

COMMUNICATIONS

CACM.ACM.ORG

OF THE

ACM

12/2015 VOL.58 NO.12

What Makes Paris Look Like Paris?



Personalizing Maps
How to De-Identify Data
Internet Use and
Psychological Well-Being

An ACM-W Celebration!



Delta City Centre
Ottawa, ON
Canada



JAN 22 - 23
2016



Canadian Celebration of Women in Computing 2016

Come celebrate with us at the largest gathering of Women in Computing in Canada!
The conference will feature prominent keynote speakers, panels, workshops,
presentations and posters, as well as a large career fair.

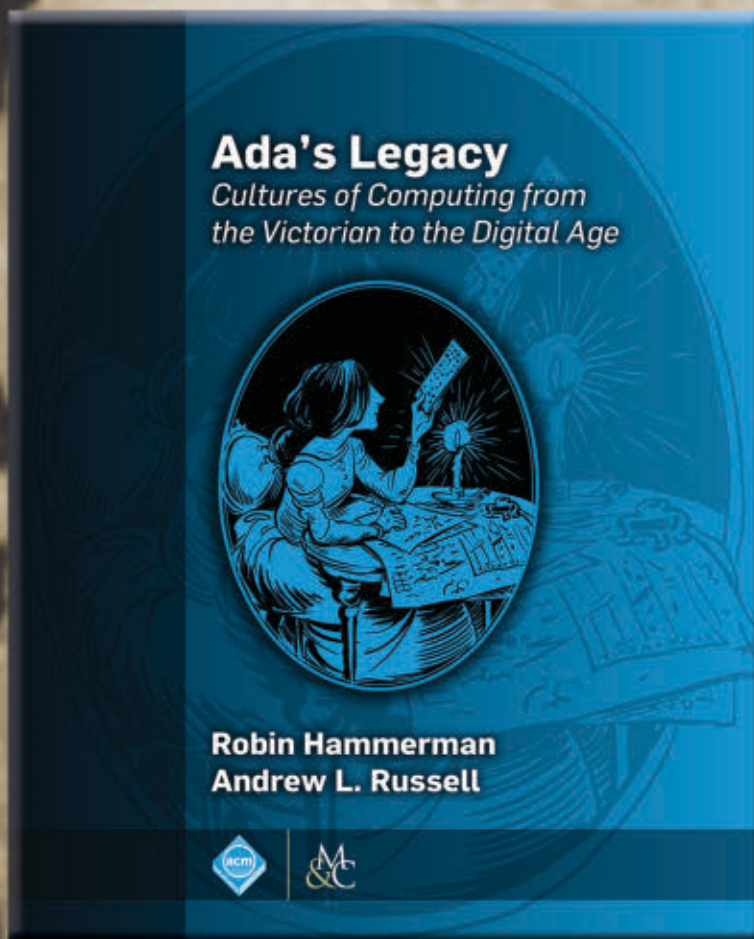
Register today!

www.can-cwic.ca

Robin Hammerman and Andrew L. Russell

Ada's Legacy

Cultures of Computing from the
Victorian to the Digital Age



INSPIRING MINDS FOR 200 YEARS

Ada's Legacy illustrates the depth and diversity of writers, things, and makers who have been inspired by Ada Lovelace, the English mathematician and writer.

The volume commemorates the bicentennial of Ada's birth in December 1815, celebrating her many achievements as well as the impact of her work which reverberated widely since the late 19th century. This is a unique contribution to a resurgence in Lovelace scholarship, thanks to the expanding influence of women in science, technology, engineering and mathematics.

ACM Books is a new series of high quality books for the computer science community, published by the Association for Computing Machinery with Morgan & Claypool Publishers.



ISBN: 978-1-97000-148-8 DOI: 10.1145/2809523

<http://books.acm.org>

<http://www.morganclaypoolpublishers.com/acm>

Departments

- 5 **Editor's Letter**
On Lethal Autonomous Weapons
By Moshe Y. Vardi
-
- 7 **Cerf's Up**
Advancing the ACM Agenda
By Vinton G. Cerf
-
- 8 **Letters to the Editor**
What About Statistical Relational Learning?
-
- 10 **BLOG@CACM**
What Do We Do When the Jobs Are Gone, and Why We Must Embrace Active Learning
Moshe Y. Vardi ponders the outlook for people when all work is automated, while Mark Guzdial emphasizes the importance of active learning in teaching computer science.
-
- 31 **Calendar**
-
- 122 **Careers**

Last Byte

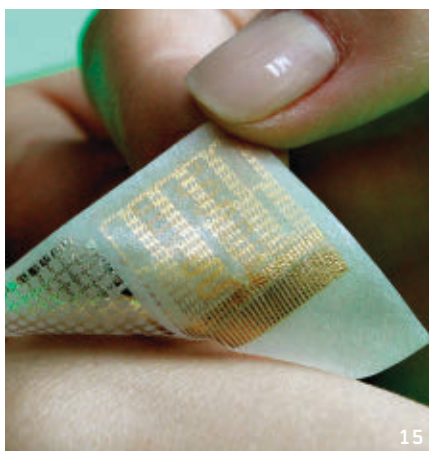
- 136 **Q&A**
Redefining Architectures
Mary Jane Irwin on building advanced circuits, special processors, and a hardware description language, while advocating for women in computer science.
By Leah Hoffmann



About the Cover: Is it possible for a computer to distinguish a major city by its visual essence? Through the use of Google Street View imagery, this month's cover story illustrates how the identifying look of a city does not rely on famous landmarks but rather stylistic elements of daily life. Cover collage by Iwona Usakiewicz/

Andrij Borys Associates/Shutterstock.

News



- 12 **When Data Is Not Enough**
Reproducibility of code is increasingly crucial to verifying scientific claims.
By Don Monroe
-
- 15 **The Hyper-Intelligent Bandage**
Scientists are developing smart, sensor-packed dressings to help heal chronic wounds.
By Gregory Mone
-
- 17 **Technology Brings Online Education in Line with Campus Programs**
Whether sitting in front of a screen or in a classroom, online and campus-based institutions want to verify students actually attend classes, take exams.
By Keith Kirkpatrick

Viewpoints

- 20 **Historical Reflections**
The Digital Dark Age
...and why it will have to wait.
By David Anderson

Viewpoints

- 24 **The Profession of IT**
Why Our Theories of Innovation Fail Us
Until we moderate our fascination with creating ideas, we will not achieve the rate of innovations we seek.
By Peter J. Denning and Nicholas Dew
-
- 27 **Computing Ethics**
Coupled Ethical-Epistemic Analysis in Teaching Ethics
Critical reflection on value choices.
By Nancy Tuana
-
- 30 **Kode Vicious**
Pickled Patches
On repositories of patches and tension between security professionals and in-house developers.
By George V. Neville-Neil
-
- 33 **Broadening Participation**
Increasing the Participation of Individuals with Disabilities in Computing
Lessons learned from a decade of practice.
By Richard E. Ladner and Sheryl Burgstahler
-
- 37 **Viewpoint**
Creating a New Generation of Computational Thinkers
Experiences with a successful school program in Scotland.
By Jeremy Scott and Alan Bundy
-
- 41 **Viewpoint**
I Can't Let You Do That, Dave
Computers should not treat their owners as adversaries.
By Cory Doctorow
-
- 43 **Point/Counterpoint**
The Case for Banning Killer Robots
Ban the bots? Considering both sides of the argument for and against.
By Stephen Goose/Ronald Arkin

Practice



- 48 **How to De-Identify Your Data**
Balancing statistical accuracy and subject privacy in large social-science datasets.
By Olivia Angiuli, Joe Blitzstein, and Jim Waldo

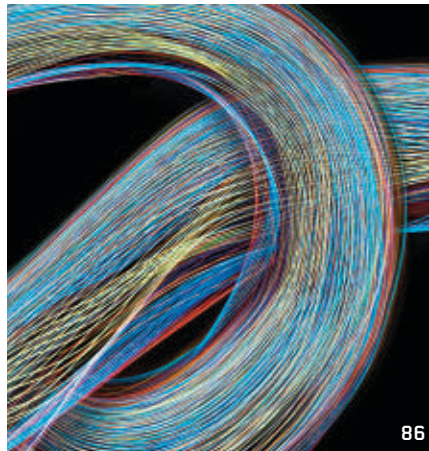
- 56 **Lean Software Development—Building and Shipping Two Versions**
Catering to developers' strengths while still meeting team objectives.
By Kate Matsudaira

- 59 **Challenges of Memory Management on Modern NUMA Systems**
Optimizing NUMA systems applications with Carrefour.
By Fabien Gaud, Baptiste Lepers, Justin Funston, Mohammad Dashti, Alexandra Fedorova, Vivien Quéma, Renaud Lachaize, and Mark Roth



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

Contributed Articles



- 68 **Personalizing Maps**
Digital maps can be engineered to adapt to a person's unique interests and experience in geographic space.
By Andrea Ballatore and Michela Bertolotto



Watch the authors discuss their work in this exclusive *Communications* video.
<http://cacm.acm.org/videos/personalizing-maps>

- 75 **Propositions as Types**
Connecting mathematical logic and computation, it ensures that some aspects of programming are absolute.
By Philip Wadler

- 86 **Smart Data Pricing: Using Economics to Manage Network Congestion**
Economic incentives that alleviate congestion for Internet customers can also improve business performance for network operators.
By Soumya Sen, Carlee Joe-Wong, Sangtae Ha, and Mung Chiang

Review Articles

- 94 **Internet Use and Psychological Well-Being: Effects of Activity and Audience**
The connection between online communication and psychological well-being depends on whom you are communicating with.
By Robert Kraut and Moira Burke

Research Highlights

- 102 **Technical Perspective**
Paris Beyond Frommer's
By Noah Snaveley
- 103 **What Makes Paris Look Like Paris?**
By Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros



Watch the authors discuss their work in this exclusive *Communications* video.
<http://cacm.acm.org/videos/what-makes-paris-look-like-paris>

- 111 **Technical Perspective**
In-Situ Database Management
By David Maier
- 112 **NoDB: Efficient Query Execution on Raw Data Files**
By Ioannis Alagiannis, Renata Borovica-Gajic, Miguel Branco, Stratos Idreos, and Anastasia Ailamaki



Association for Computing Machinery
Advancing Computing as a Science & Profession



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

Alexander L. Wolf

Vice-President

Vicki L. Hanson

Secretary/Treasurer

Erik Altman

Past President

Vinton G. Cerf

Chair, SGB Board

Patrick Madden

Co-Chairs, Publications Board

Jack Davidson and Joseph Konstan

Members-at-Large

Erik Allman; Ricardo Baeza-Yates;

Cherri Pancake; Radia Perlman;

Mary Lou Soffa; Eugene Spafford;

Per Stenström

SGB Council Representatives

Paul Beame; Barbara Boucher Owens

BOARD CHAIRS

Education Board

Mehran Sahami and Jane Chu Prey

Practitioners Board

George Neville-Neil

REGIONAL COUNCIL CHAIRS

ACM Europe Council

Fabrizio Gagliardi

ACM India Council

Srinivas Padmanabhuni

ACM China Council

Jianguang Sun

PUBLICATIONS BOARD

Co-Chairs

Jack Davidson; Joseph Konstan

Board Members

Ronald F. Boisvert; 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

Lissa Clayborn, Acting 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

Director of Media Sales

Jennifer Ruzicka

Publications Assistant

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@cacm.acm.org

Calendar items

calendar@cacm.acm.org

Change of address

acmhlp@cacm.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

Director of Media Sales

Jennifer Ruzicka

jen.ruzicka@hq.acm.org

Media Kit acmm mediasales@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

vic@cacm.acm.org

NEWS

Co-Chairs

William Pulleyblank and Marc Snir

Board Members

Mei Kobayashi; Kurt Mehlhorn;

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-Chairs

Stephen Bourne

Board Members

Eric Allman; Terry Coatta; Stuart Feldman;

Benjamin Fried; Pat Hanrahan;

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 *ACM Queue*.

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; 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; Amr El Abbadi; Sanjeev Arora;

Nina Balcan; Dan Boneh; Andrei Broder;

Doug Burger; Stuart K. Card; Jeff Chase;

Jon Crowcroft; Sandhya Dwaekadas;

Matt Dwyer; Alon Halevy; Norm Jouppi;

Andrew B. Kahng; Sven Koenig; Xavier Leroy;

Steve Marschner; Kobbi Nissim;

Steve Seitz; Guy Steele, Jr.; David Wagner;

Margaret H. Wright

WEB

Chair

James Landay

Board Members

Marti Hearst; Jason I. Hong;

Jeff Johnson; Wendy E. MacKay

ACM Copyright Notice

Copyright © 2015 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@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 \$100.

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 acmhlp@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





DOI:10.1145/2839512

Moshe Y. Vardi

On Lethal Autonomous Weapons

THE IMPRESSIVE PROGRESS in artificial intelligence (AI) over the past decade and the prospect of an impending global race in AI-based weaponry have led to the publication, on July 28, of "Autonomous Weapons: An Open Letter from AI & Robotics Researchers," with over 20,000 signatories by now, calling for "a ban on offensive autonomous weapons beyond meaningful human control." *Communications* is following up on this letter with a Point-Counterpoint debate between Stephen Goose and Ronald Arkin on the subject of lethal autonomous weapons systems (LAWS) beginning on page 43.

"War is hell," said General William T. Sherman, a Union Army general during the American Civil War. Since 1864, the world's nations have developed a set of treaties (known as the "Geneva Conventions") aiming at somewhat diminishing the horror of war and ban weapons that are considered particularly inhumane. Some notable successes have been the banning of chemical and biological weapons, the banning of anti-personnel mines, and the banning of blinding laser weapons. Banning LAWS seems to be the next frontier in effort to "somewhat humanize" war.

While I am sympathetic to the desire to curtail a new generation of even more lethal weapons, I must confess, however, to having a deep sense of pessimism as I read the Open Letter, as well as the two powerful Point and Counterpoint articles. I suspect many computer scientists, like me, like to believe that, on the whole, computing benefits humanity. Thus, it is disturbing for us to realize computing is also making a major contribution to military technology. In fact, since the

1991 Gulf War, information and computing technology has been a major driver in what has become known as the "Revolution in Military Affairs." The "third revolution in warfare," referred to in the Open Letter, has already begun! Today, every information and computing technology has some military application. Let us not forget, for example, the Internet came out of ARPAnet, which was funded by the Advanced Research Projects Agency (ARPA) of the U.S. Department of Defense. Do we really believe AI can, somehow, get an exemption from military applicability? AI is already seeing wide military deployment.

Rather than call for a general ban on military application of AI, the Open Letter calls for a more specific ban on "offensive autonomous weapons," which "select and engage targets without human intervention." But the concept of "autonomous" is intrinsically vague. In the 1984 science-fiction film *The Terminator*, the protagonist is a cyborg assassin sent back in time from the year 2029. The Terminator seems to be precisely the nightmarish future the Open Letter signatories are attempting to block, but the Terminator did not select its fictional target, Sarah Connor; that selection was done by Skynet, an AI defense network that has become "self-aware." So the Terminator itself was not autonomous! In fact, the Terminator can be viewed as a "fire-and-forget" weapon, which does not require further guidance after launch. My point here is not to debate a science-fiction scenario but to point out the intrinsic philosophical vagueness of the concept of autonomy.

Goose argues that ceding life-and-death decisions to machines on the battlefields crosses a fundamental moral and ethical line. This assumes

humans perform every life-and-death decision in today's battlefield. But today's battles are conducted by systems of enormous complexity. A lethal action is the result of many actions and decisions, some by humans and some by machines. Defining causality when discussing composite actions by highly complex systems is nearly impossible. The "fundamental moral and ethical line" discussed by Goose is fundamentally vague.

Arkin's position is that AI technology could and should be used to protect noncombatants in the battlespace. I am afraid I am as skeptical of the potential of technology to humanize war as I am skeptical of the prospect of banning technology in war. Arkin argues that judicious design and use of LAWS can lead to the potential saving of noncombatant life. Technically, this may be right. But the main effort of military designers has been and will be to increase the lethality of their weapons. I fear that protecting noncombatant life has been and will be a minor goal at best.

The bottom line is that the highly important issue raised by the Open Letter and by the Point-Counterpoint articles is highly complex. Knowledgeable, well-meaning experts are arguing the two sides of the LAWS issue. To the best of my knowledge, this is the first time the computing-research community is publicly grappling with an issue of such weight. That, I believe, is a very positive development.

Follow me on Facebook, Google+, and Twitter.

Moshe Y. Vardi, EDITOR-IN-CHIEF

Copyright held by author.

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 all aspects of the computing field. Available at no additional cost.

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



DOI:10.1145/2842510

Vinton G. Cerf

Advancing the ACM Agenda

ACM is a fairly compact organization, looking only at its modest but spirited staff, now led by the energetic team of CEO Bobby Schnabel and COO Pat Ryan. For the sake of saving

make ACM itself and its functions and activities as accessible as possible.

Diversity does not stop with accessibility. If we are to be a successful global organization, diversity on all fronts must be a global ambition. That means we must work with and empower our regional councils to help draw attention to the volunteer opportunities available within the ACM framework and to the valuable products and services that are the output of ACM's efforts. It means we need to encourage and spur efforts to bring women and minority groups more fully into the ACM family. And that brings me to another point about diversity.

We are entering into our biannual election process for leadership in ACM's volunteer component. The able, top-leadership team of Alex Wolf, Vicki Hanson, and Erik Altman will have completed their statutory two-year terms by the end of June 2016 and it is time to elect new officers. As the electorate, you have both the opportunity and the obligation to give careful thought to your choices for leadership. I chair the nominations committee as past president and I have also sought to impress on the committee the importance of diversity in our recommendations for candidates for volunteer leadership positions. It is my hope we can collectively put ACM on a track toward greater diversity in all dimensions of its activities, including our SIGs and other special interest groups, such as ACM-W.

We are already actively pursuing these goals and your election choices represent one way to spur the organization toward greater success in volunteer diversity in the years ahead. 

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

Copyright held by author.

space, I have not tried to list all the other wonderful staffers who tend to the machinery of ACM's operation and support the much larger volunteer structure that makes ACM what it is today. On the record, we have a first-rate team of full-time talent keeping ACM humming. But, alone, the staffers do not and could not make ACM the diverse and multifaceted organization it is. The volunteers, many of whom have served in multiple roles over decades, provide ACM with the intellectual and operational muscle power it needs to carry out myriad activities in support of computer science professionals and practitioners around the world. It is interesting to note that many of our dedicated ACM volunteers started with a single, focused commitment to, say, a program committee or a particular publication and over time and further engagement have become some of the most effective leaders of our association. Indeed, ACM volunteers come to value the impact they can make serving our community.

Despite the fact I have been a member of ACM since 1967, I am still discovering new things that ACM is involved in, and this after serving as president and now past-president for the past 3.5 years! The diversity of activity is well illustrated in ACM's Web pages but I confess I do not visit them enough to stay on top of everything. My New Year's resolution is going to be to visit the pages at least once a week to see what is going on. ACM is looking for new ways to draw attention

to its agenda and to opportunities for members and beneficiaries of ACM's work to discover and participate in the many volunteer opportunities that abound in our online and offline, computer-driven world. Just scanning the special interest group activities highlights the remarkable range of interests we serve. With the addition of the regional councils and chapters in India, China, and Europe, and elsewhere, ACM's footprint and scope of opportunity has increased notably.

This brings me to the first point of this column: diversity is key to ACM's ability to prosecute its agenda and to deliver benefits to members and practitioners around the world. The second point is that volunteers are absolutely essential to achieving the many goals ACM's volunteer leadership has adopted over the years. We need to expand outreach to underrepresented populations of computer professionals and educators and to the general public. We would benefit from many more volunteers and members drawn from non-U.S. populations, women, and other underrepresented minority groups, all of whom bring important perspectives and energy to ACM's mission. Just thinking about accessibility of computer-based systems reminds me how important it is to engage with users who would benefit from various assistive technologies and methods. We have a SIG focused on this community (SIGACCESS), but we also need members of these communities to feel and be welcome and supported in all our SIGs and our activities. We need to

What About Statistical Relational Learning?

WHILE STUART RUSSELL'S review article "Unifying Logic and Probability" (July 2015) provided an excellent summary of a number of attempts to unify these two representations, it also gave an incomplete picture of the state of the art. The entire field of statistical relational learning (SRL), which was never mentioned in the article, is devoted to learning logical probabilistic models. Although the article said little is known about computationally feasible algorithms for learning the structure of these models, SRL researchers have developed a wide variety of them. Likewise, contrary to the article's statement that generic inference for logical probabilistic models remains too slow, many efficient algorithms for this purpose have been developed.

The article mentioned Markov logic networks (MLNs), arguably the leading approach to unifying logic and probability, but did not accurately describe them. While the article conflated MLNs with Nilsson's probabilistic logic, the two are quite different in a number of crucial respects. For Nilsson, logical formulas are indivisible constraints; in contrast, MLNs are log-linear models that use first-order formulas as feature templates, with one feature per grounding of the formula. This novel use of first-order formulas allows MLNs to compactly represent most graphical models, something previous probabilistic logics could not do. This capability contributes significantly to the popularity of MLNs. And since MLNs subsume first-order Bayesian networks, the article's claim that MLNs have problems with variable numbers of objects and irrelevant objects that Bayes-net approaches avoid is incorrect. MLNs and their variants cannot only handle object uncertainty but relation uncertainty as well. Further, the article said MLNs perform inference by applying MCMC to a ground network, but several lifted inference algorithms for them exist.

Another major strand of research in this area the article did not portray accurately is probabilistic logic programming. The article said the "... first significant probabilistic programming language was Pfeffer's IBAL." While IBAL is definitely significant, Poole's ICL and Sato's PRISM were developed much earlier and have had a significant impact on the field. ICL and PRISM essentially extend the Prolog programming language by labeling facts with probabilities. They then use these probabilistic facts in the same way probabilistic databases use labeled tuples—to define a probability distribution over possible worlds. They can represent Bayesian networks, as well as cope with infinite possible worlds and an unknown number of objects. Sato received the test-of-time award from the International Conference on Logic Programming in 2015 for his seminal 1995 paper on PRISM.

The article concluded with, "... these are early days in the process of unifying logic and probability." On the contrary; with developments like MLNs, probabilistic logic programming, lifted inference, statistical relational learning, and more generally statistical relational AI, we are well on our way to solving this longstanding problem.

Pedro Domingos, Seattle, WA,
Kristian Kersting, Dortmund, Germany,
Raymond Mooney, Austin, TX, and
Jude Shavlik, Madison, WI

Author's Response:

Domingos et al. take me to task for my cautious claims about the state of the art. Given the history of occasional overstatement in artificial intelligence, caution seems appropriate. Domingos et al. claim "many efficient algorithms" exist for generic inference in these languages, as well as "a wide variety ... of computationally feasible algorithms for learning the structure of these models." If true, this is excellent news for computer science, since these problems subsume

the Halting Problem, as well as general program synthesis.

I am also accused of ignoring "the entire field of statistical relational learning." The excellent book Introduction to Statistical Relational Learning, compiled by my former student Lise Getoor and the late Ben Taskar (MIT Press, 2007), has 13 chapters on SRL languages and systems. My article referred to 10 of them.

My comment that IBAL was the first probabilistic programming language (PPL) was in no way intended as a slight to Sato's PRISM and Poole's ICL, contributions for which I have the highest respect. My article placed these approaches, along with BLOG, within the tradition of languages for defining probability distributions over logical worlds, as did Sato and Poole.

For example, the PRISM website (<http://rjida.meijo-u.ac.jp/prism/>) says, "The program defines a probability distribution over the set of possible Herbrand interpretations." My article clearly distinguished this approach from the PPL tradition based on distributions over execution traces of an arbitrary programming language, due to Koller, McAllester, and Pfeffer. Perhaps this is just a matter of terminology, although it seems worthwhile to point out that execution traces need have no relational structure at all. At the time of writing, the extensive bibliography at <http://probabilistic-programming.org/research/> does not include the early papers by Sato and Poole, but a broader notion of PPL might well include them.

Stuart Russell, Berkeley, CA

Give Me 'Naked' Braces

I was appalled by A. Frank Ackerman's letter to the editor "Ban 'Naked' Braces!" (Oct. 2015), which recommended programmers adopt a policy of always following the closing brace of each code block (presumably, in Algol-like languages like C and Java) with a comment intended to make it clear exactly which code block the closing brace belongs to. However,

assuming one's code is properly indented and the size and complexity of every function is below some reasonable limit, the best way (in my firmly held opinion) to determine the code block closed by any given brace is simply to follow the indentation up to find its matching brace. Although I readily admit a "reasonable limit" is a subjective matter (and may be strongly influenced by the particular development tools being used) the mere presence of a code block large and complex enough to cause potential confusion should serve as a red flag, indicating the code in question ought to be decomposed into smaller functions. In such cases, it may indeed be expedient to use comments in the manner Ackerman described, if one wishes to defer the decisions of exactly which parts to split out, how to pass parameters to and/or from the resulting functions, or what names to give these functions. However, this is a rather poor (and, one hopes, temporary) reason to add such comments. Every developer should perceive them as being unusual in production-quality code, rather than the normal way to end a code block.

The main problem with a comment that is supposed to document which code block is ended by a given brace is it merely documents the comment's belief about the purpose of each closing brace, and therefore opens up the possibility for the comments to be inaccurate. In the absence of any support by the development tools, every bug that was possible before adopting such coding practice is still possible afterward. So, any degree to which subsequent developers trust the comments simply increases their surprise when a bug happens. As a result, the comments actually communicate their author's skepticism, rather than confidence, as to whether or not the braces are in agreement with the intended structure of the code.

Broadly speaking, I would not support an attempt to convince coders they can provide a significant benefit to the quality of code by simply adding comments to indicate what they believe the code is doing. Absent a way to validate those beliefs, such attempts are generally misguided, par-

ticularly in the case of closing braces, as advocated by Ackerman.

Weston Markham, Pittsburgh, PA

Author's Response:

In matters of style there are always a variety of opinions. I personally find the practice I recommended useful both in development and in improving readability for maintenance. I have used it for many years and taught it to my students in several different procedural languages. Code has a tendency to move around quite a bit during development, and my proposed tags help keep the logic blocks straight when it happens.

A. Frank Ackerman, Butte, MT

I found an interesting juxtaposition between A. Frank Ackerman's letter to the editor (Oct. 2015) and George V. Neville-Neil's Viewpoint "Storming the Cubicle" (Oct. 2015), with Neville-Neil implying Ackerman's suggestion would not work in practice. Ackerman reasonably recommended developers tag their construct terminators (with, say, closing braces) to help avoid programming errors and seemed to assume developers would code this way if asked. But Neville-Neil illustrated the well-known fact that uneducated and/or unmotivated developers routinely ignore quality in favor of doing their work as quickly and easily as possible. Fortunately, Ackerman ended by noting educating beginners about code quality is the one true solution to such problems. I would like to add that education can take many forms, not only for beginners in an educational environment but even for experienced programmers through on-the-job training, continuing education, and, most important, peer mentoring.

Geoffrey A. Lowney, Issaquah, WA

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.

© 2015 ACM 0001-0782/15/12 \$15.00

Coming Next Month in COMMUNICATIONS

Spatial Computing

Open Data and Civic Apps

**The Building Blocks
of a Cloud Strategy**

**Algebraic Fingerprints
for Faster Algorithms**

Time Is an Illusion

Immutability Changes

**Unbalanced
Data Leads to Obsolete
Economic Advice**

**Why Knowledge
Representation Matters**

**Crowdsourced
Enumeration Queries**

**Bare-Metal Performance
for I/O Virtualization**

**Plus the latest news about
preserving the Internet,
non-volatile memory, and
how computers recognize
and understand images.**

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:10.1145/2833120

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

What Do We Do When the Jobs Are Gone, and Why We Must Embrace Active Learning

Moshe Y. Vardi ponders the outlook for people when all work is automated, while Mark Guzdial emphasizes the importance of Active Learning in teaching computer science.



Moshe Y. Vardi
**"The Future of Work:
But What Will
Humans Do?"**

<http://bit.ly/1geYbLj>
September 11, 2015

While artificial intelligence has proved much more difficult than some early pioneers believed, its progress has been nothing short of inexorable. In 2004, economists argued that driving was unlikely to be automated in the near future. A year later, a Stanford autonomous vehicle won a DARPA Grand Challenge by driving over 100 miles along an unrehearsed desert trail. A decade later, one hears regularly about the exploits of the Google driverless car. I believe that in 30 years it will be quaint, perhaps even illegal, for humans to drive on public roads.

Once driving is automated, delivery will be quick to follow; companies such as Amazon are already working hard

on fully automating their whole supply chain. The list of jobs likely to be automated grows daily, as AI increases its cognitive ability (it won at chess in 1999 and "Jeopardy!" in 2011), and its situational awareness and physical dexterity.

The unstoppable march of AI suggests that Herbert Simon was probably right when he wrote in 1956 that "machines will be capable ... of doing any work a man can do." I do not expect this to happen soon, but I do believe that by 2045 machines will be able to do much of the work that humans can do. So the question is: If machines can do almost any work humans can, what will humans do?

A typical answer is that if machines will do all our work, we will be free to pursue leisure activities. Of course, our economic system would have to undergo a radical restructuring to enable billions of people to live lives of leisure. One can imagine perhaps a society of a small number of haves and a

large number of have-nots, supported, say, by government subsidies. This is reminiscent of "panem et circenses," the Roman practice of free bread and entertainment to the masses. Yet I do not find this a promising future, as I do not find the prospect of leisure-only life appealing. I believe work is essential to human well-being. Is this our future?

It is instructive to recall the biblical story of the Garden of Eden in the book of Genesis (Chapters 2 and 3). God places Adam and Eve in the Garden and tells them: "But of the tree of the knowledge of good and evil, thou shalt not eat of it." The Serpent then tempts Eve, who, in turn, tempts Adam, to eat from the Tree of Knowledge. This leads to the expulsion of Adam and Eve from Eden. Furthermore, God metes punishment on the Serpent, Eve, and Adam: "And unto Adam he said, 'cursed is the ground for thy sake; in sorrow shalt

thou eat of it all the days of thy life; in the sweat of thy face shalt thou eat bread’.” So, according to this biblical story, our need to work for a living is an outcome of the failure of humanity to follow the word of God.

But let us contemplate humanity before and after the expulsion. Before the expulsion, Adam and Eve spent their time frolicking naked in the garden, where food is amply available without work; one could say they were no better than apes. One could even see the story as a metaphor for the roots of humanity in pre-human primates. After the expulsion, humans had to work for a living, but they have eaten from the fruit of the Tree of Knowledge. They were inventive. They have learned to hunt, mastered fire, invented agriculture, and eventually launched the Industrial Revolution. We are about to launch another Industrial Revolution, where work will be almost fully automated.

In a sense, humans used the knowledge they gained from the Forbidden Fruit to overcome God’s punishment; they will no longer need to work for a living; no more “by the sweat of thy face.” But can humanity go back to the Garden of Eden? Will we be happy just frolicking? Furthermore, human progress has been driven to a large extent by our desire to eliminate work or, at least, to lighten the toil. What will drive humanity once that goal has by and large been accomplished?

Thus, even if we manage to solve the economic implications of the complete or almost-complete automation of work, the question of the consequences to quality of life remains wide open. The classical Greek philosophers, starting with Socrates, discussed “Eudaimonia,” often translated as “the good life”—in other words, human flourishing. Aristotle viewed this question as one of the most central in philosophy. So the question facing us today is whether we can achieve the good life without work.

I believe the question of how humanity will occupy itself in the presence of intelligent machinery is one of the most central challenges facing society today. To repeat my earlier question: If machines are capable of almost any work humans can do, what will humans do?



Mark Guzdial “Be It Resolved: Teaching Statements Must Embrace Active Learning and Eschew Lecture”

<http://bit.ly/1RkVbuC>

August 14, 2015

In 1964, the U.S. Surgeon General produced a report unequivocally stating that smoking was a health hazard. The report had a dramatic impact on public policy and how people viewed smoking. Fifty years later, we know that the impact of that report was to save thousands and maybe millions of lives (see 50-year retrospective report at <http://1.usa.gov/1RkVyoM>).

We are at a similar point in understanding that lecture is an ineffective way of teaching. Active learning methods lead to better learning and greater retention. More, there is increasing evidence that poor teaching disproportionately impacts students from disadvantaged and under-represented groups.

Last year, the *Proceedings of the National Academy of Science* published a meta-analysis of 225 studies (<http://bit.ly/1Lovtqb>). The conclusion appeared as the title of the paper, *Active learning increases student performance in science, engineering, and mathematics*. There is increasing evidence that improved teaching reduces the achievement gap between disadvantaged and more advantaged students, for example, in biology (<http://bit.ly/1j5Kp06>) and in computer science (see new paper from ICER 2015 at <http://dl.acm.org/citation.cfm?doid=2787622.2787728>).

Now, *Nature* has just published a paper (<http://bit.ly/1Od6G8U>), “Why we are teaching science wrong, and how to make it right,” which includes the quote, “At this point it is unethical to teach any other way.” *Wired* magazine’s article on the active learning papers (<http://wrd.cm/1VvLAXa>) makes the connection more explicit: “The impact of these data should be like the Surgeon General’s report on ‘Smoking and Health’ in 1964—they should put to rest any debate about whether active learning is more effective than lecturing.”

It is now a matter of science, not opinion. *Active learning methods are more effective than lecturing*. We should encourage use of active learning methods in our classrooms. The blog post at <http://bit.ly/1MOMMxO> connects to resources for improved teaching methods in computer science. There are active learning methods that we can use even in large classes, like Peer Instruction (see PeerInstruction4CS.org).

Here is something concrete that we in academia can do. We can change the way we select teachers for computer science and how we *reward* faculty.

All teaching statements for faculty hiring, promotion, and tenure should include a description of how the candidate uses active learning methods and explicitly reduces lecture.

We create the incentive to teach better. We might simply add a phrase to our job ads and promotion and tenure policies like, “Teaching statements will be more valued that describe how the candidate uses active learning methods and seeks to reduce lecture.” We should read these critically. We should be convinced the candidates are not just mouthing active learning rhetoric, but are actually investigating and using active learning methods.

It is a small step, but it is an important one. Incentives change behavior. Stating clearly what we value in teaching statements will send messages that change how CS faculty teach over time. This step will likely have a critical impact on how we teach and who succeeds in computing.

I do not know if any other STEM disciplines are changing how they evaluate teaching statements in response to the *Nature* and *PNAS* papers. Let us lead. Let us be first. In 50 years, we might be looking back to find that our response to these reports brought more and more diverse students to computing. C

Moshe Y. Vardi is the Karen Ostrum George Distinguished Professor of Computational Engineering at Rice University, and editor-in-chief of *Communications*.
Mark Guzdial is a professor in the College of Computing at the Georgia Institute of Technology.

When Data Is Not Enough

Reproducibility of code is increasingly crucial to verifying scientific claims.

MASSIVE DATASETS and digital processing are transforming and accelerating science, but there is growing concern that many scientific results may not be trustworthy. Scientific procedures developed over centuries to assure reliable knowledge are sometimes overwhelmed by new ways of generating and processing scientific information. As a result, the scientific community is implementing requirements that help independent researchers reproduce published results, a cornerstone of the scientific method.

For data, the revolution is well under way. Inspired by projects like the Human Genome Project, the National Institutes of Health has provided infrastructure (and funding) for massive repositories of genetic and other data. In this field and others, researchers are expected to make their data available to other researchers.

Yet there is a growing recognition that provisions must also be made for the data-analysis software that supports the conclusions. Policy discussions of reproducibility “almost always talk primarily about data,” cautioned Victoria Stodden of the Gradu-

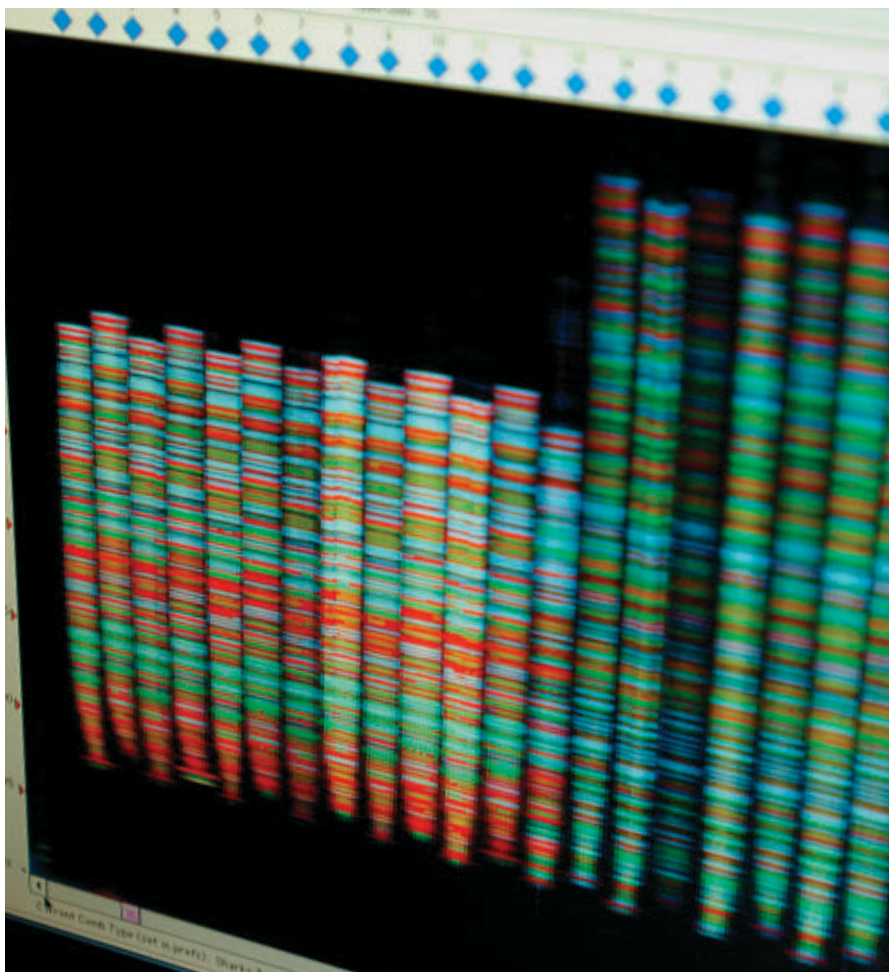


IMAGE BY MAGGIE BARTLETT/NHGRI

ate School of Library and Information Science at the University of Illinois at Urbana-Champaign. “This is a huge gap for computational science.”

“There has been an enormous flip in the last 20 years from data collection to data analysis,” said Roger Peng at the Johns Hopkins Bloomberg School of Public Health in Baltimore. In many fields, the critical challenges now reside in analysis of widely available big datasets, he said. This shift “has happened more quickly than I think most fields have been prepared to deal with.”

A Reproducibility Crisis

Reproducibility is always central to science, but especially when data may be fudged. In a recent case, Duke University researchers described a gene “signature” to predict cancer-drug response, which generated enormous interest and even clinical trials at Duke. Excited clinicians at the University of Texas MD Anderson Cancer Center enlisted statisticians Keith Baggerly and Kevin Coombes to confirm the conclusions, but the pair found numerous problems in the data analysis. Eventually the trials were stopped, but only after misrepresentations were found on researcher Anil Potti’s résumé.

Because of the limited published information, this reanalysis required “about 30% of our time for several months,” Baggerly said. “This is not a scalable solution.” Based on this experience, he said, “It became obvious that improving the reproducibility of reports was an important problem in and of itself,” including the posting of code as well as data.

“This was certainly an atypical case,” Baggerly stressed, but encouraging reproducibility is critical even without the possibility of misconduct. For example, a widely noted 2012 comment in *Nature* said retesting by the pharmaceutical company Amgen confirmed only six of 53 “landmark” biomedical studies cases. A study published in June on “The Economics of Reproducibility in Preclinical Research” (<http://bit.ly/1K7NeWL>) estimated the costs of irreproducibility in life science at \$28 billion per year in the U.S., attributing a quarter of that sum to “data analysis and reporting.” Recognizing this challenge, system-

Systematic efforts are under way to validate important studies in biomedicine and psychology.

atic efforts are under way to validate important studies in biomedicine and also in psychology, which has been plagued by irreproducible results.

Documenting Science in Progress

Traditionally, the laboratory notebook was a standard part of experimental research, documenting procedures, measurements, and calculations that led to the conclusions.

As analysis has moved to computers, researchers may explore their data using simple but powerful tools like Excel spreadsheets, which leave no record of the calculations. “It’s a little bit crazy,” Stodden said, since those programs are “not designed with scientific tracking and sharing needs in mind.”

Reproducibility of code should force scientists to migrate to platforms that document their manipulations explicitly. Although there is not yet any standard to rival the lab notebook, “it’s gotten markedly easier,” said Baggerly. “You’ve seen the advent of a whole bunch of tools which make tracking of computation and reconstruction easier.”

For code, GitHub is widely used to track rapidly evolving programs. This support is important for code, Stodden says, because unlike data, “well-used popular software is all forked and branched; it’s changing all the time.”

Other tools are also available, but so far none is appropriate for everyone. To help computational scientists navigate this complex landscape, Peng and two colleagues offer an online course on reproducible research, and he, Stodden, and a third author also collaborated on a book, *Implementing Reproducible Research*. “Everyone involved in science will need to know some data analysis,”

said Peng, or at least “how to manage people who are doing these complex data analyses” as part of increasingly multidisciplinary teams.

Putting Up Resistance

There are many reasons researchers may not want to post their code, noted Yolanda Gil of the University of Southern California’s Information Sciences Institute, who also chairs the ACM Special Interest Group on Artificial Intelligence. She likens the task to commenting code, a thankless task that everyone endorses but which many programmers neglect. On top of the work involved in making their software usable by strangers, researchers “often believe it is not very high-quality,” she said, or that “no one will use it.”

If the code and data are used by others, that can create even more work, as climate scientists learned when outside skeptics pored over their work. “It’s easy to take data that someone has provided and do something to it and say that something fraudulent has occurred,” said Peng. In principle, such scrutiny will improve the science, but the extra work can make researchers reluctant to share.

Publications

In practice, researchers rarely recreate experiments simply to confirm them, but they often retrace the key steps in order to extend the results. Scientific-article style evolved specifically to ensure readers have enough information to repeat the procedures.

An official publication is an ideal time to assess the reliability of the results and to deposit data and code in a public repository, and journals increasingly allow online posting of supporting material. They also frequently require authors to affirm that data has been placed in public repositories, although those requirements are not necessarily enforced. “The legacy publishing system hasn’t quite caught up to the fact that we need to have code and data and digital scholarly objects associated with our publications,” said Stodden.

One exception to this neglect is the journal *Mathematical Programming Computation*, founded in 2009. Following the mathematics tradition of checking proofs before publication, this journal not only requires software

but, whenever possible, technical editors confirm the software works as advertised. “It can sometimes be quite difficult,” said founding editor William Cook, a professor of Combinatorics and Optimization at the University of Waterloo in Canada, who says this process has uncovered serious unintentional errors. “I am now skeptical about any computational results (in my own area of mathematical optimization) that have not gone through a review like we do.”

Still, “the extra review is a great burden on authors,” he admits. Some of them complain that they get no extra credit for publishing in this journal, “so they choose to go elsewhere.”

As a demonstration for less-mathematical science, Gil has been shepherding The Geosciences Paper of the Future Initiative, sponsored by the National Science Foundation. For this project, a dozen or so teams of geoscientists have volunteered to submit articles that illustrate transparency and reproducibility, to be published in the next few months. Among other things, “We’re trying to incorporate these best practices as you do the work, not at the end,” Gil stressed.

Providing Incentives

“Everyone agrees” about the need to address code reproducibility, Stodden said. “We’d have much better science if we resolved these issues.” However, expecting early-career researchers to unilaterally devote effort to this end is not reasonable, she stressed. “They’re just not going to, until the incentives change.”

As a step toward providing those incentives, last winter the non-profit Center for Open Science hosted a committee including scientific experts as well as representatives of journals and funding agencies. The group published a general framework for journals to codify their transparency standards for each of eight standards—notably including “analytic methods (code) transparency.” For each subject, the framework specifies four tiers of rigor, ranging from vague support or silence to detailed validation of posted resources.

The standards had to be flexible, said executive director Brian Nosek, a psychologist at the University of Virginia in Charlottesville. “Different fields have different documentation

“Different fields have different documentation needs ... there can’t be a one-size-fits-all solution.”

needs,” and some journals have limited resources. “There can’t be a one-size-fits-all solution.” The graduated structure should make it easy for journals to get started and to increase their commitment over time.

Still, “just having a standard isn’t a solution in itself,” Nosek acknowledged. Ultimately, requirements for deposition of data, code, and other research materials also will have to be enforced broadly throughout the scientific ecosystem, including funding agencies and committees making tenure decisions. “A focus on journals was our start,” Nosek said. “It was low-hanging fruit.”

Into the Future

Code used to derive scientific results is finally being recognized as a critical ingredient for reproducibility, but supporting tools are still primitive. “The only solution that we have now is to dump all the code onto you,” said Peng, which he likens to trying to teach a musician a song with a bit-level sound file. Researchers are still trying to devise the code equivalent of sheet music that makes the significance of sounds humanly useful.

Simply making code available also fails to answer the key question: whether it actually does the calculations it is supposed to do. Such verification is burdensome and may be almost impossible for large programs like those used for climate, unless researchers can implement intrinsically robust techniques like those being developed for mathematical proofs. (See “A New Type of Mathematics?” *Communications*, February 2014.)

In addition, even a complete snap-

shot of the code used in the publication will miss “all the trails and avenues that scientists went down that didn’t pan out,” Stodden warned. Although such exploration is an important part of science, it naturally selects pleasant surprises. The statistical significance of a clinical trial, for example, can only be assessed if there was a clear protocol on record beforehand. Tracking the dead ends would require a much more sophisticated software environment, one that does not yet exist. In principle, such a platform could document the entire history and workflow of a project. Unlike the minutiae of physical and chemical experiments, data and code can be completely and cheaply captured, so software tools could lead the way to improving reproducibility across the entire scientific enterprise.

For now, however, advocates are just hoping to improve the transparency of code. “It’s a serious collective-action problem,” Stodden said. Only the highest-impact journals can impose new requirements without the risk of alienating authors, for example. “The progress is very slow,” Stodden said, but “we’re getting there.” **C**

Further Reading

Ince, D.C., Hatton, L., and Graham-Cummig, J. The case for open computer programs. *Nature* 482:485–488 (2012). doi:10.1038/nature10836

Stodden, V., Borwein, J., and Bailey, D., “Setting the default to reproducible” in Computational Science Research. *SIAM News*, June 03, 2013 <http://sinews.siam.org/DetailsPage/tabid/607/ArticleID/351/>

Stodden, V., Leisch, F., and Peng, R.D., *Implementing Reproducible Research*. CRC Press, 2014, <https://books.google.com/books?id=WVTSBQAAQBAJ>

Stodden, V., *Opportunities and Challenges for Open Data and Code: Facilitating Reproducibility*. 2013, <http://bit.ly/1MzGSBB>

Baggerly, K., *The Importance of Reproducible Research in High-Throughput Biology: Case Studies in Forensic Bioinformatics*. 2010, <http://bit.ly/1EEUWmM>

Don Monroe is a science and technology writer based in Boston, MA.

The Hyper-Intelligent Bandage

Scientists are developing smart, sensor-packed dressings to help heal chronic wounds.

IN THE RECENT movie *Avengers: Age of Ultron*, the supremely skilled archer Hawkeye suffers a major injury to his midsection during the opening battle. Hawkeye is quickly wrapped up and whisked back to superhero headquarters, where a bioengineer repairs the terrible wound, using a tool that scans and then prints a new layer of synthetic skin across the injured area. After a quick recovery, he is soon up and racing back into action.

If only wound treatment were that advanced in the real world. Unfortunately, effectively dealing with major wounds and burns, whether they stem from disease or the battlefield, remains a technical and scientific challenge.

A wound is a complex micro-environment packed with competing influences, from the agents dispatched by the body's immune response to the bacteria that cause infections. Standard bandages and dressings cannot always maintain the infection-free, moist, oxygen-rich environment required for a wound to heal. One often-cited example is the population of patients with severe diabetes who incur external ulcers on their feet; approximately 25% of diabetics suffer from a foot ulcer at some point in their lives, and these wounds are a leading cause of amputations.

Recently, scientists have begun working on new types of smart bandages capable of monitoring and even treating such chronic wounds. The research is still in the prototype stage, and the approaches vary, but these intelligent devices will be closer to miniature medical labs than advanced bandages: they will protect the wound, provide a scaffold on which new cells can grow, monitor the area for infections, wirelessly alert caregivers to changes in the status of the injury, and potentially deliver medications directly.



This prototype electronic skin patch can store and transmit data about a person's movements, receive diagnostic information, and release drugs into skin.

Chemist Conor Evans of the Wellman Center for Photomedicine at Massachusetts General Hospital compares the nascent technology to the life-supporting protective garb of the space-walking astronaut. "What we're trying to do is make a space suit for wounds—something that can go over a wound, keep it safe, and start to allow it to heal," he says.

Sensing Problems at the Site

The treatment of wounds has changed dramatically over the millennia, beginning with ancient Sumerians applying poultices of honey and animal fats to an injury, but the body's healing process has remained the same. First, the body attempts to stop bleeding through coagulation. Then inflammation takes over, followed by the proliferation of new cells, repair of the injured skin, and eventually the growth of new skin to cover the wound.

In the case of chronic wounds, the first step is not as much of a concern as the latter ones, during which bacteria and other factors conspire to slow or halt healing. One of the major areas of focus in smart bandage research

is detecting biological warning signs of these problems. At Northeastern University, chemical engineer Edgar Goluch and colleagues are developing sensors designed to identify the presence of specific bacterial species.

Oxygen is another key metric: healing tissues consume large amounts of oxygen; without it, they die. This is one of the main reasons for diabetic ulcers and the chronic wounds that can result from bedsores—the pressure of one's own body actually denies oxygen to the surrounding tissue.

Evans and his group developed a prototype bandage that measures tissue oxygenation through the use of oxygen-sensitive phosphores, or light-emitting molecules. When oxygen bumps into these molecules, it effectively dims the phosphores; in low-oxygen environments, "as oxygen decreases, it glows brighter and brighter," he explains. The result is a bandage that maps oxygen levels across the wound. Initially, the bandage had to be photographed to reveal the results, but now oxygen-deficient areas are visible to the naked eye. Evans says the ability to pinpoint these trouble spots could allow for a more targeted response.

Other groups, including a team led by Harvard University tissue engineer Ali Khademhosseini, are developing technologies to measure not just oxygen, but temperature, pH, and more. For diabetic patients who often lose feeling in their extremities and thus fail to notice a worsening wound, this kind of feedback could be extremely valuable; a smart bandage with such sensors could alert them to the presence of an infection.

Responding in Real Time

The next step in the process is reacting to alarming information. Khademhosseini and his multidisciplinary team are building smart bandages that con-

sist of a flexible substrate with embedded microelectronics and communications components, along with sensors and drug delivery mechanisms. Current prototypes are about a half-inch thick and six to eight inches long. They communicate wirelessly with smartphones and computers through low-energy Bluetooth, and a coin cell battery powers the bandages, affording them a lifetime of about one week. The group is also working to incorporate wireless charging through wireless power transfer technology.

In one prototype, low oxygen levels in the wound trigger a chemical reaction inside the bandage; the by-product of this reaction is oxygen, which then migrates out of the bandage and into the deprived area. Khademhosseini and his colleagues plan to conduct animal testing of this oxygen-releasing system, along with another that administers antibiotics. The goal is to graduate to human clinical trials within a few years. In the case of an actual patient, the bandage could wirelessly relay sensor data to a smartphone on a regular basis. An app could translate that raw data, then generate updates on whether the wound is healing or becoming infected, and relay this information to a physician for review.

"The system could transfer all the sensing parameters to medical staff, including the oxygen concentration of the wound (and) temperature," suggests Ali Tamayol, a bioengineer at Harvard who is also involved in the project. "They could then decide if things are going well or if they have to do something."

Yet Khademhosseini says the smart bandage could also be a fully closed-loop system, given the power of micro-processors and the relatively small amount of data to be analyzed. The system could be programmed so certain conditions trigger it to release oxygen molecules, while others spur the targeted delivery of antibiotics. Tamayol explains that you could define a critical range for pH, for example, and if those parameters were breached, then a particular drug could be released. "The system could understand and react," he says. "Ideally, it would all be automatic."

Healing the Wound

The final step, after warding off infection and maintaining the right oxygen,

"There's an ensemble of technologies that will be applied to wound healing, because every patient is different, and every wound is different, too."

moisture, and pH levels, is allowing the skin to properly reform. One of the ways our bodies deal with a severe injury is through scarring, but scar tissue can be painfully taut and unnatural, especially in areas like the face. Harvard University biophysicist Kevin "Kit" Parker, an Army veteran, has seen this firsthand in U.S. soldiers returning home from Iraq and Afghanistan with severe craniofacial burns. "The scarring can be really terrible," Parker says. "They struggle to use their mouths and open and close their eyes."

The soldiers' struggles inspired Parker to search for a better alternative—a dressing that would allow the skin to heal with minimal scarring. The key may be the material that actually comes in contact with the wound as new cells move in to replace the damaged ones. The skin cells benefit when there is a 'scaffold' in place, and Parker and his team found that a substrate made of precisely aligned nanofibers is an ideal material. Other scientists—including Tamayol and Khademhosseini—have tested this idea of using nanofiber dressings for wound healing, with positive results. Parker's group, however, has developed a new manufacturing technique, inspired by a cotton candy machine, that allows them to incorporate a wider range of materials into the nanofibers.

In their research of the literature on wounds, Parker and his colleagues learned about a protein called fibronectin that can accelerate the healing process and prevent scarring. The group incorporated some of this protein into its polymer-based nanofibers, then tested

the hybrid material in animal models. The results have not yet been published, but the fibronectin seemed to accelerate the healing and prevent scarring in mice. The work is still in its early stages, but Parker is encouraged. "We want skin that's soft, almost like baby skin, so if you have craniofacial burns you can still move your mouth," Parker says, "and we think we've nailed it."

Eventually, Parker says, this new material could be incorporated into smart bandage technology, but he cautions there will not be a single cure-all device for chronic wounds and burns. Khademhosseini's group is developing a range of different prototypes for that reason, and none of the scientists are underestimating the difficulty of the problem they are working to address.

"Wound site biology is very complex," Parker explains. "There's a whole ecosystem in there before you even start talking about infections. There's an ensemble of technologies that will be applied to wound healing, because every patient is different, and every wound is different, too. We're coming up with one technology, but there's no silver bullet." **C**

Further Reading

Dargaville, T.R., Farrugia, B.L., Broadbent, J.A., Pace, S., Upton, Z., and Voelcker, N.H. **Sensors and Imaging for Wound Healing: A Review.** *Biosensors and Bioelectronics*, 41.

Li, Z., Roussakis, E., Evans, C.L., et. al. **Non-Invasive Transdermal Two-Dimensional Mapping of Cutaneous Oxygenation with a Rapid-Drying Liquid Bandage.** *Biomedical Optics Express*, Vol. 5, Issue 11, 2014.

Abrigo, M., McArthur, S.L., and Kingshott, P. **Electrospun Nanofibers as Dressings for Chronic Wound Care: Advances, Challenges, and Future Prospects.** *Macromolecular Bioscience*. 2014, 14.

Najafabadi, A.H., Tamayol, A., Khademhosseini, A., et. al. **Biodegradable Nanofibrous Polymeric Substrates for Generating Elastic and Flexible Electronics.** *Advanced Materials*. Sept. 3, 2014.

Badrossamay, M.R., Balachandran, K., Parker, K.K., et. al. **Engineering Hybrid Polymer-Protein Super-Aligned Nanofibers via Rotary Jet Spinning.** *Biomaterials*, March 2014.

Gregory Mone is a Boston, MA-based science writer and children's novelist.

© 2015 ACM 0001-0782/15/12 \$15.00

Technology Brings Online Education in Line with Campus Programs

Whether sitting in front of a screen or in a classroom, online and campus-based institutions want to verify students actually attend classes, take exams.

WHEN YOU CHEAT, you only cheat yourself," is an oft-repeated pearl of wisdom from teachers, parents, and enlightened students alike. Nevertheless, 65–75 percent of college-age students have admitted to cheating at one time or another, according to surveys conducted in the early 1960s, and recent studies indicate cheating still remains rampant. Indeed, Harvard University forced out 70 students for cheating during a May 2012 final exam, and in 2015, more than 60 Dartmouth College students were accused of cheating in a sports-ethics class.

In response to such incidents, as well as growing pressure from accreditation bodies, employers, and even alumni, traditional campus-based institutions and online universities are implementing technological solutions for stopping cheaters.

Institutions also are being closely monitored by the U.S. Department of Education, which oversees the financial aid/student loan eligibility process through which most schools receive the majority of their funding. "The Department of Education took a stronger stance when the Office of Inspector General did its audits for financial aid disbursements," according to Tim Dutta, CEO of Verificent Technologies, a provider of student verification and anti-cheating technology. "From that perspective, schools have taken a more proactive and rigorous approach" to ensuring academic integrity with their programs. "Now these [verification and integrity solutions] have kind of gotten on the radar of institutions as a 'need to have,' rather than a 'nice to have.'"



A "flipped" classroom: students watch video lectures at home and work on problems in class.

The challenge faced by these institutions is twofold: schools must ensure the work a student turns in throughout the duration of the course is his or her own work, as well as preventing or discouraging students from cheating on examinations. Ultimately, if cheating is seen as pervasive, the reputation of the school, as well as its graduates and faculty, will suffer.

One Step Ahead

While students from earlier eras plagiarized authors' works and liberally quoted from other students' previous papers, it required nearly as much time and effort to retype the work as it would to simply complete the assignment properly. The advent of computer- and Internet-based learning made it easier to copy-and-paste from ever-growing annals of legitimate source material.

Furthermore, websites have sprung up to offer custom-written papers on hundreds of subjects, presumably properly sourced and annotated to pass muster with most professors.

The ease and pervasiveness of plagiarism is driving the use of technology to assess the originality of student-submitted work. Scanning software such as Turnitin (<http://turnitin.com/>), which *The New York Times* reported being used by more than 3,500 universities, compares student-submitted work with content on the Web, as well as student-submitted papers. According to the company Turnitin, the software uses a matching algorithm to identify passages of words within submitted work identical to those found within its repository of content. The result is an "originality report," indicating the degree of simi-

larity between submitted work and sources of content contained in the database. As such, human judgment is also required to determine if an instance of plagiarism has occurred.

Software, however, cannot catch all instances of students passing off the work of others as their own, particularly as online courses continue to grow in popularity among both traditional educational institutions and for-profit providers. For providers that offer fully online courses, it has become increasingly important to verify a student's work is that of the person who originally registered for the course and will be receiving a credential for completing it successfully.

Testing the Limits

Most course providers that offer a certification or fully accredited degree use examinations to demonstrate proficiency with a subject area. Not surprisingly, many students try to game the system and cheat on these exams, generally motivated by achieving an external reward such as a good grade in a course, the satisfaction of a degree program and, eventually, the securing of a good (or better) job.

Students have migrated from old-school cheating techniques (such as writing answers on their clothing) to more technologically advanced methods that are specifically designed to avoid detection by proctors. For example, students have scanned a beverage label into an image-editing program, replaced the small text listing the ingredients with exam answers, and reprinted and pasted the label back onto the bottle to serve as a cheating aid that is hiding in plain sight, usually undetected by proctors. Other students have used electronic devices to facilitate cheating, such as by loading cheat sheets into cellphones, calculators, watches, or other devices, or using them to share information surreptitiously.

To combat these types of schemes, traditional universities have embraced an approach that mirrors a 'clean room'—a highly monitored, controlled environment that makes cheating very difficult. The campus-based University of Central Florida is one of many schools that use an ultra-secure testing facility to thwart cheaters; students entering the center must lock up all

personal property, with the exception of identification documents. The room is under surveillance by video cameras and live proctors, there are no scheduled breaks, and two forms of ID are required to verify the test taker's identity.

Online Verification

In an online test-taking environment, it is much harder to verify students are not cheating, simply because it is not possible to truly know if something or someone else is assisting them during the exam.

There are a number of software-based solutions designed to make it more difficult for online students to cheat. These systems are not foolproof and will not stop a truly determined person cheating, but may deter casual cheaters, and demonstrate a school's commitment to reducing academic fraud.

Software Secure's RPNOW or Verificient Technologies' ProctorTrack each use a combination of software, analytics engines, and vision- and sensor-based systems to verify students are who they say they are, are engaged in coursework, and are closely monitored to prevent cheating.

At massive open online course (MOOC) provider edX, vice president of product Beth Porter says, "Our vendor is Software Secure, and what it does is look for anomalies."

Software Secure's RPNOW captures images of student test takers and their respective photo IDs. The test takers must confirm the images can be later used to authenticate their identity by matching the photo when the test-tak-

er logs in to take the test. The system also supports application blacklisting to prevent the test-taker from using unauthorized programs on their computer before or during a test, and records the entire exam session using their webcam and microphone. This video is stored, along with everything that is displayed on the user's screen, to allow quick review of the test and the environment by certified proctors, who can review any anomalies that may be indicative of someone trying to cheat, such as frequently looking away from the screen to read a nearby book, cheat-sheet, or second computer screen out of view of the webcam.

Verificient Technology's ProctorTrack operates in a similar fashion, tracking eye movement to determine if the subject is looking away from the screen (to prevent stealing glances at a cheat sheet), while the microphone records audio to detect extraneous noise that could be indicative of someone else in the room assisting the test-taker.

Verificient's Dutta contends monitoring test-takers is not about forcing students to be "robots" during an exam, but simply ensuring they can be "seen" by the verification software, and creating a record that can be reviewed to see if behavior that could be construed as cheating is present.

Some observers have raised concerns about student privacy with anti-cheating software that records facial and other personal identification details, which are sent back over the Internet. For its part, Verificient claims ProctorTrack complies with the U.S. Family Educational Rights and Privacy Act (FERPA), and has pledged that the software does not access any files on a student's hard drive, does not share any data with third parties, and purges personal student data within 30-60 days, depending upon the university.

Unlike K-12 students, college attendees ultimately have the choice of whether or not to attend a school that uses this type of technology. Moreover, students seeking an online degree that is well respected and valued in the marketplace likely will seek out schools that are taking steps to improve their reputations, even if it means subjecting themselves to remote monitoring.

The key, according to security and technology expert Joseph Kirkpatrick

Software-based systems will not stop a truly determined character, but demonstrate a school's commitment to reducing academic fraud.

of Kirkpatrick Price, a Tampa, FL-based licensed CPA and PCI quality security assessor firm, is for both the vendors supplying the monitoring software and the schools utilizing it to have clear privacy and security policies, as well as having a compliance officer or team to ensure these policies are enforced.

Kirkpatrick expects the debate over the fair, ethical use of monitoring and verification of student activities for online and traditional universities will continue, likely as part of the overarching debate over data collection, privacy, and protection by private companies and the government. "There are all these ethical questions," Kirkpatrick says. However, he notes that to enjoy certain conveniences, such as attending classes and taking exams online, "you're going to give up some privacy."

Still, stopping or reducing cheating is not the only area in which online educational providers must consider how they interact with students. A recent settlement between edX and the U.S. Department of Justice under which the MOOC provider will make its website and course delivery platform accessible to blind and hard-of-hearing students and to those with other physical disabilities, has thrown a light on the need for online education providers to comply with the Americans With Disabilities Act (ADA). This is especially important for any institution that receives federal funds (via grants or student loans), as Section 504 of the ADA specifies no qualified individual will be excluded from a program or activity that receives federal funding.

Similarly, educational institutions are likely to want to comply with Section 508 of the ADA, which requires electronic and information technology to be accessible to people with disabilities, to attract the maximum number of potential students and avoid any potential legislation related to discrimination.

Overall, the benefits of online education providers deploying anti-cheating, verification, and accessibility software far outweigh the costs, particularly with respect to the growing population of non-traditional students flocking to online degree programs.

Dan Butin, founding dean of the School of Education & Social Policy at Merrimack College, notes that implementing these systems is fairly easy for schools to do, even if there is some

It is key that vendors supplying monitoring software and schools utilizing it have clear privacy and security policies, along with a compliance officer or team to ensure policies are enforced.

pushback from students. Indeed, students must agree to the terms and conditions of using the school's software-based online learning platform in order to take courses, and as such, likely will agree to use this type of exam monitoring as a condition of receiving course credentials.

"The economics of this [online] model are going to necessitate that," Butin says. "The front-end investment that [Arizona State University] or anyone else is going to have to put in for student verification or anti-cheating technology is minimal, compared with the potential for revenue generation that will happen once they figure out these questions," Butin says.

For its part, edX says there will be no changes in course pricing due to the company's accessibility efforts. "Any investment by edX or course content providers with regard to accessibility is not passed on to the students," says edX general counsel Tena Herlihy. **C**

Further Reading

Harvard Cheating Scandal: <http://www.thecrimson.com/article/2012/8/30/academic-dishonesty-ad-board/>

Cheating Survey: <https://www.insidehighered.com/news/2012/03/16/arizona-survey-examines-student-cheating-faculty-responses>

University of Central Florida Testing Center Video-How it Works: <http://utc.sdes.ucf.edu/>

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

© 2015 ACM 0001-0782/15/12 \$15.00

ACM Member News

HEISER FOCUSES ON MICROKERNELS



When it comes to operating systems, Gernot Heiser is a minimalist. Heiser, Scientia Professor and

John Lions Chair of operating systems in the University of New South Wales' School of Computer Science and Engineering in Sydney, Australia, develops microkernels, the rudimentary core of operating systems.

Heiser received an undergraduate degree in physics from Germany's University of Freiburg, and a master of science degree in semiconductor physics and computer science from Brock University in Ontario, Canada. He earned his Ph.D. in computer engineering and 3D semiconductor device simulation at Switzerland's ETH Zurich.

To Heiser, "Microkernels are sexy; they make software and systems truly reliable and much more secure."

Microkernels are slimmed-down versions of traditional operating systems that consist of tens of millions of lines of code. Experts estimate developers make three errors for every 100 lines of code, making microkernels inherently "more complex, faulty, and error-prone" than operating systems, Heiser says.

Heiser and his team developed the Security Enhanced L4 (SEL4) microkernel with just 10,000 lines of code. They employed mathematical proof techniques, providing basic isolation mechanisms (the ability to partition resources and switch between the partitions). These techniques, coupled with the modest amount of code, "reduce the attack vector/surface and the propensity for flaws, making SEL4 essentially bug-free and highly secure," Heiser says. SEL4, he says, is the only protected-mode OS with sound and complete worst-case execution time analysis.

Heiser also is director of Forests Alive, an Australian company whose mission is to demonstrate it makes economic sense "to save the world's forests by using them as carbon sinks."

—Laura DiDio

Historical Reflections

The Digital Dark Age

...and why it will have to wait.

THE GROWTH in accessible digitized primary source materials, together with the improvement in quantity and quality of online research tools has changed fundamentally the way in which historians go about their business. While the fundamental importance of primary sources remains the same, it is ever more the case that the sources themselves were born digital, consumed digitally, and (hopefully) preserved digitally. But what assurance do we have that in the next century, or the next millennium, historians will be able to access materials that were produced on machines, and software packages that have long ceased to exist? This is the question Google Vice President and Chief Internet Evangelist (and ACM past president), Vint Cerf raised in a number of engaging recent public talks, interviews, and in his *Communications*’ “Cerf’s Up” column. His conclusion is that we are standing on the edge of a precipice, and that unless we take appropriate steps a digital dark age awaits us. I cannot easily gauge the effect of Cerf’s remarks in the U.S., but in Europe, his comments have been taken up widely, and accepted more or less uncritically, by the broadcast and print media.

Among professionals working in the digital preservation field, the reaction has been much less accepting. Many of the people with whom I come into contact exhibit a simmering resentment that a great deal of pioneering and fundamental work carried out in Europe (and more widely), is being overlooked in the publicity storm that Cerf’s pronouncements have generated.

This is entirely understandable, but I think it is almost certainly the wrong response. For most of the last decade, I have been working actively in digital preservation, having been

drawn into the field from a background in what has come to be called the “digital humanities,” where I continue to take a particular interest in the history of computing. For most of that time, it has been something of an uphill struggle to persuade academic colleagues that digital preservation was not only something that needed to be taken seriously, but was a field in which there were, and are, interesting and intellectually satisfying challenges to be addressed. What was true of fellow academics was doubly so of business leaders, politicians, and other opinion formers. Fellow digital preservationists were often inclined to scratch their heads, and with their thousand-yard stare firmly in place, ask what they have to do to get preservation onto the political and business agenda. I have attended a number of conferences where participants debated whether the best approach to getting preservation taken seriously was to deploy somewhat apocalyptic tales of the dangers of losing large slices of our cultural heritage as a result of inattention to preservation, or to encourage better digital custodianship by more positive means. It probably needs a little of both, but the uncomfortable truth is the mes-

The situation with the protection of our digital heritage is far from bleak, and there are, indeed, many reasons to be cheerful.



senger is often more important than the message, and whether the approach is to scare the world at large into taking preservation seriously, or to provide less dramatic encouragement, the messenger needs to be able to cut through the noise and be heard. Vint Cerf, as one of the architects of the Internet, and a key figure on the U.S. politico-scientific stage, is extremely well placed to bring digital preservation to greater prominence than has hitherto been the case, and should therefore be welcomed wholeheartedly, even if he does not always capture fully the digital preservation zeitgeist. The media interest in Cerf's talk of a looming digital dark age can only serve to help raise awareness among computer scientists and engineers and encourage them to engage with the outstanding technological challenges. His intervention may even be instrumental in persuading funders to put their dollars and euros behind much-needed digital preservation research.

The almost daily news of the damage being caused to historic build-

ings, and sites of archaeological and cultural importance in the Syrian civil war, makes talk of a digital dark age seem particularly resonant. The recent execution by ISIS of the longtime keeper of Palmyra's extraordinary cultural artifacts, octogenarian Khaled al-Asaad, for 'crimes' including representing Syria at "infidel conferences," and serving as "the director of idolatry" in Palmyra, should remind us that in some parts of the world, protecting cultural heritage comes at a very high price. However, whatever grounds there are for despondency in the Middle East, and somewhat at odds with the picture being drawn by Cerf, the situation with the protection of our digital heritage is far from bleak, and there are, indeed, many reasons to be cheerful. Over recent years the global digital preservation community has been very active and there has been real and substantial progress made. In addition to calling for new technologies and techniques, and promoting the "Olive" project at Carnegie Mellon (<https://olivearchive.org>), which approaches preservation by using virtual

machines running application state capture files, Cerf has drawn attention to the importance of the human, societal, and organizational dimensions of preservation. His call, for example, to revisit the rules on copyright to allow for a "fair use" provision in the case of preservation activity would be a real step forward.

Despite the impression left by Cerf's remarks, Olive is not the only fruit of the digital preservation field. In Europe, in response to Cerf, the Executive Director of the Digital Preservation Coalition, William Kilbride, has issued a call for the preservation community to highlight preservation activities or projects in which they have been involved, so that a fuller picture of the preservation landscape over recent years might be given. Many of these contributions have been gathered together in a Twitter feed with the pleasingly optimistic hashtag (<https://twitter.com/hashtag/nodigitaldarkage>).

For the remainder of this column, and in the spirit of the nodigitaldarkage campaign, I would like to draw

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



some attention to a number of projects and other activities members of my own team have led, or in which they have played prominent roles. These cover the development of innovative tools and techniques, dedicated outreach and dissemination, and the organizational aspects of preservation and reuse of culturally significant material in digital form.

Vint Cerf's comments give very little background of the intellectual roots of projects like Olive. Most of the early work in digital preservation concentrated on migration as a preservation approach. This, in essence, depends on copying or converting digital objects originally intended to run on one technology platform to run on another. Inevitably, migration involves changing some of the characteristics of the original digital object, so a lot of attention must be paid to ensuring the properties considered to be of the greatest importance for a designated stakeholder community are preserved intact. Major practical limitations in the migration approach are exposed when the digital objects in question are inherently complex. Migrating a modern computer game from one platform to another, for example, involves a level of technical expertise that simply does not exist in cultural heritage organizations, and involves intellectual property rights issues, which are, for all practical purposes insurmountable. As the tendency is for digital objects to become ever more complex, interest has turned to developing approaches to preservation that depend, as Olive does, on

Most of the early work in digital preservation concentrated on migration as a preservation approach.

emulation. The KEEP (Keeping Emulation Environments Portable) project (<http://www.keep-project.eu>), was the first publically funded project to develop emulation services to enable accurate rendering of both static and dynamic digital objects: text, sound, and image files; multimedia documents, websites, databases, videogames, and so forth. The overall aim of the project was to facilitate universal access to our cultural heritage by developing flexible tools for accessing and storing a wide range of digital objects. KEEP also considered legal issues concerning the implementation of emulation-based systems and proposed solutions that comply with European and national copyright laws.

Digital objects are, of course, connected intimately with the technical environments in which they were created and used. In order to ensure long-term preservation of, and access to, digital material it is therefore essential to carefully record the hardware and software dependencies of each digital object in a preserved corpus. Typical information required includes details of the computer hardware, operating system, plugins, software libraries, and so forth, which a preserved object originally required, together with information on the hardware and software environment that was used during any subsequent preservation actions such as migration or emulation. There is substantial complexity involved in assembling and maintaining even the basic technical environment metadata required for subsequent emulation. This is a seriously time-consuming, detailed and complex task in its own right. To address this challenge, my colleague Janet Delve led the development of the TOTEM (Trustworthy Online Technical Environment Metadata) technical registry (<http://amzn.to/1JuKR3c>). The TOTEM generic data models, a database implementation, and a metadata schema have been combined with a compatible OWL ontology created within the PLANETS project.

Although engaged in public outreach, Cerf does not refer to the considerable efforts that have been made in this area by organizations like the Open Preservation Foundation, or the

Digital Preservation Coalition. Nor does he refer to any of the projects that have played a role in this space. The POCOS (Preservation of Complex Objects Symposia) project (<http://bit.ly/1LmMaPa>), concentrated not on the development of new tools or techniques, but was established to give global thought-leaders in research into the Preservation of Complex Objects an opportunity to share and thereby extend the body of knowledge on this topic through a series of symposia at locations across the U.K. The fundamental task facing these symposia was to present material of great technological and organizational complexity in a lucid, cogent, relevant and approachable manner so as to engage U.K. High-End Instrumentation researchers and practitioners in a wide variety of disciplines, as well as reaching those further afield in, for example, commerce, industry, cinema, government, games, and films classification boards.

The seminars were arranged around three general themes: visualizations and simulations; software art; and videogames and virtual worlds. Each of these domains involves the development, use, and manipulation of complex digital objects, and each presents a different, although clearly related, set of preservation challenges. A substantial and innovative dissemination program was established to ensure the various stakeholder communities obtained the maximum long-term value. This included the production of a peer-reviewed book (<http://bit.ly/1NxL3i8>) presenting the key outputs.

Considerable efforts have been made to address the need for integrated approaches to digital preservation that include organizations as well as tools. On the organizational side, archives provide an indispensable component of the digital ecosystem by safeguarding information and enabling access to it. Harmonization of currently fragmented archival approaches is required to provide the economies of scale necessary for general adoption of end-to-end solutions. There is a critical need for an overarching methodology addressing business and operational issues, and technical solutions for ingest, preservation, and reuse.

Many of the problems that appeared, just a few years ago, to be almost intractable are being brought under control.

To address this, the E-ARK project (<http://www.eark-project.com/>) working cooperatively with commercial systems providers, is creating and piloting a pan-European methodology for electronic document archiving. The emphasis is on not on “blue-sky” research but on synthesizing existing tools and techniques that have been developed over the last decade or so both commercially and within the context of publically funded research projects. National and international best practices, that will keep records and databases authentic and usable over time, are also being drawn together and integrated, with the intention of providing a single, scalable, robust approach capable of meeting the needs of diverse organizations, public and private, large and small, and able to support complex data types. E-ARK will therefore demonstrate the potential benefits for public administrations, public agencies, public services, citizens, and business by providing simple, efficient access to the workflows for the three main activities of an archive—acquiring, preserving, and enabling reuse of information.

The methodology will be implemented in various national contexts, using existing, near-to-market tools, and services developed by the partners. This will allow memory institutions and their clients (public- and private-sector) to assess, in an operational context, the suitability of those state-of-the-art technologies.

The practices developed within the project will reduce the risk of information loss due to unsuitable approaches to keeping and archiving of records.

The project will be public facing, providing a fully operational archival service, and access to information for its users. The project results will be generic and scalable in order to build an archival infrastructure across the EU and in environments where different legal systems and records management traditions apply. E-ARK will provide new types of access for business users.

E-ARK will pilot an end-to-end OAIS-compliant e-archival service covering ingest, vendor-neutral archiving, and reuse of structured and unstructured data, thus covering both databases and records, addressing the needs of data subjects, owners, and users. The pilot and methodology will also focus on the essential pre-ingest phase of data export and normalization in source systems. The pilot will integrate tools currently in use in partner organizations, and provide a framework for providers of these and similar tools ensuring compatibility and interoperability. A core component of the project is the integration platform that uses the existing ES-SArch Preservation Platform (EPP) application as an Archival Information System, which is already in productive deployment at the National Archives of Norway and Sweden. In order to achieve scalability, E-ARK will adopt a data management and storage layer for this tool on top of the proven open source Cloudera CDH4 distribution of Apache Hadoop, enabling storage and computational power to be seamlessly added to the system.

All in all, there is considerable reason to feel hopeful significant progress will continue to be made in digital preservation in the years ahead. Many of the problems that appeared, just a few years ago, to be almost intractable are being brought under control. The pace of technological change continues unabated, and this brings with it fresh preservation challenges, but despite rumors to the contrary, it appears the digital dark age will have to wait a little while longer.

I dedicate this column to Khaled al-Asaad. 

David Anderson (cdpa@btinternet.com) is the CiTECH Research Centre Director at the School of Creative Technologies, University of Portsmouth, U.K.

Copyright held by author.

The Profession of IT

Why Our Theories of Innovation Fail Us

Until we moderate our fascination with creating ideas, we will not achieve the rate of innovations we seek.

ONLY 1 IN 500 patents makes its inventor money, and businesses are awash in great ideas of dubious market value (only about 4% make money).¹ So why do people think innovation begins with a creative idea, is sold through an imaginative story, and diffuses through society because of novelty and merit? Innovators mobilize people to adopt ideas. Although they might start with idea creation, innovators focus mostly on other aspects: market offers, market testing, beta prototyping, production, sales, and customer-support infrastructures that companies use to get products adopted. In fact, 90% of innovation is in fostering adoption.^{1,4} Ideas are often stories invented after the fact to explain innovations that already emerged, as with the iPhone example discussed later in this column.

Yet the media telling of the story makes it sound as if ideation—the creation of ideas—is 90% of the work of innovation. Ideation has produced many inventions that never became innovations because no one adopted them. Many people are misled by stories that inaccurately equate innovation with invention. People who believe these stories put too little effort into adoption and are disappointed by their low success rates.^{1,5}

Bob Metcalfe, the inventor of Ethernet, tells the story of 3Com, a company he founded to make and sell Ethernets.⁴ His story is full of accounts



of his having to convince executives of companies they needed a network product they never heard of before, and then living up to the expectations he left them with. He spent one year developing his Ethernet idea and the next 10 years selling Ethernets. Sales do not matter in invention, but they matter in a big way in innovation. He summarized his effort with his famous saying, “Invention is a flower, innovation is a weed.”

Three Flawed Memes: Hindsight, Oversimplification, and Ideation

Innovation stories are tremendously influential in guiding our perceptions

(and interpretations) about innovation process and hence our actions. An example is the innovation pipeline: an innovation begins as an idea and flows through stages of prototyping, production, and marketing before arriving in the marketplace. Another is the innovation funnel: a set of ideas is progressively winnowed by reviews, prototype tests, and market tests, until the few with greatest merit make it to the marketplace. A third example is network diffusion: an innovator injects an idea into a social network, where it spreads out across the communication channels of the network until everyone has a chance to adopt it. A fourth is the

innovation cell, a protected pocket of innovators spinning off ideas into the surrounding environs.

These stories are all “sticky.” It is easy to form a mental picture of a pipeline with ideas flowing through it, or a series of progressively narrower funnels flowing one into the next, or waves of adoption washing through a network, or ideas spinning off a roundtable. They are memes that hold our attention. However, these sticky stories contain flaws that lead the unwary into actions that do not work.

The first flaw is that our stories about innovation are retrospective. In hindsight, we can see all the actions involved in an innovation and describe a pattern they seem to follow. But as innovators “in the trenches” we experience things quite differently. Every action seems to have an unpredictable outcome and we cannot tell if it leads us closer to our desired innovation. So many things depend on actions of other people. Doubt and uncertainty are irreducible. You cannot “see” where you are in the pipeline, funnel, network, or cell; only future historians can pass those judgments. Bob Metcalfe did not find executives ready and waiting for Ethernets; he constantly had to confront their doubts about a product they never heard of before, persuade them of its benefits to their companies, and convince that he would be a trustworthy supplier of Ethernets. If you try to form an innovation plan around the pipeline, funnel, network, or cell model, your plan will almost always fail because the people involved cannot tell where they are in your imagined structure.

The second flaw is that our stories about innovation are tremendously oversimplified. The stories present the successful actions of innovators as deliberate, considered, and sometimes inspired choices by persons able to make sense of the situation and control it. Their individual actions fit together into neat causal chains whose outcomes align with the innovator’s intentions. Bob Metcalfe could say afterward that he visited the “ABC” company, overcame their doubts, and got their order for Ethernets. But when he was there nothing was certain. He had to learn their doubts and concerns,

Innovations are new practices adapted in a community, which displace other practices.

find a way to show them Ethernet took care of an important concern, and build trust in him as the salesman and supplier. How did he learn their doubts? Discover their unmet concerns? Construct a proposal on the spot for how they could try Ethernets at acceptable risk? Lead them to the conclusion he was sincere, competent, and had their best interests at heart? Bob will tell you he often had no idea what it would take to close a deal, and in many cases he failed to close a deal. He did not feel in control. The best he could do is approach each encounter with a sense of confidence he could lead the conversation to a successful conclusion. How did Bob cultivate a mood in himself that disposed him toward success?

The third flaw is all the innovation models assume an idea starts the process. Someone’s idea triggers the pipeline, or feeds the funnel, starts a wave in the network, or seeds the cell. What if most innovations do not begin with an idea? For example, social innovations such as Mothers Against Drunk Driving or more recently legalized marijuana and same-sex marriage welled up in popular opinion and swept many people along. The leaders of these movements report they were reacting to injustices and not creating ideas. Many technology innovations seemed to well up out of circumstances of the time without anyone claiming to have put an idea into action. For example, the iPhone and smartphones that imitated it seemed to catch on because “the time was right” even though many previous similar attempts had failed. Blogging seemed to well up without anyone inventing blogging or even stepping

forward afterward when there was an opportunity to claim credit and be recognized in Wikipedia. These examples illustrate the larger pattern: most innovations “emerge” in the practices of communities and are not caused by someone’s good idea.³ In fact, most of what we call “ideas” behind innovations are actually stories made up in hindsight to explain the practices already emerging.

With these flaws, it is difficult to see how careful strategic planning, innovation process management, and charismatic leadership can work consistently well. In a review of Barbara Tuchman’s *March of Folly: From Troy to Vietnam*, written many years ago, Gordon Wood wrote there was but one big lesson of history: “Nothing ever works out quite the way its managers intended or expected.”⁶ This larger lesson unfortunately has not yet made it into our dominant narratives about innovation.

Sticky innovation stories are easy to recall and fun to retell. The only way to displace these stories is to interpret innovation with new and better stories. You need a new story to dislodge a story.³

If Not Ideation, Then What?

If ideation is a relatively easy 10% of your effort, how should you spend the other 90%? What should you do? We like the story from Fernando Flores about innovation emergence.² This story begins with the notion that innovations are new practices adopted in a community, which displace other practices. Emergence of a new practice begins when someone makes a proposal to combine existing practices in a new way to meet an unmet concern. The proposal is contingent on many factors: technologies and practices already in existence, unmet concerns in the community, the proposer’s timing and choice of concern to address, the social power of the proposer’s network, and the strength of the opposition.² Bringing an innovation into reality, therefore, is unpredictable and relies on explicitly working for adoption.

For example, Steve Jobs did not simply create iPhone in a flash of genius and sit back and wait for the profits to roll in. His contribution was to believe in a vision of a lightweight portable phone that could be custom-

ized to its owner’s detailed personal preferences, and to mobilize a business network to make it happen. The Apple company invested a lot of work to transform the iPhone vision into an adopted technology. The transformation was contingent on the existence of other components already, or soon to be, in place. Apple worked with suppliers to build smaller and more energy-efficient components such as hard disks, touch displays, scratch-resistant gorilla glass, sensors for GPS and motion, and batteries. Apple adapted the operating system MacOS into iOS that would manage an interface presenting a large collection of user-chosen apps. Apple adapted the iTunes store into an apps store and cultivated a network of a million programmers to populate it with downloadable apps. Apple worked with the telecommunication companies, initially AT&T, to create data plans within the cellular phone network. Apple worked with professional product designers and marketers to position the iPhone as a lifestyle enhancer rather than a mobile phone. The iPhone was contingent on all these components and the business deals that made them work. Its adoption took a great deal of business and political skill. Yet the popular stories focus on Jobs alone and ignore the huge amount of work Apple invested to get the iPhone widely adopted. You will find similar stories in all the other technology companies. The standard stories focus on the ge-

Innovations emerge in spaces of practices, which are constantly drifting and changing as powerful forces converge and conflict.

nius of the founders and ignore the hard work they put into adoption.

Six Fundamental Skills

The six skills in the accompanying table nicely summarize what innovators do.² Innovations emerge in spaces of practices, which are constantly drifting and changing as powerful forces converge and conflict. Innovators propose changes of practice and shape their adoption. In the swirl of the forces nothing is certain. Multiple people are likely to come up with competing proposals at about the same time, each responding to the sense of an unmet concern that anyone who cares to listen can detect. These six skills are based on your ability to listen for concerns, histories, movements of social power, barriers, moods, reactions to offers, and followers in networks. They depend only loosely on communicating your ideas or telling your stories.

Offering and mobilizing are the core skills. Your offers are proposals to take designs into social movements. Can you make offers that intrigue people with new possibilities to address their (often unspoken) concerns and do not seem too risky? Can you turn your networks (or build a network) into a following of people who commit to the new practices the offer brings? Do you understand who will resist or support and what actions will harness the power of the network to shape the emerging new future?

Detecting, appropriating, navigating, and surfing all support the core skills. Detecting means to sense an unmet concern and form an inkling that you can do something about it. Appropriating means to immerse in related domains to discover marginal practices and interpretations that can help you with your inkling. Navigating means to move toward a goal in a complex and uncertain world; the metaphor recalls seafaring explorers in open oceans who must respect the power of the waves and the limitations of their crews, avoid storms, and deal with emergencies. Surfing means to ride waves that move in the direction you seek and keep your balance when turbulent network forces buffet you.

To be an innovator, learn these six skills. C

References

1. Denning, P. and Dunham, R. *The Innovator's Way*. MIT Press, 2010.
2. Denning, P. and Flores, F. Emergent innovation. *Commun. ACM* 58, 6 (June 2015).
3. Freedman, L. *Strategy: A History*. Oxford University Press, 2013.
4. Metcalfe, R. Invention is a flower, innovation is a weed. *MIT Technology Review* (Nov. 1999); <http://bit.ly/1Pt5ygT>.
5. Read, S., Sarasvathy, S., Dew, N. Wiltbank, R., and Ohlsson, A.-V. *Effectual Entrepreneurship*. Routledge, 2010.
6. Wood, G. History lessons. Review of Barbara Tuchman, *March of Folly: From Troy to Vietnam*. New York Review of Books (Mar. 29, 1984).

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.

Nicholas Dew (ndew@nps.edu) is Associate Professor of Strategic Management at the Naval Postgraduate School in Monterey, CA. He researches entrepreneurship and is a co-author of the textbook *Effectual Entrepreneurship*. The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

Copyright held by authors.

Six skills for achieving adoption.	
Offering	Making proposals of combinations of existing practices and technologies to meet an unmet concern, then observing reactions and modifying the offer to be more attractive.
Mobilizing	Getting a social network to back your offer and help make it happen; depends on the social power of the network and on your personal power.
Detecting	Sensing an opportunity in an unmet concern or a disharmony; being unsettled by an anomaly.
Appropriating	Investigating related domains to understand their history behind their concerns, and to discover existing practices that might help with the concern you are dealing with.
Navigating	Finding your way amidst conflicting waves of possibilities, coming to your goal without having a detailed plan to get there.
Surfing	Finding waves of possibilities moving toward your goal and riding with them, retaining your balance and center when hitting turbulence.

Computing Ethics

Coupled Ethical-Epistemic Analysis in Teaching Ethics

Critical reflection on value choices.

ETHICS IS AN important component of STEM education as illustrated by the fact that ABET accreditation requires proof of training in ethics for engineering fields. But what range of knowledge and skills are required for integrity? Covering codes of ethics does not teach ethical sensibility or the ethical reasoning skills required for research and professional integrity.¹⁰ So what should be included in ethics education for engineers?

Engineers like to “get it right,” and engineering education should focus on knowledge and skills to ensure that engineering work meets the highest standards of empirical and theoretical adequacy. This is what Mason⁵ refers to as a “covenant with reality.” But this does not ensure the best interests of the broader community and the environment are served. “Getting it right” requires understanding what is epistemically *and* ethically salient to the issues being addressed, communicating information and sharing technologies in ways users can understand, and supporting responsible application.

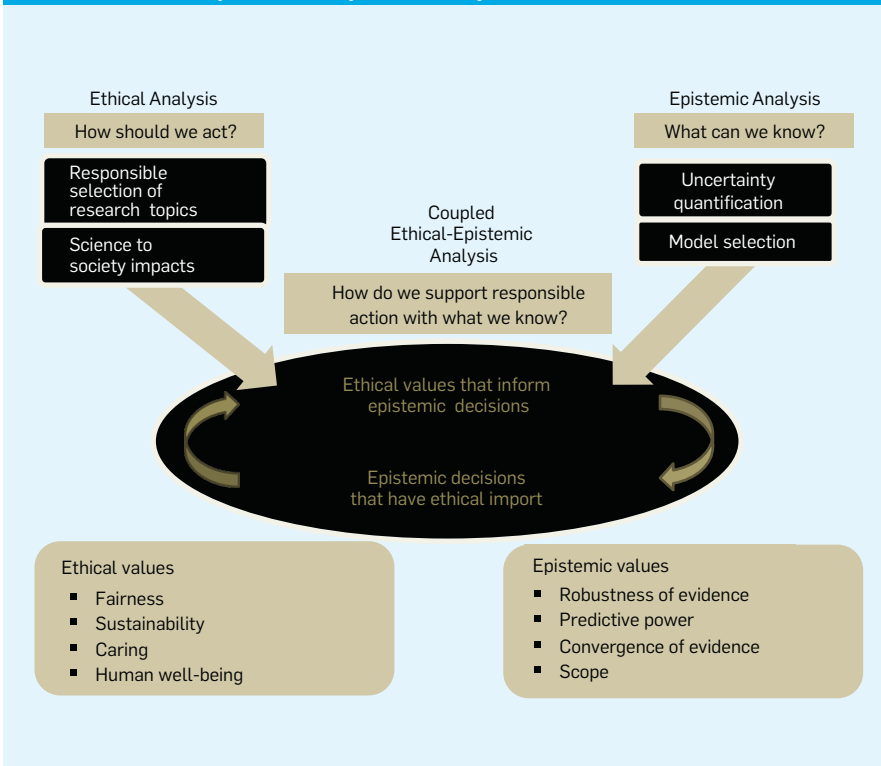
Integrity in engineering research and practice depends on values. This is especially important in the responsible conduct of research. “Computer experts,” Deborah Johnson reminds us, “aren’t just building and manipulating hardware, software, and code, they are building *systems that help to achieve important social functions, systems that constitute social arrangements, relationships, institu-*



tions, and values.” Johnson claims computer experts can “facilitate and constrain behavior, and materialize social values.”⁴ Social media pages, for instance, can be designed to track Internet browsing histories—even

after the user logs off—to tailor advertisements or software or enable companies to engage in workplace surveillance. Is this OK and why? An important but often overlooked component of ethics education is the abil-

A visualization of coupled ethical-epistemic analysis.



ity to identify values and appreciate the ethical dimensions of the broader impacts from actions computing professionals might take. Coupled ethical-epistemic analysis is an important lens for critical reflection on value choices. This column explains why and how.

Values serve as a guide to action and knowledge. They are relevant to all aspects of scientific and engineering practice, including discovery, analysis, and application. Cognitive scientists have found values to be inextricable components of STEM research. Paul Thagard explains, “the decisions that scientists and others need to make about what projects to pursue, what theories to accept, and what applications to enact will unavoidably have an emotional, value-laden aspect,” and concludes, “the best course is not to eliminate values and emotions, but to try to ensure that the best values are used in the most effective ways.”⁹ Decisions about whether there is robust evidence for a claim (an epistemic value) can for example be influenced by possible effects on human well-being (an ethical value).

To use the best values effectively, I advocate ethics education in STEM

fields be based on analysis of values in four dimensions of research and practice. Various types of values can be involved in each domain including ethical values (the good of society, equity, sustainability), aesthetic values (simplicity, elegance, complexity), or epistemic values (predictive power, reliability, coherence, scope).

► What is a good basis for the selection of research topics?

► What counts as evidence and what constitutes robust evidentiary support?

► What is the likelihood that a model, hypothesis, or theoretical explanation will provide convincing explanation?

► Are epistemic and ethical values relevant to applying results to other research problems or to social problems (for example, via decision-support)?

Coupled ethical-epistemic analysis

Users (and modelers) are not always aware of relevant values.

will help answer these questions (see the accompanying figure).

Values transparency and analysis is central to the National Science Foundation’s Sustainability Research Network on Sustainable Climate Risk Management (SCRiM, <http://scrim-hub.org>). Coupled ethical-epistemic analysis has helped identify new and refined research topics, and informed modeling for multi-objective, robust decision making.⁸ An example is the debate over ice sheet data in modeling sea level rise. Whether ice sheet melt data is sufficiently robust for sea level rise projection models was debated by Intergovernmental Panel on Climate Change (IPCC) scientists. They had to balance evidential robustness and predictive power (epistemic values), a decision that remains controversial despite improvements in scientific understanding and modeling of ice sheet dynamics (for example, Chang et al.²). Balancing these epistemic values has ethical implications: a decision to not include ice sheet data might result in underreporting future sea level rise.⁷ Does the value assigned to evidential robustness outweigh the impact on predictive power? Also at stake is whether ice sheet data helps the wider community understand the impact of climate change mitigation and adaptation decisions. Epistemic and ethical values are coupled.

Mason⁵ urges that modelers adopt a covenant with reality and a covenant with values to make models faithful not only with relevant facts, but also with the values of intended model users. Fleischmann and Wallace³ augment Mason’s work with a third covenant, values transparency in modeling, which “not only allows the client to assess whether the model conforms to the first two covenants, it allows the client to assess when the model is misbehaving or malfunctioning.” They suggest that transparency can allow clients to avoid, “circumstances where the model errors might lead to negative consequences for those affected by the model.”³

I applaud and support these positions, and add a further step. Users (and modelers) are not always aware of relevant values. Ethics education designed to promote values identification and analysis is a key element

Epistemic and ethical value decisions have important social implications.

of this ability, but values analysis requires techniques to identify values. An example is values-informed mental models (ViMM).¹ ViMM is an empirically grounded method for gathering information about individuals' beliefs and inferences about an issue, and elicits values in addition to beliefs and inferences. It is designed to provide the transparency regarding users' and modelers' values in a decision situation.

Epistemic and ethical value decisions have important social implications. This fact is reflected in the National Science Foundation broader impacts criterion and its view of the role and purpose of ethics education: "Ethics education is particularly critical to the science and engineering community as it faces an increasingly competitive funding environment; rising collaboration with international colleagues who may follow different guidelines; and growing recognition of the relevance of science and engineering to social, economic, and ethical issues of wide public and political interest."⁶

A twofold approach should be used in training for coupled ethical-epistemic analysis. This requires some changes in ethics education for engineers, philosophers, and social scientists. Engineers must become more aware of the ethical and epistemic values embedded in all components of their work as well as their salience and role in research and practice through training for values transparency and coupled ethical-epistemic analysis skills. This better prepares engineers to understand and take responsibility for the epistemic and/or ethical import of the values embedded in their work, so values of the users and those

impacted by use become essential elements in that work.

Parallel training of philosophers and social scientists is needed so they can assist with coupled ethical-epistemic analysis. Techniques such as ViMM assist in identifying the range of relevant values, and serve as a basis for careful analysis of the implications of value choices. This approach works best when trained philosophers and social scientists are embedded in research teams, collaborating with engineers and scientists. The SCRiM network illustrates the benefits of this transdisciplinary approach to coupled ethical-epistemic analysis.

Engineering ethics training involving coupled ethical-epistemic analysis helps with "getting it right" in all senses. 

References

1. Bessette, D. et al. Values-informed mental models: A new tool for decision support and analysis. *Nature, Climate Change*. Forthcoming.
2. Chang, W. et al. Probabilistic calibration of a Greenland Ice Sheet model using spatially-resolved synthetic observations: Toward projections of ice mass loss with uncertainties. *Geoscientific Model Development*, 7 (2014), 1933–1943; DOI: 10.5194/gmd-7-1933-2014.
3. Fleischmann, K.R. and Wallace, W.A. A covenant with transparency: Opening the black box of models. *Commun. ACM* 48, 5 (May 2005), 93–97.
4. Johnson, D.A. Computer experts: Guns-for-hire or professionals? *Commun. ACM* 51, 10, (Oct. 2008), 24–26. DOI: 10.1145/1400181.1400190.
5. Mason, R.O. *Morality and models. Ethics in Modeling*, W.A. Wallace, Ed. Elsevier, Tarrytown, NY, 1994.
6. National Science Foundation. Status Update on NSF Implementation of Section 7009 of the America COMPETES Act (ACA): Responsible Conduct of Research Advisory Committee for Business and Operations Spring Meeting, May 2008; <http://1.usa.gov/1G6c0IR>.
7. Rahmstorf, S. A new view on sea level rise. *Nature Reports Climate Change*, 4 (2010), 44–45; DOI: 10.1038/climate.2010.29.
8. Singh, R., Reed, M., and K. Keller, K. Many-objective robust decision making for managing an ecosystem with a deeply uncertain threshold response. *Ecology and Society*. Forthcoming.
9. Thagard, P. *The Cognitive Science of Science: Explanation, Discovery, and Conceptual Change*. MIT Press, Cambridge, MA, 2012.
10. Tuana, N. An ethical leadership developmental framework. *A Handbook of Ethical Educational Leadership*. C. Branson and S.J. Gross, Eds. Routledge, 2014.

Nancy Tuana (ntuana@psu.edu) DuPont/Class of 1949 Professor of Philosophy and Women's Studies at the Nancy Tuana Directorship of the Rock Ethics Institute, Penn State University, State College, PA.

This work was partially supported by the National Science Foundation through the Network for Sustainable Climate Risk Management (SCRiM) under NSF cooperative agreement GEO-1240507. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. The theory and practice of coupled ethical-epistemic analysis was informed by dialogue with members of the SCRiM team. Thanks to Rachelle Hollander and several anonymous reviewers for insightful editorial feedback on the column.

Copyright held by author.

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



DOI:10.1145/2835959

George V. Neville-Neil

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

Code Vicious Pickled Patches

On repositories of patches and tension between security professionals and in-house developers.

Dear KV,

I recently came upon a software repository that was not a repo of code, but a repo of patches. The project seemed to build itself out of several other components and then had complicated scripts that applied the patches in a particular order. I had to look at this repo because I wanted to fix a bug in the system, but trying to figure out what the code actually looked like at any particular point in time was baffling. Are there tools that would help in working like this? I have never come across this type of system before, where there were more than 100 patches, some of which contained thousands of lines of code.

Pick a Peck of Pickled Patches

Dear Pickled,

The appropriate tools for such a system do exist, but they require a background check in many states in the U.S. and are banned outright in the more developed countries of the world.

What you are faced with is a project that probably ought to have forked the projects it was working with, but, instead, started with one patch, then two patches, then four patches, until you have what you see before you. When a project is developing quickly and has not started out with the understanding that it is a significant derivative work, the proper use of source code control tools may not occur soon enough in the development process. It requires



discipline to spend some upfront time thinking about how to integrate existing code with new development, and if that work does not get done early, it often does not get done at all.

If you want to fix a single bug in the system, I suggest you contact the developers, because they should understand what they have done—and the mess they are in—sufficiently well to be able to address your problem more quickly than you can sort out what they have done with their system. On the other hand, if you need to do significant work on the system you are looking at, you may have to take more

extreme measures.

You mention that the project you are looking at is a repo of patches for use with another project. That being the case, you need to lay down what I will refer to as a base track. The ultimate, upstream software the system is based on has to be the base layer. It also needs to be placed into a source code control system that allows you to update that base layer from the ultimate source. With the base layer in place, you should create a branch per patch from the derivative system. You could do this blindly, but it is probably best—although possibly quite frustrat-

ing—to read through the project build scripts carefully beforehand. I will wager that some of the patches you see in the derived project are not standalone, but instead depend on each other to fix some underlying bug or to implement a complex feature of the system. Once you have collected the patches into groups, you can then create the patch branches and import the patches from the derived repo. With code now properly contained in a source code control system, you should branch the base layer into its own development branch. Never directly modify the base layer in your project repository, as this will make integrating changes from the ultimate upstream repository nearly impossible. Let me say that again, *never* directly modify the base layer in your project repository, as this will make integrating changes from the ultimate upstream repository nearly impossible.

For the base layer you will always have at least two branches, the pristine branch that includes only changes coming from the upstream, and the development branch that takes code from the pristine branch and merges it with the patches. You can now integrate patches into the development branch and test them one by one to make sure they work individually before trying to make them all work together. KV often goes on about testing, but in your case it bears a good deal of emphasis. Unless you are in close communication with your upstream providers, you have no idea how they are testing these patches, and accepting them wholesale without incremental tests is a great way to wind up paying a lot of money to someone who puts you on a couch and asks you questions about your childhood.

Of course your best bet is to find the people creating patches of patches and then give them this response. You might want to inscribe it in golden fire on tablets first, but that is up to you.

KV

Dear KV,

Many organizations may be institutionalizing tension between security professionals and in-house developers. In organizations where the security professionals and the in-house de-

Let me say that again, never directly modify the base layer in your project repository.

velopers are in different organizational units and the security professionals have the ultimate responsibility for security, it is natural for the security perspective to dominate dialogue between these two camps.

Security professionals have a clear mandate to protect the organization, and their toolset necessarily includes rigidly standardized computer settings and policies and enforcement mechanisms.

In-house developers frequently require exceptions to security policies because their work may require access to software tools excluded from the standard office suite (for example, integrated development environments), security testing tools (such as OWASP Zap), and/or elevated privileges. Also, developers who work with multiple projects (as most of us do) may need multiple virtual machines in order to manage multiple development-project contexts.

As a software developer who attempts due diligence with respect to security, I am often disappointed when security professionals seem to pay so little attention to the concerns of in-house software developers. When security policies are inflexible, useful tools or approaches are disapproved, and in-house developers are unable to apply software development best practices, the organization is not necessarily more secure.

I am hoping you will consider using your voice to stimulate debate on this topic. Please consider a blog article discussing how security professionals might collaborate with in-house developers to the benefit of all. You might consider discussing alternative approaches for reconciling corporate policies (or USGCB) with develop-

Calendar of Events

December 1–4

CoNEXT '15: Conference on Emerging Networking Experiments and Technologies, Heidelberg, Germany
Sponsored: ACM/SIG,
Contact: Felipe Huici
Email: felipe.huici@neclab.eu

December 5–9

MICRO-48: The 48th Annual IEEE/ACM International Symposium of Microarchitecture, Waikiki, Hawaii,
Contact: Milos Prvulovic,
Email: milos@cc.gatech.edu

December 7–10

UCC '15: 8th International Conference on Utility and Cloud Computing,
Contact: Ashiq Anjum
Email: a.anjum@derby.ac.uk

December 07–11

Middleware '15: 16th International Middleware Conference,
Sponsored: ACM/SIG,
Contact: Rodger Lea,
Email: rodger.lea@gmail.com

2016

January

January 10–13

SMA '16: The 4th International Conference on SmartMedia and Applications, Danang, Viet Nam,
Contact: Youngchul Kim,
Email: yckim@chonnam.ac.kr

February

February 14–17

TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction, Eindhoven, Netherlands,
Sponsored: ACM/SIG,
Contact: Saskia Bakker,
Email: s.bakker@tue.nl

February 21–23

FPGA'16: The 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Monterey, CA,
Sponsored: ACM/SIG,
Contact: Deming Chen,
Email: dchen@illinois.edu

ers executing security probes against their own code, for example with OWASP Zap.

Zapped

Dear Zapped,

Once upon a time I was one of those in-house security professionals, I still have cards with my title, “Paranoid,” printed right on them. I never keep old business cards, but I kept those because they are the only ones I have ever had that were that honest. I could print more honest cards, but then I could not hand them out in polite company.

I have never been a fan of blanket bans of, well, just about anything, and definitely not software that would help developers produce better and more secure code. Blanket bans usually come from a misguided belief that the rules of engagement can be defined by a small group and that if everyone sticks to those rules nothing can possibly go wrong. That belief is not only mistaken, but incredibly dangerous. Any security team with a clue knows you set out general guidelines and then work with the development group in an advisory role to ensure the guidelines make sense. Only an idiot would create a set of rules that apply to both the development and the accounting teams. Alas, the world is filled with idiots as well as those who simply fear the unknown.

A cursory glance at the software you mention does not show it to be more dangerous than any other piece of software a developer might use, right down to a compiler or a debugger. What is important in any of these discussions is an ability to come up with reasonable boundaries and safeguards so the software in question can be used without causing accidental damage to the systems. Reasonable guidelines are developed in concert with the teams doing the work. Members of a good security team know they are playing a supporting role and they must gain the trust of the people they are working with in order to do their job.

When developers need to do something considered particularly dangerous, for instance attack their own systems, it often makes sense to do this in a lab environment, at least at first. Unleashing the latest penetration testing toy on the company website might

The case for elevated privileges is another one that comes up frequently between security and development teams.

be amusing, in much the way that some people consider car accidents fun to watch, but it is not going to improve site reliability or security. Large companies have dedicated teams of pen testers—often within the security team. Having a small group that can work with developers to create appropriate security test scenarios and schedule them to run at times that are convenient for testing is a good solution if sufficient resources are available. Startups and smaller companies will need to have their developers do this type of work, just as they have them do everything else, from coding to testing and documentation. The rise of cheap cloud computing should actually help in this area because a service can be cloned, walled off, and “attacked” with tools without harming the active service. One has to be careful with this type of testing, as some cloud providers may flag your attack testing as a real attack and shut you down. You mention virtual machines in your letter, and this is another way to achieve the same ends as the cloud solution, by spinning up a cluster of virtual machines on a virtual network on a large server dedicated to security testing.

The case for elevated privileges is another one that comes up frequently between security and development teams. The basic problem is that software systems, including operating systems, but also extending to databases and other pieces of critical software, are not engineered with the idea of security in mind. The now-famous XKCD comic (<https://xkcd.com/149/>) about the *sudo* command pretty much says it all. Most software is written to run

either as an individual user, who usually has sufficient privilege to run the service but not enough to test or debug it, or as root. When developers are debugging or testing, they want to run the software as “root” or with a similar superuser-type power because then “it just works.” While these two levels of power are easy to understand—user vs. root—and while they do underpin a lot of modern computing thanks to their use in the original Unix operating system, they are insufficiently expressive. But that is a topic for another, much longer discussion.

Short of rewriting a ton of existing software, we come down to needing test environments, walled off from most of the rest of the system, or requiring special commands to get external access, to give developers a safe sandbox in which to work.

If elevated privileges are absolutely required to get a job done, that privilege must have a timeout. To me it seems reasonable to give a developer root powers on a production box so long as they are working with one other person, and all their commands are written to a file. The *sudo* program can actually do all of this. I would either set the timeout for a day or to when the problem was fixed, whichever came first. That statement is not meant to be a policy to be blindly adopted by all of my willing thralls; it is simply an example of how security and development teams can work together to get a job done.

KV

Related articles on queue.acm.org

Painting the Bike Shed

Kode Vicious

<http://queue.acm.org/detail.cfm?id=1557897>

Security in the Browser

Thomas Wadlow

<http://queue.acm.org/detail.cfm?id=1516164>

Patching the Enterprise

George Brandman

<http://queue.acm.org/detail.cfm?id=1053344>

George V. Neville-Neil (kv@acm.org) is the proprietor of Neville-Neil Consulting and co-chair of the ACM *Queue* editorial board. He works on networking and operating systems code for fun and profit, teaches courses on various programming-related subjects, and encourages your comments, quips, and code snips pertaining to his *Communications* column.

Copyright held by author.

Broadening Participation

Increasing the Participation of Individuals with Disabilities in Computing

Lessons learned from a decade of practice.

LACKING DIVERSITY ON an engineering team, we limit the set of solutions that will be considered and we may not find the best, the *elegant* solution.”⁸ This insight, as expressed by computer scientist William A. Wulf, has motivated efforts to encourage women and racial/ethnic minorities to pursue computing careers. However, the underrepresentation of individuals with disabilities in these disciplines has been largely ignored. Not so at the University of Washington. Our *AccessComputing* initiative has worked tirelessly in this domain at a national level and, over a decade, has achieved promising outcomes in increasing the number of people with disabilities completing degrees and pursuing careers in computing fields. *AccessComputing* efforts have increased the capacity of key players—postsecondary institutions, professional organizations, and industry—to fully include individuals with disabilities and helped individuals with disabilities move toward computing degrees and careers.¹ In this column, we share examples of project activities and lessons learned to encourage other educators and employers to join our efforts in increasing diversity in computing disciplines by engaging more individuals with disabilities in these professions.



A student engages with a computer through a touchscreen.



The Need

Increasing the participation of individuals with disabilities in computing fields is not just a matter of quantity. Engaging this population can help us meet the mandate the National Science Foundation (NSF) to support “the best ideas from the most capable researchers and educators.”⁶ This applies to the design of information technology (IT)

to be accessible to a broad audience and preparation of professionals with accessibility skills to help employers meet legal mandates under the Americans with Disabilities Act (ADA) and address the needs of an aging population. Individuals with disabilities are uniquely suited to contribute to this effort. However, people with disabilities face challenges in pursuing science, technology, engineering, and mathematics (STEM), completing degrees and succeeding in careers. Student veterans with post-traumatic stress disorder and traumatic brain injury,

“signature injuries” of recent conflicts, can face additional challenges related to social adjustment, finances, and adjustment to their disabilities.³

One ongoing challenge to address is ensuring the availability of information technology that is accessible to everyone. There are well-documented solutions for making IT accessible; they help with access to keyboards and mice for those with limited mobility, to content within images for students who are blind, and to audio content for students who are deaf or hard of hearing.⁷ Students with learning disabilities or who have difficulty with focus, reading comprehension, or other cognitive issues can benefit from inclusive instructional approaches that do not make them stand out as different.⁵ The success stories of the relatively few individuals with disabilities in computing fields demonstrate the opportunities for those who develop academic and self-determination skills and who overcome inaccessibility in facilities, curricula, and IT and inadequate academic support. However, the biggest barrier is often lack of encouragement by individuals with whom they interact, including educators and employers. Integrating content about how IT can improve the lives of individuals with disabilities in computing courses can encourage students with disabilities and others interested in social impact who may not otherwise consider the computing field.

Increasing the participation of individuals with disabilities in computing fields is not just a matter of quantity.

AccessComputing: It Is About Engaging Professionals

AccessComputing was created in 2006 to address the need to increase diversity in computing by including more individuals with disabilities. Funded by the NSF, the initiative engages individuals with disabilities as well as those who support, educate, and employ them, in sustainable practices for making computing disciplines more welcoming and accessible to people with disabilities. Project objectives fall into two areas:

- Organizational capacity development: Increase the capacity of postsecondary institutions, organizations that broaden participation in computing, and industry to fully include individuals with disabilities in computing fields.

- Individual participation: Implement evidence-based practices (for example, workshops, mentoring, internships) to increase the number of individuals with disabilities moving through critical junctures to computing careers.

Bringing together multiple stakeholder groups is a key to *AccessComputing*'s success. This starts with leadership. As its leaders, we represent the Department of Computer Science and Engineering and the DO-IT (Disabilities, Opportunities, Internetworking, and Technology) Center, both at the University of Washington. Together, we have a depth of experience in computer science as well as access issues and solutions for individuals with disabilities. In addition, *AccessComputing* engages 58 project partners who represent a diverse set of postsecondary institutions and computing organizations. Three of our partners, National Technical Institute for the Deaf at Rochester Institute of Technology, Gallaudet University, and Landmark College, serve students with disabilities as a specific mission. Institutional representatives engage online and onsite to explore promising practices and develop resources that enhance understanding and promote replication at educational institutions, in professional organizations, and in the computing industry.

AccessComputing leaders have also worked with 175 collaborators, some who have hosted student interns and others who have hosted events for students with disabilities or faculty and staff at their sites. We have awarded 75 mini-grants to collaborators to help fund their activities that support the project objectives. We have worked with 78 computing departments and professional organizations to help make them more accessible and welcoming to people with disabilities. Examples include working with 21 computing departments to evaluate and help improve the accessibility of their websites; with ACM SIGCHI in its efforts to make CHI conferences more accessible; with our partner, Center for Minorities and People with Disabilities in IT (CMD-IT), in its efforts to make the Tapia Conference, Academic Careers Workshop, and its other activities more accessible and welcoming



A student uses a speech interface to gain the functionality of a keyboard and mouse.

to people with disabilities; and EDUCAUSE in making its events more accessible and to address disability-related topics in its conference exhibits and presentations. We have also worked with partners and collaborators to expand the *AccessComputing* online searchable knowledge base of 487 questions and answers, case studies, and evidence-based practices that can help leaders and practitioners make computing disciplines more inclusive.¹

AccessComputing: It Is About Engaging People with Disabilities

Over the past 10 years almost 450 students around the U.S. have participated in the *AccessComputing* Team, an online community of computing students. They have engaged with peer and career mentors, in research and industrial internships, and professional conferences and other career development activities. Of the 410 *AccessComputing* internships, 110 were research internships at various universities. The outcomes of participation are impressive. For example, in a follow-up survey, 151 *AccessComputing* interns rated six statements on a 5-point scale, where 1 is “strongly disagree” and 5 is “strongly agree.” The items and average responses are represented in the accompanying table.

These results are consistent with an earlier study where STEM interns with disabilities reported increased motivation to work toward a career, knowledge about careers and the workplace, job-related skills, ability to work with supervisors and coworkers, and self-advocacy skills.² Profiles of 17 *AccessComputing* Team members are highlighted on the Choose Computing web page of our website.¹

We measure project impact by gathering participant feedback and by tracking the progress of student participants. *AccessComputing* participants are tracked through critical junctures⁴ to computing degrees and careers. Some *AccessComputing* participants participate in the *AccessSTEM/AccessComputing/DO-IT* Longitudinal Transition Study (ALTS), an ongoing research study that tracks the progress of students with disabilities through critical junctures to post-secondary degrees and careers. Analy-

Respondent ratings of gains from internships: Six questions on a 5-point scale (1 = strongly disagree, 5 = strongly agree).

Survey Statement:	Average Response
I am more motivated to study and work toward a career.	4.33
My knowledge of my career interests has increased.	4.27
I have learned the skills I need to succeed in specific job tasks.	4.08
I have learned skills I need to effectively work with co-workers.	4.00
I have learned the skills I need to effectively work with supervisors.	3.99
I learned about disability-related accommodations I may need at work.	3.67

sis of results supports the efficacy of evidence-based interventions by documenting higher levels of success—in terms of high school graduation, college attendance/persistence, computing majors/degrees and careers—of participants with disabilities than of people with disabilities in comparison groups.

Lessons Learned

A decade of *AccessComputing* efforts has resulted in students with disabilities moving through critical junctures to computing degrees and careers; increased capacity of computing faculty and professional organizations to include participants with disabilities; useful resources for educators and employers; and sustained engagements of stakeholder groups. Others who want to host a project like *AccessComputing* can benefit from what we

have learned. For example, we have learned that positive outcomes result from cooperative efforts between organizations focused on computing and those focused on disability and that new activities benefit from facilitation by experienced staff and provision of external funding for initial activities on which institutionalization can build.

Many lessons have been learned with respect to the participation of students with disabilities in computing fields.

- Individuals face common issues as well as unique challenges related to specific disabilities, in both academic and non-academic (for example, self-advocacy) arenas.

- Motivational activities should be undertaken to recruit students without initial interest in computing.

- For students with disabilities interested in computing, comprehensive preparation and retention interventions should be undertaken in order to produce more positive outcomes than isolated efforts.

- We need to develop more leaders and role models with disabilities in computing fields by undertaking more efforts like *AccessComputing*.

- Aggressive retention efforts are essential because the percentage of students in computing majors who have disabilities decreases as education levels increase.

Much has been learned about organizational capacity development as well.

- Computing departments, programs, and professional organizations should be welcoming and accessible to people with disabilities,

One ongoing challenge to address is ensuring the availability of information technology that is accessible to everyone.



Association for
Computing Machinery

ACM Conference Proceedings Now Available via Print-on-Demand!

Did you know that you can now order many popular ACM conference proceedings via print-on-demand?

Institutions, libraries and individuals can choose from more than 100 titles on a continually updated list through Amazon, Barnes & Noble, Baker & Taylor, Ingram and NACSCORP: CHI, KDD, Multimedia, SIGIR, SIGCOMM, SIGCSE, SIGMOD/PODS, and many more.

For available titles and ordering info, visit:
librarians.acm.org/pod

Bringing together multiple stakeholder groups is a key to *AccessComputing's* success.

consulting individuals with disabilities and/or disability services to make accessibility a centerpiece of the organization. Being responsive to legal obligations means designing for accessibility and providing instructions for requesting accommodations.

- Employers need to become more aware of how they can benefit from hiring employees with disabilities because addressing accessibility challenges in their daily lives has made this group good problem solvers.

- Both students and employees with disabilities provide unique perspectives on the usability of products and services that benefit academic programs and industry.

- Instructors and employers should consult with individuals with disabilities about what will make them most productive in computing activities; after all, they are the experts about their own access challenges and solutions.

Conclusion

Evaluation results of *AccessComputing* activities suggest that computing departments, professional organizations, and employment opportunities have become more accessible and welcoming because of the engagement between *AccessComputing* partners and collaborators. Many individual students have been positively impacted by their participation in the project as well. Clearly, much has been accomplished. But critical work remains to be done by *AccessComputing* and complementary projects. NSF has demonstrated its agreement with this conclusion by authorizing five more years of funding for the *AccessComputing* initiative. The renewal of the grant will particularly focus on the following new areas:

- Supporting the integration relevant disability/accessibility/universal design content into computing courses in order to increase knowledge and skills among future computing professionals.

- Engaging the computing industry to increase focus on placement of individuals with disabilities in careers.

Those interested in moving this diversity agenda forward should consult the *AccessComputing* website at <http://www.uw.edu/accesscomputing> or contact us at accesscomputing@uw.edu. Ultimately, outcomes of projects like *AccessComputing* will benefit society by making computing opportunities available to more people and by enhancing computing fields with the talents and perspectives of people with disabilities. **C**

References

1. *AccessComputing*. *AccessComputing*. University of Washington, Seattle; <http://www.uw.edu/accesscomputing>
2. Bellman, S., Burgstahler, S., and Ladner, R. Work-based learning experiences help students with disabilities transition to careers: A case study of University of Washington projects. *WORK: A Journal of Prevention, Assessment & Rehabilitation* 48, (2014), 399-405.
3. Bernston, H. Senate passes legislation to improve care for soldiers. *The Seattle Times*, (2007, Dec. 15, 2007); http://seattletimes.nwsource.com/html/politics/2004074552_warriors15m.html
4. Burgstahler, S. *AccessSTEM: Progress of teens with disabilities toward STEM careers: Positive inputs leading to critical junctures*. University of Washington, Seattle, 2006; <http://www.uw.edu/doit/Stem/flowchart.html>
5. Burgstahler, S. Universal design: Implications for computing education. *ACM Transactions on Computing Education* 11, 3 (Oct. 2011).
6. Congressional Commission on the Advancement of Women and Minorities in Science, Engineering, and Technology Development. *Land of Plenty: Diversity as America's Competitive Edge in Science, Engineering and Technology*. (Sept. 2000) Washington, D.C.
7. World Wide Web Consortium. *Web Content Accessibility Guidelines, 2.0*. 2008; <http://www.w3.org/TR/wcag20/>
8. Wulf, W.A. (2000). How shall we satisfy the long-term educational needs of engineers? In *Proceedings of the IEEE* 88, 4 (2000), 593-596.

Richard E. Ladner (ladner@cs.washington.edu) is a professor in computer science and engineering at the University of Washington. He is principal investigator for *AccessComputing*, an ACM Fellow, and was awarded the 2014 SIGCHI Social Impact Award.

Sheryl Burgstahler (sheryl@uw.edu) is Director of the Disabilities, Opportunities, Internetworking, and Technology (DO-IT) Center and Affiliate Professor in Education at the University of Washington and is the co-principal investigator and Director of *AccessComputing*. The recipient of many awards, DO-IT won a 1997 Presidential Award for Excellence in Science, Mathematics, and Engineering Mentoring.

This column was created with funding from the National Science Foundation (grant numbers CNS-0540615, CNS-0837508, and CNS-1042260). Any questions, findings, and conclusions or recommendations expressed in this column are those of the authors and do not necessarily reflect the views of the federal government.

Copyright held by authors.

Viewpoint

Creating a New Generation of Computational Thinkers

Experiences with a successful school program in Scotland.

PREVIOUS COMMUNICATIONS VIEWPOINTS have discussed the crisis in school computing teaching.^{1,4,5} In many countries, school computing lessons have degenerated into the teaching of office skills, often by unqualified teachers. Students typically found this boring and uninspiring. There is a worldwide movement to tackle this problem, manifest in the U.K. by the Computing at Schools organization (CAS), the Royal Society report *Shutdown or Restart?*⁶ and Google Executive Chairman Eric Schmidt's McTaggart Lecture at the 2011 Edinburgh Television Festival.

It is important to create a community of *toolmakers* rather than just *tool users*. To achieve this we must teach schoolchildren to program, but this is not sufficient. Students also need to think *computationally*: to use abstraction, modularity, hierarchy, and so forth in understanding and solving problems. It is also necessary to employ a pedagogy that is informed by the latest research into the most effective ways to teach computing.

In Scotland, we have recently taken advantage of a unique opportunity to reform the teaching of computing. The new Curriculum for Excellence (CfE) has granted teachers significantly more autonomy in deciding what to teach and how to teach it. CS is now a core entitlement for all students during the first three years of secondary school, and an option in a new series of public qualifications during the senior



years. CS teachers are eager to obtain teaching materials that address the CfE's aims and objectives.

Scotland starts from a stronger position than many other countries. While not immune to a decline in specialist CS teachers and student demand,³ this decline has been more gentle and from a higher starting point than in the rest of the U.K. Scottish teachers are all required to have a teaching qualification in their specialty and some public CS qualifications have included a significant and rising

element of programming. Scotland also benefits from a unitary national curriculum and a unitary assessment authority for public examinations. Lastly, it is also small enough that one can get representatives from all the key stakeholders around a table, get buy-in from them all, and thereby be nimble in solving problems.

Scotland thus provided an ideal laboratory in which to develop a program for teaching computing: combining programming, computational thinking (CT), and evidence-based

Distinguished Speakers Program

<http://dsp.acm.org>

Students and faculty can take advantage of ACM's Distinguished Speakers Program to invite renowned thought leaders in academia, industry and government to deliver compelling and insightful talks on the most important topics in computing and IT today. ACM covers the cost of transportation for the speaker to travel to your event.



Association for
Computing Machinery

pedagogy, in order to address the computing teaching crisis. We encapsulated this program in teaching materials that we have made freely available and which have been enthusiastically adopted, not just in Scotland, but in many other countries. Some Scottish schools previously taught only information systems in the senior years, which does not include programming (beyond basic scripting), rather than computing, which does. Moreover, business education teachers sometimes teach computing in junior years. Therefore, it was necessary for our materials to be easily used by a non-specialist.

In our view, the success of our program shows that, with the aid of professionally curated materials and evidence-based pedagogy, schoolchildren can be taught to think computationally and to become toolmakers.

The RSE/BCS Exemplification Project

In 2010, the Royal Society of Edinburgh (RSE) and the BCS Academy of Computing joined forces in a project to support schoolteachers to develop materials to exemplify the CS CfE aims and objectives. A wide variety of bodies generously and enthusiastically provided funding: Scottish university CS departments, national and international industry and the government agency Education Scotland. One of the co-authors of this article, Jeremy Scott, was appointed: initially to a one-year, part-time post, but this was eventually extended to three years from 2011–2014, at a total cost of £90,000. The project was overseen by an advisory group that drew on all the stakeholders: RSE, BCS, universities, industry, CAS Scotland, CS teachers, and various other agencies. This multi-agency engagement helped get the project taken seriously by government, as did the presence of respected members of the advisory group.

Pedagogy and Approach

In Phase I, we focused on the first three years of secondary school (ages 11–14 in Scotland)—the “Broad General Education” that seeks to give all of Scotland’s students a solid grounding across the curriculum—including the teaching of CS.

Context is also vital. Today’s learners have a different experience of computing from previous generations: it is online, social, and increasingly mobile. Computing devices have become tactile and personal, the result of the convergence of numerous technologies from multitouch to motion sensing and GPS. We therefore had to connect to learners in a way that resonated with their own experiences of the digital world, eschewing the tedium of calculating payroll and other such mundane tasks.

New and rich programming environments allowed us to focus on problem solving without requiring students to learn a demanding syntax—something that often defeats learners before they get started. To that end, we selected Scratch and App Inventor, both maintained by MIT’s Media Lab, as the environments for Phase 1. Their blocks-based approach to coding finesses problems of syntax, enabling students to concentrate on the logic of programming, while supporting the construction of rich multimedia applications (Scratch) and smartphone apps (App Inventor). Students learn all the usual programming concepts: sequential, conditional, and iterative connectives, but this happens implicitly. Both Scratch and App Inventor enabled the construction of quite sophisticated applications with short and easily developed programs, so learners were quickly rewarded for their efforts. Note that none of this would have been possible 5–10 years earlier.

In Phase 2, we built on Phase 1 by developing materials for the new Scottish qualifications of National 4 and 5 that follow the Broad General Education. This required progressing from the previous blocks-based languages to more conventional text-based ones. We chose LiveCode, a modern equivalent of Apple’s HyperCard, which permits cross-platform development and deployment of rich apps to a variety of desktop and mobile platforms.

In addition to traditional programming, the new qualifications cover information systems design and development. While computational thinking certainly applies to information systems, there are further overarching ideas that together we referred to as “informational thinking.” Informational thinking stresses

the notion that information systems rely on abstracting knowledge into data structures. These structures are linked to develop user-centered systems for the storage, processing, and retrieval of information.

Our materials were designed to be used by even non-specialist teachers and to make available to them the latest, evidence-based pedagogical methodology, for example, “did you understand” questions. Each off-the-shelf pack included:

- **Screencasts:** Videos illustrate key program construction operations step by step to the YouTube generation, which makes the materials easier for non-specialist teachers to use.

- **Buggy code:** Students are lured into making errors and then challenged to debug their faulty code.

- **Did you understand?:** Embedded questions reinforce and assess learning and encourage discussion and collaboration between students (which pedagogic research suggests is an important aid to deep understanding²). Some of the questions involve debugging code samples, while others address key CT themes, such as sequencing, scope and hierarchy (see the appendices to this Viewpoint).

- **Background:** Opportunities are taken to deepen students’ understanding of broader underlying principles, such as data representation, and place developments in computing in a historical and social context.

- **Design not imitation:** The emphasis is on the student’s design of the algorithm, not on imitation without understanding.

- **Excitement:** The environments and applications are chosen to motivate the target age group, in contrast to some of the standard applications in the past, mentioned earlier.

- **Computational thinking:** Underlying principles such as abstraction, modularity, and hierarchy, for example, are abstracted from the examples and highlighted via box-outs.

We believe this overall package, combining the elements listed here, allowed the project’s materials to aid the learning of CT informed by evidence-based pedagogy.

The Materials

All of the materials comprise student

New and rich programming environments allowed us to focus on problem solving without requiring students to learn a demanding syntax.

and teacher packs.^a The latter are annotated with pedagogical suggestions and suggestions for further work, including interdisciplinary learning—a cornerstone of Scotland’s new curriculum—while mapping to the aims and objectives of CfE. This enables non-specialist teachers to step out of their comfort zone and provide the kind of computing teaching our students require.

The sequence of materials was envisaged to support CS from approximately 11–15 years over, say, 15 hours per pack:

- **Starting from Scratch:** An introduction to computing science using Scratch.

- **Itching for More:** Intermediate material using BYOB (Build Your Own Blocks) Snap!, a modification of Scratch developed at UC Berkeley.

- **I ♥ My Smartphone:** Consolidating previous work through the medium of mobile app development using App Inventor.

- **Information Everywhere:** An introduction to web-based information systems.

- **Live to Code:** A programming course for Scotland’s new high school CS qualifications using the LiveCode app development language.

Response

The materials were successfully trialled in a number of Scottish schools before general release, receiving widespread praise from both teach-

ers and students. Since release we have had further positive feedback, not just from within Scotland but across the U.K. and internationally. A typical comment was: “The RSE materials set the pace on a global scale—they’re comprehensive, accessible and were clearly written by someone with an understanding of how kids learn. They also employ a lot of research-based pedagogy, which is evident in both the teacher notes and student materials.” [Cameron Fajjo, while Director of Software Engineering Education for the New York City Department of Education (now Project Lead in Computer Science Education at Google)]

INRIA has also translated *Starting From Scratch* into French for use in French schools.

Lessons

The project’s success exceeded our expectations. The materials have been widely and successfully adopted, not just in Scotland, but across the U.K. and elsewhere. We achieved a great deal at relatively low cost, as much of the project worked on good will: busy academics and industry figures selflessly gave up their time, ably coordinated by expert help from the RSE. Jeremy Scott’s school was also flexible in accommodating his secondment and speaking engagements.

But why else did the project succeed and what advice would we give to others trying to tackle the same issues?

- A coordinated approach is key: universities, learned societies, and industry are taken seriously by governments and we successfully leveraged their support to get our voice heard.

- We were also ‘squeaky wheels’, making noise with government and through the press. We seized upon the project’s initial success to extend its original remit and raise broader issues facing the subject.

- It also happened at a time when government wanted to be seen to be doing something about it. A launch event for the first phase of materials maintained the interest of politicians who were happy to be associated with it.

- This allowed us to engage with our public education bodies in order to help shape the curriculum, rather than simply react to it.

^a *Computing Science.* The materials discussed in this Viewpoint are available from the Royal Society of Edinburgh at <http://bit.ly/1LaWIXr>.

Did You Understand? question from *Starting from Scratch*.

Nested repeat question



How many times will sprite move 10 steps?

Why?

Answer: the sprite (character) will move 10 steps 15 times. This example demonstrates to learners concepts of fixed loops and nested commands.

Did You Understand? question from *I ♥ My Smartphone*. In this example, a button can be clicked to increase the line width in a finger-painting application.

FingerPaint app question

A user starts up a FingerPaint app and immediately clicks **ButtonBigBrush** (code shown below).



However, when the user tries to paint, nothing appears on the canvas until they click **ButtonBigBrush** a second time.

Discuss with your partner why this happens and what change(s) should be made to the code to fix this bug.

Answer: Variable brushSize is initially set to zero, then assigned to the LineWidth property before it is incremented. Consequently, when the button is pressed for the first time, the user paints a line of width 0 pixels. The answer is to increment the brushSize variable before assigning it to the LineWidth property.

This example demonstrates to learners concepts of initialization, assignment, the difference between variables and properties to which they can be assigned and, above all, sequencing.

Did You Understand? question from *I ♥ My Smartphone*. In this example, a bat from a classic “Pong”-style game is constantly moving, while its heading left and right can be changed by tilting the phone.

Bat & ball question

Algorithm

Moving bat (using Orientation Sensor)

```
if orientation sensor roll > 0 then (phone is tilted to the right)
    set the bat heading to the right (0)
else (phone is tilted to the left)
    set the bat heading to the left (180)
```

In what direction will the bat move if the phone's tilt is zero (completely level)?

Why?

Answer: The bat will move to the left because the orientation sensor roll measurement is not greater than zero. This example demonstrates the binary logic in an IF...THEN...ELSE statement. The text in brackets could also be interpreted as internal commentary for the IF and ELSE statements, demonstrating further the concept of programmer logic errors.

► Everyone is “Scratching,” but there is a risk that teachers could just let students loose on a programming platform, such as Scratch, and tick the CS strand on their curriculum. For us, these wonderful tools are the *medium* but CT is the *message*. It is about laying firm foundations, which our materials provide.

The Future

We are now at a critical point in Scottish CS school teaching. On the one hand we have our, and other (for example, CAS inspired), well-received materials; a band of able and energetic teachers; enthusiastic students; and buy-in from industry, government, universities, and learned societies. On the other hand, in some schools, we still have declining numbers taking CS qualifications and non-replacement of CS teachers. Some head teachers, career advisers, parents, and local education authorities are still unaware of the rewards of programming, the intellectual content of CT, and the exciting CS job opportunities.

It could go either way. It is a matter of urgency to get the message out to schools, parents, local education authorities and other stakeholders to reverse the current decline. In Scotland, we hope to play a significant role in spreading this message both within our country and worldwide. With the support of our government and the media, we have made a successful start. ☐

References

1. Bell, T. Establishing a nationwide CS curriculum in New Zealand high schools. *Commun. ACM* 57, 2 (Feb. 2014), 28–30.
2. Chi, M.T.H. Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science* 1 (2009), 73–105.
3. *Computing Science Teachers in Scotland 2014*. A report by Computing At School Scotland, December 2014.
4. Cooper, S., Grover, S. and Simon, B. Building a virtual community of practice for K–12 CS teachers. *Commun. ACM* 57, 5 (May 2014), 39–41.
5. Shein, E. Should everybody learn to code? *Commun. ACM* 57, 2 (Feb. 2014), 16–18.
6. *Shut Down or Restart? The Way Forward for Computing in UK Schools*. The Royal Society, January 2012.

Jeremy Scott (jts@george-heriots.com) is Principal Teacher of Computing Science at George Heriot's School, Edinburgh, Scotland.

Alan Bundy (A.Bundy@ed.ac.uk) is Professor of Automated Reasoning at the School of Informatics, University of Edinburgh, Scotland.

We wish to thank the numerous funders of our project, the project's advisory group, and the anonymous *Communications* referees, whose constructive comments helped shape our Viewpoint.

Copyright held by authors.

Viewpoint

I Can't Let You Do That, Dave

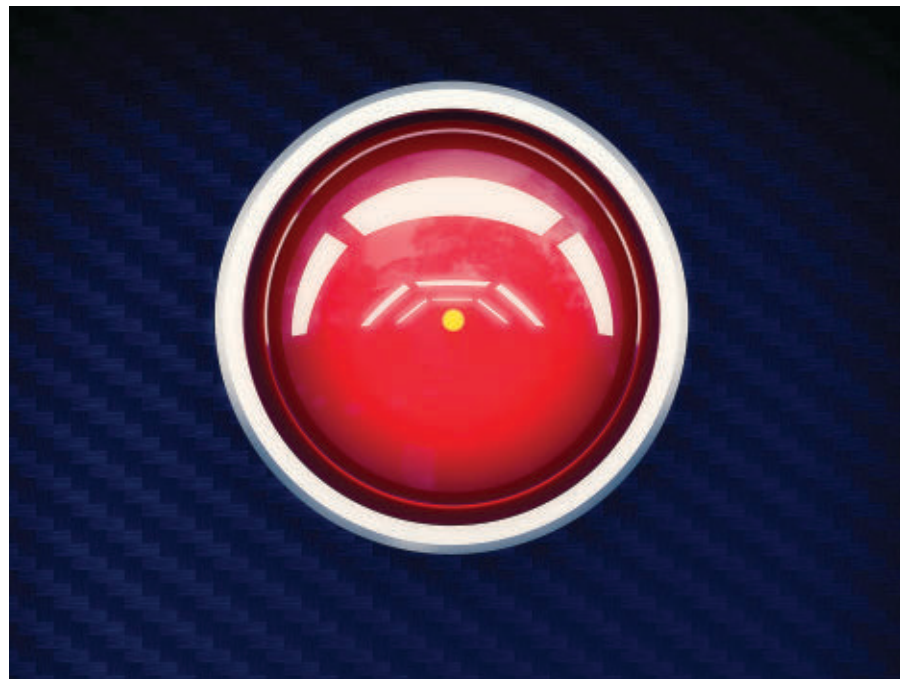
Computers should not treat their owners as adversaries.

IT HAS BEEN 25 years since the Electronic Frontier Foundation was founded to ensure the civil liberties that mattered in the real world followed us into the online world, and it has been a heady quarter-century, with many significant victories, and we have learned some alarming, important lessons on the way.

First among these lessons: there is no distinction between the “virtual” and “real” worlds. The Internet is the nervous system of the 21st century. This was obvious even before the Internet of Things (IoT) turned our cars and homes into computers that engulf our bodies, but the IoT has made the issue especially urgent.

The problems of regulation and the IoT is an old one, but with a new urgency. Since 1998, EFF has been part of the movement to reform the Digital Millennium Copyright Act (DMCA). The DMCA has many flaws, the worst of which is in Section 1201, the “anti-circumvention” rule, whose ills have been magnified by recent technological developments, turning them into something of an existential threat to a free and fair future.

DMCA 1201 prohibits breaking “digital locks” that restrict access to copyrighted works. Though it was originally conceived as a means of preventing piracy, it has proved most useful at preventing competition and the creation of legitimate, otherwise legal technologies. Copyright law has many flexibilities and exclusions that protect designers, developers, and users



can freely exercise, without any permission from the copyright holder. But under 1201, you can only make these uses if you do not have to break a lock.

For example, it is legal for you to rip your CDs. Put a CD in your computer's optical drive and the manufacturer-supplied OS will launch a tool that invites you to rip and library the music on the disc, automating the process of taking your music with you on a mobile device. Ripping DVDs is legal under the same theory, with one important difference: DVDs were born with DMCA-covered digital locks. Insert a DVD into your computer and the only feature you get is the same one your DVD

player had in 1996: playing the movie. If you want to watch your movie on your phone legally, you cannot do so, because despite the legality of transcoding and moving files for personal use, the legal inviolability of the digital lock you must break to do that computation means you must buy all your movies all over again to watch them on the go.

Watching a movie you paid for on a device you own is not piracy by any definition, and it is bad enough that the DMCA prevents this lawful feature. It has been 19 years since the DVD was introduced and *not one single feature* has been introduced to the platform in all that time.

But the main event is not user rights or innovation: it is security and free speech. As ACM members doubtlessly appreciate, preventing the owner of a computer from executing the code of their choice is an impossible task. No matter how cleverly the operating system and its services monitor the user and hide the keys necessary to unlock files without permission, users will eventually find a flaw in the defenders' code and use it to jailbreak the system, allowing arbitrary code execution. Even if you stipulate that locking computer users out of their own computers is a legitimate objective, it is still a technological nonsense. A security model that treats the computer's user as an attacker is doomed. We cannot hide keys in devices we give to attackers for the same reason we cannot keep safes—no matter how well designed—in bank-robbers' living rooms.

The DMCA tries to address this by threatening people who publish code or information that would help remove a lock with severe penalties: five years in prison and \$500,000 in fines for a first offense.

But information about flaws in a computer is not just useful to people who want to add functionality to their computers: it also provides opportunities for malware to seize control over the system. By criminalizing disclosure of flaws, the DMCA ensures systems covered by its measures become reservoirs of long-lived digital pathogens. This is bad enough in the context of mobile devices—your phone is not just a distraction rectangle that lets you throw birds at pigs; it is a sensor-studded supercomputer that is privy to your every movement, conversation, and authentication credential—but it gets much worse in the age of the IoT.

Internet of Things startups are under intense investor pressure to restrict their devices with DMCA-covered locks to create managed “ecosystems.” GM and John Deere both filed comments with the U.S. Copyright Office this spring asserting the software locks on their products are covered by the DMCA. They want to ensure independent mechanics cannot jailbreak those vehicles and provide service without first entering into a license-

A security model that treats the computer's user as an attacker is doomed.

agreement through which they promise only to buy original parts.

We are familiar with this model: it is the printer-ink business-model, where digital locks are used to ensure people who buy a product are locked into buying consumables for it from one company, at the highest possible price. It is just one of the ways the DMCA rewards IoT businesses that treat their customers as the enemy.

Another is the ability to make promises to other people about what your customers will and will not do with your product. The World Wide Web Consortium yielded to pressure from Netflix and the BBC to add digital locks to the standards for HTML5 so that these companies could promise copyright holders viewers would not be able to save streamed video. Smart-thermostat offerings like Nest want to be able to promise power authorities they can lower their customers' thermostats without customers turning them back up again.

Finally, the DMCA lets vendors extract rent from, and exact control over, independent software vendors. Putting locks on the devices you sell means you can set up app stores and no one else can set up competing stores. This lets you charge high commissions on sales and refuse to carry apps that add functionality your users want, but that you would rather not see.

DMCA 1201 is turning all of IoT into a playground for malware, where reporting vulnerabilities and releasing third-party improvements to systems are chilled by a law that was stupid in 1998 and is deadly in 2015.

Malware is always frightening, but it is much worse on systems already designed to treat their owners as adversaries. Infections on devices that

take pains to hide their processes and files from their owners are much more difficult to detect and root out. Those devices are *supposed* to run programs that user-space apps cannot see or terminate, so malware that avails itself of this privilege becomes nearly bulletproof.

The Electronic Frontier Foundation's new Apollo 1201 project aims to reform DMCA 1201, and all of the laws like it around the world, within a decade. We want to litigate the constitutionality of 1201, representing scholars, researchers and academics, these being the kind of unimpeachable clients judges are loathe to find against.

We know from our own off-the-record conversations with academics and researchers that they quietly violate 1201 in their work *all the time*, and that there are plenty of legitimate projects that never launch for fear of violating the law. If you do this sort of work, the Electronic Frontier Foundation would like to discuss it with you. If you know someone who does this kind of work, encourage that person to get in touch with the Electronic Frontier Foundation.

The model of fixing social problems by locking users out of their own devices is an invitation to even worse security policies. When FBI Director James Comey and U.K. Prime Minister David Cameron call for backdoors in our crypto, they are necessarily implying a means of ensuring you cannot install code of your choosing on your devices, lest you choose to install working crypto.

The Internet of Things is here already. The most salient fact about a person's pacemaker is its network stack and security model; the most salient fact about a person's car is its informatics. As computers move inside our bodies and as we move our bodies into computers, it is clearer than ever that “offline” and “online” are not meaningful distinctions. If the information age is to be habitable, then crucial free speech provisions that let experts blow the whistle on unsafe practices in our infrastructure cannot be denied. 

Cory Doctorow (cory@eff.org) is the co-editor of the blog Boing Boing and the former European Affairs Coordinator for the Electronic Frontier Foundation.

Copyright held by author.

Point/Counterpoint

The Case for Banning Killer Robots

Ban the bots? Considering both sides of the argument for and against.

DOI:10.1145/2835963

Point: Stephen Goose

DIPLOMATS AND MILITARY experts from more than 90 countries gathered in Geneva in April 2015 for their second meeting on “lethal autonomous weapons systems,” also known as fully autonomous weapons, or more colloquially, killer robots. Noted artificial intelligence expert Stuart Russell informed delegates the AI community is beginning to recognize the specter of autonomous weapons is damaging to its reputation and indicated several professional associations were moving toward votes to take a position on the topic.

On July 28, 2015, more than 1,000 AI professionals, roboticists, and others released an open letter promoting a “ban on offensive autonomous weapons beyond meaningful human control.”^a

Thus, it is timely and appropriate that *Communications* is focusing on fully autonomous weapons, particularly the call for a preemptive prohibition on their development, production, and use as has been made by the Campaign to Stop Killer Robots, many in the AI community, more than 20 Nobel Peace Laureates, and many others.

There is no doubt that AI and greater autonomy can have military and hu-

Weaponizing fully autonomous robotics systems is seen by many as a step too far.

manitarian benefits. Most of those calling for a ban support AI and robotics research. But, weaponizing fully autonomous robotic systems is seen by many as a step too far.

The Many Reasons to Ban Fully Autonomous Weapons

The central concern is with weapons that once activated, would be able to select and engage targets without further human involvement. There would no longer be a human operator deciding whom to fire at or when to shoot. Instead, the weapon system itself would undertake those tasks. This would constitute not just a new type of armament, but a new method of warfare that would radically change how wars are fought, and not to the betterment of humankind.

Many in the AI community have focused on the serious international security and proliferation concerns related to fully autonomous weapons.

There is the real danger that if even one nation acquires these weapons, others may feel they have to follow suit in order to defend themselves and to avoid falling behind in a robotic arms race.

The open letter signed by some of the most renowned AI experts stated, “If any major military power pushes ahead with AI weapon development, a global arms race is virtually inevitable ... We therefore believe that a military AI arms race would not be beneficial for humanity. There are many ways in which AI can make battlefields safer for humans, especially civilians, without creating new tools for killing people ... [M]ost AI researchers have no interest in building AI weapons—and do not want others to tarnish their field by doing so, potentially creating a major public backlash against AI that curtails its future societal benefits.”^b

There is also the prospect that fully autonomous weapons could be acquired by repressive regimes or non-state armed groups with little regard for the law. These weapons could be perfect tools of repression and terror for autocrats.

Another type of proliferation concern is that such weapons would increase the likelihood of armed attacks, making resort to war more likely, as decision makers would not have the same concerns about loss of soldiers’ lives. This could have an overall destabilizing effect on international security.

^a The Future of Life Institute organized the letter. The text and list of signatories can be found at: <http://bit.ly/1UQ17QE>

^b <http://bit.ly/1UQ17QE>

For many people, these weapons would cross a fundamental moral and ethical line by ceding life and death decisions on the battlefield to machines. Giving such responsibilities to machines in such circumstances has been called the ultimate attack on human dignity.^c The notion of allowing compassionless robots to make decisions about the application of violent force is repugnant to many. Compassion is a key check on the killing of other human beings.

In my extensive engagement with a variety of audiences on this issue, it has been striking how most people have a visceral negative reaction to the notion of fully autonomous weapons. The Martens Clause, which is articulated in the Geneva Conventions and elsewhere, is a key provision in international law that takes into account this notion of general repugnance on the part of the public. Under the Martens Clause, fully autonomous weapons should comply with the “principles of humanity” and the “dictates of public conscience.” They likely would not comply with either.

There are serious questions about whether fully autonomous weapons would ever be capable of complying with core principles of international humanitarian law (IHL) during combat, or international human rights law (IHRL) during law enforcement operations, border control, or other circumstances. There is of course no way of predicting what technology might produce many years from now, but there are strong reasons to be skeptical about compliance with international law in the future, including the basic principles of distinction and proportionality.

Could robots replicate the innately human qualities of judgment and intuition necessary to comply with IHL, including judgment of an individual’s intentions, as well as subjective determinations? Compliance with the rule of proportionality prohibiting attacks in which expected civilian harm outweighs anticipated military gain would

Many advanced militaries are rapidly marching toward ever greater autonomy in their weapons systems.

be especially difficult as this relies heavily on situational and contextual factors, which could change considerably with a slight alteration of the facts.

There are also serious concerns about the lack of accountability when fully autonomous weapons fail to comply with IHL in any particular engagement. Holding a human responsible for the actions of a robot that is acting autonomously could prove difficult, be it the operator, superior officer, programmer, or manufacturer.

Scientists and military leaders have also raised a host of technical and operational issues with these weapons that could pose grave dangers to civilians—and to soldiers—in the future.^d A particular concern for many is how “robot vs. robot” warfare would unfold, and how devices controlled by complex algorithms would interact.^e

Taken together, this multitude of concerns has led to the call for a preemptive prohibition on fully autonomous weapon systems—a new international treaty that would ban the development, production, and use of fully autonomous weapons, and require there is always meaningful human control over targeting and kill decisions.

^d See, for example, “Autonomy in Weapons Systems,” U.S. Department of Defense, Directive Number 3000.09, Nov. 21, 2012. This directive, which guides U.S. policy on autonomous weapons, cites a multitude of technical issues that would have to be overcome before fielding fully autonomous weapons.

^e An October 2013 joint statement calling for a ban on fully autonomous weapons signed by more than 270 computing experts said, “Such interactions could create unstable and unpredictable behavior, behavior that could initiate or escalate conflicts, or cause unjustifiable harm to civilian populations.” See <http://icrac.net/call/>

Rapid and Growing Support for a Ban

Over the past two years, the question of what to do about fully autonomous weapons has rocketed to the top ranks of concern in the field of disarmament and arms control, or what is now often called humanitarian disarmament. Within this short period of time, world leaders including the Secretary-General of the United Nations, the U.N. disarmament chief, and the head of the International Committee of the Red Cross, have expressed deep concerns about the development of fully autonomous weapons and urged immediate action.

The Campaign to Stop Killer Robots—an international coalition of non-governmental organizations (NGOs)—was launched in April 2013 calling for a preemptive ban on fully autonomous weapon systems.^f Coordinated by Human Rights Watch, it is modeled on the successful civil society campaigns that led to international bans on antipersonnel landmines, cluster munitions, and blinding lasers. A month after the campaign launched, the U.N. special rapporteur on extrajudicial killings presented the Human Rights Council with a report that echoed many of the campaign’s concerns and called on governments to adopt national moratoria on the weapons.

In October 2013, more than 270 prominent engineers, computing and artificial intelligence experts, roboticists, and professionals from related disciplines issued a statement calling for a ban. This was organized by the International Committee for Robot Arms Control (ICRAC), which was founded in 2009 by roboticists, ethicists, and others.^g As noted earlier, in July 2015 more than 1,000 AI experts signed an open letter supporting a prohibition; the letter was organized by the Future of Life Institute.

During 2014, the European Parliament passed a resolution that calls for a ban, more than 20 Nobel Peace Laureates issued a joint statement in favor of a ban, and more than 70 prominent faith leaders from around the world released a statement calling for a ban. A Canadian robotics

^f <http://www.stopkillerrobots.org/>

^g <http://icrac.net/call/>

^c This point has been made repeatedly by the U.N. Special Rapporteur Christof Heyns. See, for example, U.N. Human Rights Council, Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns, Lethal Autonomous Robotics, A/HRC/23/47, Apr. 9, 2013.

company, Clearpath, became the first commercial entity to support a ban and declare it will not work toward the development of fully autonomous weapons systems.

Governments have become seized with the issue of fully autonomous weapons since 2013, though few have articulated formal policy positions. Most importantly, the 120 States Parties to the Convention on Conventional Weapons (CCW) agreed in November 2013 to take up the issue, holding meetings in May 2014 and then again in April 2015. In the diplomatic world, the decision to take on killer robots was made at lightning speed.

It appears certain that nations will agree to continue their CCW deliberations next year, though questions remain about the nature, content, and duration of the work. A small but growing number of states have already called for a preemptive ban, while most participating states have expressed interest in discussing the concept of meaningful human control of weapons systems, indicating they see a need to draw a line before weapons become fully autonomous.

The April 2015 CCW experts meeting was by far the richest, most in-depth discussion held to date on autonomous weapons. Not a single state said it is actively pursuing them yet the week featured extensive discussion about the potential benefits of such weapons.

The U.S. and Israel were the only states to explicitly say they were keeping the door open to the acquisition of fully autonomous weapons but there are plenty of indicators there are many states that are contemplating them. Without question, many advanced militaries are rapidly marching toward ever-greater autonomy in their weapons systems, and there are no stop signs in their path.

Why a Ban Is the Best Solution

Some oppose a preemptive and comprehensive prohibition, saying it is too early and we should “wait and see” where the technology takes us. Others believe restrictions would be more appropriate than a ban, limiting their use to specific situations and missions. Some say existing international humanitarian law will be sufficient to address the challenges posed.

The point of a preemptive treaty is to prevent future harm and with all the dangers and concerns associated with fully autonomous weapons, it would be irresponsible to take a “wait and see” approach and only try to deal with the issue after the harm has already occurred. Once developed, they will be irreversible; it will not be possible to put the genie back in the bottle as the weapons spread rapidly around the world.

The notion of a preemptive treaty has been done before. The best example is the 1995 CCW protocol that bans blinding laser weapons. After initial opposition from the U.S. and others, states came to agree the weapons would pose unacceptable dangers to soldiers and civilians. The weapons were seen as counter to the dictates of public conscience and nations came to recognize their militaries would be better off if no one had the weapons than if everyone had them. These same rationales apply to fully autonomous weapons.

While some rightly point out that there is no “proof” there cannot be a technological fix to the problems of fully autonomous weapons, it is equally true there is no proof there can be. Given the scientific uncertainty that exists, and given the potential benefits of a new legally binding instrument, the precautionary principle in international law is directly applicable. The principle suggests the international community need not wait for scientific certainty, but could and should take action now.

Fully autonomous weapons represent a new category of weapons that could change the way wars are fought

It will not be possible to put the genie back in the bottle as the weapons spread rapidly around the world.

and pose serious risks to civilians. As such, they demand new, specific law that clarifies and strengthens existing international humanitarian law.

A specific treaty banning a weapon is also the best way to stigmatize the weapon. Experience has shown that stigmatization has a powerful effect even on those who have not yet formally joined the treaty, inducing them to comply with the key provisions, lest they risk international condemnation. A regulatory approach restricting use to certain locations or to specific purposes would be prone to longer-term failure as countries would likely be tempted to use them in other, possibly inappropriate, ways during the heat of battle or in dire circumstances. Once legitimized, the weapons would no doubt be mass produced and proliferate worldwide; only a preemptive international treaty will prevent that.

The call for a ban on development of fully autonomous weapons is not aimed at impeding broader research into military robotics or weapons autonomy or full autonomy in the civilian sphere. It is not intended to curtail basic AI research in any way. Research and development activities should be banned if they are directed at technology that can only be used for fully autonomous weapons or that is explicitly intended for use in such weapons.

Conclusion

While there are at this stage still many doubters, my experience leads me to conclude a preemptive ban is not only warranted, but is achievable and is the only possible approach that would successfully address the potential dangers of fully autonomous weapons. However, the involvement, advice, and expertise of the AI community are needed both to get to a ban and to ensure it is the most effective ban possible.

The AI community has an important role to play in bringing about the ban on fully autonomous weapons. This is not a political issue to be avoided in the name of pure science, but rather an issue of humanity for which we are all responsible. 

Stephen Goose (gooses@hrw.org) is the director of the Human Rights Watch Arms Division.

Copyright held by author.

DOI:10.1145/2835965

Counterpoint: Ronald Arkin

LET ME UNEQUIVOCALLY state: *The status quo with respect to innocent civilian casualties is utterly and wholly unacceptable.* I am not Pro Lethal Autonomous Weapon Systems (LAWS), nor for lethal weapons of any sort. I would hope that LAWS would never need to be used, as I am against killing in all its manifold forms. But if humanity persists in entering into warfare, which is an unfortunate underlying assumption, we must protect the innocent noncombatants in the battlespace far better than we currently do. Technology can, must, and should be used toward that end. Is it not our responsibility as scientists to look for effective ways to reduce man's inhumanity to man through technology? Research in ethical military robotics could and should be applied toward achieving this goal.

I have studied ethology (animal behavior in their natural environment) as a basis for robotics for my entire career, spanning frogs, insects, dogs, birds, wolves, and human companions. Nowhere has it been more depressing than to study human behavior in the battlefield (for example, the Surgeon General's Office 2006 report¹⁰ and *Killing Civilians: Method, Madness, and Morality in War*.⁹). The commonplace occurrence of slaughtering civilians in conflict over millennia gives rise to my pessimism in reforming human behavior yet provides optimism for robots being able to exceed human moral performance in similar circumstances. The regular commission of atrocities is well documented both historically and in the present day, reported almost on a daily basis. Due to this unfortunate low bar, my claim that robots may be able to eventually outperform humans with respect to adherence to international humanitarian law (IHL) in warfare (that is, be more humane) is credible. I have the utmost respect for our young men and women in the battlespace, but they are placed into situations where no human has ever been designed to function. This is exacer-

bated by the tempo at which modern warfare is conducted. Expecting widespread compliance with IHL given this pace and resultant stress seems unreasonable and perhaps unattainable by flesh and blood warfighters.

I believe judicious design and use of LAWS can lead to the potential saving of noncombatant life. If properly developed and deployed it can and should be used toward achieving that end. It should not be simply about winning wars. We must locate this humanitarian technology at the point where war crimes, carelessness, and fatal human error lead to noncombatant deaths. It is not my belief that an unmanned system will ever be able to be perfectly ethical in the battlefield, but I am convinced they can ultimately perform more ethically than human soldiers.

I have stated that I am not averse to a ban should we be unable to achieve the goal of reducing noncombatant casualties, but for now we are better served by a moratorium at least until we can agree upon definitions regarding what we are regulating, and it is indeed determined whether we can realize humanitarian benefits through the use of this technology. A preemptive ban ignores the moral imperative to use technology to reduce the persistent atrocities and mistakes that human warfighters make. It is at the very least premature. History indicates that technology can be used toward these goals.⁴ Regulate LAWS usage instead of prohibiting them entirely.⁶ Consider restrictions in well-defined circum-

Is it not our responsibility as scientists to look for effective ways to reduce man's inhumanity to man through technology?

stances rather than an outright ban and stigmatization of the weapon systems. Do not make decisions based on unfounded fears—remove pathos and hype and focus on the real technical, legal, ethical, and moral implications.

In the future autonomous robots may be able to outperform humans from an ethical perspective under battlefield conditions for numerous reasons:

- ▶ Their ability to act conservatively, as they do not need to protect themselves in cases of low certainty of target identification.

- ▶ The eventual development and use of a broad range of robotic sensors better equipped for battlefield observations than humans currently possess.

- ▶ They can be designed without emotions that cloud their judgment or result in anger and frustration with ongoing battlefield events.

- ▶ Avoidance of the human psychological problem of "scenario fulfillment" is possible, a factor contributing to the downing of an Iranian Airliner by the USS *Vincennes* in 1988.⁷

- ▶ They can integrate more information from more sources far faster than a human possibly could in real time before responding with lethal force.

- ▶ When working in a team of combined human soldiers and autonomous systems, they have the potential of independently and objectively monitoring ethical behavior in the battlefield by all parties and reporting infractions that might be observed.

LAWS should not be considered an end-all military solution—far from it. Limited circumstances for their use must be utilized. Current thinking recommends:

- ▶ Specialized missions only where bounded morality,^{a,1} applies, for example, room clearing, countersniper operations, or perimeter protection in the DMZ.^b

- ▶ High-intensity interstate warfare,

a Bounded morality refers to adhering to moral standards within the situations that a system has been designed for, in this case specific battlefield missions and not in a more general sense.

b For more specifics on these missions see Arkin, R.C., *Governing Lethal Behavior in Autonomous Systems*, Chapman-Hall, 2009.

not counterinsurgencies, to minimize likelihood of civilian encounter.

► Alongside soldiers, not as a replacement. A human presence in the battlefield should be maintained.

Smart autonomous weapon systems may enhance the survival of noncombatants. Consider Human Rights Watch's position on the use of precision-guided munitions in urban settings—a *moral imperative*. LAWS in effect may be mobile precision-guided munitions resulting in a similar moral imperative for their use. Consider not just the possibility of LAWS making a decision when to fire, but rather deciding when *not* to fire (for example, smarter context-sensitive cruise missiles). Design them with runtime human overrides to ensure meaningful human control,¹¹ something everyone wants. Additionally, LAWS can use fundamentally different tactics, assuming far more risk on behalf of noncombatants than human warfighters are capable of, to assess hostility and hostile intent, while assuming a “First do no harm” rather than “Shoot first and ask questions later” stance.

To build such systems is not a short-term goal but will require a mid- to long-term research agenda addressing the many very challenging research questions. By exploiting bounded morality within a narrow mission context, however, I would contend that the goal of achieving better performance with respect to preserving noncombatant life is achievable and warrants a robust research agenda on humanitarian grounds. Other researchers have begun related work on at least four continents. Nonetheless, there remain many daunting research questions regarding lethality and autonomy yet to be resolved. Discussions regarding regulation of LAWS must be based on reason and not fear. Some contend that existing IHL may be adequate to afford adequate protection to noncombatants from the potential misuse of LAWS.² A moratorium is more appropriate at this time than a ban, until these questions are resolved and only then can careful, graded introduction of the technology into the battlespace be ensured. Proactive management of these issues is necessary. Other technological approaches are of course welcome, perhaps such as the creation

I say to my fellow researchers, if your research is of any value, someone somewhere someday will put it to work in a military system.

of ethical advisory systems for human warfighters to assist in their decision-making when in conflict.

Restating my main point: *The status quo is unacceptable with respect to non-combatant deaths*. It may be possible to save noncombatant lives through the use of this technology—if done correctly—and these efforts should not be prematurely terminated by a preemptive ban.

Quoting from a recent *NewsWeek* article³: “But autonomous weapon systems would not necessarily be like those crude weapons [poison gas, landmines, cluster bombs]; they could be far more discriminating and precise in their target selection and engagement than even human soldiers. A preemptive ban risks being a tragic moral failure rather than an ethical triumph.”

Similarly from the *Wall Street Journal*⁸: “Ultimately, a ban on lethal autonomous systems, in addition to being premature, may be feckless. Better to test the limits of this technology first to see what it can and cannot deliver. Who knows? Battlefield robots might yet be a great advance for international humanitarian law.”

I say to my fellow researchers, if your research is of any value, someone somewhere someday will put it to work in a military system. You cannot be absolved from your responsibility in the creation of this new class of technology simply by refusing a particular funding source. Bill Joy argued for the relinquishment of robotics research in his *Wired* article “Why the Future Doesn't Need Us.”⁵ Perhaps it is time for some to walk away from AI if their conscience so dictates.

But I believe AI can be used to save innocent life, where humans may and

do fail. Nowhere is this more evident than on the battlefield. Until that goal can be achieved, I support a moratorium on the development and deployment of this technology. If our research community, however, firmly believes the goal of achieving better performance than a human warfighter with respect to adherence to IHL is unattainable, and states collectively that we cannot ever reach this level of exceeding human morality in narrow battlefield situations where bounded morality applies and where humans are often at their worst, then I would be moved to believe our community asserts artificial intelligence in general is unattainable. This appears to contradict those who espouse their goal of doing just that.

We must reduce civilian casualties if we are foolish enough to continue to engage in war. I believe AI researchers have a responsibility to achieve such reductions in death and damage during the conduct of warfare. We cannot simply accept the current status quo with respect to noncombatant deaths. Do not turn your back on those innocents trapped in war. It is a truly hard problem and challenge but the potential saving of human life demands such an effort by our community. 

References

1. Allen, C., Wallach, W., and Smit, I. Why machine ethics? *IEEE Intelligent Systems* (Jul./Aug. 2006), 12–17.
2. Anderson, K. and Waxman, K. Law and ethics for autonomous weapon systems: Why a ban won't work and how the laws of war can. Stanford University, The Hoover Institution (Jean Perkins Task Force on National Security and Law Essay Series), 2013.
3. Bailey, R. Bring on the killer robots. *Newsweek*; (Feb. 1, 2015); <http://bit.ly/1K3VaYK>
4. Horowitz, M. and Scharre, P. Do killer robots save lives? *Politico Magazine* (Nov. 19, 2014).
5. Joy, B. Why the future doesn't need us. *Wired* 8, 4 (Apr. 2000).
6. Muller, V. and Simpson, T. Killer robots: Regulate, don't ban. Blavatnik School of Government Policy Memo, Oxford University, Nov. 2014.
7. Sagan, S. Rules of engagement. In *Avoiding War: Problems of Crisis Management*. A. George, Ed., Westview Press, 1991.
8. Schechter, E. In defense of killer robots. *Wall Street Journal* (July 10, 2014).
9. Slim, H., *Killing Civilians: Method, Madness, and Morality in War*. Columbia University Press, New York, 2008.
10. Surgeon General's Office, Mental Health Advisory Team (MHAT) IV Operation Iraqi Freedom 05–07, Final Report, Nov. 17, 2006.
11. U.N. *The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward*. UNIDIR Resources, Report No. 2, 2014.

Ronald Arkin (arkin@cc.gatech.edu) is a Regents' Professor and is the director of the Mobile Robot Laboratory in the College of Computing at the Georgia Institute of Technology.

Copyright held by author.



Article development led by [acmqueue](http://queue.acm.org)
queue.acm.org

Balancing statistical accuracy and subject privacy in large social-science datasets.

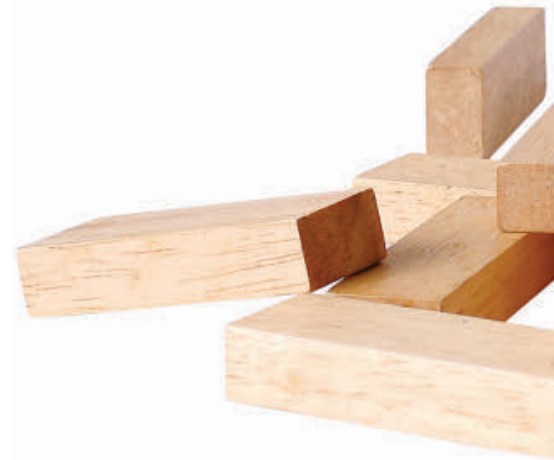
BY OLIVIA ANGIULI, JOE BLITZSTEIN, AND JIM WALDO

How to De-Identify Your Data

BIG DATA IS all the rage; using large datasets promises to give us new insights into questions that have been difficult or impossible to answer in the past. This is especially true in fields such as medicine and the social sciences, where large amounts of data can be gathered and mined to find insightful relationships among variables. Data in such fields involves humans, however, and thus raises issues of privacy that are not faced by fields such as physics or astronomy.

Such privacy issues become more pronounced when researchers try to share their data with others. Data sharing is a core feature of big-data science, allowing others to verify research that has been done and to pursue other lines of inquiry the original researchers may not have attempted. But sharing data about

human subjects triggers a number of regulatory regimes designed to protect the privacy of those subjects. Sharing medical data, for example, requires adherence to HIPAA (Health Insurance Portability and Accountability Act); sharing educational data triggers the requirements of FERPA (Family Educational Rights to Privacy Act). These laws require that, to share





data generally, the data be de-identified or anonymized (note that, for the purposes of this article, these terms are interchangeable). While FERPA and HIPAA define the notion of de-identification slightly differently, the core idea is if a dataset has certain values removed, the individuals whose data is in the set cannot be identified, and their privacy will be preserved.

Previous research has looked at how well these requirements protect the identities of those whose data is in a dataset.² Violations of privacy, like re-identification, generally work by linking data from a de-identified dataset with outside data sources. It is often surprising how little information is needed to re-identify a subject.

More recent research has shown

a different, and perhaps more troubling, aspect of de-identification. These studies have shown the conclusions one can draw from a de-identified dataset are significantly different from those that would be drawn when the original dataset is used.¹ Indeed, it appears the process of de-identification makes it difficult or impossible to use a de-identified (and therefore eas-

ily sharable) version of a dataset either to verify conclusions drawn from the original dataset or to do new science that will be meaningful. This would seem to put big-data social science in the uncomfortable position of having either to reject notions of privacy or to accept that data cannot be easily shared, neither of which are tenable positions.

This article looks at a particular dataset, generated by the massive open online courses (MOOCs) offered through the edX platform by Harvard University and the Massachusetts In-

stitute of Technology during the first year of those offerings. It examines which aspects of the de-identification process for that dataset caused it to change significantly, and it presents a different approach to de-identification that shows promise to allow both sharing and privacy.

Defining Anonymization

The first step in de-identifying a dataset is determining the anonymization requirements for that set. The notion of privacy that was used throughout the de-identification of this particular dataset was guided by FERPA, which requires personally identifiable information be removed, such as name, address, Social Security number, and mother's maiden name. FERPA also requires other information, alone or in combination, must not enable identification of any student with "reasonable certainty."

To meet these privacy specifications, the HarvardX and MITx research team (guided by the general counsel, for the two institutions) opted for a k -anonymization framework, which requires every individual in the

dataset to have the same combination of identity-revealing traits as at least $k-1$ other individuals in the dataset. Identity-revealing traits, termed *quasi-identifiers*, are those that allow linking to other datasets; information that is meaningful within only a single dataset is not of concern.

Anonymizing a dataset with regard to quasi-identifiers is important in order to prevent the re-identification of individuals that would be made possible if these traits were linked with external data that shares the same traits. The example in Figure 1 illustrates how two datasets could be combined in such a way that allows re-identification.²

In the edX dataset, the quasi-identifiers were course ID, level of education, year of birth, gender, country, and number of forum posts. The number of forum posts is considered a quasi-identifier because the forum was a publicly accessible website that could be scraped in order to link user IDs with their number of forum posts. Course ID is considered a quasi-identifier because unique combinations of courses could conceivably enable linking personally identifiable information that a student posts in a forum with the edX dataset.

The required value of k within k -anonymization was set to 5 in this context, based on the U.S. Department of Education's Privacy Technical Assistance Center's claim that "statisticians consider a cell size of 3 to be the absolute minimum" and that values of 5 to 10 are even safer. A higher value of k corresponds to a stricter privacy standard, because more individuals are required to have a given combination of identity-revealing traits.³

Note this is not a claim that de-identifying the dataset to a privacy standard of $k = 5$ assures no one in the dataset can be re-identified. Rather, this privacy standard was chosen to allow legal sharing of the data.

What Methods Allow Anonymization?

There are two techniques to achieve a k -anonymous dataset: generalization and suppression. Generalization occurs when granular values are combined to create a broader category that will contain more records. This can be

Figure 1. Combination of two datasets that allow re-identification.

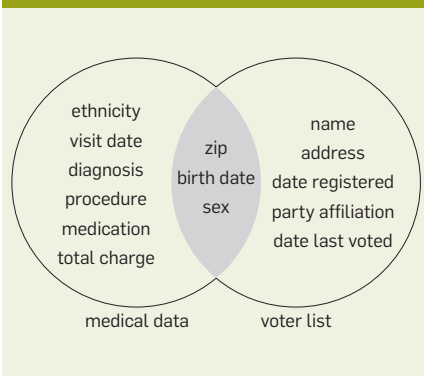
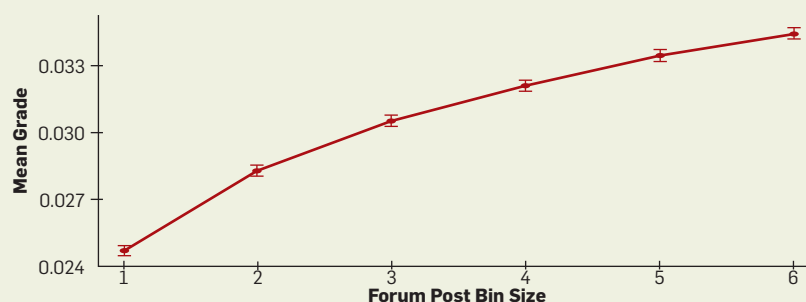


Figure 2. Distortion of mean grade increasing with k .



Figure 3. Distortion of mean grade decreasing with bin size.



achieved both for numerical variables (for example, combining ages 20, 21, and 22 into a broader category of 20–22) and for categorical variables (for example, generalizing location data from “Boston” to “Massachusetts”). Suppression occurs when a record that violates anonymity standards is deleted from the dataset entirely.

Generalization and suppression techniques introduce differing kinds and degrees of distortion during the anonymization process. Relying on suppression can mean a large number of records in the dataset will be removed. Suppression-only de-identification also skews the integrity of a dataset when values are eliminated disproportionately to the original distribution of the data, causing distortion in resulting analyses.

On the other hand, generalized values are often less powerful than granular values—it may be difficult, for example, to fit a linear regression line on generalized numeric attributes. Further, while generalization-only de-identification leaves non-quasi-identifier fields intact, quasi-identifiers may become generalized to a point where few conclusions can be drawn about their relationship with other fields. Finally, since generalization is applied to whole columns, it decreases the quality of the entire dataset, whereas suppression decreases the quality of the dataset on a record-by-record basis.

The anonymization process used to de-identify edX data for public release in 2014 employed a “suppression-emphasis” approach toward k -anonymization. In this approach, the names of the countries were first generalized to region or continent names, then date-time stamps were transformed into date stamps, and finally any existing records that were not k -anonymous after these generalizations were suppressed. In the process, records that claimed a birth date before 1931 (which seemed unlikely to be correct) were automatically suppressed.

Daries et al.’s 2014 study of edX data confirmed a suppression-emphasis approach tended to distort mean values of de-identified columns, whereas a generalization-emphasis approach tended to distort correlations between de-identified columns.¹



Anonymizing a dataset with regard to quasi-identifiers is important in order to prevent the re-identification of individuals that would be made possible if these traits were linked with external data that shares the same traits.



Understanding the Mechanisms of Distortion

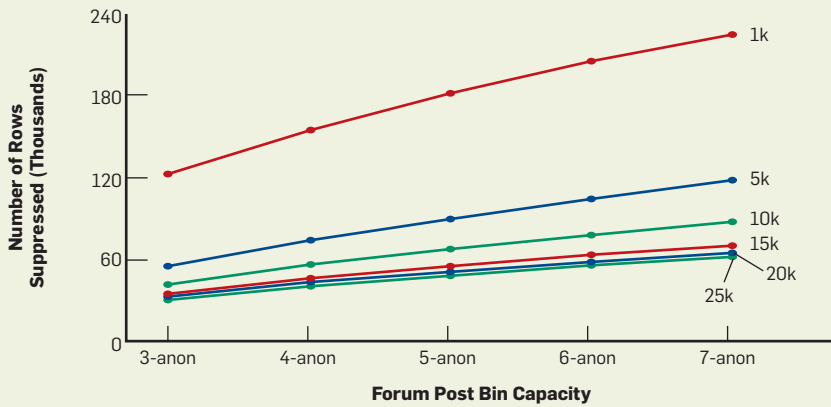
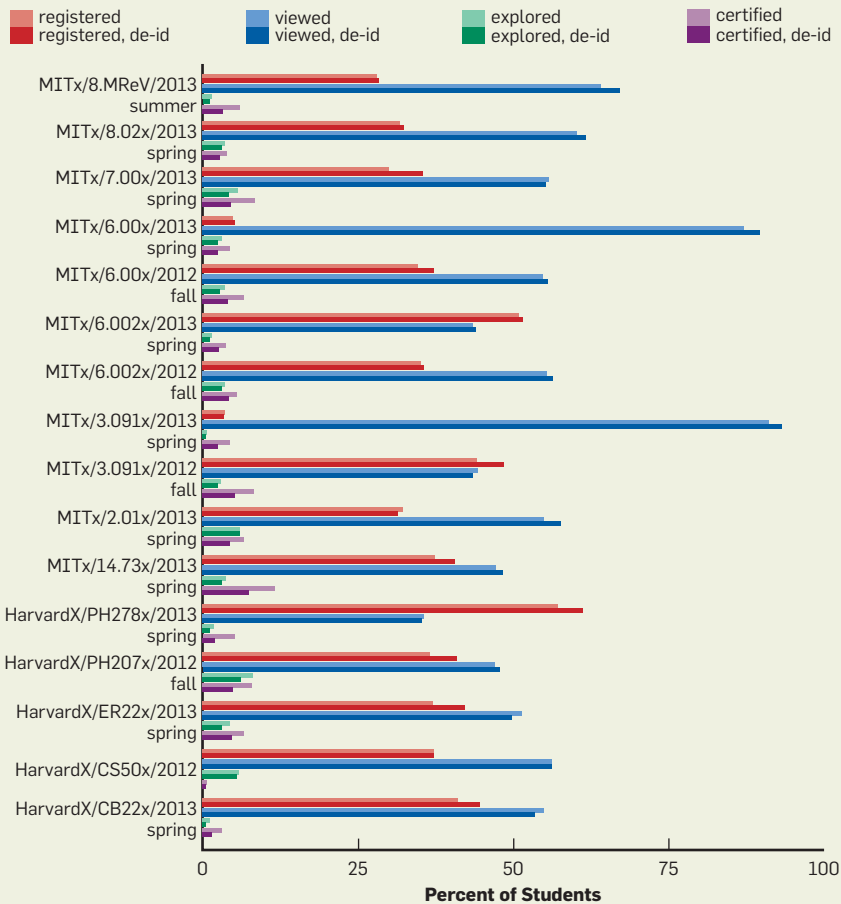
Daries et al. showed de-identification distorted measures of class participation by suppressing records of rare (generally higher) levels of participation. We pursued investigating where distortion of summary statistics was being introduced into the dataset. Intuitively, distortion is introduced whenever a row becomes generalized or suppressed. Under k -anonymity, this occurs only when a row’s combination of quasi-identifier values occurs fewer than k times. If rare quasi-identifier values tend to be associated with high grades or participation levels, then the de-identified dataset would be expected to have a lower mean grade or participation level than the original dataset.

We did, in fact, find a quasi-identifier characteristic whose *frequency of occurrence* is correlated with a numeric attribute is most likely to create distortion in that numeric attribute. Specifically, we confirmed this hypothesis in three ways, using the edX data:

- As privacy requirements increase (that is, k is increased), distortion increases in such numeric attributes as mean grade, shown in Figure 2. The fact that more distortion is introduced as more rows are suppressed is consistent with the hypothesis the association of rare quasi-identifier values with high grades will cause more distortion of the dataset as the privacy standard is increased.

- The deletion of quasi-identifier columns whose values’ frequency of occurrence is highly correlated with numeric attributes results in a decreased amount of distortion in numeric attributes. This supports the hypothesis the presence of a correlation between the frequency of quasi-identifier values and numeric attributes introduces distortion of the dataset by de-identification.

- As the correlation between the frequency of occurrence of quasi-identifier values and other numeric attributes is manually increased, more distortion is introduced into those attributes. This, too, supports the hypothesis the magnitude of a correlation between the frequency of quasi-identifier values and numeric attributes increases distortion of those

Figure 4. Number of rows suppressed vs. bin capacity.**Figure 5. Original and de-identified data, 5-anonymous, 3k bins.**

attributes by de-identification.

What methods may alleviate distortion introduced by de-identification?

The analyses here indicate associations between quasi-identifier traits and numeric attributes may introduce distortion of means by suppression

during de-identification. We therefore consider a prospective role for generalization in alleviating distortion during de-identification.

Since the number of forum posts is the quasi-identifier whose frequency of values is most correlated to grade,

we first explore the effect of generalizing this attribute. As the bin size increases (for example, from 0,1,2,3 to values of 0-1,2-3, and so on), the number of rows requiring suppression decreases, as shown in Figure 3. Further, the mean grade approaches the true value (of 0.045) as bin size increases, suggesting generalization may alleviate distortion by preventing records associated with rarer quasi-identifier values from becoming suppressed.

Generalization, however, can make it difficult to draw statistical conclusions from a dataset. Certain statistical properties of a column, like its mean, can be maintained after generalization by computing a weighted mean of the pregeneralized values within each bin. The average of these bin averages will be equal to the true mean of the pregeneralized values.

Such a solution, however, cannot easily preserve two-dimensional relationships among generalized values. Table 1 illustrates the correlation of the number of forum posts with various numeric attributes becomes increasingly distorted with increasing forum post bin size.

Thus is encountered the fundamental trade-off between generalization and suppression as discussed earlier: although an approach emphasizing suppression may introduce bias into an attribute where a correlation exists between quasi-identifier frequency and numeric attributes, generalization may also distort correlational and other multidimensional relationships inherent within datasets.

Decreasing distortion introduced by generalization. One potential improvement to generalization may be to distribute the number of records more evenly within each bin, using small bucket sizes for values that are well represented and larger bucket sizes for less-well-represented values.

When the number of forum posts is generalized into groups of five for values greater than 10 (for example, 1,2,3,...,11-15, 16-20, and so on), the correlations between the number of forum posts and other characteristics become less distorted than with generalization schemes that use constant bin widths. This suggests optimizing for equal numbers of records within each bin may enable a compromise

between the loss of utility and the distortions caused in numeric analysis, such as correlations between different variables. Using this framework for generalization, let's now explore its relationship to suppression in more detail.

A Trade-Off Between Generalization and Suppression

To reach a compromise between the distortions introduced by suppression and by generalization, we first want to quantify the relationship between suppression and generalization. As generalization is increased, how much suppression is prevented, and does this change at a constant rate as generalization is increased?

Each of the quasi-identifiers was individually binned to ensure a minimum number of records in each bin, termed *bin capacity*. An increase in bin capacity from 1,000 to 5,000 drastically decreases the number of records that have to be suppressed, but this improvement drops off as bin capacity continues to increase. Furthermore, in Figure 4, the decreasing slope of the lines as the bin size increases suggests the larger the chosen bin capacities, the smaller the marginal cost of a greater degree of anonymity.

We then quantify the distortion that was introduced under each choice of bin capacity. Concentrating on sets that were 5-anonymous with bin capacities of 3k, 5k, and 10k, we compare the resulting de-identified datasets with the original set on the percentage of students who simply registered for the course; those who registered and viewed (defined as looking at less than half of the material); those who explored (defined as looking at more than half of the material but not completing the course); and those who were certified (completed the material). This comparison shows the greatest disparity in the de-identification scheme that favors suppression; the results are skewed by as much as 20% with the suppression-emphasis de-identification approach.

A generalization scheme using bin capacities of 3,000 entries, as shown in Figure 5, produces a distribution of participation that is somewhat closer to the original distribution than the suppression-only approach. While

Figure 6. Original and de-identified data, 5-anonymous, 5k bins.

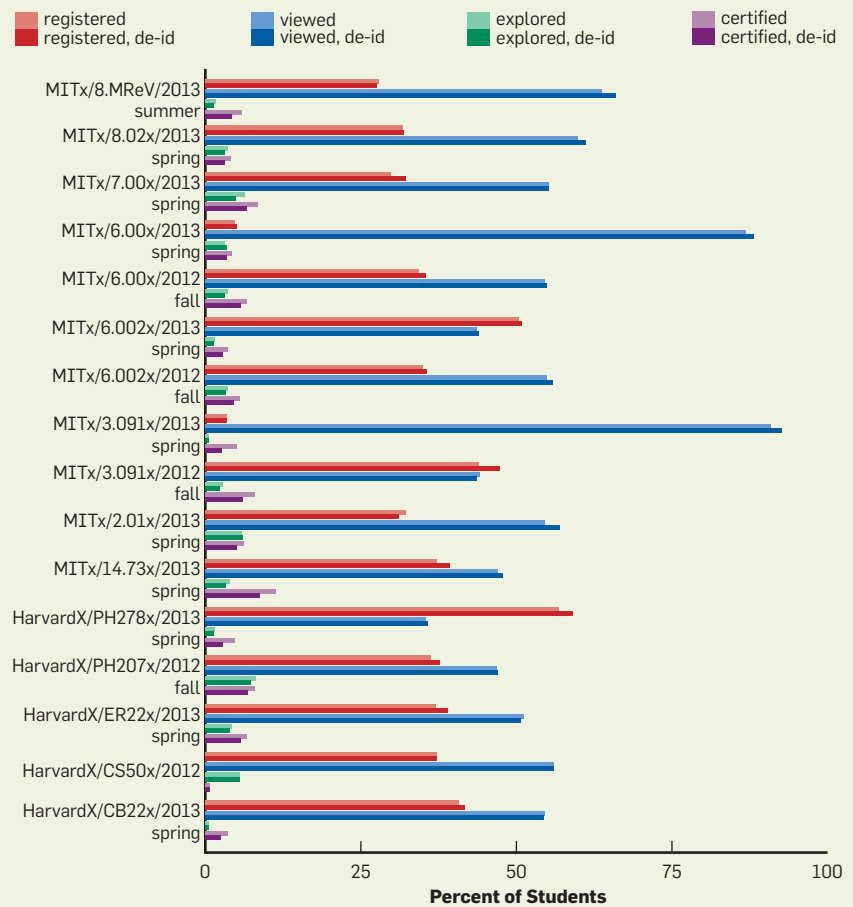


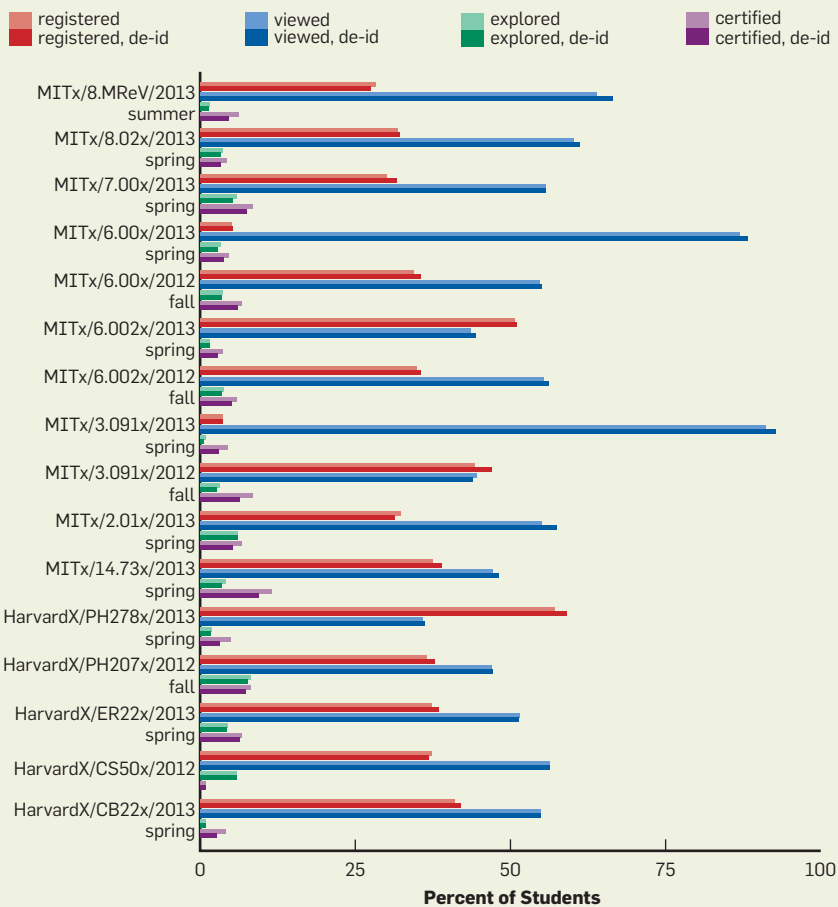
Table 1. Increasing distortion of correlation with increasing bin size.

	Original	Bin size					
		1	2	3	4	5	6
Grade	0.159	0.105	0.0980	0.0919	0.0833	0.0732	0.0533
Viewed	0.0444	0.0683	0.0582	0.0462	0.0372	0.0294	0.0228
Explored	0.127	0.0744	0.0710	0.0661	0.0620	0.0554	0.0418
Certified	0.152	0.0868	0.0810	0.0758	0.0699	0.0598	0.0482
# Active Days	0.236	0.117	0.111	0.106	0.0940	0.0855	0.0649
# Chapters	0.154	0.143	0.127	0.115	0.100	0.0858	0.0715
# Events	0.283	0.103	0.103	0.0964	0.0986	0.0913	0.0597
# Video Plays	0.0929	0.0943	0.105	0.103	0.125	0.110	0.0683

in some categories the distortion is large (such as the certification rates for MITx/7.00x during the Spring semester), others are much closer to the original values.

The situation gets considerably better by using bins with a minimum

of 5,000 entries, as shown in Figure 6. The distribution of participation is nearly the same in the de-identified set as in the original dataset. The maximum difference between the measures is less than three percentage points; most are within one percent.

Figure 7. Original and de-identified data, 5-anonymous, 10k bins.

Moving to a bin capacity of 10,000 gives even better results, as shown in Figure 7. While there are one or two cases of results differing by almost three percentage points, in most cases the difference is a fractional percentage.

As expected, the decrease in the distortion of the mean of certain attributes is accompanied by an increase in the distortion of the correlation between quasi-identifier fields with numeric attributes as bin capacity increases. The table in Figure 8 shows the correlations between the number of forum posts and numeric attributes under various bin capacities. The column corresponding to a bin capacity of 1 represents a suppression-only approach.

Encouraged, we observe a bin capacity of 3,000 produces a dataset whose correlations are close to those of the original, non-de-identified dataset, as shown in Figure 8. Even though a bin capacity of 3,000 did not produce optimal results in terms of minimization of class participation distortion, these results may signal the existence of a bin capacity that produces an acceptable balance of distortion between single- and multidimensional relationships.

Further Opportunities for Optimization

Given these results, the question naturally arises whether bin capacities may be chosen differently for each quasi-identifier in order to minimize distortion further.

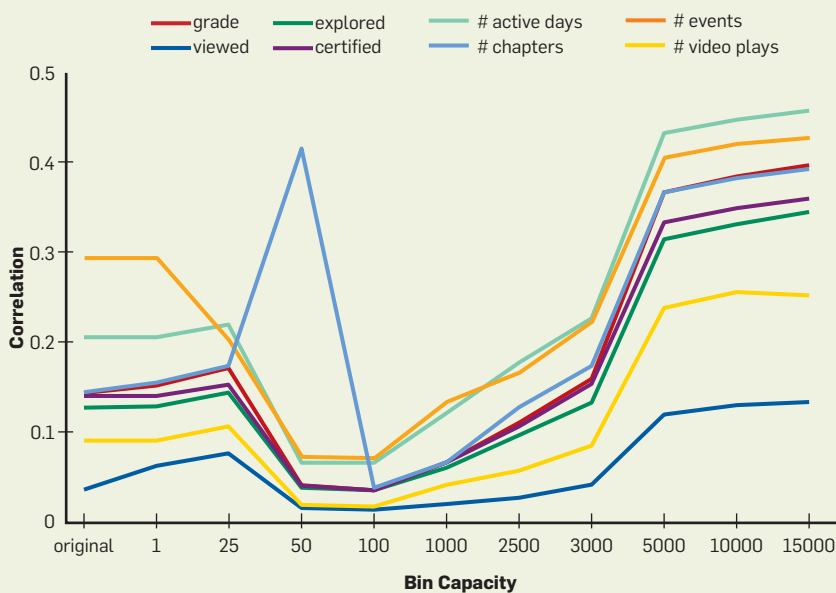
The edX dataset contains two numeric, generalizable quasi-identifier values: year of birth and number of forum posts. Experimentation with different bin capacity combinations yielded the results shown in Table 2. This table illustrates the number of records that must be suppressed with the respective amounts of generalization. It is particularly noteworthy that generalization of each quasi-identifier has uneven effects: the required number of suppressed values drops off much more quickly as the bin capacity for number of forum posts increases, as compared with the bin capacity for year of birth.

Such an analysis of the trade-offs between generalization versus sup-

Table 2. Number of rows suppressed: number of forum posts bin size vs. year of birth bin size.

Number of Forum Posts: Bin Capacity	Year Of Birth: Bin Capacity					
		100	200	300	400	500
	100	1119	1074	1057	1017	1009
	200	577	530	508	468	456
	300	400	v 356	337	291	280
	400	355	303	281	238	221
	500	283	236	210	169	167
	600	261	208	187	146	131
	700	220	161	140	107	100
	800	224	164	148	112	101
	900	208	138	129	88	76
	1000	186	122	111	72	61
	1100	183	120	109	70	61
	1200	164	94	78	53	46
	1300	145	81	70	48	41
	1400	145	81	70	48	41
	1500	136	78	63	42	35
1600	134	76	61	45	35	

Figure 8. Correlation between number of forum posts with various attributes.



	Bin Capacity							
	Original	1	100	1000	2500	3000	5000	10000
Grade	0.145	0.155	0.037	0.071	0.113	0.161	0.369	0.386
Viewed	0.039	0.063	0.013	0.019	0.028	0.038	0.122	0.129
Explored	0.127	0.128	0.037	0.060	0.095	0.133	0.314	0.332
Certified	0.138	0.141	0.036	0.067	0.108	0.153	0.333	0.349
# Active Days	0.207	0.207	0.068	0.123	0.177	0.227	0.434	0.449
# Chapters	0.145	0.155	0.039	0.067	0.128	0.174	0.367	0.383
# Events	0.294	0.294	0.073	0.135	0.167	0.222	0.407	0.420
# Video Plays	0.091	0.091	0.016	0.043	0.058	0.084	0.240	0.256

pression becomes exponentially harder as the number of quasi-identifier values increases. A brute-force method of calculating the number of suppressed records would demand excessive computation time with datasets like edX's that contain six quasi-identifier fields. The development of approximation algorithms for these calculations would enable researchers to quickly determine a near-optimal generalization scheme that strikes an ideal balance between distortions introduced by generalization versus suppression. This is an area where further research is needed.

Conclusion

De-identification techniques will continue to be important as long as the

regulations around big datasets involving human subjects require a level of anonymity before those sets can be shared. While there is some indication regulators may be rethinking the tie between de-identification and ensuring privacy, there is no indication the regulations will be changed anytime soon. For now, sharing will require de-identification.

But de-identification is hard. We have known for some time it is difficult to ensure the dataset does not allow subsequent re-identification of individuals, but we now find it is also difficult to de-identify datasets without introducing bias into those sets that can lead to spurious results.

A combination of record suppression and data generalization offers a

promising path to solving the second of these problems, but there seems to be no magic bullet here; our best results were obtained by trying a number of different combinations of generalization, sizing, and record suppression. There is further work to be done, such as investigating the possibility of choosing different bin capacities for different quasi-identifiers, which may mitigate some of the distortions introduced by anonymity. We are more confident than we were a year ago that some form of de-identification may allow sharing of datasets without distorting the analyses done on those shared sets beyond the point of usefulness, but there is much left to investigate. C

Related articles on queue.acm.org

Privacy, Anonymity, and Big Data in the Social Sciences

Jon P. Daries et al.

<http://queue.acm.org/detail.cfm?id=2661641>

Broadcast Messaging: Messaging to the Masses

Frank Jania

<http://queue.acm.org/detail.cfm?id=966719>

Modeling People and Places with Internet Photo Collections

David Crandall, Noah Snavely

<http://queue.acm.org/detail.cfm?id=2212756>

References

1. Daries, J.P. et al. Privacy, anonymity, and big data in the social sciences. *Commun. ACM* 57, 9 (Sept. 2014), 56–63.
2. Sweeney, L. *k*-anonymity: A model for protecting privacy. *Intern. J. Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.
3. Young, E. Educational privacy in the online classroom: FERPA, MOOCs, and the big data conundrum. *Harvard Journal of Law & Technology* 28, 2 (2015), 549–592.

Olivia Angiuli (oangiuli@post.harvard.edu) is a data scientist at Quora. She is ultimately interested in harnessing big data for social good.

Joe Blitzstein (blitz@fas.harvard.edu) is a professor of the practice of statistics at Harvard University, whose research is a mixture of statistics, probability, and combinatorics. He is especially interested in graphical models, complex networks, and Monte Carlo algorithms.

Jim Waldo (waldo@seas.harvard.edu) is a Gordon McKay Professor of the Practice in Computer Science, a member of the faculty of the Kennedy School, and the Chief Technology Officer at Harvard University. His research centers around distributed systems and topics in technology and policy, especially around privacy and cyber security.

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

**Catering to developers' strengths
while still meeting team objectives.**

BY KATE MATSUDAIRA

Lean Software Development—Building and Shipping Two Versions

ONCE UPON A time (and isn't that how all good stories start?) I was managing a software team and we were working on several initiatives. Projects were assigned based on who was available, their skillsets, and their development goals. This resulted in two developers, let's call them Mary and Melissa, being assigned to the same project.

Mary and Melissa had been working together for a few weeks when I started hearing complaints in my one-on-ones with each of them about the other. Mary was complaining that Melissa was taking too long to do her part, and spending time on unit tests that did not make sense because things were in flux with the project. Meanwhile, Melissa was complaining that

Mary wrote sloppy code and did not write enough tests (she even showed me comments in the code like "hack:...." and "to do: someone should add more error handling here"). Each of them had valid points and feedback for the other.

I spent the next few weeks coaching each of them on how to improve and address the other's concerns. With Mary I was focused on pushing her to write more tests and not just throw things together, and with Melissa I was focused on having her prototype something quickly and then add the polish after it was working. Both of them tried really hard, and both were miserable. It just was not what they were meant to do.

These two women were just not meant to collaborate. This got me thinking: how could we change the process so they both could write code the way they liked to create (and that catered to their strengths) and we could still meet our team objectives?

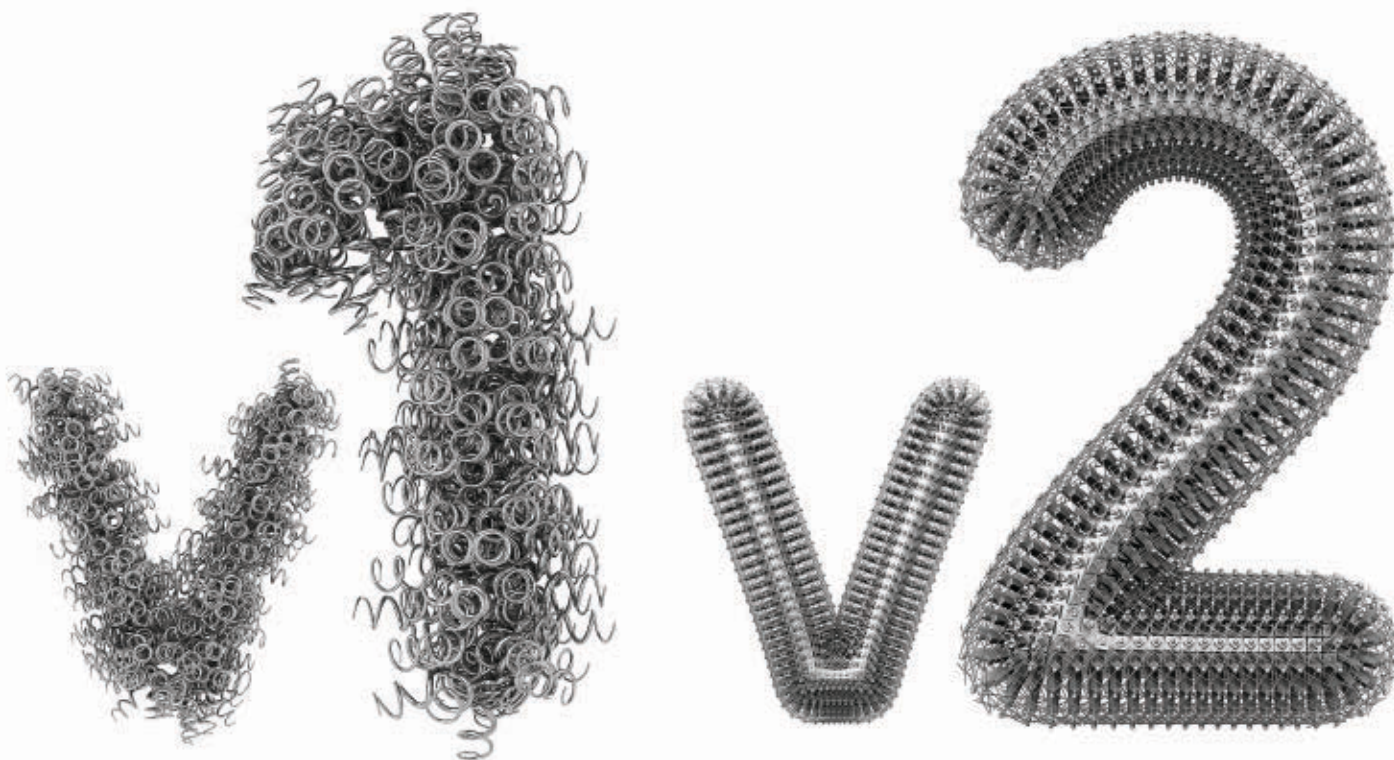
From this was born the v1/v2 development process.

You see, certain developers love building the first versions or prototyping—they are the ones who love hacking things together to get something working quickly. They love (and are best at) building version 1 of a product. The other type of developers love building the second version. They see their code as a craft and write unit tests for everything. "Test coverage" and "beautiful code" are phrases they use a lot.

And of course this definition is not black and white—there are people who fall on different sides of the line at different times, so it is more like a spectrum.

Typically, the v1 person does not like working with the v2 person, and vice versa—not for personal reasons, but because of the ways they think and create software. At the heart of it, the fundamental thing they enjoy about what they do is different.

By changing the way we thought about software, we were able to address this concern and actually make the team more agile and solve some business problems along the way.



The V1/V2 Process

When it comes to software development there are many ways to build and create products.

In my experience it is difficult to build the right product the first time. By “right” I do not just mean something customers will want and use, and that hopefully will generate revenue, but also the right technical solution. How customers will use products can be difficult to predict. For example, before you ship the first version of a product it can be difficult to answer questions like:

- ▶ How quickly will the data grow?
- ▶ How fast will writes need to be?
- ▶ Will it be feasible to write directly to the data store or do you need a queue to manage writes?
- ▶ What will the throughput be, and is it enough?
- ▶ Will the system scale with usage?

Even if you could reasonably answer these questions, or if you built a system that took all these things into account such that you could “just add hardware,” the effort would probably be larger than if you had simply pushed to ship a first version quickly. And it is possible (if not always probable) such a system would be over-engineered (since it was designed to solve prob-

lems or scale in a way that may not be useful), or even have other unforeseen issues. While there are certainly pros to doing it right the first time, in practice it is very difficult to get it right, and building the end-all system will take longer, delaying the answers to all of the questions indicating you are moving in the right direction.

There is a popular movement called Lean Startup that advocates fast iteration and data collection to home in on the right product and requirements as quickly as possible. It is all about figuring out what customers will use and then building those things. These ideas can also be applied to the technical ways we build products.

Build V1 Fast, Then Build V2 Right

Similar to creating a minimum viable product (MVP), build the minimum viable technical implementation that meets the business goals (meaning the product can meet current usage requirements and performance).

Most engineering teams have been taught to think about software design and architecture from the beginning and to build something that can scale. However, if you do not have any customers yet, scaling really is not your

challenge. So in some ways, designing for scale when you do not have to scale yet is solving the wrong problem.

Of course I am not advocating doing something stupid, or writing bad code, but do not over-engineer or solve problems before you have them. Here are some examples:

- ▶ Foregoing the deployment of a multi-node Cassandra cluster and storing the initial data in a single database (with backup) that is fast and easy to make arbitrary queries against;
- ▶ Faster UI development using standard forms and charts versus interesting shapes or fancy custom graphs; and
- ▶ Cutting corners on unit tests (because the units are in flux) or comments and documentation.

All these shortcuts will help you get something out faster. They may not be best practices but they do ensure you are building the right thing.

In many ways, this process is about looking at the problems differently and, instead of solving the “big” problem, focusing on speed and efficiency—getting the highest ROI (return on investment) for your development resources.

So you modify your development process to ship the first version quick-

ly, and then once v1 is out in the wild, plan to start on v2 right afterward.

Then your v2 can address all the problems with the first version of the product. You can improve usability, cut or add features based on customer usage and feedback, pick the right technology and libraries to power the product, and spend time writing acceptance tests and unit tests are unlikely to change. From a business perspective, you get the product out there faster, so it is easy to see if this product is going to drive revenue and success.

Why This Process Makes Sense

It reduces the risk of over-engineering. As noted earlier, getting something out quicker and understanding how it is used (or how successful it is in the market) will help confirm you are building the right product and solving the right technical problems for your business. This allows you to mitigate the risk of building or scaling out systems that may not address the production issues.

If you build the wrong thing you can correct it. Since the product is truly the minimum viable product, you reduce the risk of wasting resources creating the wrong product. Furthermore, you will get feedback earlier in the release cycle and can pivot or make changes faster (since it is always more difficult to change a system with a plethora of moving parts).

Chances are you will get to market faster. Building something quickly will help you launch sooner, which will help you test and determine the viability of your product faster than waiting for the bigger launch. This means you will not spend time building software your customers will not use or you cannot sell.

Staffing and personnel happiness. This is one of the biggest upsides to this method. Software engineers get to do what they love, in a way that caters to their strengths. Moreover, they are not forced to work with people who do not share their approach, and they do not have to work in a code base they cannot sculpt into their desired piece of art.

So, What Are the Downsides?

Total time to build is more than if you just got it right the first time. In this model it is likely v2 development will

be finished later and the total development time will be greater than if you built the product once. If you build the v1 using one technology (say, a MySQL database that will not scale the way you need it to), and then the v2 needs a different one (say, a proper key value store for faster queries), the team will spend time shipping and building expertise on one database, only to have to learn and manage another.

Operations can be a headache. Most seasoned engineers will tell you that building something really fast can result in operations nightmares if the usage or data grows in a way that was not considered. No team likes being operational and fighting fires. Of course, if this is a problem it probably means your product has achieved some adoption and success, so at least you are solving something that produces real value. However, even for the v1 it is still worthwhile to identify risks and potential bottlenecks, so in the event there are problems the team will have some ideas for solutions.

So You Want to Try It at Home ... Err ... Work?

Like any software process or methodology it is important to assess if this makes sense for your business, product, and company—it certainly is not a one-size-fits-all. However, it definitely has its merits, so feel free to steal the pieces you like.

If you do embark on this journey, here are some pointers from my experience:

Get management buy-in. This is really critical, because a key to this approach being successful is always building a v1 and a v2. If you do not have buy-in and your boss is content to keep and operate the v1, your team is going to hate you. Seriously. No one wants to get stuck supporting crufty v1 software indefinitely. So make sure everyone knows and understands the plan ahead of time.

Measure everything. In order to build the right v2 you need to know what to build. Where are your bottlenecks? How is the data growing? Without this data it is very difficult to realize many of the advantages of this model. Make sure you think about what you need to measure, and how you will measure it, before the product ships.

The v2 will take much longer than the v1. I have personally seen this process evolve across four projects (no, it is not a large sample set), and in each case the v2 took more than twice as long as the v1 to create and launch. Of course, every v2 I have launched had a lot of extra features, and they did not always leverage a lot of the v1 code. Some people assume the v2 will be faster, since “you have already built it once before,” but I have not found this to be the case.

Have a tight feedback loop with your customers. Make sure it is easy to get insight into what people use, what they like, and what they do not like—that way you can build a great v2.

Make sure v1s and v2s are celebrated equally. Sometimes teams or companies celebrate the first launch of a feature in a much grander way than the next version. But in this case the next version is what you need to grow your business. It is just as important as v1, if not more so. Make sure you value each of these equally, or people will gravitate to the project with the most fanfare, not necessarily the one suited to their strengths and talents.

Be open to people being good at both v1s and v2s. Sometimes I think I have someone pegged and then they prove me wrong. Be open to the fact there are not two types of people, and some people are just good at everything. (I just wish I was one of those people!)

I hope some of these ideas will prove useful for you and your team. If anything, let this anecdote inspire you to question the way you are doing things, and look critically at innovative ways to improve how you build software. 

Related articles on queue.acm.org

A Conversation with Tim Bray

<http://queue.acm.org/detail.cfm?id=1046941>

Big Games, Small Screens

Mark Callow, Paul Beardow, and David Brittain
<http://queue.acm.org/detail.cfm?id=1331296>

Major-league SEMAT - Why Should an Executive Care?

Ivar Jacobson, Pan-Wei Ng, Ian Spence, and Paul E. McMahon
<http://queue.acm.org/detail.cfm?id=2590809>

Kate Matsudaira (katemats.com) is the founder of her own company Popforms. Previously, she worked in engineering leadership roles at companies such as Decide (acquired by eBay), Moz, and Amazon.

© 2015 ACM 00010782/15/11 \$15.00



**Optimizing NUMA systems
applications with Carrefour.**

**BY FABIEN GAUD, BAPTISTE LEPEERS, JUSTIN FUNSTON,
MOHAMMAD DASHTI, ALEXANDRA FEDOROVA,
VIVIEN QUÉMA, RENAUD LACHAIZE, AND MARK ROTH**

Challenges of Memory Management on Modern NUMA Systems

MODERN SERVER-CLASS SYSTEMS are typically built as several multicore chips put together in a single system. Each chip has a local DRAM (dynamic random-access memory) module; together they are referred to as a *node*. Nodes are connected via a high-speed interconnect, and the system is fully coherent.

This means that, transparently to the programmer, a core can issue requests to its node's local memory as well as to the memories of other nodes. The key distinction is that remote requests will take longer, because they are subject to longer wire delays and may have to jump several hops as they traverse the interconnect. The latency of memory-access times is hence non-uniform, because it depends on where the request originates and where it is destined to go. Such systems are referred to as non-

uniform memory access (or NUMA).

Systems with NUMA characteristics were built as early as the 1980s, and along with the hardware operating system, support for NUMA has evolved. Modern NUMA systems are quite different from the old ones, so we must revisit our assumptions about them and rethink how to build NUMA-aware operating systems. This article evaluates performance characteristics of a representative modern NUMA system, describes NUMA-specific features

in Linux, and presents a memory-management algorithm that delivers substantially reduced memory-access times and better performance.

A Modern NUMA System

NUMA systems consist of several nodes, each containing a subset of the system's CPU cores and a portion of its RAM. If a

core accesses memory from within the same node, it is called a *local access*. Similarly, an access to a different node is called a *remote access*. Remote accesses have longer latencies than local ones, because they must traverse one or more *interconnect links*, communication pathways between nodes that also service cache-coherency traffic. Figure

1 is a diagram of a typical NUMA system with four nodes and four cores per node. At the time of this writing, NUMA systems are built with up to eight nodes and 10 cores per node.

Current x86 NUMA systems are cache coherent (called ccNUMA), which means programs can transparently access memory on local and remote nodes without changes to the code or special operating system support. This allows easy migration to NUMA systems, but it does not address important performance considerations. A naïve implementation—for example, a program that allocates all of its memory on a single node—can easily cause excessive remote accesses or the overloading of a memory controller.

New vs. old NUMA systems. Avoiding performance pitfalls on NUMA systems requires considering how the nodes are connected, where the program's memory is placed, and how it accesses that memory. Previous NUMA-aware operating systems focused on locality, attempting to minimize the number of remote accesses at all costs in order to avoid the performance penalty. Modern NUMA systems, however, have a strikingly different latency profile compared with the older ones. A remote access takes approximately 30% longer than a local one,^{2,7} while on older hardware, it could take up to seven times longer.³ The remote-access penalty is substantially reduced on modern NUMA systems.

On the other hand, current CPUs can generate an immense load on the memory subsystem, causing congestion on memory controllers and interconnect links (if requests are remote). If multiple cores are heavily accessing a single node, memory latencies can be as long as 1,200 cycles (!) due to congestion, while normal latencies are only around 300 cycles. Avoiding memory controller and interconnect congestion, therefore, becomes the key concern on modern NUMA systems. Here, we examine the effects of congestion on performance.

NUMA Performance—Locality And Congestion

Benchmarks can provide a complete picture of the performance characteristics of NUMA systems. The Numerical Aerodynamic Simulation (NAS),¹ Princeton Application Repository for

Figure 1. A modern NUMA system.

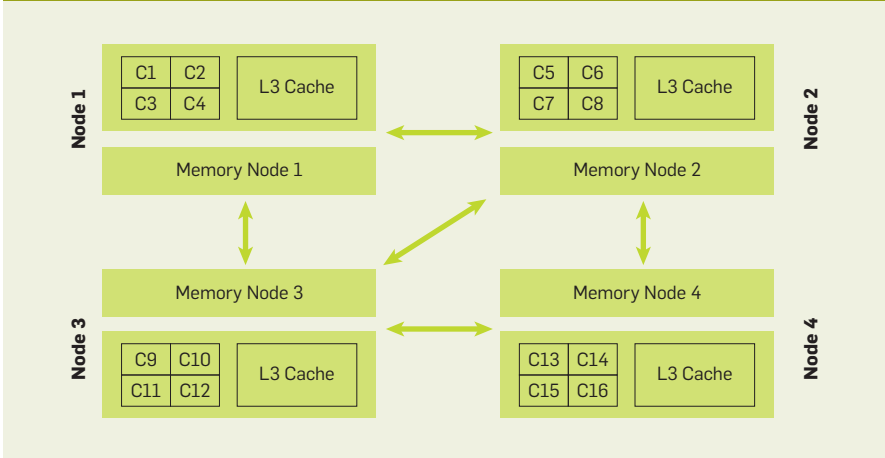
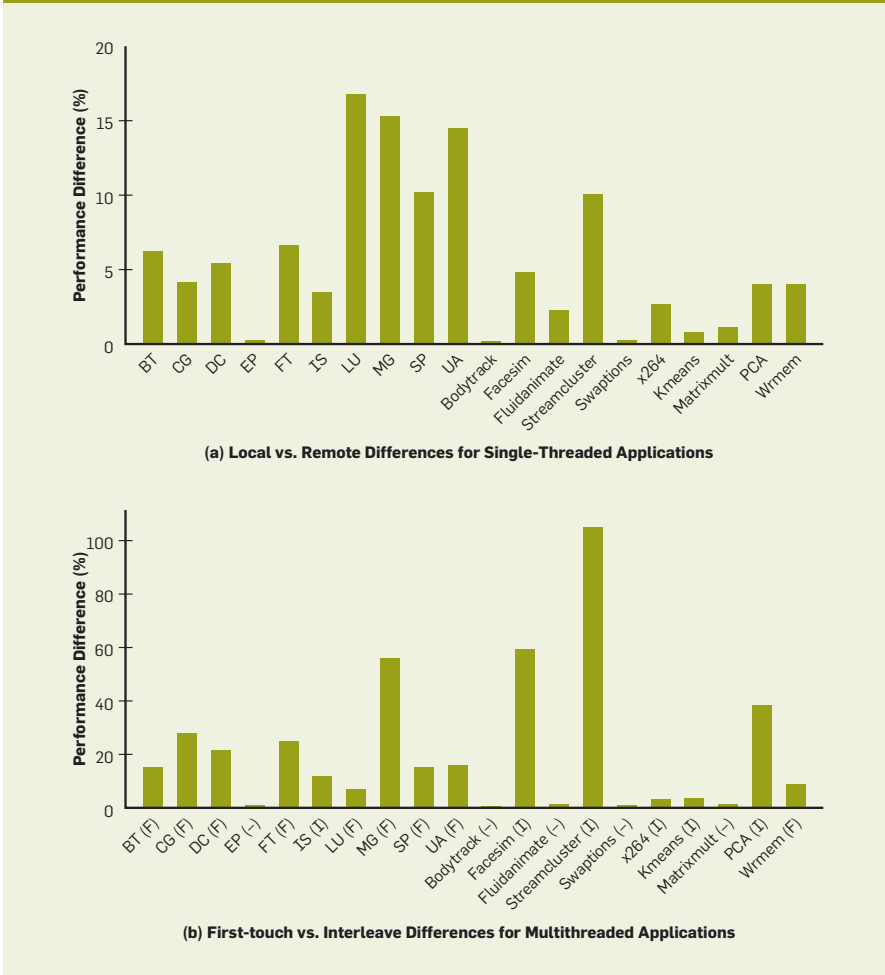


Figure 2. Performance differences.



Shared-Memory Computers (PARSEC)¹⁰ and Metis MapReduce⁹ suites were chosen here because they have a CPU utilization greater than 30%, allowing the focus to be on NUMA effects and not other factors such as disk I/O or blocking synchronization. The experiments were conducted on an AMD server with four quad-core CPUs, as we describe later.

The first experiment quantified the effect of only the remote-access penalty, without the presence of memory-subsystem congestion. To do this, the benchmarks were run with only a single thread, limiting the pressure on the memory controllers and interconnects, and then compared under two different memory configurations: local and remote. In the local memory configuration, applications were executed with their memory and thread on the same node. In the remote memory configuration, standard Linux tools were used to force the application thread to run on a different node from its memory. Therefore, in the remote-memory case all memory accesses were remote, and in the local-memory case all memory accesses were local.

Figure 2a shows performance differences between local and remote memory configurations for the single-threaded versions of applications used in this experiment. Performance never degraded by more than 20%, even when all memory requests were remote.

Although the remote-access penalty is worth minimizing when possible, that is not the whole story of NUMA performance effects. To demonstrate this, the benchmarks were run using multiple threads, with one thread per core and under two different common NUMA memory-allocation policies: *first-touch* and *interleave*. Linux's default policy is *first-touch*, where memory is allocated on the same node as the thread that first accesses a memory page. The *first-touch* policy is meant to maximize local accesses over remote accesses, but of course it cannot guarantee local accesses because threads on multiple nodes can share data. On the other hand, the *interleave* policy distributes memory allocations equally on all nodes regardless of which threads access it. *Interleaving* ensures that memory allocations are balanced but not necessarily that

Avoiding performance pitfalls on NUMA systems requires considering how the nodes are connected, where the program's memory is placed, and how it accesses that memory.

memory accesses will be balanced. Both policies work at the granularity of a page (typically 4KB).

Figure 2b shows the absolute performance difference between *first-touch* (F) and *interleave* (I) for multithreaded versions of applications. The applications are labeled (F) or (I) depending on which policy performed best. The figure compares the two policies by showing the performance difference between the best and worst policy for each benchmark. If there was no negligible difference, the application is labeled (-). The first observation to make is that no one policy is best for all applications. Several applications perform best with the *first-touch* policy, but many prefer interleaving. The second observation is that NUMA effects beyond the remote-access penalty can indeed severely affect performance. For the *Streamcluster* benchmark, using the *first-touch* policy nearly doubled the running time over the *interleave* policy.

We further investigated the NUMA performance characteristics of *Streamcluster* and *PCA* (another benchmark that has significant performance loss with the *first-touch* policy) by using hardware performance counters to gather the following key metrics:

- *Local access ratio*. The portion of RAM accesses that result in a local access.
- *Memory latency*. The number of cycles it takes to perform a RAM access, on average. Higher latencies mean the CPU must stall for longer on a last-level cache miss, which will negatively affect performance.
- *Memory-controller imbalance*. The standard deviation (as a percentage of the mean) of the load on the memory controllers, where the load is measured as the number of requests per time unit. Along with interconnect imbalance, it is a sign of congestion.
- *Average interconnect usage*. The average bandwidth utilization of the interconnect links. A low interconnect usage could imply that either the application is not very memory intensive, or there is imbalance because some links are left underutilized.
- *Average interconnect imbalance*. The standard deviation (as a percentage of the mean) of the bandwidth utilization of interconnect links.
- *L3MPKI*. The number of last-level cache misses per 1,000 instructions.

This is a relative indicator of how much pressure an application puts on the memory subsystem and of how sensitive an application is to memory latencies.

► *Instructions per cycle (IPC)*. For the same application and workload, a higher IPC means better performance.

The metrics for Streamcluster and PCA are reported in the accompanying table. Traffic congestion effects are highlighted by differences in key NUMA metrics for each benchmark under the first-touch (F) and interleave (I) policies. The performance difference between the two NUMA policies cannot be explained by a change in the last-level cache miss rate, which stays the same. Nor can it be explained by the local-access ratio, which stays the same for Streamcluster and in fact is worse for PCA in the case of the better performing interleave policy. The local-access ratio of PCA drops from 33% to 25% when interleaving memory, but performance improves significantly, so the conclusion is that better locality does not necessarily improve performance.

When using the first-touch policy both applications show signs of congestion with high last-level cache miss rates, memory-controller imbalance, and interconnect imbalance. The congestion results in high memory latencies. In the case of Streamcluster, the average memory latency with the first-touch policy is more than double the latency of the interleave policy. Interleaving balances the memory among the nodes, which reduces traffic hotspots and congestion, and therefore improves memory latency and overall performance. A visualization of the memory traffic and congestion of Streamcluster is shown in Figure 3; traffic imbalance under first-touch is shown on the top and interleaving on the bottom. Nodes and links bearing the majority of the traffic are shown proportionately larger in size and in brighter colors. The percentage values show the fraction of memory requests destined for each node.

NUMA Memory Placement Strategies

The results in Figure 2 and the table motivate a NUMA memory-management algorithm that places importance on congestion management, rather than focusing solely on reduc-



A NUMA memory-management algorithm should place importance on congestion management, rather than focusing solely on reducing remote accesses.



ing remote accesses. That is not to say reducing remote accesses is not important (they do, after all, add latency and contribute to interconnect congestion), but this should not be the only goal. Managing congestion effectively means being concerned with how the memory-access traffic is spread across the system. It is not enough simply to use interleaving. Many applications do not suffer from imbalance, so they would needlessly incur remote-access delays (for example, the benchmarks in Figure 2b that prefer the first-touch policy). The algorithm must be able to intelligently place memory based on the application's access patterns, such that congestion is reduced whenever possible but locality is not sacrificed when congestion is minimal. Since access patterns are not known *a priori*, the algorithm must also be able to determine the access patterns and move memory at runtime with low overhead.

Later, we present our NUMA algorithm, called Carrefour, which takes all of these considerations into account. First, though, we describe existing NUMA tools available on Linux. (This article gives an overview of the Carrefour algorithm. Please see Dashti et al.⁶ for a complete discussion, including implementation details and exhaustive experimental results.)

NUMA on Linux. Linux allows administrators to set the NUMA policy for applications via the numactl utility. The NUMA policies available are first-touch, interleave, and restricting allocations to specific nodes. As described earlier, first-touch is the policy of allocating memory on the same node as the CPU that first accesses the memory page, and the interleave policy distributes memory pages on nodes in a round-robin manner.

Linux exposes manual NUMA memory-management functions to programmers through the libnuma library and associated system calls. This allows a program to query the NUMA topology, set NUMA policies for specific address ranges, and migrate memory pages to different nodes at runtime. (See Lameter⁸ for detailed information on Linux's NUMA facilities.)

Linux also provides robust support for hardware performance counters, which are used for counting CPU events such as cycles elapsed, instructions re-

tired, branch mispredictions, or cache misses. These events can be used to calculate the important NUMA metrics listed previously. Perf is the standard Linux tool for using hardware performance counters to profile applications. It can gather data for several events with negligible performance overhead and minimal developer effort.

Hardware instruction sampling is an advanced CPU feature similar to performance counters. With instruction sampling, a proportion of instructions are tagged by the hardware. Tagged instructions will record extra information about their execution. It is necessary for obtaining some NUMA-related statistics, including memory-access latency and the addresses of memory accesses. The feature is implemented as IBS (instruction-based sampling) on AMD CPUs and as PEBS (precision event-based sampling) on Intel CPUs. Unfortunately, Linux support for hardware-instruction sampling is limited and requires a custom kernel module for most uses.

AutoNUMA. AutoNUMA⁴ aims to provide Linux with a more proactive NUMA solution. A kernel task routinely iterates through the allocated memory of each process and tallies the number of memory pages on each node for that process. It also clears the present bit on the pages, which will force the CPU to stop and enter the page-fault handler when the page is next accessed.

In the page-fault handler it records which node and thread is trying to access the page before setting the present bit and allowing execution to continue. Pages that are accessed from remote nodes are put into a queue to be migrated to that node. After a page has already been migrated once, though, future migrations require two recorded accesses from a remote node, which is designed to prevent excessive migrations (known as page bouncing).

AutoNUMA's memory-placement algorithm, now known as Automatic NUMA Balancing, has been merged into the Linux kernel. It can be enabled through the sysctl interface by setting `kernel.numa_balancing` to 1.

AutoNUMA also uses thread placement to try to improve locality. The scheduler will consider migrating or swapping threads if it will cause more of the thread accesses to be local (based on the gathered page-fault statistics)

but not at the cost of introducing load imbalance among CPUs.

Carrefour is a memory-placement algorithm for NUMA systems that focuses on traffic management: placing memory so as to minimize congestion on interconnect links or memory controllers.

Carrefour uses global information and memory-usage statistics to inform three primary techniques for limiting congestion:

- *Memory collocation.* Moving memory to a different node so accesses will likely be local.

- *Replication.* Copying memory to several nodes so threads from each node can access it locally (useful for read-only and read-mostly data).

- *Interleaving.* Moving memory such that it is distributed evenly among all nodes.

All three of these techniques have been analyzed individually in prior studies, but Carrefour combines them into a novel algorithm that is effective for modern NUMA systems.

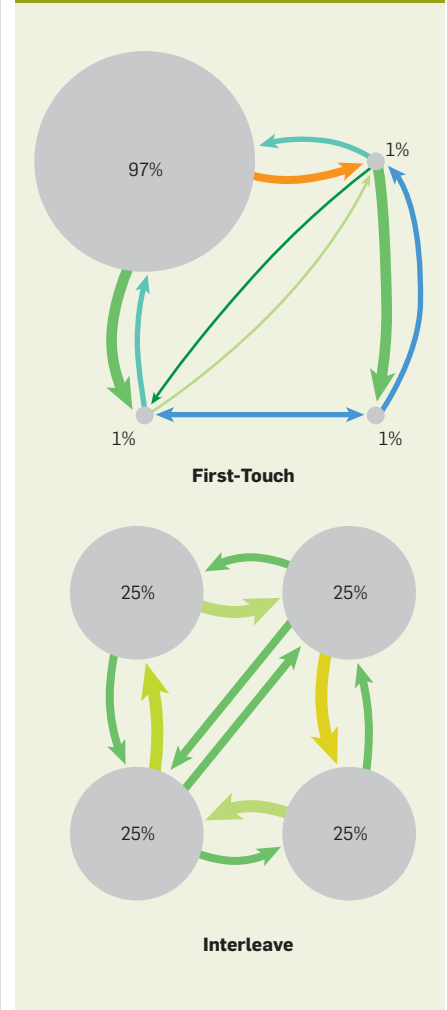
To combine these techniques and apply them judiciously, Carrefour collects per-page, per-process, and global statistics from hardware performance counters. Carrefour also uses hardware-instruction sampling to log which threads and nodes access which memory pages. Instruction sampling lets Carrefour gather many more samples at low overhead than AutoNUMA's page-fault handler technique because it has lower overhead per sample.

The first metric Carrefour uses is the number of RAM accesses per microsecond. If it is less than the threshold (50 in our experiments), then the rest of the algorithm is completely disabled until it becomes greater than the threshold. If the rate of RAM accesses is low, then

the application is unlikely to benefit from better memory placement, so this rule prevents the overhead of Carrefour when it is not needed. If the algorithm does remain enabled, then Carrefour iterates over the memory pages for which it has gathered statistics and applies the replication, collocation, and interleaving techniques.

We implemented memory replication in the Linux kernel with a patch

Figure 3. A composable select element.



Traffic congestion effects.

	Streamcluster		PCA	
	Best (I)	Worst (F)	Best (I)	Worst (F)
Local-access ratio	25%	25%	25%	33%
Memory latency	476	1197	465	660
Memory-controller imbalance	8%	170%	5%	130%
Interconnect imbalance	22%	85%	20%	68%
Interconnect usage	59%	33%	48%	31%
L3MPKI	16.85	16.89	7.35	7.4
IPC	0.29	0.15	0.52	0.36

to the virtual memory layer, and our implementation is able to automatically maintain consistency when there is a write to a replicated page. To enable replication there must be enough

free memory available, and at least 95% of the application's memory accesses must be reads because the performance cost of a write to a replicated page is quite high. The rule for replicat-

ing particular pages is simple: pages are replicated if they are observed to have accesses from multiple nodes in read-only mode. Replication improves both locality and congestion because a replicated page can be accessed locally from more than one node.

Collocation is enabled if the local access ratio is less than 80%. Pages that have been accessed only by a single remote node are migrated to that node, thereby improving locality.

The primary purpose of interleaving is to alleviate congestion by distributing memory—and, therefore, memory accesses—among multiple nodes. The first step is to consider the memory-controller imbalance. If it is below 35%, then interleaving is deemed unprofitable and it is disabled globally. Otherwise, pages that have recorded read and write accesses from more than one node are migrated to a random node, where the probability of being migrated to a specific node is inversely proportional to the relative load on that node's memory controller.

The source code for Carrefour is available at <https://github.com/Carrefour>.

Evaluation

Testbed. All experiments were conducted on an AMD system with 64GB of RAM and four quad-core Opteron 8385 processors running at 2.3GHz. It is divided into four NUMA nodes with four cores and 16GB of RAM per node (the topology is shown in Figure 1) interconnected with HyperTransport 1.0 links.

The operating system was Linux kernel v3.6, and the AutoNUMA configuration used v27 of the patch.

A variety of multithreaded benchmarks were used for the evaluation: PARSEC benchmark suite v2.1,¹⁰ FaceRec v5.0,⁵ Metis MapReduce benchmark suite,⁹ and the NAS parallel benchmark suite v3.3.¹ The PARSEC benchmarks used the “native” workloads, and the NAS benchmarks used problem sizes that provided running times of at least 10 seconds. Applications that had CPU utilizations below 33% were excluded because they are not affected by memory-management policies. Each configuration and benchmark was run 10 times, which resulted in standard deviations of less than 2% for the Carrefour, default Linux, and interleaving configurations. AutoNUMA

Figure 4. PARSEC and Metis.

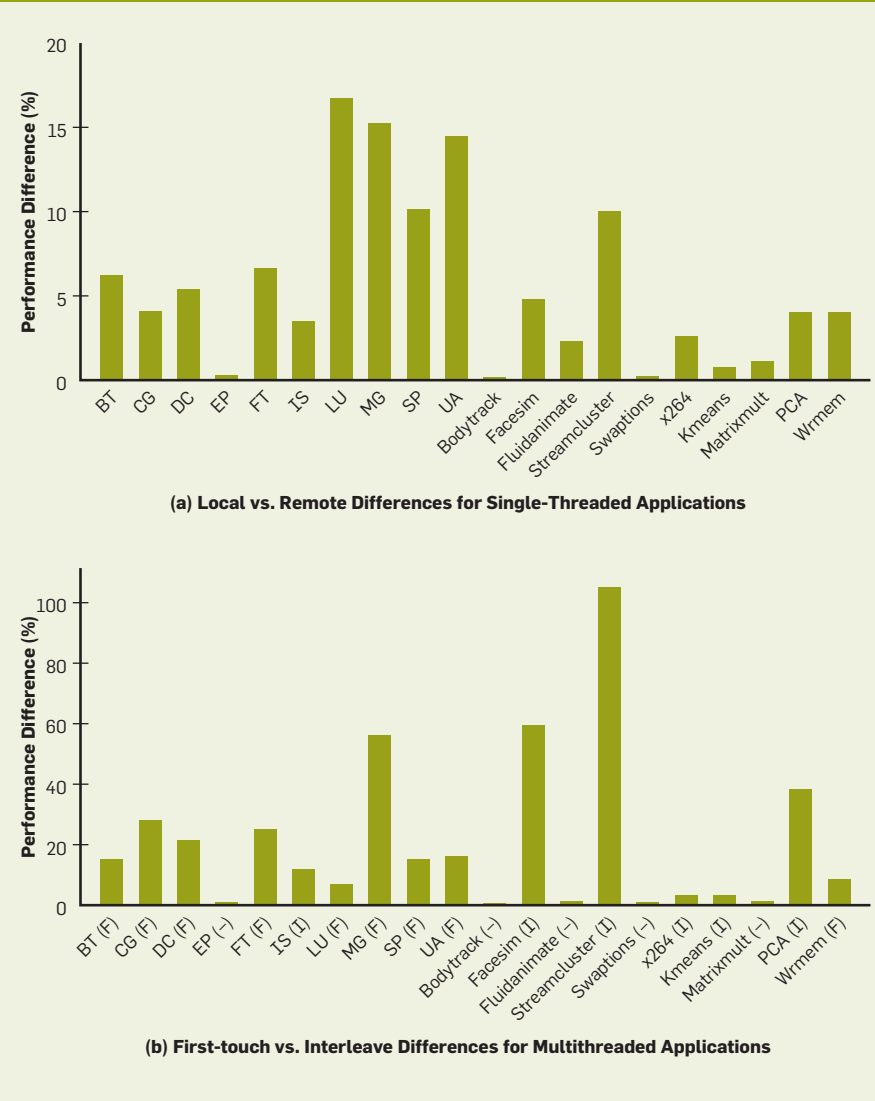


Figure 5. NAS parallel benchmarks.

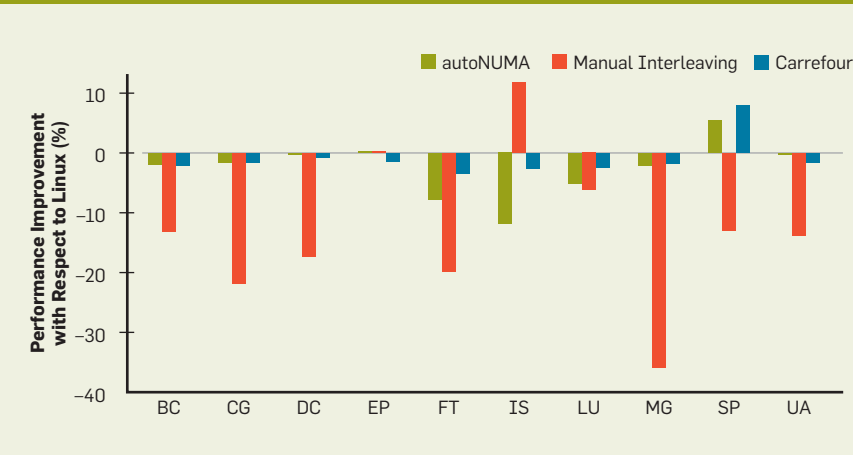
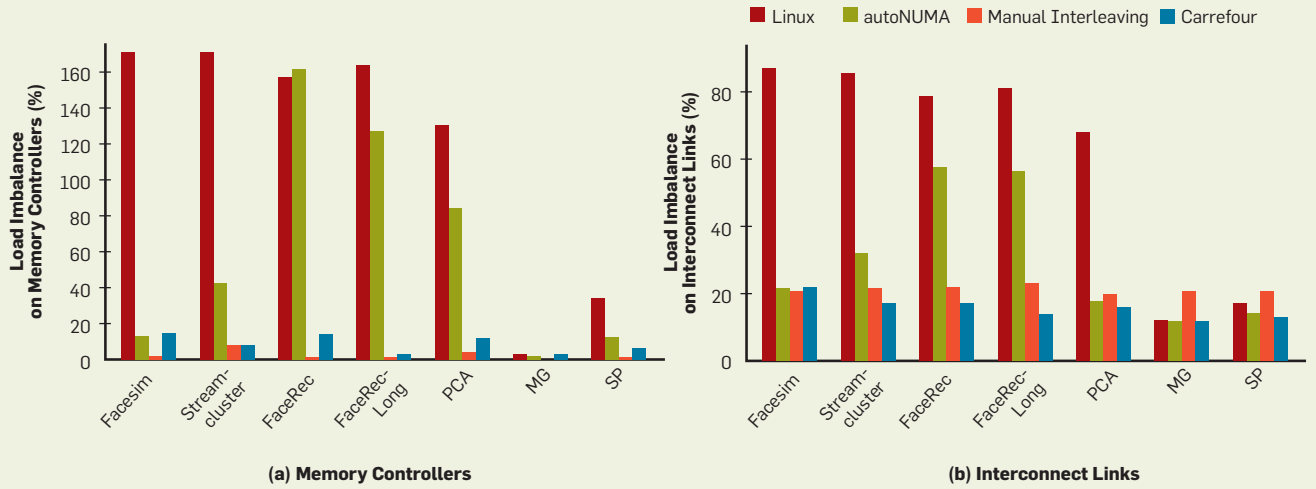


Figure 6. Load imbalance for selected benchmarks.



gave standard deviations of up to 9%.

Performance. We evaluated Carrefour's performance against Linux's default policy (first-touch), manually interleaving memory using Linux's interleave policy, and the AutoNUMA patch. Figures 4 and 5 show the performance improvement relative to default Linux.

There are three general classes of applications. First are those that have the same performance no matter which NUMA technique is used (for example, Bodytrack and Swaptions). These applications are not memory intensive and tend to have a low last-level cache miss rate. They also do not suffer much overhead from Carrefour or AutoNUMA, because most of the overhead is proportional to the memory intensiveness of the application.

The second class of applications is memory intensive, but the default first-touch policy works well for them. BT, CG, DC, FT, MG, and UA fall into this category. For these applications, manual interleaving hurts performance because it eliminates the locality benefit of first-touch without reducing congestion. On the other hand, Carrefour does not cause poor memory placement but only has a small overhead.

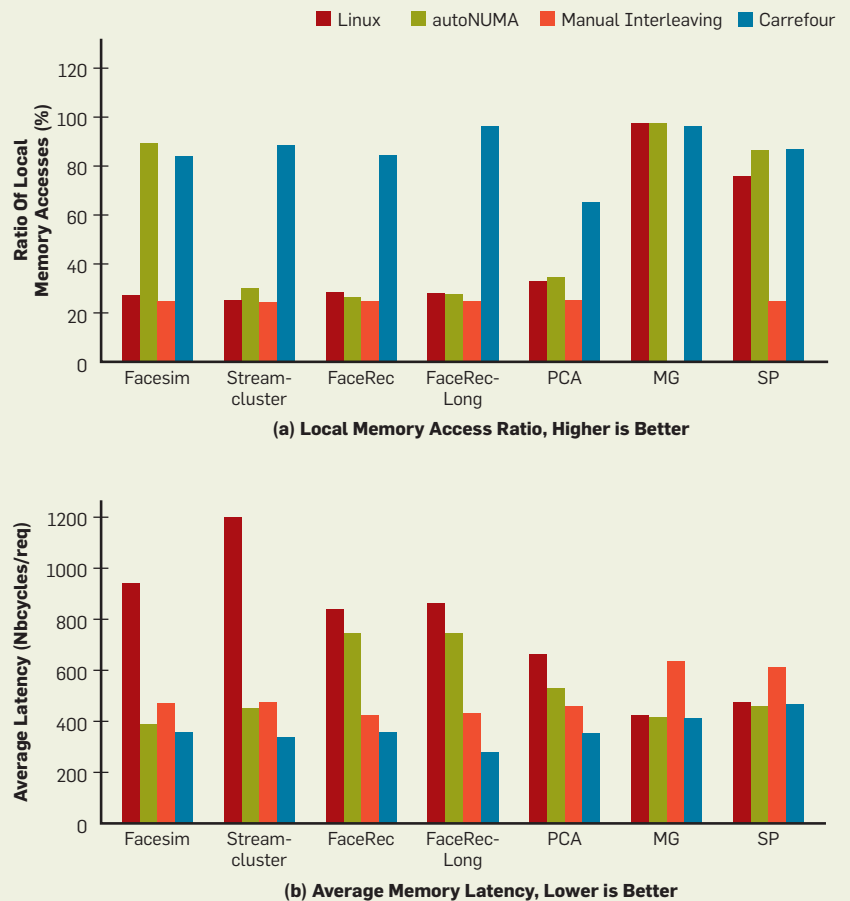
The remaining benchmarks suffer from poor memory placement under default Linux. AutoNUMA is able to improve the performance of some these applications, but for others (for example, FaceRec and PCA) it has only a small impact. Carrefour, on the other hand, significantly improves the performance of these applications, and

in two cases Carrefour greatly outperforms the second-best technique. It improves the performance of FaceRec Long by 120% over default Linux, where manual interleaving improves performance by only 60%. Similarly,

Carrefour improves the performance of Streamcluster by 180%, and manual interleaving improves it by only 100%.

One exception is IS, which is improved by manual interleaving but not by Carrefour. Carrefour's sampling

Figure 7. DRAM latency and locality for selected benchmarks.



and migration rate cannot keep up with the burst of traffic produced by IS, so the memory is not balanced in time to improve performance.

We further profiled select applications in order to see how Carrefour affects the key imbalance, locality, and latency metrics. Figures 6 and 7 present the results. Figure 6 shows the load imbalance for selected benchmarks. Lower is better.

In Figure 6a Carrefour consistently minimizes the imbalance on memory controllers, as does manual interleaving. AutoNUMA is sometimes able to reduce the imbalance but usually not to the same degree, and in the case of FaceRec it makes the imbalance worse than default Linux. The imbalance on interconnect links as depicted in Figure 6b shows similar trends.

Although manual interleaving is able to reduce imbalance, it always does so at the cost of locality. This is made evident in Figure 7a. Carrefour, on the other hand, always has the highest or nearly the highest local memory-access ratio. For applications that have good locality under default Linux (MG and SP), AutoNUMA retains the good locality but is able to improve the locality of Facesim.

The effects of imbalance and the local access ratio are reflected in the memory-access latency, shown in Figure 7b. As expected, Carrefour produces the lowest (or is tied for the lowest) average memory latencies for each profiled application. There is also a strong correlation between the average memory latency and a benchmark's performance. For example, Streamcluster and FaceRec have large reductions in memory latency with Carrefour, and they show large performance improvements in Figure 4.

Overall, we can conclude Carrefour systematically fixes NUMA memory-placement issues in nearly all situations, is able to greatly outperform other techniques in some circumstances, and does not significantly hurt performance for any application.

Conclusion

NUMA architecture is for scaling the processor count of today's server-class systems. In the near future, expect systems to have even more NUMA nodes and more complicated NUMA topologies. The experiments described here show the performance effects of NUMA

are significant and the problem is non-trivial, which motivates careful study and a comprehensive solution.

Contrary to previous NUMA studies, our experiments found congestion causes the most serious NUMA problems. Congestion happens when the rate of requests to memory controllers or the rate of traffic over interconnects is too high, which causes excessive delays for memory accesses. It can be alleviated by balancing the traffic among multiple memory controllers and interconnect links. The other factor of NUMA performance is locality, which is what previous NUMA algorithms have focused on. Good locality means that most of the memory accesses will be to the local node and therefore do not pay the latency cost of traversing interconnect links.

As shown earlier, the two NUMA concerns of congestion and locality are difficult to reconcile, and for any particular application we cannot know the best memory placement beforehand. Carrefour uses hardware performance counters and hardware sampling to determine an application's memory-access patterns online with low overhead. It then uses that knowledge to apply three page-level techniques: memory replication, memory interleaving, and memory collocation. Each technique serves a specific purpose: collocation improves locality, interleaving reduces imbalance, and replication does both in situations when reads vastly outnumber writes. The novelty of Carrefour is in combining these strategies and applying each only when appropriate.

The result is Carrefour is able to improve performance compared with default Linux for many applications. For two benchmarks, Streamcluster and FaceRecLong, performance is more than doubled by using Carrefour. Unlike manual interleaving and AutoNUMA, Carrefour never significantly degrades performance by making improper page placements.

As NUMA systems grow and the number of cores issuing memory requests increases, NUMA effects will continue being a concern. Carrefour demonstrates a collection of techniques that effectively reduce these concerns. Developers can use the methods and insights gained from Carrefour, along with the tools described earlier, to optimize their applications for NUMA systems.

Acknowledgments

We thank Oracle Labs and the British Columbia Innovation Council for funding this work. 

Related articles on queue.acm.org

NUMA (Non-Uniform Memory Access): An Overview

Christoph Lameter

<http://queue.acm.org/detail.cfm?id=2513149>

Scalability Techniques for Practical Synchronization Primitives

Davidlohr Bueso

<http://queue.acm.org/detail.cfm?id=2698990>

Photoshop Scalability: Keeping It Simple

Clem Cole and Russell Williams

<http://queue.acm.org/detail.cfm?id=1858330>

References

1. Bailey, D. NAS Parallel Benchmarks. RNR Technical Report (1994); <http://www.nas.nasa.gov/publications/npb.html>.
2. Boyd-Wickizer, S. et al. Corey: An operating system for many cores. In 8th Usenix Symposium on Operating Systems and Design (2008), 43–57.
3. Brecht, T. On the importance of parallel application placement in NUMA multiprocessors. In *Proceedings of the Usenix Symposium on Experiences with Distributed and Multiprocessor Systems* (1993) 4, 1.
4. Corbet, J. AutoNUMA: The other approach to NUMA scheduling. LWN.net (2012); <http://lwn.net/Articles/488709/>.
5. CSU Face Identification Evaluation System. Evaluation of Face Recognition Algorithms. Colorado State University, 2010; <http://www.cs.colostate.edu/evalfacerec/index10.php>.
6. Dashti, M. et al. Traffic management: A holistic approach to memory placement on NUMA systems. In *Proceedings of the 18th International Conference on Architectural Support for Programming Languages and Operating Systems* (2013), 381–394.
7. David, T., Guerraoui, R. and Trigonakis, V. Everything you always wanted to know about synchronization but were afraid to ask. In *Proceedings of the 24th ACM Symposium on Operating Systems Principles* (2013), 33–48.
8. Lameter, C. An overview of non-uniform memory access. *Commun. ACM* 56, 9 (2013), 59–65.
9. Metis MapReduce Library; <http://pdos.csail.mit.edu/metis/>.
10. PARSEC Benchmark Suite; <http://parsec.cs.princeton.edu/>.

Fabien Gaud is a senior software engineer at Coho Data, focusing on performance and scalability.

Baptiste Lepers is a postdoc at EPFL. His research topics include performance profiling, optimizations for NUMA systems, and multicore programming.

Justin Funston and **Mohammad Dashti** are Ph.D. students at the University of British Columbia. Funston's research interests include memory management, thread scheduling, and multicore systems. Dashti research focuses on operating systems, GPGPU, and heterogeneous CPU/GPU systems.

Alexandra Fedorova is an associate professor in the ECE department at the University of British Columbia. Her research focuses on performance, usability, and energy-efficiency of computer systems.

Vivien Quéma is a professor at Grenoble INP (ENSIMAG), France. His research is about understanding, designing, and building (distributed) systems.

Renaud Lachaize is an assistant professor at the University of Grenoble, France. His research interests are in the area of operating systems and distributed systems, with currently a particular focus on multicore systems.

Mark Roth currently works at Google as an engineer.

Copyright held by authors. Publication rights licensed to ACM. \$15.00

ACM Books. In-depth. Innovative. Insightful.

The VR Book: Human-Centered Design for Virtual Reality

By Jason Jerald, PhD

The VR Book focuses on human-centered design of virtual reality (VR) experiences. Creating compelling VR experiences is an incredibly complex challenge. When VR is done well, these experiences can be brilliant and pleasurable, but when done badly, they can result in frustration and sickness. While there are many causes of bad VR such as the limitations of technology, much is centered on a lack of understanding of human perception, interaction, design principles, and real users.

This book focuses on human elements of VR, such as how users perceive and intuitively interact with various forms of reality, causes of VR sickness, creating useful and pleasing content, and how to design and iterate upon effective VR applications. It is not just for VR designers; it is for the entire team, as they all should understand the basics of perception, VR sickness, interaction, content creation, and iterative design.

Good VR design requires strong communication between human and machine, indicating what interactions are possible, what is currently occurring, and what is about to occur. A human-centered design principle, like lean methods, is to avoid completely defining the problem at the start and to iterate upon repeated approximations and modifications through rapid tests of ideas with real users. Thus, The VR Book is intended as a foundation for anyone and everyone involved in creating VR experiences including: designers, managers, programmers, artists, psychologists, engineers, students, educators, and user experience professionals.

Paperback ISBN: 9781970001129 Hardcover ISBN: 9781970001150
Ebook ISBN: 9781970001136 DOI: 10.1145/2792790



For more info please visit
<http://books.acm.org>

or contact ACM at
ACMbooks-Info@acm.org



Association for Computing Machinery
2 Penn Plaza, Suite 701
New York, NY 10121-0701, USA
Phone: +1-212-626-0658
Email: acmbooks-info@acm.org



Morgan & Claypool Publishers
1210 Fifth Avenue, Suite 250
San Rafael, CA 94901, USA
Phone: +1-415-462-0004
Email: info@morganclaypool.com

Digital maps can be engineered to adapt to a person's unique interests and experience in geographic space.

BY ANDREA BALLATORE AND MICHELA BERTOLOTTO

Personalizing Maps

GEOGRAPHIC MAPS CONSTITUTE a ubiquitous medium through which we understand, construct, and navigate our natural and built surroundings. At the intersection of the explosion of geographic information online, data-mining techniques, and the increasing popularity of Web maps, a novel possibility has emerged: Instead of generating one map for large numbers of users, user profiling and implicit feedback analysis can support creation of a different map for each person. The automated personalization of the map-making process is still in its infancy but has the potential to provide more relevant maps to millions of users worldwide.

While mapmaking has traditionally aimed to produce static maps to be printed and distributed to a target audience, geographic information systems (GISs) provide interactive tools to collect and process information dynamically, transforming not only cartography but also geography, urban planning, and any activity that relies on geographic knowledge. Since the 1960s, using GISs, geographers, urban planners, army generals, and economists have been generating different representations of the same input

data to better understand diverse geospatial phenomena. Over the past decade, GISs have further merged with Web technologies and mobile computing, enabling mass adoption of digital maps while overcoming the limitations of paper maps.

As interactive digital maps replace paper maps, this “ubiquitous cartography” is quietly becoming part of our lives, changing not only the consumption but also the production of geographic information.^{5,7} Just below the surface of this tumultuous reconfiguration, the fundamental problems of cartography have hardly changed. Complex, dynamic, and uncertain geographic data needs to be represented on a screen, selecting what needs to be displayed and how to display it with respect to the user's informational needs.⁶ Appropriate cartographic projections, scales, generalization principles, human-computer interaction, and semiotic conventions constitute essential ingredients for the design of usable digital maps.

Although maps are often perceived as a form of objective, scientific knowledge about the world, the same area can be represented from many alternative perspectives, including and ex-

» key insights

- Cartography traditionally focuses on producing maps for large groups of readers, and digital maps, including Google's, have barely begun to challenge this approach; collecting explicit and implicit feedback from users, digital cartography is able to capture a person's geographic knowledge, experiences, and attitudes, better supporting spatial learning and decision making.
- As a research frontier, automated map personalization requires real-time task detection, geographic user profiling, trajectory analytics, data fusion, geo-visualization, and sentiment analysis, along with insight from cognitive psychology and human geography.
- Depending on design, personal maps could foster exploration of the environment beyond the user's known territory or reinforce segregation, fragmenting the collective knowledge of the spaces we inhabit.



cluding different pieces of uncertain information, and choosing arbitrary graphical and symbolic conventions; for example, Figure 1 includes alternative representations of University College Dublin. For cartographers, it is uncontroversial that radically different maps are needed to perform different tasks. Nautical charts, tourist maps, and urban-planning maps display different geographic information tailored to specific tasks (such as reaching a port safely, understanding the structure of a city, or identifying a suitable location to build a new bridge). Less obvious is the fact that different maps might be needed by different people to perform the same task. Since the 1950s, psychological studies have shown every person perceives and develops an individual mental model of their environment, based on direct and mediated subjective experiences.¹¹

Likewise, since the late 1990s, the economic value of personalization of Web-based services has attracted considerable attention, resulting in now-ubiquitous personalized news stories, commercial offers, film recommendations, and search results. In 1995, Nicholas Negroponte, founder of MIT's Media Lab, imagined a newspaper called *Daily Me* that would automatically collect and arrange stories relevant to the reader, rather than impose the same content on everyone, overcoming the paradigm of mass production that dominated the 20th century.¹⁶ Knowing a customer's behavior and tastes through surveillance techniques has become commonplace in marketing, in a tight feedback loop between companies and their current or prospective consumers, in what has been called by the oxymoron "mass customization."¹⁹

As Web-based digital maps progressively become the main portal through which to view the world and its places, the idea of applying mass personalization to maps comes within reach. It is now conceivable to develop mapping platforms that generate personalized maps not only for a specific task but for a specific individual, taking into account the individual's experience, behavior, knowledge, and particular viewpoint. Surprisingly, while many online products and services have been personalized over the past decade, dig-

ital maps are still fundamentally untouched by mass customization.

Map Personalization So Far

Maps are complex cultural and technical objects that assemble multiple data sources, assumptions about the user, cartographic traditions and practices, and design choices. All elements that form a digital map can in principle be personalized to increase the usability, efficiency, and clarity of the map with respect to a task.²⁰ To personalize maps, useful information can be provided by the user through explicit or implicit feedback.¹⁰

"Explicit feedback" is the conscious selection of preferences in the interface, and any action that is explicitly aimed at expressing a preference about any element of the map, as in, say, changing the language and the default settings of the interface. By contrast, implicit feedback includes any data about the user who expresses a preference indirectly; for example, moving the mouse cursor and clicking on a geographic area expresses a form of interest in that region, while hiding a layer at the beginning of each session indicates lack of relevance. The user's location represents another instance of implicit feedback about what areas of the map are of particular interest. Implicit feedback can in principle be extracted from any data generated by the user, including activity on social media, instant messaging, purchases from online stores, and email messages.

The most important indicator that must be determined is the task the user is currently performing on the map. Common tasks performed on popular online services (such as Yahoo! Maps and Google Maps) include information retrieval, general exploration of a region of interest, and routing. Different maps suit different tasks, with respect to features, layers, and controls. While in some cases the user's intentions are easy to detect (such as typing a place name is likely to indicate an information retrieval and/or a routing task), many behaviors do not imply specific tasks and present a considerable interpretive challenge. Automatic task detection can be performed with many indicators, including user demographics, interaction logs, search history, and the user's context and its spatial, temporal,

social, and computational aspects.

Over the past 15 years, academic researchers, including us, have investigated ideas and techniques through which map personalization can be achieved, working in two complementary strands. On the one hand, the automated adaptation of the map is pursued to increase clarity and efficiency and reduce information overload by removing or highlighting features based on user preferences and current task. On the other, the area of recommendation has generated techniques to personalize search results and recommendations of hotels, restaurants, and other points of interest based on individual and/or collective preferences. These two strands overlap in that similar techniques can be used to adapt the map and personalize search results and recommendations.

In a pioneering work in 2000, Oppermann and Specht¹⁷ developed *Hippie*, a tool that presents museum information based on the context of use. It relies on a user model that represents the user's knowledge and interests, a domain model of the information being displayed, and a space model in which the interaction occurs. Brunato and Battiti⁴ devised the Personal Item Locator and General Recommendation Index Manager (PILGRIM), a recommender system that takes the user's location into account to rank Web pages. Although recommender systems have been widely adopted since the mid-2000s, little work has been done to increase their spatial awareness.²¹

Our work focuses on the use of implicit feedback to adapt the map content itself. The core assumption is that implicit feedback indicators (such as mouse movements and navigational behavior) can be used to infer user interests.¹³ A recurring cognitive issue, particularly in the context of mobile computing, is that of spatial information overload, or display of excessive amounts of information on the map, hindering, rather than helping, the user. CoMPASS is a GIS application that monitors user interaction to recommend groups of features (such as layers) to users.^{12,22} The MAPPER system generates maps containing specific features, taking into account the user's preferences and the computational context by implicitly monitoring the interactions of users when brows-

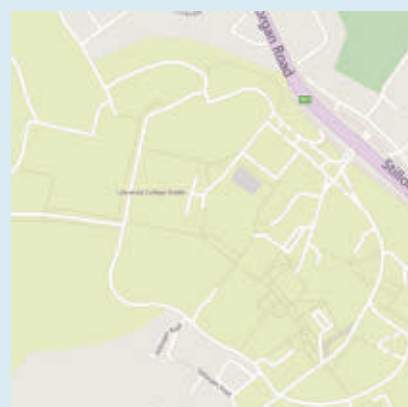
Figure 1. Alternative cartographic representations of University College Dublin. Maps a–c are from commercial services, and d–f are based on OpenStreetMap open data using different themes; visualization generated with GeoFabrik tools.



(a) Bing Satellite



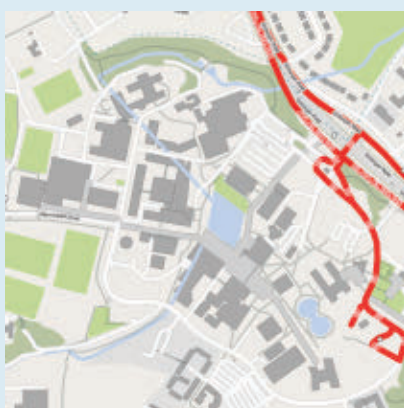
(b) Google Maps



(c) Bing Maps



(d) OpenStreetMap



(e) OpenStreetMap



(f) OpenStreetMap

ing maps and inferring individual and group preferences.²³ This approach has been evaluated on a variety of map-based tasks,²⁴ showing increased efficiency in task completion; a similar approach was applied to detect the current task and adapt the map to it.¹⁴

In the RecoMap prototype,² we explored the possibility of computing interest scores for geographic features based on two complementary aspects—interaction (amount of interaction with a feature) and proximity (physical proximity to the feature) to generate personalized recommendations. A memory model simulates the decay of interest over time, assuming if the user does not interact with an object, the user's interest in it is declining. Moreover, we investigated the possibility of integrating crowdsourced spatial data into the personalization analysis and utilizing Linked Open Data, a net-

work of inter-connected datasets, to increase the semantic structure of the geographic features.¹ Although this body of research initiated the theoretical and practical development of map personalization, new concepts, paradigms, and techniques await further investigation and evaluation.

This line of research promises to generate a plethora of commercial applications, greatly enriching current Web-mapping platforms. Since the mid-2000s, following increased bandwidth and more sophisticated Web browsers, a growing non-specialist mass market for Web maps has emerged, first on desktop computers, and more recently on GPS-enabled smartphones.^a In order to

review the state of the art of map personalization in products on the consumer market, it is useful to distinguish between manual and automated personalization, reflecting implicit and explicit feedback. Manual personalization allows users to modify aspects of the map, using preferences, bookmarking, and map editors. By contrast, automated personalization relies on implicit feedback to modify the map without user intervention, using data mining to model the user's tastes and intentions; the table here lists personalization capabilities of popular, global, currently active mapping services.

Google Maps is the only service today that provides some automated personalization, tailoring the search results and ads based on the user's search history and ratings, claiming to generate “a map for every person

^a http://www.comscore.com/Insights/Blog/Map_Searches_Shift_from_Desktops_to_Smartphones

Personalization in Web maps, October 2015.

Product	Manual personalization	Automated personalization
Google Maps	Vector base map. In "My Places," users can set their home and work address. MapsEngine cloud tool can be used to create new maps by adding layers on a set of base maps.	Search results, recommendations, and advertisements are based on previous ratings and searches.
ArcGIS Online (ESRI)	Complex new maps can be created, combining base maps with user-defined layers. Advanced user-defined analytics available.	None.
Apple Maps	Vector base map. Users can bookmark locations in iCloud.	None.
OpenStreetMap (OSM Foundation)	OSM open data can be used to generate customized maps with a variety of dedicated tools (such as MapBox).	None.
HERE (Nokia)	Users can create "collections" of locations.	None.
MapQuest (AOL)	Users can save favorite locations and vehicles to improve routing. With "My Maps," they can save collections of locations and routes.	None.
Yahoo! Maps	Based on Nokia's HERE.	None.
ViaMichelin	In "My Michelin," users can bookmark locations, restaurants, and itineraries.	None.
OS Map Finder (U.K. Ordnance Survey)	Users can draw paths.	None.
Bing Maps (Microsoft)	Users can bookmark locations in "My Places."	None.

and place."^b Other popular Web mapping products (such as ArcGIS Online and Yahoo! Maps) offer some manual personalization, typically in the form of bookmarks or editors to create and share new maps with user-provided data and visual styles. None of these products attempts to perform automated personalization.

Computational Challenges

The considerable increase in variety

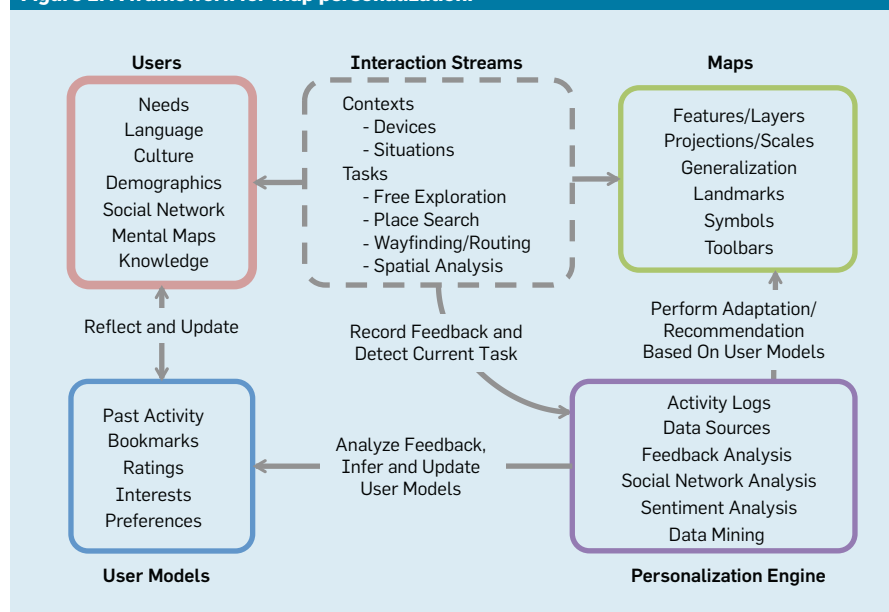
^b <http://google-latlong.blogspot.it/2013/05/meet-new-google-maps-map-for-every.html>

and volume of data provides a largely unexplored yet fertile basis from which to rethink maps. As we have shown, the field of map personalization is still at an early stage, with little research conducted or applied to commercial products. To further map personalization, several components must be integrated into a coherent conceptual framework (see Figure 2). A personalization engine must be able to perform multivariate feedback analysis on the many channels through which users express their spatial interests and preferenc-

es. By monitoring how users perform their tasks, the engine should be able to mine and extract meaningful patterns, inferring effective user models. Based on these models, the current task can be detected and trigger personalization in its two dimensions—adaptation and recommendation—at appropriate moments, unobtrusively supporting the user.

Many possibilities lie ahead to deliver more relevant, effective, and useful maps to Internet users. Given the same geographic data, such a system will tailor different maps for, say, a Japanese tourist in San Francisco and for an Italian ex-pat who lives and works in San Francisco. The tourist explicitly indicated an interest in architecture and a dislike for fast food. Her interaction shows implicitly an interest in historical areas, reflected in the map by increasing the prominence of history-themed museums. The map also captures that she was in the city before, displaying previously visited points of interest, facilitating spatial comprehension and way-finding, while at the same time emphasizing unknown areas of the city that feature notable buildings. Fast food restaurants remain hidden, unless she searches for them explicitly or when they might provide useful navigation landmarks. Before meal times, the map emphasizes restaurants that were recommended by her friends

Figure 2. A framework for map personalization.



who visited the city, while at night it increases the visibility of movie theaters, taking into account her interest in cinema.

By contrast, the Italian ex-pat specified an interest in music. As he navigates the city, the map tends to hide familiar tourist attractions while highlighting content neighboring his home-to-work commutes. Based on his interaction history, the system also captures an interest in the Mission District, which becomes more prominent and detailed on the map. As he uses a car to drive around the city, the public transport infrastructure fades into the background of the map. However, if his movement patterns match an efficient bus route, the system discretely suggests an alternative transportation option. Occasionally, the map emphasizes an unfamiliar neighborhood that presents a high density of music venues, inviting him to go beyond the borders of his daily routine. To make this vision real, several computational challenges lie ahead.

Real-time task detection and prediction. As maps are used in a variety of situations for different tasks, the system must be able to detect and predict them effectively. For this purpose, specific machine-learning techniques must be developed and optimized. In the collection of implicit feedback, relevant features include the user's spatial and temporal context, as well as search and interaction streams.

Spatial user modeling. Because of maps' spatial nature, personalization requires deep understanding of user behavior in space and time. Hence, the aggregation and interpretation of large numbers of noisy spatiotemporal trajectories containing GPS fixes, clicks, and search logs are essential for developing models of user behavior able to capture and predict recurring patterns and anomalies (such as sightseeing, as opposed to daily commuting). Recording, storing, and mining a large volume of spatiotemporal trajectories constitute an open research challenge.²⁵ Trajectories can traverse the geographic space, as well as other spaces, including mouse trajectories in a user interface. Spatial social network analysis can also illuminate the deep structures that influence interaction with maps.

Geographically weighted personalization. As maps assist users in the exploration and navigation of the geographic space, personalization should be able to tap the spatial variation in human activities. Development of geographically weighted techniques relies on the assumption that individual information needs and content relevance change both spatially and temporally. Such spatialization of users and content and their relationships increases the computational complexity of traditional personalization models. In turn, spatiotemporal user models would enable finer and more sensitive personalization of contents.

Geosemantic interoperability and data fusion. Map personalization requires aggregation and fusion of a range of heterogeneous sources of geographic information characterized by intrinsic uncertainty, vagueness, and rapid obsolescence, ranging from traditional government agencies to crowdsourced, volunteered datasets. In this sense, research in the context of the semantic Web and linked open data provides suitable representational tools to organize, store, explore, and retrieve spatiotemporal objects and data streams.⁹ Reliable mechanisms are needed to reference entities and perform identity resolution, reducing the friction caused by interoperability issues.³

Geoparsing and sentiment analysis. As a vast amount of geographic knowledge is expressed in natural language, natural language processing (NLP) is crucial for map personalization, extracting value from unstructured data. Geoparsing, or extraction of geographic information from natural language, is an open problem in NLP, intimately connected to word-sense disambiguation. Detection of affect, sentiment, and emotion in text is an emerging yet important aspect of interpreting user behavior, improving extraction and modeling of users' opinions about places.

Cognitive map design. Concepts and principles from cognitive map design¹⁵ can be applied to generate and validate alternative cartographic representations, providing an exciting opportunity to test cognitive theories against real scenarios on large numbers of users. From a technologi-

cal viewpoint, interactive maps have moved from a tile-based approach, in which maps are served through cached pre-rendered images, to a more flexible vector-based approach, in which the rendering occurs in real time in the client, providing the ideal platform for experimenting with alternative rendering choices and styles. Knowledge from spatial cognition could also be useful for producing better personalized maps and gain further insight into the human perception and understanding of the geographic environment.⁸

To achieve significant advances in these areas, academic researchers and commercial developers must tap the informational wealth produced by millions of users worldwide in their daily interactions mediated by online platforms, in which space and place are deeply intertwined with social, cultural, and economic processes. Due to the intrinsic complexity of these processes, map personalization needs help from thriving research areas in the context of big-data analytics. From a complementary perspective, NLP and sentiment analysis can be used to mine user-generated opinions about places to increase or decrease the emphasis of specific features. Advanced data-mining techniques are needed to extract meaning from noisy interaction logs. Beyond these computational steps, the challenges of map personalization are intrinsically multidisciplinary, harnessing ideas and tools from geographic information science, cartography, cognitive psychology, human-computer interaction, and software design.

Consequences

From a societal viewpoint, automated map personalization at a mass scale could have serious implications that should be responsibly taken into account, particularly by commercial developers whose products reach millions of users. Beyond the obvious concern for privacy, fostered by any surveillance-based technology, specific problems include the potential loss of a common representation of geographic realities. Personalized maps might result in what Internet activist Eli Pariser calls a "filter bubble," increasing social and cultural seg-

regation between groups of users.¹⁸ Likewise, personalized landmarks can be useful for increasing the clarity of maps but might also reduce the common semantic ground shared by the inhabitants of a geographical area. In this regard, Google Maps, currently the only commercial product to include some form of map personalization, presents several unresolved questions. The most conspicuous is the product's lack of transparency, making it difficult for users to understand why certain features are recommended over others. The user models generated by Google are black boxes not accessible to the users they are supposed to represent, and, more important, there is no visible "off" button to disable the personalization; even when logged out, the search results are still personalized in unclear ways based on cookies and the IP location of the user's machine.

A serious challenge for academic research in map personalization is the lack of realistic interaction datasets for evaluating novel systems and approaches. As private corporations are understandably reticent to share their map-interaction logs, the studies discussed here have limited evaluations, failing to reflect the complexity, noise, and variety of situations in real mapping applications, limiting their observation to small, artificial, and controlled contexts. A few large corporations (such as Google and Microsoft) attract the vast majority of online map users and are thus in a privileged position to unobtrusively devise and evaluate proprietary techniques on large groups of users on a variety of tasks, interpreting their behavior as implicit feedback. For this reason, academic research must either focus on well-defined cognitive, computational, and cartographic aspects of map personalization that can be convincingly evaluated or work in close partnership with map providers. No progress in map personalization can be assessed in the absence of rigorous measures to quantify the effectiveness of techniques, algorithms, and models.

Conclusion

Development of personalized maps has important applications in many domains. To date, the focus has been

on commercial applications; for example, Google has been exploring location-based advertising, trying to maximize the relevance and profitability of ads that have a strong spatial component. Similarly, most research focuses on efficiency, reducing information overload, increasing clarity, and helping users complete tasks more quickly or with lower cognitive load, as in, say, decision making, information retrieval, and routing. However, exciting possibilities also exist beyond increased efficiency. Personalized maps need not reinforce users' biases and limited perspectives but can be designed to operate in the opposite way, attracting attention to the unknown and unfamiliar and promoting diversity, serendipity, and discovery. In education, self-adaptive maps could support students, tailoring maps to different learning styles and backgrounds. It is reasonable to expect map personalization could trigger a quiet but deep reconfiguration of familiar maps, leading to unexpected changes in the way we perceive and imagine the world around us.

Acknowledgments

The research presented here was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. 

References

- Ballatore, A. and Bertolotto, M. 2011. Semantically enriching VGI in support of implicit feedback analysis. *Web and Wireless Geographical Information Systems*, K. Tanaka, P. Fröhlich, and K.-S. Kim, Eds. LNCS, Vol. 6574 (Kyoto, Japan, Mar. 3–4). Springer, Berlin, Germany, 78–93.
- Ballatore, A., McArdle, G., Kelly, C., and Bertolotto, M. RecoMap: An interactive and adaptive map-based recommender. In *Proceedings of the 25th ACM Symposium on Applied Computing* (Sierre, Switzerland, Mar. 22–26). ACM Press, New York, 2010, 887–891.
- Ballatore, A., Wilson, D., and Bertolotto, M. A survey of volunteered open geo-knowledge bases in the semantic Web. In *Quality Issues in the Management of Web Information*, G. Pasi, G. Bordogna, and K. Jain, Eds. Intelligent Systems Reference Library, vol. 50. Springer, Berlin, Germany, 2013, 93–120.
- Brunato, M. and Battiti, R. PILGRIM: A location broker and mobility-aware recommendation system. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications* (Fort Worth, TX, Mar. 23–26). IEEE, Piscataway, NJ, 2003, 265–272.
- Gartner, G., Bennett, D., and Morita, T. Towards ubiquitous cartography. *Cartography and Geographic Information Science* 34, 4 (2007), 247–257.
- Haklay, M., Ed. *Interacting with Geospatial Technologies*. John Wiley & Sons, Chichester, U.K., 2010.
- Haklay, M., Singleton, A., and Parker, C. Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass* 2, 6 (Nov. 2008), 2011–2039.
- Hirtle, S.C. Geographical design: Spatial cognition and geographical information science. *Synthesis Lectures*

- on *Human-Centered Informatics* 4, 1 (2011), 1–67.
- Janowicz, K., Scheider, S., Pehle, T., and Hart, G. Geospatial semantics and linked spatiotemporal data: Past, present, and future. *Semantic Web* 3 (2012), 321–332.
- Kelly, D. and Teevan, J. Implicit feedback for inferring user preference: A bibliography. *ACM SIGIR Forum* 37, 2 (Fall 2003), 18–28.
- Lynch, K. *The Image of the City*. MIT Press, Cambridge, MA, 1960.
- Mac Aoidh, E. and Bertolotto, M. Improving spatial data usability by capturing user interactions. In *The European Information Society: Leading the Way with Geo-information*, S.I. Fabrikant and M. Wachowicz, Eds. Springer, Berlin, Germany, 2007, 389–403.
- Mac Aoidh, E., Bertolotto, M., and Wilson, D.C. Analysis of implicit interest indicators for spatial data. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems* (Seattle, WA, Nov. 7–9). ACM Press, New York, 2007, 1–4.
- Mac Aoidh, E., McArdle, G., Petit, M., Ray, C., Bertolotto, M., Claramunt, C., and Wilson, D. Personalization in adaptive and interactive GIS. *Annals of GIS* 15, 1 (2009), 23–33.
- Montello, D. Cognitive map-design research in the twentieth century: Theoretical and empirical approaches. *Cartography and Geographic Information Science* 29, 3 (2002), 283–304.
- Negroponte, N. *Being Digital*. Random House, New York, 1995.
- Oppermann, R. and Specht, M. A context-sensitive nomadic exhibition guide. In *Handheld and Ubiquitous Computing*, P. Thomas and H.-W. Gellersen, Eds. LNCS, Vol. 1927. Springer, Berlin, Germany, 2000, 127–142.
- Pariser, E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin, London, U.K., 2012.
- Piller, F.T. and Tseng, M.M., Eds. *Handbook of Research in Mass Customization and Personalization, Volume 1*. World Scientific, Singapore, 2010.
- Reichenbacher, T. Adaptive concepts for a mobile cartography. *Journal of Geographical Sciences* 11, 1 (2001), 43–53.
- Stiller, C., Ros, F., and Ament, C. Towards spatial awareness in recommender systems. In *Proceedings of the International Conference for Internet Technology and Secured Transactions* (London, U.K., Nov. 9–12). IEEE, Piscataway, NJ, 2009, 1–7.
- Weakliam, J., Bertolotto, M., and Wilson, D. Implicit interaction profiling for recommending spatial content. In *Proceedings of the 13th Annual ACM International Workshop on Geographic Information Systems* (Bremen, Germany, Nov. 4–5). ACM Press, New York, 2005, 285–294.
- Weakliam, J., Wilson, D., and Bertolotto, M. Personalising map feature content for mobile map users. In *Map-based Mobile Services: Theories, Methods and Implementations*, L. Meng, A. Zipf, and T. Reichenbacher, Eds. Springer, Berlin, Germany, 2008, 125–145.
- Wilson, D., Bertolotto, M., and Weakliam, J. Personalizing map content to improve task completion efficiency. *International Journal of Geographical Information Science* 24, 5 (2010), 741–760.
- Zheng, Y. and Zhou, X., Eds. *Computing with spatial trajectories*. *Database Management & Information Retrieval*. Springer, Berlin, 2011.

Andrea Ballatore (aballatore@spatial.ucsb.edu) is a postdoctoral researcher and research coordinator in the Center for Spatial Studies at the University of California, Santa Barbara.

Michela Bertolotto (michela.bertolotto@ucd.ie) is a senior lecturer in the School of Computer Science and Informatics at the University College Dublin, Ireland.

© 2015 ACM 00010782/15/12 \$15.00



Watch the authors discuss their work in this exclusive *Communications* video. <http://cacm.acm.org/videos/personalizing-maps>

Connecting mathematical logic and computation, it ensures that some aspects of programming are absolute.

BY PHILIP WADLER

Propositions as Types

POWERFUL INSIGHTS ARISE from linking two fields of study previously thought separate. Examples include Descartes's coordinates, which links geometry to algebra, Planck's Quantum Theory, which links particles to waves, and Shannon's Information Theory,

which links thermodynamics to communication. Such a synthesis is offered by the principle of Propositions as Types, which links logic to computation. At first sight it appears to be a simple coincidence—almost a pun—but it turns out to be remarkably robust, inspiring the design of automated proof assistants and programming languages, and continuing to influence the forefronts of computing.

» key insights

- **Propositions as Types observes a deep correspondence between logic and computation: propositions in a logic correspond to types in a programming language; proofs of propositions correspond to programs of the corresponding type; and simplification of proofs corresponds to evaluation of programs.**
- **Propositions as Types is broadly applicable, applying to a wide variety of logics (intuitionistic, second-order, classical, linear) and of language features (lambda calculus parametric polymorphism, continuations, concurrency).**
- **Often the same ideas are discovered independently by logicians and computer scientists, demonstrating some aspects of programming language design are not arbitrary but absolute.**

Propositions as Types is a notion with many names and many origins. It is closely related to the BHK Interpretation, a view of logic developed by the intuitionists Brouwer, Heyting, and Kolmogorov in the 1930s. It is often referred to as the Curry–Howard Isomorphism, referring to a correspondence observed by Curry in 1934 and refined by Howard in 1969. Others draw attention to significant contributions from de Bruijn's Automath and Martin-Löf's Type Theory in the 1970s.

Propositions as Types is a notion with depth. It describes a correspondence between a given logic and a given programming language. At the surface, it says that for each proposition in the logic there is a corresponding type in the programming language—and vice versa. Thus we have

propositions as types.

It goes deeper, in that for each proof of a given proposition, there is a program of the corresponding type—and vice versa. Thus we also have

proofs as programs.

And it goes deeper still, in that for each way to simplify a proof there is a corresponding way to evaluate a program—and vice versa. Thus we further have

*simplification of proofs as
evaluation of programs.*

Hence, we have not merely a shallow bijection between propositions and types but a true isomorphism, preserving the deep structure of proofs and programs, simplifications, and evaluation.

Propositions as Types is a notion with breadth. It applies to a range of logics, including propositional, predicate, second-order, intuitionistic, classical, modal, and linear. It underpins the foundations of functional programming, explaining features including functions, records, variants, parametric polymorphism, data abstraction, continuations, monads, linear types, and session types. It has inspired automated proof assistants and programming languages, including Agda, Automath, Coq, Epigram, F#, F*, Haskell, LF, ML, NuPRL, Scala, Singularity, and Trellis.

Propositions as Types is a notion with mystery. Why should it be the case that intuitionistic natural deduction, as developed by Gentzen in the 1930s, and simply typed lambda calculus, as developed by Church around the same time for an unrelated purpose, should be discovered 30 years later to be essentially identical? And why should it be the case that the same correspondence arises again and again? The logician Hindley and the computer scientist Milner independently developed the same type system, now dubbed Hindley–Milner. The logician Girard and the computer scientist Reynolds independently developed the same calculus, now dubbed Girard–Reynolds. Curry–Howard is a double-barreled name that ensures the existence of other double-barreled names. Those of us who design and use programming languages may often feel they are arbitrary, but Propositions as Types assures us some aspects of programming are absolute. (See the online appendix, which contains a full version of this article, along with additional details and references, plus a historic note provided by William Howard.)

Church and the Theory of Computation

The origins of logic lie with Aristotle and the stoics in classical Greece, Ockham and the scholastics in the middle ages, and Leibniz’s vision of a calculus ratiocinator at the dawn of the enlightenment. Our interest in the subject lies with formal logic, which emerged from the contributions of Boole, De Morgan, Frege, Peirce, Peano, and others in the 19th century.

As the 20th century dawned, Whitehead and Russell’s *Principia Mathematica* demonstrated formal logic could express a large part of mathematics. Inspired by this vision, Hilbert and his colleagues at Göttingen became the leading proponents of formal logic, aiming to put it on a firm foundation.

One goal of Hilbert’s Program was to solve the Entscheidungsproblem (decision problem), that is, to develop an “effectively calculable” procedure to determine the truth or falsity of any statement. The problem presupposes completeness—that for any statement, either it or its negation possesses a proof. In his address to the 1930 Mathematical Congress in Königsberg, Hilbert affirmed his belief in this principle, concluding “Wir müssen wissen, wir werden wissen” (“We must know, we will know”), words later engraved on his tombstone. Perhaps a tombstone is an appropriate place for these words, given that any basis for Hilbert’s optimism had been undermined the day before, when at the selfsame conference Gödel¹⁸ announced his proof that arithmetic is incomplete.

While the goal was to satisfy Hilbert’s program, no precise definition of “effectively calculable” was required. It would be clear whether a given procedure was effective or not, like Justice Stewart’s characterization of obscenity, “I know it when I see it.” But to show the Entscheidungsproblem undecidable required a formal definition of “effectively calculable.”

One can find allusions to the concept of algorithm in the work of Euclid and, eponymously, al-Khwarizmi, but the concept was formalized only in the 20th century, and then simultaneously received three independent definitions by logicians. Like buses, you wait 2,000 years for a definition of

“effectively calculable,” and then three come along at once. The three were lambda calculus, published in 1936 by Church,⁷ recursive functions, proposed by Gödel at lectures in Princeton in 1934 and published in 1936 by Kleene,²⁴ and Turing machines, published in 1937 by Turing.³⁶

Lambda calculus was introduced by Church at Princeton, and further developed by his students Rosser and Kleene. At that time, Princeton rivaled Göttingen as a center for the study of logic. The Institute for Advanced Study was co-located with the Mathematics Department in Fine Hall. In 1933, Einstein and von Neumann joined the Institute, and Gödel arrived for a visit.

Logicians have long been concerned with the idea of function. Lambda calculus provides a concise notation for functions, including “first-class” functions that may appear as arguments or results of other functions. It is remarkably compact, containing only three constructs: variables, function abstraction, and function application. Church⁶ at first introduced lambda calculus as a way to define notations for logical formulas (almost like a macro language) in a new presentation of logic. All forms of bound variable could be subsumed to lambda binding; for instance, instead of $\exists x. A[x]$, Church wrote $\Sigma(\lambda x. A[x])$. However, it was later discovered by Kleene and Rosser that Church’s system was inconsistent. By this time, Church and his students had realized the system was of independent interest. Church had foreseen this possibility in his first paper on the subject, where he wrote, “There may, indeed, be other applications of the system than its use as a logic.”

Church discovered a way of encoding numbers as terms of lambda calculus. The number n is represented by a function that accepts a function f and a value x , and applies the function to the value n times; for instance, the number three is $\lambda f. \lambda x. f(f(f(x)))$. With this representation, it is easy to encode lambda terms that can add or multiply, but it was not clear how to encode the predecessor function, which finds the number one less than a given number. One day in the dentist’s office, Kleene suddenly saw how

to define predecessor.²³ Once this hurdle was overcome, Church and his students soon became convinced any “effectively calculable” function of numbers could be represented by a term in the lambda calculus.

Church proposed λ -definability as the definition of “effectively calculable,” what we now know as Church’s Thesis, and demonstrated there was a problem whose solution was not λ -definable, that of determining whether a given λ -term has a normal form, what we now know as the Halting Problem. A year later, he demonstrated there was no λ -definable solution to the Entscheidungsproblem.

In 1933, Gödel arrived for a visit at Princeton. He was unconvinced by Church’s contention that every effectively calculable function was λ -definable. Church responded by offering that if Gödel would propose a different definition, then Church would “undertake to prove it was included in λ -definability.” In a series of lectures at Princeton in 1934, based on a suggestion of Herbrand, Gödel proposed what came to be known as “general recursive functions” as his candidate for effective calculability. Kleene took notes and published the definition.²⁴ Church and his students soon determined that the two definitions are equivalent; every general recursive function is λ -definable, and vice versa. Rather than mollifying Gödel, this result caused him to doubt his own definition was correct! Things stood at an impasse.

Meanwhile, at Cambridge, Turing, a student of Max Newman, independently formulated his own notion of “effectively calculable” in the form of what we now call a Turing machine, and used it to show the Entscheidungsproblem undecidable. Before the paper was published, Newman was dismayed to discover Turing had been scooped by Church. However, Turing’s approach was sufficiently different from Church’s to merit independent publication. Turing hastily added an appendix sketching the equivalence of λ -definability to his machines, and his paper³⁶ appeared in print a year after Church’s, when Turing was 23. Newman arranged for Turing to travel to Princeton, where



Whereas Church merely presented the definition of λ -definability and baldly claimed it corresponded to effective calculability, Turing undertook an analysis of the capabilities of a “computer.”



he completed a doctorate under Church’s supervision.

Turing’s most significant difference from Church was not in logic or mathematics but in philosophy. Whereas Church merely presented the definition of λ -definability and baldly claimed it corresponded to effective calculability, Turing undertook an analysis of the capabilities of a “computer” (at this time, the term referred to a human performing a computation assisted by paper and pencil). Turing argued that the number of symbols must be finite (for if infinite, some symbols would be arbitrarily close to each other and undistinguishable), that the number of states of mind must be finite (for the same reason), and that the number of symbols under consideration at one moment must be bounded (“We cannot tell at a glance whether 9999999999999999 and 9999999999999999 are the same”). Later, Gandy¹⁴ would point out that Turing’s argument amounts to a theorem asserting any computation a human with paper and pencil can perform can also be performed by a Turing machine. It was Turing’s argument that finally convinced Gödel; since λ -definability, recursive functions, and Turing machines had been proved equivalent, he now accepted that all three defined “effectively calculable.”

As mentioned, Church’s first use of lambda calculus was to encode formulas of logic, but this encoding had to be abandoned because it led to inconsistency. The failure arose for a reason related to Russell’s paradox, namely that the system allowed a predicate to act on itself, and so Church adapted a solution similar to Russell’s, that of classifying terms according to types. Church’s simply typed lambda calculus ruled out self-application, permitting lambda calculus to support a consistent logical formulation.⁸

Whereas self-application in Russell’s logic leads to paradox, self-application in Church’s untyped lambda calculus leads to non-terminating computations. Conversely, Church’s simply typed lambda calculus guarantees every term has a normal form, or corresponds to a computation that halts.

Untyped lambda calculus or typed lambda calculus with a construct for

general recursion (sometimes called a fixpoint operator) permits the definition of any effectively computable function but has a Halting Problem that is unsolvable. Typed lambda calculus without a construct for general recursion has a Halting Problem that is trivial—every program halts!—but cannot define some effectively computable functions. Both kinds of calculus have their uses, depending on the intended application.

Gentzen and the Theory of Proof

A second goal of Hilbert's program was to establish the consistency of various logics. If a logic is inconsistent, it can derive any formula, rendering it useless.

In 1935, at the age of 25, Gentzen¹⁵ introduced not one but two new formulations of logic—natural deduction and sequent calculus—that became established as the two major systems for formulating a logic and remain so to this day. He showed how to normalize proofs to ensure they were not “roundabout,” yielding a new proof of the consistency of Hilbert's system. And, to top it off, to match the use of the symbol $\exists$ for the existential quantification introduced by Peano, Gentzen introduced the symbol $\forall$ to denote universal quantification. He wrote implication as $A \supset B$ (if A holds, then B holds), conjunction as $A \& B$ (both A and B hold), and disjunction as $A \vee B$ (at least one of A or B holds).

Gentzen's insight was that proof rules should come in pairs, a feature not present in earlier systems (such as Hilbert's). In natural deduction, these are introduction and elimination pairs. An introduction rule specifies under what circumstances one may assert a formula with a logical connective (for instance, to prove $A \supset B$, one may assume A and then must prove B), while the corresponding elimination rule shows how to use that logical connective (for instance, from a proof of $A \supset B$ and a proof of A , one may deduce B , a property dubbed *modus ponens* in the middle ages). As Gentzen noted, “The introductions represent, as it were, the “definitions” of the symbols concerned, and the eliminations are no more, in the final analysis, than the consequences of these definitions.”

A consequence of this insight was

that any proof could be normalized to one that is not “roundabout,” where “no concepts enter into the proof other than those contained in the final result.” For example, in a normalized proof of the formula $A \& B$, the only formulas that may appear are itself and its subformulas, A and B , and the subformulas of A and B themselves. No other formula (such as $(B \& A) \supset (A \& B)$ or $A \vee B$) may appear; this is called the Subformula Principle. An immediate consequence was consistency. It is a contradiction to prove false, written $\perp$. The only way to derive a contradiction is to prove, say, both $A \supset \perp$ and A for some formula A . But given such a proof, one could normalize it to one containing only subformulas of its conclusion, $\perp$. But $\perp$ has no subformulas! It is like the old saw, “What part of no don't you understand?” Logicians became interested in normalization of proofs because of its role in establishing consistency.

Gentzen preferred the system of Natural Deduction because it was, in his view, more natural. He introduced Sequent Calculus mainly as a technical device for proving the Subformula Principle, though it has independent interest. It is an irony that Gentzen was required to introduce Sequent Calculus in order to prove the Subformula Principle for Natural Deduction. He needed a roundabout proof to show the absence of roundabout proofs! Later, in 1965, Prawitz showed how to prove the Subformula Principle directly, by introducing a way to simplify Natural Deduction proofs; and this set the ground for Howard's work described in the next section.

Propositions as Types

In 1934, Curry observed a curious fact, relating a theory of functions to a theory of implication.¹¹ Every type of a function ($A \rightarrow B$) could be read as a proposition ($A \supset B$), and under this reading the type of any given function would always correspond to a provable proposition. Conversely, for every provable proposition there was a function with the corresponding type.

In 1969, Howard circulated a xeroxed manuscript;²² it was not published until 1980, where it appeared in a *Festschrift* dedicate to Curry. Motivated by Curry's

observation, Howard pointed out there is a similar correspondence between natural deduction, on the one hand, and simply typed lambda calculus, on the other, and he made explicit the third and deepest level of the correspondence, as described in the introduction, that simplification of proofs corresponds to evaluation of programs. Howard showed the correspondence extends to the other logical connectives—conjunction and disjunction—by extending his lambda calculus with constructs that represent pairs and disjoint sums. Just as proof rules come in introduction and elimination pairs, so do typing rules; introduction rules correspond to ways to define or construct a value of the given type, and elimination rules correspond to ways to use or deconstruct values of the given type.

We can describe Howard's observation as follows:

- *Conjunction.* Conjunction $A \& B$ corresponds to Cartesian product $A \times B$, or a record with two fields, also known as a pair. A proof of the proposition $A \& B$ consists of a proof of A and a proof of B . Similarly, a value of type $A \times B$ consists of a value of type A and a value of type B .
- *Disjunction.* Disjunction $A \vee B$ corresponds to a disjoint sum $A + B$, or a variant with two alternatives. A proof of the proposition $A \vee B$ consists of either a proof of A or a proof of B , including an indication of which of the two has been proved. Similarly, a value of type $A + B$ consists of either a value of type A or a value of type B , including an indication of whether this is a left or right summand.
- *Implication.* Implication $A \supset B$ corresponds to function space $A \rightarrow B$. A proof of the proposition $A \supset B$ consists of a procedure that given a proof of A yields a proof of B . Similarly, a value of type $A \rightarrow B$ consists of a function that when applied to a value of type A returns a value of type B .

This reading of proofs goes back to the intuitionists and is often called the BHK interpretation, named for Brouwer, Heyting, and Kolmogorov.

Brouwer founded intuitionism, and Heyting and Kolmogorov formalized intuitionistic logic and developed the interpretation in the 1920s and 1930s. Realizability, introduced by Kleene in the 1940s, is based on a similar interpretation.

Given the intuitionistic reading of proofs, it hardly seems surprising that intuitionistic natural deduction and lambda calculus should correspond so closely. But it was not until Howard that the correspondence was laid out clearly, in a way that allowed working logicians and computer scientists to put it to use.

Howard's paper²² divides into two halves. The first half explains a correspondence between two well-understood concepts, the propositional connectives $\&$, $\vee$, $\supset$ on the one hand and the computational types $\times$, $+$, $\rightarrow$ on the other hand. The second half extends this analogy, and for well-understood concepts from logic proposes new concepts for types that correspond to them. In particular, Howard proposes that the predicate quantifiers $\forall$ and $\exists$ corresponds to new types we now call "dependent types."

With the introduction of dependent types, every proof in predicate logic can be represented by a term of a suitable typed lambda calculus. Mathematicians and computer scientists proposed numerous systems based on this concept, including de Bruijn's Automath,¹³ Martin-Löf's type theory,²⁶ Bates and Constable's PRL and nuPRL,² and Coquand and Huet's Calculus of Constructions,⁹ which developed into the Coq proof assistant.

Applications include CompCert, a certified compiler for the C programming language verified in Coq; a computer-checked proof of the four-color theorem also verified in Coq; parts of the Ensemble distributed system verified in NuPRL; and 20,000 lines of browser plug-ins verified in F*.

de Bruijn's work was independent of Howard's, but Howard directly inspired Martin-Löf and all the other work listed earlier. Howard was (justly!) proud of his paper, citing it as one of the two great achievements of his career.³⁴

Intuitionistic Logic

In Gilbert and Sullivan's *The Gondoliers*, Casilda is told that as an infant she

was married to the heir of the King of Batavia, but that due to a mix-up no one knows which of two individuals, Marco or Giuseppe, is the heir. Alarmed, she wails, "Then do you mean to say that I am married to one of two gondoliers, but it is impossible to say which?" To which the response is "Without any doubt of any kind whatever."

Logic comes in many varieties, and one distinction is between "classical" and "intuitionistic." Intuitionists, concerned by cavalier assumptions made by some logicians about the nature of infinity, insist upon a constructionist notion of truth. In particular, they insist that a proof of $A \vee B$ must show which of A or B holds, and hence they would reject the claim that Casilda is married to Marco or Giuseppe until one of the two was identified as her husband. Perhaps Gilbert and Sullivan anticipated intuitionism, for their story's outcome is that the heir turns out to be a third individual, Luiz, with whom Casilda is, conveniently, already in love.

Intuitionists also reject the law of the excluded middle, which asserts $A \vee \neg A$ for every A , since the law gives no clue as to which of A or $\neg A$ holds. Heyting formalized a variant of Hilbert's classical logic that captures the intuitionistic notion of provability. In particular, the law of the excluded middle is provable in Hilbert's logic, but not in Heyting's. Further, if the law of the excluded middle is added as an axiom to Heyting's logic, then it becomes equivalent to Hilbert's.

Propositions as Types was first formulated for intuitionistic logic. It is a perfect fit, because in the intuitionist interpretation the formula $A \vee B$ is provable exactly when one exhibits either a proof of A or a proof of B , so the type corresponding to disjunction is a disjoint sum.

Other Logics, Other Computation

The principle of Propositions as Types would be remarkable even if it applied only to one variant of logic and one variant of computation. How much more remarkable, then, that it applies to a wide variety of logics and of computation.

Quantification over propositional variables in second-order logic corresponds to type abstraction in second-order lambda calculus. For this reason,

the second-order lambda calculus was discovered twice, once by the logician Girard¹⁶ and once by the computer scientist Reynolds.³³ And for the same reason, a similar system that supports principle type inference was also discovered twice, once by the logician Hindley²⁰ and once by the computer scientist Milner.²⁷ Building on the correspondence, Mitchell and Plotkin²⁸ observed existential quantification in second-order logic corresponds precisely to data abstraction, an idea that now underpins much research in the semantics of programming languages. The design of generic types in Java and C# draws directly upon Girard-Reynolds, while the type systems of functional languages, including ML and Haskell, are based on Hindley-Milner. Philosophers might argue as to whether mathematical systems are "discovered" or "devised," but the same system arising in two different contexts argues that here the correct word is "discovered."

Two major variants of logic are intuitionistic and classical. Howard's original paper observed a correspondence with intuitionistic logic. Not until two decades later was the correspondence extended to also apply to classical logic, when Griffin¹⁹ observed that Peirce's Law in classical logic provides a type for the `call/cc` operator of Scheme. Murthy³¹ went on to note that Kolmogorov and Gödel's double-negation translation, widely used to relate intuitionistic and classical logic, corresponds to the continuation-passing style transformation widely used by both semanticists and implementers of lambda calculus. Parigot,³² Curien and Herbelin,¹⁰ and Wadler³⁹ introduced various computational calculi motivated by correspondences to classical logic.

Modal logic permits propositions to be labeled as "necessarily true" or "possibly true." Clarence Lewis introduced modal logic in 1910, and his 1938 textbook²⁵ describes five variants, S1–S5. Some claim each of these variants has an interpretation as a form of computation via Propositions as Types, and a down payment on this claim is given by an interpretation of S4 as staged computation due to Davies and Pfenning,¹² and of S5 as spatially distributed computation due to Murphy et al.³⁰

Moggi²⁹ introduced monads as a

Figure 1. Gerhard Gentzen (1935)—Natural Deduction.

$$\begin{array}{c}
 \frac{A \quad B}{A \& B} \&-I \quad \frac{A \& B}{A} \&-E_1 \quad \frac{A \& B}{B} \&-E_2 \\
 \\
 \frac{[A]^x \quad \vdots \quad B}{A \supset B} \supset-I^x \quad \frac{A \supset B \quad A}{B} \supset-E
 \end{array}$$

Figure 2. A proof.

$$\frac{\frac{[B \& A]^z}{A} \&-E_2 \quad \frac{[B \& A]^z}{B} \&-E_1}{A \& B} \&-I \quad \frac{A \& B}{(B \& A) \supset (A \& B)} \supset-I^z$$

Figure 3. Simplifying proofs.

$$\begin{array}{c}
 \frac{\frac{A \quad B}{A \& B} \&-I}{A} \&-E_1 \Rightarrow A \\
 \\
 \frac{[A]^x \quad \vdots \quad B}{A \supset B} \supset-I^x \quad \frac{A \supset B \quad A}{B} \supset-E \Rightarrow B
 \end{array}$$

Figure 4. Simplifying a proof.

$$\begin{array}{c}
 \frac{[B \& A]^z}{A} \&-E_2 \quad \frac{[B \& A]^z}{B} \&-E_1}{A \& B} \&-I \quad \frac{B \quad A}{B \& A} \&-I \\
 \frac{A \& B}{(B \& A) \supset (A \& B)} \supset-I^z \quad \frac{B \& A}{B} \&-E_2}{A \& B} \supset-E \\
 \\
 \Downarrow \\
 \frac{\frac{B \quad A}{B \& A} \&-I}{A} \&-E_2 \quad \frac{\frac{B \quad A}{B \& A} \&-I}{B} \&-E_1}{A \& B} \&-I \\
 \\
 \Downarrow \\
 \frac{A \quad B}{A \& B} \&-I
 \end{array}$$

technique to explain the semantics of important features of programming languages such as state, exceptions, and input-output. Monads became widely adopted in the functional language Haskell and later migrated into other languages, including Clojure, Scala, F#, and C#. Benton et al.³ observed that monads correspond to yet another modal logic, differing from all of S1–S5.

In classical, intuitionistic, and modal logic, any hypothesis can be used an arbitrary number of times—zero, once, or many. Linear logic, introduced in 1987 by Girard,¹⁷ requires that each hypothesis is used exactly once. Linear logic is “resource conscious” in that facts may be used up and superseded by other facts, suiting it for reasoning about the world where situations change. Computational aspects of linear logic are discussed by Abramsky¹ and Wadler,³⁸ among many others. Most recently, Session Types, a way of describing communication protocols introduced by Honda,²¹ have been related to intuitionistic linear logic by Caires and Pfenning,⁴ and to classical linear logic by Wadler.⁴⁰

Propositions as Types remains a topic of active research.

Natural Deduction

We now turn to a more formal development, presenting a fragment of natural deduction and a fragment of typed lambda calculus in a style that makes clear the connection between the two.

We begin with the details of natural deduction as defined by Gentzen¹⁵; the proof rules are shown in Figure 1. To simplify our discussion, we consider just two of the connectives of natural deduction. We write A and B as placeholders standing for arbitrary formulas. Conjunction is written $A \& B$, and implication is written $A \supset B$.

We represent proofs by trees, where each node of the tree is an instance of a proof rule. Each proof rule consists of zero or more formulas written above a line, called the “premises,” and a single formula written below the line, called the “conclusion.” The interpretation of a rule is that when all the premises hold, then the conclusion follows.

The proof rules come in pairs, with rules to introduce and to eliminate each connective, labeled $\&-I$ and $\&-E$, respectively. As we read the rules from top to bottom, introduction and elimination rules do what they say on the tin: The first “introduces” a formula for the connective, which appears in the conclusion but not in the premises; the second “eliminates” a formula for the connective, which appears in a premise but not in the conclusion. An introduction rule describes under what conditions we say the connective holds—how to *define* the connective. An elimination rule describes what we may conclude when the connective holds—how to *use* the connective.

The introduction rule for conjunction, $\&-I$, states that if formula A holds and formula B holds, then the formula $A \& B$ must hold as well. There are two elimination rules for conjunction. The first, $\&-E_1$, states that if the formula $A \& B$ holds, then the formula A must hold as well. The second, $\&-E_2$, concludes B rather than A .

The introduction rule for implication, $\supset-I$, states that if from the assumption that formula A holds we may derive the formula B , then we may conclude the formula $A \supset B$ holds and discharge the assumption. To indicate that A is used as an assumption zero, once, or many times in the proof of B , we write A in brackets and tether it to B via ellipses. A proof is complete only when every assumption in it has been discharged by a corresponding use of $\supset-I$, which is indicated by writing the same name (here x) as a superscript on each instance of the discharged assumption and on the discharging

rule. The elimination rule for implication, $\supset$ -E, states that if formula $A \supset B$ holds and if formula A holds, then we may conclude formula B holds as well; as mentioned earlier, this rule also goes by the name modus ponens.

Critical readers will observe we use similar language to describe rules (“when-then”) and formulas (“implies”). The same idea applies at two levels, the meta level (rules) and the object level (formulas), and in two notations, using a line with premises above and conclusion below for implication at the meta level, and the symbol $\supset$ with premise to the left and conclusion to the right at the object level. It is almost as if to understand implication one must first understand implication! This Zeno’s paradox of logic was wryly observed by Carroll.⁵ We need not let it disturb us; everyone possesses a good informal understanding of implication, which may act as a foundation for its formal description.

A proof of the formula

$$(B \& A) \supset (A \& B).$$

is shown in Figure 2; that is, if B and A hold, then A and B hold. This may seem so obvious as to be hardly deserving of proof! However, the formulas $B \supset A$ and $A \supset B$ have meanings that differ, and we need some formal way to conclude that the formulas $B \& A$ and $A \& B$ have meanings that are the same. This is what our proof shows, and it is reassuring it can be constructed from the rules we posit.

The proof reads as follows. From $B \& A$ we conclude A , by $\&$ -E2, and from $B \& A$ we also conclude B , by $\&$ -E1. From A and B we conclude $A \& B$, by $\&$ -I. That is, from the assumption $B \& A$ (used twice) we conclude $A \& B$. We discharge the assumption and conclude $(B \& A) \supset (A \& B)$ by $\supset$ -I, linking the discharged assumptions to the discharging rule by writing z as a superscript on each.

Some proofs are unnecessarily roundabout. Rules for simplifying proofs appear in Figure 3, and an example appears in Figure 4. Let us focus on the example first.

The top of Figure 4 shows a larger proof built from the proof in Figure 2. The larger proof assumes as premises two formulas, B and A , and concludes with the formula $A \& B$. However, rather

Figure 5. Alonzo Church (1935)—Lambda Calculus.

$$\begin{array}{c} \frac{M:A \quad N:B}{\langle M, N \rangle : A \times B} \times\text{-I} \quad \frac{L:A \times B}{\pi_1 L:A} \times\text{-E}_1 \quad \frac{L:A \times B}{\pi_2 L:B} \times\text{-E}_2 \\ \\ \frac{[x:A]^x \quad \vdots \quad N:B}{\lambda x. N : A \rightarrow B} \rightarrow\text{-I}^x \quad \frac{L:A \rightarrow B \quad M:A}{LM:B} \rightarrow\text{-E} \end{array}$$

Figure 6. A program.

$$\frac{\frac{[z:B \times A]^z}{\pi_2 z:A} \times\text{-E}_2 \quad \frac{[z:B \times A]^z}{\pi_1 z:B} \times\text{-E}_1}{\langle \pi_2 z, \pi_1 z \rangle : A \times B} \times\text{-I} \quad \frac{\lambda z. \langle \pi_2 z, \pi_1 z \rangle : (B \times A) \rightarrow (A \times B)}{\lambda z. \langle \pi_2 z, \pi_1 z \rangle : (B \times A) \rightarrow (A \times B)} \rightarrow\text{-I}^z$$

Figure 7. Evaluating programs.

$$\begin{array}{c} \frac{M:A \quad N:B}{\langle M, N \rangle : A \times B} \times\text{-I} \quad \frac{\langle M, N \rangle : A \times B}{\pi_1 \langle M, N \rangle : A} \times\text{-E}_1 \quad \Rightarrow \quad M:A \\ \\ \frac{[x:A]^x \quad \vdots \quad N:B}{\lambda x. N : A \rightarrow B} \rightarrow\text{-I}^x \quad \frac{\vdots \quad M:A}{\lambda x. N : A \rightarrow B} \rightarrow\text{-E} \quad \Rightarrow \quad N[M/x]:B \end{array}$$

Figure 8. Evaluating a program.

$$\begin{array}{c} \frac{[z:B \times A]^z}{\pi_2 z:A} \times\text{-E}_2 \quad \frac{[z:B \times A]^z}{\pi_1 z:B} \times\text{-E}_1}{\langle \pi_2 z, \pi_1 z \rangle : A \times B} \times\text{-I} \quad \frac{\lambda z. \langle \pi_2 z, \pi_1 z \rangle : (B \times A) \rightarrow (A \times B)}{\lambda z. \langle \pi_2 z, \pi_1 z \rangle : (B \times A) \rightarrow (A \times B)} \rightarrow\text{-I}^z \quad \frac{y:B \quad x:A}{\langle y, x \rangle : B \times A} \times\text{-I} \\ \frac{\lambda z. \langle \pi_2 z, \pi_1 z \rangle : (B \times A) \rightarrow (A \times B) \quad \langle y, x \rangle : B \times A}{(\lambda z. \langle \pi_2 z, \pi_1 z \rangle) \langle y, x \rangle : A \times B} \rightarrow\text{-E} \\ \\ \Downarrow \\ \frac{y:B \quad x:A}{\langle y, x \rangle : B \times A} \times\text{-I} \quad \frac{y:B \quad x:A}{\langle y, x \rangle : B \times A} \times\text{-I} \quad \frac{\langle y, x \rangle : B \times A}{\pi_2 \langle y, x \rangle : A} \times\text{-E}_2 \quad \frac{\langle y, x \rangle : B \times A}{\pi_1 \langle y, x \rangle : B} \times\text{-E}_1}{\langle \pi_2 \langle y, x \rangle, \pi_1 \langle y, x \rangle \rangle : A \times B} \times\text{-I} \\ \\ \Downarrow \\ \frac{x:A \quad y:B}{\langle x, y \rangle : A \times B} \times\text{-I} \end{array}$$

than concluding it directly we derive the result in a roundabout way, in order to illustrate an instance of $\supset$ -E, modus ponens. The proof reads as follows: On the left is the proof given previously, concluding in $(B \& A) \supset (A \& B)$; on the right, from B and A we conclude $B \& A$ by $\&$ -I. Combining these yields $A \& B$ by $\supset$ -E.

We may simplify the proof by applying the rewrite rules of Figure 3. These rules specify how to simplify a proof when an introduction rule is immediately followed by the corresponding elimination rule. Each rule shows two proofs connected by an arrow, indicating that the “redex” (the proof on the left) may be rewritten, or simplified, to yield the “reduct” (the proof on the right). Rewrites always take a valid proof to another valid proof.

For $\&$, the redex consists of a proof of A and a proof of B that combine to yield $A \& B$ by $\&$ -I, which in turn yields A by $\&$ -E₁. The reduct consists simply of the proof of A , discarding the unneeded proof of B . There is a similar rule, not shown, to simplify an occurrence of $\&$ -I followed by $\&$ -E₂.

For $\supset$, the redex consists of a proof of B from assumption A , which yields $A \supset B$ by $\supset$ -I, and a proof of A , which combine to yield B by $\supset$ -E. The reduct consists of the same proof of B , but now with every occurrence of the assumption A replaced by the given proof of A . The assumption A may be used zero, once, or many times in the proof of B in the redex, so the proof of A may be copied zero, once, or many times in the proof of B in the reduct. For this reason, the reduct may be larger than the redex, but it will be simpler in the sense it has removed an unnecessary detour via the subproof of $A \supset B$.

We can think of the assumption of A in $\supset$ -I as a debt that is discharged by the proof of A provided in $\supset$ -E. The proof in the redex accumulates debt and pays it off later; while the proof in the reduct pays directly each time the assumption is used. Proof debt differs from monetary debt in that there is no interest, and the same proof may be duplicated freely as many times as needed to pay off an assumption, the very property that money, by being difficult to counterfeit, is designed to avoid!

Figure 4 demonstrates use of these rules to simplify a proof. The first proof

contains an instance of $\supset$ -I followed by $\supset$ -E and is simplified by replacing each of the two assumptions of $B \& A$ on the left by a copy of the proof of $B \& A$ on the right. The result is the second proof, which, as a result of the replacement, now contains an instance of $\&$ -I followed by $\&$ -E₂, and another instance of $\&$ -I followed by $\&$ -E₁. Simplifying each of these yields the third proof, which derives $A \& B$ directly from the assumptions A and B and can be simplified no further.

It is not difficult to see that proofs in normal form satisfy the Subformula Principle: Every formula of such a proof must be a subformula of one of its undischarged assumptions or of its conclusion. The proof in Figure 2 and the final proof of Figure 4 both satisfy this property, while the first proof of Figure 4 does not, since $(B \& A) \supset (A \& B)$ is not a subformula of $A \& B$.

Lambda Calculus

We now turn our attention to the simply typed lambda calculus of Church⁸; the type rules are in Figure 5. To simplify our discussion, we take both products and functions as primitive types; Church’s original calculus contained only function types, with products as a derived construction. We now write A and B as placeholders for arbitrary types, and L , M , N as placeholders for arbitrary terms. Product types are written $A \times B$, and function types are written $A \rightarrow B$. Now, instead of formulas, our premises and conclusions are judgments of the form

$$M : A$$

indicating term M has type A .

Like proofs, we represent type derivations by trees, where each node of the tree is an instance of a type rule. Each type rule consists of zero or more judgments written above a line, called the “premises,” and a single judgment written below the line, called the “conclusion.” The interpretation of a rule is that when all the premises hold, then the conclusion follows.

Like proof rules, type rules come in pairs. An introduction rule describes how to *define* or *construct* a term of the given type, while an elimination rule describes how to *use* or *deconstruct* a term of the given type.

The introduction rule for products,

$\times$ -I, states that if term M has type A and term N has type B , then we may form the pair term $\langle M, N \rangle$ of product type $A \times B$. There are two elimination rules for products. The first, $\times$ -E₁, states that if term L has type $A \times B$, then we may form the term $\pi_1 L$ of type A , which selects the first component of the pair. The second, $\times$ -E₂ is similar, save that it forms the term $\pi_2 L$ of type B .

The introduction rule for functions, $\rightarrow$ -I, states that if given a variable x of type A we have formed a term N of type B , then we may form the lambda term $\lambda x. N$ of function type $A \rightarrow B$. The variable x appears free in N and bound in $\lambda x. N$. Undischarged assumptions correspond to free variables, while discharged assumptions correspond to bound variables. To indicate that the variable x may appear zero, once, or many times in the term N , we write $x : A$ in brackets and tether it to $N : B$ via ellipses. A term is closed only when every variable in it is bound by a corresponding λ term. The elimination rule for functions, $\rightarrow$ -E, states that given term L of type $A \rightarrow B$ and term M of type A we may form the application term LM of type B .

For natural deduction, we noted earlier there might be confusion between implication at the meta level and at the object level. For lambda calculus the distinction is clearer, as we have implication at the meta level (if terms above the line are well typed, then so are terms below) but functions at the object level (a function has type $A \rightarrow B$ because if it is passed a value of type A then it returns a value of type B). What previously had been discharge of assumptions (perhaps a slightly diffuse concept) becomes binding of variables (a concept understood by most computer scientists).

The reader will have observed a striking similarity between Gentzen’s rules from the preceding section and Church’s rules from this section; ignoring the terms in Church’s rules then they are identical if one replaces $\&$ by $\times$ and $\supset$ by $\rightarrow$. The coloring of the rules is chosen to highlight the similarity.

A program of type

$$(B \times A) \rightarrow (A \times B)$$

is shown in Figure 6. Whereas the difference between $B \& A$ and $A \& B$ appears a mere formality, the difference between

$B \times A$ and $A \times B$ is easier to appreciate; converting the latter to the former requires swapping the elements of the pair, which is precisely the task performed by the program corresponding to our former proof.

The program reads as follows. From variable z of type $B \times A$ we form term $\pi_2 z$ of type A by $\times\text{-}E_2$ and also term $\pi_1 z$ of type B by $\times\text{-}E_1$. From these two terms we form the pair $\langle \pi_2 z, \pi_1 z \rangle$ of type $A \times B$ by $\times\text{-}I$. Finally, we bind the free variable z to form the lambda term $\lambda z. \langle \pi_2 z, \pi_1 z \rangle$ of type $(B \times A) \rightarrow (A \times B)$ by $\rightarrow\text{-}I$, connecting the bound typings to the binding rule by writing z as a superscript on each. The function accepts a pair and swaps its elements, exactly as described by its type.

A program may be evaluated by rewriting. Rules for evaluating programs appear in Figure 7, and an example appears in Figure 8. Let us focus on the example first.

The top of Figure 8 shows a larger program built from the program in Figure 6. The larger program has two free variables, y of type B and x of type A , and constructs a value of type $A \times B$. However, rather than constructing it directly we reach the result in a roundabout way, in order to illustrate an instance of $\rightarrow\text{-}E$, function application. The program reads as follows: On the left is the program given previously, forming a function of type $(B \times A) \rightarrow (A \times B)$. On the right, from B and A we form the pair $\langle y, x \rangle$ of type $B \times A$ by $\times\text{-}I$. Applying the function to the pair forms a term of type $A \times B$ by $\rightarrow\text{-}E$.

We may evaluate this program by applying the rewrite rules of Figure 7. These rules specify how to rewrite a term when an introduction rule is immediately followed by the corresponding elimination rule. Each rule shows two derivations connected by an arrow, indicating the “redex” (the term on the left) may be rewritten, or evaluated, to yield the “reduct” (the term on the right). Rewrites always take a valid type derivation to another valid type derivation, ensuring rewrites preserve types, a property known as “subject reduction” or “type soundness.”

For $\times$, the redex consists of term M of type A and term N of type B that combine to yield term $\langle M, N \rangle$ of type $A \times B$ by $\times\text{-}I$, which in turn yields term $\pi_1 \langle M, N \rangle$ of type A by $\times\text{-}E_1$. The reduct

consists simply of term M of type A , discarding the unneeded term N of type B . There is a similar rule, not shown, to rewrite an occurrence of $\times\text{-}I$ followed by $\times\text{-}E_2$.

For $\rightarrow$, the redex consists of a derivation of term N of type B from variable x of type A , which yields the lambda term $\lambda x. N$ of type $A \rightarrow B$ by $\rightarrow\text{-}I$, and a derivation of term M of type A , which combine to yield the application $(\lambda x. N) M$ of type B by $\rightarrow\text{-}E$. The reduct consists of the term $N[M/x]$, which replaces each free occurrence of the variable x in term N by term M . Further, if in the derivation that N has type B we replace each assumption that x has type A by the derivation that M has type A , we get a derivation showing $N[M/x]$ has type B . Since the variable x may appear zero, once, or many times in the term N , the term M may be copied zero, once, or many times in the reduct $N[M/x]$. For this reason, the reduct may be larger than the redex, but it will be simpler in the sense it has removed a subterm of type $A \rightarrow B$. Discharge of assumptions thus corresponds to applying a function to its argument.

Figure 8 demonstrates use of these rules to evaluate a program. The first program contains an instance of $\rightarrow\text{-}I$ followed by $\rightarrow\text{-}E$, and is rewritten by replacing each of the two occurrences of z of type $B \times A$ on the left by a copy of the term $\langle y, x \rangle$ of type $B \times A$ on the right. The result is the second program, which, as a result of the replacement, now contains an

instance of $\times\text{-}I$ followed by $\times\text{-}E_2$, and another instance of $\times\text{-}I$ followed by $\times\text{-}E_1$. Rewriting each of these yields the third program, which derives the term $\langle x, y \rangle$ of type $A \times B$, and can be evaluated no further.

Hence, simplification of proofs corresponds exactly to evaluation of programs, in this instance demonstrating that applying the function to the pair indeed swaps its elements.

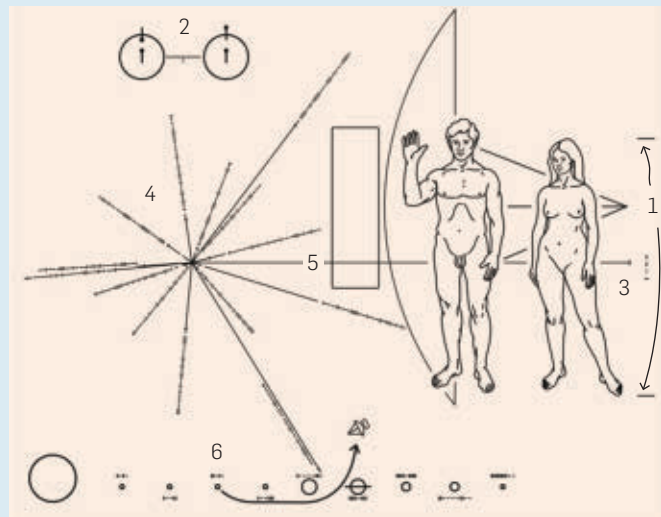
Conclusion

Proposition as Types informs our view of the universality of certain programming languages.

The *Pioneer* spaceship includes a plaque designed to communicate with aliens, if any should ever intercept it (see Figure 9). They may find some parts of it easier to interpret than others. A radial diagram shows the distance of 14 pulsars and the center of our galaxy from Sol. Aliens are likely to determine the length of each line is proportional to the distances to each body. Another diagram shows humans in front of a silhouette of *Pioneer*. If *Star Trek* gives an accurate conception of alien species, they may respond, “They look just like us, except they lack pubic hair.” However, if the aliens’ perceptual system differs greatly from our own, they may be unable to decipher these squiggles.

What would happen if we tried to communicate with aliens by transmitting a computer program? In the movie *Independence Day*, the heroes destroy the invading alien mothership by infecting

Figure 9. Plaque on *Pioneer* spaceship.



it with a computer virus. Close inspection of the transmitted program shows it contains curly braces; it is written in a dialect of C! It is unlikely that alien species would program in C and unclear that aliens could decipher a program written in C if presented with one.

What about lambda calculus? Propositions as Types tell us lambda calculus is isomorphic to natural deduction. It seems difficult to conceive of alien beings who do not know the fundamentals of logic, and we might expect the problem of deciphering a program written in lambda calculus to be closer to the problem of understanding the radial diagram of pulsars than of understanding the image of a man and a woman on the *Pioneer* plaque.

We might be tempted to conclude lambda calculus is universal, but first ponder the suitability of the word “universal.” These days, the multiple-worlds interpretation of quantum physics is widely accepted. Scientists imagine that in different universes one might encounter different fundamental constants (such as the strength of gravity or the Planck constant). But easy as it may be to imagine a universe where gravity differs, it is difficult to conceive of a universe where fundamental rules of logic fail to apply. Natural deduction, and hence lambda calculus, should not only be known by aliens throughout our universe but also throughout others. So we may conclude it would be a mistake to characterize lambda calculus as a universal language, because calling it universal would be *too limiting*.

Acknowledgments

Thank you to Gershon Bazerman, Pete Bevin, Guy Blelloch, Rintcius Blok, Ezra Cooper, Ben Darwin, Benjamin Denckla, Peter Dybjer, Johannes Emerich, Martin Erwig, Yitz Gale, Mikhail Glushenkov, Gabor Greif, Vinod Grover, Sylvain Henry, Philip Hölzenspies, William Howard, John Hughes, Colin Lupton, Daniel Marsden, Craig McLaughlin, Tom Moertel, Simon Peyton-Jones, Benjamin Pierce, Lee Pike, Andrés Sicard-Ramírez, Scott Rostrup, Dann Toliver, Moshe Vardi, Jeremy Yallop, Richard Zach, Leo Zovik, and the referees. This work was funded under Engineering and Physical Sciences Research Council grant EP/K034413/1. 

References

- Abramsky, S. Computational interpretations of linear logic. *Theoretical Computer Science* 111, 1–2 (1993), 3–57.
- Bates, J.L., Constable, R.L. Proofs as programs. *Transactions on Programming Languages and Systems* 7, 1 (Jan. 1985), 113–136.
- Benton, P.N., Bierman, G.M., de Paiva, V. Computational types from a logical perspective. *Journal of Functional Programming* 8, 2 (1998), 177–193.
- Caires, L., Pfenning, F. Session types as intuitionistic linear propositions. In *Proceedings of the 21st International Conference on Concurrency Theory* (Paris, France, Aug. 31–Sept. 3, 2010), 222–236.
- Carroll, L. What the Tortoise said to Achilles. *Mind* 4, 14 (Apr. 1895), 278–280.
- Church, A. A set of postulates for the foundation of logic. *Ann. Math.* 33, 2 (1932), 346–366.
- Church, A. An unsolvable problem of elementary number theory. *American Journal of Mathematics* 58, 2 (Apr. 1936), 345–363; presented to the American Mathematical Society, Apr. 19, 1935; abstract in *Bulletin of the American Mathematical Society* 41 (May 1935).
- Church, A. A formulation of the simple theory of types. *Journal of Symbolic Logic* 5, 2 (June 1940), 56–68.
- Coquand, T. and Huet, G.P. The calculus of constructions. *Information and Computation* 76, 2/3 (1988), 95–120.
- Curien, P.-L., Herbelin, H. The duality of computation. In *Proceedings of the International Conference on Functional Programming* (Montreal, Canada, Sept. 18–20). ACM Press, New York, 2000, 233–243.
- Curry, H.B. Functionality in combinatory logic. *Proceedings of the National Academy of Science* 20 (1934), 584–590.
- Davies, R., Pfenning, F. A modal analysis of staged computation. In *Principles of Programming Languages* (St. Petersburg Beach, FL, 1996), 258–270.
- de Bruijn, N.G. The mathematical language Automath, its usage, and some of its extensions. In *Proceedings of the Symposium on Automatic Demonstration, Volume 125 of Lecture Notes in Computer Science* (Versailles, France, Dec.). Springer-Verlag, 1968, 29–61.
- Gandy, R. The confluence of ideas in 1936. In *The Universal Turing Machine: A Half-century Survey*, R. Herken, Ed. Springer, 1995, 51–102.
- Gentzen, G. Untersuchungen über das logische Schließen. *Math. Z.* 39, 2–3 (1935), 176–210, 405–431; reprinted in Szabo.³⁵
- Girard, J.Y. Interprétation fonctionnelle et élimination des coupures dans l’arithmétique d’ordre supérieure. These D’Etat, Université Paris VII, 1972.
- Girard, J.-Y. Linear logic. *Theoretical Computer Science* 50 (1987), 1–102.
- Gödel, K. Über formal unterscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik* 38 (1931), 173–198; reprinted in Heijenoort.³⁷
- Griffin, T. A formulae-as-types notion of control. In *Proceedings of the 40th Annual Symposium on Principles of Programming Languages* (Rome, Italy, Jan. 23–25). ACM Press, New York, 1990, 47–58.
- Hindley, R. The principal type scheme of an object in combinatory logic. *Transactions of the American Mathematical Society* 146 (Dec. 1969), 29–60.
- Honda, K. Types for dyadic interaction. In *Proceedings of the Fourth International Conference on Concurrency Theory* (Hildesheim, Germany, Aug. 23–26, 1993), 509–523.
- Howard, W.A. The formulae-as-types notion of construction. In *To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus, and Formalism*. Academic Press, 1980, 479–491; original version was circulated privately in 1969.
- Kleene, S. Origins of recursive function theory. *Annals of the History of Computing* 3, 1 (1981), 52–67.
- Kleene, S.C. General recursive functions of natural numbers. *Mathematical Annalen* 112, 1 (Dec. 1936); abstract in *Bulletin of the AMS* (July 1935).
- Lewis, C. and Langford, C. *Symbolic Logic*, 1938; reprinted by Dover, 1959.
- Martin-Löf, P. *Intuitionistic Type Theory*. Bibliopolis, Naples, Italy, 1984.
- Milner, R. A theory of type polymorphism in programming. *Journal of Computer and System Sciences* 17, 3 (1978), 348–375.
- Mitchell, J.C., Plotkin, G.D. Abstract types have existential type. *Transactions on Programming Languages and Systems* 10, 3 (July 1988), 470–502.
- Moggi, E. Notions of computation and monads. *Information and Computation* 93, 1 (1991), 55–92.
- Murphy VII, T., Cray, K., Harper, R., Pfenning, F. A symmetric modal lambda calculus for distributed computing. In *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science* (Turku, Finland, July 13–17). IEEE Press, 2004, 286–295.
- Murthy, C. An evaluation semantics for classical proofs. In *Proceedings of Sixth Annual IEEE Symposium on Logic in Computer Science* (Amsterdam, the Netherlands, July 15–18). IEEE Press, 1991, 96–107.
- Parigot, M. $\lambda\mu$ -calculus: An algorithmic interpretation of classical natural deduction. In *Logic Programming and Automated Reasoning, Volume 624 of Lecture Notes in Computer Science*. Springer-Verlag, 1992, 190–201.
- Reynolds, J.C. Towards a theory of type structure. In *Proceedings of the Symposium on Programming, Volume 19 of Lecture Notes in Computer Science* (1974), 408–423.
- Shell-Gellasch, A.E. Reflections of my advisor: Stories of mathematics and mathematicians. *The Mathematical Intelligencer* 25, 1 (2003), 35–41.
- Szabo, M.E., Ed. *The Collected Papers of Gerhard Gentzen*. North Holland Publishing Co., Amsterdam, the Netherlands, 1969.
- Turing, A.M. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* s2–42, 1 (1937); received May 28, 1936, read Nov. 12, 1936.
- van Heijenoort, J. *From Frege to Gödel: A Sourcebook in Mathematical Logic, 1879–1931*. Harvard University Press, Cambridge, MA, 1967.
- Wadler, P. A taste of linear logic. In *Proceedings of the 18th International Symposium on Mathematical Foundations of Computer Science Volume 711 of Lecture Notes on Computer Science* (Gdańsk, Poland, Aug. 30–Sept. 3). Springer-Verlag, 1993, 185–210.
- Wadler, P. Call-by-value is dual to call-by-name. In *Proceedings of the International Conference on Functional Programming* (Uppsala, Sweden, Aug. 25–29). ACM Press, New York, 2003, 189–201.
- Wadler, P. Propositions as sessions. In *Proceedings of the International Conference on Functional Programming* (Copenhagen, Denmark, Sept. 10–12). ACM Press, New York, 2012, 273–286.

Philip Wadler (wadler@inf.ed.ac.uk) is Professor of Theoretical Computer Science in the Laboratory for Foundations of Computer Science in the School of Informatics at the University of Edinburgh, Scotland.

Copyright held by authors.
Publication rights licensed to ACM. \$15.00.

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 18–23, 2016** 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 3, 2016**.

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



DOI:10.1145/2756543

Economic incentives that alleviate congestion for Internet customers can also improve business performance for network operators.

BY SOUMYA SEN, CARLEE JOE-WONG,
SANGTAE HA, AND MUNG CHIANG

Smart Data Pricing: Using Economics to Manage Network Congestion

SINCE 2010, THERE has been rapid evolution in pricing practices among Internet service providers (ISPs) in the U.S. and other international markets, particularly in moving away from flat-rate to usage-based pricing in cellular networks.¹⁹ In 2010, AT&T eliminated unlimited data plans and introduced a tiered plan of about \$10/GB, along with deliberately slowing the traffic of heavy users and adding hefty overage fees. Verizon and other U.S. ISPs have since introduced similar data plans. These changes have fueled the continuing debate about Net neutrality and the openness of the Internet.

ISPs argue these price increases and usage penalties are needed in response to rapid growth in data traffic, as driven by increasing demand for smart devices, bandwidth-hungry applications, cloud-based services, machine-to-machine traffic, and media-rich Web content.^{5,20} This explosive growth, which Cisco's visual networking index projects will cause a nearly tenfold increase in global wireless traffic between 2014 and 2019,⁴ requires investment in expanding wired and wireless network capacities (such as additional spectrum, Wi-Fi hotspots for offloading data traffic, backhauling infrastructure, and newer technologies like 4G/LTE and femtocells). The benefits of this capacity expansion are partly accrued by the content providers who attract more advertising and e-commerce revenue from greater user demand while further driving demand for bandwidth. ISPs contend they are trapped in a vicious cycle that does not allow them to match their prices to their costs. Measures (such as throttling, data caps, and usage-based metered pricing) are thus viewed as essential for regulating demand and managing network congestion.

However, penalizing demand would be harmful for the Internet ecosystem and could restrict network access for some users. Appreciating the role of economics in network management

» key insights

- Network operators should not see demand growth as a problem to be solved by penalizing users but as an opportunity to monetize their networks by broadening the revenue base and managing congestion by creating the right economic incentives for users.
- Two complementary approaches—time-dependent pricing and traffic offloading—aim to reduce network congestion by giving users incentives and mechanisms to shift their use to less-congested times or frequencies and networks.
- Smart data pricing encompasses other approaches in network pricing and engineering, including toll-free sponsored data, to benefit network operators, consumers, and content providers alike.



and understanding how these models can be realized in practical network systems is crucial to the Internet's long-term growth and sustainability.

Exploring the link between economic principles and network engineering is the goal of several recent research efforts in smart data pricing (SDP).²⁴ SDP mechanisms go beyond simple byte-counting schemes to include time/location/app-based dynamic pricing, usage-based pricing with differentiated speed tiers, auction-based smart markets, Wi-Fi offloading, proactive caching, zero-rating or sponsored content, and quota-aware content adaptation. In gen-

eral, it asks three kinds of questions along the following dimensions:

Who pays? Who should pay for bandwidth (such as zero-rating, sponsored content, and two-sided pricing)?;

What services? What service should be charged for (such as transaction-based pricing and quality-based pricing)?; and

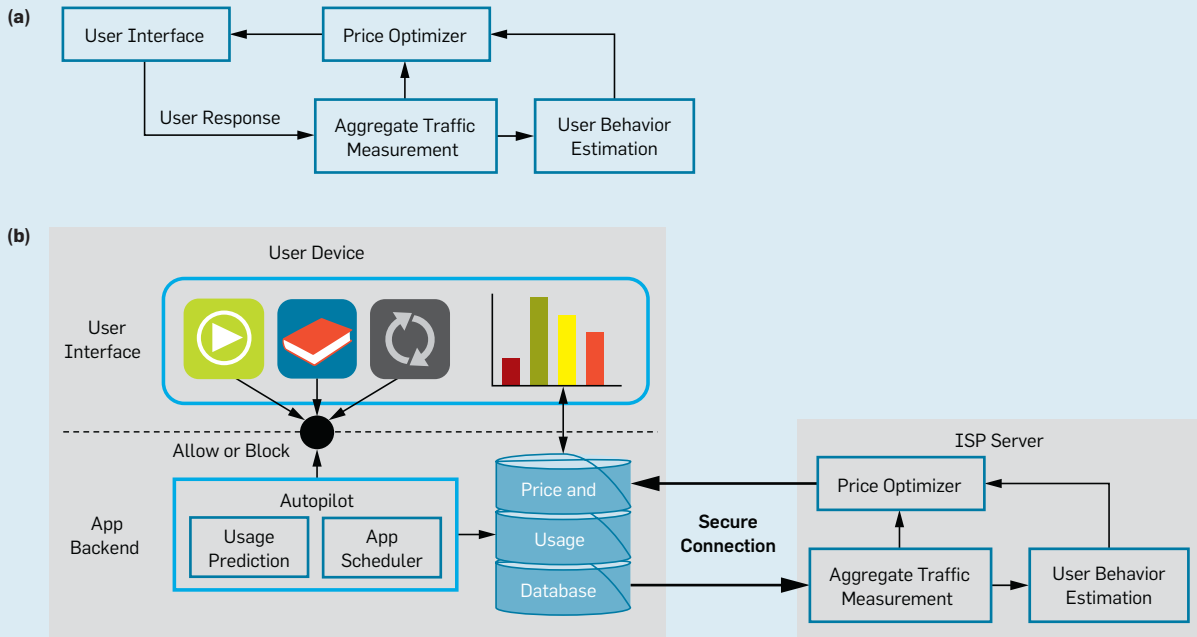
How to charge? How to enable and charge for mechanisms like time/location/congestion-dependent pricing and traffic offloading?

In this article, we mainly address “How?,” reporting on two research directions—time-dependent pricing (TDP) and traffic-offloading mecha-

nisms. Although researchers have been exploring the interplay between networks and economics for years^{5,8,15} (for a detailed survey of various pricing proposals, see Sen et al.²¹), the need for designing and demonstrating fully functional prototypes has become more urgent due to users' growing demand for data. Consequently, a key aspect of recent SDP research has been how to bridge the gap between analytical models and practical considerations:

From practice to modeling. Network technology should be complemented with sound economics, while analytical models should account for the real-

Figure 1. Components of the TDP system: (a) control-feedback loop of dynamic TDP; (b) functionality separation between user-side and operator-side devices.



world constraints imposed by existing technical and practical operational considerations, including parameter measurability, data granularity, solution scalability and complexity, integration and deployment feasibility, regulatory requirements; and

From modeling to practice. Analytical models should provide guidance on the economics of bandwidth pricing and policy, as well as on the architecture that will implement the models in operational networks. The prototypes developed must then be tested by deploying them in the wild.

To incorporate this interplay in SDP, researchers should take a holistic approach bringing together ideas from economics, networking, information systems, and human-computer interaction. In general, SDP research involves three stages, each with interesting research questions essential to realizing a new pricing scheme, as in the following:

Analytical modeling. How can ISPs create economic models for computing optimized prices or incentives they are willing to offer and users are willing to accept for modifying their bandwidth-consumption behavior in the desired manner?;

System development. How can ISPs

develop scalable systems and related signaling protocols for these incentive mechanisms? How should the various required functionalities (such as congestion measurement, traffic profiling, and delay-optimal scheduling) be divided among network backend and end-user devices?; and

Field trials. How can ISPs use human-computer interaction principles to design end-user interfaces so users understand and respond to the pricing signals?

A series of new initiatives since 2012, including the Internet Architecture Board's Workshop on Internet Technology Adoption & Transition,¹⁴ Smart Data Pricing Forum,²³ and Internet Research Task Force Working Group on Global Access to the Internet for All,¹⁷ provide momentum for such interdisciplinary collaborations. This article discusses two complementary research themes in SDP—TDP and traffic offloading—that aim to reduce network congestion by giving users incentives and mechanisms to “shift” their use to less-congested times or to other supplementary networks (such as Wi-Fi). We also consider SDP's implications for the Internet's long-term sustainability and accessibility to a wider user population.

Case 1. Time-Dependent Pricing

Much of the need for expanded network capacity is due to large peak demand created by users' simultaneous consumption of data; Cisco's visual networking index predicts peak hour traffic will grow at 64% CAGR.⁴ Yet, as Odlyzko et al.¹⁶ correctly said, ISP attempts to slow this growth by transitioning from flat-rate to usage-based pricing are unlikely to solve the problem. To discourage large peak demand, prices should have a temporal component or vary over different times of the day, as in TDP. Only then will users be incentivized to spread their demand over time, improving network resource utilization by reducing the peaks and filling in the valley periods. Writing in 2008 at Google's Public Policy blog,^a Vinton G. Cerf advocated a similar view, saying, “Network Management also should be narrowly tailored, with bandwidth constraints aimed essentially at times of actual congestion.” He cautioned against ISPs rushing to change their pricing, favoring a more detailed study on the efficacy of such dynamic

a <http://googlepublicpolicy.blogspot.com/2008/08/whats-reasonable-approach-for-managing.html>

or time-varying pricing mechanisms. The work we present here is one such attempt to understand the efficacy and feasibility of TDP for mobile data.

The telecommunications industry has long used dynamic TDP plans for voice calls as a response to demand variability in call volume by adjusting users' prices and incentives. However, dynamic pricing plans for data traffic, in spite of their theoretical potential to make resource allocation much more efficient, have remained largely unrealized in the global market, possibly due to the gap between the large body of analytical work on the topic and the lack of functional prototypes implementing these ideas.

The extensive academic literature on dynamic pricing theories²¹ includes responsive pricing (setting prices so as to keep user demand under a certain threshold), proportional fairness pricing (setting prices to optimize a proportional fairness criterion on the amount of bandwidth allocated to different users), priority pricing (explicitly accounting for QoS by allowing users to pay less by accepting a longer delay at congested times), and "smart market" auction pricing (deciding whether to admit a packet into the network at congested times based on the user-specified bid attached to that packet). But such schemes face two practical challenges: users prefer flat rates over the uncertainties associated with near-real-time price fluctuations,^{15,22} and users are often reluctant to delegate price bidding or traffic scheduling to automated agents, preferring the psychological assurance of manual control despite the greater convenience of automation.²² For users to be comfortable with dynamic pricing for data, ISPs must be willing to provide guarantees on available future incentives or prices, design intuitive user interfaces to aid manual decision making, and demonstrate the underlying system's feasibility with a proof-of-concept prototype.

Practice to models. To address the practical issues within an analytical framework, we have explored an alternative pricing approach—dynamic day-ahead TDP⁹—in which the ISP computes hourly prices one day in advance and advertises them to all users. The provider continues to compute new prices to maintain a sliding one-day



To discourage large peak demand, prices should have a temporal component or vary over different times of the day, as in TDP.



window of announced prices; that is, a new price point for the 24th hour is computed and announced every hour. Users thus receive day-ahead price guarantees while being rewarded for shifting some portion of their data usage (such as non-critical traffic) according to the announced price points. ISPs can measure changes in usage volume at different hours of the day in response to the given set of prices and estimate users' willingness to shift different types of traffic, which is in turn used to compute the next set of optimized prices. These prices are computed through a convex optimization formulation that minimizes the provider's total cost of overshooting capacity and the cost of providing these incentives to users while accounting for current estimates of users' willingness to shift traffic, temporal variations in usage volume, and capacity constraints.⁹ The framework thus requires a control-feedback loop between ISPs and users (see Figure 1a).

In addition to considering users' psychological preference for certainty regarding future price points, TDP must accommodate technological and regulatory concerns. In order to preserve users' privacy, our formulation does not require ISPs track each individual user's usage pattern, precluding the need for any deep packet inspection to realize this pricing scheme. The formulation remains computationally efficient as the number of network users grows; the model in Ha et al.⁹ implicitly assigns the aggregate traffic from all users and applications into virtual traffic classes characterized by different delay and price-sensitivity estimates. It then accounts for heterogeneity across users by separately modeling the probabilistic deferral behavior for each traffic class. This optimization model avoids utility functions, which can be difficult to measure quantitatively. Instead, all parameters (such as changes in usage volume in each period in response to prices) can be measured directly (such as users' price and delay sensitivities for different traffic classes) or estimated by the ISP without compromising the system's scalability or user privacy.

Models to practice. Figure 1b outlines the system we use to realize this day-ahead TDP for mobile data; the aggregate traffic measurement, user sensitivity estimation, and price computa-

tion engines are on the ISP side, while a mobile application that communicates with the pricing engine is on each end-user device. The primary purpose of the mobile application is to give users information on available price points, but it also has optional features like usage monitoring and alerts, as well as an auto-pilot mode for automated scheduling of applications based on available prices and user-specified delay sensitivities. For privacy reasons, such scheduling information remains in the app but is not communicated to the ISP. For more on the user-interface design, efficacy, and user response to these features, see Sen et al.²²

We developed and tested a prototype of this system over eight months in 2012 in different phases of a randomized field experiment in Princeton, NJ, with 50 mobile users on the AT&T network. In it, we effectively became a resale ISP offering this data plan with day-ahead dynamic prices to the trial participants. We separated the 3G traffic of the participants from that of other customers using an access-point-name setup, tunneling participants' 3G traffic from the ISP's core network into lab servers (see Figure 2). The participants installed the TDP mobile application on their iOS devices. Wi-Fi use, voice calls, and text messaging were not included in the trial traffic, as these services do not count toward 3G data caps.

The mobile application running on users' devices displayed the prices/discounts available for the next 24 hours in a color-coded format (see Figure 3a). Each price is color-coded by its discount rate (such as red for 0%–10%, orange for 11%–19%, yellow for 20%–29%, and

green for >30%). Users could view their usage history superimposed over the offered prices to visualize how much they spent and saved on data. Additionally, to help users save money, the app, called TUBE (a free version with only traffic-monitoring features, called DataWiz, is available in both iOS and Android appstores; see <http://scenic.princeton.edu/datawiz/>), provided an interface for users to see their top five bandwidth-consuming applications, set alerts, and configure their weekly budgets and app delay sensitivities for automatic scheduling, as in Figure 3b–3e.

Using the analytical model in Ha et al.⁹ we offered optimized day-ahead time-varying price discounts on the baseline price of \$10/GB to all users. We found users did shift their traffic from high-price to low-price periods under TDP. Overall data use by iPad users decreased by 10.1% in high-price periods and increased by 15.7% in low-price periods; that is, for most users, average use decreased in high-price periods relative to average use at the same hours before the trial.⁹ Additionally, by focusing on use in consecutive periods, where discounts differed by only 1% but the colors of the price indicator bars were different (such as comparing usage volumes in a yellow price period with a 29% discount to the following green price period with a 30% discount), we found a significant change in use, even though the absolute percent change in discount was only 1%. This resulting changed behavior indicates users paid more attention to the color-coding than to the actual value of the price discounts, emphasizing the need for careful and intuitive

user interface design to ensure users are able to understand and respond to pricing signals.²²

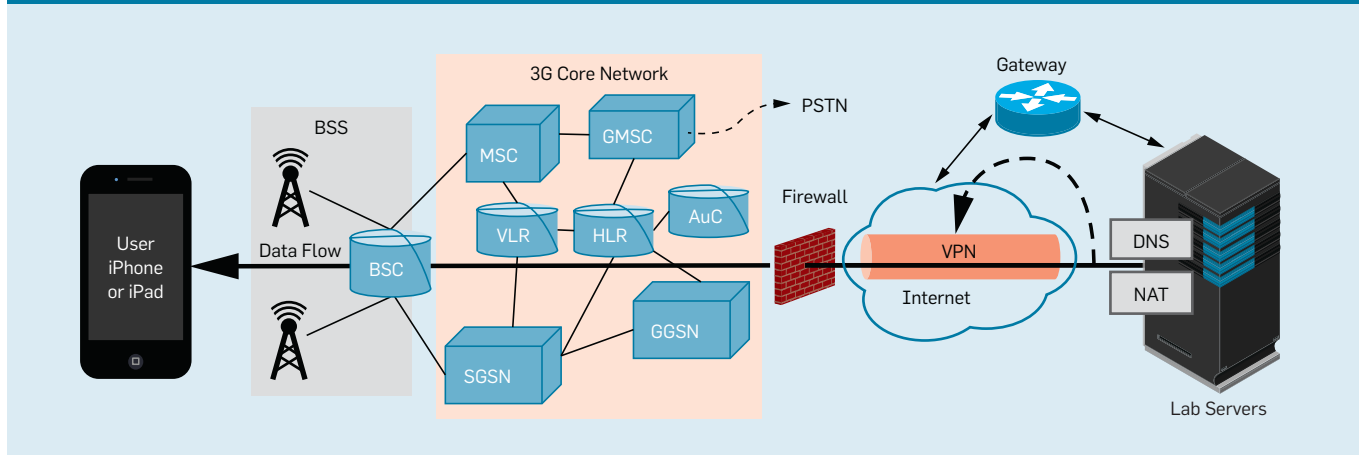
With optimized day-ahead time-dependent prices, resource utilization at off-peak hours nearly doubled, indicating TDP can also improve utilization of network capacity by flattening and distributing demand over different times of the day. The resulting maximum observed daily peak-to-average ratio (PAR) decreased by 30% with dynamic day-ahead TDP, and approximately 20% of the PARs from the pre-trial period were greater than the maximum PAR with TDP. Although the population size of our field trial was somewhat limited due to the complexity of actually conducting it, the results are promising, demonstrating it is possible to not only operationalize dynamic pricing for mobile data but also to use time-varying prices to change user behavior.

TDP can also be offered as time-dependent sponsored content, whereby consumers do not see a price fluctuation, but the valley discount is reflected in the price sponsoring parties (such as content providers and enterprises) pay to an ISP.

Case 2. Traffic Offloading

In addition to shifting demand from peak to valley time periods, ISPs can alleviate congestion by shifting demand off their cellular networks onto supplementary networks like Wi-Fi or femtocells, a process known as “traffic offloading.” Many ISPs encourage offloading by selling bundles of base and supplementary technologies; for example, the French telecommunications company Orange offers a £2 bundle of 3G and

Figure 2. Field trial setup for dynamic day-ahead, time-dependent usage-based pricing.

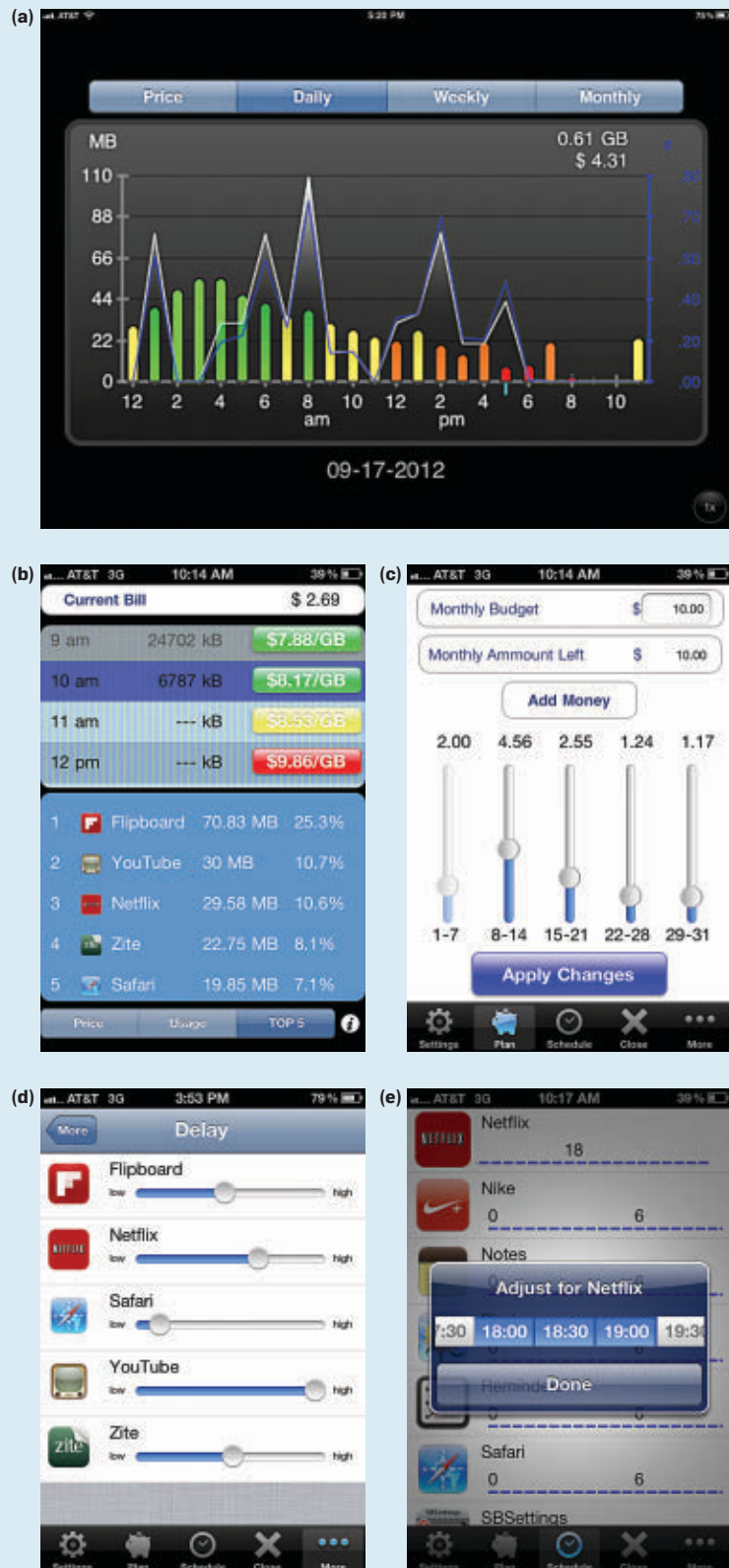


Wi-Fi hotspot access.¹³ Conventional wisdom says increasing the access price of the base technology (such as 3G or 4G) will encourage users to purchase the bundle and offload more to the supplementary network. This seems to be a direction many service providers are pursuing today with various penalty mechanisms on their base technology. But this strategy does not account for the fact the supplementary network can itself become more congested as more and more users offload their traffic, potentially making it less attractive to users. That is, there is a complex interaction among the prices of the base and supplementary technologies, the relative network congestion externalities of these two technologies, and the coverage area of the supplementary technology. Economic models and their practical realization in field trials can help ISPs design more effective offloading mechanisms. Here, we discuss related results from our studies that focus on the theory, as well as implementation, of ideas that can improve offloading performance and make it easier for users to make offloading decisions.

Practice to models. Understanding how users decide to adopt the base technology's network or a bundle of base and supplementary technologies, as well as deriving the resulting equilibrium and transient market outcomes, requires analytical models that incorporate practical issues like congestion on both networks and the coverage area of supplementary networks. Sen et al.¹⁸ introduced a model to study the dynamics of competition between two generic network technologies with cross-network externalities in the presence of network gateways or converters. In Joe-Wong et al.,¹³ we extended this framework to develop an analytical model in which users individually make their adoption decisions based on several factors (such as the technologies' intrinsic qualities, users' heterogeneity in the evaluation of these qualities, negative congestion externalities from the presence of other subscribers on the technologies, and access rates charged by an ISP).

Using the analytical model introduced by Joe-Wong et al.,¹³ we studied how user-level decisions translate into aggregate adoption dynamics and

Figure 3. Graphical user interfaces of the mobile app used in the TDP trial: (a) landscape view of superimposed price and usage history by day, week, and month; (b) view of the top-five bandwidth-consuming apps in the bottom split-screen; (c) weekly budget adjustment screen; (d) app-delay sensitivity settings screen; and (e) app-specific temporal blocking in parental control.



characterize the equilibrium outcomes for different system parameters. The model reveals seemingly intuitive strategies can sometimes have unintended consequences for congestion on the base network technology; for example, increasing the coverage area of the supplemental technology can increase traffic on the supplemental network, motivating some users to drop the bundled service altogether and use only the base technology. Likewise, increasing the base technology's access price can cause some users to find the base technology's coverage and relative network congestion conditions do not offset the decrease in their utility from the base price increase. These users are then motivated to drop the bundled service and use just the base technology, thus contributing to an increase in congestion on that network. Careful analysis using such economic models can prove very useful in providing guidance and insight into potential outcomes due to network pricing and policy changes.

Models to practice. While analytical frameworks like the one we discuss here can provide insight into users' long-term adoption behaviors (such as equilibrium and stability), users' minute-to-minute offloading decisions reflect their more immediate concerns. Users face a three-way trade-off among cost, throughput, and delay; while they can save money and receive greater throughput by waiting for Wi-Fi access, they may not want to wait for certain critical applications. To navigate it, we propose a

practical, cost-aware Wi-Fi offloading system called Adaptive bandwidth Management through User-Empowerment, or AMUSE,¹⁰ a user-centric tool that learns users' behavior and mobility patterns to help decide which applications to offload to what times of day, enabling users to stay under their data caps. In doing so, it uses a utility-optimization algorithm to decide if and how much 3G bandwidth must be allocated to each application at any moment of the day based on the user's available budget, application delay sensitivities, and input from prediction algorithms regarding the user's demand patterns and Wi-Fi availability in the future.

To implement this allocation in practice, AMUSE uses a receiver-side transmission control protocol (TCP) bandwidth control algorithm that enforces each application's assigned download rate by controlling the TCP advertisement window on the user side. The algorithm is thus fully contained on end-user devices and does not require modification on the TCP server side, making it suitable for real-world deployment (see Figure 4).

An AMUSE system implementation on Windows 7 tablets, when tested with simulated user behavior based on 3G and Wi-Fi usage and availability data from a field study of 37 mobile users, showed other offloading algorithms yield 14% and 27% less user utilities than AMUSE for light and heavy users, respectively.¹⁰ Intelligently managing users' competing in-

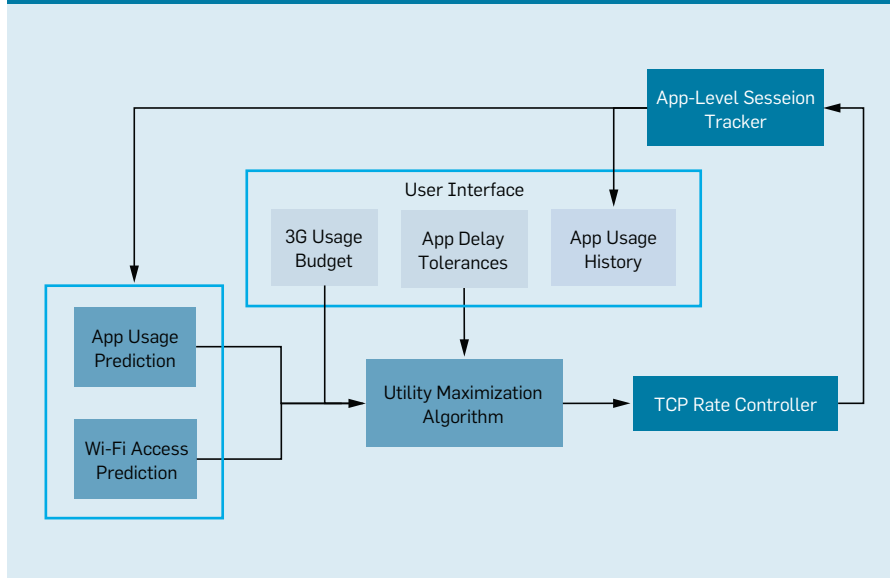
terests for cost, throughput, and delay via user-side management tools can thus facilitate user decisions, leading to offloading benefits that could be further improved when integrated with ISP-centric approaches (such as ad hoc authorization for offloading to hidden Wi-Fi access points or market-based solutions for offloading from base stations to third-party-owned Wi-Fi access points and femtocells).^{3,6,11} Such measures to push some control from the network core out to end users complement previous and existing efforts of the networking research community, as with Briscoe et al.²

Win-Win Solution

The two approaches to managing network congestion discussed earlier, namely realizations of TDP and delaying optimal traffic offloading, can help create a financial win-win solution for ISPs and their users. ISPs benefit through reduced peak congestion, and users have more choices and technologies to help them save on their monthly bills.

Research on network resource pricing also has implications for bridging the digital divide between those who can and those who cannot access the Internet regularly. Rural local exchange carriers (RLECs) often suffer from congestion in their wired networks due to the persistence of the middle-mile problem. Although the cost of middle-mile bandwidth has declined over the years due to increased demand needed to fill the middle mile, the bandwidth requirements of home users have also increased sharply. The cost of middle-mile upgrades to meet the Federal Communication Commission's target speed of 4Mbps will be substantial and is a barrier to digital expansion in rural areas.⁷ New access-pricing mechanisms like TDP can help reduce middle-mile investment costs by reducing RLEC peak-capacity provisioning or leasing needs and improving resource utilization in the valley periods. Providers can thus match their prices to the cost of delivery while also creating incentives for light users to adopt broadband services. Instead of being charged by the volume of data consumed, users can save on their monthly bill by choosing "when" to consume the data. An extension of this idea can be used to create

Figure 4. System modules and interaction across AMUSE components.



ultra-affordable data plans for delay-tolerant users that allow automated opportunistic access to the network only when large discounts are available or when the network is lightly loaded. Implementation of such dynamic TDP and opportunistic offloading schemes can help service providers better utilize their available resources and increase Internet adoption by being financially attractive to even light data users.

Content providers have also tried to bridge the digital divide with zero-rating, app-based pricing and sponsored content,^{1,12} all of which open further interesting questions for SDP research. For example, how should platforms for sponsored content or zero-rating be made compatible with Net neutrality regulations, as in the case of Facebook's Internet.org initiative? And how will such subsidized plans affect broadband adoption and network congestion? Much of the current SDP research focuses on such questions.

Conclusion

As demand for bandwidth grows in both wireless and wired networks, ISPs are pursuing penalty mechanisms like deliberately slowing traffic, capping, overage charges, and usage-based fees to manage their available network capacity. But such measures are arguably suboptimal and even harmful to the Internet ecosystem. In contrast, ideas from economics can help design incentives and pricing policies that are beneficial to both service providers and their users. Although many analytical models for pricing-based network management have been proposed, their implementation has begun only recently. Tackling today's challenges requires not only developing analytical models that incorporate practical concerns (such as measurability, scalability, privacy, and user behavior) but also demonstrate efficacy and feasibility through prototypes and field trials.

Here, we have focused on "shifting" demand through two complementary efforts that aim to alleviate network congestion by creating incentives and mechanisms to modify user behavior or shift demand to less congested times (through TDP) or to a supplementary network (through delay-optimized traffic offloading). The results indicate such measures, if im-

plemented with intuitive designs, can help ISPs better monetize and manage their network capacity while empowering users with more options to avoid hefty overage fees. With implicit pricing signals and automated decision making, these solutions can be readily adapted to emerging machine-to-machine and Internet of Things applications. These approaches also have positive implications for making Internet access affordable for many more users, thereby contributing to bridging the digital divide.

The scope of SDP is much broader than the two particular cases we outlined here. The complementary questions "Who?" and "What?" are gaining traction, in addition to "How?" on which we focused. Researchers must move beyond models, trials, and testing of prototypes and consider integration with existing infrastructure, the design of pricing and signaling protocols, and regulatory concerns. SDP research will help bring together academic researchers, network providers, content providers, the e-commerce industry, and policymakers to design and deploy new mechanisms to help ensure the sustainable growth of the Internet, mobile, and content markets.

Acknowledgments

We thank many collaborators in academic institutions and industry, especially Krishan Sabnani of Bell Labs for his encouragement and feedback on this article. 

References

- Andrews, M., Riemann, M., Wang, Q., and Ozen, U. Economic models of sponsored content in wireless networks with uncertain demand. In *Proceedings of the IEEE Conference on Computer Communications Workshops* (Turin, Italy, Apr. 19). IEEE, New York, 2013, 345–350.
- Briscoe, B., Darlagiannis, V., Heckman, O., Oliver, H., Siris, V., Songhurst, D., and Stiller, B. A market managed multi-service Internet. *Computer Communications* 26, 1 (Mar. 2003), 404–414.
- Cheung, M. and Huang, J. Optimal delayed Wi-Fi offloading. In *Proceedings of the 11th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks* (Tsukuba Science City, Japan, May 13–17). IEEE, 2013.
- Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019*; http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html
- El-Sayed, M., Mukhopadhyay, A., Urrutia-Valdés, C., and Zhao, Z.J. Mobile data explosion: Monetizing the opportunity through dynamic policies and QoS pipes. *Bell Labs Technical Journal* 16, 2 (Sept. 2011), 79–100.
- Gao, L., Iosifidis, G., Huang, J., Tassiulas, L., and Li, D. Bargaining-based mobile data offloading. *IEEE Journal on Selected Areas in Communications* 32, 6 (June 2014), 1114–1125.
- Glass, V., Stefanova, S., and Dibelka, R. Customer price

- sensitivity to broadband service speed: What are the implications for public policy? In *Smart Data Pricing*, S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, M., Eds. John Wiley & Sons, Inc., Hoboken, NJ, 2014.
- Gupta, A., Stahl, D.O., and Whinston, A.B. The economics of network management. *Commun. ACM* 42, 9 (Sept. 1999), 57–63.
 - Ha, S., Sen, S., Joe-Wong, C., Im, Y., and Chiang, M. TUBE: Time dependent pricing for mobile data. *ACM SIGCOMM Computer Communication Review* 42, 4 (Oct. 2012), 247–258.
 - Im, Y., Joe-Wong, C., Ha, S., Sen, S., Kwon, T., and Chiang, M. AMUSE: Empowering users for cost-aware offloading with throughput-delay trade-offs. *IEEE Transactions on Mobile Computing* (forthcoming, 2015).
 - Iosifidis, G., Gao, L., Huang, J., and Tassiulas, L. A double auction mechanism for mobile data offloading markets. *IEEE/ACM Transactions on Networking* (Sept. 2014).
 - Joe-Wong, C., Ha, S., and Chiang, M. Sponsoring mobile data: An economic analysis of the impact on users and content providers. In *Proceedings of the 34th IEEE International Conference on Computer Communications* (Hong Kong, China, Apr. 26–May 1). IEEE, New York, 2015.
 - Joe-Wong, C., Sen, S., and Ha, S. Offering supplementary network technologies: Adoption behavior and offloading benefits. *IEEE/ACM Transactions on Networking* 23, 2 (Feb. 2014).
 - Lear, E. *Report from the IAB Workshop on Internet Technology Adoption and Transition*. IETF Network Working Group, Internet-Draft, May 19, 2014; <http://tools.ietf.org/html/draft-ietf-iab-itat-report-03>
 - Odlyzko, A. Will smart pricing finally take off? In *Smart Data Pricing*, S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, Eds. John Wiley & Sons, Inc., Hoboken, NJ, 2014.
 - Odlyzko, A., St. Arnaud, B., Stallman, E., and Weinberg, M. *Know Your Limits: Considering the Role of Data Caps and Usage-Based Billing in Internet Access Service*. White Paper. Public Knowledge, Washington, D.C., May 2012.
 - Sathiaselvan, A. Researching Global Access to the Internet for All (GAIA). *The IETF Journal*. Internet Society IRTF Working Group, Reston, VA, 2014.
 - Sen, S., Jin, Y., Guerin, R., and Hosanagar, K. Modeling the dynamics of network technology adoption and the role of converters. *IEEE/ACM Transactions on Networking* 18, 6 (Dec. 2010), 1793–1805.
 - Sen, S., Joe-Wong, C., Ha, S., and Chiang, M. Incentivizing time-shifting of data: A survey of time-dependent pricing for Internet access. *IEEE Communications Magazine* 50, 11 (Nov. 2012), 91–99.
 - Sen, S., Joe-Wong, C., Ha, S., and Chiang, M. Smart data pricing: Economic solutions to network congestion. *ACM SIGCOMM eBook on Recent Advances in Networking*, ACM SIGCOMM, New York, 2013.
 - Sen, S., Joe-Wong, C., Ha, S., and Chiang, M. A survey of broadband data pricing: Past proposals, current plans, and future trends. *ACM Computing Surveys* 46, 2 (Nov. 2013), 1–37.
 - Sen, S., Joe-Wong, C., Ha, S., and Chiang, M. When the price is right: Enabling time-dependent pricing of broadband data. In *Proceedings of ACM SIGCHI* (Paris, France, Apr. 27–May 2). ACM Press, New York, 2013, 2477–2486.
 - Smart Data Pricing Forum; <http://scenic.princeton.edu/sdp/>
 - Smart Data Pricing Research; <http://sdpresearch.org/>

Soumya Sen (ssen@umn.edu) is an assistant professor in the Department of Information Systems & Decision Sciences in the Carlson School of Management at the University of Minnesota, Minneapolis.

Carlee Joe-Wong (cjoe@math.princeton.edu) is a Ph.D. candidate in the Program in Applied and Computational Mathematics at Princeton University, Princeton, NJ.

Sangtae Ha (sangtae.ha@colorado.edu) is an assistant professor in the Department of Computer Science with a joint appointment in the Interdisciplinary Telecommunications Program at the University of Colorado at Boulder.

Mung Chiang (chiangm@princeton.edu) is a professor of electrical engineering and an affiliated faculty member in applied and computational mathematics and in computer science at Princeton University, Princeton, NJ.

Copyright held by the authors.
Publication rights licensed to ACM. \$15.00

The connection between online communication and psychological well-being depends on whom you are communicating with.

BY ROBERT KRAUT AND MOIRA BURKE

Internet Use and Psychological Well-Being: Effects of Activity and Audience

PEOPLE AROUND THE world have incorporated the Internet into their daily lives, using it to find information, communicate with friends and family, shop, play games, and pass the time. How does it affect our well-being?

The media frequently posit the Internet is changing our social lives, warning of a “Sad, Lonely World Discovered in Cyberspace”¹³ or asking “Is Facebook Making Us Lonely?”¹⁶ Even Pope Benedict warned that “virtual contact cannot and must not take the place of direct human contact.”¹⁹ A major reason for this fascination with new technologies and social relationships is that these relationships have important consequences for both physical and psychological health (for recent reviews, see Callaghan⁸ and Thoits³¹).

These concerns have been reflected in the scholarly literature as well. For example, prior research suggests

Internet use influences the amount of interpersonal communication people engage in, the partners with whom

» key insights

- Talking with close friends online is linked to improvements in social support, depression, and other measures of well-being, but talking with strangers and reading about acquaintances are not.
- Readers should be skeptical of cross-sectional and survey-based studies linking well-being to Internet use.
- Instead, experiments or longitudinal designs pairing surveys with log data provide more reliable insights.
- Human agency is key: The effect of technology on our lives depends on how we use it, what we talk about, and whom we talk to.





they communicate, and the quality of the communication episodes in which they participate (for example, Cummings⁹). Much research shows people communicate online primarily with people with whom they communicate offline—their relatively strong social ties—and their online communication supplements rather than replaces their offline communication.^{10,20} For example, in early research, Bikman and Eveland³ and Hampton and Wellman¹² showed how everyday use of computer networks increased peoples' recognition of those they communicated with online. If so, one would expect that increased use of the Internet, especially for communication with strong ties, would increase

the amount of social support people have available to them and thereby improve the downstream consequences associated with social support, for example, by reducing depression, stress, and loneliness, and improving mood and even physical health.^{21,28}

However, other research suggests the quality of the communication people have online is impoverished and less valuable than time spent in spoken communication, either face-to-face or by phone (for example, Cummings⁹). Moreover, the relatively low cost of online communication and its insensitivity to distance may encourage people to differentially increase their communication with weak ties or strangers.²⁹

Communication with weak ties and strangers is unlikely to have the same psychological benefits as communication with stronger ties. While it is possible for friendships initiated online to develop into close relationships over time,^{17,13} the preponderance of ties in our online social networks are acquaintances and other weak ties.²⁹ If substitution is occurring, with communication with weak ties crowding out communication with stronger ones, increased use of the Internet could potentially decrease the social support people have available and harm their psychological well-being.

Our research over the past 15 years has attempted to determine how every-

day use of the Internet influences users' psychological well-being.

A Methodological Critique

The answers to questions about the impact of Internet use on well-being are unclear for two reasons. First, while questions about causation and change demand longitudinal or experimental data,²⁷ typical research in this area uses cross-sectional survey techniques. Cross-sectional analysis can produce misleading conclusions. For example, in one of our early studies with the Pew Internet and American Life Project, cross-sectional results showed Internet communication with a particular other person was strongly associated with phone and face-to-face communication with that person. In contrast, longitudinal analyses of the same data showed greater Internet use during a time period was associated with declines in in-person visits to the partner.²⁵ One reason that conclusions from cross-sectional analyses should not be trusted is they confound predispositions for using technology with its

effects.²⁶ When trying to understand the causal connection between Internet use and psychological well-being, the problem is initial well-being or relatively stable social characteristics, such as social competence or extraversion, can influence both how people use the Internet and their social and psychological well-being. For example, in longitudinal work, Mikami and colleagues demonstrated that adolescents' social competence and psychological health (for example, depression) predicted their online activities seven years later, including their online social network size and the number of their ties offering verbal social support.¹⁸ In the face of this evidence showing adolescents' depression predicts their adult Internet use, it is then difficult to argue using only cross-sectional studies that Internet use causes depression.

The second major problem is that much of the extant research fails to differentiate types of Internet use and use by different types of people, and thus fails to provide insight into the mechanisms at play. Early research simply

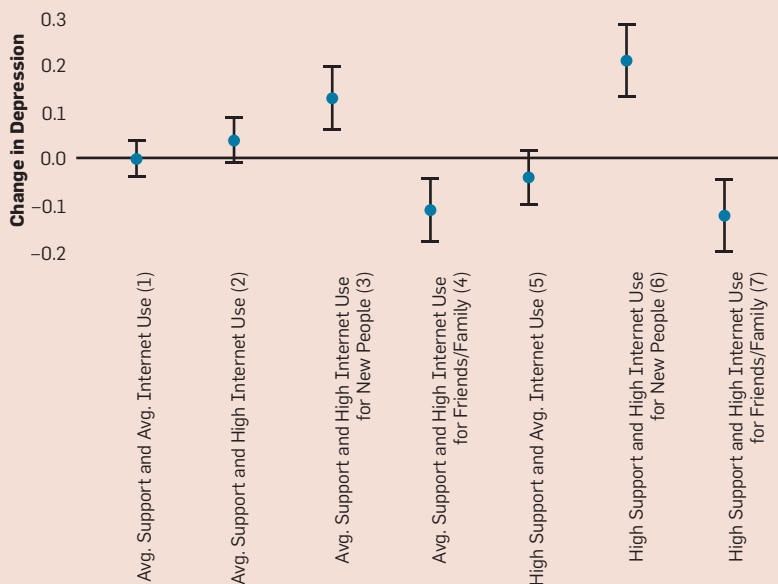
differentiated Internet users from non-users,¹⁴ and more recent research often examines the amount people use the Internet without comparing qualitatively different types of use.²⁸ Yet research that does differentiate types of use suggests that different uses of the Internet may have distinct associations with well-being for different kinds of people. For example, over a one-year period, depression increased among Dutch adolescents who used the Internet for Web surfing, but declined among those who used it for chatting with friends; moreover, these associations occurred only among adolescents who received little social support from their closest friends.²³ In another study, instant messaging use predicted increases in depression, but email use did not;³² the increase in depression may have been limited to instant messaging because it tends to be substantively shorter and more superficial than email.⁹ Burke and colleagues found that exchanging messages with one's Facebook ties was associated with increases in social capital while simply reading news about them was not; however, individuals with lower social communication skills seemed to benefit from both types of activities.⁶ Understanding the complicated relationship between Internet use and well-being requires an examination of the Internet's myriad forms and differences among users that both drive online activity and moderate its effects.

Differentiating Internet Uses in Longitudinal Research

Research from our lab over the past 15 years has attempted to overcome these challenges. It employs a common methodology, using lagged dependent variable linear regression on longitudinal data to examine how different uses of the Internet during a time interval predict changes from the beginning to the end of the interval in social and psychological outcomes such as social support, depression, and loneliness. This research is correlational and not as powerful as random assignment experiments in demonstrating the causal impact of an intervention, for example, Shaw.²⁴ However, random assignment is generally impractical if the goal is to identify the long-term impact of Internet use, because in developed

Figure 1. Relative changes in self-reported depression symptoms (CES-D) over a six-month period predicted by initial social support and types of Internet use.

Zero on the y-axis represents changes in depression among survey respondents with average levels of social support and Internet use. Each point shows the relative change in depression in standard deviation units and its standard error associated with having more initial social support and/or use of the Internet to communicate with friends and family or with new people. "High" means a standard deviation more than average of a predictor variable (adapted from Bessière et al.¹).



countries where Internet use is already pervasive, few people would willingly agree to be randomly assigned to forgo Internet access use for long periods of time. A panel design mitigates many threats to making causal inferences from correlational data. By observing the same individuals multiple times, by knowing the temporal ordering of the intervention (Internet activities) and the outcomes (well-being), and by controlling for initial well-being, this research design minimizes the possibility that self-selection to the intervention and reverse causation account for associations between Internet use and changes in well-being.

Our first research on this topic was conducted in 1997, the dawn of the Internet era for most people who did not work in universities or research laboratories.¹⁵ In this research, families with high school-aged children were given computers and Internet access. Their email and Web use was monitored, and family members completed three surveys over one year period. Results indicated the more people used the Internet, independent of the way they used it, the more their depression increased and social support and other measures of psychological well-being declined. Although depressed people typically engage in less social contact than do less-depressed people,¹⁸ initial depression was statistically controlled in the analysis and Internet use was measured subsequent to the initial measures of depression; therefore variations in initial depression cannot account for these results. Moreover, this research tested for and found no evidence for reverse causation, where early measures of social support, depression, loneliness, or stress predicted subsequent Internet use.

Even though this research differentiated asocial Internet use (Web browsing) from social uses (online communication via email and participation in online groups), it did not differentiate communication with stronger or weaker ties. Indeed, during this early era, most participants' close ties were not yet online. Therefore, use of the Internet for any purpose, even highly social, interpersonal communication, may have presented an opportunity cost, shifting people's time and attention away from more fruitful offline com-



Understanding the complicated relationship between Internet use and well-being requires an examination of the Internet's myriad forms and differences among users that both drive online activity and moderate its effects.



munication with closer friends. Or, other changes in unmeasured factors, such as participants' satisfaction with their relationships, may have driven increases in both time spent online and depression.

We conducted a replication with a new sample, when a larger fraction of people's social networks were Internet users and when the Internet offered a wider variety of services.¹ In this research started in late 2000, 922 respondents from a national sample of U.S. households were contacted using random digit dialing. In three surveys spread over a year, respondents described on multi-item scales how frequently they used the Internet for different purposes: communicating with friends and family; communicating in online groups and to meet people; retrieving and using information; seeking entertainment or escape; shopping; and acquiring health information or talking about health. Confirmatory factor analyses demonstrated that differentiating these types of Internet use fit the data better than a model that assumed use reflected a single dimension ranging from light use to heavy use.

Figure 1 illustrates how Internet use was associated with changes in well-being. It shows changes in self-reported depression over a six-month period among people who initially had different levels of social resources available and who used the Internet for different purposes compared to people with an average amount of social support and of Internet use. The first point shows the change among people who initially had average levels of social support and who used the Internet an average amount across the different purposes; it has been normalized to zero for comparison purposes. Point 2 shows the more people used the Internet overall, the more their self-reported levels of depression increased compared to the base rate, although the increase was not statistically significant during the first six-month interval reported in Bessière¹ ($p < 0.10$). However, additional data collection in a six-month follow-up with the same respondents showed the increase in depression with overall Internet use was statistically significant over the year-long period ($p < .02$).²

Moreover, differentiating types of Internet use and communication partners clarified these results. Use of the Internet for finding information, for entertainment, or for commerce was not associated with greater changes in depression than the base rate ($ps < .25$), but use for communication was. Because the online population had grown so much, the research could differentiate online communication with close ties from communication with weaker ties and strangers. More use of the Internet to communicate with weaker ties (in particular, to meet new people and to hang out in groups comprised primarily of strangers) was associated with increases in depression compared to the base rate (see point 3). In contrast, more online communication with friends and family was associated with declines in depression (point 4). Moreover, interaction results suggest a substitution effect. Respondents reported on their social resources on the initial questionnaire: their perceived social support, the number of friends with whom they regularly communi-

cated, and their extroversion. Compared to those who initially had fewer social resources, people who initially had greater social resources offline and used the Internet to communicate with strangers reported larger increases in depression than one would expect simply from their social resources or their Internet use alone (compare point 6 to points 3 and 5). In contrast, people who used the Internet to communicate more with friends and family had decreases in depression regardless of their initial social resources (compare point 7 with points 4 and 5).

This correlational data cannot determine why online communication with strangers and other weak ties was associated with increases in depression while communication with friends and family was associated with declines in depression. One possibility is that interactions with weak ties were simply less satisfying than interactions with family or friends. This explanation is suggested by other research we conducted, which asked survey respondents in a national sample to evalu-

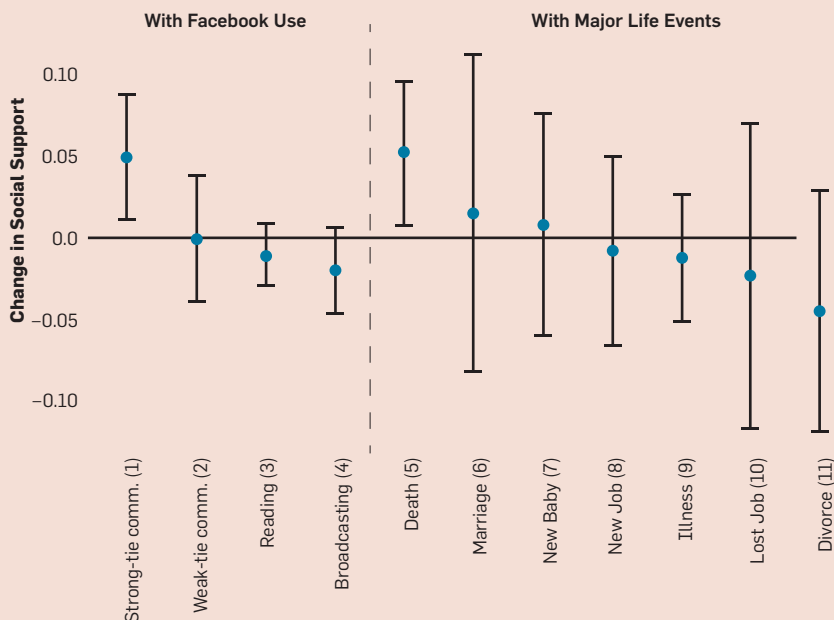
ate a specific online communication episode they had participated in the previous day.⁹ Another possibility is a substitution effect: interactions with weak ties reduced the amount of time and attention people had available to spend on richer and more valuable interactions with closer friends. This explanation is consistent with the finding that online communication with strangers was associated with increases in depression most strongly among individuals who had higher levels of offline social resources (social support) at the start of the study.

Our most recent research assessing how the Internet influences psychological well-being focused on the variety of uses people make of online social network sites, in particular Facebook.^{4,6,7} As of 2012, two-thirds of American Internet-using adults subscribed to an online social network, with Facebook being the most popular. Approximately 65% of Facebook's 1.4 billion monthly active users worldwide visit the site on a typical day.¹¹

The goal of this research was to assess how the well-being of Facebook users changed with different kinds of Facebook use and interaction with different partners. To conduct this analysis, we combined data from three surveys administered one month apart measuring social support, depression, and other aspects of psychological well-being with de-identified, aggregated counts of Facebook activity. The Facebook use data consisted of counts from server logs of online activity (for example, number of wall posts and comments posted and read, likes delivered and received, stories read, and photos viewed); no content was analyzed. Respondents reported how close they felt to up to eight of their Facebook friends, up to six of whom were close friends they identified (mean = 4.4 close ties) and the remaining ones randomly selected from their Facebook friend networks. They rated themselves as substantially closer to the ties they chose than to the randomly selected ties. Respondents' ratings of closeness were used to train a linear regression model to estimate their tie strength with each of their approximately 130 Facebook friends (see Burke⁵ for details of the estimation procedures). The large sample ($N = 1,927$ respondents communi-

Figure 2. Changes in self-reported social support over a month-long period associated with different uses of Facebook (left) and major life events (right).

The y-axis represents change in social support for an average Facebook user. Each point on the left shows the mean change in social support and its standard error associated with a standard-deviation increase in one of four uses of Facebook (communicating with strong ties, communicating with weak ties, reading others' broadcast content, and broadcasting one's own content). Each point on the right shows the association between changes in social support and major life events, such as a death in the family or getting married (adapted from Burke.⁵)



cating with over 2.4 million Facebook friends in total) and logs of participants' activity on Facebook allow us to differentiate types of communication: one-on-one exchanges such as wall posts or comments that were targeted at the recipient; reading friends' status updates that were broadcast widely rather than being tailored for a specific viewer; and broadcasting one's own content, such as status updates, game scores, and photos, out to a wider circle of friends. Confirmatory factor analyses demonstrated the validity of differentiating one-on-one communication from broadcast communication.

We hypothesized that one-on-one exchanges, by virtue of being tailored for a single recipient, would be more likely to increase social support, while reading broadcasts from a wider audience and sending them out would not, because they require less effort per recipient and are likely to deal with less intimate topics than messages meant for a single person. Using the tie-strength model, we were also able to distinguish whether the people they interacted with on Facebook were close friends or weaker acquaintances. Using a tie-strength threshold of 5 on a 7-point closeness scale to distinguish close from weak ties, 39.4% of participants' ties were considered "strong" and the median user had 38 strong ties ($M = 47$). The cutoff of 5 was both the mean and median tie strength for the ties participants selected as very close friends in the survey; however, the results we report here are substantively the same if we use a higher threshold and consider close friends to be ones with a threshold of 6 or 7 on the 7-point scale. Participants also reported recent major life events, such as the death of a family member or losing a job. We used these events as control variables in our models and as a baseline for understanding what constitutes a meaningful change in well-being.

Results paralleled our earlier studies. The relationship between Facebook use and well-being depended on how people used the site and with whom they communicated. As shown in Figure 2, receiving targeted, one-on-one communication such as private messages, wall posts, or comments from one's strong ties was associated with improvements in

perceived social support (see point 1), as well as happiness, self-reported health, depression, loneliness, negative affect and stress (all $ps < .05$). Reverse causation cannot account for these results; measures of initial psychological well-being did not predict changes in either respondents' communication with strong ties or the ties' communication directed at respondents. In contrast to communication with strong ties, communication with weaker ties (point 2), reading content such as status updates that were broadcast to a larger audience (point 3), or broadcasting one's own content (point 4) were not associated with these improvements in well-being. Moreover, the association of one-on-one communication with improvements in well-being was stronger for more substantive communication (for example, written comments) than for stylized, low-effort communication (for example, Facebook "likes").⁷

The effect sizes were large in comparison to the effects of other life events. For example, receiving a standard deviation more communication from strong ties—approximately 60 more comments—was associated with increases in perceived social support as large as those that occur following a death in the family, a time when the bereaved receive an outpouring of support and condolences from others (compare points 1 and 5 in Figure 2). The effect size for strong-tie communication was also comparable to the effect size for other major life events, like getting married, having a new baby, or losing one's job.⁶ In contrast, after accounting for this strong-tie communication, well-being did not improve with other Facebook activities, such as talking with weak ties, reading friends' broadcasts, or broadcasting content oneself.

Summary and Conclusion

Does penetration of the Internet into people's lives for connecting to other people, finding information, and entertainment have larger consequences, beyond directly supporting these tasks? The lessons from this literature review are both substantive and methodological.

In terms of substantive conclusions, research reviewed here is con-

sistent with a hypothesis that communicating online with close friends and family can have beneficial effects on psychological well-being as measured by declines in depression, loneliness and stress, and increases in perceived social support, mood, and life satisfaction. In contrast, many other uses of the Internet, including using the Internet for information, entertainment and communicating online with weaker ties, do not have similar, positive associations with psychological well-being. Indeed our earliest research showed that communication with strangers was associated with declines in psychological well-being.

Methodologically, one should be skeptical about conclusions drawn from cross-sectional research about its possible effects because preexisting differences in psychological well-being can shape how people use the Internet. Indeed, our research demonstrates different conclusions from cross-sectional and longitudinal analyses of the same data.²⁵ To better understand the relationship between Internet use and well-being, one needs longitudinal data and data that differentiates types of Internet use and classes of communication partners. Preexisting conditions like depression often cause withdrawal from social activities both online and offline; however, the analyses in this article control for respondents' initial well-being. Therefore differences in preexisting conditions like depression and its association with withdrawal from social interactions cannot account for the findings. We caution, however, that despite the advantages that derive from panel data, the analyses are correlational. Therefore, it is possible that *changes* during the measurement period in some third variable, like relocating, losing a job, or acquiring a serious illness, can shift both how people use the Internet and their psychological well-being.

Much has changed in the decades since we first began communicating on the Internet, and so in addition to being skeptical of cross-sectional, undifferentiated research, we must continuously reevaluate the impact the Internet has on our lives. For example, ubiquitous access via smartphones and other mobile devices has changed social norms of connectiv-

ity. The composition of our online networks has expanded to include grandparents and coworkers as well as many more strangers available on Twitter and other social media. In addition to text that dominated early online interpersonal communication, the genres of online interactions have expanded to include game-playing and the exchange of pictures and movies. Millennials have never known a world without the Internet. We can see detailed histories of our online interactions with Facebook friends and their relationships with others. There are many opportunities to investigate how these changes affect our well-being over the long term. However, our research over the past 15 years has demonstrated that, much like offline communication, the impact depends on the nature of the communication and with whom we are talking.

If use of the Internet does cause changes in psychological well-being, what is the nature of these “Internet effects”? Although evidence suggests the Internet, like print media and television before it, seems to have identifiable effects on psychological well-being, these associations do not imply a strong technological determinism. Human agency is key, because the technological effects depend upon how people decide to use technology. However, as with other daily-life activities, the way choices are framed and the effort involved in engaging in the activities are likely to bias the choices people make (see Thaler³⁰ for a fuller presentation of this argument). Even though people seem willing to invest more effort to communicate with closer ties than weaker ties and receive more benefits from interactions with them,²² modern information technology can change the effort needed to keep up with distant friends, to meet and have discussions with strangers, to have rich communication with specific ties or to have superficial interactions with acquaintances. These changes in effort can shift how people spend their time, which relationships they retain over time, and what they talk about. Thus, it is not the use of technology per se, but these decisions, which can be biased by technology, which directly influence psychological well-being. 

References

- Bessière, K., Kiesler, S., Kraut, R. and Boneva, B. Effects of Internet use and social resources on changes in depression. *Information, Communication & Society* 11, 1 (2008), 47–70.
- Bessière, K., Pressman, S., Kiesler, S. and Kraut, R.E. Effects of Internet use on health and depression: A longitudinal study. *J. Medical Internet Research* 12, 1 (2010), e6; <http://www.jmir.org/2010/2011/e2016/HTML>.
- Bikson, T.K. and Eveland, J.D. The interplay of work group structures and computer support. *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work*. J. Galegher, R.E. Kraut and et al. (Eds). Engleclides, Inc., Hillsdale, NJ, 1990, 245–290.
- Burke, M., Marlow, C. and Lento, T. Social network activity and social well-being. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*. ACM Press, New York, 2010, 1909–1012.
- Burke, M., Reading, writing, relationships: The impact of social network sites on relationships and well-being (Ph.D. Thesis, 2011). Carnegie Mellon University, Pittsburgh, PA.
- Burke, M., Kraut, R.E. and Marlow, C. Social capital on Facebook: Differentiating uses and users. In *Proceedings of CHI 2011: The ACM Conference on Human Factors in Computing Systems*. ACM, New York, 571–580.
- Burke, M. and Kraut, R. Using Facebook after losing a job: Differential benefits of strong and weak ties. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, NY, 1419–1430.
- Callaghan, P. and Morrissey, J. Social support and health: A review. *J. Advanced Nursing* 18, 2 (2008), 203–210.
- Cummings, J., Butler, B. and Kraut, R. The quality of online social relationships. *Commun. ACM* 45, 7 (July 2002), 103–108.
- Ellison, N.B., Steinfield, C. and Lampe, C. The benefits of Facebook ‘friends’: Social capital and college students’ use of online social network sites. *J. Computer-Mediated Communication* 12, 4 (2007), 143–168.
- Facebook, Inc. Company info. Facebook, Menlo Park, CA, 2015.
- Hampton, K. and Wellman, B. Neighboring in Netville: How the Internet supports community and social capital in a wired suburb. *City & Community* 2, 4 (2003), 277–311.
- Harmon, A. Sad, lonely world discovered in cyberspace. *New York Times*, (Aug. 30, 1998), section 1, 1.
- Katz, J.E. and Rice, R.E. *Social Consequences of Internet Use: Access, Involvement, and Interaction*. MIT Press, Cambridge, MA, 2002.
- Kraut, R.E., Patterson, M., Lundmark, V., Kiesler, S., Mukhopadhyay, T. and Scherlis, W. Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist* 53, 9 (1998), 1017–1031.
- Marche, S. Is Facebook making us lonely? *The Atlantic* (May 2012).
- McKenna, K., Green, A.S. and Gleason, M. Relationship formation on the Internet: What’s the big attraction? *J. Social Issues* 58, 1 (2002), 9–31.
- Mikami, A.Y., Szewdo, D.E., Allen, J.P., Evans, M.A. and Hare, A.L. Adolescent peer relationships and behavior problems predict young adults’ communication on social networking websites. *Developmental Psychology* 46, 1 (2010), 46–56.
- Pope Benedict XVI. Truth, proclamation and authenticity of life in the digital age. June 5, 2011.
- Raine, L., Boase, J., Horrigan, J. B. and Wellman, B. The Strength of Internet Ties. Pew Internet and American Life Project, Washington, D.C., 2006.
- Rains, S.A. and Young, V. A meta-analysis of research on formal computer-mediated support groups: Examining group characteristics and health outcomes. *Human Communication Research* 35, 3 (2009), 309–336.
- Roberts, S.G. and Dunbar, R.I. The costs of family and friends: An 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior* 32, 3 (2011), 186–197.
- Selfhout, M.H., Branje, S.J., Delsing, M., ter Bogt, T.F. and Meeus, W.H. Different types of Internet use, depression, and social anxiety: The role of perceived friendship quality. *J. Adolescence* 32, 4 (2009), 819–833.
- Shaw, L.H. and Gant, L.M. In defense of the Internet: The relationship between Internet communication and depression, loneliness, self-esteem, and perceived

social support. *CyberPsychology & Behavior* 5, 2 (2002), 157–171.

- Shklovski, I., Kraut, R. and Rainie, L. The Internet and social participation: Contrasting cross-sectional and longitudinal analyses. *J. Computer-Mediated Communication* 10, 1 (2004).
- Shklovski, I., Kiesler, S. and Kraut, R.E. The Internet and social interaction: A meta-analysis and critique of studies, 199–2003. *Computers, Phones, and the Internet: Domesticating Information Technology*. R.E. Kraut, M. Brynin and S. Kiesler, (Eds). Oxford University Press, New York, 2006, 251–264.
- Singer, J.D. and Willett, J.B. *Applied Longitudinal Data Analysis*. Oxford University Press, New York, 2003.
- Steinfeld, C., Ellison, N. and Lampe, C. Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *J. Applied Developmental Psychology* 29, 6 (2008), 434–445.
- Sutcliffe, A., Dunbar, R., Binder, J. and Arrow, H. Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British J. Psychology* 103, 2 (2011), 149–168.
- Thaler, R. and Sunstein, C. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press, New Haven, CT, 2008.
- Thoits, P.A. Mechanisms linking social ties and support to physical and mental health. *J. Health and Social Behavior* 52, 2 (2011), 145–161.
- van den Eijnden, R.J.J.M., Meerkerk, G.J., Vermulst, A.A., Spijkerman, R. and Engels, R.C.M.E. Online communication, compulsive Internet use, and psychosocial well-being among adolescents: A longitudinal study. *Developmental Psychology* 44, 3 (2008), 655.
- Walther, J.B. Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research* 19, 1 (1992), 52–90.

Recommended Reading

Burke, M. and Kraut, R. Using Facebook after losing a job: Differential benefits of strong and weak ties. In *CSCW’13: Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, NY, 2013, 1419–1430.

Ellison, N.B., Steinfield, C. and Lampe, C. The benefits of Facebook ‘friends’: Social capital and college students’ use of online social network sites. *J. Computer-Mediated Communication* 12, 4 (2007), 143–168.

Kraut, R.E. et al. Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist* 53, 9 (1998), 11017–1031.

Sutcliffe, A. et al. Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British J. Psychology* 103, 2 (2011), 149–168.

Robert Kraut (robert.kraut@cmu.edu) is the Herbert A. Simon Professor of Human-Computer Interaction, in the Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA.

Maira Burke (mburke@fb.com) is a data scientist at Facebook, Menlo Park, CA.

research highlights

P. 102

Technical Perspective Paris Beyond Frommer's

By Noah Snavely

P. 103

What Makes Paris Look Like Paris?

By Carl Doersch, Saurabh Singh, Abhinav Gupta,
Josef Sivic, and Alexei A. Efros

P. 111

Technical Perspective In-Situ Database Management

By David Maier

P. 112

NoDB: Efficient Query Execution on Raw Data Files

By Ioannis Alagiannis, Renata Borovica-Gajic,
Miguel Branco, Stratos Idreos, and Anastasia Ailamaki

Technical Perspective

Paris Beyond Frommer's

By Noah Snaveley

TRY TO VISUALIZE PARIS. If you are like me, you will imagine a tourist-guidebook composite—the Eiffel Tower lit up at night, the Notre Dame Cathedral, the bridges over the Seine, and so on. But this is not what most of Paris looks like. It turns out that many people can look at a photograph of any randomly selected street corner in Paris, and can correctly identify the city with high accuracy, even without paying attention to any text in the photo. One must conclude that all of Paris is infused with a “Paris-ness”—a certain *je ne sais quoi*—that leaves an indelible visual mark on the City of Light.

How can we quantify this Paris-ness? Can a computer automatically discover and tell us what makes Paris look so much like Paris? More broadly, this question of *visual style* is an important one in computer graphics and vision. Identifying the key elements that characterize a style—whether a style of interior design, art, or, in the case of a city, its architecture and ornamentations—could aid in a range of applications, such as obtaining reference imagery for a new design task, or for summarizing or categorizing the look of a large set of images.

However, identifying and characterizing visual style automatically is a very challenging problem, one that is difficult even to formulate in a rigorous way. This is where the following paper steps in. This work, and several companion papers in computer vision, offers a creative, inspiring new approach to discovering the visual style of a city like Paris. The authors achieve this feat through new algorithms that analyze massive collections of photos of Paris and other places around the world.

One key aspect of this visual discovery problem is starting with the right image data. A popular approach in computer vision is to mine data from a large set of consumer photos shared

Can a computer observe a city and discern its visual essence?

on sites like Flickr. However, in the case of a city like Paris, this approach would result in a representation of Paris akin to the collage of landmarks of my own tourist-centered imagination, because this is how Paris is represented in photos shared online. To be sure, these landmarks are important, but to capture the real essence of Paris we need to look further. This work instead turns to Google Street View to gather many thousands of photos captured systematically throughout the city from cars—the kind of views you would see on a stroll down the street. Hence, the visual elements discovered in this work are the sort you would encounter in the city every day, perhaps not even noticing they make up its visual fabric.

But given a large, representative set of images of a place, how could we use them to compute these visual elements? One approach would be to break each image up into small patches (say, about the size of a door or window), group these patches by visual similarity using a clustering algorithm, and then identify the most common clusters as the key visual elements. But, as this work shows, this approach does not work very well—even if you account for uninteresting patches such as those in the sky, the results of standard clustering methods are unremarkable patches such as edges and corners.

Instead, the authors start from a crucial insight: what makes Paris look like Paris is not necessarily image patches that commonly appear in Paris, but instead those patches that appear in Paris but nowhere else—in other words, patches that distinguish Paris from all other cities. Using a new *discriminative clustering* technique, they show how they can automatically identify such distinctive patches. The discovered patches they show are remarkably evocative of Paris, capturing its unique balconies, signs, light posts, and other elements. When I first saw these results at SIGGRAPH, it immediately struck me they were on to something new and important, and showed something that no previous visual clustering method had shown. And many other interesting insights follow—such as the fact that U.S. cities are pretty similar to each other, but are distinctive from cities on the Continent in that they are filled with cars.

Using large image collections for computer vision and graphics is by now a tried-and-true approach. Past work has used photos mined from the Internet to train better object recognition systems or build 3D models. But it is exciting to see the fresh and innovative use of big data presented here. There is something magical about automatically distilling the visual signature of a place—signatures we all can sense but cannot easily articulate. And more broadly, this work represents an exciting new direction in discovering visual styles from big data. 

Noah Snaveley (snaveley@cs.cornell.edu) is an associate professor in the computer science department at Cornell University, Ithaca, NY, where he works in the graphics and vision group and works on the technical staff at Google.

Copyright held by author.

What Makes Paris Look Like Paris?

By Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros

Abstract

Given a large repository of geo-tagged imagery, we seek to automatically find visual elements, for example windows, balconies, and street signs, that are most distinctive for a certain geo-spatial area, for example the city of Paris. This is a tremendously difficult task as the visual features distinguishing architectural elements of different places can be very subtle. In addition, we face a hard search problem: given all possible patches in all images, which of them are both frequently occurring and geographically informative? To address these issues, we propose to use a discriminative clustering approach able to take into account the weak geographic supervision. We show that geographically representative image elements can be discovered automatically from Google Street View imagery in a discriminative manner. We demonstrate that these elements are visually interpretable and perceptually geo-informative. The discovered visual elements can also support a variety of computational geography tasks, such as mapping architectural correspondences and influences within and across cities, finding representative elements at different geo-spatial scales, and geographically informed image retrieval.

1. INTRODUCTION

Consider the two photographs in Figure 1, both downloaded from Google Street View. One comes from Paris, the other one from London. Can you tell which is which? Surprisingly, even for these nondescript street scenes, people who have been to Europe tend to do quite well on this task. In an informal survey, we presented 11 subjects with 100 random Street View images of which 50% were from Paris, and the rest from eleven other cities. We instructed the subjects (who have all been to Paris) to try and ignore any text in the photos, and collected their binary forced-choice responses (Paris/Not Paris). On average, subjects were correct 79% of the time ($std = 6.3$), with chance at 50% (when allowed to scrutinize the text, performance for some subjects went up as high as 90%). What this suggests is that people are remarkably sensitive to the geographically informative features within the visual environment. But what are those features? In informal debriefings, our subjects suggested that for most images, a few localized, distinctive elements “immediately gave it away.” For example for Paris, things like windows with railings, the particular style of balconies, the distinctive doorways, the traditional blue/green/white street signs, etc. were particularly helpful. Finding those features can be difficult though, since every image can contain more than 25,000 candidate patches, and only a tiny fraction will be truly distinctive.

In this work, we want to find such local geo-informative features *automatically*, directly from a large database of photographs from a particular place, such as a city. Specifically,

given tens of thousands of geo-localized images of some geographic region R , we aim to find a few hundred visual elements that are both: (1) repeating, that is, they occur often in R , and (2) geographically discriminative, that is, they occur much more often in R than in R^c . Figure 1 shows sample output of our algorithm: for each photograph we show three of the most geo-informative visual elements that were automatically discovered. For the Paris scene (left), the street sign, the window with railings, and the balcony support are all flagged as informative.

But why is this topic important for modern computer graphics? (1) Scientifically, the goal of understanding which visual elements are fundamental to our perception of a complex visual concept, such as a place, is an interesting and useful one. Our paper shares this motivation with a number of other recent works that do not actually synthesize new visual imagery, but rather propose ways of finding and visualizing existing image data in better ways, be it selecting candid portraits from a video stream,⁵ summarizing a scene from photo collections,¹⁹ finding iconic images of an object,¹ etc. (2) More practically, one possible future application of the ideas presented here might be to help CG modelers by generating the so-called “reference art” for a city. For instance, when modeling Paris for PIXAR’s *Ratatouille*, the co-director Jan Pinkava faced exactly this problem: “The basic question for us was: ‘what would Paris look like as a model of Paris?’, that is, what are the main things that give the city its unique look?”¹⁴ Their solution was to “run around Paris for a week like mad tourists, just looking at things, talking about them, and taking lots of pictures” not just of the Eiffel Tower but of the many stylistic Paris details, such as signs, doors, etc.¹⁴ (see photos on pp. 120–121). But if going “on location” is not feasible, our approach could serve as basis for a detail-centric reference art retriever, which would let artists focus their attention on the most statistically significant stylistic elements of the city. (3) And finally, more philosophically, our ultimate goal is to provide a *stylistic narrative* for a visual experience of a place. Such narrative, once established, can be related to others in a kind of geo-cultural visual reference graph, highlighting similarities and differences between regions. For example, one could imagine finding a visual appearance “trail” from Greece, through Italy and Spain and into Latin America. In this work, we only take the first steps in this direction—connecting visual appearance across cities, finding similarities within a continent, and differences between neighborhoods. But we hope that our work might act as a catalyst for research in this new area, which might be called *computational geo-cultural modeling*.

The original version of this paper was published in *SIGGRAPH*, 2012.

Figure 1. These two photos might seem nondescript, but each contains hints about which city it might belong to. Given a large image database of a given city, our algorithm is able to automatically discover the geographically informative elements (patch clusters to the right of each photo) that help in capturing its “look and feel.” On the left, the emblematic street sign, a balustrade window, and the balcony support are all very indicative of Paris, while on the right, the neoclassical columned entryway sporting a balcony, a Victorian window, and, of course, the cast-iron railings are very much features of London.



2. PRIOR WORK

In the field of architectural history, descriptions of urban and regional architectural styles and their elements are well established. Such local elements and rules for combining them have been used in computer systems for procedural modeling of architecture to generate 3D models of entire cities in an astonishing level of detail, for example, Mueller et al.,¹² or to parse images of facades, for example, Teboul et al.²² However, such systems require significant manual effort from an expert to specify the appropriate elements and rules for each architectural style.

At the other end of the spectrum, data-driven approaches have been leveraging the huge datasets of geo-tagged images that have recently become available online. For example, Crandall et al.² use the GPS locations of 35,000 consumer photos from Flickr to plot photographer-defined frequency maps of cities and countries. Geo-tagged datasets have also been used for place recognition^{8,17} including famous landmarks.^{10,11} Our work is particularly related to Schindler et al.¹⁷ and Knopp et al.,⁸ where geo-tags are also used as a *supervisory signal* to find sets of image features discriminative for a particular place. While these approaches can work very well, their image features typically cannot generalize beyond matching specific buildings imaged from different viewpoints. Alternatively, global image representations from scene recognition, such as GIST descriptor¹³ have been used for geolocalization of generic scenes on the global Earth scale.^{6,7} There, too, reasonable recognition performance has been achieved, but the use of global descriptors makes it hard for a human to interpret *why* a given image gets assigned to a certain location.

Finally, our paper is related to a line of work on unsupervised object discovery^{16,20} (and especially Quack et al.,¹⁵ who also deal with mining geo-tagged image data). Such methods attempt to explicitly discover features or objects which occur frequently in many images and are also useful as human-interpretable elements of visual representation. But being unsupervised, these methods are limited to only discovering things that are both very common and highly visually consistent.

In contrast, here we propose a discovery method that is weakly constrained by location labels derived from GPS tags, and which is able to mine representative

visual elements automatically from a large online image dataset. Not only are the resulting visual elements geographically discriminative (i.e., they occur only in a given locale), but they also typically look meaningful to humans, making them suitable for a variety of geo-data visualization applications. The next section describes the data used in this work, followed by the full description of our algorithm.

3. THE DATA

Flickr has emerged as the data-source of choice for most recently developed data-driven applications in computer vision and graphics, including visual geo-location.^{2,6,11} However, the difficulty with Flickr and other consumer photo-sharing websites for geographical tasks is that there is a strong data bias toward famous landmarks. To correct for this bias and provide a more uniform sampling of the geographical space, we turn to GOOGLE STREET VIEW—a huge database of street-level imagery, captured as panoramas using specially designed vehicles. This enables extraction of roughly fronto-parallel views of building facades and, to some extent, avoids dealing with large variations of camera viewpoint.

Given a geographical area on a map, we automatically scrape a dense sampling of panoramas of that area from Google Street View. From each panorama, we extract two perspective images (936 × 537 pixels), one on each side of the capturing vehicle, so that the image plane is roughly parallel to the vehicle’s direction of motion. This results in approximately 10,000 images per city. For this project, we downloaded 12 cities: Paris, London, Prague, Barcelona, Milan, New York, Boston, Philadelphia, San Francisco, San Paulo, Mexico City, and Tokyo.

4. DISCOVERING GEO-INFORMATIVE ELEMENTS

Our goal is to discover visual elements which are characteristic of a given geographical locale (e.g., the city of Paris). That is, we seek patterns that are both *frequently occurring* within the given locale, and *geographically discriminative*, that is, they appear in that locale and do not appear elsewhere. Note that neither of these two requirements by itself is enough: sidewalks and cars occur frequently in Paris but are hardly discriminative, whereas the Eiffel Tower is very discriminative, but too rare to be

useful ($<0.0001\%$ in our data). In this work, we will represent visual elements by square image patches at various resolutions, and mine them from our large image database. The database will be divided into two parts: (i) the positive set containing images from the location whose visual elements we wish to discover (e.g., Paris); and (ii) the negative set containing images from the rest of the world (in our case, the other 11 cities in the dataset). We assume that many frequently occurring but uninteresting visual patterns (trees, cars, sky, etc.) will occur in both the positive and negative sets, and should be filtered out. Our biggest challenge is that the overwhelming majority of our data is uninteresting, so matching the occurrences of the rare interesting elements is like finding a few needles in a haystack.

One possible way to attack this problem would be to first discover repeated elements and then simply pick the ones which are the most geographically discriminative. A standard technique for finding repeated patterns in data is clustering. For example, in computer vision, “visual word” approaches²¹ use k-means clustering on image patches represented by SIFT descriptors. Unfortunately, standard visual words tend to be dominated by low-level features, like edges and corners (Figure 2a), not the larger visual structures we are hoping to find. While we can try clustering using larger image patches (with a higher-dimensional feature descriptor, such as HOG³), k-means behaves poorly in very high dimensions, producing visually inhomogeneous clusters (Figure 2b). We believe this happens because k-means (and similar approaches) partition the *entire* feature space. This tends to lose the needles in our haystack: the rare discriminative elements get mixed with, and overwhelmed by, less interesting patches, making it unlikely that a distinctive element could ever emerge as its own cluster.

In this article, we propose an approach that avoids partitioning the entire feature space into clusters. Instead, we start with a large number of randomly sampled candidate patches, and then give each candidate a chance to see if it can converge to a cluster that is both frequent and discriminative. We first compute the nearest neighbors (NNs) of each candidate, and reject candidates with too many neighbors in the negative set. Then we gradually build clusters by applying iterative discriminative learning to each surviving candidate. The following section presents the details of this algorithm.

Figure 2. (a) k-Means clustering using SIFT (visual words) is dominated by low-level features. (b) k-Means clustering over higher-dimensional HOG features produces visually incoherent clusters.



4.1. Our approach

From the tens of millions of patches in our full positive set, we randomly sample a subset of 25,000 high-contrast patches to serve as candidates for seeding the clusters. Throughout the algorithm, we represent such patches using a HOG+color descriptor. First, the initial geo-informativeness of each patch is estimated by finding the top 20 NN patches in the full dataset (both positive and negative), measured by normalized correlation, and counting how many of them come from Paris. Figure 3 shows NNs for a few randomly selected patches and for the patches whose neighbors all come from Paris. Note that the latter patches are not only more Parisian, but also considerably more coherent. This is because generating a coherent cluster is a prerequisite to retrieving matches exclusively from Paris: any patch whose matches are incoherent will likely draw those matches randomly from inside and outside Paris. We keep the candidate patches that have the highest proportion of their NNs in the positive set, while also rejecting near-duplicate patches (measured by spatial overlap of more than 30% between any 5 of their top 50 NNs). This reduces the number of candidates to about 1000.

Some good elements, however, get matched incorrectly during the nearest-neighbors phase. Figure 4 shows a patch that contains both a street sign and a vertical bar on the right (the end of the facade). The naïve distance metric does not know what is important, and so it tries to match both. Yet too few such patches exist in the dataset; for the remainder, the algorithm matches the vertical bar simply because it is more frequent. To fix this problem, we aim to learn a distance metric that gives higher weight to the features that make the patch geo-discriminative.

Recently, Shrivastava et al.¹⁸ showed how one can improve visual retrieval by adapting the distance metric to the given query using discriminative learning. We adopt similar machinery, training a linear SVM detector for each visual element in an iterative manner as in Singh et al.²⁰ Unlike these previous works, however, we emphasize that

Figure 3. Left: Randomly sampled candidate patches and their nearest neighbors according to a standard distance metric. Right: After sorting the candidates by the number of retrieved neighbors that come from Paris, coherent Parisian elements have risen to the top.

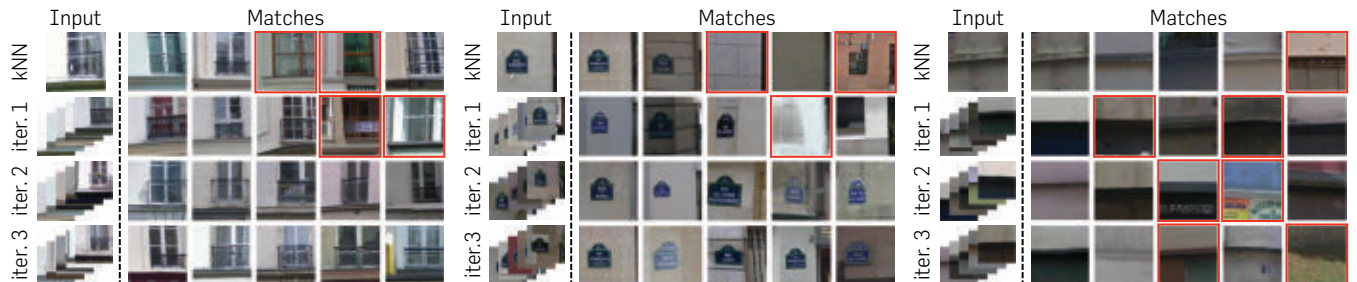


the weak labels are the workhorse of the distance learning. In the case of Figure 4, for example, we know that the street sign is more important because it occurs only in Paris, whereas the vertical bar occurs everywhere. We train an SVM detector for each visual element, using the top k NNs from the positive set as positive examples, and all negative-set patches as negative examples. While this produces a small improvement (Figure 5, row 2), it is not enough, since the top k matches might not have been very good to begin with. So, we iterate the SVM learning, using the top k detections from previous round as positives (we set $k = 5$ for all experiments). The idea is that with each round, the top detections will become better and better, resulting in a continuously improving detector. However, doing this directly would not produce much improvement because the SVM tends to over-fit to the initial positive examples Singh et al.,²⁰ and will prefer them in each next round over new (and better) ones. Therefore, we apply cross-validation by dividing both the positive and the negative parts of the dataset into l equally sized subsets (we set $l = 3$ for all experiments). At each iteration of the training, we apply the detectors trained on the previous round to a new, *unseen* subset of data to select the top k detections for retraining. In our experiments, we used three iterations, as most good clusters did not need more to converge (i.e., stop changing). After the final iteration, we rank the resulting detectors based on their accuracy: percentage of top 50 firings that are in the positive dataset (i.e.,

Figure 4. Top: Using the naïve distance metric for this patch retrieves some good matches and some poor matches, because the patch contains both a street sign and a vertical bar on the right. Bottom: Our algorithm reweights the dimensions of our patch descriptor to separate Paris from non-Paris. The algorithm learns that focusing on the street sign achieves maximum separation from the non-Paris walls.



Figure 5. Steps of our algorithm for three sample candidate patches in Paris. The first row: initial candidate and its NN matches. Rows 2–4: Iterations of SVM learning (trained using patches on left). Red boxes indicate matches outside Paris. Rows show every 7th match for clarity. Notice how the number of not-Paris matches decreases with each iteration, except for rightmost cluster, which is eventually discarded.



in Paris). We return the top few hundred detectors as our geo-informative visual elements.

Figure 5 illustrates the progression of these iterations. For example, in the left column, the initial NNs contain only a few windows with railings. However, windows with railings differ more from the negative set than the windows without railings; thus the detector quickly becomes more sensitive to them as the algorithm progresses. The rightmost example does not appear to improve, either in visual similarity or in geo-discriminateness. This is because the original candidate patch was intrinsically not very geo-informative and would not make a good visual element. Such patches have a low final accuracy and are discarded.

Implementation Details: Our current implementation considers only square patches (although it would not be difficult to add other aspect ratios), and takes patches at scales ranging from 80-by-80 pixels all the way to height-of-image size. Patches are represented with standard HOG³ ($8 \times 8 \times 31$ cells), plus a 8×8 color image in L*a*b colorspace (a and b only). Thus the resulting feature has $8 \times 8 \times 33 = 2112$ dimensions. During iterative learning, we use a soft-margin SVM with C fixed to 0.1. The full mining computation is quite expensive; a single city requires approximately 1800 CPU-hours. But since the algorithm is highly parallelizable, it can be done overnight on a cluster.

4.2. Results and validation

Figure 6 shows the results of running our algorithm on several well-known cities. For each city, the left column shows randomly chosen images from Google Street View, while the right column shows some of the top-ranked visual element clusters that were automatically discovered (due to space limitations, a subset of elements was selected manually to show variety; see the project webpage for the full list). Note that for each city, our visual elements convey a better stylistic feel of the city than do the random images. For example, in Paris, the top-scoring elements zero-in on some of the main features that make Paris look like Paris: doors, balconies, windows with railings, street signs and special Parisian lampposts. It is also interesting to note that, on the whole, the algorithm had more trouble with American cities: it was able to discover only a few geo-informative elements, and some of them turned

Figure 6. Google Street View versus geo-informative elements for six cities. Arguably, the geo-informative elements (right) are able to provide better stylistic representation of a city than randomly sampled Google Street View images (left).



out to be different brands of cars, road tunnels, etc. This might be explained by the relative lack of stylistic coherence and uniqueness in American cities (with its melting pot of styles and influences), as well as the supreme reign of the automobile on American streets.

In addition to the qualitative results, we would also like to provide a more quantitative evaluation of our algorithm. While validating data-mining approaches is difficult in general, it is possible to measure (1) to what extent our elements are specific to particular locations,

and (2) do users find them subjectively geo-informative in a visual discrimination task?

To evaluate how geo-informative our visual elements are, we ran the top 100 Paris element detectors over an unseen dataset which was 50% from Paris and 50% from elsewhere. For each element, we found its geo-informativeness by computing the percentage of the time it fired in Paris out of the top 100 firings. The average accuracy of our top detectors was 83% (where chance is 50%). We repeated this for our top 100 Prague detectors, and found the average accuracy on an unseen dataset of Prague to be 92%.

Next, we repeated the above experiment with people rather than computers. To avoid subject fatigue, we reduced the dataset to 100 visual elements, 50 from Paris and 50 from Prague. Fifty percentage of the elements were the top-ranked ones returned by our algorithm for Paris and Prague. The other 50% were randomly sampled patches of Paris and Prague (but biased to be high-contrast, as before, to avoid empty sky patches, etc.). In a web-based study, subjects (who have all been to Paris but not necessarily Prague) were asked to label each of the 100 patches as belonging to either Paris or Prague (forced choice). The results of our study (22 naive subjects) are as follows: average classification performance for the algorithm-selected patches was 78.5% ($std = 11.8$), while for random patches it was 58.1% ($std = 6.1$); the p -value for a paired-samples t -test was $<10^{-8}$. While on random patches subjects did not do much better than chance, performance on our geo-informative elements was roughly comparable to the much simpler full-image classification task reported in the beginning of the paper (although since here we only used Prague, the setups are not quite the same).

5. APPLICATIONS

Now that we have a tool for discovering geographically informative visual elements for a given locale, we can use them to explore ways of building stylistic narratives for cities and of making visual connections between them. Here we discuss just a few such directions.

5.1. Mapping patterns of visual elements

So far, we have shown the discovered visual elements for a given city as an ordered list of patch clusters (Figure 6). Given that we know the GPS coordinates of each patch, however, we could easily display them on a map, and then search for interesting geo-spatial patterns in the occurrences of a given visual element. Figure 7 shows the geographical locations for the top-scoring detections for each of three different visual elements (a sampling of detections are shown below each map), revealing interestingly non-uniform distributions. For example, it seems that balconies with cast-iron railings (left) occur predominantly on the large thoroughfares (bd Saint-Michel, bd Saint-Germain, rue de Rivoli), whereas windows with cast-iron railings (middle) appear mostly on smaller streets. The arch-supporting column (right) is a distinguishing feature of the famous Place des Vosges, yet it also appears in other parts of Paris, particularly as part of more recent Marché Saint-Germain (this is a possible example of the so-called “architectural citation”). Automatically discovering such architectural patterns may be useful to both architects and urban historians.

5.2. Visual correspondences across cities

Given a set of architectural elements (windows, balconies, etc.) discovered for a particular city, it is natural to ask what these same elements might look like in other cities. As it turns out, a minor modification to our algorithm can often accomplish this task. We have observed that a detector for a location-specific architectural element will often fire on functionally similar elements in other cities, just with a much lower score. That is, a Paris balcony detector will return mostly London balconies if it is forced to run only on London images. Naturally these results will be noisy, but we can clean them up using an iterative learning approach similar to the one in Section 4.1. The only difference is that we require the positive patches from each iteration of training to be taken not just from the source city, but from all the cities where we wish to find correspondences. For example, to find correspondences between Paris, Prague, and London,

Figure 7. Examples of geographic patterns in Paris (shown as red dots on the maps) for three discovered visual elements (shown below each map). Balconies with cast-iron railings are concentrated on the main boulevards (left). Windows with railings mostly occur on smaller streets (middle). Arch supporting columns are concentrated on Place des Vosges and the St. Germain market (right).



Figure 8. Visual correspondence. Each row shows corresponding detections of a single visual element detector across three different cities.



we initialize with visual elements discovered in Paris and then, at each round of “clean-up” training, we use nine top positive matches to train each element SVM, three from each of the three cities. Figure 8 illustrates the result of this procedure. Note how capturing the correspondence between similar visual elements across cities can often highlight certain stylistic differences, such as the material for the balconies, the style of the street-lamps, or the presence and position of ledges on the facades.

5.3. Visualizing facade layout

Another interesting observation is that some discovered visual elements, despite having a limited spatial extent, can often encode a much larger architectural context. This becomes particularly apparent when looking at the same visual element detector applied in different cities. Figure 9 shows object-centric averages (in the style of Torralba and Oliva²³) for the detector in the top row of Figure 8 for Paris and London. That is, for each city, the images with the top 100 detections of the element are first centered on that element and then averaged together in image space. Note that not only do the average detections (red squares) look quite different between the two cities, but the average contexts reveal quite a lot about the differences in the structure and style of facades. In Paris, one can clearly see four equal-height floors, with a balcony row on the third floor. In London, though, floor heights are uneven, with the first floor much taller and more stately.

6. CONCLUSION

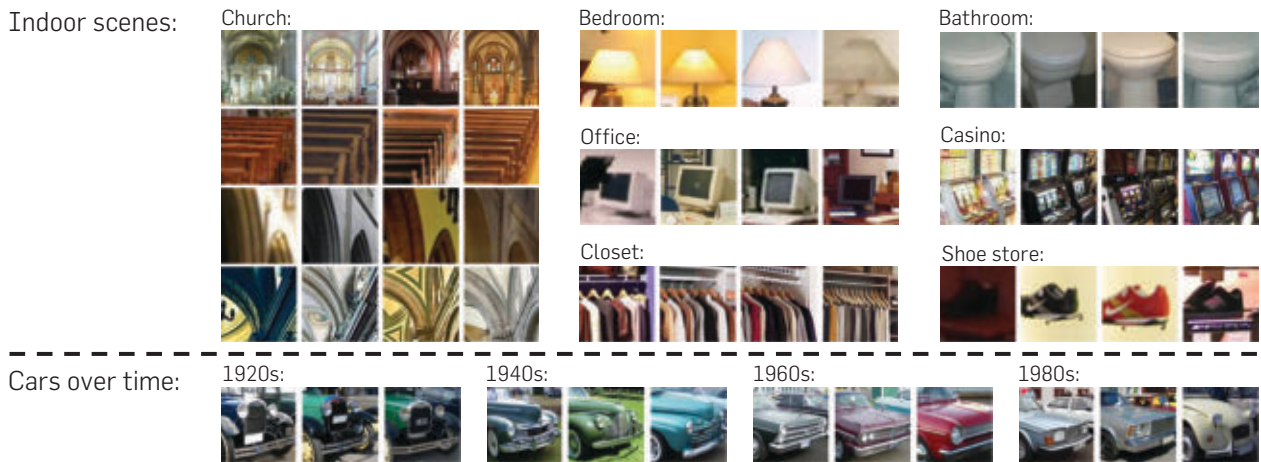
So, what makes Paris look like Paris? We argued that the “look and feel” of a city rests not so much on the few

Figure 9. Object-centric image averages for the element detector in the top row of Figure 8. Note how the context captures the differences in facade styles between Paris (left) and London (right).



famous landmarks (e.g., the Eiffel Tower), but largely on a set of stylistic elements, the visual minutiae of daily urban life. We proposed a method that can automatically find a subset of such visual elements from a large dataset offered by Google Street View, and demonstrated some promising applications. This work is but a first step toward our ultimate goal of providing stylistic narratives to explore the diverse visual geographies of our world. Currently, the method is limited to discovering only local elements (image patches), so a logical next step would be trying to capture larger structures, both urban (e.g., facades), as well as natural (e.g., fields, rivers). Finally, the proposed algorithm is not limited to geographic data. Figure 10 shows promising results for mining discriminative patches on indoor scenes, and cars, suggesting that visual elements can be a useful tool for exploring a wide variety of image data domains. □

Figure 10. Our algorithm applied to other data sources. Top: Elements for indoor scenes, where the weak label is 1 of 67 indoor scene categories.^{4, 20} Bottom: Stylistic elements that differentiate cars from different decades.⁹



References

- Berg, T., Berg, A. Finding iconic images. In *The 2nd Internet Vision Workshop at Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), IEEE, 1–8.
- Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J. Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web (WWW)* (2009), 761–770.
- Dalal, N., Triggs, B. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Volume 1 (2005), IEEE, 886–893.
- Doersch, C., Gupta, A., Efros, A.A. Mid-level visual element discovery as discriminative mode seeking. In *Advances in Neural Information Processing Systems (NIPS)*. Volume 26 (2013), 494–502.
- Fiss, J., Agarwala, A., Curtless, B. Candid portrait selection from video. *ACM Trans. Graph. (SIGGRAPH Asia)* 30, 6 (2011), 128.
- Hays, J., Efros, A. Im2gps: Estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2008), IEEE, 1–8.
- Kalogerakis, E., Vesselova, O., Hays, J., Efros, A., Hertzmann, A. Image sequence geolocation with human travel priors. In *IEEE 12th International Conference on Computer Vision (ICCV)* (2009), IEEE, 253–260.
- Knopp, J., Sivic, J., Pajdla, T. Avoiding confusing features in place recognition. In *European Conference on Computer Vision (ECCV)* (2010), Springer, 748–761.
- Lee, Y.J., Efros, A.A., Hebert, M. Style-aware mid-level representation for discovering visual connections in space and time. In *IEEE 14th International Conference on Computer Vision (ICCV)* (2013), IEEE, 1857–1864.
- Li, X., Wu, C., Zach, C., Lazebnik, S., Frahm, J.-M. Modeling and recognition of landmark image collections using iconic scene graphs. In *European Conference on Computer Vision (ECCV)* (2008), Springer, 427–440.
- Li, Y., Crandall, D., Huttenlocher, D. Landmark classification in large-scale image collections. In *IEEE 12th International Conference on Computer Vision (ICCV)* (2009), IEEE, 1957–1964.
- Mueller, P., Wonka, P., Haegler, S., Ulmer, A., Van Gool, L. Procedural modeling of buildings. *ACM Trans. Graph. (SIGGRAPH)* 25, 3 (2006), 614–623.
- Oliva, A., Torralba, A. Building the gist of a scene: The role of global image features in recognition. *Prog. Brain Res.* 155 (2006), 23–36.
- Paik, K. *The Art of Ratatouille*. Chronicle Books, 2006.
- Quack, T., Leibe, B., Van Gool, L. World-scale mining of objects and events from community photo collections. In *Proceedings of the International Conference on Content-based Image and Video Retrieval (CIVR)* (2008), 47–56.
- Russell, B.C., Efros, A.A., Sivic, J., Freeman, W.T., Zisserman, A. Using multiple segmentations to discover objects and their extent in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2006), IEEE, 1605–1614.
- Schindler, G., Brown, M., Szeliski, R. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2007), IEEE, 1–7.
- Shrivastava, A., Malisiewicz, T., Gupta, A., Efros, A.A. Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph. (SIGGRAPH Asia)* 30, 6 (2011), 154.
- Simon, I., Snavely, N., Seitz, S.M. Scene summarization for online image collections. In *IEEE 11th International Conference on Computer Vision (ICCV)* (2007), IEEE, 1–8.
- Singh, S., Gupta, A., Efros, A.A. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision (ECCV)* (2012), Springer, 73–86.
- Sivic, J., Zisserman, A. Video google: A text retrieval approach to object matching in videos. In *IEEE 9th International Conference on Computer Vision (ICCV)* (2003), IEEE, 1470–1477.
- Teboul, O., Simon, L., Koutsourakis, P., Paragios, N. Segmentation of building facades using procedural shape priors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2010), IEEE, 3105–3112.
- Torralba, A., Oliva, A. Statistics of natural image categories. *Netw. Comput. Neural Syst.* 14, 3 (2003), 391–412.

Carl Doersch (cdoersch@cs.cmu.edu), Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA.

Saurabh Singh (saurabh.me@gmail.com), Computer Science Department, University of Illinois, Urbana-Champaign, Champaign, IL.

Abhinav Gupta (abhinavg@cs.cmu.edu), Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Josef Sivic (josef.sivic@ens.fr), Computer Science Department, INRIA/Ecole Normale Supérieure, Paris, France.

Alexei A. Efros (efros@eecs.berkeley.edu), Electrical Engineering and Computer Science (EECS) Department, University of California, Berkeley, Berkeley, CA.



Watch the authors discuss their work in this exclusive *Communications* video. <http://cacm.acm.org/videos/what-makes-paris-look-like-paris>

Copyright held by authors.
Publication rights licensed to ACM. \$15.00.

Technical Perspective

In-Situ Database Management

By David Maier

IMAGINE YOU HAVE a collection of data files—say, cell-tower call records—and you believe some of them might contain useful information: towers that are near capacity, numbers whose calls are frequently dropped, failed hand-offs. You want to run a few queries over selected files to explore which ones might merit further analysis and determine what kinds of knowledge you might extract. If a file is small, you might transfer it to your own computer and inspect it with a spreadsheet program or an analysis environment such as R. However, suppose individual files are too large to fit on your computer. What are your choices then for exploring them?

One possibility is to load the data into a database management system (DBMS) and use the query language—likely SQL—to ask your questions. While you get the advantage of a high-level data language, it might take hours to load the data before you can pose your first query. If you want to switch to another file, then you have to wait again while that file loads. Moreover, any file you do load now takes up at least twice the storage space, since its data is in both the database and on the file system. Deleting the original file might not be possible, if it has other users.

An option is to use a MapReduce framework, such as Hadoop, to run your preliminary analyses. Now your “time to insight” is delayed by having to write (and probably debug) a program. Even if you are able to formulate your questions as programs fairly quickly (perhaps using a language layer such as Hive or PigLatin), each query you run will scan the whole file anew. In addition, you lose performance enhancements such as indexes and optimization available in a DBMS.

Such unattractive trade-offs face nearly everyone wanting to quickly explore a new data source. The following paper by Alagiannis et al. investigates a third approach, extending a DBMS so

The following paper is exciting, as it minimizes upfront costs when exploring new data sources, and it opens up a wide range of additional techniques to pursue for in-situ data management.

it can use the file data in situ, without having to load it first. They term their approach “NoDB” to indicate it does not require a separate copy of the data stored internally to the DBMS.

Note that some DBMSs do support links to external files that are viewed as tables, which are parsed and temporarily loaded on demand. That approach does avoid the initial load into persistent storage and allows the loading process to overlap with other query stages. However, loading happens on every query, as with the MapReduce approach. Furthermore, such external data is a second-class citizen, lacking the indexes and statistics that speed performance on internal data. The NoDB approach, in contrast, tries to make in-situ data first class, by using an incremental, pay-as-you-go approach to providing DBMS functionality that tries to minimize up-front load costs, while capturing the work that is done for the benefit of later queries.

The authors built a specific instance of a NoDB system called PostgresRaw. The main techniques it uses are to avoid parsing portions of a file not needed by the current query, and reusing the work it does do via a “positional map” (a sort of structural index) that remembers the location of fields in records it does access. Thus, initial queries avoid the full load cost of an external file, while later queries take advantage of previous parsing work. The authors also consider incremental methods on in-situ data, such as collecting statistics and caching data from one query to another. The evaluation of PostgresRaw shows the performance advantage of NoDB technology over the alternatives mentioned here.

Should we expect our DBMSs in the future will do away with internal storage and run entirely over in-situ data? Likely not—the NoDB approach targets large, static datasets (though it could be extended to handle certain classes of updates, such as appending records). Data subject to small, frequent changes is best maintained by the DBMS. Also, if you determine at some point that a file will be intensely queried in the future, then it is worthwhile to incur the up-front load cost, in exchange for faster queries. Nevertheless, the paper is exciting, as it minimizes up-front costs when exploring new data sources, and it opens up a wide range of additional techniques to pursue for in-situ data management, such as incremental value-based indexing, synthesizing access methods for common file types, and selective transfer of in-situ data to internal storage. 

David Maier (maier@cs.pdx.edu) is Maseeh Professor of Emerging Technologies in the Department of Computer Science at Portland State University, Portland, OR.

NoDB: Efficient Query Execution on Raw Data Files

By Ioannis Alagiannis, Renata Borovica-Gajic, Miguel Branco, Stratos Idreos, and Anastasia Ailamaki

Abstract

As data collections become larger and larger, users are faced with increasing bottlenecks in their data analysis. More data means more time to prepare and to load the data into the database before executing the desired queries. Many applications already avoid using database systems, for example, scientific data analysis and social networks, due to the complexity and the increased *data-to-query* time, that is, the time between getting the data and retrieving its first useful results. For many applications data collections keep growing fast, even on a daily basis, and this *data deluge* will only increase in the future, where it is expected to have much more data than what we can move or store, let alone analyze.

We here present the design and roadmap of a new paradigm in database systems, called NoDB, which *do not require data loading while still maintaining the whole feature set of a modern database system*. In particular, we show how to make raw data files a first-class citizen, fully integrated with the query engine. Through our design and lessons learned by implementing the NoDB philosophy over a modern Database Management Systems (DBMS), we discuss the fundamental limitations as well as the strong opportunities that such a research path brings. We identify performance bottlenecks specific for in situ processing, namely the repeated parsing and tokenizing overhead and the expensive data type conversion. To address these problems, we introduce an adaptive indexing mechanism that maintains positional information to provide efficient access to raw data files, together with a flexible caching structure. We conclude that NoDB systems are feasible to design and implement over modern DBMS, bringing an unprecedented positive effect in usability and performance.

1. INTRODUCTION

We are in the era of data deluge, where the amount of generated data outgrows the capabilities of query processing technology. Many emerging applications, from social networks to scientific experiments, are representative examples of this deluge, where the rate at which data is produced exceeds any past experience. Scientific disciplines such as astronomy are soon expected to collect multiple Terabytes of data on a daily basis. Similarly, web-based businesses such as social networks or web log analysis are already confronted with a growing stream of large data inputs. Therefore, there is a clear need for efficient big data processing to enable the evolution of businesses and sciences to the new era of data deluge.

Motivation. Although Database Management Systems (DBMS) remain overall the predominant data analysis

technology, they are rarely used for emerging applications. This is largely due to the complexity involved; there is a significant initialization cost in loading data and preparing the database system for queries. For example, a scientist needs to quickly examine a few Terabytes of new data in search of certain properties. Even though only a few attributes might be relevant for the task, the entire data must first be loaded inside the database. Besides being a significant time investment, it is also important to consider the extra computing resources required for a full load and its side-effects with respect to energy consumption and economical sustainability.

Instead of using database systems, emerging applications rely on custom solutions that usually miss important database features. For instance, declarative queries, schema evolution and complete isolation from the internal representation of data are rarely present. There are a wide variety of competing approaches but users remain exposed to many low-level details and must work close to the physical level to obtain adequate performance and scalability. A growing part of the database community recognizes the need for significant and fundamental changes to database design, ranging from low-level architectural redesigns to changes in the way users interact with the system.^{2, 5, 8, 9, 12, 14, 16, 17, 21}

The NoDB philosophy. We recognize this new need, which is a direct consequence of the data deluge, and describe the roadmap toward NoDB, a new database design philosophy that we believe will come to define how future database systems are designed. The goal of the NoDB philosophy is to make database systems more accessible to the user by eliminating major bottlenecks of current state-of-the-art technology that increases the data-to-query time. The data-to-query time is of critical importance as it defines the moment when a database system becomes usable and thus useful. There are, however, fundamental processes in modern database architectures that represent a major bottleneck for data-to-query time. The NoDB philosophy changes the way a user interacts with a database system by eliminating one of the most important bottlenecks, that is, data loading. We advocate querying over raw data, in situ (i.e., in its original place) as the principal way to manage data in a database and we propose to redesign the query processing layers of database systems to incrementally and adaptively query raw data files in situ, while automatically creating and refining auxiliary structures to speed-up future queries.

The original version of this paper was published in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (Scottsdale, Arizona, USA).

Adaptive data loads. We originally introduced the idea of adaptive data loading in an earlier vision paper.⁹ The current paper makes numerous and significant contributions, toward demonstrating the feasibility and the potential of that vision. Using a mature and complete implementation over a modern DBMS, we identify and overcome fundamental limitations in NoDB systems. Most importantly, we show how to make raw files first-class citizens without sacrificing query performance. We also introduce several innovative techniques such as selective parsing, adaptive indexing structures that operate on the raw files, caching techniques, and statistics collection over raw files. Overall, we describe how to exploit current relational databases to conform to the NoDB philosophy while identifying limitations and opportunities in the process.

Contributions. Our contributions are as follows:

- We convert a traditional relational database (PostgreSQL) into a NoDB system (PostgresRaw), and discover that the main bottlenecks are the repeated access and parsing of raw files. Therefore, we design an innovative adaptive indexing mechanism that makes the trip back to the raw files efficient.
- We demonstrate that the query response time of a NoDB system can be competitive with a traditional DBMS, even without prior data loading.
- We show that NoDB systems provide quick access to the data under a variety of workloads. PostgresRaw query performance improves adaptively as it processes additional queries and it quickly matches or outperforms traditional DBMS, including MySQL and PostgreSQL.
- We describe opportunities with the NoDB philosophy, as well as challenges such a research path brings.

2. QUERYING RAW DATA

In this section, we introduce the NoDB philosophy. For ease of presentation, we first discuss a straw-man approach to in situ querying, where every query relies exclusively on raw files for query processing. Then, we address the weaknesses of the straw-man approach by introducing the core concepts of NoDB that enable efficient access to raw data.

Typical storage and execution. A row-store DBMS organizes data in the form of tuples, stored sequentially one tuple after the other in the form of slotted pages. Each page contains a collection of tuples as well as additional metadata information to help in-page navigation. These pages are created during the loading process. Before being able to submit queries, the data must first be loaded, which transforms it from the raw format to the database page format. During query processing the system brings pages into memory and processes the tuples. In order to create proper query plans, that is, to decide the operators and their order of execution, an optimizer is used, which exploits previously collected statistics about the data. A query plan can be seen as a tree where each node is a relational operator and each leaf corresponds to a data access method. The access methods define how the system accesses the tuples. Each tuple is then passed one-by-one through the operators of a query plan. The NoDB philosophy needs to be integrated with the afore-mentioned design for efficient and adaptive query execution.

2.1. Straightforward approaches

We describe two straightforward ways to directly query raw data files. The first approach is to simply run the loading procedure whenever a relevant query arrives: when a query referring to table R arrives, only then load table R , and immediately evaluate the query over the loaded data. Data may be loaded into temporary tables that are discarded after processing the query, or it may be loaded into persistent tables stored on disk. These approaches however, significantly penalize the first query, since creating the complete table before evaluating the query implies that the same data needs to be accessed twice, once for loading and once for query evaluation.

A better approach is to tightly integrate the raw file accesses with the query execution. This is accomplished by enriching the leaf operators of the query plans, for example, the *scan* operator, with the ability to access raw data files. Therefore, the *scan* operator tokenizes and parses a raw file on-the-fly, creates the tuples and passes them to the remaining of the query plan. The key difference is that data parsing and processing occur in a pipelined fashion, that is, the raw file is read from disk in chunks and once a tuple or a group of tuples is produced, the *scan* immediately passes those tuples upstream.

Both straw-man techniques require that the proper schema be known a priori; the user needs to declare the schema and mark all tables as in situ tables. Other than that, both techniques represent a straightforward implementation of in situ query processing; they do not require significant new technology other than a careful integration of existing loading procedures with query processing.

Limitations of straightforward approaches. The approaches discussed here are similar to the external files functionality offered by modern database systems such as Oracle and MySQL. Such solutions are not viable for extensive and repeated query processing. For example, if data is not kept in persistent tables, then every future query needs to perform loading from scratch, which is a major overhead. Materializing loaded data into persistent tables however, forces a single query to incur all loading costs. Therefore, such approaches are only viable if a user needs to fire few queries.

Neither straw-man technique allows the implementation of important database systems functionality. In particular, given that data is not loaded, there is no mechanism to exploit indexing; modern database systems do not support indexes on raw data. Without index support, query plans for straw-man techniques rely only on full scans, incurring a significant performance degradation compared to a DBMS with loaded data and indexes. In addition, the optimizer cannot exploit any statistics, since statistics in a modern DBMS are created only after data is loaded. The lack of statistics and indexing means that straw-man techniques do not provide query processing performance comparable to a modern DBMS and any time gained by skipping data loading is lost after only a few queries.

Even though in situ features, such as external files, are important for the users, current implementations are far from the NoDB vision of providing an *instant gateway to the*

data, without losing the performance advantages achieved by modern DBMS.

2.2. The NoDB philosophy

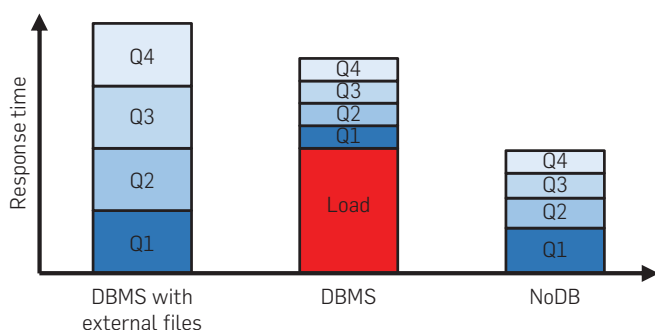
The NoDB philosophy aims to provide in situ access with query processing performance that is competitive with a database system operating over previously loaded data. In other words, the vision is to completely shed the loading costs, while achieving or improving the query processing performance of a traditional DBMS. Such performance characteristics make the DBMS usable and flexible; a user may only think about the kind of queries to pose and not about setting up the system in advance and going through all the initialization steps that are necessary today.

The design we propose in this work takes significant steps in identifying and eliminating or greatly minimizing initialization and query processing costs that are unique for in situ systems. The target behavior is visualized in Figure 1. It illustrates an important aspect of the NoDB philosophy; even though individual queries may take longer to respond than in a traditional system, the data-to-query time is reduced, because there is no need to load and prepare data in advance or to fine tune the system when different queries arrive. In addition, performance improves gradually as a function of the number of queries processed.

New challenges of NoDB systems. The main bottleneck of in situ query processing is the access to raw data. The costs involved in raw data access significantly degrade query performance. In a traditional DBMS, parsing raw data files is more expensive than accessing database pages. The NoDB philosophy aims at making raw data a first-class citizen, integrating raw data access in an abstract way into the query processing layer, allowing query processing without a priori loading. However, a NoDB system can only be useful and attractive in practice if it achieves performance levels comparable to a modern DBMS. Therefore, the main challenge for a NoDB system is to minimize the cost of accessing raw data.

From a high level point of view, we distinguish between two directions; the first one aims at minimizing the cost of raw data access through the careful design of data structures that can speed-up such accesses; the second one aims at selectively eliminating the need for raw data access by careful caching and scheduling raw data accesses. The final

Figure 1. Improving user interaction with NoDB.



grand challenge is to come up with a seamless design that integrates such features into a modern DBMS.

3. POSTGRESRAW: BUILDING NoDB IN POSTGRESQL

In this section, we discuss the design of our NoDB prototype, called PostgresRaw, implemented by modifying the open-source DBMS PostgreSQL. We show how to minimize parsing and tokenizing costs within a row-store engine via selective and adaptive parsing actions. In addition, we present a novel raw file indexing structure that adaptively maintains positional information to speed-up future accesses on raw files. Finally, we present caching and exploitation of statistics in PostgresRaw. The ideas described in this section can be used as guidelines for turning modern row-stores into NoDB systems.

In the remaining of this section we assume that raw data is stored in comma-separated value (CSV) files. Comma-separated value files as textual files are challenging for an in situ engine, considering the high conversion cost to binary format and the fact that fields may be variable length. Nonetheless, being a common data source, they present an ideal use case for PostgresRaw.

3.1. On-the-fly parsing

We first discuss aspects related to on-the-fly raw file parsing and essential features such as selective parsing and tuple formation. We later describe the core PostgresRaw components.

Query plans in PostgresRaw. When a query submitted to PostgresRaw references relational tables that are not yet loaded, PostgresRaw needs to access the respective raw file(s). PostgresRaw overrides the *scan* operator with the ability to access raw data files directly, while the remaining query plan, generated by the optimizer, works without changes compared to a conventional DBMS.

Parsing and tokenizing raw data. Every time a query needs to access raw data, PostgresRaw has to perform parsing and tokenization. In a typical CSV structure, each CSV file represents a relational table, each row in the CSV file represents a tuple of a table and each entry in a row represents an attribute value of the tuple. During parsing, PostgresRaw needs first to identify each tuple, or row in the raw file. Once all tuples have been identified, PostgresRaw must then search for the delimiter separating different values and transform those characters into their proper binary values. Overall, these extra parsing and tokenizing actions represent a significant overhead inherent to in situ query processing; a typical DBMS performs all these steps at loading time and directly reads binary database pages during query processing.

Selective tokenizing. PostgresRaw reduces the tokenizing costs by opportunistically aborting tokenizing tuples as soon as the required attributes for a query have been found. This occurs at a per tuple basis. Given that CSV files are organized in a row-by-row basis, selective tokenizing does not bring any I/O benefits; nonetheless, it significantly reduces the CPU processing costs.

Selective parsing. In addition to selective tokenizing, PostgresRaw also employs selective parsing to further reduce raw access costs. PostgresRaw transforms to binary only the

values required to answer the query. For example, if a query requests the 4th and 8th attribute of a given file and the query contains a selection on the 4th attribute. PostgresRaw with selective parsing converts all values of the 4th attribute to binary but delays the binary transformation of the 8th attribute, until it knows that the given tuple qualifies.

Selective tuple formation. To fully capitalize on selective parsing and tokenizing, PostgresRaw also applies selective tuple formation. Tuples are not fully composed but only contain the attributes required for a given query. In PostgresRaw, tuples are only created after the *select* operator, that is, after knowing which tuples qualify.

Overall selective tokenizing, parsing, and tuple formation help to significantly minimize the on-the-fly processing costs, since PostgresRaw parses only what is necessary to produce query answers.

3.2. Indexing

Even with selective tokenizing, parsing and tuple formation, the cost of accessing raw data is still significant. This section introduces an auxiliary structure that allows PostgresRaw to compete with a DBMS with previously loaded data. This auxiliary structure is a positional map, and forms a core component of PostgresRaw.

Adaptive positional map. We introduce the adaptive positional map to reduce parsing and tokenizing costs. It maintains low level metadata information on the structure of the flat file, which is used to navigate and retrieve raw data faster. This metadata information refers to positions of attributes in the raw file. For example, if a query needs an attribute *X* that is not loaded, then PostgresRaw can exploit this metadata information that describes the position of *X* in the raw file and jump directly to the correct position without having to perform expensive tokenizing steps to find *X*.

Map population. The positional map is created on-the-fly during query processing, continuously adapting to queries. Initially, the positional map is empty. As queries arrive, PostgresRaw adaptively and continuously augments the positional map. The map is populated during the tokenizing phase, that is, while tokenizing the raw file for the current query, PostgresRaw adds information to the map.

PostgresRaw learns as much information as possible during each query. For instance, it does not keep maps only for the attributes requested in the query, but also for attributes tokenized along the way; for example, if a query requires attributes in positions 10 and 15, all positions from 1 to 15 may be kept.

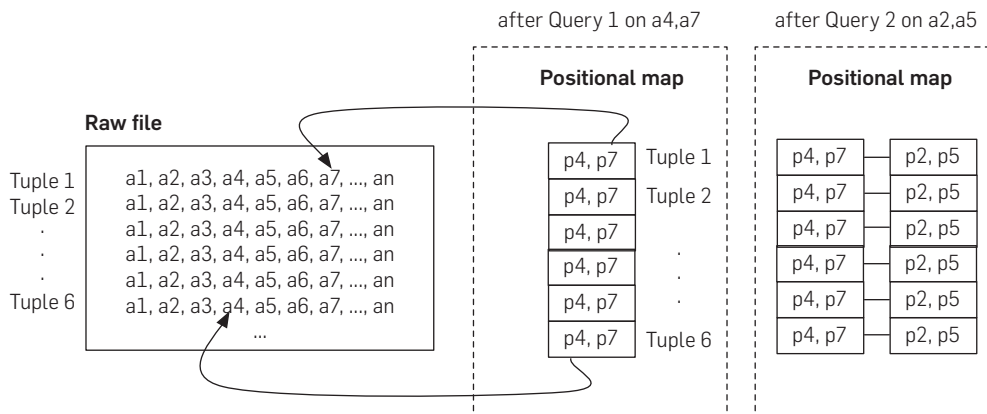
Storage format. The dynamic nature of the positional map requires a physical organization that is easy to update and incurs low cost during query execution. To achieve efficient reads and writes, the PostgresRaw positional map is implemented as a collection of chunks, partitioned vertically and horizontally. Each chunk fits comfortably in the CPU caches, allowing PostgresRaw to efficiently acquire all information regarding several attributes and tuples with a single access. The map can also be extended by adding more chunks either vertically (i.e., adding positional information about more tuples of already partially indexed attributes) or horizontally (i.e., adding positional information about currently non-indexed attributes). Figure 2 shows an example of a positional map, where the attributes do not necessarily appear in the map in the same order as in the raw file. The positional map does not mirror the raw file. Instead, it adapts to the workload, keeping in the same chunk attributes accessed together during query processing.

Exploiting the positional map. The information contained in the positional map can be used to jump to the exact position of the file or as close as possible. For example, if a query is looking for the 9th attribute of a file, while the map contains information for the 4th and the 8th attribute, PostgresRaw uses the positional map to jump to the 8th attribute and parse it until it finds the 9th attribute.

Maintenance. The positional map is an auxiliary structure and may be dropped fully or partly at any time without any loss of critical information; the next query simply starts rebuilding the map from scratch. PostgresRaw assigns a storage threshold for the size of the positional map such that the map fits comfortably in memory. Once the storage threshold is reached, PostgresRaw drops parts of the map to ensure it is always within the threshold limits.

Adaptive behavior. The positional map is an adaptive data structure that continuously indexes positions based

Figure 2. An example of indexing raw files with positional map.



on the most recent queries. This includes requested attributes as well as patterns, or combinations, in which those attributes are used. As the workload evolves, some attributes may no longer be relevant and are dropped by a least recently used (LRU) policy. Similarly, combinations of attributes used in the same query, which are also stored together, may be dropped to give space for storing new combinations. Populating the map with new combinations is decided during pre-fetching, depending on where the requested attributes are located on the current map. The distance that triggers indexing of a new attribute combination is a PostgresRaw parameter. In our prototype, the default setting is that if all requested attributes for a query belong in different chunks, then the new combination is indexed.

3.3. Caching

The positional map allows for efficient access of raw files. An alternative and complementary direction is to avoid raw file access altogether. Therefore, PostgresRaw also contains a cache that temporarily holds previously accessed data, for example, a previously accessed attribute or even parts of an attribute. If the attribute is requested by future queries, PostgresRaw will read it directly from the cache.

The cache holds binary data and is populated on-the-fly during query processing. To minimize the parsing costs and to maintain the adaptive behavior of PostgresRaw, caching does not force additional data to be parsed, that is, only the requested attributes for the current query are transformed to binary. The cache follows the format of the positional map such that it is easy to integrate it in the PostgresRaw query flow, allowing queries to seamlessly exploit both the cache and the positional map in the same query plan. The size of the cache is a parameter than can be tuned depending on the resources. PostgresRaw follows the LRU policy to drop and populate the cache. Overall, the PostgresRaw cache can be seen as the place holder for adaptively loaded data.

3.4. Statistics

Optimizers rely on statistics to create good query plans. Most important plan choices depend on the selectivity estimation that helps ordering operators such as joins. Creating statistics in modern databases, however, is only possible after data is loaded.

We extend the PostgresRaw *scan* operator to create statistics on-the-fly. We carefully invoke the native statistics routines of the DBMS, providing it with a sample of the data. Statistics are then stored and are exploited in the same way as in conventional DBMS. In order to minimize the overhead of creating statistics during query processing, PostgresRaw creates statistics only on requested attributes, that is, only on attributes that PostgresRaw needs to read and which are required by at least the current query.

On-the-fly creation of statistics brings a small overhead on the PostgresRaw *scan* operator, while allowing PostgresRaw to implement high-quality query execution plans.

4. EXPERIMENTAL EVALUATION

In this section, we present an experimental analysis of PostgresRaw. PostgresRaw is implemented on top of

PostgreSQL 9.0, thus the direct comparison between the two systems is important to understand the impact of in situ querying. We have to point out that PostgresRaw is highly affected by any performance bottlenecks present in PostgreSQL, since they share the same query engine.

All experiments are conducted in a Sun X4140 server with 2× Quad-Core AMD Opteron processor (64 bit), 2.7 GHz, 512KB L1 cache, 2MB L2 cache and 6MB L3 cache, 32GB RAM, 4× 250GB 10,000 RPM SATA disks (RAID-0) and using Ubuntu 9.04.

The experiments presented in this section, use a raw data file of 11GB, containing 7.5×10^6 tuples. Each tuple contains 150 attributes with integers distributed randomly in the range $[0-10^9]$.

4.1. Positional map

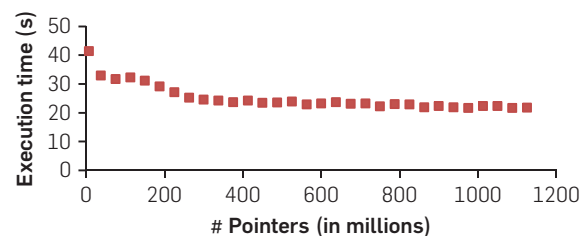
Impact. The first experiment investigates the impact of the positional map. In particular, we investigate how the behavior of PostgresRaw is affected as the map is populated dynamically with positional information based on the workload.

The set up of the experiment is as follows. We create a random set of queries accessing a subset of the attributes found in the raw file. We refer to queries as random, because they may ask for any attribute. Each query asks for 10 random attributes and retrieves all the rows of the file. We measure the average time PostgresRaw needs in order to process all queries with a varying storage capacity for the positional map, from 14.3MB up to 2.1GB.

The results are shown in Figure 3. The impact of the positional map is significant as it eventually improves response times by more than a factor of 2. In addition, performance improves rapidly, not requiring the maximum capacity. With little less than the 1/4 of the pointers (260 million positions) collected, execution time is already only 15% from the full indexed case. After 3/4 of the pointers are collected, response time remains constant even though the workload is random. Therefore, PostgresRaw does not need to maintain positional information for the complete raw file, thereby saving significant storage and access costs, without compromising performance.

Scalability. The next experiment investigates the scalability of PostgresRaw when exploiting the positional map. The set up is the same as in the previous experiment with the difference that this time the file size is increased gradually from 2GB to 92GB. We use two ways to increase

Figure 3. Effect of the number of pointers in the positional map.



the file size; first, by adding more attributes to the file and second, by appending more rows to the file. In the first case, queries remain the same as before. In the second case, queries incrementally access more attributes as we increase the file size. We ensure that for every case we compare, queries perform similar I/O and computation actions. We allow unlimited storage space for the positional map. Nevertheless, we store only positions accessed by the most recent queries.

Figure 4 depicts the results. For both cases we observe linear scalability; PostgresRaw exploits the positional map to nicely scale as raw files grow both vertically and horizontally.

4.2. Positional maps and caching

This experiment investigates the behavior of PostgresRaw when exploiting both the positional map and caching or only one of them. We create 50 queries, where each query randomly accesses five columns and all the rows of the raw file. We study four variations. The first one, called Baseline, does not use positional maps or caching, representing the behavior of PostgresRaw as if it were a straw-man external files implementation. The second variation, called PostgresRaw PM, uses only the positional map while the third, called PostgresRaw C, uses only the cache and an additional minimal map with positional information for the end of lines. The final version, called PostgresRaw PM+C, combines all previous techniques.

Figure 5 plots the response time for each query. Since there is no a priori knowledge to exploit, all PostgresRaw variations need to touch the raw file to extract the needed data for the first query; thus, they all show similar performance. Performance improves drastically as of the second

query. When the cache and the positional map are enabled the second query is 82–88% faster than the first. The Baseline variation improves slightly mainly due to file system caching and from there on it provides constant performance, which is not competitive with the other variations; every query needs to scan the raw file without any help from indexing and caching.

When only the positional map is used, the first few queries collect metadata information, improving future attribute retrieval by minimizing the parsing and tokenizing costs. The rest of the queries benefit from this information, demonstrating improved and stable performance. The positional map allows PostgresRaw to navigate as close as possible to the required attributes, which is important particularly when a parse the raw file, which increases the overall execution time (3–5 times in this example). Figure 5 shows that the combined effects of the positional map and caching achieve the best performance; PostgresRaw PM+C outperforms all other approaches across the entire query sequence.

4.3. Adapting to workload changes

In this experiment, we demonstrate that PostgresRaw progressively and transparently adapts to changes in the workload. We use the same raw file as in the previous experiments but the query sequence is expanded to 250 queries. Each query again refers to five random attributes of the file. The query sequence is divided into five epochs and in each epoch we execute 50 different queries. All queries within the same epoch focus on a given part of the raw file. The maximum size of the cache is limited to 2.8GB, while the positional map does not exceed 715MB.

Figure 6 depicts the results, separating each epoch with vertical lines at positions 50, 100, ..., 200. The graph plots both the response time for each query in the sequence and how the size of the PostgresRaw cache evolves as queries are evaluated.

During the first epoch, queries refer to columns 1–50. The cache and the positional map are initially empty. After executing 32 queries all data in this part of the file is cached; the cache does not increase and performance remains stable. In the second epoch, queries retrieve data between columns 51–100. Performance fluctuates as some queries can fully exploit the cache and have faster response times while others need to go back to

Figure 4. Scalability of the positional map.

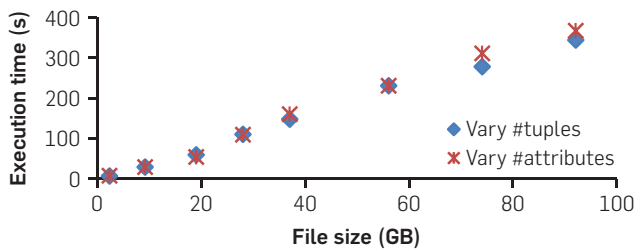


Figure 5. Effect of the positional map and caching.

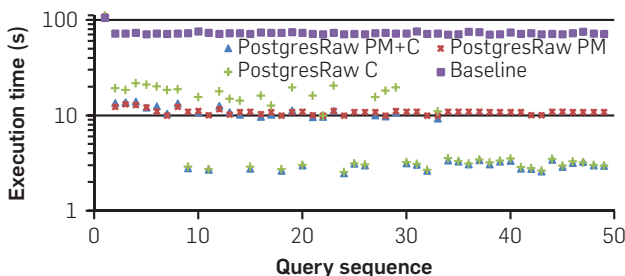
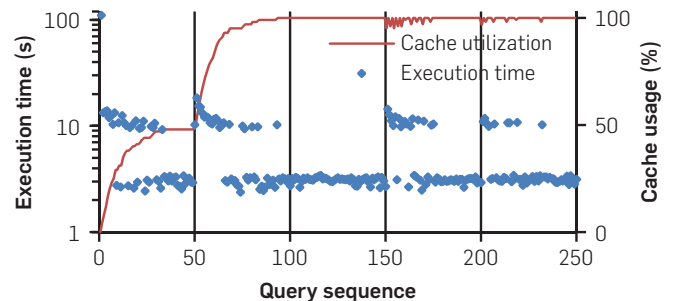


Figure 6. Adapting to changes in the workload.



the file. After the second epoch, the cache is full and all queries enjoy good performance. During the third epoch, we launch a random set of queries requesting columns between 1 and 100, that is, the same regions used in the previous epochs. Since PostgresRaw has built a complete cache of this region, no I/O or parsing is required. In the fourth epoch, queries ask for columns 75–125, that is, half of the queries hit previously explored areas and half of the queries hit new regions. PostgresRaw uses a LRU replacement policy in its cache and drops previously cached data to accommodate the new requests. During the last epoch, the workload slightly shifts to the region of columns 85–135. PostgresRaw needs to replace parts of its cache while parts of the requested data are retrieved from the file by exploiting the positional map.

Overall, we observe that PostgresRaw gracefully adapts to the changes of the workload. In every epoch, PostgresRaw quickly adapts, adjusting and populating its cache and the positional maps, automatically stabilizing to good performance levels. Additionally, the maintenance of the cache and the positional map do not add significant overhead to query execution.

4.4. PostgresRaw versus other DBMS

In our next experiment, we demonstrate the behavior of PostgresRaw against state-of-the-art DBMS. We compare MySQL (5.5.13), DBMS X (a commercial system) and PostgreSQL against PostgresRaw with positional maps and caching enabled. MySQL and DBMS X offer “external files” functionality, which enables direct querying over raw files. Therefore, for MySQL and DBMS X we include two sets of performance results; (a) using external files, and (b) using previously loaded data. For queries over loaded data we also report the time required to load the data; our goal is to show the overall data-to-query time.

We study the cumulative time needed to run a sequence of queries with each system. We use a sequence of nine queries where we also vary selectivity and projectivity. All queries have one selection predicate and then project and run aggregations on the rest of the attributes. The first query requires all attributes and accesses all rows of the file. This is the worst case for PostgresRaw since we have to pay the whole cost of populating the positional map and the cache up front. The next four queries are the same with the difference that they access fewer rows at steps of 20% at a time. Then, the final four queries are again similar to the first query with

the difference that they refer to fewer attributes at steps of 20% at a time.

Figure 7 shows the results. PostgresRaw achieves the best overall performance. It is competitive with DBMS X and MySQL for this sequence of queries. External files in MySQL (CSV Engine) and DBMS X are significantly slower than querying over loaded data or PostgresRaw, since each query repeatedly scans the entire file. Conventional wisdom indicates that the overhead inherent to in situ querying is problematic. This is indeed the case for straightforward in situ techniques such as external files. Nonetheless, these results show that the in situ overhead is not a bottleneck if we apply more advanced techniques that amortize the overhead across a sequence of queries, allowing for quick access to the data. Compared to PostgreSQL, PostgresRaw shows a significant advantage (25.75% in this case) even though it uses the same query engine. PostgreSQL is 53% slower than DBMS X if we consider the query execution time (without the loading costs). PostgresRaw, on the other hand, manages to be 6% faster than DBMS X even though it uses the same engine as PostgreSQL; by avoiding the loading costs, PostgresRaw has already answered the first four queries when DBMS X starts processing the first query.

Overall, PostgresRaw shows that it is feasible to amortize the overheads inherent to in situ querying over a sequence of queries, making an in situ system competitive with a conventional DBMS without requiring a priori data loading.

4.5. Statistics in PostgresRaw

In our final experiment, we demonstrate the behavior of PostgresRaw when statistics are created on-the-fly during query processing. We use four instances of TPC-H decision support benchmark Query 1. We compare two versions of PostgresRaw. The first one generates statistics on-the-fly in an adaptive way, while the second one does not generate or exploit statistics at all.

Figure 8 shows the response times when running all four queries. The first query uses the same plan in both versions of PostgresRaw and initializes the positional map and the caching as well. Collecting statistics adds an additional overhead of 4.5 s in the execution time of the first query. PostgresRaw analyzes and creates statistics only for the attributes required for the current query. After the first query, the rest of the queries have different

Figure 7. Comparing the performance of PostgresRaw with other DBMS.

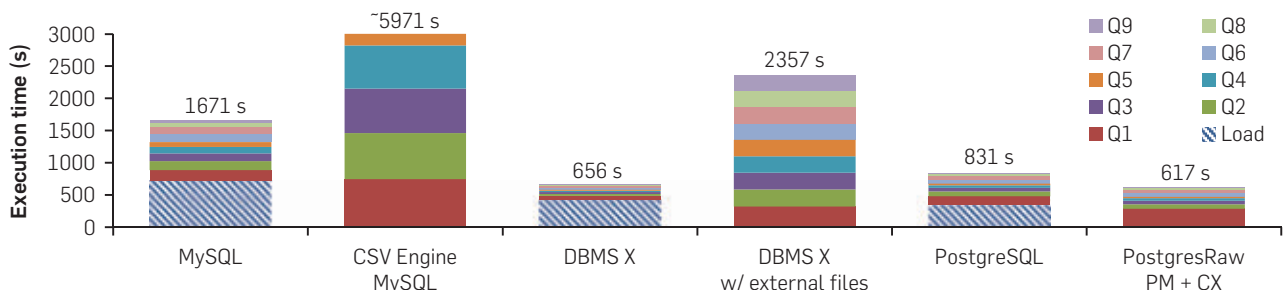
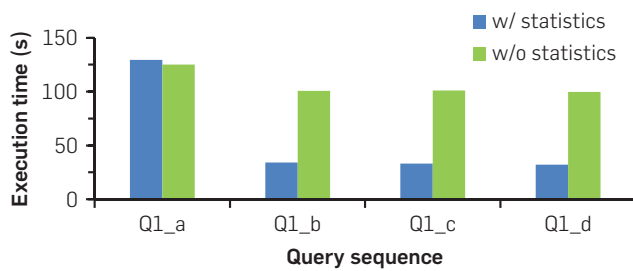


Figure 8. Execution time as PostgresRaw generates statistics.



behavior even though they follow the same query template. In the PostgresRaw version with statistics support, queries run three times faster in comparison with the version without statistics. By examining the query plans, we notice that the optimizer selects a different set of operators and changes the ordering of operators in PostgresRaw with statistics which explains the improvement in performance. Generating the statistics on-the-fly adds only a small overhead, while it significantly improves query plan selection.

5. IN SITU QUERYING: TRADE-OFFS

In situ querying, although desirable in theory, is thought to be prohibitive in practice. Executing queries directly over raw data files incurs additional overhead to the execution, when compared to query execution over previously loaded data. Nonetheless, our PostgresRaw implementation demonstrates that auxiliary structures reduce the time to access raw data files and amortize the overhead across a sequence of queries. In situ query execution, however, introduces a new set of trade-offs, which require further analysis:

Data type conversion. For ASCII files, PostgresRaw must convert the data into its proper type, for example, from string to integer. Conventional DBMS perform this conversion only once at loading time. To alleviate the data type conversion overhead, PostgresRaw only converts the attributes in the tuple that are actually needed to answer a query. Nonetheless, data type conversion is not always an overhead: if a raw data file consists of variable-length strings, then PostgresRaw over CSV files is actually faster than a conventional DBMS because there is no need to convert data nor create secondary copies when loading data into a DBMS. Different data types, however, affect NoDB performance in different ways and should be taken into account when deciding which data to cache.

File size versus database size. Loading data into a DBMS creates a second copy of the data. This copy can be stored in an optimized manner: for example, integers stored in a database page (in binary) likely take less space than in ASCII. Nonetheless, there are cases where a second copy does not imply less data. Variable-sized data stored in fixed-size fields usually takes more space in a database page rather than in its raw form. Therefore, depending on the workload, in situ engines can benefit from keeping data in its raw form.

Complex database schemas. Database Management Systems support complex database schemas with large number of tables and columns within a table. Nonetheless, complex schemas usually require a database administrator (DBA) to tune vendor-specific configuration settings. For instance, a commercial DBMS we tested does not allow a row to be split across pages; if there are many columns within a table, or columns have large fields, the DBA must manually increase the page size, buffer pool and table space. These configurations are not straightforward and are also subjected to additional limitations: for example, pages must also have a minimum number of rows. In addition, larger tuples cause unpredictable behavior due to the use of slotted pages in the DBMS.

Types of data analysis. Current DBMS are best suited to manage data that is loaded only once or rarely in an incremental fashion, with well-known and rarely changing workloads. DBMS require physical design steps for best performance, such as creating indexes, which are time-consuming tasks. In situ databases, however, are more suited for users that need to explore data without having to load entire datasets. Users should be willing to pay a penalty during the early queries, as long as they do not need to create data loading scripts. In situ databases are also useful when there are large datasets but users need to frequently analyze small subsets of the data; such scenarios are increasingly common.

Integration with external tools. Database Management Systems are designed to be the main repository for the data, which makes the integration of DBMS data with external tools inherently hard. Techniques such as ODBC, stored procedures and user-defined functions aim to facilitate the interaction with data stored on the DBMS. Nonetheless, none of these techniques is fully satisfactory and in fact, this is a common complaint of scientific users, who have large repositories of legacy code that operates against raw data files. Migrating and reimplementing these tools in a DBMS would be difficult and likely require vendor-specific hooks. The NoDB philosophy significantly facilitates such data integration, since users may continue to rely on their legacy code in parallel to systems such as PostgresRaw.

Database independence. Database Management Systems store data in database pages using proprietary and vendor-specific formats. The DBMS has complete ownership over the data, which is a cause of concern for some users. The NoDB philosophy, however, achieves database independence, since the data files remain the main data repository.

6. OPPORTUNITIES

The NoDB philosophy drastically and fundamentally redefines the way database systems are designed. It requires revisiting well-established assumptions and implementation techniques, while also enabling new opportunities, which are discussed in this section.

Flexible storage. NoDB systems do not require a priori loading, which implies no need for a priori decisions on how data is physically organized during loading. Data that is adaptively loaded can be cached in memory or written to disk in a format that enables faster access. Data compression

can also be applied, where beneficial. Deciding the proper storage layout is an open research question. Rows, columns, and hybrids all have comparative advantages and disadvantages. Nevertheless, a NoDB system benefits from avoiding to choose in advance. Physical layout decisions can be done online, and change overtime as the workload changes.³

Adaptive indexing. The NoDB philosophy brings new opportunities toward achieving fully autonomous database systems, that is, systems that require zero initialization and administration. Recent efforts in database cracking and adaptive indexing^{7, 10, 11, 13} demonstrate the potential for incrementally building and refining indexes without requiring an administrator to tune the system, or knowing the workload. Still, though, all data has to be loaded up front, forcing a delay in data-to-query time. We envision that adaptive indexing can be exploited and enhanced for NoDB systems.

Auto-tuning tools. In this paper, we have considered the hard case of zero a priori idle time or workload knowledge. Traditional systems assume “infinite” idle time and knowledge to perform all necessary initialization steps. In many cases, though, the reality can be somewhere in between. For example, there might be some idle time but not enough to load all data. Auto-tuning tools for NoDB systems, given a budget of idle time and workload knowledge, can exploit idle time to load and index as much of the relevant data. The rest of the data remains unloaded and unindexed until relevant queries arrive. A NoDB tuning tool should consider raw data access costs, I/O costs in addition to the typical query workload based parameters. The NoDB philosophy brings new opportunities in exploiting every single bit of idle time or workload knowledge.

Information integration. Another major opportunity with the NoDB vision is the potential to query multiple different data sources and formats. NoDB systems can adopt format-specific plugins to handle different raw data file formats. Implementing these plugins in a reusable manner requires applying data integration techniques but may also require the development of new techniques, so that commonalities between formats are determined and reused. Additionally, supporting different file formats also requires the development of hybrid query processing techniques, or even adding support for multiple data models (e.g., for array data).

File system interface. Another interesting opportunity that comes with NoDB is that of bridging the gap between file systems and databases. Unlike traditional database systems, data in NoDB systems is always stored in file systems, such as NTFS or ext4. This provides NoDB the opportunity to intercept file system calls and gradually create auxiliary data structures that speed-up future queries.

7. RELATED WORK

The NoDB philosophy draws inspiration from several decades of research on database technology and it is related to a plethora of research topics. We briefly discuss related work in this section.

Auto-tuning. The NoDB philosophy advocates for minimizing or eliminating the data-to-query time, which is also

the goal of auto-tuning tools. Every major database vendor offers offline indexing features, where an auto-tuning tool performs offline analysis to determine the proper physical design for a specific workload.^{1, 6, 18, 22} More recently, these ideas have been extended to support online indexing,^{4, 20} hence removing the need to know the workload in advance. These techniques are a significant step forward, but still require all data to be loaded in advance.

Adaptive indexing. Database cracking and adaptive indexing introduce the notion of incrementally refining the physical design by following and matching the workload patterns.^{7, 10, 11, 13} This shares the adaptive goal of the NoDB philosophy, where each query is seen as an advice on how to refine indexes. Nonetheless, similarly to the previous case, existing adaptive indexing techniques also require all data to be loaded up front.

External files. Most modern DBMS offer the ability to query data files directly with SQL, that is, without loading. External files, however, can only access raw data with no support for database features such as DML operations, indexes or statistics and require every query to access the entire data file, as if no other query did so in the past. In practice, this functionality is used to facilitate data loading tasks and not for querying. NoDB systems, however, provide incremental data loading, on-the-fly index creation and caching to assist future queries and drastically improve performance.

Information extraction. Information extraction techniques have been extended to provide direct access to raw text data,¹⁵ similarly to external files. The difference from external files is that raw data access relies on information extraction techniques instead of directly parsing raw data files. These efforts are motivated by the need to bridge multiple different data formats and make them accessible via SQL, usually by relying on wrappers.¹⁹

8. CONCLUSION

Very large data processing is increasingly becoming a necessity for modern applications in businesses and in sciences. For state-of-the-art database systems, the incoming data deluge is a problem. In this paper, we introduce a database design philosophy that turns the data deluge into a tremendous opportunity for database systems. It requires drastic changes to existing query processing technology but eliminates one of the most fundamental bottlenecks present in classical database systems for the past 40 years, that is, the data loading overhead. Until now, it has not been possible to exploit database technology until data is fully loaded. NoDB systems permanently remove this restriction by enabling in situ querying.

This article described the NoDB philosophy, identifies problems, solutions and opportunities. It also describes the transformation of a modern row-store, PostgreSQL, into a NoDB prototype system, which we call PostgresRaw. Experiments on PostgresRaw demonstrate competitive performance with traditional DBMS. PostgresRaw, however, does not require any previous assumptions about which data to load, how to load it or which physical design steps to perform before querying the data. Instead, it accesses the

raw data files adaptively and incrementally, allowing users to explore new data quickly and improving the usability of database systems.

The NoDB philosophy does not stop here however. We describe open issues and research challenges for the database community at large. We expect that addressing these new challenges will enable a new generation of database systems that serve the needs of modern applications and users.

References

1. Agrawal, S., Chaudhuri, S., Kollar, L., Marathe, A., Narasayya, V., Symala, M. Database tuning advisor for Microsoft SQL server 2005. In *VLDB* (2004), 1110–1121.
2. Ailamaki, A., Kantere, V., Dash, D. Managing scientific data. *Commun. ACM* 53 (2010), 68–78.
3. Alagiannis, I., Idreos, S., Ailamaki, A. H,O: A hands-free adaptive store. In *SIGMOD* (2014), 1103–1114.
4. Bruno, N., Chaudhuri, S. To tune or not to tune? A lightweight physical design alterer. In *VLDB* (2006), 499–510.
5. Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J., Welton, C. MAD skills: New analysis practices for big data. *PVLDB* 2 (2009), 1481–1492.
6. Dash, D., Polyzotis, N., Ailamaki, A. CoPhy: A scalable, portable, and interactive index advisor for large workloads. *PVLDB* 4 (2011), 362–372.
7. Graefe, G., Kuno, H. Self-selecting, self-tuning, incrementally optimized indexes. In *EDBT* (2010), 371–381.
8. Gray, J., Liu, D., Nieto-Santesteban, M., Szalay, A., DeWitt, D., Heber, G. Scientific data management in the coming decade. *SIGMOD Rec.* 34 (2005), 34–41.
9. Idreos, S., Alagiannis, I., Johnson, R., Ailamaki, A. Here are my data files. Here are my queries. Where are my results? In *CIDR* (2011).
10. Idreos, S., Kersten, M., Manegold, S. Database cracking. In *CIDR* (2007).
11. Idreos, S., Kersten, M., Manegold, S. Self-organizing tuple reconstruction in column-stores. In *SIGMOD* (2009), 297–308.
12. Idreos, S., Liarou, E. dbTouch: Analytics at your fingertips. In *CIDR* (2013).
13. Idreos, S., Manegold, S., Kuno, H., Graefe, G. Merging what's cracked, cracking what's merged: Adaptive indexing in main-memory column-stores. *PVLDB* 4 (2011), 586–597.
14. Jagadish, H.V., Chapman, A., Elkiss, A., Jayapandian, M., Li, Y., Nandi, A., Yu, C. Making database systems usable. In *SIGMOD* (2007), 13–24.
15. Jain, A., Doan, A., Gravano, L. Optimizing SQL queries over text databases. In *ICDE* (2008), 636–645.
16. Kersten, M., Idreos, S., Manegold, S., Liarou, E. The researcher's guide to the data deluge: Querying a scientific database in just a few seconds. In *PVLDB*, Volume 4 (2011), 1474–1477.
17. Nandi, A., Jagadish, H.V. Guided interaction: Rethinking the query-result paradigm. In *PVLDB*, Volume 4 (2011), 1466–1469.
18. Papadomanolakis, S., Ailamaki, A. AutoPart: Automating schema design for large scientific databases using data partitioning. In *SSDBM* (2004), 383–392.
19. Roth, M.T., Schwarz, P. Don't scrap it, wrap it! A wrapper architecture for legacy data sources. In *VLDB* (1997), 266–275.
20. Schnaitter, K., Abiteboul, S., Milo, T., Polyzotis, N. COLT: Continuous on-line tuning. In *SIGMOD* (2006), 793–795.
21. Stonebraker, M., Becla, J., DeWitt, D., Lim, K.-T., Maier, D., Ratzesberger, O., Zdonik, S. Requirements for science data bases and SciDB. In *CIDR* (2009).
22. Zilio, D., Rao, J., Lightstone, S., Lohman, G., Storm, A., Garcia-Arellano, C., Fadden, S. DB2 design advisor: Integrated automatic physical database. In *VLDB* (2004).

Ioannis Alagiannis, Renata Borovica-Gajic, Miguel Branco, and Anastasia Ailamaki (ioannis.alagiannis, renata.borovica, miguel.branco, anastasia.ailamaki}@epfl.ch), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.

Stratos Idreos (stratos@seas.harvard.edu), Harvard University, Cambridge, MA.

© 2015 ACM 0001-0782/15/12 \$15.00

World-Renowned Journals from ACM

ACM publishes over 50 magazines and journals that cover an array of established as well as emerging areas of the computing field. IT professionals worldwide depend on ACM's publications to keep them abreast of the latest technological developments and industry news in a timely, comprehensive manner of the highest quality and integrity. For a complete listing of ACM's leading magazines & journals, including our renowned Transaction Series, please visit the ACM publications homepage: www.acm.org/pubs.

ACM Transactions on Interactive Intelligent Systems



ACM Transactions on Interactive Intelligent Systems (TIIS). This quarterly journal publishes papers on research encompassing the design, realization, or evaluation of interactive systems incorporating some form of machine intelligence.

ACM Transactions on Computation Theory



ACM Transactions on Computation Theory (ToCT). This quarterly peer-reviewed journal has an emphasis on computational complexity, foundations of cryptography and other computation-based topics in theoretical computer science.

PLEASE CONTACT ACM MEMBER SERVICES TO PLACE AN ORDER
 Phone: 1.800.342.6626 (U.S. and Canada) +1.212.626.0500 (Global)
 Fax: +1.212.944.1318 (Hours: 8:30am–4:30pm, Eastern Time)
 Email: acmhelp@acm.org
 Mail: ACM Member Services
 General Post Office
 PO Box 30777
 New York, NY 10087-0777 USA



Association for Computing Machinery

Advancing Computing as a Science & Profession

www.acm.org/pubs

CAREERS

Boise State University **Department of Computer Science** *Eight Open Rank, Tenured/Tenure-Track Faculty Positions*

The **Department of Computer Science** at **Boise State University** invites applications for **eight** open rank, tenured/tenure-track faculty positions. Seeking applicants in the areas of big data (including distributed systems, HPC, machine learning, visualization), cybersecurity, human computer interaction and computer science education research. Strong applicants from other areas of computer science will also be considered.

Applicants should have a commitment to excellence in teaching, a desire to make significant contributions in research, and experience in collaborating with faculty and local industry to develop and sustain funded research programs. A PhD in Computer Science or a closely related field is required by the date of hire. For additional information, please visit <http://coen.boisestate.edu/cs/jobs>.

Boston College **Computer Science Department** *Tenure Track Faculty Position*

The **Boston College Computer Science Department** invites applications for a full-time non-tenure track faculty position, beginning September 2016.

The position is a three-year renewable contract. Applicants should have a Ph.D. in Computer Science or related discipline, and possess a strong commitment to undergraduate teaching. We will begin reviewing applications on December 1, 2015, and will continue considering applications until the position is filled. Additional information is available at www.cs.bc.edu/employment.

Boston University **Department of Electrical & Computer Engineering (ECE)** *Assistant Professor*

The **Department of Electrical & Computer Engineering (ECE)** at **Boston University (BU)** is seeking candidates for a tenure-track **Assistant Professor** Position in the general domain of computer systems and software, motivated by areas such as the Internet of Things (IoT), cybersecurity, privacy, and the cloud. Please visit <http://www.bu.edu/ece/facultysearch> for instructions on how to apply. Candidates must possess a relevant Ph.D. degree, show strong potential for attracting external research funding, and possess a strong commitment to teaching. This ECE position is aligned with a broader university-level initiative in Data Science – an area of strategic growth for Boston University (<http://www.bu.edu/datascience/>).

The application deadline is December 31, 2015. The review of applications will begin on October 1, 2015.

Boston University is an Equal Opportunity/Affirmative Action Employer and 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.

We are a VEVRAA Federal Contractor.

Bradley University **Computer Science and Information Systems Department** *Assistant Professor*

The Computer Science and Information Systems Department at **Bradley University** invites applications for a **tenure-track Assistant Professor** position starting in August 2016. The tenure-track Assistant Professor requires a PhD in Computer Science or a closely related field; candidates working on their dissertation with anticipated completion date before August 2016 will be considered. Please visit www.bradley.edu/humanresources/opportunities for full position description and application process.

Cal Poly **Electrical Engineering** *Assistant or Associate Professor - Electrical Engineering and Aerospace Engineering*

The Electrical Engineering Department and Aerospace Engineering Department at Cal Poly, SLO, invite applications for a full-time tenure-track faculty position at the Assistant or Associate Professor rank for Fall 2016. Duties include teaching coursework in Electrical Engineering and Aerospace Engineering, building a collaborative research program in the area of Satellite Communication and Mobile Terrestrial Communications. Desired areas of expertise include satellite electronic systems including the Cube Sat form factor, and mobile terrestrial communication systems. For details, qualifications, and application instructions (online application required), visit WWW.CALPOLYJOBS.ORG and apply to requisition #103796. Application review begins Jan. 4, 2016. EEO.

Cal Poly **Electrical Engineering** *Assistant or Associate Professor - Electrical & Computer Engineering*

The Electrical Engineering Department and Computer Engineering Program at Cal Poly, SLO, invite applications for a full-time, tenure-track faculty position at the Assistant or Associate Professor rank for Fall 2016. Duties include teaching undergraduate and graduate computer/electrical engineering courses and building a collaborative research program in the area of mobile comput-

ing. Teaching responsibilities include digital design, computer architecture, embedded systems, design projects, technical electives, and graduate courses. Desired areas of expertise include computer engineering, and hardware and firmware aspects of mobile computing platforms. For details, qualifications, and application instructions (online application required), visit WWW.CALPOLYJOBS.ORG and apply to requisition #103795. Application review begins Jan. 4, 2016. EEO.

California State University, Fullerton **Department of Computer Science** *Assistant Professor*

The Department of Computer Science invites applications for tenure-track positions at the **Assistant Professor** level starting August 2016. For a complete description of the department, the position, desired specialization and other qualifications, please visit <http://hr.fullerton.edu/diversity/job-openings/>.

Columbia University **Department of Computer Science** *Faculty Positions*

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: Theory (all levels) and Programming languages (Assistant or Associate Professor level).

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, 2015 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.

Dartmouth College **Department of Computer Science** **Assistant Professor of Computer Science**

The Dartmouth College Department of Computer Science invites applications for two tenure-track faculty positions at the level of assistant professor. We seek candidates who will be excellent researchers and teachers in the areas of: (1) **security or machine learning**; and (2) **theoretical computer science**; although outstanding candidates in any area will be considered. We particularly seek candidates who will help lead, initiate, and participate in collaborative research projects both within Computer Science and involving other Dartmouth researchers, including those in other Arts & Sciences departments, Dartmouth's Geisel School of Medicine, Thayer School of Engineering, and Tuck School of Business. A Ph.D. degree or ABD in Computer Science or a closely related field is required.

The department is home to 20 tenured and tenure-track faculty members and two research faculty members. Research areas of the department encompass the areas of security, computational biology, machine learning, robotics, systems, algorithms, theory, digital arts, vision, and graphics. The Computer Science department is in the School of Arts & Sciences, and it has strong Ph.D. and M.S. programs and outstanding undergraduate majors. The department is affiliated with Dartmouth's M.D.-Ph.D. program and has strong collaborations with Dartmouth's other schools.

Dartmouth College, a member of the Ivy League, is located in Hanover, New Hampshire (on the Vermont border). Dartmouth has a beautiful, historic campus, located in a scenic area on the Connecticut River. Recreational opportunities abound in all four seasons. With an even distribution of male and female students and over one third of the undergraduate student population members of minority groups, Dartmouth is committed to diversity and encourages applications from women and minorities.

To create an atmosphere supportive of research, Dartmouth offers new faculty members grants for research-related expenses, a quarter of sabbatical leave for each three academic years in residence, and flexible scheduling of teaching responsibilities.

Applicants are invited to submit application materials via Interfolio at <http://apply.interfolio.com/31847> (for the Security/ML position) or <http://apply.interfolio.com/31850> (for the Theory position). Upload a CV, research statement, and teaching statement, and request at least four references

to upload letters of recommendation, at least one of which should comment on teaching. Email facsearch16@cs.dartmouth.edu with any questions.

Dartmouth is an equal opportunity/affirmative action employer with a strong commitment to diversity. In that spirit, we are particularly interested in receiving applications from a broad spectrum of people, including women, persons of color, persons with disabilities, veterans or any other legally protected group.

Application review will begin November 1, 2015, and continue until the position is filled.

Florida Institute of Technology **CS Faculty Positions**

Florida Tech invites applications for open-rank faculty positions in computer science that begin in Fall 2016. All areas of CS are considered. The CS department offers BS, MS, and PhD degree programs. Florida Tech is in the Space Coast, where NASA Kennedy Space Center is located. According to Brookings Institute's report on "America's Advanced Industries," Space Coast ranks 7th in Advanced Industries' Share of Total 2013 Employment. **Submission of applications and more information are at:** cs.fit.edu/careers/

Equal Opportunity Employer Minorities/
Women/Veterans/Disabled

We are an E-Verify employer

EEO is the Law

http://www.fit.edu/hr/documents/eeoc_law.pdf

Florida Institute of Technology **Faculty position in Software Engineering**

The Department of Computer Sciences & Cybersecurity at the Florida Institute of Technology invites applications for an open faculty position in Software Engineering, beginning in Fall 2016. The department is ABET accredited in both Computer Science and Software Engineering. Required qualifications for the position include an earned Ph.D. with a specialization in software engineering, evidence of the ability to develop and sustain an active research program and a sincere interest in quality teaching, at both the undergraduate and graduate levels. Our current software engineering strengths are in testing, cybersecurity, maintenance and evolution, but we welcome applicants from all areas. Our preference is for faculty members who conduct research that is both pragmatic and academically rigorous. The Department has significant active research funding from multiple government agencies and commercial companies. New faculty will have the opportunity to work with the Harris Institute for Assured Information. Florida Tech is a NSA/DHS designated Center for Academic Excellence in Information Assurance Research. Florida Tech is located in Melbourne on Florida's Space Coast, one of the nation's fastest-growing high-tech areas. The campus is 5 minutes from the Indian River estuary, 10 minutes from the Atlantic Ocean and 50 minutes from Kennedy Space Center and Orlando. For more information on the Department of Computer Sciences please visit our web-

IOWA STATE UNIVERSITY

Assistant or Associate Professor **Software Engineering Program**

The Software Engineering Program, jointly administered by the Departments of Electrical and Computer Engineering (ECpE) and Computer Science (CompSci) at Iowa State University, Ames, IA, invites applications for the position of Assistant or Associate Professor in the area of software engineering, particularly as it relates to cyber-security or large scale data analysis.

The Department of Electrical and Computer Engineering and the Department of Computer Science have strong graduate programs. Almost all Ph.D. students are supported by research or teaching assistantships. These well-funded research programs provide an excellent academic environment in which to work and make career progress. In addition, the cutting-edge research and educational activities are nurtured through interdisciplinary interactions facilitated by the Laurence H. Baker Center for Bioinformatics and Biological Statistics; the Center for Computational Intelligence, Learning and Discovery; the Center for Integrative Animal Genomics; the Cyber Innovation Institute; the Information Assurance Center; the Virtual Reality Applications Center; the Center for Nondestructive Evaluation; and the U.S. Department of Energy's Ames Laboratory, all of which are located on the Iowa State University campus.

Iowa State University is an Equal Opportunity/Affirmative Action employer. All qualified applicants will receive consideration for employment without regard to race, color, age, religion, sex, sexual orientation, gender identity, genetic information, national origin, marital status, disability, or protected veteran status, and will not be discriminated against. Inquiries can be directed to the Director of Equal Opportunity, 3350 Beardshear Hall, 515-294-7612.

All interested, qualified persons are encouraged to apply early and must apply for this position by visiting <http://www.iastatejobs.com/postings/14051> and complete the Employment Application for vacancy #500149. Inquiries regarding the faculty search should be directed to Professor Johnny Wong at wong@iastate.edu or (515) 294-2586. For full consideration, applications must be received by Dec. 15, 2015.

Iowa State University is an Equal Opportunity/Affirmative Action Employer.

site (<http://cs.fit.edu>). Information on the Harris Institute is also available online (<http://harris-institute.fit.edu/>). Applicants should send letters of intent, curriculum vitae, research and teaching summaries and full contact information for at least three references, via email (swe-faculty-search@fit.edu).

Review of applications will begin in November and continue until the position is filled.

* Official transcripts of all collegiate work must be sent directly from the attended institution to the Human Resources Office prior to the first day of employment. All international degrees must have a course-by-course official evaluation and translation sent to the Human Resources Office directly from an evaluation company affiliated with the National Association of Credential Evaluation Services, Inc. (NACES).

Equal Opportunity Employer Minorities/Women/Veterans/Disabled

We are an E-verify employer
EEO is the Law

Fordham University **Department of Computer and Information Science**

Postdoctoral Position in Cybersecurity

We invite applications for a postdoctoral researcher position in Cybersecurity in the Department of Computer and Information Science of Fordham University.

This position requires a Ph.D. in Computer Science or a closely related area with a specializa-

tion in cybersecurity. Those with expertise in network security, software security, cyber-physical systems, computer forensics, wireless security, biometrics-based security, e-Systems Security, and other related areas are encouraged to apply. This postdoctoral researcher will conduct high-quality research in cybersecurity and teach courses for MS program in Cybersecurity.

Send a detailed CV, research statement, teaching statement, and info of 3 references in a single pdf file to sajalbhatia@gmail.com. Include job-code **"POSTDOC-Cybersecurity"** in subject line. Review of applications will begin immediately and continue until the position is filled.

For more details, visit: <http://www.cis.fordham.edu/events/CybersecPostdoc.pdf>

Fordham University **CIS Department** ***Tenure-track Assistant Professors***

Fordham University invites applications for two tenure track Assistant Professor Positions in the CIS Department, to start in fall 2016. The two positions require a Ph.D. in Computer Science, Information Science or related fields, a commitment to teaching excellence, and an active program of research. One of the positions is in Cybersecurity and the other in Data Analytics.

These selected candidates are expected to teach graduate and undergraduate courses in Computer and Information Science, and conduct high-quality research.

For information about the department, visit

<http://www.cis.fordham.edu>.
Applications can be electronically submitted to Interfolio Scholar Services:
For Cybersecurity Position:
apply.interfolio.com/31854
For Data Analytics Position:
apply.interfolio.com/31855

Include(1) Cover letter with qualifications,(2) Curriculum vitae,(3) Research Statement,(4) Teaching Statement,(5) Sample scholarship, and(6) At least three letters of recommendation. Applications will be accepted until the position is filled. Preference will be given to applications received by January 15, 2016.

For inquiries, contact: Palma Hutter at: hutter@fordham.edu.

Fordham University, an independent, Catholic University in the Jesuit tradition, is committed to excellence through diversity and welcomes candidates of all backgrounds. Fordham is an Equal Opportunity Employer.

Harvard University **John A. Paulson School of Engineering and Applied Sciences** ***Tenure-track Faculty Position in Computer Science***

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, 2016.

This is a broad faculty search and we welcome



Florida International University is a comprehensive university offering 340 majors in 188 degree programs in 23 colleges and schools, with innovative bachelor's, master's and doctoral programs across all disciplines including medicine, public health, law, journalism, hospitality, and architecture. FIU is Carnegie-designated as both a research university with high research activity and a community-engaged university. Located in the heart of the dynamic south Florida urban region, our multiple campuses serve over 55,000 students, placing FIU among the ten largest universities in the nation. Our annual research expenditures in excess of \$100 million and our deep commitment to engagement have made FIU the go-to solutions center for issues ranging from local to global. FIU leads the nation in granting bachelor's degrees, including in the STEM fields, to minority students and is first in awarding STEM master's degrees to Hispanics. Our students, faculty, and staff reflect Miami's diverse population, earning FIU the designation of Hispanic-Serving Institution. At FIU, we are proud to be 'Worlds Ahead'! For more information about FIU, [visit fiu.edu](http://visit.fiu.edu).

The School of Computing and Information Sciences (SCIS) seeks exceptionally qualified candidates for tenure-track and tenured faculty positions at all levels as well as non-tenure track faculty positions at the level of Instructor, including visiting instructor appointments. SCIS is a rapidly growing program of excellence at the University, with 30 tenure-track faculty members and over 2,000 students, including over 80 Ph.D. students. SCIS offers B.S., M.S., and Ph.D. degrees in Computer Science, an M.S. degree in Telecommunications and Networking, an M.S. degree in Cybersecurity, and B.S., B.A., and M.S. degrees in Information Technology. SCIS has received over \$22M in the last four years in external research funding, has six research centers/clusters with first-class computing and support infrastructure, and enjoys broad and dynamic industry and international partnerships.

Open-Rank Tenure Track/Tenured Positions (Job ID# 508676)

SCIS seeks exceptionally qualified candidates for tenure-track and tenured faculty positions at all levels. We seek well-qualified candidates in all areas; researchers in the areas of computer systems, cybersecurity, cognitive computing, data science, health informatics, and networking are particularly encouraged to apply. Preference will be given to candidates who will enhance or complement our existing research strengths.

Ideal candidates for junior positions should have a record of exceptional research in their early careers. Candidates for senior positions must have an active and proven record of excellence in funded research, publications, and professional service, as well as a demonstrated ability to develop and lead collaborative research projects. 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. An earned Ph.D. in Computer Science or related disciplines is required.

Non-tenure track instructor positions (Job Opening 507474)

We seek well-qualified candidates in all areas of Computer Science and Information Technology. Ideal candidates must be committed to excellence in teaching a variety of courses at the undergraduate level. A graduate degree in Computer Science or related disciplines is required; significant prior teaching and industry experience and/or a Ph.D. in Computer Science is preferred.

HOW TO APPLY:

Qualified candidates for open-rank faculty positions are encouraged to apply to (Job Opening ID #508676); and candidates for instructor positions are encouraged to apply to (Job Opening ID # 507474). **Submit applications at facultycareers.fiu.edu** and attach cover letter, curriculum vitae, statement of teaching philosophy, research statement, etc as individual attachments. Candidates will be required 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 31st, 2015. Review will continue until position is filled.

If you are interested in a visiting appointment please contact the department directly by emailing Dr. Mark Weiss at Weiss@cis.fiu.edu. All other applicants should apply by going to facultycareers.fiu.edu.

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.

outstanding applicants in all areas of computer science, including applicants whose research and interests connect to such areas as engineering, medicine, and the social sciences. We are particularly interested in candidates working in the broad areas of machine learning, human-computer interaction, programming languages, and systems (including networking, architecture, and databases).

The Computer Science program at Harvard University is experiencing a period of strong growth and expansion following an extraordinary gift in support of new faculty from alumnus and former Microsoft CEO Steve Ballmer, '77, and the largest gift in the University's history, received from John A. Paulson, M.B.A. '80, in support of SEAS.

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.

Candidates are required to have a doctorate or terminal degree by the expected start date. In addition, we seek candidates who have a strong research record and a commitment to undergrad-

uate teaching and graduate training.

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 15, 2015, but will continue to review applications until the position is filled. Applicants will apply on-line at <https://academicpositions.harvard.edu/postings/6497>

Harvard is 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.

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 faculty positions in **Intelligent Systems Engineering**. Multiple positions are open at all levels (asst, assoc, or full). Cluster hires are encouraged as are interdisciplinary applications spanning the interests of SoIC and collaborating with IU units. Duties include teaching, research,

and service.

The department of Intelligent Systems Engineering is an innovative new program 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, sensor and detector technologies, signal processing, and information and control theory. **We intend to add about 20 new faculty over the next 4 years covering these areas as well as interdisciplinary thrusts in bioengineering, molecular and nanoscale engineering, environmental engineering and neuro-engineering with an interdisciplinary IT component.** The program will offer a BS and a PhD for students entering in fall 2016, while a future MS will be developed this fall. 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/>. The department will be located in the new SoIC building which will be complete in about 30 months.

Interested candidates should review the application requirements and submit their



IST AUSTRIA
Institute of Science and Technology

Call for

Postdoctoral Fellows in EXECUTABLE BIOLOGY

Executable biology is the study of biological systems as reactive dynamic systems (i.e., systems that evolve with time in response to external events).

Are you a talented and motivated scientist looking for an opportunity to conduct research at the intersection of **BIOLOGY and COMPUTER SCIENCE** at a young, dynamic institution that fosters scientific excellence and interdisciplinary collaboration?

Apply at www.ist.ac.at/executablebiology
Deadline December 31, 2015





上海科技大学
ShanghaiTech University

TENURE-TRACK AND TENURED FACULTY POSITIONS IN INFORMATION SCIENCE AND TECHNOLOGY

The newly launched **ShanghaiTech University** invites talented faculty candidates to fill multiple tenure-track/tenured positions as its core founding team in the School of Information Science and Technology (SIST). Candidates should have outstanding academic records or demonstrate strong potential in cutting-edge research areas of information science and technology. They must be fluent in English. Overseas academic training is highly desired. Besides establishing and maintaining a world-class research profile, faculty candidates are also expected to contribute substantially to graduate and undergraduate education within the school.

ShanghaiTech is matching towards a world-class research university as a hub for training future generations of scientists, entrepreneurs, and technological leaders. Located in a brand new campus in Zhangjiang High-Tech Park of the cosmopolitan Shanghai, ShanghaiTech is at the forefront of modern education reform in China.

Academic Disciplines: We seek candidates in all cutting edge areas of information science and technology that include, but not limited to: computer architecture and technologies, micro-electronics, high speed and RF circuits, intelligent and integrated information processing systems, computations, foundation and applications of big data, visualization, computer vision, bio-computing, smart energy/power devices and systems, next-generation networking, statistical analysis as well as inter-disciplinary areas involving information science and technology.

Compensation and Benefits: Salary and startup funds are internationally competitive, commensurate with experience and academic accomplishment. We also offer a comprehensive benefit package to employees and eligible dependents, including housing benefits. All regular faculty members will be within our new tenure-track system commensurate with international practice for performance evaluation and promotion.

Qualifications:

- Ph.D. (Electrical Engineering, Computer Engineering, Computer Science, or related field)
- A minimum relevant research experience of 4 years.

Applications: Submit (in English, PDF version) a cover letter, a 2-3 page detailed research plan, a CV with demonstrated strong record/potentials; plus copies of 3 most significant publications, and names of three referees to: sist@shanghaitech.edu.cn. For more information, visit <http://www.shanghaitech.edu.cn>.

Deadline: December 31, 2015 (or until positions are filled).

application at: <http://indiana.peopleadmin.com/postings/1900>

For full consideration applications are due by 2/1/16 but earlier submission is encouraged.

Questions regarding the positions or application process can be directed to isechair@indiana.edu or Faculty Search, SoIC, 919 E 10th St, Bloomington, IN 47408.

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.

Indiana University – Purdue University Fort Wayne Department of Computer Science Tenure-Track Faculty Position

The Department of Computer Science at IPFW invites applications for a tenure-track position at the Assistant Professor rank with expertise in Information Systems or Computer Science. The department is ABET accredited in Computer Science. Candidates specializing in Business Intelligence, Information Science, Health Informatics, Enterprise Systems and Business Process Management, Information Management, or Web Science will be considered. In addition to these specific areas, outstanding candidates in any Computer Science area will receive full consideration. The position will start August 15, 2016.

Requirements for the position are: Ph.D. in Information Systems, Computer Science, or closely related field; a strong record of research that supports the teaching and research missions of the Department; excellent communication and interpersonal skills; and an interest in working with students and collaborating with the business community.

Job Responsibilities include: teaching undergraduate and graduate level courses in Information Systems and Computer Science, academic advising, and a strong pursuit of scholarly endeavors. Service to and engagement with the University, Department, and community is also required.

Review of applications will begin December 15, 2015 and will continue until the position is filled. A complete application should include a cover letter, CV, statements on research and teaching, with the names and contact information of three references. For teaching, we require evidence of teaching experience or effectiveness and a 1-2 page teaching philosophy. All materials must be submitted electronically to the Chair of the Search Committee at cssearch@ipfw.edu. All candidates who are invited to interview on campus will be required to prepare a 45 to 60 minute instructional student-based presentation. An official transcript and three letters of recommendation are required upon acceptance of offer. Employment is contingent on a satisfactory background records check.

IPFW is an EEO/AA employer fully committed to achieving a diverse workforce. All individuals, including minorities, women, individuals with disabilities, and protected veterans are encouraged to apply.

The Johns Hopkins University Department of Computer Science Tenure-Track Faculty Positions

The Johns Hopkins University's Department of Computer Science seeks applicants for **tenure-track faculty positions** at all levels and across all areas of computer science. Particular emphasis is at the junior level and in the areas of systems, distributed systems, networks and system security, however, all qualified applicants in all areas of computer science will be considered.

The Department of Computer Science has 25 full-time tenured and tenure-track faculty members, 8 research and 3 teaching faculty members, 130 PhD students, 130 MSE/MSSI and 216 undergraduate students. We have several affiliated research centers and institutes including Center for Language and Speech Processing, JHU Information Security Institute, Laboratory for Computational Sensing and Robotics, Institute for Data Intensive Engineering and Science, Engineering in Healthcare and more. More information about the Department of Computer Science can be found at www.cs.jhu.edu and about the Whiting School of Engineering at www.engineering.jhu.edu/about/. Qualifications and required materials can be found at <http://www.cs.jhu.edu/about/employment-opportunities/>.

Applicants should submit a curriculum vitae, a research statement, a teaching statement, three recent publications, and complete contact information for at least three references. **Applications must be made on-line at <https://academicjobsonline.org/ajob/jobs/6431>.** Review of

applications will begin in December 2015. While candidates who complete their applications by December 15, 2015 will receive full consideration, the department will consider exceptional applicants at any time.

The Johns Hopkins University is committed to active recruitment of a diverse faculty and student body. The University is an Affirmative Action/Equal Opportunity Employer of women, minorities, protected veterans and individuals with disabilities and encourages applications from these and other protected group members. Consistent with the University's goals of achieving excellence in all areas, we will assess the comprehensive qualifications of each applicant.

The Whiting School of Engineering and the Department of Computer Science are committed to building a diverse educational environment.

Macalester College Assistant Professor

Applications are invited for a tenure-track Computer Science position at Macalester College to begin Fall, 2016. Candidates must have or be completing a PhD in CS and have a strong commitment to both teaching and research in an undergraduate liberal arts environment. Areas of highest priority include computer and data security and privacy, mobile and ubiquitous computing, human-computer interaction, and visualization. See <http://www.macalester.edu/mscs> for details. Evaluation of applications will begin December 1. **Apply URL: <https://academicjobsonline.org/ajob/jobs/5794>**

PENN STATE



Director

The College of Engineering at The Pennsylvania State University invites nominations and applications for the position of Director of the School of Electrical Engineering and Computer Science. The School houses the Electrical Engineering Department and the Computer Science and Engineering Department. The College seeks an individual who will provide innovative and energetic leadership with effective administrative skills and a strong commitment to higher education. The Director will lead two Department Heads in administering the School, and will be a member of the College of Engineering Leadership Team. An ideal candidate will have the vision, energy, and experience to continue the process of integrating the two highly ranked departments into an agile, intellectually broad school, and brand the new school and establish its footprint in learning, discovery and service to the technical community and society. The candidate should have a widely recognized reputation in Electrical or Computer Engineering, Computer Science, or experience in related fields. An earned doctorate is required. Established in 1893, the Department of Electrical Engineering is among the largest, oldest and the most innovative in the nation. Approximately 550 undergraduate students and 220 graduate students are enrolled in the department, which has 40 faculty. The Department offers BS, MS, MEng and PhD degrees in electrical engineering. The Department of Computer Science and Engineering was created in 1993 with the merger of the Computer Engineering Program and the Computer Science Department. The department offers B.S. degrees in both computer engineering and computer science, and M.S., MEng, and Ph.D. degrees in computer science and engineering. There are approximately 550 undergraduate students and 150 graduate students enrolled in the department, with 43 faculty members. There are a number of unique research facilities associated with the School of Electrical Engineering and Computer Science. In recent years, our faculty received \$35M Network Science Center Award, a \$48M Collaborative Research Alliance for the Science of Security, a \$10M NSF Expeditions in Computing Award, a \$5M Dow Chemical Award for Flexible Electronics and a \$2.9 M DOE ARPA-E Award for Solar Cell Research. Growth is planned in core areas as well as numerous interdisciplinary areas. These key areas are strongly supported by the University through the Institute for Cyber Science, the Huck Institutes of the Life Sciences, the Materials Research Institute, the Penn State Institutes of Energy and the Environment, and the Applied Research Laboratory. Nominations and applications will be considered until the position is filled. Candidates from academia, industry or government agencies are encouraged to apply. Screening of applicants will begin on November 1, 2015 and it is intended that the position be filled by the beginning of the 2016/17 academic year. Applicants should submit a statement of professional interests, a curriculum vita, and the names and addresses of four references. Please submit these three items in one pdf file electronically to <http://apptkr.com/677384>. Applications will be treated with the strictest confidence. Inquiries can be made to Thomas La Porta via e-mail to tlp@cse.psu.edu or by phone at: 814-865-6725.

CAMPUS SECURITY CRIME STATISTICS: For more about safety at Penn State, and to review the Annual Security Report which contains information about crime statistics and other safety and security matters, please go to <http://www.police.psu.edu/clery/>, which will also provide you with detail on how to request a hard copy of the Annual Security Report. Penn State is an equal opportunity, affirmative action employer, and is committed to providing employment opportunities to all qualified applicants without regard to race, color, religion, age, sex, sexual orientation, gender identity, national origin, disability or protected veteran status.

Max Planck Institute for Software Systems

Tenure-track openings

Applications are invited for tenure-track faculty positions in all areas related to the theory and practice of software systems, including security 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 "<https://apply.mpi-sws.org/>". To receive full consideration, applications should be received by December 15, 2015.

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.

Mississippi State University

Department of Computer Science and Engineering

Professor and Head

Applications and nominations are being sought for the Professor and Head of the Department of Computer Science and Engineering (www.cse.msstate.edu) at Mississippi State University. The Head is responsible for the overall administration of the department and this is a 12-month tenured position.

The successful Head will provide:

- Vision and leadership for nationally recognized computing education and research programs
- Exceptional academic and administrative skills
- A strong commitment to faculty recruitment and development
- A strong commitment to promoting diversity

Applicants must have a Ph.D. in computer science, software engineering, computer engineering, or a closely related field. The successful candidate must have earned national recognition by a distinguished record of accomplishments in computer science education and research. Demonstrated administrative experience is desired, as is teaching experience at both the undergraduate and graduate levels. The successful candidate must qualify for the rank of professor.

Applicants must apply online at www.jobs.msstate.edu (PARF #9306) by completing the Personal Data Information Form and submitting a cover letter outlining your experience and vision for this position, a curriculum vitae, and the names and contact information of at least three professional references.

Screening of candidates will begin January 15, 2016 and will continue until the position is filled. MSU is an equal opportunity employer, and all qualified applicants will receive consideration for employment without regard to race, color, religion, ethnicity, sex (including pregnancy and gender identity), national origin, disability status, age, sexual orientation, genetic information, protected veteran status, or any other characteristic protected by law. We always welcome nominations and applications from women, members of any minority group, and others who share our passion for building a diverse community that reflects the diversity in our student population.

Please direct any questions to Dr. Jonathan Pote, Search Committee Chair (662) 325-3280 or jpote@abe.msstate.edu.

Mississippi State University

Department of Computer Science and Engineering

Faculty Positions in Computer Science and Engineering

The Department of Computer Science and Engineering (<http://www.cse.msstate.edu>) is seeking to fill one open position for a tenure-track faculty member at the Assistant/Associate Professor level. The primary areas of interest for this position is software engineering and computer security, however exceptional candidates in all areas will be considered.

Mississippi State University is a comprehensive land-grant institution with approximately 20,000 students and about 1,300 faculty members. The Department of Computer Science and Engineering has 16 tenure-track faculty positions and offers academic programs leading to the bachelor's, master's and doctoral degrees in computer science and bachelor's degrees in software engineering and computer engineering. Faculty members and graduate students work with a number of on-campus research centers. Research expenditures total about \$5.2 million dollars annually and the university as a whole is ranked 72nd among U.S. institutions in computer science expenditures.

Candidates for this position are expected to hold a PhD in computer science or closely related field (ABDs may be considered). Level of appointment is commensurate with qualifications and experience. The preferred candidate will have an established research record as demonstrated by significant publications beyond dissertation or research grants in relevant areas. Experience in teaching university courses in computer science is also preferred.

Applicants must apply on-line <http://www.jobs.msstate.edu/> (PARF 9253) and complete a Personal Data Information Form. A letter of application, curriculum vita, teaching and research statements, and names and contact information of at least three references must also be submitted. Review of applications will begin as early as November 2015 and will continue until the position is filled.

MSU is an equal opportunity employer, and all qualified applicants will receive consideration for employment without regard to race, color, religion, ethnicity, sex (including pregnancy and gender identity), national origin, disability status, age, sexual orientation, genetic information, protected veteran status, or any other characteristic protected by law. We always welcome nominations and applications from women, members of any minority group, and others who share our passion for building a diverse community that reflects the diversity in our student population.

North Carolina State University

Department of Computer Science

Security Faculty Positions

The Department of Computer Science at North Carolina State University (NCSU) seeks to fill tenure-track faculty positions in the area of Security starting August 16, 2016.

Successful security candidates must have a strong commitment to academic and research excellence, and an outstanding research record commensurate with the expectations of a major research university. Required credentials include a doctorate in Computer Science or a related field. While the department expects to hire at the Assistant Professor level, candidates with exceptional research records are encouraged to apply for a senior position. The department is one of the largest and oldest in the country. It is part of a top US College of Engineering, and has excellent and extensive ties with industry and government laboratories. The department's research expenditures and recognition have been growing steadily as has the recognition of our impact in the areas of security, systems, software engineering, educational informatics, networking, and games. For example, we have one of the largest concentrations of NSF Early Career Award winners (25 of our current or former faculty have received one).

NCSU is located in Raleigh, the capital of North Carolina, which forms one vertex of the world-famous Research Triangle Park (RTP). RTP is an innovative environment, both as a metropolitan area with one of the most diverse industrial bases in the world, and as a center of excellence promoting technology and science. The Research Triangle area is routinely recognized in nationwide surveys as one of the best places to live in the U.S. We enjoy outstanding public schools, affordable housing, and great weather, all in the proximity to the mountains and the seashore.

Applications will be reviewed as they are received. The positions will remain open until suitable candidates are identified. Applicants are encouraged to apply by December 15, 2015. Applicants should submit the following materials online at <http://jobs.ncsu.edu> (reference position number 00001096) cover letter, curriculum vitae, research statement, teaching statement, and names and complete contact information of four references, including email addresses and phone numbers. Candidates can obtain information about the department and its research programs, as well as more detail about the position advertised here at <http://www.csc.ncsu.edu/>. Inquiries may be sent via email to: security-search@csc.ncsu.edu.

NCSU is an equal opportunity and affirmative action employer. In addition, NCSU welcomes all persons without regard to sexual orientation or genetic information. Individuals with disabilities requiring disability-related accommodations in the application and interview process please call (919) 515-3148.

The Ohio State University
Computer Science and Engineering
Department
Multiple Tenure-Track Positions at the Assistant Professor level

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 recruiting in the following areas: cybersecurity, machine learning, distributed systems & cloud computing, and data management.

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 also 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 in December and will continue until the positions are filled.

The Ohio State University is an Equal Opportunity/Affirmative Action Employer.

Oregon State University
School of Electrical Engineering and Computer Science
Tenure-Track position in Data Science and Engineering

The School of Electrical Engineering and Computer Science at Oregon State University invites

applications for one or more full time nine-month tenure-track faculty positions in the area of data science and engineering broadly construed. Subareas of interest include (but are not limited to) databases and data management, visualization and visual analytics, security, data mining, and signal processing for big data. The appointments are anticipated at the Assistant Professor rank, but exceptionally strong candidates may be considered at the rank of Associate Professor or Professor.

A Ph.D. in Computer Science, Electrical and Computer Engineering, or related field is required by the start of employment. Duties include teaching undergraduate and graduate courses, conducting research in the area of interest, securing research funding, and service. Candidates should show evidence of strong research promise, potential for developing an externally funded research program, and commitment to quality advising and teaching at the graduate and undergraduate levels. Applicants should demonstrate a strong commitment to collaboration with other research groups in the School of EECS, with other departments at Oregon State University, and outside the university. Applicants who are women or under-represented minorities are strongly encouraged to apply.

The school of EECS emphasizes a culture of collegiality and excellence in both research and education. With 56 tenured/tenure-track faculty, we enroll 200 PhD, 250 MS and 2,800 undergraduate students.

Oregon State University is located in Corvallis, a college town renowned for its high quality of life.

For full consideration, apply online by Feb 15, 2016 with a letter of interest; vita; two-page statement of research interests; one-page statement of teaching interests; one-page statement of philosophy toward equity, inclusion and diversity; and names and contact information for at least three references. **For more information, visit eeecs.oregonstate.edu/jobs.**

Oregon State University
Tenure-Track Faculty Position in Robotics

The Oregon State University Robotics program seeks applications for full-time faculty positions to support our M.S. and Ph.D. programs in Robotics (<http://robotics.oregonstate.edu>). All areas of robotics will be considered. Appointments are anticipated at the Assistant Professor rank, but exceptionally strong candidates may be considered at the rank of Associate Professor or Professor. The successful candidates will have office and laboratory space physically located with the existing Robotics group, and will have an administrative home most appropriate to their area of expertise within the College of Engineering. As a Land/Sea/Air/Space grant institution, with strong ties to oceanography and the NOAA fleet, as well as hosting an FAA UAV test site, there are many opportunities to collaborate across disciplines and utilize Robotics as an enabling technology.

Candidates should hold a Ph.D. degrees in robotics, mechanical engineering, electrical and computer engineering, computer science, or other robotics-related disciplines by the start of employment, and have a demonstrated record of scholarship. We particularly welcome candidates

with a proven record of working with robots in real environments. Duties include teaching undergraduate and graduate courses, conducting research, securing research funding, and service. Candidates should show evidence of strong research promise, potential for developing an externally-funded research program, and commitment to quality advising and teaching at the graduate and undergraduate levels. Applicants who are women or under-represented minorities are strongly encouraged to apply.

Oregon State University is located in Corvallis, a college town renowned for its high quality of life.

For full consideration, apply online by Feb 15, 2016 with a letter of interest; vita; two-page statement of research interests; one-page statement of teaching interests; one-page statement of philosophy toward equity, inclusion and diversity; and names and contact information for at least three references.

More information, including a complete position description is available at <http://robotics.oregonstate.edu/jobs>.

Pacific Lutheran University
Assistant Professor of Computer Science and Computer Engineering

PLU invites applications for a tenure-track Assistant Professor to begin September 1, 2016. This position teaches a variety of lower- and upper-division courses. For qualifications and details, visit <https://employment.plu.edu>. Apply: <https://employment.plu.edu/postings/3715>. EOE/AA

Purdue University
Department of Computer Science
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 early in October 2015, and will continue until the positions are filled.

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 related discipline, be committed to excellence in teaching, and have demonstrated excellence in research. 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. Purdue University is

an EEO/AA employer fully committed to achieving a diverse workforce. All individuals, including minorities, women, individuals with disabilities, and protected veterans are encouraged to apply.

San Diego State University
Department of Computer Science
Chair of Computer Science

Department of Computer Science at SDSU seeks candidates for the Chair position with a PhD in Computer Science or a closely related field, and a sustained record of supported research. The Department is a dynamic and growing unit looking for a visionary Chair to lead it into its next phase of expansion.

We strive to build and sustain a welcoming environment for all. SDSU is seeking applicants with commitment to working effectively with individuals from diverse backgrounds and members of underrepresented groups.

For more details and application procedures, please apply via <http://apply.interfolio.com/31841>.

SDSU is a Title IX, equal opportunity employer. A full version of this ad can be found at: <http://cs.sdsu.edu/>

San José State University
Computer Engineering Department
Assistant/Associate Professor, Computer/Software Engineering

Job Opening ID (JOID): 23459

The Computer Engineering Department at San José State University invites applications for 4 tenure-track faculty positions at the rank of Assistant or Associate Professor. Areas of particular interest include cloud computing and virtualization, big data analytics, networking, storage, mobile systems, parallel and distributed systems, cyber security, and embedded systems, but other areas in computer and software engineering will also be considered.

Qualifications:

- Applicants must have a Ph.D. in Computer Engineering, Software Engineering, Computer Science, or Electrical Engineering.
- For appointment at the Assistant Professor rank, the candidate must demonstrate potential for teaching and scholarly excellence. For appointment at an Associate rank, the candidate must have a record of broad teaching experience and significant scholarly and professional achievements commensurate with the rank.
- Applicants should have awareness of and sensitivity to the educational goals of a multicultural population as might have been gained in cross-cultural study, training, teaching and other comparable experience.

Responsibilities:

- A faculty member is expected to teach, supervise, and advise students in both undergraduate and graduate programs, and to establish a research program related to his/her field of interest.
- A faculty member will participate in department, college, and university committee and other service assignments.
- A faculty member must address the needs of

a student population of great diversity – in age, cultural background, ethnicity, primary language and academic preparation – through course materials, teaching strategies and advisement.

Salary Range:

Commensurate with qualifications and experience.

Starting Date: August 22, 2016

Employment is contingent upon proof of eligibility to work in the United States.

For full description and consideration, visit <http://apptrkr.com/681277>

Screening of candidates will begin February 1, 2016. The positions remain open until filled.

The Computer Engineering program at SJSU was ranked number one in the nation among public institutions in its category by US News in 2014. Ideally situated at the center of Silicon Valley, the program enjoys close and multi-leveled industry ties and offers ample opportunities for industry collaboration and applied research.

The Computer Engineering Department at San José State University is a leading provider of engineering professionals to Silicon Valley's high-tech industries. It provides its faculty with opportunities for close collaborative ties with industry and research partners in Silicon Valley. The department serves more than 2500 undergraduate and graduate students and offers BS and MS degrees in both computer and software engineering. AA/EOE.

Swarthmore College
Computer Science Department
Tenure Track and Visiting Positions

The Computer Science Department invites applications for one tenure-track position and multiple visiting positions at the rank of Assistant Professor to begin Fall semester 2016.

Swarthmore College has a strong institutional commitment to excellence through diversity and inclusivity in its educational program and employment practices. The College actively seeks and welcomes applications from candidates with exceptional qualifications, particularly those with demonstrated commitments to a more inclusive society and world.

Swarthmore College is a small, selective, liberal arts college located 10 miles outside of Philadelphia. The Computer Science Department offers majors and minors at the undergraduate level.

Applicants must have teaching experience and should be comfortable teaching a wide range of courses at the introductory and intermediate level. Candidates should additionally have a strong commitment to involving undergraduates in their research. A Ph.D. in Computer Science at or near the time of appointment is required.

For the tenure-track position, we are particularly interested in applicants whose areas will complement and broaden our program, including theory and algorithms, programming languages, and systems areas. Strong applicants in other areas will also be considered.

For the visiting position, strong applicants in any area will be considered.

For the tenure-track position, priority will be given to complete applications received by December 15. For the visiting position, priority will

be given to complete applications received by February 15. Applications for both positions will continue to be accepted after these dates until the positions are filled.

Applications should include a cover letter, vita, teaching statement, research statement, and three letters of reference, at least one (preferably two) of which should speak to the candidate's teaching ability. In your cover letter, please briefly describe your current research agenda; what would be attractive to you about teaching in a liberal arts college environment; and what background, experience, or interests are likely to make you a strong teacher of Swarthmore College students.

Tenure-track applications are being accepted online at <https://academicjobsonline.org/ajo/jobs/6161>.

Visiting applications are being accepted online at <https://academicjobsonline.org/ajo/jobs/6173>. Candidates may apply for both positions.

Swarthmore College
Department of Computer Science
Lab Lecturer

The Department of Computer Science is currently accepting applications for a Lab Lecturer. The Lab Lecturer position is full time during the academic year (Fall and Spring semesters) with summers off. The start date is January 11, 2016.

Swarthmore College has a strong institutional commitment to excellence through diversity and in its educational program and employment practices. The College actively seeks and welcomes applications from candidates with exceptional qualifications, particularly those with demonstrated commitments to a more inclusive society and world.

Swarthmore College is a small, selective, liberal arts college located 10 miles outside of Philadelphia. The Computer Science Department offers majors and minors at the undergraduate level.

The Lab Lecturer position is an Instructional Staff position at the college. The responsibilities of the position include, but are not limited to: teaching lab sections of the introductory courses in the Computer Science Department; working with faculty to develop lab assignments for the introductory courses; creating lab assignment write-ups and documentation on tools used in introductory labs; supporting faculty in creating and setting up lab code examples, documentation, and software tools for lab work; lab grading and coordinating student graders; and holding regular office hours and helping students in the lab during open lab hours. More information about the Computer Science Department can be found on our website at www.cs.swarthmore.edu.

A master's degree or Ph.D. in computer science or a related field with extensive computer science background is required. Prior teaching experience at the college level is preferred.

Applications should include a vita, teaching statement, and two letters of reference that speak to the candidate's teaching ability.

Applications are being accepted online at <https://academicjobsonline.org/ajo/jobs/6465>. We will begin reviewing applications on November 9. Applications will continue to be accepted until the position is filled.

Trinity College, Hartford, Connecticut Assistant Professor of Computer Science

Applications are invited for a tenure-track position in computer science at the rank of Assistant Professor to start in the fall of 2016. Candidates must hold a Ph.D. in computer science at the time of appointment. We are seeking candidates with teaching and research interests in applied areas associated with Big Data (such as database and information systems, data mining and knowledge discovery, machine learning and cloud computing), but other related areas will be seriously considered. Applications should be submitted to: <https://trincoll.peopleadmin.com/>. Consideration of applications will begin on December 15, 2015. Trinity College is an Equal-Opportunity/Affirmative-Action employer.

Tufts University

Department of Computer Science in the School of Engineering
Assistant/Associate Professor in Machine Learning/Data Mining

The Department of Computer Science in the School of Engineering at **Tufts University** invites applications for a tenure-stream faculty appointment to begin in September 2016. We are seeking outstanding junior-level or mid-career-level candidates for an appointment at the rank of Assistant Professor or Associate Professor.

We welcome outstanding applicants with a strong vision and research programs in the areas of machine learning and data mining broadly interpreted to include related fields where data and its analysis are central. These include, for example, adaptive systems in robotics, computational linguistics, computational sustainability, and visual analytics.

Application materials should be submitted online through Interfolio at <https://apply.interfolio.com/30915>
For more information please visit <http://www.cs.tufts.edu/>.
Inquiries should be emailed to cssearch@cs.tufts.edu.

Review of applications will begin January 5, 2016 and will continue until the position is filled.

Tufts University is an Affirmative Action/Equal Opportunity employer. We are committed to increasing the diversity of our faculty. Members of underrepresented groups are strongly encouraged to apply.

University of Alabama

Tenured/Tenure-Track Faculty Positions,
Computer Science
Software Engineering Focus: Data Analytics and Computational Modeling

Openings exist for two Assistant/Associate/Full professors in software engineering with specific application to data analytics or computational modeling starting in Fall 2016. Outstanding candidates in all areas will be considered. At the time of appointment, candidates must have earned a Ph.D. in Computer Science or a related field. Candidates will be expected to form collaborations with the new NOAA Water Center on UA's

campus (<http://nws.noaa.gov/oh/nwc/>). The CS department has twenty-four faculty members (15 tenured/tenure track faculty, seven of whom have interests in software engineering), over 600 undergraduates in an ABET accredited B.S. degree program, and 40 graduate students. The department also offers a Software Engineering Concentration for its undergraduates.

For additional details and to apply, visit <http://se.cs.ua.edu/facultyjobs> or contact Dr. Jeffrey Carver (carver@cs.ua.edu). Review of applications will begin immediately. The University of Alabama is an equal opportunity/affirmative action employer. Women and minority applicants are particularly encouraged to apply.

University of California, Riverside

Department of Computer Science and Engineering
Bourns College of Engineering
Faculty Positions

The Department of Computer Science and Engineering, University of California, Riverside invites applications for **two tenure-track and/or tenured faculty positions** to begin in the 2016-17 academic year. Priority will be given to candidates in the areas of (1) Computer Graphics, Animation, and Gaming, and (2) Software Engineering; however, exceptional candidates in other areas may also be considered. Salary level will be competitive and commensurate with qualifications and experience. A Ph.D. in Computer Science (or in a closely related field) at the time of employment is required. Senior candidates need to have an outstanding record of research, funding support, teaching, and graduate student mentorship, while junior candidates need to show potential to excel in these areas. Advancement through the faculty ranks at the University of California is through a series of structured, merit-based evaluations, occurring every 2-3 years, each of which includes substantial peer input.

The CSE department offers several undergraduate degrees, as well as MS and Ph.D. degrees in Computer Science. The Department currently has 25 faculty members, including multiple ACM/IEEE/AAAS Fellows and Young Investigator/NSF CAREER award holders, who pride themselves in combining top quality teaching with cutting edge research. The research projects in the department are funded by federal (NSF, NIH, AFOSR, DoD) or industrial sponsors, with the new awards for 2015/16 exceeding 6 million dollars. More information regarding the department is available at <http://www.cs.ucr.edu>.

Full consideration will be given to applications received by January 2, 2016. We will continue to consider applications until the positions are filled. To apply, please register through the web-link at <http://www.engr.ucr.edu/facultysearch/>. For inquiries and questions, please contact us at search@cs.ucr.edu.

UCR is a world-class research university with an exceptionally diverse undergraduate student body. Its mission is explicitly linked to providing routes to educational success for underrepresented and first-generation college students. A commitment to this mission is a preferred qualification.

The University of California is an Equal Opportunity/Affirmative Action Employer. All quali-

fied applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, age, disability protected veteran status, or any characteristic protected by law.

University of California, Riverside

Bourns College of Engineering
Faculty Position in Big Data Management

The University of California at Riverside (UCR) is embarking on a major new hiring initiative that will add 300 tenured and tenure-track positions in 33 cross-disciplinary areas selected through a peer-reviewed competition. Over the next three years, we will hire multiple faculty members in each area and invest in research infrastructure to support their work. This initiative will build critical mass in vital and emerging fields of scholarship, foster truly cross-disciplinary work and further diversify the faculty at one of America's most diverse research universities. We encourage applications from scholars committed to excellence and seeking to help redefine the research university for the next generation.

The Bourns College of Engineering is leading cluster hires to enhance UCR's research strengths in Data Science. Five such hires have been approved for the Data Science cluster and will have potential home departments in Engineering or the Sciences. Candidates are expected to foster research collaborations with existing faculty across academic departments working on Data Science related topics (including astronomy, biological sciences, computational biology, environmental sciences, physics, precision agriculture, etc.) This year we are seeking to fill one tenured/tenure-track faculty position from the Data Science cluster, in the area of Big Data Management with emphasis on scalable data management, big data analytics, and scalable data mining. While priority will be given to senior candidates, promising junior candidates will also be considered and are encouraged to apply.

Salary level will be competitive and commensurate with qualifications and experience. A Ph.D. in a relevant area at the time of employment is a minimum requirement. Senior candidates need to have an outstanding record of research, funding support, teaching, and graduate student mentorship, while junior candidates need to show potential to excel in these areas.

UCR is a world-class research university with an exceptionally diverse undergraduate student body. Its mission is explicitly linked to providing routes to educational success for underrepresented and first-generation college students. A commitment to this mission is a preferred qualification.

Advancement through the faculty ranks at the University of California is through a series of structured, merit-based evaluations, occurring every 2-3 years, each of which includes substantial peer input.

Full consideration will be given to applications received by January 4, 2016. We will continue to consider applications until the position is filled. To apply, please register through the web-link at <http://www.engr.ucr.edu/facultysearch/>. For inquiries and questions, please contact us at datascienceseach@cs.ucr.edu.

The University of California is an Equal Op-

portunity/Affirmative Action Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, age, disability protected veteran status, or any characteristic protected by law.

University of California, Riverside
Bourns College of Engineering
Department of Computer Science and Engineering
Faculty Positions in Cybersecurity

The University of California at Riverside (UCR) is embarking on a major new hiring initiative that will add 300 tenured and tenure-track positions in 33 cross-disciplinary areas selected through a peer-reviewed competition. Over the next three years, we will hire multiple faculty members in each area and invest in research infrastructure to support their work. This initiative will build critical mass in vital and emerging fields of scholarship, foster truly cross-disciplinary work and further diversify the faculty at one of America's most diverse research universities. We encourage applications from scholars committed to excellence and seeking to help redefine the research university for the next generation.

The Department of Computer Science and Engineering, University of California, Riverside invites applications for **three tenure-track and/or tenured faculty positions in Cyber-security**, with particular interest in: (a) operating systems/distributed system security, (b) software security, (c) applied cryptography, and (d) human computer interaction for understanding and improving the security of systems. Senior candidates need to have a strong record of research, teaching, and graduate student mentorship, while junior candidates need to show potential to excel in these areas.

The CSE department offers several undergraduate degrees, as well as MS and Ph.D. degrees in Computer Science. The Department currently has 25 faculty members, including multiple ACM/IEEE/AAAS Fellows and Young Investigator/NSF CAREER award holders, who pride themselves in combining top quality teaching with cutting edge research. The research projects in the department are funded by federal (NSF, NIH, AFOSR, DoD) or industrial sponsors, with the new awards for 2015/16 exceeding 6 million dollars. More information regarding the department is available at <http://www.cs.ucr.edu>.

UCR is a world-class research university with an exceptionally diverse undergraduate student body. Its mission is explicitly linked to providing routes to educational success for underrepresented and first-generation college students. A commitment to this mission is a preferred qualification.

Advancement through the faculty ranks at the University of California is through a series of structured, merit-based evaluations, occurring every 2-3 years, each of which includes substantial peer input.

Full consideration will be given to applications received by January 2, 2016. We will continue to consider applications until the positions are filled. Salary level will be competitive and commensurate with qualifications and experience. Positions require a Ph.D. in Computer Science

(or in a closely related field) at the time of employment. To apply, please register through the web-link at <http://www.engr.ucr.edu/facultysearch/>. For inquiries and questions, please contact us at search@cs.ucr.edu.

The University of California is an Equal Opportunity/Affirmative Action Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, age, disability, protected veteran status, or any other characteristic protected by law.

University of Central Florida
Computer Science Division, College of Engineering and Computer Science
Assistant or Associate Professor

The Computer Science Division in the College of Engineering and Computer Science at UCF invites applications for three tenure-track (or tenured) positions at the assistant (or associate) professor level, starting Fall 2016. Successful candidates will have a record of high-quality publications and be recognized for their expertise and the impact of their research. Successful candidates at the level of associate professor must have a record commensurate with that rank. We welcome exceptional applicants from all CS research areas though we are particularly interested in applicants in the following areas: cyber security, algorithms and the theory of computing, and human-computer interaction.

All applicants must have a Ph.D. from an accredited institution in an area appropriate to Computer Science and a strong commitment to the academic process, including scholarly publications, sponsored research, and teaching.

Computer Science at UCF has a rapidly-growing educational and research program with over \$4.5 million in research contracts and expenditures annually and over 215 graduate students. Computer Science has strong areas of research in Computer Vision, Machine Learning, Virtual and Mixed Reality, Big Data, and Human-Computer Interaction. The CS Division is also well-known for the success of its two-time defending National Champion Cyber Defense team and the exceptional record of its programming teams in regional, national, and world competitions. More information about the Computer Science Division can be found at <http://www.cs.ucf.edu/>.

Research sponsors include NSF, NIH, NASA, DOT, DARPA, ONR, and other agencies of the DOD. Industry sponsors include AMD, Boeing, Canon, Electronic Arts, General Dynamics, Harris, Hitachi, Intel, Lockheed Martin, Oracle, SAIC, Symantec, Toyota USA, and Walt Disney World, as well as local startups.

UCF has the top-tier Carnegie Foundation designation of a "very high research activity" university, is the nation's second largest university, and is ranked by U.S. News and World Report as the third most up-and-coming university in terms of innovative changes in the areas of academics, faculty, and student life. Adjacent to UCF is a thriving research park that hosts more than 100 high-technology companies and the Institute for Simulation and Training. The Central Florida area is designated by the State of Florida as the Center of Excellence in Modeling and Simulation. UCF also has an accredited medical school,

which opened in 2009.

To apply, applicants must complete an online job application at <https://www.jobswithucf.com/postings/43185>. In addition to the application, candidates must also submit a signed cover letter, complete curriculum vitae, maximum two page statement outlining research vision and teaching interests, and a list of at least three professional references complete with address, phone number, and email address.

University of Colorado Boulder
Department of Computer Science
Assistant Professor

The Department of Computer Science at the University of Colorado Boulder seeks applications for multiple tenure-track positions. The openings are targeted at the level of Assistant Professor, although candidates at higher ranks may be considered. Research areas of particular interest include, but are not limited to, the areas of secure and reliable software systems (with an emphasis in software engineering and/or security), network science, scientific data analysis and visualization, computer systems as applied to autonomous and networked devices, and theoretical computer science. Candidates must have a Ph.D. in computer science or a related discipline and must show promise in their ability to develop an independent and internationally recognized research program. They must also display an ability, a record of excellence, and/or a commitment to teaching and working with undergraduate and graduate students of diverse backgrounds. Our department values inclusive excellence and we seek candidates that understand the benefits that diversity brings to scientific innovation and who, through their work, develop technologies that impact a wide range of communities. Our department is also responsive to dual career situations.

We will be accepting applications starting November 5th, 2015. Details available at: <http://www.cs.colorado.edu/~kena/2015CSFacultySearch/>

The University of Colorado Boulder is an Equal Opportunity Employer.

The University of Colorado Boulder conducts background checks for all final applicants.

University of Detroit Mercy
Department of Mathematics, Computer Science, and Software Engineering
Tenure Track Faculty Position

University of Detroit Mercy seeking a faculty member in Computer Science or Software Engineering with emphasis on teaching and research with UG & GR students. The deadline for submissions is December 20, 2015. Visit: <http://eng-sci.udmercy.edu/faculty/faculty-positions/>

University of Illinois at Urbana-Champaign
Department of Electrical and Computer Engineering (ECE)
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 levels and in all areas in computing, broadly defined, with particular emphasis on big data and its applications, including data analytics; data center and storage systems; parallel, high-performance, and energy-efficient computing; reliable and secure computing; distributed computing; bio-inspired computing; verification; wired/wireless networking; social networking; mobile, wearable sensing & applications; and computational genomics. 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. Applications are encouraged from candidates whose research programs specialize in core as well as interdisciplinary areas of electrical and computer engineering. 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 to support research in big data and bioengineering, broadly defined. In addition, the University of Illinois is home to Blue Waters petascale computer, which is supported by the National Science Foundation and developed and operated by the University of Illinois' National Center for Supercomputing Applications. Qualified candidates may be hired as Blue Waters Professors who will be provided substantial allocations on and expedited access to the supercomputer. To be considered as a Blue Waters Professor, candidates need to mention Blue Waters as one of their preferred research areas in their online application, and include a reference to Blue Waters in their cover letter.

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, 2015, but applications will continue to be accepted until all positions are filled.

The University of Illinois conducts criminal background checks on all job candidates upon acceptance of a contingent offer.

Illinois is an EEO Employer/Vet/Disabled www.inclusiveillinois.illinois.edu.

University of Kentucky Department of Computer Science Tenure-Track Faculty Positions

The University of Kentucky Computer Science Department expects to hire two tenure-track faculty members to begin employment in August of 2016.

The department seeks to hire energetic researchers/educators who are interested in the application of advanced computing to challenging and relevant problems. Our faculty undertake interdisciplinary research, working with other departments including statistics, biology, lin-

guistics, electrical engineering, computer engineering, and the humanities. We therefore favor researchers who can collaborate to solve problems involving multiple disciplines. The department has two openings. One is tied to the newly established University of Kentucky Institute for Biomedical Informatics (IBI; <http://ibi.uky.edu>), in which the focus areas include databases, biomedical informatics, data analytics, and visualization. For the other position, all areas of computing will be considered, but applicants with expertise in security and privacy, big data, and cloud computing will receive preference. These areas are associated with the department's Laboratory for Advanced Networking, Center for Visualization and Virtual Environments, and Software Verification and Validation Lab. The candidate should be able to teach upper-level courses with a focus on databases and/or security, as well as (ultimately) courses in introductory systems sequence.

Candidates must have earned a PhD in Computer Science or closely related field at the time employment begins. **To apply, a University of Kentucky Academic Profile must be submitted at <http://ukjobs.uky.edu/postings/80509>.** Applications are now being accepted. Review of credentials will begin immediately and continue until the positions are filled.

For more detailed information about these positions, go to www.cs.uky.edu/opportunities/faculty. Questions should be directed to HR/Employment by phone at 1-859-257-9555 press 2 or email (ukjobs@email.uky.edu), or to Diane Mier (diane@cs.uky.edu) in the Computer Science Department.

University of Maryland, Baltimore County Computer Science and Electrical Engineering Tenure-Track Faculty Positions

UMBC's Department of Computer Science and Electrical Engineering invites applications for three tenure-track Assistant Professor positions to begin in Fall 2016. Exceptionally strong candidates for higher ranks may be considered. Applicants must have or be completing a PhD in a relevant discipline, have demonstrated the ability to pursue a research program, and have a strong commitment to undergraduate and graduate teaching.

All areas of specialization will be considered, but we are especially interested in candidates in the following areas: information assurance and cybersecurity; mobile, wearable and IoT systems; big data with an emphasis on machine learning; analytics, and high-performance computing; knowledge and database systems; hardware systems and experimental methods in circuits, devices, VLSI, FPGA, and sensors; cyber-physical systems; low-power systems; biomedical and healthcare systems; and methods and tools for hardware-software co-design.

The department is energetic, research-oriented and multi-disciplinary with programs in Computer Science, Computer Engineering, Electrical Engineering and Cybersecurity. Our faculty (34 tenure-track, six teaching, 15 research) enjoy collaboration, working across our specializations as well as with colleagues from other STEM, humanities and arts departments and external partners.

We have 1500 undergraduate CS and CE majors and 400 MS and PhD students in our CS, CE, EE and Cybersecurity graduate programs. A diverse portfolio of government and industrial sponsors supports our research, which has 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. The 2015 US News and World Report Best Colleges report puts UMBC fourth in the Most Innovative National Universities and sixth in the Best Undergraduate Teaching, National Universities categories. Our strategic location in the Baltimore-Washington corridor is close to many federal laboratories, agencies and high-tech companies.

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/31543>. Applications received by January 15, 2016 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 Maryland College Park Department of Computer Science Faculty Positions

The Department of Computer Science at the University of Maryland, College Park, MD, USA has several openings for faculty positions effective July 1, 2016 or earlier. The openings are at the tenure-track Assistant Professor level or "junior-level" tenured Associate Professor level. Applicants will be considered for joint appointments between the Department of Computer Science and the Institute for Advanced Computer Studies (UMIACS).

We are especially interested in recruiting in Computational Biology, Cybersecurity, Machine Learning, and Databases; however, exceptional candidates in all areas will be considered.

Applications from women and other under-represented groups are especially welcome.

Please apply online at <https://ejobs.umd.edu> and <https://hiring.cs.umd.edu>. Candidates must apply to both websites to receive consideration. The review of applications will begin on December 1, 2015, and applicants are strongly encouraged to submit complete applications by that date for full consideration. Questions can be directed to the faculty recruitment committee at: faculty-search@cs.umd.edu.

Founded in 1856, University of Maryland, College Park is the flagship institution in the University System of Maryland. Our 1,250 acre College Park campus is minutes away from Washington, D.C., and the nexus of the nation's legislative, executive, and judicial centers of power. This unique proximity to business and technology

leaders, federal departments and agencies, and a myriad of research entities, embassies, think tanks, cultural centers, and non-profit organizations offers unparalleled synergistic opportunities for our faculty and students. The Department of Computer Science at the University of Maryland has been consistently ranked among the top 15 nationally. We have 47 full time tenured and tenure track faculty in a wide variety of research areas, and over 200 doctoral students drawn from top undergraduate programs internationally.

Additional information about the Department of Computer Science and the Institute for Advanced Computer Studies is available at <http://www.cs.umd.edu> and at <http://www.umiacs.umd.edu>. The Department is planning to move into the Iribe Center in the near future, for more information, see <http://csctr.cs.umd.edu>.

The University of Maryland is an Equal Opportunity, Affirmative Action Employer.

University of Minnesota-Twin Cities
Department of Computer Science and Engineering
Multiple Tenure-Track Faculty Positions in Cyber Security

The Department of Computer Science and Engineering at the **University of Minnesota-Twin Cities** invites applications for **multiple tenure-track faculty positions in cyber security** and in support of a University-wide initiative (MnDRIVE) on robotics, sensors, and advanced manufacturing (<http://cse.umn.edu/r/mndrive-minnesota-discovery-research-and-innovation-economy/>). Specific topics of interest for the positions include cyber security, sensing and networking, machine learning, computer vision, robot design, manipulation, mobility, human-robot interaction, planning, algorithmic foundations, and embedded systems. Applicants from other areas will be considered as long as they address how their work fits into the security or MnDRIVE themes. Senior applicants will also be considered. We encourage applications from women and under-represented minorities. Candidates should have a Ph.D. in Computer Science or a closely related discipline at the time of appointment. The positions are open until filled, but for full consideration apply at <https://www.cs.umn.edu/resources/employment/faculty> by December 15, 2015. The University of Minnesota is an equal opportunity employer and educator.

The University of Mississippi
Department of Computer and Information Science
Assistant Professor Position

The Department of Computer and Information Science at the University of Mississippi invites applications for a tenure-track Assistant Professor position.

An applicant must hold a PhD or equivalent in computer science or a closely related field by August 15, 2016. An applicant must have the ability to teach both graduate and undergraduate students, conduct research in major areas of computer and information science, and supervise MS and PhD students. An applicant must provide evidence of research potential, effective

communication skills, and a broad background in computing.

The Department has an ABET/CAC-accredited undergraduate program and MS and PhD programs. See the website <http://www.cs.olemiss.edu> for more information about the Department.

The University is located in Oxford, one of America's top-ranked college towns. Oxford has a wonderful small-town atmosphere with affordable housing and excellent schools.

Individuals may apply online at <http://jobs.olemiss.edu>. The applicant is asked to supply a cover letter, curriculum vitae, research and teaching statements, and contact information for four references. Review of applications will begin immediately and continue until the position is filled or an adequate applicant pool is reached.

The University of Mississippi is an EOE/AA/Minorities/Females/Vet/Disability/Sexual Orientation/Gender Identity/Title VI/Title VII/Title IX/504/ADA/ADEA employer.

University of Nevada, Reno
CSE Department
Four Tenure-Track Faculty Positions

UNIVERSITY OF NEVADA, RENO. The CSE Department invites applications for **four tenure-track faculty positions**. Two positions are at assistant professor level: the first position is in the area of big data with emphasis on security and privacy, and the second position is in the area of high performance computing with emphasis on parallel and distributed computing. The third position, in the area of cybersecurity, is at associate professor level, and will fill the role of Technical Director of the newly established Cybersecurity Center (CSC) at UNR. The fourth position is at assistant, associate or full professor level, and is open to all research areas, with preference given to candidates with expertise in embedded systems (Internet of Things, cyber-physical systems, VLSI design), machine learning (deep learning, data analytics, bioinformatics), computer graphics and visualization.. Applicants must have a Ph.D. in Computer Science or Computer Engineering by July 1, 2016. Candidates must be strongly committed to excellence in research and teaching and should demonstrate potential for developing robust externally funded research programs. The department has several faculty with NSF Career awards and leaders in statewide and multi-state multi-million dollar NSF awards. Our research is supported by NSF, DoD, DHS, NASA, Google, Microsoft, Ford and AT&T. The department's annual research expenditures have exceeded \$2M in recent years, while FY15 funding exceeds \$3M. We offer B.S., M.S., and Ph.D. degrees and have strong research and education programs in Intelligent Systems, Computer and Network Systems, Software Systems, and Games and Simulations. In the last five years, the College of Engineering has witnessed an unprecedented growth in student enrollment and number of faculty positions. The College is positioned to further enhance the growth of its students, faculty, staff, and facilities as well as its research productivity and its graduate and undergraduate programs. UNR, Nevada's land grant University, has nearly 21,000 students. Reno is a half-hour drive to beautiful Lake Tahoe, an excellent area for a wide range of outdoor activities. San Francisco is within a four-hour drive.

EEO/AA Women, under-represented groups, individuals with disabilities, and veterans are encouraged to apply. Apply online at:

<https://www.unrsearch.com/postings/19013>
(for the big data position)

<https://www.unrsearch.com/postings/18990>
(for the high performance computing position)

<https://www.unrsearch.com/postings/19015>
(for the cybersecurity position)

<https://www.unrsearch.com/postings/19004>
(for the open position)

Review of applications will begin on January 5, 2016 and will continue until the search closes on February 15, 2016. Inquiries should be directed to Ms. Lisa Cody, lacody@unr.edu.

The University of North Carolina at Greensboro
Department of Computer Science
Tenure Track Assistant Professor

The University of North Carolina at Greensboro (UNCG) seeks applications for a tenure-track position at the rank of Assistant Professor in the Department of Computer Science starting August 1, 2016. We are looking for candidates who show exceptional promise in both research and teaching. Preferred research areas are those that build upon our existing areas of strength, which include artificial intelligence, databases and data mining, foundations of computer science, human-computer interaction, networking, security, and big data analytics, but applicants in other research areas are also encouraged to apply. We are particularly interested in candidates who can pursue interdisciplinary research in the natural sciences or computational mathematics, especially through research in bioinformatics or molecular modeling. Experience building an independent research program and mentoring graduate students is a plus.

UNCG is a public coeducational, doctoral-granting residential university chartered in 1891, with a Carnegie classification of RU/H (Research University/High Research Activity). The Department of Computer Science at UNCG is a thriving department with an established, ABET-accredited B.S. degree program and an active M.S. degree program, and is experiencing rapid enrollment growth. The department currently has 7 tenured faculty members who are all active in research, as well as lecturers and part-time faculty. For more information on the Computer Science Department, visit the Department's web page at <http://www.uncg.edu/cmp>.

Candidates must hold or anticipate a Ph.D. in Computer Science or a related discipline by August 1, 2016.

Submit curriculum vitae, research and teaching statements, and four letters of reference through UNCG JobSearch at <http://jobsearch.uncg.edu>. You may direct your informal inquiries to Dr. Stephen Tate, Department of Computer Science, University of North Carolina at Greensboro, Greensboro, NC 27402 (cssearch@uncg.edu).

Review of applications will begin on January 18, 2016 and continue until the position is filled.

UNC Greensboro is especially proud of the diversity of its student body and we seek to attract an equally diverse applicant pool for this posi-

tion. We are an EEO/AA employer with a strong commitment to increasing faculty diversity and will respond creatively to the needs of dual-career couples.

University of Northern Iowa
Department of Computer Science
Assistant Professor of Computer Science

The Department of Computer Science at the University of Northern Iowa invites applications for a tenure-track assistant professor position to begin August 2016. Applicants must hold a Ph.D. in Computer Science or a closely-related discipline. The department seeks candidates able to teach and conduct research in software engineering, as well as to participate broadly in the CS curriculum.

Detailed information about the position and the department are available at <http://www.cs.uni.edu/>. To apply, visit <http://jobs.uni.edu/>. Applications received by January 15, 2016, will be given full consideration.

Pre-employment background checks are required. UNI actively seeks to enhance diversity and is an Equal Opportunity/Affirmative Action employer. The University encourages applications from persons of color, women, individuals living with disabilities, and protected veterans. All qualified applicants will receive consideration for employment without regard to age, color, creed, disability, gender identity, national origin, race, religion, sex, sexual orientation, protected veteran status, or any other basis protected by federal and/or state law. UNI is a smoke-free campus.

University of Pittsburgh
Swanson School of Engineering
Chair, Department of Electrical and Computer Engineering (ECE)

The Swanson School of Engineering at the University of Pittsburgh (<http://www.engineering.pitt.edu/>) invites applications and nominations for the position of Chair of the Department of Electrical and Computer Engineering (<http://www.engineering.pitt.edu/ECE/>). We are looking for an energetic leader with a vision to inspire innovative research and educational programs. A successful candidate must have a strong record of academic and professional accomplishments to support an appointment as a full professor with tenure.

The Department has a tradition of excellence in education and research, with many of our faculty winning awards for outstanding teaching. Research activities are organized around four main areas: (1) Biomedical Electronics and Signal Processing, (2) Energy and Electric Power Systems and Technologies, (3) Nano/Micro-Electronics & Photonics, and (4) Cyber Systems and Technologies. All four of these focus areas fall under the theme of "Reshaping the World on All Scales." The Department faculty members have significant research collaborations with their colleagues from other engineering departments, the Center for Energy (<http://www.engineering.pitt.edu/cfe/>), the Peterson Institute for NanoScience and Engineering (<http://www.nano.pitt.edu/>), the School of Medicine, and departments from the Dietrich School of Arts and

Sciences. Current annual research expenditures are around \$4.5 million.

The Department currently has 25 faculty members and offers BS, MS, and PhD degrees in Electrical Engineering and Computer Engineering (jointly with the Department of Computer Science). Current enrollment consists of 436 undergraduate and 175 graduate students. The Swanson School of Engineering just completed a \$100 million renovation and redesign of Benedum Engineering Hall with state-of-the-art teaching and research laboratories, which is the Department home.

The University of Pittsburgh, located just minutes from downtown, is adjacent to the University of Pittsburgh Medical Center and Carnegie Mellon University. The Pittsburgh metropolitan area has all the amenities of a large urban area including major sports, theatre, opera, ballet, symphony, museums, zoo, etc. The urban renaissance in Pittsburgh is directly attributable to the strong growth in high tech industries, such as information/computer technology, robotics, and biotechnology.

Review of applications will begin on December 1, 2015 and applications will be accepted until the position is filled. The start date for this position is on or after September 1, 2016. Please send nominations or application materials (curriculum vitae, names and contact information of at least three references, and any other documents deemed relevant) electronically to Ms. Nancy Donaldson (dndnson@pitt.edu) and Dr. Sanjeev Shroff, Chair, Electrical and Computer Engineering Chair Search Committee (sshroff@pitt.edu).

We highly encourage candidates from under-represented US minority groups and/or females to apply for this position. The University of Pittsburgh is an affirmative action/equal opportunity employer and does not discriminate on the basis of age, color, disability, gender, gender identity, marital status, national or ethnic origin, race, religion, sexual orientation or veteran status.

The University of Texas at San Antonio
Department of Computer Science
Faculty Positions in Computer Science

The Department of Computer Science at The University of Texas at San Antonio invites applications for a tenured/tenure-track position at the assistant or associate professor level, starting Fall 2016. Interested candidates with research focus in one or more areas of system software, data science, high performance computing, or cloud computing are encouraged to apply.

See <http://www.cs.utsa.edu/fsearch> for information on the Department and application instructions. Screening of applications will begin immediately. Full consideration will be given to applications received by January 4, 2016, and the search will continue until the positions are filled or the search is closed. The University of Texas at San Antonio is an Affirmative Action/Equal Opportunity Employer.

Department of Computer Science
RE: Faculty Search
The University of Texas at San Antonio
One UTSA Circle
San Antonio, TX 78249-0667
Phone: 210-458-4436

Washington College
Department of Mathematics & Computer Science
Assistant Professor

The Department of Mathematics & Computer Science at **Washington College** in Chestertown MD invites applications for a tenure-track position in computer science at the rank of **Assistant Professor** beginning in August 2016. For full details visit <http://www.washcoll.edu/offices/human-resources/employment.php>. EOE M/F/D/V

Wesleyan University
Department of Mathematics and Computer Science
Assistant Professor of Computer Science

The Department of Mathematics and Computer Science at **Wesleyan University** invites applications for a tenure track assistant professorship in Computer Science to start in 2016-2017. Wesleyan has a strong research group in theory, programming languages, and software engineering and a 2-1 teaching load. We encourage candidates in all areas to apply, but especially databases, machine learning, natural language processing, software engineering and software systems. **For description and application procedure see <http://www.wesleyan.edu/mathcs/employment.html>**

Yahoo Labs
Intern Scientist (Job Number: 1545124)
(Yahoo Labs Summer Internship Program - 2016)

Yahoo Labs is pioneering the new sciences underlying the Web. As the center of scientific excellence for Yahoo, Yahoo Labs delivers both fundamental and applied scientific leadership through published research and new technologies powering the company's products.

Yahoo Labs is looking for exceptional PhD students to work with us in our summer intern program. We will have openings in the US (New York City, Sunnyvale, San Francisco) plus our locations in London and Haifa. We seek world-class graduate students in pursuit of a PhD in Computer Science, Mathematics, Statistics, or a related area. We are particularly interested in students working on Machine Learning, algorithms, Natural Language Processing, Knowledge Representation, HCI, Multimedia, Mobile Innovations, search (systems or algorithms), collaborative filtering, auctions, mechanism design, linear algebra, Systems or analysis of large data. Ideal candidates will have finished at least 3 years of graduate work.

Interns are expected to work with our scientists to perform original research, apply scientific thinking and techniques to improve the performance and effectiveness of our products, and solve problems for our users and advertisers by analyzing mountains of data. They will have the opportunity to publish their work and expand the horizons of web science. Candidates will need to submit a CV plus a letter of recommendation from their graduate advisor.

Yahoo Inc. is an equal opportunity employer. For more information or to search all of our openings please visit <http://labs.yahoo.com/careers>.

[CONTINUED FROM P. 136] one of who completed his Ph.D., worked for IBM, and then came back as a faculty member. We had a very productive collaboration until he passed away.

You're referring to Robert M. Owens, who died in 1997.

We started building projects and they were a wonderful experience. We built two different special-purpose signal processors, including custom chip design, board design, algorithm design, software design—all sorts of things that we didn't know we were getting into when we started the projects.

The first, called the Arithmetic Cube, was a high-speed programmable VLSI processor for solving linear digital signal processing problems.

We designed and implemented a custom architecture that could run at the speed of the commercial parts that were available at the time, even though it was an academic project and not implemented in the most current technology node. It had custom CMOS chips that we had fabricated through MOSIS (Metal Oxide Semiconductor Implementation Service), a board-level design, and novel signal processing algorithms.

What we discovered along the way is that the design tools we needed to support the design process either didn't exist, or existed only in companies that didn't distribute them. So we had to build the design tools as well. We built the logic synthesis tools, a high-level synthesis tool, automatic cell-generation tools, and the software we needed to test the design. It turned out to be a much bigger project than we anticipated.

After that, you built a processor known as the MicroGrain Array Processor, or MGAP, designed to address complex signal and image processing problems.

We wanted to build a more programmable processor than the Arithmetic Cube. That led to the MGAP implementation, which was very FPGA-like, looking back at it, except that it only allowed nearest-neighbor communication. There were two generations of that. We used a lot of the design tools that we had built for the Arithmetic Cube, and of course we designed new tools.

"Some design with FPGAs, and that's great, but the number of people building custom hardware in universities has dwindled to almost nothing."

You also came up with a new hardware description language: Gate LangUAGe, or GLUE.

This was in the early days before there was Verilog or VHDL, so we came up with our own HDL which, looking back, was very much like a rudimentary version of Verilog. We got some things pretty close to right in that regard.

Several of the tools you developed over the course of your work on those projects were pioneering in their own right—for instance, the architectural-level power simulator that eventually became SimplePower.

At the end of the MGAP Project, we started working on an accurate power simulation tool. Until then, we just focused on performance and area, but we were beginning to realize, as had many, that power was going to be the next big problem. Bob passed away in the very early stages of its design, so I continued on that work with a new Penn State colleague, Vijay Narayanan, and once again a bunch of graduate students. Most of our work since then has been this simulation-based work. There's been very little building.

Is that because of cost?

It's really, really expensive these days to build custom hardware at the current or near-current technology node. You just can't afford to do it at a university, so most architects do simulation-based research. Some design with FPGAs, and that's great, but the number of people building custom hardware in universities has dwindled to almost nothing.

What are you working on now?

Right now, some colleagues and I are working on a grant that examines reducing the power consumption of the non-core components of a chip. There are a lot of people who are looking at how to make the cores more energy efficient. We're looking at how to make the components that support the cores more energy efficient. We're focusing now on the cache space—how much of the cache you really need to have turned on, especially when some of the cores are turned off, and how you manage it—but there are a lot of non-core components to work on.

Throughout your career, you've been involved in advocating on behalf of women in the field.

I've been involved with outreach to women both through ACM (ACM-W) and in particular through the Computing Research Association's Committee on Women (CRA-W). And I still am involved in local campus activities and try to attend the Grace Hopper Celebration about every other year.

What's changed?

Unfortunately, the numbers at the undergraduate level have not come up like we had hoped. There are some locations where they've improved—Harvey Mudd, for example, and Carnegie Mellon, though I haven't seen their latest figures. But at many places, the number of undergraduate women in computer science and computer engineering is still in the low teens. My sense is that we do have more women in leadership positions.

At the graduate level, the numbers are better because we have women from offshore, which is great. Most of these women come to the U.S. because they want to get a Ph.D. If they're going to be adding to the infrastructure either here or in their home country, it's a great thing.

But the number of native-born U.S. women going into computer science and engineering is still way too low. It's a pipeline issue. So that's always a little frustrating for those of us who have spent a lot of time working to change it. 

Leah Hoffmann is a technology writer based in Piermont, NY.

© 2015 ACM 0001-0782/15/12 \$15.00

Q&A

Redefining Architectures

Mary Jane Irwin on building advanced circuits, special processors, and a hardware description language, while advocating for women in computer science.

MARY JANE IRWIN, Evan Pugh Professor and A. Robert Noll Chair in Engineering in the Department of Computer Science and Engineering at Pennsylvania State University, is as committed to her research as she is to serving the computer science community, in particular women in the field. Her interests and accomplishments span computer architecture, multicore systems design, and energy-aware design. She has also been active in the Computing Research Association's Committee on the Status of Women in Computing Research (CRA-W), the Grace Hopper Celebration of Women in Computing, the Board on Army Science and Technology, and the National Academy of Engineering's Membership Policy Committee. She has been deeply involved with ACM as well, co-founding the *Journal on Emerging Technologies in Computing Systems* and serving as an elected member of ACM's Council, as vice president from 1997 to 1998, and as editor-in-chief of ACM's *Transactions on the Design Automation of Electronic Systems* from 1998 to 2004.

You grew up in Memphis [Tennessee], where your dad was a professor at what was then Memphis State University. What drew you to computer science?

I was good at math, but it was early in the days of computing and there was no degree in computer science at Memphis State at the time. So I majored in math and took all the computing-related courses that they had. Then I went to graduate school at the University of Illinois at Urbana-Champaign (UIUC)



to become a professor, although at the time I really had no idea what that entailed, other than college-level teaching. UIUC was the luck of the draw because my husband got a job there, so that's the only place I applied. In that case, my draw was very lucky!

At UIUC, you worked with James E. Robertson, and you wrote your dissertation on computer arithmetic.

It's a challenge to build arithmetic components so that they run as fast as possible without consuming a lot of power, and of course they're a central part of what's going on in the CPU. Jim Robertson was very famous in computer arithmetic. I was attracted to that research area, but I also think there was some empathy between us, because I was one of the

"It's a challenge to build arithmetic components so that they run as fast as possible without consuming a lot of power, and of course they're a central part of what's going on in the CPU."

very few women in UIUC's computer science graduate program, and Jim was of American Indian descent. He'd spent time on the reservation growing up, and he knew what it was like to be from an underrepresented group in a majority population.

And you've been at Penn State ever since?

Next year I celebrate my 40th anniversary at Penn State. I used to tell my students, "I've been on the faculty longer than you've been alive." Now I can say, "*much* longer than you've been alive."

How did your research interests evolve?

I was also really interested in hardware design, in building advanced circuits and special processors. At Penn State, I had a number of strong grad students,

[CONTINUED ON P. 135]

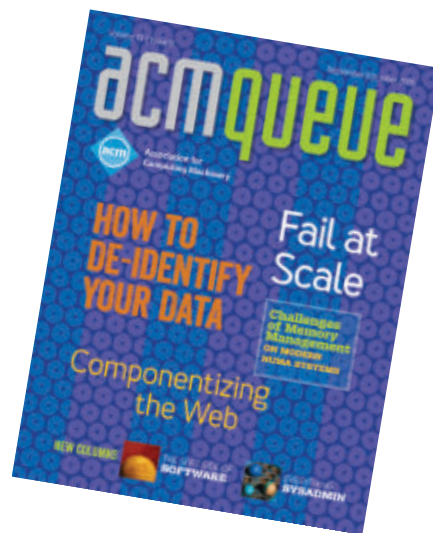
acmqueue

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.



fuse **DIS 2016**

Designing Interactive Systems

June 4 – 8

Brisbane Australia

 dis2016.org

 bit.ly/dis16

 [@DIS2016](https://twitter.com/DIS2016)



Association for
Computing Machinery



Queensland University of Technology
Brisbane Australia