www.morganclaypoolpublishers.com/text>



DOI:10.1145/3014064

Vinton G. Cerf

When Email Isn't Private

The title of this column might just have well be “Is email ever private?” In the first place, email is usually sent to someone other than yourself, so its privacy is dependent on the path it took

from your desktop, laptop, tablet, or mobile to your email service provider, the service provider's security, the path to the destination email provider, and the path taken to reach the intended recipient. There are more paths for exposure when there are multiple recipients served by yet additional email providers. Some systems, such as Google's Gmail, try to keep messages encrypted while in transit from the user's originating device to the Google cloud server(s) and encrypted while stored in the Google cloud. But the path to the destination mail relay may not be protected unless the source and destination relay agents have an agreed cryptographic protocol for transport. Moreover, the recipient of the email may not have or use an encrypted channel to pick up the email from the destination relay agent. This means that the conscientious sender of email has little or no control over the practices of the recipient.

The originator and the recipient of an email message may use weak methods to secure access to their email services. The email service provider may only support simple user name and password access control. Many of the break-ins into email services are a consequence of password guessing, dictionary attacks against one-way hashed password lists, “lost my password” processing in which the answers to secret questions may be found with some searching of the World Wide Web, and exposure of passwords in unencrypted password files or on Post-Its™ attached to laptops.

Efforts to allow users to encrypt their email on an end-to-end basis have generally not fared well in part owing to the awkwardness of maintaining lists of cryptovars for destination users, registering cryptovars (or certificates) in the first place, dealing with lost or compromised certificates, and dealing with multiparty recipients of encrypted email. One of the most popular technologies, PGP,^a has been in use for many years, but has seen deployment largely in the technical community.

By the time you read this column, the U.S. presidential election will have taken place (and, hopefully been decided). What is clear, however, is that email exposure has been in the recent news on a regular basis, ranging from Hillary Clinton's email messages while in service as Secretary of State, or the email of the Democratic National Committee or of high-ranking Russian officials and their correspondents, to say nothing of intelligence agency demands for access. To the extent that we care, collectively, to secure improved privacy in our communications by electronic means, it seems essential to adopt and deploy a variety of methods to protect access to our online email accounts and to protect their contents when in transit or at rest.

As awkward as two-factor authentication may be, it is still among the most effective mechanisms for protecting access to online accounts. Over time, biometrics may prove to be a useful substitute, but I do worry that the digi-

tized representation of a fingerprint or an eye scan may be captured or fabricated in such a way as to be injected into the authentication system at the right point to penetrate an account.

It seems clear that we need standards and agreements to protect email while in transit between mail transfer agents and between users and transfer agents. Encryption at rest and end-to-end encryption are also very useful practices. But there are other vulnerabilities such as phishing attacks at the application level that cause users to ingest malware. Source email identifiers (for example, “From” field) are not reliable and can be easily spoofed. Among the popular mechanisms for validating email sources is Domain Keys Identified Mail (DKIM)^b that allows a source or intermediary relay to validate its identity in the sequence from origin to destination email server. Composing these various mechanisms improves email security but they have to be widely implemented and used by all users and providers of the service. Anything less leaves traps in the system into which users (or their email at any rate) may fall.

Email is such a useful service, it is difficult to imagine abandoning it although there is a growing attachment to other mechanisms including mobile texting and messaging built into social network services. Even though some of them promise that messages evaporate after a time, reports have been made that they can be captured, saved, and re-injected into the Internet via other communication channels. Those of us who are in the computer science profession owe ourselves and the general public a much better experience and that's one of many challenges for this online 21st century. **C**

^b <http://www.dkim.org/>

Vinton G. Cerf is vice president and Chief Internet Evangelist at Google. He served as ACM president from 2012–2014.

Copyright held by author.

^a “Pretty Good Privacy,” <http://openpgp.org/>

Blog Ubiquity

INFORMATION EVERYWHERE



The newest ACM forum.

Contributions
that cover the vast
information-rich world
where computing
is embedded
everywhere.

ACM's *Ubiquity* is
the online magazine
oriented toward the
future of computing
and the people who
are creating it.

We invite you to
participate: leave
comments, vote for
your favorites, or
submit your own
contributions.

Captivating topics.

*Net Neutrality
and the Regulated
Internet*

*The End of Life
As We Know It*

*A Shortage of
Technicians*

*The Fractal
Software
Hypothesis*

*Your Grandfather's
Oldsmobile—NOT!*

*Superscalar
Smart Cities*



Association for
Computing Machinery

Visit us at
<http://ubiquity.acm.org/blog/>

Reclaim the Lost Promise of the Semantic Web

I WAS EAGER to learn about the latest developments in the Semantic Web through the lens of a “new kind of semantics” as Abraham Bernstein et al. explored in their Viewpoint “A New Look at the Semantic Web” (Sept. 2016), but by the end I had the impression the entire vision of a Semantic Web was somehow at risk.

If I understand it correctly, semantics is a mapping function that leads from manifest expressions to elements in a given arbitrary domain. Based on set theory, logicians have developed a framework to set up such mapping for formal languages like mathematics, provided one can fix an interpretation function. On the other hand, 20th-century logicians (notably Alfred Tarski) warned of the limits of the framework when applied to human languages. Now, to the extent it embraces a set-theoretic semantics (as in the W3C’s Ontology Web Language), the Semantic Web seems to be facing exactly such limitations or experiencing, dealing with, and suffering them.

Most Web content is expressed as natural language, and it is not easy for programmers to bring it into clean logical form; meanwhile, Percy Liang’s article “Learning Executable Semantic Parsers for Natural Language Understanding” (also Sept. 2016) gave an idea of the early stage of “semantic parsing,” or the task of obtaining a formal representation of the meaning of a given text. It seems the “new semantics” in Bernstein et al., albeit not formally characterized, was an attempt to outline a better approach to tapping the linguistic nature of the Web, which is indeed remarkable.

In taking a language-oriented view, however, Bernstein et al. seemed to neglect a key feature of formal semantics—transparency. They seem comfortable with the relaxation of logic as a conceptual framework for the Semantic Web, which is typical of modern Knowledge Graphs (such as the one Google uses). But one of the consequences of such relaxation is that part of data semantics ends up being embedded in

algorithms. Not only practitioners but also common users are aware that algorithms that work on Web data are embedded in only a few monolithic, private platforms that are far from open, transparent, and auditable.

Isn’t keeping meanings in a handful of proprietary algorithms exactly the opposite of what the Semantic Web was meant to be?

Guido Vetere, Rome, Italy

Authors Respond:

As we mentioned in the Viewpoint, the Semantic Web is not just about texts but also about myriad data, images, video, and other Web resources. While a formal logic that could be both transparent enough for all such resources and yet usable by Web developers is a noble ambition, current logics are simply not up to the task. The transparency of “some semantics” is the best to hope for and would allow all potential developers to build Web-scale, best-effort applications.

Abraham Bernstein,

Zürich, Switzerland,

James Hendler, Troy, NY, and

Natalya Noy, Mountain View, CA

More Than a Red Flag for AIs

Toby Walsh’s Viewpoint “Turing’s Red Flag” (July 2016) proposed a legislative remedy for various potential threats posed by software robots that humans might mistake for fellow humans. Such an approach seems doomed to fail. First, unless a “red flag” law would be adopted by all countries, we humans would have the same problems identifying and holding accountable violators we have with cybercrime generally. Second, though Walsh acknowledged it would take a team of experts to devise an effective law, it would likely be impossible to devise one that would address all possible interactions with non-humans or not lead to patently silly regulations, as with the original 19th-century Red Flag Act. How would a law handle algorithm-based securities trading? How about if one human is dealing with another human, but

that human has an AI whispering in his or her ear (or implanted in his or her brain) telling them what to say or do?

More important, the most significant potential harms from bots or sophisticated AIs generally would not be mitigated by just knowing when we are dealing with an AI. The harm Walsh proposed to address seemed more aimed at the “creep factor” of mistaking AIs for humans. We have been learning to deal with that since we first encountered a voicemail tree or political robocall. Apart from suffering less emotional shock, what advantage might we gain from knowing we are not dealing with a fellow human?

Learning to live with AIs will involve plenty of consequential challenges. Will they wipe us out? Should an AI that behaves exactly like a human—emotional responses and all—have the legal rights of a human? If AIs can do all the work humans do, but better, how could we change the economic system to provide some of the benefits of abundance made possible by AI-based automation to the 99% whose jobs might be eliminated? Moreover, what will we humans do with our time? How will we even justify our existence to ourselves? These sci-fi questions are quickly becoming real-life questions, requiring we have more than a red flag to address them.

Martin Smith, McLean, VA

Author Responds:

This critique introduces many wider and orthogonal issues like existential risk and technological unemployment. Yes, it will be difficult to devise a law to cover every situation. But that is true of most laws and does not mean we should have no law. However, actions speak loudest, and the New South Wales parliament has just recommended such a law in Australia; for more, see <http://tinyurl.com/redflaglaw>.

Toby Walsh, Berlin, Germany

Communications welcomes your opinion. To submit a Letter to the Editor, please limit yourself to 500 words or less, and send to letters@cacm.acm.org.

© 2016 ACM 0001-0782/16/12 \$15.00

The *Communications* Web site, <http://cacm.acm.org>, features more than a dozen bloggers in the BLOG@CACM community. In each issue of *Communications*, we'll publish selected posts or excerpts.



Follow us on Twitter at <http://twitter.com/blogCACM>

DOI:10.1145/3005674

<http://cacm.acm.org/blogs/blog-cacm>

ICER 2016, and Star Trek at 50

Mark Guzdial reports on promising papers, and Daniel Reed recalls a television show that continues to inspire innovation.



Mark Guzdial
14 Years of a
Learner-Centered
Python IDE

<http://bit.ly/2aNXAnC>
August 10, 2016

Early September saw the 2016 International Computing Education Research (ICER) conference in Melbourne, Australia. The conference was terrific, as is usual for ICER.

Two of its papers were meta-papers studying the ICER community itself. They found the community has healthy levels of newcomers and collaboration, and features methodological rigor and strong theoretical foundations.

I will report on three papers: two award recipients, and one of my favorite papers at the conference. ICER has two paper awards: a “people’s choice” (voted on by attendees) *John Henry* award for innovation and new directions, and a “Chairs” award selected by the conference chairs based on the paper reviews.

The people’s choice award was won by Elizabeth Patitsas (with Jesse Berlin, Michelle Craig, and Steve Easterbrook) from the University of Toronto for “Evidence that Computer Science Grades are not Bimodal” (<http://bit.ly/2dsBo3L>).

Elizabeth’s paper had two studies in it and a provocative discussion section.

Many CS teachers believe grades in CS classes are bimodal: some students have innate ability and just “get it”; others don’t. There are research papers presenting explanations for the bimodality effect, but is the effect really there?

In her first study, Patitsas did an analysis of 18 years’ worth of grade data from a large university CS department, and found less than 5% of the courses had signs of non-normality. Her second study was a “deception study” (debriefed at <http://bit.ly/2dKM48u>). She asked 60 CS teachers (mostly from the SIGCSE members list) to judge if a number of grade distributions were bimodal; the reality was that none of them were. She also asked teachers if they agreed with the statements “Some students are innately predisposed to do better at CS than others” and “Nearly everyone is capable of succeeding in computer science if they work at it.” Both of these statements strongly correlated with “seeing bimodality” in the distributions, the first positively and the second negatively. If teachers believed in a “Geek Gene” (that some students are innately gifted at program-

ming), they saw bimodality, even if it was not there.

The provocative explanation is that CS teachers see bimodality because they do not teach well. Patitsas used a social defense theory to explain “seeing-bimodality.” Patitsas suggests our overconfidence in CS teaching leads to seeing bimodality when there is none.

The Chairs award was particularly exciting because it was won by a team led from a School of Education. Computing education research has been dominated by CS researchers, and it’s terrific to see the Education side playing a more prominent role. The paper was “Learning to Program: Gender Differences and Interactive Effects of Students’ Motivation, Goals, and Self-Efficacy on Performance” (<http://bit.ly/2d3UKJA>) by Alex Lishinski, Aman Yadav, Jon Good, and Richard Enbody from Michigan State University. Self-efficacy is one’s own rating of their ability to succeed or perform in a particular discipline. We knew from prior work women tend to have low self-efficacy ratings at the start of CS classes, while men have high ratings. What hasn’t been studied previously was how these changed with feedback. As students get grades back on homework and exams, what changes? Lishinski showed women more quickly adapt self-efficacy ratings compared to men; the scores rise dramatically. It takes a long time (more feedback) to get men to downgrade their overestimated skills to match their performance.

One of my favorite papers at ICER 2016 was “Some Trouble with Transparency: An Analysis of Student Errors

with Object-oriented Python” (<http://bit.ly/2dKNqji>) by Craig S. Miller and Amber Settle. Anyone who writes object-oriented programs in Python knows methods in Python classes must explicitly state a parameter self. Miller and Settle call that “transparency.” References to the receiving object are available in Java (for example) methods, too, but not as an explicit parameter. Is that a problem? Settle presented evidence that it really is. In a study of object-oriented programming in Python (where students were asked to code a particular method for a given class), some errors (like returning too early from a method, or forgetting to loop through all items in a list) occurred relatively frequently—19% and 31% of all errors, respectively. The self-related errors were far more common; 53% of all errors were due to missing the self parameter in the method declaration, 39% were due to missing self in an object reference, and 39% used self incorrectly. That’s a cost of using Python for novice students not previously measured.

There were lots of other great papers I’m not going to discuss here. I recommend Andy Ko’s excellent ICER 2016 trip report (<http://bit.ly/2dSpWhd>) for another take on the conference.



Daniel A. Reed
“Star Trek”@50:
Inspiring Discovery
and Innovation
<http://bit.ly/2cFzx9X>
 September 8, 2016

On September 8, 1966, 50 years ago, “Star Trek” (<http://bit.ly/2dKQRY5>) premiered in the U.S. on the NBC television network. By the standards of the day, it was not a great success. The ratings were mediocre, the reviews were mixed, and it was canceled after three seasons despite a fan-driven letter-writing campaign. Defying this inauspicious beginning, “Star Trek” has become an international cultural phenomenon, with multiple series, movies, and casts since its television premiere.

Much has been written about why “Star Trek” has lived long and prospered. I suspect much of its enduring appeal lies in the personal relations of the three original starring characters (Kirk, Spock, and McCoy), along with technological optimism and the ethical questions and conundrums posed.

“Star Trek” has also entered the cultural lexicon in deep ways, “Beam me up, Scotty!” and “Set phasers to stun” being just two of many examples. The town of Riverside, IA, even has a commemorative monument to the “future birthplace” of James T. Kirk. More importantly, “Star Trek” has inspired generations to pursue science and technology careers, not a few of whom have transformed part of that television science fiction into technological and commercial fact.

Through a B&W Broadcast, Darkly

Alas, I missed the premiere and the original run of “Star Trek” broadcasts because my family did not own a television. Today that seems incredible, given television’s nearly universal market penetration, the plethora of cable channels, and ubiquitous streaming media services. At the time, there were only three broadcast television networks in the U.S., and cable service was not available in many rural areas. In the 1960s, channel surfing often involved climbing on the roof to rotate the antenna while responding to guidance through an open window. Instead, it was best to pick one of the broadcast networks and stick with it, particularly in winter.

When “Star Trek” entered syndication a few years later, I was able to watch it in B&W; when I saw it later in color, I was amazed by the bright, garish colors. Whether in gray scale or color, the series immediately engaged both my teenage angst and my scientific aspirations. Mr. Spock’s stoicism and logic comforted many who felt the pain and loneliness of cultural isolation, including this geek. The series also gave hope that a better world was possible, one where we could celebrate our differences as strengths while embracing the common core of our shared humanity. During the height of the Vietnam War and our ongoing struggle for civil rights, this was a powerful message of hope.

From Fiction to Reality and Back

Though one might also rightly argue convergent evolution (<http://bit.ly/2dfrqgM>), cellphones (communicators) and tablet computers both owe some elements of their form and function to “Star Trek”’s vision of ubiquitous computing and communications. The Tricorder X Prize competition to create

a portable, wireless health monitoring device, is a direct homage to “Star Trek.”

My former colleagues at Microsoft Research often reference the “Star Trek” universal translator as an inspiration for their work on real-time language translation. (Thank you, Rick Rashid.) More generally, our community’s work on deep learning and intelligent assistants is inspired not only by technical goals, but by a motivating vision of artificial intelligence (AI) that runs deeply through “Star Trek” and the science fiction world. From weak AI to ambitions of strong AI, we yearn to build a machine that will be proud of us.

Across computing, we ponder issues of AI ethics and their instantiation in autonomous vehicles; consider the limits of silicon-based computing and quantum alternatives, and debate the future of ubiquitous sensors and digital privacy. These and other technical challenges inform our imaginations and dreams of the future. This virtuous cycle of invention and imagination drives us forward.

The Frontier Awaits

The original “Star Trek” series opened with Captain Kirk intoning the Enterprise would “boldly go where no man has gone before.” It was a powerful message of exploratory optimism, though I have to admit the split infinitive has always bothered me. But the gender-specific pronoun was the wrong message, one Captain Picard rightly corrected:

Space, the final frontier. These are the voyages of the starship Enterprise. Its continuing mission: to explore strange new worlds, to seek out new life and new civilizations, to boldly go where no one has gone before.

Both versions echo the enduring words of Vannevar Bush in his seminal essay “Science: The Endless Frontier” (<http://bit.ly/2e2RMqi>), and speak to something deep in our nature:

...without scientific progress no amount of achievement in other directions can insure our health, prosperity, and security as a nation in the modern world.

We yearn to discover. Engage. Make it so. □

Mark Guzdial is a professor in the School of Interactive Computing at the Georgia Institute of Technology. **Daniel Reed** is vice president for Research and Economic Development and University Computational Science and Bioinformatics Chair at the University of Iowa.

© 2016 ACM 0001-0782/16/12 \$15.00

XRDS

Seeks Student Volunteers

Are you a student who enjoys staying up to date with the latest tech innovations, and is looking to make an impact on the ACM community?

Editorial positions are now open for the following positions:
feature editor and department editor.

XRDS is a quarterly print magazine for students by students that examines cutting edge research in computer science, viewpoints on technology's impact in the world today, and works to support a strong, positive, and inclusive community of students interested in computer science. Our goal is to make the magazine accessible to anyone with an interest in computer science and technology. XRDS focuses on interesting work being conducted at different universities, research centers and labs around the world. Our editors represent a team of students with diverse interests who are undergrads and graduate students from around the globe.

For more information and to apply visit:
<http://xrds.acm.org/volunteer.cfm>



**Association for
Computing Machinery**



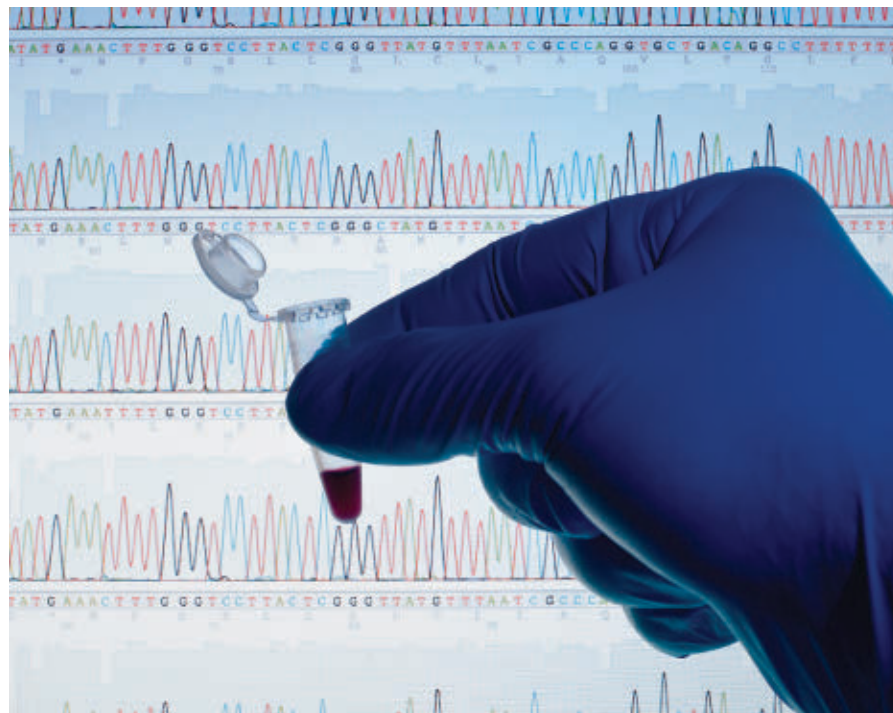
Cracking the Code on Biology

Genetic engineering promises to revolutionize everything from agriculture and biofuels to medicine and food safety.

AS SCIENTISTS LOOK TO unravel the secrets of life and address a diverse array of challenges involving medicine, agriculture, and more, they increasingly are tossing aside their test tubes and microscopes and diving directly into the genome.

Over the last decade, synthetic biology has emerged as an intriguing, if not remarkable, tool that could fundamentally change the world. The nascent concept, based on creating new sequences of DNA from scratch and rearranging existing ones, already is being used to mimic biological systems, create synthetic biological systems, and actually program biological systems.

It is no small matter. “In the latter half of the 20th century, scientists realized that DNA was the software of living systems and it’s possible to transfer code from one organism to another in order to program cells with new and different functions,” explains Timothy Lu, an associate professor in the Department of Electrical and Computer Science at the Massachusetts Institute of Technology (MIT). However, since the turn of the century, he says, “Researchers have incorporated more complex elements of computing and



memory. By assembling genes in a specific way and creating toggles and circuits, it’s possible to produce desired behaviors, and program specific functions into cells.”

Make no mistake, DNA synthesis and sequencing are advancing at a rapid rate. “These technologies are match-

ing or exceeding Moore’s Law,” Lu says. What is more, the ability to program genes will likely change things in profound ways. It could knock out afflictions such as Malaria, cancer, and Huntington’s disease, but also offer a way to steer clear of controversial genetically modified organisms (GMOs)

in plants and introduce designer animals and humans. Says Daniel Gibson, vice president of DNA Technologies at Synthetic Genomics, Inc. and associate professor at the J. Craig Venter Institute, a leading genome research center: “As with any new area of science and technology, synthetic genomics has the potential for great societal benefit, but also the potential for harm.”

Beyond the Test Tube

Since the beginning of time, humans have focused on altering the genetic code of plants and animals through selective breeding, hybridization, and other techniques. Modern computing and science have taken the concept to an entirely new level, allowing researchers to achieve results that were previously impossible, or to compress years, decades, and even centuries of gene editing research into weeks or months by rearranging genes and re-designing biological pathways.

Modern synthetic biology is rooted in research from the turn of the current century. At that time, researchers Jim Collins and Tim Gardner of Boston University found a way to create a “toggle switch” with *E. coli* bacteria; essentially, the engineered cells could hold and maintain a single bit of memory. Around the same time, researchers Michael Elowitz and Stan Leibler of Princeton University built a ring-oscillator that allowed a cell to operate like a mini-clock. “This showed that it was possible to program complex, artificial functions into cells,” Lu points out.

Unlike silicon computing, which relies on ones and zeros to build code into software, biological systems require circuits engineered from DNA, RNA, and proteins. “The DNA interacts with proteins and RNA inside the cell, and the collision between these different parts is what leads to our ability to compute information,” Lu explains. “One challenge with biological systems is that, unlike silicon where you can wire together components and ensure they are isolated from the rest of the world, it’s not easy to do this with biological systems. In addition, they aren’t modular by default; it becomes more of an analog programming task.”

To be sure, programming biology is inherently challenging. “Building circuits inside cells is not a highly scal-

“Ultimately, every ‘transistor’ in biology has to be unique, so this limits the size of things you can do.”

able approach. The amount of energy that a cell needs to drive different circuits is quite significant,” says Herbert Sauro, an associate professor of bioengineering at the University of Washington. “Unlike electrical systems, there’s a scalability issue and the fact that you need more and more independent proteins to build larger and larger circuits. Evolution did not build large computing systems inside a cell. It had to go beyond the cell and create a multi-cell system, such as the brain, to gain enough computing power. Ultimately, every ‘transistor’ in biology has to be unique, so this limits the size of things you can do.”

Yet, knowledge of synthetic biology is advancing at a rapid pace. Better heuristic methods, new design tools, more powerful computers, and increasing big data capabilities are transforming the field. One of the most important breakthroughs came with a gene-editing tool called CRISPR, which “finds the right place in the genome—out of 3 billion places—and cuts it,” says George Church, a professor of genetics at Harvard Medical School. Essentially, the technology allows researchers to expand their toolkit beyond merely adding genes; it lets them delete or change genes to suit their needs. Today, CRISPR is widely used for research on humans, animals, and plants. It is inexpensive, typically costing less than \$100 in a lab that is already set up for standard molecular biology, and many government agencies around the world have accepted the technology. “They are saying it doesn’t need to be regulated,” Church adds.

The combination of CRISPR technology, its low cost, and the ability to avoid government regulation has im-

portant ramifications. For example, in the agricultural space, CRISPR is “a non-GMO way to do very precise editing,” Church explains. This could reduce the need for lengthy, and sometimes expensive, testing. Genetically designed foods, such as corn, soybeans, and potatoes, could be commercially available within five years. Many of these crops would be easier to grow, more drought-resistant, and more resistant to insects and diseases. They might also grow faster and produce larger yields.

However, the concept goes far beyond field crops. A U.S. company, Recombinetics, is currently using CRISPR and a related technology dubbed TALEN to produce dairy cattle without horns, making them easier to handle. Meanwhile, synthetic biology is now being used to identify spoiled or infected food, and researchers in China have engineered beagles with about double the typical muscle mass, making them suitable for police and military work. Guangzhou General Pharmaceutical Research Institute boasts it produces between 2,000 and 2,500 of these dogs annually.

Yet, much of the work in synthetic biology centers on humans. For instance, Church is working on the complex task of programming pig genomes so pig organs such as livers, kidneys, hearts, and lungs would be compatible with humans, thus expanding the current boundaries of transplantation. He and others are also focusing on developing virus-resistant cells that could be used to eliminate mosquitos or their ability to carry Malaria, which currently kills upward of 650,000 humans per year. “It would target a very small fraction of about 3,500 species of mosquitos,” he explains.

In fact, the possibilities are limited only by the creativity and ingenuity of research teams and further advances in the science, Church says. “The technology could have a profound impact on health and medicine. It’s likely that we will begin to see clinical trials in programmable biology and, after testing gene therapies in animals, move forward to a small set of humans. If it is effective within that group, it will then move onto a larger set of humans.”

DNA by Design

Synthetic biology and genetic re-engineering also introduce ethical and moral questions. Designer dogs many elicit concern, but designer humans fall into an entirely different category. Should parents be allowed to choose characteristics they desire for children—say, blue eyes, or a more muscular physique? Should some parents (essentially those who are more affluent) have access to a technology that allows them to produce babies that are more likely to live longer? What if someone decides to weaponize or sabotage genes in order to introduce disease or cause other problems?

“We should not be dismissive of the potential consequences, including ecological fallout, of the technology,” Church says. “Oftentimes, we think we understand a system and later discover that we really don’t understand it.” Nevertheless, he is optimistic about synthetic biology and its potential impact on the planet. “We slog along and often find we are better off with the technology than without it. With programmable biology, there are opportunities to make life and the world better.”

Adds Lu: “Those who work in the field understand the need to monitor and surveil the technology so that it is not used in nefarious ways. There is now a huge emphasis on ethics and safety.”

One of the biggest questions is how genetic modifications will play out over time in the real world. Although it is theoretically possible for a gene to mutate in an unintended way and introduce dangers and risks, it is far more likely that, as genetically designed genes become part of the natural environment and succumb to evolution, they will eventually fizzle out and lose their characteristics. “These organisms are usually just a little bit ‘sick’ because we are taking up resources that should be used for growth,” Lu explains. “If they can shake off the engineered piece, they’re happier. So, over multiple generations, the beautiful, engineered component may go away; the organism jettisons the engineered parts in order to grow faster.”

Another question is whether scientists can construct more elaborate systems—basically, a new and different

type of computer—from biologically engineered components. Sauro points out that all biological systems have reactants bouncing around and colliding with one another, all of which produces “noise.” Evolution exploits this noise, but it is also possible it could be used to compute information. “It’s an open issue about whether there is any advantage to building a computer from biological systems,” Sauro points out.

For now, synthetic biology is marching forward. Over the next couple of decades, Gibson says the technology will usher in huge changes in medicine, biofuel, nutrition, agriculture, and more. “We are taking 4 billion years of evolution and decoding how the biological processes work on a genome. ... Right now, we have only a small section of the overall code. In the computing world, we would not have enough to boot up the computer. But in the biological world, we have now built pieces of DNA large enough that we can produce self-replicating cells. This gives us the opportunity to bang on the problem and find tailored applications.” **C**

Further Reading

Vora, S., Tuttle, M., Cheng, J., and Church, G. **Next Stop for the CRISPR Revolution: RNA Guided Epigenetic Regulators**, *The FEBS Journal*, 10.1111/febs.13768. <http://onlinelibrary.wiley.com/doi/10.1111/febs.13768/epdf>

Chavez, A., Pruitt, B.W., Tuttle, M., Shapiro, R.S., Cecchi, R.J., Winston, J., Turczyk, B.M., Tung, M., Collins, J.J., and Church, G.M. **Precise Cas9 Targeting Enables Genomic Mutation Prevention**. <http://biorxiv.org/content/biorxiv/early/2016/06/14/058974.full.pdf>

Pires, D.P. Cleto, S., Sillankorva, S., Azeredo, J., and Lu, T.K. **Genetically Engineered Phages: A Review of Advances over the Last Decade**, *Microbiology and Molecular Biology Review*, 80:523–543. June 1, 2016. <http://mmbr.asm.org/content/80/3/523.short?rss=1>

Jusiak, B., Cleto, S., Perez-Piñera, P., and Lu, T.K. **Engineering Synthetic Gene Circuits in Living Cells with CRISPR Technology**, *Trends in Biotechnology*, Volume 34, Issue 7, July 2016, Pages 535–547. <http://www.sciencedirect.com/science/article/pii/S0167779915002747>

Samuel Greengard is an author and journalist based in West Linn, OR.

© 2016 ACM 0001-0782/16/12 \$15.00

ACM Member News

SCHRÖDER LEAVES MESSAGE FOR COMPUTER GRAPHICS



Peter Schröder, professor of computer science and applied and computational mathematics at

the California Institute of Technology (CalTech), grew up in Germany. There, he worked as a massage therapist, until a transformative experience altered his path.

“In 1984, a friend turned me on to *Tron*, the science fiction movie, if you’re old enough to remember it,” says Schröder. “I had to find out how they did that; I had to just know.”

Shortly after seeing the movie, Schröder learned of ACM SIGGRAPH. He signed up for the SIGGRAPH conference, and took a mathematics and computer graphics course there that changed the trajectory of his career.

Schröder resumed his education, first at the State University of New York at Stony Brook, then returning to Germany to earn his undergraduate degree in computer science and pure mathematics from the Technical University of Berlin. In 1990, he earned a master’s degree from the Massachusetts Institute of Technology, and then a Ph.D. in computer science from Princeton University in 1994. After a year of post-doctoral research at the University of South Carolina, he joined Caltech in 1995, and has been there ever since.

Schröder, a pioneer in geometry processing, in 2003 received the ACM SIGGRAPH Computer Graphics Achievement Award in recognition of his work in the use of hierarchical methods in geometry processing.

“The big theme of my work right now is discreet differential geometry. I focus on the underlying machinery and computer graphics, which is great for students,” explains Schröder. “They get more motivated when they see fun pictures coming out of it.”

—John Delaney

Containers Push Toward the Mayfly Server

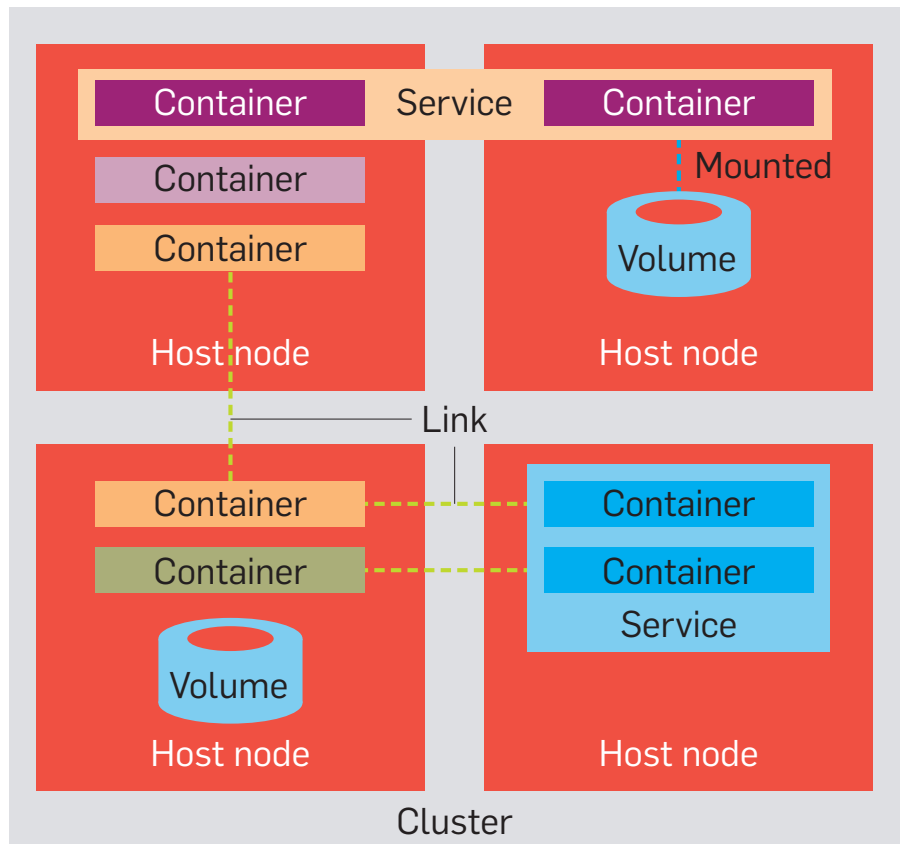
The container revolution represents a large-scale shift in thinking about multitasking systems.

THE DRIVE FOR efficiency in servers is changing the way applications and operating systems interact. The process has accelerated in just the past five years, as server-farm operators have moved on from virtual-machine technology as a way of improving hardware utilization toward even more streamlined options. The work has led as far as the operating system and application being compiled into one block of software and stripping out any unused services to reduce both memory footprint and startup times.

Speaking about a project he and fellow researcher Anil Madhavapeddy worked on to pursue more efficient server software, Richard Mortier, University Lecturer in the University of Cambridge's Computer Laboratory, says: "The original motivation that Anil and I had was that you should be able to write software for the cloud, particularly for network-connected services. But if we were to do that, what would it look like? Related to that was the idea that it should be possible to build software without having to worry about what platform it was targeted for."

To a limited extent, the move to virtualization provided an answer for the second problem. Virtualization lets completely different operating systems and their associated applications share the same processors on a server blade. A hypervisor manages and schedules the operating systems running within each virtual machine (VM).

The problem with virtualization is that each VM partition calls for a complete installation of the operating system and its support software, even if those partitions run the same versions and differ only terms of the applications they or the users who own them utilize. The container, an approach popularized by companies such as



A container-based virtualization architecture.

Docker, removes a lot of this overhead by sharing one operating system image among multiple partitions. Each container only stores the additional services and tasks required by the applications they hold, which can greatly reduce the memory footprint. Runtime also improves because full virtualization demands multiple context switches whenever I/O calls are made. Not only does the operating system need to switch into a supervisor mode to handle I/O, the hypervisor itself forces a switch to a more heavily protected mode in order to service the I/O request.

Studies by Ericsson and IBM have found containers to have little more overhead than a conventional oper-

ating system running on bare metal. Virtualized installations imposed a performance penalty for I/O-intensive applications, although improvements in hardware support for virtualization have narrowed the performance gap.

Even as the performance gap has reduced, the growing base of support software that has emerged around Docker and its competitors has bolstered market acceptance of containers. Orchestration software, such as Google's Kubernetes or Apache's Mesos, has given large users of server farms the ability to quickly start containers and to delete them just as rapidly. Chris Aniszczyk, interim executive director of the Cloud Native Computing Foundation and for-

mer engineering manager at Twitter, a major user of Mesos, says the average container-based workload at the social-media company ran for just 10 minutes of execution time.

A variety of open source projects have emerged that build on top of orchestration. Services will find the best mixture of hardware for a given group of containers and link them to data stores. Monitoring and logging services ensure the containers run correctly and trigger remedial action if things go wrong. But as the layers of software around orchestration build up, they cause a divergence between development and deployment.

Casey Bisson, director of product management at cloud computing service provider Joyent, says it has become more difficult for developers to emulate an orchestration environment that can greatly affect how containerized applications run when deployed to servers. “We have to make the orchestration software laptop more friendly,” he says.

The unikernel architecture developed at the University of Cambridge aims to help solve the problem of achieving platform independence. “Today, we write software that embeds assumptions about the execution platform. If you need to change the platform, at best you need to recompile; worst case, it calls for a rewrite,” Mortier says. “One of the concepts behind unikernels is that you are pushing these things into the toolchain.”

Mortier says the unikernel borrows from the library operating system and exokernel research of the 1990s. “It should be possible to do better than we are now. If you look at how hardware resources are handled today, you have hardware that’s abstracted through a virtual-machine hypervisor. Then it goes through the operating system kernel, the language runtime, and then more libraries on top. You have four or five layers of scheduling all trying to do the same thing. It seems a bit ridiculous,” he argues.

The unikernel bakes the application and the operating system into one executable image, removing most of the layers between them. To prepare the unikernel, a compiler analyzes the application for its dependencies so only those parts of the operating sys-

Mortier says the rapid creation and deletion of unikernel-based systems could enhance security by moving resources around the network.

tem that are needed are incorporated into the image. Mortier says the model provides better security because the unikernel has a much smaller attack surface than a full operating system and its attendant libraries.

Although in many cases the unikernel will run on directly on the host processor with no intervening software, proponents envision implementations where multiple unikernels share one processor using hardware-assisted virtualization. The lack of layering and software duplication should make the installation more efficient than traditional techniques.

Bryan Cantrill, chief technology officer of Joyent, argues the restrictions of unikernels are too great to bear, decrying the idea as a move back to the days of single-tasking operating systems such as DOS. The lack of multitasking within the unikernel makes it difficult to run standard debug tools, he says. Mortier points out it is possible to link debug and trace libraries into the executable, and additional tooling is likely to develop to support unikernels.

A further apparent downside of the first generation of unikernels is that they are designed for single languages that use strong typing, such as Haskell and OCaml. Yet unikernel projects such as RumpRun have opened the field to a wider range of languages and software by supporting software that uses the same Posix interfaces as those provided by operating systems such as BSD. The Cambridge group favored OCaml because it made sense to them to focus on more modern languages

for something that potentially can reshape how server-based computing is done. Docker has signaled its willingness to investigate the wider adoption of unikernels through its purchase of Unikernel Systems, a spinout from research at the University of Cambridge.

Mortier says unikernels are unlikely to be used in isolation, but will be aimed at particular jobs where security or performance are most important. “There is an assumption here that everything is networked,” he adds.

An experimental installation at the university is divided into microservices provided by a group of networked processors. Conventional containers host services such as MediaWiki, with unikernels used to handle redirection to HTTPS addresses and the transaction-layer security (TLS) protocol itself. The relatively low overhead of the unikernel software makes it possible to create them for single transactions.

One 2015 experiment on a system called Jitsu reduced boot time to 20ms on an Intel server processor, compared to five seconds for a conventional web-server running on a VM. Mortier says the rapid creation and deletion of unikernel-based services could enhance security by moving resources around the network. “There is no stable machine that can be targeted,” he claims.

The rapid movement of software and microservices around the network creates its own problems. Midokura systems engineer Cynthia Thomas points to issues such as “traffic tromboning,” in which traffic between microservices and their data stores crisscross the network many times, making the connections look like the folded pipes in a brass instrument.

The tromboning effect not only increases the response time as perceived by the user, but can cause the address tables in networking equipment to run out of space because of the larger number of live connections they need to maintain between services that previously would have been hosted on a single machine. The support software for orchestration software is evolving hand-in-hand with virtual networking software to create dynamic clusters of microservices that make better use of the underlying network hardware.

Although unikernels could be improve security at several levels, one po-

tential disadvantage of the containers that continue to run alongside those unikernels is the weakening of security compared to traditional VM environments. Most container platforms today use only software protection for isolation, and do not have recourse to the hardware enforcement available with a hypervisor-based VM environment.

Environments such as Docker use kernel-provided namespaces to provide software in each container with the illusion it is the sole inhabitant of the Linux system it sees. In principle, and as long there are no security vulnerabilities in the underlying container software or operating system to exploit, the containerized application has no way to alter data in other containers running alongside it on the processor blade. However, researchers have published proof-of-concept attacks that use side-channel techniques to eavesdrop on neighboring containers. An attack published by researchers at the University of North Carolina

in 2014 monitored contention in the cache to listen in on applications running in other containers.

The Cloud Native Computing Foundation's Aniszczyk points out that the rapid creation and deletion of services containers encourage makes it more feasible to run them exclusively on their target hardware. As a result, timeslicing moves from being performed on the order of tens of milliseconds to that of minutes. It supports a model where server-farm operators can dynamically allocate entire blades to an application for the seconds or minutes it needs to run.

"With availability comes flexibility and dynamism," says Mortier. "You can scale up and down quickly."

In this way, the container revolution represents a large-scale shift in thinking about multitasking systems—one that treats compute as a resource made abundant by Moore's Law, rather than the traditional view that processor capacity is scarce. ■

Further Reading

Morabito, R.

Power Consumption of Virtualization Technologies: An Empirical Investigation, Proceedings of the 8th IEEE/ACM International Conference on Utility and Cloud Computing (2015).

Madhavapeddy, A., et al

Jitsu: Just-in-Time Summoning of Unikernels, 12th USENIX Symposium of Networked System Design and Implementation (NSDI15), 559-573

Engler, D.R., Kaashoek, M. F., and O'Toole, Jr., J. *Exokernel: An Operating System Architecture for Application-Level Resource Management, Proceedings of the fifteenth ACM Symposium on Operating Systems Principles: 251-66 (1995).*

Zhang, Y., Juels, A., Reiter, M., and Ristenpart, T. *Cross-Tenant Side-Channel Attacks in PaaS Clouds, Proceedings of CCS'14, 990*

Chris Edwards is a Surrey, U.K.-based writer who reports on electronics, IT, and synthetic biology.

© 2016 ACM 0001-0782/16/12 \$15.00

Milestones

Computer Science Awards, Appointments

ACM, IEEE RECOGNIZE NISAN WITH KNUTH PRIZE

The 2016 Donald E. Knuth Prize recently was awarded to Noam Nisan of the Hebrew University of Jerusalem for fundamental and lasting contributions to theoretical computer science in areas including communication complexity, pseudorandom number generators, interactive proofs, and algorithmic game theory.

The work of Nisan, a professor of computer science in the School of Engineering and Computer Science of the Hebrew University of Jerusalem, has had a fundamental impact on complexity theory, which examines which problems could conceivably be solved by a computer under limits on its resources, whether it is on its computation time, space used, amount of randomness, or parallelism. One way in which computer scientists have explored complexity limits is through the use of randomized algorithms; Nisan has made major contributions exploring the power of randomness

in computations. His work designing pseudorandom number generators has offered many insights on whether, and in what settings, the use of randomization in efficient algorithms can be reduced.

Nisan has been a major player in Algorithmic Game Theory and laid the foundation of Algorithmic Mechanism Design (a mechanism is an algorithm or protocol designed so rational participants, motivated purely by self-interest, will achieve the designer's goals). He designed some of the most effective mechanisms by providing the right incentives to the players, and has shown that in a variety of environments, there is a trade-off between economic efficiency and algorithmic efficiency.

He is also a leading authority in communication complexity, an area of research that examines the amount of information that needs to be transferred between parties for computational problems.

The annual Donald E. Knuth Prize recognizes outstanding contributions to the foundations of computer

science by individuals for their overall impact in the field over an extended period, and includes a \$5,000 award. It is jointly bestowed by the ACM Special Interest Group on Algorithms and Computation Theory (SIGACT) and the IEEE Computer Society Technical Committee on the Mathematical Foundations of Computing (TCMF).

2 COMPUTER SCIENTISTS AMONG NEWEST MACARTHUR 'GENIUSES'

The MacArthur Foundation's recent announcement of its 2016 MacArthur Fellows, commonly known as the "genius grants," included computer scientists Subhash Khot and Bill Thies.

All of the Fellows receive a no-strings-attached \$625,000 grant for their exceptional creativity and potential for future contributions to their fields.

Thies, a computer scientist at Microsoft Research India in Bangalore, India, works to create innovative solutions to a host of socioeconomic challenges facing low-income

communities in the developing world. Thies has used affordable mobile phone technology to connect people in rural India, giving them a way to consume and create digital content through simple phone calls. His work has had wide-ranging impacts in areas of citizen journalism, mobile health applications, and higher education.

Khot, Silver Professor of Computer Science at the Courant Institute of Mathematical Sciences of New York University, is a theoretical computer scientist and the architect of the Unique Games Conjecture, which Khot and other researchers have used to make enduring discoveries in seemingly unrelated areas, such as electoral stability and the structure of foams.

The MacArthur Fellows Program awards unrestricted fellowships to talented individuals who have shown extraordinary originality and dedication in their creative pursuits and a marked capacity for self-direction.

Can We Trust Autonomous Weapons?

Nations consider using defense systems that can make their own lethal decisions.

MOST REASONABLE PEOPLE can see the benefits of using fully autonomous systems, particularly to help prevent injuries or death, as is the case with advanced driver assistance systems increasingly found in automobiles. When it comes to autonomous systems that are designed to take life rather than preserve it, there is significantly more debate.

Currently, the U.S. and other nations do not have any weapons systems that can operate fully autonomously, which is defined in military parlance as selecting, aiming, and firing at a target without a human being “in the loop,” or somehow in control of the weapon system. However, a variety of military weapons systems operate semiautonomously, requiring some human control or input to select or choose targets, but relying on pre-programmed algorithms to execute a strike.

A good example of this is the Lockheed Martin Long Range Anti-Ship Missile (LRASM) system, slated to enter service in the U.S. defense system within the next two years. The LRASM can be fired from a ship or plane and autonomously travel through the air, avoiding obstacles outside of the target area. Published reports indicate humans choose and program the algorithms to seek out and identify potential targets, thus keeping a human in the loop. While the exact factors that make up the target selection algorithm are classified, it is likely a weighting of elements such as the target’s size, location, radar signature, heat profile, or other elements that positively identify the target.

Another example of a system with semiautonomous capabilities is Samsung’s SGR-A1, a military border sentry robot in development for deployment on the border between North and South



Korea. Essentially an unarmed guard tower, the system is designed to assist border guards by scanning the area for those who might try to cross the border. The system is armed with a light machine gun that can dispense tear gas or rubber bullets, and is equipped with cameras, a laser range finder, and a pattern recognition algorithm designed to discern between people and animals. Currently, the system is designed to be operated under human

control for target verification, though developers have given it the capability to use its sensors to detect, select, and shoot at targets autonomously.

It is the last capability that has watchdogs worried. Systems such as LRASM and the SGR-A1 are now only approved for use with a human approving targets to be killed by the system, but there is considerable concern the U.S. and other world powers are on the fast track to developing machines able

to kill people independently.

"I think it's pretty clear that military mastery in the 21st century is going to depend heavily on the skillful blending of humans and intelligent machines," says John Arquilla, professor and chair of Department of Defense Analysis at the U.S. Naval Postgraduate School in Monterey, CA. "It is no surprise that many advanced militaries are investing substantially in this field."

Indeed, in late 2014, U.S. Secretary of Defense Ash Carter unveiled the country's so-called "Third Offset" strategy, essentially an attempt to offset the shrinking U.S. military force by incorporating technologies that improve the efficiency and effectiveness of weapons systems. While many of the specific aspects of the strategy are classified, industry observers agree a key tenet is increasing the level of autonomy in weapons systems, which will improve warfighting capability and reduce the number of humans required to operate weapons systems.

"For over a decade, we were the big guy on the block," explains Major General (Ret.) Robert H. Latiff, an adjunct professor at the Reilly Center for Science, Technology, and Values at the University of Notre Dame, and Values Research Professor and an adjunct professor at George Mason University. "But with the emergence of China, and the reemergence of Russia, both of whom are very, very technically capable, and both of whom have invested fairly significantly in these same technologies, it goes without saying that the DoD (the U.S. Department of Defense) feels like they need to do this just to keep up."

Military guidelines published by the U.S. Department of Defense in 2012 do not completely prohibit the development and use of autonomous weapons, but require Pentagon officials to oversee their use. That is why human rights groups such as the Campaign to Stop Killer Robots are actively lobbying the international community to impose a ban on the development and use of autonomous weapons systems.

"We see a lot of investment happening in weapons systems with various levels of autonomy in them, and that was the whole reason why we decided at Human Rights Watch back in 2012 to look at this," explains Mary Wareham, advocacy director of the Arms

Division of the Human Rights Watch, and coordinator of the Campaign to Stop Killer Robots. "We're seeking a preemptive ban on the development, production, and use of fully autonomous weapons systems in the United States and around the world."

"We're focusing quite narrowly on the point at which critical functions of the weapons system become autonomous," Wareham says. "The critical functions that matter to us are the selection and identification of the target, and the use of force." Wareham's main issue is that the fully autonomous weapons systems of the future may rely solely on algorithms to target and kill enemy targets, without a human in the loop to verify the system has made the right decision.

Human Rights Watch is pushing to have a negotiated international treaty limiting, restricting, or prohibiting the use of autonomous weapons systems written and ratified within the next two or three years. "If this becomes [a decade-long process], then we're in big trouble," she admits, noting that at present, there are no such treaties in process within the international community. "At the moment, it is just talk," Wareham acknowledges.

A key argument of groups such as Human Rights Watch is that these systems, driven by algorithms, may make mistakes in target identification, or may not be able to be recalled once deployed, even if the scenario changes. Others with military experience point out that focusing on the potential for mistakes when using fully autonomous weapons systems ignores the re-

alities of warfighting.

"I think one of the problems in the discourse is the objection that a robot might accidentally kill the wrong person, or strike the wrong target," Arquilla says. "The way to address this is to point out that in a war, there will always be accidents where the innocents are killed. This has been true for millennia, it is true now, and it is true with all of the humans killed in the Medecins Sans Frontieres hospital in Afghanistan."

Arquilla adds that while the use of artificial intelligence in weapons will not eliminate mistakes, "Autonomous weapons systems will make fewer mistakes. They don't get tired, they don't get angry and look for payback, they don't suffer from the motivated and cognitive psychological biases that often lead to error in complex military environments."

Furthermore, military experts feel an outright ban would be impossible to enforce due to the secretive nature of most militaries, and likely would not be in the best interest of any military group or nation.

Indeed, even with today's military technologies, getting the military or its contractors to discuss the exact algorithms used to acquire, select, and discharge a weapon is difficult, as disclosing this information would put them at a distinct tactical disadvantage. Therefore, even if a ban were to be put in place, devising an inspections system similar to those used for chemical and anti-personnel mines would be extremely complicated.

"A ban is typically only as good as the people who abide by it," Latiff says, noting that those who will sign and fully abide by a ban make up "a pretty small fraction of the rest of the world." In practice, he says, "When something becomes illegal, everything just goes underground. It's almost a counter-productive thing."

Work on autonomous weapons system has been going on for years, and experts insist expecting militaries to stop developing new weapons systems that might provide an advantage is foolhardy and unrealistic. As such, "There is absolutely an arms race in autonomous systems underway," Arquilla says. "We see this in both countries that are American allies, and also

"We're seeking a preemptive ban on the development, production, and use of fully autonomous weapons systems in the United States and around the world."

among potential adversaries. In particular, the Russians have made great progress. So have the British; they are putting together a fighter plane that can do everything a piloted fighter plane can, and can be built to higher performance characteristics, because you don't have a human squeezed by G-forces in the cockpit."

Others agree, even if they admit that at present, there are no significant advantages to using fully autonomous weapons, versus the semiautonomous systems already in use.

"I would imagine that [as autonomous weapons] become more capable, they will be seen to operate more effectively than systems with humans in the loop," says Lieutenant Colonel Michael Saxon, an assistant professor teaching philosophy at the U.S. Military Academy at West Point. "Once you introduce swarms or you have to respond to adversary systems that are autonomous, humans in the loop will create real disadvantages. This, of course, is all predicated on advances in these machines' capabilities."

Still, observers suggest while a ban on autonomous weapons may not be the right course of action, a deliberate approach to developing and incorporating them into the military arsenal is prudent.

"I think we're probably closer to the kind of capabilities we're talking about than most people think, and Russia and China are, too," Latiff says. "These are dangerous systems. In the wrong hands, these things could really be awful."

A ban on autonomous weapons would likely have little impact on the development of weapons systems in the near future, which still will be overseen by humans, even if the actual decision to select and strike a target is made by an autonomous system.

"An autonomous weapon operates without human control, but that does not mean that it is free from human input," Arquilla says. "There are a lot of elements to the chain [of command]; the unfortunate term is the 'kill chain.' There will be people and machines intermixed within the chain."

Whether or not a ban is put into place, the international community is likely to be faced with significant moral and legal questions surrounding

"Once you introduce swarms or you have to respond to adversary systems that are autonomous, humans in the loop will create real disadvantages."

the use of autonomous weapons, and whether they will be developed in ways that are consistent with accepted ideas about ethics and war, Saxon says.

"There are good arguments about increased moral hazard with autonomous weapons systems, that they make killing too easy," Saxon says. "I think they also have an effect on traditional military virtues that need to be examined. What does it mean to be courageous, for instance, when your machines take the risks and do the killing?"

For Latiff's part, while he does not support a ban on autonomous weapons, he would support a non-proliferation treaty allowing militaries to research and test these systems to ensure they can be made as reliable and safe as possible.

"At the end of the day, it's kind of like nuclear weapons," Latiff says. "Everybody's going to get them, and the people that don't get them are going to want them. The best we can hope for is that we slow it down." **C**

Further Reading

U.S. Department of Defense 2012 Directive on Autonomous Weapons: <http://www.dtic.mil/whs/directives/corres/pdf/300009p.pdf>

Campaign to Stop Killer Robots: <https://www.stopkillerrobots.org/>

Video: Scary Future Military Weapons Of War-Full Documentary: <https://www.youtube.com/watch?v=DDJHYEdKCBE>

Keith Kirkpatrick is principal of 4K Research & Consulting, LLC, based in Lynbrook, NY.

© 2016 ACM 0001-0782/16/12 \$15.00

Milestones

Ken Kennedy Award Goes To Gropp

ACM and IEEE Computer Society named William D. Gropp, Thomas M. Siebel Chair of Computer Science at the University of Illinois at Urbana-Champaign, as recipient of the 2016 ACM/IEEE Computer Society Ken Kennedy Award for his highly influential contributions to the programmability of high performance parallel and distributed computers.

Gropp was a leader in the development of the Message Passing Interface (MPI) standard, the de facto standard for programming distributed-memory computers in scientific applications, and a key author in the development of MPI-1, MPI-2, and MPI-3. Gropp also designed and developed MPICH, the first functional implementation of MPI.

Along with collaborators David Keyes and Xiao-Chuan Cai, Gropp developed and analyzed key scalable parallel algorithms for adaptive mesh refinement and domain decomposition methods, which are now widely used in parallel applications. This led to the development of the Portable, Extensible Toolkit for Scientific Computation (PETSc), which Gropp developed in collaboration with Barry Smith. PETSc has been used in a variety of applications, including nano-simulations, biology, fusion, geosciences, environmental modeling, fluid dynamics, and software engineering, among others.

In 2011, Gropp helped launch the ACM Special Interest Group on High Performance Computing (SIGHPC). He also served as the editor of *CONNECT*, the newsletter for SIGHPC; as chair of the Gordon Bell Prize Committee, and in various capacities for numerous conference committees.

Gropp, who also is Acting Director and Chief Scientist of the National Center for Supercomputing Applications, has been elected an ACM Fellow and an IEEE Fellow, and is a member of the U.S. National Academy of Engineering.

ACM

ON A MISSION TO SOLVE TOMORROW.

Dear Colleague,

Without computing professionals like you, the world might not know the modern operating system, digital cryptography, or smartphone technology to name an obvious few.

For over 60 years, ACM has helped computing professionals be their most creative, connect to peers, and see what's next, and inspired them to advance the profession and make a positive impact.

We believe in constantly redefining what computing can and should do.

ACM offers the resources, access and tools to invent the future. No one has a larger global network of professional peers. No one has more exclusive content. No one presents more forward-looking events. Or confers more prestigious awards. Or provides a more comprehensive learning center.

Here are just some of the ways ACM Membership will support your professional growth and keep you informed of emerging trends and technologies:

- Subscription to ACM's flagship publication ***Communications of the ACM***
- Online books, courses, and videos through the **ACM Learning Center**
- Discounts on registration fees to ACM Special Interest Group conferences
- Subscription savings on specialty magazines and research journals
- The opportunity to subscribe to the **ACM Digital Library**, the world's largest and most respected computing resource

Joining ACM means you dare to be the best computing professional you can be. It means you believe in advancing the computing profession as a force for good. And it means joining your peers in your commitment to solving tomorrow's challenges.

Sincerely,



Vicki L. Hanson
President
Association for Computing Machinery



Association for
Computing Machinery

Advancing Computing as a Science & Profession

SHAPE THE FUTURE OF COMPUTING.

JOIN ACM TODAY.

ACM is the world's largest computing society, offering benefits and resources that can advance your career and enrich your knowledge. We dare to be the best we can be, believing what we do is a force for good, and in joining together to shape the future of computing.

SELECT ONE MEMBERSHIP OPTION

ACM PROFESSIONAL MEMBERSHIP:

- ☐ Professional Membership: \$99 USD
- ☐ Professional Membership plus
ACM Digital Library: \$198 USD (\$99 dues + \$99 DL)
- ☐ ACM Digital Library: \$99 USD
(must be an ACM member)

ACM STUDENT MEMBERSHIP:

- ☐ Student Membership: \$19 USD
- ☐ Student Membership plus ACM Digital Library: \$42 USD
- ☐ Student Membership plus Print *CACM* Magazine: \$42 USD
- ☐ Student Membership with ACM Digital Library plus
Print *CACM* Magazine: \$62 USD

- ☐ **Join ACM-W:** ACM-W supports, celebrates, and advocates internationally for the full engagement of women in computing. Membership in ACM-W is open to all ACM members and is free of charge.

Priority Code: CAPP

Payment Information

Name _____

ACM Member # _____

Mailing Address _____

City/State/Province _____

ZIP/Postal Code/Country _____

Email _____

Payment must accompany application. If paying by check or money order, make payable to ACM, Inc., in U.S. dollars or equivalent in foreign currency.

☐ AMEX ☐ VISA/MasterCard ☐ Check/money order

Total Amount Due _____

Credit Card # _____

Exp. Date _____

Signature _____

Purposes of ACM

ACM is dedicated to:

- 1) Advancing the art, science, engineering, and application of information technology
- 2) Fostering the open interchange of information to serve both professionals and the public
- 3) Promoting the highest professional and ethics standards

Return completed application to:
ACM General Post Office
P.O. Box 30777
New York, NY 10087-0777

Prices include surface delivery charge. Expedited Air Service, which is a partial air freight delivery service, is available outside North America. Contact ACM for more information.

Satisfaction Guaranteed!

BE CREATIVE. STAY CONNECTED. KEEP INVENTING.



Association for
Computing Machinery

1-800-342-6626 (US & Canada)
1-212-626-0500 (Global)

Hours: 8:30AM - 4:30PM (US EST)
Fax: 212-944-1318

acmhelp@acm.org
acm.org/join/CAPP

The Profession of IT Learning to Learn

Do you get stuck when it is time to learn something new? Read this.

WHY IS IT, when we need to learn something new that will benefit our work or home life, we often find ourselves blocked by seemingly invisible forces? When learning fails we miss out on important projects, promotions, and opportunities. We end up suffering and falling short of our objectives. Somehow, for many of us, our natural capacity to learn seemed to deteriorate over time, especially in areas that we care about the most.

Educators and business leaders have used the term “learning to learn” to name a missing skill that would reverse the deterioration. The business literature is filled with tips and buzzwords about this skill: “learn from mistakes,” “fail fast and often,” “learn faster with technology,” “be curious,” “collaborate,” and “take down silos.”^{7,8} But despite these wise aphorisms, something holds us back. Something within us resists—and even gives up on—learning. What is it?

Time after time in our work with professional teams we have found that many are in the thrall of deep, automatic, and unnoticed assessments and assumptions that guide what they think is possible or appropriate to do.

Learning to learn is a navigational skill of recognizing moods that block learning and of shifting to moods that enhance learning.

These assessments and assumptions are the invisible force blocking learning. They manifest as moods.

Cultivating moods conducive for learning is not as simple as setting up a good physical environment, for example with moveable tables, calming music, and bean-bag chairs near the community watercooler. In this column, we argue here that learning to learn is a navigational skill of recognizing moods that block learning and of shifting to moods that enhance learning.

Developing the Skill

Although we understand intellectually that learning requires that we make

mistakes and learn from them, our emotions steer us away from making mistakes. In school we get bad grades if we make mistakes on tests; no one celebrates our mistakes. At work we can get a poor performance review if we make too many mistakes. Hence, even if we know that it is okay to make mistakes, we are conditioned to avoid making them and, if we do make a mistake, we feel frustrated and often embarrassed.

Similarly, we understand intellectually that technology can accelerate our learning. Simulation technologies offer to train us in skills without the risks of injury if we make mistakes—for example, in a flight simulator. Yet even in virtual environments people commonly get frustrated with their inability to learn as fast as they want and they give up. Internet and Web technologies offer to bring vast stores of information to our screen on demand via search and give us access to the world’s best courses free via MOOCs. Nonetheless, most of us discover that neither Internet searches nor MOOCs answer our questions. Instead of feeling empowered by these technologies, we often feel overwhelmed and anxious about how little we know compared to what we think we *should* know, but do not. We feel stupid if we do not know some-

thing that seems so easily accessible. In other words, access to these advanced technologies, whether for skill development or knowledge accumulation, is not a surefire path of learning.

Overwhelm, frustration, and anxiety not only cause us stress, but they also close us down to learning. Then we fall into a no-possibility mood, in which we see no chance of ever learning enough to achieve our objectives. Often, we feel like running away and quitting instead.

In short, there is an intimate connection between moods and learning. Certain moods support our learning; others block our learning. Learning to navigate away from learning-blocking moods and toward learning-enhancing moods is at the core of developing learning to learn as a skill.

Learning and Learning-to-Learn

Let us distinguish between learning a skill in a domain, such as music, sports, architecture, or programming, and learning to learn. Learning-to-learn is a disposition of openness to learning in a domain. If we are not open to learning, we will not learn.

We have found that the skill acquisition framework proposed by Stuart and Hubert Dreyfus in 1980 is excellent for understanding how we learn skills in a domain.⁴ Their framework says that a learner moves through six stages: beginner, advanced beginner, competent, proficient, expert, and master (see Table 1.) A person's progress takes time, practice, and experience. The person moves from hesitant rule-based behaviors as a beginner to fully embodied, confident, intuitive, and game-changing behaviors as a master. Hubert Dreyfus gives descriptions of these levels in his book *On the Internet*.³ A beginner, for example, does not know the domain and must rely on others to teach the basic rules and correct mistakes. A competent person knows how to do all the basic standard practices of the domain and does not need supervision to avoid common mistakes. Continued practice is necessary to guarantee progress through these levels.

But there is more to the story than practice. In working with hundreds of people as they sought to develop new skills, we have seen that, in each of these stages, people routinely fall into unproductive moods that make



them no longer want to practice, block them from progressing toward the next stage, and divert them from reaching their learning objectives. Despite an initial desire to succeed, they often disengage and ultimately give up. The key to learning to learn is to cultivate moods that will enable us to keep learning, and to shift out of those that will block learning.

Moods and Assessments that Block Learning

We work with professionals who want to learn how to build high-performing teams. As part of the training, we purposely place the teams into situations

where they must learn something new.⁵ We chose a virtual reality game (World of Warcraft) because it contains many quests that will test a team's ability to coordinate. Very few on the teams have ever played this game before. Hence most teams are rank beginners who know nothing about the game. They cannot move their players until they develop touch sensitivity with the left-hand movement keys. They cannot immediately fathom the game's undocumented complex structure or strategies for completing quests. In such a situation, you would expect them to ask for help from the coaches standing by or from their peers who may know more

Table 1. Stages of learning and their moods.

Stage	Behaviors	Mood Dispositions
Beginner	Person knows of the domain and desires to learn. Declares a commitment to learning the domain. Capable only of following rules given by a teacher or representative of the domain. Must trust the teacher. Can be very slow and tentative while learning and trying out the basic rules.	Positive: wonder, ambition, resolution, confidence in ability to learn. Negative: fear of failure, discomfort with not knowing what to do, impatience with the slow speed of advancement, distrust of the teacher, feeling "I ought to know this," frustration that even simple things do not work as wanted.
Advanced Beginner	Familiarity with common situations. Learns and applies maxims—tips and rules of thumb to be used when certain symptoms appear. Still needs help and is faster about figuring out what to do and making the moves.	Positive: resolution, ambition Negative: Boredom, apathy, distrust
Competent	Has learned the norms of the domain. Common situations all look familiar and the person knows what to do right away. Does not require supervision to avoid common mistakes and satisfy customers. Asks for help when confronted with an unfamiliar situation.	Positive: resolution, ambition, confidence in network and in capacity to ask and receive help Negative: overwhelm, anxiety, insecurity, impatience, frustration
Proficient	Has developed a high level of skill that others admire and imitate. Sets new standards of performance.	Positive: ambition, resolution, care, trust Negative: Impatience, frustration, arrogance
Expert	Has extensive experience. Quickly sees solutions to problems that baffle others. Sought out as a teacher, manager, and problem-solver.	Positive: confidence, wonder, serenity, ambition to achieve mastery, resolution Negative: arrogance, impatience
Master	Has developed a long view of the domain and knows how to intervene to change the game that everyone else is playing.	Positive: wonder, exploration, ambition to contribute to a field, serenity Negative: arrogance, boredom, disinterest

than they do. More often than not, however, that is not what happens. They become frustrated and confused and then give up. In the after-action debriefs, we ask them about their moods and ferret out their hidden assumptions. We hear assessments like these:

- Frustration: I want to contribute to my team, but I have no idea what to do. I add no value to the team.

- Confusion: I do not understand the game. It is way too complex. It is bad not to understand right away. I do not like this.

- Resignation: There is no point in me being part of this team since I do not understand this game. I am never going to understand it.

Occasionally a team has a member who is an expert gamer. You might expect their teams would make better progress. Often the exact opposite happens. The expert player cannot stop from telling everyone else what to do.

Many team members resent the expert for not giving them a chance to learn it for themselves. A few wait docilely to be told exactly what to do—at the price of being immobilized when they encounter an unfamiliar situation.

During debriefings, we assist team members to discover what holds them back from asking for help. Their hidden assessments are mostly variations of these five:

- It is important to be competent.
- It is important to be efficient and avoid waste.
- It is important to be independent and self-reliant.
- It is important to be useful.
- It is important to be prepared at all times.

In each case, they judge themselves to be incapable of doing these important things. Helping them discover these assessments and then examine the standards that they are holding

themselves to is a big step forward in enabling them to develop the skill of learning to learn.

Navigating Moods

All teams must deal with risk and uncertainty. To be an effective learner in such situations, you must learn two things: first, to *identify* the moods that you find yourself in and, second, to *shift* into moods that open you to learning. Tables 2 and 3 give examples of common productive and unproductive moods for learning.^{1,2,6} You can recognize a mood from the assessments you see yourself making about the world around you. For example, if you are experiencing awe and eagerness about what is going on, you are in a mood of wonder. If you see no possibilities for moving forward, you are in a mood of resignation.

The navigational skill of identifying and shifting moods follows the Dreyfus progression just like any other skill development. The beginner learns the names of moods and the conversations that characterize each one. The advanced beginner recognizes moods more easily and can shift some moods. The competent navigator readily identifies unproductive moods and can usually shift to productive moods alone or with assistance.

Shifting moods requires reflection and practice because you may have to unlearn some habits that you formed for coping with different situations. We have found that asking four questions and reflecting on their answers can open the door to shifting moods. We recommend that you engage these four questions with the help of a teacher or teammate. Although you can do this on your own, we have found that a different observer can help you uncover your hidden assessments, standards, and habits faster.

- What are my prevailing moods and assessments in the situation?
- What standards do I have that provoke these assessments?

- What habits do I have that keep me conforming to the standards?

- Why is reaching my learning objective important to me? What will I be able to do if I persevere?

In this investigation, it is important to notice that moods are different from emotions. Emotions are "fore-

ground” occurrences that we feel as sensations such as anger at an insult or joy at a child’s smile. Moods are “background”—they are more like an atmosphere, saturated by vapors of our history and experiences, in which things and possibilities show up in our situation. We can almost always identify the event that triggered an emotion, but we seldom notice our moods. They are automatic assessments that open or close possibilities for us outside our awareness.⁴

Sometimes moods are temporary and can shift easily. For example, you can be in a mood of resignation (no possibilities are visible) and a parent, teacher, or manager can show you a new possibility that breaks the mood. Sometimes a prevailing mood will yield when you discover a hidden standard and determine to change it. However, some moods are deeply ingrained habitual dispositions that lead you into actions you do not intend. You have adopted a set of personal practices that keep you in conformity with your hidden standards. If you try to violate a practice, you feel extremely uncomfortable and you do not even know why. To shift the long-term mood you will have to train yourself into new practices. That may take time.

Example: Delegation

Let us consider an important example: delegation. You consider yourself a poor delegator and you see that you and your team could accomplish more if you were to delegate more. Others, including potential investors in your company, agree. But, for some reason, you do not delegate despite your desire to learn to do so. In your exploration of the four questions, you discover you have these moods:

- **Distrust:** My team members lack important skills; I dare not delegate.

- **Insecurity:** If I delegate, team members are going to think that I am not capable of handling things myself.

- **Resignation:** Every delegation I have tried fails. I wind up having to do it myself anyway.

- **Frustration:** I would like other people to do as well as I do, but I cannot seem to get that to happen. I just have to do it myself and not get to the other stuff that needs my attention until later.

- **Overwhelm:** There is too much to do. I do not have time to wait for other people to do what I need them to do. I just have to do it all myself and hope that I can complete everything. No one can help me.

As you reflect on these moods, you discover you have acquired these standards:

- **The leader must be competent at all times.** Therefore, I must be competent at all times.

- **If the leader delegates and the team fails, it is the leader’s fault.** Delegating to people that are less competent than me is too risky. I might be out of business or get fired.

- **If team members are not up to the task, the leader must step in.** If people cannot perform up to my expectations, I must take over to save the project.

From this reflection you discover that you avoid delegating because of standards you acquired in the past.

Table 2. Moods that support learning.

Mood	Associated Assessments
Ambition	I really want this outcome. I am sure I can achieve it. I am committed to achieving it.
Confidence	I believe in my own capabilities to achieve my goals. I have done this before and can do it again. I know who to ask for help and who will take care of me.
Perplexity, bafflement	I am unable to make sense of what is going on, but I am intrigued. I will keep looking until I figure this out.
Resolution	I am determined to make this happen, no matter what.
Serenity, acceptance	I accept that the past has already happened and I have no control. I accept that the future is full of surprises and I cannot predict. I am grateful to life.
Trust	I am learning from experienced people who are taking care of me. I know they are competent and sincere when they make promises to me. I believe they have my interests at heart.
Wonder	I do not know what is going on, but I like it. There are so many possibilities to learn more.

Table 3. Moods that block learning.

Mood	Associated Assessments
Apathy	I don’t really care about this. I see no reason to try.
Arrogance	I already know this. There is nothing new for me here. They ought to listen to me. I don’t need to listen to other people because what they have to say is irrelevant.
Boredom	This is totally uninteresting. What a waste of time. There is no point even in pretending to be engaged.
Confusion	I do not understand what is going on here and I do not like it. It is bad to be confused. Get me out of here.
Distrust, skepticism	They are insincere or incompetent. They are not watching out for my interests. This may backfire or otherwise not work out.
Fear, anxiety	I don’t know what is going to happen. Something bad will happen to me. I will be harmed if I make a mistake. It’s better not to do anything than to try and fail.
Frustration	I keep trying and failing. I cannot do it as fast as I should. It does not work the way it is supposed to.
Impatience	This is wasting my time. There is no value here. Let’s move on, now.
Insecurity	I have no confidence in myself. I can’t do this. I am not good enough to be here. People are going to criticize me.
Overwhelm	The sheer volume of what is going on is too much for me. I cannot find a way to move forward, except to keep working hard and expect to fail.
Resignation	There are no possibilities for resolving the issue. I am never going to be able to do this. I have seen this happen over and over in the past. Nothing new will happen. There is no point in trying.

Exponential Laws of Computing Growth

Cell-Graphs: Image-Driven Modeling of Structure-Function Relationship

Resolving Conflict

Research for Practice: Web Security and Mobile Web Computing

Faucet: Deploying SDN in the Enterprise

Eulerian Video Magnification and Analysis

HACC

Effects of International Trafficking in Arms Regulations Changes on the U.S. Software Industry

Artificial Intelligence: Think Again

Rethinking the Roles of Central Information Technology Departments

Plus the latest news about truly random numbers, bias in technology, and searching for everything.

Shifting moods requires reflection and practice because you may have to unlearn some habits that you formed for coping with different situations.

Having become aware of these standards and the habits that you have acquired to support them, how do you change them? Begin by focusing on Question 4 in the list of four questions we presented earlier. Why is reaching your learning objective important to you? If you delegate more will you have a stronger team? Will you have more time to focus on strategic issues that benefit your team over time? Will people want to join your team because of their potential for growth? The answers to these questions will help you to cultivate a mood of ambition about the benefits of delegating more effectively. Supported by your ambition, you can resolve to embark on a process for learning to delegate. You start delegating and sharing your expectations for your team members' successes. You will need to cultivate some emotional fortitude during this learning period because it will not be comfortable for you—but you are confident that learning to delegate more effectively and have others grow in their responsibilities is good for you, your team, and your company. If a particular delegation does not work, you take time to evaluate why it did not work, talk to the person about it, and try again until it produces successful results. In other words, by cultivating moods of ambition and resolution you begin to shift away from moods that block you from learning to delegate and you give yourself the time

and the practice necessary to reach your objective of learning to delegate.

Conclusion

When we are blocked from learning, the chances are the blockage comes from hidden internal assessments and standards, not from external factors in the environment. Learning to learn is a skill with which we recognize our prevailing unproductive moods and shift to productive moods. The shift frequently involves retraining our automatic responses to learning situations.

Perhaps the most fundamental aspect of learning to learn is learning to be open to learning. As beginners we are more likely to learn if we are open, curious, and in a mood of wonder about what we do not know. Instead, beginners frequently experience frustration, insecurity, confusion, anxiety, and resignation—which block learning. In those moments, it is important to reflect on what may be provoking those moods so that we can see what actions will shift our moods and allow us to continue learning. Part of the learning process requires that we develop emotional fortitude to tolerate discomforts that come from mistakes and from changing our habitual patterns. If we do this, we will be much more likely to reach our final learning objectives. ■

References

1. Denning, P. Moods. *Commun. ACM* 55, 12 (Dec. 2012), 33–35.
2. Denning, P. Moods, wicked problems, and learning. *Commun.* 56, 3 (Mar. 2013), 30–32.
3. Dreyfus, H. *On the Internet*. (2nd ed. 2008), Routledge, 2003.
4. Dreyfus, S.E. and Drefus, H.L. A five-stage model of the mental activities involved in directed skill acquisition. Storming Media, 1980; <http://www.stormingmedia.us/15/1554/A155480.html>
5. Flores, G. *Learning to Learn and the Navigation of Moods*. Pluralistic Networks Publishing, 2016.
6. Flores, F. and Flores Letelier, M. *Conversations for Action and Collected Essays*. CreateSpace Independent Publishing Platform, 2013.
7. Friedman, T. *The World Is Flat*. (3rd ed.), Picador, 2007.
8. Spear, S. *The High-Velocity Edge*. McGraw-Hill Education, 2010.

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for information innovation at the Naval Postgraduate School in Monterey, CA, is Editor of *ACM Ubiquity*, and is a past president of ACM. The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

Gloria Flores (gloria@pluralisticnetworks.com) is co-founder and President of Pluralistic Networks, and co-designer of Working Effectively in Small Teams (WEST), a four-month immersive course designed to enable participants to develop skills that will enable them to work more effectively in teams, such as coordination of action, listening, building trust, and managing morale.

Copyright held by authors.

Historical Reflections

Nailing Smoke

Curation at the bleeding edge of technology.

“COMPUTERS MAY BE the best repository of all time for information—as long as the operating system or storage medium is not out of date—but they are unable to record or reproduce the sensual presence of a material work of art. Unlike the qualities of material works of art, games and arbitrary interaction do not qualify the computer as a medium for memories and recollections.”¹

In common with most senior academics, I am required from time to time to offer assessments on applications for research funding. Over recent years, I have been pleased to see funding agencies increasingly asking applicants to pay formal attention to the means by which the outputs of publically funded research may be preserved over the medium to long term. A second, although clearly related, concern is how applicants plan to ensure that ongoing access to research results may be achieved. These are good questions, and the fact they are being asked represents real progress, but unfortunately, for the most part, they are not well answered.

Based on my experience, a number of misunderstandings appear to be very widespread. Chief among these is that preservation may be said to have been achieved if a proportion (which may or may not be a significant proportion) of the digital outputs of a project are backed up and stored at some point during the project's lifetime. A related misunderstanding is the notion that all that is required to assure access is that the project team should develop a website and keep it



active for a year or so after the conclusion of a project. There is no obvious consensus on what constitutes the 'medium' or 'long' term, and very little comprehension that preservation and access need to be considered even before a project begins. The choices we make about file formats, data models, hardware, and software all impact on how easy it will be to actively preserve information for future use over the years and decades ahead. Not all the material generated by research proj-

ects is digital, of course, and digital preservation techniques often need to be supplemented by longer established techniques which are widely understood in the gallery, library, archive, and museum (GLAM) domains, but are perhaps less well known elsewhere. The more obvious lacunae in preservation awareness among applicants for public funding are, in principle at least, relatively easy to address, and call for little more than taking advantage of the numerous sources of in-

formation which publically and freely available. However, as one begins to engage seriously with preservation issues, it does not take long to come up against problems that are much more difficult to address, and which are increasingly common.

I have recently been concerned by the preservation challenges presented when dealing with interactivity and ephemerality. The drive toward providing ever more complex and nuanced forms of computer interactivity is not new. Indeed, it has been around right from the earliest days of computing, Ivan Sutherland's Sketchpad represented a significant early step, and Sutherland is usually now remembered as the founding father in computer interaction. The history can be traced through Doug Engelbart's visionary effort during the 1950s to "augment the human intellect" by making available a vast amount of human knowledge via highly responsive workstations. Engelbart's so-called NLS (oNLine System), which was developed courtesy of an ARPA grant facilitated by Sutherland (working under the direction of J.C.R. Licklider), was first demonstrated in 1968 in what came to be called the "Mother of All Demos."^a This remarkable event featured the introduction of the computer mouse, video conferencing, teleconferencing, hypertext, word processing, hypermedia, object addressing and dynamic file linking, bootstrapping, and a collaborative real-time editor.

During the 1970s, Apple did a great deal of work to move interactivity technology out of the laboratory into the living room, and today the ideas put forward by Licklider, Sutherland, and Engelbart are considered commonplace. Development continues of course, and Virtual Reality systems are just a recent manifestation of the increasingly sophisticated ways we are able to interact with computers.

From the perspective of the arts, Oliver Grau¹ has argued that the development of what he calls 'illusory visual space' is part of the overall art history of illusion and immersion, and draws connections with interactive art, interface design, agents, telepresence, and image evolution.

^a Ten clips from the Mother of All Demos are available at <http://bit.ly/2epzIs2>.

The drive toward providing ever more complex and nuanced forms of computer interactivity is not new.

The existence of complex interactivity affordances in modern computer systems has a significant impact on how we think, and can think about computers, and this in turn has consequences for our conception of what we are trying to preserve and later make accessible for future generations.

Among computer scientists, the tendency has traditionally been to concentrate on the tangible or the physical.^b Thus, preserving (say) the financial records of a company would typically be treated as a series of tasks involving ensuring the bits that comprise the company accounts package and its associated data are saved on a stable medium, and stored in a safe place. These in turn would need to be periodically updated to run on new generations of hardware when they become available (migration) or might subject to virtualization or emulation. While the implied processes of careful analysis, storage, preservation management, and access provision, usually work well for financial systems and text documents, they serve us much less well when human-computer interaction offers users the opportunity to influence fundamentally the performance of software. Grau's notion of illusory visual space draws attention to the fact that sometimes the most significant features of an object, and those that we would be the most concerned to preserve, are not tangible

^b Jerry McDonough has written a number of pieces showing how preservation of bits is in many ways the least difficult of the preservation challenges before us. See, for example, McDonough.⁵

features at all, but rather lie outside the materiality of the object.

Brenda Laurel suggests that interactivity can be thought of as sitting on the three axes, of 'frequency', 'range', and 'significance'.³ Frequency represents the number of occasions on which the user interacts with the computer. Range represents the number of distinct choices that are available, while significance represents the degree to which the choices made by the user alters the outcome. Thus: "A not-so interactive computer game judged by these standards would only let you do something once in a while, only give you a few things to choose from, and the things you could choose wouldn't make much difference to the whole action (or produce significant changes to the state of the underlying system). A very interactive computer game (or desktop or flight simulator) would let you do something that really mattered at any time, and it could be anything you could think of."³

As we reflect on the complexities of preserving the key characteristics of artifacts with a high degree of meaningful interactivity, it becomes clear that interaction gives the operation of computers a certain degree of ephemerality, and variability, which in turn produces highly individualized, and potentially unique, behaviors in computer systems. Thus, the behavior of computers may well be impossible to reproduce in full on subsequent occasions, even supposing the software designer/author were motivated to do so. Laurel characterizes the operation of computers as 'performative'¹² and comparable to theatre. So just as no two performances in a theater are (or can be) exactly alike, computer systems whose core behaviors are dependent on the variable input of users may also be unique. The degree to which this is actually so in practice depends crucially on the sort of interaction the computer system in question permits.

In many cases it is more sensible to think of software in terms of what it does, rather than in terms of the lines of code that constitute a program, or the hardware on which the software runs. Construed this way, the human component of human-computer

interaction significantly affects the process of preservation. For example, if we are trying to preserve a flight simulator, we need to consider not only the program as written but the affordances which it offers, and how these are experienced in practice, not all of which may be clear even to the programmer. Even the notion that there is a single ‘object’ of preservation cannot be treated as a given. Just as there is no single, entirely authoritative performance of *Hamlet* that can stand for all, we can say there is no single authoritative instance of the running of a flight simulator that having been preserved means we require no other. In both cases, serious curatorial skill needs to be brought to bear to decide what exactly ought to be preserved. In the language of mainstream preservation activity, we need to establish, as far as it is possible to do so, the “significant properties” of the object of preservation, and capture those for future generations.

Although he was writing about artworks Grau could just as easily have been thinking about highly interactive computer systems, as he expressed well the essence of the problem: “The strength of material works of art, both past and present, lies principally in their function as illuminating and vibrant testimonies of the social memory of humankind. For only fixed artworks are able to preserve ideas and

As the complexity of digital objects increases, and the digital impinges on more and more of our lives, so does the need to see preservation and curation as an interdisciplinary team effort.

concepts enduringly and conserve the statements of individuals or an epoch. An open work, which is dependent on interaction with a contemporary audience, or its advanced variant that follows game theory—the work is postulated as a game and the observers, according to the ‘degrees of freedom’, as players—effectively means that images lose their capability to be historical memory and testimony. In its stead, there is a durable technical system as framework and transient, arbitrary, non-reproducible, and infinitely manipulable images. The work of art as a discrete object disappears.”¹

For anyone charged with the preservation of highly interactive computer systems, Grau’s view of matters, particularly as expressed in the quotation with which I began this column, is rather too bleak.

There are a number of approaches we can take that do much to ameliorate matters. First, we should, where appropriate, abandon the idea that there is a single authoritative ‘object’ of preservation, in favor of recognizing that sometimes the essence of an object, its most significant features, if you will, lie outside the materiality of the object. That is not to say the hardware and software do not matter—they clearly do—but there is, in addition, something intangible that must also be preserved and documented. There is no simple formula for determining which aspects of user “experience” need to be captured, as this will vary from situation to situation. This is an appraisal and documentation task of the greatest complexity and one which takes a great deal of skill to perform effectively. This might involve not only recording archetypical examples of interaction but potentially also noting how things might have been.

It may not always be possible to preserve complex interactive computer systems in such a way that they can be reproduced perfectly in the future. This might be due financial, technical or other reasons. In such cases, careful documentation of what are considered as the most important elements of the interaction represents a valuable resource for future generations. It is not possible to reproduce and experience at firsthand David Garrick’s first performance of *Hamlet*

in Dublin during the 1742 season, but it remains instructive and valuable to know that his performance(s) aroused such excitement, the Irish capital was said to have been gripped with “Garrick Fever.”

The corollary to abandoning the notion of a single authoritative ‘object’ of preservation, is to abandon the idea of there being a single authoritative act of preservation. As the complexity of digital objects increases, and the digital impinges on more and more aspects of our lives, so does the need to see preservation and curation as an interdisciplinary team effort.^c The make-up of preservation teams will vary depending on the object in question, but might involve social scientists, computer scientists, artists,^d and ethnographers, in addition to traditional curators and historians. Digital preservation involves much more than simply saving bits, or putting up a website. Dealing with complex interactive software highlights the role of nuanced, detailed and patient documentation in passing on to future generations a rich understanding not only of human-machine interaction, but also human-human interaction past and present.

As someone who spends most of his working life dealing with curators of one kind or another, I find it rather satisfying that as computer systems become ever more complex and nuanced, the prospects for complete automation of the preservation process are, at the bleeding edge of technology at least, getting further away, if anything. ■

c See for example <http://bit.ly/2eqc63D>.

d An interesting discussion on the topic of “archive” as it relates to feminism and technology, is available at <http://bit.ly/2enNRE9>

References

1. Grau, O. *Virtual Art: From Illusion to Immersion*. The MIT Press, 2003, 207.
2. Laurel, B. *Computers as Theatre*. 1st edition Addison-Wesley Longman Publishing Co., Boston, MA, USA, 1991.
3. Laurel, B. *Computers as Theatre*. 2nd edition. Addison-Wesley Longman Publishing Co., Boston, MA, 2014.
4. McDonough, J. A tangled web: Metadata and problems in game preservation. In J. Delve and D. Anderson, *Preserving Complex Digital Objects*, Facet Publishing, 2014.

David P. Anderson (cdpa@btinternet.com) is Professor of Digital Humanities at the Centre for Research & Development (Arts)/Cultural Informatics Research Group, University of Brighton, U.K.

Copyright held by author.

Broadening Participation Mentoring Female Assistant Professors Enhances Their Success

A randomized controlled trial validates many of the practices used to retain women in academia.

THE AMERICAN ECONOMIC Association Committee on the Status of Women in the Economics Profession (CSWEP) has tracked the representation of women at various ranks in the profession since the early 1970s. The statistics indicate a “leaky pipeline” from Ph.D. programs into tenured academic jobs. In 1998, for example, even though 30% of new Ph.D.’s in economics were earned by women, 26%, 14%, and 6% of the assistant, associate, and full professors of economics, respectively, were women in Ph.D.-granting departments in the U.S. Using multiple datasets including the NSF Survey of Earned Doctorates, researchers find a 14–21 percentage point gender gap in the probability of promotion to tenure in Economics, controlling for publications and citations.³ Further analyses point to a lack of research networks, role models, and mentors as potential causes of women’s failure to advance in economics.^{2,4} Although this column describes a study about women in the academic economics profession, the results of the study likely apply to women in the academic computer science profession as well. The current percentages of women in computer science graduating with Ph.D.’s and in



the three ranks follow a similar pattern, although less severe, to those in economics in 1998: new Ph.D.’s 25%; assistant 24%; associate 22%; and full 14%.⁵

Randomized Controlled Trials (RCT)

To expose junior women to role models (senior women in economics)

and to build peer networks of junior women in similar research areas, the National Science Foundation funded a pilot mentoring program in 1998, “CCOFFE: Creating Career Opportunities for Female Economists.” The two-day workshop brought eight senior women economists and 40 junior women economists from the top uni-

versities to participate in a series of career development panels, and to work cooperatively on each other's projects as teams.

While many junior women appreciated the opportunities offered through CCOFFE, it was difficult to evaluate its effects since the mentees were handpicked from top universities. To address the evaluation issue, CSWEP initiated a new national mentoring program in 2004 with the support of the National Science Foundation and the American Economic Association. The program is called CeMENT, which stands for CSWEP Mentoring Program. What differentiates this new mentoring program from its predecessor (CCOFFE) and many others is the random assignment of applicants into treatment and control groups in each cohort, thus creating a panel of randomized controlled trials. Starting in 2004, every other year, more than 80 female assistant professors in economics have applied to the program. Each cohort of applicants is divided into groups by research area, and then randomly assigned to treatment or control status within each group. A treatment group typically comprises of 4–5 mentees and 1–2 mentors who are senior women in that field. Random assignment enables researchers as well as funding agencies to evaluate the effectiveness of the mentoring program.

The treatment. Junior women randomized into the treatment participate in a two-day workshop in conjunction with the American Economic Association annual meetings. Prior to the workshop, participants circulate working papers within their group. During the workshop, each group meets and discusses the working papers. Each cohort also participates in a series of panel discussions on important career-development topics, such as publishing, networking, earning tenure, grant application, teaching, and work-life balance. The two-day face-to-face meeting creates opportunities for junior women to become familiar with each other's research and exposes them to senior female mentors and role models in their field.

After the two-day workshop, mentoring groups typically stay connected by sending each other monthly email

Random assignments enables researchers as well as funding agencies to evaluate the effectiveness of the mentoring program.

updates and meeting at professional conferences. Activity levels vary across groups. Members of successful groups read and comment on each other's working papers, grant proposals, and tenure statements, organize joint sessions at conferences, and host each other as seminar speakers at their institutions. Based on my own experience, first as a mentee in CCOFFE and then as a mentor in CeMENT, these groups communicate by email at least once a month to update each other on research progress, tasks accomplished in the past month, and goals for the following month. Members often seek and give advice. Monthly updates create both peer pressure and opportunities for direct or generalized reciprocity.

Communication and mentoring can become more intense around professional events like submission of a paper or grant application (when group members read and comment on the work), submission of the tenure dossier (when group members offer feedback and help to strategize), when individuals receive competing offers from other institutions (when group members help the individual identify the trade-offs involved in moving), or even life events like the birth of a child.

Monthly email messages typically stop when all members of a group receive tenure, although most group members continue to interact and help each other professionally after their mentoring group dissolves.

Interim results. Using data from the first three cohorts (2004, 2006, and 2008), researchers find significant

Calendar of Events

December 3–4

VRCAI '16: The 15th International Conference on Virtual-Reality Continuum and its Applications in Industry, Zhuhai, China, Sponsored: ACM/SIG, Email: myycail@ntu.edu.sg

December 11–14

WSC '16: Winter Simulation Conference, Arlington, VA, Contact: Todd Huschka, Email: huschka.todd@mayo.edu

December 12–15

CoNEXT '16: The 12th International Conference on emerging Networking EXperiments and Technologies, Irvine, CA, Sponsored: ACM/SIG, Contact: Athina Markopoulou, Email: amarko@stanfordalumni.org

December 12–16

Middleware '16: 17th International Middleware Conference, Trento, Italy, Sponsored: ACM/SIG, Contact: Gian Pietro Picco, Email: gianpietro.picco@unitn.it

January 2017

January 12–15

FOGA '17: Foundations of Genetic Algorithms XI, Copenhagen, Denmark, Sponsored: ACM/SIG, Contact: Carsten Witt, Email: cawi@dtu.dk

February

February 4–8

PPOPP '17: 21st ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, Austin, TX, Sponsored: ACM/SIG, Contact: Vivek Sarkar, Email: vsarkar@rice.edu

treatment effects in terms of the three main predictors of successful tenure in Economics, the total number of publications, the likelihood of top-tier publications, and the total number of federal grants in treated women relative to those in the control condition.¹

By three years after the intervention, participants in the treatment have two more publications than those in the control condition, and by the fifth year, those in the treatment group have three additional publications. When evaluating the magnitude of the treatment effects, one must bear in mind the differences in publication norms between economics and computer science. Unlike computer scientists who typically publish several short papers (5–10 pages) each year, an economist typically publishes 1–2 longer (40–50 pages) and more substantive papers each year. This thus represents an extra year of productivity for those who were mentored.

Another measure of career success in economics is top-tier publications, which refer to the top-five general-interest journals among more than 300 economic journals. At least one top-tier publication is often necessary for tenure promotion in many economics departments in research universities. Participants in the mentoring workshop were 20 percentage points more likely to have a top-tier publication after three years and 25 percentage points more likely after five years.

Lastly, obtaining a NSF or NIH grant is quite unusual among junior economists compared to computer scientists who are expected to fund their own Ph.D. students. By the fifth year, workshop participants had 0.4 more NSF or NIH grants than those in the control condition.¹

Lessons learned. Although junior women in the treatment group on average are more successful than their control counterparts, considerable heterogeneity in activity levels exists among different groups. We summarize several features that might have contributed to the success of the program and characterize some of the more active groups:

- Initial programwide co-located training workshop: this workshop creates opportunities for group members to get to know each other's work.

Women-to-women mentoring has been in the fabric of the computing field for a long time.

- Regular email updates, or Skype discussion sessions: successful groups typically designate a mentee who starts the email conversation on a certain day of each month, which generates a thread of updates.

- Face-to-face meetings at conferences that most group members attend: successful groups typically schedule their own work sessions to give feedback on each other's papers or grant proposals during a conference.

- Substantive interaction around research, setting norms and expectations of continued input as careers progressed: more-productive members set a standard for the less-productive ones.

- Procedural interaction around norms of the profession, explicitly communicating what is often unspoken.

- Mentorship and sponsorship by senior mentors: effective senior mentors often introduce junior women to other senior colleagues and initiate workshop or seminar invitations.

Impact of CeMENT in Academia

The interim results of CeMENT suggest that putting women in touch with other women with similar research interests across institutional boundaries is helpful, and that this type of mentoring program improves women's grant and publication records. As a result of the publication of the CeMENT evaluation, which establishes a benchmark for the effects of mentoring female assistant professors in the economic profession, similar mentoring programs have been established by various professional associations, such as the American Philosophical Association, the Chinese Women Economists Network,

the Japanese Women Economists Network, and Association for the Advancement of African Women Economists. Subsequent programs do not use RCT to select a treatment and control group. Instead all eligible applicants are treated.

Application to Academic Computer Science

Women-to-women mentoring has been in the fabric of the computing field for a long time. Systems, an email listserv, was founded by Anita Borg in 1987 and is still active with over 6,000 members today. Its members include academics, computing professionals, and students. Since 2010, the Computing Research Association Women (CRA-W) has held early career mentoring workshops for women and since 2013 mid-career workshops for women. Mentoring is a major feature of the annual Grace Hopper Celebration of Women in Computing, sponsored by the Anita Borg Institute. One focus of the Academic Alliance in the National Center for Women and Information Technology (NCWIT) is the retention of women in academia. Many of the lessons learned in CeMENT are already practiced in the computer science community. It is encouraging to see that many of the practices employed in the computer science community to improve academic retention have been validated in the CeMENT controlled study. 

References

1. Blau, F.D. et al. Can mentoring help female assistant professors? Interim results from a randomized trial. *American Economic Review* 110, 2 (May 2010), 348–352.
2. Blau, F.D., Ferber, M. and Winkler, A. *The Economics of Women, Men and Work*. 6th ed. Pearson/Prentice-Hall, Upper Saddle River, NJ, 2010.
3. Ginther, D.K. and Kahn, S. Women in economics: Moving up or falling off the academic career ladder? *Journal of Economic Perspectives* 18, 3 (2004), 193–214.
4. McDowell, J.M., Singell, L.D., and Slater, M. Two to tango? Gender differences in the joint decision to publish and coauthor. *Economic Inquiry* 44, (2006), 153–168.
5. Zweben, S. and Bizot, B. Taulbee Survey. *Computing Research News* 28, 5 (2016), 2–60.

Yan Chen (yanchen@umich.edu) is the Daniel Kahneman Collegiate Professor of Information at the University of Michigan, and Distinguished Visiting Professor of Economics at Tsinghua University.

The author thanks Rachel Croson, Janet Currie, Richard Ladner, and Laura Razzolini for their helpful comments.

Copyright held by author.

Viewpoint

Assessing Abstraction Skills

What makes a good question?

ABSTRACTION IS CONSIDERED to be a key skill underlying most activities in computer science (CS) and software engineering (SE).^{3,4} As a central concept, abstraction is taught and utilized in various guises in every CS and SE course: in requirements specifications, problem solving, and modeling through to programming and debugging. Given its importance, can one assess an individual's abstraction ability? Are abstraction skills assessable at all? If no, why? If yes, how? We decided to conduct a survey to investigate this topic. As far as we know, this is one of the first trials to address this challenge.¹

In 2007, Kramer proposed to develop a test that could assess abstraction skills in the context of CS and SE, and which would be more multifaceted than that used in psychometric tests.⁴ In this spirit, we decided to consult experts in CS and SE research and teaching concerning the suitability of various question patterns (or templates) for assessing abstraction skills. We specifically used patterns rather than specific questions in order not to limit the experts' line of thought, and, at the same time, to provide a template that each instructor could adjust and populate according to his or her needs.

Our data analysis reveals that expert instructors tend to agree about the suitability of a pattern for checking abstraction ability when it asks students to construct the abstraction rather than when applying abstraction principles. We explain this approach by the



constructionism learning theory of Papert and Harel.⁵

Survey Structure

We argue that directly asking experts the question “Are abstraction skills assessable?” would be unlikely to elicit a clear indication of how to go about constructing such an assessment tool. As a concrete proposal, we therefore developed a set of 10 patterns of questions and asked the experts to assess each pattern on a 1–10 scale accord-

ing to its suitability for assessing abstraction skills. In addition, for each pattern we presented three open questions regarding:

- ▶ What specific abstraction skills does Pattern X measure?
- ▶ Can you suggest an example that best fits Pattern X?
- ▶ Any additional comments on Pattern X?

In this way, we gained both a quantitative and qualitative perspective of the experts' opinion.

COMMUNICATIONS APPS

Access the latest issue, past issues, **BLOG@CACM**, News, and more.



Available for iPad, iPhone, and Android



Available for iOS, Android, and Windows

<http://cacm.acm.org/about-communications/mobile-apps>



Association for Computing Machinery

Table 1. Survey participants' demographic data.

Where do the experts teach?

Research university	23	74%
Teaching college	5	16%
High school	0	0%
Other*	3	10%
	31	

* One indicated that he or she is a post doc who taught in the past and one indicated that she or he works in a research institute.

Teaching Experience

0–5 years	2	6%
6–10 years	2	6%
11–15 years	3	10%
16–20 years	3	10%
More than 20 years	21	68%

Students' major/minor CS/SE taught by the experts (more than one option can be checked)

Computer science major students	25
Computer science minor students	10
Prospective computer science high school teachers	1
Software engineering students	18
Electrical engineering students	5
Other engineering students	9
Other	7

Courses the experts teach (more than one option can be checked)

Introductory computer science courses	18
Advanced computer science courses	20
Software engineering courses	21
Engineering courses	5
Other	6

Survey Participation

We approached two groups with similar characteristics: 20 experts were approached based on personal relationships and 11 responded to a message that we sent to a mailing list.^a $N = 31$ is a reasonable number for such a preliminary analysis. The demographic data presented in Table 1 indicates that the majority of the experts work at research universities, have many years of teaching experience of both introductory and advanced CS and SE courses to both CS major and SE students. This background clearly indicates they understand issues related to the challenge of teaching the topic of abstraction to good CS and SE students as well as assessing their abstraction skills.

a From the questionnaire sent to the first group (<http://bit.ly/2f9GWgr>), two questions were excluded in the subsequent questionnaire sent to the second group (<http://bit.ly/2eQJH7N>): a) Question 1 was a specific case of question 2; and b) Question 9 that the first group assessed lowest concerning its ability to assess abstraction skills. In addition, to the second questionnaire, we added the open question "Can you suggest an example which best fits Pattern X?"

Data Analysis: When Do CS and SE Instructors Agree and Disagree about the Appropriate Means to Assess Abstraction?

For each pattern, we calculated the average evaluation score and the standard deviation (see Table 2). Though an agreement was not reached with respect to any pattern, we can see that:

► The *highest agreement* (Average = 7.53; SD = 2; seven different scores) with respect to a pattern's suitability for measuring abstraction skills was reached for Pattern 5: Given a system representation, students are asked to give one representation that is more abstract than the given one and one representation that is less abstract than the given one.

► The *lowest agreement* (Average = 6, SD = 3.05; 10 different scores) with respect to a pattern's suitability for measuring abstraction skills was reached for Pattern 2: Given several representations of a specific system, students are asked to rank them according to their level of abstraction and to consider the purpose of each abstraction.

Disagreement: What was surprising in the data analysis was the vast range

Table 2. CS and SE instructors' assessment of the survey patterns.

	Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5	Pattern 6	Pattern 7	Pattern 8	Pattern 9	Pattern 10
Average	6.26316	6	6.464286	7.285714	7.535714	6.518519	6.785714	6.666667	4.789474	6.1923077
SD	2.90291	3.05505	2.545616	2.565543	2.008909	2.517177	2.424522	2.557042	2.323287	3.2163223
N	19	28	28	28	28	28	28	27	19	27
# of scores	10	10	10	8	7	8	9	9	8	9

of scores and disagreement exhibited with respect to almost all patterns concerning their suitability to assess abstraction skills. For example, patterns 1–3 received all scores. These three patterns provided the students with a set of object representations and asked the students to categorize them according to the student's choice of abstraction. In other words, in these patterns, students were *not* asked to create/construct an object while considering abstraction.

Agreement: As mentioned, Pattern 5 has the highest average score with the lowest distribution. In this pattern and in another one (Pattern 4, which also was assessed high), the students were asked to *create some system representations by referring to abstraction*. We may conclude that instructors tend to agree that constructing abstraction is a suitable way to assess abstraction skills. One of our experts explains: “This is a nice concrete type of pattern question. Looks at abstraction generation skill but in a much more constrained setting.”

Psychological Explanation

Theoretically, we explain the experts' ranking by the theory of constructionism, inspired by the constructivist theory. According to constructivism, learners construct mental models in order to understand the world around them. Constructivism advocates student-centered, discovery learning. One outcome of the constructivism framework is that one has to be actively involved with the material he or she learns.² Accordingly, “The simplest definition of constructionism evokes the idea of learning by making ⁵

The scores presented in Table 2 indicate that the experts, who answered our questionnaire, saw the potential contribution of Pattern 5 to students' understanding. From a constructionism perspective, this can be explained by the fact that Pattern 5 asks students

to *construct* two representations—a less and a more abstract one than a given one, rather than to categorize, compare, or explain some aspect of a given object while referring to abstraction. Here are several instructors' answers to the question: “What specific abstraction skills does Pattern 5 measure?” which reflect this explanation [italics were added by the authors]:

- Ability to *invent* new abstractions at different levels of abstraction.

- Ability to “*do*” abstraction—to throw away detail, while keeping the essential structure. Also, ability to “*do*” refinement—to add detail.

- Ability to *CREATE* (rather than just assess), from an external knowledge source, different levels of abstraction.

- The ability to *maneuver* between abstraction levels, as needed, by adding/removing details. Also, to identify the entities that comprise an abstraction level, and, constructively, detailing or abstracting them in order to achieve the desired level of abstraction.

- ... the ability to *devise* abstractions at various levels.

Conclusion

The question: “Are abstraction skills assessable?” is part of the discussion on teaching soft CS concepts. The experts' responses, which reflect a constructionism perspective, indicate that they value students' active engagement and creativity in the learning process. One of the instructors suggested that Pattern 5 “... seems too open ended to me for a test question—although great for a studio-style discussion.”

Based on the experts' answer, this Viewpoint has focused on distinguishing between the creative construction of abstractions as against the passive assessment of abstraction level. Another particularly crucial aspect in abstraction is that of purpose and the consideration of fit-for-purpose in activities such as modeling and problem solving. This includes both the creation of new abstract

formalisms and languages to express abstractions—as is created by researchers and language designers, and the use of these to create specific abstractions for specific applications. We believe that abstraction purpose needs to be more directly and explicitly considered and discussed by instructors, students, and practitioners when reflecting on their use of abstraction. In this spirit, one of the experts who responded to the survey suggested to highlight “[...] what the system needs to achieve and what are the criteria for better solution. It will improve the quality of the answers and will provide wider range of good answers.” Another one added: “I would recommend to couple it with a specific purpose to guide what kind of aspects that can be left out or needs to be included.”

We see this Viewpoint as a starting point of the discussion of this topic and invite the ACM community to join and contribute to the discussion. Such a discussion is especially relevant today when a variety of teaching methods and frameworks are available for students who seek the highest return for their effort investment. 

References

1. Bennedsen, J.B. and Caspersen, M.E. Abstraction ability as an indicator of success for learning object-oriented programming? *SIGCSE Bulletin inroads* 38, 2 (2006), 39–43.
2. Davis, R., Maher, C., and Noddings, N. Introduction: Constructivist views on the teaching and learning of mathematics. In R. Davis, C. Maher, and N. Noddings, Eds. *Constructivist Views on the Teaching and Learning of Mathematics*. National Council of Teachers of Mathematics, Reston, VA, 1990, 7–18.
3. Denning, P.J., et al. Computing as a discipline. *Commun. ACM* 32, 1 (Jan. 1989), 9–23.
4. Kramer, J. Is abstraction the key to computing? *Commun. ACM* 50, 4 (Apr. 2007), 37–42.
5. Papert, S. and Harel, I. *Constructionism*. Ablex Publishing Corporation, 1991.

Orith Hazzan (oritha@technion.ac.il) is a Professor of the Faculty of Education in Science and Technology at Technion—Israel Institute of Technology.

Jeff Kramer (j.kramer@imperial.ac.uk) is Professor of Distributed Computing in the Department of Computing at Imperial College London.

We would like to thank the experts who answered our survey, invested the time, and shared their thoughts with us.

Copyright held by authors.

Viewpoint

(Mis)Managing Parallel Computing Research through EU Project Funding

Considering the possible negative long-term effects of EU projects on European scientific culture.

MUCH HIGH-PERFORMANCE AND parallel computing research in Europe is funded by the European Union through the so-called “EU projects.” These projects are highly competitive and often considered highly attractive and prestigious because of their scientific quality, involvement of outstanding researchers, contributions to the education of new Ph.D.’s for academia and the European workforce, and opportunities for collaboration with industry in Europe.

I will argue a dissenting opinion that EU projects in this field may not serve their stated purposes of advancing science and education, and are not effective in subsidizing European industry. On the contrary, they are a time-consuming effort with often mediocre outcome that tend to perpetrate a narrow, self-referential, and trivialized view of the scientific process with possibly detrimental long-term effects on the scientific culture in Europe and the quality of the workforce. The problems are not exclusive to the EU funding system, and also not to parallel computing research, but to a large extent caused by excessive managerial overheads and the specific stated requirements and unstated expectations in the EU



project conception. This Viewpoint from the perspective of an academic researcher is based on long, personal experience with EU projects in parallel computing and related areas, in different roles, and with different degrees of “success.” Since these reflections are not meant as criticism of specific projects or activities, I will not cite concrete

project figures or outcomes, but contend that my claims can be backed up by publicly available material.

EU Projects: “Novel Ideas for Radically New Technologies”

EU projects are set up in response to regular calls by the European Commission that outline the issues to be

investigated and the kind of projects that may receive what kind of funding under which conditions. The calls ask for novelty, excellence, radicalism and breakthroughs^a and also outline the composition of eligible consortia that will consist of “leading experts” from European university research groups and (often, but not in all call types) European industry or business, preferably small and medium-sized enterprises (so-called SMEs). Funded projects typically run for three years, and are expected to contribute to education by producing excellent, new Ph.D.’s for the European workforce. Exploitation and dissemination are of the utmost importance to demonstrate uptake and impact. Progress is documented in extensive “deliverables” and “prototypes,” and constantly monitored and assessed by EU “experts.” Deviation from work plan is an exception and needs to be negotiated. In not-so-large EU projects, five to 10 groups can be involved, each with a number of people that meet as a whole or in smaller groups several times a year. As part of the dissemination activities it is important for all consortium members to show regular presence at various associated, often self-initiated workshops and conferences, including meetings and workshops of other projects; high publication output is encouraged.

There are several points to stress. First, research areas are determined extrinsically by the European Commission (in consultation with “leading experts”) which is in itself not an intellectually contributing peer of the scientific (parallel computing) community and therefore only in an indirect, not independent position to define scientific research agenda. Second, projects are to be tightly managed, documented, and overseen by the appointed Commission experts. Third, project proposals themselves, which after approval turn into detailed “Descriptions Of Work” (DOW: work-packages, which deliverables when, “milestones,” month-by-month work, exploitation and dissemination plans),

It is doubtful whether truly groundbreaking research can be planned as envisioned by the Commission managers and thrive under the EU project conditions.

are themselves considerable pieces of work (easily 50 or more pages) taking weeks of intense effort to prepare. Fourth, dissemination of results and industrial exploitation are key factors in successful proposals and projects.

Now, it is doubtful whether truly groundbreaking research can be planned as envisaged by the Commission managers and thrive under the EU project conditions; but “normal science” is rarely called for. The effort in proposal writing is wasted (unless recycled) if a project is not funded, which lead proposals to promise what is called for, namely a meticulously documented, steady flow of high-quality breakthroughs that can readily be taken up by industry. And although projects can have large, total volume in terms of money and personnel, the managerial overhead and enforced geographical and thematic spread among consortium members in reality means that time and personnel per group is often too small to pursue serious and realistic research. Fortunately, since projects define their own “verifiable success criteria” and “key performance indicators,” success can nevertheless be guaranteed, either because results to be produced were safely known in advance, and therefore per definition not groundbreaking, or because goals, despite the inflated language, are set so low or made so vague that failure is impossible by design, as long as deliverables are delivered, milestones met, prototypes produced, workshops and dissemination activities organized, and consortium management business attended to. Whether proposers actually believe what they propose, regarding

both goals and execution, is an open question. However, proposal-writing skill is certainly cultivated by the EU system, and those in command of the language needed for successful proposals can have a welcome role as partners, or consultants in the proposal phase.

Since running an EU project and preparing for the next one takes time, the leading experts who formulated the project in the first place may not be factually contributing to the project in any major way, nor be supervising inexperienced junior personnel carrying out the projected work. Instead, young researchers in need of guidance for their Ph.D. may find themselves left alone with the considerable responsibility of advancing the project from which they are paid, including managerial duties and the writing of deliverables whose sole merit and purpose is to fulfill what was promised in the DOW. From a scientific point of view, EU project deliverables are often of a sub-par standard with little of value to the student or to the scientific community. Since the project must be successful, there is severe temptation to willingly or innocently sell trivialities as breakthroughs, effectively installing and reinforcing a belief in the student and the interested public (such as via newspapers) that trivialities are major breakthroughs. The same can be observed for the prototypes that are important as “proofs-of-concept” and means to “transfer knowledge” from research to industry, which are often of low quality and designed only to survive a single, superficial Commission expert review. Thus, EU projects may seem not to contribute to develop scientific attitude, conduct and practice in the Ph.D. student. A more experienced postdoctoral researcher may likewise be left to herself with the day-to-day running of EU projects without having enough time to concentrate on the research and lasting results that will be needed to eventually secure a permanent position.

The role of the industrial partner in an EU project is sometimes that of a peer, sometimes that of an advisor, sometimes of a provider (in the case of high-performance computing: of access to technologies), but often that of a politically necessary appendix. Especially SMEs can be as dependent on the EU funding as small university groups, but may not have the capabilities and

^a Bertrand Meyer addressed this paradigm-shift mania and its effects on scientific attitude in his “Long Live Incremental Research” blog posting, see <http://bit.ly/2dFOV75>; prose from recent FET Open calls shows how up to date this still is.

resources to contribute significantly to research-oriented deliverables, and often do not, thereby leaving more of the project burden on the academic partners. On the other hand, SMEs sometimes expect concrete product development to be carried out in the project for which the academic partners typically have neither the skill nor the manpower. Peculiarly, large (non-European) multinational companies can receive subsidy as EU project partners. It is not always obvious that this contributes to the projects, neither why these companies should be recipients of EU funding.

Exploitation and dissemination being important factors in EU projects lead to a wealth of activities serving to document successful “take-up”: publications, workshops and meetings at conferences, interproject workshops, user training, contributions to standardization bodies, and so forth. The primary purpose of many of these activities is to meet the dissemination plans, and has led to a proliferation of workshops presenting and publishing project-relevant trivialities. Apart from the time this consumes, the apparent authority of a workshop masquerading as a scientific event at a well-established conference may reinforce docility and low standards in the naive Ph.D., and appall and deter the more observant one. As harmless as many of these time-consuming activities may be, premature or incompetent standardization is not, and can bind considerable resources to either comply with or fight a bad standard, thus hampering both scientific and industrial progress.

Not-So Novel Ideas for Radically Better Funding

European research groups in high-performance and parallel computing are materially dependent on external project funding, and EU projects give a comfortable headroom with the attractive possibility to collaborate closely with other European research groups. However, the drain by the management of projects and the effort required for the overly detailed project proposals, are stunting. It would be liberating and productive to get rid of both.

First, since there is no proof that strict management procedures actually lead to better and more lasting scientific outcome, why not drop the Com-

It would be more transparent and less ambiguous if industrial participation in EU projects was not directly funded through the projects.

mission involvement entirely from the project execution (as in some national schemes), and trust responsible projects to select adequate means for managing themselves, tracking progress by short summaries and actual outcome instead of excessive technical deliverables? Likewise, the overemphasis on dissemination and take-up activities should be moderated; well-run projects are themselves capable of deciding when and what makes sense.

Second, it is an equally unproven assumption that scientific breakthroughs can be predicted by the promises made in project proposals. It would save much painful effort and dear time to ask instead for short, simple, matter-of-factly, problem-oriented funding applications without inflated promises, without excessively detailed work plans and without politically motivated constraints on consortium composition. With clear, and not too restrictive criteria for funding eligibility that can realistically be checked by the available Commission experts, it would perhaps even be acceptable to acknowledge the fact—given the current oversubscription and low success rates—that funding decisions among eligible proposals are . . . random. Criteria could emphasize previous scientific results of the involved researchers, along the lines sketched by Thorup,¹ and in general focus more on actual instead of projected outcome.² This would imply that proposals and projects to a larger extent be driven by individual groups with a clear vision and consortium members selected by their specific project-relevant expertise. It would make long-term

scientific, and perhaps even commercial sense to make it possible and easy to extend projects that produce actual, great results or valuable prototypes; currently, such extensions are difficult and often depending on a fitting, next call. Since the three years that EU projects typically run is in many cases too short to complete a Ph.D. thesis and gain the necessary breadth and depth beyond the horizon of the subsidizing EU project, means of after-project follow-up funding are especially important for Ph.D. students; the practice of putting the student on whatever next project may be available is hazardous.

EU projects in high-performance and parallel computing sit between many chairs, and in many cases industrial involvement makes a lot of sense. However, it seems confusing at best to (ab)use scientific research projects to subsidize European (SME) industry. Can't this be done much more effectively by direct means? In any case, it would be more transparent and less ambiguous if industrial participation in EU projects was not directly funded through the projects. Strong projects would be interesting enough that industry would want to participate out of its own volition, which in particular should be the case for large businesses.

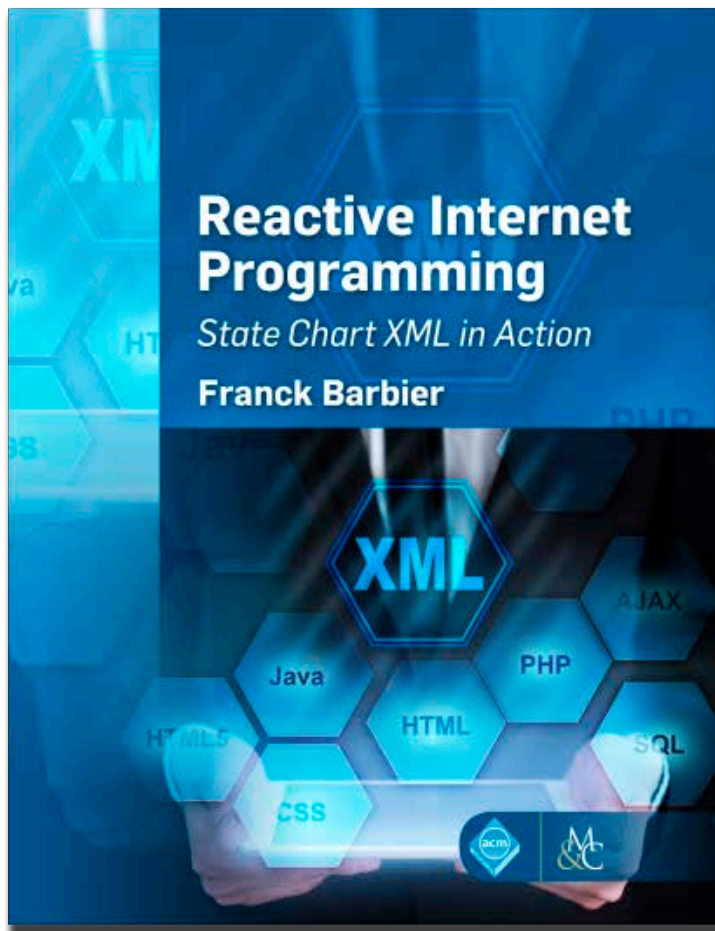
Finally, it seems that the role of European-level funding should be to support what cannot be funded nationally, which on the other hand implies that strong national funding of research groups and projects is still needed. In addition to existing, large-scale European high-performance computing infrastructure with personnel for support and development, more possibilities for travel grants and lightweight working groups to foster contacts between European research groups, and more EU scholarships for Ph.D. students doing their Ph.D. at different European universities would also be welcome additions. 

References

1. Thorup, M. Funding successful research. *Commun. ACM* 56, 3 (Mar. 2013), 38–39.
2. Vardi, M.Y. Incentivizing quality and impact in computing research. *Commun. ACM* 58, 5 (May 2015), 5.

Jesper Larsson Träff (traff@par.tuwien.ac.at) is Professor for Parallel Computing in the Faculty of Informatics at TU Wien (Vienna University of Technology), Austria.

Copyright held by author.



You need effective means to put reactive programming into practice.

THEY'RE
RIGHT HERE

Is Internet software so different from “ordinary” software? This book practically answers this question through the presentation of a software design method based on the State Chart XML W3C standard along with Java. Web enterprise, Internet-of-Things, and Android applications, in particular, are seamlessly specified and implemented from “executable models.”

Internet software puts forward the idea of event-driven or reactive programming, as pointed out in Bonér et al.’s “Reactive Manifesto”. It tells us that reactivity is a must. However, beyond concepts, software engineers require effective means with which to put reactive programming into practice. *Reactive Internet Programming* outlines and explains such means.

The lack of professional examples in the literature that illustrate how reactive software should be shaped can be quite frustrating. Therefore, this book helps to fill in that gap by providing in-depth professional case studies that contain comprehensive details and meaningful alternatives. Furthermore, these case studies can be downloaded for further investigation.

Internet software requires higher adaptation, at run time in particular. After reading *Reactive Internet Programming*, you will be ready to enter the forthcoming Internet era.



<http://books.acm.org>
<http://www.morganclaypoolpublishers.com/xml>

Article development led by **acmqueue**
queue.acm.org

Applying functional programming principles to distributed computing projects.

BY MARIUS ERIKSEN

Functional at Scale

MODERN SERVER SOFTWARE is demanding to develop and operate: It must be available at all times and in all locations; it must reply within milliseconds to user requests; it must respond quickly to capacity demands; it must process a lot of data and even more traffic; it must adapt quickly to changing product needs; and, in many cases, it must accommodate a large engineering organization, its many engineers the proverbial cooks in a big, messy kitchen.

What's more, the best computers for these applications—whether you call them clouds, datacenters, or warehouse computers—are really bad. They are complex and unreliable, and prone to partial failures. They have asynchronous interconnects and deep memory hierarchies, and leave a lot of room for operator error.¹

Cloud computing thus forces us to confront the full complexity of distributed computing, where seemingly simple problems require complicated solutions. While much of this complexity is inherent—the very nature of the problem precludes simpler solutions—much of it is also incidental, a simple consequence of using tools unfit for the purpose.

At Twitter, we use ideas from functional programming to tackle many of the complexities of modern server software, primarily through the use of higher-order functions and effects. *Higher-order functions*, or those that return other functions, let us combine simpler functions to define more complex ones, building up application functionality in a piecemeal fashion. *Effects* are values that represent some side-effecting operation; they are used in conjunction with higher-order functions to build complex effects from simpler ones.⁷

Higher-order functions and effects help build scalable software in two ways: First, building complex software from simple parts makes it easy to understand, test, reuse, and replace individual components; second, effects make side-effecting operations tractable, promoting modularity and good separation of concerns.

This article explores three specific abstractions that follow this style of functional programming: *futures* are effects that represent asynchronous operations; *services* are functions that represent service boundaries; and *filters* are functions that encapsulate application-independent behavior. In turn, higher-order functions provide the glue used to combine these to create complex systems from simple parts.

Futures, services, and filters are thus combined to build server software in a piecemeal fashion. They let programmers build up complex software while preserving their ability to reason about the correctness of its constituent parts.

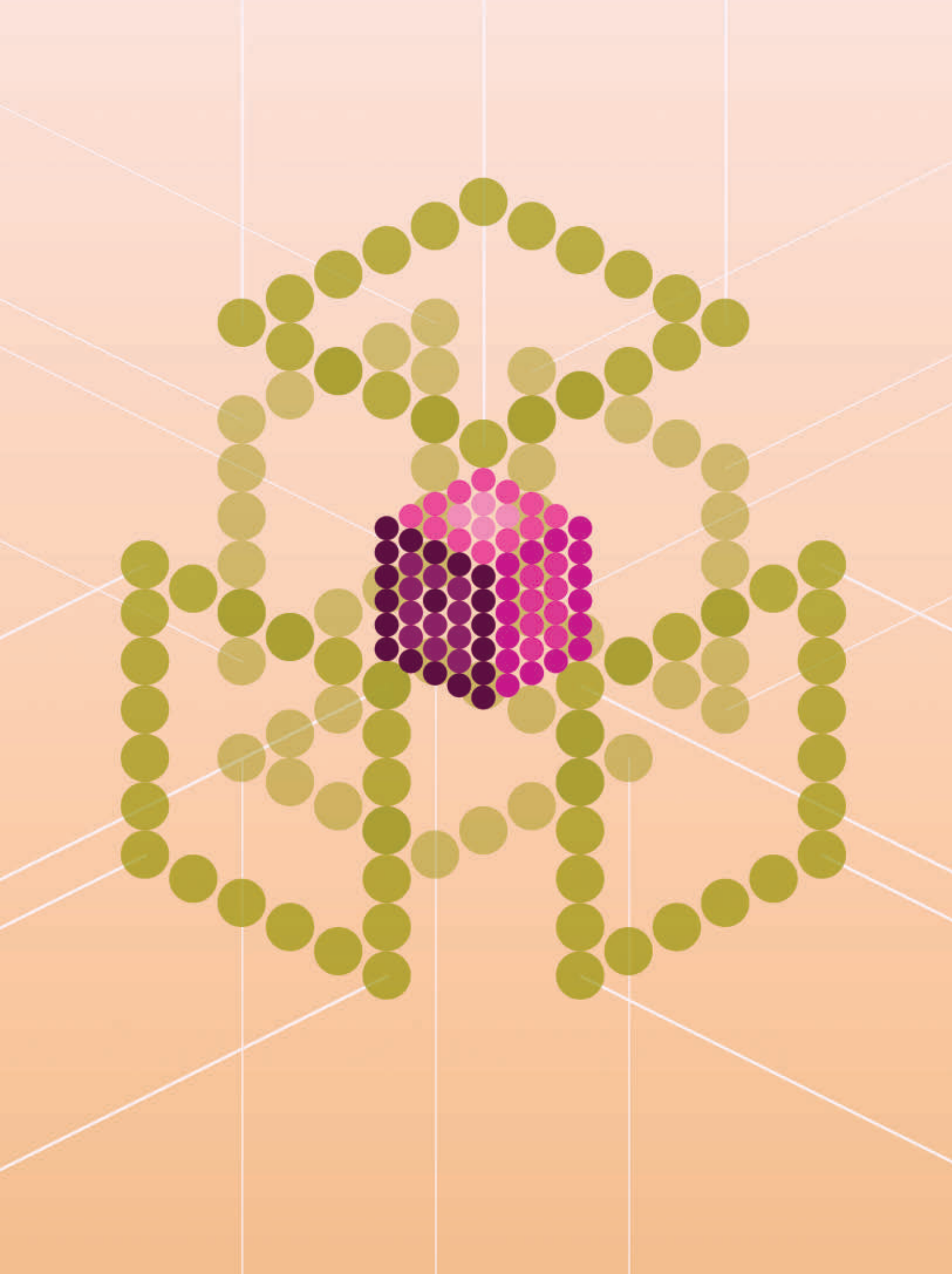
By consistently applying these principles, programmers can construct systems that are at once simpler, more flexible, and performant.

Concurrent Programming with Futures

"Concurrent programs wait faster."

—Tony Hoare

Concurrency is a central topic in server-software development.¹² Two sources of concurrency prevail in this



type of software. First, scale implies concurrency. For example, a search engine may split its index into many small pieces (shards) so the entire corpus can fit in main memory. To satisfy queries efficiently, all shards must be queried concurrently. Second, communication between servers is asynchronous and must be handled concurrently for efficiency and safety.

Concurrent programming is traditionally approached by employing threads and locks³—threads furnish the programmer with concurrent threads of execution, while locks coordinate the sharing of (mutable) data across multiple threads.

In practice, threads and locks are notoriously difficult to get right.⁹ They are hard to reason about, and they are a stubborn cause of nasty bugs. What's more, they are difficult to compose: you cannot safely and arbitrarily combine a set of threads and locks to construct new functionality. Their semantics of computation are wrapped up in the mechanics of managing concurrency.

At Twitter, we instead structure concurrent programs around futures. A future is an effect that represents the result of an asynchronous operation. It's a type of reference cell that can be in one of three states: *incomplete*, when the future has not yet taken on a value; *completed with a value*, when the future holds the result of a successful operation; and *completed with a failure*, when the operation has failed. Futures can undergo at most one state transition: from the incomplete state to either the success or failure state.

In the following example, using Scala, the future count represents the result of an integer-valued operation. We respond to the future's completion directly: the block of code after `respond` is a callback that is invoked when the future has completed. (As you will see shortly, we rarely respond directly to future completions in this way.)

```
val count: Future[Int] = getCount()
count.respond {
  case Return(value) =>
    println(s"The count was $value")
  case Throw(exc) =>
    println(s"getCount failed with $exc")
}
```

Futures represent just about every asynchronous operation in Twitter



Futures, services, and filters are combined to build server software in a piecemeal fashion. They let programmers build up complex software while preserving their ability to reason about the correctness of its constituent parts.



systems: RPC (remote procedure call), timeout, reading a file from disk, receiving the next event from an event stream.

With the help of a set of higher-order functions (called *combinators*), futures can be combined freely to express more complex operations. These combinations usually fall into one of two composition categories: sequential or concurrent.

Sequential composition permits defining a future as a function of another, such that the two are executed sequentially. This is useful where data dependency exists between two operations: the result of the first future is needed to compute the second future. For example, when a user sends a Tweet, we first need to see if that user is within the hourly rate limits before writing the Tweet to a database. In the Figure 1 example, the future `done` represents this composite operation. (For historical reasons, the sequencing combinator is called `flatMap`.)

This also shows how failures are expressed in futures: `Future.exception` returns a future that has already completed in a failure state. In the case of rate limiting, `done` becomes a failed future (with the exception `RateLimitingError`). Failure short-circuits any further composition: if, in the previous example, the future returned by `isRateLimited(user)` fails, then `done` is immediately failed; the closure passed to `flatMap` is not run.

Another set of combinators defines concurrent composition, allowing multiple futures to be combined when no data dependencies exist among them. Concurrent combinators turn a list of futures into a future of a list of values. For example, you may have a list of futures representing RPCs to all the shards of a search index. The concurrent combinator `collect` turns this list of futures into a future of the list of results.

```
val results: List[Future[String]] = ...
val all: Future[List[String]] =
  Future.collect(results)
```

Independent futures are executed concurrently by default; execution is sequenced only where data dependencies exist.

Future combinators never modify the underlying future; instead, they return a *new* future that represents

the composite operation. This is an important tool for reasoning: composite operations do not change the behavior of their constituent parts. Indeed, a single future can be used and reused in many different compositions.

Let's modify the earlier `getCount` example to see how composition allows building complex behavior piecemeal. In distributed systems, it is often preferable to degrade gracefully (for example, where a default value or guess may exist) than to fail an entire operation.⁸ This can be implemented by using a timeout for the `getCount` operation, and, upon failure, returning a default value of 0. This behavior can be expressed as a compound operation among different futures. Specifically, you want the future that represents the winner of a timeout-with-default and the `getCount` operation (see Figure 2).

The future `finalCount` now represents the composite operation as described. This example has a number of notable features. First, we have created a composite operation using simple, underlying parts—futures and functions. Second, the constituent futures preserve their semantics under composition—their behavior does not change, and they may be used in multiple different compositions. Third, nothing has been said about the mechanics of execution; instead, the composite computation is expressed as the combination of a number of underlying parts. No threads were explicitly created, nor was there any explicit communication between them—it is all implied by data dependencies.

This neatly illustrates how futures liberate the application's semantics (what is computed) from its mechanics (how it is computed). The programmer composes concurrent operations but need not specify how they are scheduled or how values are communicated. This is a good separation of concerns: application logic is not entangled with the minutiae of runtime concerns.

Futures can sometimes free programmers from having to use locks. Where data dependencies are witnessed by composition of futures, the implementation is responsible for concurrency control. Put another way, futures can be composed into a dependency graph that is executed in the

Figure 1. Admit request only when the user is within their rate limits.

```
def isRateLimited(user: Long): Future[Boolean] = ...
def writeTweet(user: Long, tweet: String): Future[Unit] = ...

val user = 12L
val tweet: String = "just setting up my twitter"

val done: Future[Unit] =
  isRateLimited(user).flatMap {
    case true => Future.exception(new RateLimitingError)
    case false => writeTweet(user, tweet)
  }
```

Figure 2. Degrade with a default value after timeout.

```
val count: Future[Int] = getCount()
// Future.sleep(x) returns a future which completes
// after the given amount of time.
val timeout: Future[Unit] = Future.sleep(5.seconds)

// The map method on Future constructs a new Future
// which, after successful completion, applies the given
// function to the result. It returns a new Future
// representing this composite operation. In this case,
// we simply return a default value of 0 after the timeout.
val default: Future[Int] = timeout.map(unit => 0)

// Select composes two Futures, returning a new
// Future which represents the first future to complete.
val finalCount: Future[Int] = Future.select(count, default)
```

manner of dataflow programming.¹¹ While we need to resort to explicit concurrency control from time to time, a large set of common use cases are handled directly by the use of futures.

At Twitter, we implement the machinery required to make concurrent execution with futures work in our open-source `Finagle`^{4,5} and `Util`⁶ libraries. These take care of mapping execution onto OS threads through a pluggable scheduler mechanism. Some teams at Twitter have used this capacity to construct scheduling strategies that better match their problem domain and its attendant trade offs. We have also used this capability to add features such as maintaining runtime statistics and Dapper-style RPC tracing to our systems, without changing any existing APIs or modifying existing user code.

Programming with Services and Filters

"We should have some ways of coupling programs like garden hose—screw in an-

other segment when it becomes necessary to massage data in another way. This is the way of I/O also."

—Doug McIlroy

Modern server software abounds with essential complexity—that which is inherent to the problem and cannot be avoided—as well as complexity of the more self-inflicted variety (did we really need that product feature)? Anyhow, since we can't seem to keep our applications simple, we must instead cope with their complexities.

Thus, the modern software engineering practice is centered on how to contain and manage complexity—to package it up in ways that allow us to reason about our application's behavior. In this endeavor the goal is to balance simplicity and clarity with reusability and modularity.

What's more, server software must account for the realities of distributed systems. For example, a search engine might simply omit results from a failing index shard so that it can return a partial result rather than failing the query in

Figure 3. A simple dictionary service.

```
// This is the dictionary data, encoded in a
// string-to-string map.
val map: Map[String, String] = Map(...)

val dictionary: Service[String, String] =
  // Construct a new service from an
  // anonymous function, invoked
  // for every request issued.
  Service(key: String => {
    // Return the lookup results. (Map.get
    // returns an optional string.) Future.value
    // creates a Future that is already
    // satisfied with the given value.
    if (map.contains(key))
      Future.value(map(key))
    else
      Future.value(null)
  })
```

Figure 4. Scatter-gather across multiple dictionaries.

```
def multipleLookup(services: Seq[Service[String, String]])
  : Service[String, String] =
  Service(key => {
    val results: Seq[Future[String]] =
      services.map(_.apply(key))
    val all: Future[Seq[String]] =
      Future.collect(results)
    all.map { results: Seq[String] =>
      results.indexWhere(_ != null) match {
        case -1 => null
        case i => results(i)
      }
    }
  })
```

its entirety (as seen in the `getCount` example). In this case, the user would not usually be able to tell the difference—the search engine is still useful without a complete set of results. Applying application-level knowledge in this way is often essential to creating a truly resilient application.

Distributed applications are therefore organized around *services* and *filters*. Services represent RPC service endpoints. They are a function of type `A => Future[R]` (that is, a function that takes a single argument of type `A` and returns an `R`-typed future). These functions are asynchronous: they return immediately; deferred actions are captured by the returned future. Services are symmetric: RPC clients call services to dispatch RPC calls; servers implement services to handle them.

The example in Figure 3 defines a simple dictionary service that looks up a

key in a map, which is stored in memory. The service takes a `String` argument and returns a `String` value, or else null if the key is not found.

Note that the service, `dictionary`, is again a value like any other. Once defined, it can be made available for dispatch over the network through a favorite RPC protocol. On the server, services are exported as

```
Rpc.serve(dictionary, ":8080")
```

while clients can bind and call remote instances of the service:

```
val service: Service[String, String]
=
  Rpc.bind("server:8080")
val result: Future[String] =
  service("key")
```

Observe the symmetry between cli-

ent and server: both are conversing in terms of services, which are location transparent; it is an implementation detail that one is implemented locally while the other is bound remotely.

Services are easily composed using ordinary functional composition. For example, the service in Figure 4 performs a scatter-gather lookup across multiple dictionary services. The dictionary can therefore be distributed over multiple shards.

This example has a lot going on. First, each underlying service is called with the desired key, obtaining a sequence of results—all futures. `Future.collect` composes futures concurrently, turning a sequence of futures into a single one containing a sequence of the values of the (successful) completion of the constituent futures. The code then looks for the first non-null result and returns it.

Filters are also asynchronous functions that are combined with services to modify their behavior. Their type is `(A, Service[A, R]) => Future[R]` (that is, a function with two arguments: a value of type `A` and a service `Service[A, R]`); the filter then returns a future of that service's return type, `R`. The filter's type indicates that it is responsible for satisfying a request, given a service. It can thus modify both how the request is dispatched to the service (for example, it can modify it) and how it is returned (for example, add a timeout to the returned future).

Filters are used to implement functionality such as timeouts and retries, failure-handling strategies, and authentication. Filters may be combined to form compound filters—for example, a filter that implements retry logic can be combined with a filter implementing timeouts. Finally, filters are combined with a service to create a new service with modified behavior.

Building on the previous example, Figure 5 illustrates a filter that downgrades failures from lookup services to null. This kind of behavior is often useful when constructing resilient services—it's sometimes better to return partial results than to fail altogether.

The dictionary lookup service `resilientService` implements scatter-gather lookup in a resilient fashion. The functionality was built from individual well-defined components that are com-

posed together to create a service that behaves in a desirable way.

As with futures, combining filters and services does not change the underlying, constituent components. It creates a new service with new behavior but does not change the meaning of either underlying filter or service. This again enhances reasoning since the constituent components stand on their own; we need not reason about their interactions after they are combined.

Futures, services, and filters form the foundation upon which server software is built at Twitter. Together they support both modularity and reuse: we define applications and behaviors independently, composing them together as required. While their application is pervasive, two examples nicely illustrate their power. First, we implemented an RPC tracing system à la Dapper¹⁰ as a set of filters, requiring no changes to application code. Second, we implemented backup requests² as a small, self-contained filter.

Conclusion

“Don’t tie the hands of the implementer.”

—Martin Rinard

Functional programming promotes thinking about building complex behavior out of simple parts, using higher-order functions and effects to glue them together. At Twitter we have applied this line of thinking to distributed computing, structuring systems around a set of core abstractions that express asynchrony through effects and that are composable. This allows building complex systems from components with simple semantics that, preserved under composition, makes it easier to reason about the system as a whole.

This approach leads to simpler and more modular systems. These systems promote a good separation of concerns and enhance flexibility, while at the same time permitting efficient implementations.

Functional programming has thus furnished essential tools for managing the complexity that is inherent in mod-

ern software—untying the hands of the implementer.

Acknowledgments

Thanks to Jake Donham, Erik Meijer, Mark Compton, Terry Coatta, Steve Jensen, Kevin Oliver, Ruben Oanta, and Oliver Gould for their guidance on earlier drafts of this article. Early versions of the abstractions discussed here were designed together with Nick Kallen; numerous people at Twitter and in the open source community have since worked to improve, expand, and solidify them. 

Related articles on queue.acm.org

Scalable SQL

Michael Rys

<http://queue.acm.org/detail.cfm?id=1971597>

Evolution and Practice: Low-latency Distributed Applications in Finance

Andrew Brook

<http://queue.acm.org/detail.cfm?id=2770868>

Distributed Development Lessons Learned

Michael Turnlund

<http://queue.acm.org/detail.cfm?id=966801>

Figure 5. Scatter-gather lookup with resiliency measures.

```
val downgradeToNull: Filter[String, String] =
  Filter((key, service) => {
    service.apply(key).transform {
      case Return(value) => value
      case Throw(exception) => null
    }
  })
```

Another filter short-circuits requests for useless keys (these are known as stopwords in search engines), so that they don't inflict needless load on the underlying service:

```
val stopwords: Set[String] ...
val stopwordFilter: Filter[String, String] =
  Filter((key, service) => {
    if (stopwords.contains(key))
      Future.value(null)
    else
      service.apply(key)
  })
```

Finally, the two filters are combined and then applied to all of the services.

```
val resiliencyFilter: Filter[String, String] =
  stopwordFilter.andThen(downgradeToNull)

def resilientMultipleLookup(services: Service[String, String]) = {
  val resilientServices =
    services.map { service => resiliencyFilter.andThen(service) }
  multipleLookup(resilientServices)
}

val resilientService: Service[String, String] =
  resilientMultipleLookup(
    Rpc.bind("server1:8080"),
    Rpc.bind("server2:8080"),
    ...
  )
```

References

- Barroso, L. A., Clidaras, J. and Hölzle, U. The datacenter as a computer: an introduction to the design of warehouse-scale machines. *Synthesis Lectures on Computer Architecture* 8, 3 (2013), 1–154.
- Dean, J. and Barroso, L.A. The tail at scale. *Commun. ACM* 56, 2 (Feb. 2013), 74–80.
- Dijkstra, E.W. Solution of a problem in concurrent programming control. *Commun. ACM* 8, 9 (Sept. 1965), 569.
- Eriksen, M. Your server as a function. In *Proceedings of the 7th Workshop on Programming Languages and Operating Systems*. ACM (Nov. 2013), 5.
- Eriksen, M. and Kallen, N. Finagle, 2010; <http://twitter.github.com/finagle>.
- Eriksen, M. and Kallen, N. Util, 2010; <http://twitter.github.com/util/>.
- Hughes, J. Why functional programming matters. *The Computer Journal* 32, 2 (1989), 98–107.
- Netflix. Hystrix; <https://github.com/Netflix/Hystrix>.
- Ousterhout, J. Why threads are a bad idea (for most purposes). In presentation given at the Usenix Annual Technical Conference, 1996.
- Sigelman, B.H., Barroso, L.A., Burrows, M., Stephenson, P., Plakal, M., Beaver, D., Jaspan, S. and Shanbhag, C. Dapper, a large-scale distributed systems tracing infrastructure. Technical report, Google, 2010, 36.
- Smolka, G. The Oz programming model. *Computer Science Today*. Jan van Leeuwen, ed. *Lecture Notes in Computer Science 1000* (1995), 324–343. Springer-Verlag, Berlin.
- Sutter, H. The free lunch is over: a fundamental turn toward concurrency in software. *Dr. Dobbs's J.* 30, 3 (2005): 202–210.

Marius Eriksen (marius@twitter.com) is a principal engineer in Twitter's systems infrastructure group. He works on all aspects of distributed systems and server software and is currently working on data management and integration systems; he also chairs Twitter's architecture group. Twitter @marius.

Copyright held by owner/author.
Publication rights licensed to ACM. \$15.00

Article development led by **acmqueue**
queue.acm.org

A discussion with Pete Hunt, Paul O'Shannessy, Dave Smith, and Terry Coatta

React: Facebook's Functional Turn on Writing JavaScript

ONE OF THE long-standing ironies of user-friendly JavaScript frontends is that building them typically involved trudging through the DOM (Document Object Model), hardly known for its friendliness to developers. But now developers have a way to avoid directly interacting with the DOM, thanks to Facebook's decision to open source its React library for the construction of user interface components.

React essentially manages to abstract away the DOM, thus simplifying the programming model while also—in a somewhat surprising turn—improving

performance. The key to both advances is that components built from standard JavaScript objects serve as the fundamental building blocks for React's internal framework, thus allowing for greatly simplified composability. Once developers manage to get comfortable with building front ends in this way, they typically find they can more readily see what is going on while also enjoying greater flexibility in terms of how they structure and display data.

All of which caused us to wonder about what led to the creation of React in the first place and what some of its most important guiding principles were. Fortunately for us, **Pete Hunt**, who at the time was an engineering manager at Instagram as well as one of the more prominent members of Facebook's React core team, is willing to shed some light on React's beginnings. Hunt has since gone on to cofound Smyte, a San Francisco startup focused on security for marketplaces and social networks.

Also helping to tell the story is **Paul O'Shannessy**, one of the first engineers at Facebook to be dedicated to React full time. He came to that role from Mozilla, where he had previously worked on the Firefox front end.

The job of asking the probing questions that drive the discussion forward falls to **Dave Smith** and **Terry Coatta**. Smith is an engineering director at HireVue, a Salt Lake City company focused on team-building software, where he has had an opportunity to make extensive use of both Angular and React. Coatta is the CTO of Marine Learning Systems, where he is building a learning management system targeted at the maritime industry. He is also a member of the *acmqueue* editorial board.

DAVE SMITH: What is it exactly that led to the creation of React?

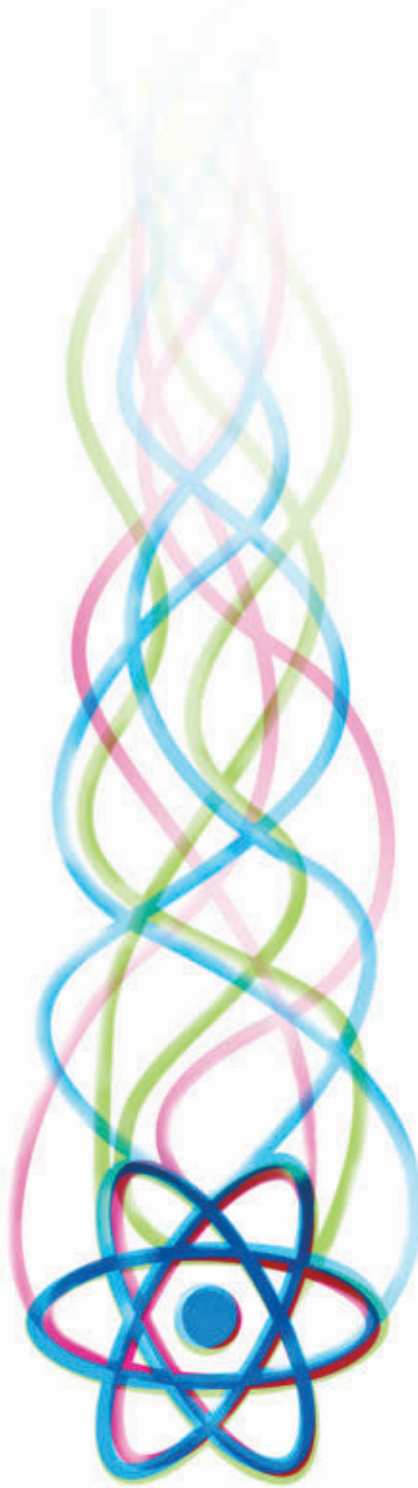
PETE HUNT: Of all the Web apps at Facebook, one of the most complex is what we use to create ads and manage ad accounts. One of the biggest problems is keeping the UI (user interface) in sync with both the business logic and the state of the application. Tradi-

tionally, we've done that by manually manipulating the DOM using a centralized event bus, whether by putting events into the queue or by having listeners for the event and then letting them do their thing.

That proved to be really cumbersome, so a few years ago we implemented what we then considered to be a state-of-the-art, DOM-monitoring system called Bolt. It was kind of like Backbone with observables, where you would register for computed properties that would eventually get flushed to the DOM. But then we found that also was pretty hard to manage since you could never be sure when your properties were going to be updated—meaning that if you changed a value, you couldn't be sure whether it was going to cause a single update, cascading updates, or no updates at all. Figuring out when those updates might actually occur also proved to be a really hard problem.

The whole idea behind React initially was just to find some way to wire up those change handlers such that engineers could actually wrap their heads around them. That hadn't been the case with Bolt, and as a consequence we ended up with lots of bugs nobody could solve. So the engineers who started working on a way to remedy that ended up going wild for a couple of months and came out with this weird-looking thing nobody thought had any chance whatsoever of working. If you're even vaguely familiar with React, you already know that whenever there's a change in your underlying data model, it essentially re-renders the whole application and then does a diff to see what actually changed in the rendered result. Then it's only those parts of the page that get updated.

Some people here had some performance concerns about that, so an early version of React ended up being run through a serious gauntlet of engineering tests where it got benchmarked against pretty much everything that could be thrown at it. As part of that, of course, we looked



at how this new programming model fared against both the Bolt model and our old event model. React ended up really surprising a lot of people—enough so, in fact, that it was shipped almost immediately as part of our “liking and commenting” interface on News Feed. That was the first big test for React, and that came a few years ago.

Then we tried it out on Instagram.com, which is where I entered the picture since I was the person at Instagram responsible for building a few things using React. We were really happy with it since it proved to be capable of running our whole page instead of just one small widget here or there. That gave us a pretty good indication it was actually going to work. Since then, it has essentially become the de facto way people write JavaScript at Facebook.

TERRY COATTA: I've heard React takes a different approach to data binding. What sets React apart there?

PH: The way I think about data binding in a Web context is that you've got some sort of observable data structure down to the DOM nodes. The challenge is that when you're implementing some sort of observable system, you're obliged to observe this data structure wherever your application touches the data model.

For example, if you use something like Ember, everything you do is going to use getters and setters, meaning you're going to need to remain aware of this observable abstraction throughout the entire application. So if you want to compute a value, you're not going to use a function only; you're going to use a computed property number, which is a domain-specific thing for Ember.

Angular, I think, does a much better job of this since it uses dirty checking, which means you can actually take advantage of regular JavaScript objects. The problem with Angular, though, is that it makes it difficult to compose your application. That's because, instead of using regular functions or objects to build up abstractions (as you would do with JavaScript), you have to



PETE HUNT

At a very high level, React essentially treats your user code as a black box while also taking in whatever data you tell it to accept. That basically allows for any structure.



pass everything through a scope in order to observe those changes. Then you end up with this data binding that couples different parts of your program in ways that aren't necessarily all that clear or obvious.

For example, let's say we're looking to sort a list of your top friends—which is the kind of thing we do all of the time here. In order for us to do that with an observable system, we would have to set up an observer for every one of the thousand friends you've listed, even if all we're really looking to do is to render the top 10. So, as you can imagine, it's going to take a good chunk of memory to maintain that whole representation.

Obviously, there are ways to get around that, but people typically just break out of the data binding abstraction altogether at that point so they can proceed manually. Now, I generally hate to say something isn't going to scale, but it's fairly obvious this is going to present some scaling issues. It's clear that the bigger your application gets, the more you're going to run into this sort of edge case.

TC: I agree completely about the Angular situation since I also find composition there to be tricky for just the reason you mentioned—that is, you end up having different parts of your application essentially coupled silently via two-way data binding. But I see that React also has data binding, so I'm curious about how you've managed to provide for better composability despite that coupling.

PH: Let me zoom out a little here to observe that, at a very high level, React essentially treats your user code as a black box while also taking in whatever

data you tell it to accept. That basically allows for any structure. It could be something like Backbone. It could be plain JSON. It could be whatever you want. Then your code will just go ahead and do whatever it's supposed to do, backed by the full power of JavaScript.

At the end of that, however, it will return a value, which we call a virtual DOM data structure. That's basically just a fancy handle for JavaScript objects that tell you which kinds of elements they are and what their attributes are. So if you think of data binding as a way to keep your UI up to date with your underlying model, you can accomplish that with React just by signaling, "Hey, something in my data model may have just changed." That will prompt React to call the black-box user code, which in turn will emit a new virtual DOM representation. Then, having kept the previous representation, React will look at the new version and the old version and do a diff of the two. Based on that, it might conclude, "Oh, we need to build a className attribute at this node."

The advantage of this approach is that it involves no actual tracking of your underlying data model. You don't have to pay a data binding cost up front. Most systems that require you to track changes within the data model and then keep your UI up to date with that are faced with a data binding cost driven by the size of the underlying data model. React, on the other hand, pays that cost relative only to what actually gets rendered.

TC: If I understand you correctly, you're saying React is in some sense a highly functional environment that takes some arbitrary input, renders an output, and then computes the difference between the two to determine what it ought to be displaying on the screen.

PH: Exactly. I like to describe this as "referentially transparent UI." Which is to say your user interface is generally a pure function of some set of inputs, and it emits the same kind of virtual DOM structure every single time for some given data input.

TC: So the data bindings that have caused us grief in Angular run in the other direction here in the sense that they reflect the value of DOM elements that are bound to underlying model ob-

jects or scope variables. Any changes there effectively become visible at multiple locations throughout your code at much the same time, meaning the composability issues surface since different locations in your code are made aware almost simultaneously of changes that propagate backwards from the UI.

PH: Another problem is that you might have multiple bindings to the same data source. So then which piece of code is going to be treated as the authoritative source for determining what the value ought to be?

This is why, with React, we emphasize one-way data flow. As I said earlier, data in our model first goes into this application black box, which in turn emits a virtual DOM representation. Then we close the loop with simple browser events. We'll capture a KeyUp event and command, "Update the data model in this place based on that KeyUp event." We've also architected the system in such a way as to encourage you to keep the least possible mutable state in your application. In fact, because React is such a functional system, rather than computing a value and then storing it somewhere, we just recompute the value on demand with each new render.

The problem is that people sometimes want to have a big form that includes something like 20,000 fields that then bind to some simple keys and data objects. The good news is that it's actually very easy for us to build an abstraction on top of a simple event loop that basically captures all the events that might possibly update the value of this field, and then sets up an automatic handler to pass the value down from the data model into the form field. The form and the data model essentially get updated at the same time. This means you end up with a system that looks a lot like data binding, but if you were to peel it back, you would see that it's actually only simple syntactic sugar on top of an event loop.

TC: One of the things I've observed about React is that it seems to be what people would call fairly opinionated. That is, there's a certain way of doing things with React. This is in contrast to Angular, which I'd say is not opinionated since it generally lets you do things



in several different ways. Do you think that is an accurate portrayal?

PH: It depends. There are certain places where React is very opinionated and others where it's quite unopinionated. For example, React is unopinionated in terms of how you express your view logic since it treats your UI as a black box and looks only at the output. But it's opinionated in the sense that we really encourage idempotent functions, a minimal set of mutable state, and very clear state transitions.

I've built a lot of stuff with React, and I have a team that's run a lot of stuff with it. From all that experience, I can tell you that whenever you run into a bug in a React application, nine times out of 10 you're going to find it's because you have too much state in there. We try to push as much mutable state as possible out of applications to get to what I like to call a fully normalized application state. In that respect, yes, we're very opinionated, but that's just because a lot of React abstractions don't work as well if you have too much mutable state.

I think Angular is actually less opinionated in that regard, but it certainly has opinions about how you need to compose your application. It's very much a model-view-presenter type of architecture. If you want to create reasonable widgets, you're going to have to use directives, which are very opinionated.

TC: Another thing I noticed right away about React is that it's very component-oriented. What was the reason for going in that direction?

PH: We actually think of a component as being quite similar to a JavaScript function. In fact, the only

“

PAUL O'SHANNESSY

We end up writing good code pretty much across the board since there are fewer people going off into crazy land writing CSS. Basically, this just gives us a way at the top level to control all that.

”



DAVE SMITH

When I started out with React, one of the hardest things for me to grasp was this idea that everything is a component... I found myself getting lost in relationships between components.



difference between a function and a component is that components need to be aware of a couple of lifecycle hooks about themselves, since it's important they know when they get added to or removed from the DOM as well as when they're going to be able to get their own DOM node. The component is a fundamental building block on top of which we've built our own internal framework. Now a lot of other people out in the open source world are also building on top of it.

We emphasize it because it's composable, which is the one thing that most separates React components from Angular directives and Web components like partials and templates. This focus on composability—which I see as the ability to build nested components on multiple layers—not only makes it easier to see what's actually going on, but also gives you flexibility in terms of how to structure and display data, while also letting you override behaviors and pass data around in a more scalable and sensible way.

PAUL O'SHANNESSEY: This also has a lot to do with how we build applications on the server, where we have a core library of components that any product team can use as the basis for building their own components. This idea of using components is really just a natural extension of the core way we build things in PHP and XHP, with the idea simply being to compose larger and larger components out of smaller components.

PH: Those product teams tend to be made up of generalists who work in all kinds of different languages, which is to say they're not necessarily experts in JavaScript or CSS (Cascading Style

Sheets). We strongly discourage the average product engineer from writing much CSS. Instead, we suggest that they take these components off the shelf, drop them into whatever it is they're doing, and then maybe tweak the layout a little. That has worked really well for us.

PO: That way we end up writing good code pretty much across the board since there are fewer people going off into crazy land writing CSS. Basically, this just gives us a way at the top level to control all that.

For all the ways in which React simplifies the creation of user interfaces, it also poses a learning curve for developers new to the environment. In particular, those who have worked primarily on monolithic systems in the past may find it challenging to adopt more of a component-oriented mindset. They will also soon find that React is opinionated about how state should be handled, which can lead to some hard lessons and harsh reminders whenever people stray.

TC: There's a lot about React that's appealing, but where are the sharp edges that people ought to look out for before diving in? What kinds of mistakes are likely to make their lives more painful?

PH: Most of the pain points are almost certain to be about state. State is the hardest part of building applications anyway, and React just makes that super explicit. If you don't think about state in the right way, React is going to highlight those bugs for you very early in the process.

TC: Give me a concrete example of how people might think about state in the wrong way.

PH: OK, I'm looking at a site powered by React that was launched earlier today. It looks like the page has four main components: side navigation, a search-results list, a search bar, and a content area containing both the search bar and the search-results list.

When you type in the search bar, it filters the results to be shown in the results grid. If I were to ask you where that filter state should live, there's a good chance you would think, "Well, the search-results list is what's doing the filtering, so the state probably

ought to live there.” That’s what intuitively makes sense.

But actually the state should live in the common ancestor between the search box and the search-results list, sort of like a view controller. That’s because the search box has the state of the search filter as well as the search results. Still, the search-results list needs access to that data as well. React will quickly let you know, “Hey, you actually need to put that in a common ancestor.”

PO: If you were building that same UI with Angular and used a directive for the search box and then another directive for the search results, you would be encouraged in that case as well to put your state in a common ancestor. This would mean having a controller hold the scope variable, since that’s where you’ll find the search text to which both of those directives would then bind. I think you’re actually looking at a pretty similar paradigm there.

PH: Good catch. But I think there’s still a distinction to be made in that React components are building blocks that can be used to construct a number of conceptually different components or objects. You could use a React component to implement a view controller or some pure view-only thing—whereas with Angular, the controller is distinct from a directive, which in turn is distinct from the “service,” which is how Angular describes those things you shove all the other logic into. Sometimes it makes sense just to make all those things React components.

DS: In this case, if you were building the UI with React, what would be the common ancestor? A React component?

PH: Yes. I would use React components for everything.

DS: When I was starting out with React, I think one of the hardest things for me to grasp was this idea that everything is a component. Even when I walked through an example on the React website that included a comment box and a comment list, I was surprised to learn that even those were treated as components. I also found myself getting lost in the relationships between those components. I wonder if you find that to be a common problem for other new React developers.

PO: For people who are used to build-



TERRY COATTA

React is in some sense a highly functional environment that takes some arbitrary input, renders an output, and then computes the difference between the two to determine what it ought to be displaying on the screen.



ing more monolithic things, that often proves to be a problem. At Facebook, where we’ve always coded in PHP, we’re accustomed to building micro-components and then composing them, so that hasn’t proved to be such a huge problem here. Anyway, what I think we’ve always encouraged is that, whenever you’re thinking about reusing something, break it down into its smallest elements. That’s why, in the example you cited, you would want to separate the comment box from the comment list, since you can reuse both of those things in other parts of your application. We really do encourage people to think that way.

PH: We also encourage that you make stuff stateless. Basically, I like to think people are going to feel really bad about using state. I know there are times when it’s a necessary evil, but you should still feel dirty whenever you have to resort to doing that. That’s because then you’ll start thinking, “OK, so I really want to put this search state in only one place in my app.” Generally, that means you’ll find the right spot for it since you’re not going to want to deal with having to synchronize states throughout your application. And you won’t have to if it lives in only one canonical place.

DS: What other major differentiators set React apart from other JavaScript frameworks?

PH: We haven’t yet talked about the idea that React, as a general way of expressing user interface or view hierarchies, treats the DOM as just one of many potential rendering back ends. It also renders to Canvas and SVG (Scalable Vector Graphics), for example. Among other things, this means

React can render on the server without booting up like a full-browser DOM. It doesn't work like it's just some other domain-specific language on top of the DOM. Basically, React pretty much hates the DOM and wants to live outside a browser as much as possible. I definitely see that as a huge differentiator between React and the other JavaScript frameworks.

PO: We've basically seen the same thing happen with WebGL or any other generic rendering platform. It just goes back to the question of immediate vs. retained mode, where you soon discover that as long as you can output something, it really doesn't matter. You just blow away whatever was there before.

DS: I'm also curious about the functional programming aspects of React. In particular, I'm interested in knowing more about which specific functional principles you've adopted.

PH: The truth is, we're actually a bunch of functional programming geeks. In part, that's because if you truly subscribe to the Church of Functional Programming, you can get a lot of performance benefits for free. For example, if your data model is serializable and you treat your render method as a pure function of your properties, you get server-side rendering and client-side rendering for free since both of those end up being pure functions of the same data on both sides of the wire. That way, you can guarantee that when your application initializes, it will get into the same state on both sides automatically. That can be really important if you have a very stateful kind of object-oriented mutative system, since then it becomes much, much harder to synchronize those two states otherwise.

The other advantage has to do with optimizing your apps. We have a hook called `shouldComponentUpdate`, where you can replace React's diff algorithm with a faster custom one. Also, many functional programmers really like to use immutable data structures since that lets them quickly figure out whether something has changed—just another example of how you can get free performance benefits this way.

TC: In the immutable data structures vein, one really powerful library I've heard about is David Nolen's Om.

PH: That's a very cool piece of tech-

nology. It's for ClojureScript, the version of Clojure that compiles to JavaScript. What makes Clojure really cool is its persistent data structures, which basically are really fast and easy-to-use immutable data structures.

What that means for us is that if you have a post on Facebook and somebody likes it, that gives you a new *like* event that should be reflected on the *like* count appearing on that post. Normally, you would just mutate that, but then you would have no way of detecting whether the change actually happened or not, which means you would basically need to re-render the whole thing and then diff it. From that diff, you would learn that only that particular part of the UI actually changed. But if you were using immutable persistent data structures, instead of mutating the *like* count, you could just copy the story object and, within that copy, update the *like* count.

Normally, that would be a very expensive way to go, but in Clojure the copy isn't expensive since it has a way of doing it where it shares the pointers with all the other parts of that data structure and then allocates new objects only for whatever actually changed. That's a good example of an abstraction that's quite complicated under the hood and yet manages to present a very, very simple user interface—something that's extremely easy for people to reason about.

TC: I assume that could also help with undo/redo capabilities.

PH: Right. When everything is immutable, everything gets simpler. Om undos and redos basically just keep around pointers to the previous state and the next state. When you want to undo, you just pass the old object into React, and it will update the whole UI accordingly.

TC: The whole thing?

PO: When your state is serialized into one object at the top level, all you do is pass that through and re-render it—and you're done. With some of the Om examples I've seen, it just snapshots the state at every point and then gives you a UI that indicates how many states you have. Then you can just drag back and forth on that. Or you could start doing some fancier things with the help of trees to produce a really advanced undo system.

I should also point out that React clearly is not purely functional. We also have some very imperative steps and hooks that let you break out of the functional paradigm. But in an ideal world, you don't have any other sources of data, so everything is at the top and just flows through—meaning everything ends up being a very pure output of these render functions.

DS: A bit earlier, you used the term “referential transparency” to describe the way React renders UI. Can you explain what that means?

PH: Basically, React components have props, parameters that can be used to instantiate those components. You might think of them as function parameters. In order to say, “I want to create a type-ahead with these options,” you can just pass in the options list as a prop.

The idea is that if you render a component using the same props and states, you'll always render the same user interface. This can get a little bit tricky, though. For example, you can't read from the random-number generator because that would change the output. Still, if you handle this as a pure function of props and state and make sure you don't read from anything else, you can probably see that this is going to make testing really fast and easy. You basically say, “I just want to make sure my component looks this certain way when it gets this data.” Then, since you don't have to take the Web-driver approach of clicking on every single button to get the app into the right state before double-checking to make sure you've got everything right... well, it becomes pretty obvious how this makes testing a whole lot easier—which, of course, makes debugging easier as well. **C**

Related articles on queue.acm.org

Dismantling the Barriers to Entry

Rich Harris

<http://queue.acm.org/detail.cfm?id=2790378>

The Flame Graph

Brendan Gregg

<http://queue.acm.org/detail.cfm?id=2927301>

Componentizing the Web

Taylor Savage

<http://queue.acm.org/detail.cfm?id=2844732>

Copyright held by owners/authors.
Publication rights licensed to ACM. \$15.00



Essence is instrumental in moving software development toward a true engineering discipline.

BY IVAR JACOBSON, IAN SPENCE, AND ED SEIDEWITZ

Industrial-Scale Agile—From Craft to Engineering

IN HIS PAPER entitled “Industrial Scale Agile,”⁹ Roly Stimson characterizes industrial-scale agile as:

- ▶ “Agile at any scale.”
- ▶ “Agile as the rule, not the exception.”
- ▶ “Agile sustainably, forever,” not just as an unrepeatable “one-off.”

This means being able to sustainably apply agile strategies appropriately to anything and everything that can benefit from them. This includes:

- ▶ Being able to do “agile at scale” as and when appropriate.
- ▶ Doing small-scale agile as and when possible/appropriate.
- ▶ Evolving the entire application landscape and not just individual applications.

Although it is important, and a necessary precursor to industrial-scale agile, scaling agile is not the challenge here. Rather, it’s about how to achieve sustainability of the following:

- ▶ The way of working in the face of ever-changing teams.

- ▶ The systems in the face of rapid change.

- ▶ The application landscape as a whole.

- ▶ Individuals and their careers, and the development organization as a whole.

- ▶ Long-term investment in IT.

There are many ways to illustrate how fragile IT investments can be. You just have to look at the way that, even after huge investments in education and coaching, many organizations are struggling to broaden their agile adoption to the whole of their organization—or at the way other organizations are struggling to maintain the momentum of their agile adoptions as

their teams change and their systems mature.

Another frequent example of unsustainability is in the way many companies are facing an uncontrolled explosion in the number of applications they have to support and the overall cost of ownership of IT as a whole.

So industrial-scale agile requires much more than just being able to scale agile. It also means taking a disciplined approach to ensuring IT investments result in sustainable benefits for both the producing organization and its customers.

This involves adopting a different approach to many aspects of agility. We need to look beyond small-scale agile, beyond independent competitive islands of agile excellence, beyond individual craftsmanship and heroic teams, and beyond the short-term instant gratification that seems to be the focus of many well-intentioned but self-centered agile teams. It is this adoption of a more holistic approach that we call moving from craft to engineering. (See Jacobson³ for more background.)

From Craft to Engineering

The move toward agility has led to many benefits for the software industry. It has broken the tyranny of the prescriptive waterfall approach to software engineering, an approach that was causing more and more large project failures, and it has allowed software developers to keep up with the ever-increasing demand for more innovative IT solutions.

It has enabled many companies to do great things but in many cases has led to a culture of entitlement, heroic programming, and short-term thinking that threatens the sustainability of the parent companies and the IT solutions on which they depend. Little or no thought is put into maintainability, the heroes become potential single points of failure, and the cost of keeping the lights on just keeps growing and growing.

What is needed is a way to maintain the values of agility while making software development more an engineering discipline than a craft—a new form of agile software engineering fit for the Internet age.

What are craft and engineering? The term *craft* is usually applied to people occupied in small-scale production

of bespoke goods and trades where skills are passed in person from master to apprentice. *Engineering*, on the other hand, is defined by Wikipedia as “the application of mathematics, empirical evidence and scientific, economic, social, and practical knowledge in order to invent, innovate, design, build, maintain, research, and improve structures, machines, tools, systems, components, materials, and processes.”

There have been many discussions about whether or not the term *engineering* should be applied to software development and whether or not software engineers are actually engineers. With the rise of cloud computing, big data, and the Internet of Things, however, it is clear there are many types of software and many aspects of software development that would benefit from an engineering approach.

In her 1990 seminal paper, “Prospects for an Engineering Discipline of Software,”⁷ Mary Shaw suggested a definition of software engineering would include these clauses: “Creating cost-effective solutions ... to practical problems ... by applying scientific knowledge ... building things ... in the service of mankind.” She also said about software work that “most tasks are routine and not innovative,” but it “is treated more often as original than routine,” implying that there is a lot of potential for improving quality and shortening time to market “if we captured and organized what we already know” by codifying our knowledge, possibly even automating it.

Her observations are still highly relevant; at the GoTo Amsterdam 2015 conference on software development, she talked about the progress made toward establishing a software engineering discipline. According to Shaw, the characteristics of engineering are as follows:

- ▶ Limited time, knowledge, and resources force decisions on trade-offs.
- ▶ The best-codified knowledge, preferentially science, shapes design decisions.
- ▶ Reference materials make knowledge and experience available.
- ▶ Analysis of design predicts properties of implementation.

Although software development shares many of the characteristics of an engineering discipline, we are not there yet. The rise of agile is not a problem unless this is where we stop.

Why is it important to move from craft to engineering?

Doing so will help us cope with the ever-increasing challenges of a more automated, more interconnected world—where small improvements in software performance can make the difference between profit and loss; where a reputation for robustness, scalability, and security can add millions to the share price; and where software is more and more the public face of the business.

The codified knowledge and professionalism of an engineering discipline are necessary for:

- ▶ Sustaining and growing delivery capability through changes in technologies, teams, and suppliers.
- ▶ Predictably scaling operations from early prototypes to global rollouts.
- ▶ Taking control of investments and knowing when to pivot to solutions more likely to deliver favorable returns.
- ▶ Systematically growing the levels of reuse and interoperability of solution components and systems.
- ▶ Producing long-lived solutions with affordable costs of ownership.

Perhaps not everybody needs to move from craft to engineering. As Mary Shaw says, “The greatest need for engineering discipline exists for software systems that are fully automated and are operating unattended and where the consequences of failure are catastrophic. Examples are telecom equipment, nuclear safety devices, medical implants, self-driving cars, and stock-trading programs. The need for engineering in software development depends upon how serious the consequences are when things go wrong and whether human beings can take action in time to minimize the consequences.” There is also a strong need for engineering systems used for e-commerce, finance, electronic medical records, and even human resources. The consequences of failures in such systems may not include the immediate loss of life, but they can still be “catastrophic” to either the businesses or the individuals affected.

Thus, for many organizations and software systems craft is not enough.

The good news is there is a way forward that maintains the values of agile while making software development more of an engineering discipline than a craft. It involves:

Engineering of software. This means the holistic engineering of all software to improve the application landscape as a whole, as well as the individual point solutions. Practices are needed that help teams engineer their software for capturing requirements and for developing software designed for engineering great products. It also means encouraging innovation in the large as well as the small—innovation of new business and new product opportunities as well as innovation that addresses the total cost of ownership impacting the whole organization, rather than just individual users and applications.

Engineering of methods. Methods should be engineered to support the full range of development challenges faced today and in the future. The emerging best practice should be captured and codified in a way that makes it easy to communicate and share among teams, and enables each team to compose the method they need from this growing set of reusable, proven practices.

Furthermore, moving from craft to engineering provides a robust platform for encouraging, establishing, and sustaining true organizational agility.

Engineering of software. How would software be developed if the craft were already a real engineering discipline? As in other engineering disciplines, it would be engineered by using commonly accepted, consistent practices that would be supported by models and analysis based on a common ground of foundational knowledge.

In the past, such an engineering mindset has been misinterpreted as meaning “big upfront design,” with everything downstream of this being akin to manufacturing rather than engineering. Upfront blueprinting is, indeed, often necessary for the engineering of physical artifacts such as buildings, bridges, and cars. This is done so that proper analysis can be carried out on the upfront models and blueprints, because of the capital cost required to build those things and the difficulty of changing them once built. Software, however, is a different kind of artifact—one that does not require manufacturing in the physical sense.

Agility in software development takes advantage of this characteristic,

Craftsmanship and Engineering

Related to the idea of *craft* is *craftsmanship*, performed by a person who practices or is highly skilled in a craft. Software development will always need craftsmanship that can stand on more or less science, more or less engineering, and more or less structured knowledge. We would, for example, describe an engineer as a craftsman using engineering practices in developing software. The new software craftsmanship movement is supportive of many engineering practices—for example, they are strong supporters of design and architecture patterns, domain-driven design, among others. They take pride in using the right tools, techniques, and design methods to achieve high-quality software. They do not believe in heroics, but in quality of work and in tools. They believe in sustainability, and in keeping the system “clean” and able to absorb change and rework. The craftsmanship movement, however, doesn’t fully address the whole engineering space and, in particular, how to systematically grow knowledge about the discipline.

allowing software to be developed in a rapid and incremental, but still reliable, way; however, there is a place for disciplined design within an agile development approach. It is just that, with software, developers can also carry out analysis and evolve designs incrementally, as they build the software system itself.

What is needed is, in fact, a merger of the agile mindset with the engineering mind-set, combining incremental development with the disciplined application of foundational knowledge. In such an approach, not everyone will necessarily be an engineer, but developers will continue to be treated as skilled craftsmen, not factory workers. (See the sidebar “Craftsmanship and Engineering.”)

It is common in agile approaches to talk of the *emergence* of the design of a software system as that system is iteratively developed. This is the very embodiment of evolutionary design as opposed to big upfront design. It can be very effective in allowing a team to explore alternatives creatively, while still converging on a good solution with a clear overall design.

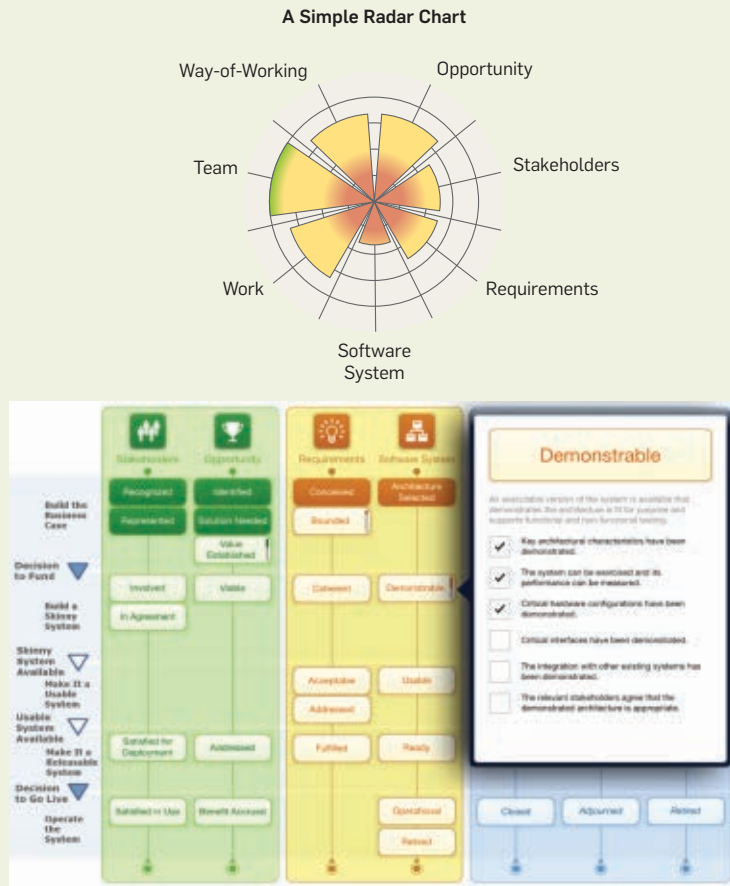
Such emergent design, however, tends to produce point solutions for specific teams. Serious software development organizations, though, are almost always dealing with multiple teams working on multiple projects within an overall enterprise-level application landscape. Various project-level solutions need to fit into this evolving landscape. Indeed, the development of a large software system often requires multiple teams whose products are components that must fit together to create the overall system.

Dealing with design at this level is the province of *software architecture*, which, at both the system and enterprise levels, can and should still be evolutionary. Rather than being entirely emergent, however, key architectural decisions, presented in a development roadmap, often need to be made in advance of the corresponding development work in order to provide common guidance across projects and teams. This is where engineering practices can be particularly important, allowing for innovations that benefit the organization as a whole, based on careful analysis of business benefit versus engineering cost.

Engineering of methods. Moving from craft to engineering relies on the codification and sharing of knowledge. What is needed is for organizations to *engineer their methods* in order to be more effective at engineering their software.

Most methods in use today are at the extremes, either monolithic or tacit. The agile space is experiencing the rise of a number of competing, monolithic scaled agile methods, such as DAD (disciplined agile delivery), SAFe (Scaled Agile Framework), LeSS (Large-scale Scrum), and SPS (Scaled Professional Scrum). All these methods have their special strengths and weaknesses. They have their own camps of supporters, but their monolithic nature doesn’t make it practical to borrow ideas from one another, even less to borrow complete codified practices. This situation is very similar to what we had in the past with methods such as RUP (Rational Unified Process), Open, Structured Analysis and Design, and

Figure 1. Simple Essence-based health monitors.

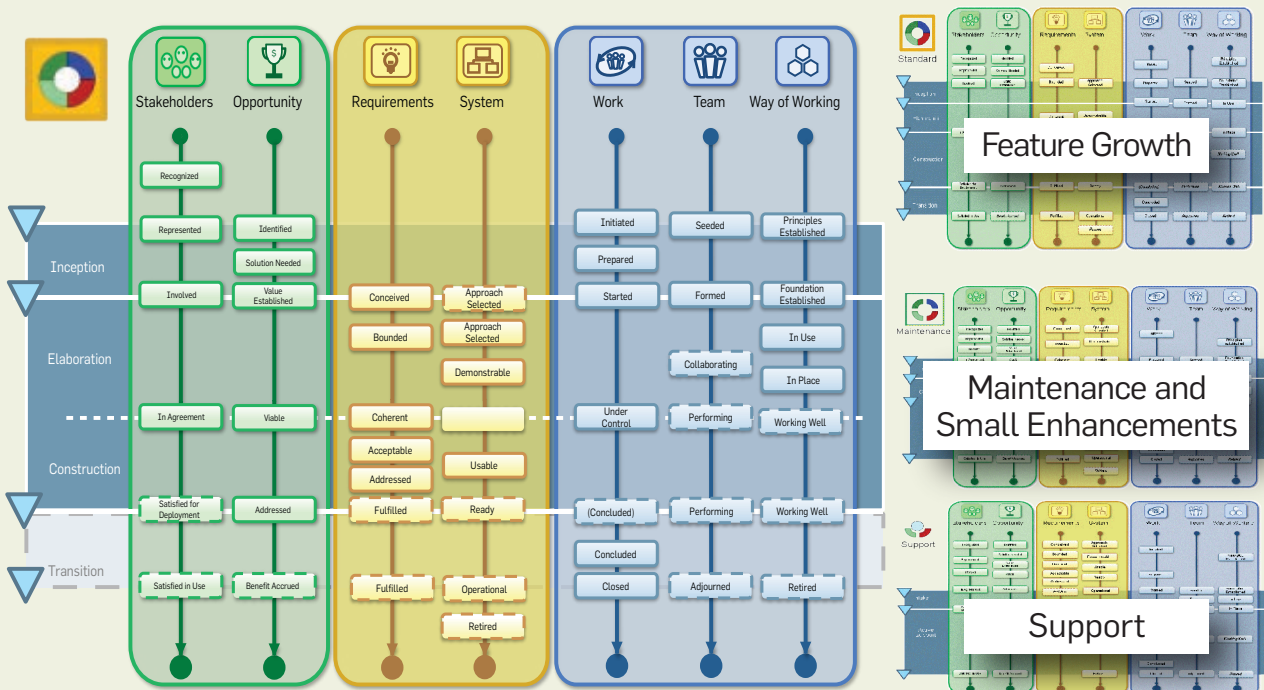


so on. We have also observed that, at many large organizations, the success of tacitly applied agile practices has led to a situation where the previously used and codified (documented) methods have been replaced by undocumented agile folklore.

The paradox here is the discomfort caused by not having a documented method causes many organizations to seek to replace their tacit agile methods with one of the new monolithic methods. What they don't realize is that it will end up being rejected in the same way as the original method as teams seek to innovate and meet the day-to-day challenges inherent in their systems and circumstances. This often leads to a constant churn as method replaces method with little or no rhyme or reason. The industry's habit of constantly switching between no methods and the latest "one true way" (an affliction that is sadly affecting even the agile community) is not the way forward.

Instead, organizations need an effective way of using what they learn from effort to effort, applying and adapting it to new projects. Moving blindly from one fad method to another provides no consistent basis for building common

Figure 2. Four compatible governance life cycles defined using the Essence kernel.



knowledge. Mandating a one-size-fits-all process for all projects does not support the need for continual learning and adaptation, however, and suppresses craftsmanship and creativity.

The move from craft to engineering requires first freeing the practices, presenting them in an accessible, reusable way that allows engineers confidently and predictably to select the right engineering practices for their context and the problems they are trying to solve.

Essence: A Hope for a Better Future

Software development is a multidimensional endeavor—where human ingenuity meets human need meets collective endeavor meets codified knowledge—that would benefit from the judicious application of engineering practices. The path from craft to engineering progresses—from ad hoc practice to codified professional engineering practices—through scientific learning.

The key to this transformation is the ability to readily capture, share, and improve the practices.

Essence is a simple intuitive language and kernel of foundational elements for the capture, description, and assembly of practices and methods. Work on Essence has been going on for more than 10 years—for the last six years within the SEMAT (Software Engineering Method and Theory) community—resulting in a new international standard, adopted by OMG (Object Management Group) in 2014.⁵ It goes beyond just providing syntax and notation for describing practices to establishing a solid common ground—a kernel—that enables teams to:

- ▶ Describe their practices on top of a universal, shared kernel.
- ▶ Easily share, adapt, and plug and play with their practices to create the innovative ways of working they need to excel and continuously improve.
- ▶ Understand and visualize the progress and health of their endeavors, regardless of their way of working.

Essence has several roles to play in the move from craft to engineering: It is helping to achieve the right balance in software engineering endeavors; helping to codify and capture engineering practices; and acting as the basis for a new kind of engineering community.

The use of Essence alone won't turn craftspeople into engineers, but

its adoption will help an organization make this important transition and, moreover, help the industry prepare for the future.

Balancing progress and health. Essence provides a kernel of elements that establishes a common ground for carrying out software engineering endeavors. This can be used in a number of ways to increase the effectiveness of software engineering teams.

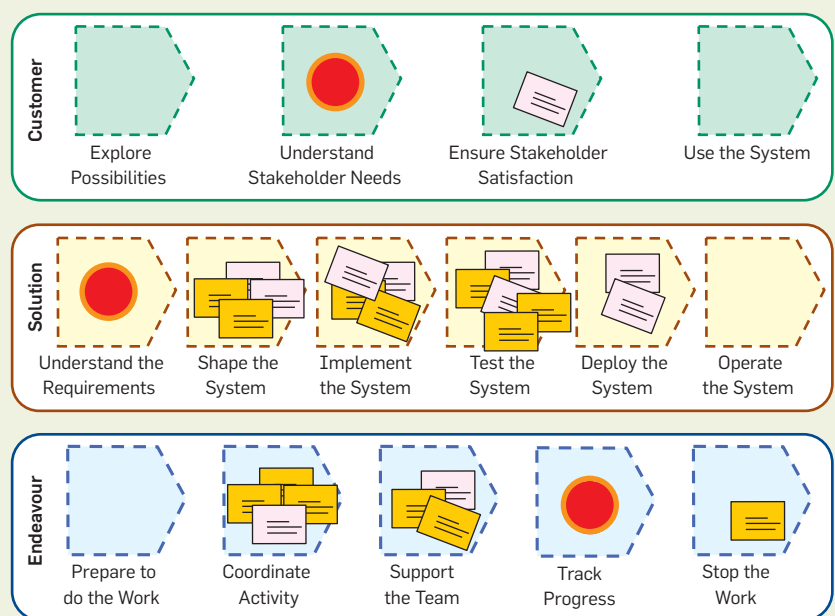
Actively monitoring the health of an endeavor. The kernel defines seven aspects of concern for any software engineering endeavor: opportunity, stakeholders, requirements, team, work, way of working, and the software system itself. For each of these elements Essence defines a series of states, with checklists, representing healthy progress. As shown in Figure 1, these can be used to create practice-independent health monitors that can be used to check that the endeavor is on course and proceeding in a healthy manner. On the top, the radar chart (an interactive, online version complete with checklists is available¹) represents progress as growth from the center; and on the bottom, on the milestone map (available from the App Store as the Alpha State Explorer app by Ivar Jacobson International), all the states are laid

out in order from top to bottom, with achieved states shown filled in. The second example also shows the checklist used to confirm the achievement of the Software System Demonstrable state.

The kernel can also be used to create lightweight governance and compliance practices to help ensure the team achieves the required level of engineering rigor. By basing the governance and compliance on the kernel itself, this can be achieved in a practice-independent fashion allowing the teams to safely innovate and own their own ways of working. Figure 2 shows the four different governance life cycles that were at the heart of Munich Re Essentials, the modern practice-based software development method created by Munich Reinsurance.⁴ The four life cycles are Exploratory, Feature Growth, Maintenance, and Support, and the checkpoints (milestones) of each life cycle are defined by the states to be achieved for each alpha.

Assessing the effectiveness of methods. Essence at its roots gives a detailed definition of software engineering. In the search for a GTSE (general theory of software engineering),⁶ several researchers use Essence as such a definition, and more is expected to come out of this work. A key aspect provid-

Figure 3. Simple Essence-based activity map.



Some Definitions

The terms used in the method space are often ill-defined or confused. For example, what is the difference between a method and a methodology? A practice and a process?

Composition: The process of merging practices into practices and methods. It is important to understand that practices are separate concerns composed through a merge operation and not components interacting through messages.

Essence kernel: An actionable reference model of software engineering that provides a framework for the definition of practices and the assembly of methods.

Essentialization: The process of rendering a method or practice down to its essence and capturing it using the Essence language.

Method: The documentation of a team's way of working. A method may or may not be documented using Essence. If it is documented in Essence, then it is the composition of the Essence kernel and a set of practices to fulfill a specific purpose. A method could belong to a single team or be shared among teams.

Methodology: A collection of practices known to share a common set of values and work well together. It's a form of practice library.

Practice: A repeatable approach to doing something with a specific objective in mind.

Practice library: A collection of potentially competing practices. For example, a requirements management practice library could contain many different competing practices such as declarative requirements, use cases, and user stories.

Starter pack: A partially built, often incomplete method that a team can use as a framework to seed their own method.

Way of working: This is what a team actually does. It may or may not match their method.

ed by such a theory is the capability to be *predictive*.

A construction engineer can use material science and the theory of structures to understand at an early stage whether a proposed building is likely to stand or fall. Similarly, using Essence, one can understand whether a proposed method is well constructed, whether or not there are any gaps or overlaps in its practices, and if there are gaps or overlaps, how to resolve them.

The kernel has many mechanisms for method analysis, the simplest of which is provided by its high-level activity map. This is a set of 12 activity spaces organized into three areas of concern. An activity space is a generic placeholder for method-specific activities. These activity spaces, as shown in Figure 3, can be used to assess the spread of the team's activities. In this example, the team has added notes to the map to indicate their activities and red circular markers to highlight the danger areas.

Note: Without understanding the meaning of the Essence language, the symbol of a pointed arrow to represent an activity space can make them

appear sequential, which is not the intended meaning. The activity spaces and the activities that they contain can of course be applied iteratively, concurrently, or in any order the practices require.

These are just a few simple examples of the kernel's capabilities, but they illustrate the many ways it can help teams and organizations assess the effectiveness of their methods.

Codifying and capturing engineering practices. In addition to the kernel, Essence provides a language for creating practices on top of the kernel and then composing methods from those practices. This is extremely important for moving from craft to engineering.

As discussed previously, most current practices are embedded in monolithic methods that aggressively compete with one another. Rather than admit that they share practices and encourage reuse and cross-pollination, they willfully slander and steal from one another. Even worse, from an engineering perspective, they are all concerned with the design and sustainability of the development organization rather than

the design and sustainability of the systems produced. This is not to say the former isn't important. Indeed, it is crucial to the success of any development organization. As the integration of software into the fabric of our daily lives grows, however, the need for proven, reusable engineering practices grows as well.

Practices are needed that help teams engineer their software: practices for working with requirements, such as use cases, features, and stories; for developing components and services; for applying an appropriate pattern or framework; for testing complex, distributed systems; that encourage reuse; and that help engineers code with confidence. In particular, practices are needed for dealing with architectural concerns such as concurrency, security, user experience, microservices, and data protection, as well as for addressing broader architectural concerns such as enterprise architecture, product-line architecture, service-oriented architecture, and the architecture of systems. Many of these practices already exist codified in the Essence language (see the section "Sharing practice: Methods and practice libraries").

These engineering practices need to be streamlined (lean), agile, and, most importantly, composable into complete methods to provide guidelines for teams working with a multitude of practices for complex systems. They are needed to help deal with the complexities of modern software engineering. They need to be available to all engineers whether they are working alone, in small teams, or in larger teams of teams, regardless of the style of team working or work-management practices adopted.

Globally, we want a robust and flexible library of codified professional engineering practices that reflect the multidimensional nature of software development and that can be used to support the many different types of software being developed today and in the future.

These practices can only come from engineering teams working on the cutting edge of technology, and these teams need a better way to capture, communicate, and share their practices.

A new method architecture. With

the Essence kernel as common ground, you can use the Essence language to describe any practices, including engineering practices, in a way that allows them to be composed seamlessly together to form methods. Figure 4 illustrates a three-layer method architecture with the kernel as the foundation, generic practices in the middle, and domain-specific practices at the top.

Starting from the bottom of the stack, the three layers are:

The Essence kernel. This provides the common ground for all practices and methods and the underlying foundation for the definition and composition of the practices.

Generic practices. These are practices that are applicable across many software engineering domains. Examples of generic practices include Scrum, use cases, user stories, test-driven development, and acceptance-test-driven development. Many engineering practices will be generic, but many of the most valuable will be domain-specific.

Domain-specific practices. These practices are explicitly targeted to a specific domain such as business intelligence, data warehousing, or telecommunications. Domain-specific practices are equally as important as generic practices, if not more so. For example, many domain-specific practices are needed to develop solutions for the Internet of Things; these practices cater to things such as asset integration architecture and different technology profiles. Just as generic practices extend the kernel to provide specific guidance, domain-specific practices are often extensions/specializations of the generic practices. For example, an asset integration architecture practice could be presented as an extension to a generic agile architecture practice.

The separation of generic practices from domain-specific practices helps teams find the practices that they need and helps organizations establish common ways of organizing and tracking their work. It is not uncommon for an organization to standardize on a small set of generic practices as the foundation for all of its teams' methods.

Liberating practices in this way is very powerful. Once practices are codi-

fied in Essence, teams can take ownership of their ways of working and start to assemble their own methods. This can start with even a simple library of practices, as shown in Figure 5.

This capturing and sharing of engineering practices, both generic and domain-specific, in a way that lets them be applied alongside popular

management practices (agile or otherwise), provides the codified knowledge needed to support a true software engineering discipline. It is also the key to moving away from monolithic management methods and isolated engineering practices.

Sharing practice: Methods and practice libraries. It's easy to say that teams

Figure 4. The Essence method architecture.

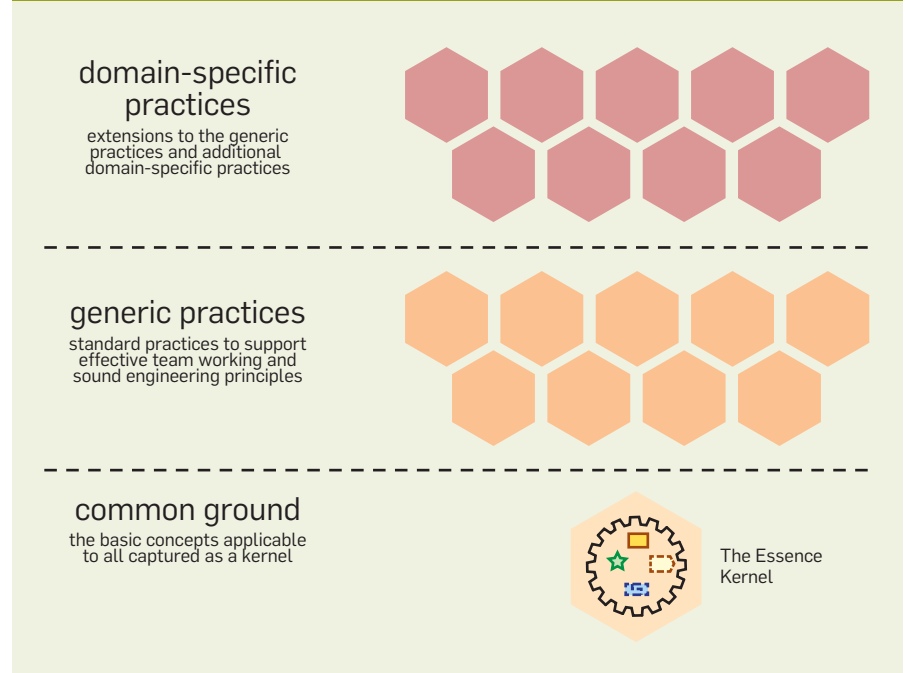
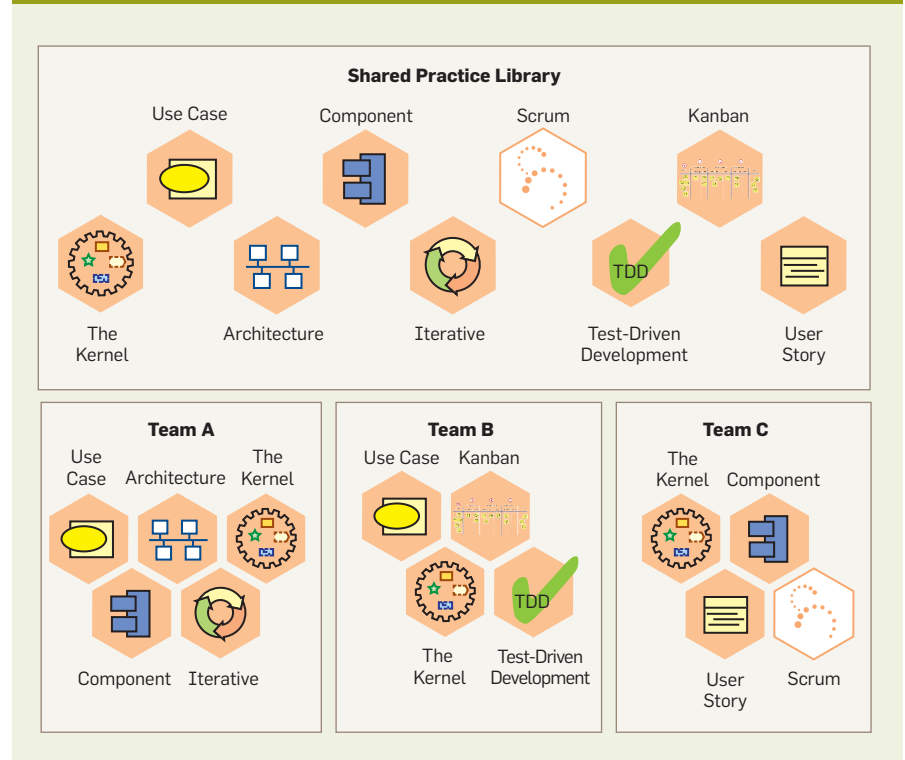


Figure 5. Three teams sharing a simple practice library.



will be able to plug and play with sets of practices to build their own methods and take ownership of their way of working. But where are the practices going to come from?

Let's take a look at two concrete examples.

Agile methods. The industry has

seen an explosion in the number of generic agile practices being published and promoted. Unfortunately, most of these “belong” to one method or another and, even though they share the same values, are rarely presented in a way that lets them play well together. This is particularly true in the area

of scaled agile methods, where each method contains many of the same practices tangled up with a few new, unique, and innovative practices in such a way that the safe separation of the new practices for use with another method is nearly impossible.

In contrast, Figure 6 outlines a starter pack of agile practices based on Essence, called Agile Essentials.² It includes practices from Scrum, Kanban, and XP (extreme programming).

This is a small library of seven practices, which, when composed together, form a starting point for a team's agile method. Scrum, user stories, and use cases have also been “essentialized” and can be used alongside the Agile Essential practices.

Thus, with Essence, a library of generic, reusable practices can be created, from which a team can select the ones they want to use and they can compose together to kick-start their own method.

The Ignite Internet of Things methodology. Ignite is a methodology developed for the Internet of Things.⁸ It supports a number of different approaches and attempts to bridge the gaps between “machine guys” and “Internet guys,” and between “five-year thinking” and “continuous beta.” Ignite can easily be described as a set of practices on top of the Essence kernel. Figure 7 demonstrates what Ignite looks like when presented using Essence.

This picture readily illustrates a number of key points:

- ▶ Ignite clearly contains and reuses a number of generic practices that are applicable in many more domains than the Internet of Things, including those already available as part of the Agile Essentials practice library.

- ▶ Successful development for the Internet of Things requires many domain-specific engineering practices.

- ▶ Whenever anyone wants to create a new method, they currently have to rewrite, re-present, and, in many cases, rebrand already established generic practices.

- ▶ The more comprehensive the approach, the less likely it is that anyone will use all of it. For example, no one is ever going to use all of these practices at the same time. Clearly, there are many methods that could be built from the practices contained within the Ig-

Figure 6. Agile essentials with its seven practices.

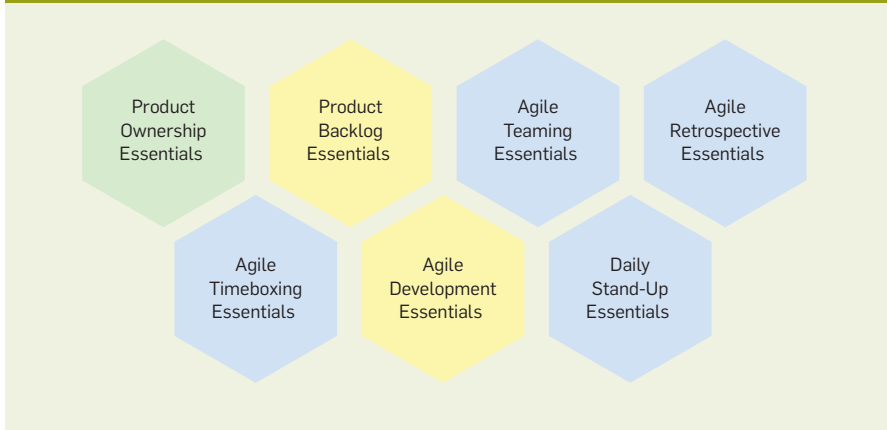
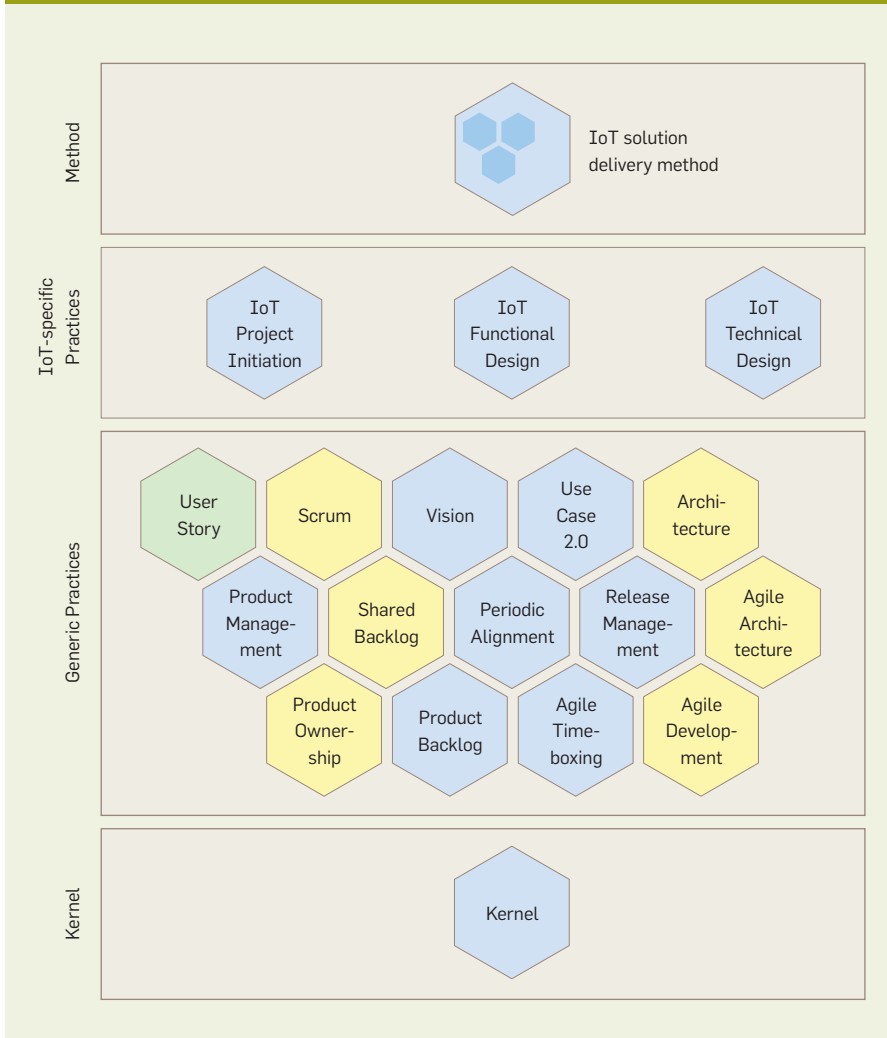


Figure 7. Ignite expressed as a set of Essence practices.



nite methodology, but without the use of Essence this will be very difficult, if not impossible, for teams to do.

Many other practices could be useful for teams developing for the Internet of Things. Some of these will be innovations unknown at the time of writing this article or the creation of Ignite. This can easily lead to the approach becoming out of date and unfashionable. The presentation of Ignite as a practice library allows the practice set to respond to the needs of the users, who may regularly add new practices and retire those that are no longer needed.

The process of extracting the practices from an existing method or methodology is called *essentialization*. Essence is designed to allow people to extract the essence of any method or practice, so essentialization of a method means identifying the method's practices and practice architecture. Moreover, each practice is described/codified in terms of the elements in Essence and the Essence language with new practice-specific elements added as needed. As of this writing, the Unified Process has been essentialized, and DSDM (dynamic systems development method) is in process. Several other methods are in the planning stages to become essentialized. Many companies around the world are now using the Essence standard to essentialize their methods.

The value of essentialization is that people can easily learn what really matters about a practice, compare it with other practices, compose it into a method (with many other proven practices), and easily modify/change the method as new knowledge becomes available. Applying Essence also makes it easier to govern the methods you have in your organization, so you create an effective learning organization. Moreover, an essentialized method is not just a static description, but helps the team while they actually use the method, allowing them to measure progress and health at any moment during their endeavor.

Less work has been done to capture the domain-specific practices needed to bridge the gap between craft and engineering. As discussed earlier, the concepts can be illustrated using Ig-

nite⁸ and other popular methodologies, but a vibrant and committed engineering community must flesh out and complete the necessary set of engineering practices.

There are two ways to accelerate the transition:

- ▶ Slice the popular methods into practices and design these practices so that they can be composed in any reasonable way teams want, maybe resulting in a method with practices from several existing methods such as DAD, SAFe, LESS, and SPS.

- ▶ Codify existing or new practices so that they can be composed with other practices to form complete methods. There are already hundreds of practices in the world, but they are not described in a way that allows them to be easily composed. Now this can be done without having to describe a complete method.

In both cases Essence is key as it provides the foundation for this work and for the industry to transition successfully from craft to engineering.

Conclusion

As software becomes more essential to the world's day-to-day activities, it is time for software development to move beyond a craft-based approach to become a true engineering discipline.

This will require a shared base of codified engineering practices that can be reused across various technical domains and various types of software; this set of practices will grow and adapt as better ways of developing software come along.

This is not going to happen overnight, but it is a challenge to which our industry needs to rise as it matures and evolves into something beyond agile and other current practices.

We still need the dedication, innovation, and invention of craft, embodied in:

- ▶ Skilled professionals, passionate about their subjects and committed to mastering new, complex, fast-moving technologies.

- ▶ Local experts who understand complex problems in depth and respond rapidly to changing needs, perceptions, and challenges.

However, we also need the codified knowledge and professionalism of an engineering discipline to be able to:

- ▶ Sustain and grow delivery capability through changes in technologies, teams, and suppliers.

- ▶ Predictably scale operations from early prototypes to global rollouts.

- ▶ Take control of investments and know when to pivot to solutions more likely to deliver favorable returns.

- ▶ Systematically grow the levels of reuse and interoperability of solution components and systems.

- ▶ Produce long-lived solutions with affordable costs of ownership.

This is what we mean by moving from craft to engineering—a journey that must be made practice by practice, domain by domain. Thanks to Essence, that journey can start today for all of us. C

References

1. Graziotin, D., Abrahamsson, P. A Web-based modeling tool for the SEMAT Essence theory of software engineering. *J. Open Research Software* 1, 1 (2013), e4; <http://dx.doi.org/10.5334/jors.ad>.
2. Ivar Jacobson International. Agile Essentials, 2015; https://www.ivarjacobson.com/sites/default/files/field_jji_file/article/agile_essentials_paper.pdf.
3. Jacobson, I., Seidewitz, E. A new software engineering. *acmqueue* 12, 10 (2014); <http://queue.acm.org/detail.cfm?id=2693160>.
4. McDonough, A. Munich Re and ESSENCE—Kernel and Language for Software Engineering Methods: A Case Study. Object Management Group, 2014; http://www.omg.org/news/whitepapers/Munich_Re_Essence_Case_Study-2014-12-01_JP.pdf.
5. Object Management Group. Essence—Kernel and language for software engineering methods (Essence), 2014; <http://www.omg.org/spec/Essence/>.
6. Ralph, P., Johnson, P., Jordan, H. Report on the first SEMAT workshop on general theory of software engineering. *ACM SIGSOFT Software Engineering Notes* 38, 2 (2012): 26–28.
7. Shaw, M. Prospects for an engineering discipline of software. *IEEE Software* 7, 6 (1990), 15–24.
8. Slama, D., Puhlmann, F., Morrish, J., Bhatnagar, R. *Enterprise Internet of Things: Strategies and Best Practices for Connected Products and Services*. O'Reilly, 2015.
9. Stimson, R. Industrial scale agile—challenges and solution strategies. Ivar Jacobson International, 2015; <https://www.ivarjacobson.com/publications/white-papers/industrial-scale-agile-challenges-and-solution-strategies>.

Ivar Jacobson is a father of components and component architecture, modern business engineering, the Unified Modeling Language, and the Rational Unified Process. His latest contribution is a formal practice concept that promotes practices as the “first-class citizens” of software development. Jacobson is also one of the founders of the SEMAT (Software Engineering Method and Theory) community.

Ian Spence is CTO at Ivar Jacobson International and the team leader for the development of the SEMAT kernel. An experienced coach, he has introduced hundreds of projects to iterative and agile practices.

Ed Seidewitz is the former CTO, Americas, for Ivar Jacobson International and is currently chair of the ongoing Essence Revision Task Force. With Ivar Jacobson International, he has led agile system architecture and development engagements in both the commercial and government sectors and participated in practice development.

Copyright held by owners/authors.
Publication rights licensed to ACM. \$15.00

DOI:10.1145/2950040

A full-body virtual autopsy of an ancient Egyptian mummy showed visitors he was likely murdered.

BY ANDERS YNNERMAN, THOMAS RYDELL, DANIEL ANTOINE, DAVID HUGHES, ANDERS PERSSON, AND PATRIC LJUNG

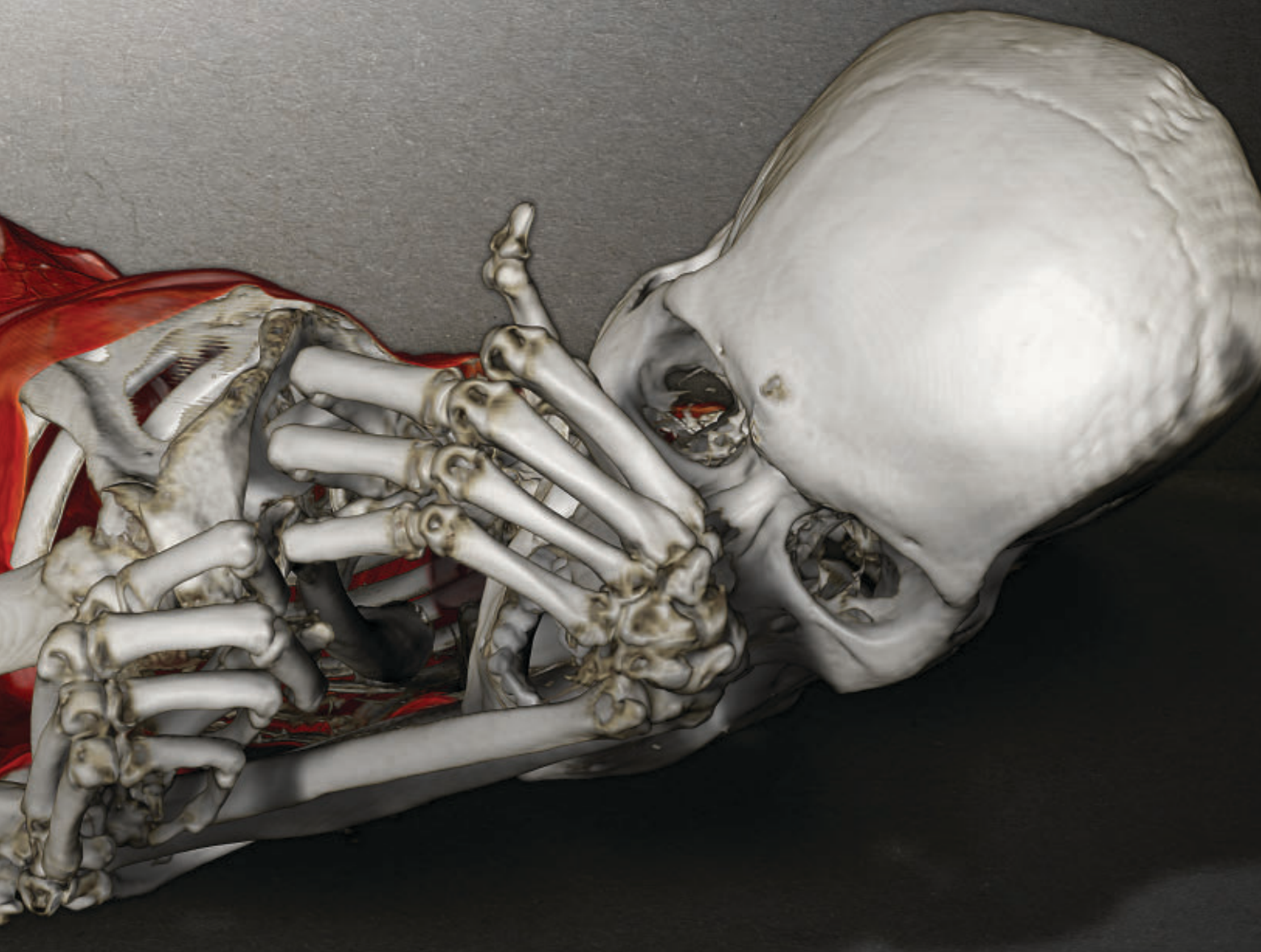
Interactive Visualization of 3D Scanned Mummies at Public Venues

BY COMBINING VISUALIZATION techniques with interactive multi-touch tables and intuitive user interfaces, visitors to museums and science centers can conduct self-guided tours of large volumetric image data. In an interactive learning experience, visitors become the explorers of otherwise invisible interiors of unique artifacts and subjects. Here, we take as our starting point the state of the art in scanning technologies, then discuss the latest research on high-quality interactive volume rendering and how it can be tailored to meet the specific demands



» key insights

- Volumetric visualization, together with CT scanning, creates opportunities for science communication, as visitors to public venues interactively explore the hidden interiors of rare museum artifacts and biological remains.
- Curating data and producing stories are bottlenecks for curators, interpreters, and exhibition designers; software solutions for supporting this workflow are lacking.
- Social context and engagement at public venues is available through touch tables with high resolution, image quality, and graphics performance together with user-interface design, balancing free exploration and guided storytelling.



of public venues. We then describe our approach to the creation of interactive stories and the design principles on which they are based and interaction with domain experts. The article is based on experience from several application domains but uses a 2012 public installation of an ancient mummy at the British Museum as its primary example. We also present the results of an evaluation of the installation showing the utility of the developed solutions.

Visitors walk into Gallery 64, the Early Egypt Gallery at the British Museum, eager to see and learn about one of the most famous and oldest

mummies in the collection. Known as the Gebelein Man, he was buried in a crouched position in a shallow grave during the late pre-dynastic period at the site of Gebelein in Upper Egypt. Unlike later Egyptian mummies, he was not artificially embalmed and his body was naturally mummified by the arid environment and direct contact with the hot dry sand more than 5,500 years ago. The mummy is safely protected behind glass and within a carefully controlled environment, beyond physical reach (see Figure 1). As visitors continue through the gallery a large touch-table display draws their

attention. Gathered around the table are several other visitors gazing down at the glass screen.

The same mummy is shown on the surface of the table, but, as a visitor moves a virtual slider on the table, the muscles, organs, and skeleton reveal themselves as the skin is gradually peeled away. Another visitor turns the mummy around to explore the other side. Someone touches the information icon near the left shoulder blade to discover that the cut on his back, as well as the damaged underlying shoulder blade and fourth rib, are the result of a single penetrating wound. The his-

Figure 1. The Gebelein Man (identified as human mummy EA 32751) is a pre-dynastic natural mummy and thus not prepared for artificial mummification, as later tradition directed. After being on display at the British Museum for more than 100 years, it was carefully transported to the Bupa Cromwell Hospital, London, U.K., for CT scanning.



Photo courtesy of the Trustees of the British Museum.

Figure 2. Visitors to the Mediterranean Museum in Stockholm can explore and virtually touch the coffins, wrapping, and body of the Neswaiu mummy, an Egyptian priest from Thebes, circa 300 B.C. This realistic tangible interaction is made possible through the integration of a range of technical developments, including high-resolution CT X-ray scans, laser scanning, photogrammetry, and modern GPUs.



Photo courtesy of Ove Kaneberg/Medelhavsmuseet.

tory of the Gebelein Man unfolds for the visitor in the same way researchers have used visualization to establish the cause of death; the Gebelein Man was murdered, with a metal blade the most likely weapon.²

The technology allowing visitors to the British Museum to explore the Gebelein Man is an image-generation technique called “volume rendering” that is efficiently executed on graphics processing units (GPUs). A stack of thousands of 2D images, as generated by modern computed tomography (CT) X-ray scanners, is processed in parallel to interactively create images. During the past decade, various museums have begun to scan human remains and artifacts from their collections. Not only are mummies scanned, so too are meteorites, bee flies, gecko lizards embedded in amber, and much more in 3D; see a 2013 report from the Smithsonian Institution.⁴

This article describes the underlying research and development of algorithms and technologies that have become the basis for a software solution for multi-touch tables. It also describes how the workflow of scanning, curating, and integrating the data into the overall creation of stories for the public can lead to engaging installations at museums around the world. Our work is an example of how computing technologies, especially in computer graphics and visualization, allow a general audience to interact with and explore the very same data that only scientists and domain experts have previously been able to study. Figure 2 shows children exploring the Neswaiu mummy at the Mediterranean Museum in Stockholm.

State-of-the-Art Scanning

The first commercial CT machinery was invented by Sir Godfrey Hounsfield (who was co-winner of the 1979 Nobel Prize for Physiology or Medicine) and made public in 1972. A tomography image of 80×80 pixels initially took 2.5 hours to compute, image-acquisition time not included. Fast-forwarding through four decades of development of detector materials, computer hardware, and image-reconstruction algorithms, a modern CT scanner can scan an entire body in fewer than five seconds, generating up to 10,000 slices, with down-to-0.3mm slice thickness.

The latest technology enables scanning of the body with multiple simultaneous energy levels. Multiple energy levels enable better tissue separation and, at the same time, give rise to new challenges related to visualization of this new type of data. Also, the radiation dose has been reduced approximately 70-fold over the past 10 years. A heart scan today results in less than one month worth of average background radiation.

The increased availability of CT scanners in modern hospitals and radiology clinics has provided an opportunity to scan the fragile mummies and learn about and reveal their interior content and structure without having to open or damage them. Not even the wrapping has to be touched. The first report on CT scanning of mummies was presented in 1979 by Harwood-Nash.⁷ Scanning protocols for mummies require custom settings, as the body is completely dehydrated and regular CT protocols assume naturally hydrated tissues, as discussed by Conlogue.⁵ Furthermore, standard protocols aim to minimize the radiation dose, and parameters (such as slice thickness) are set to the largest value that still yields sufficient image quality for clinical diagnosis. The scanning settings need a custom configuration to yield the best image quality. For mummies, the radiation dose is generally viewed as less of an issue, but some scholars have begun to raise concerns about DNA degradation.⁶

The presence of metallic objects (such as amulets) can pose a challenge to CT scanning, as these objects potentially block X-ray radiation from reaching the sensor array, and the signal reconstruction suffers from artifacts. Such artifacts can now be reduced with state-of-the-art reconstruction algorithms but may also be reduced with modified scanning protocols and multi-energy CT. As museums around the world are increasingly digitizing their collections, a shift toward dedicated CT scanners (such as industrial and micro-CT scanners) is a possible development.

Rendering What Hides Inside

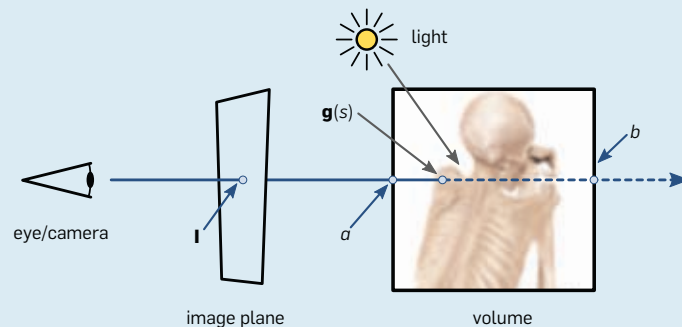
The data generated from a CT scan after reconstruction is simply a stack of tomographic images, where the pix-

Historical Notes on Virtual Autopsy

What would eventually become the Inside Explorer visualization table began in 2004, when the Research Group in Scientific Visualization at the Department of Science and Technology, Campus Norrköping, Linköping University, Sweden, began a collaboration with the Center for Medical Imaging and Visualization at Linköping University on a newly started project on virtual autopsies. The need to scan and visualize an entire cadaver before the forensic autopsy posed technical challenges. The most critical were how to deal with the large amounts of data and how to interactively view it when the available memory in a typical workstation was not enough to hold the data of an entire body. This challenge drove our thinking and development of fundamental data-handling components and rendering techniques to support full-body virtual autopsy visualization.¹² The resulting software laid the groundwork for the initial versions of the Inside Explorer Visualization table. It essentially addressed the challenges by adjusting the local level of detail based on the perceived visual impact on the final image and adapting image quality at runtime during interactive exploration. Much research has been devoted over the past 20 years to developing and refining DVR, as well as managing increasingly large and complex datasets, as surveyed by Beyer et al.³ in 2015.

In the virtual autopsy project, we put each forensic case with casualties through a full-body CT scan prior to the traditional clinical autopsy of the cadaver. Since inception of the virtual autopsy project in 2003, approximately 800 cases have been processed at the Center for Medical Imaging and Visualization, and all cadavers now routinely undergo postmortem full-body CT imaging when there is suspicion of murder. A large body of knowledge has been established, and various studies have been conducted for specific types of autopsy examinations, including classification of dental fillings. This work has been beneficial, if not pivotal, for the subsequent scanning of mummified cadavers. The knowledge gained from scanning and visualizing cadavers could be translated and adapted into scanning and visualizing mummies and other interesting artifacts at the world's premier museums.

Figure 3. The volume-rendering integral. On a modern GPU, thousands of these ray integrals can be computed in parallel, enabling full HD resolution rendering at 30Hz–60Hz.



els within each image slice become samples, or voxels, in a 3D volume. The rendering of images of the volumetric dataset on a computer screen is known as volume rendering,¹³ or direct volume rendering (DVR), as the image is generated directly from the data without any intermediate (polygonal) representation. The most common approach to DVR is raycasting, where the stack of images is traversed by simulated rays that collect the color contribution, $g(s)$, from each sampled point as the rays traverse the stack (see

Figure 3). This approach represents an excellent case of a massively data parallel algorithm that fits modern GPU architectures, with thousands of parallel processors, that execute threads of ray casting programs. Each ray estimates how simulated light is reflected and attenuated as the ray propagates from the user's eye through the volume,^a and the final footprint of

^a This follows the principle of reverse ray tracing and allows for several optimizations of the computation.

Figure 4. The four different TFs designed for the Gebelein Man are shown together in this blend of four screenshots, where angled lines separate the different settings, from left to right: skin, soft tissue, soft tissue and skeleton, and skeleton. As the user slides the bottom bar, the TF changes gradually in the volume as well, unveiling the next layer.

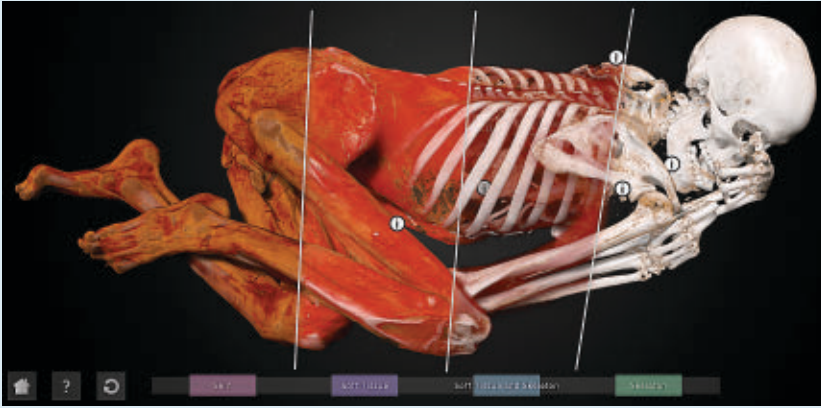


Figure 5. Including complex shadows and lighting effects can provide superior cues for depth, structure, and texture of the data. Interactive performance requires simplifications and approximations in the rendering algorithms. This image shows a state-of-the-art interactive volume illumination and rendering of the Gebelein Man mummy based on the photon mapping algorithm.⁹



Dataset courtesy of the Trustees of the British Museum; images by Daniel Jönsson.

Figure 6. Illustration of the simplified user interaction with the touch table: rotations are restricted to the horizontal axis; clipping planes can be used from top and bottom of the screen; zooming is achieved with the familiar two-finger pinch/stretch; a simple slider selects presets of the TF; and embedded points of interest are displayed according to the current context.

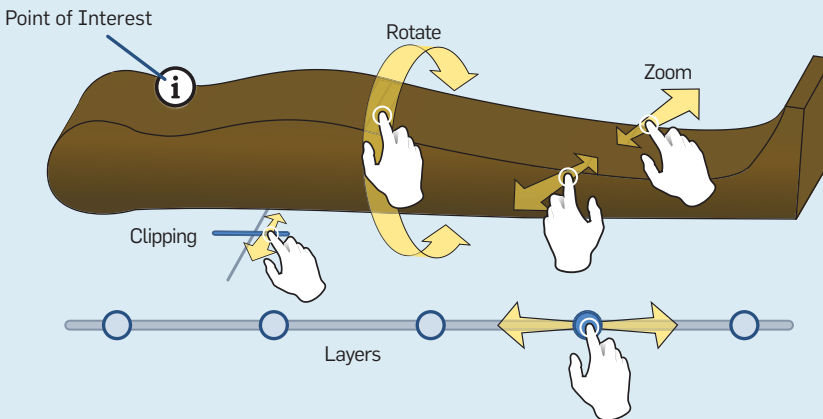


Illustration by Martin Falk.

color and opacity, I , is then deposited onto a pixel on the screen

$$I = \int_a^b g(s) e^{-\int_a^s \tau(u) du} ds$$

Computing such a ray for every pixel on the screen showing the image, and doing so at least 30 times per second, results in an interactive visualization of the interior of the scanned body or artifact. However, what is revealed depends on the user's settings (such as viewpoint and orientation) and, most important, how the X-ray absorption values are translated into colors and opacities. The function that describes this translation is referred to as the “transfer function” (TF), essentially included in $g(s)$ for the emission/absorption of light contribution and $\tau(u)$ for the density-causing attenuation. The TF is one of the most central concepts in volume rendering. Here, it controls whether bones, muscles, fat, or various soft tissues will be visible or not, and to what degree (see Figure 4). As different tissues have varying degrees of overlapping ranges in the data domain, the design of TFs is also one of the major algorithmic challenges, as discussed in Ljung et al.¹¹

For a realistic high-quality final result, the ray-casting process must consider how light sources affect every sample and how samples globally affect each other in a participating medium. This multiple scattering effect is referred to as “volumetric illumination” and is technically embedded within $g(s)$ and can thus become highly computationally demanding to compute due to the recursive multiple scattering events occurring in reality. The implementation must therefore exploit several simplifications and approximations to maintain an interactive experience. There are numerous approaches to volumetric illumination, ranging from effective methods based on approximations of light transport to more physically based methods with much higher computational demand; see Figure 5 for an example of a more complex lighting technique. The methods used in the work presented here were all discussed through a 2014 survey on volumetric illumination techniques for interactive DVR.¹⁰

Touching Data: Creating an Interactive Narrative

To bring volume rendering into the galleries of public venues we used several building blocks to design and implement a prototype of a visualization table for volumetric data termed “Inside Explorer” (see the sidebar “Historical Notes on Virtual Autopsy”). As starting points we used: state-of-the-art DVR software based on our research and development for high-quality visualization of large-scale volume data; large display screens with multi-touch interfaces; interaction design targeted at minimizing visual clutter and hiding of advanced concepts (such as the TF); and challenging large-scale volumetric data obtained from the scanning of humans and animals, as well as various artifacts.

It should be noted that interactive DVR has rarely been deployed at public venues to enable a general audience to explore scientific content. It is primarily a tool for research and limited professional use, requiring complex settings and domain-specific knowledge. The freedom and flexibility provided in professional medical workstations and experimental research software pose interesting challenges. For example, see Yu et al.¹⁵ and Jian et al.⁸ who both use multi-touch tables but primarily target expert users. These tools allow free exploration desired by expert domain users. However, for the general public of novice users, built-in guidance, simplified visual representations, and interaction are all necessary for efficiently supporting interactive narrative exploration. Here, we provide several insights gained with respect to the constraints and demands imposed on visualization at public venues. Our discussion focuses on four critical areas.

Performance. In a public installation, all aspects of performance must be considered in order to optimize the user experience. Such installations require fast loading times, as case-specific data is being loaded at runtime, and fast rendering times, with at least 30 frames per second at consistent high image quality. There are several reasons behind this requirement. In contrast with the medical- and scientific-visualization domains, where accuracy is more important than fast rendering and responsiveness, users at public



A design constraint is the limitation of interactions that cause rapid context switches and dramatic scene changes.



venues are eager to get started, not necessarily technically knowledgeable, and can potentially lose patience and interest. They are also not aware of the demanding nature of the datasets that are being displayed and often compare performance with the latest computer games. However, the key realization for software developers with respect to performance is the extreme sensitivity of human perception to large-scale touch interfaces and lagging visual updates. While the system is capable of creating the notion of touching a real artifact, that experience is lost as soon as users perceive lag in the interaction or rendering is perceived.

The Gebelein Man dataset has a size of $512 \times 512 \times 1,834$ voxels, 16-bit integers, and rendering image size of $1,920 \times 1,080$ (full HD). To ensure high frame rates with sufficient image quality, the renderer employs a set of optimizations: generation of a tight bounding surface for ray entry points and skipping empty blocks of data; termination of ray integration when opacity is saturated along the ray; and use of fast simplified shadowing using a “planesweep” technique that diffuses shadows from the light source and creates a shadow volume.¹⁴ The frame rates achieved range from 32Hz to 60Hz for the four different TFs used, as in Figure 4.

Robustness. From both a hardware and a software perspective a public environment poses particular robustness challenges in terms of number and type of user, hours of operation, and physical handling. The software must be almost as robust as high-availability systems and not provide users any opportunity to bypass the user interface and access lower layers of data or software. During hours of operation, the equipment must be able to endure handling by a wide range of visitors and sustain substantial wear and tear, as well as intentional damage caused by some of the visitors. Our experience is that multi-touch interfaces are indeed very robust, with all integrated technology hidden behind tempered glass and difficult to damage, even in environments like the British Museum, the leading visitor attraction in the U.K., with more than six million visitors per year.

Interaction. At public venues, the threshold time for “walk up and use”

Figure 7. By providing interactive narrative exploration and information at key points of interest, museum visitors are free to explore the narrative at their own pace and in their own sequence. The broken bones in the shoulder of the Gebelein Man were not the only injury visitors discovered. The posterior part of the fourth rib was also damaged. This discovery revealed a 5,500-year-old murder.



Photo courtesy of Interactive Institute Swedish ICT / Interspectral.

Figure 8. Workflow of Inside Explorer, from initial data acquisition and curation in the collection phase, to exploration, where scientific discoveries and annotations are added to the data. Production designers then provide their interpretations in the production stage, further enhancing the data with stories and visual design. User feedback and interaction patterns are then collected during exhibition.

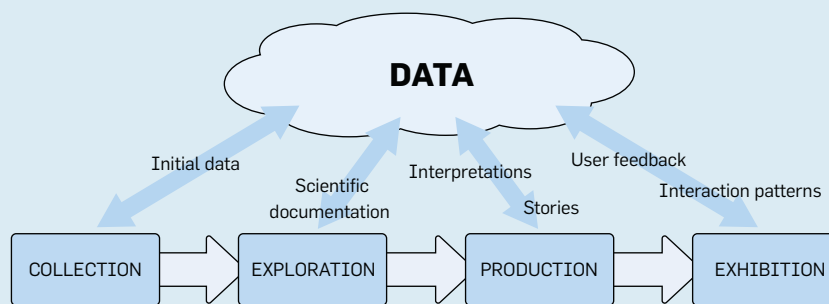


Illustration by Martin Falk.

must be close to zero. In a matter of a few seconds the user must be able to grasp how to interact, how to rotate the objects without losing control, zoom in and out, and change rendering parameters to reveal the data of interest. During the course of the project's development, touch interfaces have rapidly become commonplace, and the potential user base spans all of modern society. This combination is beneficial since the same user interface must accommodate young

children, seniors, and people from varying cultural backgrounds and physical ability. In view of this universal public expectation, we chose multi-touch surfaces as our primary interaction methodology at an early stage of the project. Our ambition with the user interface design was to enable intuitive and explorative interaction with a virtual representation while still allowing users to maintain a mental connection with the corresponding real artifact or subject (see

Figure 6). This has led us to apply design principles that can be characterized by the following parameters:

Object focus. The main focus of the representation on the screen should always be the rendered object, thus placing serious constraints on the choice of interaction paradigms and the implementation;

Judicious interaction constraints and micro-level interaction freedom. Interaction must be perceived as free but still constrained to noninvasively guide the user. We thus chose to constrain global interaction but increase freedom as points of interest are reached;

Minimalistic icons. Icons should be sparse and blend into the scene so as not to interfere with the object focus; they should also be associated with interaction points on the touch surface; and

Limiting multiuser interaction. When multiple eager and uninitiated users are interacting simultaneously, registration of unwanted interaction can occur frequently; a design constraint is thus the limitation of interactions that cause rapid context switches and dramatic scene changes. An example of the application of our principles is reflected in the TF interface design. In general, such design is a challenging task and, as noted earlier, is a major topic in the visualization research community. To hide the complexity of TF interaction, we use a range of predefined static TF settings from which the user can choose. By arranging these fixed TF settings along a slider it becomes intuitive and simple for users to peel away layer by layer of the content. The metaphor also maps easily onto polygonal geometries like coffins and chests.

A common technique used in volume rendering is known as "clipping planes" to make cutouts and reveal the interior of an object. From the user perspective, the planes are depicted as a pair of virtual scissors that enable virtual dissection of the subject/object. It is a simple yet powerful technique that is both informative and visually exciting when the user is able to discover content inside the skull or underneath the skin. To minimize usage complexity, we limit the interaction to a few fixed clipping directions with reduced flexibility. This limited interaction allows the user to explore while still making sure that focus and context are not lost due to excessive clipping and an

overly complex user interface. For ease of use, rotation is limited to a single axis, and when zooming and clipping the volume, that axis is always automatically centered on the screen.

Narrative. While adhering to the design principles described in the earlier section on interaction, the system must be able to support narratives, sometimes at multiple levels—the narrative of the case being shown, as well as the narrative of the scientific discovery and the user’s own experience of discovering the content.

At the heart of the interactive narrative exploration are the results of the work conducted after the capture and reconstruction of an object or body when curators analyzed and annotated the volumetric data. They can make

discoveries that need to be included, sometimes leading to the definition of a region or point of interest and a description of what findings have been made. The curators provide the fundamental scientific attribution to the data, documenting it such that exhibit producers can tailor the narrative and adapt the content to the public audience; see Figure 7 for an example. It is important for the software developer to realize that the narratives are implicitly embedded in the whole application, from visual settings to enabled interaction modes. For explicit information, we employ the notion of volumetric information hot spots that provide contextualized information about the scanned artifacts and subjects. During exploration, information spots

appear in the different layers, providing the user with contextual information, as well as details about particular findings. This information enables guided exploration through the datasets, as hints to other points of interest can be given.

Lessons Learned

One key to the successful installation of Gebelein Man is careful iterative interaction between the museums and the development team of programmers, modelers, and interaction designers. For example, in the design process for the Gebelein Man, we closely collaborated with the British Museum curatorial teams and the Interpretation and Learning Departments, as well as access managers; see Figure 8 for

Figure 9. The digitization of Neswaiu included multiple CT scans of the body as well as photogrammetry and laser scanning of the wooden sarcophagi in which he was buried. On the visualization table, both the textured polygonal models and the volume-rendered remains of the mummy are shown as different layers that can be peeled away, one at a time.



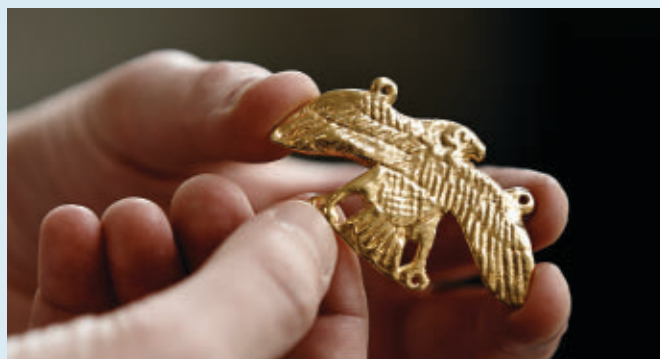
Left photo courtesy of Ove Kaneberg/Medelhavsmuseet;
right image courtesy of Interactive Institute Swedish ICT/Interspectral.



Figure 10. Reproducing hidden artifacts is another benefit of the process of digitizing museum collections. Here, a falcon amulet from the Neswaiu mummy in Stockholm is shown. It was produced through inexpensive 3D printing of a plastic model used to create a mold in which a gold-plated bronze copy was cast.



Image and photo courtesy of Interactive Institute Swedish ICT/Interspectral.



a schematic of the adopted workflow. The iterative approach was based on frequent videoconferencing, as well as regular on-site meetings and design reviews, leading to refinement of the software, as well as the physical exhibit design, to improve aspects of the interactive public experience (such as accessibility to visitors in wheelchairs).

The final Gebelein Man exhibit in the Early Egypt Gallery at the British Museum was subject to an evaluation by the museum's Interpretation Department prior to, as well as after, installation of the Inside Explorer visualization table to determine its contribution to the visitor experience.¹ In the pre-table evaluation, we tracked 114 visitors and interviewed more than 50 of them. In the post-table evaluation, we tracked 133 visitors and interviewed 95 of them. Here are some of the main findings:

Time spent in the gallery. Prior to table installation, the median time spent in the gallery was two minutes. After table installation, visitor dwell time in the gallery increased by 40%, and visitors made an average of three stops (instead of two) in the gallery. An important finding was that the fraction of visitors stopping to view other displays in the gallery increased from 3% to 20%;

Visitor attention to original display. We were initially concerned that the virtual visual exploration table would divert interest from the physical artifacts and mummy in the gallery, but the opposite occurred, where the fraction of visitors viewing the physical display increased from 59% to 83%. The total visitor attention time given to Gebelein Man was shared between the table and the mummy, leading to a decrease in time spent observing the physical subject alone; and

Interaction with the table. The table saw heavy visitor interaction and was in use 95% of the evaluation time. Of the visitors to the gallery, 83% stopped at the Gebelein Man display, 59% stopped at the Inside Explorer table, and 36% engaged with the table. The average interaction time was two minutes, nine seconds. In our interviews, the display was scored nine out of 10 on average for ease of use and enjoyment and an average of 8.5 for informativeness. Approximately 60% of table users discovered the cause of death.



The amulet of Neswaiu, still present within the mummy wrappings and behind glass, can now be physically reproduced in high-resolution detail through a process of digital virtual extraction.



Overall, the visualization table had a positive effect on the display of Gebelein Man and the gallery as a whole. It increased the attention-grabbing power of the display by 24% over surveys before table installation. Our visitor interviews revealed that the Inside Explorer visualization table had contributed to visitors' understanding of the mummy, with many returning to observe it after having discovered information on his age, preservation, and cause of death. The Inside Explorer table also appears to have had a positive effect on the other physical displays and cases in the gallery. Based on these conclusions, the British Museum decided to turn the temporary installation into a permanent exhibit.

Even though these lessons learned are described primarily in the context of our work with the British Museum, we see similar results in many other installations worldwide and regard our work as only the beginning of the exciting possibilities promised by interactive visualization at public venues.

The Future of Science Visualizations at Public Venues

The Inside Explorer visualization table described here is an example of an ongoing trend that is fundamentally changing the visitor experience in museums and science centers.⁴ Instead of passively absorbing information and studying static objects or subjects on display, visitors are allowed to engage with digital objects and explore the secrets hidden within an object in a manner similar to the way researchers would process and explore the real object and its digital representation. We curated and annotated the datasets and provided additional information to furnish a complete story for the user experience with the table.

Based on the work presented here, we now intend to take another step, aiming to develop a solution that addresses the bottlenecks in the process to enable rapid and efficient conversion of a scanned subject into a compelling visitor experience. The first bottleneck is in the deeper initial exploration of scans of the subject to facilitate discovery. Deep technical knowledge of the process should not be required, as it is today, making it possible to address

a wider group of domain experts, as well as non-experts. The second main production bottleneck is the process of closely linking any scientific or curatorial findings with the process of interpretation and exhibit production. The envisioned solution will thus support the whole chain, from scientific data collection and documentation of scientific discoveries, design, and interpretation of the information into non-linear stories, to production of end-user applications for exhibits. Figure 8 outlines how we refined, curated, and annotated the data, though the original data is the basis of every step in the process of developing the visitor experience in the gallery.

We have also taken initial steps to explore other opportunities offered by the digitization of museum artifacts, including laser scanning and photogrammetry, employed for the mummy of Neswaiu, as in Figure 9, and also 3D printing technologies of the scanned data. For example, the amulet of Neswaiu, still present within the mummy wrappings and behind glass, can now be physically reproduced in high-resolution detail through a process of digital virtual extraction (see Figure 10). Visitors can clasp and feel it and explore it physically, even though the mummy has never been unwrapped.

What does this development mean for the future of science visualization at public venues? One important aspect of sharing visualizations with the public may be authenticity of the science and resulting data. The public can interact with and explore the real underlying data, taking the visitor experience to a new level of engagement. The scientific story can be embedded in the interactive narrative on the table using points of scientific interest and told through the framing of the table installation with additional information about the process, from scan to installation.

Conclusion

Our findings demonstrate that science visualization has the potential to narrow the gap between the general public and research, as it allows scientists and curators to share the methods used to interpret and analyze the collections with visitors (such as determination of the age and sex of a mummy). Bringing

the original research data to the public and providing tools enabling learning and exploration is an exciting and challenging scientific adventure that creates many new research opportunities involving not only computer science as enabling technology but also learning and communication, social science studies, and domain-specific areas from which the objects of study originate.

Acknowledgments

The authors wish to thank the Trustees of the British Museum. The Gebelein Man dataset is courtesy of the Trustees of the British Museum. We also wish to acknowledge the help and input from the Department of Ancient Egypt and Sudan at the British Museum. We also thank Medelhavsmuseet/the Museum of Mediterranean and Near Eastern Antiquities in Stockholm, Sweden. We also thank the team from Interactive Institute Swedish ICT/Interspectral, specifically David Karlsson, Claes Ericson, Karl Lindberg, and Kristofer Jansson. Finally, we thank the Norrköping Visualization Center C and the Center for Medical Image Science and Visualization, Linköping.

This work is supported by Swedish eScience Research Center, Excellence Center at Linköping-Lund in Information Technology, Knut and Alice Wallenberg Foundation, Foundation for Strategic Research, The Knowledge Foundation, Swedish Research Council, and Vinnova. 

References

1. Antoine, D. *Evaluation of the Virtual Autopsy Table (Room 64: Early Egypt Gallery)*. Internal Report, Supplemental Material, Interpretation Department, The British Museum, London, U.K., Feb. 2016.
2. Antoine, D. and Ambers, J. The scientific analysis of human remains from the British Museum collection: Research potential and examples from the Nile Valley. In *Regarding the Dead: Human Remains in the British Museum*, A. Fletcher, D. Antoine, and J. Hill, Eds. The British Museum, London, U.K., 2014, 20–30; https://www.britishmuseum.org/PDF/Regarding-the-Dead_02102015.pdf
3. Beyer, J., Hadwiger, M., and Pfister, H. State of the art in GPU-based large-scale volume visualization. *Computer Graphics Forum* 34, 8 (Dec. 2015), 13–37.
4. Clough, G.W. *Best of Both Worlds—Museums, Libraries, and Archives in a Digital Age*. e-book. Smithsonian Institution, Washington, D.C., 2013; <http://www.si.edu/bestofbothworlds>
5. Conlogue, G. Considered limitations and possible applications of computed tomography in mummy research. *The Anatomical Record* 298, 6 (June 2015), 1088–98.
6. Grieshaber, G.M., Osborne, D.L., Doubleday, A.F., and Kaestle, F.A. A pilot study into the effects of X-ray and computed tomography exposure on the amplification of DNA from bone. *Journal of Archaeological Science* 35, 3 (Mar. 2008), 681–687.
7. Harwood-Nash, D. Computed tomography of ancient

Egyptian mummies. *Journal of Computer Assisted Tomography* 3, 6 (Dec. 1979), 768–773.

8. Jian, W., Tu, H.-W., Hua Han, X., Tateyama, T., and Chen, Y.-W. A preliminary study on multi-touch-based medical image analysis and visualization system. In *Proceedings of the Sixth International Conference on Biomedical Engineering and Informatics (BMEI 2013)* (Hangzhou, China, Dec. 16–18). IEEE, 2013, 797–801.
9. Jönsson, D., Kronander, J., Ropinski, T., and Ynnerman, A. Historygrams: Enabling interactive global illumination in direct volume rendering using photon mapping. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2364–2371.
10. Jönsson, D., Sundén, E., Ynnerman, A., and Ropinski, T. A survey of volumetric illumination techniques for interactive volume rendering. *Computer Graphics Forum* 33, 1 (Feb. 2014), 27–51.
11. Ljung, P., Krüger, J., Gröller, E., Hadwiger, M., Hansen, C.D., and Ynnerman, A. State of the art in transfer functions for direct volume rendering. *Computer Graphics Forum* 35, 3 (June 2016).
12. Ljung, P., Winskog, C., Persson, A., Lundström, C., and Ynnerman, A. Full-body virtual autopsies using a state-of-the-art volume-rendering pipeline. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (Sept.–Oct. 2006), 869–876.
13. Max, N. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics* 1, 2 (June 1995), 99–108.
14. Sundén, E., Ynnerman, A., and Ropinski, T. Image plane sweep volume illumination. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec. 2011), 2125–2134.
15. Yu, L., Svetachov, P., Isenberg, P., Everts, M.H., and Isenberg, T. FI3D: Direct-touch interaction for the exploration of 3D scientific visualization spaces. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (Nov. 2010), 1613–1622.

Anders Ynnerman (anders.ynnerman@liu.se) is a professor focusing on scientific visualization and head of the Media and Information Technology Division of Linköping University, Linköping, Sweden, and director of the Norrköping Visualization Center C, Norrköping, Sweden.

Thomas Rydell (thomas.rydell@interspectral.com) is the CEO and co-founder of Interspectral AB, Norrköping, Sweden, and formerly studio director of the Interactive Institute Swedish ICT, Norrköping, Sweden.

Daniel Antoine (dantoine@thebritishmuseum.ac.uk) is the curator of physical anthropology at The British Museum and an honorary senior research fellow in the Institute of Archaeology, University College London, U.K.

David Hughes (david.hughes@interspectral.com) is a technical advisor and strategic consultant at Interspectral AB, Norrköping, Sweden, and previously held the same position at the Interactive Institute Swedish ICT, Norrköping, Sweden.

Anders Persson (anders.s.persson@liu.se) is a professor of radiology at Linköping University, Linköping, Sweden, and the director of the Center for Medical Image Science and Visualization, Linköping, Sweden.

Patric Ljung (patric.ljung@liu.se) is a senior lecturer in immersive visualization at Linköping University, Linköping, Sweden, and research coordinator at Norrköping Visualization Center C, Norrköping, Sweden.

Copyright held by the authors.



Watch the authors discuss their work in this exclusive *Communications* video. <http://cacm.acm.org/videos/interactive-visualization-of-3d-scanned-mummies-at-public-venues>

DOI:10.1145/2950041

Biologically based computational modeling promises virtual characters capable of face-to-face human interaction.

BY MARK SAGAR, MIKE SEYMOUR, AND ANNETTE HENDERSON

Creating Connection with Autonomous Facial Animation

OF ALL THE experiences we have in life, face-to-face interaction fills many of our most meaningful moments. The complex interplay of facial expressions, eye gaze, head movements, and vocalizations in quickly evolving “social interaction loops” has enormous influence on how a situation will unfold. From birth, these interactions are a fundamental element of learning and lay the foundation for successful social and emotional functioning through life.

What are the underlying processes from which this most human form of interaction emerges? Will we be able to interact with computers in a face-to-face way that feels natural? This article discusses the unique challenges of realistically simulating the appearance and behavior of the face to create interactive autonomous virtual human models that support naturalistic learning and have the “illusion of life.” We describe our recent progress toward this goal with “BabyX,” an

autonomously animated psychobiological model of a virtual infant. While we explore drivers of facial behavior, we also expect this foundational approach has the potential for more “human” computer interfaces. We also describe our work on our “Auckland Face Simulator” we are developing to broaden this work beyond infants and give a more realistic face and a greater biological basis to adult conversational agents.

Simulating the face has great potential for human-computer interaction (HCI), as it increases the available communication channels between humans and machines in an intuitive, accessible way. But it is also a vehicle with which to explore our own nature. Akin to developmental robotics,⁶ which explores ways of learning and mental development through child-like robots, simulating the underlying processes driving the face during social interaction will enable HCI researchers to explore behavioral and learning models involving naturalistic face-to-face interaction.

There is a trend in the game and visual-effects industries to create ever more realistic animated characters, especially humans, but it turns out to not be a straightforward transition from the stylized faces of traditional animation. For these industries, realism of appearance and movement is very important, evidenced by the large financial investment going toward creating the most realistic illusion they can achieve. This is done presumably because the experience becomes more immersive and powerful the closer it is

» key insights

- The expressive and communicative power of the face has been untapped in HCI but can indeed create deeper human-machine connection and engagement.
- The holistic interplay of biologically based computational behavioral models driving a virtual character can give rise to an emotionally affecting experience.
- We are developing such a psychobiological modeling framework for autonomous characters and related HCI with realistic yet virtual faces.



Screenshot of BabyX version 4 (under development) looking at a user; image rendered in real time.

to reality. Realism lessens the leap required in the suspension of disbelief; it is closer to personal experience. Studies comparing children viewing realistic vs. cartoon depictions of aggression show realism to have an increased affective effect on subsequent behavior.³⁵

As characters move to being more human-like, discrepancies in appearance and behavior tend to alarm us, with an increasing chance of falling into the “uncanny valley.”²² The most critical component to “humanness” is the face. Achieving the illusion of life in the face of a realistic digital character is challenging in a passive medium (such as film) but even more so in interactive games and simulations. Concerning realism in computer-generated faces, Gopnik¹² noted: “It made sense to think that the ability to reason and speak was at the heart of the human mind. Turing’s bet was that a computer that could carry on a conversation would be convincingly human. But the real imitation game of digital-effects movies suggests that the ability to communicate your emotions may be even more important. The ineffable, subtle, unconscious movements that

tell others what we think and feel are what matter most. At least that’s what matters most to other human beings.”

In an interactive scenario involving unscripted life-like interactions, the problem is even more challenging. The issue is not only how to represent the complex appearance and movement of the synthetic face in real time, but, for an embodied agent’s behavior to be believable, it must be consistent and contextually appropriate.¹ If the simulated human can be affected by the interaction, and respond in a way that can affect the user and vice versa, then each partner is more invested in how the interaction unfolds, creating engagement and emotive connection.

What creates all these fleeting movements that communicate so much? And how can a simulation keep them consistent, appropriate, and adaptive? Everything that happens on the face reflects a brain-body state. Because the behavior of the face is affected by so many factors—cognitive, emotional, and physiological—we explore a more-detailed and lower-level biologically based approach than has previously been attempted in facial animation.

Here, we introduce our general approach and design of a modeling framework we call “Brain Language” (BL) to create autonomous expressive embodied models of behavior driven by neural-system models based on affective and cognitive neuroscience theories. Our goal is to integrate different current theories and models to create a holistic “large functioning sketch” of basic aspects of human behavior, with a focus on the face and interactive learning.

Autonomous Animation

There is long-term interest in creating self-animating agents that project the illusion of life. In 1994, Bates² described the importance of appropriately timed and clearly expressed emotion to make a character seem alive. Terzopolous³⁶ introduced a holistic simulation of fish behavior in which each such behavior was driven by an abstracted brain. Terzopolous’s work was a closed system, based on initial environmental state, whereas in Maes’s ALIVE system,²⁰ the user was included in the loop with autonomously animated animals, including a virtual

dog “Silas” developed by Blumberg³ using sophisticated ethological models to simulate how animals are able to organize and adapt.

These and similarly inspired works are important on many levels, as they explain how behavior can emerge, made observable through animation with constraints. Blumberg³ suggested for a creature to appear alive it must react, have goals, make choices, convey its intentionality, emotionally respond to events, adapt, and vary its movement and response.

For autonomous animation of the face in real time, Terzopolous and Lee³⁷ developed a physics-based face model driven by a basic behavioral animation model. Despite this pioneering work, few other virtual human studies have focused on this lower level of detail in real-time facial animation.

Most research in building autonomous human agents has been as “embodied conversational agents” (such as in Allbeck¹ and in Cassell⁷) at a generally more phenomenological and higher level, not specifically focused on the subtler details of facial expression and nonverbal behavior; Vinciarelli et al.³⁹ included a survey of social-signal processing in computer interaction. Simulating these signals is getting greater attention from the interdisciplinary “intelligent virtual agent” community,⁵ exploring agents that are capable of real-time perception, cognition, and actions in the social environment; Marsella and Gratch²¹ discussed simulations of psychological theories of emotion. Emotion-oriented APIs (such as SEMAINE) have been developed.³³ And Scherer³² showed how cumulative effects of sequential checks of an eliciting event, mediated by autonomic and somatic components, might combine to create compound facial expressions.

Much of the work on virtual humans has an unfortunately robotic “feeling,” particularly with facial interaction. This is possibly due to most virtual human models not focusing on the microdynamics of expression or on facial realism. These microdynamics are considered particularly critical in learning contexts. Rohlfling and Deak²⁷ stated: “When infants learn in a social environment, they do not simply pick up information passively. They respond



The dynamic behavior of the face emerges from many systems interacting on multiple levels, from high-level social interaction to low-level biology.



to, and learn from, the interaction as they jointly determine its content and quality through real-time contingent and reciprocal coaction.”

Another important factor in the acceptance of a virtual human face is its visual quality. However, realistically simulating a face—even when still—has proved to be formidable.

Challenges of Modeling the Human Face

The way a digital face moves and appears can cause unwanted effects. Rather than aiding the appearance of life, a partially realistic solution can elicit a negative response—the uncanny valley effect. This response is thought to be triggered by any number of non-expected responses, alarming the viewer’s perceptual system.¹⁹ To avoid this response, many factors must be taken into account, including the physical appearance of the face and the eye-gaze movement of the skin, any of which can trigger a form of dissonance that interferes with the affinity of the perceived face.

Appearance. The ability to “read” faces is so important that several different parts of the brain play a role in face perception. We are sensitive to many factors that act as signs of health and vitality. People often refer to someone being “as white as a sheet,” “red faced,” or “sickly looking”; it is thus important to render physically plausible healthy skin with correct surface properties, detail, and subsurface scattering of light that provides diffuse properties of skin.

This challenge has been approached in two broadly different ways. First, by using “image-based methods” that sample the face under different lighting and viewing conditions⁹ and then render the face through a combination of weighted image-blended sets, photogrammetry, and/or image projection. Second, by using “parametric methods” that fit the captured data to a face and material model used during rendering, allowing for more flexibility but at the cost of potentially increased rendering complexity.¹⁶ Given the constraints producing imagery fast enough for user interaction, adding further to the complexity of achieving an effective interactive face, a simplified implementation of the second approach is

typically used for real-time rendering, as in Jimenez.¹⁵

Deformation. Achieving coherent movement of the skin is especially challenging due to the complex deformations in broad expressions and the highly non-linear motion of skin. Many computer-generated faces in games and films do not address these characteristics; for example, the lips on a character may move while the surrounding areas of the face remain static, causing an unnatural effect. Unlike skeletal muscles, facial muscles are embedded in the mobile facial tissue, meaning facial muscle activation must be treated as a system. Arguably the most coherent and generally useful way to drive facial animation is through parameterization of individual muscle activity (such as in Ekman and Friesen's Facial Action Coding System¹⁰).

The facial deformations used in animatable faces are typically represented through deforming geometry using weighted joints or weighted shape combination ("blendshape") methods.²⁵ While effective, these methods can suffer from combinatorial explosion in representing the complex range of facial expressions. The highest-quality models used in the visual-effects industry incorporate a large number of blendshapes to form linear approximations to non-linear deformations. Creating these models is labor intensive, so a number of researchers have approached the problem using flesh simulations.^{34,37,40}

Facial Motor System

To design an autonomous digital facial system, it is important to understand how faces are controlled. Traveling inward from the facial nerves, we reach the facial nucleus in the brainstem. The facial nucleus receives its main inputs from both subcortical and cortical areas through different pathways. Both a person's emotional and voluntary facial expressions seem to arise from different neural circuits.^{8,13} The implication is that the voluntary expression cannot access a genuine emotional motor pattern and is why it is not possible to fully produce a genuine emotional expression through volition. Similarly, stroke patients with damage to certain primary motor and pre-motor areas cannot produce a sym-

metrical voluntary smile yet can smile normally in response to jokes.¹³

Expressions are generated by neural patterns in both the subcortical and cortical regions. In the subcortical area, circuits include those for laughing and crying. Evidence suggests certain basic emotional expressions like these do not have to be learned. In comparison, voluntary facial movements (such as those involved in speech and culture-specific expressions) are learned through experience and predominantly rely on cortical motor control.

Our psychobiological facial framework aims to reflect that facial expressions consist of both innate and learned elements and are driven by quite independent brain-region simulations.

Building a Holistic Model

The human face mirrors both the brain and the body, revealing mental state (such as through mental attention in eye direction) and physiological state (such as through position of eyelids and color of the skin). The dynamic behavior of the face emerges from many systems interacting on multiple levels, from high-level social interaction to low-level biology.

To drive a biologically based life-like autonomous character, one would need to model multiple aspects of a nervous system. Depending on the level of implementation, a non-exhaustive list includes models of the sensory and motor systems, reflexes, perception, emotion and modulatory systems, attention, learning and memory, rewards, decision making, and goals. We seek to define an architecture that is able to interconnect all of these models as a virtual nervous system.

Several biologically inspired cognitive architectures have been developed; see Goertzel et al.¹¹ for a survey. Most are non-graphical, focusing on cognition over affect or physiological states. It makes sense that the more biologically based the architecture and the more realistic, the more it is ultimately likely to represent biological behavior. An example is the "Leabra" framework,²³ which constructs low-level biologically based neural network models and connects them to model higher-level aspects of cognition. This modeling approach is appealing for its ability to suggest how low-level biologi-

cal explanations link to high-level behavior (such as goal setting).

Building embodied nervous systems that can learn through real-time sensorimotor interaction is being explored in the field of developmental robotics.⁸ Social-interaction models have been explored with anthropomorphized "social robots" (such as in Leonardo and Kismet⁴). Developmental robotics, in particular, seeks to explore the theory of embodied cognition—how the mind develops through real-time sensorimotor interaction.

Our approach to building live interactive virtual agents takes a similar direction whereby we embody, through realistic computer graphics, a biologically based model of behavior. We ground experience through interaction and place particular emphasis on the importance of face-to-face interaction, which is difficult to achieve in robotics due to mechanical constraints. The result is a system that can be reduced to more biological detail, as well as expandable to incorporate higher-level complex systems. As there are many competing theories on how different brain and behavioral systems function, our choice is to opt for flexibility and develop a "system to build systems" in a Lego-like manner.

Brain Language

BL²⁸ is a modular simulation framework we have been developing for the past five years to integrate neural networks with real-time computer graphics and sensing. It is designed for maximum flexibility and can be connected with other architectures through a simple API. It consists of a library of time-stepping modules and connectors. It is designed to support a wide range of computational neuroscience models, as in Trappenberg.³⁸ Models supported by BL range from simple leaky integrators to spiking neurons to mean field models to self-organizing maps. These can be interconnected to form larger neural networks (such as recurrent networks and convolutional networks like those used in deep learning). Our main interest is in online learning, in which the network learns during live interaction from both spatial and temporal data. A key strength of BL is its tight integration with computer graphics as a visualization tool. Complex dynamic

Figure 1. BabyX's interactive brain: (left) superior colliculus activity driving visual attention is visible (green) in the brainstem; (middle) BL raster plot of neural activity and scrolling display of modulatory activity; (right) basal ganglia circuit and interactive dopamine level modification (green) affecting cortico-thalamic feedback and eye movement.

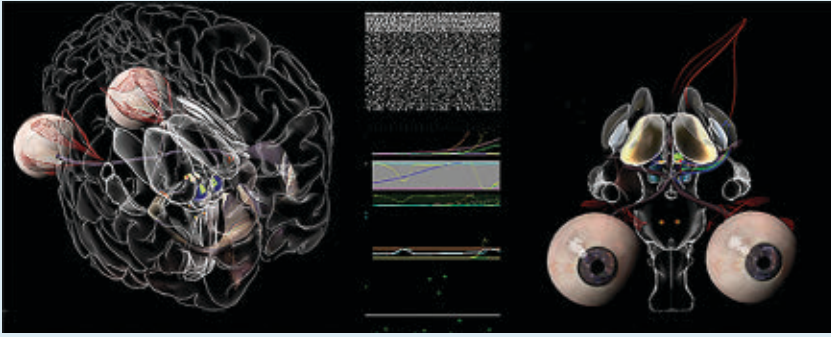
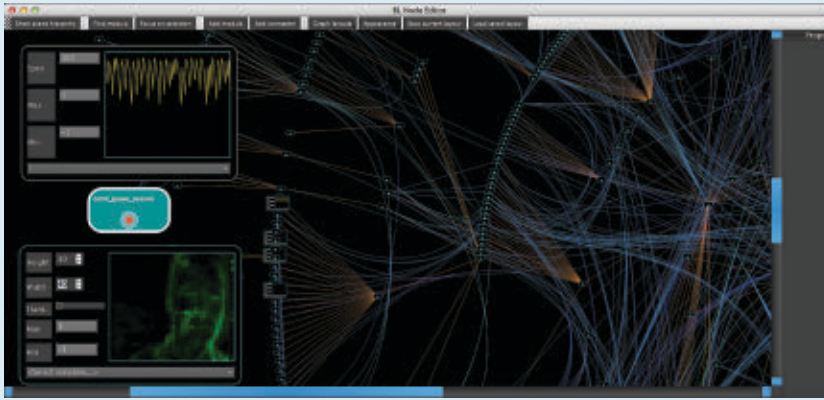


Figure 2. Screenshot of interactive BL viewer: (left) BL scopes viewing activity of a single neuron (top) or an array of retinotopic neurons (bottom) during a live interaction; (right) partial view of BabyX's virtual connectome, which can be explored interactively; connections light up (green or red) when activated.



networks can be visually investigated in multiple ways, either through 3D computer graphics (see Figure 1) or through a 2D schematic interface showing activity on the virtual connectome (see Figure 2). Individual variable activity (such as neuron voltage or firing rate) can be inspected with scopes during a simulation, as in Figure 2 left. The simulation can be interactively modified (such as by changing neural-model parameters while viewing the effect on the animation) (see Figure 3 right).

Sensory input is typically through camera, microphone, and keyboard to enable computer vision audition and “touch” processing, but data can be input from any arbitrary sensor or output to an effector through the API. Computer graphics output is through OpenGL and the OpenGL Shading Language. A key feature of BL is that any variable in the neural network system

can be shared and drive any aspect of a sophisticated 3D animation system; for more detail on BL, see Sagar et al.²⁸

BabyX Project

To illustrate how these concepts come together, we describe an experimental psychobiological simulation of an infant we call “BabyX,”^{29,31} that aims to embody models of interactive behavior and social learning to create an autonomous virtual infant one can interact with naturally.

Facial expression. At a conceptual level, BabyX's computer-graphic face model is driven by muscle activations generated from motor-neuron activity. The facial expressions are created by modeling the effect of individual muscle activations and their non-linear combination forming her range of expressions, as in Figure 3. The modeling procedures involve biomechanical

simulation, scanning, and geometric modeling. Fine details of visually important elements (such as the mouth, eyes, eyelashes, and eyelid geometry) are painstakingly constructed for life-like reality (see Figure 4).

A highly detailed biomechanical face model, as in Figure 3, has been constructed from MRI scans and anatomic reference, akin to Wu.⁴⁰ Skin deformation is generated by individual or grouped-muscle activations. We have modeled the deep and superficial fat, as well as muscle, fascia, connective tissue, and their various properties. We have used large-deformation finite-element elasticity⁴⁰ to deform the face from rest position through simulated muscle activation. Individual and combined muscle activations were simulated to form an expression space,³⁰ interpolated on the fly in BL as the face animates. The response to muscle activation is consistent skin deformation and motion.

Nervous system. BabyX's biologically inspired nervous system consists of an interconnected set of neural system and subsystem models. The models implemented so far are sparse yet span the neuroaxis and generate muscle-activation-based animation as motor output from continuously integrated neural network models. Due to the Lego-like nature of BL, we can have a closed-loop functioning system allowing experimental interchange of components while exploring different theoretical models. In total, the models aim to form a “large functioning sketch” of interconnected mechanistic systems contributing to behavior, containing both top-down and bottom up mechanisms interacting as an integrated system.

BabyX's neural networks and circuits implemented so far cover basic elements of motor control, behavior selection, reflex actions, visual attention, learning, salience, emotion, and motivation. An architectural diagram relating some of the key functional components, neuroanatomical structures, and functional loops is included in Figure 5; note cortical and subcortical input to the facial nucleus. A characteristic of this modeling approach is the representation of subcortical structures (such as the basal ganglia) and brainstem nuclei (such as the oculomotor nuclei). The structures are functionally implemented as neural

network models with particular characteristics (such as the amygdala and hippocampus as “hebbian associators” and the pulvinar as a topographically organized array of neurons forming a saliency map). Activation of the hypothalamus releases virtual hormones. Cortical regions use recurrent and multi-layer neural networks and self-organizing maps. Due to the Lego-like nature of BL, simple models can be replaced by more sophisticated models as they become available.

One of the goals of BabyX is to visually represent functional neural-circuit models in their appropriate anatomical positions. For example, our Basal Ganglia model (based on Redgrave et al.²⁶) controls motor actions and has an appropriate 3D location and geometry, as in Figure 1, and the activity of the specific neurons form inputs to the shaders to show the circuit in action as it processes.

Emotions in BabyX are, in fact, coordinated brain-body states that modulate activity in other circuits (such as increasing the gain on perceptual circuits). Emotional states modulate the sensitivity of behavioral circuits. For example, stress lowers the threshold for triggering a brainstem central pattern generator that, in turn, generates the motor pattern of facial muscles in crying.

Neurotransmitters and neuromodulators play many key roles in BabyX’s learning and affective systems.²⁴ An example of a physiological variable that affects both the internal and external state of BabyX is dopamine, which provides a good example of how modeling at a low level interlinks various phenomena. In BabyX, virtual dopamine plays a key role in motor activity and reinforcement learning. It can also modulate plasticity in the neural networks and have subtle behavioral effects such as pupil dilation and blink rate. The use of such low-level models means the user can adjust BabyX’s behavioral dynamics, sensitivities, and even temperament by adjusting virtual neurotransmitter levels.

Sensory input. BabyX takes audio-visual input Web camera and microphone, and “touch” from keyboard or touchscreen and is designed to work without special hardware. BL can interface to different devices, and the BabyX project exists separately from choice

of display systems (such as virtual reality, or VR, or augmented reality, or AR). Advances in AR, particularly in systems that allow for facial-expression tracking, accurate eye tracking, and depth gaze registration of the user, mean an obvious possible implementation

could be as a virtual agent in AR. Such an additional level of engagement would enhance the experience but also benefit from the tight emotional signaling feedback we have developed.

Learning through interaction. One motivation for the BabyX project is to ex-

Figure 3. BabyX (version 1). Detailed biomechanical face model simulating expressions generated from muscle activations.

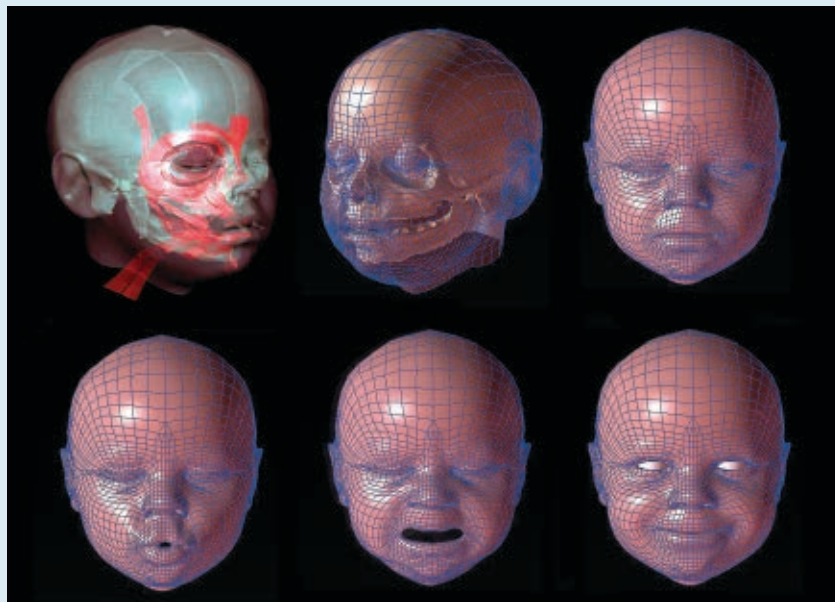
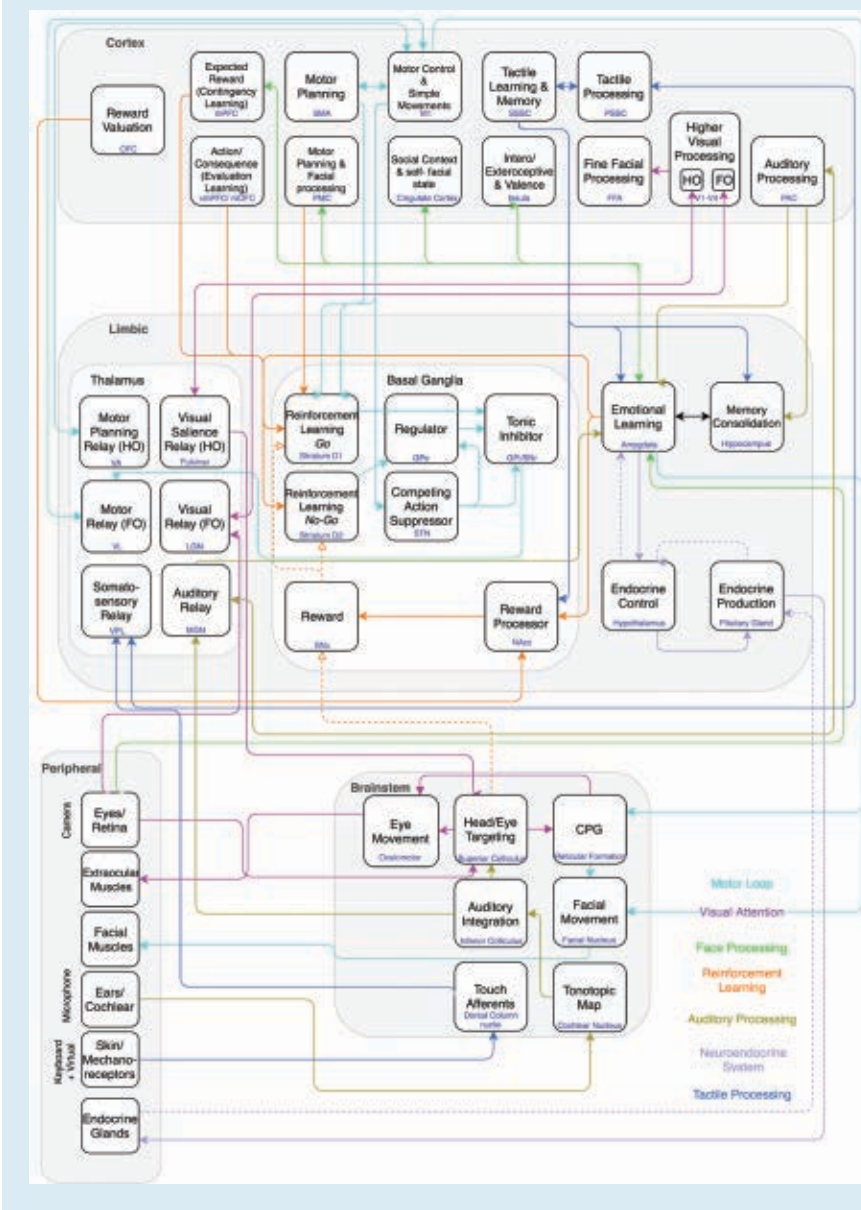


Figure 4. BabyX (version 4, under development). Screenshot from real-time interactive psychobiological virtual infant simulation.



Figure 5. Architectural diagram showing several key functional components and processing loops and their neuroanatomical equivalents (blue text).



ulated by physiological parameters leading to activation of facial motor pattern generators. BabyX may spontaneously make a puckered mouth shape, which is seen and then mirrored by the user or “caregiver.” For BabyX, the activity and feedback from this facial movement is associated with a delayed visual sensory pattern from the caregiver’s facial response. Associative learning is modulated by phasic dopamine. Strengthening bi-directional associative connections results in the caregiver’s expression being able to elicit a similar expression in BabyX. On successful imitation of an expression, the caregiver may praise with positive affect, releasing yet more dopamine, which further strengthens synaptic connections. The sensed positive affect and dopamine release activate and modulate affective circuits, causing neural-pattern generators to contract the muscles for smiling. For BabyX to learn an association of her muscle activity to the caregiver’s expressions, she must attend to the caregiver’s face. The caregiver’s actions cause activity on BabyX’s superior and inferior colliculus, where sensory events compete to drive the oculomotor network to move the eyes. A subregion of the camera input is automatically mapped to the virtual “fovea,” which maps to where the eyes are focusing.

In sum, the various circuits driving BabyX’s facial expressions converge on the facial nucleus in the brainstem that then activates BabyX’s animated facial muscles. Because the various inputs to the facial nucleus arise from the activity of independent yet interconnected networks, BabyX’s facial expressions can be understood in the context of internal activity and external factors.

While BabyX’s “default” expressions are associated with reflexes and basic affective states (such as newborns crying), the learned expressions here are voluntary. Dating to Charles Darwin, the “facial feedback hypothesis” states posing an expression can influence emotional state. By adding bidirectional connections to the affective networks that generated them, mimicked expressions can cause activity in BabyX’s affective circuits, functioning perhaps as a basis for “virtual empathy.”

explore how high-level social interactions might interact with models of lower-level biological mechanisms. An example illustrating such interaction we have investigated is facial mimicry (see Figure 6); for a related video, see <https://vimeo.com/123986611>. This may be simple to program at a high level in an embodied agent (such as by copying inputs to outputs), but when exploring how the interaction may emerge in a general sense, through biologically plausible intrinsic learning mechanisms, facial mimicry raises complex questions. It is a key example cited in the mirror-neuron debate.¹⁴ How do we learn to map other people’s actions to our own? The

mirror system is fundamental to learning, but how could the mapping between an infant’s expressions and those of a caregiver occur?

One possible mechanism is through “associative sequencing”¹⁴ in which spontaneous motor activity causes a facial action that becomes associated with sensory input caused by the caregiver’s response. Spontaneous “motor babbling” activity is thought to be a fundamental way to bootstrap exploration of motor space and considered fundamental to the development of autonomous agency.¹⁸ In our model, babbling is generated by spontaneous neural network activity, mod-

Action discovery. An example showing learning through interaction with the environment in order to demonstrate autonomous action discovery²⁶ was to aim to have BabyX learn to play the classic video game “Pong” (see Figure 7). We thus connected motor neurons in BabyX to the bat controls and overlaid the visual output of the game on the camera’s input. Motor babbling causes the virtual infant to inadvertently move the bat, much like a baby might flail its arms about. Trajectories of the ball are learned as spatiotemporal patterns on neural network maps. If the bat hits the ball, a rewarding reaction results, reinforcing the association of the current motor state with the trajectory. This association further results in the bat being moved in anticipation of where the ball is going. Without further modification to the model, it is possible for the user to actively encourage BabyX’s choices (releasing virtual dopamine), providing a nice example of “naturally supervised” reinforcement learning.

These basic examples show BabyX learning through interaction with a human user and the shared environment. While basic, these examples of intrinsic action discovery, association, and reinforcement learning (unsupervised and “naturally” supervised) are fundamental to developing generalized autonomous learning systems.

Observations. As interaction is central to the phenomena, we have demonstrated and tested BabyX in several public forums where we have observed an extension of emotion from BabyX to a shared experience with a “passive” audience reacting as vicarious participant. Audience behavior is absorbed into the simulation and is not apart from it, making the experience different from a film, game, or pre-rendered simulation. The demonstrator elicits behavior from BabyX through visual and vocal activity and tries to direct her attention. Affective expressions and voice stimulate reward and affective circuits. If BabyX is abandoned, depending on oxytocin levels, her stress system can activate a cascade of virtual hormones, and she becomes increasingly distressed. BabyX can be trained to recognize certain images that can be associated with vocalizations. When

the demonstrator gains BabyX’s attention and shows her a “First Words Book,” if an image causes a strong enough activation, the image will trigger BabyX to voice an associated word.

Observing in a real, unscripted environment, people anticipate and seek emotional responses from BabyX. As such engagement happens, they are often transformed from observers to engaged participants. An example of this was seen at the 2015 SIGGRAPH conference where Sagar³¹ demonstrated BabyX. While the audience was mainly informed professionals, their reaction was audible and visceral to BabyX. Responses repeatedly observed included a sharp negative reaction when the demonstrator offered to demonstrate

the pain response, as if someone was about to “hurt” the baby. There was no sense that the audience rationally thought the baby was real, though it immediately reacted as if she were. Even within a formal academic presentation, and with the pain response a valid part of any brain model, the audience reacted as if the demonstrator was about to be cruel. Interestingly, this was followed by an emotional display of relief in the form of laughter. As laughter is infectious, the demonstrator laughed, which was registered by BabyX’s sensory inputs, causing her to be “happier.” The audience thus became a part of the feedback loop that changed both parties’ emotional states. The implication is that a witness to a BabyX session is to

Figure 6. BabyX (version 3). Screenshot of sensorimotor online learning session in which multiple inputs and outputs of the model can be viewed simultaneously, including scrolling displays, spike rasters, plasticity, activity of specific neurons, camera input, animated output.

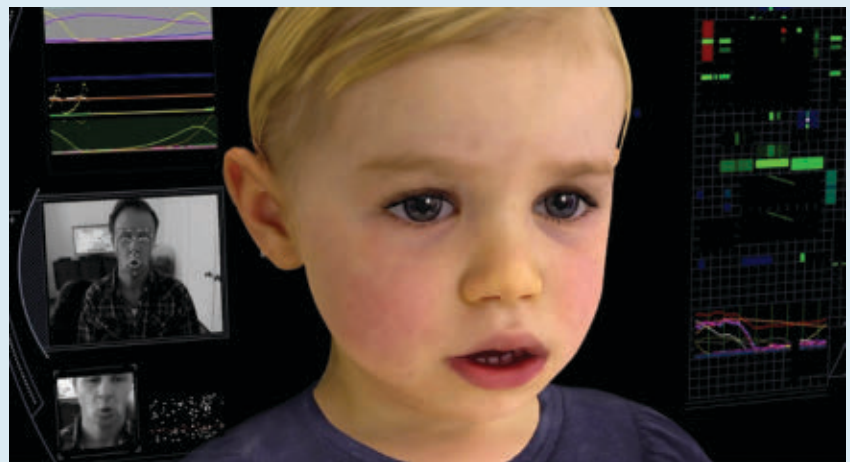


Figure 7. BabyX (version 3). Learning to play the video game “Pong” through action discovery and online reinforcement learning.



become a part of the holistic environment and the interactive experience. There may be ethical implications as well, and further research is needed to investigate the co-defined dynamic interaction that allows such strong “in the moment” emotional responses, as such responses may have long-term interface implications.

Further development and validation. We are currently working on BabyX version 4, as in Figure 4, which has a virtual body and is able to control her

limbs, with initial focus on learning to reach and grasp. BabyX version 4 is intended to interact with the public in exhibitions, performing basic learning tasks (such as label learning). For speech, BabyX babbles with a synthesized voice sampled from phonemes produced by a real child. We are implementing techniques so BabyX can learn an acoustic mapping from any arbitrary voice to construct new words using her own voice. Lip shapes are pre-associated with acoustic elements.

Our aim for BabyX is that she should be capable of learning arbitrary sensorimotor sequences, theorized to map to sentence construction.¹⁷

In an ongoing developmental psychology study, we are conducting a detailed quantitative characterization of the microdynamics of early social learning between parents and their infants. As a first step to validate the effectiveness of BabyX’s behavior at a high level, we will be exploring how well the model elicits naturalistic responses from parents in a social interaction loop, compared to their own or another child. If the model is successful, we will have a new way to study coordinated interaction, and how the way in which we teach infants may play a critical role in learning. Introducing synthetic lesions could be an effective way to explore lower-level validation.

The Auckland Face Simulator

A developing infant is certainly not the easiest approach for creating an embodied conversational agent for HCI tasks. For this purpose, we are building on the same underlying computational platform the “Auckland Face Simulator” (see Figure 8 and Figure 9, as well as Figure 4) also demonstrated at SIGGRAPH³¹ to produce highly realistic avatars capable of real-time interaction; for a related video, see <https://vimeo.com/128835008>. These faces are designed to be used as stimuli for psychological research but also to provide a realistic interface for third-party virtual-agent and AI applications. The avatars can be “told what to say” using text to speech (TTS), and the nonverbal behavior can be specified in a simple API or custom TTS markup language to add further meaning. Wrinkling the nose or raising the upper lip while speaking can dramatically change the perceived meaning. BL allows internal variables of the avatar’s nervous system to be controlled at any level, from muscles to affective circuits.

Conclusion

Engaging face to face with an interactive computer model requires autonomy with contextual responsiveness. If visually consistent, realistic appearance and movement seem to increase the sensory intensity of the experience. Internally consistent generative models enable cognitive, affective, and physiological

Figure 8. The Auckland Face Simulator is being developed to create realistic and precisely controllable real-time models of the human face and its expressive dynamics for psychology research and real-time HCI applications.



Figure 9. The Auckland Face Simulator enables autonomously animated faces to be used for cinematic-like extreme close-up shots.



factors that drive facial behavior to be produced coherently, justifying a lower-level more biologically based modeling approach than has previously been taken with virtual human faces. Exploring these elements together allows new yet familiar phenomena to occur. New, because we do not normally experience this sort of interaction with computers, familiar because we do with people.

Being able to simulate the underlying drivers of behavior, realistic appearance and real-time interaction together deliver three aspects of interaction, but virtually:

Explore. Allows us to explore how the interplay of biologically based systems can give rise to an emotionally affecting experience on a visceral, intuitively relatable human level;

Include movements. Applies an embodied-cognition approach to include the subtle and unconscious movements of the face as a crucial part of mental development and social learning; and

Understand key requirements. Gives a basis for understanding the key requirements for more natural and adaptive HCI in which the interface has a face.

The virtual infant BabyX is not an end unto itself but allows researchers to study and learn about the nature of human response. There is a co-defined dynamic interaction where one can adjust to BabyX no longer as a simulation but as a personal encounter.

In summary, the enormous complexity of modeling human behavior and dyadic interaction cannot be over-estimated, but naturalistic autonomous virtual humans who embody and process theoretical models of our behavior and reflect them back at us may give us new insight into core aspects of our nature and interaction with other people—and future machines.

Acknowledgments

This work was supported in part by the University of Auckland Vice-Chancellor's Strategic Development Fund, Cross Faculty Research Initiative Fund, Strategic Research Investment Fund, and Ministry of Business Innovation and Employment "Smart Ideas" program. We also thank Kieran Brennan, Stephanie Khuu, Kai Riemer, and John Reynolds. 

References

1. Allbeck, J. and Badler, N. Consistent communication with control. In *Proceedings of the Workshop on Multimodal Communication and Context in Embodied Agents at the Autonomous Agents Conference* (2001).
2. Bates, J. The role of emotion in believable agents. *Commun. ACM* 37, 7 (July 1994), 122–125.
3. Blumberg, B.M. *Old Tricks, New Dogs: Ethology and Interactive Creatures*. Ph.D. thesis, MIT, Cambridge, MA, 1996.
4. Breazeal, C. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies* 59, 1 (2003), 119–155.
5. Brinkman, W.P., Broekens, J., and Heylen, D., Eds. *Proceedings of the Intelligent Virtual Agents: 15th International Conference* (Delft, The Netherlands, Aug. 26–28). Springer, 2015.
6. Cangelosi, A., Schlesinger, M., and Smith, L.B. *Developmental Robotics: From Babies to Robots*. MIT Press, Cambridge, MA, 2015.
7. Cassell, J. *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2015.
8. Cattaneo, L. and Pavesi, G. The facial motor system. *Neuroscience & Biobehavioral Reviews* 38 (2014), 135–159.
9. Debevec, P., Hawkins, T., Tchou, C., Duijck, H.P., Sarokin, W., and Sagar, M. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (New Orleans, LA, July 23–28). ACM Press/Addison-Wesley Publishing Co., New York, 2000, 145–156.
10. Ekman, P. and Friesen, W.V. *Facial Action Coding System: Investigator's Guide Part I*. Consulting Psychologist Press, Palo Alto, CA, 1978.
11. Goertzel, B., Lian, R., Arel, I., De Garis, H., and Chen, S. A world survey of artificial brain projects. Part II: Biologically inspired cognitive architectures. *Neurocomputing* 74, 1 (2010), 30–49.
12. Gopnick, A. Why digital-movie effects still can't do a human face. *The Wall Street Journal* (Jan. 8, 2015).
13. Gothard, K. The amygdalo-motor pathways and the control of facial expressions. *Frontiers in Neuroscience* 8 (2014).
14. Heyes, C. Where do mirror neurons come from? *Neuroscience & Biobehavioral Reviews* 34, 4 (2010), 575–583.
15. Jimenez, J., Sunstedt, V., and Gutierrez, D. Screen-space perceptual rendering of human skin. *ACM Transactions on Applied Perception* 6, 4 (2009), 23.
16. Klehm, O. et al. Recent advances in facial appearance capture. *Computer Graphics Forum* 34, 2 (2015), 709–733.
17. Knott, A. *Sensorimotor Cognition and Natural Language Syntax*. MIT Press, Cambridge, MA, 2012.
18. Lee, M.H. Intrinsic activity: From motor babbling to play. In *Proceedings of the IEEE International Conference on Development and Learning* (Frankfurt am Main, Germany, Aug. 24–27). IEEE Press, 2011.
19. MacDorman, K. and Entezari, S. Individual differences predict sensitivity to the uncanny valley. *Interaction Studies* 16, 2 (2015), 141–172.
20. Maes, P. Artificial life meets entertainment: Lifelike autonomous agents. *Commun. ACM* 38, 11 (Nov. 1995), 108–114.
21. Marsella, S. and Gratch, J. Computationally modeling human emotion. *Commun. ACM* 57, 12 (Dec. 2014), 56–67.
22. Mori, M., MacDorman, K.F., and Norri, K. 'The uncanny valley'. *Robotics & Automation Magazine* 19, 2 (2012), 98–100.
23. O'Reilly, R., Hazy, T., and Herd, S. The leabra cognitive architecture: How to play 20 principles with nature and win! In *Oxford Handbook of Cognitive Science*, S. Chipman, Ed., Oxford University Press, Oxford, U.K., 2012.
24. Panksepp, J. *Affective Neuroscience: The Foundations of Human and Animal Emotions*. Oxford University Press, 1998.
25. Parke, F. and Waters, K. *Computer Facial Animation*. CRC Press, 2008.
26. Redgrave, P., Gurney, K., and Reynolds, J. What is reinforced by phasic dopamine signals? *Brain Research Reviews* 58, 2 (2008), 322–339.
27. Rohlfing, K. and Deak, G. Microdynamics of interaction: Capturing and modeling infants' social learning. *IEEE Transactions on Autonomous Mental Development* 5, 3 (Sept. 2013), 189–191.
28. Sagar, M., Robertson, P., Bullivant, D., Efimov, O., Jawed, K., Kalarot, R., and Wu, T. BL: A visual computing framework for interactive neural system models of embodied cognition and face-to-face social learning. In *Proceedings of the 14th International Conference on Unconventional Computation and Natural Computation* (Auckland, New Zealand). Springer, Heidelberg, Germany, 2015, 71–88.
29. Sagar, M., Bullivant, D., Robertson, P., Efimov, O., Jawed, K., Kalarot, R., and Wu, T. A neuro-behavioural framework for autonomous animation of virtual human faces. In *Proceedings of SIGGRAPH Asia Autonomous Virtual Humans and Social Robots for Telepresence* (Shenzhen, China, Dec. 3–6). ACM Press, New York, 2014.
30. Sagar, M. Facial performance capture and expressive translation for King Kong. In *Proceedings of ACM SIGGRAPH 2006 Sketches* (Boston, MA, July 30–Aug. 3). ACM Press, New York, 2006, 26.
31. Sagar, M. BabyX and the Auckland Face Simulator. In *Proceedings of ACM SIGGRAPH Computer Animation Festival* (Los Angeles, CA, Aug. 9–13). ACM Press, New York, 2015, 183–184.
32. Scherer, K., Mortillaro, M., and Mehu, M. Understanding the mechanisms underlying the production of facial expression of emotion: A componential perspective. *Emotion Review* 5, 1 (2013), 47–53.
33. Schröder, M. The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems. *Advances in Human-Computer Interaction* (2010).
34. Sifakis, E., Neverov, I., and Fedkiw, R. Automatic determination of facial muscle activations from sparse motion-capture marker data. *ACM Transactions on Graphics* 24, 3 (July 2005), 417–425.
35. Stone, R. and Papkiewicz, W. The effect of realistic versus imaginary aggressive models on children's interpersonal play. *Child Development* 42, 5 (1971), 1583–1585.
36. Terzopoulos, D. et al. Artificial fishes with autonomous locomotion, perception, behavior, and learning in a physical world. In *Proceedings of the Artificial Life IV Workshop*, P. Maes and R. Brooks, Eds. (Cambridge, MA, July 6–8). MIT Press, Cambridge, MA, 1994.
37. Terzopoulos, D. and Lee, Y. Behavioral animation of faces: Parallel, distributed, and real-time facial modeling and animation. In *Proceedings of ACM SIGGRAPH* (Los Angeles, CA, Aug. 8–12). ACM Press, New York, 2004, 119–128.
38. Trappenberg, T. *Fundamentals of Computational Neuroscience*. Oxford University Press, New York, 2010.
39. Vinciarelli, A. et al. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing* 3, 1 (2012), 69–87.
40. Wu, T. *A Computational Framework for Modeling the Biomechanics of Human Facial Expressions*. Ph.D. thesis, The University of Auckland, Auckland, New Zealand, 2014.

Mark Sagar (m.sagar@auckland.ac.nz) is an associate professor in the Auckland Bioengineering Institute and director of the Laboratory for Animate Technologies at the University of Auckland, Auckland, New Zealand, and CEO/founder of Soul Machines Ltd., Auckland, New Zealand.

Mike Seymour (mike.seymour@sydney.edu.au) is a lecturer in information systems at the University of Sydney, Sydney, Australia.

Annette Henderson (a.henderson@auckland.ac.nz) is a developmental psychologist and senior lecturer in the School of Psychology at the University of Auckland, Auckland, New Zealand.

BabyX and Auckland Face Simulator research and development contributors:

David Bullivant (d.bullivant@auckland.ac.nz), **Paul Corballis** (p.corballis@auckland.ac.nz), **Oleg Efimov** (oefi712@auckland.ac.nz), **Khurram Jawed** (mjaw002@auckland.ac.nz), **Ratheesh Kalarot** (rk018@auckland.ac.nz), **Paul Robertson** (prob014@auckland.ac.nz), **Werner Ollewagen** (wo0627@auckland.ac.nz), and **Tim Wu** (twu051@auckland.ac.nz), all at the University of Auckland, Auckland, New Zealand.

@ 2016 ACM 0001-0782/16/12 \$15.00



Watch the authors discuss their work in this exclusive Communications video. <http://cacm.acm.org/videos/creating-connection-with-autonomous-facial-animation>

Considering the social impact of a proposed future Internet architecture.

BY KATIE SHILTON, JEFFREY A. BURKE,
KC CLAFFY, AND LIXIA ZHANG

Anticipating Policy and Social Implications of Named Data Networking

THE INTERNET HAS become a critical platform for economic, political, cultural, and social activity. The technology behind the Internet continues to evolve, with ramifications for not only the technologies that govern network and application functions, but also for social, economic, and legal concerns. Internet protocols impact not only the basic performance and reliability of Internet services, but also impact debates about fairness issues in content delivery, free speech, trust and cybersecurity, privacy and intellectual property, and control over content.

This article discusses a proposed future Internet architecture that changes how data is delivered over the Internet. Named Data Networking (NDN) is

a prominent example within the broader research field of information-centric networking (ICN). We cannot fully predict how changing protocols will change policy outcomes: social impacts of technology are caused by an interdependent mix of technological decisions, user decisions, and social and policy contexts.^{4,24} But if we take seriously the notion that running code shapes rights, behavior, and governance,^{16,22} then analyzing how NDN would alter that code—the technical infrastructure we rely on every day—is an important challenge.

This article addresses this challenge by beginning a conversation about the social impacts of NDN, with a particular focus on content producers and consumers. We describe the building blocks of NDN; its request-response data exchange is inspired by the Web, but functions at a more fundamental level in the protocol stack. NDN uses data names for routing and forwarding, provides per-packet data signatures, and leverages in-network storage.^a We provide a scenario to illustrate the interactions of these building blocks and describe how the proposed changes could expand options for free speech,

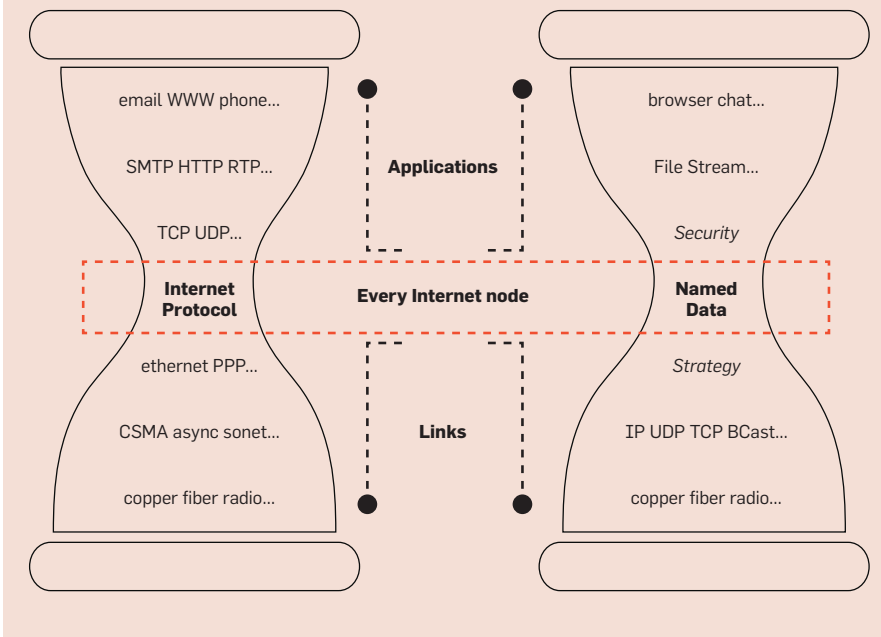
a Many of these techniques are implemented in the application layer of today's Internet. NDN enables them at the network layer, which encourages applications to comport with them.

» key insights

- **NDN is a proposed future Internet architecture that changes the technical protocols that support applications, with implications for social, economic, and policy dimensions of today's Internet ecosystem.**
- **These implications affect a range of stakeholders, including content producers, consumers, regulators, and network operators.**
- **For consumers, NDN can expand options for free speech, security, privacy, and anonymity, while raising new challenges for data retention and forgetting. For governments and content industries, NDN raises new challenges and possibilities for control of content, and for ensuring neutrality across public networks.**



Figure 1. NDN (right) replaces the “thin waist” of the Internet; in its design, the common protocol is the exchange of named, signed data packets instead of IP packets (left).



security, privacy, and anonymity, while raising new challenges regarding data retention and forgetting. We will address impacts for governments and content industries caused by changing the way networked data is identified, handled, and routed as well as examine how these changes raise new challenges and possibilities for ensuring neutrality across public networks. Taken together, this anticipatory analysis suggests research questions and areas of technical focus for ongoing NDN research, and helps us better understand the potential consequences of information-centric networking.

Fundamental Architectural Components of NDN

A team led by Principal Investigators (PIs) from UCLA, and involving Co-PIs, staff, and students from U.S. institutions and international collaborators,^b is designing and evaluating the NDN architecture, which could serve as a new foundational layer of the Internet (see Figure 1). Today, the Internet Protocol (IP) relies on host addresses to route packets across the network. In contrast, NDN delivers based on data names directly, without using host ad-

resses of either source or destination. Rather than forwarding packets based on the *where* of IP, NDN focuses on the *what*: the named data itself. NDN relies on four key architectural components to achieve secure, efficient data delivery: names, request/response data exchange, data signatures, and in-network storage, described in detail in Zhang.³¹

Names: The crux of NDN. In NDN, applications name data at packet granularity. For example, `/edu/ucla/cs/CS217/video1/v2/s3` could refer to segment 3 of version 2 of “video1” published by the teacher of course CS217 in the UCLA-provided namespace.^c The NDN design assumes that application developers will develop standard naming conventions, such as content versioning and segmenting, to aid interoperability and code reuse. NDN also supports hierarchical name structures to facilitate trust management and scalable routing, similar to how hierarchical IP address allocation has enabled global scaling of Internet routing. Globally unique names will require coordinated management and governance,^d but the architecture also

^c Our examples show hierarchical, human-readable NDN names, though the architecture supports arbitrary byte sequences.

^d Just as IP address governance is not a part of the IP architecture, global namespace governance is not an explicit part of the NDN architecture.

supports local names intended for local use (for example, to refer to “the light switch in this room.”) So while all communication in NDN relies on data names, name mechanics will vary based on application context.

Request/response data exchange for multicast delivery. NDN dictates a closed-loop communication model based on packet-by-packet request and response (Figure 2). The model resembles Web semantics but at a per-packet granularity. A consumer sends an Interest packet specifying the name of data she wishes to receive. NDN routers may be able to use cached data to answer that Interest. All data that has previously passed through an NDN router can be cached in its Content Store. (IP routers also have packet buffers due to statistical multiplexing, however a buffered packet is removed from the buffer once it is forwarded to the intended destination). If an Interest cannot be answered with data from an NDN node’s Content Store, the node’s Forwarding Interest Base (FIB) defines where to send the Interest. Nodes use *longest prefix matching* to match data names requested in Interests to data names in the Content Store, and then forward Interests toward nodes that have registered data name prefixes, analogous to IP forwarding.

Each node also uses a Pending Interest Table (PIT) to record the interface, or *face* in NDN parlance, from which it received the Interest. Unlike the FIB and Content Store, the PIT is a fundamentally new entity without an analogy in IP. PIT entries track Interest packets that have been forwarded, to enable data to be returned along the path taken by the Interests. Each PIT entry records the requested data name, the incoming face(s) of the Interest(s), and the outgoing face(s) to which the Interest has been forwarded. Interest propagation creates a hop-by-hop trail of “breadcrumbs” back to the consumer for each path the Interest takes. When the Interest packet reaches a node with matching data, the node responds with a data packet, which is forwarded back along the trail, consuming (such as, deleting) the PIT breadcrumbs along the way.

The request/response model of NDN enables inherent multicast data delivery, as requests for the same data packet from multiple consumers are

^b For a full list of participants and collaborators, see the Named Data Networking website: <http://named-data.net/>.

collapsed into a single PIT entry when they flow through the same router. For example, if a router receives Interests with the same name from five of its faces, the router only forwards the first Interest for that name while recording the incoming faces for the other four Interests in its PIT. When the corresponding data packet comes back, the router forwards that matching data back out to all five faces.

PIT state enables control of traffic load by limiting the number of pending Interests to achieve flow balance. (Only one Interest and one data need to traverse any link for all requestors to be satisfied.) The PIT state can also be used to mitigate DDoS attacks by setting an upper bound on the number of PIT entries allowed.

An NDN network is *loop-free* because each node keeps an entry for each outstanding Interest in its PIT, detecting and discarding duplicates. Each node forwards an Interest to multiple upstream nodes simultaneously and uses the feedback loop created by the request/response structure to evaluate packet delivery performance across its faces—for example, different networks peering with a router or different wireless links on a mobile handset.

Data signatures for provenance and security. Another fundamental aspect of NDN is its use of cryptographic signatures within data packets. NDN requires each data packet to be signed by a key that binds the content to its name. A key locator field encodes the name of the packet's signing key. NDN does not dictate how the consuming application evaluates whether to trust the key. This *data-centric* approach secures the data packet independently of how it is communicated, in contrast with channel-based models such as TLS/SSL on the current Internet.

An active area of research focuses on defining a set of well-understood trust models from which application developers can choose. Within a given trust model, signatures enable determination of data packet provenance, and serve as the basic building block of security in NDN,²¹ including encryption-based access control.³⁰ A valid signature by a trusted key is a strong indication that the data is what it purports to be, regardless of from

where the data was retrieved. The NDN research team is experimenting with a variety of hierarchical, web-of-trust, and evidentiary trust models that use features of NDN for efficient key dissemination and evaluation of trust relationships.²⁹

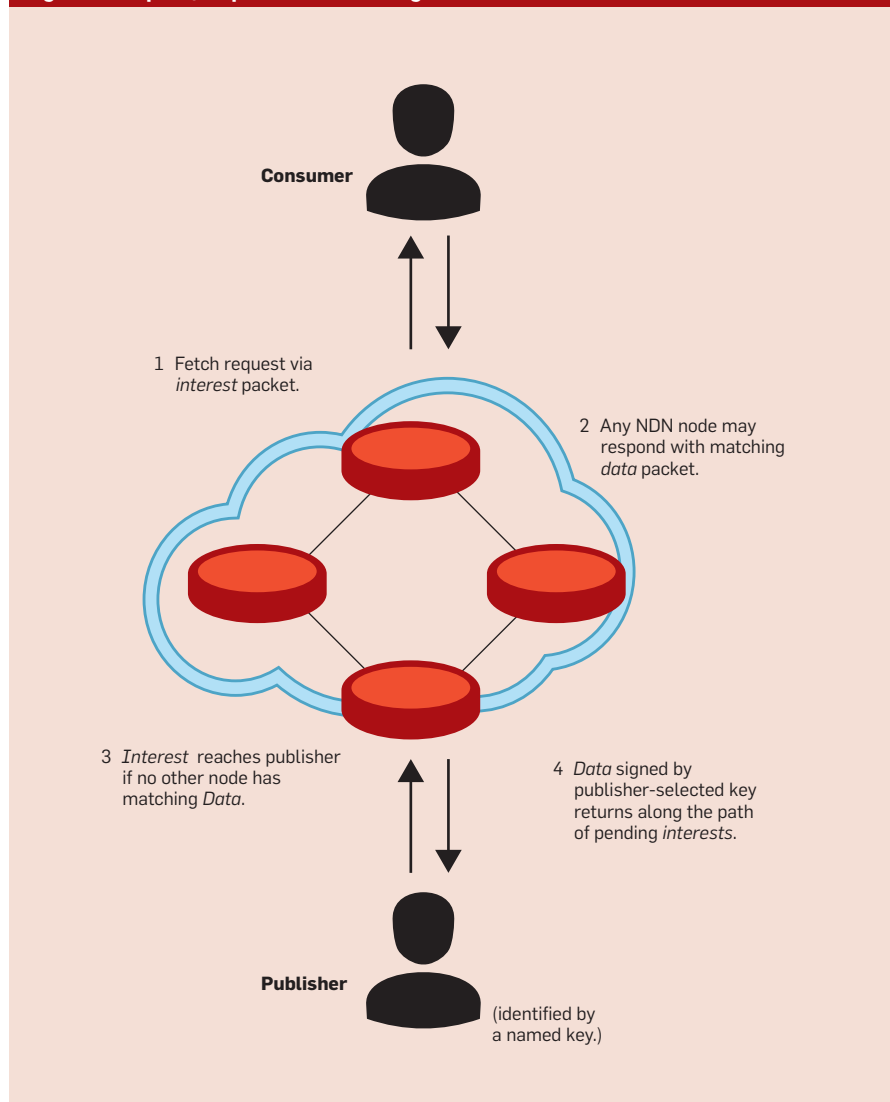
Diverse and pervasive storage at the network layer. Because NDN applications do not care from where requested data is retrieved, any NDN node can answer an Interest if it has corresponding data. This feature enables an NDN network to take advantage of diverse and pervasive forms of storage to yield performance and scalability enhancements, and also provides support for disruption-tolerant networking (DTN). NDN networks can republish data from the local storage of any nearby device, use router memory as data caches, and deploy persistent repositories that

work with any NDN content. Through these means, NDN provides features similar to today's content distribution networks^e (CDNs), but at the network layer, and thus available consistently for *all* data, without contractual agreements between content producers and CDN providers. This is an active area of exploration; for example, NDN researchers are developing new primitives to interact with repos and support efficient synchronization among named data collections.¹⁸

These four abstractions combine and interact to form an NDN network. Naming data necessitates the request/

e CDN services replicate data across a geographically distributed network connected to the IP Internet, moving content close to high concentrations of users to provide faster data access over a broader area (often globally) than a traditional Web hosting model.

Figure 2. Request/response data exchange.



response data exchange. Stored named data can serve future requests, unlike destination-specific IP packets. And because data can be served from anywhere, it must be signed to protect its provenance and integrity.

An NDN Scenario: The Internet of Things

A use case that illustrates the possibilities of NDN is the Internet of Things (IoT). The IoT concept envisions every device, and many objects, as network-enabled, context-aware (to varying extents), and often integrated with Web and mobile applications. We introduce this case, which we will draw on throughout the article, to orient readers to the ways in which NDN's technical changes shape a wide variety of social issues in a realistic application environment.

In an NDN IoT, names provide a richer and more versatile approach to addressing potentially billions of devices across the world, and the architecture's use of cryptographic signatures for each packet provide a valuable security building block not present in IP. NDN enables the Internet-connected "things," and the data they create and consume, to be addressed by one or more application-specific names at the network layer, often without requiring further middleware or gateways.^{3,8,28} For example, a manufacturer-assigned name, such as `/local/appliance/kitchen/toaster/Black&Decker/<serial_number>`, might be used to address a kitchen appliance from another device in the same smart home. That appliance would be configured in this namespace at the factory and respond to Interests in its prefix `/local/appliance` using a power line or wireless interface. In a simple scenario, other devices in a home (for example, a user's phone) could issue Interests on a regular basis. Interests for `/local/appliance` would be used to discover the device when first plugged in; then, its more specific name could be used for direct communication. In this case, NDN enables applications to use the network layer directly to discover nearby devices in these well-known namespaces (for example, `/local/appliance`), without needing the devices to be connected to the global Internet. At the same time, they share the same

network layer protocol as all other NDN Internet applications, providing opportunities for straightforward integration with local or global Web applications, using data signatures and encryption-based access control for security. This example in the IoT domain illustrates that semantic classification can facilitate discovery of new devices on a network—from a new lightbulb to a digital television—using names.

Policy and Social Implications Of NDN's Components

By fundamentally altering the concepts used to design networked applications and the components available to build them, a transition from IP to NDN could impact policy issues including free speech, security, privacy, content regulation, and network neutrality. Some changes are difficult to predict because Internet infrastructure purposefully provides adaptable mechanisms and interpretive flexibility.¹² But even during the design stage, we can articulate a few important ways NDN would likely change the nature of Internet interactions. Here, we explore how the NDN architecture could improve free speech; improve trust and security; both improve and challenge privacy; complicate content regulation by governments and industry; and introduce open questions for network neutrality.

Improvements to free speech. As the IoT example illustrates, NDN facilitates the development of environments where local devices can transmit content without reliance upon global infrastructure providers. Data packets can be stored and republished by anyone using any device, expanding the options for data dissemination and enhancing and expanding opportunities for communication and free speech.

Consider a regime with authoritarian tendencies that allows Internet access but constrains what is published. NDN makes it easier than IP to share data via alternative communications paths and opportunistic connectivity (toasters and phones as well as laptops and routers), without global infrastructure or complex intermediate services providing indirection or anonymization. Users moving in cars or planes or people with ad hoc wireless on their mobile devices can exchange

data via NDN by leveraging storage on their devices and intermittent connectivity to pass content around, without leaving traces of where the data originated. Any NDN node with access to multiple networks, for example, wireless and wired connections, can bridge those networks by forwarding and/or satisfying Interests, increasing the number of paths data can take to a consumer.³¹ Moreover, namespaces can be locally scoped or encrypted, which can render NDN's data exchange mechanisms and decentralized communication capabilities even more tolerant of disrupted connectivity than IP.

Today, blocking a small number of well-known websites is an effective censorship scheme.⁶ Enabling decentralized communication at the lowest layers of the network can allow users to route around censorship, creating positive impacts for free speech. For example, NDN would enable a group of phones at a protest to use data *muling*—a combination of data storage and direct device-to-device communication in which the phones carry data (and keys) from place to place rather than relying on infrastructure that might be subject to global surveillance. Individually signed packets of a sensitive video, or the keys to verify that video, can be reassembled by any device based on common naming conventions, and verified as being from the same publisher using data signatures. Such peer-to-peer muling can occur in IP networks, but is more complicated at the network and application layers. In addition, NDN content producers could encapsulate or encrypt data names to hide traffic and thwart attempts to block content based on its name.

NDN's emphasis on data signatures could complicate a social mechanism often relied upon to protect free speech: anonymous content production. Therefore, NDN's improvements for free speech must be weighed against its challenges to anonymous speech.

Improvements for trust and security. NDN requires all data be signed so that applications can verify the publisher of received content. In the IoT scenario, each networked device in a home would sign content, enabling applications such as lighting control or energy monitoring services to ver-

ify that data they receive, including commands, originates from a trusted source. Because per-packet signatures are part of the architecture and therefore not dependent on an application or domain, NDN will increase recognition of, and reliance on, data provenance to improve data security and thus consumer trust in content. In the IP Internet, provenance must be established on a per-application basis, and is currently established intermittently and inconsistently. NDN's signature mechanisms can help verify provenance even for orphaned data (data with no online application). Content signatures can also reduce risks such as spoofed data and phishing. Including such provenance explicitly in packets mitigates concerns about data tampering en route.

To take advantage of NDN's security features (in particular, the per-packet cryptographic signatures), application developers will require new trust models that can be used by classes of applications, as well as frameworks for establishing, exchanging, and revoking keys within data-centric networks. These challenges, as discussed earlier, are the most significant for NDN architecture development. Fortunately, we believe there will be increasing incentives to develop such trust models and key distribution mechanisms over time, as they are necessary not only for NDN, but for better security in all networked communications.

Finally, NDN's request/response data exchange provides benefits for network security by mitigating common problems in today's IP Internet, such as distributed denial of service (DDoS) attacks. Since each Interest retrieves at most one data packet, a router can use the PIT (as described previously) to control the number of pending Interests to achieve flow balance, mitigating volumetric DDoS attacks. Techniques for NDN DDoS mitigation have been explored extensively in other work.^{1,14}

Improvements and challenges for privacy. NDN's four fundamental architectural departures from IP have implications that can both challenge and benefit user privacy. The request/response data exchange increases *anonymity of information seekers*, while content signatures and names compli-



Because NDN applications do not care from where requested data is retrieved, any NDN node can answer an Interest if it has corresponding data.



cate *anonymity for content producers*. The architectural emphasis on in-network storage presents new challenges for *limiting data retention*.

Support for anonymous information-seeking. NDN's request/response data exchange improves support for anonymous information-seeking: there is no source address in an Interest. Though Interest packets create a trail in the PIT as they travel toward a data packet, each router's table indicates only the next hop and these PIT entries are erased as soon as a data packet satisfies the outstanding Interest(s). Although routers could log such trails of breadcrumbs, users are not likely to have their Interests traced back to them unless an actor (an authoritarian regime, for example) can access and correlate state across all routers in the (possibly many) paths that data packets have taken. The IoT scenario illustrates how difficult enacting this level of control would be: those paths would likely include privately-owned devices in homes and buildings, in addition to routers owned by Internet Service Providers (ISP)s. So while ISPs might log Interests and forward them to governments, decreased reliance on ISPs as the sole source of connectivity would circumvent such logging. Providing anonymous data retrieval could substantially benefit privacy, allowing individuals to consume controversial content without fear of embarrassment or harm.^{13,26}

Challenges for anonymous content production. Compared to *consumer anonymity*, *content producer anonymity* in an NDN network is difficult to achieve. Data producers can be identified in more than one way—for example, by the key used to sign the data, the namespace in which the data or key are published, or by the content itself.

While NDN data must be signed, it may be signed with ephemeral keys or keys unlinked to real-world identities. Encryption of both names and data can be used to provide confidentiality. But NDN's pervasive use of signatures may make it easier for infrastructure providers and content consumers to *require* signatures that use verified, real-world identities. For example, online forum moderators struggling with trolls and sock puppet accounts—or trying to discriminate against certain

users—might not accept comments sent in packets without verified, real-world signatures. Namespace ownership records may also reveal publisher identities, similar to today's WHOIS database. Thus, another important area of NDN research is trust schemes that provide alternatives to real-world identity for content authentication.

NDN researchers have explored special routing approaches to preserve content source anonymity.¹⁷ Content producers might desire anonymity to participate in free speech, evade censorship, and experiment with multiple online identities.²⁵ Unfortunately, anonymity is also used to evade prosecution for criminal behavior or support mob behavior and hate crimes.¹⁰ Though designing a network architecture to prevent all criminal behavior is an impossible (and, we believe, undesirable) goal, it is worthwhile to consider the benefits and costs of measures to increase content producer anonymity as the project goes forward.

Improvements for content access control. As mentioned earlier, the NDN architecture encourages applications to secure data by encrypting it rather than relying on channel-based security over which the data flows, as is currently done through secure sockets layer/transport layer security, (SSL/TLS), virtual private networks (VPN)s, and similar schemes on IP networks. In the IoT example, there is no need to set up secure connection between two communicating devices, because any potentially sensitive data is encrypted by the application. Securing the data directly should reduce the impact of now-common perimeter and channel security compromises, while still leveraging NDN caching for group communication.

Once published, encrypted data can be replicated and hosted in many (potentially hostile) locations, although only those with access to the right keys can decrypt the information. In this way, NDN makes explicit what is already implicit in schemes like SSL/TLS: encrypted data in transit can be sniffed and stored by others. NDN makes it easier to request a chunk of someone's encrypted data (for example, by sending Interests for common namespaces like /local/appliance), and that encrypted data might be

cached anywhere. Encrypted data may be widely available for extended periods of time, increasing the long-term potential for unauthorized decryption. Content access control will thus require careful design and integration of modern encryption mechanisms and techniques, such as forwarding secrecy and long-term encrypted storage. Further, NDN's integrated use of cryptography also will require navigating open challenges such as the computational burden of encryption in resource-constrained environments (like the IoT) and the challenges of key distribution and revocation.⁹

Challenges for the right to be forgotten. As personal data proliferates on the Internet, there is increasing concern that such data cannot be erased or forgotten. The specter of total accountability for our past actions is considered unpleasant at best and potentially limiting to social interaction and democracy at worst.^{7,23} International privacy scholars as well as policymakers in Europe have been paying increased attention to data retention and disposal, or the "right to be forgotten."^{7,23,27} More recently, California adopted Senate Bill 568, which requires websites to enable minors to easily remove their own posts from websites.

IP routers purge data from buffers as soon as it leaves the routers. That is, they default toward "forgetting" at the infrastructure level, with substantial data retention occurring at the application layer, to support targeted advertising and other purposes. In contrast, NDN routers default toward remembering at the infrastructure level, via content stores and repos. In IP, parties can request that publishers remove data from hosting sites at the edges of the network. Although copies may proliferate elsewhere on user machines, any new request to the hosting site will go unsatisfied. Returning to the IoT example, in NDN, cached copies of data from baby monitors or mobile devices may proliferate on routers, repositories, as well as application-specific stores, and thus remain accessible to Interests. Architectural support for "forgetting" in an NDN world will require mitigation measures, such as time-to-live information in packets, protocols that respect those limits, and further research into self-destructing data.

Challenges for law enforcement and content regulation. The Internet's vital role in cross-border commerce means it contends with diverse national and international policies regulating publication and use of content. Content produced by illegal activities may be restricted (for example, bans on the sale of Nazi memorabilia in France have led to restrictions on content listed in online marketplaces); other forms of content may have use restrictions designed to guarantee a profit to content creators. Enforcing publication and use regulations on content across the global Internet is a challenging task in today's IP Internet. Corporate interests often use the *where* of IP source addresses to enforce market-based restrictions on content access via IP geolocation heuristics. Law enforcement uses a range of tactics—ranging from IP address tracing to deep packet inspection—to track and prosecute both producers and consumers of illegal or pirated content. A transition to NDN will change the tools needed for tracing individuals and monitoring and restricting communications, making current forms of content regulation more challenging, but also potentially more equitable.

Complications for law enforcement. NDN's emphasis on semantic names and data signatures may make certain types of law enforcement easier. For example, keys used to sign data provide strong provenance. In the IoT scenario, the publisher of critical content might be traced by matching the key to identifiable (perhaps registered) devices. And if clear-text data names reflect actual content (for example, data prefixed with /local/PIR was known to be generated by passive infrared security sensors), network-level packet-sniffing and therefore, network regulation could become less computationally intensive. On the other hand, encryption of both NDN names and packet content could mitigate the risk of packet-sniffing. A social shift toward widespread data encryption would raise new challenges for law enforcement. Police and regulatory regimes have long been wary of widespread use of encryption, while developers have resisted providing back doors for law enforcement to inspect or wiretap communications. Encryption would limit the capabili-

Social and policy impacts of NDN for content producers, consumers, regulators, and network operators.

	Free speech	Privacy	Control of content	Network neutrality
Named, signed data available from any node willing to provide it	<i>Improvement for content producers and consumers: can route data around censorship attempts</i>	<i>Improvement for content consumers: surveillance of content more difficult to achieve</i>	<i>Challenge for content producers and regulators: complicates geographic content restrictions</i>	<i>Improvement for content producers & consumers: diversifies interests in tussle over Internet resources</i> <i>Challenge for network operators: diversifies competition</i>
Strong provenance built on data signatures and straightforward key distribution	<i>Improvement for content producers and consumers: increases trust in provenance of speech</i> <i>Improvement for network operators: Increases information available for network strategies</i>	<i>Challenge for content producers: may identify content producers</i>	<i>Improvement for content producers and regulators: may help identify infringing content</i>	<i>Challenge for content producers and consumers: may enable discrimination based on data type or origin.</i> <i>Improvement for network operators: Increases information available for network strategies</i>
Data persistence via uniformly accessed, pervasive storage	<i>Improvement for content producers and consumers: data persists even when subject to takedowns</i>	<i>Challenge for content producers and consumers: may increase likelihood of decryption by unauthorized parties</i>	<i>Challenge for content producers and regulators: complicates content control</i>	<i>Improvement for content producers and consumers: diversifies interests in tussle over Internet resources</i> <i>Challenge for network operators: Incentives for hosting caching unclear¹</i>
Request/response model of data exchange	<i>Improvement for content consumers: ensures anonymity for content seekers; can route requests around censorship attempts</i>	<i>Improvement for content consumers: ensures anonymity for content seekers</i>	<i>Improvement for content regulators: may suppress requests by name</i>	<i>Improvement for network operators: can control traffic load by controlling the number of pending interests</i>

ties of deep packet inspection, used for everything from security concerns to managing traffic flow.⁵

NDN will also change how governments assert regional jurisdiction on the Internet. Today, IP addresses are often used to target law enforcement action.¹⁵ Countermeasures to such targeting in IP include content encryption, encapsulation, and use of third-party resources such as botnets. NDN further disassociates communication from location, as demonstrated by the IoT scenario, which allows communication between devices without any reference to physical geography. This disassociation complicates the identification and geolocation of suspicious activity based on network data. By complicating the use of network data for identification and geolocation, NDN may encourage law enforcement methods that are more effective, such as following financial trails rather than Internet traffic.

Digital rights management. Law enforcement personnel are not the only stakeholders that rely on IP address geolocation capabilities. Sports franchises use them to restrict subscribers

in local markets from watching games online. Gambling operations restrict participation from countries in which such operations are illegal. Search results are tailored to locations. However, one level of indirection, such as Virtual Private Networks (VPNs), can often circumvent IP-address-based control. In NDN, stakeholders might need to rely on application-layer identity and location information to enforce such content restrictions. Although Interests can come from anywhere, stakeholders could build systems of encryption and key distribution based on location-verified subscribers.

Digital Rights Management (DRM) typically involves controlling distribution of content, and controlling whether consumers can redistribute that content. NDN supports the first kind of DRM well, but makes republishing easier than it is with IP. As in the IP Internet, copyright holders can distribute verified, encrypted media, and consumers can access the content with the proper key. However, widespread encryption challenges the benefits of in-network caching, reducing economic

incentives to provide such caching.² Reliance on encryption for copyright enforcement also hinders legitimate reuses of content, such as fair use in educational contexts, critique, and parody. Content producers might enable fair use by giving copies of keys to libraries, or providing portions of the content in the clear for scholarship, critique, parody, or other protected fair uses. But once consumers have received and decrypted verified content, they may distribute unauthorized versions in the clear, a task made easier by NDN.

NDN's in-network storage and caching means that many segments of both licensed (presumably encrypted) and unlicensed (presumably decrypted) media could reside on routers and repos. A world where countless copies proliferate across the Internet challenges assumptions embedded in copyright law, as well as the current mechanisms of copyright enforcement, such as the Digital Millennium Copyright Act (DMCA) takedown notice.¹⁹ On the IP Internet, videos are commonly hosted by major providers such as YouTube or Hulu, which

respond to DMCA takedown notices. However, even these major providers struggle with the scale: as of August 2015, Google was receiving over 12 million URLs requested to be removed from search per week.²⁰ Who would be responsible for taking down an infringing video distributed on thousands of routers by thousands of different organizations across the world? The political economy of repos—who owns them, and in what jurisdictions—will impact the future efficacy of such takedown notices. NDN's in-network storage may increase pressure on lawmakers to redefine intellectual and political understandings of copyright already challenged by pervasive digital duplication.

Network neutrality: An uncertain outcome. The network neutrality debate focuses on what actors pay for Internet resources such as bandwidth and storage, and whether those actors providing resources (for example, ISPs) may throttle or privilege traffic to increase revenue. Consideration of NDN's impact on network neutrality motivates a deeper discussion of NDN node operation. The algorithms and parameters for configuring network-forwarding policy for given data prefixes and links in NDN are typically referred to as *strategies*.^f Strategies are an evolving part of NDN research. In every NDN device, strategies control the operation of three tables—the FIB, the Content Store, and the PIT. The strategies for these tables affect performance by enabling node owners to express traffic shaping policies in terms of namespaces and faces to other nodes.

Interest forwarding strategies. Routing protocols and/or manual setup of static routes are used to configure forwarding strategies in the FIB. The resulting configuration expresses the policies of router administrators, who may choose to discriminate based on data types (indicated within data names, for example, /local/toaster) or namespace of publication (ucla/cs/local/toaster). Such traffic discrimination may occur in IP

but at higher layers, for example, HTTP or via the Domain Name System (DNS) names. NDN routers will be capable of such choices at the network layer.

Content store strategies. All data that passes through an NDN router can be cached in the content store, and persists according to a router's configured caching policy. NDN spreads caching and its costs across the Internet infrastructure, which democratizes content storage functions and introduces new stakeholders into the tussle over Internet resources. Researchers are considering economic incentives for deploying caches and markets for cache participation.² In-network storage will also impact the political economy of content dissemination. Given enough in-network caching, content producers on an NDN Internet can use a cheap server and low-bandwidth connection to make their viral videos reachable by millions of interested viewers, with the network providing scalability to handle content requests. Thus, NDN could reduce dependence on third-party services to scale content distribution. Users could share content on their own terms, rather than being subject to a third-party provider or hosting service's terms.

Pending interest table (PIT) strategies. Because the PIT records which Interest packets have been forwarded, and then waits for data packets to return, policies that modify how long to retain Interests in the PIT could impact data retrieval performance. (While a field in each Interest specifies a lifetime, it is up to each forwarder to obey that field.) Whether consumers or namespace providers are able to influence the quality of service through longer Interest storage in the PIT, or more aggressive re-issuing of Interests across multiple outgoing faces, are strategy configuration questions that could impact a node's neutrality.

Neutrality implications of NDN node strategies. The actors controlling NDN traffic routing decisions are likely to be more diverse than on an IP Internet. Nonetheless, an NDN-based Internet's ISPs will continue to have incentives (if not obligations) to author strategy modules to manage the tables in their routers and prioritize data with certain types or names. Data names may reveal types of content, such as IoT, video, scientific data, or emergency response data. Globally routable name prefixes

expressing data origin, such as /edu/ucla or /com/nytimes, may indicate institutional power or status. Signing keys may also reveal data origin. Future standards might use names or signing keys to prioritize particular interest/data exchanges, such as emergency response traffic. Executing such prioritization across multiple providers would bear the same policy complexity and risk as with attempts to do multi-provider QoS on today's IP networks.¹¹

In an NDN network, routing provides only one of the input factors for forwarding decisions; locally configured forwarding strategies make the final decision on which Interest is forwarded along which path, or forwarded at all. In this way, NDN's inherent support for mobility and disruption-tolerant networking could mitigate the threat of harmful traffic discrimination. Even if prioritized networking evolves using semantically-meaningful names or pay-for-retention policies on routers, NDN's ability to forward requests around providers that do not respond efficiently will give consumers more options for data transmission. Small ISPs can use multipath forwarding to choose forwarding paths based on performance measurements. NDN enables providers to route around ISPs that throttle traffic based on certain names, which will provide a disincentive for such throttling.

Conclusion

NDN brings some of the semantics of the current Internet's application layers to the network layer, providing technical benefits to application developers, network operators, and end users. NDN's architectural design decisions are also likely to have implications for social and policy issues in both layers, including some of today's most pressing challenges: free speech, privacy, control of content, and network neutrality. Whether these impacts are benefits or drawbacks depends upon stakeholder perspective. The accompanying table maps NDN's technical implications to social and policy impacts for a range of stakeholders: content producers, consumers, regulators, and network operators.

By diversifying the nodes that can provide data, NDN will likely improve conditions for free and anonymous speech and information seeking for consumers and producers. NDN's

^f Here, we use strategies more broadly than the NDN architects have so far, using this term to cover any policy choice that can be made in an NDN node that does not violate the "thin waist" of the architecture as currently understood.

strong provenance built on data signatures will identify content producers. Strong provenance will enhance trust in, and security of, content, while simultaneously complicating anonymous information production. Strong provenance may also help content producers identify infringing content, and signatures provide a mechanism to help producers secure content with encryption-based access control. But pervasive storage and request/response data exchange will challenge producers interested in content control and geographic access restrictions. Finally, network neutrality is a complicated outcome to predict. Future decisions in naming and routing may hinder network neutrality, as the use of names for routing could facilitate new forms of traffic discrimination. At the same time, NDN will promote increased competition among network operators by enabling applications to efficiently route around infrastructure that constrains their traffic.

This article has sought to address policy and social implications of the network that are significant departures from today's IP Internet. As such, it has not addressed Internet policy topics that remain closely tied to existing challenges in IP, which are areas for future work. For example, NDN faces challenges in globally routable naming rights management similar to those of IP. We have also not addressed application-level policy issues such as the relationship between advertising data collection and privacy or application-level regulations such as accessibility requirements or required geolocation services such as E911.

However, identifying open questions relevant to the network layers illustrates an advantage of anticipatory policy studies. Analyzing potential social and policy impacts of the NDN architecture can help prioritize research questions within the NDN project and broader content-centric networking initiatives. The practical impact of NDN will depend on future directions in several open research areas: (1) balancing meaningful names to simplify application development with opaque names to protect privacy; (2) standardizing mechanisms for cryptographic key assignment, distribution and revocation; (3) developing usable design patterns for managing trust in a broad range of applications; (4) providing usable, secure imple-

mentations of more complex multi-participant encryption schemes; and (5) creating fair congestion management to enable network neutrality.

Most of NDN's potential policy impacts are speculative, in part because we are exploring them while the architecture design is still evolving. Yet imagining the social changes NDN might encourage during the design process provides opportunities for pro-social computing research. We hope this work will spark continuing discussion of the current and future Internet's impact on society. Thinking creatively about changes can help us better understand the relationship between infrastructure and our world.

Acknowledgments. The authors thank Van Jacobson, David D. Clark, Paul Ohm, Charles Duan, Steven Bellovin, and anonymous reviewers for feedback and ideas that shaped this work. This research was supported by the National Science Foundation under grant numbers CNS-1040868, CNS-1421876, and CNS-1345318. 

References

- Afanasyev, A., Mahadevan, P., Moiseenko, I., Uzun, E. and Zhang, L. Interest flooding attack and countermeasures in Named Data Networking. *IFIP Networking Conference*, 2013, 1–9.
- Agyapong, P.K. and Sirbu, M. Economic incentives in information-centric networking: Implications for protocol design and public policy. *IEEE Commun. Mag.* 50, 12 (2012), 18–26; <http://doi.org/10.1109/MCOM.2012.6384447>
- Bannis, A. and Burke, J.A. *Creating a secure, integrated home network of things with Named Data Networking*. UCLA, 2015, Los Angeles, CA.
- Baxter, G. and Sommerville, I. 2011. Socio-technical systems: From design methods to systems engineering. *Interacting with Computers* 23, 1 (2011), 4–17; <http://doi.org/10.1016/j.intcom.2010.07.003>
- Bendrath, R. and Mueller, M. The end of the net as we know it? Deep packet inspection and Internet governance. *New Media & Society* 13, 7 (2011), 1142–1160.
- Best, M.L. and Wade, K.W. Democratic and anti-democratic regulators of the Internet: A framework. *The Information Society* 23, 5 (2007), 405–411; <http://doi.org/10.1080/01972240701575684>
- Blanchette, J.-F. and Johnson, D.G. Data retention and the panoptic society: the social benefits of forgetfulness. *The Information Society* 18 (2002), 33–45.
- Burke, J.A., Gasti, P., Nathan, N. and Tsudik, G. Securing instrumented environments over content-centric networking: the case of lighting control. In *Proceedings of IEEE INFOCOMM 2013 NOMEN Workshop*. Retrieved Aug. 7, 2015; <http://named-data.net/publications/nomen13/>
- Chaabane, A., De Cristofaro, E., Kaafar, M.A. and Uzun, E. 2013. Privacy in content-oriented networking: Threats and countermeasures. *SIGCOMM Comput. Commun. Rev.* 43, 3 (2013), 25–33; <http://doi.org/10.1145/2500098.2500102>
- Citron, D.K. Civil rights in our information age. In *The offensive internet: privacy, speech, and reputation*. Harvard University Press, Cambridge, MA and London, 2010, 31–49.
- Clark, D.D., Bauer, S., Lehr, W. et al. Measurement and analysis of Internet interconnection and congestion. In *Proceedings of the 42nd Research Conference on Communication, Information, and Internet Policy*. Social Science Research Network, 2014. Retrieved Aug. 17, 2015; <http://papers.ssrn.com/proxy-um.researchport.umd.edu/abstract=2417573>
- Clark, D.D., Wroclawski, J., Sollins, K.R. and Braden, R. Tussle in cyberspace: Defining tomorrow's Internet. In *Proceedings of the 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, ACM, 347–356; <http://doi.org/10.1145/633025.633059>
- Cohen, J.E. A right to read anonymously: A closer look at "copyright management" in cyberspace. *Connecticut Law Review* 28 (1996), 981–1039.
- Compagno, A., Conti, M., Gasti, P. and Tsudik, G. Poseidon: Mitigating interest flooding DDoS attacks in Named Data Networking. In *Proceedings of the 2013 IEEE 38th Conference on Local Computer Networks*, 630–638; <http://doi.org/10.1109/LCN.2013.6761300>
- Cooke, L. 2007. Controlling the net: European approaches to content and access regulation. *Journal of Information Science* 33, 3 (2007), 360–376; <http://doi.org/10.1177/0165551506072163>
- DeNardis, L. Hidden levers of internet control. *Information, Communication & Society* 15, 5 (2012), 720–738; <http://doi.org/10.1080/1369118X.2012.659199>
- DiBenedetto, S., Gasti, P., Tsudik, G. and Uzun, E. ANDaNA: Anonymous Named Data Networking application. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium*, Internet Society, Retrieved June 26, 2012; <http://arxiv.org/abs/1112.2205>
- Fu, W., Abraham, H.B. and Crowley, P. Synchronizing namespaces with invertible bloom filters. In *Proceedings of the 11th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*. IEEE Computer Society, 123–134. Retrieved Aug. 17, 2015; <http://dl.acm.org/citation.cfm?id=2772722.2772740>
- Gillespie, T. *Wired Shut: Copyright and the Shape of Digital Culture*. The MIT Press, 2009.
- Google transparency report. 2005. Retrieved Aug. 17, 2015; <http://www.google.com/transparencyreport/removals/copyright/>
- Jacobson, V., Smetters, D.K., Thornton, J.D., Plass, M., Briggs, N. and Braynard, R. Networking named content. *Commun. ACM* 55, 1 (Jan. 2012), 117–124; <http://doi.org/10.1145/2063176.2063204>
- Lessig, L. *Code: Version 2.0*. Basic Books, New York, NY, 2006.
- Mayer-Schoenberger, V. *Useful Void: The Art of Forgetting in the Age of Ubiquitous Computing*. Harvard University, Cambridge, MA, 2007.
- Pfaffenberger, B. Technological dramas. *Science, Technology & Human Values* 17, 3 (1992), 282–312.
- Phillips, D.J. From privacy to visibility: Context, identity, and power in ubiquitous computing environments. *Social Text* 23, 2 (2005), 95–108.
- Richards, N.M. The perils of social reading. *Georgetown Law Journal* 101, 3 (2013). Retrieved Apr. 15, 2012; http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2031307
- Rosen, J. The right to be forgotten. *Stanford Law Review Online* 64 (2012), 88.
- Shang, W., Ding, Q., Marianantoni, A., Burke, J. and Zhang, L. Securing building management systems using named data networking. *IEEE Network* 28, 3 (2014), 50–56; <http://doi.org/10.1109/MNET.2014.6843232>
- Yu, Y., Afanasyev, A., Clark, D., Claffy, K., Jacobson, V. and Zhang, L. Schematizing and automating trust in Named Data Networking. In *Proceedings of the 2nd ACM Conference on Information-Centric Networking*. ACM, 2015.
- Yu, Y., Afanasyev, A. and Zhang, L. *Name-based access control*. University of California Los Angeles, CA, 2015.
- Zhang, L., Estrin, D., Burke, J.A. et al. *Named Data Networking Tech Report 001*. University of California Los Angeles, CA, 2010. Retrieved Sept. 8, 2014; <http://named-data.net/publications/techreports/tr001ndn-proj/>

Katie Shilton (kshilton@umd.edu) is an assistant professor of information studies at the University of Maryland, College Park.

Jeffrey A. Burke (jburke@remap.ucla.edu) is the assistant dean, technology and innovation, in the School of Theater, Film and Television at the University of California, Los Angeles.

kc claffy (kc@caida.org) is the founder and director of the Center for Applied Internet Data Analysis (CAIDA) at the University of California, San Diego.

Lixia Zhang (lixia@cs.ucla.edu) is the Jonathan B. Poste; Professor of Computer Science at the University of California, Los Angeles.

Copyright held by authors.

Social computing benefits from mathematical foundations, but research has barely scratched the surface.

BY YILING CHEN, ARPITA GHOSH, MICHAEL KEARNS,
TIM ROUGHGARDEN, AND JENNIFER WORTMAN VAUGHAN

Mathematical Foundations for Social Computing

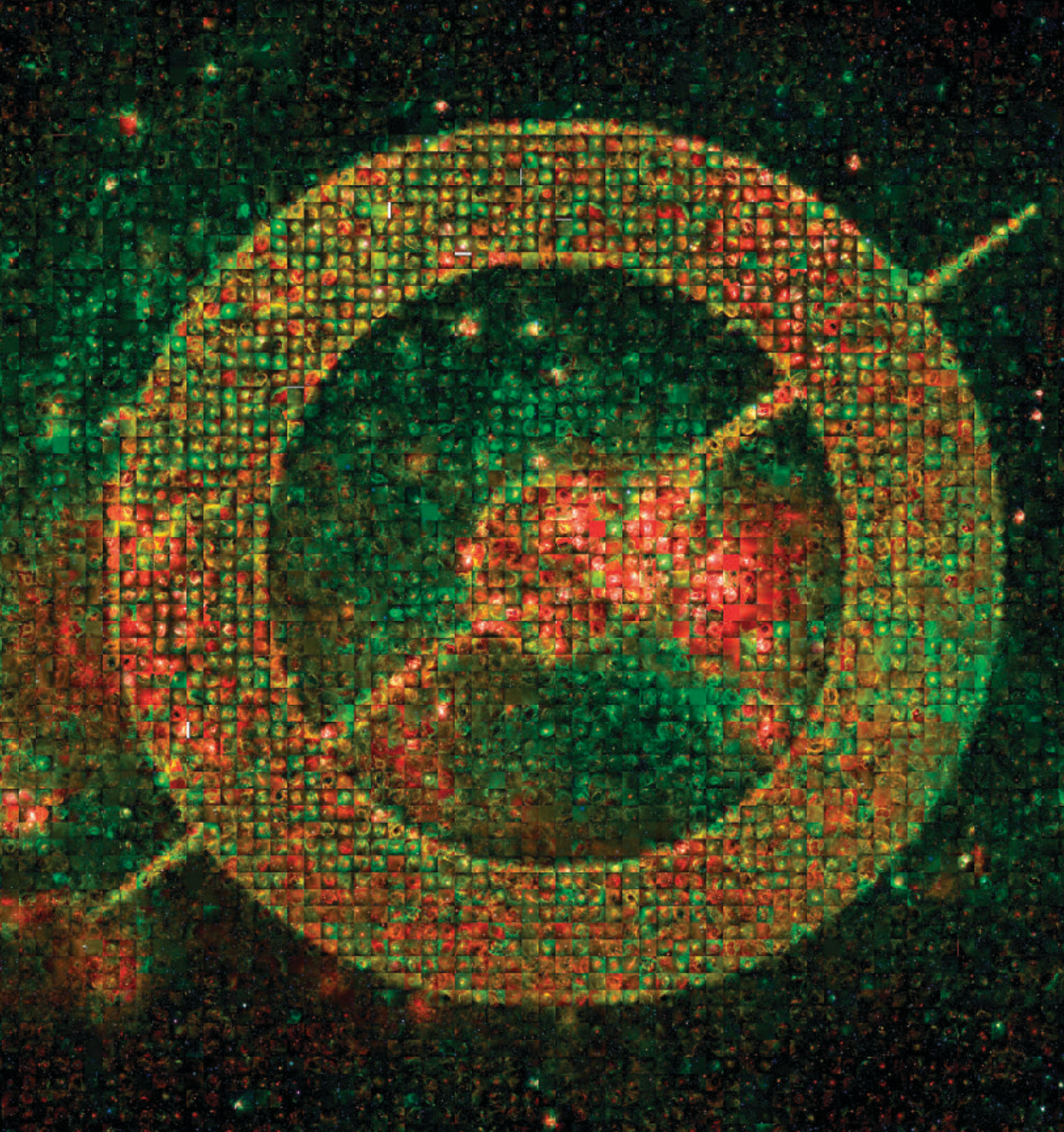
SOCIAL COMPUTING ENCOMPASSES the mechanisms through which people interact with computational systems: crowdsourcing systems, ranking and recommendation systems, online prediction markets, citizen science projects, and collaboratively edited wikis, to name a few. These systems share the common feature that humans are active participants, making choices that determine the input to, and therefore the output of, the system. The output of these systems can be viewed as a joint computation between machine and human, and can be richer than what either could produce alone. The term social computing is often used as a synonym for several related areas, such as “human computation” and subsets of “collective intelligence;” we use it in its broadest sense to encompass all of these things.

Social computing is blossoming into a rich research area of its own, with contributions from diverse disciplines including computer science, economics, and other social sciences. The field spans everything from systems research directed at building scalable platforms for new social computing applications to HCI research directed toward user interface design, from studies of incentive alignment in online applications to behavioral experiments evaluating the performance of specific systems, and from understanding online human social behavior to demonstrating new possibilities of organized social interactions. Yet a broad mathematical foundation for social computing is yet to be established, with a plethora of under-explored opportunities for mathematical research to impact social computing.

In many fields or subfields, mathematical theories have provided major contributions toward real-world applications. These contributions often come in the form of mathematical models to address the closely-related problems of analysis—why do existing systems exhibit the outcomes they do?—and design—how can systems be engineered to produce better outcomes? In computer science, mathematical research led to the development of commonly used practical machine learning methods such as

» key insights

- The output of a social computing system can be viewed as a joint computation between humans and machines, and can be richer than what either can produce alone.
- Mathematical research has led to innovations in social computing such as crowdsourced democracy, prediction markets, and fair division are all examples of social computing applications.
- Social computing systems cannot be properly modeled or analyzed without accounting for the behavior of their human components, which requires a dialog between theoretical, experimental, and empirical research and across disciplinary boundaries.



A mosaic for the Milky Way Project that was created from data analyzed by citizen scientists at Zooniverse.

boosting and support vector machines, public-key cryptography including the RSA protocol, widely used data structures such as splay trees and techniques like locality-sensitive hashing, and more. Well-known examples in economics include the analysis and design of matching markets that have en-

abled kidney exchanges and have led to significant successes in public school admissions and residence matching for doctors and hospitals, the influence of auction theory on the design of the FCC spectrum auctions, and the design and redesign of the auctions used in online advertising markets.

As in other fields, there is great potential for mathematical work to influence and shape the future of social computing. There is a small literature using mathematical models to analyze and propose design recommendations for social computing systems including crowdsourcing markets,^{7,18,24,25,30,37}

prediction markets,^{1,2,33} human computation games,^{28,39} and user-generated content sites;^{12,15,17,29} see, for example, Ghosh¹⁴ for a survey of one facet of this work. However, we are far from having the systematic and principled understanding of the advantages, limitations, and potentials of social computing required to match the impact on applications that has occurred in other fields.

We note that social computing enjoys a close relationship with another emerging discipline, which is computational social science.^{19,34,a} But it is also distinct from that field. While human and social behavior, ability, and performance are central to both, computational social science focuses primarily on the use of modern technology, data, and algorithms to understand and describe social interactions in their “natural habitats.” In contrast, social computing (as the name suggests) has a much more deliberate focus on engineering systems that are hybrids of humans and machines, which may often entail shaping collective behavior in unfamiliar environments. Nevertheless, we anticipate a continued close relationship and even blurring of the two efforts. As an example, one should expect the vast theoretical and experimental literature on the diffusion of influence and behavior in social networks to be relevant to any effort to design a social computing system that relies on such dynamics to recruit and engage workers.

In June 2015, we brought together roughly 25 experts^b in related fields at a CCC-sponsored Visioning Workshop on the Theoretical Foundations of Social Computing^c to discuss the promise and challenges of establishing mathematical foundations for social computing. This document captures several of the key ideas discussed.

Success Stories

We begin by describing some examples in which mathematical research has led to innovations in social computing.

Crowdsourced democracy. YouTube competes with Hollywood as

an entertainment channel, and also supplements Hollywood by acting as a distribution mechanism. Twitter has a similar relationship to news media, and Coursera to universities. But Washington has no such counterpart; there are no online alternatives for making democratic decisions at large scale as a society. As opposed to building consensus and compromise, public discussion boards often devolve into flame wars when dealing with contentious socio-political issues. This motivates the problem of designing systems in which crowds of hundreds, perhaps millions, of individuals collaborate together to come to consensus on difficult societal issues.

Mathematical research has recently led to new systems implementing crowdsourced democracy.²¹ This work builds upon a body of research on social choice that examines how to best take the preferences of multiple agents (human or otherwise) and obtain from them a social decision or aggregate social preference, typically accomplished through some form of voting.^d

Consider situations where a highly structured decision must be made. Some examples are making budgets, assigning water resources, and setting tax rates. Goel et al.²¹ made significant progress towards understanding the “right” mechanisms for such problems. One promising candidate is “Knapsack Voting.” Recall that in the knapsack problem, a subset of items with different values and weights must be placed in a knapsack to maximize the total value without exceeding the knapsack’s capacity. This captures most budgeting processes—the set of chosen budget items must fit under a spending limit, while maximizing societal value. Goel et al.²¹ prove that asking users to compare projects in terms of “value for money” or asking them to choose an entire budget results in provably better properties than using the more traditional approaches of approval or rank-choice voting. Inspired by these mathematical results, Goel et al. designed a participatory budgeting platform that is fast be-

coming the leader for such processes in the U.S.^e For example, this platform was recently used to decide how to spend \$250,000 of infrastructure funds to improve Long Beach (CA) Council District 9, and how to allocate \$2.4 million of Vallejo’s capital improvement budget. Looking forward, it is an interesting and open research challenge to understand if these algorithms and systems yield near-optimal aggregations of societal preferences, or decisions that are near-optimal in terms of overall societal utility.

Automated market makers for prediction markets. A prediction market is a financial market designed to extract and aggregate predictions from a crowd. In a typical prediction market, traders buy and sell securities with payments that are contingent on the outcome of a future event. For example, a security may yield a payment of \$1 if a Democrat wins the 2016 U.S. Presidential election and \$0 otherwise. A trader who believes the true probability of a Democrat winning the election is p maximizes his expected utility by purchasing the security if it is available at a price less than Sp and selling the security if it is available at a price greater than Sp . The market price of this security is thought to reflect the traders’ collective belief about the likelihood of a Democrat winning.

Prediction markets have been shown to produce forecasts at least as accurate as other alternatives in a wide variety of domains, including politics, business, disease surveillance, entertainment, and beyond, and have been widely cited by the press during recent elections. However, markets operated using traditional mechanisms like continuous double auctions (similar to the stock market) often suffer from low liquidity. Without liquidity, a market faces a chicken-and-egg problem: potential traders are dissuaded from participating due to lack of counterparties, which contributes to an even greater reduction in future trading opportunities.

Low liquidity can also lead to high price volatility and large spreads, both of which cause the market price to yield a less meaningful prediction.

To combat this problem, Hanson²³ proposed the idea of operating mar-

a There are also clear connections to, and influence from, older topics and models in the classical mathematical sociology literature.⁶

b Participant list and bios available at <http://bit.ly/1Vy9Ck7>.

c <http://cra.org/ccc/events/theoretical-foundations-for-social-computing/>

d A significant research community concerns itself primarily with computational social choice.⁴ This area has particular promise for social computing because of the problems of scale that are associated with group decision-making online, such as in crowdsourced democracy.

e <https://pbstanford.org/cambridge/approval>

kets using an automated market maker called a market scoring rule. This market maker is an algorithmic agent that is always willing to buy or sell securities at current market prices that depend on the history of trade. Hanson's ideas build on the extensive literature on proper scoring rules,²⁰ payment rules that elicit honest predictions from agents. Market scoring rules ensure the market maker has bounded risk and that traders are unable to engage in arbitrage. Because of these desirable properties, Hanson's market scoring rules have become the prediction market implementation of choice used by companies including Consensus Point, Inkling, and Augur, and large-scale academic projects including SciCast (<http://scicast.org>) and the Good Judgment Project.³⁸

Recently there has been interest in further tapping into the informational efficiency of prediction markets and using them to obtain accurate predictions on more fine-grained events. For example, instead of viewing a Presidential election as having two possible outcomes (Democrat wins or Republican wins), we could view it as having 2^{50} potential outcomes, with each outcome specifying a winner in each U.S. state. Traders could then trade securities on events (combinations of outcomes) to profit on their unique knowledge, such as whether or not the same candidate will win in both Ohio and Florida, or whether or not the Republican candidate will win in at least one of Ohio, Pennsylvania, and Virginia. Such a prediction market is called a combinatorial prediction market. Unfortunately, due to the difficulty of keeping prices logically consistent across large outcome spaces, running market scoring rules off-the-shelf is computationally intractable for many natural examples of combinatorial markets.⁵

In search of pricing rules that are tractable and preserve the logical relationships between security payoffs, Abernethy, Chen, and Vaughan¹ proposed a general framework for the design of efficient automated market makers over very large or infinite outcome spaces. They took an axiomatic approach, defining a set of formal mathematical properties that correspond to economic properties that any reasonable market should satisfy (such as “no arbitrage”



Recently there has been interest in further tapping into the information efficiency of prediction markets and using them to obtain accurate predictions of more fine-grained events.



and an “information incorporation” property) and fully characterized the set of pricing mechanisms that satisfy these properties. Then, using techniques from convex analysis, they provided a method for designing specific market makers that satisfy these properties. The framework enables formal reasoning of trade-offs between different economic features of these market makers as well as evaluating the computational efficiency of the pricing algorithms.

This framework is particularly exciting because it offers a way to think about approximate pricing in combinatorial markets when exact pricing is still intractable. Approximate pricing for markets is challenging because approximation errors may be exploited by traders to cause the market maker to incur a large or even infinite loss. The framework of Abernethy, Chen, and Vaughan¹ characterizes deviations from exact pricing that won't add additional cost to the market maker. Building upon this understanding, Dudík et al.⁹ further developed a computationally tractable method to run a large-scale prediction market that allows participants to trade almost any contract they can define over an exponentially large outcome space. This method is starting to gain traction in industry where it has been used in the PredictWise election market¹⁰ and previous and upcoming iterations of the Microsoft Prediction Service.^f

Fair division for the masses. Social computing systems can be used to help groups of people make decisions about their day-to-day lives. One particularly innovative example is Spliddit,^g a website that provides tools that help groups of people achieve fair allocations. Spliddit currently offers tools to allocate rooms and divide rent payments among roommates, split taxi fares among passengers, assign credit in group projects, divide sets of (divisible or indivisible) goods among recipients, or split up tasks among collaborators. It has been featured in the *New York Times*^h and had tens of thousands of users as of 2014.²²

Spliddit's website boasts “indisputable fairness guarantees.” Indeed, each of the division mechanisms employed on the site stems from the body of re-

^f <http://nyti.ms/1sL06wt>

^g <http://www.spliddit.org/>

^h <http://nyti.ms/1o0TUtO>

search on (computational) fair division³⁶ and comes with provable mathematical guarantees. For example, the algorithm used for room assignment and rent splitting relies on the fact that there always exists an assignment of rooms and a corresponding set of prices that are envy-free: every roommate prefers the room he is assigned to any other room given the prices. Each roommate submits her own value for each of the rooms, under the constraint that the total value of all rooms matches the total rent for the apartment; viewed another way, each roommate is essentially submitting a proposed set of prices for each room such that she would be equally happy obtaining any room at the specified price. The algorithm then maximizes the minimum utility (value of room minus price) of any roommate subject to the constraint that envy-freeness is satisfied. The solution is also Pareto efficient, meaning there is no other allocation that would increase the utility of any roommate without decreasing the utility of another.

As another example, the credit assignment problem is solved using an algorithm of de Clippel et al.⁸ Each collaborator reports the relative portion of credit that he believes should be assigned to each of the other collaborators. For example, on a project with four collaborators, collaborator A might report that collaborators B and C should receive equal credit while D should receive twice as much credit. The algorithm takes these reports as input and produces a credit assignment that is impartial, meaning that an individual's share of credit is independent of his own report, and consensual, meaning that if there is a division of credit that agrees with all collaborators' reports then this division is chosen. While these conditions may not sound restrictive, de Clippel et al.⁸ show they are not simultaneously achievable with three collaborators. Their algorithm therefore requires at least four.

In addition to providing a useful set of tools, part of Spliddit's mission is to "communicate to the public the beauty and value of theoretical research in computer science, mathematics, and economics, from an unusual perspective." Indeed, the project has inspired some members of the public to take an interest in algorithms with prov-



Mathematical and experimental research are complementary and both are needed to develop relevant mathematical foundations for social computing.



able fairness properties. As one example, a representative of one of the largest school districts in California approached the Spliddit team about a problem he was tasked with solving: fairly allocating unused classrooms in public schools to the district's charter schools. This led the Spliddit team, in collaboration with the California school district, to design a practical new approach to classroom allocation that guarantees envy-freeness as well as several other desirable properties.³²

A Challenge Problem: The Crowdsourcing Compiler

A concrete challenge problem for future research in social computing is what might be called the "Crowdsourcing Compiler:"ⁱ the development of high-level programming languages for specifying large-scale, distributed tasks whose solution requires combining traditional computational and networking resources with volunteer (or paid) human intelligence and contributions. The hypothetical compiler would translate an abstract program into a more detailed organizational plan for machines and people to jointly carry out the desired task. In the same way that today's Java programmer is relieved of low-level, machine-specific decisions (such as which data to keep in fast registers, and which in main memory or disk), the future crowdsourcing programmer would specify the goals of their system, and leave many of the implementation details to the Crowdsourcing Compiler. Such details might include which components of the task are best carried out by machine and which by human volunteers; whether the human volunteers should be incentivized by payment, recognition, or entertainment; how their contributions should be combined to solve the overall task; and so on. While a fully general Crowdsourcing Compiler might well be unattainable, significant progress toward it would imply a much deeper scientific understanding of crowdsourcing than we currently have, which in turn should have great engineering benefits. Noteworthy research efforts which can be viewed as steps on the path to the Crowdsourcing Compiler include Emery Berger's AutoMan Proj-

ⁱ See <http://bit.ly/20juYEX> and <http://bit.ly/1nIyc3P>.

ect (<http://emeryberger.com/research/automan/>),³ as well as both academic and commercial efforts to automate workflow in crowdsourcing and social computing systems (see, for example, <http://groups.csail.mit.edu/uid/turkit/> and <http://www.crowdflower.com/>).

We note the organizational schemes in most of the successful crowdsourcing examples to date share much in common. The tasks to be performed (for example, building an online encyclopedia, labeling images for their content, creating a network of website bookmark labels, finding surveillance balloons) are obviously parallelizable, and furthermore the basic unit of human contribution required is extremely small (fix some punctuation, label an image, and so on). Furthermore, there is usually very little coordination required between the contributions. The presence of these commonalities is a source of optimism for the Crowdsourcing Compiler—so far, there seems to be some shared structure to successful crowdsourcing that the compiler might codify. But are such commonalities present because they somehow delineate fundamental limitations on successful crowdsourcing—or simply because this is the “low-hanging fruit?”

Today, the Crowdsourcing Compiler is clearly a “blue sky” proposal meant more to delineate an ambitious research agenda for social computation than to serve as a guide to short-term steps. But we believe that such an agenda would both need and drive research on theoretical foundations. First steps toward developing the mathematical foundations of a Crowdsourcing Compiler include formally addressing the following questions:

- For a given set of assumptions about the volunteer force, and given the nature of the task, what is the best scheme for organizing the volunteers and their contributions? For instance, is it a “flat” scheme where all contributors are equal and their contributions are combined in some kind of majority vote fashion? Or is it more hierarchical, with proven and expert contributors given higher weight and harder subproblems? Which of these (or other) schemes should be used under what assumptions on the nature of the task and what assumptions on the volunteers?

- How can we design crowdsourced

systems for solving tasks that are much more challenging and less “transactional” than what we currently see in the field—for instance, complex problems where there are strong constraints and interdependencies between the contributions of different volunteers? Behavioral research in recent years has shown that groups of humans can indeed excel on such tasks,³¹ but we are far from understanding when and why.

Finally, we note that while the comparison to traditional compilers might be a useful guide and metaphor, a crowdsourcing analogue would have to face a variety of issues that simply do not arise with standard hardware and software. In addition to the aforementioned challenges of deciding how to organize and incentivize human contributions, there may also be the potential for malicious or deceptive behavior by workers, and the need for error correction of crowd work (which is currently largely handled by redundancy and voting techniques).

Challenges to Overcome

We have argued that mathematical research has the potential to make great contributions to social computing. However, before this potential is fully realized, there are several challenges that must be addressed.

Blending mathematical and experimental research. Mathematical and experimental research are complementary and both are needed to develop relevant mathematical foundations for social computing. The strengths of mathematical work include:

1. Mathematical modeling and analysis can be used to cleanly formulate and answer many questions about system behavior without requiring that we build a complete system, providing us with a tool to evaluate the impact of design decisions before committing to any particular design. For example, such models can provide guidance on how to increase participation (such as, comparing a leaderboard to badges^{16,27}), predict whether a social computing system will achieve critical mass, and perhaps understand how the behavior of groups of users change as the system scales.

2. Mathematical guarantees are desirable for properties like user privacy (that can be obtained, for example, us-

ing techniques from the extensive and growing literature on differential privacy¹¹), correctness of a system’s output, or the scalability of a social computing system.

3. Theoretical work in computer science provides tools for designing and analyzing new algorithms that could lie at the heart of social computing applications, answering questions like how to aggregate noisy and unstructured estimates or information from crowds,^{25,30} how to optimally divide a community into subgroups, or how to bring people together in moments of spare time to achieve a common goal.

4. Mathematical models can be used to explore counterfactual analysis, something that is notoriously difficult to do through experiments alone.

Needless to say, mathematical modeling should not and cannot replace experimental work. A mathematical theory can only be truly tested through experiments, and discrepancies between the theory and experimental results provide guidance about how to revise the theory. For example, the ability of mathematical models to make valuable predictions about system behavior depends on an accurate model of system users, which is generally best developed through experimental work.

Learning from the social sciences.

Computer scientists cannot develop the mathematical foundations of social computing in isolation. Social computing systems are fundamentally social. These systems cannot be properly modeled or analyzed without accounting for the behavior of their human components. Much of the literature thus far uses standard models of economic agents and corresponding assumptions about agent preferences, but a growing literature based on experimental work on online platforms suggests that human behavior in several online settings might deviate from these models,^{26,35,40} and these deviations can have significant consequences for how to optimally design social computing systems.^{13,26}

In order for mathematical foundations to provide useful practical results, it is necessary to base it on models that better reflect human behavior. This is most effectively achieved via a dialog between theoretical and experimental and empirical research, with studies of human behavior informing mathematical

modeling choices, as well as mathematical results suggesting the most important agent characteristics to understand via experimental research. It will be important to understand and incorporate relevant research from psychology, economics, sociology, and other fields. For example, behavioral economics and psychology provide insight into how humans respond to incentives.

Generalization. Most of the existing mathematical work on social computing focuses on a single application. What does the research on prediction market design tell us about recommendation systems or citizen science? Models will have the most potential for impact if they incorporate reusable components, allowing results to generalize to many systems. (This is one motivation for the Crowdsourcing Compiler discussed earlier.)

A related issue is the lack of consensus and understanding of the “core social computing problems,” or even if such a set of core problems exists. Mathematical theories are typically developed with one or more such core problems in mind.

Such problems should capture challenges that span a wide range of applications and be robust to small changes in the applications to be sure that they are capturing something “real.” Clearly, the identification of such problems requires a dialog between practitioners building real systems and theoreticians to identify the most pressing problems requiring mathematical study.

Transparency, interpretability, and ethical implications. One final challenge to overcome is the potential need to make social computing algorithms and models transparent and interpretable to the users of social computing systems. Users are becoming increasingly sophisticated and are aware the algorithms employed online impact both their day-to-day user experience and their privacy. When faced with the output of an algorithm, many will question where this output came from and why. It is already difficult to explain to users why complex probabilistic algorithms and models produce the results they do, and this will only become more difficult as algorithms integrate human behavior to a larger extent.

The issue of algorithmic transparency is often tied to ethical concerns

such as discrimination and fairness. Examining and avoiding the unintended consequences of opaque decisions made by algorithms is a topic that has been gaining interest in the machine learning and big data communities.^j Such concerns will undoubtedly need to be addressed in the context of social computing as well.

Acknowledgments. We thank the participants of the Visioning Workshop on Theoretical Foundations for Social Computing for their contributions. We also thank Ashish Goel, Vince Conitzer, David McDonald, David Parkes, and Ariel Procaccia for their feedback. C

j For example, see the series of recent workshops on Fairness, Accountability, and Transparency in Machine Learning (<http://www.fatml.org/>).

References

1. Abernethy, J., Chen, Y. and Vaughan, J.W. Efficient market making via convex optimization, and a connection to online learning. *ACM Trans. Economics and Computation* 1, 2 (2013), Article 12.
2. Abernethy, J., Frongillo, R., Li, X. and Vaughan, J.W. A general volume-parameterized market-making framework. In *Proceedings of the 15th ACM Conference on Economics and Computation*, 2014.
3. Barowy, D., Cutsinger, C., Berger, E.D. and McGregor, A. Automan: A platform for integrating human-based and digital computation. In *Proceedings of the Object-Oriented Programming, Systems, Languages, and Applications*, 2012.
4. Brandt, F., Conitzer, V., Endriss, U., Lang, J. and Procaccia, J.D., Eds. *Handbook of Computational Social Choice*. Cambridge University Press. Forthcoming.
5. Chen, Y., Fortnow, L., Lambert, N., Pennock, D. and Wortman, J. Complexity of combinatorial market makers. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, 2008.
6. Coleman, J. *Introduction to Mathematical Sociology*. Free Press of Glencoe, 1964.
7. Dasgupta, A. and Ghosh, A. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International World Wide Web Conference*, 2013.
8. de Clippel, G., Moulin, H. and Tideman, N. Impartial division of a dollar. *J. Economic Theory* 139, 1 (2008), 176–191.
9. Dudík, M., Lahaie, S. and Pennock, D.M. A tractable combinatorial market maker using constraint generation. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 2012.
10. Dudík, M., Lahaie, S. and Pennock, D.M. and Rothschild, D. A combinatorial prediction market for the U.S. elections. In *Proceedings of the 14th ACM Conference on Electronic Commerce*, 2013.
11. Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 34 (2014), 211–407.
12. Easley, D. and Ghosh, A. Incentives, gamification, and game theory: An economic approach to badge design. In *Proceedings of the 14th ACM Conference on Electronic Commerce*, 2013.
13. Easley, D. and Ghosh, A. Behavioral mechanism design: Optimal crowdsourcing contracts and prospect theory. In *Proceedings of the 16th ACM Conference on Economics and Computation*, 2015.
14. Ghosh, A. Game theory and incentives in human computation. *Handbook of Human Computation*. P. Michelucci, ed. Springer, 2014.
15. Ghosh, A. and Hummel, P. A game-theoretic analysis of rank-order mechanisms for user-generated content. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, 2011.
16. Ghosh, A. and Kleinberg, R. Behavioral mechanism design: Optimal contests for simple agents. In *Proceedings of the 15th ACM Conference on Economics and Computation*, 2014.
17. Ghosh, A. and McAfee, P. Incentivizing high-quality user generated content. In *Proceedings of the 20th*

18. Ghosh, A. and McAfee, P. Crowdsourcing with endogenous entry. In *Proceedings of the 21st International World Wide Web Conference*, 2012.
19. Giles, J. Computational social science: Making the links. *Nature* 488, 7412 (2012).
20. Gneiting, T. and Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. American Statistical Association* 102, 477 (2007), 359–378.
21. Goel, A., Krishnaswamy, A., Sakshuwong, S. and Altamurto, T. Knapsack voting. *Collective Intelligence*, 2015.
22. Goldman, J. and Procaccia, A.D. Spliddit: Unleashing fair division algorithms. *SIGecom Exchanges* 13, 2 (2014), 41–46.
23. Hanson, R. Combinatorial information market design. *Information Systems Frontiers* 1 (2003), 105–119.
24. Ho, C.-J. and Vaughan, J.W. Online task assignment in crowdsourcing markets. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.
25. Ho, C.-J., Jabbari, S. and Vaughan, J.W. Adaptive task assignment for crowdsourced classification. In *Proceedings of the 26th International Conference on Machine Learning*, 2013.
26. Ho, C.-J., Slivkins, A., Suri, S. and Vaughan, J.W. Incentivizing high quality crowdwork. In *Proceedings of the 24th International World Wide Web Conference*, 2015.
27. Immorlica, N., Stoddard, G. and Syrgkanis, V. Social status and badge design. In *Proceedings of the 24th International World Wide Web Conference*, 2015.
28. Jain, S. and Parkes, D.C. A game-theoretic analysis of the ESP Game. *ACM Trans. Economics and Computation* 1, 1 (2013), 3:1–3:35.
29. Jain, S., Chen, Y. and Parkes, D.C. Designing incentives for online question-and-answer forums. *Games and Economic Behavior* 86 (2014), 458–474.
30. Karger, D., Oh, S. and Shah, D. Iterative learning for reliable crowdsourcing systems. *Advances in Neural Information Processing Systems*, 2011.
31. Kearns, M. Experiments in social computation. *Commun. ACM* 55, 10 (Oct. 2012).
32. Kurokawa, D., Procaccia, A.D. and Shah, N. Leximin allocations in the real world. In *Proceedings of the 16th ACM Conference on Economics and Computation*, 2015.
33. Lambert, N.S., Langford, J., Vaughan, J.W., Chen, Y., Reeves, D., Shoham, Y. and Pennock, D.M. An axiomatic characterization of wagering mechanisms. *J. Economic Theory* 156 (2015), 389–416.
34. Lazer, D. et al. Computational social science. *Science* 323, 5915 (2009), 721–723.
35. Mason, W. and Watts, D. Financial incentives and the “performance of crowds.” In *Proceedings of the First Human Computation Workshop*, 2009.
36. Procaccia, A.D. Cake cutting: Not just child’s play. *Commun. ACM* 56, 7 (July 2013), 78–87.
37. Slivkins, A. and Vaughan, J.W. Online decision making in crowdsourcing markets: Theoretical challenges. *ACM SIGecom Exchanges* 12, 2 (2013).
38. Ungar, L., Mellors, B., Satopää, V., Baron, J., Tetlock, P., Ramos, J. and Swift, S. The good judgment project: A large-scale test of different methods of combining expert predictions. AAAI Technical Report FS-12-06, 2012.
39. Waggoner, B. and Chen, Y. Output agreement mechanisms and common knowledge. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing*, 2014.
40. Yin, M., Chen, Y. and Sun, Y.-A. The effects of performance-contingent financial incentives in online labor markets. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, 2013.

Yiling Chen (yiling@seas.harvard.edu) is Gordon McKay Professor of Computer Science at Harvard University, Cambridge, MA.

Arpita Ghosh (arpitaghosh@cornell.edu) is an associate professor of information science at Cornell University, Ithaca, NY.

Michael Kearns (mkearns@cis.upenn.edu) is a professor and National Center Chair of Computer and Information Science at the University of Pennsylvania, Philadelphia, PA.

Tim Roughgarden (tim@cs.stanford.edu) is an associate professor of CS at Stanford University, Stanford, CA.

Jennifer Wortman Vaughan (jenn@microsoft.com) is a senior researcher at Microsoft Research, New York, NY.

Copyright held by owners/authors.
Publication rights licensed to ACM. \$15.00.

research highlights

P. 110

**Technical
Perspective**
**The Chemistry of
Software-Defined
Batteries**

By Srinivasan Keshav

P. 111

Software-Defined Batteries

By Anirudh Badam, Ranveer Chandra, Jon Dutra, Anthony Ferrese, Steve Hodges, Pan Hu, Julia Meinershagen, Thomas Moscibroda, Bodhi Priyantha, and Evangelia Skiani

P. 120

**Technical
Perspective**
**3D Image Editing
Made Easy**

By Takeo Igarashi

P. 121

Extracting 3D Objects from Photographs Using 3-Sweep

By Tao Chen, Zhe Zhu, Shi-Min Hu, Daniel Cohen-Or, and Ariel Shamir

Technical Perspective

The Chemistry of Software-Defined Batteries

By Srinivasan Keshav

ABOUT FOUR DECADES ago, the widespread availability of low-cost, general-purpose microprocessors allowed them to be incorporated into many real-world systems, where they replaced analog or purpose-built digital controls. For instance, small microprocessors are embedded in systems as diverse as laser printers, DVD players, and digital watches. However, these microprocessors are incapable of hosting a general-purpose operating system, so programming them has required specialized technically adept programmers. In recent years, with the advent of low-cost multi-GHz microprocessors that can host a general-purpose operating system, such as Linux or Android, even naïve programmers can use a sophisticated high-level development environment to engage with the physical world. Under the broad banner of cyber-physical systems, this has allowed computation to make inroads into areas as diverse as the electrical grid, water distribution systems, building heating, ventilation and air conditioning systems, and electric vehicles.

How should such systems be designed? A time-tested principle in computer systems design is to use an interface to separate an abstraction from its implementation.^a This separation allows independent evolution of each component and for complex policies to be layered over underlying simple mechanisms, which in turn could be composed from yet simpler ones. A remarkable success of this design philosophy is the architecture of the Internet, where each layer of abstraction adds a limited level of additional complexity, and the top level of the layered protocol stack allows even naïve programmers to access the entire power of the Internet: a novice Python programmer can, in a few lines of code, obtain a reliable, byte-ordered, end-to-end connec-

tion from any Internet endpoint to any other endpoint, a task that requires a considerable amount of complex protocol machinery to accomplish.

The following paper represents a convergence of these two evolutionary threads. It combines the use of a general-purpose operating system to manage a new category of physical devices (storage batteries) along with the design of an interface that cleanly separates the complex operating policies for the battery system from the physical mechanisms that implement this interface. This allows programmers to come up with new, user- or application-specific storage management policies without having to understand the complexities of a battery management system.

This paper is significant along three dimensions. First, in order to provide the interface, it was necessary to redesign battery management hardware. The new hardware (a switched mode regulator) allows fine-grained control over the amount of power entering or leaving each battery, something heretofore impossible. It is interesting to see how the need for a good software interface forced hardware innovation. Second, the type of battery system discussed in this paper can be found in a dizzying array of modern devices: lap-

tops, mobile phones, smart glasses, tablets, drones, vacuum cleaners, electric cars, and even electrical grid substations. Thus, this work has potential for enormous real-world impact. Finally, from a commercial perspective, it is fascinating to note the authors work at Microsoft and Tesla. Microsoft has long viewed itself as an operating systems and applications provider for desktop, laptop, and mobile computers, and it now appears to wish to be the operating system for vacuum cleaners and cars, as well.

The paper offers some key insights. To start, batteries are not a single, homogeneous technology, but complex devices that represent a balance between power density, energy density, cost, volume, and mass, amongst other things. So, choosing the right battery for a device is a non-trivial task. Second, the paper drives home the point that, just as a memory system combines heterogeneous technologies, so should an ideal battery system. Third, the use of a simple but powerful software-hardware interface allows the development of arbitrarily complex operating strategies without require hardware to be rebuilt. This enables rapid innovation in storage systems. Finally, the evaluation of the system uses both a prototype hardware system and a validated software emulator. This combination allows rapid exploration of the design space without sacrificing realism.

In summary, although at first glance the paper seems to be about a dry and somewhat technical topic—the management of heterogeneous systems—when viewed from a different perspective, it represents the advancement of computational thinking, general-purpose computing, and Microsoft into new realms. Enjoy! 

Srinivasan Keshav is a professor and Cisco Chair at the David R. Cheriton School of Computer Science, University of Waterloo, Canada.

Copyright held by author.

The following paper drives home the point that just as a memory system combines heterogeneous technologies, so should an ideal battery system.

^a Lampson, B.W. Hints for computer system design. *ACM SIGOPS Review* 17, 5 (1983).

Software-Defined Batteries

By Anirudh Badam, Ranveer Chandra, Jon Dutra, Anthony Ferrese, Steve Hodges, Pan Hu, Julia Meinershagen, Thomas Moscibroda, Bodhi Priyantha, and Evangelia Skiani

Abstract

Different battery chemistries perform better on different axes, such as energy density, cost, peak power, recharge time, longevity, and efficiency. Mobile system designers are constrained by existing technology, and are forced to select a single chemistry that best meets their diverse needs, thereby compromising other desirable features. In this paper, we present a new hardware–software system, called Software Defined Battery (SDB), which allows system designers to integrate batteries of different chemistries. SDB exposes application programming interfaces (APIs) to the operating system, which controls the amount of charge flowing in and out of each battery, enabling it to dynamically trade one battery property for another depending on application and/or user needs. Using micro-benchmarks from our prototype SDB implementation, and through detailed simulations, we demonstrate that it is possible to combine batteries which individually excel along different axes to deliver an enhanced collective performance when compared to traditional battery packs.

1. INTRODUCTION

The utility of a mobile device is often constrained by the capabilities of its battery. While integrated circuit performance has doubled every 18 months according to Moore’s law, the same is far from true for battery technology. Battery performance can be evaluated in many different ways (see Table 1), but no matter which metric we look at, it has taken more than a decade to double performance.

Furthermore, the various properties of batteries are often at odds with each other. For example, batteries with higher

power densities tend to have lower volumetric and gravimetric energy densities, and vice versa. Similarly, making a conformable battery that fits a particular industrial design compromises its performance characteristics.

Such tradeoffs are present even within a given physical battery. For example, energy delivered by a battery in a single charge–discharge cycle (energy capacity) is inversely related to the rate at which the battery is drained (discharge rate). This is because the resistance losses inside a battery are proportional to the square of the current. Similarly, a battery’s longevity—its ability to perform consistently following many charge–discharge cycles—is inversely related to the discharge and recharge rates. This is because higher currents speed up the creation of fissures in the electrodes that reduce the amount of energy a battery can store.

In summary, no single battery type can deliver the ever-growing list of requirements of modern devices: fast charging, high capacity, low cost, less volume, light weight, less heating, better longevity, and high peak discharge rates.

A growing range of battery chemistries are under development, each of which delivers a different set of benefits. We believe that combining multiple of these heterogeneous batteries instead of using a single battery chemistry can allow a mobile system to dynamically trade between their capabilities and thereby offer attractive tradeoffs.

However, traditional methods of integrating multiple batteries are not suitable for heterogeneous batteries. Connecting them in series or parallel does not provide enough control over usage: batteries connected in series can only supply the same amount of current; batteries connected in parallel must operate at the same voltage and can only supply currents that are inversely proportional to their internal resistances.

We propose a new system, called Software Defined Battery (SDB), that allows heterogeneous batteries with different chemistries to be integrated in a mobile system. SDB consists of hardware and software components. The hardware enables fine-grained control of the amount of power going in and out of each battery using smart switching circuitry. The software, which resides in the operating system (OS), computes how much power to draw from each battery, and how to recharge each battery.

Deciding how much power to draw from and how to charge each battery is nontrivial. It depends on the efficiency of each battery under different workloads, the age of each battery, and also the usage profile. For example, if a high power workload is anticipated in the future, then it could be worth-while

Table 1: A number of battery characteristics.

Battery characteristics	Units
Energy capacity	J
Volume	mm ³
Mass	kg
Discharge rate	W
Recharge rate	W
Gravimetric energy density	J/kg
Volumetric energy density	J/L
Cost	\$/J
Discharge power density	W/kg
Recharge power density	W/kg
Cycle count	Number of discharge/recharge cycles
Longevity	% of original capacity after N cycles
Internal resistance	Ohm
Efficiency	% of energy turned into heat
Bend radius	mm

These are often in tension with each other, for example, increasing recharge rate compromises longevity.

The original version of this paper was published in *Proceedings of ACM SOSP’15*.

conserving charge on the battery that is more capable of handling such a workload in an efficient manner.

The SDB software component that resides in the OS implements a set of policies, and uses simple application programming interfaces (APIs) to communicate with the SDB hardware. The algorithms implemented by this software use various metrics to decide the ratios in which to discharge and charge each battery, such that the charge–discharge duration of the device is increased, and degradation of the batteries is reduced. We present the details of the APIs and policies in Section 3.3.

The SDB design is cross-layer and involves new chemistries, additional hardware, and new OS components. This approach opens up new battery parameters, previously unavailable to OS designers, for resource optimization. In existing mobile devices, the battery is usually treated as a black box, and is simply assumed as a reservoir of charge. As we show in Section 5, OS techniques yield substantial gains in battery usage. This design also allows a system designer to select any combination of batteries for an optimal design, including new chemistries as they are developed using just software updates.

Even with existing batteries, SDB enables several new scenarios, such as: (i) fast-charging devices that can gain a significant percentage of their charge in just a few minutes without causing unexpected battery degradation, (ii) long-lived wearables created by combining flexible bendable batteries with traditional batteries, and (iii) efficient 2-in-1 laptop-tablet convertible devices with battery usage tailored to the user's behavior.

2. BATTERY BACKGROUND

A Li-ion battery contains a negative electrode (the anode), which is usually made of graphite and a positive electrode (the cathode), which is typically a metal oxide. A separator ensures physical separation between the anode and the cathode to prevent shorting, and the battery is filled with an electrolyte composed of a lithium-based salt whose ions can easily pass through the separator. Current is discharged when the electrodes are connected externally over a resistive load while positive lithium ions flow from the anode to the cathode through the electrode and the separator. During charging, Li-ion batteries store energy by trapping positive lithium ions in the anode when an external potential is applied.

Li-ion battery capabilities, such as longevity, energy density, and internal resistance, are largely determined by the materials used for the electrodes and the separator. The battery's gravimetric and volumetric energy densities are affected by the strength of the separator. The resistance of the battery, and hence its inefficiencies, depend on the resistance of the separator, which typically increases with the age of the battery. The power density of the battery is also affected by aging. The structural integrity of the electrodes determines how much energy they can store—some lithium ions get permanently trapped in the anode. The anodes can develop cracks as they age, which can ultimately reduce both energy and power densities.

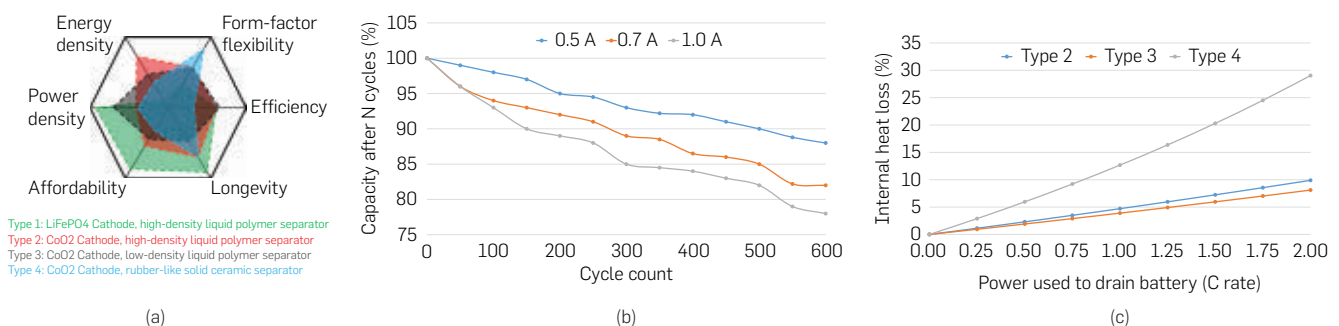
Figure 1a demonstrates the capabilities of four different Li-ion batteries, which differ in the chemistry of materials used for the cathode and the separator. Batteries of Type 1 are typically used in powered tools that need to charge quickly and provide high power for a short duration of time. Such batteries are a poor choice for mobile devices because of their poor energy density—a Type 1 battery is usually double the volume of a Type 2 battery with the same energy capacity. Type 2 batteries are commonly used in most mobile devices today. We measure the loss in capacity with respect to number of charge–discharge cycles for a sample Type 2 battery, and observe that the battery degrades much faster when discharged at higher current (Figure 1b).

Type 3 batteries are an emerging variation over Type 2 that have a slightly higher power density at the expense of some energy density. This is achieved by making the separator less dense allowing more lithium ions to pass through per unit time. This usually leads to decreased energy density as separators cannot store energy—only the electrodes can. Finally, Type 4 is another emerging battery that is flexible and bendable because of the physical properties of the rubber-like (ceramic-based) separator used—while the electrodes are implemented by coating material along the cell's walls. Unfortunately, such separators increase the resistance to passage of ions and thereby result in higher inefficiency, as shown in Figure 1c.

2.1. Typical power management

Figure 2 shows a block diagram of the typical power management hardware. It consists of a (i) battery, (ii) fuel gauge, (iii) battery charger, and (iv) voltage regulator.

Figure 1. Li-ion battery properties. (a) Li-ion batteries compared. (b) Charging rate affects longevity. (c) Discharging rate versus lost energy.



A *Battery pack* has one or more battery cells. Multiple cells are used to achieve higher voltage or higher capacity. While such multi-cell configurations exist today, for example in the Surface Pro, Galaxy Tab, and iPad, these cells have the same chemistry and are either connected in series, parallel, or a combination thereof. They are treated as a single monolithic battery by the OS. Our aim is to use a heterogeneous set of cells and achieve wide dynamic characteristics by exposing the cells directly to the OS.

The *Fuel gauge* keeps track of the state of charge (SoC) of the battery by measuring the voltage across the battery terminals, and the coulombs flowing in and out of it. This information is exposed to the OS.

The *Battery charger* charges the battery with an appropriate charging current profile based on the battery's SoC, the terminal voltage (the potential difference between the anode and the cathode), and the capability of the power source.

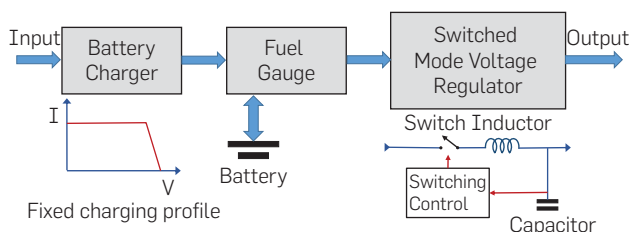
Due to the battery's internal resistance R , the battery terminal voltage changes with the load current I due to the IR voltage drop. The internal resistance and the battery voltage themselves change with the SoC. The job of the *Voltage Regulator* is to hide these terminal voltage variations due to changing potentials at the electrodes and the changing internal resistance and present a constant voltage to the load. Mobile devices use switched mode voltage regulators due to their high efficiency. As the name implies, a switch mode power supply contains a switch that opens and closes to transfer packets of energy (Figure 2). A control loop maintains a constant voltage under varying load currents by changing the energy per packet or the packet switching frequency.

Typically, all these modules are contained in a single power management integrated circuit (PMIC), which communicates with the OS over a serial bus. In current designs, the interactions between the OS and PMIC are limited to *query* operations, such as inquiring about remaining SoC. However, none of these APIs allow the OS to set the battery parameters, and in particular to change the amount of charge to be drawn from or provided to each cell within a battery pack. Through the SDB system, we propose enabling fine grain control by exposing a richer software API to the OS to dynamically change the amount of charge to be drawn from or provided to each battery.

3. SDB DESIGN

SDB allows a device to use diverse batteries through fine-grain control of the amount of charge flowing in and out

Figure 2. Traditional power management hardware.



of each battery. SDB provides APIs to the OS to change the aforementioned power values based on user workload. We describe the SDB system in detail in this section.

3.1. System overview

The SDB system spans components across three layers: the batteries and their chemistry, the battery management circuit, and the OS. We outline these components and their interactions in Figure 3.

SDB allows a system designer to combine diverse batteries. The particular batteries chosen depend on the scenario, such as a fast charging battery and a high energy battery for a tablet, or a bendable battery and high energy battery for a smart-watch.

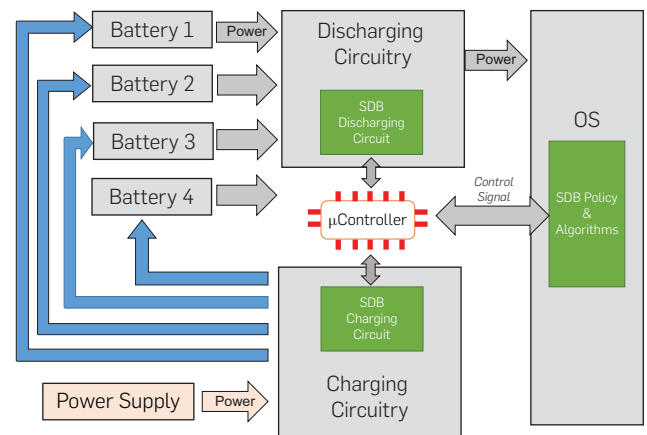
However, combining different battery types is not trivial. These batteries might have different capacities and different terminal voltages. Therefore, we design a new power distribution circuit for fine-grain control of how multiple batteries are discharged to support system load. A microcontroller interfaces between this power distribution circuitry and the mobile device OS to control the charging and discharging of batteries accordingly.

To enable flexibility in design, and to allow quick changes in policy, we only implement the mechanisms in hardware, and all policies are managed and set by the OS. A runtime component in the OS monitors the charging and discharging behavior and accordingly sets policies that meet user expectations in terms of daily battery life and longevity of the battery-pack.

3.2. SDB hardware

The SDB hardware needs to support discharging and charging across multiple, heterogeneous batteries. For discharging, it has to provide a flexible mechanism for fine-grain control of how the load current is supplied from each battery. This should support two things: coarse grain *switching* of the load across multiple batteries where the total load is supplied by a particular battery for an extended period of time and the fine grain *sharing* of the load where a certain fraction of the load is drawn from each battery.

Figure 3. SDB system overview.



For charging, the SDB hardware has to support control over how batteries are charged. In contrast to existing solutions where batteries are charged according to a fixed charging profile, SDB requires setting of charging currents and charging profiles dynamically based on OS policies. Under certain circumstances, it should even be possible to charge a battery from another one.

Designing these flexible charging and discharging circuits are challenging for two reasons. First, due to the high currents that flow in these circuits, any electronic component in series with the current flow will cause energy losses. Hence, these circuit designs should introduce as few of these components as possible. Second, each extra component we introduce can increase the weight, volume, and bill of material (BoM) cost of the device, which will make the proposed solution unattractive in the competitive hardware market.

SDB discharging circuit design. A simple discharging circuit can be implemented using a combination of an electronic switch and a capacitor as shown in Figure 4a. The microcontroller achieves load switching by connecting the appropriate battery to the load. To achieve load sharing, the load is switched between the batteries at a high frequency in round-robin fashion. The ratio of the current draw is determined by the fraction of time the switch is connected to a particular battery. The capacitor acts as an energy store to smooth out the discontinuities due to switching. Parasitic battery capacitance and external capacitors smooth out the high frequency battery current.

However, this naive implementation has two main drawbacks. First the switch, typically implemented using a Field Effect Transistor (FET), has a finite on resistance that causes significant power loss at high load currents. Second, a switch with high power handling capability and the necessary capacitors increase the BoM cost and space required.

To overcome these shortcomings, we designed a new switched mode regulator architecture that integrates fine-grain battery switching into the regulator itself. As shown in Figure 4c, we restructure the built-in switch to achieve voltage regulation and support switching between multiple batteries—by drawing packets of energy from the batteries in a weighted round-robin fashion. We reuse the storage

capacitor to smooth out the load current variations due to switching. We have evaluated the correctness of the proposed solution under different battery voltages and load conditions by running LTSPICE¹¹ simulations.

SDB charging circuit design. The SDB charging circuit should have the ability to charge batteries at a configurable rate and also charge them from each other. Given that such charging should be possible irrespective of the battery voltage, the batteries should be connected through a buck-boost regulator, such that the energy source is at the input and the energy sink is at the output of the regulator. A buck-boost regulator is a particular form of switching regulator where the regulator output voltage can be either less than or greater than its input voltage.

Apart from different charging configurations, SDB requires dynamic fine-grain control over the charging profile. This is achieved by instrumenting each switched mode regulator with multiple charging profiles where the SDB microcontroller dynamically selects the appropriate charging profile based on OS policy decisions.

Figure 4b shows how these modules can be combined to implement a flexible charging circuit. However, a major drawback of this configuration is the large number of switching regulators ($O(N^2)$ for N batteries) required, which negatively impacts the device BoM cost and space requirements.

Instead, we design an optimized charging circuit as shown in Figure 4c, which requires only $O(N)$ switched mode regulators to charge N batteries. When an external supply is present, the microcontroller configures both R_1 and R_2 in *buck* mode to charge the batteries. When external power is removed, R_1 and R_2 are disabled. When B_2 is to be charged from B_1 , R_1 operates in *reverse buck* mode while R_2 operates in *buck* mode and vice versa.

3.3. SDB policies and APIs

Our current SDB software architecture is illustrated in Figure 5. An *SDB Runtime* encapsulates the SDB microcontroller from the rest of the OS. The SDB Runtime is responsible for all scheduling decisions affecting the charging and discharging of batteries. It takes clues from the rest of the OS, and communicates the charging and discharging scheduling decisions to the SDB controller.

Figure 4. (a) A simple switch and capacitor-based battery switching solution. (b) A naive implementation of a flexible charging circuit consisting of two buck regulators and two buck-boost regulators with dynamic charging parameters. (c) SDB hardware architecture for an example two-battery system. The switched mode regulator implements discharge across multiple batteries and the reverse buck regulators implement charge.

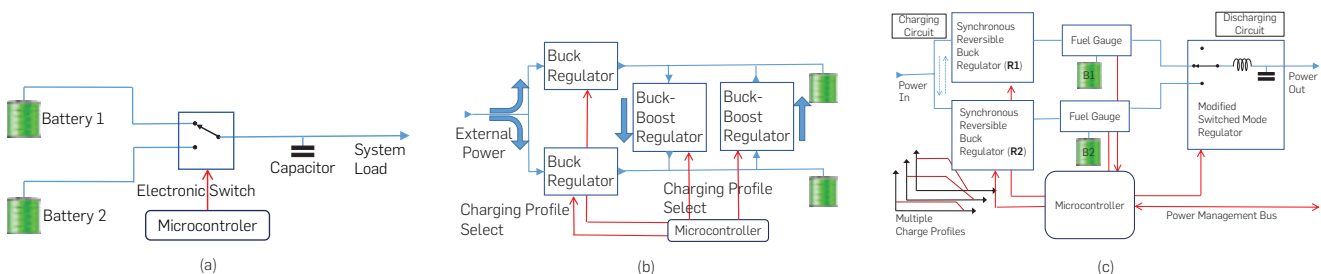
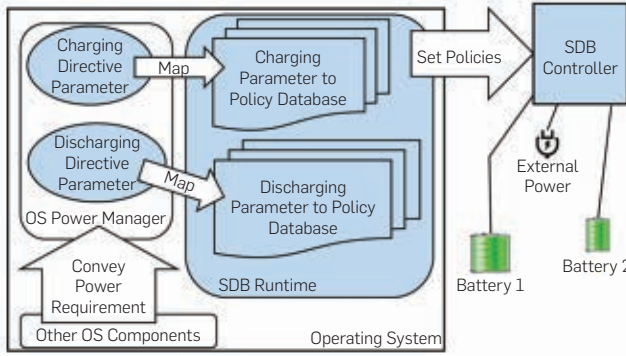


Figure 5. SDB software architecture.



APIs: For a system with N batteries, the SDB Runtime maintains two N -tuples $(c_1, \dots, c_N)$ and $(d_1, \dots, d_N)$ of non-negative values, one for charging and one for discharging. In both cases, the N values add up to one and represent power ratios, that is, the numbers represent the fraction of power that must go in and out of each of the N batteries. The runtime communicates with the SDB microcontroller using the following four APIs:

- **Charge** $(c_1, c_2, \dots, c_N)$: Charge N batteries in proportion to $c_1, c_2, \dots, c_N$, when being charged from an external source.
- **Discharge** $(d_1, d_2, \dots, d_N)$: Discharge N batteries in proportion to $d_1, d_2, \dots, d_N$, when being discharged.
- **ChargeOneFromAnother** (X, Y, W, T) : Charge battery Y from battery X with a power of W for time T .
- **QueryBatteryStatus** $()$: Returns an array with SoC, terminal voltages and cycle counts for each battery.

The SDB Runtime affects changes in the charging and discharging behavior by adapting the $2N$ numbers and sending them to the microcontroller using the above APIs, which enforces the ratios. Such changes can be triggered for example by a change of the user's needs, the battery state, workload patterns, or external factors such as a change in device temperature, etc. Determining optimal battery charging/discharging policies, is nontrivial, and the underlying algorithmic problems are deep and interesting. Often, various battery properties are in tension with one another. For example, fast-charging a battery all the time can greatly accelerate its aging. In this paper, we only scratch the surface of these algorithmic problems and instead describe a set of natural policy heuristics that exhibit good albeit non-optimal performance.

Metrics: Two key metrics any charging/discharging policy seeks to optimize are *Cycle Count Balance* (CCB) and *Remaining Battery Lifetime* (RBL). The RBL metric simply captures the remaining battery lifetime of the device, assuming that no further charging occurs in the future. In other words, RBL is the amount of useful charge in the batteries. The CCB metric reflects that—ideally—the charging and discharging policies should maximize longevity of the device, by balancing the charging cycles of each battery. In

a heterogeneous battery system, each battery is a unique precious resource that excels on a few metrics of interest described in Table 1. Therefore, having a metric-like the CCB ensures that these batteries are aging such that the properties of a battery that the user is most interested in are preserved over time. For example, a battery that has the ability to charge fast must be treated as a precious resource for a user who relies on fast charging during low-battery situations.

Concretely, let χ_i be the number of charging cycles tolerable by battery i before its capacity drops below some acceptable threshold, and let cc_i be the number of charging cycles of battery i . The *wear-ratio* $\lambda_i = cc_i / \chi_i$ describes what fraction of the tolerable recharge cycles have already been consumed by battery i . We define CCB as the ratio $CCB = \max_i \lambda_i / \min_j \lambda_j$, that is, the ratio between the most worn-out battery, normalized to each battery's total tolerable cycle count. A device's longevity is maximized by balancing CCB.

Charge/discharge algorithms. The heuristics currently driving our SDB Runtime are simple and driven by the following observation: It is possible to derive charging and discharging algorithms that, in isolation, optimize the CCB and the instantaneous RBL metric. We use these four “optimal” algorithms (CCB-Charge, RBL-Charge, CCB-Discharge, and RBL-Discharge) and weigh them by means of two parameters—Charging and Discharging Directive Parameters—handed to the SDB Runtime by the rest of the OS. Essentially, these parameters guide the SDB Runtime to weigh one of the algorithms more heavily at any moment in time. For example, a low value of the Charging Directive Parameter indicates that the user is in no hurry (e.g., charging at night), and that the Runtime should prioritize the use of the CCB-Charge algorithm. On the other hand, a high value of this parameter would lead the Runtime to prioritize the RBL-Charge algorithm in order to increase the useful charge (and thus the remaining battery lifetime) in the batteries as quickly as possible—say just before boarding an airplane. The discharge scenario is similar.

The CCB-Charge and CCB-Discharge algorithms are simple. These policies essentially enforce the controller to schedule the batteries (either for charging or discharging) in such a way that the resulting CCB is minimized, that is, is as close to 1 as possible. Our RBL-algorithms are more complex. Consider the discharge case, and let $y_1, \dots, y_N$ be the amount of current drawn from each of the batteries. The key underlying insight is that we can maximize the instantaneous RBL of the battery system by minimizing the total resistance losses across all the batteries. This can be achieved *if the resistances of the batteries are proportional to the square-root of their DCIR-to-SoC ratios*. Thus, the RBL-Discharge algorithm seeks to allocate the currents $y_1, \dots, y_N$ in such a way that the effective resistances of batteries are as much as possible proportional to the square-root of their DCIR-to-SoC rates minimizing the total energy wasted through resistive losses. Mathematically speaking, let δ_i be the instantaneous derivative of battery i 's DCIR curve, and let R_i be the current resistance. Then, the RBL-Discharge algorithm balances $R'_i = \sqrt{\delta_i / \lambda}$, where $R'_i = R_i + \delta_i y_i$ and λ is a

Lagrangian multiplier constant. Again, the case for charging (RBL-Charge) is similar. The SDB runtime calculates these power values at coarse granular time steps and updates the ratios based on the DCIR-SoC curves given by the manufacturer of the batteries.

A word of caution is necessary. The above RBL-algorithms are “optimal” only in an instantaneous sense. They minimize the instantaneous decrease of RBL (when discharging), or maximize the instantaneous increase of RBL (when charging). However, they are not globally optimal. Across the length of an entire workload, these algorithms might not actually maximize battery lifetime as we show in Section 5, that is, if we had knowledge of the future workload, we could improve upon the above instantaneously optimal algorithms by making temporarily suboptimal choices from which the system can profit later, for example, keeping a battery fully charged, if we know that this battery will be particularly helpful in the way of CCB or RBL for a future workload. For example, the overall cycle life or daily battery life may be improved when compared to using instantaneous mechanisms all the time.

Exploring these and other algorithmic nuances is interesting, but beyond the scope of this paper. We just note that the SDB resource optimization problem differs from traditional resource scheduling mechanisms, such as for Big.Little processors, hybrid storage, and SSD wear leveling, because of the resource in question—batteries. The main focus of traditional resource management algorithms is to multiplex a resource efficiently across a number of entities, such as users, processes, virtual machines, or erase blocks in case of SSDs over some fixed periods of time. The challenge of battery resource scheduling is threefold: daily battery life cannot be simply extended by minimizing instantaneous power losses; their long-term cycle life cannot be simply extended by balancing cycle life across batteries. Knowledge of impending workload can be used to improve the latter two metrics by picking strategies that may not be an instantaneous optimums as we demonstrate in Section 5. We hope that exposing the appropriate APIs will help system and algorithm designers to customize the scheduling algorithms for their battery configuration, and user workloads based on predicted as well as expected user behavior.

4. PROTOTYPE AND MICROBENCHMARKS

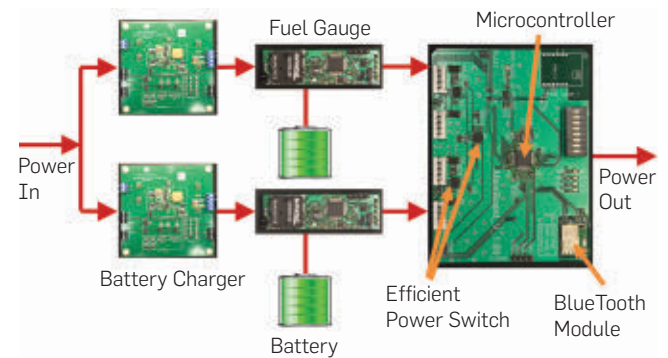
In this section, we describe the implementation of SDB and present microbenchmarks to evaluate it.

We built a hardware prototype of the SDB hardware architecture in Figure 4c. Figure 6 shows the components of our prototype.

We built a custom controller board with a ARM Cortex M3 microcontroller and a low-loss switching circuit. We also built a custom fuel gauge module that consists of a coulomb counter and a controller. We modified an off-the-shelf battery-charger evaluation board to enable dynamic charge current setting by the microcontroller on the control board. These hardware modules were interconnected as shown in Figure 6.

We used an ideal diode to switch between the batteries. The switching between batteries is extremely fast, and

Figure 6. SDB prototype implementation.



hence the battery sees a constant, smooth current draw. We note that the small power-loss due to this switch underestimates the efficiency achievable by the proposed solution. As mentioned in Section 3, the extra power-loss and the high component cost can be eliminated by augmenting existing switching regulators to switch across multiple batteries.

The boards were designed with Altium Designer,¹ a circuit board development package. The firmware was written in C using the IAR for ARM V7.40 tool chain. The board firmware contains ≈ 3500 lines of code. We also developed the prototype SDB Runtime shown in Figure 5 with 1200 lines of code.

We conducted simulations and microbenchmark experiments to evaluate the efficiency and accuracy of our hardware design, as well as to evaluate the correctness of the firmware and the runtime. Circuit simulations were done in LTSPICE,¹¹ a simulation program with integrated circuit emphasis (SPICE). We generated the circuits as shown in Figure 4, in LTSPICE, and conducted extensive simulations at various power loads to validate system correctness, stability, and responsiveness.

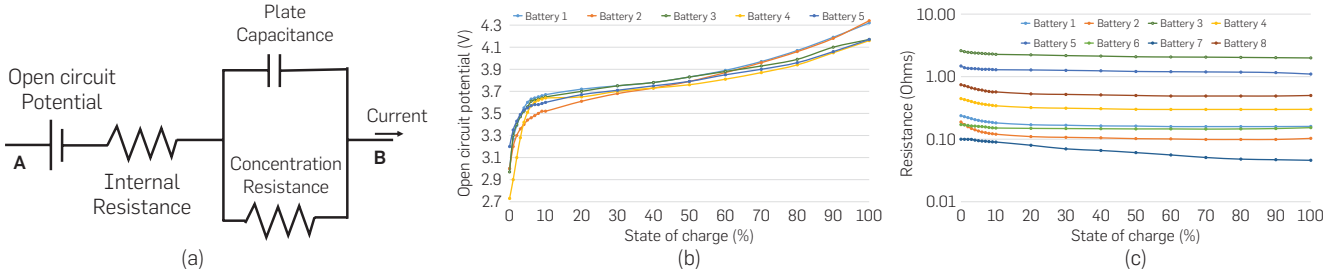
4.1. SDB emulator

We build a model for batteries based on Thevenin’s model as built by other battery researchers^{3–6,9} to simulate batteries used in production devices. The simplified Thevenin model is reproduced in Figure 7a. The model has four parameters: open circuit potential, internal resistance, concentration resistance, and plate capacitance.

The open circuit potential of a battery is the voltage across the terminals of the battery when no load is applied. It increases with the amount of energy left in a given battery. The internal resistance of a battery is the resistance across the terminals of the battery when a load is applied. It decreases with the amount of energy left in a given battery. In Figure 7b and c, we plot the open circuit potential and resistance, respectively, of a few batteries as the energy left in them increases.

The concentration resistance and the plate capacitance of a battery are fixed values for a given battery. We measure the open circuit potential, internal resistance, concentration resistance, and the plate capacitance for several kinds of batteries. We use the industry standard Arbin BT-2000²

Figure 7. Battery simulator: (a) Battery modeled with four variables that are learned using experimentation: open circuit potential, internal resistance, concentration resistance, and plate capacitance. This model allows us to conduct experiments in a scalable manner. (b) The open circuit potential of a battery increases with the state of charge (amount of energy left) of the battery. (c) The internal resistance of a battery decreases with the state of charge.



and Maccor 4200¹² battery cycling and testing hardware for measuring the battery properties.

The model takes the initial SoC, OCP versus SoC, resistance versus SoC, concentration resistance, and plate capacitance to emulate a battery. At each time step, based on the SoC, it estimates OCP, and resistance. Using the updated values, it calculates the values for the SoC after the time step.

We build the battery model using a few batteries and validate the models against other batteries of the same type. The validation results for one of the batteries are shown in Figure 8. The results show that our model is accurate to 97.5%. We modeled 15 batteries in total: two of Type 4, two of Type 3, eight of Type 2, and three more of other types (refer to Figure 1a).

We implement a simple software layer that takes the input power requirement and splits it across a given number of batteries according to the power policies set by an OS. The model and the SDB emulator are integrated into the OS using 4800 lines of code across modules written in C#.

We focus on two hardware platforms: a tablet and a watch. The tablet is a “2-in-1” development device with Intel Core i5 CPU, 4GB DRAM, 128GB SSD, and 12 inch display. The watch is a Qualcomm Snapdragon 200 development board with hardware similar to several smart-watches. Devices are instrumented to obtain fine grained (100 Hz) power-draw measurements. The power-draw is then fed into the emulator to calculate the energy drawn from the batteries.

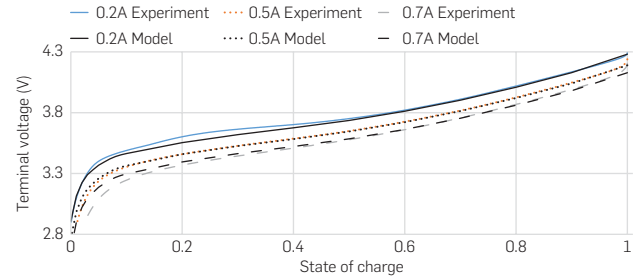
5. SDB APPLICATIONS

In this section, we describe two scenarios that benefit from using SDB with heterogeneous batteries. We also show how SDB policies can be customized for the different scenarios, and demonstrate the benefits of integrating future workload knowledge in the SDB system.

5.1. Adopting flexible batteries

Flexibility and bendability are important structural properties for wearable devices, for example, a watch-strap that is flexible and bendable tends to be easier to wear. Coincidentally, there are a few emerging battery chemistries that enable bendability. The bendability, unfortunately, comes at the cost of other battery properties. Such batteries use a solid (rubber-like) electrolyte in place of a traditional liquid (polymer) electrolyte. Unfortunately, the solid

Figure 8. Validating the model against battery testing hardware reveals that our model is 97.5% accurate.



(elastic) state of the electrolyte increases the resistance for the Li-ions and therefore, such batteries have higher internal losses. Several prototype bendable batteries we tested are excellent at handling low power workloads but often are very inefficient for high power workloads.

SDB can enable a scenario where a small traditional Li-ion battery in smart-watches is augmented with bendable batteries. This helps design better wearables that utilize the strap space to increase capacity but are still able to execute high power workloads like GPS tracking while running and cycling. The reduction in the size of the rigid Li-ion battery also allows for the design of a less bulky watch body.

The bendability of the battery in the strap is a boon, but its low efficiency is a bane that has to be intelligently managed to maximize effective battery life of the device. It is important to preserve energy in the efficient battery for times when the user is expected to perform power-intensive tasks. For example, the user may exercise, run or bicycle during certain times of the day, which all require high power. Therefore, the SDB policies should preserve the efficient battery for such times.

Since smart-watch usage will vary across users, we compare two extreme parameter values to demonstrate the benefits of SDB: One that minimizes instantaneous losses by drawing appropriate amounts of power from both the batteries and one that draws higher amounts of power from the inefficient battery to conserve the efficient battery.

Figure 9 demonstrates the setting and the results. We use a 200mAh Li-ion battery in combination with a

200mAh bendable battery for the setting. For a typical user who spends the entire day checking messages on his smart-watch and goes for a run in the evening, we plot the workload and the instantaneous losses in the batteries. We find that the latter method minimizes the total losses and therefore increases overall battery life by over an hour. These results provide evidence that mobile OSes that are aware of a user's day-to-day schedule may be able to provide better battery life by setting the right parameter. On the other hand, it is interesting to note that if the user had not gone for a run then the first policy would have given better battery life suggesting that the knowledge of an impending workload can help save energy in heterogeneous battery settings.

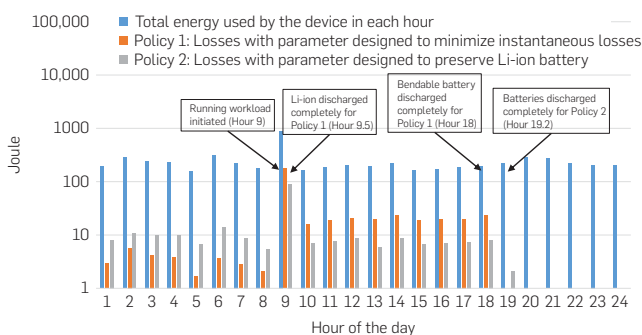
5.2. Battery management for 2-in-1s

2-in-1 devices are tablets that have a detachable keyboard. Some such devices have another battery under the keyboard. In such a setting, there are two batteries exposed to the OS, often with different capacities but the same internal chemistry—traditional Li-ion. However, efficiency of the battery in the base is less as it is used solely to charge the battery in the tablet. Significant amount of energy is lost in charging the internal battery with the external one, yet the reason why device manufacturers have chosen this route is to simplify design.

SDB via the OS can improve the battery life of a combined internal and external battery by understanding user behavior and expectations. The power drawn from an external battery can either be used toward running the system, for charging the main battery or both. For a user who rarely unplugs an external battery, the better solution would be to draw power simultaneously from both batteries as the internal losses are proportional to the square of the current (resistive losses = I^2R). Splitting the power draw across the two batteries, therefore, reduces the internal losses and increases the energy delivered to the system.

However, this strategy may not be ideal for a user who mostly operates in tablet-only mode. For such users, it makes more sense to draw as much power for as long as possible from the external battery to handle system load and also for charging the internal battery.

Figure 9. Fixed priority levels are bad. Priority levels have to be changed according to expected user schedules and workloads.



The OS sets a low parameter value for times when the external battery is expected to be plugged in for longer duration while high parameter values are for times when the external battery is plugged during battery crises.

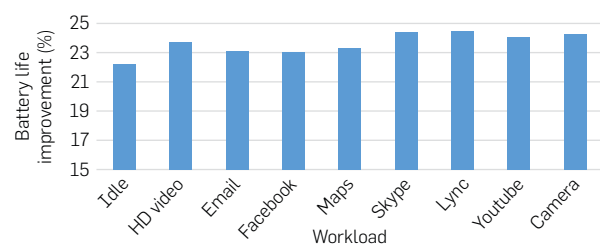
Figure 10 shows the comparison of two extreme parameters for various application workloads on a development 2-in-1 device with two equal sized traditional Li-ion batteries. Results show that the parameter causing simultaneous power draw from both batteries provides 22% more battery life than the parameter that causes one battery to charge another. However, this gain is not realizable for a user who only keeps the base with the secondary battery plugged in for short periods of time. The OS must, therefore, learn, predict and adapt to user behavior to set appropriate parameters.

6. CONCLUSION AND FUTURE WORK

Device requirements are typically hard to meet with a single battery since these requirements are often in conflict with each other. We present the SDB system that allows a device to use multiple heterogeneous batteries, and get the best of all of them. The SDB hardware is designed to be low cost, and provides rich functionality to the OS. The SDB APIs allow an OS to dynamically route charge to, and from the batteries based on application workload such that the overall goals (battery life, cycle count, fast charge, etc.) are met. We show several new scenarios that can be enabled with SDB, and demonstrate its feasibility using a prototype, and detailed emulations.

Moving forward, we are taking the SDB work in two main directions. First, we are tying personal assistants like Siri, Cortana, and Google Now understand user behavior and the user's schedule and by using this information, an OS can perform better parameter selection. For example, if the user's profile suggests that the user plays video games in the evening, then it SDB could preserve a higher power-density battery for that workload. Second, we are working on additional devices that would benefit from this technology, such as drones, smart glasses, and electric vehicles (EVs). Each would require a different combination of battery chemistries, and the SDB logic might be different too. For example, we are building on the techniques proposed for Hybrid Energy Storage in the Grid^{7, 10} or hybrid power sources in Data Centers,⁸ to improve the lifetime of EVs. An EV's NAV system could provide the vehicle's route as a hint to the

Figure 10. Drawing power simultaneously from internal and external batteries is more energy efficient than depleting the external battery for conserving and charging the internal one.



SDB Runtime, which could then decide the appropriate batteries based on traffic, hills, temperature, and other factors. Our preliminary analysis shows that SDB might help these systems achieve tradeoffs that until now were considered to be at odds with each other.

Acknowledgments

Insightful comments from Jon Crowcroft, Srinivasan Keshav, Matthew Lentz, and Vasuki Narasimha Swamy greatly improved the final version of this paper. 

References

- Altium Designer. <http://www.altium.com/altium-designer/overview>.
- Arbin BT-2000 Battery Testing Equipment. <http://www.arbin.com/products/battery>.
- Chen, M., Rincon-Mora, G.A. Accurate electrical battery model capable of predicting runtime and IV performance. *IEEE Trans. Energy Convers.* 21, 2 (2006), 504–511.
- Chiasserini, C.-F., Rao, R.R. Energy efficient battery management. *IEEE J. Sel. Areas Commun.* 19, 7 (2001), 1235–1245.
- Erdinc, O., Vural, B., Uzunoglu, M. A dynamic lithium-ion battery model considering the effects of temperature and capacity fading. In *Proceedings of the IEEE International Conference on Clean Electrical Power* (Capri, Italy, June 2009).
- Gao, L., Liu, S., Dougal, R.A. Dynamic lithium-ion battery model for system simulation. *IEEE Trans. Compon. Pack. Technol.* 25, 3 (2002), 495–505.
- Ghiassi-Farrokhfal, Y., Rosenberg, C., Keshav, S., Adjaho, M.-B. Joint optimal design and operation of hybrid energy storage systems. *IEEE J. Select. Areas Commun.* 34, 3 (Nov. 2016), 639–650.
- Govindan, S., Sivasubramaniam, A., Urgaonkar, B. Benefits and limitations of tapping into stored energy for datacenters. In *International Symposium on Computer Architecture* (2011).
- He, H., Xiong, R., Zhang, X., Sun, F., Fan, J. State-of-charge estimation of the lithium-ion battery using and adaptive extended Kalman filter based on an improved Thevenin model. *IEEE Trans. Veh. Technol.* 60, 4 (2011), 1461–1469.
- Kim, Y., Chang, N. *Hybrid Electrical Energy Storage Systems Design*. Springer International Publishing, Cham, Switzerland, 2014, 19–25.
- LTSpice: Linear Technologies Simulator Program with Integrated Circuit Emphasis.
- Maccor 4200 Battery Testing Equipment. <http://www.maccor.com/Products/Model4200.aspx>.

Anirudh Badam, Ranveer Chandra, Jon Dutra, Julia Meinershagen, and Bodhi Priyantha ([anbadam, ranveer, jodutra, juliam, bodhip]@microsoft.com), Microsoft, Redmond, WA.

Anthony Ferrese (anthony.ferrese@gmail.com), Tesla Motors, Palo Alto, CA.

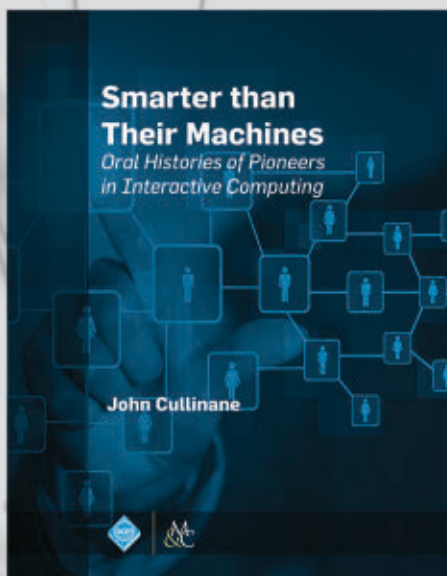
Steve Hodges (shodges@microsoft.com), Microsoft, Cambridge, U.K.

Pan Hu (lghupan@gmail.com), University Massachusetts Amherst, MA.

Thomas Moscibroda (moscitho@microsoft.com), Microsoft, Beijing, China.

Evangelia Skiani (valia@ee.columbia.edu), Department of Electrical Engineering, Columbia University, New York, NY.

© 2016 ACM 0001-0782/16/12 \$15.00



A personal walk down the computer industry road. BY AN EYEWITNESS.

Smarter Than Their Machines: Oral Histories of the Pioneers of Interactive Computing is based on oral histories archived at the Charles Babbage Institute, University of Minnesota. These oral histories contain important messages for our leaders of today, at all levels, including that government, industry, and academia can accomplish great things when working together in an effective way.



ISBN: 978-1-62705-550-5 DOI: 110.1145/2663015
<http://books.acm.org>
<http://www.morganclaypoolpublishers.com/acm>

Technical Perspective

3D Image Editing Made Easy

By Takeo Igarashi

NUMEROUS IMAGES APPEAR each day on smartphone screens, computer displays, and printed materials. News articles and advertisements attract the attention of viewers by using appealing images. People often take photographs with their smartphones and immediately share them with hundreds of viewers via social media. Moreover, users include images in documents and presentations to communicate messages. These images are typically edited to appear more aesthetically pleasing or achieve other objectives. Thus, there is an increasing need for advanced image editing tools.

Image editing has significantly advanced over the past several years to address consumer demand. Color adjustment tools now more readily facilitate simpler image sharpening, softening, brightening, and darkening. Furthermore, advanced tools can automatically remove blur and noise. Users can easily cut an object in an image and paste it onto another background. Moreover, it is easy to readily deform shapes as needed and seamlessly blend two images. It is even possible to erase objects in a scene by removing the object and filling the hole by automatically synthesizing appropriate background images.

Despite these advancements, most image editing techniques are two-dimensional (2D). Three-dimensional (3D) image editing has been strongly desired; however, it is still difficult. Even the most advanced current image editing tools lack 3D editing capabilities. This is because it is necessary for a computer to infer the 3D structure of the scene for 3D editing, which remains a difficult, ill-posed problem. Inference of a meaningful 3D structure requires abundant knowledge about the physical world, which remains missing in current systems. Moreover, 3D editing is difficult for users because the user must provide 3D control information to a computer using 2D input

devices, such as a touchpad or mouse. It is tedious and difficult for inexperienced users to specify the 3D shapes of an object in a scene and thereby manipulate the object.

Nonetheless, the importance of 3D editing is obvious. We exist in a 3D world, and 3D image editing opens countless possibilities. With the availability of 3D information, we can easily view objects from different angles and compose novel scenes by three dimensionally combining 3D objects. Even basic cutting and pasting of an object is not easy with purely 2D editing tools because the viewing angle changes if the object is moved. A 3D editing capability would make the cut-and-paste result much more convincing.

The authors of the following paper present an important step toward achieving 3D editing. To address this difficult problem, human perception and computational analysis are both required. Therefore, the authors devised an interaction technique, called 3-sweep, which is comprised of three simple mouse strokes. With this interaction, the user provides guidance for the computer to segment an object in the scene and simultaneously infer the 3D geometry of the object. The system then executes segmentation and 3D reconstruction by inferring details using image analysis methods. Using the 3D reconstruction results, the user can rotate the object to view it from multiple perspectives. In addition, the user can cut and paste the object into different scenes while preserving 3D consistency. Interested readers are strongly encouraged to watch the authors' impressive demonstration video (<https://www.youtube.com/watch?v=Oie1ZXWceqM>).

As with most technologies, this 3D editing technology is not the only one of its kind; previous efforts exist. A notable one is the photo editing tool presented by Zheng et al.² They similarly combine intuitive user interaction and computa-

tional analysis to facilitate 3D manipulation of objects in a photograph. However, their tool is specifically designed for cuboids and only works for objects made of boxes, rectangular plates, and square pillars. The present technology, on the other hand, can handle a significantly larger variety of objects with curved surfaces by introducing a sophisticated gestural interaction.

It should be noted, however, that the present technology currently supports only one class of primitive: generalized cylinders. This representation is highly versatile and covers many human-made objects; nonetheless, it is not sufficient for representing complicated shapes perceived in nature. Tools are needed to create a more diverse set of shape primitives to represent complicated shapes, which would enable manipulation of arbitrary objects viewed in photographs. One promising approach is to use a large collection of known 3D geometries.¹ Nevertheless, the core concept presented in this paper—combining human perception and computational analysis using clever interactive design—is broad and applicable to the development of future tools. Rapid developments in 3D image editing tools inspired by this work can be foreseen, and it will become increasingly easier for even casual users to readily edit photographs. Images will no longer always depict unaltered reality. Is this good or bad? The answer may not be self-evident; however, such a future is surely coming. 

References

1. Kholgade, N., Simon, T., Efros, A., and Sheikh, Y. 3D object manipulation in a single photograph using stock 3D models. *ACM Trans. Graph.* 33, 4 (July 2014), Article 127.
2. Zheng, Y., Chen, X., Cheng, M.M., Zhou, K., Hu, S.M., Mitra, N.J. Interactive images: Cuboid proxies for smart image manipulation. *ACM Trans. Graph.* 31, 4 (2012), 99:1–99:11.

Takeo Igarashi is a professor in the Computer Science Department at the University of Tokyo, Japan.

Copyright held by author.

Extracting 3D Objects from Photographs Using 3-Sweep

By Tao Chen, Zhe Zhu, Shi-Min Hu, Daniel Cohen-Or, and Ariel Shamir

Abstract

We introduce an interactive technique to extract and manipulate simple 3D shapes in a single photograph. Such extraction requires an understanding of the shape's components, their projections, and their relationships. These cognitive tasks are simple for humans, but particularly difficult for automatic algorithms. Thus, our approach combines the cognitive abilities of humans with the computational accuracy of the machine to create a simple modeling tool. In our interface, the human draws three strokes over the photograph to generate a 3D component that snaps to the outline of the shape. Each stroke defines one dimension of the component. Such human assistance implicitly segments a complex object into its components, and positions them in space. The computer reshapes the component to fit the image of the object in the photograph as well as to satisfy various inferred geometric constraints between components imposed by a global 3D structure. We show that this intelligent interactive modeling tool provides the means to create editable 3D parts quickly. Once the 3D object has been extracted, it can be quickly edited and placed back into photos or 3D scenes, permitting object-driven photo editing tasks which are impossible to perform in image-space.

1. INTRODUCTION

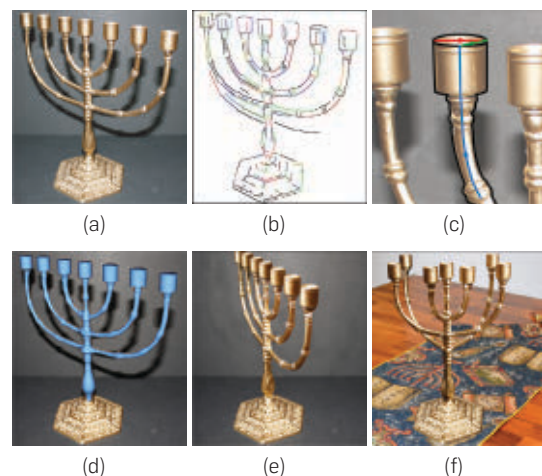
Extracting three dimensional objects from a single photo is still a long way from reality given the current state of technology, since it involves numerous complex tasks: the target object must be separated from its background, and its 3D pose, shape, and structure should be recognized from its projection. These tasks are difficult, even ill-posed, since they require some degree of semantic understanding of the object. To alleviate this difficulty, complex 3D models can be partitioned into simpler parts that can be extracted from the photo. However, assembling parts into an object also requires further semantic understanding and is difficult to perform automatically. Moreover, having decomposed a 3D shape into parts, the relationships between these parts should also be understood and maintained in the final composition.

In this paper, we present an interactive technique to extract 3D man-made objects from a single photograph, leveraging the strengths of both humans and computers. Human perceptual abilities are used to partition, recognize, and position shape parts, using a very simple interface based on triplets of strokes, while the computer performs tasks which are computationally intensive or require accuracy. The final object model produced by our method includes its geometry and structure, as well as some of its semantics. This allows the extracted model to be readily available for

intelligent editing, which maintains the shape's semantics (see Figure 1).

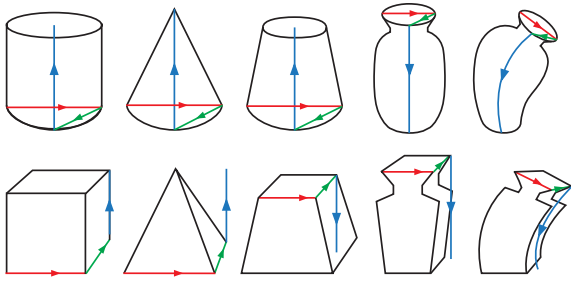
Our approach is based on the observation that many man-made objects can be decomposed into simpler parts that can be represented by generalized cylinders, cuboids, or similar primitives. A generalized cylinder is a cylindrical primitive shape where the central axis is a curve instead of a line, the shape's profile can be any 2D closed curve and not just a circle, and this shape can also change along the curve. In this work, we use just circular and cuboid profiles. The key contribution of our method is an interactive tool that guides and assists the user in the creation of a 3D editable object by defining its primitive parts. The tool is based on a rather simple modeling gesture we call *3-Sweep*. This gesture allows the user to explicitly define the three dimensions of a geometric primitive using three sweeps. The first two sweeps define the first and second dimension of a 2D profile and the third, usually longer, sweep is used to define the main curved axis of the primitive (see Figure 2).

Figure 1. 3-Sweep Object Extraction. (a) Input image. (b) Extracted edges. (c) 3-Sweep modeling of one component of the object. (d) The full extracted 3D model. (e) Changing the object viewpoint. (f) Editing the model by rotating each arm in a different direction, and pasting onto a new background. The base of the object is transferred by alpha matting and compositing.



The original version of this paper is entitled “3-Sweep: Extracting Editable Objects from a Single Photo” and was published in *ACM Transactions on Graphics*, Volume 32, Issue 6—*Proceedings of ACM SIGGRAPH Asia 2013*, November 2013 Article No. 195.

Figure 2. The 3-Sweep paradigm is used to define general cylinder and cuboid parts.



As the user sweeps the primitive, the program dynamically adjusts the progressive profile by sensing the pictorial context in the photograph and automatically snapping to it. Furthermore, relationships between the various primitive parts are automatically recognized and preserved by the program. Using several such 3-Sweep operations, the user can model 3D parts consistent with the object in the photograph, while the computer automatically maintains global constraints linking them to other primitives comprising the object. Using 3-Sweep technology, non-professionals can extract 3D objects from photographs. These objects can then be used to build a new 3D scene, or to alter the original image by changing the objects or their parts in 3D and pasting them back into the photo.

2. BACKGROUND

Modeling from images. Images have always been an important resource for 3D modeling. Many techniques use multiple images or videos to model 3D shapes and scenes.²⁰ However, our focus is on modeling objects from a single photograph. This task is challenging because of the inherent ambiguity of the mapping from a 3-dimensional world to a 2-dimensional image. To overcome this, methods use both constraints on the type of images that can be used—such as non-oblique views, and prior assumption on the possible geometry that can be extracted.

Fully automatic methods use assumption such as symmetry,²³ smoothness,¹⁸ or the existence of a prior 3D model similar to the one in the photograph.²² Some limit the geometry to planes or smooth surfaces, while others limit the application range (e.g., to architectural models^{3,10}). There are methods that automatically fit edges or regions in an image or sketch using optimization.¹⁵ These are basically 2D methods, generating either 2D models or 3D models by extrusion. Our method can directly generate complex oblique shapes in 3D space such as the menorah in Figure 1 and the telescope in Figure 7. Automatic methods would fail on such examples since their prior assumptions are not met, or because they rely on region color or clear edges, which can be missing or occluded.

Other methods use a human-in-the-loop for modeling,¹⁶ but usually require extensive manual interaction, and are more time consuming compared to the 3-Sweep approach. They either require a complete sketch of the object or tedious labeling before the machine takes control of an optimization process, while in 3-Sweep user guidance and

automatic snapping are interlaced.

Sketch-based modeling. The task of 3D modeling from a single image is closely related to the reconstruction or definition of a 3D shape from a sketch.¹⁷ The user either directly draws the curves of the object¹ or fits parts or primitives to a predefined sketch.¹² Our work was inspired by a tool for modeling simple 3D objects from sketches presented by Shtof et al.¹⁹ In that work geo-semantic constraints were used between primitive parts to define their semantic and geometric relationships, and connect them to form the final object. However, their approach is geared towards sketches and uses a drag and drop interface for primitives that need to fit the sketch contours.

Computer aided design. The use of constraints in computer-aided design has been studied extensively and allows the definition of semantic information that relates different geometric parts in an object. Automatically inferring constraints from the object or its parts' geometry has been used for reverse engineering⁵ and object deformation and editing.^{11,21} Similarly, sweep-based models have been defined and used since the beginning of the field.⁹ While we cannot report all computer-aided design work aiming at modeling sweep-based primitives, to our knowledge, none of these methods have been applied to modeling from photographs, nor they have paired sweeping with snapping to image edges.

Object-based image editing. Apart from modeling a 3D object, 3-Sweep allows the application of object-based image editing operations, that previously required extensive user interaction⁸ or massive data collection.^{13,14}

3. OVERVIEW

Our interactive modeling approach takes as input a single photo such as the one in Figure 1a. Our goal is to extract a 3D model whose projection exactly matches the object in the image. Using the *3-Sweep* modeling technique, the user constructs the whole object gradually from its parts. This means that the object is implicitly decomposed into simple parts, which are typically semantically meaningful. We define two types of primitives: both use a piecewise linear centerline to sweep a cross-section which is assumed to be perpendicular to the centerline at every location. One type of primitive uses a circular cross-section that can vary in radius along the sweep, and the other uses a rectangle cross-section that can change its aspect ratio along the sweep.

Such decomposition is both easy and intuitive for users, but provides the computer significant information for reconstructing a coherent 3D man-made object from its parts' projections. The parts are expected to have typical geometric relationships that can be exploited to guide the composition of the whole object. Although the user interacts with the given photo, she does not need to exactly fit the parts to the photo or connect them to each other. 3-Sweep automatically snaps primitive parts to object outlines created from edges, and connects them to previously defined 3D parts.

To create a single 3D primitive part on the given photo, the designer uses three strokes. The first two strokes define a 2D profile of the part and the third stroke defines its main axis, which is either straight or curved (see Figure 2). Defining the profile and sweeping the axis are simple operations since

they do not demand accuracy. The profile dimensions are guided by the object's outlines. While sweeping, the 3D extent of the part is also defined by snapping to these outlines. To compensate for perspective distortion, during this process, the camera's angle of view is estimated. Therefore, the part need only be sketched quickly and casually by the user. Figure 1c shows the result of sweeping along the tubes of the menorah (more examples could be seen in an online video at <https://vimeo.com/148236679>). We elaborate on the 3-Sweep operation in Section 4.

As more model parts are added, geometric relationships between them serve (i) to assist in disambiguating and defining the depth dimension and (ii) to optimize the positioning of the parts. These geometric relationships include parallelism, orthogonality, collinearity, and coplanarity. We use optimization to satisfy these geo-semantic constraints, while taking into account the snapping of the 3D geometry to the object's outlines and the user's sweeping input. A complete model with geo-semantic relationships is shown in Figure 1d. These geo-semantic relationships not only help define the 3D model, but become part of the 3D representation of the model, allowing smart editing of the 3D model later, as demonstrated in Figure 1f and in other figures in the paper.

Our interface also supports several operations for more effective modeling. For example, the user can copy and paste similar parts to other positions on the image. Although many geo-semantic constraints are automatically inferred, the user can also manually specify constraints between selected parts, constrain the parts to have uniform or linearly changing radii etc.

4. SINGLE PRIMITIVE FITTING

In this section, we first describe the 3-Sweep technique for creating one generalized cylinder. Simpler primitives such as spheroids or simple cubes are also supported by direct modeling in our system.

Preprocessing. In a preprocessing stage, we extract image edges and build candidate object outlines. We adopt a method for hierarchical edge feature extraction based on spectral clustering.² We then apply a technique to link the detected edge pixels into continuous point sequences,⁷ each shown in a different color in Figure 1b. An edge orientation computed over a 5×5 neighborhood is associated with each

edge pixel.

Profile definition. In the first stage, the user draws the 2D profile of the generalized cylinder, usually at one end of the shape. This is illustrated in Figure 3, where black curves are outlines detected in the input image. The task is to draw a 2D profile correctly oriented in 3D. This can be regarded as positioning a disk in 3D by drawing its projection in 2D. To simplify this task, we assume that the disk is a circle, thus reducing the number of unknown parameters. Later, the circular disk can be warped into an elliptical disk based on the 3D reconstruction. The drawing of a circular disk is accomplished by drawing two straight lines S_1S_2 and S_2S_3 over the image, as shown in Figure 3a (red and green arrows). The first line defines the major diameter of the disk, and the second line is then dragged to the end of the minor diameter. This forms an ellipse in image space that matches the projection of a circular disk: see the dark blue circle in Figure 3a. The depth of the disk is set to 0. The normal direction and radius of the disk are assigned according to the length and orientation of the two diameters of the elliptical projection.

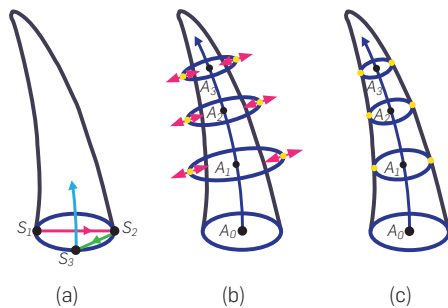
Generalized cuboids are modeled in a similar way. The two strokes that define the profile of a cuboid follow the two edges of the base of the cuboid instead of the diameters of the disk, as shown by the red and green lines in the bottom row of Figure 2.

Sweeping. After completing the base profile, the user sweeps it along a curve that approximates the main axis of the 3D part. In general, this curve should be perpendicular to the profile of the 3D primitive (see blue arrow in Figure 3a). As the curve is drawn, copies of the profile are placed along the curve, and each of them is snapped to the object's outline.

While sweeping, the axis curve is sampled in 2D image space at a uniform spacing of five pixels to produce 3D sample points $A_0, \dots, A_N$, that lie on one plane. At each sampled point A_i , a copy of the profile is created, centered around the curve. Its normal is aligned with the orientation of the curve at A_i , and its diameter is adjusted so that its projection on the image will fit the object's outline. Together, the adjusted copies of the profile in 3D form a discrete set of slices along the generalized cylinder, as shown in Figure 3c.

At each point A_i , we first copy the profile from A_{i-1} and translate it to A_i . We then rotate it to take into account the bending of the curve. We then consider the two ends of the major axis (yellow points in Figure 3b) of the profile, denoted by p_i^0, p_i^1 . For each contour point $p_i^j, j \in [0, 1]$ we cast a 2D ray from point A_i along the major axis, seeking an intersection with the object outline. Finding the correct intersection is somewhat challenging as the image may contain many edges in the vicinity of the new profile. The closest edge is not necessarily the correct one, for example, when hitting occluding edges. In other cases, the correct outline may be missing altogether. To deal with these cases, we first limit the search for an intersection to a fixed range, which limits the major axis of adjacent profiles not to vary by more than 20% in length. Secondly, we search for an intersecting edge that is not collinear to the ray (creates an angle larger than $\pi/4$). Although this method cannot guarantee that it will find the correct intersections, the

Figure 3. Modeling a primitive by defining a 2D profile and sweeping it along the main axis of the object (a). The profile is copied along the centerline (b), and the copies are snapped to the image edges (c).



subsequent profile propagation step can tolerate a limited number of missing intersections.

When an intersection point is found, we snap the contour point p_i^j to it. If both contour points of the profile are snapped, we adjust the location of A_i to lie at their midpoint. If only one side is successfully snapped, we mirror the length of this side to the other side and move the other contour point respectively. Lastly, if neither contour points is snapped, the size of the previous profile is retained.

Post-processing. The above modeling steps closely follow user gestures, especially when modeling the profile. This provides more intelligent understanding of the shape but it is less accurate. Therefore, after modeling each primitive, we apply a post-snapping stage to better fit the primitive to the image as well as to correct the view. We search for small transformations ($\pm 10\%$ of primitive size) and changes of vertical angle of view ($\pm 10^\circ$) that create a better fit of the primitive's projection to the edge curves it was snapped to in the editing process.

In many cases, the modeled object type has special properties that can be used as priors to constrain the modeling. For example, if we know that a given part has a straight spine, we can constrain the sweep to progress along a straight line. Similarly, we can constrain the sweep to preserve a constant or linearly changing profile radius. In this case, the detected radii are averaged or fitted to a line along the sweep. We can also constrain the profile to be a square or a circle. In fact, a single primitive can contain segments with different constraints: it can start with a straight axis and then bend, or use a constant radius only in a specific part. Such constraints are extremely helpful when the edge detection provides poor results.

To further assist in modeling interaction, we also provide a copy and paste tool. The user can drag a selected part that is already snapped over to a new location in the image and snap it again in the new position. While copying, the user can rotate, scale, or flip the part.

5. COMPOSITE OBJECT CONSTRUCTION

The technique described above generates parts that fit the object outlines. The positions of these parts in 3D are still ambiguous and inaccurate. However, the assumption is that these parts are components of a coherent man-made object, and semantic geometric relationships exist among them. Constraining the shape to satisfy such relationships allows creation of meaningful models.

Since each component has many degrees of freedom, direct global optimization of the positions of parts while considering their geo-semantic relationships is computationally intensive and vulnerable to trapping in local minima. In our setting, the modeled components are also constrained to agree with the outlines of the object in the image. These constraints can significantly reduce the degrees of freedom for each part, reducing the dimensionality of the optimization space and avoiding local minima. In the following discussion, we describe how we simplify the general positioning problem and ensure that geo-semantic constraints are satisfied among the 3-swept parts.

The key idea is that if the projection of a part is fixed, its position and orientation can be determined by only one or two depth values. We first describe the method for simple parts that can be modeled by a single parameter, namely parts which were modeled using a *straight* axis. General cylinders and cuboids with curved axes will be later approximated using two arbitrarily connected straight axis primitives at the start and end of the whole part.

Determining straight shapes. The position and orientation of a generalized cylinder i with a straight-axis can be determined by two points we call *anchors*, $C_{i,1}$ and $C_{i,2}$, on its main axis (see Figure 4). Similarly, a cuboid part can be represented by six anchors $C_{i,j}, j \in [1, 6]$ positioned at the center of each face. Every opposite pair of anchors defines one main axis of the cuboid. Even though four anchors are enough to fix the position and orientation of a cuboid, we use six to simplify attaching various geo-semantic constraints from other parts to each side of the cuboid.

We define a local 3D orthogonal coordinate system for each part using the three strokes defined by the user for the three dimensions of the part. First, we define the origin of the coordinate system of part i at a reference point R_i on the part's projection. For a cuboid part, we pick the point connecting the first and second user strokes, and for a cylinder we pick the point connecting the second and third strokes. Due to the internal orthogonality of the straight part, the profile of the part is perpendicular to the main axis. Therefore, we can use the endpoints of the user's strokes (after snapping them to the image edges) to define three points that together with R_i create an orthogonal system (red points and lines in Figure 5). Note that this coordinate system is defined

Figure 4. Three examples where we infer geo-semantic constraints from primitives where such relationships “almost” hold: Collinear axes (left), Parallel axes (top right), and Perpendicular axes (bottom right).

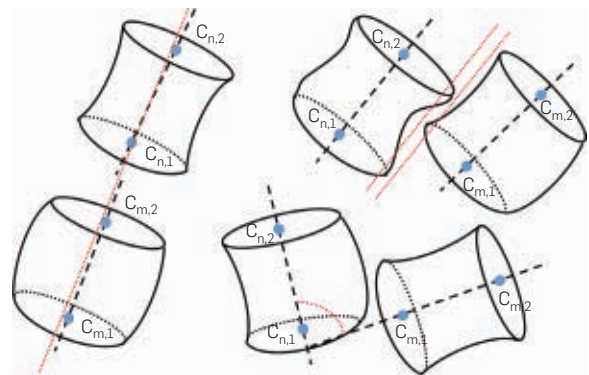
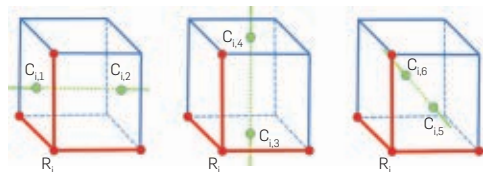


Figure 5. Determining coordinates $C_{i,j}$ for axis endpoints of a cuboid from the depth value z_i of the reference point R_i .



in camera coordinates. The x and y values of the end points are determined by the projection and their depth values can be found as a function of z_i , the z value of $\mathbf{R}_i$, by using three orthogonality constraint equations.

Next, the positions of the anchor points $\mathbf{C}_{i,j}$ in world coordinates are defined using the local orthogonal axes, giving the structure of part i . Since the local axes depend only on the depth value z_i of the point $\mathbf{R}_i$, we can parameterize the positions of $\mathbf{C}_{i,j}$ as a function of z_i : $\mathbf{C}_{i,j} = \mathbf{F}_{i,j}(z_i)$: the position and orientation of the whole part become a function of a single unknown z_i . $\mathbf{F}_{i,j}$ has the form $F_{i,j}(z_i) = b/(a(z_i + v))$ for each coordinate component, where a depends only on the x and y coordinates of the endpoints of the local axes, and b , v are decided by perspective parameters. They are different for each axis endpoint and for each coordinate component (see Ref.⁶ for the full definition).

Defining geo-semantic constraints. We use the anchor points to define the geo-semantic relationships between parts. Specifically, we support six types of constraints: parallelism, orthogonality, and collinearity of primitive axes, overlapping endpoints of axes, coplanar axis endpoints, and coplanar axes. During modeling, for each type, we test whether a pair of components is close to satisfying one of the above geo-semantic constraints. If so, we add this constraint to the definition of the object (see Figure 4). For example, for two cylinders with indices m and n , if the angle between vector $(\mathbf{C}_{m,1} - \mathbf{C}_{m,2})$ and $(\mathbf{C}_{n,1} - \mathbf{C}_{n,2})$ is less than 15° , we add a parallelism constraint $(\mathbf{C}_{m,1} - \mathbf{C}_{m,2}) \times (\mathbf{C}_{n,1} - \mathbf{C}_{n,2}) = 0$ to the system of constraints. Similarly if any three of the four anchors of two cylinders form a triangle containing an angle larger than 170° , we add a collinear axes constraint: $(\mathbf{C}_1 - \mathbf{C}_2) \times (\mathbf{C}_1 - \mathbf{C}_3) = 0$. Internal constraints such as orthogonality and concentricity of a cuboid's axes are also added to the definition, and some editing operations such as copying and pasting parts (see Section 6) can impose same-size constraints. Finally, we also allow the user to manually enforce or revoke any constraint for selected primitive parts.

Establishing an objective function. Suppose we have found p geo-semantic constraints G_k for a set of n components. Together with the objective function for fitting the image outline, we define the following optimization system:

$$\text{minimize } E = \sum_{i=1}^n w_i \left(\sum_{j=1}^{m_i} \|\mathbf{C}_{i,j} - \mathbf{F}_{i,j}(z_i)\|^2 \right), \quad (1)$$

$$\text{subject to } G_k(\mathbf{C}_{1,1}, \dots, \mathbf{C}_{n,m_n}), \quad k=1, \dots, p, \quad (2)$$

where m_i is the number of axes of the i th primitive part. We add weights w_i proportional to the radius of the base profile of each part and the length of its axis. Larger parts have more impact on the solution since typically larger parts are modeled more accurately. Intuitively, the first equation tries to fit projection of the part's geometry $(\mathbf{C}_{i,j})$ to the image outline and the user's gestures, while the second set of equations imposes the geo-semantic constraints.

Two-step solution. Solving for $\mathbf{C}_{i,j}$ and z_i together is a nonlinear non-convex optimization problem with nonlinear constraints. Directly solving such a system without becoming trapped in a local minimum is very difficult. Hence, we decompose the solution into a two step procedure. The first step tries

to find a good initial position for all parts at once, by changing only their depths (governed by z_i) to meet the geo-semantic constraints. In the second step, the full system is solved, allowing the shapes of the parts $(\mathbf{C}_{i,j})$ to change as well.

In the first step, we modify the soft constraint in Equation (1) to a hard one, and replace $\mathbf{C}_{i,j}$ by $\mathbf{F}_{i,j}(z_i)$ in all equations. This means that Equation (1) is trivially true and we are left with just the constraints in Equation (2). In effect, this means we fix the projection and find the optimal z_i meeting the geo-semantic constraints. This reduces the number of variables to n ($z_i, 1 \leq i \leq n$) and changes Equation (2) into a potentially over-determined system, where each equation only contains two different variables. We find the least squares solution $\bar{z}_i$ by using the conjugate gradient method, with all z_i values initialized to 0.

This first step provides a good initialization to find the optimal solution for $\mathbf{C}_{i,j}$, which should be close to $\mathbf{F}_{i,j}(\bar{z}_i)$, requiring only small inconsistencies to be fixed with the geo-semantic constraints. Hence, in the second step, we carry out full optimization of Equation (1) with the set of constraints in Equation (2) by an augmented Lagrangian method. Both steps are fast, and we are able to avoid local minima by the initialization of the first step. This also permits optimization to be carried out at interactive speed (see examples at <https://vimeo.com/148236679>). Note that the nonlinearity of $\mathbf{F}_{i,j}()$ arises due to the assumption of a perspective projection. However, we can linearly approximate this projection since we assume the change in z_i is small. This further increases the speed and stability of our solution.

Curved shapes. To handle parts with a non-straight axis, we first simplify the problem by assuming that the general axis lies in a plane. We define a non-straight part as a blend of two straight-axis sub-parts, placed at the two ends of the part. The position of each of these sub-parts is determined by a single depth value in the optimization above. The blend between these parts for the whole part is defined by connecting the two subparts with a piecewise linear curve where the axis points are determined when constraining the profile while snapping it during the user's sweep. More details can be found in Ref.⁶

Texturing. Once the object has been modeled, we can map the texture from the image onto the object, as shown in Figure 6. By projecting a vertex of the mesh to the image plane, we can get the 2D coordinates of the vertex in the image, which are then used as texture coordinates to map the corresponding part of the image onto the model. Since there is no information regarding the back of the object, we simply use a symmetry assumption and mirror the front texture to the back. In each half of the model (front and back), we assign the same texture coordinate for the two vertices that are mirrored symmetrically about the central plane. On the two sides of the object (left and right), there are vertices whose normal is perpendicular, or almost perpendicular, to the image plane. To deal with such vertices, we treat the texture associated with them as holes, and use the image completion technique⁴ to fill them.

6. EXPERIMENTAL RESULTS

We implemented the 3-Sweep interactive technique in C++ inside a simple modeling system. The system provides an

Figure 6. Modeling different objects: a table (a), a lamp (b), a monument (c), a samovar (d) and a menorah (e). Top: input photos. Middle: extracted 3D models (blue) are rotated and repositioned. Bottom: modified objects inserted into the same or a new environment, with their textures.



outline view for 3-Sweep interaction, a solid model view, and a texture view for checking the model and for image editing. The user can choose between cuboid, cylinder, and sphere primitives using a button or key shortcut. The system also provides conventional menu selection, view control and deformation tools. The 3-Sweep technique has been tested and evaluated on a large number of photos as we demonstrate in this section and in the online video (see <https://vimeo.com/148236679>). As shown in the video, most of the examples were modeled in a few minutes or less. The modeling process is intuitive and fluent, and can be used by unskilled people following very little training. Editing and repositioning an object after modeling requires an effort similar to using other parametric editing techniques.

6.1. Modeling from single image and editing

In the following examples, we show how the acquired 3D textured-model allows semantic image editing. Before editing, the image of the 3D model is cut out from the photo, leaving a black hole which is filled using a standard image completion technique.⁴

Figure 1e demonstrates the modeling of a menorah, and then rotating its arms to a different angles preserving the inter-part relationships of the object. Note that all candle holders have the same size, but due to the oblique view, their

sizes appear different in the photo. During modeling, we modeled one candle holder and copied it to fit all other holders in the image, while requiring that they all lie on the same plane and that their 3D sizes be the same. This efficiently recovered the true 3D position and shape of each part.

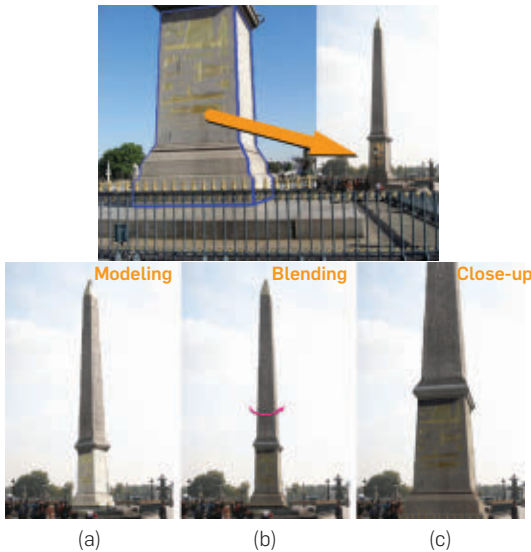
Figure 6 shows several modeling results. In the top row we show the input photos, in the middle row we show the extracted and repositioned 3D models, and in the third row, they are inserted with their textures into the same or a new environment. The rightmost column shows the modeling and repositioning of three objects in one complex photo. Note that the menorah has been rotated, and translated on the ground plane.

Figure 7 shows three examples of modeling and editing at the part-level, where some parts of the objects (highlighted in gold) are replicated, copied, and optionally rotated to enhance and enrich the shape. At top left is a tap, whose handle is augmented to be four-sided, and also rotated. The whole tap is also copied and attached to the other side of the wall. The top right shows a street lamp with duplicated lamps moved to a lower position and rotated. The whole lamp pole is also copied to other positions on the street. The bottom row shows different editing operations carried out on a telescope after modeling it. Note that different scaling factors have been applied to the different telescope parts.

Figure 7. Top: modeling and replicating parts for image editing. Orange parts are replicated or deformed. Bottom: editing a telescope. The leftmost images are the original photos. Note that different parts have been scaled differently.



Figure 8. Modeling the Obelisk in Paris from two photos. Top: the base of the Obelisk is modeled from a closer view which captures more details. Bottom: (a) The partial 3D model is transported to a more distant view (in which part of the base is occluded). (b) A rotated textured Obelisk; the texture of the transported part is blended into the region it occupied. (c) Details of the base are visible in the close-up of the new view.



In Figure 8, we show a case where two input photos are used to create one model of an object: the Obelisk in Paris. Firstly, the base of the Obelisk is modeled from a close up view in (a), allowing more detail to be captured. The partial 3D model is then moved to another photo where the entire Obelisk is visible, but the base is occluded. Similar to a copy and paste procedure, the user positions the extracted base inside the image, and it snaps to the image contours in (b). The user then continues the modeling process with other parts. The texture of the transported part is blended to match

the shading of the region in the new image, to maintain consistency: see the rotated view (c). Details of the base can be seen in the close up view (d) of the final model of the Obelisk.

In Figure 9, we show a photograph with a collection of objects that were modeled and copied from other photos. The online video (<https://vimeo.com/148236679>) shows the process of modeling and editing these objects. The modeling and editing time for each example is shown in Table 1, as well as the number of manually provided geo-semantic constraints. Objects in oblique views typically need more manual constraints, most of which designate coplanar axes, which are difficult to infer automatically.

6.2. Comparison to sketch based modeling

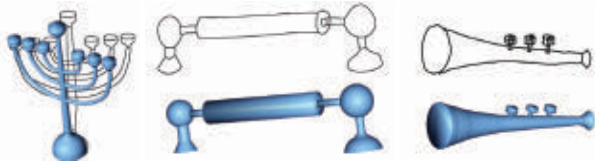
As discussed in Section 2, our method shares some similarities with the one in Shtof et al.,¹⁹ which models objects from sketches. We previously discussed the main differences between the methods. Their system is based on sketches rather than photographs, which makes it easier to assume that parts have sufficient bounding curves around them. It relies on labeling, using a drag and drop metaphor for choosing and positioning the primitives before snapping with the sketches. We make a comparison based on their sketch inputs (see Figure 10), since their method cannot handle the examples presented in this paper. Comparing the modeling time shows that sketch labeling and drag and drop snapping steps are significantly less efficient and less intuitive compared to our 3-Sweep method. Our modeling time (60s on average) is significantly lower than the time they report for their technique (180s on average).

6.3. Limitations

Our work has several limitations. First, many shapes cannot be decomposed into generalized cylinders and cuboids, and cannot be modeled using our framework (e.g., the base of the menorah in Figure 1). It would be desirable to extend the types of primitives which can be modeled using similar principles. 3-Sweep also relies on the fact that the object modeled

Figure 9. A rendered scene using source images from the top image strip.**Table 1. Modeling + Editing times (in seconds), and the number of manually provided geo-semantic constraints (added or removed) for each example.**

Figure	1	6						8	7	9			
Example	Menorah	(a)	(b)	(c)	(d/e)	(f)	Obelisk	Tap	Lamp	Telescope	Trumpet	Handle	Horn
Time (s)	80 + 25	75 + 15	20	35	30	65 + 35	20	30 + 25	40 + 50	100 + 30	80	30	60
Constraints	2	4	2	1	1	1	0	2	1	2	1	1	1

Figure 10. Modeling from sketches. Input sketches are taken from Ref.¹⁹

includes typical relationships among parts such as symmetry, parallelism, and collinearity. More free-form objects, or parts, that differ from this would need to be positioned by hand.

Even for a generalized cylinder there is sometimes ambiguity. We assume that the profiles of the cylinder are uniformly scaled and do not rotate around the main axis. This assumption is not always satisfied, as demonstrated in Figure 11b. We further assume that the main axis is mostly visible and parallel to the viewing plane. Using a perspective assumption, we can handle a small amount of skew, but not a large one.

The snapping algorithm relies on good detection of object edges and assumes the main part of the object is not occluded. Objects that are too small, such as the cross in Figure 6e, or objects that have fuzzy edges such as the example in Figure 11c are hard to model accurately. The photographs themselves often have some distortions from an ideal perspective projection (see Figure 11a), where corrections should be applied before modeling. Lastly, our editing assumes a simple illumination model without shadows. Relighting and shadow computations are not currently supported by our system.

Figure 11. Failures. (a) Due to perspective projection, the table legs cannot be snapped to the image under a parallelism constraint. (b) Due to the assumption of uniformly scaling profiles, the bottom of the toothpaste tube is not flattened. (c) Snapping fails due to an ill-defined edge caused by the shadow cast by the bottle.

7. CONCLUSION

We have presented an interactive technique which can model 3D man-made objects from a single photograph by combining the cognitive ability of humans with the computational accuracy of computers. The 3-Sweep technique is designed to allow extracting an editable model from a single image. The range of objects that our technique can support are objects that consist of simple parts, without much occlusion. As we demonstrated, this range is surprising large to be useful for interactive modeling—our tests show that our method can model a large variety of man-made objects in photographs, as well as objects in sketches. The modeled objects can be edited in a semantically meaningful way, both in the original image, or for use in composing new images. In the future, we hope to extend the range of primitives, and to allow modeling of the

freer shapes of natural objects. We also wish to add symmetry and smoothness constraints on the shapes. 3-Sweep could also be extended to allow modeling from multi-view images or video, without the help of depth data. The applications demonstrated mainly show editing and manipulation of geometry, but the recovered 3D models and surface normals can be used for re-lighting and material editing. ■

References

1. Andre, A., Saito, S. Single-view sketch based modeling. In *Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling (SBIM '11)* (2011). ACM, New York, NY, USA, 133–140.
2. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J. Contour detection and hierarchical image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.* 33, 5 (2011), 898–916.
3. Arkan, M., Schwärzler, M., Flöry, S., Wimmer, M., Maierhofer, S., O-snap: Optimization-based snapping for modeling architecture. *ACM Trans. Graph. (TOG)* 32, 1 (2013), 6.
4. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.-TOG* 28, 3 (2009), 24.
5. Benko, P., Kós, G., Várady, T., Andor, L., Martin, R. Constrained fitting in reverse engineering. *Comput. Aided Geom. Des.* 19, 3 (2002), 173–205.
6. Chen, T., Zhu, Z., Shamir, A., Hu, S.-M., Cohen-Or, D. 3sweep: Extracting editable objects from a single photo. *ACM Trans. Graph. (TOG)* 32, 6 (Nov. 2013), 1–195.
7. Cheng, M. Curve structure extraction for cartoon images. In *Proceedings of the 5th Joint Conference on Harmonious Human Machine Environment* (2009), 13–25.
8. Cheng, M., Zhang, F., Mitra, N., Huang, X., Hu, S. Repfinder: Finding approximately repeated scene elements for image editing. *ACM Trans. Graph. (TOG)* 29, 4 (2010), 83.
9. Choi, B., Lee, C. Sweep surfaces modelling via coordinate transformation and blending. *Comput.-Aided Des.* 22, 2 (1990), 87–96.
10. Debevec, P.E., Taylor, C.J., Malik, J. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96)* (1996). ACM, New York, NY, USA, 11–20.
11. Gal, R., Sorkine, O., Mitra, N., Cohen-Or, D. iwires: An analyze-and-edit approach to shape manipulation. *ACM Trans. Graph. (TOG)* 28 (2009) 33.
12. Gingold, Y., Igarashi, T., Zorin, D. Structured annotations for 2d-to-3d modeling. *ACM Trans. Graph. (TOG)* 28 (2009) 148.
13. Goldberg, C., Chen, T., Zhang, F., Shamir, A., Hu, S. Data-driven object manipulation in images. *Comput. Graph. Forum* 31 (2012) 265–274.
14. Lalonde, J., Hoiem, D., Efros, A., Rother, C., Winn, J., Criminisi, A. Photo clip art. *ACM Trans. Graph. (TOG)* 26 (2007) 3.
15. Mille, J., Boné, R., Cohen, L.D. Region-based 2d deformable generalized cylinder for narrow structures segmentation. In *Proceedings of Computer Vision—ECCV 2008: 10th European Conference on Computer Vision*, Forsyth, D., Torr, P. and Zisserman, A. (eds) (Marseille, France, October 12–18, Part II, 2008). Springer, Berlin, Heidelberg, 392–404.
16. Oh, B., Chen, M., Dorsey, J., Durand, F. Image-based modeling and photo editing. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '01)* (2001). ACM, New York, NY, USA, 433–442.
17. Olsen, L., Samavati, F., Sousa, M., Jorge, J. Sketch-based modeling: A survey. *Comput. Graph.* 33, 1 (2009), 85–103.
18. Oswald, M.R., Toppe, E., Cremers, D. Fast and globally optimal single view reconstruction of curved objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012). IEEE, 534–541.
19. Shtof, A., Agathos, A., Gingold, Y., Shamir, A., Cohen-Or, D. Geosemantic snapping for sketch-based modeling. Volume 32. In *Eurographics* (2013), 245–253.
20. Snavely, N. Scene reconstruction and visualization from internet photo collections: A survey. *IPSPJ Trans. Comput. Vision Appl.* 3 (2011), 44–66.
21. Xu, K., Zhang, H., Cohen-Or, D., Chen, B. Fit and diverse: Set evolution for inspiring 3d shape galleries. *ACM Trans. Graph. (TOG)* 31, 4 (2012), 57.
22. Xu, K., Zheng, H., Zhang, H., Cohen-Or, D., Liu, L., Xiong, Y. Photo-inspired model-driven 3d object modeling. *ACM Trans. Graph. (TOG)* 30 (2011) 80.
23. Xue, T., Liu, J., Tang, X. Symmetric piecewise planar object reconstruction from a single image. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2011). IEEE, 2577–2584.

Tao Chen and Daniel Cohen-Or
(taochen@ee.columbia.edu)
Tel Aviv & Tsinghua University,
Israel & Tsinghua University,
Beijing, China.

Zhe Zhu and Shi-Min Hu (ajex1988@gmail.com, shimin@tsinghua.edu.cn)
Tsinghua University, Beijing, China.

Ariel Shamir (arik@idc.ac.il) The
Interdisciplinary Center, Herzliya, Israel.

© 2016 ACM 0001-0782/16/12 \$15.00

ACM Transactions on Parallel Computing

Solutions to Complex Issues in Parallelism

Editor-in-Chief: Phillip B. Gibbons, Intel Labs, Pittsburgh, USA



ACM Transactions on Parallel Computing (TOPC) is a forum for novel and innovative work on all aspects of parallel computing, including foundational and theoretical aspects, systems, languages, architectures, tools, and applications. It will address all classes of parallel-processing platforms including concurrent, multithreaded, multicore, accelerated, multiprocessor, clusters, and supercomputers.

Subject Areas

- Parallel Programming Languages and Models
- Parallel System Software
- Parallel Architectures
- Parallel Algorithms and Theory
- Parallel Applications
- Tools for Parallel Computing



Association for
Computing Machinery

Advancing Computing as a Science & Profession

For further information or to submit your manuscript, visit topc.acm.org

Subscribe at www.acm.org/subscribe

CAREERS

Auburn University Department of Computer Science and Software Engineering (CSSE)

CSSE invites applications for multiple tenure-track faculty positions to begin in Fall 2017. Candidates at the Assistant Professor level will be given preference, however outstanding candidates at the Associate and Full Professor levels will also be considered. Those whose research expertise can connect with other disciplines of engineering, science and arts will be given special consideration. While we encourage all qualified candidates to apply, we are especially interested in candidates specializing in artificial intelligence, bioinformatics, cyber-infrastructure, data science & engineering, distributed systems, high performance computing, machine learning, parallel processing, privacy, programming languages, security, and software engineering.

One position is part of the Auburn University Cluster Hires Initiative in Omics & Informatics that cuts across agriculture, engineering, pharmacy, science and veterinary medicine. This position is for a researcher specializing in big data (analytics, learning, mining, visualization etc.) or informatics with the potential to collaborate with researchers from the disciplines represented in the cluster.

CSSE is home to the Auburn Cyber Research Center (cyber.auburn.edu), and is affiliated with the newly established McCrary Institute for Critical Infrastructure Protection and Cyber Systems (mccrary.auburn.edu). Both offer opportunities for faculty to engage in jointly funded research and development. The Department currently has 17 faculty members and supports strong graduate (M.S. and Ph.D.) and undergraduate programs in computer science, software engineering and wireless engineering. Enrollment for Fall 2016 is 765 undergraduate and 142 graduate students.

Faculty research areas include artificial intelligence, computer architecture, computer science education, database systems, energy-efficient systems, high performance computing, human-computer interaction, internet of things, learning science, machine learning, modeling & simulation, multi-agent systems, networks, security, software engineering, software visualization and wireless engineering. A full advertisement and application instructions may be found at the Department's home page (www.eng.auburn.edu/csse).

Boston College Assistant Professor of the Practice in Computer Science (2 positions available)

The Boston College Computer Science Department invites applications for two full time non tenure-track Assistant Professorships of the Prac-

tice starting September 2017. Teaching load is 3 courses per semester. A PhD in Computer Science and a commitment to excellence in undergraduate teaching are required. Interest in interdisciplinary curricular innovation is desirable. See www.cs.bc.edu for more information. Application review begins October 15, 2016. Apply at apply.interfolio.com/38103.

California State University Long Beach

The Computer Engineering/Computer Science Department at California State University Long Beach has multiple open tenure track positions in Computer Science (Machine Learning) and Computer Engineering (Embedded Systems and Networking, Multimedia or Hardware Security). For more information please go to <http://web.csub.edu/divisions/aa/personnel/jobs/coe/>.

Calvin College seeks applicants for a tenure track faculty position in Computer Science, pending administrative approval. We feature supportive colleagues, excellent facilities, hard-working students, and strong undergraduate programs in computer science, data science, digital communications, and information systems, including a BCS major accredited by ABET (abet.org). Responsibilities include teaching computing-related courses, maintaining an active scholarly life, and mentoring students. Applicants should have a PhD (or be near completion) in computer science or a related area. To apply, please send a C.V., a statement of teaching philosophy and scholarly interests, contact information for three professional references, and a cover letter to:

Prof. Joel Adams
Chair, Department of Computer Science
Calvin College
3201 Burton SE
Grand Rapids, MI 49546
(adams@calvin.edu)

Calvin College seeks faculty members who affirm the Christian faith as expressed by the Reformed confessions and who have academic and personal qualifications for teaching and scholarship. Calvin is building a tradition of diversity and accessibility and welcomes applications from persons whose personal characteristics will further that commitment. Calvin does not discriminate in the employment of individuals on the basis of race, color, national or ethnic origin, disability, sex, or age. Calvin College is an educational agency of the Christian Reformed Church and, in compliance with Title VII and other applicable law, reserves the right to give preference in employment based upon religion. Calvin is one of the largest Christian colleges in North America, and was named the #1 regional college in the Midwest for 2017 by *U.S. News & World Report*. Additional information about the college and department is available at: <http://cs.calvin.edu/>.

Columbia University in the City of New York Faculty Positions in the Department of Computer Science

Columbia Engineering invites applications for faculty positions in the Department of Computer Science at Columbia University in the City of New York. Applications at the assistant professor, and in exceptional cases, at the associate professor and full professor levels, will be considered.

Applications are sought in all areas of computer science, with particular emphasis on, but not limited to, the following areas: Artificial Intelligence (all areas) and Programming languages.

Candidates must have a Ph.D. or its professional equivalent by the starting date of the appointment. Applicants for this position at the Assistant Professor and Associate Professor without tenure levels must demonstrate the potential to do pioneering research and to teach effectively. Applicants for this position at the tenured level (Associate or Full Professor) must have a demonstrated record of outstanding research accomplishments, excellent teaching credentials and established leadership in the field.

The successful candidate is expected to contribute to the advancement of their field and the department by developing an original and leading externally funded research program, and by contributing to the undergraduate and graduate educational mission of the Department. Columbia fosters multidisciplinary research and encourages collaborations with academic departments and units across Columbia University. The Department is especially interested in qualified candidates who can contribute, through their research, teaching, and/or service, to the diversity and excellence of the academic community.

For additional information and to apply, please see: <http://engineering.columbia.edu/faculty-job-opportunities>.

Applications should be submitted electronically and include the following: curriculum-vitae including a publication list, a description of research accomplishments, a statement of research and teaching interests and plans, contact information for three experts who can provide letters of recommendation, and up to three pre/reprints of scholarly work. All applications received by December 15, 2016 will receive full consideration.

Applicants can consult www.cs.columbia.edu for more information about the department. Columbia is an affirmative action/equal opportunity employer with a strong commitment to the quality of faculty life.

Concordia College

The Offutt School of Business invites applicants with a background and qualifications in Computer Science for a full-time tenure eligible position, beginning in August 2017.

For a detailed description and to apply, visit <https://hr.cord.edu/postings/2514>

Florida State University
Department of Computer Science
Tenure-Track Assistant Professor Positions

The Department of Computer Science at the Florida State University invites applications for two tenure-track Assistant Professor positions to begin August 2017. The positions are 9-month, full-time, tenure-track, and benefits eligible. We are seeking outstanding applicants with strengths in the broad areas of Data Sciences or Trustworthy Computing. The Data Sciences area may involve Databases, Computer Graphics and Visualization, Artificial Intelligence, Data Mining, Data Engineering, Data Analytics, and Big Data Applications and systems as well as other Data Sciences related research. The Trustworthy Computing area includes Formal Methods and Verification, Programming Languages and Compilers, Cyber-Physical Systems, Embedded Systems, Computer Architecture, and Digital Forensics among others. Applicants should hold a PhD in Computer Science or closely related field at the time of appointment, and have excellent research and teaching accomplishments or potential. The department offers degrees at the BS, MS, and PhD levels. The department is an NSA Center of Academic Excellence in Information Assurance Education (CAE/IAE) and Research (CAE-R).

FSU is classified as a Carnegie Research I university. Its primary role is to serve as a center for advanced graduate and professional studies while emphasizing research and providing excellence in undergraduate education. Further information can be found at: <http://www.cs.fsu.edu>.

Screening will begin January 1, 2017 and will continue until the positions are filled.

Please apply online with curriculum vitae, statements of teaching and research philosophy, and the names of three references, at: <http://www.cs.fsu.edu/positions/apply.html>.

Georgia Institute of Technology
Tenure Track Faculty Positions

The School of Interactive Computing at Georgia Tech invites applications for tenure-track faculty positions at all levels. Although the areas of primary interest are Robotics and Machine Learning, we will consider truly outstanding candidates in other areas that support the school's mission or expand it in strategic directions. Deadline: Dec 15, 2016. <https://academicjobsonline.org/ajob/jobs/8083>.

Harvard University

The Harvard John A. Paulson School of Engineering and Applied Sciences seeks applicants for a position at the tenure-track level in Computer Science, with an expected start date of July 1, 2017.

We are accepting applications in all areas of Computer Science, especially machine learning and programming languages.

We seek candidates who have a strong research record, and a commitment to undergraduate and graduate teaching and training. Candidates are required to have a doctorate or terminal

degree in Computer Science or a related area by the expected start date.

Required application documents include a cover letter, CV, a statement of research interests, a teaching statement, and up to three representative papers. Candidates are also required to submit the names and contact information for at least three and up to five references, and the application is complete only when three letters have been submitted. We encourage candidates to apply by December 1, 2016, but will continue to review applications until the position is filled. Applicants will apply on-line at <http://academicpositions.harvard.edu/postings/7129>

Computer Science at Harvard benefits from outstanding undergraduate and graduate students, world-leading faculty, an excellent location, significant industrial collaboration, and substantial support from the Harvard Paulson School. Information about Harvard's current faculty, research, and educational programs in computer science is available at <http://www.seas.harvard.edu/computer-science>. The associated Institute for Applied Computational Science (<http://iacs.seas.harvard.edu>) fosters connections among computer science, applied math, data science, and various domain sciences at Harvard through its graduate programs and events.

We are an equal opportunity employer and all qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, disability status, protected veteran status, or any other characteristic protected by law.

DEPARTMENT CHAIR • COMPUTER SCIENCE AND ENGINEERING

Michigan State University (MSU) invites nominations and applications for the position of **Chair of the Department of Computer Science and Engineering (CSE)** in the College of Engineering, with a preferred starting date of August 16, 2017.

The CSE department has a strong research and educational program that is renowned for its multidisciplinary nature. The Chair of CSE will oversee department operations, personnel and resources. The Chair will advance the departmental vision of academic leadership and excellence, and will represent the department to academia, industry and government for its continued vitality. The Chair will actively work with faculty in the department and across the university to identify and pursue innovations in research, education, and service. The Chair will be actively involved in strategic planning that would expand the influence of the department in the broader academic community both within and outside the university, and will promote diversity within the department and the institution.

The Department has 31 tenure system faculty members and 5 teaching specialists, including a National Academy of Engineering member and 9 NSF CAREER awardees. The Department has strong research programs in all major areas of Computer Science and Engineering, with an annual research expenditure of over \$7M. The department has accredited B.S. degree programs in Computer Science and Computer Engineering. There are currently over 140 graduate students and 1,100 undergraduate students in the department. Department faculty are actively involved in several national-level centers, including BEACON, an NSF Center that is headquartered at MSU for the study of evolution in action.

MSU enjoys a park-like campus with outlying research facilities and natural areas. The campus is adjacent to the city of East Lansing and the capital city of Lansing. The Lansing metropolitan area has a diverse population of approximately 470,000. Local communities have excellent school systems and place a high value on education. Michigan State University is proactive in exploring opportunities for employment for dual career couples, both inside and outside the University. Information about MSU's dual career support can be found at <http://miwin.msu.edu/>. Information about WorkLife at MSU and the College of Engineering can be found at <http://www.egr.msu.edu/WE>.

The successful candidate must have an earned PhD in Computer Science or a closely related field, and be qualified to receive an annual appointment at the rank of Full Professor with tenure. The candidate should have a distinguished track record of research, classroom teaching, student mentorship, and funding from diverse sources. The candidate must provide evidence of scientific and organizational leadership, educational innovation, and have outstanding communication, interpersonal, and administrative skills.

Interested individuals should submit an application for this position through: <https://jobs.msu.edu> and refer to position #4095. Applicants must submit a cover letter summarizing qualifications and leadership approach, a vision statement for the department, a detailed curriculum vitae, and the names and contact information of five references. Applications will be reviewed on a continuing basis until the position is filled. For full consideration, applications should be received before December 1, 2016. Nominations or questions are welcome by contacting the search committee chair at cse-chair-search@cse.msu.edu.

Additional information about the university, college and department is available at: CSE Department - <http://www.cse.msu.edu> | College of Engineering - <http://www.egr.msu.edu/> | MSU - <http://www.msu.edu/>

MSU has been advancing the common good with uncommon will for more than 160 years. A member of the Association of American Universities, MSU is a research-intensive institution with 17 degree-granting colleges.

MICHIGAN STATE
UNIVERSITY

MSU is committed to achieving excellence through cultural diversity.

The University actively encourages applications and/or nominations of women, persons of color, veterans and persons with disabilities.

MSU is an Affirmative Action, Equal Opportunity Employer.

Indiana University

School of Informatics and Computing
Faculty Positions in Intelligent Systems
Engineering

The School of Informatics and Computing (SoIC) at Indiana University (IU) Bloomington invites applications for multiple open rank tenure/tenure track faculty positions to begin in Fall 2017 in **Intelligent Systems Engineering**. Cluster hires are encouraged as are interdisciplinary applications spanning the interests of SoIC and collaborating with IU units. Duties include teaching, research, and service.

Applicants should have an established record (for senior level) or demonstrable potential for excellence (for junior level) in research and teaching, and a PhD in a relevant field or (for junior level) expected before 8/2017.

The department of Intelligent Systems Engineering is an innovative new program, currently with 12 faculty, that focuses on the engineering of systems of smaller-scale, often mobile devices that draw upon modern information technology techniques including intelligent systems, big data and user interface design. Its foundation also includes computer engineering, cyber-physical systems, robotics, sensor and detector technologies, signal processing, and information and control theory. **We intend to add about 20 new faculty over the next 4 years for these areas as well as bioengineering, synthetic biology, molecular and nanoscale engineering, environmental engineering and neuro-engineering with an interdisciplinary IT component.** The program offers a BS and a PhD, which started in fall 2016, while a future MS will be developed. Artificial Intelligence, Modeling and Simulation and the Internet of Things are integrated into core curriculum. The program will draw upon IU Bloomington's considerable education and research strengths in biology, chemistry, computer science, environmental science, informatics, physics, network science, psychological and brain sciences, business and law. New faculty will have considerable opportunity and responsibility to shape the development of curricula and research. There will be a strong emphasis on world-class research, built around a few strong focused laboratories and proactively involving undergraduates. More information can be found at <https://www.engineering.indiana.edu/>.

Apply online at: <http://indiana.peopleadmin.com/postings/2730>

For full consideration applications are due by 1/2/17 but earlier submission is encouraged.

Questions regarding the positions or application process can be directed to isechair@indiana.edu

Indiana University is an equal employment and affirmative action employer and a provider of ADA services. All qualified applicants will receive consideration for employment without regard to age, ethnicity, color, race, religion, sex, sexual orientation or identity, national origin, disability status or protected veteran status.

Max Planck Institute for Software Systems and Aircloak

Applications are invited for tenure-track faculty positions in all areas related to the theory and practice of software systems, including secu-

FACULTY CLUSTER HIRE IN CYBERSECURITY

The University of Texas at San Antonio (UTSA) has embarked on a focused cluster hiring plan under the Gold Star Initiative, to recruit top-tier researchers over a four year period. The plan will focus on strategic areas of research excellence, to include cybersecurity, cloud computing and data analytics. UTSA is currently looking for candidates to fill six faculty positions to foster collaborative research, education and outreach and to create interdisciplinary areas of knowledge that will advance the field of cybersecurity.

UTSA is a recognized leader in the field of infrastructure assurance and security by the National Security Agency and the Department of Homeland Security and is a designated Center of Academic Excellence in Information Assurance Education (CAE). In spring 2014, UTSA was ranked #1 nationally for Cyber Security Programs according to a national survey of certified information technology security professionals conducted for Hewlett-Packard. UTSA is home to the Institute for Cyber Security (ICS), which conducts basic and applied cybersecurity research in partnership with academia, government and industry. The Center for Infrastructure Assurance and Security (CIAS), also located at UTSA, has developed the world's foremost center for multidisciplinary education and development of operational capabilities in the areas of infrastructure assurance and security. In complement to the ICS and CIAS, the Center for Education and Research in Information and Infrastructure Security (CERIS) conducts high impact research, as well as educates the cybersecurity workforce within the San Antonio area and beyond. In partnership with Rackspace, UTSA houses the largest open cloud infrastructure in academia. The Open Cloud Institute is an initiative to develop degree programs in cloud computing and foster collaboration with industry, positioning UTSA and San Antonio as world leaders in open cloud technology.

Required Qualifications and Responsibilities:

The successful applicant will have a Doctoral degree (Ph.D.) and publications commensurate with appointment levels in the department of interest. Successful candidates will be expected to develop and maintain externally funded research programs, engage in both undergraduate and graduate education, and contribute their leadership and innovative thinking towards global prominence in cybersecurity. Teaching opportunities will vary by department and teaching qualifications will be a consideration for fit within their respective department.

Application Process:

Applicants must submit their full application package via the respective link to each position. For more information about this cluster and to access the position links, please visit <http://research.utsa.edu/research-news/cyber/>. For general questions or additional information on the Gold Star Initiative, please contact: Bernard Arulanandam, Interim Vice President for Research at Bernard.arulanandam@utsa.edu or 210-458-8176.

Enterprise Security

Tenure-track Assistant Professor in the Department of Communication. This position is targeted towards faculty with expertise and interest in the areas of

situational awareness and decision-making, cyber data analysis, attack and response, human-machine interactions, organizational communication, information networks, and cybersecurity training.

Cyber Analytics

Tenure-track Assistant, Tenured/tenure-track Associate Professor in the Department of Information Systems and Cyber Security. This position is targeted towards faculty with expertise and interest in conducting transformative research and developing tangible "big data" solutions to cyber analytics challenges with interests in the following domains: embedded system security, cloud security, enterprise security situational awareness and decision making; and/or digital forensics.

Cyber Decision Support

Tenure-track Assistant, Tenured/tenure-track Associate Professor in the Department of Information Systems and Cyber Security. This position is targeted towards faculty with expertise and interest in conducting transformative research that enables organizations to make cyber related decisions quickly, effectively, and accurately. We are particularly interested in the following areas: cyber resiliency; enterprise security situational awareness and decision making; and/or risk assessment and management.

Cloud Computing Security

Tenure-track Assistant Professor in the Department of Electrical and Computer Engineering. This position is targeted towards faculty with expertise and interest in security and privacy issues in cloud computing. Specific topics of interest include secure and privacy-aware data analytics in cloud, data analytics techniques to enhance cloud security, secure software defined networking and network function virtualization, cloud monitoring, dependability issues (availability, assurance and recover) in cloud, secure multi-tenancy, hardware architectures to improve cloud security, etc.

Embedded Systems Security

Tenure-track Assistant Professor in the Department of Electrical and Computer Engineering. This position is targeted towards faculty with expertise and interest in embedded systems security. Areas of particular interest are: security of embedded systems with applications to cyber physical systems such as the Internet of Things, energy, transportation, building design, automation, healthcare and manufacturing.

Privacy and Data Protection

Tenure-track Assistant Professor in the Department of Computer Sciences. This position is targeted towards faculty with expertise and interest in privacy protection and security. Applicants with expertise in software engineering, programming languages and compilers, or big data analytics are particularly encouraged to apply.

As an Equal Employment Opportunity and Affirmative Action employer, it is the policy of The University of Texas at San Antonio to promote and ensure equal employment opportunity for all individuals without regard to race, color, religion, sex, national origin, age, sexual orientation, gender identity, disability, or veteran status. The University is committed to the Affirmative Action Program in compliance with all government requirements to ensure nondiscrimination. The UTSA campus is accessible to persons with disabilities.

city and privacy, embedded and mobile systems, computational social science, legal, economic, and social aspects of computing, NLP, machine learning, information and knowledge management, programming languages, verification, parallel and distributed systems.

A doctoral degree in computer science or related areas and an outstanding research record are required. Successful candidates are expected to build a team and pursue a highly visible research agenda, both independently and in collaboration with other groups.

MPI-SWS, founded in 2005, is part of a network of over 80 Max Planck Institutes, Germany's premier basic research facilities. MPIs have an established record of world-class, foundational research in the sciences, technology, and the humanities. The institute offers a unique environment that combines the best aspects of a university department and a research laboratory: Faculty enjoy academic freedom, receive institutional funding and attract additional third-party funds to build and lead a team of graduate students and post-docs; they supervise doctoral theses, and have the opportunity to teach graduate and undergraduate courses. The institute offers outstanding technical infrastructure and administrative support, as well as internationally competitive compensation.

The institute is located in Kaiserslautern and Saarbrücken, in the tri-border area of Germany, France and Luxembourg. We maintain an international and diverse work environment and seek applications from outstanding researchers worldwide. The working language is English;

knowledge of the German language is not required for a successful career at the institute.

Qualified candidates should apply at the secure application form (apply.mpi-sws.org). To receive full consideration, applications should be received by December 15, 2016.

The institute is committed to increasing the representation of minorities, women and individuals with physical disabilities in Computer Science. We particularly encourage such individuals to apply. The initial tenure-track appointment is for five years; it can be extended to seven years based on a midterm evaluation in the fourth year. A permanent contract can be awarded upon a successful tenure evaluation in the sixth year.

Michigan State University Computer Science and Engineering Faculty Position in Biometrics

Join Michigan State University's Global Impact Initiative, designed to address the grand challenges through the creation of over 100 new faculty positions in some of the most promising and exciting fields of research. We welcome applicants from diverse backgrounds. MSU offers an inclusive and collaborative work environment.

To learn more visit research.msu.edu/global-impact

The Department of Computer Science and Engineering (CSE) at Michigan State University (MSU) invites applications for a faculty position in the area of biometrics. While the position is primarily for a junior faculty, candidates at other

ranks may be considered. The successful candidate will be expected to develop an externally-funded interdisciplinary research program of international prominence that includes fundamental research, publications in high-impact journals and conferences, and training graduate students. Multidisciplinary research is strongly encouraged and is being actively pursued by the faculty members at MSU. Leadership is expected in the development of innovative educational programs that provide state-of-the-art knowledge to both undergraduate and graduate students. Candidates should have a Ph.D. in Computer Science or a closely related field, with demonstrated evidence of research accomplishments, teaching skills, and ability to work effectively with other researchers within the Department and colleagues on campus. Appointments will start in August 2017.

The campus is in the city of East Lansing and adjacent to the capital city of Lansing. The Lansing metropolitan area has a diverse population of approximately 450,000 residents. Local communities have excellent school systems and place a high value on education. Michigan State University is pro-active in exploring opportunities for employment for dual career couples, both inside and outside the University: <http://miwin.msu.edu/>. Information about work and life at MSU and the College of Engineering can be found at <http://www.egr.msu.edu/WE>.

Applicants should submit a cover letter, curriculum vitae, the names of at least three references, and statements of their research and teaching interests through <https://jobs.msu.edu>,



Florida International University is classified by Carnegie as a R1: Doctoral Universities - Highest Research Activity and recognized as a Carnegie engaged university. It is a public research university with colleges and schools that offers 196 bachelor's, master's and doctoral programs in fields such as engineering, computer science, international relations, architecture, law and medicine. As one of South Florida's anchor institutions, FIU contributes almost \$9 billion each year to the local economy. FIU is Worlds Ahead in finding solutions to the most challenging problems of our time. FIU emphasizes research as a major component of its mission. FIU has awarded more than 220,000 degrees and enrolls more than 54,000 students in two campuses and three centers including FIU Downtown on Brickell, FIU@I-75, and the Miami Beach Urban Studios. FIU's Medina Aquarius Program houses the Aquarius Reef Base, a unique underwater research facility in the Florida Keys. FIU also supports artistic and cultural engagement through its three museums: Patricia & Phillip Frost Art Museum, the Wolfsonian-FIU, and the Jewish Museum of Florida-FIU. FIU is a member of Conference USA and has more than 400 student-athletes participating in 18 sports. For more information about FIU, visit <http://www.fiu.edu/>.

FIU's School of Computing and Information Sciences (SCIS) is a rapidly growing program of excellence at Florida International University (FIU). The School has 29 tenure-track faculty members and over 2,000 students, including over 90 Ph.D. students. The School is engaged in on-going and exciting new and expanding programs for research, education and outreach. The School offers B.S., M.S., and Ph.D. degrees in Computer Science, and M.S. degrees in Telecommunications and Networking, Cyber-security, and Information Technology as well as B.S./B.A degrees in Information Technology. NSF ranks FIU 43rd nationwide in externally-funded research expenditures. SCIS has six research centers/clusters with first-class computing and support infrastructure, and enjoys broad and dynamic industry and international partnerships.

We invite applications from exceptionally qualified faculty at all levels with particular emphasis on networking, cyber-security, computer systems or data sciences, and other related areas. Ideal candidates for junior positions should have a record of exceptional research in their early careers and a demonstrated ability to pursue and lead a research program. Candidates for senior positions must have an active and sustainable record of excellence in funded research, publications and professional service as well as demonstrated leadership in collaborative or interdisciplinary research. In addition to developing or expanding a high-quality research program, all successful applicants must be committed to excellence in teaching at both the graduate and undergraduate levels. Applications are encouraged from candidates with highly transformative research programs and seminal ideas that extend the frontiers of computing and networking across other disciplines. A Ph.D. in Computer Science or related disciplines is required.

HOW TO APPLY: Qualified candidates are encouraged to apply to Job Opening ID (Job Opening ID # 512441) at facultycareers.fiu.edu and attach a cover letter, curriculum vitae, statement of teaching philosophy, research statement, etc as *individual attachments*. Candidates will be requested to provide names and contact information for at least three references who will be contacted as *determined by the search committee*. To receive full consideration, applications and required materials should be received by December 31, 2016. Review will continue until position is filled.

FIU is a member of the State University System of Florida and an Equal Opportunity, Equal Access Affirmative Action Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, national origin, disability status, protected veteran status, or any other characteristic protected by law.

refer to posting #4074. Applications will be reviewed on a continuing basis until the position is filled. Review of applications begins January 2, 2017. For questions about this position, contact the search committee chair at search@cse.msu.edu.

Additional information about the university, college and department is available at: CSE Department - <http://www.cse.msu.edu>; College of Engineering - <http://www.egr.msu.edu/> and MSU-<http://www.msu.edu/>.

Michigan State University has been advancing the common good with uncommon will for more than 160 years. A member of the Association of American Universities, MSU is a research-intensive institution with 17 degree-granting colleges.

MSU is an affirmative-action, equal opportunity employer.

Princeton University Assistant Professor

The Department of Computer Science at Princeton University invites applications for tenure track faculty positions at the Assistant Professor level. We are accepting applications in all areas of Computer Science. Applicants must demonstrate superior research and scholarship potential as well as teaching ability.

A PhD in Computer Science or a related area is required. Candidates should expect to receive their PhD before Fall 2017. Successful candidates are expected to pursue an active research program and to contribute significantly to the teaching programs of the department. Applicants should include a CV and contact information for at least three people who can comment on the applicant's professional qualifications.

For full consideration we recommend that applicants apply by December 1, 2016, though we will continue to review applications past that date.

Princeton University is an equal opportunity employer and all qualified applicants will receive consideration for employment without regard to age, race, color, religion, sex, sexual orientation, gender identity or expression, national origin, disability status, protected veteran status, or any other characteristic protected by law.

This position is subject to the University's background check policy.

Requisition Number: 1600774. You may apply online at: <https://jobs.cs.princeton.edu/junior/>

Princeton University Associate and Full Professor level

The Department of Computer Science at Princeton University invites applications at the Associate and Full Professor level. We are accepting applications in all areas of Computer Science. Applicants must demonstrate superior research and scholarship as well as an excellent teaching record.

A PhD in Computer Science or a related area is required. Successful candidates are expected to pursue an active research program and to contribute significantly to the teaching programs of the department. Applicants should include a CV and contact information for at least three people who can comment on the applicant's professional qualifications.



THE UNIVERSITY OF TEXAS
AT DALLAS

Erik Jonsson School of
Engineering and Computer Science

COMPUTER SCIENCE/SOFTWARE ENGINEERING

TENURED/TENURE-TRACK FACULTY POSITIONS

The Department of Computer Science of The University of Texas at Dallas invites applications from outstanding applicants for multiple tenure-track positions in computer science. Candidates in all areas of Computer Science will be considered though the department is particularly interested in areas of software engineering, machine learning, data science, cyber security, and IoT. Candidates must have a Ph.D. in Computer Science, Software Engineering, Computer Engineering or equivalent. The positions are open for applicants at all ranks. Candidates for senior positions must have a distinguished record of research, publication, teaching and service, and demonstrated leadership ability in developing and expanding (funded) research programs. An endowed chair may be available for highly qualified senior candidates. Junior candidates must show outstanding promise.

The Department offers B.S., M.S., and Ph.D. degree programs both in Computer Science and Software Engineering, as well as in interdisciplinary fields of Telecom Engineering and Computer Engineering. Currently the Department has a total of 48 tenure-track faculty members and 36 senior lecturers. The department is housed in a spacious 150,000 square foot facility and has excellent computing equipment and support. The department houses a number of centers and institutes, particularly in areas of **cyber security**, **human language technology**, and **net centric software**.

The University is located in the most attractive part of the Dallas metropolitan area. There are over 800 high-tech companies within a few miles of the campus, including Texas Instruments, Alcatel, Ericsson, Hewlett-Packard, AT&T, Fujitsu, Raytheon, Rockwell Collins, and Cisco. Almost all the country's leading telecommunication companies have major research and development facilities in our neighborhood. Opportunities for joint university-industry research projects are excellent. The Department averages \$9 million in extramural research funding annually putting it in 34th rank nationally in the ASEE survey of research expenditures. The University and the State of Texas are also making considerable investment in commercialization of technology developed in University labs; a new start-up business incubation center was opened in September 2011.

The CS Department has experienced very rapid growth in recent years. The University and the State of Texas are investing significant resources in the pursuit of excellence. We welcome inquiries.

To apply, please visit <http://jobs.utdallas.edu/postings/6764>.

For more information, contact Dr. Gopal Gupta, Department Head, via email (gupta@utdallas.edu), or send e-mail to cs-search@utdallas.edu, or view the internet web page at <http://cs.utdallas.edu>. The search committee will begin evaluating applications on January 10th. Applications received on or before January 31st will get highest preference.

The University of Texas at Dallas is an Equal Opportunity/Affirmative Action employer and strongly encourages applications from candidates who would enhance the diversity of the University's faculty and administration.

For full consideration we recommend that applicants apply by December 1, 2016, though we will continue to review applications past that date.

Princeton University is an equal opportunity employer and all qualified applicants will receive consideration for employment without regard to age, race, color, religion, sex, sexual orientation, gender identity or expression, national origin, disability status, protected veteran status, or any other characteristic protected by law. This position is subject to the University's background check policy.

Requisition Number: 1600784

You may apply online at: <https://jobs.cs.princeton.edu/senior/>

Princeton University Research

The Department of Computer Science at Princeton University, in collaboration with the Institute for Advanced Study, offers two 3-year positions for outstanding new Ph.D.'s working in theoretical aspects of computer science, including machine learning. Combining research with teaching duties, these positions come with attractive benefits and working conditions. Typically, the first and third years of these appointments are spent at Princeton University and the second year is spent conducting research (without teaching duties) at the Institute for Advanced Study. These arrangements are flexible.

The typical teaching load in the first year of the position is a single one-semester course. The IAS has no teaching duties.

For full consideration applicants should apply by December 1, 2016, though late applications may be reviewed when needed.

Candidates are expected to complete a PhD in Computer Science or related field by July 1, 2017 for a September 2017 start. Please include a CV, research statement, teaching statement, and 3 letters, of which one should address teaching abilities.

Candidates who have already applied for postdoctoral positions at Princeton University or the Institute for Advanced Study need to apply separately for these positions.

This position is subject to the University's background check policy. Princeton University is an equal opportunity employer and all qualified applicants will receive consideration for employment without regard to age, race, color, religion, sex, sexual orientation, gender identity or expression, national origin, disability status, protected veteran status, or any other characteristic protected by law.

Apply online at: <https://jobs.princeton.edu>
Requisition number: 1600891

Princeton University Postdoctoral Position

Department of Computer Science at Princeton University is seeking applications for postdoctoral or more senior research positions in theoretical computer science and theoretical machine learning.

Positions are for one year starting in September

2017 with the possibility of renewal. Candidates must have a PhD in Computer Science or a related field.

For full consideration, we recommend that candidates apply (including letters of recommendation) by December 1, 2016, though we will continue to review applications past that date.

Applicants should apply to: <http://jobs.princeton.edu/> and submit a cover letter, CV, research statement, and contact information for three references.

This position is subject to the University's background check policy.

Princeton University is an equal opportunity employer and all qualified applicants will receive consideration for employment without regard to age, race, color, religion, sex, sexual orientation, gender identity or expression, national origin, disability status, protected veteran status, or any other characteristic protected by law.

Requisition Number: 1600886

Purdue University Tenure-Track/Tenured Faculty Positions

The Department of Computer Science at Purdue University is in a phase of significant growth. Applications are solicited for seven tenure-track and tenured positions at the Assistant, Associate and Full Professor levels. Outstanding candidates in all areas of computer science will be considered. Review of applications and candidate interviews will begin in October 2016, and will continue until the positions are filled.

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN Positions in Computing

The Department of Electrical and Computer Engineering (ECE) at the University of Illinois at Urbana-Champaign invites applications for faculty positions at all areas and levels in computing, broadly defined, with particular emphasis on reliable and secure computing; networked and distributed computing; high-performance, energy-efficient, and scientific computing; data center and storage systems; data science, machine learning and its applications; complex data analysis and decision science; bio-inspired computing; computational genomics; and health informatics, among other areas. Applications are encouraged from candidates whose research programs specialize in core as well as interdisciplinary areas of electrical and computer engineering. From the transistor and the first computer implementation based on von Neumann's architecture to the Blue Waters petascale computer – the fastest computer on any university campus, ECE Illinois faculty have always been at the forefront of computing research and innovation. The department is engaged in exciting new and expanding programs for research, education, and professional development, with strong ties to industry. The ECE Department has recently settled into its new 235,000 sq. ft. net-zero energy design building, which is a major campus addition with maximum space and minimal carbon footprint.

Qualified senior candidates may also be considered for tenured full Professor positions as part of the Grainger Engineering Breakthroughs Initiative (<http://graingerinitiative.engineering.illinois.edu>), which is backed by a \$100-million gift from the Grainger Foundation.

Please visit <http://jobs.illinois.edu> to view the complete position announcement and application instructions. Full consideration will be given to applications received by December 15, 2017, but applications will continue to be accepted until all positions are filled.

Illinois is an EEO Employer/Vet/Disabled
www.inclusiveillinois.illinois.edu

The University of Illinois conducts criminal background checks on all job candidates upon acceptance of a contingent offer

udjobs
Employer of Choice

Faculty Positions in Strategic Computing & Information Sciences

The Department of Electrical and Computer Engineering at the University of Delaware invites applications for a Tenure-track Assistant Professor position in the area of strategic computing, including high performance computing, big data analytics and cybersecurity. ECE initiatives are supported by over 40,000 square feet of departmental facilities and fueled with over \$10M/year in research expenditures.

The University of Delaware combines a rich history in engineering with a commitment to undergraduate education and scholarly excellence. The newly opened 194,000-square-foot Harker Interdisciplinary Science and Engineering Laboratory greatly expands opportunities and resources for interdisciplinary research and active learning, and the recently acquired 272-acre Science, Technology and Advanced Research (STAR) campus offers additional opportunities for academic and commercial collaborative research and development.

Candidates must have a Ph.D. degree in electrical or computer engineering, computer science or a related discipline. Applicants should submit a curriculum vita, a one-page statement of teaching experience and interests, a 3-6-page research proposal and a list of at least four references at <https://apply.interfolio.com/38818>. The University of Delaware values diversity and is supportive of the needs of dual-career couples; women and minorities are especially encouraged to apply.

The University of Delaware is an Equal Opportunity Employer which encourages applications from Minority Group Members, Women, Individuals with Disabilities and Veterans. The University's Notice of Non-Discrimination can be seen at: www.udel.edu/aboutus/legalnotices.html

Dare to be first.
UNIVERSITY OF
DELAWARE

The Department of Computer Science offers a stimulating academic environment with research programs in most areas of computer science. Information about the department and a description of open positions are available at <http://www.cs.purdue.edu>.

Applicants should hold a PhD in Computer Science or a related discipline, have demonstrated excellence in research, and strong commitment to teaching. Successful candidates will be expected to conduct research in their fields of expertise, teach courses in computer science, and participate in other department and university activities. Salary and benefits are competitive, and Purdue is a dual-career friendly employer. Applicants are strongly encouraged to apply online at <https://hiring.science.purdue.edu>. Alternatively, hardcopy applications can be sent to: Faculty Search Chair, Department of Computer Science, 305 N. University Street, Purdue University, West Lafayette, IN 47907. A background check will be required for employment in this position.

Purdue University is an EEO/AA employer. All individuals, including minorities, women, individuals with disabilities, and veterans are encouraged to apply.

Rutgers University

The Department of Computer Science at Rutgers University invites applications for one or more positions as Teaching Professor or Professor of Practice in the area of Data Science, at the level of Assistant Professor, although exceptional candi-

dates may be appointed at the rank of associate or full professor. These appointments may begin in either spring or fall semester 2017.

Main Responsibilities will include Teaching, Coordination of our Capstone project series, Design of short term tutorials dedicated to Data Science topics of current interest and Coordination of Data Science Workshops.

Questions can be directed to: James Abello (abello@dimacs.rutgers.edu)

To apply, please go to <http://apply.interfolio.com/37381>.

Saint Bonaventure University Director of Cybersecurity Program

Saint Bonaventure University invites applications for a tenure track position as Director of the Cybersecurity program. The Director will work with current faculty to further develop the new cybersecurity major that is offered in conjunction with Hilbert College. Candidates should have extensive experience in the field. Details can be found at: <http://www.sbu.edu/cybersecuritydirector>.

Southern University of Science and Technology (SUSTech)

The Southern University of Science and Technology (SUSTech) invites applications for faculty positions in the Department of Computer Science and Engineering. It is seeking to appoint up to eight teaching-track positions at all ranks. Candidates with

teaching interests in all core fields of Computer Science and Engineering will be considered. These positions are full-time and tenure track posts although they are teaching focused. SUSTech adopts the tenure track system, which offers the recruited faculty members a clearly defined career path.

Candidates should have demonstrated excellence in and commitment to teaching and research. A doctoral degree in Computer Science or a closely related discipline is required at the time of appointment. Candidates for senior positions must have an established track record of outstanding teaching at universities and previous research experience.

Main Teaching Needs

We are particularly interested in candidates who have expertise in teaching some of the following courses: Introduction to Computer Science, Foundation of Computer Programming, Java Programming, Object-Oriented Design and Programming, Databases, Data Structures, Computer Networks, Operating Systems, Compiler Design & Principles, Software Engineering, etc.

Candidates may be requested to teach other courses than those listed above.

Applications

To apply, please provide a cover letter identifying the primary areas of teaching, previous teaching experiences (including subjects, number of students, student feedback, etc), curriculum vitae, teaching statements, previous research achievements, and arrange for at least three recommendation letters, all forward to cshire@sustc.edu.cn.



Collegiate Faculty Position – Dept. of Computer Science

The Department of Computer Science at Virginia Tech (www.cs.vt.edu) seeks applicants for a collegiate faculty position at the assistant professor level. Candidates must have a Ph.D. in computer science or related field at the time of appointment. Collegiate faculty members have a primary commitment to the instructional mission of the department, including graduate and undergraduate teaching, curricular and program development, and the design and integration of innovative and inclusive pedagogy. Successful candidates should give evidence of potential to take a lead role in enhancing curricula and promoting teaching excellence. In addition to teaching, candidates will be expected to participate in research and scholarship, whether on teaching and learning or on other computer science research topics of interest. Candidates will have the opportunity to collaborate with a wide range of research groups in the department, including a thriving group in CS education research. Preference will be given to candidates with demonstrated knowledge of CS education research topics such as education-related software systems, analysis of student data analytics, CS education for non-majors or at the K-12 level, cybersecurity education, distance education, or encouraging diversity in CS.

The department has 39 tenured and tenure-track faculty, over 700 undergraduate majors, 68 MS students, and 181 PhD students. The department is in the College of Engineering, whose undergraduate program ranks 15th and graduate program ranks 21st among U.S. engineering schools (*USN&WR*, 2015). The department plays a central role in several university-wide initiatives (see provost.vt.edu/destination-areas). Successful candidates will have the opportunity to participate in new multidisciplinary programs and curricula in Data Analytics and Decision Sciences, Integrated Security, Intelligent Infrastructure for Human-Centered Communities, and Creative Technologies and Experiences.

The collegiate faculty rank is a non-tenure-track position that offers a clear promotion path with increasingly long-term contracts. Collegiate faculty are full members of the department faculty, and are expected to participate in sponsored research, mentor graduate students, participate in department and professional service, etc. This position is located at the main campus in Blacksburg, VA, a region consistently ranked among the country's best places to live. The position requires occasional travel to professional meetings. Successful candidates should give evidence of sensitivity to issues of diversity in a campus community. Virginia Tech is committed to building a culturally diverse faculty and strongly encourages applications from women and minorities. Selected candidates must pass a criminal background check prior to employment.

Applications must be submitted online to jobs.vt.edu for posting **TR0160133**. Applicant screening will begin on December 10, 2016 and continue until the position is filled. Inquiries should be directed to Dr. Cliff Shaffer, Search Committee Chair, shaffer@cs.vt.edu.

Virginia Tech is an Equal Opportunity/Affirmative Action Employer



ADVERTISING IN CAREER OPPORTUNITIES

How to Submit a Classified Line Ad: Send an e-mail to acmm mediasales@acm.org. Please include text, and indicate the issue/or issues where the ad will appear, and a contact name and number.

Estimates: An insertion order will then be e-mailed back to you. The ad will be typeset according to CACM guidelines. NO PROOFS can be sent. Classified line ads are NOT commissionable.

Rates: \$325.00 for six lines of text, 40 characters per line. \$32.50 for each additional line after the first six. The MINIMUM is six lines.

Deadlines: 20th of the month/2 months prior to issue date. For latest deadline info, please contact:

acmm mediasales@acm.org

Career Opportunities Online: Classified and recruitment display ads receive a free duplicate listing on our website at <http://jobs.acm.org>

Ads are listed for a period of 30 days.

For More Information Contact:

**ACM Media Sales
at 212-626-0686 or
acmm mediasales@acm.org**

Temple University
**Data Science Faculty Positions (Assistant/
Associate/Full Professor)**
**Department of Computer and Information
Sciences**

The Department of Computer and Information Sciences (CIS) invites applications for multiple tenured and tenure track data science faculty. We are interested in data scientists who are developing methods and systems to collect, process, and analyze large amounts of data, as well as techniques to extract and visualize knowledge for various application domains. Successful candidates will contribute to the newly introduced undergraduate and graduate programs in data science and join world-class investigators in the Center for Data Analytics and Biomedical Informatics, Center for Networked Computing, Institute for Computational Molecular Sciences, and Institute for Genomics and Evolutionary Medicine, co-located with the CIS Department in a state-of-the-art \$140M building. Applicants are expected to have an outstanding research track record and are expected to demonstrate excellence in teaching.

Recently celebrating its 50th year anniversary, CIS is one of the oldest computer science departments in the country and is experiencing growth both in research and academic programs. Over the last seven years, the CIS annual research revenue has grown more than 300%, and both the undergraduate and graduate student enrollment has doubled. CIS is among the top 60 computer science departments in the country based on the

ARWU Shanghai ranking. CIS offers a highly competitive salary and start-up package, low teaching load, and institutional support for an active doctoral program.

Temple University is a Carnegie R1 (highest research activity) institution that serves more than 39,000 students and places among the top 55 public schools by the U.S. News ranking. It is the 26th largest university and 6th largest provider of graduate school education in the United States. Located in the heart of Philadelphia (the 5th largest city in U.S. known for its arts, culture and history, and affordable living), Temple University is in close proximity to many outstanding academic institutions, research centers, and industry partners in data science areas, including information technology, health care, biotechnology, and finance.

Applications should be submitted electronically at <https://academicjobsonline.org/ajo/jobs/8172>. Submitted materials should include a curriculum vitae, description of research accomplishments, research statement, teaching statement, and three letters of recommendation (names of referees are sufficient for senior-level candidates). Review of candidates will begin immediately and continue until the positions are filled. Applicants are encouraged to meet the deadline for priority consideration. For more information, send an email to zoran.obradovic@temple.edu with "TT Position" as the subject. Temple University is an affirmative action/equal opportunity employer with a strong commitment to the quality of faculty life.

**The College at Brockport, State
University of New York**
Department of Computing Sciences

Applications are invited for a tenure track Assistant Professor in the Department of Computing Sciences (home of ABET accredited programs) beginning Fall 2017. Doctoral degree in Computer Science, or in a closely related field with master's level expertise in Computer Science is required. ABD candidates considered. Hired faculty will be expected to teach, engage in research, and participate in service appropriate to rank. Preference will be given to candidates with expertise in cybersecurity, forensic computing, software security, computer networks, operating systems and computer architecture. Apply online at www.brockportrecruit.org by January 15, 2017. EOE/AA employer: M/F/DIS/VET

The Ohio State University

The Computer Science and Engineering Department at The Ohio State University seeks to fill multiple tenure-track positions at the assistant professor level. We are particularly interested in the following areas: (1) machine learning; (2) information visualization and visual analytics; (3) cybersecurity and privacy. Two positions are available in cybersecurity and privacy, including network/mobile security, systems security, security of cyber-physical systems, cryptography, data privacy, secure data management, privacy-preserving analytics, and data provenance. Highly qualified applicants at more senior levels will also be considered.

The department is committed to enhancing faculty diversity; women, minorities, and individuals with disabilities are especially encouraged to apply.

Some of these positions are partially funded by the university-wide Discovery Themes Initiative, a significant investment in key thematic areas, including the Data Analytics Collaborative which will establish a singular presence in data analytics at Ohio State. The university is responsive to dual-career families and strongly promotes work-life balance through a suite of institutionalized policies.

Applicants should hold or be completing a PhD in computer science & engineering or a closely related field, have a commitment to and demonstrated record of excellence in research, and a commitment to excellence in teaching.

To apply, please submit your application via the online database. The link can be found at:

<https://web.cse.ohio-state.edu/cgi-bin/portal/fsearch/apply.cgi>

Review of applications will begin on December 1, 2016 and will continue until the positions are filled.

The Ohio State University is an Equal Opportunity/Affirmative Action Employer.

Truman State University
**Non-Tenure Track Faculty Position in Computer
Science**

Truman State University has full-time non-tenure track computer science faculty positions beginning January 2017.

Responsibilities are to teach undergraduate computer science courses.



Faculty Positions – Department of Computer Science

The Department of Computer Science at Virginia Tech (www.cs.vt.edu) seeks applicants for multiple faculty positions at the assistant or associate professor levels, with research interests in all areas of computer science. Candidates must have a Ph.D. in computer science or related field at the time of appointment and a rank-appropriate record of scholarship and collaboration in computing research, broadly defined. Candidates should give evidence of sensitivity to issues of diversity and inclusion, and will be expected to teach graduate and undergraduate courses, mentor graduate students, and develop a high quality research program.

The department has 39 tenured/tenure-track faculty, over 700 majors (17% women), 68 MS students, and 181 PhD students. Departmental research expenditures last year were \$14 million. Candidates will have the opportunity to work with a wide range of productive research groups including computational biology and bioinformatics, computational science, computer science education, cybersecurity, data analytics and machine learning, human-computer interaction, software engineering and computer systems. Multidisciplinary research centers led by CS faculty include Center for Human-Computer Interaction (www.hci.vt.edu), Discovery Analytics Center (dac.cs.vt.edu), stack@cs (stack.cs.vt.edu), Synergistic Environments for Experimental Computing (research.cs.vt.edu/seec), Biocomplexity Institute (www.bi.vt.edu), and Institute for Creativity, Arts, and Technology (icat.vt.edu). The department also plays a central role in several university-wide research and teaching initiatives (provost.vt.edu/destination-areas), including Data Analytics and Decision Sciences, Integrated Security, Intelligent Infrastructure for Human-Centered Communities, and Creative Technologies and Experiences. The department is in the College of Engineering, whose undergraduate program ranks 15th and graduate program ranks 21st among U.S. engineering schools (*USN&WR*, 2015).

These faculty positions are located at the main campus in Blacksburg, VA, in a region that is consistently ranked among the country's best places to live. The positions require occasional travel to professional meetings. Virginia Tech is committed to building a culturally diverse community and strongly encourages women and minorities to apply. Candidates must pass a criminal background check prior to employment.

Submit applications to jobs.vt.edu, posting #TR0160116. Screening begins November 20, 2016 and continues until positions are filled. Direct inquiries to Dr. Ali Butt, Search Committee Chair, butta@cs.vt.edu.

Virginia Tech is an Equal Opportunity/Affirmative Action Employer

ACM Transactions on Reconfigurable Technology and Systems



This quarterly publication is a peer-reviewed and archival journal that covers reconfigurable technology, systems, and applications on reconfigurable computers. Topics include all levels of reconfigurable system abstractions and all aspects of reconfigurable technology including platforms, programming environments and application successes.

www.acm.org/trets
www.acm.org/subscribe



Association for
Computing Machinery

PhD (for Assistant Professor rank) or Masters (for Instructor rank) in CS or a closely related field is required, as are excellent English communication and interpersonal skills.

Opportunities at Tufts University Assistant/Associate/full Professor – General Search

Tenure-track faculty appointment at any rank to begin in September 2017. We are looking for an engaged and engaging researcher and teacher with a strong vision who can build and maintain a high-quality research program and collaborate with our faculty.

Candidates may have expertise in any area of computer science, especially but not limited to AI, data science, databases, human-computer interaction, machine learning, physical computing, robotics, security and privacy, systems, theoretical computer science, as well as research involving interdisciplinary collaborations and applications in science, engineering, and the humanities. We seek a candidate whose research will both connect with some of our current faculty and extend into new areas, but we are open to strong applicants in all areas.

Please submit your application online at <https://apply.interfolio.com/37872>. Email inquiries to ttsearch@cs.tufts.edu.

Assistant/Associate/full Professor – Robotics

Tenure-track faculty appointment at any rank in the area of robotics to begin in September 2017. This is a joint search between Computer Science (CS) and Mechanical Engineering (ME). We are looking for individuals with expertise in research and teaching, and a strong vision who can bridge the research strands in control and navigation in ME and artificial intelligence and human-robot interaction in CS.

Candidates may have expertise in any area of AI and robotics, especially but not limited to manipulation, control, and human-robot interaction.

Please submit your application online at <https://apply.interfolio.com/38832>.

Email inquiries to jointsearch@cs.tufts.edu.

Lecturer/Professor of the Practice

Full-time Lecturer or Professor of the Practice position to begin in 2017. The ideal Lecturer candidate will possess prior teaching experience, especially with undergraduate students, and have an engaging teaching style. The candidate must have a strong commitment to undergraduate teaching and mentoring. A good Professor of the Practice (POP) candidate will in addition possess a wealth of industry experience to share with our students. In the case of this position, we would consider a part-time hire if the finalist wished to maintain part-time employment in industry.

The initial appointment of either a Lecturer or POP will be for three years with the possibility of renewal.

Please submit your application online at <https://apply.interfolio.com/36571>. Email inquiries to cssearch@cs.tufts.edu.

More information on all searches is available at <http://www.cs.tufts.edu/jobs/employment-opportunities.html>.

Tufts University is an Equal Opportunity/ Affirmative Action Employer. We are committed to increasing the diversity of our faculty and fostering their success when hired. Members of under-represented groups are welcome and strongly encouraged to apply.

The University of California at Santa Barbara

The Electrical and Computer Engineering Department at UCSB invites applications for a tenure-track faculty position in the area of Computer Engineering with an effective appointment date of July 1, 2017.

Please visit:

<http://www.ece.ucsb.edu/employment/>.

The University of California is an Equal Opportunity/Affirmative Action Employer and all qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, disability status, protected veteran status, or any other characteristic protected by law.

University of Maryland, Baltimore County Computer Science Assistant Professor

UMBC's Department of Computer Science and Electrical Engineering invites applications for two tenure-track Computer Science Assistant Professor positions to begin in Fall 2017. Exceptionally strong candidates for higher ranks may be considered. Applicants must have or be completing a Ph.D. in a relevant discipline, have demonstrated the ability to pursue a research program, and have a strong commitment to undergraduate and graduate teaching. Candidates will be expected to build and lead a team of student researchers, obtain external research support and teach both graduate and undergraduate courses.

All areas of specialization will be considered, but we are especially interested in candidates in information assurance and cybersecurity; mobile, wearable and IoT systems; big data with an emphasis on machine learning, analytics, and high-performance computing; and knowledge and database systems.

The CSEE department is vibrant, research-oriented and multi-disciplinary with programs in Computer Science, Computer Engineering, Electrical Engineering and Cybersecurity. We have more than 1500 undergraduate and 600 graduate students and have awarded more than 290 PhDs since our establishment in 1986. Our research is supported by a growing and diverse portfolio of government and industrial sponsors with over \$5M in yearly research expenditures.

UMBC is a dynamic public research university integrating teaching, research and service. As an Honors University, the campus offers academically talented students a strong undergraduate liberal arts foundation that prepares them for graduate and professional study, entry into the workforce, and community service and leadership. UMBC emphasizes science, engineering, information technology and public policy at the graduate level. The 2016 US News and World Report Best Colleges guide placed UMBC fifth in

the *Most Innovative National Universities* category and one of its top 20 national universities for *Best Undergraduate Teaching*. Our strategic location in the Baltimore-Washington corridor puts us close to many important federal laboratories and agencies and high-tech companies, facilitating interactions, collaboration, and opportunities for sabbaticals and visiting appointments.

Applicants should submit a cover letter, a brief statement of teaching and research experience and interests, a CV, and three letters of recommendation at <http://apply.interfolio.com/37418>. Applications received by January 15, 2017 are assured full consideration and those received later will be evaluated as long as the positions remain open. Send questions to jobsTT@csee.umbc.edu and see <http://csee.umbc.edu/jobs> for more information.

We are committed to inclusive excellence and innovation and welcome applications from women, minorities, veterans, and individuals with disabilities. UMBC is an Affirmative Action/Equal Opportunity Employer.

University of Illinois at Chicago Teaching Faculty – Computer Science

The Computer Science Department at the University of Illinois at Chicago is seeking full-time teaching faculty members for fall 2017. Two tracks are available for applicants, clinical track (Clinical Assistant Professor, Clinical Associate Professor and Clinical Professor) for those who have a PhD in Computer Science, and lecturer track (Lecturer and Senior Lecturer) for those who don't. The department is committed to effective teaching. Candidates would be working alongside 7 full-time teaching faculty with over 100 years of combined teaching experience and 11 awards for excellence. The department seeks candidates dedicated to teaching; candidates must have evidence of effective teaching, or present a convincing case of future dedication and success in the art of teaching. Content areas of interest include introductory programming/data structures, theory/algorithms, discrete math, computer organization/systems, databases, software design, and web development. Candidates interested in Computer Science Education research are especially encouraged to apply. The teaching load is 3 undergraduate courses per semester.

The University of Illinois at Chicago (UIC) is one of the top-10 most diverse universities in the US (US News and World Report). UIC's hometown of Chicago epitomizes the modern, livable, vibrant city. Located on the shore of Lake Michigan, Chicago offers an outstanding array of cultural and culinary experiences. As the birthplace of the modern skyscraper, Chicago boasts one of the world's tallest and densest skylines, combined with an 8100-acre park system and extensive public transit and biking networks.

Minimum qualifications include a Master's degree, for Lecturer, or PhD, for Clinical Assistant Professor, in Computer Science or a closely related field, and either (a) demonstrated evidence of effective teaching, or (b) convincing argument of future dedication and success in the art of teaching. Applications are submitted online at <https://jobs.uic.edu/>. In the online application, include a curriculum vitae, names and addresses of at least three references, a statement providing evidence of effective

teaching, and a statement describing your past experience in activities that promote diversity and inclusion (or plans to make future contributions). For additional information contact Professor Mitch Theys, Committee Chair, mtheys@uic.edu.

For fullest consideration, please apply by February 1, 2017. We will continue to accept applications until the position is filled. The University of Illinois is an Equal Opportunity, Affirmative Action employer. Minorities, women, veterans and individuals with disabilities are encouraged to apply. The University of Illinois conducts background checks on all job candidates upon acceptance of contingent offer of employment. Background checks will be performed in compliance with the Fair Credit Reporting Act.

University of Minnesota-Twin Cities

The Department of Computer Science and Engineering at the University of Minnesota-Twin Cities is hiring to fill one or more tenure-track positions at the assistant professor level, although higher levels of appointment may be considered when commensurate with experience. Candidates with expertise in any area of computer science and engineering will be considered, and those with research interests in social computing, human-computer interaction, and security are particularly encouraged to apply.

The Department of Computer Science and Engineering is fully committed to a culturally and academically diverse faculty; we welcome applications from individuals who will further expand that diversity. Candidates should have a Ph.D. in Computer Science or a closely related discipline at the time of appointment. Submit materials as described at <https://www.cs.umn.edu/resources/employment/faculty/>.

Consideration of complete applications will begin December 1, 2016, and continue until all positions are filled. The University of Minnesota is an equal opportunity employer and educator.

University of Nevada Reno

Three Tenure-Track Assistant Professor Positions, and One Lecturer Position, in Computer Science and Engineering at the University of Nevada, Reno.

The Department of Computer Science and Engineering at the University of Nevada, Reno, invites applications for three tenure-track faculty positions and one Lecturer position. Three of the positions are at the Assistant Professor level with expertise in the areas of (i) high performance computing, (ii) cybersecurity, and (iii) data science and engineering with emphasis on bioinformatics. The fourth position is at the Lecturer level.

Applicants for the Tenure Track Faculty positions must have a Ph.D. in Computer Science or Computer Engineering by July 1, 2017 and must be strongly committed to excellence in research and teaching and should demonstrate potential for developing robust externally funded research programs. Candidates for the Lecturer position, must be strongly committed to excellence in teaching and must have a M.S. or Ph.D. in Computer Science or Computer Engineering by July 1, 2017.

In the last five years, the College of Engineering has witnessed an unprecedented growth in student enrollment and number of faculty posi-

tions. The College is positioned to further enhance its growth of its students, faculty, staff, facilities as well as its research productivity and its graduate and undergraduate programs.

The University of Nevada, Reno recognizes that diversity promotes excellence in education and research. We are an inclusive and engaged community and recognize the added value that students, faculty, and staff from different backgrounds bring to the educational experience.

Interested candidates must apply online.

For the high performance computing position, apply to www.unrsearch.com/postings/22243

For the cybersecurity position, apply to www.unrsearch.com/postings/22183

For the data science and engineering position, apply to www.unrsearch.com/postings/22239

For the lecturer position, apply to www.unrsearch.com/postings/22187

Application process includes: a detailed letter of application which also indicates how you would contribute to the diversity and excellence of the academic community through your research, curriculum vitae, statement of teaching philosophy, statement of research and plans, (research and plans does not apply the lecturer position) and contact information for three professional references. Review of applications will begin on January 5, 2017 and will continue until the search closes on February 15, 2017. Inquiries should be directed to Ms. Lisa Cody, lcody@unr.edu.

The University of Nevada serves over 21,000 students. The university is ranked as a Tier 1 institution by "U.S. News and World Report" and offers an array of degree programs at all levels. Reno is located in the foothills of the Sierra Nevada, a 30-minute drive from Lake Tahoe. Reno/Tahoe is recognized as a world-class outdoor recreation area. Additional nearby areas of interest include the Black Rock Desert, Sacramento, Yosemite National Park, Napa/Sonoma, the San Francisco Bay Area.

EEO/AA Women and under-represented groups, individuals with disabilities, and veterans are encouraged to apply.

University of North Dakota

The Department of Computer Science at the University of North Dakota is recruiting for a tenure-track, nine-month position at the Assistant or Associate Professor level with an anticipated start date of August 16, 2017. The position requires a Ph.D. in Computer Science or a related field. The department is seeking outstanding candidates with a research focus in the following areas (but not limited to): (a) theoretical foundations of computer science and (b) cybersecurity.

Please direct your attention to <https://und.edu/finance-operations/human-resources-payroll/careers/secure/job-openings-external.cfm?category=2000+Academic> for the complete job announcement

The University of North Texas Department of Computer Science and Engineering Assistant/Associate Professor

The University of North Texas (UNT), a Tier 1 Research Institution (Carnegie Classification as a Doctoral University: Highest Research Activity),

invites applications for the following faculty positions in the Department of Computer Science and Engineering (CSE).

Tenured or Tenure Track Positions:

Position 1: Assistant or Associate Professor in **Computer Security**, including cryptography, cloud and network security, mobile security, intrusion detection, secure hardware and software systems, and vulnerability and threat analysis. This position is affiliated with the Center for Information and Computer Security (CICS).

Position 2: Assistant or Associate Professor in **Computational Life Science**, including computational epidemiology, computational health science, smart and connected health, geo-informatics, bioinformatics, health-informatics, and computational environmental science, with focus on developing computational methods, models, or simulations. This position is affiliated with the Center for Computational Epidemiology and Response Analysis (CeCERA).

Position 3: Assistant or Associate Professor in **Computer Engineering**, including real-time systems, embedded systems, cyber physical systems, memory and storage systems, computer architecture, VLSI, reconfigurable systems, resilient systems, low-power computing, runtime systems, distributed systems, and performance measurement and tuning. This position may be affiliated with the NSF Industry/University Cooperative Research Center for Net-centric and Cloud Software and Systems (NCSS).

The candidates are expected to teach CSE undergraduate and graduate courses, develop a strong research program funded by external sources, support and mentor graduate students, and provide service to the University and the profession. Minimum qualifications include an earned doctorate in computer science, computer engineering or a closely related field. For the Assistant Professor position, a strong publication record and the potential to succeed in securing research funding and mentoring graduate students are required. For the Associate Professor position, a sustained record of providing mentoring to junior faculty, advising graduate students, providing service to the University and profession, and securing external funding for research activities are also required. Post-doctoral research experience or industrial research experience is preferred. Offers of employment for these positions will be made dependent upon available funding.

The Computer Science and Engineering department is home to 14 Professors, 9 Associate Professors, 4 Assistant Professors, 6 Lecturers, almost 100 Ph.D. students, over 200 master students, and almost 1100 bachelor students. We offer a Ph.D. degree in Computer Science and Engineering, M.S. degrees in Computer Science and Computer Engineering, ABET-accredited B. S. degrees in Computer Science and Computer Engineering, and an ABET-accredited B. A. degree in Information Technology. Additional information about the department is available at the website: computerscience.engineering.unt.edu.

Application Procedure:

All applicants must apply online to: <https://facultyjobs.unt.edu>. Submit nominations and questions regarding the tenure track position in computer security (system identification number 6001336) to Dr. Ram Dantu (Ram.Dantu@unt.edu).

unt.edu), the tenure track position in computational life science (system identification number 6001335) to Dr. Armin Mikler (mikler@unt.edu), and the tenure track position in computer engineering (system identification number 6001342) to Dr. Song Fu (Song.Fu@unt.edu).

Application Deadline:

The committee will begin its review of applications on November 1, 2016 and continue to accept and review applications until the positions are filled.

The University:

As the nation's 24th largest public university and the largest, most comprehensive in the Dallas-Fort Worth area, UNT is dedicated to providing an excellent educational experience to its 38,000 students while powering the North Texas region, state and nation through innovative education and research. UNT is strategically located in Denton, Texas, a vibrant city with a lively arts and music culture.

The University of North Texas is an AA/ADA/EOE committed to diversity in its educational programs.

University of Oklahoma

Tenure Track Faculty Position in Cybersecurity

The School of Computer Science in the Gallogly College of Engineering at the University of Oklahoma is entering an exciting period of growth opportunities in cybersecurity and data science and analytics (DSA). We are seeking applications for a tenure track faculty position at the rank of an assistant professor who has demonstrated research skills in cybersecurity, especially at the intersection of cybersecurity and DSA. We offer B.S., M.S., and Ph.D. degree programs in computer science and the M.S. degree in DSA.

Applicants should hold a Ph.D. in computer science and should be committed to excellence in teaching and research. Candidates should have demonstrated potential for outstanding research in cybersecurity and have ability to teach courses in these areas at all levels, including advising M.S. and Ph.D. students.

The application package should include: 1) single page cover letter describing the motivation in pursuing this position, 2) curriculum vitae, 3) teaching and research statements, 4) 1-2 papers that best represent research contributions and scholarship, and 5) list of at least three references. Application packages should be submitted via soonerway.ou.edu. Inquiries about the position can be addressed to Professor Sridhar Radhakrishnan, chair of the search committee, at sridhar@ou.edu. We encourage applicants to apply by November 1, 2016, though application packages will be accepted until the position is filled.

The University of Oklahoma is a Carnegie-R1 comprehensive public research university known for excellence in teaching, research, and community engagement. OU enrolls over 30,000 students and has more than 2,700 full-time faculty members. Norman is a culturally rich and vibrant town located in the Oklahoma City metro area.

The University of Oklahoma, in compliance with all applicable federal and state laws and regulations does not discriminate on the basis of race, color, national origin, sex, sexual

orientation, genetic information, gender identity, gender expression, age, religion, disability, political beliefs, or status as a veteran in any of its policies, practices, or procedures. The University of Oklahoma, recognizing its obligation to guarantee equal opportunity to all persons in all segments of University life, reaffirms its commitment to the continuation and expansion of positive programs which reinforce and strengthen its affirmative action policies. This commitment stems not only from compliance with federal and state equal opportunity laws but from a desire to ensure social justice and promote campus diversity. Our commitment to the concept of affirmative action requires sincere and cooperative efforts throughout all levels of our employment structure. We will continue to strive to reach the goals of fair and equal employment opportunities for all.

University of San Francisco

Assistant Professor Full-time Renewable

The Department of Computer Science at the University of San Francisco is accepting applications for a full-time, renewable, non-tenure track Assistant Professor position starting in August 2017. See <https://www.usfjobs.com/postings/11491> for the full job description and application instructions.

To receive full consideration applications must be received by January 2, 2017.

University of San Francisco

Assistant Professor Tenure Track

The Department of Computer Science at the University of San Francisco is accepting applications for a tenure-track Assistant Professor position starting in August 2017. See <https://www.usfjobs.com/postings/11490> for the full job description and application instructions.

To receive full consideration applications must be complete and submitted by January 2, 2017.

University of Texas at Austin

The Department of Computer Science of the University of Texas at Austin invites applications for tenured and tenure-track positions at all levels. Outstanding candidates in all areas of Computer Science will be considered, particularly in: computer vision, machine learning, cybersecurity and formal methods.

The department is ranked among the top ten computer science departments in the country. It has 44 tenured and tenure-track faculty members across all areas of computer science. Many of these faculty participate in interdisciplinary programs and centers in the University, including the Texas Advanced Computing Center (TACC), and those in Computational and Applied Mathematics, Computational Biology, and Neuroscience.

Austin, the capital of Texas, is a center for high-technology industry, including companies such as Dell, IBM, Advanced Micro Devices, 3M Corporation, National Instruments, Apple Computer, Inc., AT&T and Samsung. For more information please see the department web page <http://www.cs.utexas.edu>.

University of Wisconsin-Madison
Professor or Associate Professor
or Assistant Professor

The Department of Computer Sciences at the University of Wisconsin-Madison has embarked on a multi-year effort to significantly expand the strengths of the department. As part of this endeavor we invite applications from highly qualified candidates in all areas of computer science at the tenure-track and tenured faculty levels. Appointments will begin August 2017. We seek applicants across the spectrum of computer science research, ranging from core areas to applied and interdisciplinary research. Preference at the full professor level will be given to candidates active in the datafication of human knowledge who are potentially eligible for the prestigious DeWitt Chair.

Applicants must have a Ph.D. in Computer Science or in a closely related field prior to the start of the appointment. Successful candidates will show potential for developing an outstanding and highly visible scholarly research program, as well as excelling in undergraduate and graduate teaching. Application materials must be submitted online to: <https://academicjobsonline.org/ajo/jobs/7638>. Questions may be directed to: recruiting@cs.wisc.edu.

The Department of Computer Sciences is among the oldest and top ranked CS departments in the world, renowned for its groundbreaking research in computer architecture, database systems, distributed and grid computing, nonlinear optimization, and many other areas of computer science. It has a strong track record of building large scale compute infrastructures, and engaging in new interdisciplinary programs and centers. The CS department has also been the beneficiary of significant donations—amounting to nearly \$20 million over the past year alone—aimed at supporting the department's research and educational mission.

Madison, capital of the state of Wisconsin and home to the University of Wisconsin-Madison, consistently ranks as one of the greenest, most bike friendly, and most family friendly cities in the nation. With the headquarters of medical software company Epic, Microsoft's Jim Gray Systems Lab, Google, Amazon, and several other technology giants in the area, Madison is also among the fastest growing technology hubs in the country.

UW-Madison is an equal opportunity/affirmative action employer. We promote excellence through diversity and encourage all qualified individuals to apply.

University of Wisconsin – Stevens Point
Department of Computing
and New Media Technologies
Assistant/Associate Professor

The Department of Computing and New Media Technologies at UW-Stevens Point invites applications for a tenure-line faculty position, starting August 2017. We are particularly interested in candidates with skills in:

- ▶ Application development
- ▶ Web/mobile technologies
- ▶ Data science and management
- ▶ High performance computing

Required Masters or Ph.D. preferred in computer science or related field. There may be opportunity to teach graduate courses in an online Master's program. See <http://www.uwsp.edu/hr/jobs/Pages/default.aspx> for application instructions and additional information on our department.

Screening of applicants will begin on January 3, 2017 and continue until the position is filled. UW-Stevens Point is an Affirmative Action/Equal Opportunity Employer.

Worcester Polytechnic Institute

WPI invites applications for a tenure track (open rank) Professor positions with a research focus in Computer Science to begin in the Fall of 2017.

The Computer Science Department anticipates hiring multiple tenure-track faculty for the Fall of 2017 whose expertise is in the following areas: 1) User Interfaces and Visualization; 2) Mobile and Sensor Software Systems; and 3) Dependable Software Systems; 4) Data Science in areas such as machine learning, large-scale data management, and big data analytics, joining faculty in our interdisciplinary Data Science program; 5) Robotics, joining faculty in our interdisciplinary Robotics Engineering program; and 6) Interactive Media and Game Development, joining faculty in our interdisciplinary Interactive Media & Game Development program.

Candidates should have a PhD in Computer Science or a closely related field, and the potential for excellence in research and teaching.

Founded in 1865, WPI is one of the nation's first technological universities. A highly selective private university located within an hour of Boston, WPI is consistently ranked among the top 60 research institutions by US News & World Report. The university is home to an innovative and intensive project-based curriculum that empowers students with the knowledge and skills to address real world problems around the globe, an approach repeatedly cited for excellence by The Fiske Guide to Colleges and The Princeton Review.

Questions about the hiring process should be sent to recruit@cs.wpi.edu.

More information about the positions and instructions for applying are available at <http://apptkr.com/880744>.

You will need to include detailed research and teaching statements, vitae and contact information for at least three references.

We are an Equal Opportunity Employer and do not discriminate against applicants due to race, color, age, religion, sex, sexual orientation, gender identity, national origin, veteran status or disability. We are looking for individuals who value creativity, diversity, inclusion, and collaboration.

Worcester Polytechnic Institute
Robotics Engineering Program
Tenure Track Faculty Position – August 2017

The WPI Robotics Engineering Program invites applications for a tenure-track faculty position at all levels for fall 2017. You should hold a Ph.D. in engineering, computer science, or a closely related field, possess the desire to actively engage in teaching robotics at both the graduate and under-

graduate levels, and have a clear plan to establish a vibrant, externally funded research program.

Research areas of interest include, but are not limited to: Fundamentals (e.g. Controls, Learning, Perception), Capabilities (e.g. Human-Robot Interaction, Manipulation, Mobility), and Applications (e.g. Exploration, Health, Manufacturing).

Now in its 10th year, the WPI Robotics Engineering program has been recognized for its unique multi-disciplinary approach to robotics education and its cutting-edge robotics research. We offer the Nation's first undergraduate degree in Robotics, as well as the M.S. and Ph.D. degrees, with over 340 undergraduate and over 170 graduate students enrolled. The 12 core and 20 affiliated robotics faculty members have research support from NSF, NIH, AFRL, ONR, NASA, DARPA, and industry. The exceptionally close-knit, collaborative, and collegial faculty routinely engages in joint proposals and research, shared labs and facilities, mentoring, and teaching guidance.

Please refer to our website at <http://robotics.wpi.edu/> for more information.

Founded in 1865, WPI is one of the nation's oldest technological universities. Today, WPI is a highly selective private university with an undergraduate student body of over 4,000 and 1,900 full- and part-time graduate students enrolled in more than 50 Bachelor's, Master's and Ph.D. programs. WPI is consistently ranked among the top national universities in U.S. News & World Report. Our innovative project-enriched curriculum engages students and faculty in real-world problem solving, often at one of WPI's project centers. The university is located in the heart of Massachusetts an hour away from Boston and Providence.

Questions about the hiring process should be sent to robotics-recruit@wpi.edu. Applications should be submitted per instructions at <http://apptkr.com/883860>. Please include detailed research and teaching statements, vitae, and contact information for at least three references. Applications should be received by December 16, 2016. Applications will be reviewed as they are received; therefore, applicants are encouraged to apply early.

To enrich education through diversity, WPI is an affirmative action, equal opportunity employer.

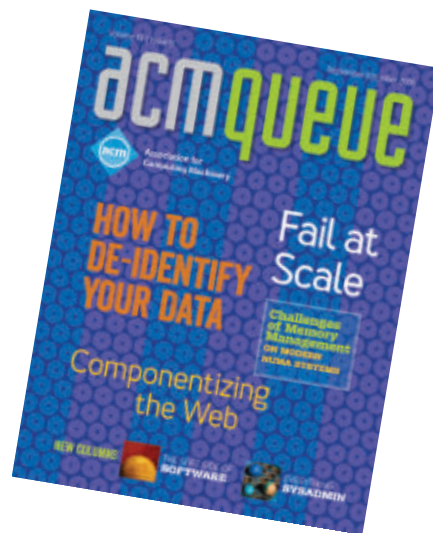
acmqueue

Check out the new acmqueue app

FREE TO ACM MEMBERS

acmqueue is ACM's magazine by and for practitioners, bridging the gap between academics and practitioners of computer science. After more than a decade of providing unique perspectives on how current and emerging technologies are being applied in the field, the new acmqueue has evolved into an interactive, socially networked, electronic magazine.

Broaden your knowledge with technical articles focusing on today's problems affecting CS in practice, video interviews, roundtables, case studies, and lively columns.



Keep up with this fast-paced world on the go. Download the mobile app.



Association for
Computing Machinery

Desktop digital edition also available at queue.acm.org.
Bimonthly issues free to ACM Professional Members.
Annual subscription \$19.99 for nonmembers.

[CONTINUED FROM P. 144] Projects Agency) to develop vision capabilities for Autonomous Land Vehicles, or ALVs. But the arrangement didn't work very well, because we didn't have a real car that gave us real input in real time. Instead, we were given video, and asked to analyze whether we could recognize this or that. But recognition is relative to what the car does. Recognition and action and the way that input changes make a loop, and you have to understand that whole loop yourself.

So you began building your own ALVs, and later, cars.

In the beginning, the work was on natural terrain, but that's harder to do in the city. So we began to use a pathway in Schenley Park, and our interests shifted to path and road following.

At one point, I understand that a tree presented something of an obstacle.

I often use the story. One of the trees in the park accidentally aligned with a path when seen from particular direction. Of course, in a three-dimensional world, the tree edge is a vertical edge and the road edge is a horizontal edge. But as a two-dimensional picture, they appeared aligned, as if they were a single object. At that time, we didn't have a good understanding of how to deal with appearance vs. reality, so we just tried to see whether a pair of edges created the appearance of a road boundary, like parallel lines.

And the edges of this particular tree trunk fooled the system.

Yes, whenever the car went the wrong direction, we had to push the "kill" button. We called it the killer tree. We once asked students how we could avoid this problem, and the best answer was, "Call up the Pittsburgh Parks Department and ask them to cut the tree."

What are you working on now?

One project I'm working on with Professor Srinivasa Narasimhan is smart headlights. When you're driving at night or on a rainy or snowy day, the raindrops or snowflakes appear as bright dots in your visible scene, because they are very reflective. But if you replace the headlight with a projector, you can turn the individual rays on and off. Then, by recognizing where the raindrops are, you

can turn off just the rays that will hit the raindrops so that you would not be able to see them, and could drive as if it's a fine day. I had this idea a long time ago, but could not do it. Professor Narasimhan's team made this magic real.

That sounds amazing.

The real use of the raindrop control thing may be still a little ways off. But our system can recognize oncoming cars and turn off just the rays that would hit driver's eyes, so that you don't have to switch between your high and low beams. This is what I call augmented reality. What people call "augmented reality" are often applications where a smartphone or a head-mount display appends additional information to the object that you're viewing. But that's not augmented reality. That's the augmented *display* of reality. With smart headlights, reality is indeed augmented, because to your naked eye, it appears to be more informative.

You're also involved with a number of Quality of Life Technology initiatives that are focused on developing systems that enable people with disabilities to live more independently.

The work has given me a different perspective on robotics. In autonomous systems, the implicit goal is to reduce or even eliminate human involvement. That's natural for applications like space and defense, where it's either difficult, undesirable, or impossible for humans to go. In Quality of Life robotics, people are a part of the system, and people want to be helped only when they need it; otherwise, they want to do things themselves. That is probably the most important aspect that we have to respect.

So your goal is to figure out how to increase human involvement.

In order for the system to work, we have to understand each component. I often say that people are the weakest link, because they are the least understood. In cars, for example, we have worked so hard to understand mechanical factors like tire traction, air pressure and resistance, and so on. But we don't understand the driver, and that's why I think we still have a lot of accidents. 

Leah Hoffmann is a technology writer based in Piermont, NY.

© 2016 ACM 0001-0782/16/12 \$15.00

INTERACTIONS



ACM's *Interactions* magazine explores critical relationships between people and technology, showcasing emerging innovations and industry leaders from around the world across important applications of design thinking and the broadening field of interaction design.

Our readers represent a growing community of practice that is of increasing and vital global importance.



To learn more about us, visit our award-winning website <http://interactions.acm.org>

Follow us on Facebook and Twitter



To subscribe: <http://www.acm.org/subscribe>

Association for
Computing Machinery



Q&A

A View to the Future

Takeo Kanade, recipient of the 2016 Kyoto Prize for Advanced Technology for pioneering contributions to computer vision and robotics, sees respect as an important part of Quality of Life Technology initiatives.

TAKEO KANADE HAS spent his career trying to help computers understand what is around them. The robotics pioneer joined the faculty of Carnegie Mellon University (CMU) in 1980, and his computer vision research underscores everything from medical robotics technology to self-driving cars. This year, to his long list of accolades, he adds the Kyoto Prize.

You've been working in computer vision since you were a graduate student at Kyoto University. What drew you to the field?

I did my undergraduate work in electrical engineering in the late 1960s, which was the dawn of computer science in Japan. There was no computer science department at that time, but I was impressed with a talk I heard from a professor who was using computers for what he called non-numerical computation. Nowadays, we might call it artificial intelligence—it was very pioneering work in speech recognition, image processing, and translation.

For your Ph.D. thesis, you developed the world's first complete face recognition system.

You said the first “complete.” Indeed, that is the word—it was a real system, including everything from the very first step digitizer to an image-processing program, all the way to a display device that I made from a Sony TV. It was written in assembler language. You know, in baseball, a complete game is when you pitch from the first to the ninth inning. In a perfect game, you



“I see the advantage of learning from biological systems, now that we understand them a little better and have the tools to recreate them.”

don't give out any hits. My Ph.D. thesis program was complete, but not perfect; that's why I'm still working on it.

To create this system, you drew inspiration from origami and geometry, rather than from human vision.

I was not against learning from hu-

mans. But I didn't feel like that the human body is an optimum system. Bodies have to deal with everything from seeing to eating to surviving. But when you isolate a particular capability, there should be a better way to do it. Moreover, at that time, we didn't know much about how the body works. I'm a practical person, so I thought that by understanding the underlying physical, geometrical, and statistical principles, we should be able to design a good system. These days, I see the advantage of learning from biological systems, now that we understand them a little better and have the tools to recreate them.

In 1980, after several years of teaching at Kyoto University, you moved to the U.S. to become a professor at Carnegie Mellon University. Why?

I felt it was a kind of intellectual adventure. At that time, computer science was very different in the U.S. than in Japan—the environment, the way people talked ... I visited CMU in 1977, and I still remember seeing Nobel Prize winner Herbert Simon in the terminal room at midnight, talking with students while he waited for his results from the printer. It was out of my imagination.

In the mid-1980s, you began your pioneering research into the field of autonomous vehicles. What sparked your interest?

It was something of a historical accident. We had a fairly big contract from DARPA (the U.S. Defense Advanced Research [CONTINUED ON P. 143]

Computing Reviews

Connect with our Community of Reviewers

“I like CR because it covers the full spectrum of computing research, beyond the comfort zone of one’s specialty. I always look forward to the next Editor’s Pick to get a new perspective.”

- Alessandro Berni



Association for
Computing Machinery

ThinkLoud

www.computingreviews.com

MobileHCI 2017



Time to move on, up, in and out!

4th - 7th September, Vienna

19th International Conference on Human-Computer Interaction
with Mobile Devices and Services

Full and Short Papers

Submissions - February 9th

Workshops

Submissions - February 23rd

Tutorials

Submissions - May 11th

Late Breaking Work

Submissions - May 11th

Demos

Submissions - May 11th

Industrial Perspectives

Submissions - May 11th

Doctoral Consortium

Submissions - May 11th

Student Volunteers

Application starts - April 27th

The Mobile HCI Conference Series has shaped research, development and practice in mobile devices and services for nearly two decades. In 2017, the Conference will forge a set of new agendas for the decades to come.

With the coming of driverless cars; drones; wearables and implantables; and an ecology of embedded devices and services in the everyday environment, there has never been a more exciting and pressing time to debate and explore what digital mobility means.

General Conference Chairs: Matt Jones (Swansea University) and Manfred Tscheligi (University of Salzburg & AIT).

mobilehci.acm.org



SIGCHI
special interest group computer human interaction



Center for
Human-Computer Interaction
University of Salzburg