

COMMUNICATIONS

CACM.ACM.ORG

OF THE

ACM

11/2017 VOL.60 NO.11



Reconfigurable
Cambits

**Previous
A.M. Turing Award
Recipients**

1966 A. J. Perlis
1967 Maurice Wilkes
1968 R.W. Hamming
1969 Marvin Minsky
1970 J.H. Wilkinson
1971 John McCarthy
1972 E.W. Dijkstra
1973 Charles Bachman
1974 Donald Knuth
1975 Allen Newell
1975 Herbert Simon
1976 Michael Rabin
1976 Dana Scott
1977 John Backus
1978 Robert Floyd
1979 Kenneth Iverson
1980 C.A.R Hoare
1981 Edgar Codd
1982 Stephen Cook
1983 Ken Thompson
1983 Dennis Ritchie
1984 Niklaus Wirth
1985 Richard Karp
1986 John Hopcroft
1986 Robert Tarjan
1987 John Cocke
1988 Ivan Sutherland
1989 William Kahan
1990 Fernando Corbató
1991 Robin Milner
1992 Butler Lampson
1993 Juris Hartmanis
1993 Richard Stearns
1994 Edward Feigenbaum
1994 Raj Reddy
1995 Manuel Blum
1996 Amir Pnueli
1997 Douglas Engelbart
1998 James Gray
1999 Frederick Brooks
2000 Andrew Yao
2001 Ole-Johan Dahl
2001 Kristen Nygaard
2002 Leonard Adleman
2002 Ronald Rivest
2002 Adi Shamir
2003 Alan Kay
2004 Vinton Cerf
2004 Robert Kahn
2005 Peter Naur
2006 Frances E. Allen
2007 Edmund Clarke
2007 E. Allen Emerson
2007 Joseph Sifakis
2008 Barbara Liskov
2009 Charles P. Thacker
2010 Leslie G. Valiant
2011 Judea Pearl
2012 Shafi Goldwasser
2012 Silvio Micali
2013 Leslie Lamport
2014 Michael Stonebraker
2015 Whitfield Diffie
2015 Martin Hellman
2016 Sir Tim Berners-Lee

ACM A.M. TURING AWARD NOMINATIONS SOLICITED

Nominations are invited for the 2017 ACM A.M. Turing Award. This is ACM's oldest and most prestigious award and is given to recognize contributions of a technical nature which are of lasting and major technical importance to the computing field. The award is accompanied by a prize of \$1,000,000. Financial support for the award is provided by Google Inc.

Nomination information and the online submission form are available on:
http://amturing.acm.org/call_for_nominations.cfm

Additional information on the Turing Laureates is available on:
<http://amturing.acm.org/byyear.cfm>

The deadline for nominations/endorsements is January 15, 2018.

For additional information on ACM's award program please visit: www.acm.org/awards/



Association for
Computing Machinery

Introducing *ACM Transactions on Human-Robot Interaction*

Now accepting submissions to ACM THRI

In January 2018, the *Journal of Human-Robot Interaction* (JHRI) will become an ACM publication and be rebranded as the *ACM Transactions on Human-Robot Interaction* (THRI).

Founded in 2012, the *Journal of HRI* has been serving as the premier peer-reviewed interdisciplinary journal in the field.

Since that time, the human-robot interaction field has experienced substantial growth. Research findings at the intersection of robotics, human-computer interaction, artificial intelligence, haptics, and natural language processing have been responsible for important discoveries and breakthrough technologies across many industries.

THRI now joins the ACM portfolio of highly respected journals. It will continue to be open access, fostering the widest possible readership of HRI research and information. All issues will be available on the ACM Digital Library.

Editors-in-Chief Odest Chadwicke Jenkins of the University of Michigan and Selma Šabanović of Indiana University plan to expand the scope of the publication, adding a new section on mechanical HRI to the existing sections on computational, social/behavioral, and design-related scholarship in HRI.

The inaugural issue of the rebranded *ACM Transactions on Human-Robot Interaction* is planned for March 2018.

To submit, go to <https://mc.manuscriptcentral.com/thri>



Departments

- 5 **From the Co-Chairs of the ACM Student Research Competition Highlights of the ACM Student Research Competition**
By Laurie Williams and Doug Baldwin

- 6 **Cerf's Up Heidelberg Laureate Forum**
By Vinton G. Cerf

- 7 **Vardi's Insights Would Turing Have Won the Turing Award?**
By Moshe Y. Vardi

- 8 **Letters to the Editor They See What You See**

- 10 **BLOG@CACM Opportunities for Women, Minorities in Information Retrieval**
Mei Kobayashi describes activities to support diversity and inclusion at the annual meeting of the ACM Special Interest Group on Information Retrieval in Tokyo this summer.

- 27 **Calendar**

- 100 **Careers**

Last Byte

- 112 **Future Tense Butterfly Effect**
But, like the weather, what can anyone do about it?
By Seth Shostak

News



- 12 **A Block on the Old Chip**
Block copolymers may help transistors shrink to tinier dimensions.
By Neil Savage
- 15 **Censoring Sensors**
Amid growing outcry over controversial online videos, tech firms grapple with how best to police online advertising.
By Alex Wright
- 17 **Overcoming Disabilities**
Brain-computer interfaces hold the promise of fully featured replacements for body parts that don't work or are missing.
By Esther Shein

Viewpoints

- 20 **Legally Speaking Disgorging Profits in Design Patent Cases**
Does the recent U.S. Supreme Court decision in the *Apple v. Samsung* case represent a quagmire?
By Pamela Samuelson
- 23 **Computing Ethics Engaging the Ethics of Data Science in Practice**
Seeking more common ground between data scientists and their critics.
By Solon Barocas and danah boyd
- 26 **Education Keeping the Machinery in Computing Education**
Incorporating intellectual and developmental frameworks into a Scottish school curriculum.
By Richard Connor, Quintin Cutts, and Judy Robertson
- 29 **Viewpoint Pay What You Want as a Pricing Model for Open Access Publishing?**
Analyzing the "Pay What You Want" business model for open access publishing.
By Martin Spann, Lucas Stich, and Klaus M. Schmidt
- 32 **Viewpoint Social Agents: Bridging Simulation and Engineering**
Seeking better integration of two research communities.
By Virginia Dignum



Association for Computing Machinery
Advancing Computing as a Science & Profession

Practice



44

36 Hootsuite: In Pursuit of Reactive Systems

A discussion with Edward Steel, Yanik Berube, Jonas Bonér, Ken Britton, and Terry Coatta

44 Breadth and Depth

We all wear many hats, but make sure you have one that fits well.
By Kate Matsudaira

46 Is There a Single Method for the Internet of Things?

Essence can keep software development for the IoT from becoming unwieldy.
By Ivar Jacobson, Ian Spence, and Pan-Wei Ng



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

Contributed Articles



54

54 Cambits: A Reconfigurable Camera System

Multiple computational cameras can be assembled from a common set of imaging components.

By Makoto Odamaki and Shree K. Nayar



Watch the authors discuss their work in this exclusive *Communications* video.
<https://cacm.acm.org/videos/cambits>

62 User Reviews of Top Mobile Apps in Apple and Google App Stores

The varying review dynamics seen in different app stores can help guide future app development strategies.

By Stuart McIlroy, Weiye Shang, Nasir Ali, and Ahmed E. Hassan

About the Cover:



This month's covers depict a simple set of blocks that can be used to build a variety of cameras with very different functionalities. Four different *Communications* covers are circulating worldwide, each depicting a different "Cambit." Our thanks to Shree Nayar and Anne Fleming of Columbia University for hosting this photoshoot. Covers by Alexander Berg.

Review Articles



68

68 Healthcare Robotics

Healthcare robotics can provide health and wellness support to billions of people.

By Laurel D. Riek



Watch the author discuss her work in this exclusive *Communications* video.
<https://cacm.acm.org/videos/healthcare-robotics>

Research Highlights

80 Technical Perspective

Solving Imperfect Information Games

By David Silver

81 Heads-Up Limit Hold'em Poker Is Solved

By Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin

89 Technical Perspective

Exploring a Kingdom by Geodesic Measures

By Marc Alexa

90 The Heat Method for Distance Computation

By Keenan Crane, Clarisse Weischedel, and Max Wardetzky

Highlights of the ACM Student Research Competition

SINCE 2003, ACM in conjunction with Microsoft have sponsored research competitions for undergraduate and graduate students in computing. The competitions provide a vehicle for these students to present their original research before a panel of judges and attendees at well-known ACM-sponsored and co-sponsored conferences. The students have the opportunity to experience a research conference, get feedback on their research, meet academic and industrial researchers and other students, and appreciate the practical applications of their research. Student competitors also have the opportunity to sharpen communication, visual, organizational, and presentation skills in preparation for the SRC. Participation by undergraduates may be literally life-changing if they alter their career path to pursue graduate studies and research careers after experiencing a conference and competition.

The following process is used to select the SRC winners:

1. Each student submits an 800-word abstract of his or her research. The abstract is evaluated by a minimum of three reviewers. Feedback on the abstract is provided to the students.

2. The students submitting the highest-evaluated abstracts are invited to attend the conference and present their work. Typically, 10 graduate students and 10 undergraduate students are invited to compete at the conference.

3. At the conference, the students present their work in front of a poster. Any conference attendee may come and ask a student about their research. A minimum of five official SRC evaluators assesses each poster.

4. The top five graduate and top five undergraduate students advance to the semifinals. In the semifinals, each student makes a 10-minute pre-

sentation of their work followed by a five-minute question-and-answer session to official SRC judges and any conference attendees who attend the open session. At least five judges are assigned to the semifinals.

5. The conference's top three finalists in each category are chosen based on these presentations.

6. The first-place winners from each conference are invited to compete in the Grand Finals. These students submit an updated 4,000-word paper on their research that is judged by a panel of experts.

The winner of the Grand Finals and their advisors are invited to the Annual ACM Awards Banquet, where they are recognized for their accomplishments and can witness other luminaries in the computing field receive prestigious society awards, such as the ACM A.M. Turing award. The first-, second-, and third-place winners of each conference competition receive cash prizes, medals, ACM student memberships, and recognition among their peers and professors—recognition that strengthens their résumés. All students who are selected to participate in the conference-level competition receive \$500 travel funding for the conference.

In 2016–2017, competitions took place at 24 participating conferences sponsored by the following ACM SIGs: SIGACCESS, SIGARCH, SIGCHI, SIGCOMM, SIGCSE, SIGDA, SIGDOC, SIGGRAPH, SIGHPC, SIGMIS, SIGMOBILE, SIGPLAN, SIGACT, SIGSAC, SIGSOFT and SIGSPATIAL. More than 330 students participated in these competitions.

Students find the SRC highly rewarding. Representative comments include the following:

“Participating in the SRC was an amazing opportunity. It was my first time attending any conference, and it really showed me how to pitch my research project, and interact with other researchers. I will carry this experience

with me in my future academic and professional endeavors.”

— Michele Hu, Cornell Tech
ASSETS 2015

“It was a great opportunity to be able to present at ACM SRC the work developed during a study abroad experience in the USA. Working with students from other places with different backgrounds was an incredible experience. That was my first time presenting at a conference and it felt great to expose our research, discuss it, and get invaluable feedback. It was an amazing chance to train my research pitch and share ideas with others.”

— Clarissa Tuxen,
Fluminense Federal University
Grace Hopper 2016

To learn more about the SRC, visit <http://src.acm.org/>.

We acknowledge the dedication of the many volunteers who make the SRC work. Each participating conference has a student research competition chair and program committee who review the papers submitted by the students. This committee and additional volunteers judge the posters and presentations at the conference venue to select the winners. Evelyn Viegas at Microsoft supports the SRC program; she replaces Judith Bishop who was a dedicated Microsoft volunteer for many years. Finally, Nanette Hernandez of ACM handles all logistical aspects of running the SRC, such as sending materials to conference sites around the world, and handles the interactions with students and conference volunteers. 

Laurie Williams, a professor of computer science in the College of Engineering at North Carolina State University, Raleigh, and Doug Baldwin, a professor of mathematics at the State University of New York at Geneseo, serve as co-chairs of the ACM Student Research Competition.

Copyright held by authors.



DOI:10.1145/3148147

Vinton G. Cerf

Heidelberg Laureate Forum

It is fall in Heidelberg and the leaves on the trees are already turning. This is the fifth year of the Heidelberg Laureate Forum (<http://www.heidelberg-laureate-forum.org/>)

and it continues to be a highlight of the year for me and for about 250 others who participate. This year, computer science was heavily represented. There were fewer mathematicians, but they made up for smaller numbers by their extraordinary qualifications. A new cohort of laureates was added this year: recipients of the ACM Prize for Computing.^a

Quantum computing and machine learning were major foci of attention and the Hot Topic sessions drew on speakers in the research and private sectors beyond the normal cohort of laureates. I took several lessons away from the quantum computing discussions: serious progress is being made in multi-qubit hardware development; the somewhat inaptly named “Quantum Supremacy” challenge^b is being met by more than one organization; and programming languages for quantum computing are being developed.^c

The other hot topic was machine learning and artificial intelligence and impressive results were outlined for image recognition, speech understanding, machine translation, and self-supervised learning. Multilayer neural networks are producing interesting results such as the ability to turn an ordinary photograph into one that inherits the *style* of various well-known artists or schools of art. The ques-

tion of transparency of machine learning in which it is made clear where the deep learning takes place and what has been learned remains unanswered, as does an easy way to explain how the system has reached its conclusions. Mathematical and computational models of biological processes found their way onto the agenda, reinforcing the notion that we might someday understand better how these processes work by modeling them and predicting some of their behaviors.

Visits to local research and business centers were arranged, such as the European Microbiology Laboratory (EMBL), European Media Laboratory, German Cancer Research Center, Heidelberg Institute for Theoretical Studies, the Max Planck Institute for Astronomy, NEC, the Mathematics Center of the University of Heidelberg, and the Interdisciplinary Center for Scientific Computing and to SAP headquarters. I spent at least two hours with a researcher at EMBL discussing in depth intra- and inter-cellular communication.

Quantum computing and machine learning were major foci of attention at this year's Heidelberg Laureate Forum.

Once again, a boat trip on the Neckar River afforded several hours of free discussion among laureates and students amidst the beautiful fall colors in the trees along the journey. A Bavarian night and a visit to the amazing Speyer Museum^d also provided ample interactive opportunities in addition to which the Forum added office hours, student working groups, and discussion groups to stimulate student/laureate interactions. A final visit to the 17th-century castle overlooking Heidelberg^e rounded out the week.

By chance (or perhaps by careful planning), the celebration of the linking of Heidelberg to Palo Alto, CA, as sister cities was hosted in the Town Hall of Heidelberg. A number of HLF participants live or have lived in the Palo Alto area and represented the city along with its mayor and other dignitaries. Of course, the Mayor of Heidelberg and the City Council were in attendance. The stained glass walls of the ceremonial room of the Town Hall had dates in the 1300s—something that is fairly mind-boggling for Americans whose sense of national history tends to start in the late 1700s!

This event continues to represent a remarkable gathering of brilliant and energetic minds and ACM and the other participating organizations owe a great debt to the Klaus Tschira Foundation and the Heidelberg Laureate Foundation for their adept and efficient support for this signature meeting. We remember gratefully Klaus Tschira, a founder of the German SAP company, who passed away far too soon in 2015. C

d <https://speyer.technik-museum.de/en/>

e https://en.wikipedia.org/wiki/Heidelberg_Castle

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

Copyright held by author.

a Formerly known as the ACM-Infosys Foundation Award in the Computing Sciences.

b That is, quantum computers in the 49–50+ qubit range that can theoretically perform functions beyond the capacity of current or projected conventional classical computing systems.

c https://en.wikipedia.org/wiki/Quantum_programming#Quantum_computing_language



Moshe Y. Vardi

DOI:10.1145/3144590

Would Turing Have Won the Turing Award?

IN 2017, WE celebrated 50 years of the ACM A.M. Turing Award, known simply as the Turing Award. The list of Turing Award winners (<http://amturing.acm.org>), starting from Alan Perlis in 1966, “for his influence in the area of advanced computer programming techniques and compiler construction,” to Sir Tim Berners-Lee in 2016, “for inventing the World Wide Web, the first Web browser, and the fundamental protocols and algorithms allowing the Web to scale,” offers a bird-eye view of the highlights of computing science and technology over the past 50 years. Justifiably, the Turing Award is often accompanied by the tagline “The Nobel Prize in Computing.” How did this prestigious award come to be?

The early history of the Turing Award is somewhat murky. The minutes of meetings of ACM Council from the mid-1960s shed some, but not complete light on this history. The Turing Award was not originally created as a “big prize,” but rather a lecture given at the annual ACM meeting. In August 1965, ACM Council considered and tabled a proposal that “the National ACM Lecture be named the Allen [sic] M. Turing Lecture.” In December 1965, ACM Council adopted the motion that “A.M. Turing be the name of the National Lectureship series.” In a 1966 meeting, ACM Council voted to name Alan Perlis as first lecturer. The minutes shed no light on *why* the lectureship was named after Alan Turing. The historical record is also not clear on how a lectureship turned into a major award. Perhaps there is a lesson here for ACM to keep better minutes of its Council’s meetings!

From today’s perspective, however, we can wonder whether ACM Council was justified in 1966 in naming its Na-

tional Lecture after Turing. Today, Turing is widely regarded as one of the most outstanding scientists of the 20th century, but that was not the case in 1966. The question, therefore, can be posed as follows: Had Turing been alive in 1966 (he died in 1954), would he have been selected for ACM’s first National Lecture?

A debate about Turing’s accomplishments has been going on for quite a while. In 1997, in an after-dinner speech in Cambridge, U.K., Maurice Wilkes, the 1967 Turing Award winner (for designing and building the EDSAC, the first stored-program computer in 1949), offered some biting comments about Turing: “However, on a technical level, of course I did not go along with his ideas about computer architecture, and I thought that the programming system that he introduced at Manchester University was bizarre in the extreme. ... Turing’s work was of course a great contribution to the world of mathematics, but there is a question of exactly how it is related to the world of computing.” (See Wilkes’s complete comments at <https://goo.gl/XkjM7n>.)

The controversy about Turing’s accomplishments flared again over the last few years. In a 2013 *Communications*’ editorial (<https://goo.gl/SpkhKw>) I argued that “The claims that Turing invented the stored-program computer, which typically refers to the uniform handling of programs and data, are simply ahistorical.” In response to this editorial, Copeland et al. argued in the 2017 *Turing Guide* (<https://goo.gl/DjC8uk>) that “Vardi is ignoring the fact that some inventions belong equally to the realm of mathematics and engineering. The Universal Turing Machine was one such, and this is part of its brilliance.” So who is right?

When it comes to historical interpretation, the same facts may lead different people to different interpretations, but one should pay attention to the facts! In August 2017, Leo Corry published an article in *Communications* on “Turing’s Pre-War Analog Computers: The Fatherhood of the Modern Computer Revisited” (<https://goo.gl/M7jCaj>) in which he carefully examined the purported connection between the “Universal Turing Machine,” as introduced in Turing’s 1936 paper and the design and implementation in the mid-1940s of the first stored-program computers. He concluded “There is no straightforward, let alone deterministic, historical path leading from Turing’s 1936 ideas on the Universal Machine to the first stored-program electronic computers of the mid-1940s.”

But the debate about how much credit Turing should get for the idea of the stored-program computer diminishes, in my opinion, from Turing’s actual contributions. The Turing Machine model offered a robust definition of computability that has been studied, refined, and debated since 1936, giving rise in the 1960s to computational complexity theory, a gem of theoretical computer science. Turing’s philosophical examination in 1950 of the possibility of machine intelligence is lucid and incisive today as it was then. Finally, we learned in the 1970s about Turing’s critical contributions to computer-aided code breaking.

Would Turing have won the Turing Award? My answer is, he should have!

Follow me on Facebook, Google+, and Twitter. 

Moshe Y. Vardi (vardi@cs.rice.edu) is the Karen Ostrum George Distinguished Service Professor in Computational Engineering and Director of the Ken Kennedy Institute for Information Technology at Rice University, Houston, TX. He is the former Editor-in-Chief of *Communications*.

Copyright held by author.

They See What You See

ANDREW CONWAY'S AND Peter Eckersley's Viewpoint "When Does Law Enforcement's Demand to Read Your Data Become a Demand to Read Your Mind?" (Sept. 2017) was an important contribution to the ongoing debate over electronic backdoors, whereby a backdoor is a means for accessing and exfiltrating user information not specifically authorized in advance by users. Here, I would like to outline several key aspects of that debate that also need to be addressed.

Although Conway and Eckersley did discuss the possibility that law enforcement could gain access to our most private thoughts, they did not mention a crucial near-term technology through which this exfiltration could happen. Within the next 10 years, "hologlasses," or holographic glasses, are projected by Apple, Facebook, Google, Microsoft, and Samsung, along with numerous startups, to become almost as common as cellphones are today, as reflected in the scale of their investment in its development. A backdoor in hologlasses could enable a "we see and hear what you see and hear" capability that would provide extraordinary insight into what users are thinking, as well as how they are behaving online and even in the physical world.

Also not mentioned was a legislative proposal that could facilitate mandatory backdoors for Internet of Things devices. In 2016, Senators Richard Burr (R., North Carolina) and Dianne Feinstein (D., California) introduced legislation—The Compliance with Court Orders Act—in the U.S. Senate to mandate providers of information products and services also provide unencrypted information on IoT devices to the government pursuant to court order. The result could be "Nothing is Beyond Our Reach," or no information is beyond the reach of law enforcement, likewise pursuant to court order. Similar legislation has been adopted in Australia, France, Germany, the U.K., and other

countries but so far has had only limited effect because these countries are not sufficiently powerful individually to enforce sanctions against large multinational foreign-domiciled IT providers. However, if Burr-Feinstein does indeed become law, then these countries might be more able to pursue mass surveillance domestically, as IT companies could lose much of the legal grounds they would need to resist.

Conway and Eckersley also did not mention a near-term technology that might be used to implement highly secure backdoors in IoT devices by requiring that each device have a different public key that could enable government security services to take over the device.¹ Even if hackers penetrated the security of a government-installed virtual machine for a device, they would gain no lasting advantage hacking additional devices.

Finally and most important, no mention was made of a technology proposal¹ that could ameliorate some of the negative effects of mass surveillance, whereby citizens' most sensitive information is stored on their own devices, provided personal IoT devices include protection against self-incrimination. By storing sensitive information on these devices, that information could be protected from the kind of efforts Conway and Eckersley identified.

Reference

1. Hewitt, C. Islets protect sensitive IoT information: Verifiably ending use of sensitive IoT information for mass surveillance can foster (international) commerce and law enforcement. Social Science Research Network WP 2836282; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2836282

Carl Hewitt, Palo Alto, CA

Authors Respond:

We generally agree with Hewitt who offers definite specific instances of the general issues we covered. Virtual reality, in particular, is, as he implies, a valuable window into the mind that also involves important technical and legislative dynamics. What to do about it is a complex question we did not attempt to answer in our Viewpoint beyond

framing its context for a wider societal discussion we consider essential.

Andrew Conway, Melbourne, Australia, and **Peter Eckersley**, San Francisco, CA

Bring 'Diseased' Software Up to Code

We agree with Vinton G. Cerf's advice in his Cerf's Up column "Take Two Aspirin and Call Me in the Morning" (Sept. 2017) that we all practice better "cyber-hygiene" but must quarrel with the continued use of public health as a metaphor for cyber security. If we as computing professionals intend to improve the cybersecurity of our critical infrastructures, rather than merely tolerate their current "diseased" state, we must think differently. We thus propose a return to an older metaphor for software, likening its structures to physical structures and its architecture to the architecture of physical buildings. Such thinking suggests we consider how to build software that will not fall over when attacked or build it from weak materials unable to bear expected stress.

Software we rely on for critical functions (such as controlling medical devices, delivering electrical power to households, and guiding automobiles) must conform to an appropriate set of constraints, just as physical structures conform to building codes before they can be occupied. A third party must be able to certify conformance to these constraints, just as building inspectors certify buildings.

These codes are best developed by those who build the systems, not by government, though governments might use them once they are in place. An industry-consensus building code, with third-party assessment of conformance, can help the marketplace reward those who build systems with fewer vulnerabilities.

Over the past few years, with support from the IEEE's Cybersecurity Initiative and the National Science Foundation, workshops have been held to begin to develop such build-

ing codes for medical-device software and for power-system software.^{1,2} In addition to these draft codes, related promising developments include *Consumers Reports'* collaboration with the Cyber Independent Testing Laboratory (<http://cyber-itl.org/>) to develop methods for publicly rating software products, and UL's (<http://www.ul.com>) development and use of a standard for certifying cybersecurity assurance of products.

Treating software security as a public health problem is not likely to lead past the decades-old ideas of aftermarket vaccines, antivirus, and quarantine. Providing evidence that software is at least free of specified classes of vulnerabilities covered by an appropriate building code can yield a more effective market incentive for companies to produce the cyberinfrastructures we all need—and that are up to code.

References

1. Haigh, T. and Landwehr, C. *A Building Code for Medical Device Software Security*. Technical Report. IEEE Computer Society, Mar. 2015; <https://www.computer.org/cms/CYBSI/docs/BCMDSS.pdf>
2. Landwehr, C.E. and Valdes, A. *Building Code for Power System Software Security*. Technical Report. IEEE Computer Society, Mar. 2017; <https://www.computer.org/cms/CYBSI/docs/BCPSSS.pdf>

Robert K. Cunningham, Lexington, MA,
Tom Haigh, Minneapolis, MN,
Carl Landwehr, New Buffalo, MI,
and **Alfonso Valdes**, Urbana, IL

Author Responds:

It is always a pleasure to hear from Carl Landwehr with whom I have had a long acquaintance and for whom I have great respect. An interesting challenge with his building/architecture metaphor relates to the way software is often constructed these days by incorporating (vast) libraries of code reflecting, perhaps, uncertain provenance. There is also the uncertainty of software interactions across the network that may never have been tested until a chance encounter leads to a breach. None of this invalidates the building-code metaphor but might make it more difficult to establish that the ensemble meets the desired code standards and properties. I am, in fact, very interested in the development of programming aids that will do a much better job of assessing source code against desirable properties of attack resistance and identifying potential sources of weakness. Until we

have reliable ways of producing the kind of code Landwehr et al. and I likely agree we need and want, we may also want to argue that infected or vulnerable devices ought not to be naively tolerated and that owners (and suppliers) bear at least some responsibility for observing diligent software hygiene.

Vinton G. Cerf, Mountain View, CA

A Neuron Net Is Also a Turing Machine

Leo Corry's article "Turing's Pre-War Analog Computers: The Fatherhood of the Modern Computer Revisited" (Aug. 2017) described the Turing machine as a purely mathematical notion. While Corry's argument was persuasive, there is indeed a direct connection from Turing's construction of the Turing machine to the Electronic Discrete Variable Automatic Computer (EDVAC) in the 1940s via the McCulloch-Pitts model of the brain. For example, during the discussion portion of a 1951 talk by John von Neumann, neuroscientist Warren S. McCulloch described the influence of Turing's original 1936 paper, saying:

"... I came, from a major interest in philosophy and mathematics into psychology with the problem of how a thing like mathematics could ever arise—what sort of a thing it was ... The attempt to construct a theory in a field like [neurophysiology], so that it can be put to any verification, is tough ... it was not until I saw Turing's paper that I began to get going the right way around, and with [logician Walter] Pitts' help formulated the required logical calculus. What we thought we were doing (and I think we succeeded fairly well) was treating the brain as a Turing machine."¹

McCulloch's and Pitts's 1943 paper emphasized the equivalence of artificial neuron nets to Turing machines, saying:

"It is easily shown: first, that every [neuron] net, ... can compute only such numbers as can a Turing machine ... This is of interest as affording a psychological justification of the Turing definition of computability and its equivalents, Church's λ -definability and Kleene's primitive recursiveness."

Moreover, in the first draft of the design for the EDVAC, von Neumann ex-

plicitly tied the computing elements for the EDVAC to McCulloch's and Pitts's model of a neuron, saying:

"Following W. Pitts and W.S. McCulloch [1943] ... we ignore the more complicated aspects of neuron functioning ... It is easily seen, that these simplified neuron functions can be imitated by telegraph relays or by vacuum tubes ... We propose to use them accordingly for the purpose described there: as the constituent elements of the devices, for the duration of the preliminary discussion ... The element which we will discuss, to be called an E-element ... which receives the excitatory and inhibitory stimuli, and emits its own stimuli along a line attached to it. ... In all this we are following the procedure of W. Pitts and W.J. MacCulloch."²

Since McCulloch and Pitts had shown that neuron nets are universal computing machines in the sense of the Church-Turing thesis, the same connection to universal computing machines would apply to the EDVAC.

References

1. von Neumann, J. The general and logical theory of automata. Chapter in *Cerebral Mechanisms in Behavior*, L.A. Jeffress, Ed. The Hixon Symposium, John Wiley & Sons, New York, 1951, 1–41.
2. von Neumann, J. *First Draft of a Report on the EDVAC* (June 30, 1945). Reprinted as a chapter in *Papers of John von Neumann on Computing and Computer Theory*, W. Aspray and A. Burks, Eds. MIT Press, Cambridge, MA, 1987, 17–82.

Brad Barber, Arlington, MA

Author Responds:

My entire argument was about Alan Turing's own views prior to the war, not about his influence on later developments. But a more general point I wanted to make was that scientific and technological ideas develop historically and that what happens later sometimes misleads us when we try to understand what happened earlier on. This may also be the case with McCulloch's very interesting, retrospective testimony, which by all means deserves a critical eye.

Leo Corry, Tel Aviv, Israel

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.

©2017 ACM 0001-0782/17/11

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:10.1145/3137627

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

Opportunities for Women, Minorities in Information Retrieval

Mei Kobayashi describes activities to support diversity and inclusion at the annual meeting of the ACM Special Interest Group on Information Retrieval in Tokyo this summer.



Mei Kobayashi
SIGIR 2017: Diversity and Inclusion

<http://bit.ly/2fFB1Uh>

August 13, 2017

こんにちは。ようこそ！

Hello and welcome!

Diversity was a central theme in the ACM SIGIR 2017 held in Shinjuku Ward in Tokyo, Japan. Upon arrival, all registrants were given a beautiful keychain and card as commemorative gifts from the local organizers to celebrate the 40th anniversary of the conference series:

"... the 40th Anniversary Logo... features Mt. Fuji, a view of Shinjuku skyscrapers, including the Tokyo Metropolitan Government (Office), as seen from Keio Plaza the conference hotel, and fireworks celebrating the 40th anniversary. The colorfulness of the fireworks and the circles within and enclosing the logo represent diversity and inclusion."

SIGIR 2017 featured a session on Women in IR (Information Retrieval) organized by Laura Dietz of the University of New Hampshire on the first day,

just before the welcome party. A week before the conference, I received an e-mail from the secretary of the session, Maram Hasanain, a graduate student in computer science (CS) at Qatar University, asking if I would like to prepare a one-minute introduction of myself for the session. I was so overwhelmed by her beautifully written e-mail, and the excitement of a first-time contact with someone from Qatar, that I immediately accepted her invitation.

The session started with one-minute presentations by:

► Vanessa Murdock, Principal Applied Researcher at Cortana Research, Microsoft, USA.

► Grace Hui Yang, associate professor at Georgetown University.

► Sahar Asadi of Spotify (digital music service), USA.

► Mei Kobayashi, an applied mathematician turned big data algorithms specialist. She is the first female manager in Customer Service of NTT Communications, Tokyo.

► Zehong Tan, Senior Software Engineer on the Search Team for eBay, USA.

► Maria Maistro, a Ph.D. student at the University of Padua, Italy, who works on IR evaluation with a focus on query log analysis to integrate user search behavior into the evaluation process. Maria is a co-author of *On Including the User Dynamic in Learning to Rank* (<http://dl.acm.org/citation.cfm?id=3080714>) at SIGIR 2017.

► Nazli Goharian, professor of computer science at Georgetown University, who works on health search and mining, including clinical and radiological reports, social media posts for mental health and adverse drug reactions, summarization, and decision support systems. She is a co-author of *Contextualizing Citations for Scientific Summarization using Word Embeddings and Domain Knowledge* (<https://arxiv.org/abs/1705.08063>) at SIGIR 2017.

► Tingting Dong, researcher at System Platform Research Laboratories, NEC, Japan, who works on diversification and summarization of video search results to provide well-organized and intuitive views for users.

► Harumi Murakami, professor of the Graduate School for Creative Cit-

ies and Vice Director of Media Center, Osaka City University.

► Xiaolu Lu, graduate student at Royal Melbourne Institute of Technology. She is a co-author of: *Can Deep Effectiveness Metrics be Evaluated Using Shallow Judgment Pools?* (<http://dl.acm.org/citation.cfm?id=3080793>) at SIGIR 2017.

► Zhuyun Dai, graduate student at Carnegie Mellon. She is a co-author of two papers at SIGIR 2017: *Learning to Rank Resources* (<http://bit.ly/2vMswxS>) and *End-to-End Neural Ad-hoc Ranking with Kernel Pooling* (<https://arxiv.org/abs/1706.06613>).

► Maram Hasanain, a Ph.D. student in computer science at the College of Engineering, Qatar University.

Murdock kicked off the lightning talks and got people chuckling with the closing line, “We’re hiring!” For better or for worse, it became a common theme among presenters from industry. Yang made everyone feel fantastic with her empowering statement, “*I just became associate professor!*” The audience broke out in a big, round of applause. I pointed out my chemistry and applied mathematics background and stated, “*Researchers in IR come from very diverse backgrounds, not just NLP.*” As it turned out, there was another mathematician in the crowd: Maistro. And Murakami, a SIGIR 2017 Committee Member, said her interest in working in IR is from a psychology perspective, her original area of expertise.

Side note: In addition to her work as a faculty member at Osaka City University, Murakami served as one of three SIGIR 2017 Social & Publicity Chairs, the other two being Yukino Baba of Kyoto University, and Falk Scholer of RMIT. During the SIGIR Business Meeting, chaired by Diane Kelly of the University of Tennessee, Knoxville, attendees learned that SIGIR 2017 had been expected to run a substantial deficit due to its expensive venue, Tokyo. However, the outstanding work of the Social & Publicity Chairs brought a record-breaking number of registrants (more than 200% of initial estimates) and drove conference coffers solidly into the black.

The closing one-minute presentation by Hasanain was a short video presentation. It was so impressive and heart-warming that I felt as though I knew her personally by

The diversity among the presenters dispelled any notions of a fixed template for success or stereotype of a woman in the sciences.

the end. Among all of the presentations, Hasanain’s received the greatest round of applause by far, proving to all the possibility of successful remote e-participation!

The second part of the session was a longer presentation by Hannah Bast of University of Freiburg, on identification of discrimination and stereotyping in the workplace. She presented types of follow-up actions an individual can take, such as, pointing out specific inappropriate behavior(s) or statistics to decision-makers, and proposing concrete methods to correct the situation. Although some in the audience may have attended similar sessions at other conferences, for some first-time attendees, the session was an eye-opener. Royal Sequeira, a graduate student at the University of Waterloo in Canada, sent me e-mail saying, “*Women in IR is one of my best experiences at SIGIR this year. While it has made me proud about the women in our community, it has also led me to introspect on several aspects of equity and diversity.*” The session was educational for me as well. I was shocked to learn that women were given the right to vote in all Swiss Cantons in 1991.

On a more positive note, it was exciting to hear about the work of women from so many different countries, backgrounds, and age groups. The diversity among the presenters dispelled any notions of a fixed template for success or stereotype of a woman in the sciences. All of us need to ensure that anyone with the interest, passion, and will to work hard can succeed. The consensus among attendees I talked with was very short presentations make a lasting impression; speakers have to distill their thoughts

into one main message and present it succinctly, in easily digestible form for the audience. Maistro noted that the short talks at the beginning of poster sessions were also effective, and they provided seeds for breaking the ice during the remainder of the conference. For myself, it was a lot of fun meeting male and female audience members during coffee breaks and evening parties who came up and introduced themselves.

A big thanks to Dietz and Hasainan for putting together this session, and to the SIGIR Organizing Committee Members who did not schedule competing parallel technical sessions so all could attend *Women in IR* without fear of missing out on learning about new technical work. Dietz had a busy week: she co-organized a full-day workshop (one of eight workshops at SIGIR 2017) with Edgar Meij (Bloomberg) and Chenyan Xiong (Carnegie Mellon University) on *Knowledge Graphs and Semantics for Text Retrieval and Analysis*. And as the first Student Affairs Chair for SIGIR, she organized a student buddy program for first-time attendees of SIGIR, and an inexpensive and enjoyable get-together featuring a karaoke stage at a nearby food court for students at the beginning of the conference.

Final Note: To promote awareness of and respect for diversity—including geographic, religious, and dietary, as well as gender diversity—this year’s SIGIR Conference featured a *Diversity and Inclusion Luncheon* with delicious and healthy vegetarian and Halal dishes. Since it took place on the third day of the conference, huge boats of fresh vegetables and fruit were welcomed by all participants. All meals during the conference—including the main banquet at the historic Hotel Chinzanso—were clearly marked vegetarian, Halal, etc. They were buffet or boxed to enable everyone to sit together, chat and mingle.

To the Conference Organizers, particularly the Social Chairs for their thoughtfulness and impeccable organization skills—Cheers! 

Mei Kobayashi is manager, Data Science/Text Analysis at NTT Communications.

© 2017 ACM 0001-0782/17/11 \$15.00

A Block on the Old Chip

Block copolymers may help transistors shrink to tinier dimensions.

FOR DECADES, COMPUTERS have grown more powerful because chipmakers have been able to make ever-smaller transistors, allowing them to cram more onto a single chip. That steady march has always depended on optics—shorter wavelengths of light allowed chipmakers to draw smaller lines for circuit paths, which then can be closer together. It has become increasingly harder, however, to reach the high resolutions needed for ever-tinier features.

The answer, or at least part of it, may lie not with optics at all, but with chemistry. Researchers in both industry and

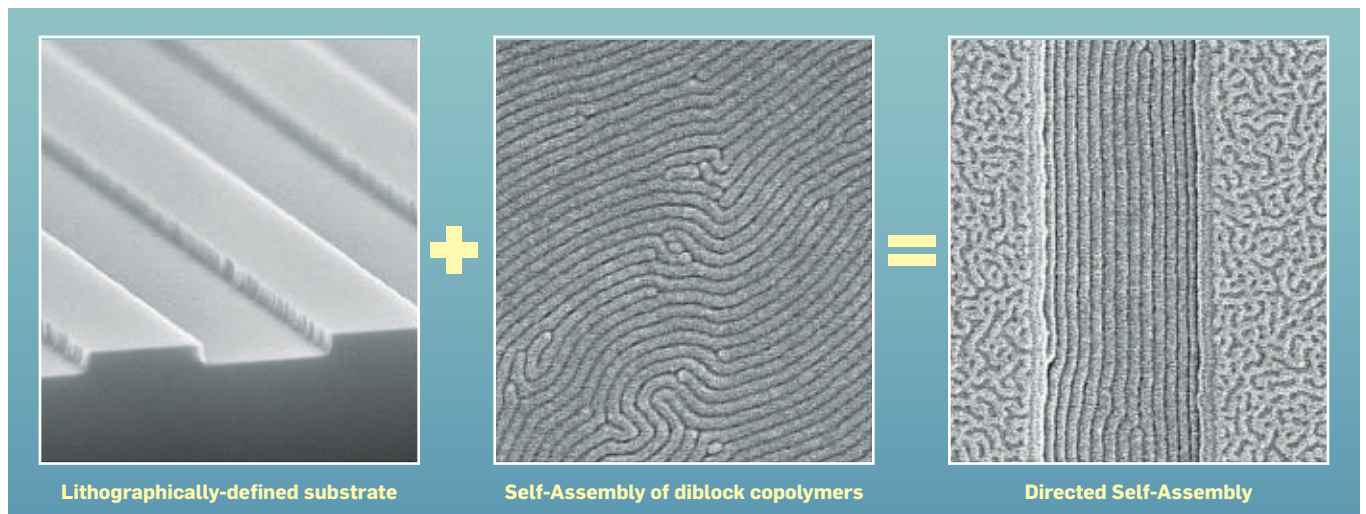
academia are trying to perfect a process that would let chemicals arrange themselves into tiny lines to serve as a pattern for the circuits. The lines and spaces in most chips now are on the order of 40 nanometers (nm) wide, but are expected to drop to less than 10 nm within a few years.

“Ten nanometers is really where they’re going to be forced to implement some new technology,” says Christopher Ober, professor of materials engineering at Cornell University in Ithaca, NY. The technology Ober and others are interested in is directed self-assembly (DSA), in which materials

called block copolymers arrange themselves into a desired pattern.

Block copolymers consist of two different materials that do not like to mix, like oil and water, but they are bonded together, so they cannot fully separate. When liquid block copolymers are heated, they form a structure in which each polymer has the least possible contact with the other. They might, for instance, form alternating stripes of each polymer, or create a checkerboard pattern, depending on how they were designed.

In chip manufacturing, that would mean coating the silicon wafer with a



Self-assembly of block copolymers on a flat substrate normally create fingerprint-like patterns having nanodomains with no long-range order (left). To fabricate well-aligned nanostructures over a large area, the self-assembly of the block copolymers can be induced in a confined space on the sub-micrometer scale.

neutral chemical to act as a base, then adding the block copolymer on top, and heating the whole thing. The polymers would form alternating lines—think of them as red stripes and blue stripes—and one set, say the red, would be washed away. The remaining blue stripes would act as a pattern for the same etching process used in current photolithography to inscribe circuits into silicon.

Photolithography also relies on chemicals—photoresists that are patterned through exposure to light, creating a pattern that is used to define which parts of the wafer to etch. Many modern designs, however, require multiple masks, driving up costs. The machines for focusing the light beams to ever-smaller sizes are also expensive, so DSA could, the thinking goes, create smaller features and do it more cheaply.

A few years ago, DSA was seen as the next big thing, and many chipmakers were investing heavily in it, says Kurt Ronse, director of the advanced lithography research program at the Inter-university Microelectronics Center (imec) in Leuven, Belgium. The focus was on a block copolymer consisting of polystyrene and polymethyl methacrylate, or PS-PMMA, which was targeted for features with a 28nm pitch, in which lines and spaces are each 14nm wide. At the time, the anticipated next photolithography technology was extreme ultraviolet (EUV), but progress toward making it commercial seemed to be stalled.

Too Many Defects

The hype over DSA did not last, says Ronse. “Very quickly it turned out that one of the big issues was going to be defects,” he says. Sometimes the copolymer would leave a gap, or the stripes would get too close to each other or even cross, a phenomenon called “line edge roughness.” The undercoating might have a pinhole or a slight bump that would throw the process off.

“It uses thermodynamics to form the patterns,” says Charles Black, director of the Center for Functional Nanomaterials at Brookhaven National Laboratory in Upton, NY. “It usually ends up with a defect here or a defect there, and microelectronics is really intolerant of that.”

Because self-assembly is a thermodynamic process, it is controlled by differences in surface energy, and the top “surface” is open air.

Over time, Ronse says, researchers managed to reduce the level of defects, while at the same time, EUV was beginning to look more promising. The combination made it seem less likely that DSA would provide the low-cost alternative the industry sought. “After a couple of years of development, the window of opportunity for 28nm pitch was starting to close,” Ronse says.

Though the hype has faded, he says, work continues. Some of the companies imec works with have ended their DSA research, while others are still going, though they have directed some of their chemistry work toward developing photoresists for EUV. “They have not stopped DSA development, but by putting less effort in it, it definitely slows down,” Ronse says.

While research into DSA is not as aggressive as it once was, it has not stopped, says Christopher Ellison, associate professor of chemical engineering at the University of Minnesota. “I don’t think it’s going away,” he says. “It’s not growing at a rapid rate.”

Part of the reason for that slowdown may be that it is not the optical physics that the chipmaking industry has always relied upon. “It makes it challenging for industry to accept because it’s so radically different,” Ellison says.

Surface Tension

While PS-PMMA may seem less attractive than it once did because its resolution limit is 11 nm, chemists are pursuing other block copolymers that could go even smaller. Some materials have a high chi, a measure of how incompatible the two polymers are. High-chi co-

polymers can perform the same self-assembly as PS-PMMA, but at a smaller size. Ellison’s group has created structures just four to five nm wide.

The trouble with high-chi copolymers is that it can be difficult to get them to stand up straight. To form parallel strips, they need to orient themselves perpendicular to the surface they are on. Often, however, they will flop over onto their sides, destroying the pattern.

Because self-assembly is a thermodynamic process, it is controlled by differences in surface energy, and the top “surface” is open air. Some researchers, such as Ellison and his colleague C. Grant Willson, a professor of chemical engineering at the University of Texas at Austin, have been working on topcoats, films of chemicals that change the surface energy on top of the copolymer so that it aligns itself correctly.

Paul Nealey, a professor of molecular engineering at the University of Chicago and one of the pioneers of DSA, says there are three or four methods to control the orientation of high-chi materials. Nealey and Karen Gleason, a professor of chemical engineering at the Massachusetts Institute of Technology, Cambridge, MA, published a paper in *Nature Nanotechnology* in March in which they reported using a technique called initiated chemical vapor deposition to create a topcoat that forced the block copolymer to line up the way they wanted it. The approach allowed them to create features only 9.3nm wide.

Another approach exposes the copolymer to vapor from a solvent to change the balance of energy. A technique explored by IBM mixes in an additive to the copolymer that changes how it responds to surface energy at the air interface.

However they are controlled, block copolymers cannot do everything on their own. Some lithography must be used to inscribe guiding lines for them to follow—that is the “directed” part of DSA. Because the material forms lines much thinner than those inscribed, they require less-advanced photolithography processes. At some resolutions, the guiding pattern could be created by the current state-of-the-art process, 193nm immersion lithography, which uses a liquid to focus the light beam to a

smaller area. It could appeal to industry to use DSA with immersion lithography equipment that is already paid for, rather than needing to invest in new equipment for EUV. The extra expense of DSA is minimal, Ober says; “It’s just an extra bottle of photoresist.”

In fact, EUV and DSA may turn out to be complementary technologies, each doing things that are difficult for the other. The first use for DSA, the researchers say, will not be for drawing lines at all, but for controlling the size of the holes through which different layers of transistors are connected. Conventional lithography makes those holes, or vias, too large. They could be filled with a block copolymer, which would assemble to form a narrower channel through which connections could be made.

Design Challenges

One shortcoming of DSA is that it cannot inscribe sharp 90-degree turns, while photolithography can. Any of these technologies will impose some constraints on chip designers, but chip architects have learned to deal with the existing ones, Ober says. “You have a set

of shapes and forms you can produce,” he says. “People have got very clever at taking these limited set of shapes and learning how to arrange them.”

Researchers will continue to push ahead with DSA, Ronse says, seeking the right blend of polymers, working on ways to keep the chemicals consistent from one batch to the next, and finding ways to bring down the number of defects to acceptable levels.

At the same time, the industry will push forward with EUV. The technology has failed to materialize despite a couple decades of development, but it may have turned a corner; “Although it looked like it was never going to happen, now it looks really, really close,” Ober says.

While it is not certain what mix of technologies will succeed, the researchers are not worried about reaching the limits of semiconductor technology any time soon. “The demise of photolithography has been predicted forever,” Black says. “Probably for 20 years people have been saying ‘the end is here,’ and those photolithography guys have been coming up with ways to get around it.”

Further Reading

Self-Assembly: Lego Blocks in Nature
Paul Nealey, University of Chicago
<https://vimeo.com/138700499>

Suh, H.S., Kim, D.H., Moni, P., Xiong, S., Ocola, L.E., Zaluzec, N., Gleason, K., and Nealey, P.F. Sub-10-nm patterning via directed self-assembly of block copolymer films with a vapour-phase deposited topcoat, *Nature Nanotechnology* 12, 2017
<http://www.nature.com/nnano/journal/v12/n6/full/nnano.2017.34.html>

Sinturel, C., Bates, F.S., and Hillmyer, M.A. High χ -Low N Block Polymers: How Far Can We Go?, *ACS Macro Letters* 4, 2015

Jiang, J., Jacobs, A., Thompson, M. O., and Ober, C. K. Laser spike annealing of DSA photoresists, *J Photopolymer Science and Technology* 28, 2015.
https://www.jstage.jst.go.jp/article/photopolymer/28/5/28_631/_article

Lane, A.P., Maher, M.J., Willson, C.G., and Ellison, C.J. Photopatterning of Block Copolymer Thin Films, *ACS Macro Letters* 5, 2016
<http://pubs.acs.org/doi/abs/10.1021/acsmacrolett.6b00075>

Neil Savage is a science and technology writer based in Lowell, MA.

© 2017 ACM 0001-0782/17/11 \$15.00

Milestones

Computer Science Awards, Appointments

ACM, IEEE POSTHUMOUSLY NAME THACKER AWARD RECIPIENT

The late Charles P. “Chuck” Thacker has been named recipient of the ACM-IEEE CS Eckert-Mauchly Award for fundamental networking and distributed computing contributions including Ethernet, the Xerox Alto, and development of the first tablet computers.

Often hailed as an “engineer’s engineer,” Thacker made fundamental contributions across the full breadth of computer development, from analog circuit and power supply design to logic design, processor and network architecture, system software, languages, and applications. He passed away June 12 at the age of 74, after a brief illness.

The ACM-IEEE CS Eckert-Mauchly Award is known as the computer architecture community’s most prestigious

award. ACM and IEEE Computer Society co-sponsor the award, which was initiated in 1979. It recognizes contributions to computer and digital systems architecture and comes with a \$5,000 prize. The award was named for John Presper Eckert and John William Mauchly, who collaborated on the design and construction of the Electronic Numerical Integrator and Computer (ENIAC), the pioneering large-scale electronic computing machine, which was completed in 1947.

TWO RECEIVE GEORGE MICHAEL MEMORIAL HPC FELLOWSHIPS
ACM, IEEE, and the SC Conference have announced the 2017 recipients of the ACM/IEEE George Michael Memorial HPC Fellowships.

Endowed in memory of George Michael, one of the founding fathers of the SC

Conference series, the ACM/IEEE-CS George Michael Memorial Fellowships honor exceptional Ph.D. students throughout the world whose research focus areas are in high performance computing, networking, storage, and large-scale data analysis. ACM, the IEEE Computer Society, and the SC Conference support this award.

Shaden Smith of the University of Minnesota, and Yang You of the University of California, Berkeley, were chosen to receive 2017 ACM/IEEE-CS George Michael Memorial HPC Fellowships. Smith is being recognized for his work on efficient and parallel large-scale sparse tensor factorization for machine learning applications. You is being recognized for his work on designing accurate, fast, and scalable machine learning algorithms on distributed systems.

Smith’s research is in the general area of parallel and high performance computing, with a special focus on developing algorithms for sparse tensor factorization, which facilitates the analysis of unstructured and high dimensional data. Smith has made several fundamental contributions that already have advanced the state of the art on sparse tensor factorization algorithms.

Your research interests include scalable algorithms, parallel computing, distributed systems, and machine learning. You has made several fundamental contributions that reduce the communications between levels of a memory hierarchy or between processors over a network.

The Fellowships include a \$5,000 honorarium and travel expenses to attend SC17 in Denver, CO, Nov. 12–17, where the Fellowships will be formally presented.

Censoring Sensors

Amid growing outcry over controversial online videos, tech firms grapple with how best to police online advertising.

FOLLOWING THE WAVE of U.K. terror attacks in the spring of 2017, prime minister Theresa May called on technology companies like Facebook and YouTube to create better tools for screening out controversial content—especially digital video—that directly promotes terrorism.

Meanwhile, in the U.S., major advertisers including AT&T, Verizon, and Wal-Mart have pulled ad campaigns from YouTube after discovering their content had been appearing in proximity to videos espousing terrorism, anti-Semitism, and other forms of hate speech.

In response to these controversies, Google expanded its advertising rules to take a more aggressive stance against hate speech, and released a suite of tools allowing advertisers to block their ads from appearing on certain sites. The company also deployed new teams of human monitors to review videos for objectionable content. In a similar vein, Facebook announced that it would add 3,000 new employees to screen videos for inappropriate content.

Yet these kinds of manual fixes won't solve the problem, especially as video becomes more and more ubiquitous thanks to smartphones, car cameras, and other embedded capture devices. Ultimately, monitoring user-generated content at scale will demand both computational and policy solutions.

At the heart of this problem lies a conflict that is deeply embedded in the history of the Internet itself: between the network's capacity for unleashing creative self-expression, and platform providers' business need to give advertisers' control over their messages. Caught between these opposing forces, companies like Google and Facebook make judgment calls about what kinds of content to deem acceptable—decisions that are often obscured from public view.

"These platforms and the firms that control them have largely been left to



their own devices in terms of developing ethics," says Sarah T. Roberts, an assistant professor in the Department of Information Studies of the University of California, Los Angeles. "The public is starting to ask questions about the power these companies have."

While certain types of content clearly violate the law (child pornography, for example), far more material falls into a vast grey area ranging from the mildly insensitive or tasteless to outright hate speech. Much of this content falls well within the free speech protections of the First Amendment in the U.S., yet remains unpalatable to advertisers who fear being associated with anything that could alienate potential consumers.

In the absence of government-enforced laws and regulations, however, what ethical obligations do these companies have toward safeguarding the public digital commons that provides the foundation for their businesses?

"It's hard to argue with a straight face about freedom of speech online when that speech is commoditized and highly lucrative," says Roberts. "This is a particular version of free speech that is deeply influenced by the politics and ethos of Silicon Valley."

Laws and regulations around freedom of expression also vary widely from country to country. For example, most European Union countries place much tighter restrictions around hate speech than the U.S. does, and a political video commentary that's considered satirical in one country might be considered

treasonous in another. Navigating this shifting terrain of international laws, regulations, and advertiser sensibilities—while continuing to provide as open a forum as possible to grow their audiences—presents companies like Google and Facebook with a complex, multidimensional challenge.

In an attempt to give the public some visibility into its internal dialogue around these questions, Facebook launched a series of blog posts called "Hard Questions" that provides a window into the company's current thinking.

"How aggressively should social media companies monitor and remove controversial posts and images from their platforms?" Facebook vice president of Public Policy and Communications Eric Schrage wrote in an introductory post. "Who gets to decide what's controversial, especially in a global community with a multitude of cultural norms?"

The company currently employs a combination of image matching, language understanding, and human monitoring to identify and remove Facebook groups that promote terrorist activity around the world.

"Although our use of AI against terrorism is fairly recent, it's already changing the ways we keep potential terrorist propaganda and accounts off Facebook," wrote Facebook's head of Global Policy Management Monika Bickert in a recent blog post.

Policy matters aside, the sheer vastness of the Internet—with over one

billion videos on YouTube (and more than 400 hours of new content being uploaded every hour)—introduces further complexities in terms of both process and technology.

Many researchers think the long-term solution to monitoring online content will inevitably involve artificial intelligence. Recent advances in big data, machine learning, and embedded Graphics Processing Units (GPUs) are starting to pave the way for more scalable approaches to computer vision, allowing neural networks to identify emerging patterns in user-generated content that may demand further human scrutiny.

“With machine learning, we can now understand a lot more about what’s going in a video or an image,” says Reza Zadeh, an adjunct professor at Stanford University and founder and CEO of Matroid, a Palo Alto, CA-based computer vision software start-up that is developing tools for video analysis.

Built atop TensorFlow (Google’s open-source library for machine intelligence), Matroid uses a video player coupled with a so-called detector program to identify similar images in a given video stream. For example, a Matroid detector could look for images of Donald Trump across five hours of video—like a few weeks’ worth of network news broadcasts, or large volumes of YouTube videos—and pinpoint the spots where those images appear. It can also easily detect images containing gore or violence, nudity and other forms of NSFW (not safe for work) content, and look for “more like this” elsewhere in other video streams.

The company offers a self-service tool for non-technical users to train the system to spot particular images, as well as a more advanced version geared toward machine learning engineers that enables them to edit the neural network architecture, explore histograms, and ultimately create their own detectors for others to use.

While deep learning approaches are yielding advances in analyzing videos and other image-based content, the wide variety in the type and quality of video across different capture devices poses additional obstacles.

“Applying machine learning techniques to analyzing video content works reasonably well when the right conditions exist,” says George Awad, project

Even if it were possible to identify visual elements across all kinds of video files with 100% accuracy, that alone wouldn’t solve the problem of screening for objectionable material.

director of TRECVID, a U.S. National Institute of Standards and Technology (NIST)-sponsored project that evaluates video search engines and explores new approaches to content-based video retrieval. “The major challenges occur when dealing with user videos in the wild that are not professionally edited.”

Even if it were possible to identify visual elements across all kinds of video files with 100% accuracy, that alone wouldn’t solve the problem of screening for potentially objectionable material. Much of the “content” of a video involves spoken words, after all, or other contextual cues that won’t be readily apparent from simply identifying an image. It’s notoriously difficult for computers to distinguish news from satire, for example—or an editorial opinion piece about terrorism from a call to arms.

In order to automate the process of screening video content at scale, researchers will likely need to apply natural language search techniques to begin parsing videos for deeper levels of nuance. “The gap between what the videos demonstrate and what an automated system would generate for a natural language description is still very big and challenging,” says Awad.

Looking ahead, Zadeh also sees plenty of opportunity on the hardware front, with semiconductor makers devising computer vision-capable chips that can work on devices like next-generation smartphones and cameras, self-driving cars, and a wide range of

other video-capable devices throughout our homes and offices.

Whereas today, machine learning happens primarily over the network—with supercomputers in datacenters analyzing big datasets stored in the cloud—eventually some of those processes will migrate toward edge-layer devices. Over time, the task of identifying objectionable content may become more diffuse, as computer vision algorithms increasingly come pre-coded on chips embedded on these devices. “The more algorithms move to the source of data capture, the more challenging it will be to cope with real-world factors,” says Awad.

Ultimately, the future of monitoring digital content may have less to do with policy-making and brute-force processing at the platform provider level, and more to do with algorithmic filters making their way into the devices all around us—for better or worse.

As Zadeh puts it: “Computers are opening their eyes.” 

Further Reading

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., and Vijayanarasimhan, S.
YouTube-8M: A Large-Scale Video Classification Benchmark. arxiv.org/abs/1609.08675

Bickert, M.
Hard Questions: How we Combat Terrorism, Facebook blog post, <https://newsroom.fb.com/news/2017/06/how-we-counter-terrorism/>

Hegde, V., Zadeh, R.
FusionNet: 3D Object Classification Using Multiple Data Representations, 3D Deep Learning Workshop at NIPS 2016. Barcelona, Spain, 2016. http://3ddl.cs.princeton.edu/2016/papers/Hegde_Zadeh.pdf

Real, S., Shlens, J., Mazzocchi, S., Pan, X., and Vanhoucke, V.
YouTube-Bounding Boxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video. (preprint) Accepted at the Conference on Computer Vision and Pattern Recognition (CVPR) 2017. [arXiv:1702.00824](https://arxiv.org/abs/1702.00824).

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., and Saenko, K.
Translating Videos to Natural Language Using Deep Recurrent Neural Networks. [arXiv:1412.4729 \[cs.CV\]](https://arxiv.org/abs/1412.4729)

Alex Wright is a writer and researcher based in Brooklyn, NY.

© 2017 ACM 0001-0782/17/11 \$15.00

Overcoming Disabilities

Brain-computer interfaces hold the promise of fully featured replacements for body parts that don't work or are missing.

IN THE MOVIE *Star Wars: The Empire Strikes Back*, Luke Skywalker is given a mechanical hand that moves and perform functions as well as his real hand. Konrad Kording, an avid Star Wars fan, has no doubt that advances in brain-machine interfaces (BMIs) will make this bit of science fiction a reality; he just doesn't know when.

"We have applications for one channel and a few channels," says Kording, a neuroscientist and professor of Physical Medicine and Rehabilitation, Physiology, and Biomedical Engineering at Northwestern University in Evanston, IL. "The question is, what are the BMI applications with hundreds of thousands of channels, and no one knows that at the moment." The channels he's referring to are electrical wires or optical connectors that can be attached to the brain and can be controlled and measured.

"The blind will see, and the amputated will move with limb replacements as well as you and me," Kording predicts.

A BMI is also sometimes referred to as a brain-computer interface (BCI), a mind-machine interface (MMI), or direct neural interface (DNI). The bottom line is such systems facilitate a direct communication pathway between an enhanced or wired brain and an external device.

BMIs are nothing new. Kording points to one of the earliest simple brain machines, the cochlear implant, which was invented in 1961. Deep brain stimulation, a surgical procedure to implant a hair-thin electrode wire in the part of the brain responsible for abnormal movement, is being used in treating diseases like Parkinson's and epilepsy. "What they all have in common," he says, "is that [they are] low-dimensional brain machine interfaces," meaning there are not many wires or channels going into the brain.

Research has been under way for sev-



Michel Fornasier, one of the presenters of the Cybathlon, uses his bionic hand prosthesis to demonstrate one of the Cybathlon disciplines.

eral years to study and facilitate BCI applications in the public, private, and educational sectors.

Most brain-computer interface research focuses on reading signals from the brain to the computer, usually EEG (electroencephalogram) or fNIRS (functional near-infrared spectroscopy), notes Robert Jacob, a computer science professor at Tufts University, Medford, MA. EEGs can be invasive, as when a patient's head is cut open to insert electrodes, whereas fNIRS is always non-invasive, he says. "There are other ways of reading information from the brain to the computer, like fMRI or MEG [magnetoencephalography], but these are generally too clumsy for reasonable BCI applications," Jacob says.

Most of this work used to be targeted at helping physically disabled people, but now, it is "spreading into more mainstream applications, where the brain information provides an extra

channel of input from the user to the computer, in addition to the keyboard, mouse, etc.," Jacob says.

The research organization BrainGate (<http://www.braingate.org>) has been working to develop and test BCI devices to restore the communication, mobility, and overall independence of people with a spinal cord injury, brain-stem stroke, or Amyotrophic lateral sclerosis (ALS). The Intelligence Advanced Research Projects Activity (IARPA), part of the U.S. Office of the Director of National Intelligence, sponsors research programs using multidisciplinary approaches "to advance understanding of cognition and computation in the brain."

In 2013, then-President Barak Obama announced \$100 million in funding for the Brain Research through Advancing Innovative Neurotechnologies (BRAIN) initiative to help neuroscientists understand the origins of cognition, percep-

tion, and other brain activities. The thought was the research could lead to new and more effective treatments for conditions like autism and mood disorders, as well as people suffering from brain injuries. As part of its mission, the BRAIN initiative will develop and deploy neurotechnologies to further understanding of the links between brain function and behavior.

This prompted federal agencies like the Defense Advanced Research Projects Agency (DARPA) to get involved, with the formation of the Neural Engineering System Design (NESD) program to enhance research capabilities in neurotechnology and provide an interface for new therapies.

So it was only a matter of time before BCI began receiving greater commercial attention from the likes of billionaire entrepreneurs like Elon Musk and Bryan Johnson. Musk, CEO of Tesla and SpaceX, is backing Neuralink, a BCI venture to create devices that can be implanted in human brains to eventually improve memory and interface with computer systems. Johnson, who founded online payments company Braintree (which was acquired by eBay in 2013 for \$800 million), has invested \$100 million into startup venture Kernel, whose goal is to enhance human intelligence with brain implants that can link people's thoughts with computers.

"Once one or two billionaires do something, everyone wants to be involved," observes Kording. "If I was a successful business person with the chance to build something really, really cool and be part of something that's really, really exciting, what's more interesting than the brain?"

That interest has spiked growth in the global BCI market, which was valued at \$723.64 million in 2014 and is expected to grow at a compound annual growth rate (CAGR) of over 10% through 2022, according to San Francisco-based Grand View Research.

"Rising occurrence of target disorders and neurodegenerative conditions are expected to propel market growth over the forecast period," the firm notes. "R&D efforts have led to pioneering in the engineering of headset development, which provides treatment for impaired cognitive function."

Healthcare applications make up by far the largest vertical segment, and ac-

counted for over 50% of the market in 2014, followed by communication and control, entertainment & gaming, and smart home control, according to the Grand View report.

Portland, OR-based Allied Market Research is forecasting the BCI market will reach \$1.46 billion by 2020, with a CAGR of 11.5% over that period.

The key factors driving this sector's growth are increasing focus on integrating the technology into various healthcare and military applications, as well as on developing communication technology for the disabled and geriatric populations, according to market research firm Future Market Insights, Valley Cottage, NY. The firm anticipates there will also be a growing focus on utilizing BCI to control Internet of Things (IoT) devices, smart home applications, and various virtual reality applications.

Fannie Liu, a Ph.D. student in the Human-Computer Interaction Institute at Carnegie Mellon University, agrees with the research, but stresses that use of BCIs right now is fairly limited.

"Industries have been focusing on gaming and medical applications, particularly for disabled or paralyzed individuals," she says, "though research in this field has shown potential in a variety of areas, such as communication, therapy, and controlling computer tools and devices."

Liu, who is working closely with her advisor, Geoff Kaufman, an assistant professor at the Institute, adds that while several applications are possible, BCI systems still face challenges in strik-

BCI systems face challenges in balancing the usability and accuracy of hardware and software before they are likely to be adopted by mainstream users.

ing a balance between the usability and accuracy of both hardware and software before they are likely to be adopted by mainstream users. "For instance, heart rate monitors, another type of biosignals system, could only become more widespread with the existence of practical and decently accurate wearables like chestbands and watches, alongside the development of a number of fitness apps," she says.

Already, systems allowing for gesturing and voice input have created new and intuitive ways for people to interact with computers without the need for the more traditional combination of keyboards and mice, Liu notes.

"Brain-computer systems push this a step further, with the goal of more directly using the brain to convey our intentions, rather than having an extra, physical step translating those intentions to text, speech, or gestures." This, she says, could make our interactions easier, faster, and ideally more natural.

Not only can these systems support human-computer interaction, but potentially human-human interaction as well, Liu says. Her work is investigating whether "we could potentially use BCI to better understand and communicate with each other," by clarifying our underlying thoughts and feelings.

In addition to Neuralink, a slew of other startups are also hoping to capitalize on the BCI market. Among them is NeuroPace, a medical device company aimed at improving the lives of patients suffering from epilepsy, which the company says affects approximately 1% of the population worldwide. NeuroPace has developed a medical device called the RNS System, which was approved by the FDA in November 2013 and which uses responsive neurostimulation (RNS) technology to monitor and control seizures in patients with drug-resistant epilepsy, according to CEO Frank Fischer. RNS is an approach to treating medically uncontrolled partial onset seizures, and works by continuously monitoring brain activity, detecting abnormal patterns, and in response, delivering imperceptible electrical pulses to normalize the activity before an individual experiences a seizure.

The RNS System can be used on individuals 18 and older to monitor and respond to brain activity in real time, preventing epileptic seizures at their source,

Liu is investigating whether “we could potentially use BCI to better understand and communicate with each other” by clarifying our underlying thoughts and feelings.

Fischer says. The system costs \$37,000, which includes the implantable device, leads, and remote monitor.

Both Kording and Liu caution that it will take years before BCIs become more widespread, with Liu observing that existing systems require a lot of setup. “Users need to be equipped either invasively through surgical implantation or noninvasively with several sensors or hardware that can take a while to calibrate.” As a result, right now, both options “are not practical for everyday use,” she says.

Although more usable consumer-grade devices have come out in the last decade, they tend to be less accurate, she observes. “Future iterations of BCI would likely continue to be noninvasive, but with new methods that can enable the recording and processing of stronger and better signals from the brain.”

Major hurdles like privacy concerns also need to be addressed. Liu says her research has found people are reluctant to allow systems to tap into their brains and understand their intentions, as they may feel their inner thoughts can be read and revealed to anyone with access to the system. “BCI systems need to protect users’ privacy if they are ever to become prevalent,” she says.

How to keep implants in the brain, particularly if a device involves wiring, also needs to be determined because the brain can reject them, adds Kording. “At the moment, we don’t have ways of interfacing with the brain in ways where devices will last a lifetime.”

Scientists and researchers also need to deal with issues involving the heat BCI devices give off, and how to analyze and control the data they produce. Un-

like the devices being developed by Musk and others, these are the higher-dimensional, more complex BCI systems, he says.

In addition to heat concerns, if the devices are not encased properly, they could potentially corrode and damage the brain, Kording adds.

“So there are lots of problems, but lots of opportunities,” and neurotechnology is helping keep technology so interesting, he says. “In some way, neurotechnology is the final frontier, because if we understand how the brain works, we can understand a lot of the things that ultimately affect us and make us be humans.” It is also probably the hardest area to figure out, Kording says.

“Understanding any other part of science is probably easier than understanding the brain and arguably, less exciting.” **C**

Further Reading

Jankowska, E.
Spinal control of motor outputs by intrinsic and externally induced electric field potentials. *Journal of Neurophysiology*. 24 May 2017. DOI: 10.1152/jn.00169.2017
<http://bit.ly/2sk6FIY>

Stevenson, I.H. and Kording, K.P.
How Advances in neural recording affect data analysis. *Nature Neuroscience*. 26 January 2011.
<http://go.nature.com/2rFboTQ>

Marblestone, A.H., Zamft, B.M., Maguire, Y.G., Shapiro, M.G., Cybulski, T.R., Glaser, J.I., Amodei, D., Stranges, P.B., Kalhor, R., Dalrymple, D.A., Seo, D., Alon, E., Maharbiz, M.M., Carmena, J.M., Rabaey, J.M., Boyden, E.S., Church, G.M., and Kording, K.P.
Physical principles for scalable neural recording. *NCBI*. 21 Oct., 2013.
<http://bit.ly/2tb11CQ>

Bansal, A.K., Truccolo, W., Vargas-Irwin, C.E., Donoghue, J.P.
Decoding 3D reach and grasp from hybrid signals in motor and premotor cortices: spikes, multiunit activity, and local field potentials. *Journal of Neurophysiology*. 1 March, 2012, Vol. 107, no. 5.
<http://bit.ly/2tw7uy3>

Pohlmeier, E.A., Oby, E.R., Perreault, E.J., Solla, S.A., Kilgore, K.L., Kirsch, R.F., and Miller, L.E.
Toward the Restoration of Hand Use to a Paralyzed Monkey: Brain-Controlled Functional Electrical Stimulation of Forearm Muscles. *PLOS Journals*. 15 June, 2009.
<http://bit.ly/2sa32XJ>

Esther Shein is a freelance technology and business writer based in the Boston area.

© 2017 ACM 0001-0782/17/11 \$15.00

ACM Member News

AT THE INTERSECTION OF PROGRAMMING LANGUAGES AND MACHINE LEARNING



“My goal is to make programmers more productive, and enable them to write programs

more easily,” says Ganesan Ramalingam, a principal researcher for Microsoft Research India. Ramalingam’s expertise is in static program analysis, or the identification of bugs and errors in code without having to execute the program being analyzed. “When the programs don’t work as they are expected to, you want to know what they are doing,” he explains.

Ramalingam earned his undergraduate degree in computer science from the Indian Institute of Technology, in Madras, and earned his Ph.D. in the discipline in 1993 from the University of Wisconsin–Madison. He then took a position at IBM’s T.J. Watson Research Center in New York, where he stayed until 2006, when he joined Microsoft Research India.

Currently, Ramalingam is primarily interested in programming languages and their interaction with machine learning, which he thinks has become more important thanks to rapid advances in this area over the past decade. “We are still in the early stages of machine learning, and it is quite challenging to build applications in this area,” he adds.

Ramalingam feels machine learning is going to be an increasingly important field, with a much wider spectrum of programmers wanting to incorporate it into their programs. “We need to make it easier for non-expert programmers to incorporate machine learning into their programs, and that these machine learning programs interact correctly with the rest of the programs these non-expert programmers write,” he says. “I think there be lots of interesting challenges here.”

—John Delaney



DOI:10.1145/3144170

Pamela Samuelson

Legally Speaking

Disgorging Profits in Design Patent Cases

Does the recent U.S. Supreme Court decision in the Apple v. Samsung case represent a quagmire?

AS REPORTED IN my March 2017 column, the U.S. Supreme Court struck down a \$399 million award against Samsung for infringing three Apple design patents. Samsung's win concerned an important but narrow issue. The Court ruled that Apple is entitled to be awarded Samsung's profits from sale of the article(s) of manufacture to which the protected designs were applied. However, lower courts erred in ruling that the relevant article of manufacture was necessarily the whole smartphone; it could instead be one or more components of the smartphones.

The *Apple v. Samsung* case has been sent back to the trial court to determine, first, to what relevant article(s) of manufacture were the patented designs applied, and second, what part of the \$399 million total profits Samsung made from sales of the infringing smartphones is attributable to the relevant article(s) of manufacture. The Court offered no guidance about how lower courts should make either assessment.

On remand, Apple is still insisting that the relevant article is the whole smartphone. (For the sake of brevity, I will use the term "article" instead of repeating "article of manufacture" or adopting AOM as an acronym.) Samsung, however, contends the relevant articles are the relatively small components that embody the three designs at issue (that is, a rectangular flat face with rounded corners, a rectangular flat face with bezel, and a colorful screen with 16 icons). Apple will be entitled to a much more modest award than \$399 million if Samsung prevails on this issue.

Samsung was not the only technology company relieved by the Supreme Court's ruling. Facebook, Google, eBay, Hewlett-Packard, and Dell, along with several high-technology industry associations, filed amicus curiae briefs in support of Samsung's appeal. The briefs argued that when design patent infringement occurs as to only one or a small number of components, it would be improper to disgorge total profits from the sale of multicomponent devices.

Samsung's victory notwithstanding, it is premature to assume the risk of excessive awards in design patent cases has subsided. The Supreme Court did not rule that the relevant article would necessarily and always be an individual component of a multicomponent device, only that it might be a component. How the relevant article and profits-from-that-article issues are resolved in *Apple v. Samsung* will have significant implications for future design patent cases involving multicomponent devices.

Recap on Design Patent Disgorgement

Disgorgement of infringer profits as a remedy in design patent cases came about because Congress got upset about a set of cases decided the 1880s. It reacted to Supreme Court rulings that the owner of a patent for a carpet design was entitled to only nominal damages (at the time, six cents, now worth about \$1.50) against two infringers. The Court rejected a plea for disgorgement, saying that the patentee had failed to prove what portion



of the profits were attributable to the infringing designs.

Soon thereafter, Congress effectively overturned the Court's ruling by creating a new remedy for design patent infringement. Under it, design patentees could elect, as an alternative to an award of actual damages (for example, lost profits or a reasonable royalty), a disgorgement of the defendant's "total profits" on the sale of any articles of manufacture embodying the protected design, or at least \$250 (approximately equivalent to \$6,100 today).

The legislative history made clear that courts should not try to determine how much of the defendant's profits were attributable to the infringing design, as opposed to the non-aesthetic attributes of the article. Instead, they should award total profits from the sale of the article of manufacture embodying the design. Congress recognized that this might overcompensate some design patentees, but that was better than undercompensating them. Moreover, the potential availability of a total profits award would also deter design patent

infringement more than an apportionment remedy would.

What's the Relevant Article?

In many, and perhaps most, design patent cases, the article embodying a patented design will be the product as a whole. The design of a carpet, wallpaper, or chair, for instance, will often be the main selling point for the product, even though the quality of the materials used in the manufacturing process, the skill of the craftsmen, and other characteristics of the product or producer may contribute to the value of the product. If, however, the design is what actually drives demand for the product, it seems fair that infringers should have to disgorge all profits.

A patented design may, of course, also be embodied in a component of a multicomponent product. Profits disgorgement should be relatively straightforward when there is a separate market for the component. Consider, for instance, a patented design for an interior light that turns on when a refrigerator door opens. Consumers are unlikely to buy refrig-

erators because they like the design of its interior light. However, if the light and switch mechanism is sold as a component to be incorporated into other refrigerator models or other products with doors, it should be possible to estimate what profits to disgorge against an infringer of this design. It would be patently unfair, though, to disgorge profits from infringer's sales of refrigerators when only the interior light embodies the patented design.

Identifying the relevant article for disgorgement purposes is more difficult, however, in cases involving complex multicomponent products, such as smartphones, when there is no separate market for components that may embody a patented design. What evidence or factors should a judge or jury consider in determining the relevant article in such cases?

The place to start a relevant article inquiry should be looking at the design patent, which is supposed to identify the article intended to embody the design. The patent must include a drawing of the patented design as it

would appear in the article. The patent may also have a textual description of the intended article. In addition, one should examine the patentee's and alleged infringer's products to discern how the patented design was embodied in the litigants' products. These types of evidence may sometimes suffice to identify the relevant article as to which disgorgement should occur.

In her argument before the Supreme Court, Samsung's lawyer recommended these steps for the relevant article inquiry. She also suggested that market studies might be useful to understand what consumers perceive the article embodying the design to be. Another indicator might be the costs incurred in developing the component embodying the design.

The Solicitor General of the United States, in a brief supporting Samsung's appeal and during oral argument, proposed consideration of four factors in making the relevant article determination: first, the scope of the patented design; second, the prominence of the design in the challenged product; third, whether the design is conceptually distinct from the product as a whole; and fourth, the physical relationship between the patented design and the rest of the product. The Supreme Court did not endorse use of these factors.

In support of its claim that the smartphone as a whole should still be considered the relevant article, Apple can be expected to argue that the patented designs are inseparable from the products embodying them, and that consumer demand for Apple products is due to its well-integrated designs. Would Samsung have sold so many millions of smartphones if it had not misappropriated the cool look of Apple designs? Apple thinks not.

Samsung will argue that the flat face of the smartphone and the flat face with bezel are two minor components of the exterior design of its smartphones. The colorful 16-icon design is similarly one small component of the screen displays of which smartphones are capable. These should be the three "articles" to which the Apple designs have been applied. Samsung will point to the very large number of components in smartphones to put the infringing components into proper perspective. Samsung could also point to consum-

One might consider how much total profit Samsung would have made from sales of smartphones if it had not infringed Apple's design.

er reports about Apple smartphones, which typically discuss features that consumers find most desirable.

What Profits to Disgorge?

Once the relevant article inquiry is resolved, the next question is what "total profits" did the infringer make from sale of that article. With a multicomponent product whose components are sold only as a package, "total profits" on one or more component-article(s) will almost certainly be some percentage of the profits made from sales of the infringing products. In *Apple v. Samsung*, \$399 million was determined as the total profit Samsung made selling the infringing smartphones. So if the relevant article is not the smartphone as a whole, how should one decide what portion of those profits are attributable to the components held to be the relevant article(s)?

Expert witnesses are likely to play a significant role in assessing a total-profits-attributable-to-the-relevant-article award in design patent cases. Experts hired by the patentee and by the alleged infringer are, of course, unlikely to agree on the bottom line. However, their assessments, as set forth in reports and testimony, will generally set the bounds within which the trier of fact, whether a judge or a jury, will make the award. Juries, in particular, are likely to consider the relative culpability of the infringer in making such awards. It is consistent with principles of unjust enrichment for them to do so.

Conjoint analysis may be a useful economic technique to contribute to a design patent profits disgorgement analysis. It is often used to analyze how consumers conceive, integrate, value, and trade off different clusters of prod-

uct and service features or attributes. By asking many people to express a preference between a few dozen pairs of designs that differ on features and attributes, product marketers can use clever mathematical techniques to estimate the importance of each feature in isolation. This technique has been used in some patent infringement cases, and would seem well suited for resolution of cases such as *Apple v. Samsung* in which design patents may cover only one or a small number of components of multicomponent products.

Alternatively, a profits disgorgement assessment might be built on a counterfactual scenario. As applied in *Apple v. Samsung*, one might consider how much total profit Samsung would have made from sales of smartphones if it had not infringed Apple's design patents and compare this estimate to the total profit Samsung actually made on infringing smartphones. The difference between the two profit scenarios would be the amount that Samsung should have to pay Apple under this model of the disgorgement remedy. This approach contrasts with a more compensatory approach that would focus on how much total profit Apple made from sales of its smartphones and how much profit it would have made if Samsung had not infringed. Economic experts could create models for undertaking these assessments.

Conclusion

Samsung won an important victory for itself and for other high technology companies in challenging the total profits award in the *Apple* case. However, uncertainty exists about how courts or juries should go about determining the relevant embodiment of a patented design to serve as the "article" on which the infringer's "total profits" should be disgorged. Because high-tech companies are utilizing design patents much more now than in the past, they have reason to worry about the legal quagmire hovering over disgorgement of profits awards in design patent cases involving multicomponent products. ■

Pamela Samuelson (pam@law.berkeley.edu) is the Richard M. Sherman Distinguished Professor of Law and Information at the University of California, Berkeley, and a member of the ACM Council.

Copyright held by author.

Computing Ethics

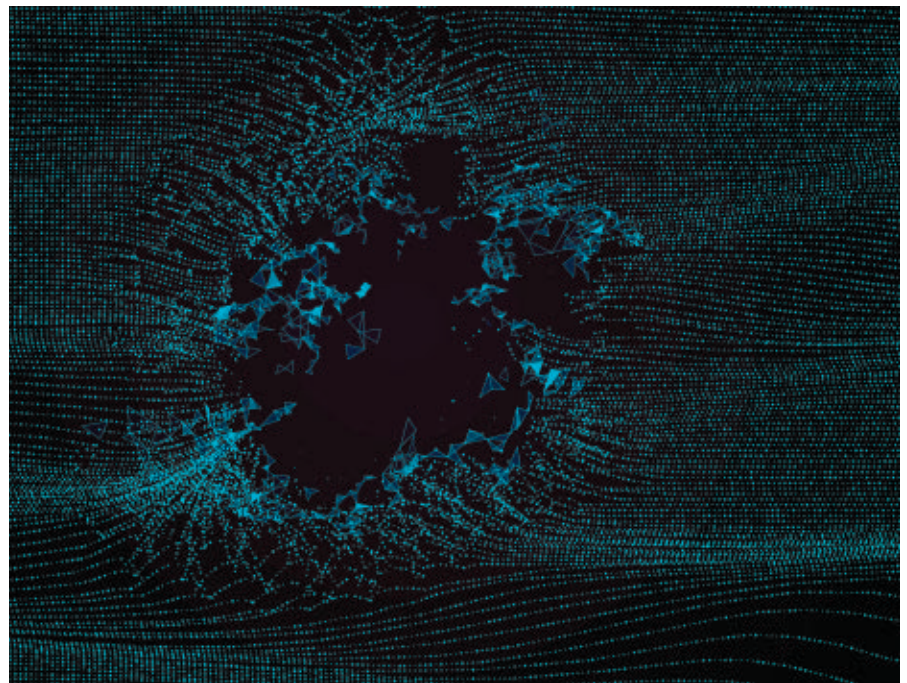
Engaging the Ethics of Data Science in Practice

Seeking more common ground between data scientists and their critics.

CRITICAL COMMENTARY ON data science has converged on a worrisome idea: that data scientists do not recognize their power and, thus, wield it carelessly. These criticisms channel legitimate concerns about data science into doubts about the ethical awareness of its practitioners. For these critics, carelessness and indifference explains much of the problem—to which only they can offer a solution.

Such a critique is not new. In the 1990s, Science and Technology Studies (STS) scholars challenged efforts by AI researchers to replicate human behaviors and organizational functions in software (for example, Collins³). The scholarship from the time was damning: expert systems routinely failed, critical researchers argued, because developers had impoverished understandings of the social worlds into which they intended to introduce their tools.⁶ At the end of the decade, however, Mark Ackerman reframed this as a social-technical gap between “what we know we *must* support socially and what we *can* support technically.”¹ He argued that AI’s deficiencies did not reflect a lack of care on the part of researchers, but a profound challenge of dealing with the full complexity of the social world. Yet here we are again.

Our interviews with data scientists give us reason to think we can avoid this repetition. While practitioners were quick to point out that common criticisms of data science tend to lack technical specificity or rest on faulty



understandings of the relevant techniques, they also expressed frustration that critics failed to account for the careful thinking and critical reflection that data scientists already do as part of their everyday work. This was more than resentment at being subject to outside judgment by non-experts. Instead, these data scientists felt that easy criticisms overlooked the kinds of routine deliberative activities that outsiders seem to have in mind when they talk about ethics.

Ethics in Practice

Data scientists engage in countless acts of implicit ethical deliberation while

trying to make machines learn something useful, valuable, and reliable. For example, dealing with dirty and incomplete data is as much a moral as a practical concern. It requires making a series of small decisions that are often fraught, forcing reflection at each step. How was this data collected? Does it capture the entire population and full range of behavior that is of interest? The same is true for validating a model and settling on an acceptable error rate. What must a data scientist do to prove to herself that a model will indeed perform well when deployed? How do data scientists decide that a reported error rate is tolerable—and defensible? Eth-

Call for Nominations for ACM General Election

The ACM Nominating Committee is preparing to nominate candidates for the officers of ACM: **President, Vice-President, Secretary/Treasurer; and two Members at Large.**

Suggestions for candidates are solicited. Names should be sent by **November 5, 2017** to the Nominating Committee Chair, c/o Pat Ryan, Chief Operating Officer, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA.

With each recommendation, please include background information and names of individuals the Nominating Committee can contact for additional information if necessary.

Alexander L. Wolf is the Chair of the Nominating Committee, and the members are Karin Breitman, Judith Gal-Ezer, Rashmi Mohan, and Satoshi Matsuoka.



ical considerations also emerge while making more fundamental decisions regarding the choice of learning algorithm, where practitioners frequently struggle to find an approach that maximizes the resulting models' performance while also providing some degree of interpretability. When is the ability to meaningfully interrogate a model sufficiently important to justify some cost in performance? What kinds of decisions—and real-world effects—drive data scientists to develop a model that they can explain, even if its decisions might be less accurate as a result?

These are difficult decisions, for which data scientists must employ carefully cultivated judgment. Yet, many data scientists do not use the language of ethics to talk about these practices. They may speak of trade-offs, but they primarily talk about what it takes to be *good* at what they do. Pressed about “ethics” directly, many data scientists say “this is not my area,” even though they draw on a wide range of values to work through difficult tensions.

Broad critiques of data science practices cannot account for the diversity of practices, concerns, or efforts among data scientists. Instead, they often presume ignorance or corrupt intentions. All too often, the data scientists we have encountered are quite sympathetic to the sentiment behind the critiques they hear, but feel maligned and misunderstood, unacknowledged for their efforts, and frustrated by vague recommendations that are not actionable. Outsiders' use of the term “ethics” suggests that normative concerns must be

Broad critiques of data science practices cannot account for the diversity of practices, concerns, or efforts among data scientists.

dealt with independently or on top of technical practice—without noticing that ethical deliberation is embedded in the everyday work of data scientists.

Even when attempting to address ethical issues more explicitly, practitioners face difficult trade-offs. One interviewee described a dilemma in choosing whether or not to “know” the gender of the individuals in his model—with that information, he could check whether his model might exhibit some kind of gender bias; without it, he could claim that this sensitive attribute did not figure into the model. Other researchers who are concerned about gender biases in data have attempted to build technical interventions to address them,⁵ but such an approach requires trading off privacy in order to construct a viable fairness remedy, a decision that presents its own challenges.⁴

Where Ethics Is Not Enough

Critics are right to emphasize the seriousness of the implications of data science. And, as Cathy O'Neil has pointed out in *The Weapons of Math Destruction*,⁸ data science is being deployed by powerful organizations to achieve goals that can magnify inequality and undermine democratic decision-making. She calls on data scientists to recognize how they are being used—and to push back against misuse of their skills.

Unfortunately, certain problems may stem from genuine value conflicts, not simply a lack of attention to the values at stake. Over the past year, a debate has unfolded over the use of data science in criminal justice, where courts rely on risk scores to make decisions about who should be released from prison while awaiting trial. The stakes are high: those given bail are more likely to keep their jobs, house, children, and spouse; those who are not are more likely to plead guilty, even when they are innocent.

A group of data scientists working with *ProPublica* established that black defendants in Broward County, FL, who did not reoffend were twice as likely to be mislabeled as posing a high risk of recidivism than white defendants.² They argued that the system exhibited a clear racial bias because errors imposed a far greater cost on black defendants, who were more likely to

be wrongly incarcerated, while white defendants were more likely to be set free but nevertheless recidivate. Northpointe (now Equivant), the company behind the risk assessment, countered that its tool was equally accurate in predicting recidivism for black and white defendants. Since then, computer scientists and statisticians have debated the different qualities that an intuitive sense of fairness might imply: that a risk score is equally accurate in predicting the likelihood of recidivism for members of different racial groups; that members of different groups have the same chance of being wrongly predicted to recidivate; or that failure to predict recidivism happens at the same rate across groups. While each of these expectations of a fair score might seem like complementary requirements, recent work has established that satisfying all three at the same time would be impossible in most situations; meeting two will mean failing to comply with the third.^{4,7} Even if Northpointe had been more sensitive to disparities in the false positive and false negative rates, the appropriate way to handle such a situation may not have been obvious. Favoring certain fairness properties over others could just as well have reflected a difference in values, rather than a failure to recognize the values at stake. One thing is for certain: this use of data science has prompted a vigorous debate, making clear that our normative commitments are not well articulated, that fuzzy values will be difficult to resolve computationally, and that existing ethical frameworks may not deliver clear answers to data science challenges.

Toward a Constructive Collaboration

The critical writing on data science has taken the paradoxical position of insisting that normative issues pervade all work with data while leaving unaddressed the issue of data scientists' ethical agency. Critics need to consider how data scientists learn to think about and handle these trade-offs, while practicing data scientists need to be more forthcoming about all of the small choices that shape their decisions and systems.

Technical actors are often far more sophisticated than critics at under-

Technical actors are often far more sophisticated than critics at understanding the limits of their analysis.

standing the limits of their analysis. In many ways, the work of data scientists is a *qualitative* practice: they are called upon to parse an amorphous problem, wrangle a messy collection of data, and make it amenable to systematic analysis. To do this work well, they must constantly struggle to understand the contours and the limitations of both the data and their analysis. Practitioners want their analysis to be accurate and they are deeply troubled by the limits of tests of validity, the problems with reproducibility, and the shortcomings of their methods.

Many data scientists are also deeply disturbed by those who are coming into the field without rigorous training and those who are playing into the hype by promising analyses that are not technically or socially responsible. In this way, they should serve as allies with critics. Both see a need for nuances within the field. Unfortunately, universalizing critiques may undermine critics' opportunities to work with data scientists to meaningfully address some of the most urgent problems.

Of course, even if data scientists take care in their work and seek to engage critics, they may not be well prepared to consider the full range of ethical issues that such work raises. In truth, few people are. Our research suggests the informal networks that data scientists rely on are fallible, incomplete, and insufficient, and that this is often frustrating for data scientists themselves.

In order to bridge the socio-technical gap that Ackerman warned about 20 years ago, data scientists and critics need to learn to appreciate each other's knowledge, practices, and limits. Unfortunately, there are few places in

which such learning can occur. Many data scientists feel as though critics only talk *at* them. When we asked one informant why he did not try to talk back, he explained that social scientists and humanists were taught to debate and that he was not. Critics get rewarded for speaking out publicly, he said, garnering rewards for writing essays addressed to a general audience. This was not his skillset nor recognized as productive by his peers.

The gaps between data scientists and critics are wide, but critique divorced from practice only increases them. Data scientists, as the ones closest to the work, are often the best positioned to address ethical concerns, but they often need help from those who are willing to take time to understand what they are doing and the challenges of their practice. We must work collectively to make the deliberation that is already a crucial part of data science visible. Doing so will reveal far more common ground between data scientists and their critics and provide a meaningful foundation from which to articulate shared values. 

References

1. Ackerman, M.S. The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Human-Computer Interaction* 15, (2–3), 2000, 179–203.
2. Angwin, J. et al. Machine bias. *ProPublica*. (May 23, 2016); <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
3. Collins, H.M. *Artificial Experts: Social Knowledge and Intelligent Machines*. MIT Press, 1993.
4. Corbett-Davies, S. et al. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post* (Oct. 17, 2016); <https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/>
5. Feldman, M. et al. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2015), 259–268.
6. Hess, D.J. Editor's Introduction, *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford University Press, 2001.
7. Kleinberg, J., Mullainathan, S. and Raghavan, S. Inherent Trade-Offs in the Fair Determination of Risk Scores. *Arxiv.org*. 2016; <https://arxiv.org/abs/1609.05807>
8. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.
9. Žliobaitė I. and Custers, B. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24, 2 (Feb. 2016), 183–201.

Solon Barocas (sbarocas@cornell.edu) is an Assistant Professor of Information Science at Cornell University.

danah boyd (danah@datasociety.net) is a Principal Researcher at Microsoft Research and the Founder/President of Data & Society.

Copyright held by authors.

Education

Keeping the Machinery in Computing Education

Incorporating intellectual and developmental frameworks into a Scottish school curriculum.

WE DO NOT think there can be “computer science” without a computer. Some efforts at deep thinking about computing education seem to sidestep the fact that there is technology at the core of this subject, and an important technology at that. Computer science practitioners are concerned with making and using these powerful, general-purpose engines. To achieve this, computational thinking is essential, however, so is a deep understanding of machines and languages, and how these are used to create artifacts. In our opinion, efforts to make computer science entirely about “computational thinking” in the absence of “computers” are mistaken.

As academics, we were invited to help develop a new curriculum for computer science in Scottish schools covering ages 3–15. We proposed a single coherent discipline of computer science running from this early start through to tertiary education and beyond, similar to disciplines such as mathematics. Pupils take time to develop deep principles in those disciplines, and with appropriate support the majority of pupils make good progress. From our background in CS education research, we saw an opportunity for all children to learn valuable foundations in computing as well, no matter how far they progressed ultimately.

Nobody knows exactly the right CS



curriculum for the average five-year old, as we have not taught them CS before, but we are unconvinced of the coherence of many current curricula: an underlying intellectual and developmental framework seems to be missing, and such a framework is our principal offering to the curriculum.

We understand both the desperate calls from industry to meet the labor market demands of the digital economy, and the extraordinary environment that will be our children's, with ever more blurring of digital and hu-

man worlds. Hence, we wanted a curriculum that properly grounds their understanding of that non-human world and gives every child the opportunity, should they wish, of a future career in our area. Our school systems have these aspirations in teaching about the natural world—why not the digital world also?

In March 2017, the new curriculum was formally adopted at government level, and its delivery has started. A teachers' guide is here—<http://www.teachcs.scot>—and we encourage inter-

ested readers to look at the full guide.

All curriculum design requires compromise. We have balanced: our initial vision of a curriculum that captures the essence of computation at the heart of the digital revolution; the practical realization that only a small amount of resources are available for teachers' professional development; the requirement to reuse a varied body of existing early-years computing educational material; and the desire from government to direct computing education down a narrow agenda to fill a perceived skills shortage.

Nonetheless, we have kept in view throughout our overarching framework consisting of three main points that we think is the real contribution of the curriculum, and the three points are the focus of this column. We will describe the essence of our proposed three-point underpinning, developing three essential strands of learning, and the way these have been eventually interpreted in the adopted curriculum.

Computational Foundations

We aimed to identify a core framework for the discipline that is equally relevant to a child, a university student, or a software engineer. The essence of computation is clear: the Church-Turing thesis. Some kind of computational mechanism—whether the Scratch programming environment, a Turing Machine, or the Lambda Calculus—can be used to model any tail-recursive numeric function ... and therefore anything that can be computed ... and furthermore all such mechanisms are somehow equivalent.

To be of interest, such mechanisms should be restricted to those that can perform some kind of *modeling* function over another domain or world. That is, they can be set up in such a way that their operation, when viewed in the context of the other domain, can be seen as simulating some aspect of that domain. Hence, a programming language can be used to model an aspect of the real world; a processor can be set up with appropriate machine code to model a computation expressed in a programming language; a lambda calculus expression, under the application of reduction, can provide the result of some recursive function.

All curriculum design requires compromise.

A deep understanding of computer science requires the following three aspects, our three-point framework, which can be neatly separated as the understanding of:

- Domains that can be modeled by computational mechanisms;
- The computational mechanisms themselves; and
- How to use the computational mechanisms to model aspects of the domains.

It is our belief that a computer scientist is habitually and implicitly aware of these, and indeed is expert at quickly assimilating new instances of them. We believe this is a core skill with many applications to a modern process- and information-driven world.

Computational thinking, as well as the learning delivered via the Unplugged approach, are, we believe, largely captured within the first aspect. The skill of programming, as taught even at university level, is mostly within the third. The second all-important aspect seems to be often neglected, at least until the later stages of a computing degree. It has long been a wry observation of the authors that, while “programming” is taught right from the start of university computing courses, more “advanced” topics such as programming language syntax and semantics are typically taught much later on. This begs the question: How can one learn to program in the absence of such knowledge? Research shows that concentrating on explaining how programs work, rather than writing them, helps students early on to learn programming. Could it be that we normally teach “by example” only, rather than ever properly defining the domain in which the modeling is performed, or even the domain being modeled?

Our Curriculum

The resulting curriculum is formally structured around these three as-

Calendar of Events

November 1–3

IMC '17: Internet Measurement Conference,
London, U.K.
Co-Sponsored: ACM/SIG,
Contact: Steve Uhlig,
Email: steve.uhlig@gmail.com

November 6–10

CIKM '17: ACM Conference on Information and Knowledge Management,
Singapore, Singapore,
Co-Sponsored: ACM/SIG,
Contact: Marianne Winslett,
Email: winslett@cs.uiuc.edu

November 7–10

SIGSPATIAL'17: 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems,
Los Angeles, CA,
Contact: Erik Hoel,
Email: ehoel@esri.com

November 8–10

MiG '17: Motion in Games, Barcelona, Spain,
Sponsored: ACM/SIG,
Contact: Nuria Pelechano,
Email: npelechano@cs.upc.edu

November 8–10

VRST '17: 23rd ACM Symposium on Virtual Reality Software and Technology
Gothenburg, Sweden,
Co-Sponsored: ACM/SIG,
Contact: Morten Fjeld,
Email: fjeld@chalmers.se

November 12–17

SC17: The International Conference for High Performance Computing, Networking, Storage and Analysis,
Denver, CO,
Contact: Bernd Mohr,
Email: b.mohr@fz-juelich.de

November 13–16

ICCAD '17: IEEE/ACM International Conference on Computer-Aided Design,
Irvine, CA,
Contact: Sri Parameswaran,
Email: sri.parameswaran@unsw.edu.au

pects. Here we outline how they are presented to non-computer scientists—see the detail at <http://www.teachcs.scot>. The vocabulary and concepts used are accessible to those who need to read them; the difficulty of this should not be underestimated, it is hard for an academic computer scientist to communicate with a teacher of early years computing.

Each of our three main aspects persists through the five defined levels of the curriculum, from ages 3–15; the text here is mostly aimed at teachers of the lower levels.

Understanding the world through computational thinking. The first aspect looks at the underlying theory in the academic discipline of Computing Science. Theoretical concepts of Computing Science include the characteristics of information processes, identifying information, classifying and seeing patterns.

This aspect is about understanding the nature and characteristics of *processes* and *information*. These can be taught through Unplugged activities (fun active learning tasks related to computing science topics but carried out without a computer) and with structured discussions with learners. There is a focus on recognizing computational thinking when it is applied in the real world such as in school rules, finding the shortest or fastest route between school and home, or the way objects are stored in collections.

Learners will be able to identify steps and patterns in a *process*, for example seeing repeated steps in a dance or lines of a song. In later stages, learners will begin to reason about properties of processes, for example considering whether tasks could be carried out at the same time, whether the output of a process is predictable, and how to compare the efficiency of two processes.

Learners will identify *information*, classify it, and see patterns. For example, learners might classify and group objects where there is a clear distinction between types or where objects might belong to more than one category.

Understanding and analyzing computing technology. This aspect aims to give learners insight into the hidden mechanisms of computers and the programs that run on them. It explores the different kinds of language, graphi-

Although solutions can be created in many ways, it is expected that all learners will experience creating solutions on computers.

cal and textual, used to represent processes and information. Some of these representations are used by people and others by machines, for example, a verbal description, a sequence of blocks in a visual programming language such as Scratch, or as a series of 1s and 0s in binary.

In this aspect, learners will learn how to ‘read’ program code (before writing it in the next aspect) and describe its behavior in terms of the *processes* they have learned about in the first aspect, processes that will be carried out by the underlying machinery when the program runs. For example, learners could read a section of code and predict what will happen when it runs or if lines of code change order. Learners will learn and explore different representations of *information* and how these are stored and manipulated in the computing system under study.

Designing, building, and testing computing solutions. The third aspect is about taking the concepts and understanding from the first two aspects and applying them. Learners will create solutions, perhaps by designing, building, and testing solutions on a computer or by writing a computational process down on paper. In doing so, they will learn about modeling process and information from the real world in programs, and what makes a good model to represent or solve a particular problem.

Learners will create representations of information. For example, learners

could make lists, tables, family trees, Venn diagrams, and data models to capture key information from the problems they are working on.

Learners will use their skills in language to create descriptions of processes that can be used by other people. For example, a computer program is a great way to describe a process.

Learners will understand how to read, write, and translate between different representations such as between English statements, planning representations, and actual computer code. For example, developing skills in writing code could be scaffolded by studying worked examples or by giving learners jumbled lines of code and asking them to put the lines into an appropriate order.

Although solutions can be created in many ways, it is expected that all learners will experience creating solutions on computers. This shows learners that computers implement exactly what they—the learners—have written, which is often not what they intended, as well as giving them practice in debugging.

Reflections

We have presented a curriculum that explicitly connects computational thinking with the more mechanical aspects of computing, with particular concentration on the explicit modeling of computational domains by computational mechanism. Not everyone needs to become a software engineer or computer scientist; the curriculum provides valuable learning at all levels, including the essential foundations for those who wish to study the subject further. While our curriculum is informed by previous educational computing research, we emphasize quite different learning outcomes via our three-point framework. 

Richard Connor (Richard.Connor@strath.ac.uk) is a Professor of Computer Science at the University of Strathclyde, Scotland.

Quintin Cutts (Quintin.Cutts@glasgow.ac.uk) is a Professor of Computer Science Education and Director of the Centre for Computing Science Education at the University of Glasgow, Scotland.

Judy Robertson (Judy.Robertson@ed.ac.uk) is a Professor of Digital Learning at the University of Edinburgh, Scotland.

Copyright held by authors.

Viewpoint

Pay What You Want as a Pricing Model for Open Access Publishing?

Analyzing the “Pay What You Want” business model for open access publishing.

THE OPEN ACCESS publishing movement has received strong support from scientists, lawmakers, and funding institutions. Many publishers are reacting to this demand by offering open access journals.⁷ However, there is an ongoing debate on how open access publishing models should be financed.^{2,4} Most open access journals that rely on the so-called “gold open access model”⁸—which makes the research output immediately available from the publisher—let authors of accepted papers pay article processing charges (APCs) of several hundred to several thousand U.S. dollars.¹ However, APCs are often criticized for potentially excluding researchers with limited funds.³

As one potential solution to this problem and also to gain a better understanding of the role of APCs in the scientific community, some publishers are starting to use *Pay What You Want* (PWYW) as a pricing model for gold open access publishing. PWYW is a pricing model where sellers delegate the full pricing power to buyers. So far, PWYW has mainly been applied in service industries (for example, restaurants, theaters), but also for the sale of digital products like software (such as <http://humblebundle.com>). More recently, several publishers of open access journals like Cogent OA (belonging to the Taylor & Francis Group), edp



Sciences, and Thieme Publishers have started to experiment with the PWYW model for APCs of open access journals. More specifically, Cogent OA empowers authors to decide how much they want to pay for their open access publication in 15 broad journals covering different domains of academic research. Likewise, *SICOT-J* (edp sciences), a multidisciplinary journal

covering the fields of surgery and engineering, and *The Surgery Journal* (Thieme Publishers), an open access journal for surgeons and trainee surgeons of all disciplines, have started to delegate pricing power to their contributing authors.

Although neoclassical economic theory predicts that buyers (that is, authors in OA publishing) pay nothing if

acm

Advertise with ACM!

Reach the innovators
and thought leaders
working at the
cutting edge
of computing
and information
technology through
ACM's magazines,
websites
and newsletters.



Request a media kit
with specifications
and pricing:

Ilia Rodriguez

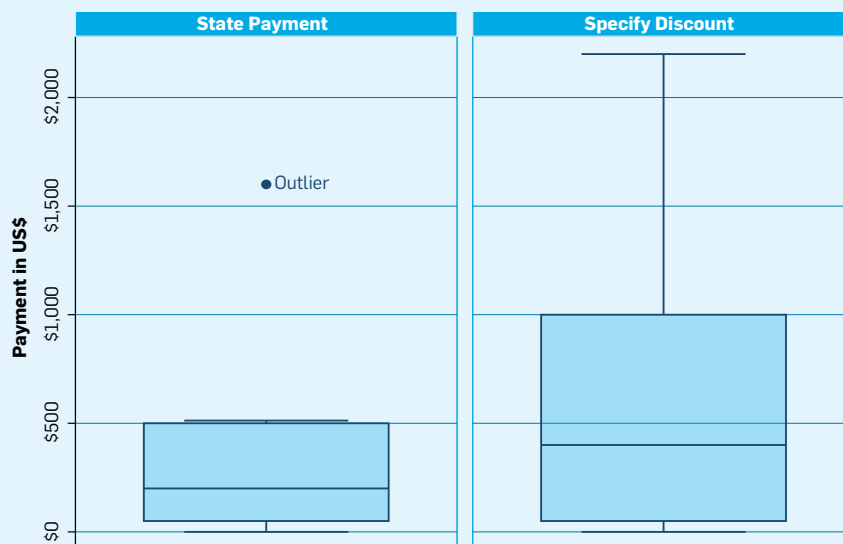
+1 212-626-0686

acmm mediasales@acm.org

acm

media

Distribution of PWYW payments according to payment form.



The box indicates the lower and upper quartile and the line within depicts the median of the payments. The whiskers extend to include all data points within the 1.5 interquartile range (IQR) of the upper/lower quartile and stop at the largest/smallest such value.

they are not forced to do so, empirical research on PWYW consistently finds that many buyers pay positive prices, often exceeding the marginal costs of the product (for example, Gneezy et al⁵). In the case of open access publishing, authors may be willing to voluntarily pay APCs for reasons of fairness and reciprocity—because they want to compensate the publisher for his costs or to reciprocate his generosity, as well as for strategic considerations, that is, because authors understand the journal will not be sustainable if the production costs are not covered.⁹ Furthermore, with their payment authors may signal to others (and potentially to themselves) the value they attach to their publication.⁶ At the same time, authors with limited funds are not ex-

cluded from publishing because they can adjust their PWYW payments to their available means.

For publishers, PWYW achieves endogenous price discrimination and higher market penetration because no author is excluded by APCs that he cannot afford. This can be especially important for the introduction of a new scientific journal. Furthermore, PWYW can initiate a debate about funding and affordability of APCs and might increase the acceptance of open access APCs in the scientific community. The obvious risk, however, is free riding of authors who do have the funds to pay for APCs but choose to pay nothing or only very little. A more specific concern in the context of open access is that some authors may be fundamentally convinced that research articles should be “free” to both readers and authors instead of just “open” and for this reason refuse a voluntary payment.

Some insights on the performance of PWYW for open access APCs at the individual level are provided by initial data from the peer-reviewed journal *The Surgery Journal* launched by Thieme Publishers in June of 2015 that exclusively uses PWYW pricing. At this peer-reviewed journal, authors are prompted *after acceptance* of their article to state the APCs.

After an article is accepted authors are directed randomly to one of two

**With their payment
authors may
signal to others
(and potentially
to themselves)
the value they attach
to their publication.**

possible forms (that is, experimental conditions) to make their PWYW payments. On both forms, all authors are informed about the recommended “regular price” of the publication fee of \$1,600. On one form they have to state their PWYW payment directly. On the other form they have to specify a discount from the regular price (a 100% discount was possible).¹⁰

So far, authors of 27 papers made their payment decisions with a mean payment of \$480 across conditions. Only four (15%) authors paid zero, and eight (30%) authors paid less than \$100. Two authors paid the recommended “regular price” for the publication fee of \$1,600 and one author even paid 50% more than the recommendation (\$2,200). Altogether, five (19%) authors paid \$1,000 or more. In the condition where authors had to specify their discount from the recommended publication fee, mean payments are substantially higher (\$616 vs. \$333; see the accompanying figure for the distribution of PWYW payments according to payment form). However, the standard deviation of the payment amounts is high (\$588.7), which might be attributed to heterogeneity in social preferences, available funds as well as cultural norms. The 27 publications came from a total of 12 different countries, which differ in terms of development. 11 publications came from developing countries, 16 from developed countries (thereof six from the U.S.). Payments from developed countries ($M = \$728.9$, $SD = \$654.4$) were significantly higher than those from developing countries ($M = \$118.2$, $SD = \$124.6$) at the 1%-level using a Mann-Whitney U test. These findings are directionally consistent with preliminary and aggregated findings provided by Cogent OA. They report that 55% of authors decided to pay something and some authors are paying even more than the recommended APC (<http://bit.ly/2wM5X8l>).

Although these constitute early empirical results and the sample size is limited, we can derive a few implications. The observed payments from our experiment together with the preliminary results from other experiments (that is, Cogent OA) indicate that PWYW may work in the context of open access publishing. The results suggest that a substantial fraction of authors do pay APCs voluntarily, in some cases

For publishers, using the PWYW pricing model can be a strategy to attract even more scientists to open access publishing.

even more than regularly asked. From discussions with the publisher, we learned the submissions exceeded their expectations compared to similar new open access journal introductions in this field. Also, the share of about 40% of publications from developing countries where authors might otherwise not be able to pay the regular APCs can be regarded as rather high. It must be noted that many publishers have introduced schemes that give authors a significant discount or the option to completely waive APCs if the authors cannot afford them. However, such arrangements often require the disclosure of the financial situation of the authors that can be uncomfortable for them. In contrast, the advantage of PWYW is that it allows authors to pay a reduced amount and thus allows them to save their face as part of the regular publication and payment process.

Clearly, PWYW is credible only if the journal takes the acceptance decision first and only then asks the authors how much they are willing to pay for the publication. One important aspect that will influence the sustainability of this pricing model for scientific publications is whether funding bodies such as the NSF permit authors to make positive payments voluntarily (within reasonable limits) despite the common requirement to use funds economically.

We believe the high motivation of scientific authors to achieve a wide dissemination of their work and an interest to support and help sustain the journal that publishes their research facilitate the applicability of PWYW in open access publishing. For publishers, using the PWYW pricing model can be a strategy to attract even more

scientists to open access publishing and to raise the visibility and the positive perception of their journals.

However, there are some points that we cannot address in our study. First, it is unclear to what extent the payments we have seen in *The Surgery Journal* depend on the journal's field (for example, medicine). It is likely that financial resources vary across different scientific disciplines and thus affect voluntarily paid APCs. Consequently, the PWYW model may be better suited for some publications and fields than for others. In this regard we would like to refer to the experiment at Cogent OA, where the comparison of payments across the 15 journals can allow for cautious conclusions in this direction. In general, we believe we need more experiments on PWYW as a pricing model for open access publishing and our contribution can be seen as a starting point in this direction. 

References

1. Beaudouin-Lafon, M. Open access to scientific publications. *Commun. ACM* 53, 2 (Feb. 2010), 32–34.
2. Boisvert, R.F. and Davidson, J.W. Positioning ACM for an open access future. *Commun. ACM* 56, 2 (Feb. 2013), 5.
3. Celec, P. Open access and those lacking funds. *Science* 303, 5663, (2004), 1467.
4. Cerf, V.G. Open access. *Commun. ACM* 56, 4 (Apr. 2013), 7.
5. Gneezy, A. et al. Shared social responsibility: A field experiment in pay-what-you-want pricing and charitable giving. *Science* 329, 5989 (2010), 325–327.
6. Gneezy, A. et al. Pay-what-you-want, identity, and self-signaling in markets. *Proceedings of the National Academy of Sciences of the United States of America* 109, 19 (2012), 7236–7240.
7. Malakoff, D. Scientific publishing. Opening the books on open access. *Science* 302, 5645 (2003), 550–554.
8. Mann, F. et al. Open access publishing in science. *Commun. ACM* 52, 3 (Mar. 2009), 135.
9. Schmidt, K.M. et al. Pay what you want as a marketing strategy in monopolistic and competitive markets. *Management Science* 61, 6 (2015), 1217–1236.
10. Schröder, M. et al. Pay-what-you-want or mark-off-your-own-price—A framing effect in customer-selected pricing. *Journal of Behavioral and Experimental Economics* 57, (2015), 200–204.

Martin Spann (spann@bwl.lmu.de) is a Professor at the Institute of Electronic Commerce and Digital Markets, LMU Munich, Germany.

Lucas Stich (stich@lmu.de) is an Assistant Professor at the Institute of Electronic Commerce and Digital Markets, LMU Munich, Germany.

Klaus M. Schmidt (klaus.schmidt@lmu.de) is a Professor at the Department of Economics, LMU Munich, Germany.

See the EC Workshop on OA publishing: <http://bit.ly/1MWDOtF>

The authors gratefully acknowledge data provision from Thieme Publishers and financial support by the German Science Foundation (grants SCHM 1196/5-1 and SP 702/2-1, and CRC TRR 190) and are also grateful for many helpful comments from the European Commission Workshop on “Alternative Open Access Publishing Models” in Brussels on Oct. 12, 2015 and the Open Access Days 2016 in Munich. There are no financial interests of the authors involved.

Copyright held by authors.

Viewpoint

Social Agents: Bridging Simulation and Engineering

Seeking better integration of two research communities.

THE USE OF the agent paradigm to understand and design complex systems occupies an important and growing role in different areas of social and natural sciences and technology. Application areas where the agent paradigm delivers appropriate solutions include online trading,¹⁶ disaster management,¹⁰ and policy making.¹¹ However, the two main agent approaches, Multi-Agent Systems (MAS) and Agent-Based Modeling (ABM) differ considerably in methodology, applications, and aims. MAS focus on solving specific complex problems using autonomous heterogeneous agents, while ABM is used to capture the dynamics of a (social or technical) system for analytical purposes. ABM is a form of computational modeling whereby a population of individual agents is given simple rules to govern their behavior such that global properties of the whole can be analyzed.⁹ The terminology of ABM tends to be used more often in the social sciences, whereas MAS is more used in engineering and technology. Although there is considerable overlap between the two approaches, historically the differences between ABM and MAS are often more salient than their similarities. For example, it is often remarked that a main difference between ABM and MAS is that ABM models are descriptive aiming at explanatory insight into the collective behavior of



agents at the macro level, whereas MAS are operational systems, acting and affecting its (physical) environment, with a focus on solving specific practical or engineering problems, and emphasizing agent architectures with sophisticated reasoning and decision processes. This has led to the development of two research communities proceeding on nearly independent tracks.

However, this division is not as black and white as it may seem. In fact, much ABM work goes beyond descriptive simulations of a situation, and, as input for decision making and policy

setting, indirectly affects the environment. And, the design of MAS is often geared to analytic insights and simulations toward the understanding of how configurations of agents behave in different circumstances. Currently, applications of MAS are broader than pure distribution problems, including interactive virtual characters, where the focus is on the cognitive, affective, and emotional characteristics of the system, and game-theoretic models, focusing on the design of incentive mechanisms that guarantee a given strategic behavior.

Social abilities are central both in ABM, where agents represent humans and their interactions, and in MAS, that enable game-theoretic analyses of decision strategies, or provide interactive virtual agents in varied situations. It is precisely in this area where the need for integration of ABM and MAS is undoubtedly the most necessary. In social simulation, the benefits of combining MAS and ABM have been advocated for many years, and are the focus of the long-lasting workshop series on Multi-Agent Based Simulation (MABS).² ABM has increasingly and successfully been used for social simulations,³ but it is in the MAS area that fundamental research on agent architectures implementing psychological traits and social concepts such as norms, commitments, emotions, identity, and social order, has been most prominent.^{4,5} Bridging these somewhat parallel tracks requires a new grounding for agent architectures.

Questioning Rationality

Traditionally, one of the most salient aspects shared by both ABM and MAS approaches is the premise of rationality. This is derived from the traditional definition of agents as autonomous, proactive, and interactive entities where each agent has bounded (incomplete) resources to solve a given problem; there is no global system control; data is decentralized; and computation is asynchronous.²¹ Agent rationality can be summarized as follows:

- Agents hold consistent beliefs;
- Agents have preferences, or priorities, on outcomes of actions; and
- Agents optimize actions based on those preferences and beliefs.

This view on rationality entails that agents are expected, and designed, to act rationally in the sense that they choose the best means available to achieve a given end, and maintain consistency between what is wanted and what is chosen.¹⁴ Even though multiple alternatives have been proposed, in both the ABM and MAS approaches, individual agents are still typically characterized as bounded rational, acting toward their own perceived interests. The main difference is that agent behaviors in ABM are used to capture the dynamics of a system for analytical purposes, grounded whenever possible on existing data about system outcomes,

Unfortunately, from a modeling perspective, real human behavior is neither simple nor rational.

whereas MAS focuses on solving specific problems using independent agents, through the formalization of the complex goal-oriented processes, such as the Beliefs-Desires and Intentions (BDI) model proposed by Bratman²⁰ or game-theoretic approaches.

The main advantages of such rationality assumptions are their parsimony and applicability to a very broad range of situations and environments, and their ability to generate falsifiable, and sometimes empirically confirmed, hypotheses about actions in these environments. This gives conventional rational choice approaches a combination of generality and predictive power not found in other approaches. In fact, rationality approaches are the basis of most theoretical models in the social sciences, including economics, political science, or social choice theories.

Unfortunately, from a modeling perspective, real human behavior is neither simple nor rational, but derives from a complex mix of mental, physical, emotional, and social aspects. Realistic applications must consider situations in which not all alternatives, consequences, and event probabilities can be foreseen. This renders rational choice approaches unable to accurately model and predict a wide range of human behaviors.

Toward Social Agents

Human sociability refers to the nature, quantity, and quality of interactions with others, including both pro-social, or cooperative, behaviors, and conflict, competitive, or dominating behaviors. Sociability is also the ability to influence others, by changing their behaviors, goals, and beliefs, the emotional reaction to others and to the environment, and how actions are affected by emo-

tions, and the ability to create, structure and 'rationalize' the environment to fit ones expectations and abilities (leading, for example, to the design of organizations, institutions, and norms).

Following an increasing number of researchers in both ABM and MAS that in recent years have come to similar conclusions,^{7,13,18,19} we claim that new models of preference and belief formation are needed that show how behavior derives from identities, emotions, motivation, values, and practices.⁶

The endeavor required to construct such agent models that are socially realistic requires the effort and the capabilities of both the MAS and ABM communities, bringing together formalization and computational efficiency, and planning techniques as in MAS, with the ABM expertise on empirical validation and on adapting and integrating social sciences theories into a unified set of assumptions,¹ furthering the fundamental understanding of social deliberation processes, and developing techniques to make these accessible for simulation platforms. This Viewpoint is therefore an appeal to join the strengths of both communities toward sociality-based agents.

Without claiming a readily available solution, we propose the concept of sociality as the leading principle of agency, as an alternative for rationality. Following the aforementioned description of rational behavior, the main characteristics of sociality-based reasoning are:

- Ability to hold and deal with inconsistent beliefs for the sake of coherence with identity and cultural background. That is, beliefs originate from other sources than observation, including ideology or culture.
- Ability to fulfill several roles, and pursue seemingly incompatible goals concurrently, for example, simultaneously aiming for comfort and environmental friendliness, or for riches and philanthropy.
- Preferences are not only a cause for action but also a result of action. Moreover, preferences change significantly over time and their ordering is influenced by the different roles being fulfilled simultaneously, which requires the need to deal with misalignment and incompatible orderings.
- Action decisions are not only geared

to the optimization of own wealth, but often motivated by altruism, fairness, justice, or by an attempt to prevent regret at a later stage.

- Understand when there is no need to further maximize utility beyond some reasonably achievable threshold.

- Understand how identity, culture, and values influence action, and use this knowledge to decide about reputation and trust about who and how to interact.

The first step toward sociality-based agents is a thorough understanding of these principles, and open discussion across disciplines on the grounds and requirements for sociality from different perspectives. This discussion will be fundamental to the development of formal models and agent architectures that make sociality-based behavior possible and verifiable.

Moreover, it is necessary to identify and formalize which mechanisms, other than imitation, can describe how agents can adapt to pressures in the environment to behave in a socially acceptable, resource-sustainable fashion. Resulting models support the understanding or predicting human behavior, including rich models of emotions, identities, culture, values, norms, and many other socio-cognitive characteristics. Such models of social reality are also needed to study the complex influences on behavior of different socio-cognitive characteristics and their relationships. The integration of psychological models of motivation and cognition, sociological theories of value and identity formation, and philosophical theories of coherence and higher-order rationality, together with different formal methods, quickly yields intractable models. However, it is important to identify what is the model being developed for. In fact, richer models are not always the most appropriate ones.

Once these characteristics are well understood, then simplified models can be developed to suit different needs. That is, implementing sociality-based agents will require other techniques than those currently used in either MAS or ABM,⁸ including the use of simpler, context-specific decision rules, mimicking how people themselves are able to deal with complex decision making, for example, using

Sociality-based agents are also fundamental to the new generations of intelligent devices.

social practices as a kind of shortcuts for deliberation.^{15,17} Where it concerns utility, satisficing can be more suitable approach than maximizing.¹² This also allows us to integrate agents of varied richness levels, for example, using rich cognitive models to zoom-in the behavior of salient agents in a simulation, whereas other agents just follow simple rules. This approach can counter the obvious criticism that sociality-based agents will become too complex for use in computational simulations.

Sociality-based agents are also fundamental to the new generations of intelligent devices, and interactive characters in smart environments. These artifacts not only must build (partial) social models about the humans they interact with, but also need to take social roles in a mixed human/digital reality. An interesting challenge would be to use the same technologies in real time mixed human/artificial interactions, and criticisms could also be on the feasibility to use these architectures (or controlled reductions/simplifications) in real time or near real time.

Moving Forward

The intent of this Viewpoint has been to appeal for a collaborative research effort toward fundamental formal theories and models that increase our understanding of the principles behind human deliberation (such as the ones listed discussed here), before deciding on which modeling techniques we need to implement them. Even though, several approaches to model social aspects in agent behavior are available, there is not sufficient consensus on which characteristics are needed for what, nor on how to specify and integrate them. We have identified an initial set of characteristics for sociability, proposed a re-

search path linking theory, model, and implementation, and suggested possible theories and techniques to develop sociality-based agents. These incorporate expertise from both ABM and MAS and require integration of both areas in order to succeed. We welcome the discussion of these issues toward a novel area of research on social agents, which take sociability as the basis for agent deliberation and enable interaction. **□**

References

1. Chai, S. *Choosing an Identity: A General Model of Preference and Belief Formation*. University of Michigan Press, 2001.
2. Conte, R., Gilbert, N., and Sichman, J. MAS and social simulation: A suitable commitment. In J. Sichman, R. Conte, and N. Gilbert, Eds, *Multi-Agent Systems and Agent-Based Simulation*, volume 1534 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 1998, 1–9.
3. Davidsson, P. Agent based social simulation: A computer science view. *Journal of Artificial Societies and Social Simulation* 5, 1 (2002).
4. Dias, J., S. Mascarenhas, and A. Paiva. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Proceedings of the International Workshop on Standards for Emotion Modeling*, 2011.
5. Dignum, F., Dignum, V., and Jonker, C.M. Towards agents for policy making. In *MABS IX*, Springer, 2009, 141–153.
6. Dignum, F. et al. A conceptual architecture for social deliberation in multi-agent organizations. *Multiaagent and Grid Systems* 11, 3 (2015), 147–166.
7. Dignum, F., Prada, R., and Hofstede, G.J. From autistic to social agents. In *AAMAS 2014*, May 2014.
8. Dignum, V. Mind as a service: Building socially intelligent agents. In V. Dignum, P. Noriega, M. Sensoy, and J. Sichman, Eds, *COIN XI: Revised Selected Papers*, Springer International Publishing, 2016, 19–133.
9. Epstein, J.M. and Axtell, R. *Growing Artificial Societies: Social Science from the Bottom Up*. The Brookings Institution, Washington, D.C., 1996.
10. Fiedrich, F. and Burghardt, P. Agent-based systems for disaster management. *Commun. ACM* 50, 3 (Mar. 2007), 41–42.
11. Ghorbani, A. Enhancing abm into an inevitable tool for policy analysis. *Policy and Complex Systems* 1, 1 (2014).
12. Gigerenzer, G. Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in Cognitive Science* 2, 3 (2010), 528–554.
13. Kaminka, G. Curing robot autism: A challenge. In *Proceedings of the AAMAS 2013*, May 2013, 801–804.
14. Lindenberg, S. Social rationality versus rational egoism. In *Handbook of Sociological Theory*, Springer, 2001, 635–668.
15. Reckwitz, A. Toward a theory of social practices. *European Journal of Social Theory*, 5, 2 (2002), 243–263.
16. Rogers, A. et al. The effects of proxy bidding and minimum bid increments within ebay auctions. *ACM Trans. Web* 1, 2 (Aug. 2007).
17. Shove, E., Pantzar, M., and Watson, M. *The Dynamics of Social Practice*. Sage, 2012.
18. Silverman, B. et al. Rich socio-cognitive agents for immersive training environments: The case of nonkin village. *Journal of Autonomous Agents and Multi-Agent Systems* 24, 2 (Mar. 2012): 312–343.
19. Vercouter, L. et al. An experience on reputation models interoperability based on a functional ontology. In *Proceedings of IJ-CAI'07*, Morgan Kaufmann Publishers Inc., San Francisco, CA, 2007, 617–622.
20. Wooldridge, M. *Reasoning about Rational Agents*. MIT Press, 2000.
21. Wooldridge, M. *An Introduction to Multiagent Systems*. Wiley, New York, 2009.

Virginia Dignum (m.v.dignum@tudelft.nl) is an associate professor with the Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands.

Copyright held by author.

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 23–28, 2018** 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?

The application process is open between **November 6, 2017** and **February 9, 2018**.

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

Successful applicants will be notified by **mid April 2018**.

[More information available on Heidelberg social media](#)



A discussion with Edward Steel, Yanik Berube, Jonas Bonér, Ken Britton, and Terry Coatta

Hootsuite: In Pursuit of Reactive Systems

BASED IN VANCOUVER, Canada, Hootsuite is the most widely used SaaS (software as a service) platform for managing social media. Since its humble beginnings in 2008, Hootsuite has grown into a billion-dollar company with more than 15 million users around the globe.

As Hootsuite evolved over the years, so did the technology stack. A key change was moving from LAMP (Linux, Apache, MySQL, PHP) to microservices. A shift to microservices didn't come without its challenges, however. In this roundtable chat, we discuss how Scala and Lightbend (which offers a reactive application development platform) were an essential part of a successful transition. The exchange includes **Jonas Bonér**, CTO of Lightbend; **Terry Coatta**, CTO of Marine Learning Systems; **Edward Steel**, senior Scala developer at Hootsuite; **Yanik Berube**, lead software developer at Hootsuite; and **Ken Britton**, senior director of software development at Hootsuite.

TERRY COATTA: I'm curious about the original set of problems Hootsuite was looking to address in the switch to microservices. Can you provide some detail?

EDWARD STEEL: Mostly, it had to do with our ability to send out notifications to user mobile devices whenever something relevant happened on Twitter. By the time we started having some concerns about how we were handling that, we were already servicing several hundreds of thousands of users, each with individual subscriptions tailored to their own specific interests. What was needed was something that could stay connected to Twitter's streaming endpoints.

COATTA: I gather that at about the same time you were making this move, you also took steps to move from PHP to Scala. What drove that?

STEEL: Initially, it had a lot to do with learning about all the success some other organizations had experienced with Scala. This was after Twitter had decided to go with Scala, for example, and that obviously lent a lot of legitimacy to it. Also, the first team here to work with Scala came from quite a varied background. We had some people who were lobbying for a more strongly typed functional language—something on the order of Haskell—and then there were some others with Clojure and Java experience. In taking all that into account, I guess Scala just seemed to check most of the boxes.

JONAS BONÉR: What would you say was the principal benefit you saw with Scala? Was it the functional nature of the language itself? Or did it have more to do with the libraries available within that ecosystem?

STEEL: The language itself was the biggest part of it. The main advocate here for Scala was working on a BlackBerry client at the time, so he had a lot of JVM (Java virtual machine) knowledge and yet also had become frustrated with Java itself. I guess he was just looking for a better way. Another aspect of our thinking had to do with building a distributed system that



could take advantage of Akka as an available library. That was a big part of the decision as well.

In fact, I think we were able to use some actors right from the start. That was with a very early version of Akka, but it still offered a lot of compelling features we found useful.

BONÉR: Were you already thinking in terms of microservices back then, even before that took off as a buzzword? Or were you more drawn by reactive principles having to do with things like a share-nothing architecture, strong isolation, and loose coupling?

STEEL: Microservices were always in the back of our minds. We already had some batch processes written in PHP that were starting to run jobs at that point. That sort of worked, but it was far from an ideal way of doing things. So I think we had already started to develop an appetite for a system on which we could build a few services, with the

thought being that perhaps we could then move toward a service-oriented architecture. The idea of microservices wasn't something that came up until a little later, and that was probably influenced by some of the buzz around the industry at the time.

COATTA: You have already mentioned Akka a couple of times, so can you speak to how that fits in here?

STEEL: At that time, at least, the main appeal of the Akka system for the JVM was that it provided people beyond the Erlang community with access to the actor model. The thing about actors is that they are both message based and highly resilient—which is to say that even when they crash, they can typically be brought back in a useful way. This probably explains why Erlang has so often been used to develop resilient telecommunications systems. Whenever you are talking about distributed systems or some system where you ex-

pect to fire a lot of messages around, you can expect the actor model to really shine.

YANIK BERUBE: Just in terms of where this fits into our current infrastructure, we should note that all our microservices are powered by Akka. Internally, we have a server-type library that handles requests and responses via Akka, and we also have at least one, if not more, back-end services that use sets of actors to accomplish work at certain intervals of time.

COATTA: One of the things that comes to mind when I think of the Erlang actor system, beyond the independence you have already mentioned, is that it's quite fine-grained. So I wonder, given your focus on notifications to user mobile devices, whether you might actually require an actor per user just to deal with that?

STEEL: In our case, no. But symptoms of that definitely showed up as we



JONAS BONÉR

Something I find very interesting is that Akka and Erlang appear to be the only platforms or libraries that put an emphasis on embracing and managing failure—which is to say they are basically designed for resilience.



were first building our system. When we were starting out, we learned more about how we should be building the system, as well as about how actors really ought to work. At first, we definitely fell into the trap of putting too much logic into single actors—for example, by putting recovery logic into each actor instead of relying on the supervision hierarchy, which would have allowed us to code less defensively. It turns out it's best just to embrace the “let it crash” philosophy, since that actually offers a lot of robustness.

We also learned the model really shines whenever you separate concerns into single-purpose actors. Besides helping to clarify the design, it opens up a lot of opportunities in terms of scalability and configuration flexibility at the point of deployment.

BONÉR: This applies to microservices as well. That is, there are plenty of opinions about what even defines a microservice. What does that term mean to you? And how does that map conceptually to how you view actors?

BERUBE: Internally, we are still trying to define what a microservice is and what the scale of that ought to be. Today, most of our services focus on accomplishing just one set of highly related tasks. Our goal is to have each of these services own some part of our domain model. Data services, for example, would each control their own data, and nothing else would have access to that. To get to that data, you would have to go through the data service itself.

Then we would also have functional services, which are the services that es-

entially glue business logic onto the data part of the logic. But in terms of the size of these things, I'd say we're still trying to figure that out, and we haven't come up with any hard and fast rules so far.

BONÉR: I'm guessing each of your microservices owns its own data store. If so—with these things being stateful—how are you then able to ensure resilience across outages?

STEEL: Each of these services absolutely owns its own data. When it comes to replacing parts of the monolithic system, it generally comes down to dealing with a table or a couple of related tables from the LAMP MySQL database. Generally speaking, the space is pretty minimal in terms of the services that need to be accounted for. It's basically just a matter of retrieving and creating data.

BERUBE: I would say we make fairly heterogeneous use of various technologies for data storage. And, yes, we do come from a LAMP stack, so there is still a heavy reliance on MySQL, but we also make use of MongoDB and other data-storage technologies.

The services typically each encapsulate some area of the data. In theory, at least, they are each supposed to own their data and rely on data storage dedicated only to them. So, we have recently started looking into storing a bit more data within the services themselves for reasons of efficiency and performance.

COATTA: To be clear, then, there's some separate data-access layer from which the services are able to pull in whatever information they need to manipulate?



TERRY COATTA

One of the things that comes to mind when I think of the Erlang actor system, beyond its independence, is that it's quite fine-grained.



STEEL: Yes, but you won't see more than one service accessing the same data store.

BONÉR: How do you manage this in terms of rolling out updates? Do you have some mechanism for deploying updates, as well as for taking them down and rolling back? Also, are these services isolated? If so, how did you manage to accomplish that? And, if not, what are you doing to minimize downtime?

BERUBE: Right now, every service uses a broker/worker infrastructure. No service can access another service without going through a pool of brokers that then will redistribute requests to multiple workers. This gives us the ability to scale by putting more workers behind the brokers. Then, when it comes to deployment, we can do rolling deploys across the target servers for those workers. In this way we are able to deploy the newest version of a service gradually without affecting the user experience or the experience of any other services that need to make use of that service as it's being redeployed.

STEEL: Another thing we have recently pulled over from the LAMP side that has proved to be useful for frequent rollouts is feature flagging. That's obviously something that was a lot easier back when we just were working with a bunch of Web servers, since we had a central place for handling it. But recently we migrated those same capabilities to HashiCorp's Consul to give ourselves a distributed, strongly consistent store, and that now lets us de-

ploy code on the Scala side with things switched off.

BONÉR: Looking back to when you were doing this along with everything else required to maintain a monolith, what do you see as the biggest benefits of having made the move to microservices?

STEEL: In terms of what it takes to scale a team, I think it has proved to be much easier to have well-defined boundaries within the system, since that means you can work with people who have only a general idea of how the overall product works but augment that with a strong, in-depth knowledge of the specific services they're personally responsible for.

I also think the microservices approach gives us a little more control operationally. It becomes much easier to scale and replace specific parts of the system as those needs arise.

BERUBE: One clear example of this is the data service my team has been working on. It's a very high-volume service in terms of the amount of data we store, and we knew it would be challenging to scale that, given the storage technology we're using. The ability to isolate all that data behind a data service makes it a lot easier for us to implement the necessary changes. Basically, this just gives us a lot more control over what it takes to change the persistent technology we're using in the background. So I certainly see this as a big win.

Some of the benefits of moving over to the reactive microservice model



EDWARD STEEL

Because of our ability to change the characteristics of actors by how we configure them, we have been able to adapt this core framework to all types of payloads and traffic profiles for the various services.



supported by the Lightbend stack surfaced almost immediately as the Hootsuite engineering team started discovering opportunities for scaling down on the underlying physical and virtual infrastructure they had run previously on their LAMP stack (where there had been a process for each request). Accordingly, it soon became apparent that operations under the reactive microservices model were going to put much less strain on their resources.

In fact, if anything, the engineers at Hootsuite quickly learned that by continuing to employ some of the practices that had made sense with a LAMP stack, they would actually be denying themselves many of the benefits available by relying to a greater degree on the Lightbend stack. For example, they found there was a real advantage to making greater use of the model classes supported by the Lightbend stack, since those classes come equipped with data-layer knowledge that can prove to be quite useful in a dynamic web-oriented system.

Similarly, they learned that by using individual actors to run substantial portions of their system instead of decomposing those components into groups of actors, they had been unwittingly depriving themselves of some of the features Akka offers for tuning parts of the system separately, parallelizing them, and then distributing work efficiently among a number of different actors capable of sharing the load.

And then there were also a few other things they learned ...

COATTA: So far, we have talked only about general issues. Now I would like to hear about some of your more specific engineering challenges.

BONÉR: One thing I would like to know is whether you're mostly doing reactive scaling, predictive scaling, or some combination of the two to optimize for your hardware.

BERUBE: For now at least, our loads don't really change a lot. Or perhaps what I should say is that they change throughout the day, but predictably so from one day to the next. And the way we have designed our services to run behind brokers means we are able just to spin off more workers as necessary. In combination with some great tooling from AWS (Amazon Web Services), we are able to adapt quickly to changing workloads.

STEEL: One thing we did decide to do was to build a framework using ZeroMQ to enable process communication between our various PHP systems. But then we saw later that we could have just as easily pulled all that into Akka.

COATTA: And I'm assuming, with Akka, it would have been easier for you to achieve your goal of adhering to the actor paradigm, while also taking advantage of better recovery mechanisms and finer-grained control.

STEEL: Yes, but I think the key is that because of our ability to change the characteristics of actors by how we configure them, we have been able to adapt this core framework to all types of payloads and traffic profiles for the various services. We can say, "This service is using a blocking database," at



YANIK BERUBE

One of the fantastic lessons to come out of all this is it has allowed me to start thinking about how the system actually works in terms of handling failures and dealing with the external agents we communicate with via messages.



which point a large thread pool will be supplied. If the service happens to be handling two different kinds of jobs, we can separate them into different execution tracks.

More recently, we have also had a fair amount of success using circuit breakers in situations where we have had a number of progressive timeouts as a consequence of some third party getting involved. But now we can just cut the connection and carry on. Much of this comes for free just because of all the tools Akka provides. We have learned that we can take much better advantage of those tools by keeping our designs as simple as possible.

BONÉR: Something I also find very interesting is that Akka and Erlang appear to be the only platforms or libraries that put an emphasis on embracing and managing failure—which is to say they are basically designed for resilience. The best way to get the fullest benefit of that is if they are part of your application from Day One. That way, you can fully embrace failure right at the core of your architecture.

But, with that being said, how did this newfound embrace of failure work out in practice for those of your developers who had come from other environments with very different mindsets? Was this something they were able to accept and start feeling natural about in fairly short order?

BERUBE: For some, it actually required a pretty substantial mindset shift. But I think Akka—or at least the actor model—makes it easier to understand the benefits of that, since you have to

be fairly explicit about how you handle failures as a supervisor. That is, as an actor that spins off other actors, you must have rules in place as to what ought to happen should one of your child actors end up failing. But, yes, people had to be educated about that. And even then, it still took a bit of getting used to.

Now, as new developers come in, we see them resort to the more traditional patterns of handling exceptions. But once you get some exposure to how much saner it is just to leave that to the supervisor hierarchy, there's generally no turning back. Your code just becomes a lot simpler that way, meaning you can turn your focus instead to figuring out what each actor ought to be responsible for.

COATTA: Talking about actor frameworks in the abstract is one thing, but what does this look like once the rubber actually hits the road? For example, how do you deal with failure cases?

BERUBE: You can just let the actor fail, which means it will essentially die, with a notification of that then being sent off to the supervisor. Then the supervisor can decide, based on the severity or the nature of the failure, how to deal with the situation—whether that means spinning off a new actor or simply ignoring the failure. If it seems the problem is something the system actually ought to be able to handle, it will just use a new actor essentially to send the same message again.

But the point is that the actor model allows you to focus all the logic related to the handling of specific failure cases in one place. Because actor systems are



KEN BRITTON

It has become apparent how critical frameworks and standards are for development teams when using microservices.



hierarchies, one possibility is that you will end up deciding the problem isn't really the original actor's responsibility but instead should be handled by that actor's supervisor. That's because the logic behind the creation of these hierarchies determines not only where the processing is to be done, but also where the failures are to be handled—which is not only a natural way to organize code, but also an approach that very clearly separates concerns.

BONÉR: I think that really hits the nail on the head. It comes down to distinguishing between what we call errors—which are things that the user is responsible for dealing with—and failures. Validation errors then naturally go back to the user while less severe errors go to the component that created the service. This creates a model that is easier to reason about, rather than littering your code with try/catch statements wherever failures might happen—since failures can, and will, happen anywhere in a distributed system.

BERUBE: One of the fantastic lessons that has come out of all this is it has allowed me to start thinking about how the system actually works in terms of handling failures and dealing with the external agents we communicate with via messages. Basically, I started to think about how we should handle the communication between services around the way we handle failures. So now that's something we always think about.

The reality is that any time we talk to external services, we should expect

some failures. They are just going to happen. This means we should not be banking on some external service responding in a short amount of time. We want to explicitly set timeouts. Then if we see that the service is starting to fail very quickly or with some high frequency, we will know it's time to trip a circuit breaker to ease the pressure on that service and not have those failures echo across all services. I have to say that came as a bonus benefit I certainly was not expecting when we first started working with Akka.

Providing for greater efficiency in the utilization of system resources by resorting to a distributed microservices architecture is one thing. But to what degree is that liable to end up shifting additional burdens to your programmers? After all, coding for asynchronous distributed systems has long been considered ground that only the most highly trained Jedi should dare to tread.

What can be done to ease the transition to a reactive microservices environment for programmers more accustomed to working within the confines of synchronous environments? Won't all the assumptions they typically make regarding the state of resources be regularly violated? And how to get a large team of coders up the concurrency learning curve in reasonably short order?

Here's what the Hootsuite team learned ...

COATTA: Let's talk a little about the impact the move to microservices has had on your developers. In particular, I would think this means you are throwing a lot more asynchrony at them than most developers are accustomed to. I imagine they probably also have a lot more data consistency issues to worry about now.

BERUBE: Although the asynchrony problem hasn't been fully addressed, Scala Futures (data structures used to retrieve the results of concurrent operations) actually make it really easy to work with asynchronous computation. I mean, it still takes some time to adjust to the fact that anything and everything can and will fail. But, with Scala Futures, it's actually quite easy for relatively uninitiated programmers to learn how to express themselves in an asynchronous world.

STEEL: If you are coming at this from the perspective of thread-based concurrency, it's going to seem much scarier for a lot of use cases than if you're coming at it from a Futures point of view. Also, when you're working directly inside actors, even though messages are flying around asynchronously and the system is doing a thousand things at the same time, you are insulated from what it takes to synchronize any state modifications, since an actor will process only one message at a time.

KEN BRITTON: I have noticed when developers first start writing Scala, they end up with these deeply nested, control-flow-style programs. You see a lot of that in imperative languages, but there is no penalty for it. In a strongly typed functional language, however, it is much more difficult to line up your types through a complex hierarchy. Developers learn quickly that they are better served by writing small function blocks and then composing programs out of those.

Akka takes this one step further by encouraging you to break up your logic with messages. I've observed a common evolution pattern where developers will start off with these very bulky, complex actors, only to discover later that they could have instead piped a Future to their own actor or any other actor. In fact, I've witnessed a number of aha moments where developers hit upon the realization that these tools actually encourage them to build smaller composable units of software.

BONÉR: That matches my experience as well. Actors are very object oriented, and they encapsulate state and behavior—all of which I think of as mapping well to a traditional approach. Futures, on the other hand, lend themselves to functional thinking—with all these small, stateless, one-off things you can compose easily. But have you found you can actually make these things derived from two very different universes work well together? Do you blend them or keep them separate?

BERUBE: We have used them together in parallel, and I think they work well that way. Ken mentioned this idea of generating Futures and then piping them either to yourself as an actor or to some other actors. I think that pattern works quite well. It's both simple and elegant.

STEEL: I have to admit I stumbled over that mental shift a bit early on. But, yes, I'd say we have been able to blend actors and Futures successfully for the most part.

BONÉR: Do you feel that certain types of problems lend themselves better to one or the other?

STEEL: In our simpler services, the routing of a request to the code is all actor based, and then the actual business logic is generally written as calls to other things that produce Futures. I suppose that when you're thinking about infrastructure and piping things around, it's very natural to think of that in terms of actors. Business logic, on the other hand, perhaps maps a little more readily to the functional point of view.

BRITTON: We are also finding that a rich object-oriented model is helpful in our messaging. For example, we have started defining richer success and failure messages containing enough detail to let an actor know exactly how to respond. So, now our message hierarchy has expanded to encapsulate a lot of information, which we think nicely aligns object-oriented concepts with the actor model.

COATTA: One thing that occurred to me as you talked about your environment is that it seems you have moved not only from a monolithic architecture, but also, in some sense, from a monolithic technology to a much wider array of technologies. So I wonder if you now find it more difficult

to operate in that environment, as well as to train people to work in it. Whereas before maybe it was sufficient just to find some new hires that were proficient in PHP, now you have ZeroMQ and actors and Futures and any number of other things for them to wrap their heads around. Without question, your environment has become more complex. But is it now in some respects also actually an easier place in which to operate?

BERUBE: I think the act of dividing the logic into a lot of different self-contained services has made it easier at some level to reason about how the system works. But we are not finished yet. There is still plenty of work to do and lots of challenging areas to continue reasoning about.

And, yes, of course, the environment has become a bit more complex. I have to agree with that. But the benefits outweigh the drawbacks of rolling in all this technology, since we now have more layers of abstraction to take advantage of. We have teams that are generally aware of the big picture but are mostly focused on just a few microservices they understand really well. That's an approach that will have huge benefits for our operations as we scale them moving forward.

BRITTON: It has become apparent how critical frameworks and standards are for development teams when using microservices. People often mistake the flexibility microservices provide with a requirement to use different technologies for each service. Like all development teams, we still need to keep the number of technologies we use to a minimum so we can easily train new people, maintain our code, support moves between teams, and the like.

We have also seen a trend toward smaller services. Our initial microservices were actually more like loosely coupled macroservices. Over time, though, we have pushed more of the deployment, runtime, and so forth into shared libraries, tooling, and the like. This ensures the services are focused on logic rather than plumbing, while also sticking to team standards. 

We all wear many hats, but make sure you have one that fits well.

BY KATE MATSUDAIRA

Breadth and Depth

I am often asked for career advice. One of the things software engineers always want to know is if they should learn some new tool or language. In fact, I cannot think of a performance review I have read for a software developer that didn't include something about growing their skills around a particular technology.

That is the nature of our work: it is constantly changing. You have to keep learning, or you will become obsolete.

But for your career, is it better to go wide and learn a lot of different things, or is it better to go deep and learn a few things really well?

Making the case for going deep. Recently, I was talking to another engineering leader about hiring and staffing. I asked which technologies he wanted

people to know, and he responded that it didn't matter—a good software engineer can work on anything.

This has been the thinking at many large software companies in the past, and there are definitely merits to it—especially when you are hiring inexperienced candidates straight out of school. As I have worked longer in the industry, however, I have started changing my thinking.

I would argue that it is important to go deep in at least one area, and it is almost always better to hire people who have a solid depth of experience in the tools and technology they are using.

Why do you need to have deep knowledge?

Really good software engineers for a particular language or technology will exhibit qualities such as these:

- ▶ **They are productive.** They produce an amount of work that is above average, and they are able to get things done. This means they know how to use the tools of their trade well and aren't slowed down by not understanding something. They use their brainpower for harder problems, not learning how to do the basics.

- ▶ **They make smart trade-offs.** They are able to understand the risks of their decisions. They have failed before and so can avoid mistakes. If there is a library or prebuilt code somewhere, they are probably aware of it—they may even have contributed to it or used it in past projects.

- ▶ **They help others.** Teammates go to them with questions because they have done this before and know how to do it right. This expertise makes them natural leaders or mentors.

- ▶ **They hit their deadlines.** Their estimates are almost always accurate. They know how long something should take them because they have experience working on similar projects.

- ▶ **They are still growing personally and professionally.** Since they are comfortable with the technology stack, they can use their time on learning skills in other areas such as lead-

ership, communication, or even new technologies.

► **They have a deep understanding.**

When engineers work with a particular technology for a long time, they learn the nooks and crannies of how it works—the good parts and the ugly parts. This can force them to think about the very constructs of how the technology was built. For example, discovering a bug in a foundational library teaches them to write better software. This depth also allows them to pick up other technologies faster.

When you hire engineers who have experience in a particular technology and can operate at this level, they are able to get up to speed quickly. They will have to learn the way your particular systems and software work, but they will not have to learn the tools and technology they were written in.

While there is definitely value in gaining that deep experience—and it should be part of your career plan—it is also important to have breadth.

Making the case for going wide.

When I look at my own career path, I can say it is my breadth that has helped me the most. As a developer, the fact that I understood operations, lower-level operating systems, and compilers helped me write better code. And as a manager, having experience with other disciplines helps me work with those people better.

For example, we have all written software tests, but when you work with really great testers and learn the way they think and approach problems, this helps you not only work better alongside them, but also write more robust code.

As a technology executive, you can't not just understand the technology—you also have to understand the business. You have to focus on customers, think about product, and be able to understand the financial implications of your decisions. Often, you are managing people doing a job you have never done yourself—but you still need to be able to measure their impact and mentor them to success.



In these cases, it isn't enough to know just one technology well; in fact, if that is the case, you won't be successful. You have to have breadth and be able to think through other aspects of the problems you are solving, and you must have a broader understanding to work with people in different roles.

So Should You Go Wide or Deep?

I once had the privilege of having a mentoring session with a VP at one of the great software companies. I asked him about his background and what he felt got him to his position. He told me that he started his career as a software engineer and soon realized that being the best software engineer would be hard; software testing, on the other hand, was much less competitive, and a lot of people didn't have the passion for it that he did.

He changed his career path to pursue software testing and went on to write several books. He became one of the best in the world at testing and was asked to speak at conferences all over the world. The VP title and responsibility came later but was a natural progression.

His advice to me: *pick something*

you can be the best in the world at. Specialize in it. Pursue it above everything else. Success will come along with it.

In many ways this is true. When you are truly exceptional at something, you build career capital, and you can trade that capital for bigger paychecks, more flexibility, or even fancy job titles.¹

When people ask me the question of where they should focus their time—should I keep learning one technology or spend time learning a new one?—I ask them that very question: What is the one thing you could be the best in the world at?

The answer might be going deep or going wide—the important thing is to spend your time on building the skills that will move you to where to you want to go. C

Reference

1. Shin, L. 7 steps to developing career capital—and achieving success. *Forbes*; <http://bit.ly/2wgoieh>.

Kate Matsudaira (katemats.com) is the founder of her own company, Popforms. Previously she worked at Microsoft and Amazon as well as startups like Decide, Moz, and Delve Networks.

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

Essence can keep software development for the IoT from becoming unwieldy.

BY IVAR JACOBSON, IAN SPENCE, AND PAN-WEI NG

Is There a Single Method for the Internet of Things?

The Industrial Internet Consortium predicts the Internet of Things (IoT) will become the third technological revolution after the Industrial Revolution and the Internet Revolution. Its impact across all industries and businesses can hardly be imagined. Existing software (business, telecom, aerospace, defense, among others) will likely be modified or redesigned, and a huge amount of new software, solving new problems, will be developed. As a consequence, the software industry should welcome new and better methods.

This article makes the case that to be a major player in this space you will need a multitude of methods, not just a single one. Existing popular approaches

such as Scrum and SAFe (Scaled Agile Framework) may be part of the future, but you will also need many new methods and practices—some of which are not even known today. Extending a single method to incorporate all that is required would result in something that is way too big and unwieldy. Instead, the new Object Management Group (OMG) standard Essence can be used to describe modular practices that can be composed together to form a multitude of methods, not only to provide for all of today's needs, but also to be prepared for whatever the future may bring.

The software world is continuously innovating and opening up new areas of opportunity and challenge. A decade ago developers were busy with trends such as service-oriented architecture and product-line architecture—still very much around, but now a commoditized part of a larger system-of-systems landscape, and also extended to cloud computing with big data and mobile applications. New software development approaches have accompanied these new trends, most of them being agile in different flavors and size: Scrum, Kanban, DAD (Disciplined Agile Delivery), SAFe, LeSS (Large-scale Scrum), and SPS (Scaled Professional Scrum) being among these approaches.

These trends have impacted the software industry in many different ways—producing more pervasive and powerful technology-based products, for example. None of them, however, has had a truly transformational or radically disruptive impact.

The Industrial Revolution in the 19th century moved us from essentially building things as a craft to manufacturing. The Internet Revolution at the end of the 20th century was another such transformation of the world or, as Bill Gates said in 1999, “A fundamental new rule for business is that the Internet changes everything.” The Internet has driven the need for faster turnaround time with less precise requirements—hence, sparking the trend toward light-



weight, empirical, and iterative methods. It has also driven the rise of social networking, which places the Internet at the heart of everyone's life.

Now the Industrial Internet Consortium³ claims that the IoT, building on the cloud, mobile Internet, big data, and so on, is a third such fundamental transformation. The Industrial Internet Consortium was founded in March 2014 by ATT, Cisco, General Electric, IBM, and Intel to remove roadblocks to widespread adoption of the IoT. Its mission is "to accelerate growth of the Industrial Internet by coordinating ecosystem initiatives to connect and integrate objects with people, processes and data using common architectures, interoperability and open standards that lead to transformational business outcomes."

The IoT touches everything. What is it then about the IoT that will dramatically change the business model for all industries? Here is an example: Traditionally, a company sells a product and, as long as all goes well,

doesn't know what happens to it once it has left the factory gate. By connecting the product via sensors to the IoT, the manufacturer can fundamentally change its value proposition. Instead of selling only assets, the manufacturer can sell services, including the assets, which enables it to build long-term relationships with its customers. For example, suppliers of aircraft engines can offer their products as a service (the fee based on the number of flying hours). The effect is that the supplier is now highly motivated to keep the machines running, since otherwise it will lose revenue, and the airline can increase its revenue since it will have reduced downtime. There are many similar examples from both government and industry. Basically, every industry will be affected, including banking, insurance, telecoms, airlines, and defense.

As voiced by Alex Sinclair, CTO of the GSMA: "We believe that with the right standards and regulation in place it will have a fundamental im-

pact on the way we live and work, reducing waste and inefficiencies and delivering major social and environmental benefits in security, health care, transportation and logistics, education and energy, amongst many other sectors of the economy."¹³ The IoT will eventually reach all areas where humans are providing products or services, both today and in the future. Moreover, it will use all of the kinds of systems in use today: communication, mobile, distributed, big data, cloud computing, among others, and it will drive new technologies not yet seen.

To be successful, companies will need to be able to respond quickly to the changing demands of the network while maintaining appropriate levels of engineering discipline, particularly for the cloud-based services upon which the distributed devices will depend. Moreover, the space to be addressed covers all levels of complexity—from very simple software running on basic sensors and other simple devices to

the high-performance, highly reliable, highly governed, secure, resilient, scalable systems needed to process, analyze, and respond to the vast amounts of data they produce, and everything in between. Not only that, the rate of change and the need for innovation will never have been higher.

The IoT Needs Everything

The IoT does not lack methods. Researching the space shows clearly, and not surprisingly, that there is not a one-size-fits-all approach. Instead, methods for waterfall and Agile, methods for small applications (apps) and for complex systems of systems, and methods for systems engineering (that is, for systems with hardware and software integrated) are all still needed. What is really new is that a larger vendor needs all this at the same time and with compressed time scales, which increases complexity significantly. Thus, for larger vendors a multitude of methods are needed. A smaller vendor needs a more specific and focused approach, but one that can grow as new products evolve and new problems emerge. Thus, methods such as Rational Unified Process (RUP) and SAFe, and practices such as Scrum, user stories, and use cases are all being applied. As always with any new trend, new branded methods are born. Literature regarding methods for the IoT is extremely sparse at the time of this writing. We have found two methods within the domain: Ignite¹³ and the IoT Methodology.²

The Ignite IoT Methodology. Ignite is an enterprise methodology for a major player in the IoT. It is a “big method” covering all aspects of developing for the IoT. It has two major practice areas. (In this article, *practice* is defined as a repeatable approach to doing something with a specific purpose in mind.⁹ Practices are the things that practitioners actually do.) These areas are *strategy execution* and *solution delivery*. Strategy execution is about agreeing what to build (that is, the solution) and involves the practices of opportunity identification, opportunity management, and initiation. Solution delivery is about delivering the solution to users, and it has a life cycle consisting of planning, building, and running (that is, operating the solution). Planning involves project initiation, whereas building and running are carried out through parallel project workstreams.

Project initiation is a set of practices that results in a number of different artifacts, including solution sketches, a milestone plan, user interface mockups, and software architecture. Project workstreams consist of a complementary set of practices (called workstreams): project management, cross-cutting, solution infrastructure and operations, back-end services, communication services, on-asset components, and asset preparation.

At a high level, these might seem to all be very general practices, but embedded within are two domain-specific practices: project dimensions (PD) and asset-integration architecture

(AIA). The intention is that the PD practice should be used to conduct project self-assessments, compare different IoT options, and select the solution architecture and technologies to be used in a project. The AIA practice is then used to identify the devices, gateways, and services, and their responsibilities for an enterprise solution. Ignite provides a set of technology patterns (such as machine-to-machine connectivity, and sensor networks, among other).

The benefit of Ignite is that it is based on real-world experience, capturing this experience and best practice in a well-thought-out and comprehensive methodology. Naturally, the first thought of the authors of the methodology was not so much about the modularity of the practices described but about the completeness and relevance of the method as a whole.

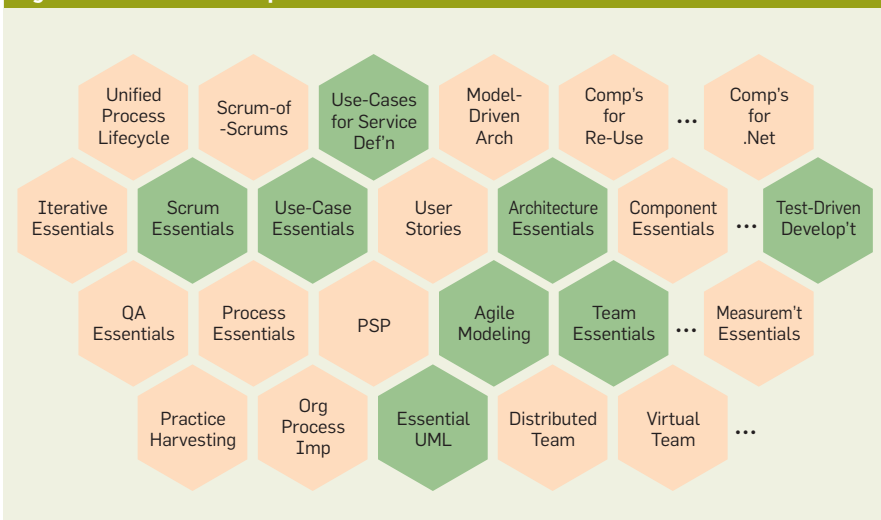
The IoT Methodology. In comparison, the IoT Methodology² is a lightweight method highly inspired by lean startup¹² and design thinking.¹ It involves the following iterative steps:

1. *Co-create.* Communicate with end users and stakeholders to identify pain problem areas in a nontechnical way.
2. *Ideate.* Simplify discussions to communicate requirements to designers, implementers, and project managers.
3. *Question and answer.* Translate soft concepts into hard requirements, analyze solutions, and brainstorm options.
4. *Map IoT OSI* (Open Systems Interconnection). Map requirements to a valid architecture, infrastructure, and business frameworks, similar to the layered approach used in the ISO/OSI model.
5. *Prototype.* Use standardized toolkits to build prototypes and iterate toward minimal viable products.

6. *Deploy* continuously to close the feedback loop and improve the products.

Like Ignite, this seems to be a very generic method at the high level. What's so special about IoT Methodology is its use of an IoT Canvas and an IoT OSI reference architecture. The IoT Canvas is an adaptation of the business model/lean canvas used in brainstorming sessions to validate minimal viable product requirements for IoT projects. The IoT OSI reference model is an adaptation of the

Figure 1. An abundance of practices.



seven-layer ISO/OSI reference model for use with IoT solutions. This IoT OSI reference model consists of five layers, with endpoints at the bottom, connectivity, middleware, IoT services, and, finally, applications at the top. Stakeholders and developers use the IoT Canvas and IoT OSI reference model to co-create and co-evolve a solution definition before prototyping.

The IoT Methodology has taken agile thinking as a starting point but is also a monolithic method.

New Practices Are Needed

It is clear from these methods, and our own experience handling emerging technologies, that new domain-specific practices will be needed to handle the very nature of the IoT—particularly practices to handle these concerns:

- *Distributed.* These systems are typically far more distributed than most other software systems. Experience from the development of telecommunication systems will come into play: new failure modes (due to communications), reliability engineering, redundant systems development, and so on.

- *Mobile.* Again telecommunication vendors have practices to develop mobile systems, which are applicable. For example, these systems have to degrade gracefully, security is critical, and they must be robust.

- *Human out-of-the-loop.* The whole idea of the IoT is to sense/analyze/activate without a human in the loop—for example, self-driving cars, automated trading systems, and population health integration systems. There may be practices to be designed here, around reliability, failure management/failover, and exception condition management.

What isn't needed are new management practices.

Both Ignite and IoT Methodology are monolithic methods that reuse many existing generic practices, combining these with new innovative practices specifically for the IoT—sadly, in a way that makes the new practices difficult to reuse and share. This issue can be easily fixed, however, by taking them to the next level by *essentializing* them and freeing their practices. This means capturing the essence



Essence provides a common framework for describing all practices and then composing them into many methods.



of a practice, which consists of the things to work with, things to do, and competencies and patterns to provide minimal explicit guidance to apply the practice effectively. This does not just make the practices more accessible, but it also makes them easier to learn, change, and use for teams that adopt them. Later, we look at how one of them—the Ignite Methodology—could be essentialized.

The IoT Needs Essence

As discussed previously, the IoT requires many methods and practices, some of them specific to the domain and others that are generally accepted good software development practices. For example, they need to deal with specific problems of distribution and mobility, yet at the same time they must be grounded in sound architecture practices.

Essence and practices. The software development world has already identified and described hundreds of different practices, some of which are shown in Figure 1. Those shaded in green are selected for an IoT team. In an ideal world teams would be able to select the set of practices they need to address their current situation and easily assemble them into a method. For example, a team building software for the IoT with a high level of engineering complexity and a high rate of change may choose to base their method on the practices highlighted in green using Use Case and Architectural Essentials to provide the required engineering rigor, and Scrum and Agile Modeling to cope with the high rates of change.

The problem is these practices come from different sources and do not share the common ground needed to allow them to be readily composed into an effective method. This isn't a problem unique to the IoT; it is a problem that has been plaguing the software industry since its inception and one that gets worse with every advance in technology.

How can teams be empowered to own and control their methods while providing them with the guidance they need to be successful, and reflecting the owning organization's need for governance and compliance? How can teams benefit from the growing

number of proven practices while continuing to innovate and rise to the new challenges that they face every day? These are issues that particularly affect companies moving into the IoT, as they will need a variety of methods.

What is needed is some concrete common ground that the practices can share, providing both a shared vocabulary for practice definition and a framework for the assembly and analysis of methods.

This will allow organizations to prepare a library of practices suitable for their industry/domain—practices that teams can easily share, adapt, and plug and play to create the innovative ways of working that they need to excel and improve.

This common ground has already been prepared in the form of the Essence kernel, part of the new OMG

standard Essence,⁹ which provides a foundation that allows teams to share and free the practices from the shackles of monolithic methods.

Essence provides the following:

- ▶ A kernel of elements that establishes a common ground for carrying out software engineering endeavors and assembling methods

- ▶ A simple, easy-to-understand, visual, intuitive language for describing practices that can be used both to represent the kernel and to describe practices and methods in terms of the kernel

By combining these capabilities, Essence provides a common framework for describing all practices and then composing them into many methods.

The power of Essence in addressing the method complexity inherent in developing software for the IoT comes from its ability to enable the composition of practices into methods; help clearly define life cycles and checkpoints, enabling practice-independent governance; and support the creation of practice libraries from which practices can be selected to be composed into methods.

Let's now look at each of these in more detail.

Composing practices into methods.

In the past, different methods have primarily been described as isolated, conceptual islands. Every method is basically a unique phenomenon, described in its own language and vocabulary and not standing on any widely accepted common ground. Any method, however, may be considered to be composed from a number of practices.

For example, the agile method of extreme programming (XP) is described as having 12 practices, including pair programming, test-driven development, and continuous integration. Scrum, on the other hand, introduces practices such as maintaining a backlog, daily scrums, and sprints. Scrum is not really a complete method, though; it is a composite practice built from a number of other practices designed to work together. Scrum can itself be composed with other practices from, say, XP, to form the method used by an agile team. This composition is typically done tacitly, as Scrum and XP are not provided in a format that allows them to be explicitly composed.

As discussed previously, Essence provides a framework and language for describing and composing practices. This framework provides a practice architecture where, as shown in Figure 2, both generic and domain-specific practices are described and assembled on top of the Essence kernel.

Now individual practices can be described using Essence. A practice can be expressed by extending the kernel with practice-specific elements, by describing the activities used to progress the work and the work products produced, and by describing the specific competencies needed to carry out these activities.

Liberating practices in this way is very powerful. Once practices are codified in Essence, teams can take ownership of their way of working and start to assemble their own methods. This can start with even a simple library of practices, as shown in Figure 3.

This capturing and sharing of practices, both generic and domain-specific, in a way that lets them be applied alongside popular management practices (agile or otherwise), provides the raw materials that teams need to compose their own ways of working.

Figure 2. The essence practice architecture.

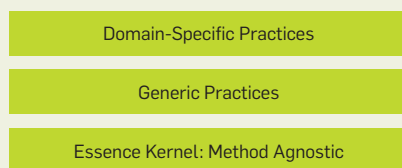
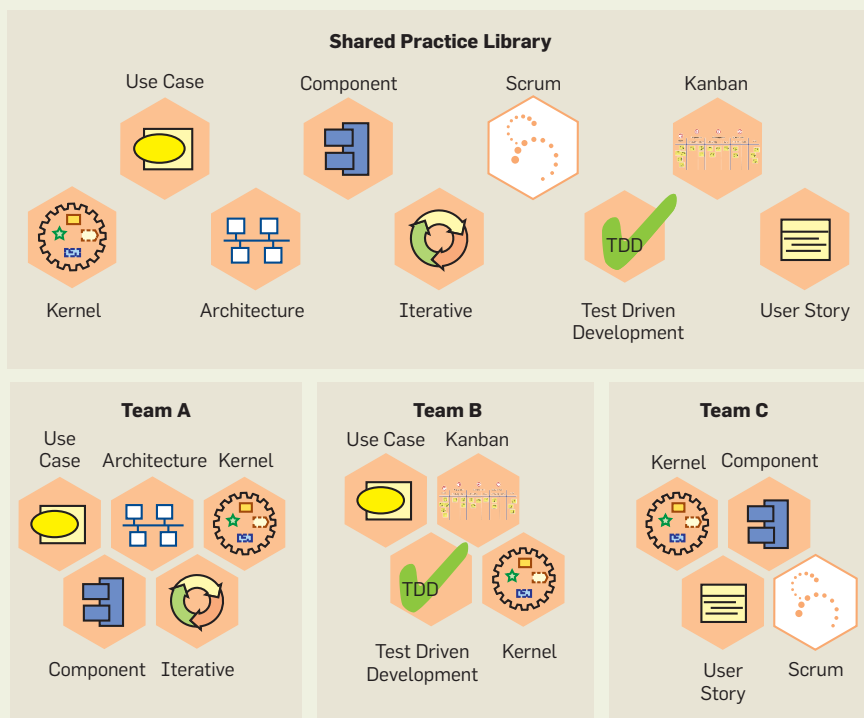


Figure 3. Three teams sharing a simple practice library.



Bringing a set of practices into this common system also allows gaps and overlaps to be more easily identified. The gaps can then be filled with additional practices and the overlaps resolved by connecting the overlapping practices together appropriately.

Governance and compliance. The Essence kernel allows you to define life cycles easily in a practice-independent way. Having a selection of different life cycles is incredibly useful when tackling a domain as complex as the IoT—particularly when the life cycles can be combined with whichever set of practices the team wants to use, ensuring that appropriate governance is applied without compromising the other aspects of the team’s way of working.

Using the Essence kernel makes it very easy to assemble a number of life cycles, each built using the same building blocks but addressing a different context and containing its own contextualized checkpoints. For example, Munich Re¹¹ defined a family of life cycles, each addressing a different context:

- *Exploratory:* A lightweight agile development life cycle for experiments, proof of concept, and small creative endeavors
- *Feature growth:* A rigorous engineering life cycle to support rapid feature growth with a strong architectural foundation
- *Maintenance and small enhancements:* A lightweight life cycle to enable the continuous flow of small enhancements and bug fixes for a fixed, funded period of time (typically a year)
- *Support:* A support-focused life cycle to aid in the transition between the development and support organizations

The ability to capture checkpoints and life cycles in a practice-independent way is incredibly powerful. It liberates the practices, allowing them to be used where appropriate and not constraining them to any predefined type or style of development. It also makes it possible to address the entire IoT methods space with a minimal, extensible, evolving set of practices, and allows teams to get the help they need without compromising their agility or engineering rigor.

Introducing Essence

The new Object Management Group (OMG) standard Essence⁹ is designed to support organizations and communities in becoming learning organizations with empowered teams that own their own ways of working and share their practices.

In addition to liberating the practices by enabling them to play well together, Essence does the following:

- Makes methods significantly lighter by focusing on the essentials.
- Helps teams measure progress and health in a method-independent way.
- Allows organizations to build a library of practices from which teams can select the ones needed for a particular solution (some teams need a “big” method, while others need only a small one).
- Helps organizations build “forever” learning organizations.

Essence provides a foundation for software engineering methods. This foundation helps in two ways: enables teams to understand and visualize the progress and health of their endeavors, regardless of their ways of working; and, allows teams to easily share, adapt, and plug and play their practices to create the innovative ways of working that they need to excel and continuously improve.^{6,7}

It guides *developers* in achieving measurable results and reusing their knowledge in systematic ways.

It helps *executives* lead programs and projects in balanced ways, without more governance than necessary, and develop learning organizations.

Note that Essence is generic enough to support a waterfall life cycle, as well as agile approaches.

Building a practice library. It is easy to see how the use of Essence would readily allow the assembly of a comprehensive practice library containing all the practices needed for a particular domain in a way that empowers teams to select just the practices they need to build their methods. Over the past few years, working in many areas of software development, including embedded systems, financial systems, telecommunications, modems, and many other areas affected by the IoT, Ivar Jacobson International (IJI; <https://practicelibrary.ivarjacobson.com/>) start has built an Essence-based practice library. Its library caters to both the craft and engineering ends of the development spectrum.

The practice library is constantly evolving as more and more practices are captured in the Essence language. At press time, IJI has essentialized close to 30 practices, including:

- Agile essentials, such as daily stand-ups, product ownership, and agile retrospectives,
- Common agile practices such as Scrum, user stories, and continuous flow,
- Proven architectural practices such as Use-Case 2.0, architectural essentials, and component-based development, and
- Life cycles such as the ones defined by Munich Re, discussed earlier.

Parallel to these efforts, existing methods such as dynamic systems development method (DSDM) and the Unified Process are being essentialized.^{8,10} An essentialized method is first structured in terms of its inherited practices, and then each practice is essentialized without changing its original idea.

All of these practices are built on top of the kernel and can be assembled to prime the pump for the methods that your teams will use. For example, organizations have used these practices to create lightweight agile methods, robust software engineering methods, pull-based flow methods, and flexible method families. They have been used to create both agile and waterfall methods that share many of the same practices but apply them with a very different emphasis.

What is powerful here is that these methods all share the same foundation and can adapt to changing circumstances by dropping and adding practices. The methods can also share practices, helping the teams—and the software they produce—to align and collaborate with one another.

To make the practices accessible and easy to learn, they are all available in card and electronic formats. Easy-to-use tools are available for practice and card creation, for method composition and publication, and for practice exchange and community building. These tools make it easy to extend ex-

isting practices to meet your needs and local standards, add your own practices, define practice-independent life cycles, and build your own frameworks and methods.

This allows you to leverage not just the industry best practice captured in the IJI practices, but also your own best practices, be they technical, financial, motivational, or managerial.

Building a Practice Library for the IoT

Examining the practices found in Ignite helps illustrate how to add domain-specific practices to a practice library.

Ignite describes a number of IoT-specific practices, including the AIA practice discussed earlier. Today, the generic practices in Ignite are not described in any detail, a gap that can easily be addressed by reusing the generic practices available in the IJI practice library.

Essentializing Ignite in this way helps distinguish the IoT-specific practices in a way that allows them to be adopted separately and applied alongside whatever generic practices the team or commissioning organization deems to be the most appropriate.

New domain-specific practices. By their very nature, the practices in the IJI practice library are very generic and applicable to many software-development domains. These generic practices are useful for many kinds of software, including for the Internet of Things.

The specific practices from Ignite and IoT Methodology are useful domain-specific practices that help address specific challenges for IoT applications. In addition, practitioners would have to work with specific technologies such as EPC (Electronic Product Code) to identify smart objects over an RFID network communicating with REST (representational state transfer) interfaces.⁵ Thus, there would be other domain-specific practices to use EPC and REST correctly.

Let's take a peek at how domain-specific practices are added to the practice architecture. A method has many aspects, such as team collaboration, how to manage requirements, architecture, and so on. In the discussion to follow, as shown in Figure 4, the focus is on architecture aspects because IoT ap-



The IoT will eventually reach all areas where humans are providing products or services, both today and in the future.



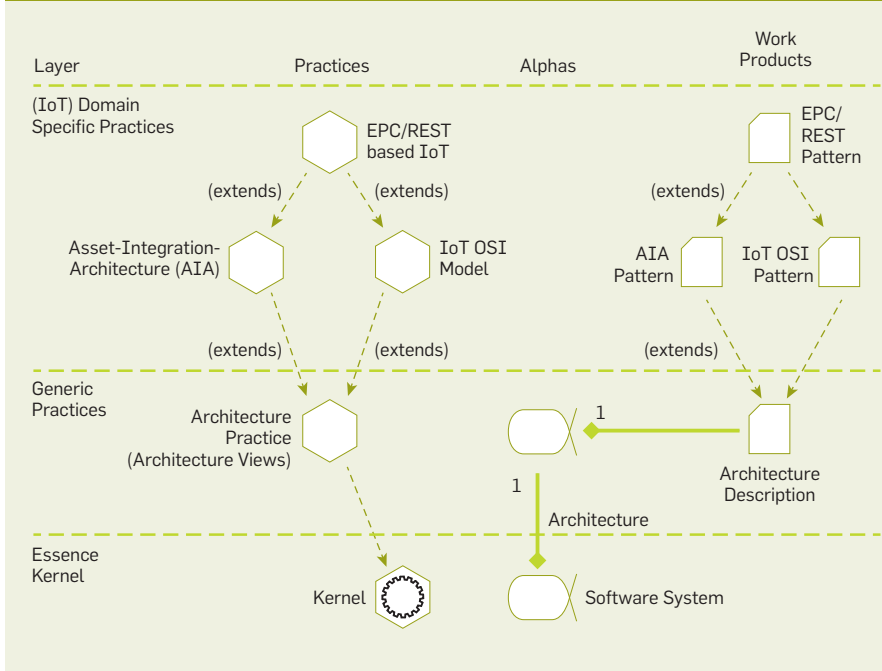
plications, with their high distribution and ubiquity, require serious attention to architecture.

At the kernel layer, Essence provides guidelines for working with the software system. IJI's generic library has a practice for working with architecture, including guidelines for creating a sound architecture description (a work product) in an agile and lightweight manner. The Ignite method recommends using AIA as a way to describe architecture, and IoT Methodology recommends using its IoT OSI model. An application that uses EPC and REST would have technology specifics about how to name products and connections and so on.

Let's dive into the practices identified in Figure 4. The Essence language specifies a number of constructs. For brevity, this article illustrates only alphas and work product. An *alpha* is "an essential element that is relevant to an assessment of the progress and health of a software engineering endeavor."⁹ The alphas provide descriptions of the kinds of things that a team will manage, produce, and use in the process of developing, maintaining, and supporting software and, as such, are relevant to assessing the progress and health of a software endeavor. "A *work product* is an artifact of value and relevance for a software engineering endeavor. A work product may be a document or a piece of software."³ Practices are a kind of package consisting of these elements.

The Essence kernel, which stands at the bottom of Figure 4, is made up of a number of elements. The figure specifically shows the Software System alpha. The Essence kernel does not have an explicit notion of architecture because in simple development, this is left for teams to define. For more sophisticated development, the architecture practice fills the gap by providing explicit guidance on creating an intentional architecture. The architecture practice introduces an Architecture alpha that is described by an architecture description work product. The Architecture alpha provides guidance on how to determine architecture goals and how to identify and validate architecture scenarios.

The two domain-specific practices—namely, AIA practice and IoT OSI practice—provide specializations on

Figure 4. Architecture practices in Ignite and IoT Methodology.

how an IoT application architecture is described. The way teams work on an IoT architecture is similar to the way they work on other kinds of architectures. Thus, they do not introduce a new Architecture alpha but reuse the Architecture alpha and description from the generic architecture practice. There are specific considerations peculiar to IoT applications, however. Hence, each of these domain-specific practices introduces a pattern for describing an IoT application. A pattern provides domain/technology-specific stereotypes to model the IoT application. In Unified Modeling Language (UML) speak, this corresponds to a UML profile.⁴ UML profiles are a common approach to describe domain-specific architectures, and IoT is one such domain. The AIA practice introduces an AIA pattern for the architecture description, whereas the IoT OSI practice introduces an IoT OSI pattern. At the very top is a technology-specific architecture practice for EPC/REST-based IoT applications. This contains a specific pattern for EPC/REST.⁵

The layering of practices helps practitioners understand what is truly different when working with IoT-based applications, as opposed to a more general application. Understanding this difference helps practitioners quickly pinpoint the specifics they need to be aware of and, hence,

learn a domain quickly. This practice separation is in contrast to monolithic methods where salient aspects of such methods often drown in the sea of generic information. It also helps practitioners differentiate methods—for example, Ignite and IoT Methodology—from the way they work with architecture and to understand if they are truly different. Practice separation also helps practitioners pick the best parts from different methods, provided they have been decomposed, as shown in Figure 4. This mix-and-match approach helps teams become innovative with methods, as well as the solutions they produce.

Thus, architecture is one area that needs special attention when building IoT applications. Security and privacy also need special consideration. The IoT opens the world to new ideas and use cases, and, as such, product idea generation and formulation also need special considerations. Each of these areas require generic practices and domain-specific practices.

Welcome to the Future

The IoT promises a new dawn for all sorts of industries, fundamentally changing the basics of everyday life. Let's make sure our software-engineering practices do not get left behind. Let's stop producing inflexible, monolithic methods that are not easy

to adopt. Instead, the focus should be on essentialized practices that provide an incremental and safe path for teams and organizations to evolve and grow their ways of working.

By using Essence as the foundation for a new practice library, we can liberate the practices and provide development teams with the guidance they need to innovate, improvise, and excel. We can avoid the traps of the past and enable software-engineering methods to evolve at Internet speeds while building on established, proven practices. ■

References

1. Brown, T. Design thinking. *Harvard Business Review* 86, 6 (2008), 84.
2. Collins, T. A methodology for building the Internet of Things; <http://www.iotmethodology.com/>
3. Evans, P.C., Annunziata, M. Industrial Internet: Pushing the boundaries of minds and machines. GE, 2012; www.ge.com/docs/chapters/Industrial_Internet.pdf.
4. Fontoura, M., Pree, W., Rumpe, B. *The UML Profile for Framework Architectures*. Addison-Wesley Longman Publishing, 2000.
5. Guinard, D., Mueller, M., Pasquier-Rocha, J. Giving RFID a REST: Building a Web-enabled EPCIS. *Internet of Things*. IEEE, 2010, 1–8.
6. Jacobson, I., Ng, P.-W., McMahon, P. E., Spence, I., Lidman, S. The Essence of software engineering: The SEMAT kernel. *Commun. ACM* 55, 12 (Dec. 2012); and [acmqueue 10, 10; http://queue.acm.org/detail.cfm?id=2389616](http://queue.acm.org/detail.cfm?id=2389616).
7. Jacobson, I., Ng, P.-W., McMahon, P. E., Spence, I., Lidman, S. *The Essence of Software Engineering: Applying the SEMAT Kernel*. Addison-Wesley, 2013.
8. Jacobson, I., Ng, P.-W., Spence, I. The Essential Unified Process. *Dr. Dobbs' Journal* (Aug. 2006); <http://www.drddobbs.com/architecture-and-design/the-essential-unified-process/191601687>.
9. Object Management Group. Essence—Kernel and language for software engineering methods, 2014; <http://www.omg.org/spec/Essence/>.
10. Page, V., Stimson, R. Essentializing the DSDM Agile Project Framework. Agile Methods Conference, London, 2016. Ivar Jacobson International; https://www.ivarjacobson.com/sites/default/files/field_jji_file/article/essentializingdsdm_1.pdf.
11. Perkens-Golomb, B., Folkjaer, P., Rauch, F., Spence, I. Ending method wars: The successful utilization of Essence at Munich Re. Ivar Jacobson International, 2015; https://www.ivarjacobson.com/sites/default/files/field_jji_file/article/essence_munichre_0.pdf.
12. Ries, E. *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. Random House, 2011.
13. Slama, D., Puhlmann, F., Morrish, J., Bhatnagar, R. *Enterprise IoT: Strategies and Best Practices for Connected Products and Services*. O'Reilly, 2015.

Ivar Jacobson, chair of Ivar Jacobson International, is a father of components and component architecture, use cases, the Unified Modeling Language, and the Rational Unified Process. He has contributed to modern business modeling and aspect-oriented software development.

Ian Spence is CTO at Ivar Jacobson International and the team leader for the development of the SEMAT kernel. An experienced coach, he has introduced hundreds of projects to iterative and agile practices.

Pan-Wei Ng coaches large-scale systems development involving many millions of lines of code and hundreds of people per release, helping them transition to a lean and agile way of working, not forgetting to improve their code and architecture and to test through use cases and aspects.

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

DOI:10.1145/3139453

Multiple computational cameras can be assembled from a common set of imaging components.

BY MAKOTO ODAMAKI AND SHREE K. NAYAR

Cambits: A Reconfigurable Camera System

THE CAMERAS IN our phones and tablets have turned us all into avid photographers, regularly using them to capture special moments and document our lives. One notable feature of camera phones is they are compact and fully automatic, enabling us to point and shoot without having to adjust any settings. However, when we need to capture photos of high aesthetic quality, we resort to more sophisticated DSLR cameras in which a variety of lenses and flashes can be used interchangeably. This flexibility is important for spanning the entire range of real-world imaging scenarios, while enabling us to be more creative.

Many developers have sought to make these cameras even more flexible through both hardware and software. For example, Ricoh's GXR camera has interchangeable lens units, each with a different type of sensor.¹² Some manufacturers make their cameras more flexible through application program interfaces (APIs) developers then use to control various camera parameters and create new image-processing tools. For example, Olympus's Open Platform Camera, released

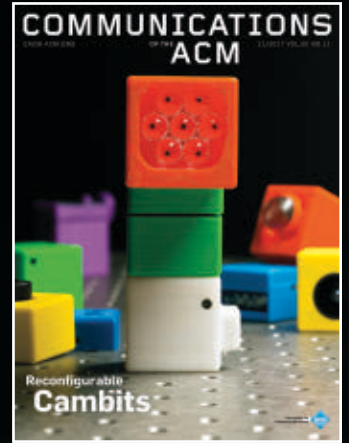
in 2015, can be controlled via Wi-Fi and Bluetooth.⁸ And at the high-performance end of the camera market, RED offers a modular camera with interchangeable parts, including lenses, battery packs, and broadcast modules.¹⁰ Although they provide some level of flexibility, such cameras are limited in the types and quality of images they are actually able to produce.

In the realm of research, Adams et al.¹ proposed a computational photography platform called Franken-camera, including API, sensor interface, and image-processing unit. That system can be used to implement various computational-imaging methods. However, its hardware is relatively rigid, limiting the extent to which it can be reconfigured. Manakov et al.⁵ proposed a camera system that can accommodate different optical add-ons, including kaleidoscope-like imaging to make optical copies of the captured image. Different filters are then used to produce high dynamic range (HDR), multispectral, polarization, and light-field images. The system provides some reconfigurability but is bulky and difficult to scale in terms of functionality. Finally, reconfigurability is a well-explored topic in the field of science education;¹⁶ for instance, Schweikardt and Gross¹⁴ developed a related robot kit, including blocks with multiple functionalities they call Cublets. And littleBits Electronics Inc. developed a modular electronic system for experiential

» key insights

- Cambits includes a set of physical blocks for building computational cameras with multiple functionalities, including high dynamic range, wide angle, panoramic, collage, kaleidoscopic, post-focus, light field, stereo imaging, and even a microscope.
- Blocks include sensors, actuators, lenses, optical attachments, and light sources attached through magnets without screws or cables.
- The configuration of the blocks can be changed without rebooting any of the related hardware or software.

IMAGES BY ALEXANDER BERG



Cambit pieces can be assembled to create a dozen different imaging systems. To celebrate this assortment, *Communications* has published four different covers, each one featuring a different Cambit configuration.



learning in which the modules can be snapped together through a magnetic interface to create circuits with various functionalities.⁴

Here, we present Cambits, a set of physical blocks that can be used to build a variety of cameras with different functionalities. Blocks include sensors, actuators, lenses, optical attachments, and light sources, assembled with magnets without screws or cables. When two blocks are attached, they are connected electrically through spring-loaded pins that carry power, data, and control signals. The host computer always knows the current configuration and automatically provides a menu of imaging functionalities from which the user can choose. Cambits is a scalable system, allowing users to add new blocks and computational photography algorithms to the current set.

Concept

Figure 1a shows the set of blocks that make up Cambits. They come in a variety of colors, each indicating a specific function: white for base, red for image sensor, blue for flash, green for actuators and spacers, yellow for lenses, and orange and purple for optical attachments. Figure 1b shows the host computer, which always knows the current Cambits configuration, using a suite of computational photography algorithms to produce a variety of images. The system reflects a number of attributes:

Ease of assembly. The blocks are attached using magnets, and the configuration of blocks can be changed without requiring a reboot of the hardware or software;

Self-identification. The host computer can detect the system's current configuration, information that is conveyed to the user through 3D visualization and a menu of functionalities it can perform;

Diverse functionality. Since there are many types of blocks, many controllable by the user, a diverse set of camera systems can be configured in which each is able to produce a different type of image; and

Scalability. The design of the system's hardware and software architecture makes it inherently scalable so new blocks and computational photography algorithms are added readily to the existing set.

Cambits is a scalable system, allowing users to add new blocks and computational photography algorithms to the existing set.



System Architecture

A number of Cambits attributes follow the design of the Cambits hardware and software architecture:

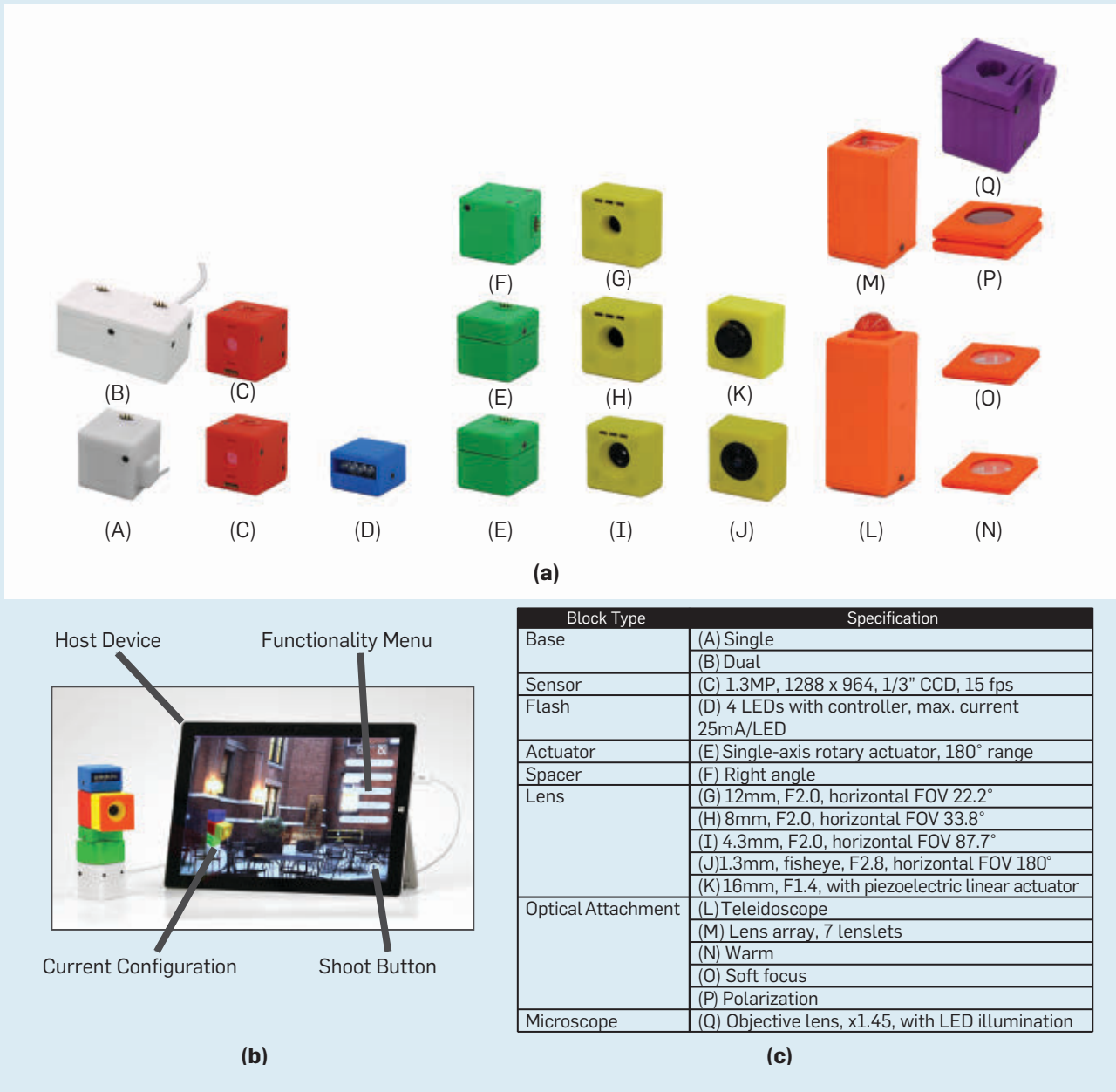
Mechanical and electrical connections. Each Cambits block is 40mm along at least two of its three dimensions. We 3D printed the chassis of each block, including sockets close to its corners designed to hold magnets; the magnets are used to attach blocks to each other, as well as to mechanically align them, as in Figure 2a; the alignment is aided by convex and concave bumps on the surface of the chassis. The polarities of the magnets in each block are also chosen such that it is not possible for the user to attach two blocks that are otherwise incompatible. For instance, a lens block cannot be directly attached to an actuator block. When two blocks are attached, a set of either four or six spring-loaded pins on one block (see Figure 2b) is aligned and electrically connected to contact pads on the other block. The system uses USB 2.0 for the data signal and I2C for the control signal.

Tree structure with bucket brigade. Each Cambits block has three types of pins for conveying power, data signals, and control signals. The data signal conveys image data from the sensor block. The control signal conveys the configuration data upstream and various commands (such as the actuator block's rotation parameters and the flash's strobing parameters) downstream.

These signals communicate through a tree structure. The host device—the root of the Cambits structure—provides electrical power to all blocks in the configuration, detects the current configuration of the entire tree structure, and controls all blocks within the tree. This design ensures each block is able to connect with multiple other blocks. Upstream is defined as the direction toward the host device and downstream as the direction in which components proceed forward from the host device (see Figure 3).

The data signal flows upstream directly from each sensor block to the host device. However, the control signals are passed from component to component in bucket-brigade fashion. Each block is able to communicate through control signals with only the blocks that are connected to it. When a

Figure 1. Cambits overview: (a) Cambits components; (b) a Cambits configuration, with host computer displaying a 3D visualization of the current configuration and a menu of functionalities it can perform; and (c) Cambits blocks and their specifications.



block is attached to the system, it scans the components downstream. If it detects any blocks, it reads the configuration data of the blocks that are downstream, adds its own identity and address to the data, and then sends the information upstream. As a result, the host device is able to detect the complete order of the configuration. If we had instead used a conventional electrical bus for the control signals, the system would not have been able to detect the order of the blocks. Moreover, a conventional bus would not be able to detect configurations in which multiple

blocks have the same address, as is possible when using the I2C interface.

When the host device seeks to control a specific block in the tree structure, it sends its command and the address of the block to the base block. The base block and subsequent blocks pass the command downstream in bucket-brigade fashion. The addressed block ultimately receives the command and executes it.

Controller board. A key aspect of the design is the controller board inside the base, actuator, spacer, and sensor blocks. It includes a microcontroller

unit (MCU) (Texas Instruments MSP430F5510) with two serial ports for the bucket brigade. The controller board has an upstream interface and a downstream interface (see Figure 4).

When a block with the controller is attached to the system, it turns on automatically, thus triggering the firmware on its MCU to start scanning downstream for approximately 100msec to communicate with its adjacent blocks. When the block is removed from the system, it loses power, and the firmware stops.

Each block has a power circuit to pre-

Figure 2. Cambits detachable connector: (a) mechanical assembly and alignment of blocks using magnets; and (b) electrical connection between blocks using spring-loaded pins that carry power, data, and control signals.

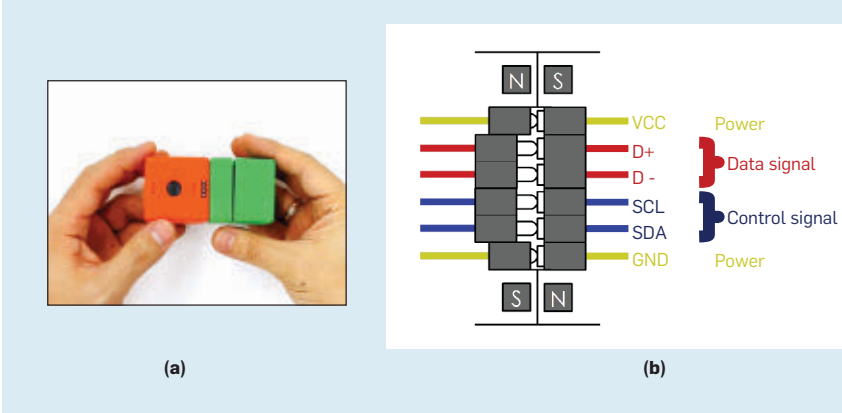


Figure 3. Tree architecture used to implement Cambits; power flows downstream, data flows upstream, and control signals are communicated in bucket-brigade fashion.

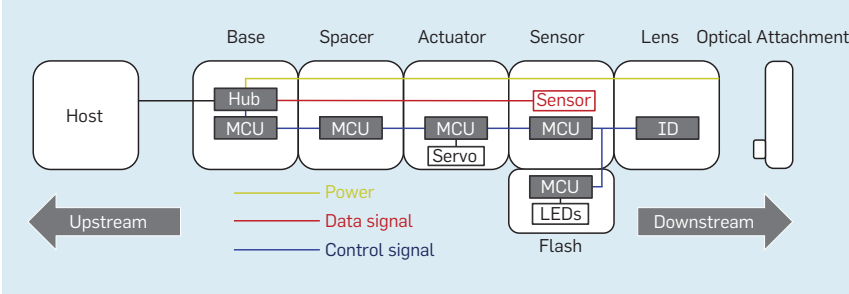


Figure 4. The base, actuator, spacer, and sensor blocks include a controller board that allows a block to communicate with its adjacent blocks.

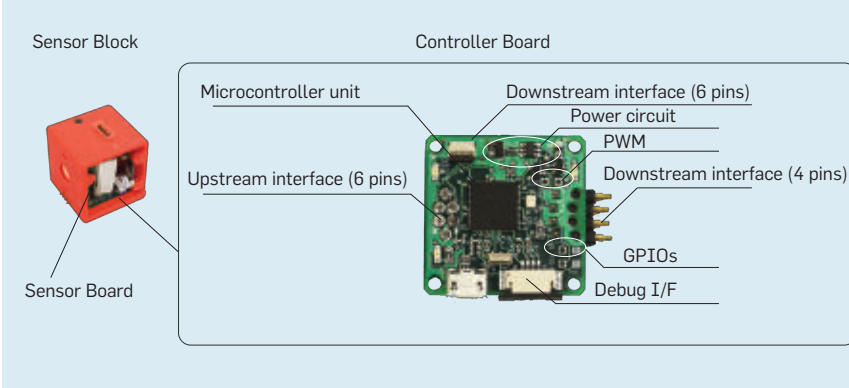
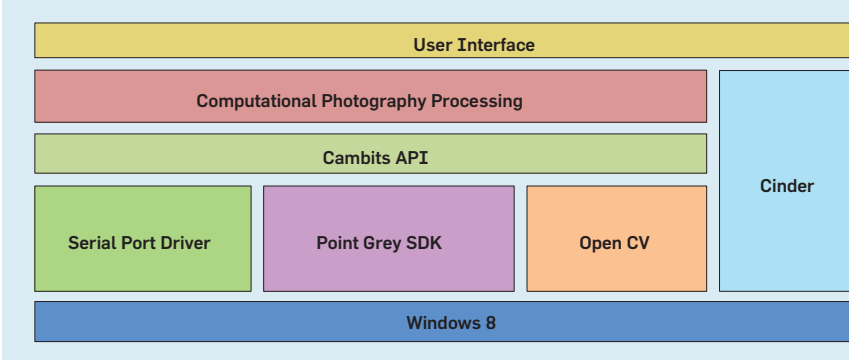


Figure 5. Cambits software architecture.



vent inrush current and voltage drop when attached, thus maintaining a steady input voltage. Due to the circuit, the system can be reconfigured without requiring a reboot of the hardware (blocks) or the software running on the host computer.

The controller board can also control other devices (such as the servo motor in the actuator block and the LED controller in the flash block) through the I2C bus, pulse-width modulation (PWM), and general-purpose input outputs (GPIO) based on commands it receives from the host device. For example, when an actuator block receives the command for rotation, the MCU generates the pulse signal needed to drive its servo motor.

The I2C interface is also useful in terms of scalability because it is widely used in the field of embedded systems, allowing us to add various extra devices (such as a light sensor, acoustic sensor, IR sensor, GPS, IMU, and multispectral light source) to the Cambits set.

Lens blocks and optical attachments. The lens block includes an identification board with I2C expander device that can detect the identification number of the lens type itself and an additional optical attachment connected to the lens (such as soft focus filter, lens array, and “teleidoscope,” or lens for creating kaleidoscope-like images). The optical attachment includes no electrical parts but does have up to three bumps that push against mechanical switches on the lens block to generate a three-bit code the lens block can use to identify the attachment. The lens block then sends this information upstream.

Sensor block. The sensor block includes a Point Grey camera board (BFLY-U3-13S2C-CS) that can produce 1.3-megapixel video in various formats (such as YUV411 and RGB8) and send the video upstream as a USB 2.0 data signal. In designing Cambits, we aimed to minimize the length of the data signal line and the number of connectors so as to enable high-frequency (480Mbps) transmission needed to preserve the integrity of the video.¹⁵ Users are able to control various imaging parameters of the sensor board (such as exposure time and gain) from the host device.

Mechanical design. As mentioned, the Cambits blocks attach to each oth-

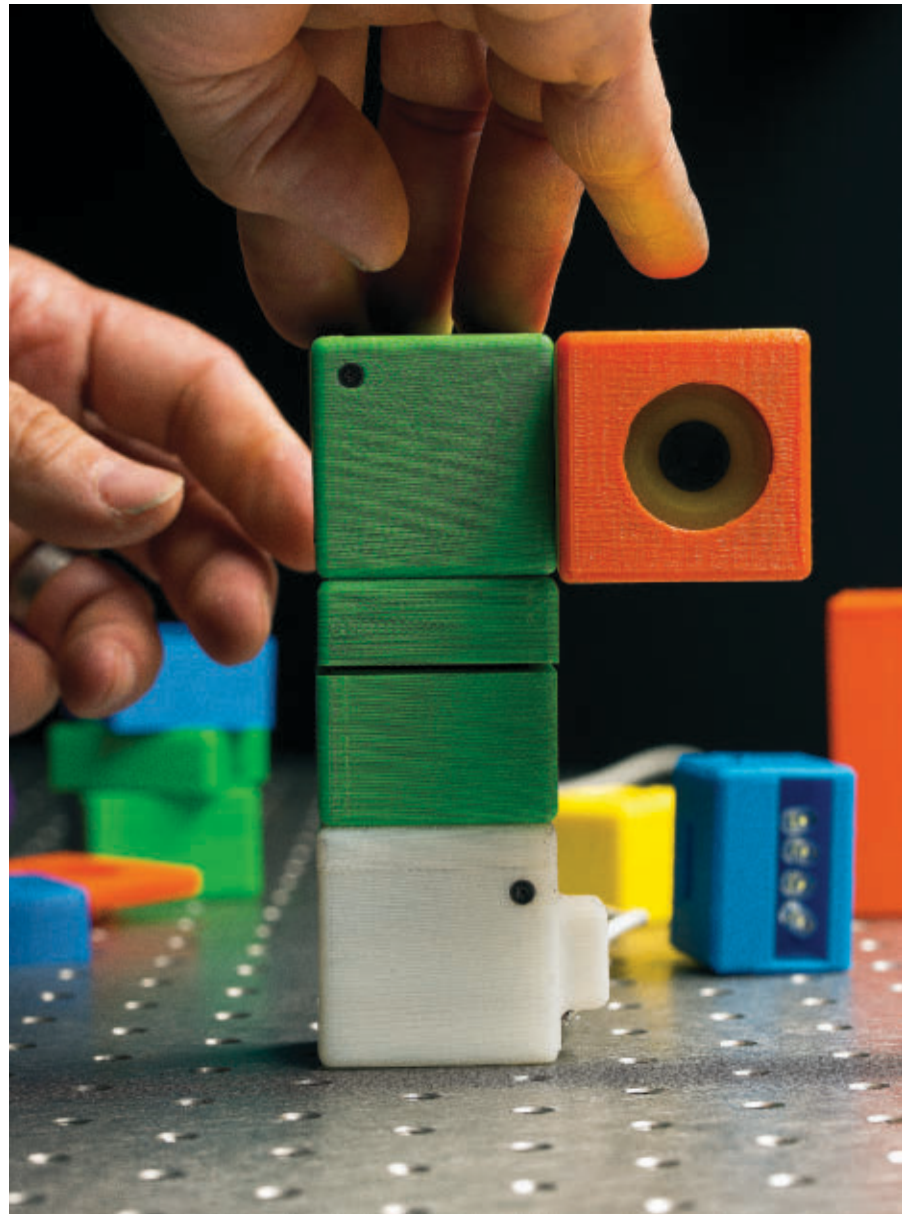
er through magnets invisible to the user, as they are embedded within the plastic (polylactic acid, or PLA) enclosing the blocks. To ensure the magnetic forces exerted through the PLA block enclosures are strong enough to keep the blocks attached, we used Neodymium block magnets in dimensions $3/8" \times 1/4" \times 1/16"$. To ensure precise optical alignment between blocks, we designed the faces of the block covers with small mechanical bumps and indentations. The dimensions of the PLA enclosures of the blocks must be precise enough to ensure that when lens and optical attachment blocks are attached to a sensor block, an image is achieved with the desired depth of field. Finally, as in Figure 4, the controller circuit board within each block includes not only the various electronic components but also the spring-loaded connectors used to electrically connect the block to the one to which it is attached. In our prototype, the PLA enclosure, magnets, and controller circuit board add approximately 5.2mm in linear dimension to the component—sensor, actuator, and flash—in a block. Detailed mechanical design files are available at <http://www.cs.columbia.edu/CAVE/projects/cambits>.

Software. The software system that runs on the host device captures images from the Cambits system, giving users a 3D visualization of the current configuration and the option to apply various computational photography methods to the captured images. The current implementation runs on Windows and is based on open source libraries (such as Open CV-v2.4.10 and Cinder v1.20) (see Figure 5). It also uses the Point Grey Flycapture2 software interface to control the image sensors.

The Cambits API is also able to receive images from the Point Grey SDK and control camera parameters (such as exposure time and gain) and various devices on the tree architecture, including servo motors, linear actuators, and LEDs, through serial ports. The API and the open libraries allow developers to add new blocks and image-processing algorithms to the system.

Functionality

We have used Cambits to assemble a range of computational cameras (see Figure 6). To construct a basic one, we



use a base, a sensor block, and 8mm lens block. In HDR mode, the camera captures six images with different exposure times—where the exposures are the geometric sequence $t, 2t, 4t, 8t, 16t$, and $32t$ seconds—computes an HDR image using a triangular weighting function,² and then “tone maps” it using Reinhard’s algorithm.¹¹

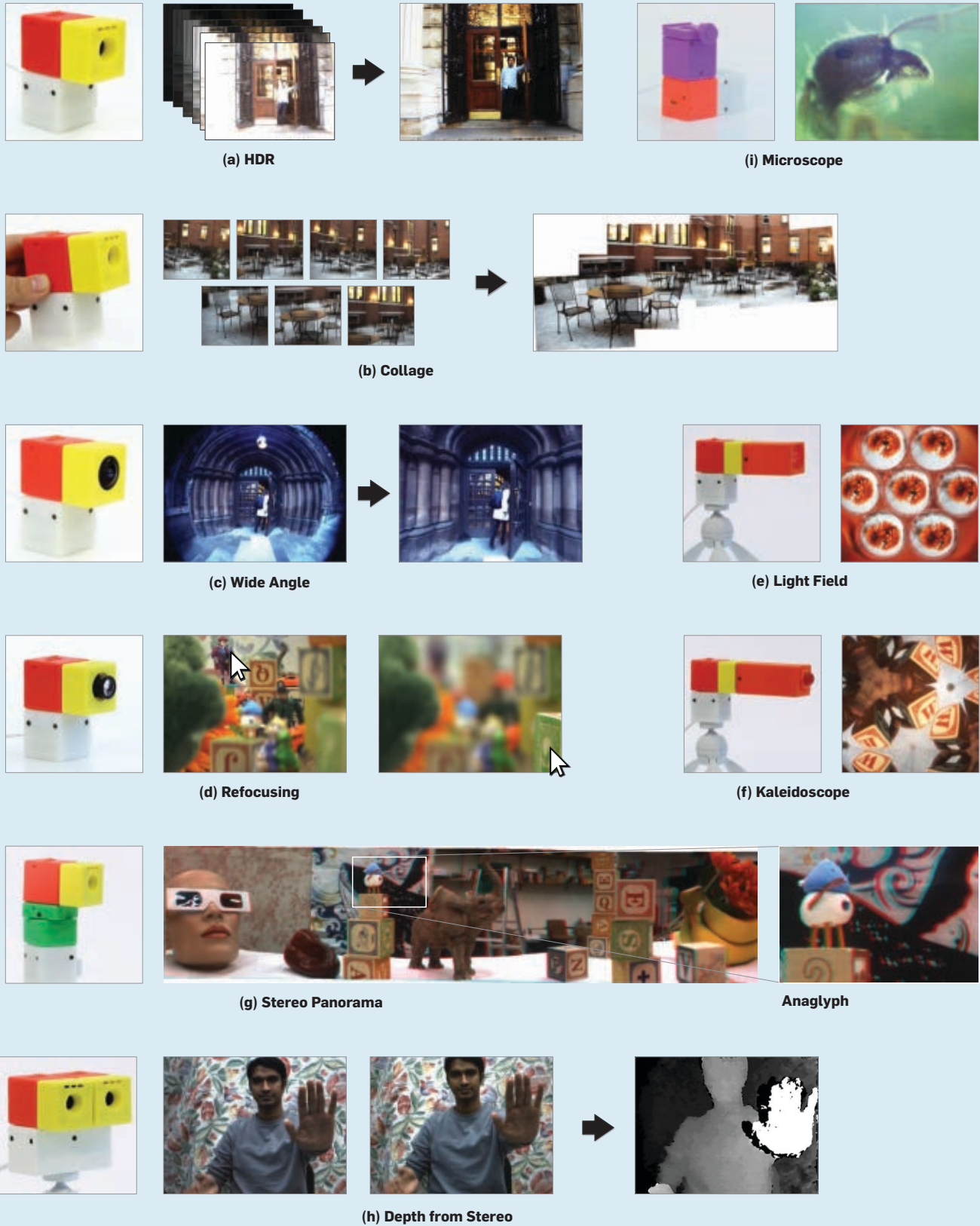
Users can also move this basic camera around to capture a set of images that can be fused to obtain a scene collage. The system is able to detect the features in each image using the scale-invariant feature transform (SIFT) algorithm, reduce the outliers using the random sample consensus (RANSAC) algorithm, and find corresponding features between pairs of images. Cambits

uses the image with the most corresponding features, with all remaining images as the center image of the collage, transforming the remaining images to align with the center image and overlay to obtain the collage.⁷

Cambits also includes a variety of lenses. For instance, a fisheye lens with a focal length of 1.3mm and f-number of 2.8 can be used to capture a wide-angle image of a scene with horizontal field of view of 180 degrees. Since the host computer knows the type of lens being used, the software automatically maps the captured image to a perspective without barrel distortion.¹³

As with the lenses, users can attach a variety of optical filters to the lens of the imaging system, including simple opti-

Figure 6. Example results; for high-resolution versions, see <http://www.cs.columbia.edu/CAVE/projects/cambits/>.



cal filters like diffusion and polarization, as well as more complex ones (such as a lens array and a teleidoscope). The Cambits lens-array attachment includes seven acrylic ball lenses to produce a 4D light-field image of the scene.³ The teleidoscope attachment produces a kaleidoscope image. An acrylic ball lens in front of the attachment captures the scene image, and a set of first-surface planar mirrors between the ball lens and the lens block creates multiple rotated copies of the image.

The focal stack lens block includes a linear actuator that physically sweeps the lens to capture a set of images corresponding to different focus settings. The linear actuator moves the lens in steps of 0.05mm, with a total travel distance up to 2.0mm, using a piezoelectric linear actuator to move the lens precisely. The captured stack of images helps compute an index map that represents the image in which each pixel is focused. The focal stack lens block then generates an interactive image that lets users click on any part of the image to bring it into focus.^{6,17}

We designed Cambits so it would be possible to insert a rotary actuator between the base and the sensor to scan a panorama of a scene. If the camera is rotated off-axis—with an offset between the rotation axis and the center of projection of the camera—users would be able to take left and right image strips from the captured sequence of images to generate a stereo panorama for creating virtual reality.⁹ In the example in Figure 6g, 120 images were taken while the actuator rotated 120 degrees and the offset between the rotation axis and the center of projection of the camera lens was 70mm.

A second rotary actuator can be added to the system to configure a pan/tilt camera system.

Cambits is not limited to a single image sensor. Its second base can be used with two sensor blocks and lenses to create a stereo camera system with a baseline of 44mm. Cambits processes the left and right video streams from this system in real time to produce a gray-coded-depth video of the scene.

Cambits can also be used to assemble a microscope that includes an objective lens, a mechanism to adjust the height of the sample slide to bring the sample into focus, and an LED light to

“bright field” illuminate the sample. The user controls the LED in terms of brightness through the host computer. Alternatively, ambient illumination in the environment can be used to back-light the sample.

Conclusion

Cambits is a versatile modular imaging system that lets users create a range of computational cameras. The current prototype is a proof of concept we use to demonstrate key aspects of Cambits: ease of assembly, self-identification, and diverse functionality. We have thus shown Cambits can be a powerful platform for computational photography, enabling users to express their creativity along several dimensions. An important aspect of Cambits is that it is designed to be an open platform that is scalable. That design allows users to add multiple hardware blocks, including structured light sources, multispectral sources, telescopic optical attachments, and even non-imaging sensors for measuring acceleration, orientation, sound, temperature, and pressure. We anticipate developing algorithms that use such a diverse set of sensors to trigger/control various image-capture-and-processing strategies. To encourage others to modify or build on the current system, we have made the details of its hardware and software design available at http://www.cs.columbia.edu/CAVE/projects/cambits/databases/cambits_supporting_database.zip.

Acknowledgments

We did this research at the Computer Vision Laboratory of Columbia University in New York while Makoto Odamaki was a Visiting Scientist from Ricoh Company, Ltd.; for the design data covered here, see <http://www.cs.columbia.edu/CAVE/projects/cambits>. We thank William Miller for designing and 3D printing the chassis of the Cambits blocks, Wentao Jiang for his contribution to the user interface, and Daniel Sims for editing the demonstration video and formatting the project webpage. Divyansh Agarwal, Ethan Benjamin, Jihan Li, Shengyi Lin, and Avinash Nair implemented several of the computational-photography algorithms. The authors also thank Anne Fleming for proofreading an early draft of the article. 

References

1. Adams, A. et al. The Frankencamera: An experimental platform for computational photography. *ACM Transactions on Graphics* 29, 4 (July 2010), article 29.
2. Debevec, P.E. and Malik, J. Recovering high-dynamic-range radiance maps from photographs. In *Proceedings of the ACM SIGGRAPH Conference* (Los Angeles, CA, Aug. 11–15). ACM Press, New York, 2008, 31.
3. Georgiev, T. et al. Spatio-angular resolution trade-offs in integral photography. In *Proceedings of the 17th Eurographics conference on Rendering Techniques* (Nicosia, Cyprus, June 26–28). Eurographics Association, Aire-la-Ville, Switzerland, 2006, 263–272.
4. LittleBits Electronics Inc., New York; <http://littlebits.cc/>
5. Manakov, A. et al. A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging. *ACM Transactions on Graphics* 32, 4 (July 2013), article 47.
6. Ng, R. et al. Light-field photography with a hand-held plenoptic camera. *Computer Science Technical Report CSTR 2, 11* (Apr. 2005), 1–11.
7. Nomura, Y., Zhang, Li, and Nayar, S.K. Scene collages and flexible camera arrays. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (Grenoble, France, June 25–27). Eurographics Association, Aire-la-Ville, Switzerland, 2007, 127–138.
8. Olympus Corporation, Tokyo, Japan; https://opc.olympus-imaging.com/en_sdkdocs/index.html
9. Peleg, S. and Ben-Ezra, M. Stereo panorama with a single camera. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (Fort Collins, CO, June 23–25). IEEE Computer Society, Los Alamitos, CA, 1999.
10. RED Digital Cinema Camera Company, Lake Forest, CA; <http://www.red.com/products>
11. Reinhard, E. Parameter estimation for photographic tone reproduction. *Journal of Graphics Tools* 7, 1 (Nov. 2002), 45–51.
12. Ricoh Company, Ltd., Tokyo, Japan; https://www.ricoh.com/r_dc/gxr/
13. Schneider, D., Schwalbe, E., and Maas, H.G. Validation of geometric models for fisheye lenses. *Journal of Photogrammetry and Remote Sensing* 64, 3 (May 2009), 259–266.
14. Schweikardt, E. and Gross, M.D. roBlocks: A robotic construction kit for mathematics and science education. In *Proceedings of the Eighth International Conference on Multimodal Interfaces* (Banff, Alberta, Canada, Nov. 2–4). ACM Press, New York, 2006, 72–75.
15. USB Implementers Forum, Inc. *High Speed USB Platform Design Guidelines Rev. 1.0*; http://www.usb.org/developers/docs/hs_usb_pdg_r1_0.pdf
16. Yim, M. et al. Modular self-reconfigurable robot systems: Grand challenges of robotics. *IEEE Robotics & Automation Magazine* 14, 1 (Apr. 2007), 43–52.
17. Zhou, C., Miao, D., and Nayar, S.K. *Focal Sweep Camera for Space-Time Refocusing*. Technical Report. Department of Computer Science, Columbia University, New York, 2012; <https://academiccommons.columbia.edu/catalog/ac:154873>

Makoto Odamaki (makoto.odamaki@nts.ricoh.co.jp) is an engineer of digital camera systems at Ricoh Company, Ltd., Tokyo, Japan.

Shree K. Nayar (nayar@cs.columbia.edu) is the T.C. Chang Professor of Computer Science at Columbia University in New York where he also heads the Columbia Vision Laboratory.

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



Watch the authors discuss their work in this exclusive *Communications* video. <https://cacm.acm.org/videos/cambits>

DOI:10.1145/3141771

The varying review dynamics seen in different app stores can help guide future app development strategies.

**BY STUART MCILROY, WEIYI SHANG,
NASIR ALI, AND AHMED E. HASSAN**

User Reviews of Top Mobile Apps in Apple and Google App Stores

ONE OF THE unique aspects of app stores is the convenience of providing user feedback.¹³ Users can effortlessly leave a review and a rating for an app, providing quick feedback for developers. Developers are then better able to update their apps. This feedback mechanism contrasts with traditional feedback mechanisms like bug-reporting systems (such as Bugzilla), which are negative in nature, as they include only bugs, unlike reviews, which can be positive. Moreover, reviews can even serve as a means for deriving additional app requirements.⁷

Developers of top apps might be overwhelmed by the large number of received reviews. Several papers

(such as by Fu et al.,⁵ Galvis Carreño and Winbladh,⁶ and Google Analytics⁷) and commercial efforts (such as Applause Analytics³) have proposed solutions to help developers cope with large numbers of reviews.

A 2013 study of reviews of iOS apps by Pagano and Maalej²⁰ found that on average a free app receives 37 reviews per day, while paid apps receive approximately seven reviews per day,²⁰ and another study of iOS apps found that 50% of studied free apps receive only 50 reviews in their first year.¹¹ Yet no prior research examined the reviews in the Google Play store, considering, say, “Is the data normally distributed or highly skewed, with only a small number of apps receiving a substantial number of reviews on a daily basis?”

Here, we explore the question of how pervasive are the frequently reviewed apps in the Google Play store. In particular, we empirically cover app reviews from the perspective of the developers of the top apps there. Through an analysis of reviews for the top 10,713 apps in the Google Play store over a period of two months—January 1 to March 2, 2014—we found:

More than 500 reviews daily. Only 0.19% of the studied apps received more than 500 reviews per day;

Majority of studied apps. Almost 88% of the studied apps received only a small number (20 or fewer) reviews per day; and

Correlates with reviews. The number of downloads and releases correlated with the number of received reviews, while the app category did not play a major role.

Some of our observations differ from other studies of user reviews of iOS

» key insights

- **The characteristics of user reviews differ depending on app store.**
- **Few mobile apps in the Google Play store attract large numbers of user reviews.**
- **More app downloads and releases correlate with more reviews in the Google Play store, whereas app category plays only a minor role.**



apps,¹¹ highlighting the need for additional in-depth investigation of the reviewing dynamics in both stores.

Mobile App Analytics

A Vision Mobile survey of 7,000 developers, also in 2014, found 40% of them made use of user-analytics tools and 18% used crash-reporting and bug-tracking tools. Other studies also found that developers need tools for app analytics. For example, a 2013 study by Pagano and Bruegge¹⁹ of how feedback occurs following initial release of a software product identified the need to structure and analyze feedback, particularly when it involves a large amount of feedback.

A number of app-analytics companies, including App Annie,¹ specialize in tools designed to help developers understand how users interact with their apps, how developers can help generate revenue (such as through in-app purchases, e-commerce, and direct buy), and how to leverage user demographics of the apps. These

companies also provide developers overviews of user feedback and crash reports. Google promotes its own extensive analytics tools for Android developers as a key competitive differentiator relative to other mobile app stores. The tools measure how users use an app (such as by identifying user locations and how users reached the app). They also track sales data (such as how developers generate revenue through in-app purchases and the effect of promotions on app sales²). However, other than crash-reporting tools, many analytics tools today are mostly sales-oriented rather than software-quality-oriented involving bugs, performance, and reliability.

Other studies have highlighted the effect of reviews of mobile apps on an app's success.^{9,15,19} Harman et al.⁹ found a strong correlation between app ratings and an app's total download numbers. User reviews include information that could help developers improve the quality of their apps and increase their revenue. Kim et al.¹⁵ interviewed app

buyers, finding reviews are a key determinant in their decisions to purchase an app. A survey by Lim et al.¹⁶ found reviews are one of the top reasons for users to choose an app. Likewise, Mudambi et al.¹⁸ showed that user reviews have a significant effect on sales of on-line products.

The importance of user reviews has motivated many studies, as well as our own work analyzing and summarizing user reviews for mobile apps (see Table 1). Iacob and Harrison¹² built a rule-based automated tool to extract feature requests from user reviews of mobile apps, an approach that identifies whether or not a user review contains a feature request. Chandy and Gu³ identified spam reviews in the Apple (iOS) App Store, using a technique that achieved high accuracy with both labeled and unlabeled datasets. Carreño and Winbladh⁶ used opinion-mining techniques and topic modeling to successfully extract requirements changes from user reviews. Fu et al.⁵ introduced an approach for discovering inconsis-

tencies in apps, analyzing the negative reviews of apps through topic analysis to identify reasons for users liking or disliking a given app. Khalid et al.¹⁴ manually analyzed and categorized one- and two-star reviews, identifying the issues (such as the hidden cost of using an app) about which users complained. Chen et al.⁴ proposed the most extensive summarization approach to date, removing uninformative reviews and prioritizing the most informative reviews before presenting a visualization of the content of reviews. Guzman and Maalej⁸ performed natural language processing techniques to identify app features in the reviews and leveraged sentiment analysis to identify whether users like such features. Our own work differs from these studies, as it aims to provide context about when the other techniques would be needed.

Pagano and Maalej²⁰ and Hoon et al.¹¹ analyzed the content of reviews of both free and paid apps in the Apple App Store, answering a similar research question as ours about the number of received reviews, but there are major differences between them and us in

terms of findings, methodologies, and context, or Android vs. iOS (see Table 2).

Studied Apps

Martin et al.¹⁷ noted that not all stores provide access to all their reviews, leading to biased findings when studying reviews. To avoid such bias, we collected all reviews on a daily basis, ensuring we would include all available reviewers. However, the Google Play Store provides access to only the 500 latest reviews for an app. If more than 500 reviews are received in the 24-hour period between daily runs of our crawler, then the crawler does not collect those reviews. This limitation means we thus offer a conservative estimate of the number of reviews for apps that receive more than 500 reviews per 24-hour time period. We based our Google Play store crawler on an open source crawler called the Akdeniz Google Play crawler (<https://github.com/Akdeniz/google-playcrawler>) to extract app information (such as app name, user ratings, and reviews). Running it meant we were simulating a mobile device over approximately two months—January 1 to March 2, 2014.

We collected review information from 12,000 free-to-download apps from the Google Play store. From among 30 different categories, including photography, sports, and education, we selected the top apps in each category in the U.S. based on app-analytic company Distimo's (acquired by App Annie) ranking of apps for a total of 12,000; Distimo ranked the top 400 apps for each of the 30 categories. We used Distimo's Spring 2013 list of top apps. Of the 12,000 top apps, 1,287 were not accessible during our two-month crawl because some of them might have been removed from the store. We thus collected data from 10,713 top apps, with a total of 11,047 different releases during the studied time period.

Our own selection of top apps might have biased our results, possibly generalizing to only the top, stable, free apps in the Google Play store. Nevertheless, we studied successful apps we felt were more likely to have a large user base and receive a large number of reviews, rather than blindly study all apps. We chose apps that had been popular one year before we began our study because we were

Table 1. Our observations on Google Play apps compared to the Pagano and Maalej²⁰ and Hoon et al.¹¹ observations on the Apple (iOS) App Store.

Item	In the Apple App Store, from Pagano and Maalej ²¹	In the Apple App Store, from Hoon et al. ¹¹	In the Google Play Store, from us	Notes
Reviews received	Average of 22 reviews per day, with 36.87 for free apps and 7.18 for paid apps	Median of 50 reviews in first year for free apps and 30 reviews in first year for paid apps	Average of seven reviews per day, with median of no reviews per day for free apps	We found fewer average and median user reviews compared to Pagano and Maalej ²¹ and more user reviews than Hoon et al. ¹¹ Reviews were skewed, with median number of received reviews at 0 and 88% of the studied apps receiving 20 reviews or fewer per day.
Number of reviews received	Facebook received 4,275 reviews in one day	(not studied)	Only 0.19% of apps received more than 500 reviews, and the top 100 most-reviewed apps had 6,000 to 43,000 reviews in the two-month study period.	Pagano and Maalej ²¹ were the first to observe that some apps (for them, the Facebook app) might receive a large number of reviews per day. We were first to explore this observation—apps receiving a large number of reviews per day—in depth, finding that while some apps might receive a large number of reviews, only 0.19% of all studied apps received more than 500 reviews per day. Most top apps might not benefit much from automated approaches that leverage sophisticated techniques (such as topic modeling) given the small number of reviews they received and their limited length.
Effect of app category	Number of daily reviews differs by category	Certain categories receive greater numbers of reviewers than others	No relation	Compared to both iOS studies, we found no relation between an app's category and number of received reviews, once we controlled for number of downloads and number of releases.
Spike in number of reviews decreases following release	Number of reviews decreases over time following a release	(not studied)	The standard deviation of received reviews deviates from the median directly following release and returns back to normal afterward.	Both stores showed evidence of spikes in number of reviews immediately following a new release.
Average length of a review	Average of 106 characters and median of 61 characters	Average of 117 characters and median of 69 characters	Average of 64 characters and median of 36 characters	Reviews in the Google Play Store were shorter than in the Apple App Store. Median length of reviews demonstrated that the distribution of review length is highly skewed, with long reviews as outliers.

interested in stable, mature apps that had not been released within the past few months to avoid the expected burst of reviews following an app's initial release.²⁰ We focused on free-to-download apps, since recent work showed that free apps receive five times as many reviews as paid apps.²⁰ Moreover, over 90% of downloaded apps were, at the time, of the free-to-download variety, according to Gartner. Such apps use other revenue models (such as freemium, in-app purchases, and ads). The developers of such apps are thus concerned about the effect of reviews on their revenue.⁹

Findings

Here, we present our findings, as in Table 2, concerning the reviews from the Google Play store while comparing our results with prior studies.

Number of received reviews. On the number of received reviews in the Google Play Store

Finding 1. Most apps (88% of those of the 10,713 we studied) received few reviews during our studied time period. The average and median number of reviews were fewer than Pagano's and

Maalej²⁰ and greater than Hoon et al.,¹¹

Finding 2. The number of user reviews were skewed; similar findings were reported by Pagano and Maalej;²⁰ and

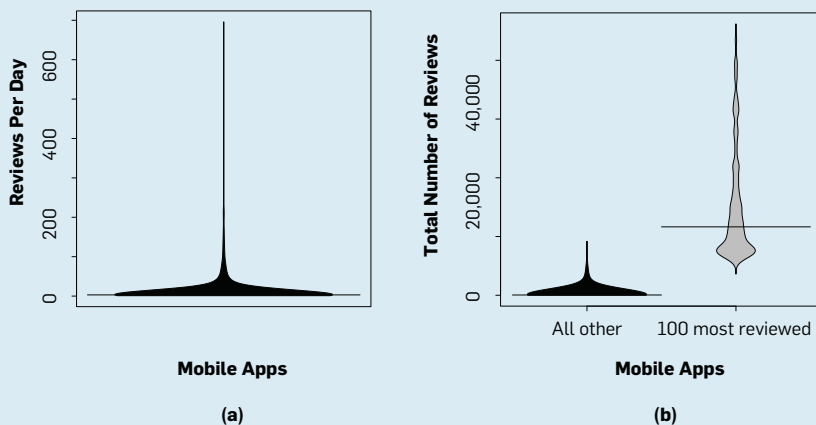
Implication. Most top apps might not benefit much from automated approaches to analyzing reviews that leverage sophisticated techniques (such as topic modeling) given the small number of received user reviews and their limited length.

We plotted the number of reviews per day, as well as total number of received reviews, using a beanplot combining a boxplot with a kernel-density-estimation function. Figure 1a reports the median number of reviews per day was 0. We found 20, or 0.19%, of the 10,713 studied apps received 500 or more reviews; as mentioned earlier, 500 would be a conservative estimate, whereas 88% of the apps in our 10,713-app dataset received fewer than 20 reviews per day. Additionally, the median total number of reviews was 0 during the study period. We also calculated the number of words in each of the received reviews, with median number of words per review at 46.

Table 2. Datasets of prior work mining reviews of mobile apps.

Paper	App Store	Apps	Reviews
Jacob and Harrison ¹²	Google Play Store	161	3,279
Galvis and Carreno ⁷	Google Play Store	2	710
Fu et al. ⁵	Google Play Store	171,493	13,286,706
Chen et al. ⁵	Google Play Store	4	169,097
Pagano and Maalej ²¹	Apple App Store	1,100	1,126,453
Hoon et al. ¹¹	Apple App Store	17,000	8,700,000

Figure 1. Beanplots showing number of reviews per day and in total.



We found fewer average reviews per day than Pagano and Maalej²⁰ possibly due to any of several factors. The first is we collected reviews from stable top apps that had been released for at least one year, whereas Pagano and Maalej²⁰ may have collected new apps and not focused on top apps. The second was that our estimates for the frequently reviewed apps were conservative; we did not count more than 500 reviews in a day. For instance, Pagano and Maalej reported that Facebook received 4,275 reviews in a day, with such large numbers increasing the overall reported average number of received reviews on a daily basis. We separated the apps into two groups: 100 most-reviewed apps and all other apps. Figure 1b reports there was a large gap in the total number of reviews among the 100 most-reviewed apps. The total number of reviews of the 100 most-reviewed apps ranged from 43,000 to 6,000 in the two-month study period. The reviews themselves were short, much shorter (approximately 40%) than the reviews in the Apple App Store. We also observed a notable skew in the length of reviews in both stores.

Influence of app category and downloads on number of reviews. In the Google Play Store

Finding 3. The number of downloads and releases correlated with the number of received reviews, whereas an app's category did not play a major role during the study period. On the other hand, Pagano and Maalej²⁰ and Hoon et al.¹¹ both reported a relation between an app's category and the number of received reviews; and

Implication. The relationship between number of received reviews and an app's category should be explored further, especially in light of the discrepancy between the two app stores.

Here, we investigate the effect of an app's number of downloads, number of releases, and app category on the number of received reviews. We built a regression model with an app's number of received reviews as the dependent variable. Due to the notable skew in the number of reviews, we log-transformed the number of reviews before building the linear-regression model.

Figure 2 plots the total number of reviews using the built-regression model. We included three plots, each keeping the median values of the other factors

Figure 2. Plots of the total number of reviews (logged) on the y -axis and three separate graphs of app categories, number of downloads, and number of releases on the x -axis; the graphs reflect the relation between the three factors and the total number of reviews.

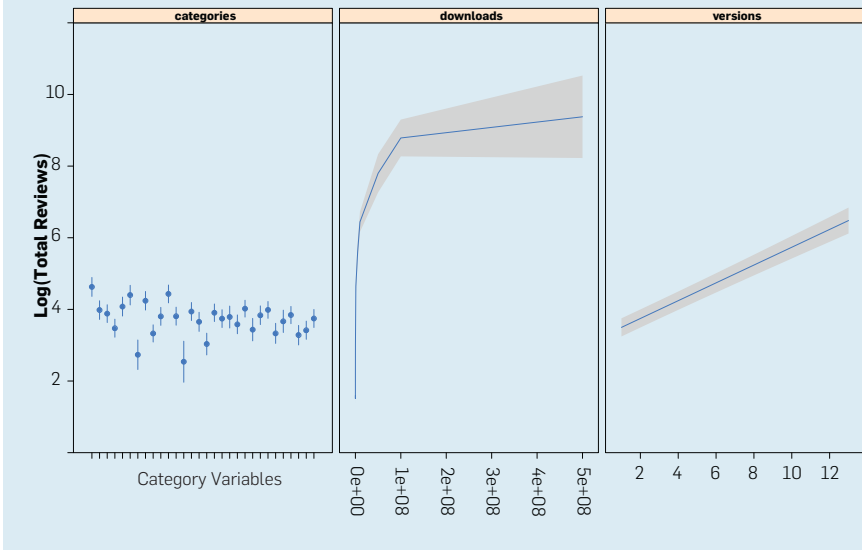


Figure 3. A nomograph of the effect of new releases, app category, and number of downloads on total number of reviews received.

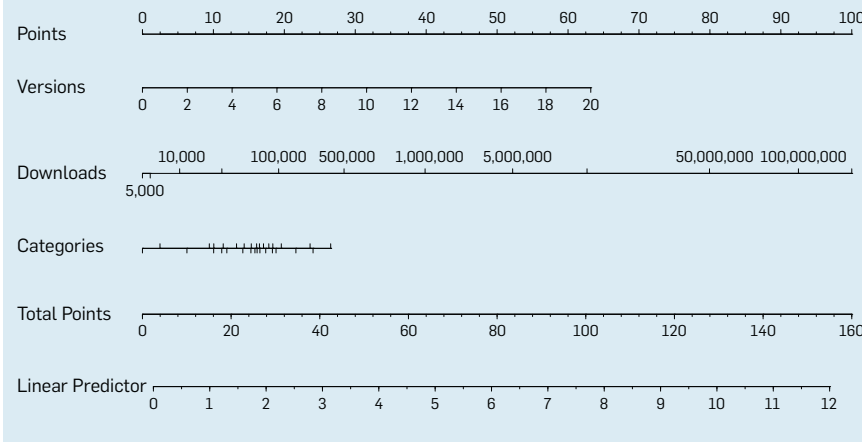
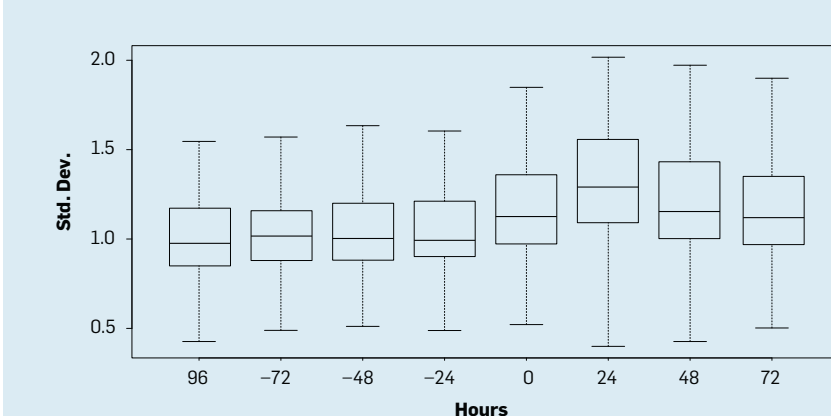


Figure 4. Standard deviation of new reviews every 24 hours before and after the first collected release for each studied app; each boxplot represents the standard deviation from the median number of reviews for each app at that time.



the same so we could see how each factor affects the total number of reviews.¹⁰ The gray bands around the plotted lines are bootstrap confidence intervals for our estimates.

We generated a nomogram (see Figure 3) to visualize the results of our regression model,¹⁰ helping us examine the effect of each factor while controlling for other factors. The nomogram consists of a series of scales. The Linear Predictor scale is the total number of reviews in log scale. To calculate the total number of reviews, we can draw a straight line from the value of the “total points” scale to the linear predictor scale. The total points are calculated by summing the points of each of the scales of the three factors: releases, downloads, and categories. To calculate the points value of each factor, we can draw a line from the value in the factor scale to the points scale. The value in the points scale becomes the points for that factor. For example, releases = 2, downloads = 100,000, and categories = tools. We found that 2-releases corresponded to approximately seven points, 100,000-downloads corresponded to approximately 20 points, and the tools category corresponded to approximately five points. The sum was 32 total points, which corresponded to approximately 2.5 log scale, or 316 total user reviews.

We found that as the number of downloads and releases increased, the total number of reviews also increased. We found no relation between individual categories (such as communications, social, tools, and review count) when we controlled for number of downloads and releases. In contrast, Pagano and Maalej²⁰ and Hoon et al.¹¹ observed a relation between categories and number of received reviews in the Apple App Store; however, neither study controlled for the other metrics in its analysis. Those studies observed a relation between categories and number of reviews that may be due to the interaction between categories and number of downloads or between categories and number of releases.

Spike in reviews following a release. Finally, concerning the spike in reviews following a release of an app in Google Play Store

Finding 4. Both the Google Play store and the Apple App Store showed evi-

dence of a spike in reviews following a release; and

Implication. Greater effort examining user reviews should follow a release in order to improve app quality.

Pagano and Maalej²⁰ reported that the number of received reviews decreased over time after a release, suggesting releases contribute to new reviews. We observed the same kind of correlation for the Google Play store. Figure 4 outlines a boxplot of the median number of reviews for each studied app across each of its releases, showing a spike in reviews directly on and after an app's release day.

However, still not clear is if these spikes were due to an app attracting new users following its release or to current users becoming more inclined to review the app. Looking closer at our nomogram, we note that many releases (more than 20) for an app has as much of an effect as an app with 10 million downloads. Frequent releases thus ensure an app's user base is more engaged as it begins providing feedback.

Conclusion

A very small percentage of the top apps we studied (0.19% of 10,713) have ever received more than 500 reviews per day, yet most studied apps received only a few reviews per day. The number of received reviews for the studied apps did not vary due to the category to which the app belonged, varying instead based on number of downloads and releases. Some of our results highlight differences between the Google Play store and the Apple App Store.

Additional studies are needed to better understand the review dynamics across both stores. Researchers should thus examine whether other empirical findings hold across them. In particular, techniques designed to assist mobile-app developers should be optimized for each store. 

References

1. App Annie Analytics; <http://www.appannie.com/app-store-analytics/>
2. Applause; <https://www.applause.com/testing/>
3. Chandy, R. and Gu, H. Identifying spam in the iOS app store. In *Proceedings of the Second Joint WICOW/AIRWeb Workshop on Web Quality* (Lyon, France, Apr. 16). ACM Press, New York, 2012, 56–59.
4. Chen, N., Lin, J., Hoi, S.C.H., Xiao, X., and Zhang, B. AR-Miner: Mining informative reviews for developers from the mobile app marketplace. In *Proceedings of the 36th International Conference on Software Engineering* (Hyderabad, India, May 31–June 7). ACM Press, New York, 2014, 767–778.
5. Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., and Sadeh,

Frequent releases ensure an app's user base is more engaged as it begins providing feedback.

- N. Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, IL, Aug. 11–14). ACM Press, New York, 2013, 1276–1284.
6. Galvis Carreño, L.V. and Winblad, K. Analysis of user comments: An approach for software requirements evolution. In *Proceedings of the 2013 International Conference on Software Engineering* (San Francisco, CA, May 18–26). IEEE Press, Piscataway, NJ, 2013, 582–591.
7. Google Analytics; <http://www.google.ca/analytics/mobile/>
8. Guzman, E. and Maalej, W. How do users like this feature? A fine-grained sentiment analysis of app reviews. In *Proceedings of the 22nd International Requirements Engineering Conference* (Karlskrona, Sweden, Aug. 25–29). IEEE Press, Piscataway, NJ, 2014, 153–162.
9. Harman, M., Jia, Y., and Zhang, Y. App store mining and analysis: MSR for app stores. In *Proceedings of the Ninth Working Conference on Mining Software Repositories* (Zurich, Switzerland, June 2–3). Piscataway, NJ, 2012.
10. Harrell, F.E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York, 2001.
11. Hoon, L., Vasa, R., Schneider, J.-G., Grundy, J. et al. *An Analysis of the Mobile App Review Landscape: Trends and Implications*. Technical Report. Swinburne University of Technology, Faculty of Information and Communication Technologies, Melbourne, Australia, 2013.
12. Jacob, C. and Harrison, R. Retrieving and analyzing mobile apps feature requests from online reviews. In *Proceedings of the 10th International Workshop on Mining Software Repositories* (San Francisco, CA, May 18–19). IEEE Press, Piscataway, NJ, 2013, 41–44.
13. Johns, T. *Replying to User Reviews on Google Play*. Android Developers Blog, June 21, 2012; <http://android-developers.blogspot.ca/2012/06/replying-to-user-reviews-on-google-play.html>
14. Khalid, H., Shihab, E., Nagappan, M., and Hassan, A. What do mobile app users complain about? *IEEE Software* 32, 3 (May–June 2015), 70–77.
15. Kim, H.-W., Lee, H.L., and Son, J.E. An exploratory study on the determinants of smartphone app purchase. In *Proceedings of the 11th International DSI Decision Sciences Institute and 16th APDSI Asia Pacific Region of Decision Sciences Institute Joint Meeting* (Taipei, Taiwan, July 12–16, 2011).
16. Lim, S.L., Bentley, P.J., Kanakam, N., Ishikawa, F., and Honiden, S. Investigating country differences in mobile app user behavior and challenges for software engineering. *IEEE Transactions on Software Engineering* 41, 1 (Jan. 2015), 40–64.
17. Martin, W., Harman, M., Jia, Y., Sarro, F., and Zhang, Y. The app-sampling problem for app store mining. In *Proceedings of the 12th Working Conference on Mining Software Repositories* (Florence, Italy, May 16–17). IEEE Press, Piscataway, NJ, 2015.
18. Mudambi, S.M. and Schu, D. What makes a helpful online review? A study of customer reviews on Amazon.com. *MIS Quarterly* 34, 1 (2010), 185–200.
19. Pagano, D. and Bruegge, B. User involvement in software evolution practice: A case study. In *Proceedings of the 2013 International Conference on Software Engineering* (San Francisco, May 18–26). IEEE Press, Piscataway, NJ, 2013, 953–962.
20. Pagano, D. and Maalej, W. User feedback in the App Store: An empirical study. In *Proceedings of the 21st IEEE International Requirements Engineering Conference* (Rio de Janeiro, Brazil, July 15–19). IEEE, Piscataway, NJ, 2013.

Stuart McIlroy (mcilroy@cs.queensu.ca) is a Ph.D. student at Dalhousie University, Halifax, Canada.

Weiyei Shang (shang@encs.concordia.ca) is an assistant professor and Concordia University Research Chair in Ultra-Large-Scale Systems in the Department of Computer Science and Software Engineering at Concordia University, Montreal, Canada.

Nasir Ali (cnali@memphis.edu) is an assistant research professor at the University of Memphis, Memphis, TN.

Ahmed E. Hassan (ahmed@cs.queensu.ca) is Canada Research Chair in Software Analytics and NSERC/ BlackBerry Software Engineering Chair in the School of Computing at Queen's University, Kingston, Canada.

Healthcare robotics can provide health and wellness support to billions of people.

BY LAUREL D. RIEK

Healthcare Robotics

THE USE OF robots in healthcare represents an exciting opportunity to help a large number of people. Robots can be used to enable people with cognitive, sensory, and motor impairments, help people who are ill or injured, support caregivers, and aid the clinical workforce. This article highlights several recent advancements on these fronts, and discusses their impact on stakeholders. It also outlines several key technological, logistical, and design challenges faced in healthcare robot adoption, and suggests possible avenues for overcoming them.

Robots are “physically embodied systems capable of enacting physical change in the world.” They enact this change with effectors, which can move the robot (locomotion), or objects in the environment (manipulation). Robots typically use sensor data to make decisions. They can vary in their degree of autonomy, from fully autonomous to fully teleoperated, though most modern systems have mixed initiative, or shared autonomy. More broadly, robotics technology includes affiliated systems, such as related sensors, algorithms for processing data, and so on.²⁸

There have been many recent exciting examples of robotics technology, such as autonomous vehicles, package delivery drones, and robots that work side-by-side with skilled human workers in factories. One of the most exciting areas where robotics has a tremendous potential to make an impact in our daily lives is in healthcare.

An estimated 20% of the world's population experience difficulties with physical, cognitive, or sensory functioning, mental health, or behavioral health. These experiences may be temporary or permanent, acute or chronic, and may change throughout one's lifespan. Of these individuals, 190 million experience severe difficulties with activities of daily living tasks (ADL).^a These include physical tasks (basic ADLs), such as grooming, feeding, and mobility, to cognitive functioning tasks (instrumental ADLs), which include goal-directed tasks such as problem solving, finance management, and housekeeping.¹⁴ The world also has a rapidly aging population, who will only add to this large number of people who may need ADL help. Of all of these individuals, few want to live in a long-term care facility. Instead,

a World Bank; <http://documents.worldbank.org/curated/en/2011/01/14440066/world-report-disability>

» key insights

- **Over 20% of the world's population experience physical, cognitive, or sensory impairments. Robots can fill care gaps and support independence.**
- **Robots can help caregivers and the clinical workforce, who are overloaded and experience high rates of injury themselves.**
- **In health, most problems are open-ended, and there is no “one-size-fits-all” solution. Every person, task, and care setting are different, and require robots to be able to robustly learn and adapt on the fly.**
- **Technologists, researchers, providers, and end users must closely collaborate to ensure successful adoption.**

At a nursing residence in Florence, Italy, a robot performs caregiving and support duties for 20 elderly guests. The robot was developed through the Robot-Era project supported by the European Union.



Figure 1. The main stakeholders for healthcare robotics, and exemplar contextualizations of their relationship to the technology.

Stakeholder	Context for Robotics	Examples of Robotics Use
Primary Stakeholders:		
Direct Robot Users (DRU): People who directly use robots to aid them with daily living or wellness activities. This may include people who experience difficulties with physical, cognitive, or sensory functions, mental health, or behavioral health. These experiences may be temporary or permanent, acute or chronic, and may change throughout one's lifespan.	A DRU may directly use robotics technology to help them accomplish daily living activities, with physical, cognitive, or social tasks.	▶ A person with a lower limb amputation uses a robotic arm to grasp objects ▶ A person with autism works with a robot to learn to read facial expressions ▶ A person who has low vision uses a smart cane to sense obstacles
Clinicians (CL): Persons who may provide healthcare or work with DRU. These individuals may be: nurses, physicians, mental healthcare providers, rehabilitation professionals, pharmacists, EMTs, among others.	A CL may use robotics technology while providing care, in the course of their training, or to help them with day-to-day administrative tasks.	▶ A therapist employs a therapeutic robotic pet in a treatment regimen ▶ A nurse uses a robot to help lift a DRU from their wheelchair to a bed ▶ A surgeon uses a robot to aid with a minimally invasive procedure ▶ A medical student uses a robotic patient simulator to learn how to treat a stroke
Care Givers (CG): Family members, neighbors, volunteers, or other unpaid persons who may support DRU.	A CG may use robotics technology to directly or indirectly support a DRU	▶ An adult child uses a telepresence robot to communicate with an older parent ▶ A friend may use a robot to perform household tasks in the DRU's home
Secondary Stakeholders:		
Robot Makers (RM): Individuals who design, build, program, instrument, or research robotics technology.	A RM may work with DRU, CL, CG, PM, and ESW to perform their work.	▶ A company builds a hospital discharge robot ▶ A student writes sensing algorithms for a robot to lift people out of a wheelchair ▶ A Maker club adapts toys to be accessible by children with motor impairments
Environmental Service Workers (ESW): Persons who provide secondary care to DRUs by helping prevent the spread of infection through cleaning services. These can include environmental service workers in hospitals, housekeeping staff in nursing homes, and so on.	An ESW may use robotics technology to ensure care environments are safe and sanitary to help prevent the spread of infection. Their use of robotics directly affects DRU's quality of care, and CL's workplace safety.	▶ An ESW teleoperates a disinfecting robot which emits UV light to kill superbugs in a hospital room ▶ An ESW uses a waste removal robot to safely transport medical waste
Health Administrators (HA): Individuals who provide leadership to a care setting by planning, coordinating, and directing care delivery.	An HA may purchase robots to support staff, patients, or visitors, or set policy on their usage.	▶ A chief medical officer reviews clinical effectiveness data of a rehabilitation robot ▶ A HA preforms a cost effectiveness study of acquiring robots for their institution
Tertiary Stakeholders:		
Policy Makers (PM): People who work for or with federal, state, and local governments to design policy regarding: how robots will be used, which robots will be used, and how their costs will be managed.	A PM may work with DRU, CL, CG, ESW, RM, and AG to understand how to best craft policy for the use of robots.	▶ A Federal Food and Drug Administration (FDA) worker establishes new policy for Home Use Devices ▶ A Federal Trade Commission (FTC) worker sets privacy policies for robot sensors
Insurers (IC): Public or private organizations who makes decisions about benefits to DRU and CG, including service payments to CL and RM.	ICs may work with PM, AG, HA, RM, and CL to establish guidelines for reimbursable robot-related services.	▶ An IC worker explores the robotic exoskeletons evidence base to establish reimbursement policy ▶ An IC worker consults with a company to understand a robot's control system
Advocacy Groups (AG): Organizations who work on behalf of DRU populations	AGs may work with DRU, CL, CG, RM, PM, and others to ensure robots are employed in ways that are of the best interest of their DRU population.	▶ An muscular dystrophy AG supports new research on exoskeletons ▶ An MS advocacy group lobbies congress to fund new robotic therapies

many people would prefer to live and age gracefully in their homes for as long as possible, independently and with dignity.²² However, for people requiring help with ADL tasks, this goal is challenging to meet for a few reasons. First, this level of care is quite expensive; in the U.S. it costs between

\$30,000 and \$85,000 per year in provider wages alone.^b

Second, there is a substantial health-care labor shortage—there are far more

people who need care than healthcare workers available to provide it.³³ While family members and friends attempt to fill these care gaps, they too have full-time jobs and other familial obligations, and thus cannot meet the need. Health-care workers are not only overburdened by this labor shortage, but face increas-

^b U.S. Department of Health and Human Services; <http://longtermcare.gov/costs-how-to-pay/costs-of-care/>

ingly hazardous work environments, and are themselves at great risk of debilitating injury and disability. According to the National Institute for Occupational Health and Safety (NIOSH), health care workers have the most hazardous industrial jobs in America, with the greatest number of nonfatal occupational injuries and illness.^c

Thus, there is an incredible opportunity for robotics technology to help fill care gaps and help aid healthcare workers. In both the research and commercial space, robotics technology has been used for physical and cognitive rehabilitation, surgery, telemedicine, drug delivery, and patient management. Robots have been used across a range of environments, including hospitals, clinics, homes, schools, and nursing homes; and in both urban and rural areas.

Before discussing these applications, it is important to first contextualize the use of robots within healthcare. This article begins by identifying who will be providing, receiving, and supporting care, where this care will take place, and key tasks for robots within these settings. Examples of new technologies aimed at supporting these stakeholders will be introduced, and key challenges and opportunities to realizing the potential use of robots in healthcare that research and industry are encouraged to consider, will be addressed. These adoption issues include a robot's capability and function (Does a robot have the required capabilities to perform its function?), cost effectiveness (What is the robot's value to stakeholders relative to its cost?), clinical effectiveness (Has the robot been shown to have a benefit to stakeholders?), usability and acceptability (How easy is the robot to use, modify, and maintain? Is the robot's form and function acceptable?), and safety and reliability (How safe and reliable is the robot?)

Stakeholders, Care Settings, and Robot Tasks

Stakeholders. For this article, stakeholders are defined as people who have a vested interest in the use of robotics technology in healthcare. Stakeholders can be: people who directly

use robots to provide assistance with daily living or wellness activities (direct robot users (DRU)), health professionals who use robots to provide care (clinicians (CL)), non-CL individuals who support DRUs (care givers (CG)), technologists and researchers (robot makers (RM)), health administrators (HAS), policy-makers (PMs), advocacy groups (AGs), and insurers (IC). Figure 1 introduces these stakeholders.

These stakeholders can be grouped into three beneficiary groups: *Primary beneficiaries*: direct robot users, clini-

cians, and caregivers, all of whom are likely to use robotics technology on a regular basis; *Secondary beneficiaries*: health administrators, robot makers, and environmental service workers, all of whom are involved in the use of robotics technology in healthcare settings but do not directly use the robots to use robots to support the health and wellness of DRUs; and *tertiary beneficiaries*: policymakers and advocacy groups, who have interest in the use of robots to provide care to their constituents, but are unlikely to use them directly.

Selected care settings where robots may be used.

Care Setting	Definition
Longer-Term	
Assistive Living Facility	"Congregate residential facility with self-contained living units providing assessment of each resident's needs and on-site support 24 hours a day, 7 days a week, with the capacity to deliver or arrange for services including some health care and other services."
Group Home	"A residence, with shared living areas, where clients receive supervision and other services such as social and/or behavioral services, custodial service, and minimal services (e.g., medication administration)."
Custodial Care Facility	"A facility which provides room, board and other personal assistance services, generally on a long-term basis, and which does not include a medical component"
Nursing Facility	"A facility which primarily provides to residents skilled nursing care and related services for the rehabilitation of injured, disabled, or sick persons, or, on a regular basis, health-related care services above the level of custodial care to other than [people with intellectual disabilities]"
Home Care	"Location, other than a hospital or other facility, where [a person] receives care in a private residence."
Shorter-Term	
Inpatient Hospital	"A facility, other than psychiatric, which primarily provides diagnostic, therapeutic (both surgical and nonsurgical), and rehabilitation services by, or under, the supervision of physicians to patients admitted for a variety of medical conditions."
On/Off Campus Outpatient Hospital	"A portion of a... hospital provider based department which provides diagnostic, therapeutic (both surgical and nonsurgical), and rehabilitation services to sick or injured persons who do not require hospitalization or institutionalization."
Urgent Care Facility	"Location, distinct from a hospital emergency room, an office, or a clinic, whose purpose is to diagnose and treat illness or injury for unscheduled, ambulatory patients seeking immediate medical attention."
Inpatient Psychiatric Facility	"A facility that provides inpatient psychiatric services for the diagnosis and treatment of mental [health disorders] on a 24-hour basis, by or under the supervision of a physician."
Hospice	"A facility, other than a patient's home, in which palliative and supportive care for terminally ill patients and their families are provided."
Substance Abuse Treatment Facility	"A location which provides treatment for substance (alcohol and drug) abuse on an ambulatory basis. Services include individual and group therapy and counseling, family counseling, laboratory tests, drugs and supplies, and psychological testing." Residential facilities also provide room and board.

Source: http://www.cms.gov/Medicare/Coding/place-of-service-codes/Place_of_Service_Code_Set.html

c National Institute for Occupational Safety and Health, <http://www.cdc.gov/niosh/topics/healthcare/>

This article will focus on primary beneficiaries; however, it is important to note that all other stakeholder groups are critical to the successful end-deployment of robotics in healthcare, and should be included when possible in decision-making.

Care settings. Another critical dimension to contextualizing the use of robotics in healthcare is to consider the location of use. This can significantly impact on how suitable different technologies are for a given setting,¹² and can affect the design of a robot and its required capabilities. For example, while a 400-lb, 5'4" dual-arm mobile manipulator may work well in a lab, it is ill-suited to an 80-sq. ft. room in an assisted living facility. While it is understandable robot makers may immediately be more concerned with achieving platform functionality than the particulars of care settings, to successfully deploy healthcare robots, setting must be considered.

The accompanying table defines different kinds of care settings, and includes longer-term care facilities in the community, as well as shorter-term care facilities, such as hospitals. For longer-term care in the U.S., the Fair Housing Act, and Americans with Disabilities Act set some general guidelines for living space accessibility; however, the majority of space guidelines is state-dependent, and can have a large degree of variation. For example, an assisted living facility in Florida must provide 35-sq. ft. per resident for living and dining, whereas in Utah it is 100-sq. ft. An in-patient psychiatric facility in Kentucky must provide 30-sq. ft. per patient in social common areas, Oregon requires 120-sq. ft. in total and 40-sq. ft. per patient.

Robots in healthcare can also affect the well-being, health, and safety of both direct robot users and clinicians. The field of evidence based healthcare design⁴⁰ has produced hundreds of studies showing a relationship between the built environment and health and wellness, in areas including patient safety, patient outcomes, and staff outcomes. When new technology such as a robot becomes part of a care setting, it is now a possible disruptor to health. HAs must balance the risks and benefits for adopting new technology, and robot makers should be aware of

these tradeoffs in how they design and test their systems.

Care tasks. Robots may be helpful for many health tasks. Robots can provide both physical and cognitive task support for both DRUs and clinicians/caregivers, and may be effective and helping reduce cognitive load. Task assistance is particularly critical as the demand for healthcare services is far outpacing available resources, which places great strain on clinicians and caregivers.³³

Physical tasks. *Clinicians.* Tasks involving the "3Ds" of robotics—dirty, dangerous, and dull—can be of particular value for clinical staff. Clinicians spend an inordinate amount of time on "non-value added" tasks, for example, time away from treating patients. The overburden of these tasks creates a climate for error; so robots, which can help clinicians effectively, surmount these challenges would be a boon. Some of these non-value added tasks include: Transportation, such as moving materials or people from one place to another, Inventory, such as patients waiting to be discharged, Search Time, such as looking for equipment or paperwork, Waiting, for patients, materials, staff, medications, and Overburdening of Staff and Equipment, such as during peak surge times in hospitals.⁴²

Two of the best tasks for robots in this task space are material transportation and scheduling, which robots can be exceptionally skilled at given the right parameters. For example, robots that can fetch supplies, remove waste, and clean rooms. Another task robots can do that will help greatly improve the workplace for clinicians is moving patients. This is a very hazardous task—hospital workers, home health workers, and ambulance workers experience musculoskeletal injuries between three and five times the national average when moving patients according to NIOSH.

Robots can also help clinicians with other dangerous tasks, such as helping treat patients with highly infectious diseases. Robot mediated treatment has become particularly pertinent after the recent Ebola outbreak, where clinicians and caregivers can perform treatment tasks via telepresence robots.¹⁷

Finally, robots may help extend the physical capabilities of clinicians. For

example, in surgical procedures, robots may provide clinicians with the ability to perform less invasive procedures to areas of the body inaccessible with existing instrumentation due to issue or distance constraints. These can include types of neurological, gastric, and fetal surgical procedures.⁴¹

Direct robot users. When designing robots for DRUs, there is great value in designing straightforward solutions to problems. At a recent workshop discussing healthcare robotics, people with Amyotrophic Lateral Sclerosis (ALS) and other conditions reported that most of all they just wanted "a robot to change the oil."³⁰ In other words: help is most needed with basic, physical ADL tasks, such as dressing, eating, ambulating, toileting, and housework. Robots that can help people avoid falling could also be incredibly beneficial, as falls cause thousands of fatal and debilitating injuries per year.

Currently, standalone robots that can successfully perform the majority of these key physical ADL tasks are a long way from reaching the consumer market. There are several reasons for this. First, the majority of these tasks remain challenging for today's robots, as they require a high degree of manual dexterity, sensing capability, prior task knowledge, and learning capability. Furthermore, most autonomous, proximate robots move extremely slowly due to safety and computational purposes, which will undoubtedly be frustrating for end users. Finally, even if robots could perform some of these more complex ADL tasks, their power budgets may make them impractical for deployment in most care settings.

However, there have been substantial gains in recent years for other tasks. For example, robots that provide DRUs with additional physical reach (for example, smart on-body prostheses, wheelchair mounted robot arms) and robots which provide multi-setting mobility capability (for example, exoskeletons, accessible personal transportation devices).²⁶ These are likely to continue to be the types of systems that reach end users first for the foreseeable future.

Cognitive tasks. *Clinicians.* Any technology that can effectively reduce clinical workload is likely to be warm-

ly embraced in healthcare. Many of these systems exist in a non-embodied fashion, for example, decision support tools to aid in emergency medicine,¹² patient logistical management, or charting. However, robotic systems may have a place within this domain, particularly if a robot is well integrated into existing workflow and able to access EHR data. For example, perhaps a medication management robot could anticipate a clinician's "next move" in treatment by prefetching a likely medicine from the pharmacy. Or perhaps a robot could deliver personalized messages to family members in waiting rooms to update them on the status of their relative while clinicians are occupied with other tasks.

Another area where robotics has been extensively used to aid clinicians with cognitive tasks is in clinical simulation and training. Robotic patient simulators are life-sized, humanoid robots that can breathe, bleed, speak, expel fluids, and respond to medications. They are the most commonly used humanoid robot worldwide, and provide learners with the ability to simultaneously practice both procedural and communication skills.^{22,23} These robots are used by inter-professional clinicians across a wide range of specialties, including acute care, perioperative care, trauma, and mental healthcare. The author and her students have been designing the next generation of these simulators, which can convey realistic facial patient pathologies, such as pain, stroke, and cerebral palsy, and are integrated with on-board physiological models.^{23,29}

Direct robot users and care givers. The ways in which robots may be able to provide cognitive task support to CGs has yet to be fully realized. However, similar to clinicians, the ability to reduce cognitive load would be greatly welcomed. CGs in particular are often overburdened when providing care; they frequently have other family members to care for, other jobs, and their own lives (and health) to manage.³ Robots might be able to cognitively support CGs by learning and anticipating their needs, prefetching items, attending to time-intensive tasks which detract from care, and so on.

For DRUs, robotics technology might be able to help facilitate inde-



There is incredible opportunity for robotics technology to help fill care gaps and aid healthcare workers. Robotics have been used for physical and cognitive rehabilitation, surgery, telemedicine, drug delivery, and patient management.



pendence by providing sensory augmentation or substitution. For example, DRUs who are blind or low vision may benefit from a robotic way finding tool, or DRUs using robotic prostheses might receive sensory feedback from a robotic finger in their shoulder.

Robots also may be able to help DRUs with regaining (or supplementing) cognitive function in neurorehabilitative settings, such as in cases of stroke, post-traumatic stress disorder, or traumatic brain injury. Robots also may provide socio-emotional support to DRUs: to provide companionship, teach people with autism to learn to read emotions, or to help reduce symptoms of dementia. However, there is a paucity of clinical effectiveness trials showing DRU benefit compared to standard treatment, so it is unclear what the future for these robots may be.²⁹

Recent Advances in Healthcare Robotics

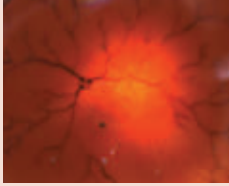
The 2016 U.S. Robotics Roadmap was recently released,¹ which frames the state of the art in robotics and future research directions in the field. Over 150 robotics researchers contributed, including the author of this article. The roadmap includes a detailed summary of advancements in robotics relating to health and wellness. Some key focus areas include: aging and quality of life improvement, surgical and interventional robotics, rehabilitative robotics, and clinical workforce support.

In general, robots used in these areas can be divided into three categories: inside the body, on the body, and outside the body. Those inside and on the body are primarily intended for direct robot users, and those outside the body for direct robot users, care givers, and clinicians. These robots have the potential to be used across a range of care settings and clinical foci, and can provide both physical and cognitive support.

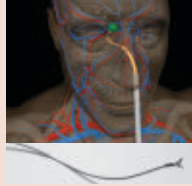
Inside the body. Recent advances for internal robots have occurred in the fields of microrobotics, surgical robotics, and interventional robotics. Microrobotics are micro-scale, untethered devices that can move through the body and can perform a range of functions, such as targeted therapy (that is, localized delivery of medicine or energy), material removal (for exam-

Figure 2. Key examples of recent advances in healthcare robotics. Those inside and on the body are primarily intended for direct robot users, and those outside the body for direct robot users, caregivers, and clinicians. These robots have the potential to be used across a range of care settings and clinical foci, and can provide both physical and cognitive support. Image credits (clockwise from upper left): B. Nelson, R. Alterovitz, Mobius, TED, Ekso Bionics, B. Smart, L. Riek, S. Sabanovic, C. Kemp.

In the body



Microrobots are micro-scale, untethered robots that can move through the body and can perform targeted therapy, material removal, structural control, and sensing.



Concentric tube robots (active cannulas) can be used as small, teleoperated manipulators or as steerable needles, and enable procedures in areas inaccessible with traditional instruments.

On the body



Robotic prostheses and exoskeletons. People with forearm-to-shoulder amputations can use wearable robot prostheses, which can provide fine-grained dexterity, reach, and strength. People with lower-limb amputations or lower-body muscle weakness can use powered-knee and ankle prostheses to do everything from running marathons to dancing. Exoskeletons have helped people with muscle weakness, movement disorders, or paralysis locomote.

Outside the body



Mobile manipulators. Clinicians can safely tele-operate mobile robots to treat patients with highly infectious diseases such as Ebola Virus Disease.



Patient simulators. Over 180,000 clinicians annually train on high fidelity robotic patient simulators, which can simulate physiological cues, and sense and respond to learners.



Mental and Behavioral Healthcare. Robots can support people with cognitive impairments, facilitate neurorehabilitation, support wellness, or provide companionship.



Physical task support. Robots can support people with motor impairments, movement disorders, and brain injuries to provide external manipulation capabilities.

ple, biopsy, ablation), structural control (for example, stent placement), and sensing (for example, determining oxygen concentrations, sensing the presence of cancer).²⁵ Recent advances in the field have enabled actuating, powering, and controlling these robots (see Nelson et al.²⁵ for a review.)

In surgical and interventional robotics, a range of advances have been made that enable clinicians to have improved dexterity and visualization inside the body and reduce the degree of movement during operations.¹ Furthermore, promising advances have been made in concentric tube (active cannula) robots. These robots are comprised of precurved, concentrically nested tubes that can bend and twist throughout the body. The robots can be used as small, teleoperated manipulators or as steerable needles. The robots can enter the body directly, such as through the skin or via a body opening, or could be used via an endoscope.¹¹

Some future research directions for in-the-body robots include new means

for intuitive physical and cognitive interaction between the user and robot, new methods for managing uncertainty, and providing 3D registration in real time while traversing both deformable and non-deformable tissue.¹

On the body. In terms of wearable robots for DRUs, there have been recent advances in the areas of actuated robot prostheses, orthoses, and exoskeletons. A prosthesis supplants a person's missing limb, and acts in series with a residual limb. An orthosis is a device that helps someone who has an intact limb but an impairment, and an exoskeleton provides either a person with intact limbs (DRU or otherwise) assistance or enhancement of existing physical capability. Orthoses and exoskeletons act in parallel to an existing limb.³⁹

All of these robots can be used to enable DRUs to perform tasks. For example, people with forearm-to-shoulder amputations can use wearable robot prostheses, which can provide dexterity, reach, and strength. Peo-

ple with lower-limb amputations or lower-body muscle weakness can use powered-knee and ankle prostheses to engage in a range of activities, including everyday locomotion to running marathons and dancing. Exoskeletons have helped people with muscle weakness, movement disorders, and paralysis locomote.

Several advances have been made recently in how people interface with these robots. For example, some robot prostheses offer neural integration to provide tactile feedback and increasingly more intuitive control of the limb.¹ Other advances include an increase in the workspace and range of motion of wearable robots, as well as improvements in user comfort.

Outside the body. Robots outside the body are being used across many clinical application spaces. For clinicians, mobile manipulators are being used to help treat patients with highly infectious diseases,¹⁷ aid in remote surgical procedures,²⁶ and help provide physical assistance to CLs when moving pa-

tients.²⁴ They are also used extensively in clinical training, as discussed earlier.

Robots are also being explored in mental and behavioral healthcare applications. Robots are being used to support people with autism spectrum disorder and cognitive impairments, to encourage wellness, and to provide companionship.

(See Riek²⁹ for a detailed review of these applications).

For physical task support, robots can provide external manipulation and sensing capabilities to DRUs. For example, wheelchair mounted robot arms can provide reach, smart wheelchairs can help facilitate safe navigation and control, and telepresence robot surrogates can enable people with severe motor impairments the ability to fly, give TED talks, and make coffee.^{6,7,38}

There are other examples of external robots that are outside the scope of this paper, but could prove highly pertinent in healthcare. For example, autonomous vehicles may provide new opportunities for DRUs to locomote, or may enable EMTs to focus on treating patients rather than driving ambulances. Telepresence may also have unforeseen applications in healthcare, such as through aerial manipulation, drone delivery of medical supplies, among others.

Healthcare Robotics Adoption: Challenges and Opportunities

While there are exciting advances in healthcare robotics, it is important to carefully consider some of the challenges inherent in healthcare robotics, and discuss ways to overcome them. Robots have the ability to enact physical change in the world, but in healthcare that world is inherently safety critical, populated by people who may be particularly vulnerable to harm due to their disability, disorder, injury, or illness. Stakeholders face five major considerations when considering deploying robots in healthcare: Usability and acceptability, safety and reliability, capability and function, clinical effectiveness, and cost effectiveness. Each is explored here.

Usability and acceptability. Robots that are difficult for primary stakeholders to use have a high likelihood of being abandoned. This phenomenon has been well documented in the Assistive Technology Community.^{5,9,20} For example, a 2010 study reported that as many

as 75% of hand rehabilitation robots were never actually tested with end users, rendering them completely unusable in practice and abandoned.²

One of the major challenges is that clinicians, even those who are well-educated and accomplished in their disciplines, often have low technology literacy levels.¹⁹ Thus, if they themselves find a robot unusable, the likelihood of them successfully training a direct robot user or caregiver to use the robot is greatly diminished.

Another challenge is that DRUs are often excluded from the robot design process, which leads to unusable and unsuitable technology. Robots with multiple degrees of freedom, such as wearable prostheses or wheelchair-mounted arms, require a high level of cognitive function to control.³⁸ However, many people needing such robots often have co-morbidities (that is, other conditions), which can make control a further exhausting process.

There are several ways to address this issue. One approach is for robot makers to reduce robot complexity. Balasubramanian et al.² argue for functional simplicity in therapeutic robot design, which will lead to robots that are easier for all primary stakeholders to use, control, and maintain. This concept is echoed in much of the reliability and fault tolerance literature; lower-complexity robots are more likely to be longitudinally reliable and fault tolerant.

Forlizzi and Zimmerman propose the idea of a service-centered design process, wherein rather than only think about a single user and a system, designers consider including the broader ecosystem surrounding a technology.¹⁰ This is a particularly beneficial idea in healthcare robotics. Rarely will there be one DRU and one robot; rather, there is a complex social structure surrounding caregiving that should be considered carefully in robot design.

Another important barrier to healthcare robot adoption is its acceptability. The morphology, behavior, and functionality of a robot play a major role in its adoption and use. When a DRU uses a robot in public, they are immediately calling attention to their disability, disorder, or illness. DRUs already face significant societal stigma, so frequently

avoid using anything which further advertises their differences, even if it provides a health benefit.^{27,32,33}

Shinohara and Wobbrock argue that in addition to designers considering the functional accessibility of system, they also consider its social accessibility, and employ a “Design for Social Acceptance” (DSA) approach.³⁵ This means going beyond purely functional designs, which may be “awkward and clunky.”³⁴ Robot makers are usually primarily concerned about a robot’s functional capabilities; for example, can the robot perform its task safely and reliably given workspace, environmental, and platform constraints. However, the aforementioned literature suggests that there may be great value in also considering a robot’s appearance and behavior to help enable technology adoption.

Safety and reliability. When robots and people are proximately located, safety and reliability are incredibly important. This is even more critical for DRUs who may rely extensively on robots to help them accomplish physical or cognitive tasks, and who may not have the same ability to recover from robot failures as easily as non-DRUs.

There has been a fair bit of work on safe physical human-robot interaction, particularly with regard to improving collision avoidance, passive compliance control methods, and new advances in soft robotics to facilitate gentle interaction.³⁷ There also have been recent advances on algorithmic verifiability for robots operating in partially unknown workspaces,¹⁸ which may prove fruitful in the future.

However, there has been little work to date on safe cognitive human-robot interaction. People with cognitive disabilities and children are particularly prone to being deceived by robots.²⁹ This is an important and under-explored question in the robotics community, though a few efforts have been made recently with regard to encouraging robot makers to employ value-centered design principles. For example, ensuring the appearance of the robot is well-aligned with its function (for example, avoiding false-advertising), enabling transparency into how a robot makes decisions, and maintaining the privacy and dignity of DRUs.^{15,31}

Another way to help bridge the

safety gap is for robot makers to employ in-depth testing and training regimens that enable direct robot users, care givers, and clinicians to fully explore the capabilities of a platform. This can help prevent people from either over-relying or under-relying on the robot, and help facilitate trust.

Capability and function. The field of robotics has seen amazing capability gains in recent years, some of which have been instrumental in healthcare. However, despite these advances, robotics is still an exceptionally difficult problem. For example, many demonstrations in robotics technology remain demos, and fail outside of highly constrained situations.⁸ This is particularly problematic when designing technology for healthcare: most problems are open-ended, and there is no “one-size-fits-all” solution.^{12,28} Every person, task, and care setting are different, and require robots to be able to robustly learn and adapt on the fly.

As discussed previously, care settings differ substantially. Even the same type of care setting, such as an emergency department or assisted living facility, have substantial differences in their environment, practices, and culture. In our prior work designing health information technology, we have demonstrated that these differences can be surmounted by conducting multi-institutional trials, and by building solutions that are adaptable to different care settings.¹³ The same approach can be taken in robotics.

Real-world, real-time, robust perception in human environments is another major challenge in robotics. While the field of computer vision has seen advances in solving still-image, fixed-camera recognition problems, those same algorithms perform poorly when both the cameras and people are moving, data is lost, sensors are occluded, and there is clutter in the environment. However, these situations are highly likely in human social settings, and it is an open challenge to sense, respond to, and learn from end users in these settings.²⁸ There have been some recent advances, however: the fields of social signal processing and human-robot interaction have moved toward multimodal sensing



Robotic patient simulators are life-sized, humanoid robots that can breathe, bleed, speak, expel fluids, and respond to medications.



approaches, which help enable more robust algorithms. Furthermore, life-long learning and longitudinal experimental approaches have also enabled researchers to surmount some of these perceptual challenges. Modeling situational context and object and environmental affordances within them can also be a useful tool in surmounting these issues.^{1,28}

Learning, too, is a challenge. It is critical that primary stakeholders, who have a wide range of physical abilities, cognitive abilities, and technology literacy levels, are able to easily repurpose or reprogram a robot without a RM present. This level of adaptability and accessibility presents robot makers with a complex technical and socio-technical challenge. As mentioned previously, simple is undoubtedly better; it helps constrain the problem space and lowers the complexity of the system. Another major aid will be the research community continuing to develop new datasets, evaluation metrics, and common platforms;⁸ these have shown to be useful in other computing domains, so are likely to be helpful here.

Cost effectiveness. When robots are being acquired in healthcare, it is important that their cost effectiveness is considered beyond the purchase, maintenance, and training costs for the system. For example, when electronic health records (EHRs) were first employed in hospitals, they were touted as a means to save clinicians and patients' time. However, because EHR systems were so poorly designed, difficult to use, and poorly integrated into existing they ended up creating substantially more non-value added work. This resulted in “unintended consequences,” including increasing costs and patient harm.¹⁶ It is critical these same pitfalls are avoided for robots.

The Agency for Healthcare Research and Quality (AHRQ) created a guide for reducing these unintended consequences for EHRs;¹⁶ the same methodological approach can be employed for robots. For example, when assessing the acquisition and deployment of a robot in a first place:

► *Are you ready for a robot (and is a robot ready for you)?* HAS must carefully consider their institution's robot readiness. Robots may solve

some problems, but may make others worse. For example, suppose a supply-fetching robot that is purchased help nurses save time. However, it has difficulties functioning at high volume times of day due to sensor occlusion, so supply deliveries end up being delayed. This causes a cascade effect, increasing the workload of nurses. Situations like these can be remedied through a careful exploration of existing workflow in a unit, and by fully understanding a robot's existing capabilities and limitations. See Gonzales et al.^{12,13} for examples on engaging in this process with clinicians in safety critical settings.

► *Why do you want a robot?* It is important stakeholders define exactly why a robot is necessary for a given task in the first place. What are the goals of the stakeholders? What is the plan for deploying the robot, and how will success be measured? These questions can also be explored through design activities while assessing workflow and institutional readiness.

► *How do you select a robot?* As mentioned previously, functionality is only one aspect to a robot; there is also: usability, acceptability, safety, reliability, and clinical effectiveness. While there are not yet definitive guidelines to aid HAS in this process, science policy is starting to be shaped within this space. The CCC recently held an event entitled "Discovery and Innovation in Smart and Pervasive Health,"^d which brought together over 60 researchers from across academia, industry, and government, many of whom are roboticists who work in health. These efforts will hopefully begin to provide guidelines in the future.

► What are the recommended practices for avoiding unintended consequences of robot deployment? Successfully deploying robots is a difficult process that may result in a disruptive care setting, and upset key stakeholders. To avoid unintended consequences, it is important that:

► The robot's scope is well-defined with clear goals;

► Key stakeholders are included and engaged in the deployment from the onset;

► Detailed deployment plans are provided but are not overly complicated;

► There are multiple ways to collect, analyze, and act on feedback from users;

► Success metrics should be determined in advance and evaluated continually; and,

► Quality improvement should be supported on an ongoing basis.

Recently, the IEEE released a document on "Ethically Aligned Design" which contains detailed suggestions for how to engage in this value-centered practice in engineering, which could be helpful for all stakeholders moving forward.^e

Clinical effectiveness. Clinical effectiveness answers the question: "Does it work?" In particular, does a given intervention provide benefit to a primary stakeholder? This question is answered by conducting thorough, evidence-based science. For robots directly affecting DRUs, this evidence comes from comparative effectiveness research (CER), which is "generated from research studies that compare drugs, medical devices, tests, surgeries, or ways to deliver healthcare."^f

CER can include both new clini-

cal studies on effectiveness, or can synthesize the existing literature in a systematic review.

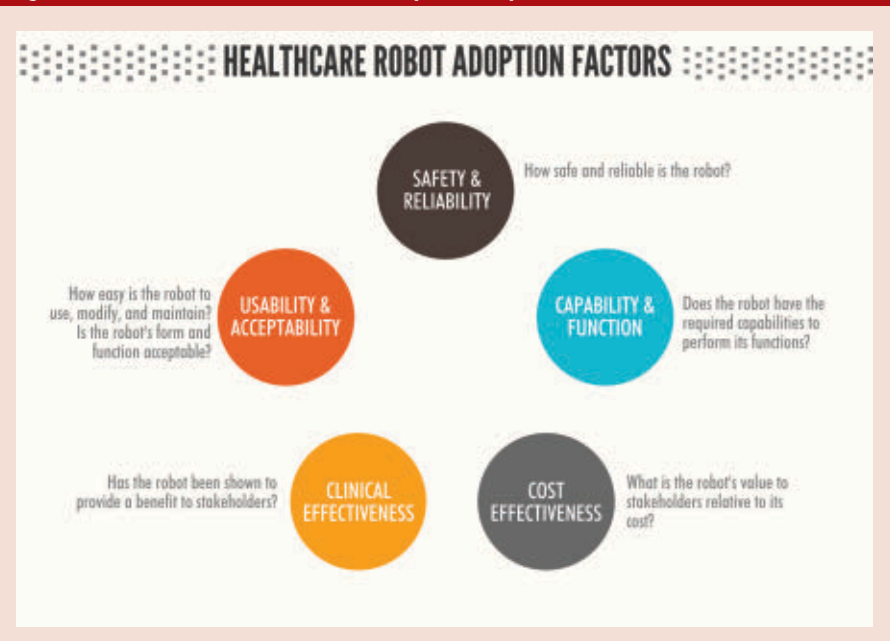
All consumer in-the-body robots and many on-the-body robots must undergo regulatory approval before they can be marketed and sold. In the United States, this approval is through the FDA, which typically requires a strong level of evidence showing the effectiveness and safety of a medical device. Outside-the-body robots typically do not need to undergo a device review process provided they fall within existing classifications; for example, Paro the robot seal (see Figure 2, bottom right) is classified by the FDA a neurological therapeutic device, and thus is exempt from premarket review. Shimshaw et al.³⁶ argue this lack of regulation of healthcare robots may be harmful to stakeholders both physically and informationally, and should be subject to premarket review on dimensions including privacy, safety, reliability, and usability.

In the meanwhile, while the policy community races to catch up with technology, the robotics community can and should engage in research that tests the clinical effectiveness of robots across care settings. Begum et al.⁴ suggest robot makers follow existing clinical effectiveness benchmarks within their intended care space and adopt them for use with robots. Furthermore, Riek²⁹ suggests that when

e http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf

f Agency for Health and Research Quality, effectivehealthcare.ahrq.gov/index.cfm/what-is-comparative-effectiveness-research1/

Figure 3. Factors that will affect the widespread adoption of robotics in healthcare.



d <http://cra.org/ccc/events/discovery-innovation-smart-health/>

conducting CER with robots, particularly in cognitive support settings, it is not sufficient to simply test robot vs. no-robot, as the morphology can affect outcomes, but to instead to test actuated vs. non-actuated.

Discussion

Healthcare robotics is an exciting, emerging area that can benefit all stakeholders across a range of settings. There have been a number of exciting advances in robotics in recent years, which point to a fruitful future. How these robots ultimately will be integrated into the lives of primary beneficiaries remains unknown, but there is no doubt that robots will be a major enabler (and disruptor) to health.

It is critical that both the research and industrial communities work together to establish a strong evidence-base for healthcare robotics. As we have learned from the large-scale deployment of EHRs, technology development and deployment cannot happen in a vacuum, or it is likely to cause grave harm to DRUs, overwhelming stress to clinicians, and astronomical unseen costs. It is wise for all stakeholders to proceed cautiously and deliberately, and consider the full context of care as much as possible.

It is also critical that direct robot users remain directly involved in the research, development, and deployment of future robots in health and wellness across the entire lifecycle of a project, as ultimately they are the ones who will be using these robots. As discussed earlier, ignoring DRU input leads to unusable, unsuitable, and abandoned robots, which benefits no one. Secondary and Tertiary stakeholders should look to the Patient Centered Outcomes Research Institute (PCORI)⁸ as a highly successful model for how-to engage with primary stakeholders in clinical research and development.

Finally, it is important that robot makers work with DRUs to help bridge technology literacy gaps and appropriately set expectations. Most people's experience with robotics comes from movies or media, which rarely reflects the true state of affairs. Robots are quite fallible in the

real world, and will remain so for the foreseeable future; however, they still have the potential to be a remarkable game changer in health.

Acknowledgments

Some research reported in this article is based upon work supported by the National Science Foundation under Grant Nos. IIS-1253935 and SES-1457307, and the Luce Foundation. 

References

1. A roadmap for US robotics: From Internet to robotics (Nov. 2016); <http://jacobsschool.ucsd.edu/contextualrobotics/docs/rm3-final-rs.pdf>, November 2016.
2. Balasubramanian, S., Klein, J., and Burdett, E. Robot-assisted rehabilitation of hand function. *Curr Opin Neurol*, 2010.
3. Bastawrous, M. Caregiver burden—a critical discussion. *Int'l J of Nursing Studies* 50, 3 (2013), 431–441.
4. Begum, M., Serna, R.W., and Yanco, H.A. Are robots ready to deliver autism interventions? A comprehensive review. *International J. Social Robotics* 8, 2 (2016).
5. Brose, S.W., Weber, D.J., Salatin, B.A., Grindle, G.G., Wang, H., et al. The role of assistive robotics in the lives of persons with disability. *Am J Phys Med*, 2010.
6. Carlson, T. and Demiris, Y. Collaborative control for a robotic wheelchair: evaluation of performance, attention, and workload. *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)* 42, 3 (2012), 876–888.
7. Chen, T.L. et al. Robots for humanity: using assistive robotics to empower people with disabilities. *IEEE Robotics & Automation* 20, 1 (2013), 30–39.
8. Christensen, H.I., Okamura, A., Mataric, M., Kumar, V., Hager, G., and Choset, H. Next generation robotics (2016); *arXiv preprint arXiv:1606.09205*.
9. Dawe, M. Desperately seeking simplicity: how young adults with cognitive disabilities and their families adopt assistive technologies. In *Proceedings of the Conference on Human Factors in Computing Systems*, 2006.
10. Forlizzi, J. and Zimmerman, J. Promoting service design as a core practice in interaction design. In *Proceedings of the 5th IASDR World Conference on Design Research*, 2013.
11. Gilbert, H.B., Rucker, D.C., and Webster III, R.J. Concentric tube robots: The state of the art and future directions. *Robotics Research*. Springer, 2016, 253–269.
12. Gonzales, M.J., Cheung, V.C., and Riek, L.D. Designing collaborative healthcare technology for the acute care workflow. In *Proceedings of the 9th Int'l Conference on Pervasive Computing Technologies for Healthcare*, 2015.
13. Gonzales, M.J., Henry, J.M., Calhoun, A.W., and Riek, L.D. Visual task: A collaborative cognitive aid for acute care resuscitation. In *Proceedings of the 10th Int'l Conference on Pervasive Computing Technologies for Healthcare*, 2016.
14. Graf, C. The Lawton instrumental activities of daily living scale. *The American J. Nursing* 108, 4 (2008).
15. Hartzog, W. Unfair and deceptive robots. *Maryland Law Review* 74, 785 (2015).
16. Jones, S.S. et al. Guide to reducing unintended consequences of electronic health records. *Agency for Healthcare Research and Quality*, 2011.
17. Kraft, K. and Smart, W.D. Seeing is comforting: effects of teleoperator visibility in robot-mediated health care. *The Proceedings of the 11th ACM/IEEE International Conference on Human Robot Interaction*, 2016, 11–18.
18. Lahijanian, M., Maly, M.R., Fried, D., Kavrakli, L.E., Kress-Gazit, H., and Vardi, M.Y. Iterative temporal planning in uncertain environments with partial satisfaction guarantees. *IEEE Trans. Robotics*, 2016.
19. Lluch, M. Healthcare professionals' organizational barriers to health information technologies—A literature review. *International J. Medical Informatics*, 2011.
20. Lu, E.C. et al. Development of a robotic device for upper limb stroke rehabilitation: A user-centered design approach. *Paladyn* 2, 4 (2011), 176–184.
21. Milligan, C. *There's no place like home: Place and care in an ageing society*. Ashgate Publishing Ltd., 2012.
22. Moosaei, M., Das, S.K., Popa, D.O., and Riek, L.D. Using facially expressive robots to calibrate clinical pain perception. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 32–41.
23. Moosaei, M., Gonzales, M.J., and Riek, L.D. Naturalistic pain synthesis for virtual patients. *International Conference on Intelligent Virtual Agents*, 2014.
24. Mukai, T., Hirano, S., Nakashima, H., Kato, Y., Sakaida, Y., et al. Development of a nursing-care assistant robot RIBA that can lift a human in its arms. *IEEE Intelligent Robots and Systems*, 2010.
25. Nelson, B.J., Kalliakatos, I.K., and Abbott, J.J. Microrobots for minimally invasive medicine. *Annual Review of Biomedical Engineering* 12 (2010), 55–85.
26. Okamura, A.M., Mataric, M.J., and Christensen, H.I. Medical and health-care robotics. *Robotics and Automation* 17, 3 (2010), 26–27.
27. Parette, P. and Scherer, M. Assistive technology use and stigma. *Education and Training in Developmental Disabilities*, 2004, 217–226.
28. Riek, L.D. The social co-robotics problem space: Six key challenges. Robotics challenges and vision. In *Proceedings of the Workshop at Robotics: Science and Systems*, 2013.
29. Riek, L.D. Robotics technology in mental health care. *Artificial Intelligence in Behavioral and Mental Health Care*. D. Luxton, (ed). Academic Press, 2015.
30. Riek, L.D., Hartzog, W., Howard, D.A., Moon, A., and Calo, R. The emerging policy and ethics of human robot interaction. *HRI (Extended Abstracts)*, 2015.
31. Riek, L.D. and Howard, D. A code of ethics for the human-robot interaction profession. In *Proceedings of We Robot*, 2014.
32. Riek, L.D. and Robinson, P. Using robots to help people habituate to visible disabilities. In *IEEE International Conference on Rehabilitation Robotics*, 2011.
33. Shi, L. and Singh, D. A. *Delivering health care in America*. Jones & Bartlett Learning, 2014.
34. Shinohara, K. A new approach for the design of assistive technologies: Design for social acceptance. *ACM SIGACCESS Accessibility and Computing*, 2012.
35. Shinohara, K. and Wobbrock, J.O. In the shadow of misperception: Assistive technology use and social interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.
36. Simshaw, D., Terry, N., Hauser, K., and Cummings, M. Regulating healthcare robots: Maximizing opportunities while minimizing risks. *Richmond J. of Law & Tech*, 2016.
37. Trivedi, D., Rahn, C.D., Kier, W.M., and Walker, I.D. Soft robotics: Biological inspiration, state of the art, and future research. *Applied Bionics and Biomechanics*, 2008.
38. Tsui, K.M., Kim, D.J., Behal, A., Kontal, D., and Yanco, H. A. 'I want that': Human-in-the-loop control of a wheelchair-mounted robotic arm. *Applied Bionics and Biomechanics* 8, 1 (2011), 127–147.
39. Tucker, M.R. et al. Control strategies for active lower extremity prosthetics and orthotics: a review. *J. of Neuroengineering and Rehabilitation*, 2015.
40. Ulrich R.S. et al. A review of the research literature on evidence-based healthcare design. *Health Environments Research & Design J.*, 2008.
41. Webster R.J., Okamura, A.M., and Cowan, N.J. Toward active cannulas: Miniature snake-like surgical robots. *IEEE/RSJ Intelligent Robots and Systems*. IEEE, 2006.
42. Wellman, J., Jeffries, H., and Hagan, P. *Leading the Lean Healthcare Journey: Driving Culture Change to Increase Value*. CRC Press, 2016.

Laurel D. Riek (lriek@ucsd.edu) is an associate professor of computer science and engineering at the University of California, San Diego. She directs the Healthcare Robotics lab and builds autonomous robots that can sense, understand, and learn from real people in the real world.

© 2017 ACM 0001-0782/17/11 \$15.00



Watch the author discuss her work in this exclusive Communications video. <https://cacm.acm.org/videos/healthcare-robotics>

research highlights

P. 80

**Technical
Perspective**
**Solving Imperfect
Information Games**

By David Silver

P. 81

Heads-Up Limit Hold'em Poker Is Solved

By Michael Bowling, Neil Burch,
Michael Johanson, and Oskari Tammelin

P. 89

**Technical
Perspective**
**Exploring
a Kingdom by
Geodesic Measures**

By Marc Alexa

P. 90

The Heat Method for Distance Computation

By Keenan Crane, Clarisse Weischedel, and Max Wardetzky

Technical Perspective

Solving Imperfect Information Games

By David Silver

THE STUDY OF GAMES is as old as computer science itself. Babbage, Turing, and Shannon devised algorithms and hardware to play the game of chess. Game theory began with questions regarding optimal strategies in card games and chess, later developed into a formal system by von Neumann. Chess subsequently became the *drosophila*—or common fruitfly, the most studied organism in genetics—of artificial intelligence research. Early successes in chess and other games shaped the emerging field of AI: many planning algorithms first used in games became pillars of subsequent research; reinforcement learning was first developed for a checkers playing program; and the performance of game-playing programs has frequently been used to measure progress in AI.

Most of this research focused on perfect information games, in which all events are observed by all players, culminating in programs that beat human world champions in checkers, chess, Othello, backgammon, and most recently, Go. However, many applications in the real world have *imperfect* information: each agent observes different events. This leads to the possibility of deception and a wealth of social strategies. Imperfect information games provide a microcosm of these social interactions, while abstracting away the messiness of the real world.

Among imperfect information games, Poker is the most widely studied—the latest *drosophila*—due to its enormous popularity and strategic depth. The smallest competitively played variant by humans, and the most widely played by computers, is the two-player game known as Heads-Up Limit Hold’Em (HULHE), in which each player holds two private cards in addition to five public cards. Two decades of research in this game has led to powerful methods, such as counter-

The methods used to solve poker are quite general, and therefore have potential applications far beyond this one game.

factual regret minimization (CFR), for approximating a Nash equilibrium. Several years ago, a program called Polaris—created by many of the authors of the following paper—defeated for the first time a human professional poker player in HULHE.

However, Polaris was still far from perfect; indeed, it turns out in retrospect that it was exploitable, due to the approximations it made, by a very large margin. The obvious remaining question was whether a “near-perfect” solution could be found—a strategy so close to a Nash equilibrium that it cannot be differentiated in a lifetime of play.

The following paper takes the CFR methods used in previous work to the next level. Using a number of innovations—and several hundred machine-years of computation—they were able to find a near-perfect solution to HULHE. Their solution also provides insights into the game itself, showing the dealer holds a significant advantage, and that seemingly poor hands should be played surprisingly often.

For the game of poker, the next step beyond HULHE is no-limit poker, which has a much larger action space. This too has recently been cracked, with the programs Libratus (from CMU) and DeepStack (also from Alber-

ta) both defeating human professionals using variants of CFR—although a near-perfect solution remains out of reach. The final challenge would be the variant most widely played by humans: multiplayer no-limit poker.

The methods used to solve poker are quite general, and therefore have potential applications far beyond this one game. Many other imperfect information games played by humans, including a wide variety of card games, board games, and video games, are tractable to these methods. Furthermore, there are many real-world applications, such as auctions, negotiations, and security, in which agents receive different information, and must make a sequence of decisions to maximize a final pay-off—and therefore belong to the same class of imperfect information games as HULHE.

Solving a problem attains perfection in one domain. The frontier of solved domains is an incontrovertible measure of current computer capabilities. That frontier has now been extended by one significant step, to include for the first time a challenging imperfect information game. 

David Silver leads the reinforcement learning research group at Google DeepMind, London, and is lead researcher on AlphaGo.

Copyright held by author.

Heads-Up Limit Hold'em Poker Is Solved

By Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin

Abstract

Poker is a family of games that exhibit imperfect information, where players do not have full knowledge of past events. While many perfect information games have been solved (e.g., Connect-Four and checkers), no nontrivial imperfect information game played competitively by humans has previously been solved. In this paper, we announce that the smallest variant of poker in-play, heads-up limit Texas hold'em, is now essentially weakly solved. Furthermore, this computation formally proves the common wisdom that the dealer in the game holds a significant advantage. This result was enabled by a new algorithm, CFR⁺, which is capable of solving extensive-form games three orders of magnitude larger than previously possible. This paper is an extended version of the original 2015 *Science* article,⁹ with additional results showing Cepheus' in-game performance against computer and human opponents.

1. INTRODUCTION

Games have been intertwined with the earliest developments in computation, game theory, and Artificial Intelligence (AI). At the very conception of computing, Babbage had detailed plans for an “automaton” capable of playing tic-tac-toe and dreamt of his Analytical Engine playing chess.⁴ Both Alan Turing⁴⁶ and Claude Shannon,⁴⁰ on paper and in hardware respectively, developed programs to play chess as validation of early ideas in computation and AI. For over a half-century, games have continued to act as testbeds for new ideas and the resulting successes have marked significant milestones in the progress of AI: For example, the checkers-playing computer program Chinook becoming the first to win a world championship title against humans,³⁸ Deep Blue defeating Kasparov in chess,¹⁴ and Watson defeating Jennings and Rutter on *Jeopardy!*¹⁷ However, defeating top human players is not the same as “solving” a game, that is, computing a game-theoretically optimal solution that is incapable of losing against any opponent in a fair game. Solving games has also served as notable milestones for the advancement of AI, for example, Connect-Four² and checkers.³⁹

Every nontrivial game played competitively by humans that has been solved to-date is a *perfect information game*.^a

^a We use the word trivial to describe a game that can be solved without the use of a machine. The one near-exception to this claim is oshi-zumo, but it is not played competitively by humans and is a simultaneous-move game that otherwise has perfect information.¹³ Furthermore, almost all nontrivial games played by humans that have been solved to-date also have no chance elements. The one notable exception is hypergammon, a three-checker variant of backgammon invented by Hugh Sconyers in 1993 which he then strongly solved, that is, the game-theoretic value is known for all board positions. It has seen play in human competitions. See <http://www.bkgm.com/variants/HyperBackgammon.html> (accessed July 4, 2014).

In perfect information games, all players are informed of everything that has occurred in the game prior to making a decision. Chess, checkers, and backgammon are examples of perfect information games. In *imperfect information games*, players do not always have full knowledge of past events (e.g., cards dealt to other players in bridge and poker, or a seller's knowledge of the value of an item in an auction). These games are more challenging, with theory, computational algorithms, and instances of solved games lagging behind results in the perfect information setting.^b And, while perfect information may be a common property of parlor games, it is far less common in real-world decision making settings. In a conversation recounted by Bronkowski, John von Neumann, the founder of modern game theory, made the same observation, “Real life is not like that. Real life consists of bluffing, of little tactics of deception, of asking yourself what is the other man going to think I mean to do. And that is what games are about in my theory.”¹²

Von Neumann's statement hints at the quintessential game of imperfect information: the game of poker. Poker involves each player being dealt private cards, with players taking structured turns making bets on having the strongest hand (possibly bluffing), calling opponent bets, or folding to give up the hand. Poker played an important role in the early developments of the field of game theory. Borel⁷ and von Neumann's^{47, 48} foundational works were motivated by developing a mathematical rationale for bluffing in poker, and small synthetic poker games^c were commonplace in many early papers.^{7, 29, 32, 48} Poker is also arguably the most popular card game in the world with over 150mn players worldwide.¹ The most popular variant of poker today is Texas hold'em. When it is played with just two-players (heads-up) and with fixed bet-sizes and number of raises (limit), it is called Heads-Up Limit Hold'Em (HULHE).¹⁰ HULHE was popularized by a series

^b For example, Zermelo proved the solvability of finite, two-player, zero-sum, perfect information games in 1913,⁵¹ while von Neuman's more general minimax theorem appeared in 1928.⁴⁷ Minimax and alpha-beta pruning, the fundamental computational algorithm for perfect information games, was developed in the 1950s, while Koller and Megiddo's first polynomial-time technique for imperfect information games was introduced in 1992.²⁶

^c We use the word synthetic to describe a game that was invented for the purpose of being studied or solved rather than played by humans. A synthetic game may be trivial, such as Kuhn poker,²⁹ or nontrivial such as Rhode Island hold'em.⁴¹

The original version of this paper was published in *Science* 347, 6218 (Jan. 2015) 145–149. Adapted with permission from AAAS.

of high-stakes games chronicled in the book *The Professor, the Banker, and the Suicide King*.¹⁶ It is also the smallest variant of poker played competitively by humans. HULHE has 3.16×10^{17} possible states the game can reach making it larger than Connect Four and smaller than checkers. However, as an imperfect information game, many of these states cannot be distinguished by the acting player as they involve information about unseen past events (i.e., private cards dealt to the opponent). As a result, the game has 3.19×10^{14} decision points where a player is required to make a decision.

While smaller than checkers, the imperfect information nature of HULHE makes it a far more challenging game for computers to play or solve. It was 17 years after Chinook won its first game against world champion Marion Tinsley in checkers that the computer program Polaris won the first meaningful match against professional poker players.³⁴ While Schaeffer et al. solved checkers in 2007,³⁹ heads-up limit Texas hold'em poker, until now, was unsolved. This slow progress is not for lack of effort. Poker has been a challenge problem for artificial intelligence, operations research, and psychology with work going back over 40 years.⁶ 17 years ago, Koller and Pfeffer²⁸ declared, “we are nowhere close to being able to solve huge games such as full-scale poker, and it is unlikely that we will ever be able to do so.” The focus on HULHE as one example of “full-scale poker” began in earnest over ten years ago,⁵ and became the focus of dozens of research groups and hobbyists after 2006 when it became the inaugural event in the Annual Computer Poker Competition,⁵³ held in conjunction with the main conference of the Association for the Advancement of Artificial Intelligence (AAAI). This paper is the culmination of this sustained research effort toward solving a “full-scale” poker game.¹⁰

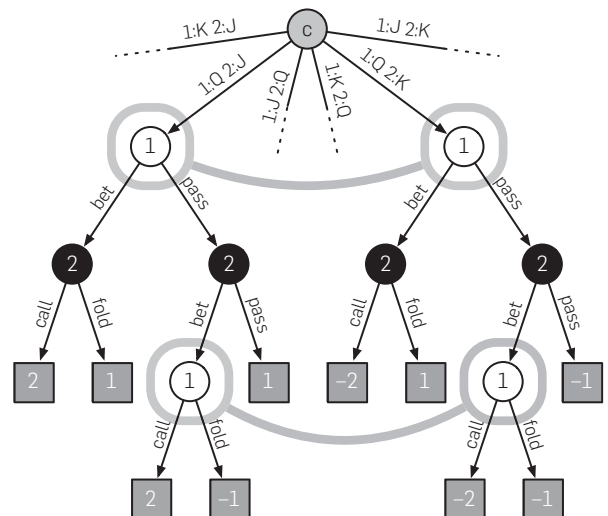
Allis³ gives three different definitions for solving a game. A game is said to be *ultra-weakly solved* if for the initial position(s), the game-theoretic value has been determined; *weakly solved* if for the initial position(s), a strategy has been determined to obtain at least the game-theoretic value, for both players, under reasonable resources; and *strongly solved* if for all legal positions, a strategy has been determined to obtain the game-theoretic value of the position, for both players, under reasonable resources. In an imperfect information game, where the game-theoretic value of a position beyond the initial position is not unique, Allis's notion of “strongly solved” is not well-defined. Furthermore, imperfect information games, due to stochasticity in the players' strategies or the game itself, typically have game-theoretic values that are real-valued rather than discretely valued (such as “win,” “loss,” and “draw” in chess and checkers), and only achieved in expectation over many playing of the game. As a result, game-theoretic values are often approximated, and so an additional consideration in solving a game is the degree of approximation in a solution. A natural level of approximation under which a game is *essentially weakly solved* is if a human lifetime of play is not sufficient to establish with statistical significance that the strategy is not an exact solution.

In this paper, we announce that heads-up limit Texas hold'em poker is essentially weakly solved. Furthermore, we bound the game-theoretic value of the game, proving that the game is a winning game for the dealer.

2. SOLVING IMPERFECT INFORMATION GAMES

The classical representation for an imperfect information setting is the *extensive-form game*. Here the word “game” refers to a formal model of interaction between self-interested agents and applies to both recreational games and serious endeavors such as auctions, negotiation, and security. See Figure 1 for a graphical depiction of a portion of a simple poker game in extensive-form. The core of an extensive-form game is a *game tree* specifying branches of possible events, namely player actions or chance outcomes. The branches of the tree split at *game states* and each is associated with one of the players (or chance) who is responsible for determining the result of that event. The leaves of the tree signify the end of the game, and have an associated utility for each player. The states associated with a player are partitioned into *information sets*, which are sets of states which the acting player cannot distinguish between (e.g., corresponding to states where the opponent was dealt different private cards). The branches from states within an information set are the player's available *actions*. A *strategy* for a player specifies for each information set a probability distribution over the

Figure 1. Portion of the extensive-form game representation of three-card Kuhn poker²⁹ where player 1 is dealt a queen (Q) and the opponent is given either the Jack (J) or King (K). Game states are circles labeled by the player acting at each state (“c” refers to chance, which randomly chooses the initial deal). The arrows show the events the acting player can choose from, labeled with their in-game meaning. The leaves are square vertices labeled with the associated utility for player 1 (player 2's utility is the negation of player 1's). The states connected by thick gray lines are part of the same information set, that is, player 1 cannot distinguish between the states in each pair since they represent a different unobserved card being dealt to the opponent. Player 2's states are also in information sets, containing other states not pictured in this diagram.



available actions. If the game has exactly two players and the utilities at every leaf sum to zero, the game is called *zero-sum*.

The classical solution concept for games is a *Nash equilibrium*, a strategy for each player such that no player can increase their expected utility by unilaterally choosing a different strategy. All finite extensive-form games have at least one Nash equilibrium. In zero-sum games, all equilibria have the same expected utilities for the players, and this value is called the *game-theoretic value of the game*. An ϵ -*Nash equilibrium* is a strategy for each player where no player can increase their utility by more than ϵ by choosing a different strategy. By Allis's categories, a zero-sum game is ultra-weakly solved if its game-theoretic value is computed, and weakly solved if a Nash equilibrium strategy is computed. We call a game essentially weakly solved if an ϵ -Nash equilibrium is computed for a sufficiently small ϵ to be statistically indistinguishable from zero in a human lifetime of played games. For perfect information games, solving typically involves a (partial) traversal of the game tree. However, the same techniques cannot apply to imperfect information settings. We briefly review the advances in solving imperfect information games, benchmarking the algorithms by their progress in solving increasingly larger synthetic poker games as summarized shown in Figure 2.

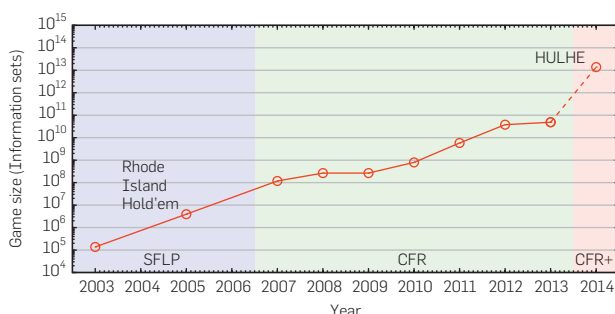
Normal-Form Linear Programming. The earliest method for solving extensive-form games involved converting it into a *normal-form game*, represented as a matrix of values for every pair of possible deterministic strategies in the original extensive-form game, and then solving it with a Linear Program (LP). Unfortunately, the number of possible deterministic strategies is exponential in the number information sets of the game. So, while LPs can handle normal-form games with many thousands of strategies, even just a few dozen decision points makes this method impractical. Kuhn poker, a poker game with three cards, one betting round, and a one bet maximum having a total of 12 information sets (see Figure 1), can be solved with this approach. But even Leduc hold'em,⁴² with six cards, two betting rounds, and a two bet maximum having a total of

only 288 information sets, is intractable having over 10^{86} possible deterministic strategies.

Sequence-Form Linear Programming. Romanovskii³⁵ and later Koller et al.^{26,27} established the modern era of solving imperfect information games, introducing the sequence-form representation of a strategy. With this simple change of variables, they showed that the extensive-form game could be solved directly as an LP, without the need for an exponential conversion to normal-form. Sequence-Form Linear Program (SFLP) was the first algorithm to solve imperfect information extensive-form games with computation time that grows as a polynomial of the size of the game representation. In 2003, Billings et al.⁵ applied this technique to poker, solving a set of simplifications of HULHE to build the first competitive poker-playing program. In 2005, Gilpin and Sandholm¹⁹ used the approach along with an automated technique for finding game symmetries to solve Rhode Island Hold'em,⁴¹ a synthetic poker game with 3.94×10^6 information sets after symmetries are removed.

Counterfactual Regret Minimization. In 2006, the Annual Computer Poker Competition was started.⁵³ The competition drove significant advancements in solving larger and larger games, with multiple techniques and refinements being proposed in the years that followed.^{36,37} One of the techniques to emerge, and currently the most widely adopted in the competition, is Counterfactual Regret Minimization (CFR).^d CFR is an iterative method for approximating a Nash equilibrium of an extensive-form game through the process of repeated self-play between two regret-minimizing algorithms.^{10,52} *Regret* is the loss in utility an algorithm suffers for not having selected the single best deterministic strategy, which can only be known in hindsight. A *regret-minimizing* algorithm is one that guarantees its regret grows sub-linearly over time, and so eventually achieves the same utility as the best deterministic strategy. The key insight of CFR is that instead of storing and minimizing regret for the exponential number of deterministic strategies, CFR stores and minimizes a modified regret for each information set and subsequent action, which can be used to form an upper bound on the regret for any deterministic strategy. An approximate Nash equilibrium is retrieved by averaging each player's strategies over all of the iterations, and the approximation improves as the number of iterations increases. The memory needed for the algorithm is linear in the number of information sets, rather than quadratic, which is the case for efficient LP methods.²⁵ Since solving large games is usually memory-bound, CFR has resulted in as dramatic an increase in the size of solved games as Koller et al.'s advance. Since its introduction in

Figure 2. Increasing sizes of imperfect information games solved over time. The shaded regions refer to the technique used to achieve the result with references in the main text. CFR⁺ is the algorithm used in this work and the dashed line shows the result established in this paper.



^d Another notable algorithm to emerge from the Annual Computer Poker Competition is an application of Nesterov's excessive gap technique³³ to solving extensive form games.¹⁸ The technique has some desirable properties, including better asymptotic time complexity than what is known for CFR. However, it has not seen widespread use among competition participants due to its lack of flexibility in incorporating sampling schemes and its inability to be used with powerful (but unsound) abstractions that employ imperfect recall. Recently, Waugh and Bagnell⁴⁹ have shown that CFR and the excessive gap technique are more alike than different, suggesting that the individual advantages of each approach may be attainable in the other.

2007, CFR has been used to solve increasingly complex simplifications of HULHE, reaching as many as 3.8×10^{10} information sets in 2012.²⁰

3. SOLVING HEADS-UP LIMIT HOLD'EM

The full game of HULHE has 3.19×10^{14} information sets. Even after removing game symmetries it has 1.38×10^{13} , that is, three orders of magnitude larger than previously solved games. There are two challenges for established CFR variants to handle games at this scale: memory and computation. During computation CFR must store the resulting solution and the accumulated regret values for each information set. Even with single-precision (four byte) floating point numbers, this requires 262TB of storage. Furthermore, past experience has shown that a three order of magnitude increase in the number of information sets requires at least three orders of magnitude more computation. In order to tackle these two challenges we employ two ideas recently proposed by Tammelin, a co-author of this paper.⁴⁴

To address the memory challenge we store the approximate solution strategy and accumulated regrets using compression. For the solution and regrets we use fixed-point arithmetic by first multiplying all values by a scaling factor and truncating them to integers. The resulting integers are then ordered to maximize compression efficiency, with compression ratios around 13-to-1. Overall, we require under 11TB of storage during the computation, which is distributed across a cluster of computation nodes. This amount is infeasible to store in main memory, and so we store the compressed strategy and regret values on each node's local disk. Each node is responsible for a set of *subgames*, that is, portions of the game tree partitioned based on publicly observed actions and cards so that each information set is associated with one subgame. The regrets and strategy for a subgame are loaded from disk, updated, and saved back to disk, using a streaming compression technique that decompresses and recompresses portions of the subgame as needed. By making the subgames large enough, the update-time dominates the total time to process a subgame. With disk pre-caching, the inefficiency incurred by disk storage is approximately 5% of the total time.

To address the computation challenge we use a variant of CFR called CFR⁺.^{10, 44} CFR implementations typically sample only portions of the game tree to update on each iteration. They also employ regret-matching at each information set, which maintains regrets for each action and chooses among actions with positive regret with probability proportional to that regret. Instead, CFR⁺ does exhaustive iterations over the entire game tree, and uses regret-matching⁺, a variant of regret-matching where regrets are constrained to be non-negative. Actions that have appeared poor (with less than zero regret for not having been played) will be chosen again immediately after proving useful (rather than waiting many iterations for the regret to become positive). Finally, in contrast with CFR, we have observed empirically that the exploitability of the players' strategies during the computation regularly converges to zero. Therefore, we skip the step of computing and storing the average strategy, instead using the players' current strategies as the CFR⁺ solution. We have

empirically observed CFR⁺ to require considerably less computation than state-of-the-art sampling CFR,²² while also being highly suitable for massive parallelization.

THEOREM 1.^e *Given a set of actions A , and any sequence of T value functions $v^t: A \mapsto \mathbb{R}$ with a bound L such that $|v^t(a) - v^t(b)| \leq L$ for all t and $a, b \in A$, an agent acting according to the regret-matching⁺ algorithm will have regret of at most $L\sqrt{|A|T}$.*

Like CFR, CFR⁺ is an iterative algorithm that computes successive approximations to a Nash equilibrium solution. The quality of the approximation can be measured by its *exploitability*: the amount less than the game value that the strategy achieves against the worst-case opponent strategy in expectation.¹⁰ Computing the exploitability of a strategy involves computing this worst-case value, traditionally requiring a traversal of the entire game tree. This was long thought to be intractable for games the size of HULHE. Recently it was shown that this calculation could be dramatically accelerated by exploiting the imperfect information structure of the game and regularities in the utilities.²³ This is the technique we use to confirm the approximation quality of our resulting strategy. The technique and implementation has been verified on small games and against independent calculations of the exploitability of simple strategies in HULHE.

A strategy can be exploitable in expectation and yet, due to chance elements in the game and randomization in the strategy, its worst-case opponent still is not guaranteed to be winning after any finite number of hands. We define a game to be *essentially solved* if a lifetime of play is unable to statistically differentiate it from being solved at 95% confidence. Imagine someone playing 200 hands of poker an hour for 12hrs a day without missing a day for 70 years. Furthermore imagine them employing the worst-case, maximally exploitive, opponent strategy, and never making a mistake. Their total winnings, as a sum of many millions of independent outcomes, would be normally distributed. Hence, the observed winnings in this lifetime of poker would be 1.64 standard deviations or more below its expected value (i.e., the strategy's exploitability) at least 1 time out of 20. Using the standard deviation of a single hand of HULHE, which has been reported to be around 5bb/g (big-blinds per game, where the big-blind is the unit of stakes in HULHE),¹¹ we arrive at a threshold of $1.64 * 5 / \sqrt{200 * 12 * 365 * 70} \approx 0.00105$. So, an approximate solution with an exploitability under 1mbb/g (milli-big-blinds per game) cannot be distinguished with high confidence from an exact solution, and indeed has a 1-in-20 chance of winning against its worst-case adversary even after a human lifetime of games. Hence, 1mbb/g is the threshold for declaring HULHE essentially solved.

4. THE SOLUTION

Our CFR⁺ implementation was executed on a cluster of 200 computation nodes each with 24 2.1GHz AMD cores, 32GB

^e Theorem 1 and others providing the theoretical support for CFR⁺ did not appear in the original version of this article, and were published in a subsequent paper.⁴⁵

of Random Access Memory (RAM), and a 1TB local disk. We divided the game into 110,565 subgames (partitioned based on preflop betting, flop cards, and flop betting). The subgames were split among 199 worker nodes, with one parent node responsible for the initial portion of the game-tree. The worker nodes performed their updates in parallel, passing values back to the parent node for it to perform its update, taking 61 min on average to complete one iteration. The computation was then run for 1,579 iterations, taking 68.5 days, and using a total of 900 core years of computation^f and 10.9TB of disk space, including file system overhead from the large number of files.

Figure 3 shows the exploitability of the computed strategy with increasing computation. The strategy reaches an exploitability of 0.986mbb/g, making HULHE essentially weakly solved. Using the separate exploitability values for each position (as the dealer and non-dealer) we get exact bounds on the game-theoretic value of the game: between 87.7mbb/g and 89.7mbb/g for the dealer, proving the common wisdom that the dealer holds a significant advantage in HULHE.

The final strategy, as a close approximation to a Nash equilibrium, can also answer some fundamental and long-debated questions about game-theoretically optimal play in HULHE. Figure 4 gives a glimpse of the final strategy in two early decisions of the game. Human players have disagreed about whether it may be desirable to “limp,” that is, call as the very first action rather than raise, with certain hands. Conventional wisdom is that limping forgoes the opportunity to provoke an immediate fold by the opponent, and so raising is preferred. Our solution emphatically agrees (see the absence of blue in Figure 4a). The strategy limps just 0.06% of the time and with no hand more than 0.5%. In other situations, the strategy gives insights beyond conventional wisdom, indicating areas where humans might improve. The strategy rarely “caps,” that is, makes the final allowed raise, in the first round as the dealer, whereas some strong human players cap the betting with a wide range of hands. Even when holding the strongest hand, a pair of aces, the

strategy caps the betting less than 0.01%, and the hand most likely to cap is a pair of twos, with probability 0.06%. Perhaps more importantly, the strategy chooses to play, that is, not fold, a broader range of hands as the non-dealer than most human players (see the relatively small amount of red in Figure 4b). It is also much more likely to re-raise when holding a low-rank pair (such as threes or fours).^g

While these observations are only for one example of game-theoretically optimal play (different Nash equilibria may play differently), they both confirm as well as contradict current human beliefs about equilibria play, and illustrate that humans can learn considerably from such large-scale game-theoretic reasoning.

5. IN-GAME RESULTS

In this extended version of the original paper,⁹ we present additional results measuring Cepheus’ in-game performance against computer agents and human opponents. HULHE has served as a common testbed for artificial intelligence research for more than a decade, and researchers have produced a long series of computer agents for the domain. This effort was largely coordinated by the Annual Computer Poker Competition (ACPC) which began in 2006 with HULHE. While each year’s top agents outperformed the older agents in the competition, and so appeared to be converging to optimal play, their actual worst-case exploitability was unknown. In 2011, an efficient best response technique was developed that made it feasible to measure a computer agent’s exploitability,²³ and for the first time researchers were able to exactly measure their progress towards the goal of solving the game. A key result in that paper was that top ACPC agents only defeated each other by tiny margins, and yet had a wide range of exploitability. Using Cepheus, we can now also evaluate these historical agents through matches against an essentially optimal strategy.

^g These insights were the result of discussions with Mr. Bryce Paradis, previously a professional poker player who specialized in HULHE.

^f The total time and number of core years is larger than was strictly necessary as it includes computation of an average strategy that was later measured to be more exploitable than the current strategy and so discarded. The total space noted, on the other hand, is without storing the average strategy.

Figure 3. Exploitability of the approximate solution with increasing computation.

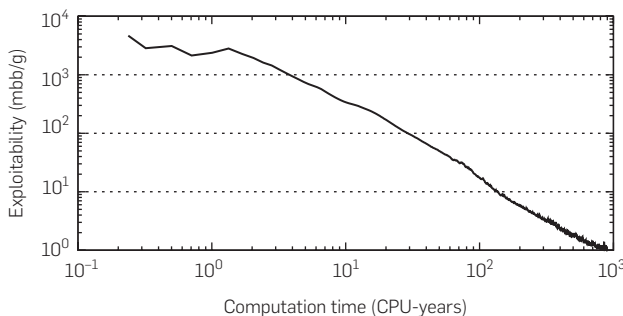


Figure 4. Action probabilities in the solution strategy for two early decisions. Each cell represents one of the possible 169 hands (i.e., two private cards) with the upper diagonal consisting of cards with the same suit and the lower diagonal consisting of cards of different suits. The color of the cell represents the action taken: red for fold, blue for call, and green for raise, with mixtures of colors representing a stochastic decision.

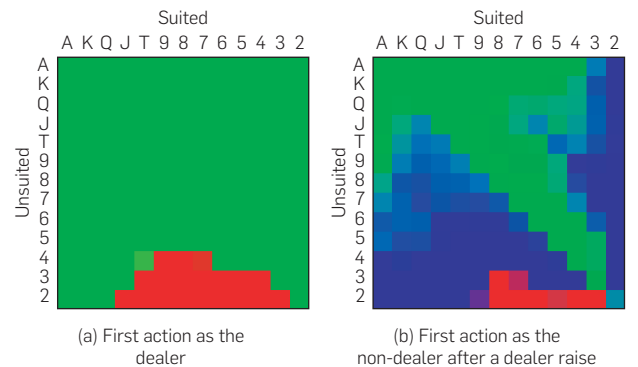


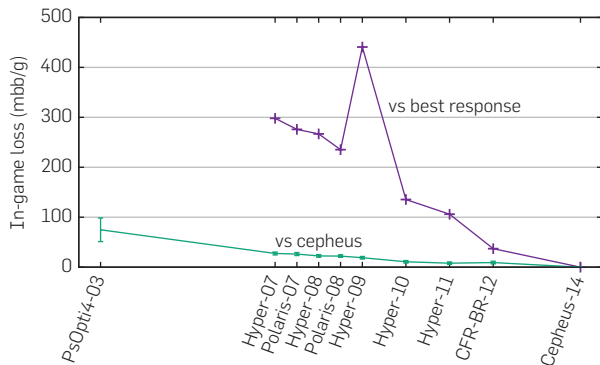
Figure 5 presents the exploitability of our historical agents and their average loss in games played against Cepheus. To reduce the impact of luck, a duplicate poker format was used where each game is played twice, using the same cards, but with the players in opposite positions. PsOpti4 was the first game theoretic strategy produced for HULHE, and was also the University of Alberta entry to the 2006 ACPC.^{5, h} The University of Alberta entries to the ACPC were named Hyperborean, and from 2007 onwards, all were created using variants of CFR.¹ The Polaris 2007 and 2008 agents were created by the University of Alberta for its two Man-vs.-Machine Poker Championship matches, in which Polaris narrowly lost in 2007 and narrowly won in 2008; an analysis of these matches is available in²⁴ [Chapter 8]. Finally, the CFR-BR agent was our closest equilibrium approximation prior to this work.²¹ It used the same abstract game as Hyperborean 2011, but used an algorithm that solved for the abstract strategy with the lowest real game exploitability.

These results show that, with the exception of Hyperborean 2009, each new generation of strategies improved in both exploitability and in loss against an essentially optimal

^h PsOpti4 acts too slowly for an exploitability calculation to be practical, or for a long match against Cepheus.

¹ In the inaugural 2006 ACPC, PsOpti4 was the core component of Hyperborean 2006.

Figure 5. Exploitability and performance against Cepheus for earlier computer strategies. Results are in mbb/g, and indicate the expected winnings by the strategy's opponent (a best response or Cepheus, respectively). The Cepheus matches involved 1mn games of duplicate poker (2mn games total), except for PsOpti4 which played 20,000 duplicate games (40,000 games total).



Name	Year	Exploitability	Cepheus
PsOpti4	2003	–	74.9 ± 23.7
Hyperborean 2007	2007	298.106	27.4 ± 2.9
Polaris 2007	2007	275.880	26.2 ± 3.0
Hyperborean 2008	2008	266.797	22.5 ± 2.7
Polaris 2008	2008	235.294	22.2 ± 2.6
Hyperborean 2009	2009	440.823	18.9 ± 2.6
Hyperborean 2010	2010	135.427	10.8 ± 2.5
Hyperborean 2011	2011	106.035	8.0 ± 2.4
CFR-BR	2012	37.113	9.2 ± 2.6
Cepheus	2014	0.986	0

strategy. However, even though many of these strategies were highly exploitable, the rate at which they lose to Cepheus is quite low. This loss is difficult to measure with statistical confidence: a 100,000 game (non-duplicate) match would have a 95% confidence interval of 31mbb/g, larger than the performance difference between Cepheus and every agent but PsOpti4. Further, Hyperborean 2009 did improve over its predecessors in terms of in-game performance against Cepheus, and regressed in exploitability due to its use of “Strategy Grafting,” an unsound solving technique that solves an abstraction as a series of fragments.⁵⁰ This technique allows for a much larger and finer grained abstraction than would otherwise be feasible, resulting in improved in-game performance, but without theoretical guarantees on exploitability. Together, these results illustrate the difficulty in evaluating a strategy only through its competition performance, instead of calculating its exploitability.

We can also measure Cepheus’ performance against human adversaries. After this article was first published in January 2015, our website allowed visitors to play against Cepheus and inspect its strategy.⁸ Each visitor chose a username and played any number of short 100-game matches against Cepheus. Over the last two years, 39,564 unique usernames have played 98,040 matches, with 3,564,094 total games played.¹ Over this set of games, Cepheus is winning at a rate of 169.9 ± 5.2 mbb/g with 95% confidence. However, most of the players did not finish a single 100-game match (only 7,878 players did so, with 20,374 completed matches in total), and so this winrate is likely not reflective of Cepheus’ performance against strong opponents.

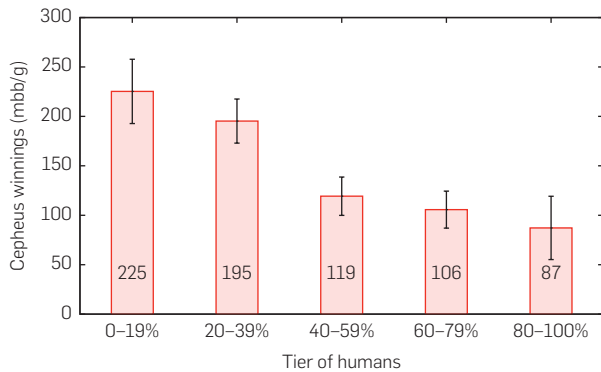
Determining which of these players are strong is non-trivial because of both variance in their matches, and the unequal amount of games played by each player. While both luck and skill contribute to a player’s performance, the highest-scoring players are more likely to be the luckiest rather than the strongest. Additionally, bias may be introduced if players keep playing while ahead, but quit if they are losing. In order to limit the impact of bias and evaluate Cepheus’ performance against different tiers of humans, we used the following method. First, we eliminated usernames with insufficient data that had played fewer than 500 games, leaving 821 usernames playing 33,752 matches with 1,765,656 games. Next, we divided each username’s games into two sets, called Rank and Test.^k Each username’s Rank games were evaluated, and the resulting winrates were used to sort the players by performance. This ordering reflected both their skill and luck. The players were then divided equally into five tiers: the bottom 20% of usernames, 21–40% etc. Within each tier, the Test game results were averaged to produce a winrate for the tier, independent from the luck that affected the Rank games.

These results are shown in Figure 6. Cepheus’ estimated winrate varies from 225 to 87mbb/g as we advance through the tiers, decreasing as the quality of the human

^j Many players quit before finishing the 100-game match.

^k In each block of four sequential games, one pair (played in each position) was assigned to each set.

Figure 6. Games played by humans and Cepheus. Humans were divided equally into five skill groups, and the column and error bar indicates the group's average loss to Cepheus in mbb/g.



players improves. Even against the top 20% tier of players in this experiment, Cepheus' winrate of 87mbb/g is higher than against any of our historical agents. It even exceeds 50mbb/g, a commonly cited benchmark for what a professional poker player seeks to win from a weaker opponent.

6. CONCLUSION

In this paper, we announced that heads-up limit Texas hold'em poker is essentially weakly solved. This is the first nontrivial imperfect information game played competitively by humans to be solved. Even still, the reader may ask what is the ultimate significance of solving poker? The breakthroughs behind this result are general algorithmic advances that make game-theoretic reasoning in large-scale models of any sort more tractable. And, while seemingly playful, game theory has always been envisioned to have serious implications, for example, its early impact on cold war politics.³¹ More recently, there has been a surge in game-theoretic applications involving security, including systems being deployed for airport checkpoints, air marshal scheduling, and coast guard patrolling.⁴³ CFR algorithms, based on those described in this paper, have been used for robust decision-making in settings where there is no apparent adversary, with potential application to medical decision support.¹⁵ With real life decision-making settings almost always involving uncertainty and missing information, algorithmic advances, such as those needed to solve poker, are needed to drive future applications. However, we also echo a response attributed to Alan Turing in defense of his own work in games, "It would be disingenuous of us to disguise the fact that the principal motive which prompted the work was the sheer fun of the thing."³⁰

Acknowledgments

The author order is alphabetical reflecting equal contribution by the authors. The idea of CFR⁺ and compressing the regrets and strategy originated with Oskari Tammelin.⁴⁴ This research was supported by Natural Sciences and Engineering Research Council (NSERC), Alberta Innovates Technology Futures (AITF) through the Alberta Innovates

Centre for Machine Learning (AICML), and was only possible due to computing resources provided by Compute Canada and Calcul Québec. The authors would like to thank all of the current and past members of the University of Alberta Computer Poker Research Group (CPRG), where the idea to solve heads-up limit Texas hold'em was first discussed; Jonathan Schaeffer, Robert Holte, Duane Szafron, and Alex Brown for comments on early drafts of this article; and Bryce Paradis for insights into the conventional wisdom of top human poker players.

References

1. Poker: A big deal. *The Economist*, Economist Newspaper Ltd., London, December 22, 31–38, 2007.
2. Allis, V.L. *A Knowledge-Based Approach to Connect-Four. The Game is Solved: White Wins*. Master's thesis, Vrije Universiteit, Amsterdam, The Netherlands, 1988.
3. Allis, V.L. *Searching for Solutions in Games and Artificial Intelligence*. PhD thesis, Vrije Universiteit Amsterdam, The Netherlands, 1994.
4. Babbage, C. *Passages from the Life of a Philosopher*. Longman, Green, Longman, Roberts, and Green, London, 1864. Chapter 34.
5. Billings, D., Burch, N., Davidson, A., Holte, R.C., Schaeffer, J., Schauenberg, T., Szafron, D. Approximating game-theoretic optimal strategies for full-scale poker. *IJCAI*, (2003), 661–668.
6. Billings, D., Davidson, A., Schaeffer, J., Szafron, D. The challenge of poker. *Artificial Intelligence* 134, 1–2 (2002), 201–240.
7. Borel, E., Ville, J. *Applications de la théorie des probabilités aux jeux de hasard*. Gauthier-Villars, 1938.
8. Bowling, M., Burch, N., Johanson, M., Tammelin, O. The cepheus website, 2015. <http://poker.srv.ualberta.ca>.
9. Bowling, M., Burch, N., Johanson, M., Tammelin, O. Heads-up limit hold'em poker is solved. *Science* 347, 6218 (Jan. 2015), 145–149.
10. Bowling, M., Burch, N., Johanson, M., Tammelin, O. Heads-up limit hold'em poker is solved: Supplementary online material, January 2015.
11. Bowling, M., Johanson, M., Burch, N., Szafron, D. Strategy evaluation in extensive games with importance sampling. *ICML*, (2008), 72–79.
12. Bronowski, J. *Ascent of man*. Documentary, 1973. Episode 13.
13. Buro, M. Solving the oshi-zumo game. *Adv. in Comp. Games* 135, (2004) 361–366.
14. Campbell, M., Hoane, A.J. Jr., Hsu, F. Deep blue. *Artificial Intelligence* 134, (Jan. 2002), 57–83.
15. Chen, K., Bowling, M. Tractable objectives for robust policy optimization. *Adv. Neural Inf. Process. Syst. (NIPS)* 25, (2012), 2078–2086.
16. Craig, M. *The Professor, the Banker, and the Suicide King: Inside the Richest Poker Game of All Time*. Grand Central Publishing, New York, NY, 2006.
17. Ferrucci, D. Introduction to "this is watson." *IBM J. Res. Dev* 56, 3.4 (May 2012) 1:1–1:15.
18. Gilpin, A., Hoda, S., Peña, J., Sandholm, T. Gradient-based algorithms for finding nash equilibria in extensive form games. *WINE*, (2007), 57–69.
19. Gilpin, A., Sandholm, T. Lossless abstraction of imperfect information games. *J. ACM* 54, 5 (2007).
20. Jackson, E. Slumbot: An implementation of counterfactual regret minimization on commodity hardware. In *Proceedings of the 2012 Computer Poker Symposium*. (2012).
21. Johanson, M., Bard, N., Burch, N., Bowling, M. Finding optimal abstract strategies in extensive form games. *AAAI*, (2012), 1371–1379.
22. Johanson, M., Bard, N., Lanctot, M., Gibson, R., Bowling, M. Efficient Nash equilibrium approximation through Monte Carlo counterfactual regret minimization. *AAMAS* (2012).
23. Johanson, M., Waugh, K., Bowling, M., Zinkevich, M. Accelerating best response calculation in large extensive games. *IJCAI* (2011), 258–265.
24. Johanson, M.B. *Robust Strategies and Counter-Strategies: From Superhuman to Optimal Play*. PhD thesis, University of Alberta, Edmonton, Alberta, Canada, 2016.
25. Karmarkar, N. A new polynomial-time algorithm for linear programming. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing* (1984), ACM, New York, NY, 302–311.
26. Koller, D., Megiddo, N. The complexity of two-person zero-sum games in extensive form. *Games Econ. Behav.* 4, 4 (1992), 528–552.
27. Koller, D., Megiddo, N., von Stengel, B. Efficient computation of equilibria for extensive two-person games. *Games Econ. Behav.* 14, 2 (1996).
28. Koller, D., Pfeffer, A. Representations and solutions for game-theoretic problems. *Artificial Intelligence* 94, (1997), 167–215.
29. Kuhn, H. Simplified two-person poker. In *Contributions to the Theory of Games*, volume 1 of *Annals of mathematics studies*. H. Kuhn and A. Tucker, eds. Princeton University Press, Princeton, New Jersey, 1950, 97–103.
30. Mirowski, P. What were von neumann and Morgenstern trying to accomplish? In *Toward a History of Game Theory*. Weintraub, ed. Duke University Press, 1992, 113–147. Mirowski cites Turing as author of the paragraph containing this remark. The paragraph appeared in [46], in a chapter with Turing listed as one of three contributors. Which parts of the chapter are the work of which contributor, particularly the introductory material containing this quote, is not made explicit.
31. Morgenstern, O. The cold war is cold poker. *N. Y. Times Mag.* (Feb. 5 1961) pages 21–22.
32. Nash, J.F., Shapley, L.S. A simple 3-person poker game. In *Contributions to the Theory of Games I*. Princeton University Press, Princeton, New Jersey, 1950, 105–116.

33. Nesterov, Y. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization* 16, 1 (2005), 233–249.
34. Rehmeier, J., Fox, N., Rico, R. Ante up, human: The adventures of polaris the poker-playing robot. *Wired* 16, 12 (Dec. 2008), 186–191.
35. Romanovskii, I.V. Reduction of a game with complete memory to a matrix game. *Soviet Mathematics* 3, (1962), 678–681.
36. Rubin, J., Watson, I. Computer poker: a review. *Artificial Intelligence* 175, (2011), 958–987.
37. Sandholm, T. The state of solving large incomplete-information games, and application to poker. *AI Mag.* 31, 4 (2010), 13–32.
38. Schaeffer, J., Lake, R., Lu, P., Bryant, M. Chinook the world man-machine checkers champion. *AI Mag.* 17, 1 (1996), 21–29.
39. Schaeffer, J., Neil Burch, Y.B., Kishimoto, A., Müller, M., Lake, R., Lu, P., Sutphen, S. Checkers is solved. *Science* 317, 5844 (2007), 1518–1522.
40. Shannon, C.E. Programming a computer for playing chess. *Philosophical Magazine, Series 7*, 41, 314 (Mar. 1950), 256–275.
41. Shi, J., Littman, M.L. Abstraction methods for game theoretic poker. In *Comp. Games*, (2000), 333–345.
42. Southey, F., Bowling, M., Larson, B., Piccione, C., Burch, N., Billings, D., Rayner, D.C. Bayes' bluff: Opponent modelling in poker. *UAI*, (2005) 550–558.
43. Tambe, M. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press, Cambridge, England, 2011.
44. Tammelin, O. Cfr+. *CoRR*, abs/1407.5042, 2014.
45. Tammelin, O., Burch, N., Johanson, M., Bowling, M. Solving heads-up limit texas hold'em. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 2015, 645–652.
46. Turing, A. Digital computers applied to games. In *Faster Than Thought*. B.V. Bowden, ed. Chapter 25. Pitman, 1976.
47. von Neumann, J. Zur theorie der gesellschaftsspiele. *Mathematische Annalen* 100, 1 (1928), 295–320.
48. von Neumann, J., Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, Second Edition, 1947.
49. Waugh, K., Bagnell, J.A. A unified view of large-scale zero-sum equilibrium computation. In *AAAI Workshop on Computer Poker and Imperfect Information*, 2015.
50. Waugh, K., Bard, N., Bowling, M. Strategy grafting in extensive games. In *Advances in Neural Information Processing Systems 22 (NIPS-09)*, 2009. <http://webdocs.cs.ualberta.ca/~games/poker/publications/NIPS09-graft.pdf>.
51. Zermelo, E. Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. In *Proceedings of the Fifth International Congress Mathematics*. Cambridge University Press, Cambridge, 1913, 501–504.
52. Zinkevich, M., Johanson, M., Bowling, M., Piccione, C. Regret minimization in games with incomplete information. *NIPS* (2008), 905–912.
53. Zinkevich, M., Littman, M. The AAAI computer poker competition. *J. Inter. Comp. Games Association* 29, (2006), News item.

Michael Bowling (bowling@cs.ualberta.ca), Department of Computing, Science, University of Alberta, Edmonton, Alberta, Canada.

Oskari Tammelin (ot@iki.fi) (<http://jeskola.net/>).

Neil Burch and Michael Johanson ([\[nburch, johanson\]@ualberta.ca](mailto:[nburch, johanson]@ualberta.ca)), Department of Computing, Science, University of Alberta, Edmonton, Alberta, Canada.

© 2017 ACM 0001-0782/17/11 \$15.00



Plato and The Nerd

The Creative Partnership of Humans and Technology

Edward Ashford Lee

How humans and technology evolve together in a creative partnership.

Foolproof, and Other Mathematical Meditations

Brian Hayes

A non-mathematician explores mathematical terrain, reporting accessibly and engagingly on topics from Sudoku to probability.

Robot Sex

Social and Ethical Implications

edited by John Danaher and Neil McArthur

Perspectives from philosophy, psychology, religious studies, economics, and law on the possible future of robot-human sexual relationships.

mitpress.mit.edu/acm

Technical Perspective

Exploring a Kingdom by Geodesic Measures

By Marc Alexa

ONCE UPON A time, there was a particularly generous king. He decided to give all of his people a share of his land. Any family owning a house would have to mark the center of their land with a stone. Then the ruling said that every point they could reach within 1,000 steps from the mark stone would belong to them.

Initially, the king was praised for the simplicity and clarity of his ruling. Yet, as most new landowners decided to build a fence around their property they soon realized that determining the location of the fence was frustratingly difficult. Any point that was reached after walking 1,000 steps was part of the property, sure; but how would one have to walk to reach a point that was as far away from the stone as possible?

Luckier landowners lived in flat regions, where walking along a straight line (for example, walking toward any landmark further away than 1,000 steps) would always reach a point on the boundary of the land. Landowners in hilly regions had a problem. They needed to find a path that was the equivalent of a straight line. Such paths are now known as geodesics—the name indeed relating to the problem of *measuring the earth*. In today's language, the land assigned to each family is a geodesic disk. And it should come as no surprise that the distance between two points in curved domains is a fundamental problem with numerous applications in science and engineering.

So, how would we solve the problem of computing geodesic paths, distances, and circles on a digital computer? Let's assume the smooth terrain is approximated as a triangulation. The shortest path from a vertex of this triangulation to all other vertices *along the edges* of the triangulation can be computed using Dijkstra's algorithm. While this provides an efficient solution to the problem

of finding a path, traveling only along the edges of the triangles will not yield the shortest possible path. A better option is tracking how a wavefront emanates from a source vertex, the so-called *fast marching method*.

Unfortunately, this method may fail to provide shortest paths for certain surfaces and shapes of triangles. It turns out the problem of shortest paths along a piecewise linear surface has a superquadratic worst-case complexity, and algorithms establishing this complexity are surprisingly complicated. The most recent algorithms of this kind have much better asymptotic complexity in common practical cases, yet they are still significantly more complex than a simple graph traversal such as Dijkstra's.

The approach by Crane, Weischedel, and Wardetzky in the following paper may be related to what the landowners could have observed and exploited: assume one would start from the mark stone, walk 1,000 random steps, and mark the endpoint of this walk with a pebble. Over many repetitions of this exercise a distribution of pebbles on the land emerges: closer to the mark stone the pattern is denser, and fewer and fewer pebbles are found further away. As it turns out, starting at the mark stone and always walking in the direction of

fewer stones will result in the desired geodesic path. Or, in more recent language, the gradient of the probability density function of a random walk is parallel to geodesics.

Now, why is this observation useful for computation? The density of random walks is related to heat diffusion, and heat diffusion is governed by a first-order partial differential equation, which can be well approximated by solving a sparse linear system. The important insight from the authors is that a wide range of linear operators can be used to compute functions with the desired gradient directions (yet different gradient lengths)

Given any such function, distances can be computed by simply normalizing the gradients and then integrating them. Integration amounts to solving another sparse linear system. Together this means geodesic distances can be computed by solving two (related) sparse linear systems, plus computing the gradients of the intermediate function and normalizing them. Note how the non-linearity of the problem is pushed into the trivial step of normalizing a set of vectors, while all global computations are linear.

And lastly, turning the problem of computing geodesic paths to the solution of a sparse linear system has another very desirable feature: the most time-consuming part lies in the step of factorizing the system matrix. This factorization can be reused for arbitrary distance computations on the domain. So *all* the king would have needed were the triangular factors of a linear system, and then each geodesic circle could have been computed almost instantaneously. 

Marc Alexa is a professor in the Faculty of Electrical Engineering and Computer Science at the Technical University of Berlin, and heads the Computer Graphics group.

Copyright held by author.

How would we solve the problem of computing geodesic paths, distances, and circles on a digital computer?

The Heat Method for Distance Computation

By Keenan Crane, Clarisse Weischedel, and Max Wardetzky

Abstract

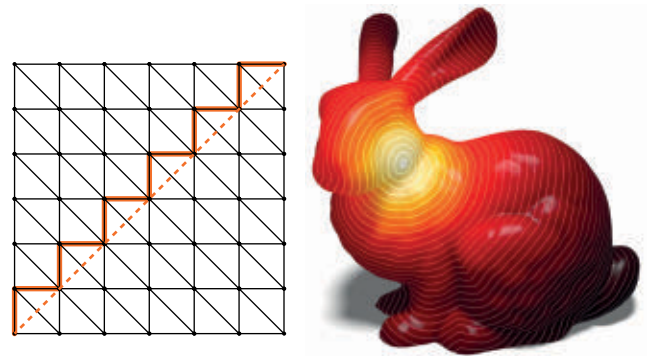
We introduce the *heat method* for solving the single- or multiple-source shortest path problem on both flat and curved domains. A key insight is that distance computation can be split into two stages: first find the direction along which distance is increasing, then compute the distance itself. The heat method is robust, efficient, and simple to implement since it is based on solving a pair of standard sparse linear systems. These systems can be factored once and subsequently solved in near-linear time, substantially reducing amortized cost. Real-world performance is an order of magnitude faster than state-of-the-art methods, while maintaining a comparable level of accuracy. The method can be applied in any dimension, and on any domain that admits a gradient and inner product—including regular grids, triangle meshes, and point clouds. Numerical evidence indicates that the method converges to the exact distance in the limit of refinement; we also explore smoothed approximations of distance suitable for applications where greater regularity is desired.

1. INTRODUCTION

The multiple-source shortest path problem seeks the distance from each point of a domain to the closest point within a given subset; different versions of this problem are fundamental to a wide array of problems across computer science and computational mathematics. Solutions date back at least to Dantzig's work on linear programs³⁵; typically the problem is formulated in terms of a weighted graph, as in Dijkstra's algorithm. Often, however, one wishes to capture the distance on a continuous domain; a key example is illustrated in Figure 1 (left) where the graph distance will overestimate the straight-line Euclidean distance, no matter how fine the grid becomes. In 2D, an important development was the formulation of “exact” algorithms, where paths can cut through the faces of a triangulation^{8, 27}; a great deal of subsequent work has focused on making these $O(n^2)$ algorithms practical for large datasets.^{40, 46} However, for problems in data analysis and scientific computing it is not clear that the cost and complexity of exact algorithms are always well-justified, since the triangulation itself is only an approximation of the true domain (see Figure 4).

A very different approach is to formulate the problem in terms of partial differential equations (PDEs), where domain approximation error can be understood via, for example, traditional finite element analysis. However, the particular choice of continuous formulation has a substantial impact on computation. The *heat method* was inspired by S.R.S. Varadhan's classic result in differential geometry⁴² relating heat diffusion and *geodesic distance*, which

Figure 1. In contrast to algorithms that compute shortest paths along a graph (left), the heat method computes the distance to points on a continuous, curved domain (right). A key advantage of this method is that it is based on sparse linear equations that can be efficiently prefactored, leading to dramatically reduced amortized cost.



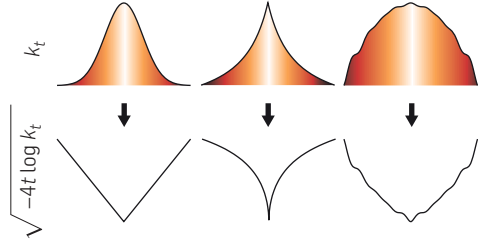
measures the length along shortest and straightest curves through the domain rather than straight lines through space. Our key observation is that one can decompose distance computation into two stages: first determine the direction along which distance increases, then recover the distance itself. Moreover, since each stage amounts to a standard problem in numerical linear algebra, one can leverage existing algorithms and software to improve the efficiency and robustness of distance computation. Although this approach can in principle be used in the context of graph distance, its real utility lies in approximating the distance on continuous, curved domains. This approach has proven effective for a diverse range of applications in computational neuroscience, geometric modeling, medical imaging, computational design, and machine learning (Section 2), and has recently inspired more accurate variations of our original method.³

1.1. Formulation

Imagine touching a scorching hot needle to a single point on a surface. Over time heat spreads out over the rest of the domain and can be described by a function $k_{t,x}(y)$ called the *heat kernel*, which measures the heat transferred from a source x to a destination y after time t . A well-known relationship between heat and distance is *Varadhan's formula*,⁴²

The original version of this paper was published in *ACM Transactions on Graphics* 32, 5 (Sept. 2013).

Figure 2. Given an exact reconstruction of the heat kernel (top left) Varadhan's formula can be used to recover geodesic distance (bottom left) but fails in the presence of approximation or numerical error (middle, right), as shown here for a point source in 1D. The robustness of the heat method stems from the fact that it depends only on the *direction* of the gradient.



which says that the distance ϕ between any pair of points x, y on a curved domain can be recovered via a simple pointwise transformation of the heat kernel:

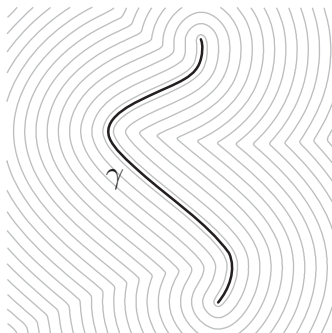
$$\phi(x, y) = \lim_{t \rightarrow 0} \sqrt{-4t \log k_{t,x}(y)}. \quad (1)$$

The intuition behind this behavior stems from the fact that heat diffusion can be modeled as a large collection of hot particles taking random walks starting at x : any particle that reaches a distant point y after a small time t has had little time to deviate from the shortest possible path. Previously, however, this relationship had not been exploited by numerical algorithms that compute distance.

Why had Varadhan's formula been overlooked in this context? The main reason, perhaps, is that it requires a precise numerical reconstruction of the heat kernel, which is difficult to obtain—applying the formula to a mere approximation of $k_{t,x}$ does not yield the correct result, as illustrated in Figures 2 and 8. The heat method circumvents this issue by working with a broader class of inputs, namely any function whose gradient is parallel to the gradient of the true distance function. We can then separate computation into two stages: first find the gradient, then recover the distance itself.

Relative to existing algorithms, the heat method offers two major advantages. First, it can be applied to virtually any

Figure 3. The heat method computes the shortest distance to a subset γ of a given domain. Gray curves indicate isolines of the distance function.



type of geometric discretization, including regular grids, polygonal meshes, and point clouds. Second, it involves only sparse linear systems, which can be prefactored once and rapidly resolved many times—this feature substantially reduces the amortized cost for applications that require repeated distance queries on a fixed geometric domain. Moreover, because the heat method is built on standard linear PDEs that are widespread in scientific computing, it can immediately take advantage of new developments in numerical linear algebra and parallelization.

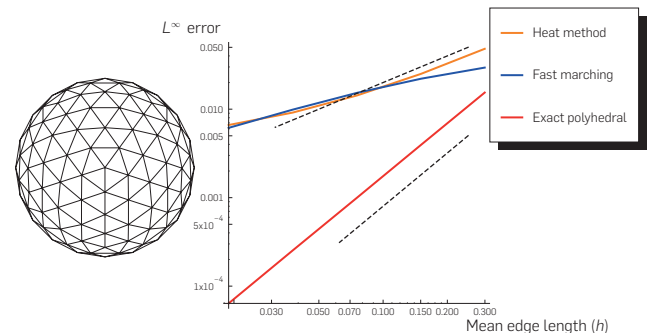
2. RELATED WORK

The prevailing approach to distance computation is to solve the *eikonal equation*

$$|\nabla \phi| = 1, \quad (2)$$

subject to boundary conditions $\phi|_{\gamma} = 0$ over some subset γ of the domain (like a point or a curve). Intuitively, this equation says something very simple: as we move away from the source, the distance function ϕ must change at a rate of “one meter per meter.” Computationally, however, this formulation is nonlinear and hyperbolic, making it difficult to solve directly. Typically one applies an iterative relaxation scheme such as Gauss-Seidel—special update orders are known as *fast marching* and *fast sweeping*, which are some of the most popular algorithms for distance computation on regular grids³⁷ and triangulated surfaces.¹⁹ These algorithms can also be used on implicit surfaces,²⁵ point clouds,²⁶ and polygon soup,⁷ but only indirectly: distance is computed on a simplicial mesh or regular grid that approximates the original domain. Implementation of fast marching on simplicial grids is challenging due to the need for nonobtuse triangulations (which are notoriously difficult to obtain) or else an iterative unfolding procedure that preserves monotonicity of the solution; moreover these issues are not well-studied in dimensions greater than two. Fast marching and fast sweeping have asymptotic complexity of $O(n \log n)$ and $O(n)$, respectively,

Figure 4. Convergence of distance approximations on the unit sphere with respect to mean edge length; as a baseline for comparison, we use the analytical solution $\phi(x, y) = \cos^{-1}(x \cdot y)$. Notice that even with a nice tessellation, the exact distance along the polyhedron converges only quadratically to the true distance along the sphere it approximates. (Linear and quadratic convergence are plotted as dashed lines for reference.)



but sweeping is often slower due to the large number of sweeps required to obtain accurate results.¹⁶

One drawback of these methods is that they do not reuse information: the distance to different source sets γ must be computed entirely from scratch each time. Also note that both sweeping and marching present challenges for parallelization: priority queues are inherently serial, and irregular meshes lack a natural sweeping order.

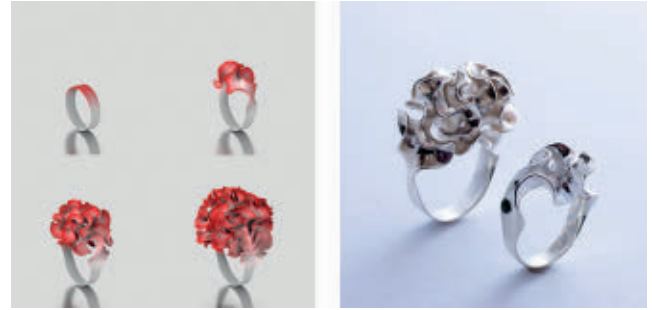
In a different development, Mitchell et al.²⁷ give an $O(n^2 \log n)$ algorithm for computing the exact polyhedral distance from a single source to all other vertices of a triangulated surface. Surazhsky et al.⁴⁰ demonstrate that this algorithm tends to run in sub-quadratic time in practice, and present an approximate $O(n \log n)$ version of the algorithm with guaranteed error bounds; Bommers and Kobbelt⁴ extend the algorithm to polygonal sources. Similar to fast marching, these algorithms propagate distance information in wavefront order using a priority queue, again making them difficult to parallelize. More importantly, the amortized cost of these algorithms (over many different source subsets γ) is substantially greater than for the heat method since they do not reuse information from one subset to the next. Finally, although⁴⁰ greatly simplifies the original formulation, these algorithms remain challenging to implement and do not immediately generalize to domains other than triangle meshes.

Closest to our approach is the recent method of Rangarajan and Gurumoorthy,³² who do not appear to be aware of Varadhan's formula—they instead derive an analogous relationship between the distance function and solutions ψ to the time-independent Schrödinger equation; this derivation applies only in flat Euclidean space rather than general curved domains. Moreover, they compute solutions using the fast Fourier transform, which limits computation to regular grids.

A slight modification of the heat method allows us to compute a smoothed distance function, useful in contexts where sharp discontinuities can cause subsequent numerical difficulties. Previous smooth distance approximations provide this regularity at the cost of poor approximation of the true geometric length^{10, 14, 21, 33}; see Section 3.3 for a comparison.

Recently, the heat method has facilitated a variety of tasks in computational science and data analysis. For example, Huth et al.¹⁵ use fast distance queries to optimize a probabilistic model of cortical organization; van Pelt et al.⁴¹ use the heat method to assist cerebral aneurysm assessment; Zou et al.⁴⁷ use the heat method for efficient tool path planning; Solomon et al.³⁸ leverage our approach to efficiently solve optimal transport problems on geometric domains; Lin et al.²⁰ apply this approach to vector-valued data in the context of manifold learning. Figure 5 shows a real-world design application of the heat method based on differential growth. Various improvements have also been made to the original algorithm; for instance, de Goes et al.¹³ and Yang and Cohen⁴⁵ describe two different ways to extend the method to accurate computation of anisotropic distance; it has also been adapted to voxelizations⁶, C^1 finite elements,²⁹ and subdivision surfaces.¹² Finally, Belyaev and Fayolle³ provide a variational interpretation of our method, observing that more accurate results can be obtained by either

Figure 5. The heat method has been applied to a diverse range of tasks that demand repeated geodesic distance queries. Here, geodesic distance drives a differential growth model (left) that is used for computational design (right). Images courtesy Nervous System/Jesse Louis-Rosenberg.



iterating the heat method, or by applying more sophisticated descent strategies.

3. THE HEAT METHOD

A useful feature of the heat method is that the basic algorithm can be described in the purely continuous setting (i.e., in terms of curved surfaces, or more generally, *smooth manifolds*) rather than in terms of discrete data structures and algorithms. In other words, at this point one should not imagine that we have chosen a particular data structure (triangle meshes, grids, point clouds, etc.) or even dimension (2D, 3D, etc.). Instead, we focus on a general principle that can be applied on many different domains in different dimensions. We will later explore several particular choices of spatial and temporal discretization (Sections 3.1 and 3.2); further alternatives have been explored in recent literature.^{13, 29, 45}

In general, the heat method can be applied in any setting where one has a gradient operator ∇ , divergence operator $\nabla \cdot$, and Laplace operator $\Delta = \nabla \cdot \nabla$ —standard derivatives from vector calculus, possibly generalized to curved domains. Expressed in terms of these operators, the heat method consists of three basic steps:

Algorithm 1 The Heat Method

- I. Integrate the heat flow $\dot{u} = \Delta u$ for some fixed time t .
- II. Evaluate the vector field $X = -\nabla u_t / |\nabla u_t|$.
- III. Solve the Poisson equation $\Delta \phi = \nabla \cdot X$.

The function ϕ approximates the distance to a given source set, approaching the true distance as t goes to zero (Equation 1). For instance, to recover the distance to a single point x we use initial conditions $u_0 = \delta(x)$, that is, a Dirac delta encoding an “infinite spike” of heat. More generally we can obtain the distance to any subset γ by letting u_0 be a generalized Dirac distribution⁴²—essentially an indicator function over γ ; see Figures 3 and 7. Note that since the solution to (III) is determined only up to an additive constant, final values are shifted such that the smallest distance is zero.

The heat method can be motivated as follows. Consider an approximation u_t of heat flow for a fixed time t . Unless u_t exhibits precisely the right rate of decay, Varadhan's

Figure 6. The three steps of the heat method. (I) Heat u is allowed to diffuse for a brief period of time (left). (II) The temperature gradient Δu (center left) is normalized and negated to get a unit vector field X (center right) pointing along geodesics. (III) A function γ whose gradient follows X recovers the final distance (right).

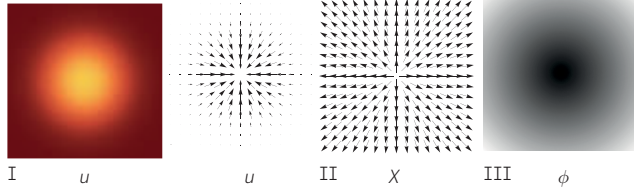


Figure 7. Distance to the boundary on a region in the plane (left) or a surface in space (right) is achieved by simply placing heat along the boundary curve.



transformation $u_i \mapsto \sqrt{-4t \log u_i}$ will yield a poor approximation of the true geodesic distance ϕ because it is highly sensitive to errors in magnitude (see Figures 2 and 8). The heat method asks for something different: it requires only that the gradient ∇u_i point in the right direction, that is, parallel to $\nabla \phi$. Magnitude can safely be ignored since we know (from the eikonal equation) that the gradient of the true distance function has unit length. We therefore compute the normalized gradient field $X = -\nabla u_i / |\nabla u_i|$ and find the closest scalar potential ϕ by minimizing $\int_M |\nabla \phi - X|^2$, or equivalently, by solving the corresponding Euler-Lagrange equations $\Delta \phi = \nabla \cdot X$.³⁶ The overall procedure is depicted in Figure 6.

This procedure is used as the starting point for a family of discrete algorithms, as outlined in Sections 3.1–3.3. Note that some details have been omitted from this manuscript, and can be found in Crane et al.¹¹

3.1. Time discretization

To translate our continuous procedure (Algorithm 1) into a discrete algorithm, we must replace derivatives in space and time with suitable approximations. The heat equation from step I of Algorithm 1 can be discretized in time using a single backward Euler step for some fixed time t —in practice, this means we simply solve the linear equation

$$(\text{id} - t\Delta)u_t = u_o, \quad (3)$$

over the entire domain M , where id denotes the identity operator. Note that at this point we still have not discretized space; spatial discretization is discussed in Section 3.2. We can get a better understanding of solutions to Equation (3) by considering the elliptic boundary value problem

Figure 8. Left: Varadhan’s formula. Right: the heat method. Even for very small values of t , simply applying Varadhan’s formula does not provide an accurate approximation of geodesic distance (top left); for large values of t spacing becomes even more uneven (bottom left). Normalizing the gradient results in a more accurate solution, as indicated by evenly spaced isolines (top right), and is also valuable when constructing a smoothed distance function (bottom right).



$$\begin{aligned} (\text{id} - t\Delta)v_t &= 0 & \text{on } M \setminus \gamma, \\ v_t &= 1 & \text{on } \gamma. \end{aligned} \quad (4)$$

which for a point source yields a solution v_t equal to u_i up to a multiplicative constant. As established by Varadhan in his proof of Equation (1), v_t also has a close relationship with distance, namely

$$\lim_{t \rightarrow 0} -\frac{\sqrt{t}}{2} \log v_t = \phi. \quad (5)$$

This relationship ensures the validity of steps II and III since the transformation applied to v_t preserves the direction of the gradient.

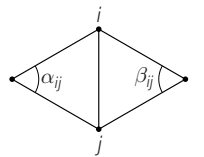
3.2. Spatial discretization

Here we detail several possible implementations of the heat method on triangle meshes, polygon meshes, and point clouds. Note that the heat method can also be used on flat Euclidean domains of any dimension by simply applying standard finite differences on a regular grid; Belyaev and Fayolle³ outline implementation on tetrahedral (3D) meshes.

Triangle meshes. Let $u \in \mathbb{R}^{|V|}$ specify a piecewise linear function on a triangulated surface with vertices V , edges E , and faces F . A standard discretization of the Laplacian at a vertex i is given by

$$(Lu)_i = \frac{1}{2A_i} \sum_j (\cot \alpha_{ij} + \cot \beta_{ij})(u_j - u_i),$$

where A_i is one third the area of all triangles incident on vertex i , the sum is taken over all neighboring vertices j , and α_{ij}, β_{ij} are the angles opposing the corresponding edge.²³ We can express this operation via a matrix $L = M^{-1}L_C$, where $M \in \mathbb{R}^{|V| \times |V|}$ is a diagonal matrix containing the vertex areas and $L_C \in \mathbb{R}^{|V| \times |V|}$ is the *cotan operator* representing the



remaining sum. Heat flow can then be computed by solving the symmetric positive-definite system

$$(M - tL_C)u = \delta_\gamma,$$

where δ_γ is a Kronecker delta (or indicator function) over γ . The gradient in a given triangle can be expressed succinctly as

$$\nabla u = \frac{1}{2A_f} \sum_i u_i (N \times e_i),$$

where A_f is the area of the triangle, N is its outward unit normal, e_i is the i th edge vector (oriented counter-clockwise), and u_i is the value of u at the opposing vertex. The integrated divergence associated with vertex i can be written as

$$\nabla \cdot X = \frac{1}{2} \sum_j \cot \theta_1 (e_1 \cdot X_j) + \cot \theta_2 (e_2 \cdot X_j),$$

where the sum is taken over incident triangles j each with a vector X_j , e_1 , and e_2 are the two edge vectors of triangle j containing i , and θ_1, θ_2 are the opposing angles. If we let $b \in \mathbb{R}^{|V|}$ be the vector of (integrated) divergences of the normalized vector field X , then the final distance function is computed by solving the symmetric Poisson problem

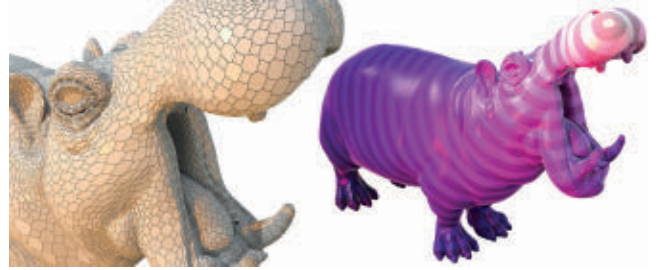
$$L_C \phi = b.$$

As noted in Section 3.1, the solution to step I is a function that decays exponentially with distance. Fortunately, normalization of small values is not a problem because floating point division involves only arithmetic on integer exponents; likewise, the large range of magnitudes does not adversely affect accuracy because gradient calculation is local. For the calculation of ϕ itself we advocate the use of a direct (Cholesky) solver in double precision; empirically we observe roughly uniform pointwise relative error across the domain.

Polygon meshes. Curved surfaces are often described by polygons that are neither planar nor convex; although such polygons can of course be triangulated, doing so can adversely affect an existing computational pipeline. We instead leverage the polygonal Laplacian of Alexa and Wardetzky¹ to implement the heat method directly on polygonal meshes—the only challenge in this setting is that for nonplanar polygons the gradient vector no longer has a clear geometric meaning. This issue is resolved by noting that we need only the magnitude $|\nabla u|$ of the gradient; see Crane et al.,¹¹ Section 3.2.2 for further details. Figure 9 demonstrates distance computed on an irregular polygonal mesh.

Point clouds. Raw geometric data is often represented as a discrete point sample $P \subset \mathbb{R}^n$ of some smooth surface M . Rather than convert this data into a polygon mesh, we can directly implement the heat method using the point cloud Laplacian of Liu et al.,²² which extends previous work by Belkin et al.² Computation of the gradient and divergence are described by Crane et al.,¹¹ Section 3.2.3. Other discretizations are certainly possible (see for instance the work of Luo et al.²³); we picked one that was simple to implement in any dimension. It is particularly interesting

Figure 9. Since the heat method is based on well-established discrete operators like the Laplacian, it is easy to adapt to a variety of geometric domains. Above: distance on a hippo composed of high-degree nonplanar (and sometimes nonconvex) polygonal faces.



to note that the cost of the heat method depends primarily on the intrinsic dimension n of M , whereas methods based on fast marching require a grid of the same dimension m as the ambient space²⁵—this distinction is especially important in contexts like machine learning where m may be significantly larger than n .

Choice of time step. Accuracy of the heat method relies in part on the time step t . In the smooth setting, Equation (5) suggests that smaller values of t yield better approximations of geodesic distance. In the discrete setting we instead observe the somewhat surprising behavior that the limit solution to Equation (3) depends only on the number of edges between a pair of vertices, independent of how we might try to incorporate edge lengths into our formulation—see Crane et al.,¹¹ Appendix A. Therefore, on a fixed mesh decreasing the value of t does not necessarily improve accuracy, even in exact arithmetic—to improve accuracy we must simultaneously refine the mesh and decrease t accordingly. Moreover, very large values of t produce an over-smoothed approximation of geodesic distance (Section 3.3). For a fixed mesh, we therefore seek an optimal time step t^* that is neither too large nor too small.

An optimal value of t^* is difficult to obtain due to the complexity of analysis involving the cut locus.²⁸ We instead use a simple estimate that works well in practice, namely $t = mh^2$ where h is the mean spacing between adjacent nodes and $m > 0$ is a constant. This estimate is motivated by the fact that $h^2 \Delta$ is invariant with respect to scale and refinement; numerical experiments suggest that $m = 1$ yields near-optimal accuracy for a wide variety of problems. In this paper the time step

$$t = h^2$$

is therefore used uniformly throughout all tests and examples, except where we explicitly seek a smoothed approximation of distance, as in Section 3.3. For highly nonuniform meshes one could set h to the maximum spacing, providing a more conservative estimate. Numerical underflow could theoretically occur for extremely small t , though we do not encounter this issue in practice.

Numerics. As demonstrated in Figures 10, 18, and 19, one does not need a particularly nice mesh or point cloud to get a reasonable distance function. However, as with any numerical

Figure 10. The heat method can be applied directly to scattered point clouds. Left: face scan with holes and noise. Right: kitten surface with connectivity removed. Yellow points are close to the source.



method, accuracy and other properties of the solution may be influenced by the quality of the mesh. For instance, in some applications one may wish to avoid “spurious minima,” that is, local maxima or minima that do not appear in the true (smooth) distance function. At present, there is no numerical scheme that guarantees the absence of spurious minima on arbitrary meshes, including exact polyhedral schemes.¹⁷ Empirically, however, we observe that the heat method produces fewer spurious minima than either fast marching or the biharmonic distance (see Figure 20), in part due to regularization from the Hodge step (step III). In cases where one wishes to avoid spurious minima altogether, we advocate the use of Delaunay meshes.

3.3. Smoothed distance

Geodesic distance fails to be smooth at points in the *cut locus*, that is, points at which there is no unique shortest path to the source—these points appear as sharp cusps in the level lines of the distance function. Non-smoothness can result in numerical difficulty for applications which need to take derivatives of the distance function ϕ (e.g., level set methods), or may simply be undesirable aesthetically.

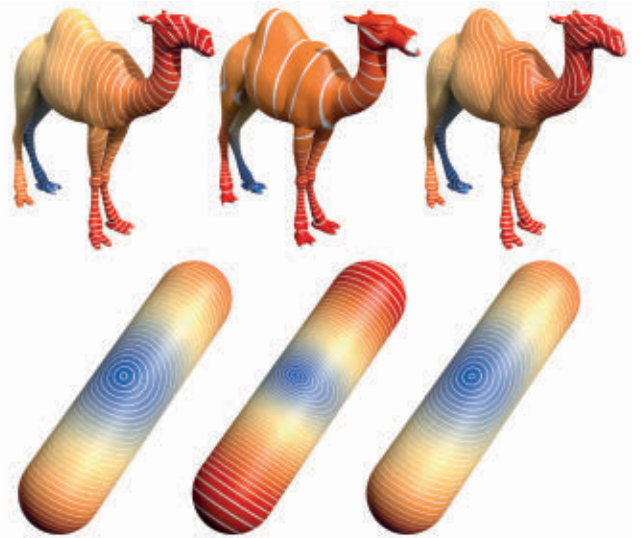
Several distances have been designed with smoothness in mind, including diffusion distance,¹⁰ commute-time distance,¹⁴ and biharmonic distance²¹ (see the last reference for a more detailed discussion). These distances satisfy a number of important properties (smoothness, isometry-invariance, etc.), but are poor approximations of true geodesic distance, as indicated by uneven spacing of isolines (see Figure 12, middle). They can also be expensive to evaluate, requiring one to either solve a linear system for each vertex, or compute a large number of eigenvectors of the Laplace matrix (~ 150 to 200 in practice).

In contrast, one can rapidly construct smoothed approximations of geodesic distance by simply applying the heat method for large values of t (Figure 11). The computational cost remains the same, and isolines are evenly spaced for

Figure 11. A source on the front of the Stanford Bunny results in nonsmooth cusps on the opposite side. By running heat flow for progressively longer durations t , we obtain smoothed approximations of geodesic distance (right).



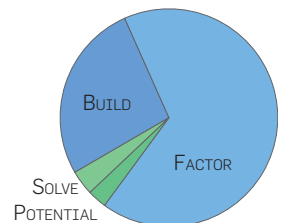
Figure 12. Top row: our smoothed approximation of geodesic distance (left) and biharmonic distance (center) both mitigate sharp “cusps” found in the exact distance (right), yet our approximation provides more even spacing of isocontours. Bottom row: biharmonic distance (center) tends to exhibit elliptical level lines near the source, while our smoothed distance (left) maintains isotropic circular profiles as seen in the exact distance (right).



any value of t due to normalization (step II); the solution is isometry invariant since it depends only on intrinsic operators. For a time step $t = mh^2$, meaningful values of m are found in the range $1 - 10^6$ —past this point the term $t\Delta$ dominates, resulting in little visible change.

3.4. Boundary conditions

To solve the equations in steps I and II, we must define the behavior of derivatives near the boundary. Intuitively, the behavior of our distance approximation should not be significantly influenced by the shape of the boundary (Figure 13)—for instance, cutting off a corner of a convex domain should not affect the distance at the points that remain. For exact distance computation, we can apply standard *zero-Neumann* or *zero-Dirichlet* boundary conditions, since this choice does not affect the behavior of the smooth limit



solution (see Renesse⁴⁴ Corollary 2 and Norri,³⁰ Theorem 1.1, respectively). Boundary conditions do however alter the behavior of our smoothed distance. Although there is no well-defined “correct” behavior for this smoothed function, we advocate the use of boundary conditions obtained by taking the mean of the Neumann solution u_N and the Dirichlet solution u_D , that is, $u = \frac{1}{2}(u_N + u_D)$. The intuition behind this behavior again stems from a random walker interpretation: zero Dirichlet conditions absorb heat, causing walkers to “fall off” the edge of the domain. Neumann conditions prevent heat from flowing out of the domain, effectively “reflecting” random walkers. Averaged conditions mimic the behavior of a domain without boundary: the number of walkers leaving equals the number of walkers returning. Figure 14 shows how boundary conditions affect the behavior of geodesics in a path-planning scenario.

Figure 13. Effect of Neumann (top-left), Dirichlet (top-right) and averaged (bottom-left) boundary conditions on smoothed distance. Averaged boundary conditions mimic the behavior of the same surface without boundary.



4. EVALUATION

4.1. Performance

A key advantage of the heat method is that the linear systems in steps I and III can be prefactored. Our implementation uses sparse Cholesky factorization,⁹ which for Poisson-type problems has guaranteed sub-quadratic complexity but in practice scales much better⁵; moreover there is strong evidence to suggest that sparse systems arising from elliptic PDEs can be solved in very close to linear time.^{34, 39} Independent of these issues, the amortized cost for problems with a large number of right-hand sides is roughly linear, since back substitution can be applied in essentially linear time. See inset for a breakdown of relative costs in our

Figure 14. For path planning, the behavior of geodesics can be controlled via boundary conditions and the time step t . Top-left: Neumann conditions encourage boundary adhesion. Top-right: Dirichlet conditions encourage avoidance. Bottom-left: small values of t yield standard straight-line geodesics. Bottom-right: large values of t yield more natural trajectories.

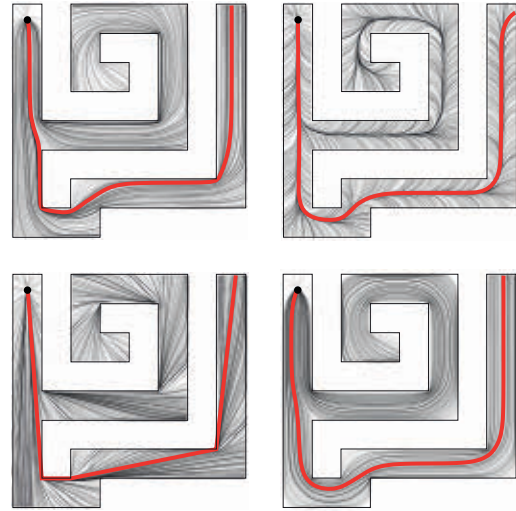


Figure 15. Meshes used in Table 1. Left to right: Bunny, Isis, Horse, Bimba, Aphrodite, Lion, Ramses.



Table 1. Comparison with fast marching and exact polyhedral distance

Model	Triangles	Heat method				Fast marching			Exact time (s)
		Precompute (s)	Solve	Max error (%)	Mean error (%)	Time (s)	Max error (%)	Mean error (%)	
Bunny	28k	0.21	0.01s (28x)	3.22	1.12	0.28	1.06	1.15	0.95
Isis	93k	0.73	0.05s (21x)	1.19	0.55	1.06	0.60	0.76	5.61
Horse	96k	0.74	0.05s (20x)	1.18	0.42	1.00	0.74	0.66	6.42
Kitten	106k	1.13	0.06s (22x)	0.78	0.43	1.29	0.47	0.55	11.18
Bimba	149k	1.79	0.09s (29x)	1.92	0.73	2.62	0.63	0.69	13.55
Aphrodite	205k	2.66	0.12s (47x)	1.20	0.46	5.58	0.58	0.59	25.74
Lion	353k	5.25	0.24s (24x)	1.92	0.84	10.92	0.68	0.67	22.33
Ramses	1.6M	63.4	1.45s (68x)	0.49	0.24	98.11	0.29	0.35	268.87

Best speed/accuracy in bold; speedup in orange.

implementation; Potential is the time taken to compute the right hand side in step III.

In practice, a number of factors affect the run time of the heat method including the choice of spatial discretization, discrete Laplacian, and geometric data structures. As a typical example, we compared the scheme from Triangle meshes section to the first-order fast marching method of Kimmel and Sethian¹⁹ and the exact algorithm of Mitchell et al.,²⁷ using the state-of-the-art fast marching implementation of Peyré and Cohen³¹ and the exact implementation of Kirsanov.⁴⁰ The heat method was implemented in ANSI C in double precision using a vertex-face adjacency list. Single-threaded performance was measured on a 2.4 GHz Intel Core 2 Duo (Table 1). Note that even for a single distance computation the heat method outperforms fast marching; more importantly, updating distance for new subsets γ is consistently an order of magnitude faster (or more) than both fast marching and the exact algorithm.

4.2. Accuracy

We examined errors in the heat method, fast marching,¹⁹ and the polyhedral distance,²⁷ relative to mean edge length h on triangulated surfaces. Both fast marching and the heat method appear to exhibit linear convergence; it is interesting to note that even the exact polyhedral distance provides only quadratic convergence. Keeping this fact in mind, Table 1 uses the polyhedral distance as a baseline for comparison on more complicated geometries—Max is the maximum error as a percentage of mesh diameter and Mean is the mean relative error at each vertex. Note that fast marching tends to achieve a smaller maximum error, whereas the heat method does better on average. Figure 16 gives a visual comparison of accuracy; the only notable discrepancy is a slight smoothing at sharp cusps, which may explain the larger maximum error. Figure 17 indicates that smoothing does not interfere with the extraction of the cut locus—here we visualize values of $|\Delta\phi|$ above a fixed threshold. Overall, the heat method exhibits errors of the same order and magnitude as fast marching (at lower computational cost) and is therefore suitable in applications where fast marching is presently used; see Crane et al.¹¹ for more extensive comparisons.

Figure 16. Visual comparison of accuracy. Left: exact polyhedral distance. Using default parameters, the heat method (middle) and fast marching (right) both produce results of comparable accuracy, here within less than 1% of the polyhedral distance—see Table 1 for a more detailed comparison.

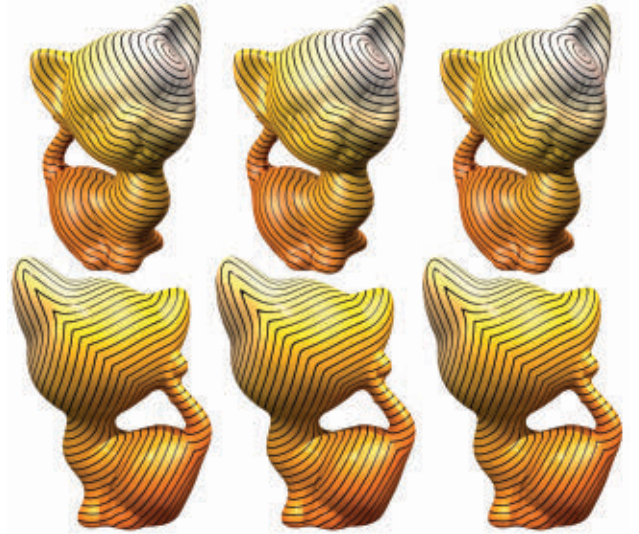
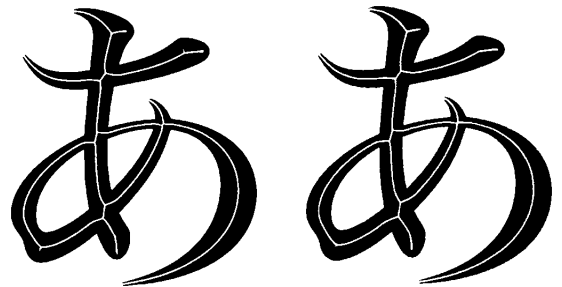


Figure 17. Medial axis of the hiragana letter “a” extracted by thresholding second derivatives of the distance to the boundary. Left: fast marching. Right: heat method.



More recent implementations of the heat method improve accuracy by using a different spatial discretization,²⁹ or by iteratively updating the solution.³ The accuracy of fast marching schemes is determined by the choice of *update*

Figure 18. Smoothed geodesic distance on an extremely poor triangulation with significant noise—note that small holes are essentially ignored. Also note good approximation of distance even along thin slivers in the nose.

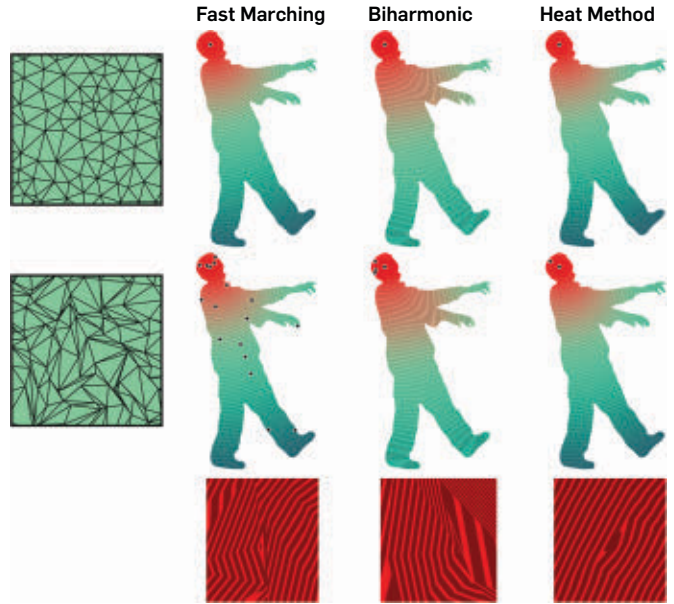


Figure 19. Tests of robustness. Left: our smoothed distance ($m = 10^4$) appears similar on meshes of different resolution. Right: even for meshes with severe noise (top) we recover a good approximation of the distance function on the original surface (bottom, visualized on noise-free mesh).



rule—a number of highly accurate rules have been developed for regular grids (e.g., HJ WENO¹⁸), but fewer options are available on irregular domains such as triangle meshes, the predominant choice being the first-order update of Kimmel and Sethian.¹⁹ Finally, the approximate algorithm of Surazhsky et al.⁴⁰ provides an interesting comparison

Figure 20. In any method based on a finite element approximation, mesh quality will affect the quality of the solution. However, because the heat method is based on solving low-order elliptic equations (rather than high-order or hyperbolic equations), it often produces fewer numerical artifacts. Here, for instance, we highlight spurious extrema in the distance function (i.e., local maxima and minima) produced by the fast marching method (left), biharmonic distance (middle), and the heat method (right) on an acute Delaunay mesh (top) and a badly degenerate mesh (bottom). Inset figures show closeup view of isolines for the bottom figure.



since it tends to produce results more accurate than fast marching at a similar computational cost. However, accuracy is measured relative to the polyhedral distance rather than the smooth geodesic distance of the approximated surface. Like fast marching, Surazhsky's method does not take advantage of precomputation and therefore exhibits a significantly higher amortized cost than the heat method; it is also limited to triangle meshes.

4.3. Robustness

Two factors contribute to the robustness of the heat method, namely (1) the use of an unconditionally stable time discretization and (2) an elliptic rather than hyperbolic formulation (i.e., relatively stable local averaging vs. more sensitive global wavefront propagation). Figure 19 verifies that the heat method continues to work well even on meshes that are poorly discretized or corrupted by a large amount of noise (here modeled as uniform Gaussian noise applied to the vertex coordinates). In this case we use a moderately large value of t to investigate the behavior of our smoothed distance; similar behavior is observed for small t values. Figure 18 illustrates the robustness of the method on a surface with many small holes as well as long sliver triangles.

5. CONCLUSION

The heat method is a simple, general method that can be easily incorporated into a broad class of algorithms. However, a

CAREERS

Auburn University

Department of Computer Science and Software Engineering (CSSE)

Multiple Faculty Positions in Cybersecurity

CSSE invites applications for multiple tenure-track faculty positions to begin in Fall 2018 or later. Candidates at the Assistant Professor level will be given preference, however outstanding candidates at senior levels will also be considered. A Ph.D. degree in computer science, software engineering or a closely related field must be completed by the start of appointment. Excellent communication skills are required. We are interested in candidates specializing in all areas related to security, such as *AI/machine learning applications to security, blockchain, cryptocurrency, cyberidentity, cyberinfrastructure and critical infrastructure protection, digital forensics, reverse engineering, secure cloud, mobile systems, networks and operating systems, secure software engineering, and securing the Internet of Things.*

CSSE is home to the Auburn Cyber Research Center (<http://cyber.auburn.edu>), and is affiliated with the McCrary Institute for Critical Infrastructure Protection and Cyber Systems (<http://mccrary.auburn.edu>). The department has 21 full-time tenure-track faculty members and supports strong M.S. and Ph.D. programs (with a new M.S. in Cybersecurity Engineering projected to start in Fall 2018). Faculty research areas include artificial intelligence, architecture, computational biology, computer science education, data science, energy-efficient systems, human-computer interaction, Internet of Things, learning science, machine learning, modeling and simulation, multi-agent systems, networks, security, software engineering and wireless engineering.

Auburn University is one of the nation's premier public land-grant institutions. It is ranked 46th among public universities in the U.S. News and World Report 2018 Rankings. It is nationally recognized for its commitment to academic excellence, a positive work environment, student engagement, and its beautiful campus. Auburn residents enjoy a thriving community, recognized as one of the "best small towns in America." The city is located on the rapidly developing I-85 corridor between Atlanta, GA, and Montgomery, AL. The Auburn City school system is ranked as one of the top school systems in the nation and the state. A nationally recognized hospital, East Alabama Medical Center, is located close by in Opelika. The Auburn-Opelika metropolitan area has a population of over 150,000.

Applicants should submit a cover letter, curriculum vita, research vision, teaching philosophy, and the names of references through the faculty hiring link on the department home page <http://www.eng.auburn.edu/csse>. There is no application deadline. Application review will begin in November. Selected candidates must be able to meet eligibility requirements to work legally in the United States at the time of appointment for the proposed term of employment. Auburn University is an EEO/Vet/Disability Employer.

Baylor University

Chairperson, Department of Computer Science

The School of Engineering and Computer Science invites nominations and applications for the position of Chair of the Department of Computer Science. The successful candidate must have an earned Ph.D. in Computer Science or a related field, leadership experience, a commitment to undergraduate and graduate education, a strong research record that includes significant external funding, and effective communication and organization skills.

The Department: Computer Science is one of three departments in the School of Engineering and Computer Science. It offers an ABET-accredited B.S. in Computer Science degree, a B.A. degree with a major in Computer Science, a B.S. in Informatics with a major in Bioinformatics, B.S. in Computing with a major in Computer Science Fellows, and M.S. and Ph.D. degrees in Computer Science. The Department has 15 full-time faculty, over 370 undergraduate majors and 20 graduate students. The Department's greatest strength is its dedication to the success of the students and each other. Interested candidates may contact any faculty member to ask questions and/or visit the departmental web site at <http://www.ecs.baylor.edu/computerscience>.

The University: Baylor University is a private Christian university and a nationally-ranked research institution, consistently listed with highest honors among The Chronicle of Higher Education's "Great Colleges to Work For." The university is recruiting new faculty with a deep commitment to excellence in teaching, research and scholarship. Baylor seeks faculty who share in our aspiration to become a tier-one research institution while strengthening our distinctive Christian mission as described in our strategic vision, Pro Futuris (<https://www.baylor.edu/profuturis>). As the world's largest Baptist University, Baylor offers over 40 doctoral programs and has over 17,000 students from all 50 states and more than 80 countries.

Appointment Date: Fall 2018.

Application Procedure: Applications, including detailed curriculum vitae, a statement demonstrating an active Christian faith, and contact information for three references should be either emailed to CSSearch@Baylor.edu or mailed to Chair Search Committee, Department of Computer Science, Baylor University, One Bear Place #97141, Waco, TX 76798-7141.

For full consideration, applications should be received by January 1, 2018. However, applications will be accepted until the position is filled.

Boston College

Assistant Professor of the Practice or Lecturer

The Computer Science Department of Boston College aims to grow substantially over the next

several years, and will seek to fill faculty positions at all levels. We invite applications for a full-time, non-tenure-track faculty position, beginning in the 2018-2019 academic year. Candidates should be committed to excellence in undergraduate education, and should be able to both teach a broad variety of undergraduate computer science courses, and to participate in the development of new courses that reflect the evolving landscape of the discipline.

Minimum requirements for the title of Assistant Professor of the Practice include a Ph.D. in Computer Science or closely related discipline. Candidates who have only attained a Master's degree would be eligible for the title of Lecturer.

Application review begins October 1, 2017. See www.cs.bc.edu for more information.

To apply go to <http://apply.interfolio.com/44984>.

Boston College

Associate or Full Professor of Computer Science

The Computer Science Department of Boston College aims to grow substantially over the next several years, and will seek to fill faculty positions at all levels. We invite applications for a senior-level (Associate or Full Professor) position, starting in the 2018-2019 academic year. The successful candidate is expected to play a leadership role in the creation of a Data Science program in conjunction with the new interdisciplinary Institute for Integrated Science and Society, recently announced at Boston College, and will also participate in shaping the future of our growing department.

Applicants must have a Ph.D. in Computer Science or closely related discipline, a record of strong research accomplishment and external funding, and a commitment to quality in undergraduate and graduate education. Preference will be given to candidates whose research is in the areas of high-performance data mining / machine learning or data visualization, particularly those with a history of interdisciplinary collaboration, but outstanding candidates in all research areas will be considered.

Application review begins October 1, 2017. See www.cs.bc.edu for more information.

To apply go to <http://apply.interfolio.com/44982>.

Boston College

Tenure Track Assistant Professor in Computer Science

The Computer Science Department of Boston College aims to grow substantially over the next several years and will seek to fill faculty positions at all levels. We invite applications for one or more tenure-track faculty positions at the rank of

Assistant Professor, beginning in the 2018-2019 academic year. Successful candidates will be expected to develop strong research programs that can attract external research funding. The search will focus on candidates who can participate in cross-disciplinary research in conjunction with the new Institute for Integrated Science and Society recently announced at Boston College, in the areas of high-performance data mining / machine learning, systems / networks, data visualization, and human-computer interaction. However, outstanding candidates in all research areas will be considered.

Minimum requirements for all positions include a Ph.D. in Computer Science or closely related discipline, an energetic research program that promises to attract external funding, and a commitment to quality in undergraduate and graduate education.

Application review begins October 1, 2017. See www.cs.bc.edu for more information.

To apply go to <http://apply.interfolio.com/44980>.

California Institute of Technology (Caltech)

Tenure-Track Faculty Position

The Computing and Mathematical Sciences (CMS) department at the California Institute of Technology (Caltech) invites applications for a tenure-track faculty position. CMS is a unique environment where innovative, interdisciplinary, and foundational research is conducted in a collegial atmosphere. Candidates in all areas of ap-

plied and computational mathematics, computer science and statistics are invited to apply. Areas of interest include (but are not limited to) scientific computing, optimization, statistics, probability, networked systems, control and dynamical systems, robotics, theory of computation, security, privacy, machine learning, and algorithmic economics. In addition, we welcome applications from candidates who have demonstrated strong connections between computer science, engineering and applied mathematics, and to other fields such as the physical, biological, and social sciences.

A commitment to world-class research as well as high-quality teaching and mentoring is expected. The initial appointment at the Assistant-Professor level is for four years and is contingent upon the completion of a Ph.D. degree in Applied Mathematics, Computer Science, Engineering, or a related field.

Applications will be reviewed beginning November 15, 2017, and applicants are encouraged to have all their application materials on file by this date. For a list of documents required and full instructions on how to apply on-line, please visit <http://www.cms.caltech.edu/search>. Questions about the application process may be directed to: search@cms.caltech.edu.

We are an equal opportunity employer and all qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, or national origin, disability status, protected veteran status, or any other characteristic protected by law.

California State University – Sacramento

Tenure-Track Assistant Professor

California State University – Sacramento, Department of Computer Science. Three tenure-track assistant professor positions to begin with the Fall 2018 semester. Applicants specializing in any area of computer science will be considered. Those with expertise in areas related to software engineering, embedded systems, or artificial intelligence are especially encouraged to apply. Ph.D. in Computer Science, Computer Engineering, or closely related field required by the time of the appointment. For detailed position information, including application procedure, please see <http://www.csus.edu/about/employment/>. Screening will begin November 19, 2017, and remain open until filled. AA/EEO employer. Clery Act statistics available. Mandated reporter requirements. Criminal background check will be required.

Carnegie Mellon University Faculty Hiring

The School of Computer Science consists of seven departments, spanning a wide range of topics in computer science and the application of computers to real-world systems. Faculty positions are specific to each department, though in certain cases, joint positions are also possible.

We are seeking tenure, research, and systems track faculty candidates with a strong interest in research, an earned Ph.D., and outstanding aca-

ETH zürich

Professor or Assistant Professor (Tenure Track) of Analog and Mixed Signal Interfaces

→ The Department of Information Technology and Electrical Engineering (www.ee.ethz.ch) at ETH Zurich invites applications for the above-mentioned position.

→ The successful candidate is expected to develop a strong and visible research program in the area of analog and mixed signal interfaces circuits and systems. He or she has a strong background in one or more of the following areas: (i) analog circuits and techniques for ultra-low power, ranging from basic building blocks (e.g. amplifiers, filters) to silicon sensors, interfaces, and novel clock generation architectures; (ii) data converters enabled by new techniques, architectures, or circuit topologies; (iii) wireless transceiver and RF circuits for low-power and energy-efficient links, cellular connectivity including machine-to-machine applications, emerging broadband networks [5G], millimeter-wave and THz systems (radar, sensing and imaging); (iv) wireline communications circuits and systems for chip-to-chip communication, including serial links, high-speed memory, accelerator, peripheral interfaces, backplanes, long-haul, and powerline communications. Candidates should hold a PhD. A track record of successfully manufactured chips and systems is highly desirable. In addition, commitment to teaching and the ability to lead a research group are expected. Generally, at ETH Zurich undergraduate level courses are taught in German or English and graduate level courses are taught in English.

→ Assistant professorships have been established to promote the careers of younger scientists. ETH Zurich implements a tenure track system equivalent to other top international universities. The level of the appointment will depend on the successful candidate's qualifications.

→ Please apply online: www.facultyaffairs.ethz.ch

→ Applications should include a curriculum vitae, a list of publications, a statement of future research and teaching interests, and a description of the three most important achievements. The letter of application should be addressed to the President of ETH Zurich, Prof. Dr. Lino Guzzella. The closing date for applications is 15 December 2017. ETH Zurich is an equal opportunity and family friendly employer and is responsive to the needs of dual career couples. We specifically encourage women to apply.

demographic credentials. Candidates for tenure track appointments should also have a strong interest in graduate and undergraduate education.

We are also seeking teaching track faculty candidates. You should have a Ph.D. in Computer Science or a related computing discipline, a background of demonstrated excellence and dedication to teaching, the ability to collaborate with other faculty in a fast-paced environment, and must be prepared to teach in a wide variety of settings, including large undergraduate lecture courses and classes delivered in non-traditional formats.

Candidates with a commitment toward building an equitable and diverse scholarly community are particularly encouraged to apply. We are very interested in applications from candidates who have a demonstrated track record in mentoring and nurturing women and students from groups traditionally underrepresented in computer science.

To ensure full consideration of your application, please submit all materials no later than December 15, 2017. In your cover letter, please indicate clearly the department(s) you are applying to. You can learn more about our hiring plans and application instructions by visiting <http://www.cs.cmu.edu/employment-scs>.

For more information about the hiring priorities in a particular department, please visit a department site below:

Computational Biology Department: <http://www.cbd.cmu.edu/tenure-track-faculty-positions/>

Computer Science Department: <https://www.csd.cs.cmu.edu/careers/faculty-hiring>

Human-Computer Interaction Institute:

<https://hcii.cmu.edu/careers/list>

Institute for Software Research: <http://www.isri.cmu.edu/jobs/index.html>

Language Technologies Institute: <http://lti.cs.cmu.edu/news/lti-hiring>

Machine Learning Department: http://www.ml.cmu.edu/Faculty_Hiring.html

Robotics Institute: <http://ri.cmu.edu/about/hiring-faculty-positions/>

Please send email to faculty-search@cs.cmu.edu with any questions.

Carnegie Mellon University shall abide by the requirements of 41 CFR §§ 60-1.4(a), 60-300.5(a) and 60-741.5(a). These regulations prohibit discrimination against qualified individuals based on their status as protected veterans or individuals with disabilities, and prohibit discrimination against all individuals based on their race, color, religion, sex, or national origin. Moreover, these regulations require that covered prime contractors and subcontractors take affirmative action to employ and advance in employment individuals without regard to race, color, religion, sex, national origin, protected veteran status or disability.

Cornell University Multiple Tenure-Track Faculty Positions

The SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING at CORNELL UNIVERSITY, Ithaca, New York, invites applications for multiple tenure-track Faculty positions in **all areas** of

electrical and computer engineering, as well as interdisciplinary areas such as robotics and cyber-physical systems, bio-ECE, microsystems, machine learning, applied mathematics, and energy. See our website, <https://www.ece.cornell.edu/academics>, for additional information on our programs.

Applicants must hold a doctorate in an appropriate field by the time their Faculty appointment would begin, must have demonstrated an ability to conduct outstanding research, and must show promise for excellence in teaching. Although we anticipate filling positions at the assistant professor level, applications at all levels will be considered; salary and rank will be commensurate with qualifications and experience.

Applicants should submit a curriculum vita, a research statement, a teaching statement, three recent publications, and complete contact information for at least three references. Applications must be submitted on-line at <https://academicjobsonline.org/ajob/jobs/9871>. Review of applications will begin immediately. Applications received by **December 4, 2017**, will receive full consideration.

The School of Electrical and Computer Engineering is especially interested in qualified candidates who can contribute, through their research, teaching, and/or service, to the diversity of the academic community and to creating a climate that attracts students of all races, genders and nationalities. We strongly encourage underrepresented minority and women candidates to apply. Cornell University actively

ETH zürich

Professor or Assistant Professor (Tenure Track) of Embedded Information Systems

→ The Department of Information Technology and Electrical Engineering (www.ee.ethz.ch) at ETH Zurich invites applications for the above-mentioned position.

→ The successful candidate is expected to develop a strong and visible research program in the area of embedded information systems. He or she has a strong background in areas such as embedded software, edge computing, embedded operating systems, real-time systems, biomedical embedded systems, security of embedded systems, as well as corresponding theoretical concepts. Candidates should hold a PhD and have an excellent record of accomplishments. In addition, commitment to teaching and the ability to lead a research group are expected. Generally, at ETH Zurich undergraduate level courses are taught in German or English and graduate level courses are taught in English.

→ Assistant professorships have been established to promote the careers of younger scientists. ETH Zurich implements a tenure track system equivalent to other top international universities. The level of the appointment will depend on the successful candidate's qualifications.

→ **Please apply online: www.facultyaffairs.ethz.ch**

→ Applications should include a curriculum vitae, a list of publications, a statement of future research and teaching interests, and a description of the three most important achievements. The letter of application should be addressed **to the President of ETH Zurich, Prof. Dr. Lino Guzzella**. The closing date for applications is **15 December 2017**. ETH Zurich is an equal opportunity and family friendly employer and is responsive to the needs of dual career couples. We specifically encourage women to apply.

seeks to meet the needs of dual career couples, has a Dual Career program, and is a member of the Upstate New York Higher Education Recruitment Consortium to assist with dual career searches.

Diversity and Inclusion are a part of Cornell University's heritage. We are a recognized employer and educator valuing AA/EEO, Protected Veterans and Individuals with Disabilities.

Duke University

Tenure-Track Faculty Positions in Computing

Duke University invites applications and nominations for five tenure-track or tenured faculty positions in computing, at all ranks, to begin July 2018. This search is a joint initiative between the Department of Computer Science and the Department of Electrical and Computer Engineering to rapidly expand the university's existing strengths and to support exciting new initiatives in computing. We encourage applications in all areas of computer science and engineering, with special interest in the following themes:

► Two positions in all disciplinary areas of computer science, including but not limited to security and privacy, distributed systems and networking, mobile and embedded systems, machine learning, algorithms, as well as interdisciplinary work that relates to social sciences or biological sciences.

► Two joint positions between computer science and electrical and computer engineering in the area of machine learning (with an applied and methodological focus).

► One position in all disciplinary areas of computer engineering, including but not limited to security and privacy, distributed systems and networking, as well as mobile and embedded systems.

Candidates are expected to have a doctoral degree in computer science, computer engineering, or a related discipline. A successful candidate must have a solid disciplinary foundation and demonstrate promise of outstanding scholarship in every regard, including research and teaching.

The Duke faculty and students communities are currently very diverse and are strongly committed to further enhancing this diversity. We seek faculty members who are committed to building a diverse and inclusive community, which fosters excellence in research and teaching. We strongly encourage applications from women and underrepresented minorities in computing. Please see www.cs.duke.edu, www.ece.duke.edu, and www.provost.duke.edu/faculty/ for information about the departments and advantages that Duke offers to faculty.

Applicants should submit their materials (cover letter, research statement, teaching statement, contacts for at least three references) electronically through AcademicJobsOnline <https://academicjobsonline.org/ajob/jobs/9908>

For full consideration, applications and references should be received by December 15, 2017.

Duke University is an Affirmative Action/Equal Opportunity Employer committed to providing employment opportunity without regard to an individual's age, color, disability, genetic information, gender, gender identity, national origin, race, religion, sexual orientation or veteran status.

Durham, Chapel Hill, and the Research Triangle of North Carolina are frequently ranked among the best places in the country to live and work. Duke and the many other universities in the area offer a wealth of education and employment opportunities for spouses and families.

Georgia Institute of Technology Tenured-Tenure Track Faculty Positions

The School of Computer Science in the College of Computing at the Georgia Institute of Technology invites applications for tenure-track faculty positions. We are primarily seeking junior-level candidates at the Assistant Professor level, but truly exceptional candidates at all levels may also be considered. Applicants should have completed or be near completion of a Ph.D. in computer science or a related field, and should demonstrate potential for excellence in research and teaching. The School of Computer Science, one of three schools in the College of Computing, focuses on research that makes computing and communication smart, fast, reliable, and secure, with research groups in computer architecture, databases, machine learning, networking, programming languages, security, software engineering, systems, and theory. We seek candidates that can add to or enhance the current research areas of the school, and we put more emphasis on excellence than on candidates' specific area of expertise.

Applications will be considered until open positions are filled. However, to receive full

consideration, applications should be submitted by December 15, 2017. Applicants are encouraged to clearly identify in their cover letter the area(s) that best describe their research interests. All applications must be submitted online at <https://www.scs.gatech.edu/content/cs-faculty-hiring>. More information about the School of Computer Science is available at <http://scs.gatech.edu>. In addition to collaborations within our school and the two other schools in the college, our faculty works closely with other world-class faculty in the university, including those in the Colleges of Engineering, and Sciences.

Georgia Tech is located in the heart of the city of Atlanta, the cultural and economic center of the Southeastern U.S. The greater Atlanta area is very cosmopolitan, home to many large corporations including Coca Cola, Delta, Home Depot, and UPS, and is listed among the top 10 metropolitan areas in the United States. Georgia Tech is an Affirmative Action/Equal Opportunity Employer. Applications from women and underrepresented minorities are strongly encouraged.

Georgia Institute of Technology, School of Computational Science and Engineering Tenure-Track Faculty

Computational Science and Engineering solves real-world problems in science, engineering, health, and social domains, by using high-performance computing, modeling and simulation, large-scale "big data" analytics, and machine learning. The School of Computational Science



TENURE-TRACK AND TENURED POSITIONS

ShanghaiTech University invites highly qualified candidates to fill multiple tenure-track/tenured faculty positions as its core founding team in the School of Information Science and Technology (SIST). We seek candidates with exceptional academic records or demonstrated strong potentials in all cutting-edge research areas of information science and technology. They must be fluent in English. English-based overseas academic training or background is highly desired.

ShanghaiTech is founded as a world-class research university for training future generations of scientists, entrepreneurs, and technical leaders. Boasting a new modern campus in Zhangjiang Hightech Park of cosmopolitan Shanghai, ShanghaiTech shall trail-blaze a new education system in China. Besides establishing and maintaining a world-class research profile, faculty candidates are also expected to contribute substantially to both graduate and undergraduate educations.

Academic Disciplines: Candidates in all areas of information science and technology shall be considered. Our recruitment focus includes, but is not limited to: computer architecture, software engineering, database, computer security, VLSI, solid state and nano electronics, RF electronics, information and signal processing, networking, security, computational foundations, big data analytics, data mining, visualization, computer vision, bio-inspired computing systems, power electronics, power systems, machine and motor drive, power management IC as well as inter-disciplinary areas involving information science and technology.

Compensation and Benefits: Salary and startup funds are highly competitive, commensurate with experience and academic accomplishment. We also offer a comprehensive benefit package to employees and eligible dependents, including on-campus housing. All regular ShanghaiTech faculty members will join its new tenure-track system in accordance with international practice for progress evaluation and promotion.

Qualifications:

- Strong research productivity and demonstrated potentials;
- Ph.D. (Electrical Engineering, Computer Engineering, Computer Science, Statistics, Applied Math, or related field);
- A minimum relevant (including PhD) research experience of 4 years.

Applications: Submit (in English, PDF version) a cover letter, a 2-page research plan, a CV plus copies of 3 most significant publications, and names of three referees to: sist@shanghaitech.edu.cn. For more information, visit <http://sist.shanghaitech.edu.cn/NewsDetail.asp?id=373>

Deadline: The positions will be open until they are filled by appropriate candidates.

and Engineering of the College of Computing at the Georgia Institute of Technology seeks tenure-track faculty at all levels. Our school seeks candidates who may specialize in a broad range of application areas including biomedical and health; urban systems and smart cities; social good and sustainable development; materials and manufacturing; and national security. Applicants must have an outstanding record of research, a sincere commitment to teaching, and interest in engaging in substantive interdisciplinary research with collaborators in other disciplines.

Georgia Tech is located in the heart of metro Atlanta, a home to more than 5.5 million people and nearly 150,000 businesses, a world-class airport, lush parks and green spaces, competitive schools and numerous amenities for entertainment, sports and restaurants that all offer a top-tier quality of life. From its diverse economy, global access, abundant talent and low costs of business and lifestyle, metro Atlanta is a great place to call "home." Residents have easy access to arts, culture, sports and nightlife, and can experience all four seasons, with mild winters that rarely require a snow shovel.

Applications should be submitted online through <https://academicjobsonline.org/ajob/jobs/9687>. For best consideration, applications are due by December 15, 2017. The application material should include a full academic CV, a personal narrative on teaching and research, a list of at least three references and up to three sample publications.

Georgia Tech is an Affirmative Action/Equal Opportunity Employer. Applications from wom-

en and under-represented minorities are strongly encouraged.

For more information about Georgia Tech's School of Computational Science and Engineering please visit: <http://www.cse.gatech.edu/>.

Le Moyne College

Assistant Professor / Professor of Practice – Cybersecurity

The College of Arts and Sciences at Le Moyne College invites applications for a full-time faculty position in its new interdisciplinary cybersecurity program to begin in August, 2018. The College is seeking a candidate with a Ph.D. in cybersecurity or cognate field, to fill a tenure-track Assistant Professor position (other ranks will be considered). Candidates with significant industry experience in cybersecurity are also encouraged to apply as a Professor of Practice.

The College's new cybersecurity program presents a holistic approach to thinking about cybersecurity issues. It is designed to give students foundational knowledge regarding the varied cybersecurity challenges that individuals and organizations face on a daily basis. It will prepare students for graduate programs and careers that include international relations, legal studies, government (federal, state and local), criminology, military, security compliance, and cybersecurity technology specialist. This new cybersecurity program includes three concentrations: Crime, Society & Culture; Information & System Security; and, Policy & Law. The responsibilities of the suc-

cessful candidate will be to develop and teach courses that support the Information & System Security concentration, as well as collaborating with faculty in computer science, political science, criminology, and sociology on introductory courses that explain the connections between these disciplines.

Le Moyne College strives for academic excellence through its comprehensive programs rooted in the liberal arts and sciences. The College supports a quality undergraduate learning experience with small class sizes, and faculty that are approachable and attentive to student advising.

For more information and how to apply visit our website at www.lemoyne.edu/employment. Preference given to applications received by December 8, 2017. Review will continue until position is filled. Le Moyne College is an equal opportunity employer and encourages women, persons of color, and Jesuits to apply for employment.

Marist College

Assistant/Associate Professor, Computing Technology

The Marist College Department of Computing Technology currently seeks applications for tenure-track positions in Computer Science and Information Technology & Systems.

We welcome candidates who have the following teaching and research interests: software development and systems analysis/design, data and information security, cloud computing and networks, or data science and analytics. Applicants with the ability to teach in multiple areas across several disciplines will receive preference. Applicants must be willing to teach undergraduate and graduate courses in both traditional on-ground and on-line environments. Required duties outside the classroom include scholarly activities that result in peer-reviewed publications as well as engagement in college and professional services such as advising/mentoring students, serving on department, school, or college committees.

Candidates must have a doctoral degree in Computer Science, Information Systems, or a closely related field. We will consider ABDs in an appropriate field who will complete their dissertation within one year of hire. Evidence of excellence in teaching and scholarly work is required. Excellent written and oral communication skills are required. Industry and/or consulting experience is highly desirable. As our programs host a diverse population, the proven ability to work effectively in a multicultural environment is highly regarded.

Marist College is an independent and comprehensive liberal arts institution located in New York's historic Hudson River Valley. Situated on 210 acres overlooking the Hudson River, it enrolls 4,962 traditional undergraduate, 953 full and part-time graduate and 412 continuing education students. Marist has a branch campus in Florence, Italy, and extension sites throughout New York. Marist is recognized for excellence by U.S. News & World Report, TIME Magazine, The Princeton Review's The Best 376 Colleges, and Barron's Best Buys in College Education and is noted for being a pioneer in the area of on-line degree programs.

To learn more or to apply, please visit <http://>



ADVERTISING IN CAREER OPPORTUNITIES

How to Submit a Classified Line Ad: Send an e-mail to acmm mediasales@acm.org. Please include text, and indicate the issue/ or issues where the ad will appear, and a contact name and number.

Estimates: An insertion order will then be e-mailed back to you. The ad will be typeset according to CACM guidelines. NO PROOFS can be sent. Classified line ads are NOT commissionable.

Deadlines: 20th of the month/2 months prior to issue date. For latest deadline info, please contact:

acmm mediasales@acm.org

Career Opportunities Online: Classified and recruitment display ads receive a free duplicate listing on our website at:

<http://jobs.acm.org>

Ads are listed for a period of 30 days.

For More Information Contact:

ACM Media Sales

at 212-626-0686 or

acmm mediasales@acm.org

jobs.marist.edu. Only on-line applications are accepted. Review of applications will begin immediately and will continue until the position is filled. Marist College is strongly committed to the principle of diversity and is especially interested in receiving applications from members of ethnic and racial minority groups, women, persons with disabilities, and persons from other under-represented groups.

AN EQUAL OPPORTUNITY/AFFIRMATIVE ACTION EMPLOYER

Max Planck Institute for Software Systems (MPI-SWS)

Tenure-Track Openings

Applications are invited for faculty positions at all career stages in computer science, with a particular emphasis on systems (broadly construed). We expect multiple positions to be filled in systems, but exceptional candidates in other areas of computer science are also strongly encouraged to apply.

A doctoral degree in computer science or related areas and an outstanding research record (commensurate for the applicant's career stage) 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 is part of a network of over 80 Max Planck Institutes, Germany's premier basic-research organisations. 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 full academic freedom, lead a team of doctoral students and post-docs, and have the opportunity to teach university courses; at the same time, they enjoy ongoing institutional funding in addition to third-party funds, a technical infrastructure unrivaled for an academic institution, as well as internationally competitive compensation.

The institute is located in the German cities of Saarbrücken and Kaiserslautern, 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 on our application website (apply.mpi-sws.org). To receive full consideration, applications should be received by December 1st, 2017.

The institute is committed to increasing the representation of minorities, women, and individuals with physical disabilities. 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.

Missouri State University

Computer Science Department Head

The Department of Computer Science at Missouri State University seeks a Department Head. As well as administrative duties, the department head

will need to participate in teaching, research and service. Missouri State University (MSU) is located in Springfield, Missouri. More information about MSU can be found at: <http://www.missouristate.edu/>. Information about the department, its programs and research endeavors can be found at: <http://computerscience.missouristate.edu/undergraduate/>.

Review of applications will begin November 15, 2017 and continue until finalists are identified.

To see the required qualifications, complete list of duties and online application go to: <https://jobs.missouristate.edu/>.

Employment will require a criminal background check at University expense.

Missouri State University is an equal opportunity/affirmative action/minority/female/veterans/disability/sexual orientation/gender identity employer and institution. We encourage applications from all interested minorities, females, veterans, individuals with disabilities, and sexual orientation/gender identity.

Purdue University

Tenure-Track and Tenured Positions at the Assistant, Associate and Full Professor Levels

The Department of Computer Science at Purdue University is in a phase of significant growth. Applications are being solicited for nine tenure-track and tenured positions at the Assistant, Associate and Full Professor levels. Outstanding candidates in all areas of computer science will be considered. Review of applications and candidate interviews will begin in September 2017, and will continue until the positions are filled.

The Department of Computer Science offers a stimulating academic environment. Information about the department and a description of open positions are available at <http://www.cs.purdue.edu>.

Applicants should hold a PhD in Computer Science or a related discipline, have demonstrated excellence in research, and strong commitment to teaching. Successful candidates will be expected to conduct research in their fields of expertise, teach courses in computer science, and participate in other department and university activities. Purdue University's Department of Computer Science is committed to advancing diversity in all areas of faculty effort, including scholarship, instruction, and engagement. Candidates should address at least one of these areas in their cover letter, indicating their past experiences, current interests or activities, and/or future goals to promote a climate that values diversity, and inclusion. Salary and benefits are competitive, and Purdue is a dual-career friendly employer. Applicants are strongly encouraged to apply online at <https://hiring.science.purdue.edu>. Alternatively, hardcopy applications can be sent to: Faculty Search Chair, Department of Computer Science, 305 N. University Street, Purdue University, West Lafayette, IN 47907.

A background check will be required for employment in this position. Purdue University is an EOE/AA employer. All individuals, including minorities, women, individuals with disabilities, and veterans are encouraged to apply.

Requirements: PhD in Computer Science, or a closely related discipline, be committed to excellence in teaching, and have demonstrated potential for excellence in research.

San José State University

Assistant or Associate Professor

The Computer Engineering Department at San José State University (SJSU) invites applications for two tenure-track faculty positions at the rank of Assistant or Associate Professor (10 mo. Apt.) Areas of interest include virtual and augmented reality, machine learning and artificial intelligence, robotics, cloud computing and virtualization, big data, networking, mobile systems, cyber security, computer systems architecture, FPGA, and embedded systems, other areas in computer and software engineering will be considered. Applicants must have a doctorate in Computer/Software Engineering, Computer Science, Electrical Engineering or a closely related field by the start of the appointment. For more information about the position and to apply, go to: <http://apptkr.com/1081339>.

SJSU is an Affirmative Action/ Equal Opportunity committed to the core values of inclusion, civility, and respect for each individual. A background check (including a criminal records check) must be completed satisfactorily before any candidate can be offered a position with the CSU.

San José State University

Four Tenure-Track Positions

Rank: Assistant Professor (tenure-track) Job Opening ID (JOID): 24229

Qualifications:

- Applicants must have earned their Ph.D. in Computer Science or a closely related field and have demonstrated excellence in teaching and scholarship.
- 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.
- Preference will be given to candidates with teaching, research, and/or industry experience in the following areas: Data Structures and Algorithms, Data Science, Databases, Computer Graphics, Human Computer Interaction, Cybersecurity, and Computer Science Education.
- Special attention will be given to candidates with teaching and research experience in interdisciplinary fields.

Responsibilities:

The department's faculty members are expected to teach, supervise, and advise students in both the undergraduate and graduate programs, and to establish a successful research program related to his/her field of interest. Additionally, faculty members will participate in department, college, and university committee and other service assignments. Candidate 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 17, 2018.

Eligibility: Employment is contingent upon proof of eligibility to work in the United States.

Application Procedures: All materials are due by Friday, December 15, 2017. For full consideration upload a letter of application, curriculum vitae, statement of teaching interests/philosophy, description of research plans, and at least three letters of reference with contact information at apply.interfolio.com/44419.

Please include Job Opening ID (JOID) on all communications.

Search Committee Chair,
Department of Computer Science
San José State University
One Washington Square
San José, CA 95192-0249

San José State University is California's oldest institution of public higher learning. The campus is located on the southern end of San Francisco Bay in downtown San José (Pop. 967,000), hub of the world famous Silicon Valley high-technology research and development center. SJSU is an Equal Opportunity/Affirmative Action Employer committed to the core values of inclusion, civility, and respect for each individual.

For further details, please visit www.cs.sjsu.edu.

Santa Clara University **Two Tenure-Track Assistant Professors in** **Computer Science**

The Department of Mathematics and Computer Science at Santa Clara University invites applications for two tenure-track assistant professor positions in computer science. Our highest preference is in candidates with research interests in an area related to cybersecurity for the first position and an area related to algorithms for the second position. Strong candidates with research interests in artificial intelligence and software aspect of data science will be considered as well. The successful candidates will demonstrate not only potential for excellent undergraduate teaching, but also promise in sustained research with opportunities to involve undergraduates, mentoring or recruiting underrepresented groups in computer science, and service to the department, College or University. **Positions available starting in September 2018.** Ph.D. or equivalent required by September 2018.

The closing date for applications is December 1, 2017 at 3 pm Pacific time. Undergraduate teaching only.

Santa Clara University, located in California's Silicon Valley, is a comprehensive, Jesuit, Catholic university, and an AA/EEO employer.

For more information or to apply, visit <http://jobs.scu.edu/postings/6211>.

Southern Methodist University **Chair and Professor, Computer Science and** **Engineering**

Southern Methodist University (SMU) invites nominations and applications for the position of Chair and Professor, Computer Science and Engineering (Position No: 5781). It is expected that the appointment will be at the tenured Full Professor level. The Chair will be the intellectual leader of the Computer Science and Engineering Department with strong interest in educational programs at the BS, MS, and PhD levels, and will

be instrumental in the continuing development of a world-renowned interdisciplinary research program synergistic with the ongoing research in the Department and the Lyle School of Engineering. He/she will possess strong administrative skills and will be an outstanding communicator representing the Computer Science and Engineering Department and the Lyle School on- and off-campus. The anticipated starting date is on or before August 1, 2018. Candidates must have a Ph.D. degree in Computer Science or a closely related field and must be qualified for a tenured appointment at the full Professor level.

The Computer Science and Engineering (CSE) department resides within the Lyle School of Engineering and is located in the Caruth Hall Engineering Building, a LEED Gold designed facility. The CSE department offers BS, MS, and Ph.D. degrees in Computer Science and in Computer Engineering, BA in Computer Science, MS in Software Engineering, MS in Security Engineering, and D.Eng in Software Engineering. To learn more about the rich cultural environment of SMU, please see: <http://smu.edu>.

While applications and nominations will be accepted until a new Chair and Professor of Computer Science and Engineering is selected, interested parties are encouraged to **submit their application by November 30, 2017** electronically to CSEChair@smu.edu. Applications should include a cover letter, curriculum vitae, list of five references and a statement of interest and capabilities related to academic leadership, education and research. The anticipated start date for this position is on or before August 1, 2018. Hiring is contingent upon the satisfactory completion of a background check. SMU will not discriminate in any employment practice, education program, or educational activity on the basis of race, color, religion, national origin, sex, age, disability, genetic information, or veteran status. SMU will not discriminate in any program or activity on the basis of race, color, religion, national origin, sex, age, disability, genetic information, veteran status, sexual orientation, or gender identity and expression. The Executive Director for Access and Equity/Title IX Coordinator is designated to handle inquiries regarding nondiscrimination policies and may be reached at the Perkins Administration Building, Room 204, 6425 Boaz Lane, Dallas, TX 75205, 214-768-3601, accessequity@smu.edu.

Swarthmore College **Tenure-Track and Visiting Faculty Positions in** **Computer Science**

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 2018.

For the tenure-track position, we are interested in applicants whose areas fit broadly into systems (including but not limited to operating systems, security, or high-performance computing) or programming languages. Priority will be given to complete applications received by November 17, 2017. For the visiting position, strong applicants in any area will be considered. Priority will be given to complete applications received by February 2, 2018.

Applications for both positions will continue to be accepted after these dates until the positions are filled.

The Computer Science Department currently has eight tenure-track faculty and four visiting faculty. Faculty teach introductory courses as well as advanced courses in their research areas. We have grown significantly in both faculty and students in the last five years. Presently, we are one of the most popular majors at the College and expect to have over 70 Computer Science majors graduating this year.

Candidates may apply for both positions.

QUALIFICATIONS: 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. The strongest candidates will be expected to demonstrate a commitment to creative teaching and an active research program that speaks to and motivates undergraduates from diverse backgrounds.

APPLICATION INSTRUCTIONS: 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 diverse students in a liberal arts college environment; and what background, experience, or interests are likely to make you a strong teacher of Swarthmore College students.

This institution is using Interfolio's Faculty Search to conduct this search. Applicants to this position receive a free Dossier account and can send all application materials, including confidential letters of recommendation, free of charge.

To apply, visit <https://apply.interfolio.com/45234>

Swarthmore College actively seeks and welcomes applications from candidates with exceptional qualifications, particularly those with demonstrable commitments to a more inclusive society and world. Swarthmore College is an Equal Opportunity Employer. Women and minorities are encouraged to apply.

University of Alabama **Computer Science Faculty Position**

The Department of Computer Science at the University of Alabama invites applications for a tenure-track faculty position at the Assistant or Associate level to begin either January or August 2018. Candidates will be expected to engage with faculty and researchers in the Alabama Water Institute (<http://awi.ua.edu/>). Research areas of interest include, but are not limited to, management and manipulation of large sensor data sets, real-time (and near real-time) data processing, resource constrained data collection and analysis, high performance computing, big data, and robotics command-and-control.

Located in Tuscaloosa, Alabama, the University of Alabama enrolls over 37,000 students and is the capstone of higher education in the State. Housed in the College of Engineering, the Computer Science Department has 24 faculty mem-

bers (16 tenured/tenure-track faculty), roughly 700 undergraduates and approximately 50 graduate students. The Department has produced 33 doctoral graduates in the last five years and is funded by agencies such as NSF, Google, Departments of Education and Commerce, various Defense agencies, multiple State agencies, and other sponsors. In 2013, the College completed construction of a \$300M Shelby Engineering and Science Complex.

Applicants should apply online at <http://facultyjobs.ua.edu> requisition # 0810845. Applicants must have an earned doctorate (Ph.D.) in computer science or a closely related field. The application package should include a cover letter, curriculum vitae, and the names of three references. The University of Alabama is an equal opportunity/affirmative action employer. Women and minority applicants are particularly encouraged to apply.

University of Central Florida **Assistant or Associate Professor in Faculty** **Cluster for Cyber Security and Privacy**

The University of Central Florida (UCF) is recruiting a tenure-track assistant or associate professor for its cyber security and privacy cluster. This position has a start date of August 8, 2018.

This will be an interdisciplinary position that will be expected to strengthen both the cluster and a chosen tenure home department, as well as a possible combination of joint appointments. The candidate can choose a combination of units from the cluster for their appointment (see <http://www.ucf.edu/faculty/cluster/cyber-security-and-privacy/>).

The ideal junior candidates will have a strong background in cyber security and privacy, and be on an upward leadership trajectory in these areas. They will have research impact, as reflected in high-quality publications and the ability to build a well-funded research program. All relevant technical areas will be considered. We are looking for a team player who can help bring together current campus efforts in cyber security or privacy. In particular, we are looking for someone who will work at the intersection of several areas, such as: (a) hardware and IoT security, (b) explaining and predicting human behavior, creating policies, studying ethics, and ensuring privacy, (c) cryptography and theory of security or privacy, or (d) tools, methods, training, and evaluation of human behavior.

Minimum qualifications include a Ph.D., terminal degree, or foreign degree equivalent from an accredited institution in an area appropriate to the cluster, and a record of high impact research related to cyber security and privacy, demonstrated by a strong scholarly and/or funding record. A history of working with teams, especially teams that span multiple disciplines, is a strongly preferred qualification. The position will carry a rank commensurate with the candidate's prior experience and record.

Candidates must apply online at <https://www.jobswithucf.com/postings/50404> and attach the following materials: a cover letter, curriculum vitae, teaching statement, research statement, and contact information for three professional references. In the cover letter candidates must address their background in cyber security and privacy,

and identify the department or departments for their potential tenure home and the joint appointments they would desire. When applying, have all documents ready so they can be attached at that time, as the system does not allow resubmission to update applications.

As an equal opportunity/affirmative action employer, UCF encourages all qualified applicants to apply, including women, veterans, individuals with disabilities, and members of traditionally underrepresented populations.

For questions, please contact the Cluster's Search Committee Chair, Gary T. Leavens, at Leavens@ucf.edu.

University of Central Florida **Cluster Lead, Cyber Security and Privacy** **Cluster**

The University of Central Florida (UCF) is recruiting a lead for its cluster on cyber security and privacy. This position has a start date of August 8, 2018. The position will carry a rank of associate or full professor, commensurate with the candidate's prior experience and record. The lead is expected to have credentials and qualifications like those expected of a tenured associate or full professor. To obtain tenure, the selected candidate must have a demonstrated record of teaching, research and service commensurate with rank.

This will be an interdisciplinary position that will be expected to strengthen both the cluster and a chosen tenure home department, as well as a possible combination of joint appointments. The candidate can choose a combination of units from the cluster for their appointment. (See <http://www.ucf.edu/faculty/cluster/cyber-security-and-privacy/>.) Both individual and interdisciplinary infrastructure and startup support will be provided.

The ideal candidate will have a strong background in cyber security and privacy and outstanding research credentials and research impact, as reflected in a sustained record of high quality publications and external funding. All relevant technical areas will be considered including: network security, cryptography, blockchains, hardware security, trusted computing bases, cloud computing, human factors, anomaly detection, forensics, privacy, and software security, as well as applications of security and privacy to areas such as IoT, cyber-physical systems, finance, and insider threats. A history of working with teams, especially teams that span multiple disciplines, is a strongly preferred qualification. A record of demonstrated leadership is highly desired, as we are looking for a leader to bring together all the current campus efforts in cyber security and privacy. This includes three cluster members already hired, as well as a pending hire for the 2017-18 academic year.

Minimum qualifications include a Ph.D. from an accredited institution in an appropriate area, and a record of high impact research related to cyber security and privacy demonstrated by a strong scholarly publication record and a significant amount of sustained funding.

Candidates must apply online at <http://www.jobswithucf.com/postings/50044> and upload the following materials: cover letter, CV, teaching and research statements, and contact information for 3 professional references. In the cover letter, can-

didates should address their background, and identify the department for their potential tenure home and any desired joint appointments.

An equal opportunity/affirmative action employer, UCF encourages all qualified applicants to apply, including women, veterans, individuals with disabilities, and members of traditionally underrepresented populations.

Questions can be directed to the search committee chair, Gary T. Leavens, at Leavens@ucf.edu.

University of Central Florida (UCF) **Assistant or Associate Professor, Computer** **Science**

The Department of Computer Science (CS) at the University of Central Florida (UCF) is seeking applicants for two faculty positions with an anticipated start date of August 8, 2018. The positions will carry the rank of assistant or associate professor. Rank (and tenure for associate professors) will be based on the candidate's prior experience and record.

The department is particularly interested in candidates with experience in the areas of human computer interaction, virtual reality, robotics, data science, algorithms, theory of computing, financial technology, and software engineering and systems. However, all relevant technical areas will be considered. The ideal candidate will have a strong research background and be on an upward leadership trajectory in their research area. They will have research impact, as reflected in high-quality publications and the ability to build a well-funded research program.

The CS Department is home to the first computer science Ph.D. program in the state. Its 38 tenured and tenure-track faculty are engaged in world-class research in Computer Vision, AI and Machine Learning, Virtual Reality, HCI, data analytics, cyber security and privacy, and several other areas. The department has both CS and IT undergraduate degrees accredited by ABET, M.S. degrees in CS, Digital Forensics, and Data Analytics and, a Ph.D. in CS. To learn more about the department see <http://www.cs.ucf.edu/>.

UCF is one of the nation's largest universities. As an economic engine, UCF attracts and supports vital industry to Orlando. UCF is located at the center of the Florida High Tech Corridor where industries include software, defense, space, simulation and training, and entertainment. Next to UCF is a thriving research park that conducts over \$2 billion in funded research. Great weather, easy access to the seashore, one of the largest convention centers in the nation, and one of the world's best airports are just a few features that make Orlando an ideal location. Learn more about UCF at <http://www.ucf.edu/faculty>.

As an equal opportunity/affirmative action employer, UCF encourages all qualified applicants to apply, including women, veterans, individuals with disabilities, and members of traditionally underrepresented populations. UCF's Equal Opportunity Statement can be viewed at: <http://www.eeo.ucf.edu/documents/PresidentsStatement.pdf>. As a Florida public university, UCF makes all application materials and selection procedures available to the public upon request.

Candidates must apply online at www.jobswithucf.com and attach the following materials:

a cover letter, curriculum vitae, teaching statement, research statement, and contact information for three professional references.

NOTE: Please have all documents ready when applying so they can be attached at that time. Once the online submission process is finalized, the system does not allow applicants to submit additional documents at a later date.

For questions regarding this opportunity, please contact the department via email at cs-search@cs.ucf.edu.

Applicants must have a Ph.D. from an accredited institution in an area appropriate to the department, including Computer Science, Computer Engineering, or Mathematics by the time of the appointment.

In order to obtain tenure, the selected candidate must have a demonstrated record of teaching, research and service commensurate with rank in the tenure department.

University of Central Missouri

Assistant Professor of Computer Science - Tenure Track

The School of Computer Science and Mathematics at the University of Central Missouri is accepting applications for one tenure-track position in Computer Science at the rank of Assistant Professor. The appointment will begin August 2018. We are looking for faculty excited by the prospect of shaping our school's future and contributing to its sustained excellence.

The Position: Duties will include teaching undergraduate and graduate courses in computer science and cybersecurity and developing new courses depending upon the expertise of the applicant and school needs, conducting research which leads toward peer-reviewed publications and/or externally funded grants, and program accreditation/assessment. Faculty are expected to assist with school and university committee work and service activities, and advising majors.

Required Qualifications:

- ▶ Ph.D. in Computer Science by August 2018
- ▶ Research expertise and/or industrial experiences in Cybersecurity
- ▶ Demonstrated ability to teach existing courses at the undergraduate and graduate levels
- ▶ Ability to develop a quality research program and secure external funding
- ▶ Supporting the school's plans to establish a National Security Agency designated Center for Academic Excellence in Information Assurance/Cyber Defense
- ▶ Commitment to engage in curricular development/assessment at the undergraduate and graduate levels
- ▶ A strong commitment to excellence in teaching, research, and continued professional growth
- ▶ Excellent verbal and written communication skills

The Application Process: To apply online, go to <https://jobs.ucmo.edu>. Apply to position #997374. The following items should be attached: a letter of interest, a curriculum vitae, a teaching and research statement, copies of transcripts, and a list of at least three professional references including their names, addresses, telephone numbers and email addresses. Official transcripts and three letters of recommendation

will be requested for candidates invited for on-campus interview. For more information, contact:

Dr. Songlin Tian, Search Committee Chair
School of Computer Science and Mathematics
University of Central Missouri
Warrensburg, MO 64093
(660) 543-4930
tian@ucmo.edu

Initial screening of applications begins November 15, 2017, and continues until position is filled.

AA/EEO/ADA. Women and minorities are encouraged to apply.

UCM is located in Warrensburg, MO, which is 35 miles southeast of the Kansas City metropolitan area. It is a public comprehensive university with about 13,000 students. The School of Computer Science and Mathematics offers undergraduate and graduate programs in both Computer Science and Cybersecurity.

University of Chicago

Assistant Professor/Associate Professor/Professor, Computer Science

The Department of Computer Science at the University of Chicago invites applications from qualified candidates for faculty positions at the ranks of Assistant Professor, Associate Professor, and Professor. The University of Chicago has embarked on an ambitious, multi-year effort to significantly expand its computing and data science activities. Candidates with research interests in all areas of computer science will be considered. However, applications are especially encouraged in the areas of AI and Machine Learning, Robotics, Data Analytics, Human-Computer Interaction, and Visual Computing.

Candidates must have demonstrated excellence in research and a strong commitment to teaching. Completion of all requirements for a Ph.D. in Computer Science or a related field is required at the time of appointment. Candidates for Associate Professor and Professor positions must have demonstrated leadership in their field, have established an outstanding independent research program and have a record of excellence in teaching and student mentorship.

Applications must be submitted through the University's Academic Jobs website.

To apply for Assistant Professor, go to <http://tinyurl.com/ya46ybql>

To apply for Associate Professor, go to <http://tinyurl.com/yd9x33eu>

To apply for Professor, go to <http://tinyurl.com/yaqpar49>

To be considered as an applicant, the following materials are required:

- ▶ cover letter
- ▶ curriculum vitae including a list of publications
- ▶ statement describing past and current research accomplishments and outlining future research plans
- ▶ description of teaching philosophy and experience
- ▶ the names of at least three references

Reference letter submission information will be provided during the application process.

Applications received by December 15, 2017 will be given priority consideration.

The University of Chicago has the highest standards for scholarship and faculty quality, is dedicated to fundamental research, and encourages collaboration across disciplines. We encourage connections with researchers across campus in such areas as bioinformatics, mathematics, molecular engineering, natural language processing, statistics, public policy, and social science to mention just a few.

The Department of Computer Science (cs.uchicago.edu) is the hub of a large, diverse computing community of two hundred researchers focused on advancing foundations of computing and driving its most advanced applications. The larger computing and data science community at the University of Chicago includes the Department of Statistics, the Computation Institute, the Toyota Technological Institute at Chicago (TTIC), the Polsky Center for Entrepreneurship and Innovation, the Mansueto Institute for Urban Innovation and the Argonne National Laboratory.

The Chicago metropolitan area provides a diverse and exciting environment. The local economy is vigorous, with international stature in banking, trade, commerce, manufacturing, and transportation, while the cultural scene includes diverse cultures, vibrant theater, world-renowned symphony, opera, jazz, and blues. The University is located in Hyde Park, a Chicago neighborhood on the Lake Michigan shore just a few minutes from downtown.

The University of Chicago is an Affirmative Action/Equal Opportunity/Disabled/Veterans Employer and does not discriminate on the basis of race, color, religion, sex, sexual orientation, gender identity, national or ethnic origin, age, status as an individual with a disability, protected veteran status, genetic information, or other protected classes under the law. For additional information please see the University's Notice of Nondiscrimination at http://www.uchicago.edu/about/non_discrimination_statement/. Job seekers in need of a reasonable accommodation to complete the application process should call 773-702-0287 or email ACOppAdministrator@uchicago.edu with their request.

University of Illinois at Urbana-Champaign

Positions in Computing

The Department of Electrical and Computer Engineering (ECE ILLINOIS) at the University of Illinois at Urbana-Champaign invites applications for faculty positions at all areas and levels in computing, broadly defined, with particular emphasis on cybersecurity and reliability; embedded systems, cyber-physical systems, and the internet of things; networked and distributed computing; data-centric computing systems and storage; quantum computing; robotics and machine vision; machine learning and AI; and bio computation, and health, among other areas. Applications are encouraged from candidates whose research programs specialize in core as well as interdisciplinary areas of electrical and computer engineering. From the transistor and the first computer implementation based on von Neumann's architecture to the Blue Waters petascale computer (the fastest computer on any university

campus), ECE ILLINOIS has always been at the forefront of computing research and innovation. ECE ILLINOIS is in a period of intense demand and growth, serving over 3000 students and averaging 7 new tenure-track faculty hires per year in recent years. It is housed in 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 (graingerinitiative.engineering.illinois.edu), which is backed by a \$100-million gift from the Grainger Foundation.

Please visit <http://jobs.illinois.edu> to view the complete position announcement and application instructions. Full consideration will be given to applications received by November 15, 2017, but applications will continue to be accepted until all positions are filled.

Illinois is an EEO Employer/Vet/Disabled www.inclusiveillinois.illinois.edu.

The University of Illinois conducts criminal background checks on all job candidates upon acceptance of a contingent offer.

The University of Michigan, Ann Arbor **Department of Electrical Engineering and Computer Science** **Computer Science and Engineering Division** **Faculty Positions**

The University of Michigan Computer Science and Engineering (CSE) Division expects strong growth in the coming years and invites applications for multiple tenure-track positions at all levels. These positions include, but are not limited to, cross-disciplinary areas and an endowed professorship (the Fischer Chair) in theoretical computer science. Exceptional candidates from all areas of computer science and computer engineering will be considered. Qualifications include an outstanding academic record, a doctorate or equivalent in computer science or computer engineering, and a strong commitment to teaching and research. The college is especially interested in candidates who can contribute, through their research, teaching, and/or service, to the diversity and excellence of the academic community.

The University of Michigan is one of the world's leading research universities with annual research funding of well over \$1 billion. It consists of highly-ranked departments and colleges across engineering, sciences, medicine, law, business, and the arts. More than a quarter of CSE faculty have sponsored collaborations with faculty in other units. The CSE Division is vibrant and innovative, with over 50 world-class faculty members, over 300 graduate students, several Research Centers, and a large and illustrious network of alumni. Ann Arbor is well known as one of the best college towns in the country. The University of Michigan has a strong dual-career assistance program.

We encourage candidates to apply as soon as possible. For best consideration for Fall 2018, **please apply by December 1, 2017**. Positions remain open until filled and applications can be submitted throughout the year.

For more details on these positions and to apply, please visit the Application Web Page at

<https://www.eecs.umich.edu/eecs/etc/csejobs/>.

The University of Michigan is a Non-Discriminatory/Affirmative Action Employer.

University of Notre Dame **Multiple Tenure-Track Faculty Positions at all Ranks**

The Department of Computer Science and Engineering at the University of Notre Dame invites applications for multiple tenure-track faculty positions at all ranks, with one position specifically in circuits, architecture, or related areas. We seek to attract, develop, and retain excellent faculty members with strong records and future promise. The Department is especially interested in candidates who will contribute to the diversity and excellence of the University's academic community through their research, teaching, and service.

The Department offers the Ph.D. degree and undergraduate Computer Science and Computer Engineering degrees. Faculty are expected to excel in classroom teaching and to lead highly-visible research projects that attract substantial external funding. More information about the department can be found at: <http://cse.nd.edu/>.

Applicants must submit a CV, a teaching statement, a research statement, and contact information for three professional references at <http://apply.interfolio.com/41330>. To guarantee full consideration, applications must be received by December 1, 2017, however, review of applications will continue until the positions have been filled.

The University is an Equal Opportunity and Affirmative Action employer; we strongly encourage applications from women, minorities, veterans, individuals with a disability and those candidates attracted to a university with a Catholic identity.

University of Rochester **Faculty Positions in Computer Science**

The Computer Science Department at the University of Rochester seeks applicants for two tenure-track positions. Outstanding candidates will be considered in any area of computer science and at any level of seniority. We are particularly eager to grow our strength in human-computer interaction and in the theory and practice of security and privacy.

Candidates must have (or be about to receive) a doctorate in computer science or a related discipline. Applications should be submitted online (at <https://www.rochester.edu/faculty-recruiting/login>) no later than January 1, 2018, for full consideration; submissions beyond this date risk being overlooked due to limited interview slots.

The Department of Computer Science (<https://www.cs.rochester.edu>) has a distinguished history of research in artificial intelligence, HCI, systems, and theory. We nurture a highly collaborative and interdisciplinary culture, with exceptionally strong external funding and with active ties to numerous allied departments, including brain and cognitive science, linguistics, biomedical engineering, electrical and computer engineering, and several departments in the medical center. Recent faculty hires have received a host of national honors,

including the NSF CAREER award, the MIT TR35 award, honorable mention in the ACM dissertation competition, multiple Google research awards, and best paper designations at top-tier conferences. In 2015 we were one of only two CS departments nationwide to secure three NSF CRII awards for junior faculty.

The department is deeply committed to building a more diverse and representative faculty, and strongly encourages applications from groups underrepresented in higher education. We have a vibrant Women in Computing community, and are a charter member of the ABI/HMC BRAID Initiative. With funding from the NSF, the CRA, and major industrial sponsors, BRAID works to increase diversity and inclusivity in the undergraduate program and to rigorously evaluate factors that contribute to change. In 2017, women constituted 33% of our BA/BS graduates, and we are actively working to improve the environment for other underrepresented groups.

The University of Rochester is a private, Tier I research institution with approximately 5,000 undergraduates and a comparable number of graduate students. It has recently committed \$50M to the multidisciplinary Goergen Institute for Data Science (GIDS), of which Computer Science is the leading departmental member — and with which it shares a newly constructed state-of-the-art facility. Ongoing hiring in GIDS provides exciting opportunities for collaboration between computing and other disciplines.

Anchoring the Finger Lakes region of western New York State, the greater Rochester area is home to over a million people, and offers unsurpassed quality of life, with a thriving arts scene, outstanding public schools, affordable housing, and a huge range of cultural and recreational opportunities. Traditionally strong in optics research and manufacturing, the area was recently selected by the Department of Defense as the hub of a \$600M Integrated Photonics Institute for Manufacturing Innovation.

The University of Rochester is an Equal Opportunity Employer:

EOE Minorities/Females/Protected Veterans/Disabled

The University of Rochester, an Equal Opportunity Employer, has a strong commitment to diversity and actively encourages applications from candidates from groups underrepresented in higher education.

EOE Minorities/Females/Protected Veterans/Disabled

University of Toronto **Assistant Professor - Electrical and Computer Engineering**

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering (ECE) at the University of Toronto invites applications for up to four tenure-stream faculty appointments at the rank of Assistant Professor. The appointments will commence on July 1, 2018.

Within the general field of electrical and computer engineering, we seek applications from candidates with expertise in one or more of the following strategic research areas: 1. Computer or Communications Engineering, with preference for a focus on machine learning, computer security and privacy, or data science; 2. Electrical

Power Systems, with preference for a focus on power systems protection; 3. Systems Control, with preference for a focus on robotics.

Applicants are expected to have a Ph.D. in Electrical and Computer Engineering, or a related field, at the time of appointment or soon after.

Successful candidates will be expected to initiate and lead an independent research program of international calibre, and to teach at both the undergraduate and graduate levels. Candidates should have demonstrated excellence in research and teaching. Excellence in research is evidenced primarily by publications in leading journals or conferences in the field, presentations at significant conferences and strong endorsements by referees of high international standing. Evidence of excellence in teaching will be demonstrated by strong communication skills, a compelling statement of teaching submitted as part of the application highlighting areas of interest and accomplishments, as well as strong letters of recommendation.

Eligibility and willingness to register as a Professional Engineer in Ontario is highly desirable.

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto ranks among the best in North America. It attracts outstanding students, has excellent facilities, and is ideally located in the middle of a vibrant, artistic, diverse and cosmopolitan city. Additional information may be found at <http://www.ece.utoronto.ca>.

Review of applications will begin after October 3, 2017, however, the position will remain open until December 11, 2017. You must submit your application online, by following the submission guidelines given at <http://uoft.me/how-to-apply>. Applications submitted in any other way will not be considered.

As part of your online application, please include a curriculum vitae, a summary of your previous research and future research plans, as well as a statement of teaching experience and interests. Applicants must arrange for three letters of reference to be sent directly by the referees (on letterhead, signed and scanned), by email to the ECE department at search2017@ece.utoronto.ca.

The University of Toronto is strongly committed to diversity within its community and especially welcomes applications from racialized persons / persons of colour, women, Indigenous / Aboriginal People of North America, persons with disabilities, LGBTQ persons, and others who may contribute to the further diversification of ideas.

As part of your application, you will be asked to complete a brief Diversity Survey. This survey is voluntary. Any information directly related to you is confidential and cannot be accessed by search committees or human resources staff. Results will be aggregated for institutional planning purposes. For more information, please see <http://uoft.me/UP>.

All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority.

University of Toronto

Associate Professor - Electrical and Computer Engineering

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering (ECE) at the Uni-

versity of Toronto invites applications for up to four tenure-stream faculty appointments at the rank of Associate Professor. The appointments will commence on July 1, 2018.

Within the general field of electrical and computer engineering, we seek applications from candidates with expertise in one or more of the following strategic research areas: 1. Computer or Communications Engineering, with preference for a focus on machine learning, computer security and privacy, or data science; 2. Electrical Power Systems, with preference for a focus on power systems protection; 3. Systems Control, with preference for a focus on robotics.

Applicants are expected to have a Ph.D. in Electrical and Computer Engineering, or a related field, and have at least five years of academic or relevant industrial experience.

Successful candidates will be expected to initiate and lead an independent, competitive and innovative research program of international calibre, and to teach at both the undergraduate and graduate levels. Candidates should have demonstrated excellence in research and teaching. Excellence in research is evidenced primarily by publications in leading journals or conferences in the field, presentations at significant conferences and a high profile in the field with strong endorsements by referees of high international standing. Evidence of excellence in teaching will be demonstrated by strong communication skills, a compelling statement of teaching submitted as part of the application highlighting areas of interest and accomplishments, as well as strong letters of recommendation.

Eligibility and willingness to register as a Professional Engineer in Ontario is highly desirable.

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto ranks among the best in North America. It attracts outstanding students, has excellent facilities, and is ideally located in the middle of a vibrant, artistic, diverse and cosmopolitan city. Additional information may be found at <http://www.ece.utoronto.ca>.

Review of applications will begin after October 3, 2017, however, the position will remain open until December 11, 2017. You must submit your application online, by following the submission guidelines given at <http://uoft.me/how-to-apply>. Applications submitted in any other way will not be considered.

As part of your online application, please include a curriculum vitae, a summary of your previous research and future research plans, as well as a statement of teaching experience and interests. Applicants must arrange for three letters of reference to be sent directly by the referees (on letterhead, signed and scanned), by email to the ECE department at search2017@ece.utoronto.ca.

The University of Toronto is strongly committed to diversity within its community and especially welcomes applications from racialized persons / persons of colour, women, Indigenous / Aboriginal People of North America, persons with disabilities, LGBTQ persons, and others who may contribute to the further diversification of ideas.

As part of your application, you will be asked to complete a brief Diversity Survey. This survey is voluntary. Any information directly related to you is confidential and cannot be accessed by search

committees or human resources staff. Results will be aggregated for institutional planning purposes. For more information, please see <http://uoft.me/UP>.

All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority.

US Air Force Academy

Assistant Professor of Computer Science

The Department of Computer Science at the US Air Force Academy seeks to fill a full-time faculty position at the Assistant Professor level. The department is particularly interested in candidates with backgrounds in artificial intelligence, computer and network security, operations research, or unmanned aerial systems, but all candidates with a passion for undergraduate computer science teaching are encouraged to apply.

The Academy is a national service institution, charged with producing lieutenants for the US Air Force. Faculty members are expected to exemplify the highest ideals of professionalism and character. USAFA is located in Colorado Springs, an area known for its exceptional natural beauty and quality of life. The United States Air Force Academy values the benefits of diversity among the faculty to include a variety of educational backgrounds, professional and life experiences.

For information on how to apply, go to usajobs.gov and search with the keyword 478670300.

York University

Assistant Professor

The Department of Electrical Engineering and Computer Science, York University, is seeking an outstanding candidate at the rank of Assistant Professor in the area of Computer Systems although exceptional applicants from other areas in Computer Science may also be considered. The successful candidate will have a PhD in Computer Science, or a closely related field, and a research record commensurate with rank. The appointment will commence on July 1, 2018, subject to budgetary approval. For full position details, see <http://www.yorku.ca/acadjobs>.

Applicants should complete the on-line process at <http://lassonde.yorku.ca/new-faculty/>. A complete application includes a cover letter indicating the rank for which the candidate wishes to be considered, a detailed CV, statement of contribution to research, teaching and curriculum development, three sample research publications and contact information for three referees. Complete applications must be received by **November 30, 2017**.

York University is an Affirmative Action employer and strongly values diversity, including gender and sexual diversity, within its community. The Affirmative Action Program, which applies to women, Aboriginal people, visible minorities and people with disabilities, can be found at <http://acadjobs.info.yorku.ca/affirmative-action/> or by calling the AA office at 416.736.5713. All qualified candidates are encouraged to apply; however, Canadian Citizens and Permanent Residents will be given priority.

[CONTINUED FROM P. 101] its effects was impossible.

But now he considered a slightly greater perturbation—not a butterfly's flutter but a disturbance as large as a tennis court. A change in the parameters of one cell in that mother of all spreadsheets. Could such a perturbation have a significant effect anywhere else?

Eliot decided on an experiment. First, he ran a model based on the latest grid data—nothing new there—predicting the weather 20 minutes into the future. Then he went back into the input matrices and manipulated the parameters in one cell, the one that covered his backyard. He then lowered the temperature by five degrees and ran the model again.

A simple program subtracted the two results, color-coding any differences. Blue indicated cells with no change, green very slight differences, and yellow and red locations that were significantly altered. Because weather, carried by the wind, doesn't move faster than the speed of sound, he looked at the results only within 250 miles. Beyond that, Eliot would need a prediction further into the future, and couldn't be sure if any effects were due to his manipulated data or not.

The difference plot was mostly just a noisy sea of blue and green pixels. A few were red, indicating, for example, the barometric pressure was 1% greater or lesser than it would have been without the backyard temperature change. His theoretical butterfly hadn't had much effect. But there was one spot that was yellow . . . where the local wind speed had increased by a factor of two. A factor of two! Eliot pushed back his chair and let out a low whistle.

Lily watched her husband inch the rented pickup down the driveway, stopping just short of the desiccated wasteland that was their yard. He had lined its bed with plastic and spent several hours filling it with water from the kitchen sink.

"Are you going to tell me what this is about, Eliot? I mean, is this your idea of making rain?"

"I'm going for a real live butterfly effect."

"Like when you feel nervous? When your tummy turns upside down?"

"Not at all. It's when a small stimulus produces a big effect. Consequences that greatly outweigh causes."

But there was one spot that was yellow . . . where the local wind speed had increased by a factor of two. A factor of two!

"You mean like the shooting of Archduke Franz Ferdinand?"

"Who?"

"1914. Franz Ferdinand was assassinated by a fervent Yugoslavian nationalist. Small event. But within months, they were digging trenches in France. The First World War.

"No trenches here, Lily. And no archdukes either." He pulled the gate on the pickup, and three tons of water spilled out onto the parched earth.

In 20 minutes, Eliot was looking at his difference plots. Despite the fact that he had lowered the surface temperature of his yard considerably more than the five degrees of his numerical test, there was still no apparent effect within 250 miles. This butterfly had flapped its wings for naught.

What Eliot couldn't see, and wouldn't understand until much later, was that six hours after he flooded the yard, a Russian military transport carrying diplomats would run into trouble on its approach to Warsaw's Bemowo Airfield. The flight forecast had predicted smooth skies all the way. And they were, except at 3,000 feet short of the runway where an unusually forceful dust devil lifted one wing and tossed the plane on its back.

Was it an accident? The Russians certainly didn't think so. This was subtle sabotage, and no degree of pleading could convince them otherwise. In the predawn skies the following day, an unmanned drone headed west, hugging the flat landscapes of Northern Europe. Someone had finally done something about the weather. 

Seth Shostak (seth@seti.org) is the senior astronomer at the SETI Institute in Mountain View, CA.

© 2017 ACM 0001-0782/17/11 \$15.00

INTERACTIONS



ACM's *Interactions* magazine explores critical relationships between people and technology, showcasing emerging innovations and industry leaders from around the world across important applications of design thinking and the broadening field of interaction design.

Our readers represent a growing community of practice that is of increasing and vital global importance.



To learn more about us, visit our award-winning website <http://interactions.acm.org>

Follow us on Facebook and Twitter  

To subscribe: <http://www.acm.org/subscribe>

Association for
Computing Machinery



From the intersection of computational science and technological speculation, with boundaries limited only by our ability to imagine what could be.

DOI:10.1145/3140960

Seth Shostak

Future Tense Butterfly Effect

But, like the weather, what can anyone do about it?

LILY'S EYES SCANNED the yard, an expansive tract of suburban real estate she called the back 40. She was not pleased.

"Eliot, this is embarrassing. Our property makes the Dust Bowl look lush. Is it ever going to rain again?"

Her husband turned toward the sky as if seeking an answer. But of course he already knew what his response would be, and so did she.

"October, Lily... It rains in October."

"I don't care about seasonal behavior or what's normal for the state. If it doesn't rain in Longmont or Loveland, well, tough for them. I just care about this backyard patch. Make it rain *here*, will you? You're the meteorologist."

"Yes, dear, I am," Eliot replied, flashing a slight smile in the interests of domestic tranquility, and went inside.

Despite the fact that Eliot had a sheepskin testifying to his meteorological chops Lily's gibe was a reminder he wasn't the man he had once hoped to be. He wasn't really a weather forecaster. In high school, he could look at the clouds, sense the temperature, and know a front was moving in. Some of his friends admired his ability to predict afternoon thunderstorms, while others thought he was just obsessive. Eliot didn't care much. He enjoyed the fact that his skills could be tested every day, his predictions verified or disproved within hours. It was like being a day trader, but without the risk.

Eliot followed his meteorological interests through grad school, but by the late 2030s technology was rendering his skills obsolete. Weather forecasts were increasingly the province of computer models, massive calculations that spit out accurate predictions for any place on the planet, down to an acre or less.



This development was inevitable, requiring only the ability to build a finer grid of weather data—wind, temperature, barometric pressure—and the compute power to crunch it all. Both were now at hand.

Improved satellites had refined the grid by a factor of 20 in all directions. The whole planet—continents, ocean surface, the entire atmosphere below six miles—was now sampled on a scale of 300 feet. Every few minutes, the weather was measured and binned into five trillion cells, the mother of all spreadsheets.

It was Eliot's job to feed this vast anthill of numbers into the models several times a day and bring to bear the compute power available in the Extended Cloud. Yes, he had to understand what he was doing, and, yes, he had to be careful. But it wasn't traditional weather forecasting, and definitely not weather manipulation.

So Lily's remarks bothered him. Was he truly helpless? All his life people had teased him with Charles Dudley Warner's bromide that everyone complains about the weather, but no one does anything about it. But he was a scientist. He knew that weather involved immense highly energetic systems. The output of 100 power plants was nothing compared to a hurricane's terawatts. How could *anyone* do anything about the weather? It was akin to moving the Rocky Mountains.

He also knew the weather was a system that was chaotic and close to equilibrium. A small change could have big consequences. The butterfly effect.

The idea that a butterfly could precipitate a storm was a popular idea, recognized centuries ago. Eliot had thought about it in school but reckoned the flapping wings of a single insect really couldn't do much. And predicting

[CONTINUED ON P. 111]



10th
Edition

CONFERENCE 27 – 30 November 2017
EXHIBITION 28 – 30 November 2017
BITEC, Bangkok, Thailand

THE CELEBRATION OF LIFE & TECHNOLOGY

The 10th ACM SIGGRAPH Conference and Exhibition
on Computer Graphics and Interactive Techniques in Asia



Register online by 15 October 2017,
& enjoy early bird discounts of up to

20%

👉 SA2017.SIGGRAPH.ORG/REGISTRATION

Sponsored by



Organized by



Middleware 2017

Dec 11 – 15, Las Vegas

We invite you to attend the 18th ACM/IFIP/USENIX International Middleware Conference 2017 covering recent scientific advances in middleware systems. The conference will showcase an exciting agenda including a single-track technical program, invited speakers, workshops, tutorials, demos & posters, doctoral symposium, and social events with researchers from academia and industry.



Keynotes



*Prof. Magdalena Balazinska, University of Washington
Jean Loup Baer Professor of Computer Science & Engineering
Performance SLAs for Cloud Data Analytics*



*Dr. Ricardo Bianchini, MSR Redmond
Chief Efficiency Strategist & Manager of the Systems Research Group
Toward intelligent cloud platforms: the Resource Central experience*

Tutorials

- High Performance Network Middleware with Intel DPDK and OpenNetVM
- Istio Service mesh for more robust, secure and easy to manage microservices
- SMACK stack 101: Building Fast Data stacks
- Trusted Execution of Software using Intel SGX

Workshops

- Active: International Workshop on Active Middleware on Modern Hardware
- ARM: Adaptive and Reflective Middleware
- DIDL: Workshop on Distributed Infrastructures for Deep Learning
- M4IoT: Middleware and Applications for the Internet of Things
- MECC: Middleware for Edge Clouds & Cloudlets
- SERIAL: Scalable and Resilient Infrastructures for distributed Ledgers
- WoSC: Workshop on Serverless Computing

2017.middleware-conference.org



@middleware2017

