

# COMMUNICATIONS

CACM.ACM.ORG

OF THE

# ACM

01/2013 VOL.56 NO.1



## **Human Mobility Characterization from Cellular Network Data**

What College Could Be Like

Who Begat Computing?

Computer Security  
and the Modern Home

ACM's FY12 Annual Report

What's a Robot?

Association for  
Computing Machinery



allrecipes 



# Cook up the next great app.

Opportunity doesn't just knock. In the Windows Store it swipes, taps and clicks, too. See how Allrecipes and others are building immersive apps for the new Windows experience and learn how you can put your app in the hands of new users everywhere.

**Build for the new Windows Store.**  
Open for business at [windowsstore.com](http://windowsstore.com)

 **Windows 8**

# Inviting Young Scientists

## Meet Some of the Greatest Minds of Mathematics and Computer Science

Young researchers in the fields of mathematics and/or computer science are invited to participate in an extraordinary opportunity to meet some of the preeminent scientists in the field. ACM has joined forces with the newly created Heidelberg Laureate Forum (HLF) to bring students together with the very pioneering researchers who may have sparked their passion for science and math. These role models include recipients of the Abel Prize, the ACM A.M. Turing Award, and the Fields Medal.

The first Heidelberg Laureate Forum will take place September 22–27, 2013 in Heidelberg, Germany.

The week-long event will focus on scientific inspiration and exchange through a series of presentations, workshops, panel discussions, and social events involving both the laureates and the young scientists.

### Who can participate?

The HLF invites new and recent Ph.D.'s, Ph.D. candidates, other graduate students involved in research and undergraduate students with solid experience in and a commitment to computing research to apply.

### How to apply:

Young researchers can apply online:

[https://application.heidelberg-laureate-forum.org/intern/reg\\_registration\\_for.php](https://application.heidelberg-laureate-forum.org/intern/reg_registration_for.php)

The materials required for a complete application are listed on the site.

### What is the schedule?

The deadline for applications is **February 15, 2013**.

We reserve the right to close the application website early should we receive more applications and nominations than our reviewers can handle.

Successful applicants will be notified by **April 15, 2013** and will receive full support to attend the Forum.

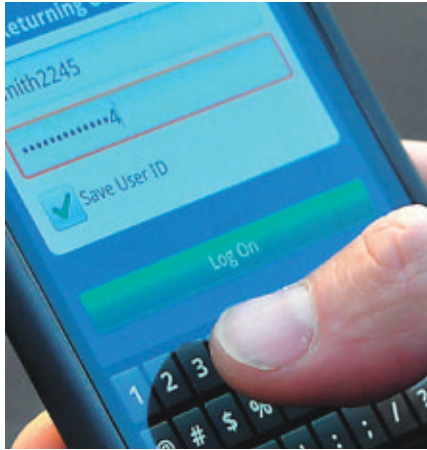
## Departments

- 5 **Editor's Letter**  
**Who Begat Computing?**  
*By Moshe Y. Vardi*
- 
- 7 **From the President**  
**What's a Robot?**  
*By Vinton G. Cerf*
- 
- 8 **Letters to the Editor**  
**Computer Science Is Not a Science**
- 
- 11 **ACM's FY12 Annual Report**
- 
- 16 **BLOG@CACM**  
**Lost in Translation**  
Daniel Reed on straddling the intellectual divide between technology experts and policymakers.
- 
- 49 **Calendar**
- 
- 125 **Careers**

## Last Byte

- 136 **Future Tense**  
**Share My Enlightenment**  
I self-publish, and you get to sail my aether wave for free.  
*By Rudy Rucker*

## News



- 19 **Stopping the Leaks**  
Side channels give out information that can be used to crack secrets, but researchers are identifying the holes and trying to close them.  
*By Neil Savage*
- 
- 22 **Beyond Hadoop**  
The leading open source system for processing big data continues to evolve, but new approaches with added features are on the rise.  
*By Gregory Mone*
- 
- 25 **Just the Facts**  
In repackaging other companies' news, some news aggregators are diverting readers and ad dollars, and, critics argue, undercutting the incentive to spend money on original reporting. It is an economic and ethical problem without a clear legal fix.  
*By Marina Krakovsky*

## Viewpoints

- 28 **Technology Strategy and Management**  
**The Apple-Samsung Lawsuits**  
In search of a middle ground in the intellectual property wars.  
*By Michael A. Cusumano*
- 
- 32 **The Business of Software**  
**How We Build Things**  
...and why things are 90% complete.  
*By Phillip G. Armour*
- 
- 34 **Law and Technology**  
**Beyond Location: Data Security in the 21<sup>st</sup> Century**  
Viewing evolving data security issues as engineering problems to be solved.  
*By Deven Desai*
- 
- 37 **Historical Reflections**  
**Five Lessons from Really Good History**  
Lessons learned from four award-winning books on the history of information technology.  
*By Thomas Haigh*
- 
- 41 **Viewpoint**  
**What College Could Be Like**  
Imagining an optimized education model.  
*By Salman Khan*
- 
- 44 **Viewpoint**  
**Conference-Journal Hybrids**  
Considering how to combine the best elements of conferences and journals.  
*By Jonathan Grudin, Gloria Mark, and John Riedl*

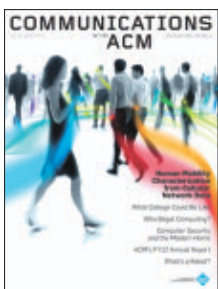


## Practice



- 50 **Condos and Clouds**  
Constraints in an environment empower the services.  
*By Pat Helland*
- 
- 60 **Browser Security: Appearances Can Be Deceiving**  
A discussion with Jeremiah Grossman, Ben Livshits, Rebecca Bace, and George Neville-Neil.
- 
- 68 **The Web Won't Be Safe or Secure Until We Break It**  
Unless you have taken very particular precautions, assume every website you visit knows exactly who you are.  
*By Jeremiah Grossman*

**Q** Articles' development led by **acmqueue** [queue.acm.org](http://queue.acm.org)



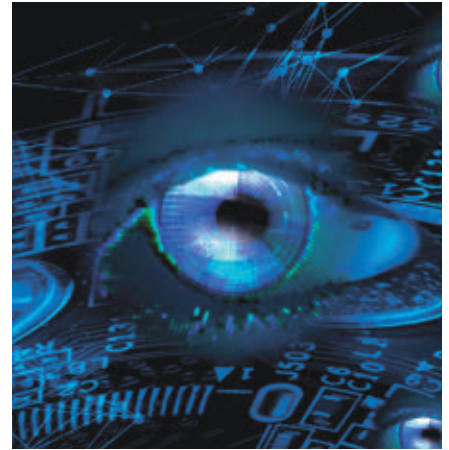
**About the Cover:** Human mobility patterns have long been studied for urban planning, traffic routing, and the spread of viruses, among many other reasons. Today, most humans travel with cellphones making valuable footprint data available via cellular networks. This month's cover story (p. 74) captures the power in the patterns.

## Contributed Articles



- 74 **Human Mobility Characterization from Cellular Network Data**  
Anonymous location data from cellular phone networks sheds light on how people move around on a large scale.  
*By Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky*
- 
- 83 **Abstractions for Genomics**  
Large genomic databases with interactive access require new, layered abstractions, including separating “evidence” from “inference.”  
*By Vineet Bafna, Alin Deutsch, Andrew Heiberg, Christos Kozanitis, Lucila Ohno-Machado, and George Varghese*

## Review Articles



- 94 **Computer Security and the Modern Home**  
A framework for evaluating security risks associated with technologies used at home.  
*By Tamara Denning, Tadayoshi Kohno, and Henry M. Levy*

## Research Highlights

- 105 **Technical Perspective**  
**Visualization, Understanding, and Design**  
*By Doug DeCarlo and Matthew Stone*
- 
- 106 **Illustrating How Mechanical Assemblies Work**  
*By Niloy J. Mitra, Yong-Liang Yang, Dong-Ming Yan, Wilmot Li, and Maneesh Agrawala*
- 
- 115 **Technical Perspective**  
**Finding People In Depth**  
*By James M. Rehg*
- 
- 116 **Real-Time Human Pose Recognition in Parts from Single Depth Images**  
*By Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore*



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

**Executive Director and CEO**

John White  
**Deputy Executive Director and COO**  
 Patricia Ryan  
**Director, Office of Information Systems**  
 Wayne Graves  
**Director, Office of Financial Services**  
 Russell Harris  
**Director, Office of SIG Services**  
 Donna Cappel  
**Director, Office of Publications**  
 Bernard Rous  
**Director, Office of Group Publishing**  
 Scott E. Delman

**ACM COUNCIL**

**President**  
 Vinton G. Cerf  
**Vice-President**  
 Alexander L. Wolf  
**Secretary/Treasurer**  
 Vicki L. Hanson  
**Past President**  
 Alain Chesnais  
**Chair, SGB Board**  
 Erik Altman  
**Co-Chairs, Publications Board**  
 Ronald Boisvert and Jack Davidson  
**Members-at-Large**  
 Eric Allman; Ricardo Baeza-Yates;  
 Radia Perlman; Mary Lou Soffa;  
 Eugene Spafford  
**SGB Council Representatives**  
 Brent Hailpern; Joseph Konstan;  
 Andrew Sears

**BOARD CHAIRS**

**Education Board**  
 Andrew McGettrick  
**Practitioners Board**  
 Stephen Bourne

**REGIONAL COUNCIL CHAIRS**

**ACM Europe Council**  
 Fabrizio Gagliardi  
**ACM India Council**  
 Anand S. Deshpande, PJ Narayanan  
**ACM China Council**  
 Jianguang Sun

**PUBLICATIONS BOARD**

**Co-Chairs**  
 Ronald F. Boisvert; Jack Davidson  
**Board Members**  
 Marie-Paule Cani; Nikil Dutt; Carol Hutchins;  
 Joseph A. Konstan; Ee-Peng Lim;  
 Catherine McGeoch; M. Tamer Ozsu;  
 Vincent Shen; Mary Lou Soffa

**ACM U.S. Public Policy Office**

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

**Computer Science Teachers Association**

Chris Stephenson,  
 Executive Director

# COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

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

**STAFF**

**DIRECTOR OF GROUP PUBLISHING**

Scott E. Delman  
 publisher@cacm.acm.org

**Executive Editor**

Diane Crawford

**Managing Editor**

Thomas E. Lambert

**Senior Editor**

Andrew Rosenbloom

**Senior Editor/News**

Jack Rosenberger

**Web Editor**

David Roman

**Editorial Assistant**

Zarina Strakhan

**Rights and Permissions**

Deborah Cotton

**Art Director**

Andrij Borys

**Associate Art Director**

Margaret Gray

**Assistant Art Directors**

Mia Angelica Balaquiot

Brian Greenberg

**Production Manager**

Lynn D'Addesio

**Director of Media Sales**

Jennifer Ruzicka

**Public Relations Coordinator**

Virginia Gold

**Publications Assistant**

Emily Williams

**Columnists**

Alok Aggarwal; Phillip G. Armour;  
 Martin Campbell-Kelly;  
 Michael Cusumano; Peter J. Denning;  
 Shane Greenstein; Mark Guzdial;  
 Peter Harsha; Leah Hoffmann;  
 Mari Sako; Pamela Samuelson;  
 Gene Spafford; Cameron Wilson

**CONTACT POINTS**

**Copyright permission**  
 permissions@cacm.acm.org

**Calendar items**  
 calendar@cacm.acm.org

**Change of address**  
 acmhlp@acm.org

**Letters to the Editor**  
 letters@cacm.acm.org

**WEB SITE**

http://cacm.acm.org

**AUTHOR GUIDELINES**

http://cacm.acm.org/guidelines

**ACM ADVERTISING DEPARTMENT**

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

**Director of Media Sales**

Jennifer Ruzicka  
 jen.ruzicka@hq.acm.org

**Media Kit** acmm mediasales@acm.org

**Association for Computing Machinery (ACM)**

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

**EDITORIAL BOARD**

**EDITOR-IN-CHIEF**

Moshe Y. Vardi  
 eic@cacm.acm.org

**NEWS**

**Co-Chairs**

Marc Najork and Prabhakar Raghavan

**Board Members**

Hsiao-Wuen Hon; Mei Kobayashi;  
 William Pulleyblank; Rajeev Rastogi

**VIEWPOINTS**

**Co-Chairs**

Susanne E. Hambrusch; John Leslie King;  
 J Strother Moore

**Board Members**

P. Anandan; William Aspray; Stefan Bechtold;  
 Judith Bishop; Stuart I. Feldman;  
 Peter Freeman; Seymour Goodman;  
 Mark Guzdial; Richard Heeks;  
 Rachele Hollander; Richard Ladner;  
 Susan Landau; Carlos Jose Pereira de Lucena;  
 Beng Chin Ooi; Loren Terveen;  
 Jeannette Wing

**PRACTICE**

**Chair**

Stephen Bourne

**Board Members**

Eric Allman; Charles Beeler; Bryan Cantrill;  
 Terry Coatta; Stuart Feldman; Benjamin Fried;  
 Pat Hanrahan; Tom Limoncelli;  
 Marshall Kirk McKusick; Erik Meijer;  
 George Neville-Neil; Theo Schlossnagle;  
 Jim Waldo

The Practice section of the CACM

Editorial Board also serves as the Editorial Board of *COMMUNIQUE*.

**CONTRIBUTED ARTICLES**

**Co-Chairs**

Al Aho and Georg Gottlob

**Board Members**

William Aiello; Robert Austin; Elisa Bertino;  
 Gilles Brassard; Kim Bruce; Alan Bundy;  
 Peter Buneman; Erran Carmel;  
 Andrew Chien; Peter Druschel; Carlo Ghezzi;  
 Carl Gutwin; James Larus; Igor Markov;  
 Gail C. Murphy; Shree Nayar; Bernhard  
 Nebel; Lionel M. Ni; Sriram Rajamani;  
 Marie-Christine Rousset; Avi Rubin;  
 Krishan Sabnani; Fred B. Schneider;  
 Abigail Sellen; Ron Shamir; Yoav Shoham;  
 Marc Snir; Larry Snyder; Manuela Veloso;  
 Michael Vitale; Wolfgang Wahlster;  
 Hannes Werthner; Andy Chi-Chih Yao

**RESEARCH HIGHLIGHTS**

**Co-Chairs**

Stuart J. Russell and Gregory Morrisett

**Board Members**

Martin Abadi; Sanjeev Arora; Dan Boneh;  
 Andrei Broder; Stuart K. Card; Jon Crowcroft;  
 Alon Halevy; Monika Henzinger;  
 Maurice Herlihy; Norm Jouppi;  
 Andrew B. Kahng; Xavier Leroy;  
 Mendel Rosenblum; Ronit Rubinfeld;  
 David Salesin; Guy Steele, Jr.; David Wagner;  
 Alexander L. Wolf; Margaret H. Wright

**WEB**

**Chair**

James Landay

**Board Members**

Gene Golovchinsky; Marti Hearst;  
 Jason I. Hong; Jeff Johnson; Wendy E. MacKay



**ACM Copyright Notice**

Copyright © 2013 by Association for Computing Machinery, Inc. (ACM). Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or fee. Request permission to publish from permissions@acm.org or fax (212) 869-0481.

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

**Subscriptions**

An annual subscription cost is included in ACM member dues of \$99 (\$40 of which is allocated to a subscription to *Communications*); for students, cost is included in \$42 dues (\$20 of which is allocated to a *Communications* subscription). A nonmember annual subscription is \$100.

**ACM Media Advertising Policy**

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

**Single Copies**

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

**COMMUNICATIONS OF THE ACM**

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

**POSTMASTER**

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



Association for Computing Machinery



Printed in the U.S.A.



Moshe Y. Vardi

DOI: 10.1145/2398356.2398357

# Who Begat Computing?

The Turing Centenary with its furious pace is now behind us and we can afford some reflection on what has transpired. What started as an idea that the centenary of one

of the founding figures of computing should be celebrated has turned into a global social phenomenon. A quick perusal of the Turing Centenary Web page (<http://www.turingcentenary.eu/>) reveals an amazing explosion of meetings, lectures, exhibitions, and volumes.

There is a risk, however, that in our focus on highlighting Turing's seminal contributions we may have gone from celebration to hagiography. Listening to so many speakers extol Turing's accomplishments, one could end up believing that Turing single-handedly begat computing, being the father of computability, universal machines, stored-program computers, cryptanalysis, and artificial intelligence. This picture is simplistic and does not do justice to the richness of the story of how computing emerged between 1930 and 1950. We do not have one founding figure, we have several, and we should recognize and celebrate all of them.

The study of computability was launched at Princeton University, where Alonzo Church, together with his students Stephen Kleene and Barkley Rosser formalized computability in the early 1930s first in terms of the lambda-calculus, and then in terms of recursive functions (proposed by Jacques Herbrand and Kurt Gödel). They also proved the equivalence of the two formalisms, which led to Church's identification of computability with recursiveness. Yet, this characterization of computability was not compelling enough and described as "thoroughly unsatisfactory" by Gödel. It was then Turing's influential analysis of com-

putability in terms of finite machines and its equivalence to the lambda-calculus and recursiveness that led to our current accepted understanding of computability, referred to as the Church-Turing Thesis. (Emil Post independently formulated another notion of machines, which turned out to be equivalent to Turing machines.)

Turing was a leading scientist in deciphering the German Enigma code at Bletchley Park in the early 1940s. Yet, unlike his computability work, which was done independently of the Princeton effort, breaking the Enigma was a collective effort. To start with, Turing was building on previous work by Polish and British code-breakers. I.J. Good played a key role in the Bayesian statistical analysis of Enigma messages and Gordon Welchman made key contributions to the design of the Bombe, the machine that used brute-force search to identify correct Enigma rotor positions. Overall, one must remember that the British code-breaking project was a huge effort; 12,000 people toiled at Bletchley Park during the war.

The claims that Turing invented the stored-program computer, which typically refers to the uniform handling of programs and data, are simply ahistorical. One can trace the kernel of the idea of handling programs and data uniformly back to Gödel's arithmetization of provability in 1931. The idea then showed up again in the lambda-calculus, recursive functions, and Turing machines. Turing invented a universal machine, a machine that can simulate all

other machines, but he was preceded by the Princeton group, who constructed a universal lambda-term and a universal recursive function. While these ideas undoubtedly influenced the efforts of John von Neumann and his collaborators at the University of Pennsylvania in the 1940s, we should not confuse a mathematical idea with an engineering design. It was the EDVAC Report of 1945 that offered the first explicit exposition of the stored-program computer. Turing's ACE Report, which elaborated on this idea and cited the EDVAC Report, was submitted in early 1946. The first embodiments of the stored-program computer were the Manchester Baby and the Cambridge EDSAC, put into operation in 1949 and preceding the Pilot ACE, which was based on Turing's design and first run in 1950.

Turing was not the first to think about artificial intelligence (AI). The philosopher Charles S. Peirce wrote in 1887: "Precisely how much the business of thinking a machine could possibly be made to perform, and what part of it must be left to the living mind is a question not without conceivable practical importance." Nevertheless, Turing's 1950 paper "Computing Machinery and Intelligence" is indeed the first deep philosophical investigation of the possibility of artificial intelligence. While the Turing Test, referred in the paper as the "Imitation Game," has been rather under-influential in the history of AI, Turing does deserve the credit for putting the question of general machine intelligence so squarely on the table.

Computing emerged during the 1930–1950 period because the time was right. Many people played key roles in this development; assigning precise credit is quite impossible. Turing was a great computing pioneer, and his place in the computing pantheon is secure, but he is not alone there.

*Moshe Y. Vardi*, EDITOR-IN-CHIEF



Association for  
Computing Machinery

Advancing Computing as a Science & Profession

# membership application & digital library order form

Priority Code: AD13

## You can join ACM in several easy ways:

### Online

<http://www.acm.org/join>

### Phone

+1-800-342-6626 (US & Canada)

+1-212-626-0500 (Global)

### Fax

+1-212-944-1318

Or, complete this application and return with payment via postal mail

### Special rates for residents of developing countries:

<http://www.acm.org/membership/L2-3/>

### Special rates for members of sister societies:

<http://www.acm.org/membership/dues.html>

Please print clearly

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State/Province \_\_\_\_\_ Postal code/Zip \_\_\_\_\_

Country \_\_\_\_\_ E-mail address \_\_\_\_\_

Area code & Daytime phone \_\_\_\_\_ Fax \_\_\_\_\_ Member number, if applicable \_\_\_\_\_

### Purposes of ACM

ACM is dedicated to:

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

I agree with the Purposes of ACM:

Signature \_\_\_\_\_

ACM Code of Ethics:

<http://www.acm.org/about/code-of-ethics>

## choose one membership option:

### PROFESSIONAL MEMBERSHIP:

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

### STUDENT MEMBERSHIP:

- ACM Student Membership: \$19 USD
- ACM Student Membership plus the ACM Digital Library: \$42 USD
- ACM Student Membership PLUS Print CACM Magazine: \$42 USD
- ACM Student Membership w/Digital Library PLUS Print CACM Magazine: \$62 USD

All new ACM members will receive an  
ACM membership card.

For more information, please visit us at [www.acm.org](http://www.acm.org)

Professional membership dues include \$40 toward a subscription to *Communications of the ACM*. Student membership dues include \$15 toward a subscription to *XRDS*. Member dues, subscriptions, and optional contributions are tax-deductible under certain circumstances. Please consult with your tax advisor.

### RETURN COMPLETED APPLICATION TO:

Association for Computing Machinery, Inc.  
General Post Office  
P.O. Box 30777  
New York, NY 10087-0777

Questions? E-mail us at [acmhelp@acm.org](mailto:acmhelp@acm.org)  
Or call +1-800-342-6626 to speak to a live representative

**Satisfaction Guaranteed!**

### payment:

Payment must accompany application. If paying by check or money order, make payable to ACM, Inc. in US dollars or foreign currency at current exchange rate.

Visa/MasterCard     American Express     Check/money order

Professional Member Dues (\$99 or \$198)    \$ \_\_\_\_\_

ACM Digital Library (\$99)    \$ \_\_\_\_\_

Student Member Dues (\$19, \$42, or \$62)    \$ \_\_\_\_\_

**Total Amount Due**    \$ \_\_\_\_\_

Card # \_\_\_\_\_ Expiration date \_\_\_\_\_

Signature \_\_\_\_\_





Vinton G. Cerf

DOI:10.1145/2398356.2398358

# What's a Robot?

I am a big science fiction fan and robots have played a major role in some of my favorite speculative universes. The prototypical robot story came in the form of a play by

Karel Čapek called “R.U.R.” that stood for “Rossum’s Universal Robots.” Written in the 1920s, it envisaged android-like robots that were sentient and were created to serve humans. “Robot” came from the Russian word “работать” (“rabotat,” which means “work”). Needless to say, the story does not come out well for the humans. In a more benign and very complex scenario, Isaac Asimov created a universe in which robots with “positronic” brains serve humans and are barred by the Three Laws of Robotics from harming humans:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

A “zeroth” law emerges later:

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

In most formulations, robots have the ability to manipulate and affect the real world. Examples include robots that assemble cars (or at least parts of them). Less facile robots might be devices that fill cans with food or bottles with liquid and then close them up. The most primitive robots might not normally even be considered robots in normal parlance. One example is a temperature control for a home heating system that relies on a piece of bi-metal material that

expands differentially causing a circuit to be closed or opened depending on the ambient temperature.

I would like to posit, however, that the notion of robot could usefully be expanded to include programs that perform functions, ingest input and produce output that has a perceptible effect. A weak notion along these lines might be simulations in which the real world remains unaffected. A more compelling example might be high-frequency stock trading systems whose actions have very real-world consequences in the financial sector. While nothing physical happens, real-world accounts are impacted and, in some cases, serious consequences emerge if the programs go out of control leading to rapid market excursions. Some market meltdowns have been attributed to large numbers of high-frequency trading programs all reacting in similar ways to inputs leading to rapid upward or downward motion of the stock market.

Following this line of reasoning, one might conclude that we should treat as robots any programs that can have real-world, if not physical, effect. I am not quite sure where I am heading with this except to suggest that those of us who live in and participate in creation of software-based “universes” might wisely give thought to the potential impact that our software might have on the real world. Establishing a sense of professional responsibility in the computing community might lead to increased safety and reliability of software products

and services. This is not to suggest that today’s programmers are somehow irresponsible but I suspect that we are not uniformly cognizant of the side effects of great dependence on software products and services that seems to increase daily.

A common theme I hear in many conversations is concern for the fragility or brittleness of our networked- and software-driven world. We rely deeply on software-based infrastructure and when it fails to function, there can be serious side effects. Like most infrastructure, we tend not to think about it at all until it does not work or is not available. Most of us do not lie awake worried that the power will go out (but, we do rely on some people who *do* worry about these things). When the power *does* go out, we suddenly become aware of the finiteness of battery power or the huge role that electricity plays in our daily lives. Mobile phones went out during Hurricane Sandy because the cell towers and base stations ran out of power either because of battery failure or because the back-up generators could not be supplied with fuel or could not run because they were underwater.

I believe it would be a contribution to our society to encourage deeper thinking about what we in the computing world produce, the tools we use to produce them, the resilience and reliability that these products exhibit and the risks that they may introduce. For decades now, Peter Neumann has labored in this space, documenting and researching the nature of risk and how it manifests in the software world. We would all do well to emulate his lead and to think whether it is possible that the three or four laws of robotics might motivate our own aspirations as creators in the endless universe of software and communications.

Vinton G. Cerf, ACM PRESIDENT

# Computer Science Is Not a Science

**T**O THE QUESTION Vinton G. Cerf addressed in his President's Letter "Where Is the Science in Computer Science?" (Oct. 2012), my first answer would be that there isn't any. Max Goldstein, a mentor of mine at New York University, once observed that anything with "science" in its name isn't really a science, whether social, political, or computer. A true science like physics or chemistry studies some aspect of physical reality. It is not concerned with how to build things; that is the province of engineering. Some parts of computer science lie within mathematics, but mathematics is not a science and is rarely claimed to be one.

What we mislabel as computer science would more aptly be named "computology"—the study of computational processes and the means by which they can be realized. Its components can broadly be grouped into three interdependent areas: software engineering, hardware engineering, and the mathematics of computation. Just as the underlying discipline of chemical engineering is chemistry, the underlying discipline of software engineering is mathematics.

But not so fast. To qualify as a subject of science, a domain of inquiry needs two qualities: regularity and physicality. Reproducible experiments are at the heart of the scientific method. Without regularity they are impossible; without physicality they are meaningless. Digital computers, which are really just very large and complicated finite-state machines, have both these qualities. But digital computers are artifacts, not part of the natural world. One could argue either way whether that should disqualify them as a subject of science. Quantum computing and race conditions complicate the picture but not in a fundamental way.

None of this detracts from Cerf's essential point—that when we design software we rarely understand the full implications of our designs. As he said, it is the responsibility of the computing community, of which ACM is a vital

part, to develop tools and explore principles that further that understanding and enhance our ability to predict the behavior of the systems we build.

**Paul W. Abrahams**, Deerfield, MA

In his President's Letter (Oct. 2012), Vinton G. Cerf wrote: "We have a responsibility to pursue the science in computer science [...and to develop] a far greater ability to make predictions about the behavior of these complex, connected, and interacting systems." This is indeed a worthwhile cause that would likely increase the reliability and trustworthiness of the whole field of computing. But, having picked up the gauntlet Cerf threw down, how do I make that cause fit the aspects of computer science I pursue every day?

Cerf discussed the problems software developers confront predicting the behavior of both software systems and the system of people developing them. As a professional developer, I have firsthand experience. Publishing a catalog of the issues I find might lead analysts to identify general problems and suggest mitigations would be subject to two limitations: probably not interesting enough for journal editors to want to publish and my employers likely viewing its content as commercially sensitive.

I could instead turn to the ACM Digital Library and similar resources, looking for ways to apply it to my professional work. However, this also has limitations; reading journal articles is a specialized, time-consuming art, and the guidance I would need to understand what and how results are relevant is often not available. Many of the "classic results" quoted by professionals turn out to be as verifiable as leprechaun sightings.<sup>1</sup>

To the extent the creation of software can be seen as "computer science," such creation is today two distinct fields: creating software and researching ways software can be created. If we would accept the responsibility Cerf has bestowed upon us, we would have to create an interface disci-

pline—call it "computer science communication"—between these fields.

**Graham Lee**, Leicester, U.K.

## Reference

1. Bossavit, L. *The Leprechauns of Software Engineering*. Leanpub, Vancouver, B.C., 2012; <https://leanpub.com/leprechauns>

## Only Portfolios Mitigate Risk Well

Peter G. Neumann's "Inside Risks" Viewpoint "The Foresight Saga, Redux" (Oct. 2012) addressed how to provide security but fell short. Though security requires long-term approaches and research advances, traditional incentives target quick rewards. I teach a graduate course on IT strategy and policy largely focused on this dilemma. When technology moved slowly, slow acquisition and delayed delivery caused minor losses. Now, however, along with improvement due to technology innovation, delays in exploiting advanced technology incur exponentially increased opportunity costs. Most businesses cannot wait for high-trust solutions or systems that significantly surpass state-of-the-art quality. Likewise, most government systems are already too costly and too late, in part because they try to address an unreasonably large number of requirements.

The risk-management problem necessitates a portfolio-management approach. In the context of IT systems for business or government, it would be more affordable and practical to create multiple alternatives and fallback options and not depend on a single system where failure would be devastating. In addition, applications should be separated from research and funded appropriately. It would be great to have a secure Internet, unbreakable systems, and uniformly trained people, but such goals are not practical today. The focus should instead be on risk mitigation, resilience, and adaptation, even though the incentives for moving quickly are often irresistible. "Ideal" systems are indeed the enemy of practical portfolios built to withstand a range of risks.

**Rick Hayes-Roth**, Monterey, CA

### Clock-Free Computing

As an undergrad at MIT in 1972, I took a course in asynchronous design from Prof. Jonathan Allen. Having some background at the time in digital circuitry, it was exciting to see this latest work as presented by Allen, and it was easy to imagine that in a few years most computers and other digital systems would operate this way. The reasoning was much like what Ivan Sutherland advocated in his Viewpoint “The Tyranny of the Clock” (Oct. 2012). Following graduation I started out in the working world designing digital hardware. Industry opens a student’s eyes to the real world, and it was clear rather quickly that the synchronous world would not in fact budge for a long time. Though my work today involves mostly software, I still see the appeal of asynchronous logic and hope the vision of asynchronous computing finally takes hold. We could use more calls-to-arms like Sutherland’s: “The clock-free design paradigm must eventually prevail.” I look forward to that day, just as I look forward to another paradigm that should eventually prevail—parallel processing.

**Larry Stabile**, Cambridge, MA

### Relational Model Obsolete

I write to support and expand on Erik Meijer’s article “All Your Database Are Belong to Us” (Sept. 2012). Relational databases have been very useful in practice but are increasingly an obstacle to progress due to several limitations:

*Inexpressiveness.* Relational algebra cannot conveniently express negation or disjunction, much less the generalization/specialization connective required for ontologies;

*Inconsistency non-robustness.* Inconsistency robustness is information-system performance in the face of continually pervasive inconsistencies, a shift from the once-dominant paradigms of inconsistency denial and inconsistency elimination attempting to sweep inconsistencies under the rug. In practice, it is impossible to meet the requirement of the Relational Model that all information be consistent, but the Relational Model does not process inconsistent information correctly. Attempting to use transactions to remove contradictions from, say, relational

medical information is tantamount to a distributed-denial-of-service attack due to the locking required to prevent new inconsistencies even as contradictions are being removed in the presence of interdependencies;

*Information loss and lack of provenance.* Once information is known, it should be known thereafter. All information stored or derived should have provenance; and

*Inadequate performance and modularity.* SQL lacks performance because it has parallelism but no concurrency abstraction. Needed are languages based on the Actor Model (<http://www.robust11.org>) to achieve performance, operational expressiveness, and inconsistency robustness. To promote modularity, a programming language type should be an interface that does not name its implementations contra to SQL, which requires taking dependencies on internals.

There is no practical way to repair the Relational Model to remove these limitations. Information processing and storage in computers should apply inconsistency-robust theories<sup>1</sup> processed using the Actor Model<sup>2</sup> in order to use argumentation about known contradictions using inconsistency-robust reasoning that does not make mistakes due to the assumption of consistency.

This way, expressivity, modularity, robustness, reliability, and performance beyond that of the obsolete Relational Model can be achieved because computing has changed dramatically both in scale and form in the four decades since its development. As a first step, a vibrant community, with its own international scientific society, the International Society for Inconsistency Robustness (<http://www.isir.ws>), conducted a refereed international symposium at Stanford University in 2011 (<http://www.robust11.org>); a call for participation is open for the next symposium in the summer of 2014 (<http://www.ir14.org>).

**Carl Hewitt**, Palo Alto, CA

#### References

1. Hewitt, C. Health information systems technologies. Stanford University CS Colloquium EE380, June 6, 2012; <http://HIST.carlhewitt.info>
2. Hewitt, C., Meijer, E., and Szyperski, C. *The Actor Model*. Microsoft Channel 9 Videos; <http://channel9.msdn.com/Shows/Going+Deep/Hewitt+Meijer+and+Szyperski-The-Actor-Model-everything-you-wanted-to-know-but-were-afraid-to-ask>

### Design Software for the Unknown

In his article “Software Needs Seatbelts and Airbags” (Sept. 2012), Emery D. Berger identified typical flaws in coding, as well as techniques that might help prevent them, addressing a major conundrum taking much of a typical programmer’s time: Much more time goes for tracking bugs than for writing useful programs.

Berger’s analogy of software techniques and automobile accessories was illuminating, though computer technology has generally outpaced the automobile by orders of magnitude. Some have said automobiles could go one million miles on a single gallon of fuel, reaching its destination in one second, if automobile engineers were only as bright as computer scientists. Others have said we would be driving \$25 cars that get 1,000 miles to a gallon if they were only designed by computer scientists instead of by automobile engineers. But the analogy should not be stretched too far. An advocate of software reliability might say seatbelts and bumpers are not intended to protect drivers from design errors but from their own errors, or bad driving; in software the analogous problem is designer error. If defects are discovered, cars are recalled and defective parts replaced. Some software products that update themselves multiple times a day crash anyway because analogous seatbelts and airbags in software are a luxury.

The defects Berger covered are more analogous to bad plumbing and crossed wires. Moreover, software developers may not even know all the components and functions in the software they deliver. Though perfect in terms of memory handling and buffer-overflow management, software can become a work of art during development, with no way to completely anticipate how it will perform under unknown circumstances.

I would like to see researchers of software code take a look at something I call “mind of software,” aiming for ways to make software more safe and predictable for common use.

**Basudeb Gupta**, Kolkata, India

*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](mailto:letters@cacm.acm.org).

© 2013 ACM 0001-0782/13/01



# MentorNet

e-Mentoring for diversity in engineering and science

## The Extra Edge for ACM Members – MentorNet Matching

### Have you ever asked yourself:

- What's it like to work in industry?
- What is graduate school like, and is it for me?
- How do I manage a career and a life?

### Check out MentorNet!

ACM partners with MentorNet to promote e-mentoring between students and professionals.

- As **protégés**, students gain invaluable career advice, encouragement and support.
- As **mentors**, professionals lend expertise to help educate and inspire young professionals.

Protégés are matched in one-on-one email relationships with mentors—from industry, academia, and government—who have applicable experience in relevant technology, engineering, and scientific fields.

### Mentors, Protégés help each other:

*“My mentor, Brian Kernighan, helped me navigate graduate school. Having learned the value of mentoring, I became a mentor myself...”*

— Mary Fernandez,  
Principal Technical Staff Member,  
ATT Labs - Research

*“I am fortunate to have a mentor who spends his time in not only answering my questions, but also in directing my career path...”*

— Emeka (Chukwuemeka  
Nwankwo), student at Nnamdi  
Azikiwe University Awka, Nigeria

### Who can be a protégé?

ACM Members who are Undergraduates, Graduates, Post-Doctoral students, or Untenured Faculty.

### Who can mentor?

ACM Professional Members

### How does the E-Mentoring Program Work?

1. Register in the MentorNet Community by providing your name, valid email address, and username and password. MentorNet Community Registration form at <http://www.mentor.net/join.aspx>.
2. Sign in and click on the appropriate Find a Mentor or Be a Mentor button.
3. The official e-mentoring relationship lasts approximately 8 months.



Association for  
Computing Machinery

*Advancing Computing as a Science & Profession*

Learn more at <http://www.acm.org/mentornet> and <http://www.mentor.net/>



During my two years of service, I was honored to be part of some key initiatives that will hopefully continue to propel ACM as the world's largest and most prestigious scientific and computing society.



DOI:10.1145/2398356.2398362

Alain Chesnais

## ACM's Annual Report

FY12 was an outstanding year for ACM. Membership reached an all-time high for the 10<sup>th</sup> consecutive year. We witnessed our global hubs in Europe, India, and China

take root and flourish. And we elevated ACM's overriding commitment to educating future generations about the wonders of computer science to a new level by joining forces with other associations, corporations, and scientific societies dedicated to the same cause.

The fiscal year culminated in a landmark event honoring the life and legacy of Alan Turing on what would have been his 100<sup>th</sup> birthday. ACM's Turing Centenary Celebration drew more than 1,000 participants, including over 30 ACM A.M. Turing Award laureates and a host of other world-renowned computer scientists and technology pioneers for a historic celebration of computer science and Turing's wide-ranging contributions to the field.

For me, the event served as an extraordinary way to close my ACM presidency. During my two years of service, I was honored to be part of some key initiatives and projects that will hopefully continue to propel ACM as the world's largest and most prestigious scientific and computing society. I was

thrilled to see membership exceed the 100,000 mark this year. Even in these economically tenuous times, ACM realized a 7.7% increase in membership over last year. Much of this growth is attributable to our international initiatives, principally China.

I am particularly proud of ACM's leadership role in Computing in the Core, a nonpartisan advocacy coalition striving to promote computer science education as a core academic subject in K-12 education. The success of this initiative is key to giving young people the college- and career-readiness, knowledge, and skills necessary in a technology-focused society.

It is a remarkable time to be part of the computing arena. Thanks to our legion of devoted volunteers, supportive industry sponsors, and steadfast members, ACM is able to make a real difference in setting the course for educating and encouraging generations to come. It was my great pleasure to serve as your president.

*Alain Chesnais*, ACM PRESIDENT

# Highlights of ACM Activities: July 1, 2011–June 30, 2012

*ACM, the Association for Computing Machinery, is an international scientific and educational organization dedicated to advancing the arts, sciences, and applications of information technology.*

## Publications

The centerpiece of ACM Publications is the ACM Digital Library (DL) serving as the primary distribution mechanism for all the association's publications as well as host to scientific periodicals and a set of conference proceedings from external organizations. The DL, now available at 2,650 institutions in 64 countries, boasts an estimated 1.5 million users worldwide. The result of this widespread availability led to more than 15 million full-text downloads in FY12.

ACM is committed to continually increasing the scope of material available via the DL. Last year, over 26,000 full-text articles were added, bringing total DL holdings to 350,000 articles. ACM's *Guide to Computing Literature* is also integrated within the DL. More than 285,000 works were added to the bibliographic database in FY12, bringing the total *Guide* coverage to more than two million works.

ACM is the publisher of 74 periodicals, including 40 journals and transactions, eight magazines, and 26 newsletters as of year-end FY12. During the year, ACM added 479 conference and related workshop proceedings to its portfolio.

In addition, a collection of 1,200 e-books was assimilated into the DL last year, available to all ACM members. Moreover, the ACM International Conference Proceedings Series (ICPS) published 150 new volumes, the most in any year since the program was launched.

Under the guidance and strategic planning of ACM's Publications Board, the ACM Author-Izer Service debuted in the Digital Library this year. This service enables authors to generate and post links on their

home page or institutional repository for visitors to download the definitive version of their articles at no charge.

Production began on two new journals: *Transactions on Interactive Intelligent Systems* and *Transactions on Economics and Computation*. The ACM Publications Board also approved a new journal—*Transactions on Spatial Algorithms and Systems*—to debut in FY13.

The ACM-W newsletter provides members an opportunity to share experiences, provoke discussion, and present research findings related to their mission. The newsletter was designed to appeal to anyone interested in promoting women in computing efforts at all stages of their career.

## Education

ACM continues to lead the computer science education community through the work of the ACM Education Board, the ACM Education Council, ACM SIGCSE, Computer Science Teachers Association (CSTA), and ACM Education Policy committee.

ACM remains at the forefront of the Computing in the Core (CinC) advocacy coalition working to promote computer science as a distinct discipline in K–12 U.S. education. This year CinC engaged both the House and Senate committees responsible for reauthorization of the Elementary

and Secondary Education Act to advocate for computer science. Indeed, the coalition's membership has grown by more than 50% over the last year.

ACM launched a new initiative to address the grand challenges of expanding K–12 computer science education in the U.S. Google, Microsoft, ACM, CSTA, the National Center for Women and Information Technology, and the National Science Foundation are partnering to lay the groundwork for scaling AP computer science reform.

ACM's Education Board is directing the formation of an annual Taulbee-like survey for non-Ph.D.-granting institutions in computing.

The CSTA continues to thrive as a key component in ACM's efforts to see real computer science count at the high school level. CSTA membership reached an all-time high of 12,000 this year. In addition, CSTA has been an active and important partner in AP computer science reform.

Several SIGs hosted innovative educational programs and special projects throughout the year. SIG Bioinformatics organized a special program—Women in Bioinformatics—that was sponsored by the U.S. National Science Foundation. SIGCAS and the Committee on Professional Ethics (COPE) jointly ran a workshop on teaching computer ethics. And SIGCOMM maintains its education website (<http://education.sigcomm.org>) where members of the community can share education-related resources.

## Professional Development

The Practitioners Board and Professional Development Committee directed many new products and initiatives designed for computing professionals and managers. ACM's Learning Center (<http://learning.acm.org>) offers products, services, videos, resources, webinars, and courses designed especially for prac-

**ACM is committed to continually increasing the scope of material available via the Digital Library.**

tioners. ACM's Learning Webinar series opened with a program on smart devices and cloud computing, followed by a webinar on cybersecurity. The first two webinars proved a huge success attracting almost 9,000 registrants collectively. More webinars are in the works for this thriving new project.

The Board also oversees the development of Tech Packs (<http://tech-pack.acm.org/>) and Learning Paths (<http://learning.acm.org/path/>).

There are currently five Tech Packs (with a sixth on the way) that are designed to provide a significant learning resource for emerging areas of computing not directly covered by an ACM SIG, conference, or publication. Learning Paths offer practical, hands-on educational training tools designed to help IT professionals extend their skill sets.

*ACM Queue*, the online magazine for professionals spirited by the Practitioners Board, again surpassed the million-pageview threshold, with over 1.2 million pages viewed over the last 12 months.

### Public Policy

ACM's U.S. Public Policy Council (US-ACM) educates policymakers in many areas of potential legislation, including bills on Internet monitoring, patent reform, e-voting, privacy, and security. This year the committee explored new ways to contribute expertise to the development of policy connected to computing, offering comments on how the National Strategy for Trusted Identities in Cyberspace should be managed, on the proposed online verification system for the Social Security Administration, and on the U.S. Commerce Department's report on cybersecurity and innovation, among many other proposals.

The Committee on Computers and Public Policy assists ACM in a variety of globally relevant issues pertaining to computers and public policy. The efforts of this committee help make the association more visible worldwide. Most notably, CCP's highly respected *ACM Forum on Risks to the Public in Computers and Related Systems* designed to share and discuss the potential and serious computer-related risks with a global audience.

## ACM Europe, ACM India, and ACM China saw significant increases in the number of chapters established in FY12.

### Students

The 36<sup>th</sup> Annual ACM International Collegiate Programming Contest (ICPC) took place in Warsaw, Poland with 122 teams competing in the World Finals. Earlier rounds of the competition included nearly 30,000 contestants representing 2,200 universities from 85 countries. Financial and systems support for ICPC is provided by IBM. The top four teams won gold medals as well as employment or internship offers from IBM.

The ACM Student Research Competition (SRC), sponsored by Microsoft Research, continues to offer a unique forum for undergraduate and graduate students to present their original research at well-known ACM-sponsored and co-sponsored conferences before a panel of judges and attendees. This year's SRC saw graduate and undergraduate winners compete against more than 213 participants in contests held at 15 ACM conferences.

ACM-W added seven new student chapters this year, bringing the total number to 37, with six being interactive chapters.

### Internationalization

ACM Europe and ACM India finished their second full year of operation. Both saw significant increases in the number of chapters established during the last 12 months. Progress was made on two ACM-Europe/Informatics-Europe joint activities. And a third successful ACM India event was held last January attended by over 600 Indian computing students, faculty, and professionals.

### ACM Council

#### PRESIDENT

Alain Chesnais

#### VICE PRESIDENT

Barbara G. Ryder

#### SECRETARY/TREASURER

Alexander L. Wolf

#### PAST PRESIDENT

Wendy Hall

#### SIG GOVERNING BOARD CHAIR

Vicki Hanson

#### PUBLICATIONS BOARD

##### CO-CHAIRS

Ronald Boisvert

Jack Davidson

#### MEMBERS-AT-LARGE

Vinton G. Cerf

Carlo Ghezzi

Anthony Joseph

Mathai Joseph

Kelly Lyons

Mary Lou Soffa

Salil Vadhan

#### SGB COUNCIL

##### REPRESENTATIVES

Joseph A. Konstan

G. Scott Owens

Douglas Terry

#### REGIONAL COUNCIL CHAIRS

##### ACM Europe

Fabrizio Gagliardi

##### ACM India

Anand Deshpande,

PJ Narayanan

##### ACM China

Jianguang Sun

##### ACM-W

Elaine Weyuker, Valerie Barr

##### USACM

Eugene Spafford

##### Education Board

Andrew McGettrick

##### Practitioners Board

Stephen R. Bourne

### ACM Headquarters

#### EXECUTIVE DIRECTOR/CEO

John R. White

#### DEPUTY EXECUTIVE DIRECTOR/COO

Patricia M. Ryan

ACM China finished its first year with an increase in the number of conferences, chapters, and memberships established in the region.

ACM's commitment to women in computing was further strengthened with the relaunch of ACM-W activities in India.

Internationalization continues to be a major focus of SIGCSE, where members are working with Informatics Europe representatives to discuss the possibility of forming a new SIGCSE-like education conference in Europe.

### Electronic Community

A website honoring ACM's A.M. Turing Award recipients launched in FY12. Curated by members of ACM's History Committee, the site (<http://amturing.acm.org/>) provides a robust collection of Turing laureates, including citations, bibliographic information, Turing lectures, résumés, and other works. The site also spotlights images and videos from ACM's recent Turing Centenary Celebration, a landmark event recognizing the computer science icon on the 100<sup>th</sup> anniversary of his birth.

Three ACM magazines debuted new websites this year. *ACM Inroads* (<http://inroads.acm.org/>), the quarterly magazine for computing educators, launched its first site last summer. *Computers in Entertainment* (<http://cie.acm.org/>) introduced a new interactive site incorporating video features and interviews, games, art, music, movies and research. *interactions*, the bimonthly magazine for professionals in CHI-related disciplines, launched a new site (<http://interactions.acm.org/>) last spring that quickly garnered a design award from Interactive Media.

By the end of the year, *Communications of the ACM* became accessible as an easy-to-use mobile application for iPhones, iPads, and Android devices. These new downloadable apps enable ACM's members to access the flagship magazine in a new way.

Many ACM SIGs introduced new and/or improved websites making significant use of social media outlets to reach out to members. SIGBED overhauled its website (<http://sigbed.blogspot.com/>), as did SIGDOC (<http://sigdoc.acm.org/>), SIGACCESS ([## Balance Sheet: June 30, 2012 \(in Thousands\)](http://www.sigac-</a></p>
</div>
<div data-bbox=)

### ASSETS

|  |                  |
|--|------------------|
| Cash and cash equivalents                                      | \$30,438         |
| Investments  | 62,741           |
| Accounts receivable and other assets                           | 6,506            |
| Deferred conference expenses                                   | 8,381            |
| Fixed assets, net of accumulated depreciation and amortization | 1,220            |
| <b>Total Assets</b>  | <b>\$109,286</b> |

### LIABILITIES AND NET ASSETS

|   |                  |
|---|------------------|
| Liabilities:  |                  |
| Accounts payable, accrued expenses, and other liabilities | \$10,934         |
| Unearned conference, membership, and subscription revenue | 23,527           |
| <b>Total liabilities</b>                                  | <b>\$34,461</b>  |
| Net assets:   |                  |
| Unrestricted  | 67,906           |
| Temporarily restricted                                    | 6,919            |
| <b>Total net assets</b>                                   | <b>74,825</b>    |
| <b>Total liabilities and net assets</b>                   | <b>\$109,286</b> |

|   |              |
|---|--------------|
| Optional Contributions Fund — Program Expense (\$000) |              |
| Education Board accreditation                         | \$95         |
| USACM Committee                                       | 18           |
| <b>Total expenses</b>                                 | <b>\$113</b> |

cess.org/), and SIGBIOINFORMATICS (<http://sigbioinformatics.org/>).

SIGMOD established a new blog site (<http://wp.sigmod.org/>) to catch the pulse of its community on exciting and controversial topics and in its first seven months of operation, SIGHPC created a Web presence (<http://www.sighpc.org/>) that details the mission and goals of this new thriving SIG.

### Conferences

FY12 closed with a milestone celebration of the 100th anniversary of Alan Turing's birth as well as of the past and future of computing. The ACM A.M. Turing Centenary Celebration

was a two-day event that drew over 1,000 attendees, including 32 of the 39 living ACM A.M. Turing Award laureates—the largest single gathering of Turing Award recipients in computing history. The landmark event recognized Turing's contribution to computer science as well as presented an array of panel discussions featuring the laureates and other industry leading lights.

SIGGRAPH 2011 welcomed 15,872 artists, research scientists, gaming experts and developers, filmmakers, students, and academics from 74 countries to Vancouver—breaking the city's previous conference attendance



**Statement of Activities: Year ended June 30, 2012 (in Thousands)**

| <b>REVENUE</b>                           | <b>Unrestricted</b> | <b>Temporarily Restricted</b> | <b>Total</b>     |
|--|---------------------|-------------------------------|------------------|
| Membership dues                          | \$8,636             |                               | \$8,636          |
| Publications                             | 19,249              |                               | 19,249           |
| Conferences and other meetings           | 24,686              |                               | 24,686           |
| Interests and dividends                  | 1,767               |                               | 1,767            |
| Net appreciation of investments          | (398)               |                               | (398)            |
| Contributions and grants                 | 3,672               | \$2,027                       | 5,699            |
| Other revenue                            | 222                 |                               | 222              |
| Net assets released from restrictions    | 1,310               | (1,310)                       | 0                |
| <b>Total Revenue</b>                     | <b>59,144</b>       | <b>717</b>                    | <b>59,861</b>    |
| <b>EXPENSES</b>                          |                     |                               |                  |
| Program:                                 |                     |                               |                  |
| Membership processing and services       | \$767               |                               | \$767            |
| Publications                             | 11,290              |                               | 11,290           |
| Conferences and other meetings           | 22,354              |                               | 22,354           |
| Program support and other                | 8,712               |                               | 8,712            |
| <b>Total</b>                             | <b>43,123</b>       |                               | <b>43,123</b>    |
| Supporting services:                     |                     |                               |                  |
| General administration                   | 10,135              |                               | 10,135           |
| Marketing                                | 1,138               |                               | 1,138            |
| <b>Total</b>                             | <b>11,273</b>       |                               | <b>11,273</b>    |
| <b>Total expenses</b>                    | <b>54,396</b>       |                               | <b>54,396</b>    |
| Increase (decrease) in net assets        | 4,748               | 717                           | 5,465            |
| Net assets at the beginning of the year  | 63,158              | 6,202                         | 69,360           |
| <b>Net assets at the end of the year</b> | <b>\$67,906*</b>    | <b>\$6,919</b>                | <b>\$74,825*</b> |

\* Includes SIG Fund balance of \$33,028K

records. SIGGRAPH's exhibition hall drew 156 industry organizations from 17 countries, half of which were outside the U.S.

The fourth annual SIGGRAPH Asia conference once again captured a wide spectrum of digital innovations. Over 7,500 professionals from 53 countries attended the conference and exhibition in Hong Kong, where over 300 presentations and panel discussions generated deep insight into the future developments in the field.

KDD remains one of the leading conferences on data mining and knowledge discovery. KDD 2011 drew an all-time record 714 submissions.

### Recognition

There were 94 new chapters chartered in FY12. Of the 17 new professional chapters, 12 were internationally based; of the 77 new student chapters, 35 were international.

The ACM Fellows Program recognized 46 members for their contributions to computing and computer science in FY12. The new inductees brought the number of ACM Fellows to over 700.

ACM also named 54 new Distinguished Members in FY12, of which there were four Distinguished Educators, one Distinguished Engineer, and 49 Distinguished Scientists, bringing the total number of Distinguished Members to 285.

### 2011 ACM Award Recipients

**ACM A.M. TURING AWARD**  
Judea Pearl

**ACM-Infosys Foundation Award in the Computing Sciences**  
Sanjeev Arora

**ACM/AAAI Allen Newell Award**  
Stephanie Forrest

**The 2012-2013 ACM-W Athena Lecturer Award**  
Nancy A. Lynch

**Grace Murray Hopper Award**  
Luis von Ahn

**ACM-IEEE CS 2012 Eckert-Mauchly Award**  
Algirdas Avizienis

**Karl V. Karlstrom Outstanding Educator Award**  
Hal Abelson

**Outstanding Contribution to ACM Award**  
Calvin C. (Kelly) Gotlieb

**Distinguished Service Award**  
William A. Wulf

**Paris Kanellakis Theory and Practice Award**  
Hanan Samet

**Software System Award**  
*Eclipse*

Gregory Adams  
John Duimovich  
Erich Gamma  
Kevin Haaland  
Julian Jones  
Philippe Mulet  
Stephen Northover  
Dave Thomson  
John Weigand

**ACM-IEEE CS Ken Kennedy Award**  
Susan L. Graham

**Doctoral Dissertation Award**  
Seth Cooper

**Honorable Mention**  
Aleksander Madry  
David Steurer

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.

twitter

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

DOI:10.1145/2398356.2398361

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

## Lost in Translation

*Daniel Reed on straddling the intellectual divide between technology experts and policymakers.*



**Daniel Reed**  
"Being Bilingual:  
Speaking  
Technology  
and Policy"

<http://cacm.acm.org/blogs/blog-cacm/110891>  
June 30, 2011

There is an old joke:

Q: What do you call someone who speaks only one language?

A: An American.

Certainly, being monolingual limits and constrains one's exposure to and understanding of the cultural and linguistic diversity that is our global human heritage. Alas, I fear the same is true in far too many domains where cross-cultural fertilization would inform and enlighten all parties.

I will spare you a meandering discourse on the theory of language origins, the Sapir-Whorf hypothesis, or the Indo-European language tree. Nor will I digress to discuss the intellectual divide that separates the sciences from the arts, for C.P. Snow has written far more eloquently about that than I can. Rather, I want to focus on a far more constrained and practical intellectual concern, the cultural gap separating technologists and policy experts.

### Divergent Ontology

Over the years, I have learned that being bilingual in matters of science and technology and in matters of strategy and policy is far rarer than I might have first hypothesized. Those of us with Ph.D's, Sc.D's, or research M.D.'s speak a particular argot largely incomprehensible to the general public and even to the learned and sophisticated in other domains. Similarly, those who live in the legislative and policy world depend on a vernacular that seems obscure and obtuse to those in technical domains.

The technical and policy communities lack shared cultural referents, created all too often by endemic pressure to differentiate. In consequence, the communities are often estranged, lacking an ontology of discourse to address their common problems and exploit their complementary skills. The power of consilience has long been known, as the tale of the Tower of Babel makes clear.

*And the Lord said, Behold, the people is one, and they have all one language; and this they begin to do; and now nothing will be restrained from them, which they have imagined to do. Go to, let us go down, and there confound their language, that they may not understand one another's speech.*

Today, we have our own Babel of misperceptions and miscommunications at the intersection of technology and technical policy. (For a few thoughts on consilience in a technological world, see my blog "Consilience: The Path To Innovation," Nov. 9, 2009.)

### Interpreting the Signs

The linguistic and cultural divergence of technologists and policy experts is no more evident than in the way they identify and select outcomes. If you have ever felt compelled to explain quantum efficiency when discussing silicon solar cells and renewable energy sources, electron mobility and leakage current when discussing the future of smartphones, Shannon's theorem and the Heaviside layer when explaining wireless communication, or transcriptional gene regulation when discussing the future of health care, you live on the technological side of the communication chasm. Conversely, if you are facile with poll dynamics and sampling error, macro- and microeconomics and their shifting theories of global economic impact, trade imbalances and structural unemployment; if you understand the distinct and important roles of the World Bank and the International Monetary Fund; and if you are adept

at reading the nuance of diplomatic language, then you live on the policy side of the communication chasm.

There are, of course, other tell-tale signs. If you own more than one T-shirt covered with Maxwell's equations (and you can explain them) and a statistically significant fraction of your wardrobe is festooned with technical conference logos, then you are definitely on the geek side. Conversely, if you own a closet full of suits, perhaps some bespoke, and you choose the suit, matching tie, shoes, and fountain pen based on your mood, those you expect to meet, and the venue, then you are likely a policy wonk.

I exaggerate, of course, on both counts to illustrate a point, though elements of the humorous stereotypes are real. I resonate with both caricatures for my closet is filled with both conference T-shirts and with a variety of suits. But—and this is important—I do not wear them at the same time.

### Bridging the Divide

How do we cross the intellectual divide, providing technical advice to policy experts in ways that they find useful and actionable? Equally importantly, how do we translate policy constraints—political, economic, and social—into contexts intelligible and actionable by technical experts?

The key in both cases is to respect the differences and value each bring, and place one's self in the other's situation. If you are a technical expert, this often means finding intuitive analogies that capture the key elements of the technical idea. For example, I recently explained the potential economic advantage of cloud computing by saying that it brought some of the efficiencies to organizational IT that big box retailers brought to consumer goods.

Had I explained the design of a cloud data center, the networking and content distribution network, and the infrastructure optimizations, I would have bewildered my audience. I simply wanted them to understand that familiar economic forces were driving the cloud transition and raise awareness of the policy implications. Any new technology can both create new jobs and spur economic development and disrupt an industry, creating un-

employment, which then strains the educational and social safety net.

Likewise, if you are a policy expert or a technical person facile in policy, you must explain the constraints and practicalities of government actions and budgets to your technical partners. As a technologist, one must respect those realities rather than disparage them. The policy world is a complex, dynamic system with deep and unexpected consequences from almost any major change.

It is a critical fallacy to believe that if a legislator or staff member had access to the same facts as a scientist or technologist, then he or she would draw the same conclusions about policy implications as the technologist. Policy discussions may begin with data gathering, but their outcomes are based on values, priorities, and trade-offs. Any government's agenda must be balanced against a myriad of social and political constraints, including education, social welfare, and national defense. Often this means finding ways to compromise and achieve part of one's goal. Ten percent of the objective can be victory and should be celebrated as such. There is often time to seek the next 10% at a future engagement.

### Moving Forward: Technical Policy Ambassadors

For those of us in science and technology, I believe we must encourage more of our colleagues to become bilingual. An increasing fraction of our world is shaped by technology, and it is incumbent on us to facilitate the discussion of technology, social welfare, economic development, environmental policy, security and privacy, health care and medicine, defense and protection, and innovation and discovery. If we are respectful of political constraints, we can help policymakers understand the pace of technological change and its possibilities and recognize that even a national legislative body cannot overturn the laws of physics.

Only by being ambassadors to both our technical colleagues and our policy partners can we constructively shape our future. We must find and expand those shared cultural referents, creating an ontology of technical

policy discourse, and we must reward our colleagues for such engagements.

Remember, it's okay to wear the conference T-shirt and the suit, just not at the same time.

### Reader Comment

*This article shows one of the many reasons that central planning actually does not work. No matter how much geekspeak a policy wonk understands, or how much economics a technogeek understands, one size does not fit all. Community organizers and career politicians make for very poor technology bosses. Even in economics, there is only one U.S. Congressman at this time that fully understands the damage that false (fiat) FED currency does to the body politic, for example.*

—Anonymous

Daniel Reed is vice president of Technology Policy at Microsoft.

© 2013 ACM 0001-0782/13/01

Coming Next Month in COMMUNICATIONS

**The Tail at Scale:  
Managing Latency  
Variability in Large-Scale  
Online Services**

**New Approaches  
to Security and  
Cloud Data**

**Symbolic Execution  
for Software Testing—  
30 Years Later**

**The Explosive Growth  
of PostDocs in  
Computer Science**

**Cloud Service  
Certification**

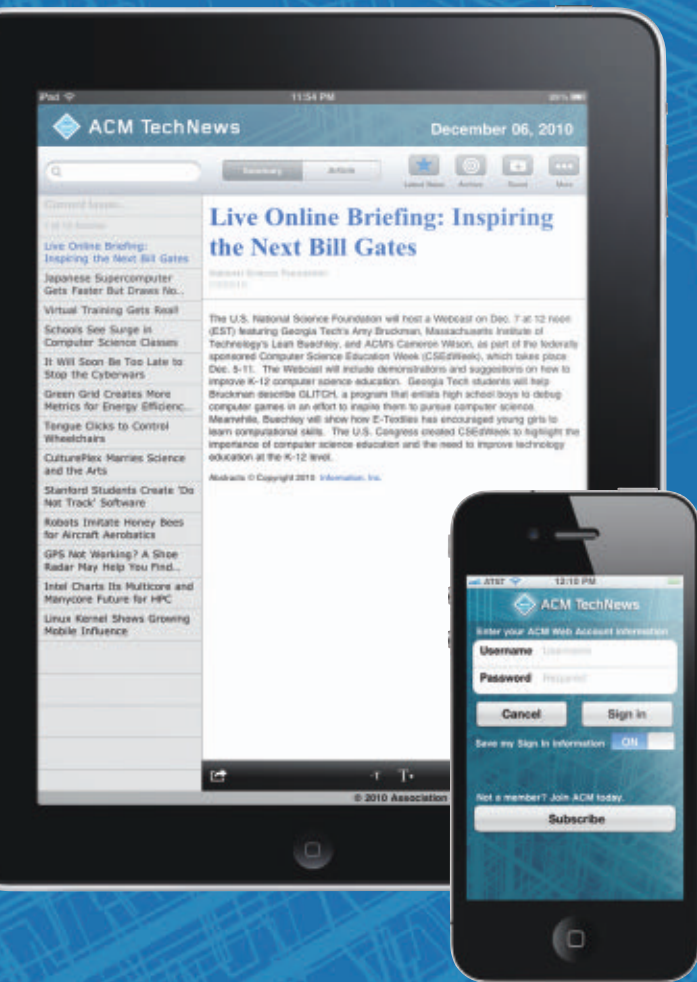
Plus the latest news about software simulations, computer models and intervention strategies, and Curiosity Rover.

# ACM TechNews Goes Mobile

## iPhone & iPad Apps Now Available in the iTunes Store

ACM TechNews—ACM's popular thrice-weekly news briefing service—is now available as an easy to use mobile apps downloadable from the Apple iTunes Store.

These new apps allow nearly 100,000 ACM members to keep current with news, trends, and timely information impacting the global IT and Computing communities each day.



### TechNews mobile app users will enjoy:

- **Latest News:** Concise summaries of the most relevant news impacting the computing world
- **Original Sources:** Links to the full-length articles published in over 3,000 news sources
- **Archive access:** Access to the complete archive of TechNews issues dating back to the first issue published in December 1999
- **Article Sharing:** The ability to share news with friends and colleagues via email, text messaging, and popular social networking sites
- **Touch Screen Navigation:** Find news articles quickly and easily with a streamlined, fingertip scroll bar
- **Search:** Simple search the entire TechNews archive by keyword, author, or title
- **Save:** One-click saving of latest news or archived summaries in a personal binder for easy access
- **Automatic Updates:** By entering and saving your ACM Web Account login information, the apps will automatically update with the latest issues of TechNews published every Monday, Wednesday, and Friday

The Apps are freely available to download from the Apple iTunes Store, but users must be registered individual members of ACM with valid Web Accounts to receive regularly updated content.

<http://www.apple.com/iphone/apps-for-iphone/> <http://www.apple.com/ipad/apps-for-ipad/>

# ACM TechNews



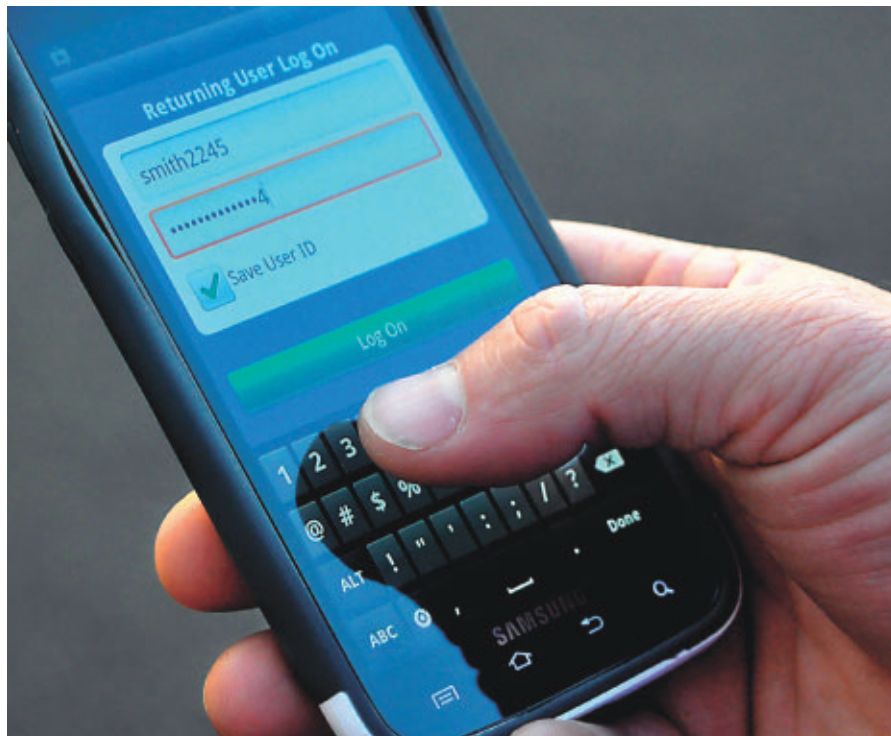
## Stopping the Leaks

*Side channels give out information that can be used to crack secrets, but researchers are identifying the holes and trying to close them.*

**C**OMPUTERS LEAK. EVEN machines running strongly encrypted programs give out information that can be used to infer secrets as electromagnetic signals, the size of packets, or the time it takes to store data in memory can all provide hints to what an application is doing, allowing attackers to deduce information and break cryptographic security.

The vulnerability of such side channels has been known for years, but as more computing moves into the cloud and onto smartphones and tablets, computer scientists are discovering new avenues of attack. And they are working on ways to thwart those attacks.

“One problem we are encountering increasingly is our computations aren’t happening in a safe, secure environment,” says Guy Rothblum, a Microsoft researcher. Many applications run on virtual machines inside servers, ostensibly walled off from neighboring programs. But other applications on the same computer can gain information about the state of that server and use it to infer what the first program is doing. And it is not only encryption keys that are vulnerable to such attacks, which are also known as physical attacks because they mea-



sure some physical characteristic of the operation.

“They are particularly devastating for cryptographic algorithms, but I would worry about it for my proprietary algorithm being run on a server or my sensitive data being stored on a server,” Rothblum says.

To combat the problem, Rothblum and Shafi Goldwasser, professor of electrical engineering and computer science at the Massachusetts Institute of Technology and Rothblum’s former advisor, developed a scheme to make computing more leakage resilient. They created a compiler that takes an

algorithm and breaks it into a series of modules that perform sub-computations. Although the sub-computations can leak, not all of their bits can be discovered. “The only assumption is that it doesn’t expose everything that’s happening in a sub-computation at once,” Rothblum says.

Each module is a cryptographically secure black box, allowing a hypothetical attacker to see the input and the output, but not the computation occurring inside the box. Data goes into the first module, which performs some computation, then passes the output to the next module, and so on down the line. “Each of them alone isn’t doing anything too sensitive,” explains Rothblum.

Even if the attacker can get some idea of what a particular module is doing, because he can only discover a small number of bits, he cannot learn enough to expose the whole algorithm. That is true even if the adversary can throw as many computing resources as he wants into the problem.

Essentially, the compiler elongates the program, and that can slow down

the computation substantially; to protect against  $k$  bits of leakage from each sub-computation, the program’s execution time expands by a factor between  $k$ -squared and  $k$ -cubed. But Rothblum is optimistic that, though it may take years, the computational overhead could be significantly reduced. One approach might be to identify a sensitive core piece of an algorithm and only protect that smaller piece. The significant fact is that the compiler is feasible. “We know it can be done, which we didn’t before,” he says. “Now it’s a matter of making it more efficient.”

### Bugging Smartphones

While Rothblum works on safer ways to compile programs, other researchers are discovering vulnerabilities in the growing universe of mobile devices. Suman Jana, a doctoral student at the University of Texas, Austin, won Best Student Paper at the IEEE Symposium on Security and Privacy in San Francisco in May for describing an attack that lets an adversary figure out which Web sites a smartphone user is

browsing. The attack takes advantage of the `proc` filesystem, a virtual file of process information, in Unix, which is the basis of Linux and Android, to uncover how much memory is allocated to a program. Other avenues can reveal similar information in Windows and iOS systems.

“Whenever a browser loads a page, it does a bunch of memory allocations,” Jana explains. “That gives away which pages you are browsing.”

The attacker would first use his own browser to visit Web pages and record how much memory the browser requires to render the page. Jana found that he could distinguish 30%–50% of the 100,000 Web sites he visited, depending on which browser he used. The process worked best for pages that did not change very much between visits.

Having disguised his malicious app as an innocuous one and tricked a smartphone user into downloading it, the attacker could watch the phone’s memory allocations and figure out which pages the user was seeing and finding, for instance, what diseases

## Security

# Emerging Cyber Threats Ring In The New Year

The coming year will experience a serious array of new and more sophisticated ways to seize and manipulate user data, according to the recently released “Emerging Cyber Threats Report for 2013.” Produced by the Georgia Tech Information Security Center (GTISC) and the Georgia Tech Research Institute (GTRI), the report forecasts several specific and ominous trends likely to occur in the months ahead.

The cloud is among the most threatening on the horizon, as it opens the potential for cloud-based botnets that could provide a way to create vast, virtual computing resources that will further convince cyber criminals to look for ways to co-opt cloud-based infrastructure for their own ends. For example, attackers could use stolen credit card information to purchase cloud-computing resources to create dangerous clusters of temporary virtual attack systems.

The report also predicts cyber criminals will continue to manipulate search engine algorithms and other automated mechanisms that control what information is presented to Internet users during a search. Indeed, researchers fear cyber criminals may use “search history poisoning” in the future to manipulate users’ search histories and use legitimate resources for illegitimate gains.

“It is easy for attackers to manipulate information on the Internet to have control of what a user sees and hence influence the user’s mind or decision process,” says Wenke Lee, director of GTISC. “It is very alarming that most users just assume personalization is all good, but what we find is that personalization algorithms provide ample opportunities to attackers/abusers to victimize users. My prediction is that what we have studied so far is only the tip of the iceberg: attackers will

come up with many more ways of information manipulation and we must race ahead to bring awareness to users and help them mitigate these threats.”

Researchers note worrisome security problems with the U.S. supply chain that are both difficult to detect and expensive to defend against. The concern is that security flaws in some of these systems not only render them vulnerable to compromise, but may in fact offer a backdoor for cyber espionage.

The proliferation of smartphones will continue to tempt attackers to exploit mobile vulnerabilities. Indeed, Lee predicts the kind of Web-based attacks experienced in the non-mobile world will show up in the mobile world this year. Researchers see browser-based attacks and attempts to subvert digital wallet apps as major threats. The developers of malicious software will employ various methods to hinder

malware detection, such as hardening their software with techniques similar to those employed in digital rights management, and exploiting the wealth of new interfaces and novel features on mobile devices.

“At the very least in the U.S., the infection rate of known mobile malware is *very* low. I think the main reasons are most users download their apps from well-vetted app stores, and it is quite hard to produce a very popular app to begin with,” says Lee, musing that if one can write a very popular app, he or she can make enough money legitimately, thus erasing any desire for malicious behavior. “On the other hand, privacy-undermining apps will continue to grow until users start to care more about privacy and better data access policies and mechanisms are developed and deployed.”

—Diane Crawford

he researches on WebMD. He could deduce even more information. For example, he can tell when someone logs into the dating site OKCupid, because a profile page uses more JavaScript than a log-in page. A free user of the site will see ads, which the attacker can infer by the use of the Flash plug-in.

Combining this process with other side-channel attacks could expose even more information, Jana says. For instance, an adversary can measure the time between keystrokes when a user is typing, and use that to guess which keys are being struck. That information can help distinguish between two pages with similar memory allocations. But if the user is on a log-in page or a shopping site, it can also help an attacker guess passwords or credit card numbers.

Hao Chen, associate professor of computer science at the University of California, Davis, developed Touchlogger, a program that can infer keystrokes on a smartphone or tablet by the way the device moves in response to pressure on the touch screen. Although apps require user permission to access sensitive data, such as location readings, the output of spatial sensors, such as the gyroscope or accelerometer, have not been considered sensitive, Chen says. It turns out those sensors can measure shifting and rotation when a user taps the onscreen keyboard, and a machine learning algorithm can predict which keys are pressed. "It gives you enough information that would improve your guessing, much better than a random guess," Chen says.

A different smartphone app, Soundcomber, also figures out what numbers a user presses, thus inferring phone and credit card numbers. The telephone system uses a combination of two tones to stand for each of the 10 digits on a keypad; capturing those tones and performing a frequency analysis allows the attacker to tease out the digits. The phone's microphone can also pick up different sounds the phone emits when the user presses a key, gaining another piece of identifying information.

The app also performs "light-weight" speech recognition, focusing


## Soundcomber can deduce where in an automated phone system a user is and record the individual's bank account number or Social Security number.

only on digits, says Kehuan Zhang, assistant professor at the Chinese University of Hong Kong, who helped develop Soundcomber as a student at Indiana University, Bloomington. The attacker would first call a series of banks and work his way through the various prompts on the automated system, from which language he preferred to different options to check a balance or make a payment, developing a model for each bank. Then, when a user called the bank, Soundcomber would deduce where in the automated system he was and record, say, his account number or Social Security number.

In all these cases, one defense is to add noise to the data being processed, making it more difficult for attackers relying on statistical analysis to tease out the signal they are after. But, like Rothblum's compiler, that increases the burden on computation, rendering a program less efficient. The memory attack might be thwarted by disabling the proc system, but Jana says that would cause a lot of legitimate programs to malfunction. Likewise, blocking Internet access to the accelerometer might break some desirable apps.

Chen has tried to address the problem with AndroidLeaks, a program that automatically examines smartphone applications and looks for potential information leaks. Users can then decide to restrict the permissions given to apps or not to

use them at all. Zhang, along with colleagues from Indiana and Shuo Chen of Microsoft Research, developed another program, Sidebuster, to examine the source code of Web applications. "We try to find out where information could be leaked and try to quantify the leaks," Zhang says. They have discovered vulnerabilities in popular online tax and investment software.

"It's very difficult to come up with general defense solutions," Zhang says. Rothblum's approach, for instance, would not work with Touchlogger, which records not the internal state of the system but its interaction with the user. Rothblum hopes his compiler can secure many algorithms from different sorts of side-channel attack. But until it is developed into something practical for widespread deployment, the best defense may be looking for potential leaks and plugging them when they are found. 

### Further Reading

*Cai, L., and Chen, H.*

**TouchLogger: Inferring keystrokes on touch screen from smartphone motion, 6th USENIX Workshop on Hot Topics in Security, San Francisco, CA, Aug. 9, 2011.**

*Goldwasser, S. and Rothblum, G.N.*

**How to computer in the presence of leakage, *Electronic Colloquium on Computational Complexity* 19, 10, Feb. 5, 2012.**

*Jana, S., and Shmatikov, V.*

**Memento: Learning secrets from process footprints, *IEEE Symposium on Security and Privacy, San Francisco, CA, May 20–23 2012.***

*Schlegel R., Zhang, K., Zhou, X.,*

*Intwala, M., Kapadia, A., and Wang, X.*

**Soundcomber: A stealthy and context-aware sound trojan for smartphones, *Proceedings of the 18th Annual Network & Distributed System Security Symposium, San Diego, CA, Feb. 6–9, 2011.***

*University of Washington Television*

**Side Channels and Clouds: New Challenges in Cryptography, <http://www.youtube.com/watch?v=6hQ5kvaEFkw>, Aug. 30, 2010.**

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

# Beyond Hadoop

*The leading open source system for processing big data continues to evolve, but new approaches with added features are on the rise.*

**W**HEN A NEW user visits the online music service Pandora and selects a station, the company's software immediately generates a playlist based on the preferences of its community. If the individual creates a Chopin station, for example, the string of songs would consist of the most popular renditions of the composer's music among Pandora's community. Once the new listener inputs a rating, clicking on either the "thumbs up" or the "thumbs down" icon, Pandora factors this preference into future selections. In effect, the service becomes smarter with the vote of each thumb.

Pandora will not discuss exactly how much data it churns through daily, but head of playlist engineering Eric Bieschke says the company has at least 20 billion thumb ratings. Once every 24 hours, Pandora adds the last day's data to its historical pool—not just thumbs, but information on skipped songs and more—and runs a series of machine learning, collaborative filtering, and collective intelligence tasks to ensure it makes even smarter suggestions for its users. A decade ago this would have been prohibitively expensive. Four years ago, though, Bieschke says Pandora began running these tasks in Apache Hadoop, an open source software system that processes enormous datasets across clusters of cheap computers. "Hadoop is cost efficient, but more than that, it makes it possible to do super large-scale machine learning," he says. Pandora's working dataset will only grow, and Hadoop is also designed for expansion. "It's so much easier to scale. We can literally just buy a bunch of commodity hardware and add it to the cluster."

Bieschke is hardly alone in his endorsement. In just a few years, Hadoop has grown into the system of choice for engineers analyzing big data in fields as diverse as finance, marketing, and bioinformatics. At the same time, the



Hadoop ecosystem components as visualized by Datameer.

changing nature of data itself, along with a desire for faster feedback, has sparked demand for new approaches, including tools that can deliver ad hoc, real-time processing, and the ability to parse the interconnected data flooding out of social networks and mobile devices. "Hadoop is going to have to evolve," says Mike Miller, chief scientist at Cloudera, a cloud database service based in Boston, MA. "It's very clear that there is a need for other tools." Indeed, inside and outside the Hadoop ecosystem, that evolution is already well under way.

## Distributing Data

Hadoop traces back to 2004, when Google published the second of a pair of papers [see the "Further Reading" list] describing two of the key ideas behind its search success. The first detailed the Google File System, or GFS, as a way of distributing data across hundreds or thousands of inexpensive computers. To glean insights from that

data, a second tool, called MapReduce, breaks a given job into smaller pieces, sends those tasks out to the different computers, then gathers the answers in one central node. The ideas were revolutionary, and soon after Google released the two papers, Yahoo! engineers and others quickly began developing open source software that would enable other companies to take advantage of the same breed of reliable, scalable, distributed computing that Google had long enjoyed.

The result, Apache Hadoop, consists of two main software modules. The Hadoop Distributed File System (HDFS) is similar to a file system on a single computer. Like GFS, it disperses enormous datasets among hundreds or thousands of pieces of inexpensive hardware. The computational layer, Hadoop MapReduce, takes advantage of the fact that those chunks of data are all sitting on independent computers, each with its own processing power. When a developer writes a program



to mine that data, the task is split up. “Each computer will look at its locally available data and run a little segment of the program on that one computer,” explains Todd Lipcon, an engineer at Palo Alto, CA-based Hadoop specialist Cloudera. “It analyzes its local data and then reports back the results.”

Although Hadoop is open source, companies like Cloudera and MapR Technologies of San Jose, CA, have found a market in developing additional services or packages around making it easier to use and more reliable. MapR, for example, has helped ancestry.com use Hadoop to carry out pattern matches on its enormous library of DNA details. After a customer sends in a saliva sample, the company can extract the basic biological code and use a Hadoop-based program to search for DNA matches—that is, potential mystery relatives—across its database.

### Analyzing Networks

For all its strengths in large-scale data processing, however, experts note MapReduce was not designed to analyze data sets threaded with connections. A social network, for example, is best represented in graph form, where in each person becomes a vertex and an edge drawn between two individuals signifies a connection. Google’s own work supports the idea that Hadoop is not set up for this breed of data: The company’s Pregel system, publicly described for the first time in 2009, was developed specifically to work with graph structures, since MapReduce had fallen short.

Along with a handful of students, University of Washington network scientist Carlos Guestrin recently released a new open source processing framework, GraphLab, that uses some of the basic MapReduce principles but pays more attention to the networked structure. The data distribution phase takes the connections into account. “If I know that you and I are neighbors in the graph, there will be some computation that needs to look at your data and my data,” Guestrin explains. “So GraphLab will try to look at our data on the same machine.”

The trick is that GraphLab partitions this data in a novel way. The standard method would have been to split the data into groups of highly connected

**“Hadoop is going to have to evolve,” says Mike Miller. “It’s very clear that there is a need for other tools.”**

vertices. In social networks, however, a few persons have a disproportionate number of connections. Those people, or vertices, could not be stuffed into single machines, which would end up forcing numerous computers to communicate. To avoid this inefficiency, Guestrin says GraphLab’s algorithm partitions data according to the edges, so that closely linked edges are on the same machines. A highly connected individual such as the pop star Britney Spears will still live on multiple pieces of hardware, but far fewer than with the standard technique. “It minimizes the number of places that Britney Spears is replicated,” Guestrin explains.

Hadoop, on the other hand, is agnostic to the structure of the data, according to Guestrin. Two pieces of information that should be analyzed on the same computer might end up in different clusters. “What ends up happening is that they have to move data around a lot,” Guestrin says. “We can be smart about how data is placed and how data is communicated between machines.”

Hadoop can often complete the same tasks as GraphLab, but Guestrin says his more efficient approach makes it much faster. On several common benchmark tests, such as a name recognition task in which an algorithm analyzes text and assigns different categories to words, GraphLab has completed the job 60 times faster, using the same hardware.

### Running in Real Time

In some cases, though, this is not fast enough. Today’s companies often want results in real time. A hedge fund might be looking to make a snap decision based on the day’s events. A

## ACM Member News

**DANIEL SPIELMAN WINS MACARTHUR ‘GENIUS’ AWARD**



If you received a \$500,000 grant for being a “genius,” how would you react? In the case of

Yale University computer scientist Daniel Spielman, he did not tell a soul other than his wife.

“I learned about the award in mid-September,” explains Spielman, who recently received a MacArthur Fellowship award. “But the MacArthur Foundation people asked me to keep it a secret until Oct. 1. Fortunately, I received an email before speaking to them. Otherwise, I would have tipped off my entire department with my screaming.”

As for spending the money? Spielman has no immediate plans. “It’s given in 20 installments over five years,” he says, “so I can’t do anything too crazy with it. But I plan to use it to provide more time to work on research.”

That is good news for the scientific community—and society. Spielman has devoted his career to the pursuit of abstract questions that address how to measure, predict, and regulate the environment and behavior. He helped develop error-correcting codes that are used in satellite video broadcasts, as well as optimization algorithms that support computational science and machine learning on massive datasets, among other applications.

“Most of my research is about the design of faster algorithms. They don’t just ‘speed things up.’ They change what is reasonable to do. You wouldn’t, say, browse the Internet on a phone if it took 10 minutes to load a page. Sophisticated algorithms compress the communications of the phone to enable it to transmit reliably without using too much power. Web pages are rarely delivered to your phone through the shortest route in the Internet. Rather, algorithms are used to manage information flow and prevent information traffic jams.”

—Dennis McCafferty

global brand might need to respond quickly to a trending topic on Twitter. For those sorts of snap decision-related tasks, Hadoop is too slow, and other tools have begun to emerge.

The Hadoop community has been building real-time response capabilities into HBase, a software stack that sits atop the basic Hadoop infrastructure. Cloudera's Lipcon explains that companies will use Hadoop to generate a complicated model of, say, movie preferences based on millions of users, then store the result in HBase. When a user gives a movie a good rating, the website using the tools can factor that small bit of data into the model to offer new, up-to-date recommendations. Later, when the latest data is fed back into Hadoop, these analyses run at a deeper level, analyzing more preferences and producing a more accurate model. "This gives you the sort of best of both worlds—the better results of a complex model and the fast results of an online model," Lipcon explains.

Cloudant, another real-time engine, uses a MapReduce-based framework to query data, but the data itself is stored as documents. As a result, Miller says, Cloudant can track new and incoming information and only process the changes. "We don't require the daily extraction of data from one system into another, analysis in Hadoop, and re-injection back into a running application layer," he says. "That allows us to analyze results in real time." And this, he notes, can be a huge advantage. "Wait-

## GraphLab, a new open source processing framework, uses some of the basic MapReduce principles, but pays more attention to the networked structure.

ing until overnight to process today's data means you've missed the boat."

Miller says Cloudant's document-oriented store approach, as opposed to the column-oriented store adopted in HBase, also makes it easier to run unexpected or ad hoc queries—another hot topic in the evolving Hadoop ecosystem. In 2009, Google publicly described its own ad hoc analysis tool, Dremel, and a project to develop an open source version, Drill, just launched this summer. "In between the real-time processing and batch computation there's this big hole in the open source world, and we're hoping to fill that with Drill," says MapR's Ted Dunning. LinkedIn's "People You May Know" functionality would be an ideal target for Drill, he notes. Currently, the results are on a 24-hour delay.

"They would like to have incremental results right away," Dunning says.

Although these efforts differ in their approaches, they share the same essential goal. Whether it relates to discovering links within pools of DNA, generating better song suggestions, or monitoring trending topics on Twitter, these groups are searching for new ways to extract insights from massive, expanding stores of information. "A lot of people are talking about big data, but most people are just creating it," says Guestrin. "The real value is in the analysis." **C**

### Further Reading

Anglade, T.

NoSQL Tapes, <http://www.nosqltapes.com>.

Dean, J. and Ghemawat, S.

MapReduce: Simplified data processing on large clusters, *Proceedings of the 6th Symposium on Operating Systems Design and Implementation*, San Francisco, 2004.

Ghemawat, S., Gobioff, H., and Leung, S.

The Google file system, *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, Lake George, NY, Oct. 19–22, 2003.

Low, Y., Gonzalez, J., Kyrola, A., Bickson, D.,

Guestrin, C., and Hellerstein, J.M.

GraphLab: A new parallel framework for machine learning, *The 26th Conference on Uncertainty in Artificial Intelligence*, Catalina Island, CA, July 8–11, 2010.

White, T.

*Hadoop: The Definitive Guide*, O'Reilly Media, Sebastopol, CA, 2009.

Gregory Mone is a Boston, MA-based writer and the author of the novel *Dangerous Waters*.

© 2013 ACM 0001-0782/13/01

### Milestones

## Supercomputing Visionaries Honored

ACM and the IEEE Computer Society honored high-performance computing innovators at the recent SC12 conference in Salt Lake City, UT. Among those honorees were the inventor of the first multicore processor, biomolecular modeling researchers, and an expert in managing software security flaws.

University of Notre Dame computer science and engineering professor Peter Kogge received the Seymour Cray Computer Engineering Award. Kogge

developed the space shuttle I/O processor, invented the Kogge-Stone-Adder process for adding numbers in a computer, and helped create the first multicore processor (EXECUBE) at IBM. He recently spearheaded DARPA's initiative to investigate a super-computer capable of a quintillion operations per second.

Klaus Schulten and Laxmikant Kale, professors at the University of Illinois at Urbana-Champaign, received the Sidney Fernbach Award for

their contributions to the development of "widely used parallel software for large biomolecular systems simulation." Schulten directs the Center for Biomolecular Modeling and was the first to demonstrate that parallel computers can be used to solve the "many-body" problem in biomolecular modeling. Kale directs the Parallel Programming Laboratory; his work has focused on enhancing performance and productivity via adaptive run-time systems.

Mary Lou Soffa of the University of Virginia received the ACM-IEEE Computer Society Ken Kennedy Award for her work in detecting and managing software security flaws. Soffa developed software tools for debugging and testing programs to eliminate or reduce false alarms and improve operating efficiency. Her research produced automatic, practical solutions in software engineering, and systems and programming languages for improving software reliability, security and productivity.

# Just the Facts

*In repackaging other companies' news, some news aggregators are diverting readers and ad dollars, and, critics argue, undercutting the incentive to spend money on original reporting. It is an economic and ethical problem without a clear legal fix.*

**L**AST YEAR, SIMON DUMENCO, a columnist for *Advertising Age*, found himself at the center of a minor media maelstrom. A blog post he had written for AdAge.com went viral, getting picked up in one form or another by several other sites, including Techmeme and the hugely popular Huffington Post. By now, the topic of his piece (Apple's launch of iCloud getting upstaged on Twitter by the "Weinergate" scandal) is old news. What remains timely is the window the incident opened on a big question for the struggling news business. Do news aggregators—which include AOL's Huffington Post, Techmeme, Google News, and the many other sites that profit from recycling news stories first reported by others—help or harm the news organizations whose material they use?

## A Broken Market

For many news people, the answer is obvious. The *Wall Street Journal's* managing editor Robert Thomson has called aggregators "tech tapeworms in the intestines of the Internet" and Bill Keller, the executive editor of *The New York Times*, has likened Huffington Post's practices to Somali piracy. Whether portraying aggregators as parasites, pirates, or petty pickpockets, critics make the same basic argument: Aggregators free-ride on the efforts of others, taking readers and ad revenue that would otherwise go to the companies who financed the labor-intensive work of actually breaking the news. "When the Huffington Post reproduces much of the article, even if it gives credit to the *Times* and the reporter, they are undermining the market for the original product," explains Joe Mathewson, a professor at Northwestern University's Medill School of Journalism.



**Arianna Huffington, co-founder, president, and editor-in-chief of the Huffington Post Media Group, a leading content aggregator.**

Business journalist Robert Levine, the author of a book about the free-rider problem in the culture industry, puts it more bluntly. "We have a broken market," he says. "If it's so easy to make money aggregating, investment is going to go to aggregators." It becomes less profitable than ever

**Do news aggregators help or harm the news organizations whose material they use?**

to fund investigative reporting, for example, or to send journalists overseas. Dumenco points out that because aggregation is far cheaper than original reporting, aggregators also drive down rates for advertising, traditionally a major source of newspapers' revenue. And although some newspapers have set up paywalls to earn subscription dollars, aggregators make it more difficult to do so successfully, Levine says, since readers can now get much the same information for free elsewhere. This system seems inherently unsustainable and bound to bankrupt the very content sources on which aggregators rely.

But defenders of aggregation, most notably Arianna Huffington, do not seem to worry. They counter that what they are doing is "curation," giving readers the most interesting stories from a broad range of news produc-

ers; and rather than stealing readers, aggregators claim they actually drive additional traffic to the originating news sites, enlarging the total readership and thus making aggregation a win-win strategy.

Which side of the aggregation debate is right is an empirical question with different answers for different cases, since aggregators are a varied lot. “Techmeme gives a headline and a couple of sentences,” says Dumenco, “and they’re doing aggregation in the purest sense—taking a bare minimum and encouraging readers to click back.” Google News is the best-known example of this species, which typically select which stories to display algorithmically, rather than through editorial decision-making. The Huffington Post, in contrast, presented far more than the opening of an article: its blogger more or less rewrote Dumenco’s entire story.

For Dumenco, that difference created a handy test: Using Google Analytics, he could compare the sites to see which one was driving traffic to his original column. Even though Huffington Post’s overall traffic is vastly greater than Techmeme’s (suggesting that more people probably saw the Huffington Post’s piece), Techmeme beat Huffington Post hands down in helping *Ad Age*, directing 746 page views to Dumenco’s original post, in sharp contrast to a mere 57 from Huffington Post. Dumenco knows his post-hoc analysis with a sample size of one item is far from conclusive, but his story touched many a nerve. After he wrote about the incident in a follow-up column, a Huffington Post editor publicly apologized and suspended the young blogger who’d rewritten Dumenco’s work. But so many journalists and bloggers complained about similar experiences that Dumenco helped form the Council on Ethical Blogging and Aggregation, a group that hopes to set voluntary standards for how to properly draw on other people’s content.

These days, the online arms of old-media companies do a great deal of blogging and aggregation, too, so the Council on Ethical Blogging and Aggregation includes representatives from a broad range of publications, including Huffington Post, *The Atlantic*,

and Slate. If these members can agree on standards for their own staffs, it is uncertain how widely these will be adopted by others, especially given the culture clash between journalists and technologists. Whereas many new-media and tech companies, such as Google, view aggregation as a natural and legitimate part of the open source movement, that philosophy is at odds with journalistic culture, says University of Iowa journalism professor Jane Singer, an authority on the ethics of online journalism. Particularly in the many countries where journalism is a commercial enterprise, the ethos has been “I profit when I have the story, I have the story first, I have the better story, it’s my story, and you have to come to me to get that story,” Singer says. “That’s not open source.”

### Scarce Legal Remedies

Federal copyright law in the U.S., however, does not support the journalists’ proprietary view of the news, extending copyright protection only to the actual words and images used to tell a story, not to the underlying facts. This idea-expression dichotomy is deliberate, meant to safeguard the First Amendment. “We want people to be able to use facts to do new and original things with them, ensuring that news gets out there to people who need it quickly,” explains Andy Sellars, a fellow at Harvard’s Berkman Center for Internet and Society and a staff attorney with the Digital Media Law Project. But the distinction between facts and their expression creates a problem. “In a lot of

**Aggregators free-ride on the efforts of others, taking away readers and ad revenue, according to their critics.**

these aggregation cases, companies aggregating the news are pretty carefully taking just the facts,” says legal scholar Joseph Liu, who studies intellectual property and Internet law at Boston College School of Law. That, says Liu, “leads to a certain lack of protection that can challenge the incentives for gathering the news in the first place.”

Even when aggregators routinely copy whole chunks of original news stories—displaying, for example, the headline and opening lines of a news story—U.S. copyright law is not necessarily on the side of content creators. When threatened with copyright infringement claims, aggregators typically hide behind the “fair use” defense, as Google did when Agence France-Presse (AFP) sued it for aggregating AFP’s stories on Google News without paying a licensing fee. In determining whether a particular case falls under fair use, judges examine four factors—most importantly, the purpose of the use and the effect on the market for the original work—so it is impossible to articulate a general rule about what kind of aggregation is legal, Sellars says. Moreover, most fair-use cases (including between AFP and Google) settle before reaching the bench, adding to uncertainty about how future judges are likely to rule. “The tricky thing about settlements is that it’s hard to draw inferences,” Liu says. “There can be all kinds of reasons to settle a case that don’t reflect what the underlying law might be.”

The only other legal remedy for victims of aggregation, stemming from a famous Supreme Court decision in the age of the telegraph, stands on even shakier ground. During the World War I, the Associated Press (AP) was stationed in Europe while a rival American news service, called INS, was forced for political reasons to return to the U.S. “The AP expended a great deal of energy gathering the news, telegraphing them, and posting them on the East Coast,” explains Sellars. “Because the INS couldn’t go, they would quickly read the stories and telegraph the facts to the West Coast.” As a result, and in a situation remarkably similar to today’s aggregation battles, the INS could beat the

AP to the punch while incurring a fraction of the AP's costs. When the case reached the Supreme Court in 1918, the justices reasoned that though the INS did not break copyright law, the company was violating the common-law doctrine of unfair competition, dubbing their particular offense "hot news misappropriation." Unfortunately for future news organizations, however, a landmark 1938 case held that federal courts have no jurisdiction over common law; "hot news," therefore, became the purview of the states, and today only five states (including New York) recognize the doctrine, which in theory gives a temporary monopoly to news creators. Even in these states, though, the hot-news doctrine may be unconstitutional if it conflicts with federal law, which trumps state law. The legal concern, Liu explains, is "Are the media companies trying to protect something with state law that federal copyright law says you can't protect?"

At least one company sees an opportunity in the midst of the crisis—and, if successful, may help fix the market for news. Newsright, a year-old start-up, aims to serve both journalists and aggregators by making it easy to use news content legally. Newsright's interim CEO Srinandan Kasi, formerly a scientist at IBM and more recently the head of legal affairs for the AP, points out that people are consuming news differently than they used to—all day long, and away from the package in which it was released by the owner. "Given that kind of model, you need an 'always-on' supply chain," he says, hinting at possible plans to offer 24/7 spot sales of individual stories. In such a supply chain, Newsright operates more like a wholesale distributor than a retailer of news, so its customers would be aggregators, rather than the ultimate consumers of news.

More than two dozen news organizations have signed on to offer content through Newsright, but attracting aggregators has been more difficult, despite the offer of analytics that reveal trending stories. Why should aggregators pay for content that they have been able to repurpose without paying? Sellars, of the Berkman Center, says if Newsright

**Defenders of aggregation say what they are doing is "curation," giving readers the most interesting stories from a broad range of news producers.**

offers a quick and painless way to pay, aggregators who are risk averse might prefer that option over the chance of a legal fight. "But it all depends on what the figures are. If the money's not right, they might pursue other options." **C**

#### Further Reading

*Athey, S. and Mobius, M.*

The impact of news aggregators on Internet news consumption: The case of localization. Working paper, 2012.

*Bavitz, C. et al.*

Saving journalism from itself? Hot news, copyright fair use and news aggregation, <http://cyber.law.harvard.edu/interactive/events/2010/04/omlnpanel1>, April 9, 2010.

*Isbell, K. and the Citizen Media Law Project*

*The Rise of the News Aggregator: Legal Implications and Best Practices*. Berkman Center Research Publication No. 2010-10, Aug. 30, 2010.

*Levine, R.*

*Free Ride: How Digital Parasites Are Destroying the Culture Business, and How the Culture Business Can Fight Back*. Doubleday, New York, NY, 2011.

*McDonnell, J.C.*

The Continuing Viability of the Hot News Misappropriation Doctrine in the Age of Internet News Aggregation. *Northwestern Journal of Technology and Intellectual Property* 10, 3, Winter 2012.

Based in San Francisco, **Marina Krakovsky** is the co-author of *Secrets of the Moneylab: How Behavioral Economics Can Improve Your Business*.

© 2013 ACM 0001-0782/13/01

#### Supercomputing

## Exascale Enthusiasm Escalates

High-performance computers still have not crossed the exaflop-per-second barrier, but recent supercomputing conferences such as SC12 and the International Supercomputing Conference (ISC12) confirm the drive toward that goal. (An exaflop/s is a quintillion floating-point operations per second.)

As SC12 Technical Program Chair Rajeev Thakur pointed out in an article in *The Exascale Report*, the terms "exascale" or "extreme scale" occur in the titles or abstracts of 75 conference items, including the keynote address by theoretical physicist Michio Kaku.

Thakur also noted that explicit mention of exascale computing first appeared (with nine occurrences) in conference materials in 2008, which was also the year that the first supercomputer passed the one-petaflop/s mark.

Thakur noted that, as those dates approach, "People are more consciously thinking of it. When exascale first appeared in the conference program in 2008, it was only in a few areas. But now it's on everyone's mind."

However, Thakur demurred to name a date himself, saying that "a lot depends on whether there is sufficient investment in research for exascale and in the acquisition of large systems. Nobody really knows when exascale will actually appear."

When exascale first appeared in the conference program in 2008, it was only in a few areas. Now it's on everyone's mind.

—Tom Geller



DOI:10.1145/2398356.2398366

Michael A. Cusumano

## Technology Strategy and Management

# The Apple-Samsung Lawsuits

*In search of a middle ground in the intellectual property wars.*

**I**MAGINE THE FOLLOWING scenario: The Cadillac Automobile Company, founded in 1902, is granted a set of patents in the U.S. and then worldwide that define the design and basic functionality of what becomes the modern automobile: four wheels, a gasoline-powered internal combustion engine, a closed steel body, and a round steering wheel. Then Henry Ford comes out with the much less expensive Model T in 1908 and quickly gains 80% of the market. After a brief lawsuit, a jury finds that Ford has violated Cadillac's patents and must cease production as well as pay extensive damages to Cadillac, now part of the General Motors Corporation.

Two additional scenarios might have unfolded. One could be that patent protection for Cadillac stimulates further innovation. Competitors develop automobiles with batteries as well as fuel cells powering electric motors, high-tech materials for convertible bodies, and joystick controls that resemble those of modern-day video game consoles. The intense competition spurs on advances in manufacturing methods as well, leading to

**As many as 250,000 patents are filed that cover the design and functionality of the iPhone and other smartphones.**

lower prices. Lots of companies make money, and consumers have choices of varying quality and prices. Another scenario, however, could be less cheerful. Maybe no company comes up with a better design. Cadillac dominates the market and becomes the most valuable company in history. Its investors are delighted and so are its affluent customers. As for everyone else, they make do with inferior three-wheel vehicles or continue to drive horses and buggies.

In some ways, the August 2012 California jury verdict against Samsung

(the world's largest maker of smartphones) in the case brought by Apple (the world's second largest maker) *could evolve* to resemble this Cadillac fantasy. I say "could evolve" because we are not there yet but patent offices, judges, and jurors around the world must exercise caution and common sense. Apple is continuing to challenge Samsung on newer products and additional patents not covered in this particular litigation. It will also probably go after Google, maker of the Android operating system used in the Samsung products, as well as other phone makers that did what Samsung did—introduce products patterned after the Apple iPhone and iPad, running Google Android software. Google gives away its software and makes money from mobile ads; it will be difficult to calculate damages compared to companies actually selling products. But the iPhone has already lost a lot of market share to Android phones, and someone will place a value on these lost sales. At the time of the case, the Apple iPhone had merely 19% of the global smartphone market, compared to 64% for Android phones.<sup>10</sup>



### Case Study

The basic facts of the case seem straightforward. The jury awarded Apple \$1.05 billion in damages (which the judge might triple because Samsung was found to have willingly copied Apple). The lawsuit covered some two dozen older devices mostly sold outside the U.S., not the current models Samsung is pushing.<sup>4</sup> Apple will have to go after the newer models and other potential patent violations in separate litigation, which it is doing. In fact, as many as 250,000 patents are filed that cover the design and functionality of the iPhone and other smartphones, and there are already dozens of lawsuits and countersuits between Apple and Samsung in 10 countries.<sup>6</sup>

Samsung has already fared better overseas. A Japanese court found

in favor of Samsung, saying it did not violate an Apple patent on technology that synchronizes music and videos between devices and servers. Though Apple had sought only \$1.3 million in damages, this is a victory for Samsung. A South Korean court also rendered a mixed decision in another Apple versus Samsung patent case.<sup>8</sup>

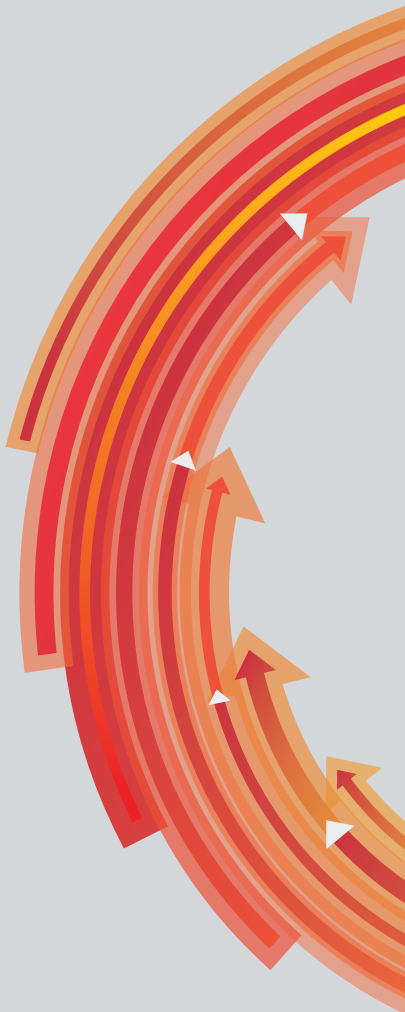
In the August 2012 case, the U.S. jurors found that all seven of Apple's patents were valid and that Samsung violated six of those that help define the look and feel of the iPhone. Four were "design patents" related to the appearance of the iPhone (the use of white and black on the devices and the rounded edges on the user-interface icons, which Samsung was found to have violated, and the tablet computer's rectangular design, which Samsung did not

violate). Three "utility patents" were more technical, involving both hardware and software controlling how the device enlarges documents when the user taps the screen, distinguishes single-touch versus multitouch gestures, and appears to bounce back when scrolling to the end of a page.<sup>9</sup>

In the wake of the Samsung verdict in the U.S., the largest smartphone and tablet market in the world, we are likely to have more innovation as well as competition. Microsoft and its partner Nokia may benefit, for example, because they have not copied Apple's iPhone designs and iOS software so explicitly. The Windows 8 phones from Nokia have received good technical reviews, though they seem less intuitive to use, there are few compelling applications, and

## Computing Reviews is on the move

Our new URL is  
**ComputingReviews.com**



**COMPUTING REVIEWS**

A daily snapshot of what is new  
and hot in computing.

customers continue to prefer Android phones that look and feel like the Apple iPhone. Not surprisingly, Windows phones as of this writing had a mere 3% of the market.<sup>1</sup>

If more court decisions come down in favor of Apple's design patents, it is likely the iPhone will remain distinctive. But it might also remain a minority platform at the higher end of the market, like the Macintosh in past years or the once-revered Cadillac. Maybe this is the best scenario for a company that wishes to maximize its profits. But is this the best scenario for consumers? Maybe not.

### Intellectual Property Protection

The broader issue, of course, involves the pluses and minuses of intellectual property protection—a hotly debated topic for decades. On one side are those who have argued—successfully—that companies will not have sufficient incentives to invest in research and product development if they are not allowed to capture the value generated by their investments for some significant period of time. On the other side are those who argue that intellectual property protection actually hurts innovation at a broader social level. Patents, because they grant temporary monopolies, might restrain the ability of other companies or individuals from making improvements or disseminating the technology to a broader user base.<sup>2</sup> Other research has found that patents issued in a particular area (for example, around a specific protein molecule in biotech research) can have a “chilling” effect on further research in this area. Patents in this scenario can actually create *disincentives* for further research and development.<sup>7</sup> In general, though, most governments have agreed with the former argument and that is why they

have created patent systems, dating back at least to the 15<sup>th</sup> century in Europe. The U.S. patent system predates even the Constitution and currently provides protection for 20 years from the date of filing.

It is also true that platform dynamics with skillful marketing can lead some companies to dominate markets without relying on patents (see my column “The Evolution of Platform Thinking,” *Communications*, Jan. 2010). Microsoft achieved a 90%-plus market share with DOS and then Windows mainly through good fortune and timing (the deal in 1980–1981 to provide the operating system for IBM's new PC) as well as deliberate efforts to cultivate hardware and software partners, and long-term contracts with PC manufacturers. It also used specific pricing tactics and volume discounts (some of which were found to be illegal in the 1990s U.S. antitrust action).<sup>3</sup> Most software companies have not viewed patents as useful to protect their products since there are often many ways to implement a particular function and work around a patented algorithm. But times have changed, and now many hardware and software companies are acquiring stocks of patents and suing each other with increasing frequency.<sup>6</sup>

Particularly worrisome are design patents, which appear broader and fuzzier than technical utility patents and might create undesirable consequences for users and competitors. The courts need to be careful as well as consistent when issuing such patents to make sure they do not overly inhibit innovations that are impor-

**If more court decisions come down in favor of Apple's design patents, it is likely the iPhone will remain distinctive.**

a For arguments in favor of intellectual property protection, see the classic articles by K. Arrow, “Economic Welfare and the Allocation of Resources for Invention,” in National Bureau of Economic Research, *The Rate and Direction of Inventive Activity: Economic and Social Factors* (Princeton University Press, 1962), 609–626; and R. Nelson, “The Simple Economics of Basic Scientific Research.” *Journal of Political Economy* 67, 3 (1959), 297–306. For a recent popular argument on the other side, see E. Von Hippel, *Democratizing Innovation* (MIT Press, 2005).



tant to disseminate as broadly as possible. Recall that Apple once sued Microsoft in 1988 for copying the “look and feel” of the Lisa and Macintosh graphical user interface (GUI). Apple had engaged Microsoft in the early 1980s to create versions of Word and Excel for the Macintosh, which came out in 1984. From this experience, Microsoft learned how to design graphical software—similar to how Samsung learned about the details of the iPhone and iPad by being the largest supplier of the microprocessors to Apple.<sup>5</sup> Microsoft had licensed some GUI design elements for Windows 1.0, a layer it built to sit on top of DOS. But Microsoft continued to use Apple’s design elements in later releases of Windows, which Apple challenged. Apple lost this case at least in part because it had licensed and copied aspects of the GUI from Xerox, which had done the pioneering work at its research lab, Xerox PARC. In its defense, Microsoft also argued that a company should not be able to protect something as vague as a “look and feel.”<sup>b</sup>

So while Apple was the first to commercialize the graphical user interface, the court did not grant Apple a monopoly on the general design (though it made Microsoft change its trash can icon because this copied the Macintosh garbage can too explicitly). Microsoft eventually made the GUI ubiquitous by broadly licensing Windows to many PC manufacturers, which brought down prices. The Macintosh remained expensive (the “Cadillac” of PCs?) and became a niche product, where it remains despite a recent revival in sales. The iPhone and the iPad have turned Apple into the world’s most valuable company, but Microsoft still generates remarkable profit levels (see “Reflecting on the Facebook IPO,” *Communications*, Oct. 2012). More importantly, thanks mainly to Microsoft and Windows PC manufacturers, the Macintosh-style graphical user interface became the dominant way to use a personal computer, elevating billions of people be-

b *Apple Computer, Inc. v. Microsoft Corporation*, 35 F.3d 1435 (9th Cir. 1994); <http://bulk.resource.org/courts.gov/c/F3/35/35.F3d.1435.93-16883.93-16869.93-16867.html>. Also see S. Manes and P. Andrews, *Gates* (Doubleday, 1993), especially pp. 357–364 and 437–438.

## Many hardware and software companies are acquiring stocks of patents and suing each other with increasing frequency.

yond character-based DOS computing (which reminds me of the horse and buggy).

How valuable Apple remains in the future has a lot to do with how effectively it can prevent other firms from copying its innovations. At the same time, whether billions of consumers will be able to buy iPhone-like smartphones and iPad-like tablets that are rectangular and have touch screens and other functions that work in similar ways has a lot to do with how patent offices, juries, and judges act in the future.

These cases are complex because there are valid arguments on the different sides. From the innovators’ point of view, strong intellectual property protection is desirable to stimulate and protect their investments. Apple may not have developed the Macintosh in 1984 or the iPod, iPhone, or iPad products in the 2000s if Steve Jobs did not believe he could prevent others from copying the designs, at least to some extent. From the competitors’ point of view, strong patent protection is not desirable when they want to “borrow” or “build on” good ideas or follow the “dominant design” established in the marketplace.<sup>c</sup> From the point of view of consumers and society at large, we all lose if companies do not have sufficient incentives to invest in research and development. We also lose when patents prevent advances in a particular technology or make it difficult or expensive for the majority of consum-

c For a discussion of the “dominant design” concept, see J. Utterback, *Mastering the Dynamics of Innovation* (Harvard Business School, 1994).

ers to adopt the most useful, usable, and elegant designs.

### Conclusion

Apple needs to be fully rewarded for its innovations. But if the company’s patents and lawsuits prevent other firms from creating elegant, simple-to-use products patterned after the iPhone and the iPad but perhaps better, faster, and cheaper, then it will be a sorry ending to the current battle between Apple and Samsung. Perhaps there could be a middle-ground solution where, for example, Samsung, as well as Google and other companies, reach agreements with Apple to make royalty payments and then cross-license some of their own patents. In fact, a recent court did this with a lawsuit filed between Apple and Motorola, and Apple and Google are already engaged in patent discussions.<sup>2,9</sup> With a negotiated outcome, companies would be able to take better advantage of the innovations they and their competitors produce, while giving proper credit to the innovators and allowing them a fair return on investment. The definition of “fair” will be another matter of negotiation and litigation. Nonetheless, even small royalties on every Android device sold could quickly produce a financial windfall for Apple that exceeds current iPhone and iPad sales. ■

### References

1. Ante, S. and Trojanovski, A. Microsoft’s mobile moment: Will consumers buy in? *The Wall Street Journal* (Aug. 29, 2012), B1.
2. Bloomberg News. Apple, Google discuss patent issues. *The Boston Globe* (Aug. 31, 2012), B6.
3. Cusumano, M. and Selby, R. *Microsoft Secrets*. Free Press, New York, 1995, 164–166, 436–437.
4. Elias, P. Apple asks judge to ban U.S. sales of 8 Samsung phones. *The Boston Globe* (Aug. 28, 2012), B6.
5. Koetsier, J. Apple stuck in bed with Samsung as exclusive CPU deal with Taiwan semi fails. *Venturebeat.com*, (Aug. 29, 2012); <http://venturebeat.com/2012/08/29/apple-samsung-taiwan-semi-cpu-iphone-ipad/>.
6. Lohr, S. A patent war in your pocket. *The New York Times* (Aug. 26, 2012), A4.
7. Murray, F. and Stern, S. Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization* 63, 4 (2007).
8. Tabuchi, H. and Wingfield, N. Tokyo court hands win to Samsung over Apple. *The New York Times* (Sept. 1, 2012).
9. Vascellaro, J. Apple wins big in patent case. *The Wall Street Journal Online* (Aug. 25, 2012).
10. Wingfield, N. In Apple’s patent case, tech shifts may follow. *The Wall Street Journal* (Aug. 20, 2012), B1.

**Michael A. Cusumano** (cusumano@mit.edu) is a professor at the MIT Sloan School of Management and School of Engineering and author of *Staying Power: Six Enduring Principles for Managing Strategy and Innovation in an Unpredictable World* (Oxford University Press, 2010).

Copyright held by author.



## The Business of Software

# How We Build Things

*...and why things are 90% complete.*

**I**T SEEMS TO be a law of software development that things always take longer than we expect. So we expect them to take longer and make appropriate allowances and they still seem to take longer. Even if we carefully account for unrealistic expectations as Tom DeMarco so testily observed in *Why Does Software Cost So Much?*<sup>2</sup> we are regularly and unpleasantly surprised when we overrun still more. And often the person doing the work is astonished that: it takes longer than it was thought, planned, and committed to; even in the middle of the work, while acknowledging it has taken longer than expected to this point, it will still take longer to finish than will be predicted right now; and we continue to repeat this behavior.

This scares project managers. Due to the rather opaque work done in the business of software it is difficult for a project manager to know what progress has been made without asking those doing the work how they are doing. The code is written, but is it the right code? Is there enough of it? The test was run, but was it the right test? Was the result what we expected? If it was not what are the consequences in terms of progress? If we unearthed a defect does that put us further ahead (one less defect to deal with) or further behind (yet another thing gone wrong)?

### Zeno's Paradox of Progress

When a project manager talks to a designer, programmer, or tester and tries to get a sense of how "complete" the assigned task is, the normal reply is "about 90%." Asking the same person the same question the following week often elicits

the same reply. Sometimes for weeks, with equal amounts of rising embarrassment and certainty that this time the estimate is good, the programmer or tester will affirm that he or she is really, truly, 90% complete. If the embarrassment gets too high, the engineer might employ the "Zeno's Paradox"<sup>a</sup> approach of halving the distance to the goal each time. So 90% becomes 95%, then 97.5% then 98.75%. We may laugh at this, but perhaps it contains a germ of truth.

A programmer describing progress as "90% complete" is not usually intentionally lying, but clearly the implied 10% remaining time to complete is not correct. So why do people say this and why do project managers continue to operate on this assumption? There are several reasons.

### Second Order Ignorance (2OI)

Systems development is primarily a knowledge discovery activity. Even in coding, the thing that takes the most time is figuring out what we don't know. If systems development were a matter of transferring what we already knew into code, we would build systems as fast as we can type. And for some aspects of development, such as testing, discovering what we don't know represents almost all the effort.

We may be quite aware of some of the things we don't know—this is lack

of knowledge or First Order Ignorance (1OI).<sup>1</sup> But sometimes there are things about the system we don't know we don't know, which means we have Second Order Ignorance (2OI). Building what is already known is easy, resolving clearly defined questions is a bit more challenging, but discovering what we are not aware of is by far the most difficult and time consuming task. So asking the question "...how complete are you?" or its corollary "...how much more work is there left to do?" is the equivalent of asking "...how much do you not yet know?" There is no easy way for someone to answer that question. Most people extrapolate from their past experience and what they have learned on the task so far to figure how much they might have left to learn. But they don't actually *know*. Then a certain amount of optimism, a modicum of professional pride, and perhaps a dash of hubris and the calculation yields the ubiquitous "90%." I have observed that years of experience have a way of tempering one's optimism and experienced engineers do tend to be more realistic about what they don't know yet. But there is another reason why we overestimate progress.

**Linear Equation.** Project managers get confused because they associate percentage complete with percentage of resources used. They assume a linear relationship between these two variables (see Figure 1). So when the quantity produced = 90%, the resources used must be = 90%. But in development this is not so.

**Rayleigh Curve.** Research has determined that the rate at which we build

<sup>a</sup> The Greek philosopher Zeno of Elea was credited with identifying an apparent paradox in movement where, in order to travel any distance, we must first move half the distance, and then half the remaining distance, and then half that remaining distance. The sum of  $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots$  never reaches 1 so therefore we should never be able to move anywhere.

Figure 1. Linear completion.

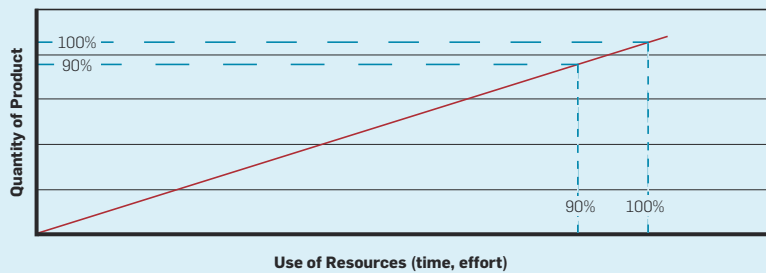


Figure 2. Rayleigh Curve.

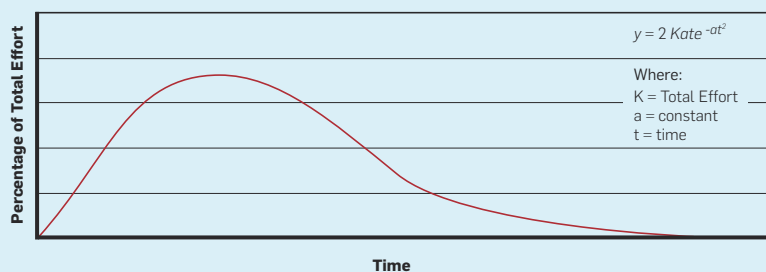
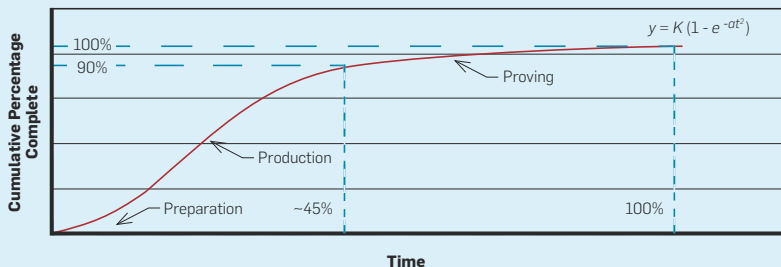


Figure 3. Cumulative Completion.



things in systems development can be described by a Rayleigh Curve<sup>3,4</sup> (see Figure 2). This curve shows the amount of work that can be done varies with time. It rises from zero to a peak, perhaps a third of the way into the activity, and then drops off back to zero. The cumulative version of this curve is shown in Figure 3.

**Non-Linear.** If we compare Figure 3 with Figure 1 we notice the cumulative curve relationship of percentage of product complete and percentage of resource used (time in this case) is not a straight line. Some percentage on one

axis does not mean the same percentage on the other. In fact, in Figure 3, when the product is 90% complete, the activity is only about halfway through its total time. This is partly where the miscalculation of progress comes from. If developers or project managers, consciously or unconsciously, apply a mental model that is linear to something that behaves non-linearly it is not surprising that they incorrectly assess progress.

### Three Stages

This model also makes intuitive sense. In Figure 3, the curve has three dis-

tinct stages and they apply as much to painting a room as they might to building software.

**Preparation.** The environment must be prepared before work can commence. For software this means setting up servers and version control systems, acquiring tools, planning and estimating, sourcing staff and other essential activities that take time but might not actually create much in the way of software product. When painting a room, we have to choose the color scheme, measure, tape up the woodwork, and buy the paint, all before we can start putting it on the wall.

**Production** is the steepest slope where the maximum rate of measurable work occurs—the most code is written, the most paint is applied.

**Proving** is the final long tail of the process. In software it is completing the exception code (the mainline code was done in production), testing it to prove it works and fixing it when it does not. This always seems to take longer than it should partly because we invariably find out things we were not expecting—which is where the Second Order Ignorance comes in. In painting, this is the detail work, the tricky corners and, of course, the cleanup—which also always seems to take longer than it should.

### Mathematical Solution

There are many reasons why people overreport progress: optimism, pride, pressure to show results, inexperience, poor resource allocation, even how they are compensated, and they all play a part in overstating development and underestimating the remaining work.

But realizing that we cannot calculate exactly how far we have left to go because we do not know exactly where we are going and simply using the correct mathematical model of progress would certainly help. □

### References

1. Armour, P.G. The Five Orders of Ignorance. *Commun. ACM* 43, 10 (Oct. 2000), 17.
2. DeMarco, T. *Why Does Software Cost So Much?* Dorset House, 1995, 4.
3. Norden, P.V. On the anatomy of development projects. *IRE Trans. Eng. Management* 7, 1 (1960), 40.
4. Putnam, L.H., and Myers, W. *Measures for Excellence*. Yourdon Press, 1992, 46.

**Phillip G. Armour** (armour@corvusintl.com) is a senior consultant at Corvus International Inc., Deer Park, IL, and a principal consultant at QSM Inc., McLean, VA.

Copyright held by author.

## Law and Technology

# Beyond Location: Data Security in the 21<sup>st</sup> Century

*Viewing evolving data security issues as engineering problems to be solved.*

**T**HE CONTINUED ATTENTION to data protection and the growth of cloud computing highlight tensions among data protection regulators, businesses, and the computer science communities. As new data protection laws are proposed, these groups have the chance to share insights and achieve their respective goals; but right now, with respect to data security, they may be passing by each other.

Data protection laws seek to protect user rights and rely in part on a certain view of data location and related security practices to ensure those rights are maintained. In simplified terms, data protection laws tend to focus on data not leaving a country or region as part of a given data protection regime. Businesses apply cloud techniques for a range of purposes. Some ends are internal such as improved network operations; some are external such as selling storage and services to other businesses. In either case, advances in, and the future of, cloud computing rely on moving data on an almost continuous basis. Thus, the political and business interests seem to be set to collide. That collision is not, however, inevitable. Does data protection require keeping data in one place? Is data security enhanced or harmed by such an approach? Does jurisdiction have to turn on data location? By parsing what is at stake for location and jurisdiction and what cloud computing may offer for security, we should be able to fashion laws that respect the political interests

in data protection and that draw on the best insights of computer science to achieve heightened data security.

### Possibly Competing Interests

Governments and businesses have legitimate, competing interests in data management. For example, they disagree about what to do with cloud computing. Sometimes the debates devolve into accusations that the other side “doesn’t get it.” Yet, if we start by stating what those interests are, we should be able to see where the interests intersect or diverge. Once that is done, we can see whether there is a way to bridge remaining gaps.

Although there are many different data protection laws, the European Union’s approach provides a way to understand government interests and possible mistakes on the horizon. Unfortunately, mandated data location serves two, conflicting purposes. On the one hand, it allows for an exercise of jurisdiction based on the idea that data stored in a particular jurisdiction is subject to the laws of that place. The need for jurisdiction is real. Governments want to be able to reach out and touch our data. They also want to enforce laws to protect their citizens and their data. On the other hand, the EU’s previous Data Protection Directive (DPD) and current, proposed General Data Protection Regulation (GDPR) seek to prevent unauthorized access to and, by extension, use of data. For example, Article 30 of the GDPR requires that those responsible for data

processing take “appropriate technical and organizational measures to ensure a level of security appropriate to the risks represented by the processing and the nature of the personal data to be protected.” It also requires those responsible for data “protect personal data against accidental or unlawful destruction or accidental loss and to prevent any unlawful forms of processing, in particular any unauthorized disclosure, dissemination or access, or alteration of personal data.”

The location problem arises because the current and proposed approaches employ complicated rules about data location, storage, and movement to achieve the protection goals. In addition, the laudable goals of Article 30 inadvertently run into the realities of the latest security advances in cloud computing. For example, in a recent decision in the EU, data location requirements interfered with a city’s ability to use modern cloud computing services.<sup>a</sup>

What then is cloud computing and how does it relate to data security? First, one can think of the security problem as the ways in which someone could gain unauthorized access to data. That view comports with Article 30. A perhaps somewhat misunder-

<sup>a</sup> See Notification of decision—New email solution within Narvik local authority (Narvik kommune)—Google Apps, Norwegian Data Inspectorate, reference number 11/00593-7/SEV, January 16, 2012 (denying a request to use Google Apps for email and other service based in part on location of data and methods for data storage concerns).



stood issue is from where the threat of unauthorized access comes. Given some recent stories about security breaches, some may think the largest threat for unauthorized access is at data centers which, like banks, might be targeted and then breached. Yet, losing computers and thumb drives is a major way that data is lost, perhaps more so than through a data center.<sup>b</sup> It is the fact of walking data—that is, data on a portable device—that leads to cloud computing as a step forward in security practices. Cloud services address many data security issues, especially by reducing threats from loss of a personal device. But not all cloud services are the same. A big problem for any data protection law is that the type of cloud service and the way data is managed varies greatly depending on how the provider manages the backend *and* what the customer is doing.

Some cloud computing is distributed computing. The data may be sharded across many servers; copies are made of documents and then split across servers. Instead of a single point of failure where all your data happens to be on the server, your data is spread out on many servers. That model can

apply to in-house data management or when providing services to others. In simplest terms the provider may manage all service in-house or use other parties as well, but pinning down exactly what mix of data management is in play requires knowing the configuration for a given service.

New work on optimization means that the best cloud services will likely *move data often, if not all the time*, to address issues such as overheating that threatens servers, bandwidth, packet loss, power, resources, and other “failure modes” as well as to be more efficient with their networks as cycles of computing fluctuate. Thus moving data is part of securing the data against loss, which is another goal of Article 30. In addition, customers may use a cloud

**Governments want to be able to reach out and touch our data. They also want to enforce laws to protect their citizens and their data.**

service on an ongoing or a temporary basis, but others run the service. For example, with Amazon’s E2C offering, the cloud is a commodity service, which allows a company to buy services for a short period to accomplish a large task. When the *New York Times* converted 11 million articles to .pdf, it used a rented cloud for a day to finish the work and at a much lower cost than having its own data center for the task. All of these variables challenge any attempt to mandate security protocols based on location, because companies, customers, and even governments will not know ahead of time what option they want to pursue or they will be forced to choose older, slower, and costlier approaches to data for the sake of compliance.

Global cloud services present additional problems. With multiple jurisdictions involved, anyone offering or using cloud services could face location requirements for each country in which they operate. To the EU’s credit, it is trying to address that criticism as it applies to the current DPD. The proposed GDPR will be binding on all members. The current state-by-state approach under the DPD would go away in favor of a harmonized approach to data with the regulation being implemented in its entirety and taking effect even without a member state taking action to put the regulation into national law.

<sup>b</sup> See, for example, 2010 Annual Study: Global Cost of a Data Breach, May 2011; [http://www.symantec.com/content/en/us/about/media/pdfs/symantec\\_cost\\_of\\_data\\_breach\\_global\\_2010.pdf](http://www.symantec.com/content/en/us/about/media/pdfs/symantec_cost_of_data_breach_global_2010.pdf).

Even if the EU harmonizes its data laws, the focus, however, is still on data as residing in one place or within a region. The EU is big enough that one might think location problem is not an issue. One could set up data centers across the EU and move the data within the system. With one law to govern, it will all work out. That view misses the fact that, like a power grid, data networks have cycles of demand and manage that demand dynamically. If there is idle capacity in a region, it may be useful for service outside the region. If there is a demand spike in the region, capacity from outside the region may be used to meet the demand. Location-based data rules clash with these realities. Furthermore, many countries are copying the EU approach to data protection. The EU is a large region and market; Singapore is not. Nor is Vietnam, Costa Rica, Egypt, Peru, Ghana, or most single countries. From a security perspective, location-based rules falter in a large area such as the EU; they fail in smaller markets. Nonetheless, governments have a real need to protect their citizens' data and to have a legal process to obtain data. Location models, however, do not achieve these goals well.

### Possible Ways Forward

Neither mandated location compliance by governments nor claims of inability to comply with laws by businesses provides satisfactory results; but some options are available. Part of a possible solution is to abandon the location-based aspects of data security laws. Indeed, forcing companies to limit data movement ignores the reality of data management and can increase the security threat rather than reduce it. Unraveling the jurisdiction question is more difficult. As Jack Goldsmith and Tim Wu point out, governments will always find a way to exert power to shape the world as they see fit, so places and borders still matter.<sup>1</sup> This point is already seen in current laws in the U.S. and Europe, which allow governments to access data when investigating national security or terrorism crimes. Thus, countries or regions may fashion new data laws that move beyond location to determine jurisdiction and demand data for areas beyond national security and terrorism. If so, companies will have to find

## From a security perspective, location-based rules falter in a large area such as the EU; they fail in smaller markets.

ways to comply. The time when a company could stick its data-head in the sand of one country and reject other countries' laws may be over precisely because of government needs, global computing services, and advances in data security and networking. If companies wish to have the flexibility to employ different data management methods and especially ones that involve continual movement of data, they cannot simultaneously argue that no law or method covers how and when a government may gain access to data. Yet, it is this need to comply that may undermine the trust of one country over another. For example, Country X may be comfortable with data moving to Country Y, but not Country Z, because Country Z has a history of forcing companies to divulge data. All of which presents an opportunity.

As data security laws evolve, governments, companies, and computer scientists will have to work together to create a data security system for the 21<sup>st</sup> century. A key hurdle is identifying when any government may demand data. Transparent policies and possibly treaties could help better identify and govern under what circumstances a country may demand data from another. Countries might work with local industry to create data security and data breach laws with real teeth as a way to signal that poor data security has consequences. Countries should also provide more room for companies to challenge requests and reveal them so the global market has a better sense of what is being sought, which countries respect data protection laws, and which do not. Such changes would allow companies to compete based not only on their se-

curity systems but their willingness to defend customer interests. In return companies and computer scientists will likely have to design systems with an eye toward the ability to respond to government requests when those requests are proper. Such solutions may involve ways to tag data as coming from a citizen of a particular country. Here, issues of privacy and freedom arise, because the more one can tag and trace data, the more one can use it for surveillance. This possibility shows why increased transparency is needed, for at the very least it would allow citizens to object to pacts between governments and companies that tread on individual rights. Nonetheless, distributed computing techniques and encryption even with some type of tracing system may be better protection than relying on data residing in one place. After all, if data is stored in one place and a government knows where the data is, a government can simply walk off with the server. So far, however, the one group that could explain whether these ideas or others are viable—computer scientists—is missing from this interplay.

The nuances and possibilities of how we choose to manage data present computer science with the chance to inform policymakers about how best to meet their interests while simultaneously meeting the needs of business and individuals. Providing a better understanding of how cloud computing, in its range of manifestations, operates can only improve how the government shapes data policy. Governments should also explain their needs to computer scientists. By presenting issues as engineering problems to be solved, governments would likely stimulate the desire to meet a challenge. Together, governments, businesses, and computer scientists ought to be able to leverage advances in technology so all may benefit. Removing the focus on data location as way to increase data security is hopefully just one step in that direction. ■

### Reference

1. Goldsmith, J. and Wu, T. *Who Controls the Internet? Illusions of a Borderless World*. Oxford University Press, 2006, 180–181.

**Deven Desai** (devenrdesai@gmail.com) is a law professor at the Thomas Jefferson School of Law in San Diego, CA, and recently completed serving as the first Academic Research Counsel at Google, Inc.

Copyright held by author.

## Historical Reflections

# Five Lessons from Really Good History

*Lessons learned from four award-winning books on the history of information technology.*

**M**Y LAST COLUMN (September 2012) explored the lessons to be found in “bad history” of the invention of email. One of the things this reminded me of is just how little many people whose work focuses on information technology know about its evolution over the past 50 years. In this column, I look at some of the very best historical writing about computing from the past few years. I highlight one big lesson from each of four books, giving four ways in which learning more about history can change the way you think about computing. A bonus lesson sums up what the books tell us about the field as a whole.

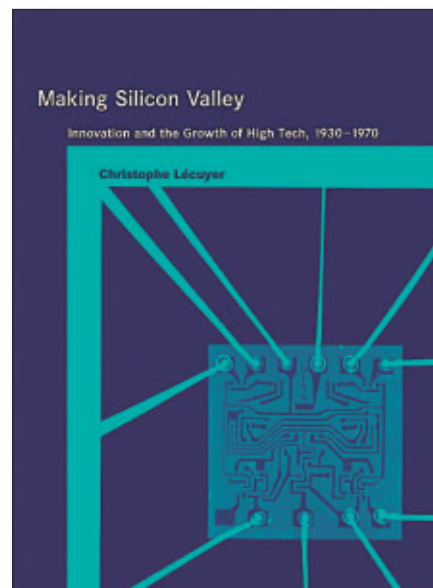
You do not just have to take my word on the “very best” part of the preceding description. The four books are the first winners of Computer History Museum Prize, given each year to the author of an outstanding book on the history of information technology. The prize was created in 2008 by SIGCIS, the organization for historians of computing, with money pledged by networking pioneer and inventor of packet switching Paul Baran. Like many other pioneers, Baran felt a keen interest in preserving and documenting the heritage of his field. He was a keen supporter of the Charles Babbage Institute, the leading academic and archival center for the history of computing, and a fellow and advisory board member of the Computer His-

tory Museum, in whose honor he suggested the name of the prize. When Baran died last year he left instructions for a gift to SIGCIS of \$25,000 to endow the prize in perpetuity.

One of the rewarding aspects of working on the history of computing is that even cutting-edge research is potentially accessible to a broad audience. Historians make an effort to write clearly, at least compared to a typical technical paper in computer science, and we generally do our best to avoid technical jargon. I highly recommend you include one or more of these books on your reading list.

### 1. Making Stuff Creatively Made Silicon Valley Creative

In *Making Silicon Valley: Innovation and the Growth of High Tech, 1930–1970* (MIT 2006, CHM Prize winner 2009) Christophe Lecuyer tackles one of the most familiar stories in the history of computing: the invention of the transistor at Bell Labs, through William Shockley’s creation of a company in California to exploit his invention and the founding of Fairchild Semiconductor by refugees from his erratic management style to the founding of Intel by some of the same people. This is the creation myth for computing in Silicon Valley and has been told and retold by journalists and biographers over the years. It explains how, over the course of a single working lifetime, transistors went from sizable handmade blobs sold at prices only



the military could afford to microscopic metal smears so cheap that we package millions of them into singing greeting cards and other disposable fripperies.

Scholarly and journalistic histories generally rely on different kinds of evidence. Journalists tend to shun endnotes and conduct their research largely by interviewing people. Scholarly historians place great emphasis on finding original written documents from the time in question. If handed a new book close to his or her own research area a historian will often go first to the endnotes, thumbing the back pages to evaluate the range and appropriateness of archival sources used before looking at the main text.

Lecuyer's book is a great example of the depth of insight and detail this approach can provide. He broadens the story out to encompass less widely celebrated firms, such as Litton Industries, National Semiconductor, and Varian Associates, and pushes earlier in time to document the importance of radio component manufacturing to the Valley. The book is based on careful research in the preserved archival records of the people and companies concerned, rather than the recycled anecdotes often used by journalists. While giving due credit to the importance of military sponsorship and Stanford University he puts the development of a pool of skilled labor and amateur electronics enthusiasts at the heart of the Valley's success.

More than anything else, Lecuyer's careful accumulation of detail shows that the early success of Silicon Valley was based on innovation in manufacturing techniques and processes, so that the design and production of its products was closely coupled. This makes one wonder how well the physical and organizational separation of the two now practiced by Apple and other modern firms will sustain long-term innovation.

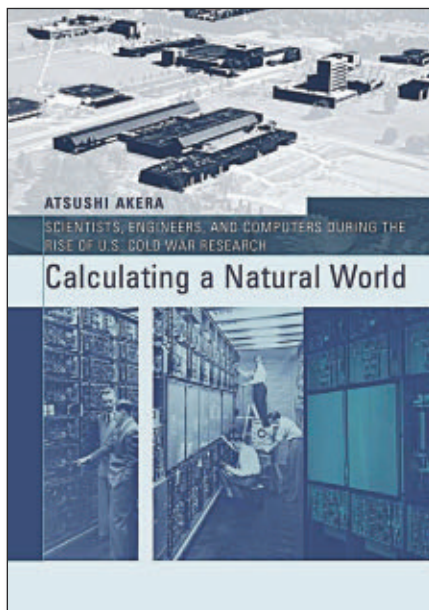
## 2. Computing Was Built at the Intersection of Many Other Fields

Perhaps the most important choice facing a historian is the question of what it is that the book they are writing is really about. History is a kind of storytelling, and stories have protagonists. These protagonists might be specific individuals, as in biography, but they might also be technologies, ideas, companies, occupations, groups of people, countries, or even the entire world. There is also the question of when to start the story and when to stop, as it is rarely possible to cover the entire lifespan of the protagonist.

The topic of Atushi Akera's book *Calculating A Natural World: Scientists, Engineers, and Computers During the Rise of U.S. Cold War Research* (MIT 2007, CHM Prize winner 2010) is difficult to sum up in a single sentence, which is deliberate on his part. Like Lecuyer, Akera is bringing a new perspective to one of the best-known stories in the history of computing:

## Historians make an effort to write clearly, at least compared to a typical technical paper in computer science.

the creation during the 1940s of the programmable electronic computer, initially as a scientific instrument, and its rapid spread into universities, companies, and government agencies over the subsequent 15 years. One long chapter is a biography of John Mauchly, a creator of ENIAC (remembered by historians as the first useful and flexibly programmable electronic computer). Other chapters explore topics as diverse as IBM's drive to sell its equipment to the new market of corporate computing centers, the role of the SHARE user group in creating programming as a new occupation, and the connection work on timesharing operating systems in university computer centers and the emergence of computer science as an academic field of study. These choices reflect in part the availability of archival source material, but Akera's shifts



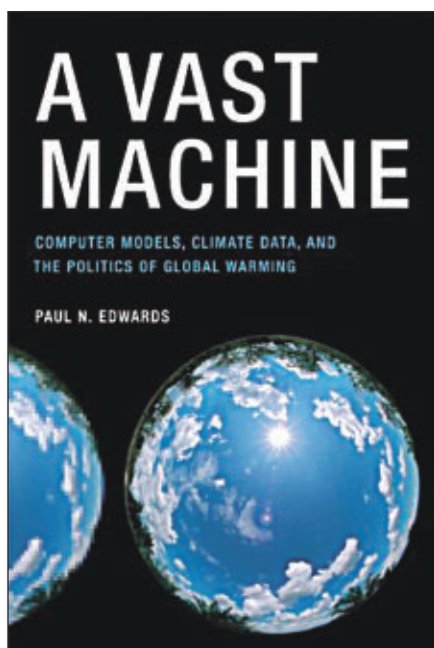
of focus and topic from individuals to institutions and technologies are also supported by his choice of the “ecology of knowledge” as an analytical framework. Put simply, this suggests that early computing developed as it did only because people with skills of many different kinds converged for reasons of their own around this new kind of technology. Because computer science, and computing more generally, emerged through the interactions of different kinds of experts and institutions we need to understand the entire intellectual “ecosystem” rather than fixating on any one part in isolation.

If that sounds daunting you might want to skip the introductory chapter. But the various stories told in the book are simply written and well researched, and the strength of Akera's holistic approach is made tangible through many unexpected insights. For example, we learn that Mauchly flitted from topic to topic in his early career, trying to turn his Ph.D. in molecular physics into a stable research career in the dreadful economic climate of the 1930s. He approached computing through statistics, meteorology, and tinkering with electronics. Akera shows how this complex background influenced his design for the ENIAC.

## 3. Scientists Know the World Through Computers

Like most other academics, when historians of computing get together we tend to bemoan the tendency of the world to completely ignore our ground-breaking work addressing vital issues. We then go back to our studies and spend years writing narrowly focused, painstakingly researched, books of intense interest to a few dozen of our colleagues. Of the thousand or so copies published by a major academic press a couple of hundred are given away as review or prize submission copies and the rest, once purchased, usually languish unread on the shelves of the ever-dwindling number of libraries that can still afford to err on the side of completeness. The problem is that writing a book the wider world might actually notice takes a lot of work. It is not easy to tackle a big topic, or to





articulate the relevance of historical work to present-day debates without falling victim to what historians call “presentism” (misinterpreting historical events in the light of present-day knowledge or perspectives).

The third CHM Prize winner, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* by Paul Edwards (MIT 2010, CHM Prize winner 2011) is an outstanding example of the potential for historians to contribute to broader public debates and give non-specialists insight into the work done by scientists and the process by which computer simulation has transformed scientific practice. Edwards tackles one of the most politically polarizing topics in U.S. science today: the connection of climate models to the real world. Without computers we could calculate average temperatures and plot trends, but only computer models can separate underlying climate trends from local or random fluctuations, project their future trend, or test explanations of the physical processes at work against the underlying data.

Our traditional idea of science, embraced by many scientists, is that scientists collect objective observations about the world and then formulate theories to explain them. Scholars in the field of science and technology studies, in which Edwards is trained, have instead stressed that nothing can be perceived except through one

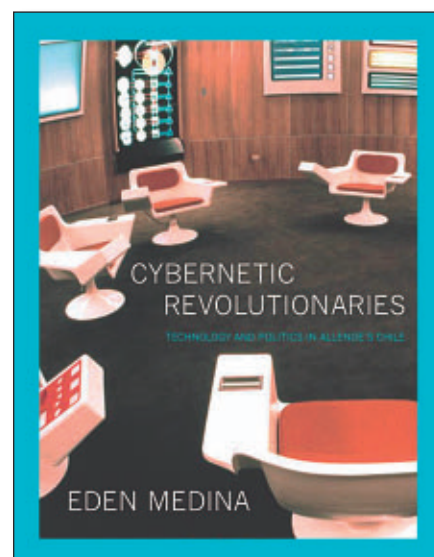
or another set of theories and assumptions. In climate science, as in much modern science, data points from the natural world become knowledge of a kind that can support or challenge a theory only after they are processed in computer models. These models are themselves based on theories. Thus, as Edwards succinctly puts in an introduction aimed at general readers, “without models there are no data.” This is not to say that Edwards is content, as some earlier radical scholars in science and technology studies were, simply to establish that the scientific knowledge he is examining was “socially constructed” and exit in triumph. We have arrived at an odd moment where this strategy, once associated with the academic left, is now a mainstay of the political right.

Instead, Edwards dives into decades of history to show the slow process by which these models were developed and explore their relationship to technological change. The first computerized weather forecasts were made in 1950 using ENIAC, by members of a team sponsored by John von Neumann. However, even this was only made possible by human networks to record and consolidate weather observations. Since then ever more powerful network, sensor, and computer technologies have been used to construct a global “information infrastructure” to collect climate data and drive ever more complex models of weather forecasting and climate change. He focuses particularly on the work needed to integrate information from different sources and the “data friction” technology imposes on its flexible use.

Edwards believes the public should understand how the “sausage” of scientific knowledge is made, to better understand its strengths and weaknesses. The fact that scientific knowledge is created by social processes and with simulation techniques does not mean all ideas about climate change are equally valid or that scientific knowledge has no special reliability. His success in this mission was confirmed when *The Economist* named *A Vast Machine* as one of just six “Books of the Year” in science and technology for 2010. It was the only one about computing.

#### 4. Computer Technologies Are Always Political

Eden Medina’s book *Cybernetic Revolutionaries: Technology and Politics in Allende’s Chile* (MIT 2011, CHM Prize winner 2012) takes a close look at the period from 1971 to 1973 when the British operations research specialist Stafford Beer was hired by Salvador Allende’s short-lived democratic Marxist government in Chile to implement his newly developed vision of cybernetic control. The boldest version of this Cybersym project imagined traditional political control replaced entirely by a new system in which decisions were influenced to the greatest possible extent by the input of ordinary citizens, industrial production was organized with the help of constantly updated computer models of the entire national economy, and decisions were based on information rather than bureaucratic self-interest. Beer modeled his plans on an abstracted view of the human nervous system in accordance with the central idea of cybernetics, which was that artificial, natural, and living systems are all governed by conceptually equivalent processes of feedback and control. Cybernetics originated in the 1940s with close ties to early work on computers and the support of many of the U.S.’s brightest minds across a range of disciplines. By the 1970s it was still fairly prominent in popular culture but was already sliding into obscure eccentricity within science as researchers favored more focused disciplinary approaches such as arti-



cial intelligence and cognitive science over the ostentatious universality of cybernetic theory.

Beer and his collaborators never came close to fulfilling their grand vision, though they did produce some economic models of little practical use and an “operations room” with an obvious debt to the bridge of the Starship Enterprise. As the revolution crumbled under a U.S. economic blockade and a series of strikes their most practical contribution to its defense was the national telex network, low tech even by the standards of the 1970s, which proved useful for centralized control of emergency responses. Beer himself was changed by his experiences in Chile, devoting himself to fixing the world rather than making money. The urbane lover of fine living gave up his Rolls Royce to spend much of his later career as a mystic, living simply in a primitive rural cottage.

This would seem to offer rich materials for a farce, or perhaps a tragicomic opera like those featuring talk show host Jerry Springer and Canadian Prime Minister Brian Mulroney. To her credit, Medina avoids mockery while doing justice to the gripping weirdness of the story. She puts the Chilean experience center stage, examining tensions between Beer’s vision and the agendas pursued by various hosts and collaborators. Media’s heart is open to the hopes for a better world that drove Allende’s revolution and the faith her characters put in Beer’s approach, but is not shy about speaking up when she catches them exaggerating its actual accomplishments or making contradictory statements. One contribution of her work is to remind us that computer technology has been in use outside the U.S. and Western Europe for a long time, and that its history in the developing world may follow a quite different path.

To me, the most fundamental lesson is that all technology is political and most new approaches to computing are promoted with utopian fantasies that later come to seem embarrassing. We instantly recognize the political nature and unhinged ambition of the Cybersym project because they are alien to our own experience

## These books demonstrate the rewards of tackling big topics of fundamental importance.

in wealthy countries during an economically liberal era. But, as I have discussed elsewhere, a similarly impractical vision of gigantic, real-time systems incorporating forecasting models was a mainstream part of corporate America in the 1960s.<sup>1</sup> Even the science fiction control room idea was already established in the business press.<sup>2</sup> Likewise, the banking industry first embraced the idea of the “cashless society” almost 50 years ago, but it still retains a futuristic allure. You may also remember all the predictions that the Internet would transform politics, revitalize democracy, and solve the problems of U.S. education. Snake oil, utopian dreams, and science fiction narratives have played a much more important role in the adoption of information technology than we would usually like to admit.

### 5. The History of Computing Is Maturing

This kind of prize plays an important role in the development of a field. By honoring excellence it helps to shape a canon of exemplary work and to build a consensus on topics and approaches of central importance. So what can we learn about the history of computing by looking at the books together?

One thing that jumps out is just how far the field has developed from its earliest days in the 1970s. The history of computing used to focus on the history of computers themselves. While many scholars continue to look closely at particular machines, such as ENIAC, there has been an unmistakable shift from hardware to applications and from narrow technical histories to broad portrayals of technologies in their social contexts.

These books, in particular, demonstrate the rewards of tackling big topics of fundamental importance such as the rise of Silicon Valley or the rise of computer use within scientific practice.

There has also been a shift in the kinds of people telling the stories. Early activity on history of computing was driven by computer scientists and pioneers such as Herman Goldstine, Brian Randell, Bernie Galler, Donald Knuth, and Nick Metropolis. In contrast, all four prize-winning authors discussed in this column have Ph.D.’s in some variety of science and technology studies or history of technology. Two also hold degrees in computer science or electrical engineering. Three hold faculty positions—one in a department of science and technology studies and two within information schools. None are appointed primarily in history departments or history of science programs.

This hiring pattern reflects the openness of other disciplines to historical scholarship in computing, but also has a negative impact on the development of the field as most of the best scholars have limited opportunities to teach in their specialist areas or to train doctoral students in historical research. I recently learned that Medina’s book is also the first history of computing book to win the annual Edelstein prize from the Society for the History of Technology, which is indisputably a good sign for the recognition of work on computing by other historical specialists. Hopefully this column and other historical commitments by the ACM and IEEE can maintain a similar connection between computer people and historians. I like to think that Paul Baran would have approved. ■

#### References

1. Haigh, T. Inventing information systems: The systems men and the computer, 1950–1968. *Business History Review* 75, 1 (2001), 15–61.
2. Widener, W.R. New management concepts: Working and profitable. *Business Automation* 15, 8 (1968), 28–34.

**Thomas Haigh** (thaigh@computer.org) is an associate professor of information studies at the University of Wisconsin, Milwaukee, and chair of the SIGCIS group for historians of computing. A guide to other outstanding historical work is at <http://www.sigcic.org/resources>.

Copyright held by author.

## Viewpoint

# What College Could Be Like

*Imagining an optimized education model.*

**B**ETWEEN 2000–2001 AND 2010–2011, prices for undergraduate tuition, room, and board at public institutions rose 42%, and prices at private not-for-profit institutions rose 31%, after adjustment for inflation.<sup>a</sup> With rising costs, large numbers of graduates are leaving college with huge debts and are struggling to find work. Fifty-three percent of recent college graduates are jobless or underemployed, the highest in 11 years.<sup>b</sup> The core value proposition of higher education is under increasing scrutiny, in part because of the disconnect between student’s expectations, the traditional classroom experience, and the ever-growing need for active creators in the marketplace.

There is a basic divide between most students’ expectations for college—a means to employment first and a good intellectual experience second; and what universities believe their value is—an intellectual and social experience first, with only secondary consideration to employment. At the same time, existing credentials carry very little information for employers to really decide who has the skills and talents they need.

a U.S. Department of Education, National Center for Education Statistics. *Digest of Education Statistics*, 2011 (NCES 2012-001), 2012, Chapter 3.

b Yen, H. Half of recent college grads underemployed or jobless, analysis says. *Cleveland.com*. Associated Press (Apr. 23, 2012); [http://www.cleveland.com/business/index.ssf/2012/04/half\\_of\\_recent\\_college\\_grads\\_u.html](http://www.cleveland.com/business/index.ssf/2012/04/half_of_recent_college_grads_u.html).



So let us face this as an open-ended design problem: Is it possible to craft a university experience that bridges the gap between students’ expectations, universities’ strengths, and employers’ needs? One that provides the rich social and intellectual atmosphere of a good existing college, while at the same time exposing students to those intellectual but

also practical fields that will make them valuable to the world. Where “microcredentials” could be earned and maintained in intellectual and vocational fields to prove to the world what a student can do. And now let’s be ambitious: Might there be a sustainable way to make this experience free or even pay the students to participate?

# ACM Transactions on Accessible Computing



This quarterly publication is a quarterly journal that publishes refereed articles addressing issues of computing as it impacts the lives of people with disabilities. The journal will be of particular interest to SIGACCESS members and delegates to its affiliated conference (i.e., ASSETS), as well as other international accessibility conferences.

[www.acm.org/taccess](http://www.acm.org/taccess)  
[www.acm.org/subscribe](http://www.acm.org/subscribe)



Association for  
Computing Machinery

Computer science is a good place to start. I know the field reasonably well and I also have a sense for the job market—which is tight and growing tighter every day. It is a field where degrees can be valuable, but the ability to design and execute on open-ended, complex projects is paramount; 17-year-olds with unusual creativity and intellect have been known to get six-figure salaries. Because of the demand for talent and the recognition that college degrees and high GPAs are not the best predictor of creativity, intellect, or passion, top employers have begun to treat summer internships as something of a farm league. They observe students *actually working* and make offers to those who perform the best. Employers know that working with a student is an infinitely better assessment than any degree or transcript.

Students have also begun to recognize something very counterintuitive: that they are more likely to get an intellectual grasp of computer science—which is really the logical and algorithmic side of mathematics—by working at companies like Google, Microsoft, or Facebook or trying to create their own mobile applications than by reading textbooks or sitting in lecture halls. They see the real-world projects as being more intellectually challenging and open-ended than the somewhat artificial projects given in classrooms. Even more, they know that the product of their efforts has the potential to touch millions of people instead of just being graded by a teaching assistant and thrown away.

So, to be clear, in software engineering, the internship and self-directed projects have become far more valu-

**Employers know  
that working  
with a student is  
an infinitely better  
assessment than any  
degree or transcript.**

able to the students, as an intellectual learning experience, than any university class. And they have become more valuable to employers, as a signal of student ability, than any formal credential, class taken, or grade point average.

When it comes to internships, I want to emphasize that these are very different from the ones many people remember having even 20 years ago. There is no getting coffee for the boss, sorting papers or doing other types of busywork. The projects are not just cute things to work on that have no impact on real people. In fact, the best way to differentiate between forward-looking, 21<sup>st</sup> century industries and old-school, backward-looking ones is to see what interns are doing. At top Internet companies, interns might be creating patentable artificial intelligence algorithms or even creating new lines of business. By contrast, at a law firm, government office, or publishing house, they will be doing paperwork, scheduling meetings, and proofreading text. This trivial work will be paid accordingly, if at all, whereas pay scales at the new-style internships reflect the seriousness of the work involved.

Given the increasing importance of real-world projects in terms of both intellectual enrichment and enhancement of job prospects, why do traditional colleges tend to limit them to summers, pushing them aside to cater to the calendar needs of lectures and homework? The answer is simple inertia—this is how it has always been done, so people have not really questioned it.

Actually, some universities have. Despite being founded not even 60 years ago, the University of Waterloo is generally considered to be Canada's top engineering school. Walk down a hallway at Microsoft or Google and you will find as many Waterloo grads as those from MIT, Stanford, or Berkeley—despite the fact that, because of work visa issues, it is a significant hassle for U.S. employers to hire Canadian nationals. And this is not some attempt to get low-cost labor from across the border—Waterloo graduates are commanding salaries as high as the very best American grads. What is Waterloo doing right?

For one thing, Waterloo recognized the value of internships long ago (they call them co-ops) and has made them

an integral part of its students' experience. By graduation, a typical Waterloo grad will have spent six internships lasting a combined 24 months at major companies—often American. The typical U.S. college graduate will have spent about 36 months in lecture halls and a mere three to six months in internships.

This past winter—not summer—all of the interns at the Khan Academy, and probably most of the interns in Silicon Valley, were from Waterloo because it is the only school that views internships as an integral part of students' development *outside of the summer*. While the students at most colleges are taking notes in lecture halls and cramming for winter exams, the Waterloo students are pushing themselves intellectually by working on real projects with experienced professionals. They are also getting valuable time with employers and pretty much guaranteeing several job offers once they graduate. On top of that, some are earning enough money during their multiple high-paying internships to pay for their tuition (which is about 1/6th to 1/3rd the cost of a comparable American school) and then some. So Waterloo students graduate with valuable skills, broad intellectual development, high-paying jobs, and potential savings after four or five years.

Compare this to the typical American college grad with tens or hundreds of thousands of dollars in debt, no guarantee of an intellectually challenging job, and not much actual experience with which to get a job.

Waterloo has already proven that the division between the intellectual and the useful is artificial; I challenge anyone to argue that Waterloo co-op students are in any way less intellectual or broad thinking than the political science or history majors from other elite universities. If anything, based on my experience with Waterloo students, they tend to have a more expansive worldview and are more mature than typical new college graduates—arguably due to their broad and deep experience base.

So let us imagine optimizing the model that schools like Waterloo have already begun. Imagine a new university in Silicon Valley—it does not have to be there but it will help to make things

## What is completely different is where and how the students spend their days.

concrete. I am a big believer that inspiring physical spaces and rich community really does elevate and develop one's thinking. So we will put in dormitories, nicely manicured outdoor spaces, and as many areas that facilitate interaction and collaboration as possible. Students would be encouraged to start clubs and organize intellectual events. So far, this is not so different from the typical residential college.

What is completely different is where and how the students spend their days. Rather than taking notes in lecture halls, these students will be actively learning through real-world, intellectual projects. A student could spend five months at Google optimizing a search algorithm. She might spend another six months at Microsoft working on human speech recognition. The next four months could be spent apprenticing under a designer at Apple, followed by a year of building her own mobile applications. Six months could be spent doing biomedical research at a startup or even at another university like Stanford. Another four months could be spent prototyping and patenting an invention. Students could also apprentice with venture capitalists and successful entrepreneurs, eventually leading to attempts to start their own businesses. One of the primary roles of the college itself would be ensure the internships are challenging and intellectual; that they truly do support a student's development. The college will also provide a scaffold of shared, physical experiences, but they will not be passive lectures. They will be active interactions between faculty and students.

All of this will be tied together with a self-paced academic scaffold through something like EdX (Harvard, MIT, and Berkeley's "MOOC") or Khan Academy. Students will also still be

expected to have a broad background in the arts and deep proficiency in the sciences; it will just be done in a more natural way. They will be motivated to formally learn about linear algebra when working on a computer graphics apprenticeship at Pixar or Electronic Arts. They will want to learn accounting when working under the CFO of a publicly traded company. Ungraded seminars will be held regularly during nights and weekends when students can enjoy and discuss great works of literature and art. If the students decide they want to prove their academic ability within a domain—like algorithms or French history—they can sign up for rigorous assessments leading to microcredentials that are valued by employers and graduate schools.

This thought experiment envisions a school focused on engineering, design, and entrepreneurship in Silicon Valley. We placed it there so that it could take advantage of the local ecosystem, but why not a school of finance or journalism located in New York or London, or a school focused on energy in Houston? Even better, why can't they all be affiliated so that a student can experience multiple cities and industries, all while having a residential and intellectual support network?

Will this be for everyone? Absolutely not. But majoring in literature or accounting at a traditional university is not for everyone either. There should be more options, and this could be one of them—an option that introduces diversity of thought and practice into a higher education world that has not changed dramatically in hundreds of years.

It also should be noted that this does not necessarily have to be a new university. Existing campuses could move in this direction by de-emphasizing or eliminating lecture-based courses, having their students more engaged in research and co-ops in the broader world; and have more faculty with broad backgrounds that show a deep desire to mentor students. □

---

**Salman Khan** (skhan@khanacademy.org) is the founder and executive director of the Khan Academy in Mountain View, CA.

---

This Viewpoint is adapted from S. Khan, *The One World Schoolhouse: Education Reimagined*. Hachette Book Group, 2012.

Copyright held by author.

## Viewpoint

# Conference-Journal Hybrids

*Considering how to combine the best elements of conferences and journals.*

**N**UMEROUS PROPOSALS AND experiments have addressed the stresses resulting from computer science's shift from a journal to a conference publication focus, discussed in over two dozen commentaries in *Communications*, three panels at CRA Snowbird conferences, the Workshop on Organizing Workshops, Conferences and Symposia for Computer Systems (WOWCS'08), and a recent Dagstuhl workshop.<sup>a</sup> We focus here on recent efforts to blend features of conferences and journals, highlighting a conference that incorporated a revision cycle without increasing the overall reviewer workload. We also survey a range of approaches to improving conference reviewing and management.

Proposals fall into three principal categories:

1. Return to the journal orientation that has long served the sciences and engineering well.
2. Develop hybrid approaches that combine features of journals and conferences.
3. Improve conference reviewing and management in other ways.

**1. Return to the traditional journal focus of the sciences.** Rolling back the clock is attractive but probably not feasible. It would require a unified effort in a famously decentralized discipline. It would be resisted by established researchers who built their careers on conference publication. It would have

to overcome the forces that motivated the shift to conference publication in the first place.<sup>4</sup>

**2. Combine journal and conference elements.** Conference proceedings have usurped two key journal functions: They are now archived and widely available. The boundary is blurred further when conferences increase reviewing rigor and journals reduce reviewing time. Article lengths are converging as reduced production costs let conferences relax or drop length limits and as demands on reader attention push journals to decrease article length.<sup>b</sup> The most significant remaining distinctions are: journals encourage more revision and are less deadline driven; conferences promote informal interaction and other community-building activities.

Most calls for change are related to these two distinctions. Conference program committees evaluate papers on different dimensions (originality, technical rigor, audience engagement, and so forth) and make binary, in-or-out quality determinations under time pressure on first drafts. Conferences struggle to foster a sense of community when rejecting the great majority of submissions. Authors perceive injustice, feel their careers could be affected, and are driven to attend or form other conferences. Table 1 lists approaches that seek a middle ground to deliver benefits of both conferences and journals.

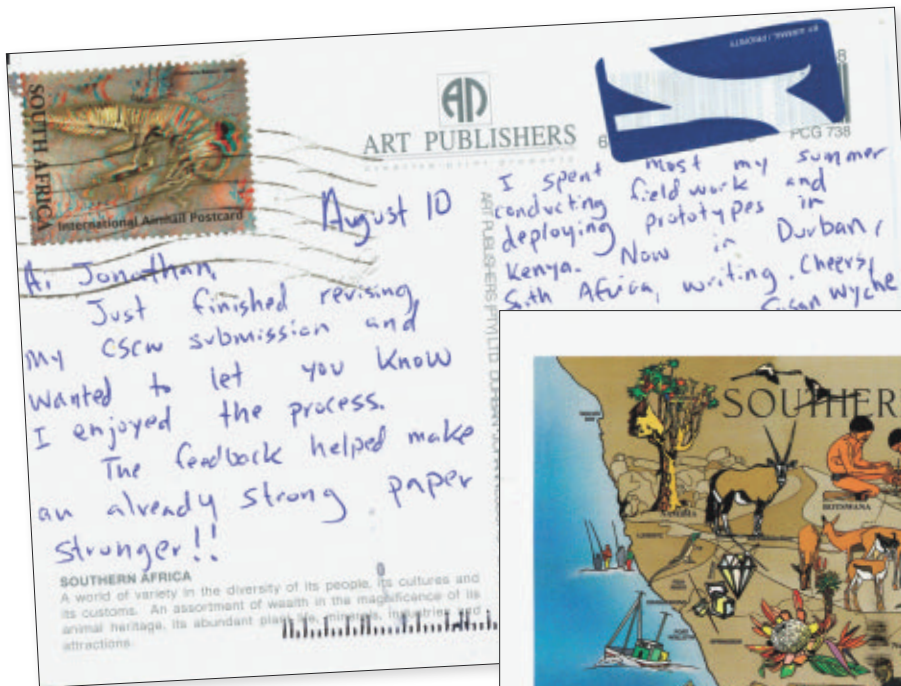
*Journal acceptance precedes conference presentation.* Articles submitted to the online monthly journal *Proceedings of the Very Large Data Bases Endowment (PVLDB)* are limited to 12 pages and receive three rapid reviews. Those that have been accepted a couple months prior to the annual fall VLDB conference are eligible for presentation. Longer versions of *PVLDB* articles can be published in the independently managed *VLDB Journal*. The same process is used by the Conference on High-Performance and Embedded Architectures and Compilers (HiPEAC), which presents papers that have been accepted by the *ACM Transactions on Code Optimization (TACO)*. Although these journals stress rapid reviewing, this undercuts the hypothesis that rate of innovation was tied to the shift to a conference focus in computer science in the U.S.

VLDB and HiPEAC no longer have separate submission and review processes. Alternatively, partial overlaps are being explored. Submissions to a special issue of the journal *Theory and Practice of Logic Programming* are also submissions to The International Conference on Logic Programming (ICLP), and granted one revision cycle if necessary, as is common to journal special issues. This approach retains much of the conference deadline and program committee structure. We will describe this approach in detail as used for conferences unaffiliated with a journal.

Similarly, the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) now will

<sup>b</sup> Haslam, N. Bite-size science: Relative impact of short article formats. *Perspectives on Psychological Science* 5, 3 (2010), 263–264; <http://pps.sagepub.com/content/5/3/263.abstract>. SIGGRAPH dropped submission length limits in the 1990s, followed by UIST and CSCW more recently. Few papers exceed 10 or 12 pages.

<sup>a</sup> Links to *Communications* commentaries, Snowbird panels, WOWCS'08 notes, and the Dagstuhl workshop are at <http://research.microsoft.com/~jgrudin/CACMviews.pdf>.



Postcard from South Africa commenting on the CSCW 2012 review process.

have a journal track comprising papers accepted by the *Machine Learning and Data Mining and Knowledge Discovery* journals. An editorial board is created to handle submissions for the journal track of the conference that is distinct from the journal editorial boards and from the program committee that will handle submissions to non-journal conference tracks.<sup>c</sup>

These ensure the preeminence of a journal. They do not provide the conference's traditional role of building community by providing authors with feedback on less polished work that is intended for subsequent journal submission: The adage that a computer science conference is "a journal that meets in a hotel" is literally true in this case. Although best suited to a small research community, the practice has been embedded in larger conferences. The Computer-Human Interaction (CHI) conferences' 1,500–2,000 submissions are more than a journal review process could cope with, but *ACM Transactions on Computer-Human Interaction* articles comprise about 5% of CHI presentations. Similarly, 23% of Intelligent User Interfaces (IUI) 2011 presentations were of papers from *ACM Transactions on Interactive Intelligent Systems*. The Association

for Computational Linguistics (ACL) also presents papers published in the *Transactions of the ACL*.

*Shepherd conference papers become journal articles.* SIGGRAPH and Infovis proceedings become issues of *ACM Transactions on Graphics (TOG)* and *IEEE Transactions on Visualization and Computer Graphics*, respectively. IJ-CAI'13 will have a track in the *Journal of AI Research*. Following a conditional acceptance, a committee member may

**Journals encourage more revision and are less deadline driven; conferences promote informal interaction and other community-building activities.**

be assigned to check each revision. In practice, the committee expressed an inclination to accept and time is limited; acceptance is all but certain. Such papers are rarely if ever rejected. Shepherding is not a journal-style revision and re-review.

Concerned that this practice will lower journal standards, ACM has a new policy<sup>2</sup> that forbids rechristening of conference papers as journal articles (exempting SIGGRAPH, although many authors list SIGGRAPH papers as peer-reviewed conference papers rather than *TOG* journal articles on their CVs). ACM may be fighting a tide that is eroding the conference-journal distinction. Outside ACM, boundaries continue to blur: the *Journal of Machine Learning Research* publishes a volume of workshop and conference proceedings, a useful collection of related work.

*Conferences without a journal affiliation that incorporate a revision cycle.* Aspect-Oriented Software Development (AOSD) offers multiple submission deadlines. For the March 2012 conference, authors could submit in April,

<sup>c</sup> See [http://www.acm.org/sigs/volunteer\\_resources/conference\\_manual/acm-policy-on-the-publication-of-conference-proceedings-in-acm-journals](http://www.acm.org/sigs/volunteer_resources/conference_manual/acm-policy-on-the-publication-of-conference-proceedings-in-acm-journals).

July, or September of 2011. The April and July review processes yield accepts, rejects, and revise and resubmits. Submissions in September are either accepted or rejected. The Asian Conference on Machine Learning (ACML2012) had two deadlines and the same model.

The ACM Computer Supported Cooperative Work 2012 (CSCW 2012) conference employed a single submission date and a five-week revision period. This is essentially a rapid version of a journal special issue process: a submission deadline followed by reviewing and one round of revision that is fully examined by the original reviewers. This process is discussed below. It has now been used by ACM Interactive Tabletops and Surfaces 2012, SIGMOD 2013, and CSCW 2013.

Adding a revision cycle resembles a time-compressed variant of the usual practice, where a rejected paper can be resubmitted in a year, but the dynamic is different. A reviewer does not have to make an often stressful, binary, in-or-out recommendation. Reviewers have more incentive to provide constructive, complete reviews, rather than primarily identify flaws that justify rejection. Authors who will get the same reviewers have more incentive to respond to suggestions than when they resubmit a rejected paper to a different conference with a new set of reviewers. Reviewers have to examine some submissions twice, but most find the second review easier and are rewarded by seeing authors take their suggestions seriously. When a rejected paper is resubmitted elsewhere, the overall reviewing burden for the community is greater and less rewarding.

**The CSCW 2012 experiment.** As program chairs for this large annual conference, we were asked to move the submission deadline forward two months to reduce the reviewing overlap with the larger CHI conference. To turn this lemon into lemonade, we inserted the revision cycle.

Computer Supported Cooperative Work conferences averaged 256 submissions and 56 acceptances (22%) from 2000–2011. In recent years, most submissions were reviewed by two program committee members (called Associate Chairs or ACs) and three external reviewers, followed by a face-to-face meeting of the ACs.

**Table 1. Approaches that merge conference and journal elements.**

|   |
|---|
| Journal acceptance precedes conference presentation.  |
| Shepherded conference papers become journal articles. |
| Conference reviewing incorporates a revision cycle.   |

**Table 2. Adding a revision cycle.**

#### Benefits

|  |
|--|
| Higher quality.  |
| Higher quantity and attendance.                                  |
| Zero-sum game eliminated.  |
| More constructive review focus. Reviewers see results of effort. |
| Few decisions passed to committee.                               |

#### Challenges

|                                  |
|----------------------------------|
| More periods of review activity. |
| Perception of acceptance rate.   |
| More conference tracks?          |
| Reviewer assignment still hard.  |

The 2012 conference received a record 415 submissions. In the first round, each submission was reviewed by an AC and two external reviewers. Forty-five percent were rejected and authors notified. The authors of the remaining 55% had five weeks to upload a revision and a document describing their responses to reviewer comments. A few withdrew; the others revised, often extensively. The revisions were reviewed by the original reviewers. After online discussion only 26 remained unresolved. The face-to-face program committee meeting may have been unnecessary, but it had been scheduled. Final decisions were made there.

In the second round, 27% of the revisions were rejected. Overall, 39.5% of the original submissions were judged to have cleared the quality bar. The traditional process would have accepted approximately 22% ‘as is’; this year all papers were revised. The consensus was that a high-quality conference had become larger and much stronger.

Analysis of the CSCW 2011 data enabled us to focus reviewing where it was needed—we did not increase the overall reviewing burden. In 2011, 60% of submissions found no advocate in the early reviews and *not one of those was ultimately accepted*. Yet each received a summary review by one AC, was read by a second AC, and was given a rebuttal opportunity. For CSCW 2012, 45% had no advocate in the first round and were summarily but politely rejected. The average workload per submission was reduced despite each “revise and resubmit” paper receiving eight reviews by four reviewers over the two rounds.

The number of reviews or serious considerations fell from an average of 5.5 per submission the previous year to 4.6, a reduction of 400. And 35% of all reviews were of revisions, aided by the authors’ documents describing the changes. The rebuttal process was not needed; dropping it led to no objections. Finally, the 60–80 papers that were accepted because of the revision process would have been prime candidates for resubmission elsewhere, so the community was spared hundreds of additional reviews downstream. (Many conferences already have more streamlined review processes than we started with, but could still realize a net reduction in effort by adopting a revision cycle.)

*Positive outcomes.* First the good news. Overall the reports were very positive. The revise-and-resubmit option for borderline papers in the first round reduced reviewer stress. Reviewers could focus on finding what might be of interest and formulating constructive guidance for revision, rather than identifying flaws that warrant rejection. Reviewers found it rewarding to see authors who had responded well to comments. An interesting benefit was that acceptance was not a zero-sum game. We did not have a quota, just the goal of keeping the same quality bar. Without the pressure to reject 75% to 80% of submissions, a four-person review team could iterate with an author without disadvantaging other authors. Some review teams engaged in protracted online discussions. With few decisions left to make, half of the face-to-face meeting was spent discussing broader issues and planning



for next year. Some program committee members remarked that for the first time, the review process left them energized rather than drained.

Conference attendance increased 80% to a record 657. Over one-third of attendees responded to a post-conference survey. Of those expressing an opinion, 94% felt the new process improved the conference. The community building that was once the *raison d'être* of conferences was arguably strengthened. Authors and reviewers sent many positive comments, exemplified by the postcard shown earlier in this column.

**Challenges.** Some reviewers found that multiple reviewing sessions for one conference were taxing, especially given that reviewing was in the summer when vacations were scheduled. An operational challenge was that despite good intentions, PC members and reviewers with strongly entrenched habits born of years of committee service could find it difficult to adjust to new ways of working.

The greatest benefit of the revise-and-resubmit approach is that rather than rejecting many papers that need to be polished or fixed, reviewers can coach the authors to produce higher-quality versions fit for publication. Ironically, a challenge to employing this approach is the perception that a higher acceptance rate signals low conference quality. Despite a widespread view that quality had risen, some researchers fear that external referees will look unfavorably on the acceptance rate. Many senior researchers discount the selectivity = quality equation, but it is built into the assessment practices of some universities and has traction among junior researchers who like to think that decisions are based on visible, objective data. Acceptance rate as a signifier of quality has a sound historical basis.

In 1999, some peer-reviewed conferences were formally recognized as sources of high-quality computer science research in the U.S.<sup>6</sup> However, not all conferences in the U.S. and few elsewhere stressed polished research; they remained more inclusive, providing authors with feedback on work in progress toward journal publication. Acceptance rate—rejecting 75% or more of its members' proffered work—signaled that a conference emphasized quality over

community-building inclusiveness.

That said, for acceptance rate to truly indicate quality across different conferences, three things must be true: the submissions must represent the same quality mix; the process used to reject submissions must accurately measure merit; and the same assessment process must be used by the conferences being compared. Unfortunately, none of these are reliably true. Conferences vary in the quality of the submissions they attract. Thomas Anderson has argued eloquently that beyond the top handful of strong papers, there is little ability for reviewers to reliably differentiate between a large fraction of submissions.<sup>1</sup> And process changes such as ours will raise the proportion of papers that attain a high level of quality.

Eventually, citation or download numbers could indicate the impact of a process change. CSCW 2012 is accumulating citations far more rapidly than its predecessors, but it is too early for definitive judgment. For now, to see the problem with the selectivity = quality equation, consider this analogy: Two organizations each admit six applicants for a probationary year. One gives them minimal attention, tests them at year's end, and retains the two who managed to learn on their own. The second invests in training and at year's end keeps four: The two who would have survived on their own are, by virtue of the training, rock stars, and two others passed. The first organization had a 33% acceptance rate, the second a 66% acceptance rate, based on the same raw materials. Low acceptance rate accompanied lower quality. The same is true with journals—those that work patiently with authors over multiple versions raise both quality and acceptance rate. Process matters.

## Acceptance rate as a signifier of quality has a sound historical basis.

Another concern is that many people enjoy single-track conferences. Raising the acceptance rate requires adding tracks, lengthening a conference, or allocating less presentation time to papers. CSCW had already moved on from being single-track. As the field grew and specializations developed, trade-offs arose. With more submissions and more varied submissions, rejection rates were pushed up, reviewing became more stressful and less uniform, and incremental advances competed for space with new ideas. CSCW shifted to multiple tracks and a tiered program committee. Other conferences have maintained a single track and driven acceptance rates into single or low double digits. Many reports of conference stress published in *Communications*<sup>d</sup> come from these fields. High rejection rates foster disaffection and drive authors to other conferences, dispersing the literature and undermining the sense of community that the single track initially created.

Finally, the perennial large-conference challenge of matching reviewers to submissions. We devised a comprehensive set of keyword/topic areas. We let associate chairs bid on submissions. Nevertheless, a match based on topic often proves to be poor due to differences in preferred method, theoretical orientation, or other factors. Some reviewers, who might be called 'annihilators,' consistently rate papers lower than others who handle the same submissions. Some reviewers are 'Santa Clauses.' Statistical normalization does not fully compensate—a submission does not get the essential advocate by adjusting the scores of annihilators. Others see some merit and some weakness in everything and rate everything borderline. Add Anderson's observation that differences in quality are very small over a broad range of submissions, and luck in reviewer assignment can be a larger factor in outcomes than submission quality. Next, we review other proposals and experiments to improve conference management.

<sup>d</sup> Links to *Communications* commentaries, Snowbird panels, WOWCS'08 notes, and the Dagstuhl workshop are at <http://research.microsoft.com/~jgrudin/CACMviews.pdf>.

### 3. Improve existing review processes.

Approaches listed in Table 3 have been noted in *Communications* commentaries. We are not endorsing them all. Some conflict. Many could be used together. Some are in regular use in some conferences; others have been tried but did not take root.

Conference size is an important variable in assessing utility. Some prestigious conferences attract fewer than 150 submissions and might have a flat program committee. Some attract more submissions and enlist a second tier of external (or ‘light’) reviewers. The largest can have three levels, with tracks or subcommittees. Potential comparisons increase nonlinearly with submission number; approaches that work for one size may not scale up or down.

*Tracking submission histories.* The International Conference on Functional Programming allows authors of papers previously rejected (by ICFP or other conferences) to append the review history and comments. Eleven percent of ICFP 2011 authors declared a history. Half of those provided annotated reviews. Half were re-reviewed by one of the original reviewers, and all except one were accepted. Although mandating that authors report a paper’s prior submission history would face practical and ethical challenges, it can be a win-win for authors and reviewers when authors opt to do it.

**Table 3. Proposals for improving conference reviewing outcomes.**

|                                   |
|-----------------------------------|
| Track submission histories.       |
| Streamline the review process.    |
| Adopt double-blind reviewing.     |
| Clarify review criteria.          |
| Improve reviewer match.           |
| Control for reviewer differences. |
| Write more constructive reviews.  |
| Reduce feedback to authors.       |
| Allow author rebuttals.           |
| Stage a shadow PC meeting.        |
| Mentor or shepherd submissions.   |
| Publish reviews.                  |
| Improve presentation quality.     |

*Streamlining.* In phased reviewing, submissions first get two or three reviews, then some are rejected and reviewers added to the others. Often all results are announced together, but EuroSys starting in 2009 has notified the authors of rejected papers following the first phase, as CSCW 2012 did after the first round, enabling authors to quickly resume work. Conferences have also experimented with various methods of ordering papers for discussion in the committee: randomly, periodically inserting highly rated or low-rated papers, starting at the top, starting at the bottom. No consensus has emerged, although some report that papers that are first discussed late in the day tend to fare poorly whatever their rating.

*Double-blind reviewing.* Evidence indicates that author anonymity is fairer, and this practice has spread. To anonymize is sometimes awkward for authors. Because anonymity can inhibit program committee members from finding duplication or extreme incrementalism, some two-tier committees only blind the less influential external reviewers.

*Clarifying review criteria.* Reviewers have been asked to rate papers on diverse dimensions: originality, technical completeness, audience interest, strengths and weaknesses, and so on. In our experience, committees working under time pressure focus on the overall rating; writing quality gets some attention and nothing else does once conferences reach a moderate size.

*Matching reviewers to papers.* In general, the smaller the conference, the more easily reviewer assignments can be tuned. Keyword or topic area lists are common. Matching semantic analyses of a reviewer’s work to submissions has been tried. Some conferences let reviewers bid on submissions based on titles and abstracts. CSCW 2013 authors could nominate ACs for their papers; although no promises were made, their choices were helpful. IUI twice let program committee members choose which submissions to review; an absence of reviewer interest was a factor in the final decision.

*Normalizing to control for consistently negative or positive reviewers.* Since 2006, Neural Information Processing Systems (NIPS) has calculated a statistical normalization to offset consistently

high or low reviewer biases. Other conferences have tried this, usually just once. It does not counter biases directed at particular topics or methods, the occasional reviewer who gives only top and bottom ratings “to make a real difference,” or those who uniformly give middling reviews. It does not produce an advocate—knowing that reviewers were inherently negative does not replace harsh critiques with positive points. Normalization may be more useful for smaller conferences with fewer papers to discuss. Another approach tried once by SIGCOMM (2006), SOSP (2009), and other conferences had reviewers rank submissions. Although the rankings were used primarily to create a discussion order, relative judgments could counter a reviewer’s overall positive or negative bias.

*More or less constructive reviews?* A high-pressure binary decision process often yields reviews that focus on identifying a fatal flaw. Some people suggest that providing less feedback could reduce reviewer load and discourage incomplete submissions, but calls for more balanced and constructive appraisal are more often heard.

*Rebuttals.* Authors are generally prohibited from promising to make changes, but many appreciate the opportunity to express themselves. This practice has spread after being introduced over a decade ago. How rebuttals affect outcomes is unknown. The art of writing an effective rebuttal must be mastered, which disadvantages the uninitiated.

*Shadow PCs.* Several conferences—NSDI, SIGCOMM, SOSP, and EuroSys—have staged full-blown mirror events for training purposes.<sup>3,5</sup> Shadow PC output does not inform PC decisions. Large differences in acceptance decisions do suggest that younger and older researchers have different orientations and support Anderson’s hypotheses about imprecision in conference reviewing.<sup>1</sup>

*Mentor or shepherd authors of tentatively accepted papers.* Assigning mentors to papers in the preparation phase can help authors, but is stressful for the mentors of struggling submissions. This practice has been tried but not taken root. More frequently, a shepherd is assigned to guide a tentative acceptance into final form. Some authors find shepherding helpful,

others report perfunctory or absentee shepherds. Assigning a shepherd is often a face-saving means to calm down a program committee member who has reservations. Shepherded papers virtually always make it into the coral. The 2011 Internet Measurement Conference gave authors a choice of a shepherd or a 'soft' open review alternative (publishing the paper with its reviews and the author's descriptions of changes). Most chose the latter.

*Publish reviews.* Reviews of accepted HotNets 2004 and SIGCOMM 2006 papers were posted publicly. Neither conference continued the practice, perhaps because of the extra effort that reviewers reported. Similar experiments are under way.

*Improve presentations.* ICME 2011 required authors of accepted papers to submit lecture videos. A subset was selected for oral presentation.

Other member support efforts include offering a free registration and a five-minute 'booster' presentation to finishing graduate students at Innovations in Theoretical Computer Science. Publicly honoring exemplary reviewers, a practice of some journals, has been encouraged for conferences.

### Conclusion: Change Is Probably Inevitable

In computer science especially, conferences and journals compete to communicate and archive results. Journal articles grow shorter and reviewing time decreases. Conference reviewing rigor increases and proceedings are more polished. Measures of impact now cover both. There are stresses, but is there a need for a major adjustment?

We think so. The wealth of proposals and experiments signal dissatisfaction with the status quo. Some involve bringing conferences and journals closer through direct ties or shared features. Adding a revision cycle led to more acceptances, but also shorter presentation times, more parallel sessions, and a shift from acceptance rates to citations and downloads as measures of impact.

At risk with conference-journal hybrids is the community building and community maintenance that conferences once provided. Many conferences decline in size even as the researchers and practitioners in the field

## At risk with conference-journal hybrids is the community building and community maintenance that conferences once provided.

increase in number. The popularity of workshops that accompany conferences reveals a need for member support and a sense of community, but a set of disjoint workshops does not signify a thriving community. Indeed, successful workshops often spin off to become stand-alone conferences.

Other changes may be coming. Globalism has made geographically anchored conferences more expensive. As real-time audio and video become more reliable, travel becomes more uncomfortable, and concern for our carbon footprint grows, community activity may move online, perhaps suddenly. We cannot predict the future, but we do know the future will not resemble the present or the past. ■

#### References

1. Anderson, T. Conference reviewing considered harmful. *ACM SIGOPS Operating Systems Review* 43, 2 (2009), 108–116.
2. Blockeel, H., Kersting, K., Nijssen, S. and Zelezny, F. A revised publication model for ECML PKDD. Computing research repository, 2012; <http://arxiv.org/pdf/1207.6324v1.pdf>
3. Feldmann, A. Experiences from the SIGCOMM 2005 European shadow PC experiment. *ACM SIGCOMM Computer Communication Review* 35, 3 (2005), 97–102; <http://dl.acm.org/citation.cfm?id=1070889>
4. Grudin, J. Technology, conferences, and community. *Commun. ACM* 54, 2, (Feb. 2011), 41–43.
5. Isaacs, R. Report on the 2007 SOS shadow program committee. *ACM SIGOPS Operating Systems Review*, 42, 3 (2008), 127–131; <http://dl.acm.org/citation.cfm?id=1368524>.
6. Patterson, D., Snyder, L., and Ullman, J. Evaluating computer scientists and engineers for promotion and tenure. *Computing Research News* (Sept. 1999), A–B.

**Jonathan Grudin** ([jgrudin@microsoft.com](mailto:jgrudin@microsoft.com)) is a principal researcher at Microsoft Research in Redmond, WA.

**Gloria Mark** ([gmark@uci.edu](mailto:gmark@uci.edu)) is a professor of information and computer science at the University of California, Irvine.

**John Riedl** ([jriedl@gmail.com](mailto:jriedl@gmail.com)) is a professor of computer science at the University of Minnesota.

Copyright held by author.

# Calendar of Events

### January 16–20

Foundations of Genetic Algorithms XII, Adelaide, Australia, Sponsored: SIGEVO, Contact: Frank Neumann, Email: [frank.neumann@adelaide.edu.au](mailto:frank.neumann@adelaide.edu.au)

### January 17–19

The 7<sup>th</sup> International Conference on Ubiquitous Information Management and Communication, Kota Kinabalu, Malaysia, Sponsored: SIGAPP, Contact: Sukhan Lee, Email: [lsh@ece.skku.ac.kr](mailto:lsh@ece.skku.ac.kr)

### January 22–25

18<sup>th</sup> Asia and South Pacific Design Automation Conference, Yokohama, Japan, Sponsored: SIGDA, Contact: Shinji Kimura, Phone: +81-93-692-5374, Email: [shinji\\_kimura@waseda.jp](mailto:shinji_kimura@waseda.jp)

### January 23–25

The 40<sup>th</sup> Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, Rome, Italy, Sponsored: SIGACT, Contact: Roberto Giacobazzi, Phone: +39-045-802-7995, Email: [roberto.giacobazzi@univ.it](mailto:roberto.giacobazzi@univ.it)

### January 26

2<sup>nd</sup> Program Protection and Reverse Engineering Workshop, Rome, Italy, Sponsored: SIGPLAN, Contact: Jeffrey Todd McDonald, Phone: 850-322-7866, Email: [jtmcdonald@southalabama.edu](mailto:jtmcdonald@southalabama.edu)

### January 28–29

Personalized Medicine World Conference (PMWC) 2013, San Antonio, TX, Sponsored: SIGCHI, Contact: Susan R. Fussell, Phone: 607-255-1581, Email: [sfussell@cornell.edu](mailto:sfussell@cornell.edu)

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

## Constraints in an environment empower the services.

BY PAT HELLAND

# Condos and Clouds

LIVING IN A CONDOMINIUM (commonly known as a condo) has its constraints and its services. By defining the lifestyle and limits on usage patterns, it is possible to pack many homes close together and to provide the residents with many conveniences. Condo living can offer a great value to those interested and willing to live within its constraints and enjoy the sharing of common services.

Similarly, in cloud computing, applications run on a shared infrastructure and can gain many benefits of flexibility and cost savings. To get the most out of this arrangement, a clear model is needed for the usage pattern and constraints to be imposed in order

to empower sharing and concierge services. It is the clarity of the usage pattern that can empower new Platform as a Service (PaaS) offerings supporting the application pattern and providing services, easing the development and operations of applications complying with that pattern.

Just as there are many different ways of using buildings, there are many styles of application patterns. This article looks at a typical pattern of implementing a Software as a Service (SaaS) application and shows how, by constraining the application to this pattern, it is possible to provide many concierge services that ease the development of a cloud-based application.

Over the past 50 years, it has become increasingly common for buildings to be constructed for an expected usage pattern. Not all buildings fit this mold. Some buildings have requirements that are *so* unique, they simply need to be constructed on demand—steel mills, baseball stadiums, and even Super Walmarts are so specialized you cannot expect to find one using a real estate agent.

Such custom buildings are becoming increasingly rare, however, while more and more buildings—whether industrial parks, retail offices, or housing—are being constructed in a common fashion and with a usage pattern in mind. They are built with a clear idea of *how* they will be used but not necessarily *who* will use them. Each has standard specifications for the occupants, and the new occupants must fit into the space.

A building's usage pattern may impose constraints but, in turn, offers shared concierge services. A condominium housing development, for example, imposes constraints on parking, noise levels, and barbecuing. Residents cannot work on garage projects or gardening projects. In exchange, someone is always on hand to accept their packages and dry cleaning. They may have a shared exercise facility and pool. Somebody else fixes things when they break.



An office building may have shared bathrooms, copy rooms, and lobby. The engineering for the building is typically common for the whole structure. To get these shared benefits, tenants may have a fixed office layout, as well as some rules for usage. Normally, people cannot sleep at work, cannot have pets at work, and may even have a dress code for the building.

A retail mall provides shared engineering, parking, security, and common space. An advertising budget may benefit all the mall tenants. In exchange, there are common hours, limits on the allowable retail activities, and constraints on the appearance of each store.

Each of these building types man-

dates constraints on usage patterns and offers concierge services in exchange. To enjoy the benefits, you need to accept the constraints.

Similarly, cloud computing typically has certain constraints and, hence, can offer concierge services in return. What can the shared infrastructure do to make life easier for a sharing application? What constraints must a sharing application live within to fit into the shared cloud?

### What is Cloud Computing?

Cloud computing delivers applications as services over an intranet or the Internet. A number of terms have emerged to characterize the layers of cloud-computing solutions.

► *SaaS*. This refers to the user's ability to access an application across the Internet or intranet. SaaS has been around for years now (although the term for it is more recent). What is new is the ability to project the application over the Web without building a data center.

► *PaaS*. This is a nascent area in which higher-level application services are supplied by the vendor with two goals: first, a good PaaS can make developing an application easier; second, a good PaaS can make it easier for the cloud provider to share resources efficiently and provide concierge services to the app. Today, the leading examples of PaaS are Salesforce's Force.com<sup>5</sup> and Google's App Engine.<sup>2</sup>

► *IaaS* (infrastructure as a service). Sometimes called *utility computing*,

this is virtualized hardware and computing available over the Web. The user of an IaaS can access virtual machines (VMs) and storage to accompany them on demand.

Figure 1 shows the relationship between the cloud and SaaS providers and users. (The figure was derived from a technical report from the University of California at Berkeley, "Above the Clouds: A Berkeley View of Cloud Computing."<sup>3</sup>) As observed in "Above the Clouds," cloud computing has three new aspects: the illusion of infinite computing resources on demand; the elimination of upfront commitment by cloud users; and the ability to pay for computing resources on a short-term basis.

Cloud computing allows the deployment of SaaS—and scaling on demand—without having to build or provision a data center.

**Public and private clouds.** Clouds are about sharing. The question is whether you share *within* a company or go to a third-party provider and share *across* companies.

In a public cloud, a cloud-computing provider owns the data center. Other companies access their computing and storage for a pay-as-you-go fee. This has tremendous advantages of scale, but it is more challenging to manage the trust relationship. Trust ensures that the computing resources are available when they are needed (this could be called an SLA, or service-level agreement). In addition, there are issues of privacy trust in which the subscribing company needs to have confidence its private data will not be accessed by prying eyes. Demonstrating privacy is easier if the company owns the data center.

A public cloud can project its shared resources as VMs and low-level storage requiring the application to build on what appears to be a pool of physical machines (even though they are really virtual). This would be a public-cloud IaaS. Amazon's AWS (Amazon Web Service)<sup>1</sup> is a leading example of this.

Alternatively, in a public-cloud PaaS, higher-level abstractions can be presented to the applications that allow finer-grained multitenancy than a VM. The shape and form of these abstractions are undergoing rapid evolution. Again, Force.com and App Engine are emerging examples.

Figure 1. Cloud computing, utility computing, and software as a service.

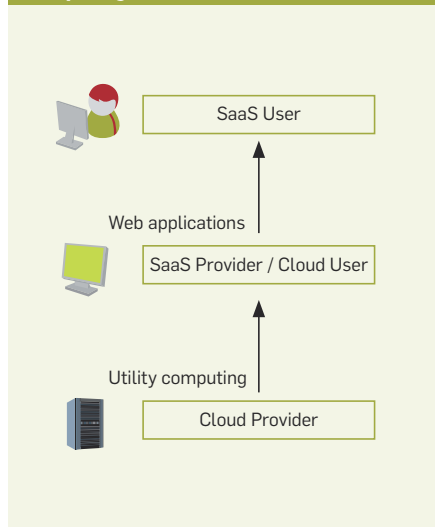
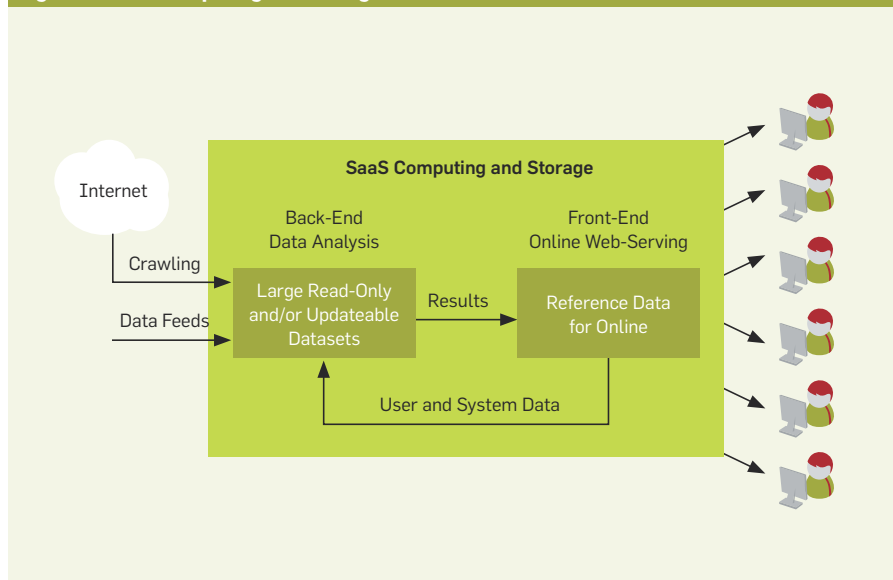


Figure 2. SaaS computing and storage.



In a private cloud the data center, physical machines, and storage are all owned by the company that uses them. Sharing happens within the company. The usage of the resources can ebb and flow as determined by different departments and applications. The size of the shared cloud is likely to be smaller than within a public cloud, which may reduce the value of the sharing. Still, it is attractive to many companies because they do not need to trust an outside cloud provider. So far, we have seen only private-cloud IaaS. The new PaaS offerings are not yet being made available for individual companies to use in their private clouds.

**Forces driving us to the cloud.** A number of forces are prompting increased movement of applications to the cloud:

*Data-center economics.* Very large data centers can offer computation, storage, and networking at a relatively cost-effective price. Power is an ever-increasing portion of data-center costs, and it can be obtained more effectively by placing the data center near inexpensive sources of electricity such as hydroelectric dams. Internet ingress and egress is less expensive near Internet main lines. Containerized servers with thousands of machines delivered in a shipping container offer lower cost for computation and storage. Shared administration of the servers offers cost savings in operations. All of this is included in the enormous price tag for the data center. Few companies can afford such a large investment. Sharing (and charging for) the large investment reduces the costs. This provides economic drive for both the cloud providers and users.

*Shared data.* Increasingly, companies are finding huge (and serendipitous) value in maintaining a “big-data” store. More and more, vast amounts of corporate data are placed into one store that can be addressed uniformly and analyzed in large computations. In many cases the value of the discoveries grows as the size of the data store increases. It is becoming a goal to store all of an enterprise’s data in a common store, allow analysis, and see surprising value.

*Shared resources.* By consolidating computation and storage into a shared cloud, it is possible to provide higher



**Power is an ever-increasing portion of data-center costs, and it can be obtained more effectively by placing the data center near inexpensive sources of electricity such as hydroelectric dams.**



utilization of these resources while maintaining strong SLAs for the higher-priority work. Low-priority work can be done during slack times while being preempted for higher-priority work during the busy times. This requires that resources are fluid and fungible so that the lower-priority work can be bumped aside and the resources re-allocated to the higher-priority work.

**SaaS: Front end, back end, and decision support.** Let’s look more closely at a typical pattern seen in a SaaS implementation. In general, the application has two major sections: the front end, which handles incoming Web requests; and the back end, which performs offline background processing to prepare the information needed by the front end. In addition to its work preparing data for the front end, the back-end application is usually shared with decision-support processing (see Figure 2).

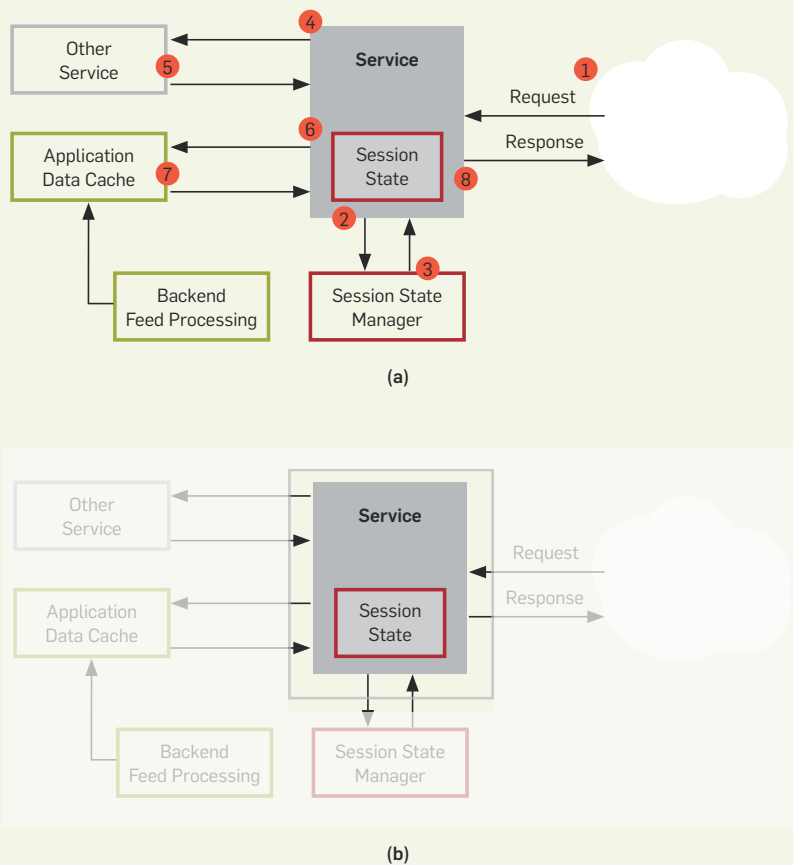
In a typical SaaS implementation the front end offers user-facing services dealing with Web services or HTML. It is normal for this Web-serving code to have aggressive SLAs, typically of only 300ms–500ms, sometimes even tighter. The back-end processing consumes crawled data, partner feeds, and logged information generated by the front end and other sources, and it generates reference data for use by the front end. You may see product catalogs and price lists as reference data, or you may see inverted search indices to support systems such as Google or Bing search. In addition to the generation of reference data, the back-end processing typically performs decision-support functions for the SaaS owner. These allow “what-if” analyses that guide the business.

### **Patterns in SaaS Apps: The Front End**

Here, I explore a common pattern used in building the front-end portion of SaaS applications. By leveraging the pattern used by these applications, a number of very useful concierge services can be supplied by the PaaS plumbing.

Many service applications fit nicely within a pattern of behavior. The goal of these applications is to implement the front end of a SaaS application. Incoming Web-service requests or HTML

**Figure 3. (a) The typical front-end SaaS application pattern. (b) The application's focus on business logic.**



requests arrive at the system and are processed with a request-response pattern using session state, other services, and cached reference data.

When a front-end application fits into the constraints of the pattern just described, a lot of *conciierge services* may be supplied to the application. These services simplify the development of the app, ease the operational challenges of the service, and facilitate sharing of cloud resources to efficiently meet SLAs defined for the applications. Some possible *conciierge services* include:

**Auto-scaling.** As the workload rises, additional servers are automatically allocated for this service. Resources are taken back when load drops.

**Auto-placement.** Deployment, migration, fault boundaries, and geographic transparency are all included. Applications are blissfully ignorant.

**Capacity planning.** This includes analysis of traffic patterns of service

usage back to incoming user workload. Trends in incoming user workload are tracked.

**Resource marketplace.** The *conciierge plumbing* automatically tracks a service's cost as it directly consumes resources and indirectly consumes them (by calling other services). This allows the cost of shared services to be attributed and charged back to the instigating work.

**A/B testing and experimentation.** The *plumbing* makes it easy to deploy a new version of a service on a subset of the traffic and compare the results with the previous version.

**Auto-caching and data distribution.** The back end of the SaaS application generates reference data (for example, product catalog and price list) for use by the front end. This data is automatically cached in a scalable way, and changes to items within the reference data are automatically distributed to the caches.

**Session-state management.** As each request with a partner is processed, it has the option to record session state that describes the work in progress for that partner. This is automatically managed so that subsequent requests can easily fetch the state to continue work. The session-state manager works with dynamically scalable and load-balanced services. It implements the application's policy for session-state survival and fault tolerance.

Each of these *conciierge services* depends on the application abiding by the constraints of the usage pattern as described for the typical front-end SaaS application.

### Stateless Request Processing

Incoming requests for a service are routed to one of many servers capable of responding to the request. At this point, no state is associated with the session present in the target server (we will get it later if needed). It is reasonable to consider this a stateless request (at least so far) and select any available server.

The *plumbing* keeps a pool of servers that can implement the service. Incoming requests are dynamically routed and load balanced. As demand increases and decreases, the *conciierge services* of the *plumbing* can automatically increase and decrease the number of servers.

### Composite request processing.

Frequently, a service calls another service to get the job done. The called service may, in turn, call another service. This composite call graph may get quite complex and very deep. Each of these services will need to complete to build the user's response. As requests come in, the work fans out, gets processed, and then is collected again. In 2007, Amazon reported that a typical request to one of its e-commerce sites resulted in more than 150 service requests.<sup>4</sup> Many SaaS applications follow the pattern shown in Figure 3a:

1. A request arrives from outside (either Web service or HTML).

2. The service optionally requests its session state to refresh its memory about the ongoing work.

3. The response comes back from the session-state manager.

4. Other services are consulted if needed.



5. The other service responds.

6. The application data cache (curated by the back-end processing) is consulted.

7. Cached reference data is returned to the service for use by its front-end app.

8. The response is issued to the caller.

**SLAs and request depth.** Requests serviced by the SaaS front end will have an SLA. A typical SLA may be a “300ms response for 99.9% of the requests assuming a traffic rate of 500 requests per second.”

It is common practice when building services to measure an SLA with a percentile (for example, 99.9%) rather than an average. Averages are much easier to engineer and deploy but will lead to user dissatisfaction because the outlying cases are typically very annoying.

### Pounding on the Services at the Bottom

To implement a systemwide SLA with a composite call graph, there is a lot of pressure on the bottom of the stack. Because the time is factored into the caller’s SLA, deeper stacks mean more pressure on the SLAs.

In many systems, the lowest-level services (such as the session-state manager and the reference-data caches) may have SLAs of 5ms–10ms 99.9% of the time. Figure 4 shows how composite call graphs can get very complex and put a lot of SLA pressure down the call stack.

### A Quick Refresher on Simple Queuing Theory

The expected response time is dependent on both the minimum response time (the response time on an empty system) and the utilization of the system. Indeed, the equation is:

$$\text{Expected Response Time} = \frac{\text{Minimum Response Time}}{1 - \text{Utilization}}$$

This makes intuitive sense. If the system is 50% busy, then the work must be done in the slack, so it takes twice the minimum time. If the system is 90% busy, then the work must get done in the 10% slack and takes 10 times the minimum time.

**Automatic provisioning to meet SLAs.** When the SLA for a service is

slipping, one answer is to reduce the utilization of the servers providing the service. This can be done by adding more servers to the server pool and spreading the work thinner.

Suppose each user-facing or externally facing service has an SLA. Also, assume the system plumbing can track the calling pattern and knows which internal services are called by the externally facing services. This means that the plumbing can know the SLA requirements of the nested internal services and track the demands on the services deep in the stack.

Given the prioritized needs and the SLAs of various externally facing services, the plumbing can increase the number of servers allocated to important services and borrow or steal from lower-priority work.

**Accessing data and state.** When a request lands into a service, it initially has no state other than what arrived with the request. It can fetch the session state and/or cached reference data if needed.

The session state provides information from previous interactions that this service had over the session. It is fetched at the beginning of a request and then

stored back with additional information as the request is completing.

Most SaaS applications use application-specific information that is prepared in the background and cached for use by the front end. Product catalog, price list, geographical information, sales quotas, and prescription drug interactions are examples of reference data. Cached reference data is accessed by key. Using the key, the services within the front end can read the data. From the front end, this data is read only. The back-end portion of the application generates changes to (or new versions of) the reference data. An example of read-only cached reference data can be seen on the Amazon.com retail site. Look at any product page for the ASIN (Amazon Standard Identification Number), a 10-character identifier usually beginning with “0” or “B.” This unique identifier is the key for all the product description you see displayed, including images.

**Managing scalable and reliable state.** The session state is keyed by a session-state ID. This ID comes in on the request and is used to fetch the state from the session-state manager.

Figure 4. Composite call graphs.

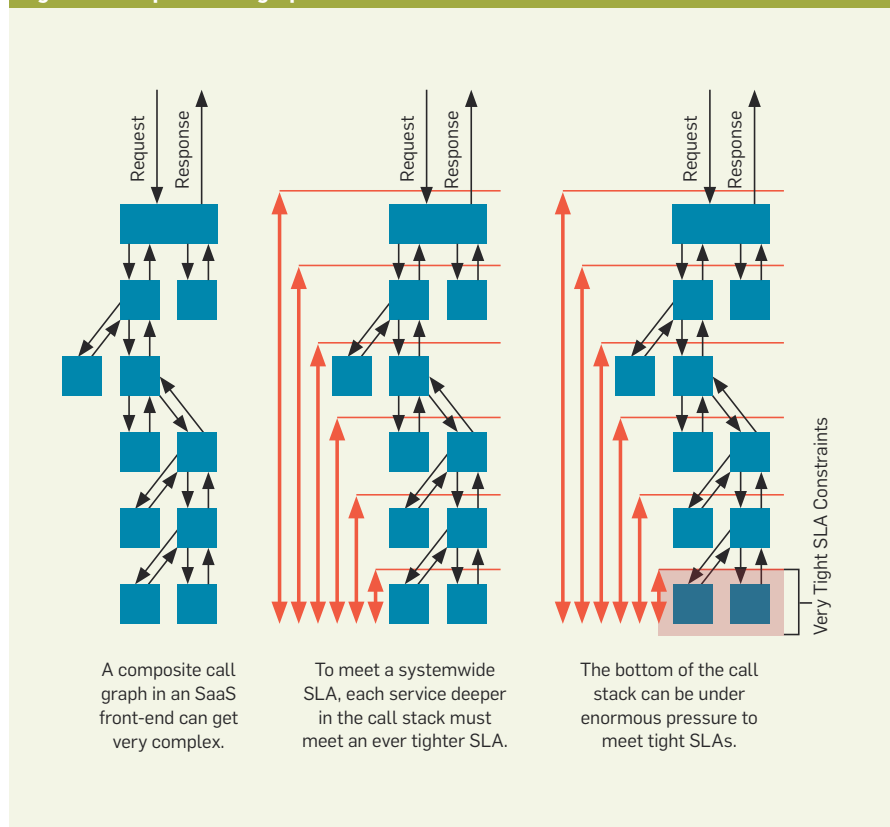
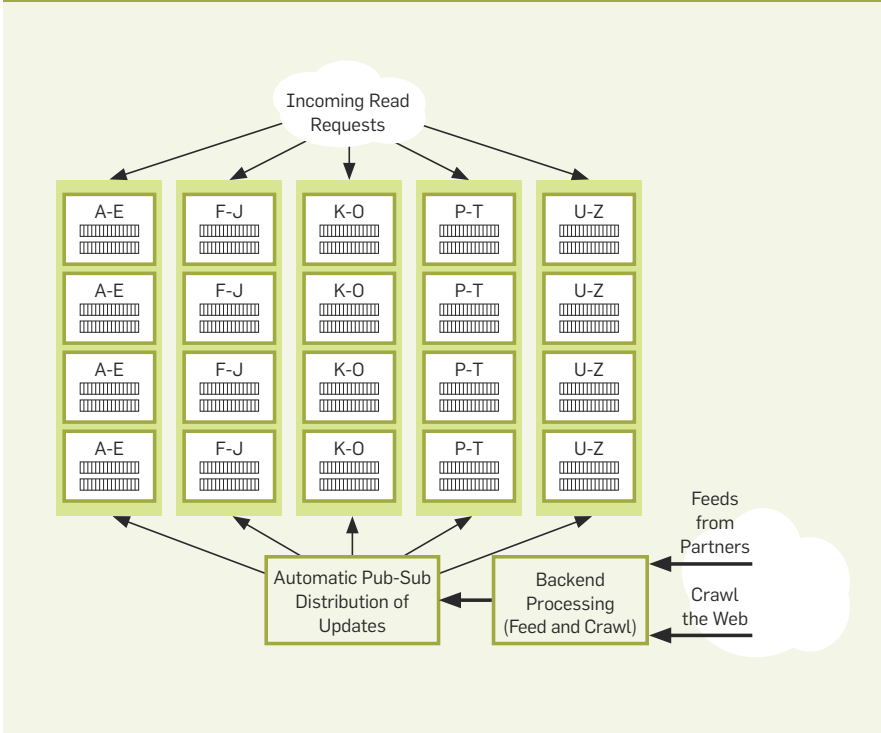


Figure 5. SaaS application interface from the back-end to the front-end.



An example of session state is a shopping cart on an e-commerce site such as Amazon.com.

The plumbing for the session-state manager handles scaling. As the number of sessions grows, the session-state manager automatically increases its capacity.

This plumbing is also responsible for the durability of the session state. Typically, the durability requirements mandate that the session state survive the failure of a single system. Should this be written to disk on a set of systems in the cloud? Should it be kept in memory over many systems to provide acceptable durability? Increased durability requires more system implementation cost and may require increased latency to ensure the request is durable.

Typically, a session-state manager is used so frequently that it must provide a very aggressive SLA for both reads and writes. A 5ms or 10ms guarantee is not unusual. This means it is not practical to wait for the session state to be recorded on disk. It is common for the session state to be acknowledged as successfully written when it is present on two or three replicas in memory. Shortly thereafter, it will likely be written to disk.

#### Applying changes to the back end.

Sometimes, front-end requests actually “do work” and apply application changes to the back end. For example, the user pushes Submit and asks for its work to be completed.

Application changes to the back end may be either synchronous, in which the user waits while the back end gets the work done and answers the request; or asynchronous, in which the work is enqueued and processed later.

Amazon.com provides an example of asynchronous back-end app changes. When the user presses Submit, a portion of the front end quickly acknowledges the receipt of the work and replies that the request has been accepted. Typically, the back end promptly processes the request, and the user receives an email message in a second or two. Occasionally, the email message takes 30 minutes or so when the asynchronous processing at the back end is busy.

#### Automatic services, state, and data.

By understanding the usage pattern of a SaaS application, the platform can lessen the work needed to develop an application and increase its benefits. As suggested in Figure 3b, the application should simply worry about its business logic and not about the system-level issues. Interfaces to call other services, access cached data, and ac-

cess session state are easy to call. This provides support for a scalable and robust SaaS application.

The application session state, application reference-data cache, and calls to other services are available as concierge services. The platform prescribes how to access these services, and the application need not know what it takes to build them. By constraining the application functionality, the platform can increase the concierge services.

### Patterns in SaaS Applications:

#### The Back End

This section explores the patterns used in the back-end portion of a typical SaaS application. What does this back end do for the application? How does it typically do it?

The back end of a SaaS application receives data from a number of sources:

- ▶ *Crawling*. Sometimes the back end has applications that look at the Internet or other systems to see what can be extracted.

- ▶ *Data feeds*. Partner companies or departments may send data to be ingested into the back-end system.

- ▶ *Logging*. Data is accumulated about the behavior of the front-end system. These logs are submitted for analysis by the back-end system.

- ▶ *Online work*. Sometimes, the front end directly calls the back end to do some work on behalf of the front-end part of the application. This may be called synchronously (while the user waits) or asynchronously.

All of these sources of data are fed into the back end where they are remembered and processed.

Many front-end applications use reference data that is periodically updated by the back end of the SaaS application. Applications are designed to deal with reference data that may be stale. The general model for processing reference data is:

1. Incoming information arrives at the back end from partners' feeds, Web crawling, or logs from system activity. Online work may also stimulate the back-end processing.

2. The application code of the back end processes the data either as batch jobs, event processing with shorter latency, or both.

3. The new entries in the reference-

data caches are distributed to the caching machines. The changes may be new versions made by batch updates or incremental updates.

4. The front-end apps read the reference-data caches. These are gradually updated, and the users of the front end see new information.

The reference-data cache is a key-value store. One easy-to-understand model for these caches has partitioned and replicated data in the cache. Each cache machine typically has an in-memory store (since disk access is too slow). The number of partitions increases as the size of the data being cached increases. The number of replicas increases initially to ensure fault tolerance and then to support increases in read traffic from the front end.

It is possible to support this pattern in the plumbing with a full concierge service. The plumbing on the back end can handle the partition for data scale (and repartitioning for growth or shrinkage). It can handle the firing up of new replicas for read-rate scale. Also, the plumbing can manage the distribution of the changes made by the back end (either as a batch or incrementally). This distribution understands partitioning, dynamic repartitioning, and the number of replicas dynamically assigned to partitions.

Figure 5 illustrates how the interface from the back end to the front end in a SaaS application is typically a key-value cache that is stuffed by the back end and read-only by the front end. This clear pattern allows for the

creation of a concierge service in a PaaS system, which eases the implementation and deployment of these applications.

Note that this is not the only scheme for dynamic management of caches. Consistent hashing (such as implemented by Dynamo,<sup>4</sup> Cassandra<sup>6</sup>, and Riak<sup>8</sup>) provides an excellent option when dealing with reference data. The consistency semantics of the somewhat stale reference data, which is read-only by the front end and updated by the back end, are a very good match. These systems have excellent self-managing characteristics.

**Styles of back-end processing.** The back-end portion of the SaaS app may be implemented in a number of different ways, largely dependent on the scale of processing required. These include:

*Relational database and normal app.* In this case, the computational approach is reasonably traditional. The data is held in a relational database, and the computation is done in a tried-and-true fashion. You may see database triggers, *N*-tier apps, or other application forms. Typically in a cloud environment, the *N*-tier or other form of application will run in a VM. This can produce the reference data needed for the front end, as well as what-if business analytics. This approach has the advantage of a relational database but scales to only a few large machines.

*Big data and MapReduce.* This approach is a set-oriented massively parallel processing solution. The underlying data is typically stored in a GFS

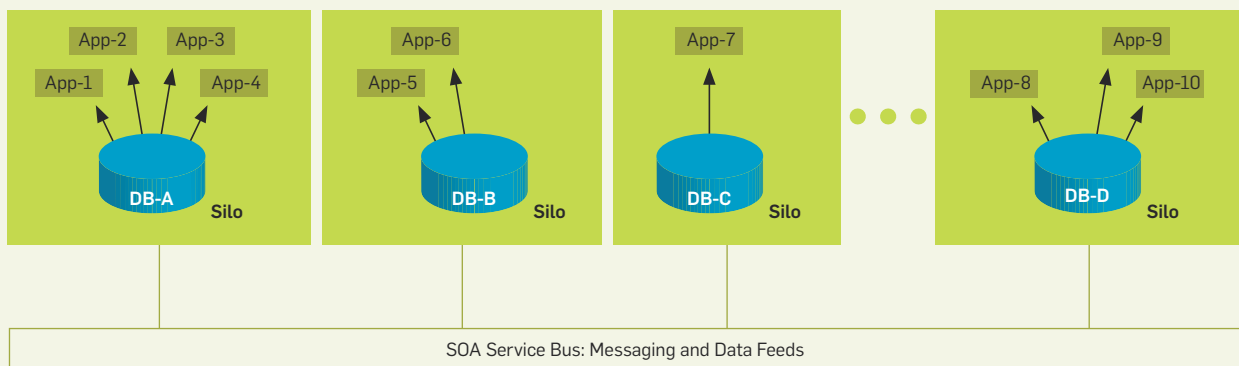
(Google File System) or HDFS (Hadoop Distributed File System), and the computation is performed by large batch jobs using MapReduce, Hadoop, or some similar technology. Increasingly, higher-level languages declaratively express the needed computation. This can be used to produce reference data and/or to perform what-if business analytics. Over time, we will see MapReduce/Hadoop over *all* of an enterprise's data.

*Big data and event processing.* The data is still kept in a scalable store such as GFS or HDFS, and batch processing is available for both the production of reference data and for what-if analysis. What is interesting here is the emergence of the ability to recognize changes from feeds or from crawling the Web within seconds and perform transactional updates to the same corpus of back-end data that is available to the MapReduce/Hadoop batch jobs. A noteworthy example of this is the Google Percolator project.<sup>7</sup> Rapid processing of events into fresh reference data offers a vibrant Web experience.

When the back-end portion of the SaaS application follows the pattern described, it is possible to create a PaaS offering that can make it much easier to build, deploy, and manage the application. By living within the constraints of the pattern, many concierge services are available such as:

*Big-data unified data access.* With unified enterprisewide (and controlled cross-enterprise) data, anything may be processed with anything (if authorized).

Figure 6. Silos and SOA.



*Fault-tolerant and scalable storage.* Cloud-managed storage for both big data and relational databases is available, with automatic intra-data-center and cross-data-center replication.

*Massively parallel batch processing.* High-level set operations perform large-scale computation.

*Event processing.* Low-latency operations with extremely high update rates for independent events are available. Transactional updates occur with more than tens of petabytes of data (see Percolator<sup>7</sup>). Event processing over the same corpus of data is available for batch processing.

*Automatically scaling reference-data caches.* The back end supplies the front end with dynamically updated application-specific data. The management and operations of the key-value cache are automatic. The plumbing ensures read SLAs by the front end are maintained by increasing replication of the caches as necessary.


*Multitenanted access control.* Both inter-enterprise (public cloud) and intra-enterprise (private cloud) require access control to the data contained within the shared stores.

*Prioritized SLA-driven resource management.* Relational databases, big-data batch, and big-data event processing compete for the same computational resources. Different organizations have different workloads, all competing for resources. The various workloads are given an SLA and a priority, and the trade-offs are then automated by the plumbing.


## The Drive Toward Commonality in Computing

Just as buildings have commonality, there is commonality within different classes of computing environments. The different classes must hook up to the community, just as happens in buildings. Increasingly, the cloud environment is facilitating these changes.

Today's enterprises glue together silos of applications using SOA (service-oriented architecture). Silos are collections of applications that work on a shared database. SOA is the application integration that subsumes both EAI (enterprise application integration) and B2B (business-to-business) communication using messaging and data feeds. Figure 6 shows



**As the pieces of the enterprise's computing move to common hosting in a private or public cloud, the behavior of the applications, interactions across them, and usage patterns of the apps can be tracked.**



how applications with shared traits will congregate in the SOA service bus. Current enterprise applications will leverage their commonality as they migrate to the cloud. Certainly, existing applications that have not been designed for cloud deployment cannot get all the same benefits as those that follow the new and emerging patterns. Still, there is a way of leveraging the existing commonality and gain new benefits via cloud deployment, within either a public or private cloud.

Cloud computing will help drive silos and SOA to a common representation:

*Relational databases.* Relational databases will be supported on standardized servers with standardized storage running over SANs (storage area networks). This will allow robust storage to underlie restartable database servers.

*Existing apps using VMs.* Because existing applications have their own expectations of the processing environment, those expectations are most easily met in a general fashion using a VM. The VM looks just like a physical machine to the application software and yet can be managed within the cloud's cluster. The application may get one machine or a pool of machines.

*SOA messaging through the big-data store.* The large cloud-based store can receive the messages and data feeds used to connect the enterprise's silos. By plumbing the SOA service bus through the big-data store, the information about the interaction of the silos is available for analysis. This class of analysis has been shown to be invaluable in the insight it brings to the enterprise.

*Standardized monitoring and analysis.* As the pieces of the enterprise's computing move to common hosting in a private or public cloud, the behavior of the applications, interactions across them, and usage patterns of the apps can be tracked. Capturing this in the big-data store allows for analysis in a fashion integrated with the application's messaging data and other enterprise knowledge.

**Database and big data.** Relational databases offer tightly controlled transactions and full relational semantics. They do, however, face chal-

lenges scaling beyond a handful of servers. Still, relational databases have more than 30 years of investment in applications, operations, and skills development that will survive for many years.

The emerging big-data stores as represented by MapReduce and Hadoop offer complementary sets of advantages. Leveraging massive file systems with highly available replicated data, these environments offer hundreds of petabytes of data that may be addressed in a common namespace. Recently, updatable key-value stores have emerged that offer transaction-protected updates.<sup>7</sup> These enormous systems are optimized for sharing with multiple users accessing both computational and storage resources in a prioritized fashion.

Increasingly, these benefits of the big-data environments will be applied to copies of the relational database data used within existing applications. The integration of the line-of-business relational data with the rest of the enterprise's data will result in a common backplane for data.

The line-of-business department in an enterprise drives the development of new applications. This is the department that needs the application and the solutions it provides to meet a business requirement. The department funds the application and, typically, is not too concerned with how the application will fit into the rest of the enterprise's computational work.

The IT department, on the other hand, has to deal with and operate the application once it is deployed. It wants the application, database, and servers to be on common ground. It needs to integrate the application into enterprise-wide monitoring and management.

This natural tension is similar to that seen between property developers and the city planning commission. Developers want to construct and sell buildings and are not too fussy about the quality. The city planners have to ensure the developers consider issues such as neighborhood plans and whether the sewage-treatment plant has enough capacity.

As we move to cloud-computing environments in which the application, database management system, and other computing resources are hosted

on a common collection of servers, we will see an increase in the standards and expectations over how they will tie to the enterprise.

What about the forces driving us to the cloud? The forces will have their way as they focus on a common and shared basis for computation and storage. First, the cost savings from cheaper electricity, networking, and server technology that is available in concentrated and expensive data centers will continue. While there are issues of managing trust, the economics are a powerful force. Second, enterprises will continue to find serendipitous value in the analysis of data. The more data available for analysis, the more often valuable surprises will happen. Finally, both computational and data-storage resources will be increasingly shared across applications and enterprises. Existing applications will get some sharing value and new cloud-aware applications even more.

These forces will encourage applications to work together atop new abstractions for sharing. When applications stick to their old models, they will get some advantages of the commonality of the cloud but not as many as those seen by new applications.

## Conclusion

The constraints in an environment are what empower the services. The usage pattern allows for supporting infrastructure and concierge services.

Shared buildings become successful by constraining and standardizing their usage. Building designers know *how* a building will be used, even if they do not know *who* will be using it. Not everyone can accept the constraints, but for those who do, there are wonderful advantages and services.

The standardization of usage for computational work will empower the migration of work to the shared cloud. With these usage patterns, supporting services can dramatically lower the barriers to developing and deploying applications in the cloud. Lower-level standards are emerging with VMs. These support a broad range of applications with less flexibility for sharing. Higher-level PaaS solutions are nascent but offer many advantages.

We must define and constrain the

usage models for important types of cloud applications. This will permit enhanced sharing of resources with important supporting services. The new PaaS offerings will bring tremendous value to the computing world. □

## Related articles on queue.acm.org

### Commentary: A Trip Without a Roadmap

Peter Christy

<http://queue.acm.org/detail.cfm?id=1515746>

### Fighting Physics: A Tough Battle

Jonathan M. Smith

<http://queue.acm.org/detail.cfm?id=1530063>

### CTO Roundtable: Cloud Computing

January 10, 2009

<http://queue.acm.org/detail.cfm?id=1551646>

## References

1. Amazon Web Services; <http://aws.amazon.com/>.
2. App Engine; <http://code.google.com/appengine/>.
3. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I. and Zaharia, M. Above the clouds: A Berkeley view of cloud computing; <http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf>.
4. DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P. and Vogels, W. Dynamo: Amazon's highly available key-value store. *ACM Symposium on Operating Systems Principles*; <http://www.allthingsdistributed.com/files/amazon-dynamo-sosp2007.pdf>.
5. Force.com; <http://www.force.com/>.
6. Lakshman, A. and Malik, P. Cassandra—A decentralized structured storage system. *Large-scale Distributed Systems and Middleware*; <http://www.cs.cornell.edu/projects/ladis2009/papers/lakshman-ladis2009.pdf>.
7. Peng, D. and Dabek, F. Large-scale incremental processing using distributed transactions and notifications. In *Proceedings of the 9th Usenix Symposium on Operating Systems Design and Implementation* (2010); <http://research.google.com/pubs/pub36726.html>.
8. Riak; <http://basho.com/products/riak-overview/>

**Pat Helland** has worked in distributed systems, transaction processing, databases, and similar areas since 1978. He was the chief architect of Tandem Computers' TMF (Transaction Monitoring Facility), which provided distributed transactions for the NonStop System. At Microsoft, he served as chief architect for Microsoft Transaction Server, SQL Service Broker, and a number of features within Cosmos, the distributed computation and storage system underlying Bing. He recently joined Salesforce.com and is leading a number of new efforts working with very large-scale data.

This paper was written before Helland joined Salesforce.com and, while there are many similarities, this is not intended to be a description of Salesforce's architecture.

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

**A discussion with Jeremiah Grossman,  
Ben Livshits, Rebecca Bace,  
and George Neville-Neil.**

# Browser Security: Appearances Can Be Deceiving

IT SEEMS EVERY day we learn of some new security breach. It is all there for the taking on the Internet—more and more sensitive data every second. As for privacy, we Facebook, we Google, we bank online, we shop online, we invest online...we put it all out there. And just how well protected is all that personally identifiable information? Not very.

The browser is our most important connection to the Web, and our first line of defense. But have the browser vendors kept up their end of the bargain in protecting users? They claim to have done so in various ways, but many of those claims are thin. From SSL (Secure Sockets Layer) to the Do Not Track initiative to browser add-ons to HTML5, attempts

to beef up security and privacy safeguards have fallen well short.

For example, many experts dismiss the notion that the most widely used protocol for providing security over the Internet, the SSL CA (certificate authority) model, actually provides adequate transport-layer security. But for all its faults, there is much resistance among vendors to changing the model.

HTML5 is waiting in the wings, viewed by many as the next step toward improving the Web experience, while retaining compatibility with existing browsers. It has been put forth with great promise, but so far it has not adequately addressed security shortcomings.

Vendors have attempted to achieve better browser security by supplying add-ons for protection, but users first must know where to find, and then download, install, and configure them. That is a lot to ask. It also means first being aware of the dangers—many businesses have never heard of cross-site request forgery or clickjacking and most users have no idea just how exposed their personal information really is. This is not an easy message to get across.

Likewise, users must be proactive to derive any protection from the Do Not Track initiative, a means of requesting Internet companies to stop following a user's every move. Though endorsed by the W3C and the Federal Trade Commission, it, too, falls short by putting the burden on generally uninformed users to opt in rather than making it a default setting.

For this case study on browser security ACM has assembled an experienced group to break down some of the mythical claims of security in today's browsers and argue the case for increased protection.

**Jeremiah Grossman** is founder and CTO at WhiteHat Security, a leading provider of Web application security services, including Sentinel, a website vulnerability management solution. A founding member of WASC (Web Application Security Consortium), he is



“

JEREMIAH GROSSMAN

Is there still anyone out there who seriously believes the CA model works? It's completely broken.

”

sought after for his expertise in Web application security. Prior to White-Hat, he was an information security officer at Yahoo!.

**Ben Livshits** is a researcher at Microsoft Research and an affiliate professor at the University of Washington. He has been focusing on improving Web 2.0 application and browser reliability, performance, and security.

Security technology expert **Rebecca Gurley Bace** is president/CEO of Infidel, a network security consulting practice, and chief strategist for the Center for Forensics, Information Technology, and Security at the University of South Alabama. Her career has included a decade overseeing security investments, founding roles in several IT security communities, and advisory roles in a number of successful security ventures, both in the public and private sectors. Previously, Bace was a senior electronics engineer at National Security Agency (NSA) and served as a charter member of NSA's Information Security (Infosec) Research and Technology Group. She left NSA to become the deputy security officer for the computing, information, and communications division of the Los Alamos National Laboratory.

Facilitating the discussion is **George Neville-Neil**, a software engineer who

builds high-speed, low-latency systems for customers in the financial-services sector. Previously, he was part of the Yahoo! Paranoids security team. From 2004 to 2008, Neville-Neil worked in Japan, where he developed a set of courses dubbed “The Paranoid University,” teaching safe and secure programming to engineers at Yahoo!. For the past 10 years he has served on the *ACM Queue* editorial board and more recently he joined the ACM Practitioner Board.

**GEORGE NEVILLE-NEIL:** In talking about the current SSL CA model, Jeremiah, you have commented previously that no SSL feature ever gets turned off. What would it mean exactly to turn off something from the CA model?

**JEREMIAH GROSSMAN:** Many security experts, including myself, consider Convergence viable and believe it should replace the CA model as soon as possible, since what we currently have clearly isn't working. But there hasn't been general acceptance of that yet. Beyond acceptance, the bigger challenge would be to manage the migration to Convergence. Let's say we just add it alongside the CA model. At what point would we turn off the CA model? It's only by doing so, after all, that we would actually realize the security benefits of Convergence. Otherwise, as with the wait to cut over to IPv6, ev-



BEN LIVSHITS

**There is a lot of fear that any real enforcement of Do Not Track might end up destroying the fundamental revenue model for the Web economy. I don't think we are likely to see enforcement in any form anytime soon.**



everyone would just continue to be stuck with the same old mess as before.

**GN-N:** How would that actually work?

**JG:** CAs would be converted into notaries, and then the browser user would choose which notaries to trust. If any of those notaries were to become untrustworthy for any reason, the user could easily remove the trust indicated for that particular notary. That's very important because in the current CA model it's very difficult—if not impossible—to withdraw trust from any one CA without breaking the Web, which makes things very challenging.

One of the major criticisms [computer security researcher] Moxie Marlinspike (a pseudonym) has raised about the CA model has to do with this lack of trust agility. That is, whomever we decide to trust, we're then obliged to trust forever. Still, Moxie and the team responsible for introducing the Convergence plugin say they have taken the idea about as far as they can, and the browser vendors now need to take it the rest of the way, but the browser vendors seem pretty disinterested.

**GN-N:** The biggest problem with the Convergence model is that it trusts the user to do the right thing, but most users will just do whatever they're told.

**JG:** Maybe this is naive on my part, but I think users have a pretty good

idea of whom they trust and whom they would like to trust and whom they know they're not about to trust.

Convergence also offers flexibility. I can trust five notaries today and then change to five different ones tomorrow. I would be able to do that without a whole lot of technical know-how.

Still, there are two major challenges facing Convergence. The first has to do with getting the browser vendors to implement it, and frankly, they just don't seem to have a lot of incentive for doing that. Just for the sake of argument, however, let's say they did. The next challenge would be to get people to run the notaries. That's a pretty big challenge since there is no obvious business model—which is to say there is no way for anybody to make any money. So, achieving a critical mass of notaries is going to be really difficult.

All that being said, is there still anyone out there who seriously believes the CA model works? It's completely broken.

**REBECCA BACE:** Even in the earliest days of the certificate model, there was a lot of criticism that it had been blindly adopted from an archaic paper-driven DoD model, without really thinking things through from a technology perspective.

**JG:** During a presentation on authentication, Moxie said he had located the



person most directly responsible for the browser SSL CA model as we know it, and that guy told him, “Oh yeah, the CA model... we just threw that in at the end. We really had no idea.”

So why do the browser vendors hold onto this obviously outmoded CA model, while making it obvious they don't want to help out with Convergence despite all the community support for that?

**GN-N:** It's probably because moving to Convergence would represent more work on their part. That's usually why people resist doing something.

Anyway, are the implementers going to have to worry about it, or are they just going to wait for the browser vendors to create it?

**JG:** As I understand the Convergence spec, the 1.8 million websites that currently have SSL enabled should not have to do anything, since the idea is for everything to work exactly as it currently does. Everything should happen over on the browser and the notary side. We should be able to carry forward the CA model through an interim period, but we would also need to have 20 or 100 notaries set up at different organizations, and the browsers would need to support that.

**GN-N:** So far, we have talked about protection. Let's look now at what is happening over on the attack side.

**JG:** I caused a bit of a furor at a conference a few years ago by talking about intranet hacking. What I meant is that you can go to a website and use it to force your browser to make basically any type of Web request of any location you want. We generally refer to that now as cross-site request forgery, but until 2006, no one had really thought about that. People knew, of course, that you could force your browser to make a request of any public website, but then Robert Hampton and I made the observation that you could force your browser to make a request of an RFC-1918 network, such as a 10.0.0.1, and then just start hacking the intranet.

We showed how you could go to a public website and force your browser to hack into your own DSL router from the inside and then move out to the Web interface and change the settings. Normally, devices on the intranet don't have very good Web security because of the understanding that you can't hack

them from the outside, which is true. But at the same time, there is nothing to prevent the browser itself from being used as an attack platform by bad guys on the outside.

I've asked various browser vendors the following question: “If I'm on a public website, why do you allow that site to force my browser to make RFC-1918 requests?” They usually raise two points in response. One is that to do otherwise might mess up certain proxy configurations—I'm not sure what they mean by that. The other point is that sometimes there's actually a legitimate use case—that is, some corporate public websites have actually referenced various resources for the benefit of their employees on RFC-1918 networks. So basically, the argument is that because some big companies have adopted some really stupid practices, the rest of us have to live with compromised security on the Internet.

**GN-N:** Somehow I doubt they would frame it in quite that way, however.

**JG:** The browser vendors are just not willing to do anything that is going to disrupt the Web because of their concerns about market share. Any feature that might break some tiny portion of the Web and lose 1% of their market share is something they are just not going to consider. This is where it's useful to remember that we, the folks who use these browsers, are not really considered to be the customers. Instead, we are the product—or at least the data related to our online behavior is the product.

**GN-N:** On the privacy front, there seem to be some stirrings now to challenge the status quo. What are your thoughts about the Do Not Track initiative, which has been promoted by the U.S. Federal Trade Commission?

**JG:** Basically, that amounts to a header that browsers can pass along to websites saying, “Please do not track this user.” It effectively puts websites on the honor system where tracking is concerned.

**GN-N:** Are you saying this is kind of like robots.txt, only in the opposite direction?

**BEN LIVSHITS:** Maybe something similar to that.

**JG:** There are no criminal sanctions to back it up, so any enforcement will have to come in the form of civil suits

once the initiative really starts to get adopted. Google was the last holdout in terms of providing browser support for it, but didn't commit to any particular date.

The other challenge is that there's no clear definition of what it means to “not track” someone. Some have taken that to mean they can track you but not advertise to you.

**BL:** It's very easy for browser vendors to implement this as a feature. Some people will then choose to turn it on, but probably not a very high percentage if the feature isn't on by default. Even if they do decide to turn on Do Not Track and are able to figure out how to do that, they still have to make sense of what it even is. What about a site that authenticates the user? Is that site not also allowed to track the user? That would be kind of ridiculous—a contradiction in terms. What about online merchants? They have to track things just to make sure your order gets delivered, right?

What's really odd is that we have browser support for this thing that's likely to become available soon pretty much across the board, and yet there's still no consensus on what it even means.

**JG:** The only place where it actually makes sense to tell the user about Do Not Track is at the browser level, and the browser guys are completely disinclined to do anything of the sort. To Ben's point, if you look at all the implementations to date, you'll find that Do Not Track is turned off in every last one of them by default and buried three clicks deep where no one is ever going to find it. There is one notable and very controversial exception: Internet Explorer 10, which effectively installs with Do Not Track enabled.

**BL:** There's also a lot of fear that any real enforcement of Do Not Track might end up destroying the fundamental revenue model for the Web economy. I don't think we are likely to see enforcement in any form anytime soon.

**JG:** It's hard to imagine how they're going to be able to enforce this in any event. How would I, as a user, find out someone had been following me around in violation of Do Not Track? How would you ever discover that?

**RB:** My own curmudgeon's view is that this is a classic example of what happens all too often when policy-

makers decide to issue some dictum just because it seems like a good idea. Then, the technology solution providers readily agree, knowing full well that the new policy will be totally unenforceable. The policy becomes nothing more than window-dressing for the industry.

Data is money, and that goes to the core of the browser-security debate. Browser users do not fully appreciate the value of their own data, but the Facebooks and Googles of the world certainly do. Introducing measures to help users protect their data gets in the way of milking that data for all its worth. That is a strong disincentive for implementing strong browser privacy protection measures.

Adding stronger security also comes with a trade-off—more security usually means less functionality. With loss of functionality comes loss of market share, which vendors fear more than anything.

Only when users begin to see the value of their data and demand more protection for it will privacy measures get their due. If the market shifts in this direction and vendors see that adding better protection to their browsers could actually increase market share, then and only then will those measures become standard operating practice.

**GN-N:** We talked a little earlier about how it's the browser users, rather than the browsers themselves, that are the real products here. Anyone care to expand?

**RB:** Well, that is the case, and it's fundamental to this whole space. I would argue that every last conundrum in the area of browser security is rooted in the fact that we are not dealing with a classic commercial model. That is, at present users don't pay browser makers for software or, for that matter, the maintenance and upkeep of that software.

**JG:** The browser makers are monetizing your data, directly or indirectly, and therefore cannot see a way to protect that data without losing money. That makes for a really difficult situation.

**BL:** I'm not sure you can actually say it's the browser makers who are "monetizing your data." If anything, it's the sites that are monetizing your data.

**JG:** Actually, there is a clear interplay there. Just look at Google Chrome; it's pretty obviously monetizing your data. The Mozilla guys derive 98% of their revenue directly from Google. Then you've got Microsoft, which you could argue is also desperate now to get into the advertising business. So that raises the question: How can you work to institute healthier business incentives when those efforts are so obviously at odds with the foundation the whole business sits upon?

**BL:** I don't know. One of the problems with privacy is that it is difficult to put a value on it. It's difficult even to convince the users that their own privacy is actually worth all that much.

**JG:** Maybe users just aren't all that aware of what they're giving up with every single mouse click.

**BL:** Right, but there are a few companies such as Allow (<http://i-allow.com>) that will sign you up quite explicitly for \$20 to \$50 for each site you're willing to share your information with. There also are various experiments under way to establish the value of each Facebook "Like," for example. They are finding that, while some users' information is quite valuable, there are many others whose information is largely useless.

**RB:** I think this question rides a bigger value wave where the age dynamic comes into play. It's hard to find anybody under the age of, say, 25 who really cares about privacy. My young nieces happily tell me they have never felt like they had any privacy to begin with, so why should they start caring now?

**GN-N:** You have also got those people who lived through the 1960s and 1970s when the stories were rampant about people having their data exposed by the government. There are plenty of people that age who have just become inured to privacy violations. They might have cared at one point in their lives, but they're over that now.

**JG:** Another aspect of this is that security and privacy have become conflated. For example, if you have decided you can trust Google with your data, then the question is no longer about privacy; it's all about security. On the other hand, if you don't trust your provider, you can distinguish between security and privacy.

Once you cross that threshold and decide to trust someone with your data, you're in kind of the same situation we were talking about earlier with regard to the CA model. That is, you're essentially stuck with trusting them forever. It's not like you can take back your data from Facebook and say, "Hey, you're not allowed to have that anymore."

**GN-N:** Yeah, just try!

**JG:** You can get a copy of your data—and, according to [WikiLeaks'] Julian Assange, that can literally run to 1,000 pages. But, guess what, I don't think they are going to delete that information.

**RB:** Violations of our trust are already common occurrences even in the holy of holies—namely, the healthcare space, where you'd like to believe the protection of personal data would be considered sacrosanct. If people's trust isn't being honored in that domain, what hope can we hold out for more faithful protection anywhere else?

**JG:** That's why the fact that Do Not Track is off by default really bothers me. By the time users figure out what it is they've given up, there's no way to undo the damage or to take back any degree of control. As [computer security specialist] Bruce Schneier once pointed out, there's no delete button in the cloud, or at least there's no guarantee that, once you've pressed delete, things are actually going to be deleted.

**GN-N:** Is there any cause for hope?

**JG:** I have a pretty good strategy for protecting my own data—at least it's good enough to improve my level of comfort. I think it's an approach other people could use. The challenge is that it takes some behavioral discipline and a bit of know-how, both of which are lacking for most users. There has also been little motivation for people to work on cleaning up their acts since, for the most part, they're not even aware of the issues we've been talking about. Still, I'd say there is some reason for hope in that there are steps you can take to protect yourself.

**GN-N:** Do you see the browser vendors helping matters at all?

**JG:** No. To give you an example: since I really don't like the whole SSL model, I've put SSL VPNs (virtual private networks) on the Amazon cloud so that, no matter where I am, I can be encrypted over a hostile or untrusted network



REBECCA BACE

**It's hard to find anybody under the age of, say, 25 who really cares about privacy. My young nieces happily tell me they have never felt like they had any privacy to begin with, so why should they start caring now?**



while also making sure no one is able to sniff on me over the last mile. That is just one small thing you can do. It's not something my mom would be able to do, but any techie certainly could handle it.

**RB:** I've had a long-running debate with [risk management specialist] Dan Geer about when people might start offering the functional equivalent of gated communities on the Internet, where you would be able to buy into a managed security environment with a ready-made Internet safety barrier capable of protecting you from breaches of privacy or revelations of personal information.

**JG:** Geer says it's not so much about who's at fault for the current mess but instead who's going to take responsibility for it. If you say, "The user is the one who ought to take responsibility"—which is kind of where we are today—well, that just doesn't work all that well, does it?

So you might say, "OK, the ISP should take responsibility for all this bad traffic," but then you're going to have to let the ISP monitor, log, and analyze all your traffic down to a very detailed degree. You could ask the government to handle the mess, but it would need that same detailed level of access to the data and so would need

to establish new powers and laws to provide for that. None of those options seems particularly attractive.

HTML5 may not be perfect, but it is inevitable and will soon be a part of all modern browsers. It adds features, particularly multimedia functions, and is meant to make the browser a richer environment. That's what Web developers want because it could lead to increased market share.

What HTML5 does not do particularly well is add security to the browser. It also leaves the door open to some Internet attacks. Many security experts thus have come to see HTML5 as an inexcusable missed opportunity. Any security work-arounds will have to be made separate from HTML5. So, yes, it's new, it's improved, but it's not going to save us.

**GN-N:** HTML5 is now with us, and some people have probably been hoping that would bring some relief on the security front. Any comments?

**JG:** The whole idea of HTML5 was to bring richer media to the browser—all native through an open standard—so you wouldn't need to add plugins such as Flash, QuickTime, and all



GEORGE NEVILLE-NEIL

**Stored procedures are a clever idea on a database, but they are a terrifying idea in a client.**



kinds of other crazy stuff. That's huge because plugins have proved to be major sources of security vulnerabilities. The missed opportunity, though, is that HTML5 fails to address some long-standing Web security issues such as cross-site scripting, clickjacking, and cross-site request forgery. HTML5 developers just sort of punted on all that.

Then they added the sandbox tag as a kind of Band-Aid to be able to say they had done their bit to provide for Web security. I could go on. There are many examples of how I think HTML5 is going to make browser security much worse.

**GN-N:** My experience with cross-site scripting and cross-site referral forgery has been that the only real way to deal with it is to handle it on the server. This generally means drilling into the heads of the people who are using the server-side code that what they need to be doing is to make sure those exploits don't happen again.

Clickjacking is something else altogether. Right now it's probably the exploit most likely to pay off in a big way for the bad guys, whereas cross-site scripting and cross-site referral hijacking are more what you would expect from someone who is just trying to cause trouble.

Most of Facebook's security effort is expended on preventing clickjacking, and it's certainly not alone in that. In fact, I think that's really the new frontier, and I don't think HTML5 is going to address that.

**JG:** That could have been addressed, but as it stands, HTML5 has no security model for safely incorporating third-party data or code into your website. That model is supposed to come later with something called "cross-site security policy" or "cross-site content security policy." Even then, it will still be separate from HTML5.

As for Facebook, clickjacking is only an issue because Facebook is looking to track you around the Web. That aspect of clickjacking is going to remain unfixable since what Facebook really wants is to put Like buttons on everybody's pages. You can always clickjack something that's meant to be framed. On its own website, Facebook has already more or less fixed the clickjacking problem.

**GN-N:** Of course, it's not just Facebook that's looking to put some sort of button everywhere.

**JG:** That's right, and that's why one of the briefings at the most recent Blue-Hat conference described a new solution that involves putting anti-click-

jacking stuff in the browser. But, again, all that work is separate from HTML5.

**GN-N:** What else concerns you about HTML5?

**JG:** Are you familiar with the use of session storage as an alternative to cookies? Basically, some Web programmers are starting to put actual executable JavaScript code into local storage in addition to data. That way, when the page loads, they can just eval that code directly rather than having to make a network call, because that gets them a performance win.

Of course, the bad guys find this attractive. If they cross-site script the site that loaded that code, they'll be able to backdoor the application and thus enjoy permanent access to any client that thing happens to get loaded onto, since that backdoor code will always run.

**GN-N:** Stored procedures are a clever idea on a database, but they are a terrifying idea in a client.

**JG:** Even once you become aware of the exploit, backing out of it will be all but impossible. You certainly wouldn't be able to override it from the server. So while the HTML5 guys will say they haven't increased the attack surface, I don't think they actually know yet what all the implications are going to be.

**GN-N:** This would really simplify the distribution of something that looks an awful lot like a virus.

**JG:** It really does, but that isn't obvious yet since use of HTML5 in that way still isn't particularly widespread. Give it a few years, though, and it will be everywhere, because it really is a lot faster.

**GN-N:** This tells me that the browser vendors ought to include a feature that lets you flush an application's program space—perhaps not from the server, but the user ought to be able at least to flush a bad application. And now I'm suddenly picturing virus scanners that run in your browser.

**JG:** Oh, yeah, that's definitely going to be the case.

**BL:** Even then, ensuring data integrity is not going to be easy. If you have complex data structures, who's to say some of those haven't been affected in some subtle ways?

**JG:** I think what the browser vendors have done—knowingly or unknowingly—is to turn the browser into a new operating system.

**GN-N:** Well, Chrome isn't called Google Chrome OS for nothing, you know.

**JG:** That's right. Actually, within that sandbox there's not all that much security buffer between applications.

**GN-N:** We keep ripping on HTML5, but is there anything people might be able to do to provide for a better and safer user experience?

**JG:** Well, let's be clear: if you are using any modern browser, you are going to end up using HTML5. There's no way to turn it off in your browser since it's not a feature. It's HTML. You can't turn off HTML in the browser.

**GN-N:** I wasn't actually thinking in terms of turning off HTML5, although it's an interesting notion. In any event, I don't think the typical user ever turns off anything. It's up to the client and server application developers to build things in such a way that, even in the face of a wide-open browser, the user won't end up getting abused constantly.

**JG:** I can share how I try to protect myself and how I've instructed my mom to do it. Take two browsers—any modern browsers that have been updated will do. The important thing is to have two of them so you can compartmentalize risk. The first of these will be the primary browser, the one you use for all your promiscuous browsing—read the news, visit your favorite websites, click on the links in your Twitter feed, and whatever else you feel tempted to do. But don't ever use the primary browser to do anything with online accounts you consider sensitive or important.

If you're using Chrome or Firefox, you should also turn on ad blocking and tracker blocking as extensions in the browser. That's not just for sanity purposes, but also to prevent a whole lot of malware, which often ends up getting propagated over advertising networks. Bonus points if you run in incognito or private mode. That might save you a little bit of privacy as well. Another thing you should do is to block plugins from playing by default. You can run them whenever you want to with a right click, but don't let them automatically run. Generally, when you get infected with a virus or a piece of malware, it's because of some invisible plugin that runs automatically.

Your secondary browser is the one you want to fire up only when it's time

to do online banking or online shopping or anything involving a credit card number, an account number, or anything else you want to protect. Once you have fired up that browser, get in and do what you need to do quickly, and then close that thing down.

If you can manage to keep those two worlds separate, when you are out surfing the Web with your primary browser, it won't even be possible to hack your bank with a cross-site request forgery request because it will be like you've never logged in at that bank. So clickjacking, cross-site request forgery, and cross-site scripting pose almost no threat, since there effectively is no cross site.

**GN-N:** What advice do you have for Web developers?

**BL:** I think CSP (content security policy) and the sandbox tag are among the best things for security-conscious Web developers to have come along in a long time.

**JG:** Also, of course, Web developers would be well advised to pay special attention to input validation, parameterized SQL statements, and output filtering. That covers about 90% of website vulnerabilities.

If you were to talk to the Facebook guys or even the Microsoft guys, you would find they usually have standard controls and libraries for printing the screen. By extension, that means removing all the nonstandard options—some of which might be unsafe—so people have no choice but to use the corporate standard version.

Then I guess the other thing is: don't ever try to roll your own crypto.

**GN-N:** That's solid advice. If you're not a cryptographer, don't try that at home.

#### Related articles on [queue.acm.org](http://queue.acm.org)

##### **Building Secure Web Applications**

*George V. Neville-Neil*

<http://queue.acm.org/detail.cfm?id=1281889>

##### **CTO Roundtable:**

##### **Malware Defense Overview**

*Mache Creeger*

<http://queue.acm.org/detail.cfm?id=1734092>

##### **Java Security Architecture Revisited**

*Li Gong*

<http://queue.acm.org/detail.cfm?id=2034639>

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

**Unless you have taken very particular precautions, assume every website you visit knows exactly who you are.**

BY JEREMIAH GROSSMAN

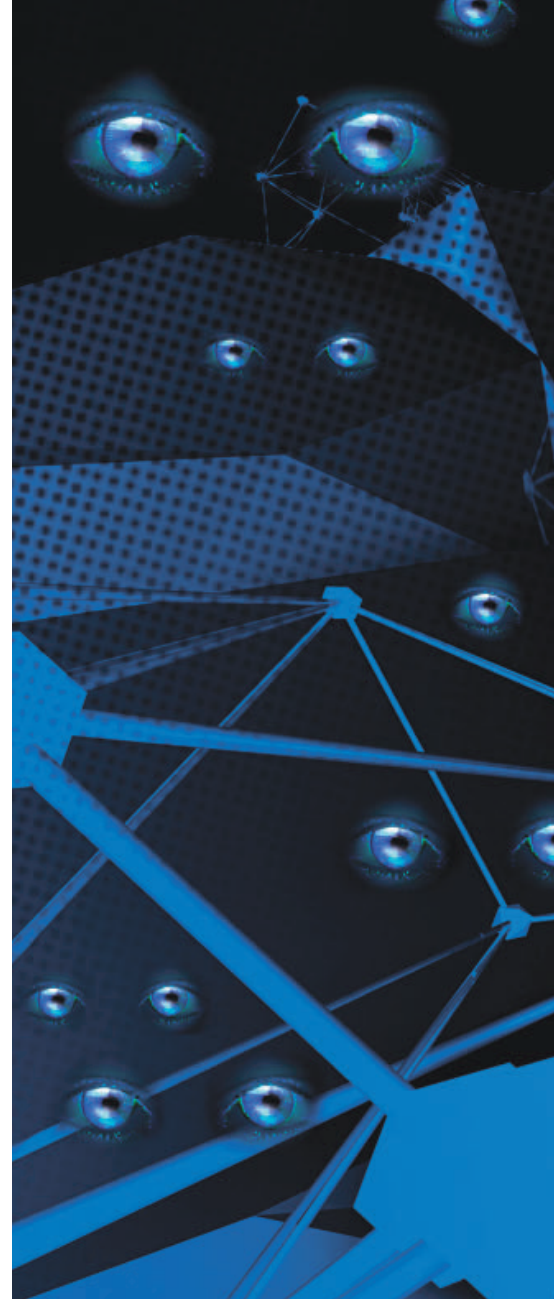
# The Web Won't Be Safe or Secure Until We Break It

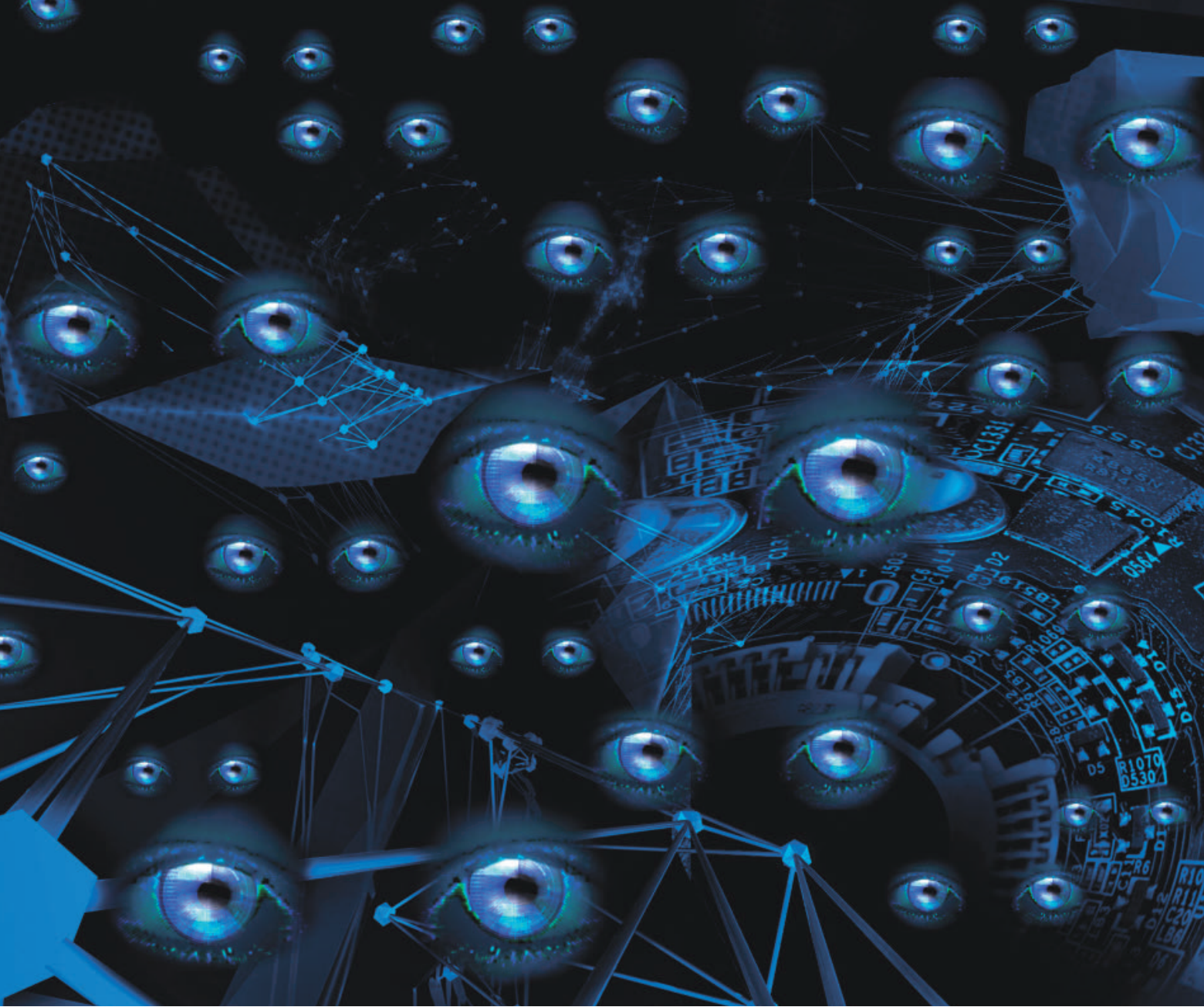
THE INTERNET WAS designed to deliver information, but few people envisioned the vast amounts of information that would be involved or the personal nature of that information. Similarly, few could have foreseen the potential flaws in the design of the Internet—more specifically, Web browsers—that would expose this personal information, compromising the data of individuals and companies.

If people knew just how much of their personal information they unwittingly make available to each and every website they visit—even sites they've never been to before—they would be disturbed. If they give that website just one click of the mouse, out goes even more personally identifiable data, including full name and address, hometown, school, marital status, list

of friends, photos, other websites they are logged in to, and in some cases, their browser's autocomplete data and history of other sites they have visited.

Obtaining all this information has been possible for years. Today's most popular browsers, including Chrome, Firefox, Internet Explorer, and Safari, do not offer adequate protection for their users. This risk of data loss seems to run counter to all the recent marketing hype about the new security features and improvements browser vendors have added to their products over the past several years such as sandboxing, silent and automatic updates, increased software security, anti-phishing and anti-malware warnings, all of which are enabled by default. While all are welcome advancements, the fact is





that these features are designed only to prevent a very particular class of browser attacks—those generally classified as *drive-by downloads*.

Drive-by downloads seek to escape the confines of the browser walls and infect the computer's operating system below with malware. Without question, drive-by-downloads are a serious problem—millions of PCs have been compromised this way when encountering infected websites—but they certainly are not the *only* threat browser users face, especially in an era of organized cybercrime and ultra-targeted online advertising.

The techniques behind attacks that obtain personal information are completely different and just as dangerous as malware, perhaps more so since the solution is far more com-

plicated than just installing antivirus software. These attack techniques have even more esoteric labels such as XSS (cross-site scripting), CSRF (cross-site request forgery), and clickjacking. These types of attacks are (mostly) content to remain within the browser walls, and they do not exploit memory-corruption bugs as do their drive-by download cousins, yet they are still able to do their dirty work without leaving a trace.

These attacks are primarily written with HTML, CSS (Cascading Style Sheets), and JavaScript, so they are not identifiable as malware by antivirus software in the classic sense. They take advantage of the flawed way in which the Internet was designed to work. The result is that these attack techniques are immune to protections that thwart

drive-by downloads. Despite the dangers they pose, they receive very little attention outside the inner circles of the Web security industry. To get a clearer picture of these lesser-known attacks, it's important to understand a common Web technology use case.

HTML allows Web developers to include remotely hosted image files on a Web page from any location across the Web. For example, a website located at <http://coolwebsite/> may contain code such as:

```
<img src= "http://someotherweb-  
site/image.png">
```

This instructs a visiting browser to send a Web request to <http://someotherwebsite/> automatically, and when returned, to display the image on the

screen. The developer may tack on some JavaScript to detect if the image file was loaded successfully or contained an error:

```

```

If the image file loaded correctly, then the “successful” JavaScript function executes. If an error occurred, then the error function executes. This code is completely typical and innocuous, but the same functionality can also be leveraged for invasive, malicious ends.

Now, let’s say `http://coolwebsite/` loaded an image file from `http://someotherwebsite/`, but that image file is accessible only if the user’s browser is currently logged into `http://someotherwebsite/`. As before:

```

```

If the user is logged in, then the image file loads successfully, which causes the executions of `loggedIn`. If the user is not logged in, then `notLoggedIn` is executed. The result is an ability to test easily and invisibly whether a visitor is logged in to a particular website that a Web developer does not have a relationship with. This login-detection technique, which leverages CSRF, can be applied to online banks, social networks, Web mail, and basically anything else useful to an attacker. The attacker behind `http://coolwebsite/` just has to find the URLs that respond in a Boolean state with respect to login.

Next, consider that a malicious website owner might want to go one step further and “deanonymize” a Web visitor, which is to say, learn the visitor’s real name. Assume from the previous example that the attacker can determine if the visitor is logged into Twitter, Facebook, Google+, among others. Hundreds of millions of people are persistently logged in to these online services every day. These websites, and many like them, are designed that way for convenience purposes.

The next thing an attacker could take advantage of is those familiar

third-party Web widgets, such as Twitter’s “Follow,” Facebook’s “Like,” and Google’s “+1” buttons.

While these buttons may seem innocent and safe enough, nothing really technically prevents websites from placing those buttons within an HTML container, such as a `div` tag, making those buttons transparent and hovering them just under a Web visitor’s mouse pointer. This is done so that when visitors click on something they see, they instead automatically Follow, Like, or +1 whatever else the bad guy wants them to. This is a classic case of clickjacking—an attack seen in the wild every day.

Here’s why this flaw in the Internet matters: since the attacker controls the objects behind those buttons, after the user clicks, the attacker can tell exactly “who” just Followed, Liked, or +1’ed on those online services (for example, Twitter: “*User X Followed you.*” Facebook: “*User X Liked Page Y.*”). To deanonymize the Web visitor, all the attacker needs to do is look at the public profile of the user who most recently clicked. That is when the fun begins for the attacker and trouble begins for the unsuspecting Internet user.

One more longstanding issue, “browser intranet hacking,” deserves attention. This serious risk, first discussed in 2006, remains largely unaddressed to this day. Browser intranet hacking allows website owners to access the private networks of their visitors, which are probably behind network firewalls, by using their browsers as a launch point. This attack technique is painfully simple and works equally well on enterprises and home users, exposing a whole new realm of data.

The attack flow is as follows: a Web user visits a malicious website such as `http://coolwebsite/`. That site instructs the visitor’s browser to make a Web request to an IP address or host name that the visitor can get to but the attacker cannot, such as `192.168.x.x` or any non-routable IP as defined by RFC-1918. Such requests can be forced through the use of `IMG` tags, as in the earlier example, or also through the use of `iframe`, `script`, and `link` tags:

```
<iframe src="http://192.168.1.1/"
onload="detection()"></iframe>
```

Depending on the detectable response given from the IP address, the attacker can use the Web visitor’s browser to sweep internal private networks for listening IP Web servers. This sweeping can locate printers, IP phones, broadband routers, firewalls, configuration dashboards, and more.

The technique behind browser intranet hacking is similar to the Boolean-state detection in the login-detection example. Also, depending on whether the user is logged in to the IP/Hostname, this type of attack can force the visitor’s browser to make configuration changes to the broadband router’s Web-based interface through well-known IPs (`192.168.1.1`, `10.10.0.1`, and so on) that can be quickly enumerated. The consequences of this type of exploitation can be devastating as it can lead to all traffic being routed through the attacker’s network first.

Beyond login detection, deanonymization, and browser intranet hacking are dozens of other attack techniques possible in today’s modern browsers. For example, IP address geolocation tells, roughly speaking, what city/town a Web visitor is from. The user-agent header reveals which browser distribution and version the visitor is using. Various JavaScript DOM (Document Object Model) objects make it trivial to list what extensions and plugins are available—to hack or fingerprint. DOM objects also reveal screen dimensions, which provides demographic context and whether the user is using virtualization.

The list of all the ways browser security can be bent to a website owner’s will goes on, but the point is this: Web browsers are not “safe”; Web browsers are not “secure”; and the Internet has fundamental flaws impacting user (personal or corporate) security.

Now here’s the punch line: the only known ways of addressing this class of problem adequately is to “break the Web” (that is, negatively impact the usability of a significant percentage of websites). These issues remain because Web developers, and to a large extent Web users, demand that certain functionality remain available, and that functionality is what makes these attacks possible.

Today’s major browser vendors, whose guiding light is market share,



are only too happy to comply. Their choice is simple: be less secure and more user-adopted, or be secure and obscure. This is the Web security trade-off—a choice made by those who do not fully understand, appreciate, or are liable for the risks they are imposing on everyone using the Web.


### Nonstarter Solutions

To fix login detection, a browser might decide not to send the Web visitor's cookie data to off-domain destinations (those different from the hostname in the URL bar) along with the Web requests. Cookies are essential to tracking login state. The off-domain destination could still get the request, but would not know to whom it belonged. This is a good thing for stopping the attack.


Not sending cookies off-domain, however, would break functionality for any website that uses multiple hostnames to deliver authenticated content. The approach would break single-click Web widgets such as Twitter's "Follow," Facebook's "Like," and Google's "+1" buttons. The user would be required to perform a second step. It would also break visitor tracking via Google Analytics, Coremetrics, and so on. This is a clear nonstarter from the perspective of many.

To fix clickjacking, Web browsers could ban iframes entirely, or at least ban transparent iframes. Ideally, browser users should be able to "see" what they are really clicking on. Suggesting such a change to iframes, however, is a losing battle; millions of websites rely upon them, including transparent iframes, for essential functionality. Notable examples are Facebook, Gmail, and Yahoo! Mail. You do not normally see iframes when they are used, but they are indeed everywhere. That level of breakage is never going to be tolerated.

For browser intranet hacking, Web browsers could prohibit the inclusion of RFC-1918 resources from non-RFC-1918 websites. This essentially creates a break point in the browser between public and private networks. One reason that browser vendors say this is not doable is that some organizations actually do legitimately include intranet content on public websites. Therefore, because some organiza-



## Dramatic improvements in browser security and online privacy are held hostage by backward compatibility requirements related to how the Internet was designed.



tions (whom you have never heard of and whose websites you'll never visit) have an odd use case, your browser leaves the private networks you are on, and that of hundreds of millions of others, wide open.

As shocking as this sounds, try looking at the decision not to fix the problem from the browser vendors' perspective. If they break the uncommon use case of these unnamed organizations, the people within those organizations are forced to switch to a competing "less-secure" browser that allows them to continue business as usual. While the security of all other users increases for the browser that makes the change, that browser vendor loses some fraction of market share.

### Security Chasm

The browser vendors' unwillingness to risk market share has led to the current security chasm. Dramatic improvements in browser security and online privacy are held hostage by backward compatibility requirements related to how the Internet was designed. Web-browser vendors compete with each other in trench-style warfare, gaining ground by scratching for a tiny percentage of new users, everyday—users who do not pay them a dime, while simultaneously trying to keep every last user they already have.

It's important to remember that mainstream browsers are essentially advertising platforms. The more eyeballs browsers have, the more ads are delivered. Ads, and ad clicks, are what pay for the whole party. Anything getting in the way of that is never a priority.

To be fair, there was one important win recently when, after years of discussion, a fix was applied to CSS history sniffing. This is the ability of a website to uncover the history of other websites a user had visited by creating hyperlinks on a Web page and using either JavaScript or CSS to check the color of the link displayed on the screen. A blue link meant the visitor had not been there; purple indicated the user had visited the site. This was a serious privacy flaw that was simple, effective, and 10,000-URLs-per-second fast to execute. Any website could quickly know where you banked, shopped, what news you read, adult websites frequented, among others.

The problem of CSS history sniffing finally got so bad and became so high profile that approximately 10 years after it first came up, all the major browser vendors finally broke the functionality required for the attack. Many Web developers who relied on the underlying functionality were vocally upset, but apparently this was an acceptable level of breakage from the browser vendors' perspective.

When the breakage is not acceptable, but the issue is still bad, new opt-in browser security features are put forth. They generally have low adoption rates. Prime examples are Content Security Policy, X-Frame-Options, Origin, Strict Transport Security, SSL (Secure Sockets Layer), Secure and HttpOnly cookie flags, and others. Website owners can implement these solutions only when or if they want to, thereby managing their own breakage. What none of these features do is to allow Web users to protect themselves, something every browser should enable its users to do. Right now, Web security is in a holding pattern—waiting for the bad guys to cause enough damage—which then should give enough juice to those with the power to take action.

### Beyond the Status Quo

The path toward a more secure Web has a few options. We could establish a brand-new World Wide Web, or an area within it. A Web platform designed to be resilient to the current laundry list of problems, however, will forever plague its predecessor. For the moment, let's assume we technically know how to make a secure platform, which is a big *if*.

The next step would be to convince the developers behind the millions, potentially hundreds of millions, of important websites to move over and/or build atop version two. Of course, the promise of a "more secure" platform would not be sufficient incentive by itself. They would have to be offered something more attractive in addition. Even if there were something more attractive, this path would only exchange our backward-compatibility problem for a legacy problem, which is likely to take years, perhaps a decade or more, to get beyond.

There is another path—one that already has a demonstrated model of

success in mobile applications. What you find there basically amounts to many tiny Web browsers connected to the mobile version of the main website. The security benefit provided by mobile platforms such as Apple's iOS and Google's Android is that the applications are isolated from one another in both memory and session state.

For example, if you launched Bank of America's mobile application, logged in, did your banking, and then subsequently launched Facebook's mobile application and logged in, neither app has access to the other app's session, as would be the case in a normal desktop Web browser. Mobile applications have little to no issues regarding login detection, deanonymization, and intranet hacking. If mobile platforms can get away with this level of application and login-state isolation, certainly the desktop world could as well.

By adopting a similar application model on the desktop using custom-configured Web browsers (let's call them DesktopApps), we could address the Internet's inherent security flaws. These DesktopApps could be branded appropriately and designed to launch automatically to Bank of America's or Facebook's website, for example, and go no further. Like their mobile application cousins, these DesktopApps would not present a URL bar or anything else making them look like the Web browsers they are on the surface, and of course they would be isolated from one another. Within these DesktopApps, attacks such as XSS, CSRF, and clickjacking would become largely extinct because no cross-domain connections would be allowed—an essential precondition.

DesktopApps would also provide an important security benefit to Chrome, Firefox, Internet Explorer, and Safari. Attacks such as login detection and deanonymization would be severely hampered. Let's say Web visitor X uses only a special DesktopApp when accessing the websites of Bank of America, Facebook, or whatever else and never uses the default Web browser for any of these activities. When X is using Chrome, Firefox, or Internet Explorer and comes across a website trying to perform login detection and deanonymization, well, X has never logged in to

anything important in that browser, so the attacks would fail.

What about intranet hacking? The answer is to break the functionality, as described earlier. Web browsers should not allow non-RFC-1918 websites to include RFC-1918 content—at least not without an SSL-style security exception. One or all of the incumbent browser vendors need to be convinced of this. If that mystery company with an odd use case wants to continue, it should have a special corporate DesktopApp created that allows for it. It would be far more secure as a result, as would we all.

This article has outlined a broad path to fix Web security, but much is left unaddressed about how to roll out a DesktopApp and get the market to adopt such practices. Beyond just the security benefits, other features are needed to make DesktopApps attractive to Web visitors; otherwise there is no incentive for browser vendors to innovate. There's also lobbying to be done with website owners and developers. All of this makes fixing the Internet a daunting task. To get past security and reach our final destination—a world where our information remains safe—we must develop creative solutions and make hard choices. 

### Related articles on [queue.acm.org](http://queue.acm.org)

#### Browser Security

Charles Reis, Adam Barth, and Carlos Pizano  
<http://queue.acm.org/detail.cfm?id=1556050>

#### Security in the Browser

Thomas Wadlow and Vlad Gorelik  
<http://queue.acm.org/detail.cfm?id=1516164>

#### Cybercrime 2.0: When the Cloud Turns Dark

Niels Provos, Moheeb Abu Rajab, and Panayiotis Mavrommatis  
<http://queue.acm.org/detail.cfm?id=1517412>

**Jeremiah Grossman** is the founder and CTO of WhiteHat Security, where he is responsible for Web security R&D and industry outreach. He is a cofounder of Web Application Security Consortium (WASC) and was previously named one of *InfoWorld's* Top 25 CTOs. He serves on the advisory boards of two start-ups, Risk I/O and SD Elements. Before founding WhiteHat, he was an information security officer at Yahoo!

The Ultimate Online Resource for Computing Professionals & Students

ACM DL DIGITAL LIBRARY

<http://www.acm.org/dl>



Association for Computing Machinery

Advancing Computing as a Science & Profession

DOI:10.1145/2398356.2398375

## Anonymous location data from cellular phone networks sheds light on how people move around on a large scale.

BY RICHARD BECKER, RAMÓN CÁ CERES, KARRIE HANSON, SIBREN ISAACMAN, JI MENG LOH, MARGARET MARTONOSI, JAMES ROWLAND, SIMON URBANEK, ALEXANDER VARSHAVSKY, AND CHRIS VOLINSKY

# Human Mobility Characterization from Cellular Network Data

AN IMPROVED UNDERSTANDING of human-mobility patterns would yield insight into a variety of important societal issues. For example, evaluating the effect of human travel on the environment depends on knowing how large populations move about in their daily lives. Likewise, understanding the spread of a disease requires a clear picture of how humans move and interact. Other examples abound in such fields as urban planning, where knowing how people come and go can help determine where to deploy infrastructure and how to reduce traffic congestion.

Human-mobility researchers traditionally rely on expensive data-collection methods (such as surveys and direct observation) to glimpse the way people move about. This cost typically results in infrequent data collection or small sample sizes; for example,

the U.S. national census produces a wealth of information on where hundreds of millions of people live and work but is carried out only once every 10 years.

In contrast, data from cellular telephone networks can help study human mobility cheaply, frequently, and on a global scale. Billions of people worldwide keep a phone near them most of the time. Since cellular networks need to know the approximate location of all active phones to provide them voice and data services, location information from these networks holds the potential to revolutionize the study of human mobility.

We have analyzed billions of anonymized Call Detail Records (CDRs) from a cellular network to characterize the mobility patterns of hundreds of thousands of people. CDRs are routinely collected by wireless-service providers for billing and to help operate their networks by, say, identifying congested cells in need of more resources. Each CDR contains information (such as the time a phone placed a voice call or received a text message, as well as the identity of the cellular antenna with which the phone was associated at the time). When joined with information about the locations and directions of these antennas, CDRs can serve as sporadic samples of the approximate locations of the associated phones' owners.

### » key insights

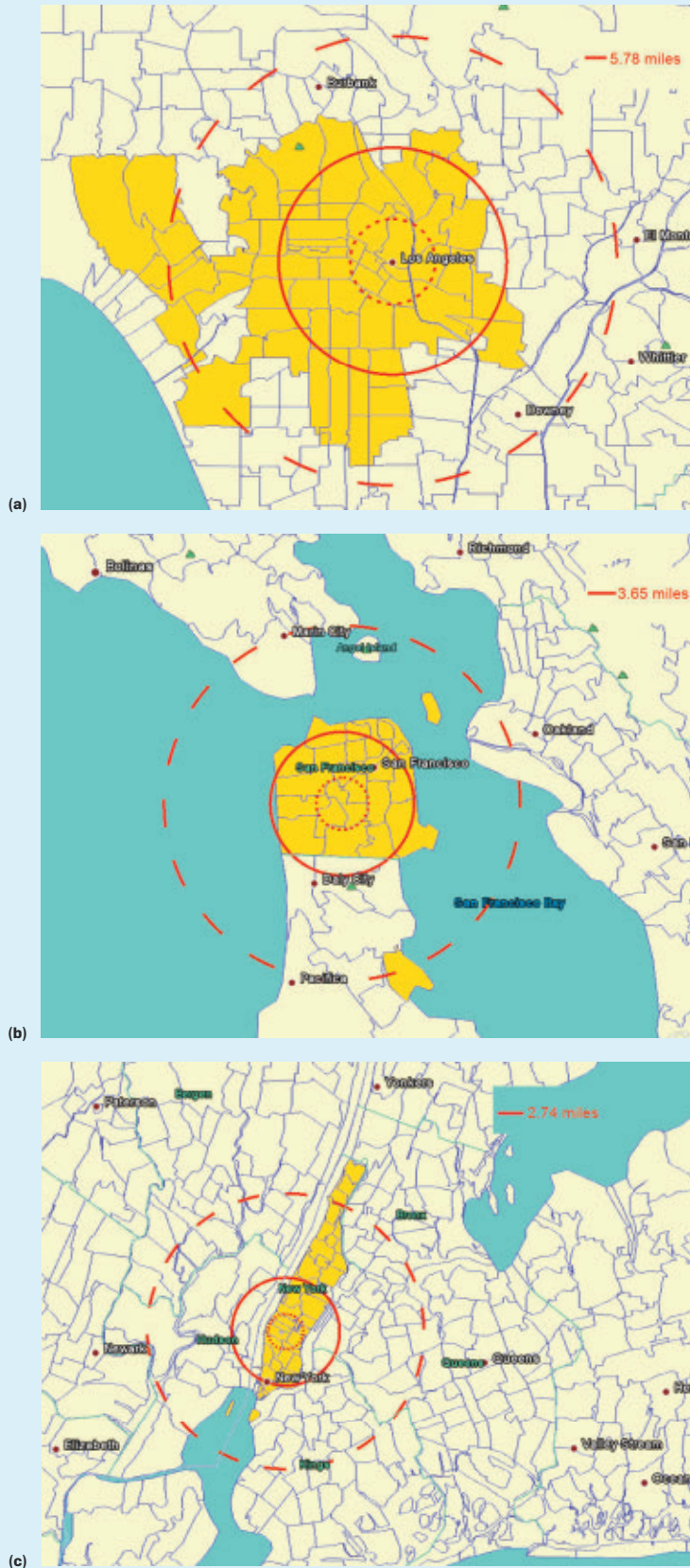
- Cellular telephone networks enable the study of human mobility at low cost and on an unprecedented scale.
- Results from such studies have broad applicability in mobile computing, urban planning, ecology, and epidemiology.
- We have developed and validated techniques for analyzing billions of anonymous location samples to determine the daily range of travel, carbon footprint of home-to-work commutes, and other mobility characteristics of hundreds of thousands of people living in the Los Angeles, San Francisco, and New York metropolitan areas.

ILLUSTRATION BY ALICIA KUBISTA / ANDRÉJ BORYS ASSOCIATES



**Figure 1. Median daily range of cellphone users living in central LA, SF, and NY (darker yellow areas).**

The radii of the inner, middle, and outer circles represent the 25th, 50th, and 75th percentiles of these ranges across all users in that area. All maps are drawn to the same scale.



CDRs are an attractive source of location information for three main reasons: They are collected for all active cellular phones, numbering in the hundreds of millions in the U.S. and billions worldwide; they are already being collected to help operate the networks, so additional uses incur little marginal cost; and they are continuously collected as each voice call and text message completes, enabling timely analysis.

At the same time, CDRs have two significant limitations: They are sparse in time because they are generated only when a phone engages in a voice call or text-message exchange; and they are coarse in space because they record location only at the granularity of a cellular antenna. Not obvious a priori is whether CDRs provide enough information to characterize human mobility in a useful way.

Since 2009, we have pursued a research program aimed at developing sound analysis techniques for exploring aspects of human mobility using CDRs and shown that CDRs are indeed useful for accurately characterizing important aspects of human mobility. Our results to date include the following:

*Daily travel.* We have determined how far anonymous populations of hundreds of thousands of people travel every day in the Los Angeles, San Francisco, and New York metropolitan areas;

*Carbon emissions.* We have calculated the carbon emissions due to the home-to-work commutes of these populations, accounting for differences in distance and modes of travel;

*Number of workers and event goers.* We have identified which residential areas contribute what relative number of workers and holiday parade attendees at a suburban city—Morristown, NJ; and

*Traffic volumes.* We have estimated relative traffic volumes on the main commuting routes into Morristown.

We validated our results by comparing them against ground truth provided by volunteers and against independent sources (such as the U.S. Census Bureau). Throughout our work, we have taken measures to preserve individual privacy. The rest of this article covers the methodologies and findings of our human-mobility studies based on cellular network data.

**Privacy and Terminology**

Though CDRs are a valuable source of data for mobility studies that could benefit society at large, cellular customers rightfully have the expectation that their individual privacy will be preserved. We take several active steps to protect privacy:

*Anonymization.* All our CDRs are anonymized by someone not involved in the data analysis; each cellular phone number is replaced with an identifier consisting of a unique integer;

*Minimal information.* We use only the minimal information needed for our studies. Our simplified CDRs consist of the anonymous phone identifier, date, and time of a voice call or text message; the elapsed time of a call (zero for a text message); the cellular antennas involved in the event; and the phone’s billing ZIP code. Our data does not include demographic information for the subscriber or any information about the other party in the communication. In some of our studies we use the billing ZIP code as a rough estimate of the phone owner’s home location. We excluded business subscribers from all our datasets because those billing ZIP codes generally do not correspond to home locations; and

*Aggregate results.* We present only aggregate results and do not focus our analysis on individual phones, aside from those of a group of volunteers who gave us permission to look at their records.

In addition to these active steps, the nature of CDRs is to give only temporally sparse and spatially coarse information about a phone. A CDR is generated only when the phone is used for a call or text message; the phone is invisible to us at all other times. We know only the location of the phone in an approximate way, based on the antennas involved with the call. Because an antenna often covers an area greater than one square mile, our spatial resolution is limited.

A brief note on terminology surrounding cellular network equipment will help in understanding the rest of the article. We refer to a cell tower as the location of equipment placed on a freestanding tower, atop a building, or on some other physical structure. In general, each tower hosts multiple antennas, each handling a particular

radio technology and frequency (such as Universal Mobile Telecommunications System at 850 MHz) and pointing in a specific compass direction (such as north). All antennas pointing in the same direction from the same tower cover what we call a sector.

**Daily Range of Travel**

How far do people travel every day? We can approximate this quantity by finding the maximum distance between any two cell towers a phone contacts in one day, calling this distance the daily range. Here, we present some of our findings regarding the daily range of people living in three major metropolitan areas in the U.S.: Los Angeles (LA), San Francisco (SF), and New York (NY).

We gathered anonymous location data for cellular phones whose owners live in the metropolitan regions of interest. We identified ZIP codes within a 50-mile radius of the LA, SF, and NY city centers, corresponding to the colored regions in Figure 2. We obtained anonymized CDRs for a random sample of phones with billing addresses in those ZIP codes. And, so as to exclude people not living near their billing address, we removed all CDRs for phones that appeared in their base ZIP code fewer than half the days they had voice or text activity.

The table here describes our most recent dataset for each region, with each dataset containing hundreds of millions of location samples for hundreds of thousands of phones over three months of activity, with 12–18 median location samples per day for each phone.

We compared our sets of phones against U.S. Census data<sup>24</sup> and confirmed the number of sampled phones in each ZIP code is proportional to the population of that ZIP code. We there-

fore believe our datasets are representative of the populations at large in the regions of interest.

We computed each phone’s daily range by calculating distances between all pairs of cell towers contacted by the phone on a given day and selecting the maximum distance between any two such towers. To validate our methodology, we recruited volunteers who logged their actual locations for one month and gave us permission to inspect their CDRs for the same period. The median difference between daily ranges computed from CDRs and those derived from the ground-truth logs was less than 1.5 miles, giving us confidence in our range-of-travel results; for more, see Isaacman et al.<sup>13</sup>

The study of daily ranges yields numerous insights about human mobility. For example, the median of a phone’s daily range values over the duration of a dataset is an approximation of the most common daily distance traveled by the phone’s owner. Similarly, the maximum daily range across a dataset corresponds to the longest trip taken during that time.

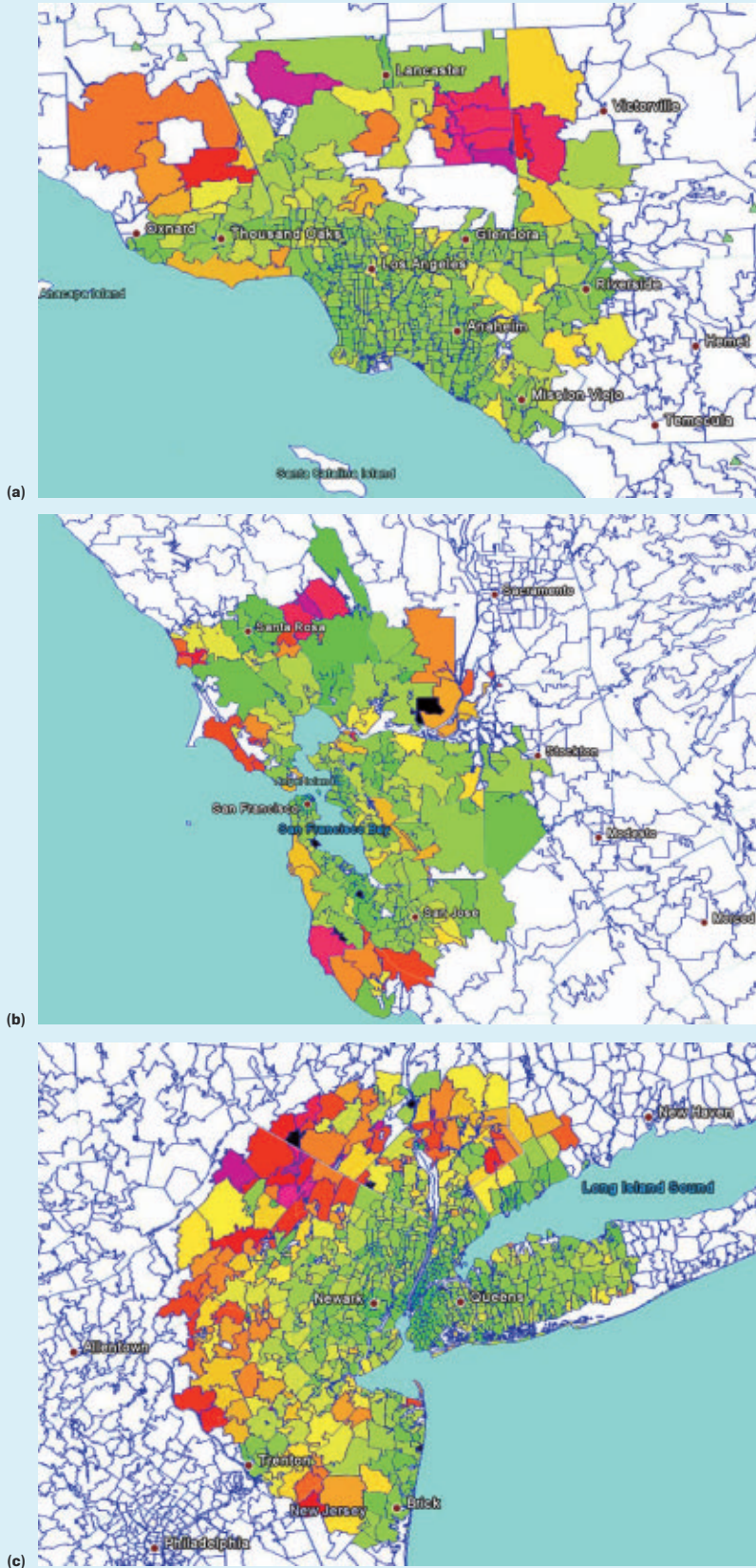
Figure 1 is a visual representation of the median daily ranges for residents of central LA, SF, and NY; the darker yellow areas correspond to ZIP codes in the City of Los Angeles, the City of San Francisco, and the Borough of Manhattan. These areas do not include the surrounding communities also represented in our complete metropolitan-area datasets. The radii of the red circles are proportional to the median daily ranges for residents of the corresponding shaded areas. As shown, people living in the city of Los Angeles travel longer distances on a typical day than people living in the city of San Francisco, who in turn travel longer distances than people living in Manhattan.

**Characteristics of CDR datasets for the LA, SF, and NY metropolitan areas, with each dataset spanning 91 consecutive days, April 1 to June 30, 2011.**

|                                | LA    | SF   | NY    |
|--------------------------------|-------|------|-------|
| Total unique phones            | 318K  | 241K | 267K  |
| Total unique CDRs              | 1395M | 701M | 1095M |
| Median CDRs per phone per day  | 18    | 12   | 18    |
| Median calls per phone per day | 6     | 5    | 7     |
| Median texts per phone per day | 6     | 3    | 5     |

**Figure 2. Median carbon emissions per home-to-work commute of cellphone users living in the LA, SF, and NY metropolitan areas.**

Greener ZIP codes denote smaller carbon footprints, ranging through yellow, orange, red, and purple as footprints grow. All these maps use the same geographic and carbon scales; emissions are scaled linearly.



By analyzing similar datasets from different time periods, we made additional spatial and temporal comparisons between the daily ranges of various populations. For example, people throughout the LA region travel farther on a typical day than people throughout the NY area. In contrast, the longest trips taken by residents of Manhattan are much longer than those taken by residents of central Los Angeles. Furthermore, people in both the LA and NY regions tend to travel shorter distances in the winter months than in the summer months, with the effect being more pronounced in NY. For a more complete description of our daily range results, see Isaacman et al.<sup>13</sup> and Isaacman et al.<sup>14</sup>

### Carbon Footprints

Evaluating the environmental impact of human travel is of urgent interest to society at large. A person's commute between home and work can account for a significant portion of his or her overall carbon footprint. We can estimate the carbon emissions due to these commutes by combining our datasets of cellphone locations with a U.S. Census dataset on mode of transport to work (such as automobile, bus, and train)<sup>24</sup> and a table of carbon emissions by mode of transport.<sup>4</sup>

We devised an algorithm that uses CDRs to identify important places in people's lives, defined as places a person visits frequently or spends a lot of time. We further identified the likely home and work locations from among these important places, then calculated the home-to-work commute distance. Our approach, described in more detail and validated in Isaacman et al.,<sup>12</sup> uses a series of clustering and regression steps to accomplish these tasks. We found our commute-distance estimates were within one mile of the ground-truth distances provided by volunteers.

We then applied this approach to our large CDR datasets for the LA, SF, and NY metropolitan areas described earlier and computed the distribution of commute distances across the population of each ZIP code in our regions of interest. We found that our estimates were within one mile of the average commute distances for these same regions as published by the U.S. Bureau of Transportation Statistics.<sup>23</sup>



Finally, we joined our distributions of commute distances with the publicly available distributions of modes of transport per ZIP code and of carbon emissions per mode of transport per passenger. Figure 2 shows our results in the form of heat maps, where color corresponds to the median carbon emission per commute across the people in each ZIP code. Colors are ordered so greener ZIP codes correspond to lower carbon emissions, with yellow, orange, red, and purple ZIP codes showing increasing emissions.

In the NY area, increasing distance from Manhattan correlates with an increasing carbon footprint; in contrast, LA is more uniform throughout, except for parts of Antelope Valley (northeast portion of the map) separated from downtown LA by a mountain range drivers must go around. The results for SF are between those for NY and LA.

These patterns match well with generally understood movement patterns in each city. Popular knowledge indicates that in NY, a great many people commute into Manhattan, while in LA, there is no single concentration of jobs. SF has at least two major job centers, one focused in the city of San Francisco proper, another in Silicon Valley approximately 40 miles to the south. Thus, unlike NY, SF has more than one strong jobs focus, but unlike LA, it has some clear areas of jobs focus.

Beyond identifying patterns of carbon emissions, we also compared raw carbon values. For instance, though difficult to see in Figure 2, Manhattan ZIP codes have the smallest carbon footprints of all ZIP codes studied, presumably due to the nearness to work of many people's homes, as well as to an extensive public transportation infrastructure.

**Laborshed and Paradeshed**

City and transportation planners are interested in knowing the home locations of people who work in and visit their city. The information is useful in, say, forecasting road-traffic volume during morning and evening rush hours. The set of residential areas that contribute workers to a city is known as the city's laborshed.

To study an example laborshed, we captured all transactions carried by the 35 cell towers located within five miles

of the center of Morristown, NJ, a suburban city with approximately 20,000 residents. These 35 towers house approximately 300 antennas pointed in various directions and supporting various radio technologies and frequencies. Our goal was to capture cellular traffic in and around the town. Choosing the five-mile radius allowed us to cover both Morristown proper and its neighboring areas. We obtained anonymized CDRs for 60 consecutive days, March 1 to April 29, 2011, thus collecting more than 17 million voice CDRs and 39 million text CDRs for more than 472,000 unique phones.

We identified Morristown's laborshed from the CDRs as follows: We classified as Morristown workers those cellphone users with significant activity inside Morristown during business hours (9 A.M. to 5 P.M., Monday to Friday). We then used billing ZIP codes to identify their places of residence. This method produced counts of Morristown workers by residential ZIP code.

We validated our results by comparing them with data from the 2000 U.S. Census, confirming that the number of workers we attributed to each ZIP code was strongly correlated with the number of workers in the same ZIP

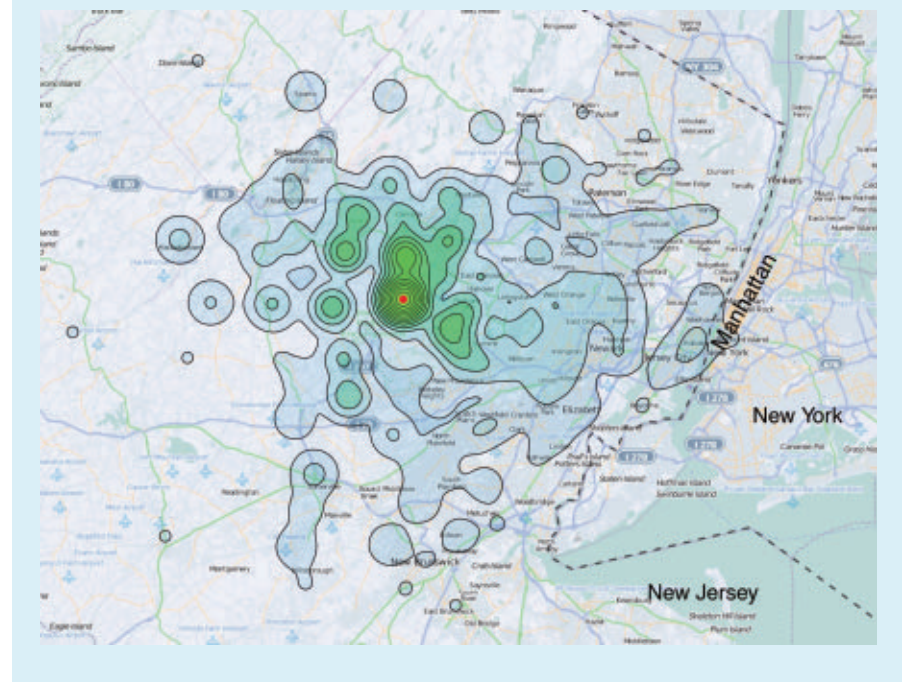
code as published in the "Journey to Work" tables of the 2000 U.S. Census Transportation Planning Package.<sup>24</sup> Our analysis and validation methodology are described in more detail in Becker et al.<sup>2</sup>

Figure 3 is a geographic representation of Morristown's laborshed, with darker colors indicating the home areas of larger numbers of Morristown workers. Interestingly, there seem to be many more workers coming from the area immediately north of Morristown than from the south. These two areas have similar population densities, so the difference may be related to geography, demographics, or transportation infrastructure. Furthermore, though population density increases dramatically to the east (as one gets closer to Manhattan), we see almost as many workers coming from the west, perhaps because Morristown is a regional center of commerce. However, there do seem to be workers making long "reverse commutes" from areas of New Jersey close to Manhattan. All these facts could be useful to policymakers deciding on future municipal and regional mass-transit investments.

Our methodology allows us to estimate the flow of people in and out

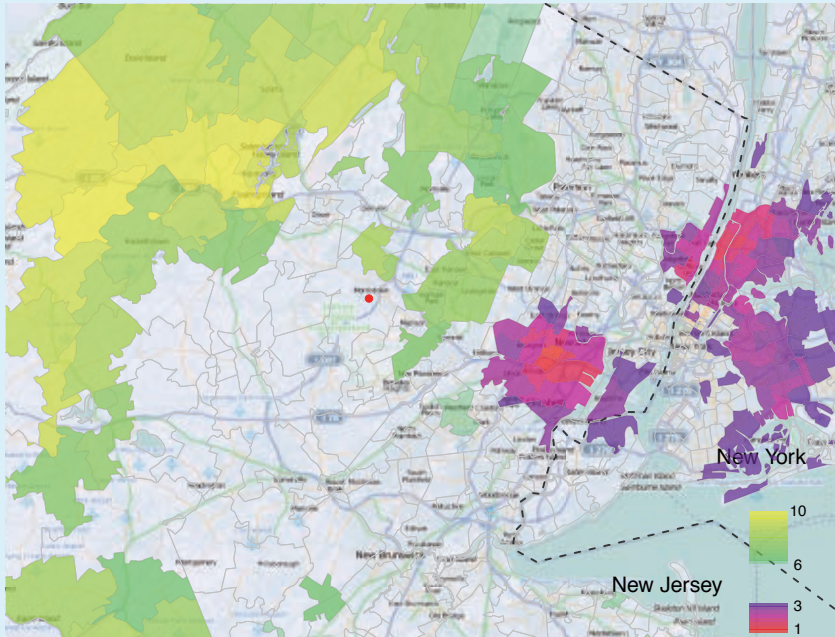
**Figure 3. Laborshed of Morristown, NJ; the red dot denotes the city center.**

Contour lines divide regions of different concentrations of workers' homes, with workers identified as those who use their cellphones in Morristown during weekday business hours. Most workers are from nearby areas, but some are from as far as 40 miles away in Manhattan.



**Figure 4. Paradeshed of Morristown, NJ; the red dot denotes the city center.**

Five times as many people were in Morristown for the St. Patrick's Day Parade as on a normal Saturday. To show the geographical distribution of parade attendees' homes, we mapped the number of people coming from each surrounding ZIP code. Green-yellow areas contributed more than the parade-day average and purple-red areas less than that average. Communities contributing near the average are not colored to highlight the outliers.



of a geographic area during arbitrary time periods. Of particular interest to city officials is how the mix of inhabitants changes during special occasions (such as extreme weather, construction projects, and regional events). Knowing where people come from can help them in advertising for the event and easing traffic congestion.

One such occasion in Morristown was the St. Patrick's Day Parade on Saturday March 12, 2011, from 11 A.M. to 3 P.M. We repeated our analysis for obtaining the laborshed but on cell-phone transactions handled during the time of the parade by the antennas pointing along the parade route. Figure 4 is the resulting paradeshed, with people coming for the parade, compared with data for the same antennas and time interval on a typical Saturday without special events. The parade is a county affair, so we would have expected the event to draw widely from other parts of the county (north and west of Morristown). Indeed, we see the areas north and west of Morristown showing large increases, while other areas south and east show smaller increases. Prior to the

advent of cellular networks, it was notably difficult for local officials to obtain this information except through expensive surveys.

**Traffic Volume**

The quality of life in any urban area is directly influenced by the frustration, pollution, time lost, and noise of traffic congestion. Efforts by planners to improve traffic flow while not sacrificing street life need a thorough understanding of existing traffic conditions. Since traditional methods of obtaining traffic data are expensive, we set out to determine whether we could estimate traffic volumes from CDRs.

To explore traffic volume on major commuting routes into Morristown, we used the same data-collection procedure we used to calculate the laborshed, as described earlier. However, in this case we recorded activity in and around Morristown from December 2009 to January 2010. We used two filters to obtain an appropriate subset of CDRs for the study: First, to retain data about moving vehicles, we used only voice CDRs including antennas on at least five towers, as indicated by our

own experiments to determine how motion was reflected in CDRs. We ignored text CDRs because text messages involve only a single location. Second, since we were interested in routes to and from the center of town, we used only CDRs with antenna sequences that began or ended at the tower handling calls for the core downtown area. After filtering, we were still left with tens of thousands of CDRs.

We began by identifying 15 common commuting routes (13 driving routes and two train routes) radiating from the town center. We obtained ground-truth data for them by driving/riding each one four times (two in each direction), using at least two phones calling each other on each drive/ride. We obtained the CDRs for these calls to both train and test our algorithms. From our training data, we determined a reference pattern of cellular sectors used by calls on each of the routes. We intentionally included some routes very close to one another and others that partially overlap, as routes do in real life. Some of our reference patterns were thus quite similar, making disambiguation a challenge.

We then developed two methods for assigning CDRs to routes: One uses a distance metric to assign a test CDR to the route with the closest reference pattern. We used a variant of Earth Mover's Distance (EMD), a measure of the difference between two arbitrary probability distributions, as a metric that takes into account common subsets of sectors, the particular sequence of sectors, how long the call is associated with each sector, and tower locations. The other method uses as reference data the radio-frequency scans routinely performed by cellular network operators to measure network coverage. The scanner data contains signal-strength measurements stamped with global-positioning system (GPS) locations from all observable antennas along major driving routes. Our classification algorithm estimates the likelihood of a given sequence of antennas being seen on a particular route and selects the most likely route. This approach has the advantage of being able to reuse data that is already available, without requiring additional data collection on every target route. It could easily be ex-

tended to larger-scale studies in other urban areas.

Both classification algorithms achieved approximately 90% accuracy on our test data, outperforming several other algorithms based purely on common subsets of towers, sectors, or antennas. Our route-classification algorithms and their accuracy are described in more detail in Becker et al.<sup>1</sup>

Figure 5 shows the result of our route assignment to moving phones in the Morristown area, using the EMD-based algorithm applied to CDRs; the signal-strength-based method yields similar results. The widths of the lines superimposed on each route are proportional to the estimated traffic volumes on each route. The two wide black lines running roughly north and south correspond to the interstate highway that passes through Morristown. The counts shown at the beginning of each route are normalized to 1,000 moving phones. We compared our relative traffic volumes to traffic counts published by the New Jersey Department of Transportation<sup>17</sup> and found a correlation coefficient of 0.77, giving us added confidence in the accuracy of our approach.

**Related Work**

The research community increasingly uses cellular network data to study human mobility, applying its findings to various domains, including urban planning,<sup>19</sup> mobility modeling,<sup>10</sup> social-relation inference,<sup>11</sup> and health care.<sup>3</sup> Here, we survey a subset of that work most similar to our own.

Several efforts have explored how cellular network data can be used for urban planning. In studies of Milan, Italy, Ratti et al.<sup>19</sup> and later Pulselli et al.<sup>18</sup> demonstrated it is possible to characterize the intensity and spatio-temporal evolution of urban activities using call volume at cell towers. Reades et al.<sup>20</sup> studied call-volume activity in six distinct locations in Rome, Italy, showing that volume varied drastically between the studied locations and between weekdays and weekends. Girardin et al.<sup>8</sup> used tagged photographs from Flickr in combination with call-volume data to determine the whereabouts of locals and tourists in Rome. They later repeated the study with only call-volume data to exam-

ine differences in behavior between tourists and locals in New York City.<sup>9</sup> Calabrese et al.<sup>6</sup> studied where people came from to attend special events in Boston, finding that people who live close to an event are more likely to attend it and that events of the same type attract people from roughly the same home locations. Though we have also studied how cellular network data can be used for urban planning, we pursued different research goals (such as calculating daily ranges, deriving and validating laborsheds, and estimating traffic volume).

In the domain of mobility modeling, Gonzalez et al.<sup>10</sup> used cellular network data from an unnamed European country to form statistical models of how individuals move, finding human trajectories reflect a high degree of spatial and temporal regularity, with each individual having a time-independent characteristic travel distance and returning often to a few characteristic locations. Song et al.<sup>21</sup> analyzed similar data to study the predictability of an individual's movements, finding a high degree of predictability across a large user base largely independent of travel

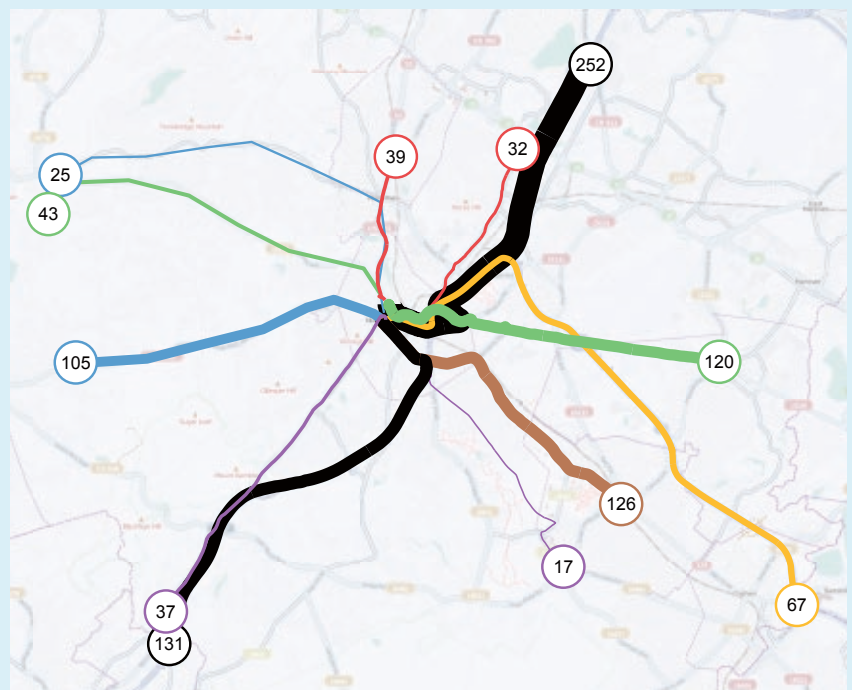
distances and other factors. Whereas these efforts modeled individuals, we focused on mobility differences between large populations in distinct geographic regions.

A complementary approach to collecting human-mobility data from cellular networks is to collect it directly from cellular phones themselves. For example, as in our route-classification work, CTrack<sup>22</sup> maps a phone's route by matching the cellular-signal-strength fingerprints seen by a phone against a database of such fingerprints. More generally, there is a growing body of work in participatory sensing that uses cellphones as sensors of location and other contexts.<sup>5,7,16</sup> Cellphone-based efforts have some attractive properties, most notably that they often have access to more varied and finer-grain sources of location information (such as GPS readings and Wi-Fi fingerprints) than the cellular antenna identities in our CDRs.

However, our network-based approach maintains important advantages: In particular, the cellphone-based approach typically requires the installation and running of spe-

**Figure 5. Relative traffic volume on 12 commuting routes to the center of Morristown, NJ, as assigned by our route-classification algorithms.**

Line widths are proportional to the estimated volumes; counts shown at the beginning of each route are normalized to 1,000 moving cellphones.



cial software on phones, consuming power on the devices and generally inhibiting truly large-scale data collection. In contrast, we use information already collected by the network for all phones and does not require additional software or consume extra power on mobile devices. As a result, our work has involved orders of magnitude more subjects than participatory-sensing efforts to date.

### Conclusion

Our goal with this work has been to make a case for the value of cellular network data to support a range of research and policy goals related to human mobility. Through several studies since 2009, we have demonstrated how CDRs—despite their temporal sparseness and spatial coarseness—offer important insights into the movement patterns of individuals and communities.

To demonstrate the broad utility of CDR data, our work comprises several types of analyses: In one case, we demonstrated techniques for identifying important places in people’s lives from CDR traces. Coupling them with other data (such as U.S. Census data on transportation use) we are able to generate estimates of home-to-work carbon footprints in a manner that can be updated much more frequently than typical census surveys, which are expensive and therefore infrequent. We also showed the use of CDR-based analysis to map laborshed statistics, helping predict how special events (such as a holiday parade) might influence commute and travel patterns.

These studies point to the great value of cellular network data for future urban-planning applications (such as traffic-congestion mitigation and mass-transit planning). Unlike expensive and infrequent census approaches, the fact that CDR-based mobility data can be collected in unobtrusive ways has the potential to make broad use both cheaper and easier.

Motivating all this work is the desire to glean useful statistics and models from the data without compromising the privacy of individual cellular telephone users. We employed various anonymization techniques to ensure privacy preservation. More broadly, we showed that a range of useful conclusions can be drawn about regional

mobility patterns based solely on anonymized, sampled, highly aggregated versions of the source mobility data.

Our most recent work seeks to provide fully synthetic models that mimic the individual and regional mobility patterns seen in the measured CDRs.<sup>15</sup> Such models will further improve the ability of scientists and planners to perform accurate, low-cost, privacy-preserving studies of human mobility.

### Acknowledgments

Part of this work was performed while Sibren Isaacman was a doctoral student at Princeton University; it was supported by the National Science Foundation under Grant Nos. CNS-0614949, CNS-0627650, and CNS-0916246. Isaacman also acknowledges support from a Wallace Memorial Fellowship in Engineering from Princeton University and research internships from AT&T Labs.

### References

1. Becker, R., Cáceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., and Volinsky, C. Route classification using cellular handoff patterns. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, 2011).
2. Becker, R., Cáceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., and Volinsky, C. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing* 10, 4 (Oct–Dec. 2011), 18–26.
3. Bengtsson, L., Lu, X., Thorson, A., Garfield, R., and von Schreeb, J. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLoS Medicine* 8, 8 (Aug. 2011).
4. M.J. Bradley and Associates. *Comparison of Energy Use & CO<sub>2</sub> Emissions from Different Transportation Modes*. Report to American Bus Association, Washington, D.C., May 2007; <http://www.buses.org/files/ComparativeEnergy.pdf>
5. Burke, J., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., and Srivastava, M.B. Participatory sensing. In *Proceedings of the Workshop on World-Sensor-Web: Mobile Device-Centric Sensor Networks and Applications* (Boulder, CO, Oct. 2006).
6. Calabrese, F., Pereira, F., DiLorenzo, G., Liu, L., and Ratti, C. The geography of taste: Analyzing cell-phone mobility and social events. In *Proceedings of the Eighth International Conference on Pervasive Computing* (Helsinki, May 2010).
7. Cuff, D., Hansen, M., and Kang, J. Urban sensing: Out of the woods. *Commun. ACM* 51, 3 (Mar. 2008), 24–33.
8. Girardin, F., Calabrese, F., Dal Fio, F., Ratti, C., and Blat, J. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing* 7, 4 (Oct–Dec. 2008), 36–43.
9. Girardin, F., Vaccari, A., Gerber, A., Biderman, A., and Ratti, C. Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management* (Hong Kong, June 2009).
10. González, M.C., Hidalgo, C.A., and Barabási, A.-L. Understanding individual human mobility patterns. *Nature* 453, 5 (June 2008), 779–782.
11. Hidalgo, C.A. and Rodríguez-Sickert, C. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications* 387, 12 (May 2008), 3017–3024.
12. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. Identifying important places in people’s lives from cellular network data. In *Proceedings of the Ninth International Conference on Pervasive Computing* (San Francisco, June 2011).
13. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., and Varshavsky, A. Ranges of human mobility in Los Angeles and New York. In *Proceedings of the Eighth International Workshop on Managing Ubiquitous Communications and Services* (Seattle, Mar. 2011).
14. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Rowland, J., and Varshavsky, A. A tale of two cities. In *Proceedings of the 11th ACM Workshop on Mobile Computing Systems and Applications* (Annapolis, MD, Feb. 2010).
15. Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., and Willinger, W. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th ACM Conference on Mobile Systems, Applications, and Services* (Lake District, U.K., June 2012).
16. Mun, M., Reddy, S., Shilton, K., Yau, N., Burke, J., Estrin, D., Hansen, M., Howard, E., West, R., and Boda, P. PEIR, the personal environmental impact report as a platform for participatory sensing systems research. In *Proceedings of the Seventh ACM Conference on Mobile Systems, Applications, and Services* (Krakow, Poland, June 2009).
17. N.J. Department of Transportation; <http://www.state.nj.us/transportation/>
18. Pulselli, R., Romano, P., Ratti, C., and Tiezzi, E. Computing urban mobile landscapes through monitoring population density based on cellphone chatting. *International Journal of Design and Nature and Ecodynamics* 3, 2 (2008).
19. Ratti, C., Pulselli, R.M., Williams, S., and Frenchman, D. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33, 5 (2006), 727–748.
20. Reades, J., Calabrese, F., Sevtsuk, A., and Ratti, C. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing* 6, 3 (July–Aug. 2007), 30–38.
21. Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. Limits of predictability in human mobility. *Science* 327, 5968 (Feb. 2010), 1018–1021.
22. Thiagarajan, A., Ravindranath, L.S., Balakrishnan, H., Madden, S., and Girod, L. Accurate, low-energy trajectory mapping for mobile devices. In *Proceedings of the Eighth USENIX Symposium on Networked Systems Design and Implementation* (Boston, Mar. 2011).
23. U.S. Bureau of Transportation Statistics. Washington, D.C.; <http://www.transtats.bts.gov>
24. U.S. Census Bureau. Washington, D.C.; <http://www.census.gov>

**Richard Becker** ([rab@research.att.com](mailto:rab@research.att.com)) is a Member of Technical Staff at AT&T Labs - Research, Florham Park, NJ.

**Ramón Cáceres** ([ramon@research.att.com](mailto:ramon@research.att.com)) is a Lead Member of Technical Staff at AT&T Labs - Research, Florham Park, NJ.

**Karrie Hanson** ([karrie@research.att.com](mailto:karrie@research.att.com)) is a Lead Member of Technical Staff at AT&T Labs - Research, Florham Park, NJ.

**Sibren Isaacman** ([isaacman@cs.loyola.edu](mailto:isaacman@cs.loyola.edu)) is an assistant professor of computer science at Loyola University Maryland, Baltimore.

**Ji Meng Loh** ([loh@njit.edu](mailto:loh@njit.edu)) is an associate professor at the New Jersey Institute of Technology, Newark, NJ.

**Margaret Martonosi** ([mrm@princeton.edu](mailto:mrm@princeton.edu)) is the Hugh Trumbull Adams ’35 Professor of Computer Science at Princeton University, Princeton, NJ.

**James Rowland** ([jrr@research.att.com](mailto:jrr@research.att.com)) is Director of Applied Data Mining Research at AT&T Labs - Research, Florham Park, NJ.

**Simon Urbanek** ([urbanek@research.att.com](mailto:urbanek@research.att.com)) is a Principal Member of Technical Staff at AT&T Labs - Research, Florham Park, NJ.

**Alexander Varshavsky** ([varshavsky@research.att.com](mailto:varshavsky@research.att.com)) is a Senior Member of Technical Staff at AT&T Labs - Research, Florham Park, NJ.

**Chris Volinsky** ([volinsky@research.att.com](mailto:volinsky@research.att.com)) is Executive Director of Statistics Research at AT&T Labs - Research, Florham Park, NJ.

## Large genomic databases with interactive access require new, layered abstractions, including separating “evidence” from “inference.”

BY VINEET BAFNA, ALIN DEUTSCH, ANDREW HEIBERG, CHRISTOS KOZANITIS, LUCILA OHNO-MACHADO, AND GEORGE VARGHESE

# Abstractions for Genomics

HUMANS ARE A product of nature and nurture, meaning our phenotype (the composite of all outward, measurable, characteristics, including our health parameters) is a function of two things: our genotype (the DNA program in all cells) and the environment (all inputs to a human, like food and medicine). This arrangement is analogous to how the output of a program (such as a search engine)

is a function of both the program and the input (keywords typed by a user). Using the same input with a different program (such as Google search vs. Bing) can result in different output. In this analogy, the role of the medical professional is to provide information that is “diagnostic” (such as, “Is there a bug in the program based on observed output?”), “prognostic” (such as, “Can output/outcome be predicted, given specific inputs, like diet?”), or “therapeutic” (such as, “Can a spe-

cific input, like a drug, lead to the desired output?”). Also, the electronic medical record (EMR) of a patient can be viewed as an archive of previously acquired inputs and outputs.

Unlike computers, the human program is largely hidden. Hence, traditional medicine is “depersonalized,” with doctors providing treatment by comparing the patient’s phenotype (symptoms) against empirical observations of outputs from a large number of individuals. Limited customization is based on coarse classes, like “race.” All this changed with the sequencing of the human genome in early 2000 and the subsequent drop in costs from hundreds of millions of dollars to \$1,000 on small desktop sequencing machines. The ability to cheaply read the program of each human underlies the great promise of personalized medicine, or treatment based on symptoms and the patient’s distinctive DNA program.

We frame this point with a clas-

### » key insights

- **Making genomics interactive is potentially transformative, similar to the shift from batch processing to time sharing.**
- **Analogous to Internet layering, genome processing can be layered into an instrument layer, an evidence layer, and an inference layer.**
- **A declarative query language we call GQL enables automatic optimization and provenance and privacy checks more readily than procedural alternatives used today.**

sic example: The blood-thinner drug Warfarin is widely prescribed to prevent blood clots. Dosage is critical; too high and the patient can bleed to death, too low and the drug might not prevent life-threatening blood clots. Often, the right dosage is established through multiple visits to the clinic and regular testing. However, recent reports<sup>16</sup> suggest that knowledge of the patient's genetic program can help establish the right dosage. We outline this approach (genetic association and discovery workflow) in three steps:

*Collect samples.* Collect a sample of affected and “genetically matched” control individuals; then sample DNA and catalog variations;

*Identify variations.* Identify and report variations that co-segregate, or correlate, with the affected/control status of the individual; and

*Follow up with studies.* Follow up on the genetic basis of the correlation through costly studies and experiments in animal models and clinical trials; then transfer knowledge to the clinic.

Even with its success, the discovery approach involves complications: First, studies are resource-intensive, requiring identifying and sequencing large cohorts of individuals with and without a disease. Second, it is unclear how to apply study results to a specific individual, especially one genetically different from the inves-

tigated cohort. Finally, data reuse is difficult; significant computation is needed to dig out data from a previous study, and much care is required to reuse it. We contrast “discovery workflow” with “personalized medicine.” Here, a physician treating individual *A* may query a database for treatments suitable for patients with genetic variations similar to those of *A* or query for patients genetically similar to *A* for treatments and dosages that worked well for these patients.

The notion of computational tools enabling personalized medicine is gaining currency. Some groups have proposed specialized databases for cancer genomes,<sup>14</sup> though details are still emerging. Here, we take a more general view, allowing for broader access to genomic information and enabling both discovery and personalized medicine.

We start with a shift in perspective implicit in the personalized-genomics community. Instead of sequencing small numbers of people on an as-needed basis, we assume individuals are sequenced at birth and their personal genome is a part of their EMR, available to be queried by researchers and medical personnel. This scenario is realistic considering how quickly sequencing costs are falling. This shift in perspective enables personalized medicine and large-scale discovery (see Figure 1).

In choosing the Warfarin dosage

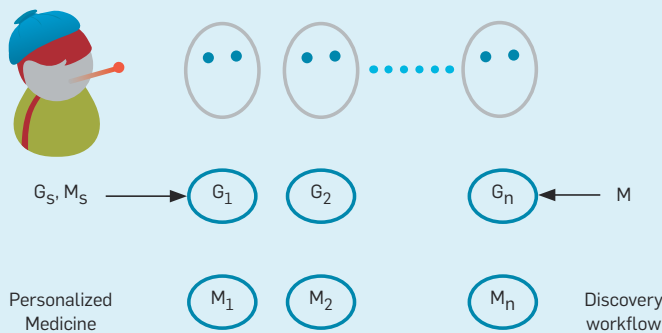
for a patient, a medical team might identify a collection of individuals genetically similar to the patient and on a Warfarin regimen; query their genomes and EMRs for genetic variation in candidate genes and Warfarin dosage, respectively; and choose the appropriate dosage based on the patient's specific genetic program. The ability to logically select from a very large database of individuals using the phenotype as a key removes the first problem with the discovery workflow. Using genetic variations specific to individual *A* as a key to return treatments (that work well for such variations) addresses the second problem. Finally, if the accompanying software system has good abstractions, then the third problem (computational burden to reuse data) is greatly eased. Here we focus on key software abstractions for genomics, suggesting that like other CS areas (such as VLSI/systems), software abstractions will enable genomic medicine.

We start with basic genetics using programming metaphors, then describe trends in sequencing and how genetic variations are called today and outline our vision for a vast genomic database built in layers; the key idea is the separation of “evidence” and “inference.” We then propose a language for specifying genome queries and end by outlining research directions for other areas of computer science to further this vision.

We limit our scope to genomics, ignoring dynamic aspects of genomic analysis (such as transcriptomics, proteomics expression, and networks). Genomic information is traditionally analyzed using two complementary paradigms: First, in comparative genomics, where different species are compared, most regions are dissimilar, and the conserved regions are functionally interesting.<sup>6,7</sup> The second is population genomics, where genomes from a single population are compared under the baseline hypothesis that the genomes are identical, and it is the variations that define phenotypes and are functionally interesting. We focus on population genomics and its application to personalized medicine and do not discuss specific sequencing technologies (such as strobe sequencing vs. color space encoding).

**Figure 1. Universal sequencing, discovery, and personalized medicine.**

Assume every individual is sequenced at birth. In discovery, clinical geneticists logically select a subset of individuals with a specific phenotype (such as disease) and another without the phenotype, then identify genetic determinants for the phenotype. By contrast, in personalized medicine medical professionals retrieve the medical records of all patients genetically similar to a sick patient *S*.




## Genetics for Computer Scientists


We start with a brief introduction to genetics for computer scientists; standard references (such as Alberts et al.<sup>1</sup>) provide more details. All living creatures consist of cells, each one like a computer running a program, or its DNA. The program uses three billion characters (nucleotides/base pairs, or bp) from a fixed alphabet  $\{A, C, G, T\}$ . Humans are diploid; that is, two programs control each cell, one inherited from the father and one from the mother. Further, each program is broken up into 23 “modules” called chromosomes, and within each chromosome are sparsely scattered small functional blocks called genes. The module pairs from the father and mother are called homologous chromosomes, with each human having a pair of (homologous) genes from each parent.

The “hardware” of the cell consists of cellular organelles and proteins—the cellular machinery. Proteins perform specific cellular functions (such as catalyzing metabolic reactions and transducing signals). The gene contains the “code” for manufacturing proteins, with each gene executing in one of many “ribosomes” (analogous to a CPU). Information travels from the nucleus (where the packaged DNA resides) to the ribosome via a “messenger” molecule (mRNA), essentially a copy of the coding DNA. The ribosome “reads” the code 3 bases (one codon) at a time; each codon is analogous to an OpCode instructing the ribosome to attach a specific amino acid to the protein sequence being constructed. The DNA program thus provides instructions for making the hardware that in turn performs all cellular functions.

A change (mutation) in the DNA can change the amino acid and, correspondingly, the cellular machinery resulting in a different phenotype (output). In the Mendelian paradigm, each of the two homologous copies of the gene controls one phenotypic trait (such as ability to clot blood). A mutation in one gene might affect the phenotype strongly (dominant), not at all (recessive mutation), or somewhere in between. Most phenotypes are complex, controlled by the paired copies



**The key to efficiency in genomics is the premise that an individual’s genetic record can be summarized succinctly by a much smaller list of individual genetic variations.**



of multiple genes. Nevertheless, DNA controls traits, so even the simplest queries on DNA are useful (such as “Compared to a ‘normal’ individual, have parts of the patient’s DNA program mutated?”).

Advances in sequencing have made it possible to cheaply scan an individual’s genetic program for mutations, or variations. First, a physical process is used to randomly shear genomic DNA into small inserts of size 500bp–10,000bp. The sequencing machine deciphers the DNA from small fragments, or reads (length  $L \approx 100$ bp) at one or both ends of the inserts. The genomic information is thus presented as a collection of small strings over A,C,G,T sampled from a random location on a (maternal or paternal) chromosome. It is natural to assume these fragments will be assembled like a giant jigsaw puzzle. However, such an assembly is complex and computationally expensive due to the large amount of repetitive portions in human genomes.

**Mapping and variation.** An alternative to assembly is to align, or map, the non-repetitive fragments of a sampled genome (the donor/patient genome) to a reference human genome. The current reference is a single (haploid) copy of each chromosome sampled from multiple individuals. Mapping involves finding a location on the reference where the genomic substring matches the query fragment up to a small number of errors. The errors might be sequencing errors or true variations in the query relative to the reference. Mapping works because string search is more tractable than assembly and any pair of human genomes is identical to one in 1,000 locations.

A true deviation in the donor sequence relative to the reference is called a variation. The simplest variation is a single nucleotide (or single character) variation (SNV). Recall that a donor genome consists of two copies; a variation that occurs in both copies is called homozygous, and a variation in only one copy is called heterozygous. The specific value of the variant is called an allele; for example, suppose the DNA at homologous chromosomes of individual *A* compared to the reference is

...ATG...GAGTA... Reference Assembly  
 ...ACG...GAGTA... Maternal chromosome 1  
 ...ATG...GAGCA... Paternal chromosome

Individual A is bi-allelic, or heterozygous, at the two SNV sites and has the genotype ...C/T...C/T..., and the genotypes are resolved into two haplotypes ...C...T..., ...T...C...

Sites containing SNVs that are prevalent in a population demarcate chromosomal positions as varying, or polymorphic. Consequently, these locations are called single nucleotide polymorphisms (SNPs). In discovery workflows, geneticists test populations to see if the occurrence of variation correlates, or associates, with the phenotype status of the individual.

So far we have discussed simple variations involving one or a small number of changes at a location. By contrast, geneticists also investigate structural variations in which large (1kbp up to several million bases) genomic fragments are deleted, inserted, translocated, duplicated, or inverted, relative to the reference.<sup>19</sup>

**Sequencing Trends**

Four technological trends are relevant for designing a genomic software architecture:

*Reduced cost.* While the Human

Genome Project (<http://www.genome.gov/>) cost \$100 million, human re-sequencing for redundant (15x) coverage now costs less than \$5,000 in the U.S., projected to fall below \$1,000. This implies universal sequencing may be realizable, and archiving and analysis, not sequencing, will dominate cost;

*Short read lengths.* New sequencing technologies largely sacrifice length and sequence quality to massive parallelism and high throughput. As “single-molecule” sequencing technologies emerge, reads may become long enough (~100Kbp) to allow de novo assembly. Nevertheless, raw reads will continue to be first-class entities;

*Prohibitive assembly costs, paired-end sequencing.* Repetitive regions cover 40% of the human genome. If the read length is shorter than the length of repetitive sequence, the genome cannot be assembled uniquely. Longer reads or reads sequenced from the ends of long clones (paired end reads) are necessary for resolving repeats and assembling sequences de novo. Sequenced reads are today mapped to a standard human reference to identify variations correlated to phenotype variation; and

*Computer system costs.* Some studies<sup>15,18</sup> have shown the cost of disk storage for genomes is now greater

(and decreasing more slowly) than the cost of sequencing.

We begin with exemplar queries on genomic data that illustrate the difficulty of genomic analysis and lack of consensus as to a best method. Abstractions must be flexible enough to handle a variety of methods.

**Variation Calling**

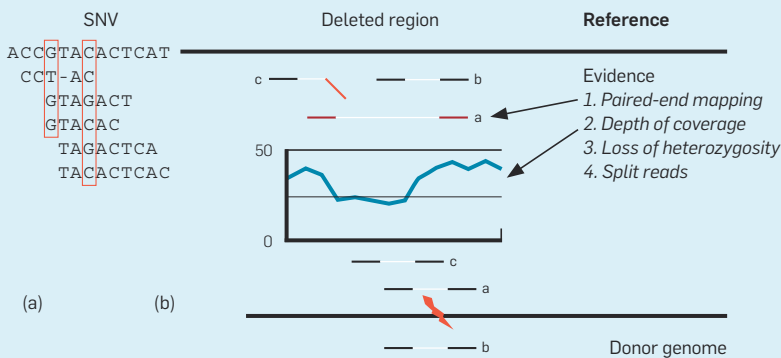
The key to efficiency in genomics is the premise that an individual’s genetic record can be summarized succinctly by a much smaller list of individual genetic variations. While we develop this premise further in our layering proposal, we provide insight as to how variants are called today; the expert should skip this section and proceed to our layering proposal. We start with querying for SNVs in the donor genome, the simplest form of variation:

**Calling SNVs.** Figure 2 outlines how a mutation may be called. Consider the reference allele C. We see two copies of the donor genome with a G allele and some copies with a C, indicating a heterozygous SNV. If the variation is homozygous, all overlapping reads would be expected to have a G in that position, though even this simple call can be confounded. Some reads may have been mapped to the wrong place on the reference (such as the top donor read in the figure). The G/T mutation may not be correct, and the alignment of the incorrectly mapped read might present many variations. Even if the read is mapped correctly, sequencing errors could incorrectly appear as heterozygous mutations.

Mutation callers use statistical methods informed by mapping the quality of the read (such as number of potential places in the genome a read can map to), the quality score of a base call, and the distribution of bases or alleles in the reads for that location. Some mutation callers use evidence based on the surrounding locations (such as an excess of insertion/deletion events nearby suggesting alignment problems). The decision itself could be based on frequentist, Bayesian inference, or other machine-learning techniques. While SNP callers use various inference techniques, all refer to the same evidence—the set

**Figure 2. Evidence for variation in the donor.**

(a) Evidence for SNVs is provided by aligning donor reads against the reference sequence; the G/T variation might be a sequencing error, as the variant reads maps with too many errors, though the G/C variation appears to be a true SNV. (b) Paired-end sequencing and mapping provides evidence for deletion in the genome; the dotted rectangle demarcates the region in the reference deleted in one of the two donor chromosomes. Read “a” samples the region around the deletion (marked with the lightning bolt), mapping “discordantly” in the reference; read “b” maps concordantly, but with coverage of only about half of neighboring regions; and read “c” is sampled from the breakpoint, mapping at only one end.





of reads overlapping the location of the SNP in question.

**Calling structural variations.** In addition to small nucleotide changes, larger structural variations involving insertion, deletion, and translocation of large genomic regions are another important source of genomic variation. Consider the example of deletions in the donor genome (see Figure 2b) in which a large segment of DNA is deleted, relative to the reference. If both copies of the donor genome are deleted, the deletion is homozygous;

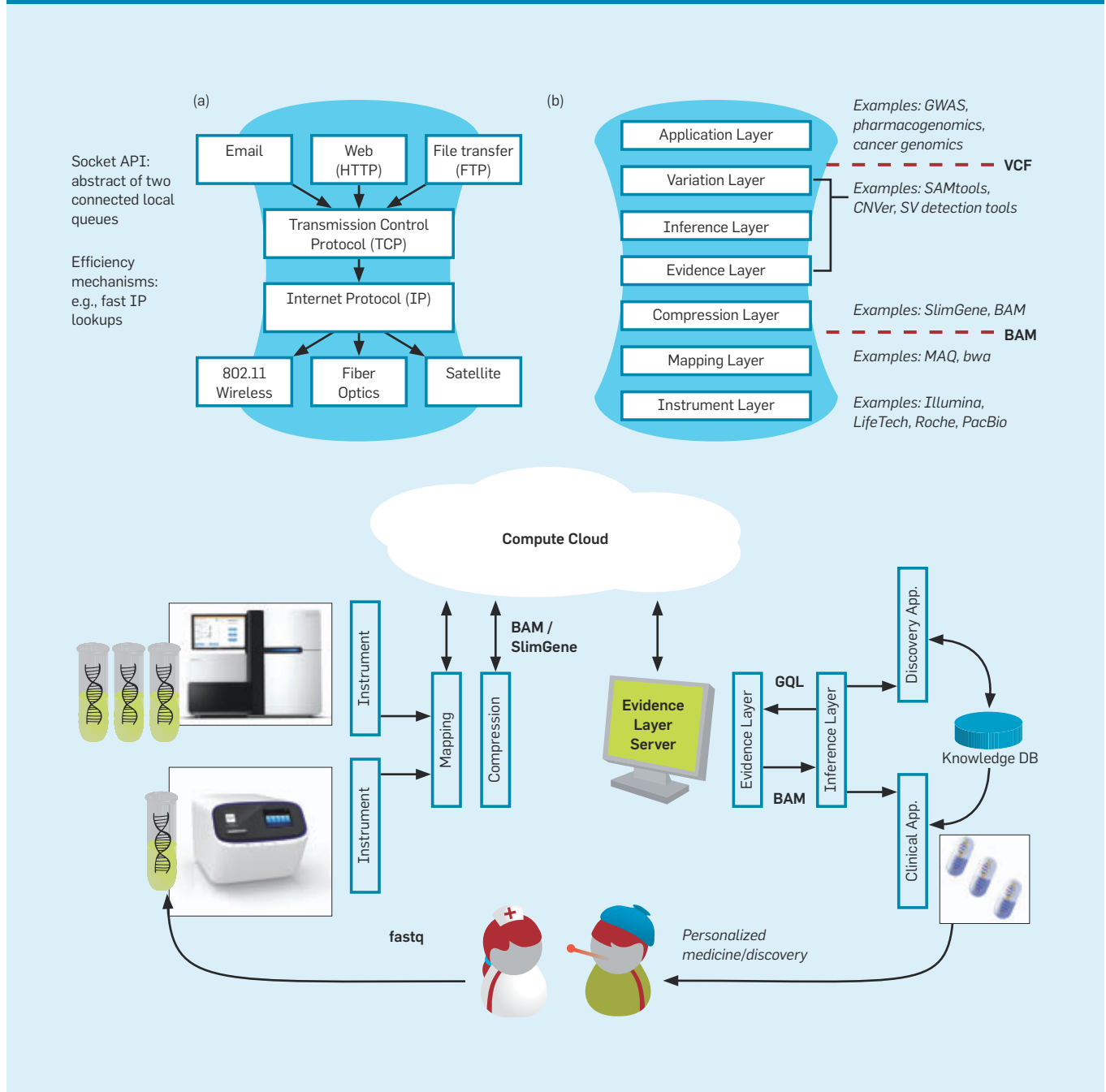
otherwise, it is heterozygous deletion. Deletions are detected through several techniques:

*Paired-end mapping.* Paired-end sequencing sequences both ends of large genomic fragments (sampled randomly from the donor genome). These fragments are size-selected to be tightly distributed around a specified length  $L$  ( $\approx 500$ ). If paired reads end up mapping much further apart than  $L$  (length discordance), a geneticist can infer a deletion in the donor relative to the reference (such

as read “a” in Figure 2b). If the deletion is heterozygous, the geneticist would see a mix of concordant and discordant reads at the breakpoints of the deletion.

*Depth of coverage.* “Depth” at a position refers to the count of reads mapped to the position. Deleted regions of the donor chromosome will have reduced coverage—roughly half for heterozygous deletions and zero for homozygous ones. Thus, read “b” in Figure 2b maps within the deleted region, but reads “a” and “c” do not.

Figure 3. Layers of genomic processing software.



*Single-end mapped and split-reads.* When a read maps to the breakpoint of the deletion on the donor it cannot be mapped back to the reference (Figure 2b, read “c”). In the case of a “clean” deletion, the prefix and suffix of the fragment can be mapped separately; such split-reads are indicative of deletion events.

*Loss of heterozygosity.* Consider the SNV locations on the donor genome. While sampling multiple polymorphic sites, a geneticist would expect a mix of heterozygous and homozygous sites. At a deletion, the single chromosome being sampled displays a loss of heterozygosity.


Even within the constraints of these four categories, a number of design decisions must be made by software tools to account for repetitive sequences and to reconcile conflicting evidence. Variant inference remains a challenging research problem.

### Layering for Genomics


Our vision is inspired by analogy with systems and networks; for example, the Internet has dealt with a variety of new link technologies (from fiber to wireless) and applications (from email to social networks) via the “hourglass” model using the key abstractions of TCP and IP (see Figure 3a).

Similarly, we propose that genomic-processing software be layered into an instrument layer, a compression layer, an evidence layer, an inference layer, and a variation layer that can insulate genomic applications from sequencing technology. Such modularity requires computer systems to forgo efficiencies that can be gained by leaking information across layers; for example, biological inferences can be sharpened by considering which sequencing technology is used (such as Illumina and Life Technologies), but modularity is paramount.

Some initial interfaces are in vogue among geneticists today. Many instruments now produce sequence data in the “fastq” format. The output of mapping reads is often represented as “SAM/BAM” format, though other compressed formats have been proposed.<sup>10</sup> At a higher level, standards (such as the Variant Call Format, or VCF) are used to describe variants (see Figure 3a).



**GQL also supports multiple types of inference, changing definitions of variation and pooling evidence across instrument types.**



We propose additional layering between the mapped tools and applications. Specifically, our architecture separates the collection of evidence required to support a query (deterministic, large data movement, standardized) from the inference (probabilistic, comparatively smaller data movement, little agreement on techniques). While inference methods vary considerably, the evidence for inferences is fairly standard. To gather it in a flexible, efficient manner, we propose a Genome Query Language (GQL). Though we do not address it here, a careful specification of a variation layer (see Figure 3a) is also important. While the data format of a variation is standardized using, say, VCF, the interface functions are not.

**The case for an evidence layer.** Genomes, each several hundred gigabytes long, are being produced at different locations around the world. To realize the vision outlined in Figure 1, individual laboratories must be able to process them to reveal variations and correlate them with medical outcomes/phenotypes at each place a discovery study or personalized medicine assay is undertaken. The obvious alternatives are not workable, as described in the following paragraphs:

*Downloading raw data.* Transporting 100Gb for each of 1,000 genomes across the network is infeasible today. Compression can mitigate (5x) but not completely avoid the problem. Massive computational infrastructure must be replicated at every study location for analysis.

*Downloading variation information.* Alternatively, the genomic repositories could run standard-variant-calling pipelines<sup>4</sup> and produce much smaller lists of variations in a standard format (such as VCF). Unfortunately, variant calling is an inexact science; researchers often want to use their own callers and almost always want to see “evidence” for specific variants. Discovery applications thus very likely need raw genomic evidence. By contrast, personalized genomics applications might query only called variants and a knowledgebase that correlates genotypes and phenotypes. However, even medical personnel might occasionally need to review the raw evidence for critical diagnoses.

Our approach provides a desirable compromise, allowing retrieval of evidence for variations on demand through a query language. The query server itself uses a large compute (cloud) resource and implements a query interface that returns the subset of reads (evidence) supporting specific variations. Some recent approaches have indeed hinted at such an evidence layer, including SRA and Samtools, but in a limited scenario useful mainly for SNV/SNP calling. The Genome Analysis Toolkit (<http://www.broadinstitute.org/gatk/>) provides a procedural framework for genome analysis with built-in support for parallelism. However, our approach—GQL—goes further, allowing declarative querying for intervals with putative structural variation (such as with discrepant reads supporting a deletion) or copying number changes. GQL also supports multiple types of inference, changing definitions of variation and pooling evidence across instrument types.

Consider this complex biological query: Identify all deletions that disrupt genes in a certain biological network and the frequency of the deletions in a natural population. For any statistical-inference algorithm, the evidence would consist of mapped reads that satisfy certain properties, including: length-discordant reads; reads with reduced depth of coverage; and reads with one end unmapped. The evidence layer supports queries to get those reads and delivers the following benefits:

*Alternate forms of evidence.* The separation allows inference-layer designers to start thinking of alternate forms of evidence to improve the confidence of their queries (such as split-end reads that map to the deletion breakpoints);

*The cloud.* The evidence layer can be a data bottleneck, as it involves sifting through large sets of genomic reads. By contrast, the inference layer may be compute-intensive but typically works on smaller amounts of data (filtered by the evidence layer). The evidence layer can be implemented in the cloud, while the inference layer can be implemented either in the cloud or on client workstations; and

*Moving target.* A standardized evidence layer gives vendors time to cre-

ate a fast, scalable implementation; by contrast, the inference layer is today a moving target.

Here, we develop this intuitive idea further by describing GQL, which is being developed to support the evidence layer:

### Querying the Genome via GQL

We wanted to develop a query language that is complete (capable of handling all evidence level queries), efficient, and easy to express, and that uses standard input/output. Ideally, the language would allow selecting from a set of reads we call READS and output a subset of reads in a standard format, like BAM. GQL uses a standard SQL-like syntax that is easy to use and familiar to most programmers. However, a standard relational database does not work well.

GQL includes two fundamental types of relations: Reads mapped to the human genome often expressed as BAM files and tables of intervals representing “interesting” (functional) regions of the genome. While simple selection queries are exactly the same as in relational languages (select from reads where mapped pairs are far apart), many useful queries require joining relations using interval intersection, not equality as the join operator; for example, a geneticist might want to join a relation consisting of READS that map to gene exons, the gene regions that translate to proteins, but where the paired ends are far apart, indicating a deleted exon

(see Figure 4).

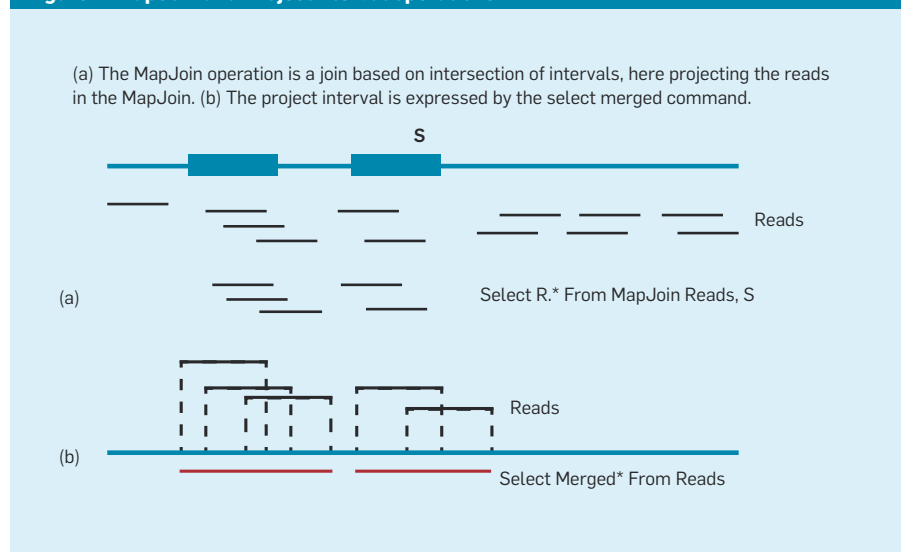
GQL defines a special MapJoin operator to allow this, and some of the newer sequencing technologies allow multiple reads to be generated from the same physical clone. While not explicitly discussed here, the relations can be extended to this case. We also find it extremely helpful to have a third operator we call “project interval” to compute the largest continuous interval containing a merged representation of what may be a number of smaller intervals; example queries using these operations are outlined in Figure 4.

**Sample queries.** GQL is based on a genome query algebra, and, here, we discuss several examples of its expressive power. In a companion technical paper (to be published elsewhere), we show GQL’s expressive power captures the language of first-order logic over the relations discussed earlier, as well as a signature of aggregation functions.

*What is the genotype at a specific position (such as SNV)?*

*Query.* Define an interval by the triple  $\langle \text{chr}, \text{beg}, \text{end} \rangle$ , signifying the beginning and ending coordinates on a specific chromosome. Let the SNV of interest be located at a point interval A ( $\langle \text{chr}, \text{beg}=i, \text{end}=i \rangle$ ). The evidence for the genotype is provided by alignments of reads that map to the location; we can either query for the mapped reads or for the alignments themselves, which are often stored as a mapped read attribute (such as

Figure 4. MapJoin and ProjectInterval operations.



R.ALIGNSTR). Thus

```
GQL: SELECT R.ID, R.ALIGNSTR
FROM MAPJOIN R, A
```

*What are the diploid haplotypes (phased genotypes) across a set of linked loci in a dataset?*

*Query.* This query is more challenging than the first. Assembling haplotypes requires a collection of reads, each (perhaps along with their paired-end reads) connecting at least two polymorphic sites. Let attribute  $R.CloneId$  denote the clone identifier so the paired-end reads  $r_1, r_2$  derived from the same clone satisfy  $r_1.CloneId = r_2.CloneId$ . Also, let relation  $S$  denote the collection of point intervals, one for each variant locus.

(a) Find a subset of reads mapping to the loci and the count of sites the reads or their paired-ends map to (call it count  $c$ )

```
GQL: RC = SELECT R.CloneId, c =
count(*)
FROM MAPJOIN R, S
GROUPBY R.CloneID
```

(b) Return IDs of reads with count  $\geq 2$

```
GQL: SELECT R.ID
FROM R, RC
WHERE R.CloneID = RC.CloneID
AND (RC.c  $\geq$  2)
```

*What genomic loci are affected by copy number variations (CNVs)?*

*Query.* If the number of donor reads mapping to a region exceeds some threshold  $T$  then the inference might be the region has been duplicated in the donor genome. Such CNVs have been implicated as an important variation for many disease phenotypes. To gather evidence, a geneticist would look to identify all intervals where the number of mapped reads exceeds, say, threshold  $t$ . Let  $G.loc$  denote a specific chromosome and location.

(a) Compute for each location the number of reads that map to the location

```
GQL: V = SELECT G.loc, c = COUNT(*)
FROM MAPJOIN R, G
GROUPBY G.loc
```

(b) Return all “merged regions”

where the read count exceeds threshold  $t$

```
GQL: SELECT MERGED RS.loc
FROM V
WHERE V.c  $>$  t
```

*Identify all regions in the donor genome with large deletions.*

*Query.* As discussed earlier, the evidence for deletion comes from several sources. Suppose a user prefers discrepant paired-end mapping. Paired-end reads from clones of, say, length 500 should map  $\approx 500$ bp apart on the reference genome. If, instead, the ends happen to map discrepantly far (such as  $\ell$  apart for some  $\ell \gg 500$ , like  $\ell \approx 10,000$ ), they support the case for a deletion in the donor genome. The goal is to identify all regions with at least  $t$  discrepant paired-end reads:

(a) Use a join in which each record contains the mapping locations of the read, as well as its paired-end.

```
GQL: Table READS already contains
this join.
```

(b) Select records containing discrepant reads.

```
GQL: H2 = SELECT * FROM READS
WHERE abs(loc - mateLoc)  $>$  10,000
```

(c) Select intervals containing at

least  $t$  discrepant reads.

```
GQL: SELECT MERGED G.loc FROM H2
GROUPBY G.loc, c = count(*)
WHERE c  $>$  t
```

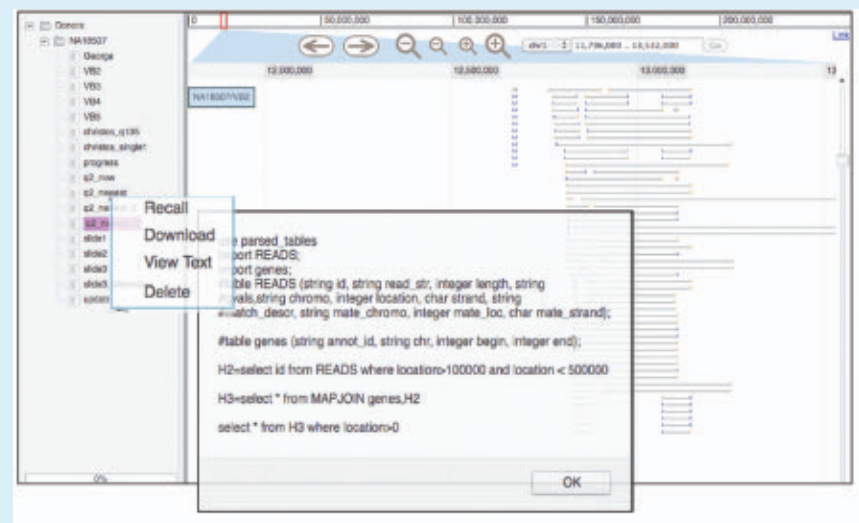
**Population-based queries.** The true power of querying genomes comes from the ability to query populations. Indeed, existing tools (such as Samtools) support the ability to extract reads from multiple individuals at specific locations corresponding to polymorphic sites. GQL extends the full power of genomic queries to interrogate populations. In the Warfarin example, the goal was to query for Warfarin dosage and genetic variation in candidate genes (identified through a discovery work flow) among individuals on the Warfarin regimen. This suggests the following query: “Report Warfarin dosage and genomic intervals and reads in individuals such that the copy number of mapped reads is at least twice the expected coverage in the interval.”

This query would be similar to that of a single individual but repeated via a “join” with a population  $P$ , using

```
GQL: SELECT * FROM P, MAPJOIN R, E
WHERE P.WD = TRUE
```

Using earlier arguments, GQL can be used to count the read depth and report high-CNV regions. A similar idea

**Figure 5. Prototype implementation of GQL, with visualization using the open-source tool jbrowse; included are discordant paired-end reads supporting deletions.**



applies to a personalized workflow, where a geneticist might be interested in patients with copy numbers similar to the specific query individual.

*Group inference without accurate individual inference.* The ability to query populations has an important benefit: Individual genomes may have little evidence for SNV calls at low-coverage sequencing. However, if a large number of affected individuals in a population (such as 800 out of 1,000) all show the same SNV while controls do not, an inference tool can reliably predict an association, even with unreliable calls for individuals. While more work is required to demonstrate the benefits of group inference, the point is that GQL provides query support for group inference.

### Prototype Implementation


We have developed a prototype implementation of a subset of GQL (see Figure 5). The uploaded genome (in BAM format) can be queried through a simple text interface that allows the user to write a GQL query. This query is compiled and executed and the output returned as a (smaller) BAM file that can be visualized through appropriate genome browsers (such as jbrowse, <http://jbrowse.org/>) or downloaded to the client for further analysis.

The implementation of “genome-query” has a customized parser that converts GQL to an intermediate representation. We included customized procedures for each of the algebraic operations, with some concessions for efficiency, mainly with respect to memory. Specifically, GQL uses interval trees to implement joins and customized indices (such as strength vectors) for efficient querying.


### Challenges

We have made the case for a set of genomic layers, including an evidence layer where evidence is retrieved through GQL. Successful implementation of this vision depends on new ideas from computer science:

**Query power (database theory).** Is GQL sufficiently powerful to address all evidence-layer queries needed in practice? The goal is to have the evidence layer handle as much data-intensive computation as possible while preserving performance; without the



**The true power of querying genomes comes from the ability to query populations.**



performance goal, any query can be trivially satisfied by passing the entire genome to the inference layer. Note that GQL’s expressive power coincides with that of first-order logic over the schema of the three relations  $R, G, P$ , a signature of aggregation functions, and a group-by operator. However, user feedback may require GQL developers to add extensions, and, in implementing extensions, care must be taken to balance expressive power with efficient evaluation.

### Query speed (database systems).

We designed a corresponding algebra GQA as an internal representation for optimization and for evaluating query plans; for example, assume queries on populations are automatically decomposed into queries on individuals. Consider queries of the general form  $\text{SELECT } \alpha \text{ FROM MAPJOIN } R, G \text{ WHERE } b$ . Two steps are needed to construct such a query:

1. Select for relations that satisfy constraints  $b$ ; and
2. Project (while removing duplicates) onto attributes  $\alpha$ .

GQL uses a location-based index  $\text{LTOR}$ , where  $\text{LTOR}(\ell)$  is a pointer to first read that maps to the point-interval  $l$ . For each individual, GQL keeps a compressed index of the mapped reads in memory. The index can be used for select operations based on specific locations (such as reads that map to specific genes).


However, many queries involve scanning the entire genome for maximal intervals; for example, find all maximal regions where there is a disproportionate increase in the number of mapped reads (high copy number). For efficient implementation of these queries, GQL constructs special indices that allow filtering for reads according to a user-defined constraint. Define a strength-vector  $S_\Theta$  for a constraint  $\Theta$  as a vector of length  $G$  (the entire genome). Any location  $l \in G$ ,  $S_\Theta[l]$  gives the strength of the evidence at that location and can be pre-computed for common constraints  $\Theta$ . To reduce memory, GQL also chooses a minimum strength cut-off and maintains  $C_{\Theta,t}$  as a sorted sequence of intervals  $i_1, i_2, \dots$  such that each  $i_j$  is a maximal interval satisfying  $S_\Theta[l] \geq t$  for all  $l \in i_j$ . The compressed vectors reduce memory and computation re-

quirements and can be recomputed on demand.


**EMRs (information retrieval).** The phenotypes associated with each sequenced individual are already in patient medical records. Initial results from the eMERGE network indicate that, for a limited set of diseases, EMRs can be used for phenotype characterization in genome-wide association studies within a reasonable margin of error.<sup>9,13</sup> We anticipate that most health-care institutions will be using EMRs by 2014, given incentives provided by the Health Information Technology for Economic and Clinical Health Act of 2009.<sup>17</sup> Increasing adherence to interoperability standards<sup>5</sup> and advances in biomedical natural language processing<sup>12</sup> make efficient querying possible. However, there is no integration of genotype and phenotype data today. GQL should be useful for both interrogating a single genome and interrogating multiple genomes across groups of individuals but will need to integrate with existing EMR systems so phenotype data can be queried together with genomes.

**Privacy (computer security).** The genome is the ultimate unique identifier. All privacy is lost once the public has access to the genome of an individual, but current regulation, based on the Health Information Portability and Accountability Act, is silent about it.<sup>2,3,11</sup> Though the Genetic Information Nondiscrimination Act addresses accountability for the use of genetic information<sup>8</sup> privacy laws must change to ensure sensitive information is available only to the appropriate agents. Checking that a given study satisfies a specific privacy definition requires formal reasoning about the data manipulations that generated the disclosed data—impossible without a declarative specification (such as GQL) of such manipulations.

**Provenance (software engineering).** GQL is an ideal way to record the provenance of genomic study conclusions. Current scripts (such as GATK) often consist of code that is too ad hoc for human readability and span various programming languages too low level for automatic analysis. By contrast, publishing the set of declarative GQL queries along with their results would



**While the work is a challenge, making genetics interactive is potentially as transformative as the move from batch processing to time sharing.**



significantly enhance the clarity and reproducibility of a study's claims.

Provenance queries also enable scientists to reuse the data of previously published computing-intensive studies. Rather than run their costly queries directly on the original input databases, these scientists would prefer to launch an automatic search for previously published studies in which provenance queries correspond to (parts of) the computation needed by their own queries. The results of provenance queries can be directly imported and used as partial results of a new study's queries, skipping re-computation. This scenario corresponds in relational database practice to rewriting queries using views.

**Scaling (probabilistic inference).** Learning the correlation between diseases and variations can be tackled differently if there are a large number of genomes. It may be less critical to accurately evaluate individual variations for such a discovery problem, as erroneous variations are unlikely to occur over a large group of randomly selected individuals. More generally, do other inference techniques leverage the presence of data at scale? As an example, Google leverages the big-data collections it has to find common misspellings. Note that accurately screening individual variations is still needed for personalized medicine.

**Crowd sourcing (data mining).** Crowdsourcing might be able to address difficult challenges like cancer,<sup>14</sup> but the query system must first have mechanisms to allow a group to work coherently on a problem. Imagine that a group of talented high-school science students are looking for genetic associations from cases and controls for a disease. A potentially useful GQL mechanism would be to select a random subset of cases and controls that are nevertheless genetically matched (arising from a single mixing population). Researchers could then query for a random subset of 100 individuals with a fraction of network bandwidth while still providing similar statistical power for detecting associations.

**Reducing costs (computer systems).** Personalized medicine must be commoditized to be successful so requires computer systems research; for example, since most genomes are

read-only, is there some way to leverage solid-state disks? Efficient decomposition between the cloud and a workstation is key to reducing data traffic in and out of the cloud. While genomics has been dominated by costly parallel computers, it makes economic sense to adapt genomic software to harness the parallelism in today's cheap multicore CPUs.

## Conclusion

Genomics is moving from an era of scarcity (a few genomes with imperfect coverage) to abundance (universal sequencing with high coverage and cheap re-sequencing when needed). This shift requires geneticists and computer scientists alike to rethink genomic processing from ad hoc tools that support a few scientists to commodity software that supports a world of medicine. The history of computer systems teaches us that as systems move from scarcity to abundance, modularity is paramount; ad hoc software must be replaced by a set of layers with well-defined interfaces. That these trends have been recognized by industry is seen in the shift from machine-specific formats (such as Illumina) to standards (such as BAM) and from vendor-specific variant formats to VCF. The 1000 Genomes Project (<http://www.1000genomes.org/>) has gained momentum, with a large number of sequences accessible today. However, much progress has involved defining data formats without powerful interface functionality; using an Internet analogy, it is as if TCP packet formats were defined without the socket interface.


We propose going beyond the layering implicit in current industry standards to enabling personalized medicine and discovery. We advocate separation of evidence from inference and individual variation from variation across groups, as in Figure 3a. We propose a specific interface between the evidence layer and the inference layer via the GQL. While GQL is based on a relational model using a virtual interval relation, further development is required beyond standard relational optimization to allow GQL to scale to large genomes and large populations.

Here, we described several benefits from separating evidence from

inference; for example, a genome repository accessible by GQL offers the ability to reuse genomic data across studies, logically assemble case-control cohorts, and quickly change queries without ad hoc programming. GQL also offers the ability to reduce the quality of inference on an individual basis when applying group inference over large populations. We also described simple ideas for scaling GQL to populations using compressed-strength indices and for doing evidence-layer processing in the cloud.

We emphasized that GQL and the evidence layer are only our initial attempt at capturing abstractions for genomics. We hope to prompt a wider conversation between computer scientists and biologists to tease out the right interfaces and layers for medical applications of genomics. Beyond abstractions much work remains to complete the vision, including better large-scale inference, system optimizations to increase efficiency, information retrieval to make medical records computer readable, and security mechanisms. While the work is a challenge, making genetics interactive is potentially as transformative as the move from batch processing to time sharing. Moreover, computer scientists only occasionally get a chance to work on large systems (such as the Internet or Unix) that can change the world.

## Acknowledgments

This work was funded in part by the National Institutes of Health-sponsored iDASH project (Grant U54 HL108460), NIH 5R01-HG004962, and a Calit2 Strategic Research Opportunity scholarship for Christos Kozanitis. We thank Rajesh Gupta, Vish Krishnan, Ramesh Rao, and Larry Smarr for their useful discussions and support. 

## References

1. Alberts, B. et al. *Molecular Biology of the Cell*. Garland Science, New York, 2007.
2. Annas, G.J. HIPAA regulations: A new era of medical-record privacy? *New England Journal of Medicine* 348, 15 (Apr. 2003), 1486–1490.
3. Benitez, K. and Malin, B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Association* 17, 2 (Mar. 2010), 169–177.
4. De Pristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA-sequencing data. *Nature Genetics* 43, 5 (May 2011), 491–498.

5. Dolin, R.H. and Alschuler, L. Approaching semantic interoperability in Health Level Seven. *Journal of the American Medical Association* 18, 1 (Jan. 2011), 99–103.
6. Haussler, D. David Haussler. *Nature Biotechnology* 29, 3 (Mar. 2011), 243–243.
7. Haussler, D. et al. Genome 10K: A proposal to obtain hole-genome sequence for 10,000 vertebrate species. *Journal of Heredity* 100, 6 (Nov. 2009), 659–674.
8. Hudson, K.L., Holohan, M.K., and Collins, F.S. Keeping pace with the times: The Genetic Information Nondiscrimination Act of 2008. *New England Journal of Medicine* 358, 25 (June 2008), 2661–2663.
9. Kho, A.N. et al. Electronic medical records for genetic research. *Science Translational Medicine* 3, 79 (Apr. 2011).
10. Kozanitis, C., Saunders, C., Kruglyak, S., Bafna, V., and Varghese, G. Compressing genomic sequence fragments using SlimGene. *Journal of Computational Biology* 18, 3 (Mar. 2011), 401–413.
11. Malin, B., Benitez, K., and Masys, D. Never too old for anonymity: A statistical standard for demographic data sharing via the HIPAA privacy rule. *Journal of the American Medical Association* 18, 1 (Jan. 2011), 3–10.
12. Nadkarni, P.M., Ohno-Machado, L., and Chapman, W.W. Natural language processing: An introduction. *Journal of the American Medical Association* 18, 5 (Sept. 2011), 544–551.
13. Pathak, J. et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: The eMERGE network experience. *Journal of the American Medical Association* 18, 4 (July 2011), 376–386.
14. Patterson, D. Computer scientists may have what it takes to help cure cancer. *The New York Times* (Dec. 5, 2011).
15. Pennisi, E. Will computers crash genomics? *Science* 331, 6018 (Feb. 2011), 666–668.
16. Schwarz, U.I. et al. Genetic determinants of response to Warfarin during initial anticoagulation. *New England Journal of Medicine* 358, 10 (Mar. 2008), 999–1008.
17. Stark, P. Congressional intent for the HITECH Act. *American Journal of Managed Care* 16, 12 (Dec. 2010), 24–28.
18. Stein, L.D. The case for cloud computing in genome informatics. *Genome Biology* 11, 5 (May 2010), 207–214.
19. E. Tuzun et al. Fine-scale structural variation of the human genome. *Nature Genetics* 37, 7 (July 2005), 727–732.

**Vineet Bafna** (vbafna@cs.ucsd.edu) is a professor in the Department of Computer Science and Engineering of the University of California, San Diego.

**Alin Deutsch** (aldeutsch@ucsd.edu) is an associate professor in the Department of Computer Science and Engineering of the University of California, San Diego.

**Andrew Heiberg** (aheiberg@eng.ucsd.edu) is a master's student in the Department of Computer Science and Engineering of the University of California, San Diego.

**Christos Kozanitis** (ckozanitis@eng.ucsd.edu) is a Ph.D. student in the Department of Computer Science and Engineering of the University of California, San Diego.

**Lucila Ohno-Machado, MD** (machado@ucsd.edu) is Associate Dean for Informatics and Technology in the School of Medicine of the University of California, San Diego, and founding chief of its Division of Biomedical Informatics and a professor of medicine.

**George Varghese** (gvarghese@ucsd.edu) works at Microsoft Research, on leave from the Department of Computer Science of the University of California, San Diego.

DOI:10.1145/2398356.2398377

## A framework for evaluating security risks associated with technologies used at home.

BY TAMARA DENNING, TADAYOSHI KOHNO, AND HENRY M. LEVY

# Computer Security and the Modern Home

COMPUTATION IS EMBEDDED throughout our homes. Some devices are obvious: desktops, laptops, wireless routers, televisions, and gaming consoles. Increasingly, however, computational capabilities are appearing in our appliances, healthcare devices, children's toys, and the home's infrastructure. These devices are incorporating new sensors, actuators, and network capabilities: a Barbie with a video camera<sup>1</sup>; a lock for your front door controlled by your cell phone; or a bathroom scale that reports readings over your wireless network.<sup>26</sup> Many of these devices are also subject to control by servers external to the home, or are mobile technologies that regularly leave the home's perimeter and interact with other networks. These trends, which we expect to accelerate in the coming



years, create emergent threats to people's possessions, well-being, and privacy. We seek to survey the security and privacy landscape for devices in the home and provide a strategy for reasoning about their relative computer security needs.

Many human assets—whether electronic, physical, or nontangible items of value to end users—can be accessed or influenced from computing devices within the home; unsurprisingly, these assets are also potentially attractive targets to adversaries. The capabilities of new electronics and their presence in

### » key insights

- Homes are becoming increasingly computerized, filled with devices ranging from the traditional (laptops and desktops) to TVs, toys, appliances, and home automation systems.
- We survey potential computer security attacks against in-home technologies and their impact on residents; some of the attacks are familiar, but the new capabilities of home technologies enable novel attacks and allow some traditional attacks to have new consequences.
- We present a framework for articulating key risks associated with particular devices in the home, which includes identifying human assets, security goals, and device features that may increase the risk posed by individual technologies.





the home facilitate traditional crimes and allow new classes of attacks. Technically savvy burglars, for example, may use technology both to identify houses with expensive, easily resold items and to better plan and execute their crimes. Adversaries can also target technologies with a wide range of new capabilities, with the goal of accessing video and audio feeds,<sup>25</sup> unlocking doors or disabling home security,<sup>27</sup> tampering with home healthcare devices,<sup>13,26</sup> or interfering with home appliances and utilities.<sup>22,24</sup>

Fortunately, there have been few “high tech” crimes to date exploiting these new capabilities. Now is the right time to develop a foundation for securing the myriad devices within the home: before these technologies become more ubiquitous, communicative, and capable, and before real adversarial pressures emerge. While progress has been made in understanding security concerns for specific home technologies or categories of technologies,<sup>3,14</sup> there is currently a lack of unified vision for evaluating security threats posed by the assortment of consumer devices within the home. There are trade-offs in the design of any security system, but without a cohesive strategy for reasoning about home device security, product manufacturers will be left to determine the

appropriate trade-offs for themselves without best-practice references.

Our goal is first to survey the landscape of potential attacks, then to provide structure and guidance for reasoning about the differing security needs of home technologies. While many of these elements will be familiar from security for traditional computers, their implications are worth reassessing in the context of the home ecosystem. This article is also complementary to an existing body of research on security for home technologies, including work on the security needs and behaviors of users<sup>4,8,10,18,19</sup> and work on centralized security technology solutions.<sup>28</sup>

Table 1 presents an overview of the topics covered by this article. We begin by presenting an overview of how the ecosystem of home technologies can enable a range of attacks with electronic and physical consequences. Building on this discussion, we present the two key components of our strategy for evaluating the potential risks with home technologies: a taxonomy of security goals for home technologies; and a set of device characteristics that can be used to estimate a device’s potential risk to users. We apply our approach to three home technologies: a webcam toy, a networked scale, and a home automation

device. Our framework is not intended to be definitive, but rather informative: our intent is that this approach will provide a useful starting point for home technology stakeholders ranging from product manufacturers to consumer advocacy groups to the research community. Moreover, by focusing on the entire home technology ecosystem, our hope is that this work will strengthen the foundations for developing secure home technologies—with the ultimate goal of creating a trustworthy home environment for users.

### **The Big Picture: Challenges and Attacks**

The home technology space is interesting and unique from other domains. In a nutshell, the new home landscape takes four challenges—challenges that are not unique in and of themselves—and combines them to create a new problem space: (1) an extremely personal, asset-filled environment where there is (2) no dedicated, professional administrator to maintain a (3) heterogeneous collection of consumer technologies that (4) are increasingly cyber-physical and sensor-rich. The combination of these factors leads to an array of attacks and complicates the design of defenses for home devices. From a technical perspective,

**Table 1. An overview of topics discussed in this article.**

| Infection Pathways             | Human Assets         | Defensive Goals          | Device Risk Axes                                   |
|--------------------------------|----------------------|--------------------------|--|
| Physical                       | The Biosphere        | Device Privacy           | Potential Exposure to Attack                       |
| In-person                      | Emotional Well-being | Device Availability      | Communication Capabilities                         |
| Secondhand via Infected Device | Financial Well-being | Device Operability       | Communication Behavior                             |
| Found                          | Personal Data        | Command Authenticity     | The Cloud  |
| Gift                           | Physical Well-being  | Execution Integrity      | Software Updates                                   |
| Infected from Manufacture      | Relationships        | Data Privacy             | Configuration Defaults, User Interfaces, and Users |
| Lent                           | Societal Well-being  | Data Integrity           | Attractiveness as a Target                         |
| Returned                       |                      | Data Availability        | Technology Market Share                            |
| Used                           |                      | Environment Integrity    | Intended Users and Usage                           |
| Technological                  |                      | Activity Pattern Privacy | Sensors  |
| Remote or In-Network           |                      | Presence Privacy         | Actuators  |
| Direct Compromise              |                      | Occupant Identities      | Power  |
| Eavesdropping                  |                      | Sensed Data Privacy      | Connectedness                                      |
| Man-in-the-Middle              |                      | Sensor Validity          | Storage and Computation                            |
| Social Engineering             |                      | Sensor Availability      |  |

the home is filled with a diverse range of technologies with varying levels of security, hybrid communication structures, and no centralized security management system. From a human perspective, the home contains private and semi-private spaces shared by children, parents, siblings, elderly, roommates, and guests. Interpersonal dynamics, varying levels of security expertise, and different social and technical preferences all contribute to complicating the home technology security landscape. In order to effectively create and evaluate defenses, it is important to first understand the threat landscape.

**Attack Scenarios.** One unique aspect of the new home technology space is the vast array of attacks that it enables—many of which differ in effect from Web or desktop attacks. The increasing presence of electronics in the home—controlling our houses and coordinating our lives—provides unique opportunities for the technically savvy criminal.

Table 2 breaks down attacks into three tiers: low-level mechanisms, intermediate goals, and high-level goals. The low-level mechanisms listed in Table 1—such as denial-of-service attacks, tampering with logs, or eavesdropping on network traffic—will be familiar to anyone who has experience with computer security. However, the additional focus on sensors and actuators is something that is not generally encountered with traditional computing devices. Similarly, the high-level goals behind the attacks (blackmail, extortion, theft,

and vandalism, among others) are the same motivations that one encounters with all criminal activities. Arguably, the most novel aspects of attacks on the home ecosystem are the intermediate goals: the ways in which the unique capabilities of devices or the assets to which they have access enable criminal opportunities.

In order to highlight some of the unique properties of the home ecosystem, we list examples of attacks that are not viable with traditional computing platforms:

- ▶ Determining the locations of lucrative home burglary targets via camera feeds or the distinctive signatures of multiple, expensive devices;
- ▶ Providing access to homes that have cyber-physical locks that are vulnerable to electronic compromise;
- ▶ Checking whether or not a home is occupied (and by whom) via: cameras; microphones; motion sensors; logs for lights, thermostats, and door locks; or HVAC air pressure sensors;<sup>2,3</sup>
- ▶ Turning up the thermostat settings while the user is away in order to increase heating bills, thereby causing financial harm;
- ▶ Electronically manipulating a washing machine to cause flooding;
- ▶ Tampering with home healthcare technologies in order to change treatment, notifications, or perform a denial-of-service attack; and
- ▶ Targeting entire communities by coordinating their devices to overload the power grid.

*Attack targets.* For many types of attacks, an adversary could either attempt to target a particular person of interest or simply take advantage of known hardware and software flaws to indiscriminately attack any vulnerable victim. Attacks on a designated person require that the adversary identify useful exploits for the target’s particular technology configuration. On the other hand, for attacks on “low-hanging” targets—attacks of exploitative opportunity—the adversary need only focus on a known exploit and locate victims who are vulnerable to that exploit.

*The physical and the electronic.* At a high level, it is interesting that the presence of actuators and sensors in the new home environment allows interactions between the physical and electronic states of devices. It is possible to perform electronic attacks with physical consequences, but it is also possible to perform physical attacks with electronic consequences, or attacks that have both physical and electronic components. As an example of a physical attack that has electronic (then physical) consequences, an adversary might apply a bright, directed light source to an external light sensor in order to trick outdoor flood lighting into turning off. Similarly, one can imagine an attack where physically tricking a system sensor causes the system to enter a fail-safe mode that is more easily compromised via electronic attack.

**Infection Pathways.** The challenges of the home environment—such as its

heterogeneous topology and the idiosyncrasies of its occupants—help enable novel or complex infection pathways. Mobile devices, infrastructure electronics, cyber-physical systems, guest devices, and machines brought home from work all commingle in one hodgepodge environment, increasing the exposure to compromise. Understanding the potential infection pathways—particularly nontraditional pathways—that malware might follow to compromise a device helps us understand its exposure to risk, which we use later in our characterization of device risk. The Infection Pathways column of Table 1 provides an overview of the kinds of pathways that malware can take to infect a device in the home.

*Entry points.* There are a number of entry points an adversary could use to attack home technologies. Electronically, a device on the home network might be compromised by a direct attack from a device external to the home, or compromised by an infected device within the home (whether stationary, mobile, or belonging to a guest). If a device is mobile and connects to an infected network, it might become infected. Physically, a device might be infected by a manual interface such as USB or CD.<sup>5,9</sup> Alternative physical attack vectors include: receiving an infected device as a gift; purchasing a used, compromised device from a source such as eBay or Craigslist; purchasing a “new” device that has previously been purchased, infected, then returned; or purchasing a device that was infected during its manufacture.<sup>11</sup> Additionally, an adversary has a number of opportunities to socially engineer a user into installing malware, such as via app stores.<sup>15,21</sup> As another vector, an adversary could take advantage of the increasing number of “prosumers”—consumers who jailbreak their devices or perform similar automated modifications—whose devices allow behaviors that go beyond the capabilities expected by the manufacturer’s typical APIs and might not receive security software updates.

*Stepping back.* As this survey of the attack scenarios and infection pathways shows, the risks with computer security vulnerabilities in home technologies are quite varied and, in some cases, significant. Here, we present a framework for more methodically identifying and

prioritizing the security risks within the home.

### Human Assets and Security Goals

To design a system for defending home technologies, it is necessary to understand the human assets that are at stake and the desired security goals. We present a casual taxonomy of goals for protecting human assets in the home (also shown in the Defensive Goals column in Table 1). The general goals of confidentiality, integrity, authenticity, and availability are familiar security concepts; we frame the goals for defending the home slightly differently in order to highlight the domain in which they are applied and the unusual consequences of security failures. This taxonomy is meant to approach security and privacy goals from a variety of perspectives, and as such items are not mutually exclusive.

Security failures can result in a variety of kinds of harm to users. It is common to consider harm to users in terms of financial assets; it is less typical to consider damaging users by, for example, wasting their time or causing them stress. We suggest considering the potential negative impact of attacks on the following assets (in the Human Assets column in Table 1): emotional well-being, financial well-being, personal data, physical well-being, and relationships. In addition to considering the assets of individuals, it can be beneficial to consider the broader assets of societal well-being and impact on the biosphere. The list is

derived in part from Value Sensitive Design<sup>12</sup>—an area of human-computer interaction that focuses on what different individuals value—and in part from the discussion sections of papers on emerging technologies.<sup>5,7,16</sup>

**Device Goals.** These are security goals that pertain to the operation of traditional or embedded computing devices.

1. *Device privacy.* A device should avoid broadcasting or otherwise disclosing its presence (for example, a wireless electronic adult toy, a device to treat a stigmatized medical condition, or an expensive device that is attractive to thieves). Example harms include: emotional harm from shame or embarrassment; or financial or physical harm if a physical break-in occurs.

2. *Device availability.* A device should not suffer malicious service interruptions. In many cases, device unavailability might only cause irritation and result in wasted time; however, consequences can range from financial (for example, the user cannot perform some time-critical transaction) to physical (if the user is unable to properly use a medical device or if a non-functioning refrigerator spoils food).

► *Device operability.* A device should have protection against operating in a manner that could damage or destroy itself since the device is an investment of time and money.

3. *Command authenticity.* A device should only accept and send authentic commands that reflect the user’s inten-

**Table 2. An overview of the structure of attacks to the home ecosystem.**

|                            | Examples                    |                              |
|----------------------------|-----------------------------|------------------------------|
| <b>Low-level Mechanism</b> | Altering logs               | Viewing data                 |
|                            | Altering or destroying data | Viewing or altering traffic  |
|                            | DoS attacks                 | Viewing sensors              |
|                            | Using actuators             |                              |
| <b>Intermediate Goals</b>  | Accessing financial data    | Gathering incriminating data |
|                            | Causing device damage       | Misinformation               |
|                            | Causing environment damage  | Planting fake evidence       |
|                            | Causing physical harm       | Viewing private data         |
|                            | Enabling physical entry     |                              |
| <b>High-level Goals</b>    | Blackmail                   | Physical Theft               |
|                            | Espionage                   | Resource Theft               |
|                            | Exposure                    | Stalking                     |
|                            | Extortion                   | Terrorism                    |
|                            | Framing                     | Vandalism                    |
|                            | Fraud                       | Voyeurism                    |
|                            | Kidnapping                  |                              |

tion. This applies both to commands that elicit immediate reactions and commands that elicit delayed reactions (for example, turn on the sprinklers at 10 A.M.).

4. *Execution integrity.* A device should not deviate from its intended operating specification. More specifically, security vulnerabilities should not allow unintended behaviors that violate other security goals.

**Digital Data Goals.** These are security goals that pertain to a user's digital data.

1. *Data privacy.* Defenses should protect the confidentiality of the user's data (for example, leaked data could result in embarrassment, loss of reputation, financial damage, or legal repercussions due to possession of information or evidence of activities incompatible with local laws).


2. *Data integrity.* Defenses should ensure that the user's data is not corrupted. Non-critical data can be an inconvenience if lost (such as minor corruption of address book), but critical or irreplaceable data can present major emotional or logistical challenges (such as losing photos of deceased family members). Alternatively, undetected, intentional changes to data or the addition of new data could have legal (for example, illicit materials), financial (for example, inaccurate tax paperwork), emotional (for example, SMSs or email messages being sent to unintended recipients), or physical (such as inaccurate medical logs) consequences.

3. *Data availability.* Defenses should ensure the user's data does not suffer from malicious access interruptions.


**Environment Goals.** We must also consider security goals that pertain to the home infrastructure and general environmental conditions.

1. *Environment integrity.* Defenses should protect against single or multiple cyber-physical devices accepting commands that maliciously change the home environment—particularly if those changes might harm the home or its occupants (for example, lowering the thermostat could result in poor sleep, increased susceptibility to illness, or damage to water pipes).

2. *Activity pattern privacy.* Defenses should protect against accidentally revealing information about the activities of home occupants. Such disclosure could be the direct result of one data



**If a device is mobile, then the chances are higher that it will come into contact with malicious or infected networks or devices.**



source, or inference and cross-referencing from multiple sensors. Activity patterns could reveal information that is embarrassing (for example, intimate habits) or informative to a miscreant (for example, whether or not occupants are asleep). We consider two special cases:

► *Presence privacy.* Defenses should protect against accidentally revealing whether or not the home is occupied, as this can facilitate physical attacks on the home and enable cyber-physical attacks that might otherwise be detected and interrupted.

► *Occupant identities.* Defenses should protect against accidentally revealing the identities and number of occupants, thereby supporting freedom and privacy of association. As an example of privileged information, one may not wish to reveal that a young child is home alone.

3. *Sensed data privacy.* Defenses should protect against confidentiality leaks of sensor data (such as audio or video feeds) of shared and private home spaces.

4. *Sensor validity.* The readings from environmental sensors should be valid and immune to technical tampering. Sensor readings generally remain susceptible to tampering in the analog channel. Altered sensors might cause financial harm (for example, inaccurate power metering) and/or physical harm (for example, disabled home intrusion sensor facilitating a break-in). Alternatively, a miscreant who is unable to alter the function of a home system directly might instead tamper with sensor readings in an effort to alter the actions of the actuator in a feedback loop. In some scenarios, homeowners themselves may be considered the adversary (such as tampering with power meter readings to reduce billing<sup>17</sup> altering medical sensor readings for health insurance fraud).

5. *Sensor availability.* Sensor readings should be available without interruption according to their regular schedule. For example, the failure of a sensor can lead to physical harm or damage (such as the burglar alarm, the smoke detector, the temperature sensor in refrigerator).

Having explored human assets and security goals, we now explore a strategy for evaluating the potential risks with home technologies.

## Evaluating Potential Risks

The risk posed by a given home tech-

nology can be broken down into three components: the feasibility of an attack on the system; the attractiveness of the system as a compromised platform; and the damage caused by executing a successful attack. The first two factors, when combined, provide some indication of the likelihood that an adversary will compromise the device in question, while the third factor helps weight the overall risk. The human assets and security goals discussed previously provide a framework for reasoning about the impacts of potentially successful attacks. Here, we provide some guidelines for how to evaluate a device's exposure to attack and the likelihood of an attack attempt based upon the rough design characteristics of a technology (also summarized in the Device Risk Axes column of Table 1). Such a strategy for evaluation could be used by product designers, policymakers, or consumer advocacy groups.

**Potential Exposure to Attack.** In order to determine the risk posed by a home technology, it is necessary to evaluate how vulnerable the device is to an attack. It is difficult to make arbitrary evaluations of a technology's vulnerability without performing a hands-on study of the device in question; nonetheless, we provide some loose guidelines for design factors that tend to increase the likelihood that a device may be vulnerable to compromise. Those devices that are most likely to be vulnerable may deserve the most security consideration.

We stress that these guidelines indicate the likelihood of a potential vulnerability absent appropriate defenses, and are not an absolute measure of risk. Second, we stress that the list here is not exhaustive: instead, it focuses on some common issues that affect a device's attack surface. One would need to conduct a full security analysis of a product in order to more accurately gauge its level of security.

*Communication capabilities.* The more communication capabilities that a device possesses (for example, Wi-Fi, Ethernet, infrared, Bluetooth, ZigBee, cellular, powerline), the more media an adversary can use to attack the device. Manual communications capabilities such as USB or CD interfaces must also be considered.

*Communication behavior.* We consider three aspects of a device's communi-

cation behavior that affect its exposure to attack: initiated communications; receptiveness to incoming communications; and mobility. If a device is designed to communicate with a server or peer external to the home network, then a remotely located adversary has increased opportunities to attempt a range of passive and active attacks such as traffic eavesdropping, man-in-the-middle attacks, relay attacks, replay attacks, and spoofing. Additionally, a device's receptiveness to acting upon or replying to incoming network communications may also increase its exposure to attack.

If a device is mobile, then the chances are higher that it will come into contact with malicious or infected networks or devices.

*The cloud.* The flexibility and affordability of storage and computation in the cloud (such as software-as-a-service, platform-as-a-service, infrastructure-as-a-service) are causing more manufacturers to rely on the cloud for storage, backup, remote access, or configuration. If data is stored on those remote servers, then we must consider the risks to users if that data is exposed, altered, rendered inaccessible, or otherwise misused. By facilitating online configuration or remote access, manufacturers expose a different surface to attack—one that should not be overlooked even though it lies outside the physical boundaries of the home.

*Software updates.* The ability or inability to perform software updates can have positive or negative implications in a security context.<sup>2</sup> A device that connects to a manufacturer's server regularly to download updates may receive patches that remove vulnerabilities; however, if the update system does not properly verify that an update is legitimate or if that verification process is flawed, then an adversary has a convenient mechanism with which to modify a device's behavior.

*Configuration defaults, user interfaces, and users.* Defaults, user interfaces, and intended users all affect a device's security configuration (for example, sharing settings, account passwords, or update settings) and therefore its ultimate vulnerability to attack. A device with more secure default settings has an advantage over devices with less secure defaults, as some users never modify default con-

figurations. Entire research venues are dedicated to tackling issues surrounding configuration models and defaults.

Some user interfaces are rich whereas others are minimal. There are advantages and disadvantages with each. Rich interfaces have the potential to be confusing but can allow greater control over security settings. Rich interfaces can also inform users of security compromises and give them the ability to respond.

Similarly, it is important to consider the characteristics of the people who are most likely to administer the device. Different users might have different levels of security caution, different levels of familiarity with computers, or different priorities. For example, if a device resembles a toy or is meant to be used by children, then parents might give it to their children to administer, despite the child's likely lack of experience with computer security and different stance on privacy issues.

**Attractiveness as a Target.** To understand the risk posed by a home technology, it is also necessary to consider how much value the device holds for an adversary. A device's attractiveness to an adversary is relevant for two reasons: first, it affects the likelihood that an adversary will attempt to compromise the device. Second, the properties that cause a device to be of interest to an adversary are most likely the same properties that make the device a potential risk to users: after all, an adversary has some goal in attacking the device, and most of those goals cause direct or indirect harm to the user. We articulate here some of the capabilities and usage scenarios that make a device more attractive as an attack target.

*Technology market share.* If an adversary is intending to perform attacks of exploitative opportunity—attacks targeted at nonspecific vulnerable people rather than specific victims—then it is most efficient for the adversary to attack a technology that is deployed in many homes. Conversely, targeted attacks may better succeed with devices that have received less scrutiny due to a smaller market share.

*Intended users and usage.* Understanding a technology's most likely usage scenario helps indicate how valuable it would be to an adversary, since it dictates the assets with which the tech-

nology will interact. For example, a nanny cam would allow an adversary to spy on children; a networked storage server might hold backups of tax records or other financial data; and an electronically controlled door lock might allow full access to a home, whereas an electronic garage door opener would only allow access to the garage. While one cannot always anticipate how a device might be repurposed, it is important to consider future usage scenarios.

**Sensors.** If a device has sensors that record data then it might be a target of increased interest. The value of a sensor depends upon how much interest the raw or mined data holds for the adversary: for example, microphones and cameras have obvious value for voyeurs, blackmailers, or even private investigators or industrial spies; accelerometers might indicate whether or not a person is awake; and devices with GPS or Wi-Fi can be used to track an individual.

**Actuators.** A device holds increased value for an adversary if it can be used to effect changes in the physical world, since cyber-physical systems are both more efficient and less risky to use than physically traveling to a home. Cyber-physical effects of interest might include: locking or unlocking doors, cutting off electricity or water, changing

thermostat temperatures, controlling lights, and turning appliances such as fireplaces on or off.

**Power.** The power reserves and power schedule of a device affects its utility to an adversary. A device with limited battery life, such as a mobile phone or a universal Wi-Fi remote, has constraints on its usefulness. Alternatively, devices that are regularly unplugged or powered off by their users are not dependably accessible to the adversary.

**Connectedness.** A target might have value for an adversary either because it is likely to interact with many devices in the future—due to mobility or high network traffic—or because it will interact with a device of particular interest to the attacker; for example, an adversary might target a mobile device with the intention that it will later be able to infect networked-attached storage that houses financial data.

**Storage and computation.** Devices with large storage capabilities might be targeted to store illegal materials. Devices with smaller storage capabilities are less useful on their own, but could be used as part of a distributed storage botnet. Devices with large computational capabilities might also be attractive to adversaries with heavy computational tasks, such as farming Bitcoins or crack-

ing passwords. While there are additional properties that might affect a device's potential exposure to attack or its attractiveness as a target for attack, we chose to list the characteristics that we judged most significant and relevant for home technologies.

**Tying Things Together**

We tie together our framework with an example of how one might use it to analyze or compare the potential risks posed by different technology designs. We present a conceptual investigation of three technologies: a mobile webcam toy, a wireless scale, and a siren for a home security system. These technologies are not meant to be specific products, but rather amalgamations of products or exemplars of product categories. They represent a range of target audiences, technical capabilities, and application scenarios.

► **Mobile webcam toy.** Consider a mobile robotic webcam designed as a telecommunications toy for children. The toy can be used to drive around the house, chat with a friend, or communicate with a parent away on business. The toy broadcasts an ad hoc Wi-Fi wireless network to which a client computer can connect to view the webcam or drive the robot; alternatively, port forwarding can

**Table 3. An approximate risk evaluation of the three example technologies via potential exposure to attack and attractiveness of the attack target. The cells are color-coded to indicate the approximate severity of the concern: dark orange (serious), light orange (moderate), and light blue (minor).**

|                   | Communication Capabilities                                 | Communication Behavior  | Software Updates | Configuration Defaults, User Interfaces, and Users  | Market Share |
|-------------------|--|---|------------------|---|--------------|
| Mobile Webcam Toy | Long-range (Internet), short-range (Wi-Fi), USB (physical) | Communication with external server; Low inter-home mobility; Accepts incoming connections | Manual via USB   | Global default password; Minimal UI inputs <sup>1</sup> ; Minimal notification of connection (LED); Children admins | Marginal     |
| Wireless Scale    | Long-range (Internet), short-range (Wi-Fi), USB (physical) | With external server; Low inter-home mobility; Rejects incoming connections               | No               | No default data protection; Minimal UI inputs <sup>1</sup> ; No visual cue when data is accessed; Adult admin       | Marginal     |
| Security Siren    | Short-range (Z-wave)                                       | Low inter-home mobility; Highly connected to other automation devices                     | No               | Manual reset required to join automation network; No UI inputs <sup>2</sup> ; No UI feedback; Adult admins          | Marginal     |

1. Configured with PC via USB. 2. Programmed over short-wave.

be set up on the home router to allow the toy to be accessed from the Internet.

► *Wireless scale.* The second example technology we consider is a scale that wirelessly connects to an access point to send users' measurements over the Internet to their accounts on a server. Users can access their data, graphs, and trends via an online Web site or a smart-phone application.

► *Security siren.* The third technology is a siren that is part of a home automation or security system. The siren receives notification from entry sensors if a suspected break-in occurs and sounds an alarm. The various components in the home automation system communicate over short-range wireless.

Tables 3 and 4 present a high-level view of how our framework might be used to evaluate the approximate risk posed by these device designs. Interpretations and rankings of different risk levels are subjective and depend upon perspective. Table 3 considers the technologies according to the characteristics presented in the section "Evaluating Potential Risks." Table 4 summarizes the consequences that can result if the security goals discussed in the section "Human Assets and Security Goals" are not met. Color-coding provides an overview of the comparative risk patterns of the

different devices.

Our goal is not to be exhaustive or predictive. Rather, our goal is to facilitate an informed discussion about the potential risks with a technology if security is not sufficiently addressed in its design. To clarify, this framework only provides a skeleton for characterizing risks; individuals not accustomed to considering attack scenarios might require additional guidance.

*Mobile webcam.* Having populated the tables, we can now quickly assess the potential security risks with each technology. With its communications capabilities, communication behaviors, and user interface design, the mobile webcam toy clearly has significant potential exposure to attack (Table 3). Furthermore, with its proximity to children and its significant sensing capabilities (camera and microphone), the webcam toy appears to be a potentially attractive target to some adversaries (Table 3); more particularly, this device might be an attractive target to adversaries seeking to compromise the privacy of home occupants (Table 4). Given the high potential exposure to compromise, the attractiveness of the target, and the importance of the corresponding security goals, we would identify the mobile toy robot as a technology that merits signifi-

cant security review by product designers before the device enters the market. Similarly, based on the data in these tables, consumer advocacy groups would likely identify this device as one deserving post-market security auditing.

Fortunately, security best practices—if deployed—could significantly harden this device against attack: for example, the ability to perform authenticated software updates could allow the manufacturer to quickly address vulnerabilities once uncovered and strong audit logs could help further dissuade attack.

*Wireless scale.* Turning to the wireless scale, we see that although it does have some technical features that increase its potential exposure to attacks (Table 3)—particularly the inclusion of Wi-Fi capabilities—it is not a particularly attractive attack target and the associated security goals are not critical (Table 4). While there are arguments for trying to harden all devices against all possible attacks, that strategy is not feasible in practice. First, increasing security may impact the usability, desirability, or utility of the product. Second, companies do not have unlimited budgets to spend on security. These tables suggest that if a single manufacturer produced both the mobile webcam toy and the scale, the company would be well advised to focus

| Intended Users and Usage                              | Sensors                  | Actuators       | Power  | Connectedness                             | Storage and Computation |
|---|--------------------------|-----------------|--|---|-------------------------|
| Webcam used in the proximity of children              | Video camera, microphone | Wheels, speaker | Several hours continuous operation before recharge | High (externally addressable)             | Medium                  |
| Used by adults to weigh themselves                    | Pressure sensor          | None            | AA batteries                                       | Medium (not externally addressable)       | Low                     |
| Used to alert home owners and neighbors of burglaries | None                     | Speaker         | Continuous (plugged in)                            | Medium (connects with automation devices) | Low                     |

its security efforts on the webcam toy over the wireless scale; nevertheless, given the scale’s potential effects on emotional well-being, eating, or exercise activities, the integrity of sensor readings might become a security priority if the product were being marketed toward users with eating disorders (Table 4).

*Security siren.* Finally, we turn to the security siren. Table 4 suggests the primary security goals for the siren are related to device operability and command authenticity. If an attacker can disable the siren, then the attacker might be able to enter a home without alerting those nearby, thereby rendering the short-term benefits of the home alarm system ineffective; the home security system might still automatically call the police, but the police will not arrive immediately. Since the market share is listed as small in Table 3, the likelihood of an attacker choosing to target this system today seems small; however, the market share may increase over time. Having identified device operability as a particularly pertinent security goal, the device manufacturer can once again implement techniques to harden the device. For example, the device could issue a distinctive alert if a denial-of-service attack renders the siren unavailable to the rest of the home automation network.

The continuous sounding of an alarm could also cause a service interruption by tempting the user into turning off or ignoring the system; therefore, it is also important for the manufacturer to deploy defenses such as transmitting logs and incident reports to a monitoring agency.

*Stepping back.* As these examples illustrate, our framework can guide the analysis of potential security risks with technologies in the home. Devices in the home will likely incorporate varying degrees of security defenses, due in part to oversights by designers and developers, but also due to the costs associated with implementing security measures. By methodically evaluating a device’s potential exposure to attack and its attractiveness to adversaries (Table 3), as well as the potential impacts on security goals and human assets if the device is compromised (Table 4), one can assess the degree to which security might be important for a given device, as well as which security goals are the most important to address. This information can help developers focus their energies on the most significant risks of a design and help consumer advocacy groups direct their attention toward the computer security properties of the most concerning home technologies.

**Conclusion**

Our homes are increasingly becoming hubs for technologies with a wide variety of capabilities. While it would be ideal to strive for “perfect” security on all consumer devices, the reality is that resources such as time and money constrain these efforts. In the coming years, it will become increasingly important to improve the efficacy, interoperability, and usability of computer security solutions for the home. It remains to be seen what such a security solution would look like. It might take the form of a centralized security console that displays and controls device permissions and traffic.<sup>28</sup> The security system could incorporate trusted hardware, network intrusion detection systems, tiered security,<sup>6,20</sup> or cryptographic trust evidence of past transactions or device state.

We need a strategy for how to secure devices in the home. We need to understand the potential risks: risks that are a function of a device’s potential exposure to attack, its attractiveness as an attack target, and the potential impacts on human assets if the device is compromised. In this article, we explored the landscape of technological attacks on the home and provided a strategy for thinking about security in the home. In particular, we have identified human

**Table 4. An approximate risk evaluation of the three example technologies considering how human assets might be impacted if defensive goals are not met. The cells are color-coded to indicate the approximate severity of the concern: dark orange (serious), light orange (moderate), and light blue (minor).**

|                   | Device Privacy  | Device Operability                                    | Device Availability                  | Command Authenticity   | Data Privacy                                    | Data Integrity   | Data Availability       |
|-------------------|---|---|--------------------------------------|--|---|--|-------------------------|
| Mobile Webcam Toy | Device is interesting target                                  | Replaceable but not cheap; Non-essential device       | Non-essential                        | Potential minor property damage; Could send spam or launch similar attacks | Videos of household, including children         | Could add disturbing images or sounds into stream                | Non-essential           |
| Wireless Scale    | Device is not sensitive; Not a theft target                   | Replaceable but not cheap; Non-essential device       | Non-essential                        | Could send spam or launch similar attacks                                  | Weights are private; Online account credentials | Inaccurate weights could cause shame, affect eating and exercise | Non-essential           |
| Security Siren    | Device is interesting target, may indicate affluent household | Replaceable; Destruction would disable security siren | If unavailable weakens home security | Continuous alarm an annoyance, could cause user to disable or ignore alarm | N/A—does not store data                         | N/A—does not store data  | N/A—does not store data |



assets at stake within the home and security goals for computational home devices. We then identified key features of devices that, in general, make them more vulnerable to attack or more attractive as attack targets. Together, these axes can be used to evaluate the level and type of security attention appropriate for different home technologies. We applied our approach to three example technologies: a wireless webcam toy, a wireless scale, and a home automation siren. With further research, we conjecture that our risk framework could be distilled into a decision tree-like structure with questions that would allow those without security expertise to deterministically assign a device to a risk category. By seeking to understand the risks posed by home technologies as a cohesive whole, our hope is that this work will strengthen the foundations for developing secure home technologies—with the ultimate goal of creating a more trustworthy home environment for users.

**Acknowledgments**

We thank Intel and the Intel Trust Evidence Program for supporting this work. We thank Dan Halperin, Greg Piper, Jesse Walker, and Meiyuan Zhao for feedback on earlier versions of this article.

**References**

1. Barbie Video Girl; <http://www.barbie.com/videoirl/>
2. Bellissimo, A., Burgess, J. and Fu, K. Secure software updates: Disappointments and new challenges. In *Proceedings of USENIX Hot Topics in Security*, (July 2006).
3. Bojinov, H., Bursztein, E. and Boneh, D. Xcs: Cross channel scripting and its impact on Web applications. In *CCS '09*.
4. Brush, A.J.B. and Inkpen, K.M. Yours, mine and ours? Sharing and use of technology in domestic environments. In *Proceedings of UbiComp '07*.
5. Checkoway, S., McCoy, D., Kantor, B., Anderson, D., Shacham, H., Savage, S., Koscher, K., Czeskis, A., Roesner, F. and Kohno, T. Comprehensive experimental analyses of automotive attack surfaces. In *Proceedings of USENIX Security '11*.
6. Cisco NAC; <http://www.cisco.com/en/US/products/ps6128/index.html>.
7. Denning, T., Matuszek, C., Koscher, K., Smith, J.R. and Kohno, T. A spotlight on security and privacy risks with future household robots: attacks and lessons. In *Proceeding of UbiComp '09*.
8. Dixon, C., Mahajan, R., Agarwal, S., Brush, A.J., Lee, B., Saroiu, S. and Bahl, V. The home needs an operating system (and an app store). In *Proceedings of Hotnets '10*.
9. Edwards, C., Kharif, O. and Rile, M. Human Errors Fuel Hacking as Test Shows Nothing Stops Idiocy (June 27, 2011); <http://www.bloomberg.com/news/2011-06-27/human-errors-fuel-hacking-as-test-shows-nothing-prevents-idiocy.html>
10. Edwards, W.K., Grinter, R.E., Mahajan, R. and Wetherall, D. Advancing the state of home networking. *Commun. ACM* 54, 6 (June 2011).
11. Fisher, D. Samsung Handsets Distributed With Malware-Infected Memory Cards (June 4, 2010); [http://threatpost.com/en\\_us/blogs/samsung-handsets-distributed-malware-infected-memory-cards-060410](http://threatpost.com/en_us/blogs/samsung-handsets-distributed-malware-infected-memory-cards-060410)
12. Friedman, B., Kahn Jr, P.H. and Borning, A. Value sensitive design and information systems: Three case studies. In *Human-Computer Interaction and Management Information Systems: Foundations*.
13. GlowCaps; <http://www.rxvitality.com/glowcaps.html>.
14. Gourdin, B., Soman, C., Bojinov, H. and Bursztein, E. Toward secure embedded Web interfaces. In *Proceedings of USENIX Security '11*.
15. Greenberg, A. iPhone Security Bug Lets Innocent-Looking Apps Go Bad (Nov. 7, 2011); <http://www.forbes.com/sites/andygreenberg/2011/11/07/iphone-security-bug-lets-innocent-looking-apps-go-bad/>
16. Halperin, D., Heydt-Benjamin, T.S., Ransford, B., Clark, S.S., Defend, B., Morgan, W., Fu, K., Kohno, T. and Maisel, W.H. Pacemakers and implantable cardiac defibrillators: Software radio attacks and zero-power defenses. *IEEE* 2008.
17. Khurana, H., Hadley, M., Lu, N. and Frincke, D.A. Smart-grid security issues. *IEEE Security and Privacy* 8 (2010), 81–85.
18. Kim, T-H.J., Bauer, L., Newsome, J., Perrig, A. and Walker, J. Challenges in access right assignment for secure home networks. In *Proceedings for HotSec'10*.
19. Mazurek, M.L., Arsenault, J.P., Bresee, J., Gupta, N., Ion, I., Johns, C., Lee, D., Liang, Y., Olsen, J., Salmon, B., Shay, R., Vaniea, K., Bauer, L., Cranor, L.F., Ganger, G.R. and Reiter, M.K. Access control for home data sharing: Attitudes, needs and practices. In *Proceedings of CHI '10*.
20. Microsoft NAP; <http://technet.microsoft.com/en-us/network/bb545879>.
21. Mills, E. More malware targeting Android (July 11, 2011); [http://news.cnet.com/8301-27080\\_3-20078606-245/more-malware-targeting-android/](http://news.cnet.com/8301-27080_3-20078606-245/more-malware-targeting-android/)
22. Nest; <http://www.nest.com/>.
23. Patel, S.N., Reynolds, M.S. and Abowd, G.D. Detecting human movement by differential air pressure sensing in HVAC system ductwork: An exploration in infrastructure mediated sensing. In *Proceedings of Pervasive '08*.
24. Rock Star in your kitchen; (Aug. 29, 2008); <http://www.gorenjegrup.com/en/news?aid=933>
25. Spykee; <http://www.spykeeworld.com/>.
26. Withings WiFi Body Scale; <http://www.withings.com/en/bodyscale>.
27. XFINITY Home Security; <http://www.comcast.com/homesecurity/>.
28. Yang, J., Edwards, W.K. and Haslem, D. Eden: Supporting home network management through interactive visual tools. In *Proceedings of UIST '10*.

**Tamara Denning** (tdenning@cs.washington.edu) is a Ph.D student at the University of Washington, Seattle.

**Tadayoshi Kohno** (yoshi@cs.washington.edu) is an associate professor at the University of Washington, Seattle.

**Henry M. Levy** (levy@cs.washington.edu) is Wissner-Stivka Chair of Computer Science and Engineering at the University of Washington, Seattle.

© 2013 ACM 0001-0782/13/01

| Environment Integrity   | Activity Pattern Privacy                             | Presence Privacy  | Occupant Identities                                       | Sensed Data Privacy      | Sensor Validity  | Sensor Availability |
|---|--|---|---|--------------------------|--|---------------------|
| Toy can cause minor physical property damage (for example, fragile objects) | Activities easily deduced from A/V feed              | Could reveal whether house is occupied and the presence of children | Occupants easily identifiable                             | Home can be very private | Could add disturbing images or sounds into stream                | Non-essential       |
| N/A   | Weighing times might indicate when occupants wake up | Could potentially reveal whether occupants are on vacation          | Could reveal profile information (for example, name, age) | Weights are private      | Inaccurate weights could cause shame, affect eating and exercise | Non-essential       |
| Continuous alarm an annoyance, user might disable or ignore alarm           | Siren may indicate unauthorized entry                | Siren may indicate unauthorized entry                               | N/A   | N/A                      | N/A  | N/A                 |

# research highlights

---

P. 105

## **Technical Perspective Visualization, Understanding, and Design**

By Doug DeCarlo  
and Matthew Stone

P. 106

## **Illustrating How Mechanical Assemblies Work**

By Niloy J. Mitra, Yong-Liang Yang, Dong-Ming Yan,  
Wilmot Li, and Maneesh Agrawala

---

P. 115

## **Technical Perspective Finding People In Depth**

By James M. Rehg

P. 116

## **Real-Time Human Pose Recognition in Parts from Single Depth Images**

By Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon,  
Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore

---

# Technical Perspective Visualization, Understanding, and Design

By Doug DeCarlo and Matthew Stone

PHOTOGRAPHS CAPTURE THE moment; paintings convey perception, impression, and feeling; illustrations tell stories. Computer graphics aims to enrich all these artistic practices through technology. The following paper is a watershed in *depiction*, creating imagery that gets ideas across. Mitra et al. describe an interactive system that analyzes the operation of mechanical devices and explains them visually to users. Their compelling results showcase an innovative synthesis of newly mature techniques for robust analysis of 3D geometry and for domain-specific information design.

Like many watershed papers in computer graphics, this paper takes its cues from the work of a master artist—in this case, RISD Professor and MacArthur Fellow David Macaulay. While Macaulay may be best known for the engaging and richly informative visual storytelling of his architectural history books *Cathedral* and *Castle*, the authors here base their system on his book *The Way Things Work*. It is a fascinating compendium of explorations of everyday artifacts, like the lock on your front door, illustrated through lucid visual explanations that communicate a deep understanding. Such pictures are the authors' inspiration.

They get their understanding of the world through geometry. They do not simulate physics directly. They use symmetries to infer how components might move; they use correspondences to recognize components that can drive one another. The geometric computations allow approximate matches, by exploiting the robust statistical principle that reliable symmetries and correspondences leave lots of evidence. So a first step searches for candidate matches, and a second step tabulates the results to find consistent patterns. Such techniques have a long history, but recent bridges to computer graphics have had an enormous in-

fluence on the practical analysis of 3D shape, as you see here.

Geometry is a powerful cue to the behavior of everyday artifacts. The reciprocating rack and pinion at the end of their paper is a mesmerizing example. But geometry goes only so far. The system must assume its input is a working machine. And since parts like levers and belts do not have a distinctive geometric signature, the user has to label them interactively.

Making an effective picture takes more than just the right representation of the mechanism. It requires the right *design principles*. Design principles are domain-specific rules for using specific visual techniques to make the information in a display easy to see and understand. Maneesh Agrawala has built a wide range of influential systems for depiction by codifying such design principles, implementing them as algorithms, and evaluating their effectiveness; his research program is described in his co-authored *Communications* article "Design Principles for Visual Communication" (Apr. 2011). For explaining mechanical assemblies, the key principles are to focus on showing causal chains of motion, and to do this by selecting key frames and annotating them with diagrammatic arrows. Here as elsewhere, the design principles come from

**The following paper is a watershed in *depiction*, creating imagery that gets ideas across.**

analysis of exemplary hand-made work, from artists' reflection on their practice, from psychological theories of how people understand these visualizations, and from the researchers' own experimentation with the possibilities of technology.

Realizing these design principles involves a judicious choice of visual techniques. Non-photorealism in computer graphics offers diverse ways to stylize appearances and guide the viewer's attention. Examples include modulations of detail and weight in rendering objects, the use of cutting and transparency to depict objects in multiple layers, and even selective choices about which elements to render at all. The use here of simple line drawings, with arrows for annotation, and a constrained set of highlighted parts and exploded views, is a choice that reliably leads to clear and uncluttered imagery. To create accessible imagery with more richly varying rendering techniques, or with visualizations of additional information (forces, for example), it might be necessary to develop much more nuanced design principles. The pictures here, however, are clearly a success.

It is never easy to endow computers with a deep and interesting understanding that they can share with their users. But that does not mean we should regard inference as hopeless or design as magic. As this work shows, general tools and methodologies are making it easier and easier for systems to communicate the understanding they have through clear and compelling visualizations. The results here thus take on particular significance as a benchmark in visual explanation, and a model for future systems. □

**Doug DeCarlo** (decarlo@cs.rutgers.edu) and **Matthew Stone** (mdstone@cs.rutgers.edu) are associate professors in the Department of Computer Science and Center for Cognitive Science at Rutgers University, Piscataway NJ. DeCarlo is currently on leave at Google, Inc.

© 2013 ACM 0001-0782/13/01

# Illustrating How Mechanical Assemblies Work

By Niloy J. Mitra, Yong-Liang Yang, Dong-Ming Yan, Wilmot Li, and Maneesh Agrawala

## Abstract

**How-things-work visualizations use a variety of visual techniques to depict the operation of complex mechanical assemblies. We present an automated approach for generating such visualizations. Starting with a 3D CAD model of an assembly, we first infer the motions of the individual parts and the interactions across the parts based on their geometry and a few user-specified constraints. We then use this information to generate visualizations that incorporate motion arrows, frame sequences, and animation to convey the causal chain of motions and mechanical interactions across parts. We demonstrate our system on a wide variety of assemblies.**

## 1. INTRODUCTION

... all machines that use mechanical parts are built with the same single aim: to ensure that exactly the right amount of force produces just the right amount of movement precisely where it is needed.

(David Macaulay, *The New Way Things Work*<sup>18</sup>)

Mechanical assemblies are collections of interconnected parts such as gears, cams, and levers that move in conjunction to achieve a specific functional goal. As Macaulay points out, attaining this goal usually requires the assembly to transform a driving force into a specific movement. For example, the gearbox in a car is a collection of interlocking gears with different ratios that transforms rotational force from the engine into the appropriate revolution speed for the wheels. Understanding how the parts interact to transform the driving force into motion is often the key to understanding how such mechanical assemblies work.

There are two types of information that are crucial for understanding this transformation process: (i) the spatial configuration of the individual parts within the assembly and (ii) the causal chain of motions and mechanical interactions between the parts. While most technical illustrations effectively convey spatial relationships, only a much smaller subset of these visualizations is designed to emphasize how parts move and interact with one another. Analyzing this subset of *how-things-work* illustrations and prior cognitive psychology research on how people understand mechanical motions suggests several visual techniques for conveying the movement and interactions of parts within a mechanical assembly: (i) *motion arrows* indicate how individual parts move; (ii) *frame sequences* highlight key snapshots of complex motions and the sequence of interactions along the causal chain; and (iii) *animations* show the dynamic behavior of an assembly.

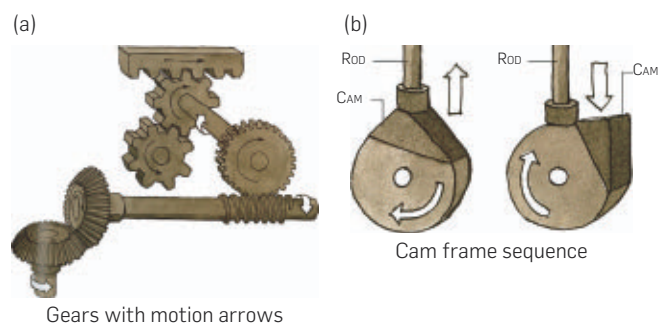
Creating effective how-things-work illustrations and animations by hand is difficult because a designer must understand how a complex assembly works and also have the skill to apply the appropriate visual techniques for emphasizing the motions and interactions between parts. As a result, well-designed illustrations and animations are relatively uncommon, and the few examples that do exist (e.g., in popular educational books and learning aids for mechanical engineers) are infrequently updated or revised. Furthermore, most illustrations are static and thus do not allow the viewer to inspect an assembly from multiple viewpoints (see Figure 1).

In this paper, we present an automated approach for generating how-things-work visualizations of mechanical assemblies from 3D CAD models, thus facilitating the creation of both static illustrations and animations from any user-specified viewpoint (see Figure 2). Our work addresses two main challenges:

(i) **Motion and interaction analysis.** Most 3D models do not specify how their parts move or interact with one another. Yet, this information is essential for creating visualizations that convey how the assembly works. We present a semi-automatic technique that determines the motions of parts and their causal relationships based on their geometry.

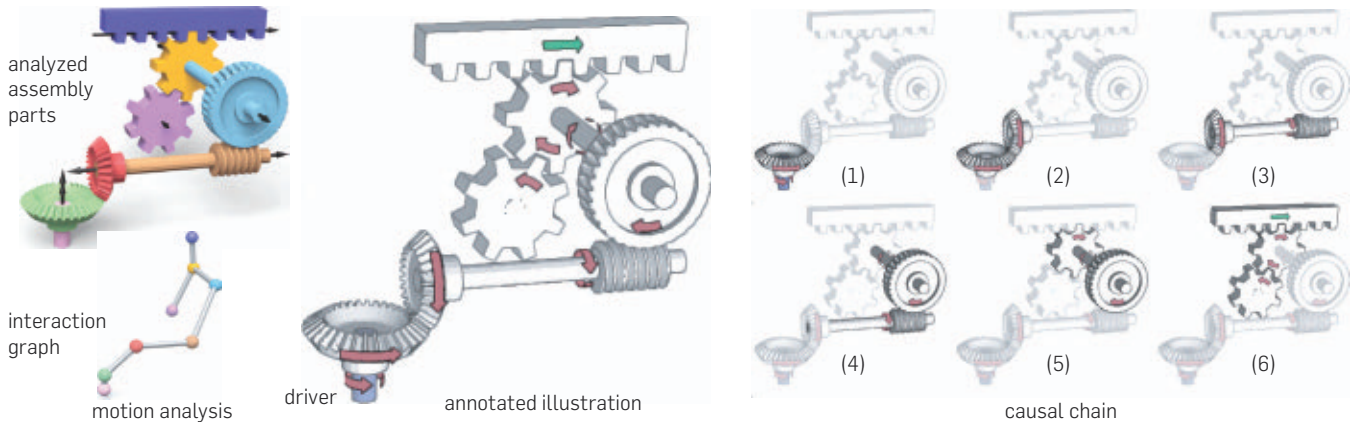
(ii) **Automatic visualization.** We present algorithms that use the motion and interaction information from our analysis to automatically generate a variety of how-things-work visualizations, including static illustrations with motion arrows, frame sequences to highlight key snapshots, and the

**Figure 1. Hand-designed illustrations. These examples show how motion arrows (a) and sequences of frames (b) can help convey the motion and interactions of parts within mechanical assemblies.**



The original version of this paper was published in *ACM Transactions on Graphics (SIGGRAPH)*, July 2010.

**Figure 2.** We analyze a given geometric model of a mechanical assembly to infer how the individual parts move and interact with each other and encode this information as a time-varying interaction graph. Once the user indicates a driver part, we use the interaction graph to compute the motion of the assembly and generate an annotated illustration to depict how the assembly works. We also produce a corresponding causal chain sequence to help the viewer mentally animate the motion.



causal chain of mechanical interactions, as well as simple animations of the assembly in motion.

## 2. DESIGNING HOW-THINGS-WORK VISUALIZATIONS

Illustrators and engineers have produced a variety of books<sup>2, 3, 15, 18</sup> and websites (e.g., [howstuffworks.com](http://howstuffworks.com)) that are designed to show how complex mechanical assemblies work. These illustrations use a number of diagrammatic conventions to highlight the motions and mechanical interactions of parts in the assembly. Cognitive psychologists have also studied how static and multimedia visualizations help people mentally represent and understand the function of mechanical assemblies.<sup>19</sup> For example, Narayanan and Hegarty<sup>24, 25</sup> propose a cognitive model for comprehension of mechanical assemblies from diagrammatic visualizations, which involves (i) constructing a spatial representation of the assembly and then (ii) producing the causal chain of motions and interactions between the parts. To facilitate these steps, they propose a set of high-level design guidelines to create *how-things-work* visualizations.

Researchers in computer graphics have concentrated on refining and implementing design guidelines that assist the first step of the comprehension process. Algorithms for creating exploded views,<sup>13,16,20</sup> cutaways,<sup>4,17,27</sup> and ghosted views<sup>6, 29</sup> of complex objects apply illustrative conventions to emphasize the spatial locations of the parts with respect to one another. In contrast, the problem of generating visualizations that facilitate the second step of the comprehension process remains largely unexplored within the graphics community. While some researchers have proposed methods for computing motion cues from animations<sup>26</sup> and videos,<sup>8</sup> these efforts do not consider how to depict the causal chain of motions and interactions between parts in mechanical assemblies.

### 2.1. Helping viewers construct the causal chain

In an influential treatise examining how people predict the behavior of mechanical assemblies from static

visualizations, Hegarty<sup>9</sup> found that people reason in a step-by-step manner, starting from an initial *driver* part and tracing forward through each subsequent part along a causal chain of interactions. At each step, people infer how the relevant parts move with respect to one another and then determine the subsequent action(s) in the causal chain. Although all parts may be moving at once in real-world operation of the assembly, people mentally animate the motions of parts one at a time in causal order.

While animation might seem like a natural approach for visualizing mechanical motions, in a meta-analysis of previous studies comparing animations to informationally equivalent sequences of static visualizations, Tversky et al.<sup>28</sup> found no benefit for animation. Our work does not seek to engage in this debate between static versus animated illustrations. Instead we aim to support both types of visualizations with our tools. We consider both static and animated visualizations in our analysis of design guidelines.

Effective *how-things-work* illustrations use a number of visual techniques to help viewers mentally animate an assembly:

**(i) Use arrows to indicate motions of parts.** Many illustrations include arrows that indicate how each part in the assembly moves. In addition to conveying the motion of individual parts, such arrows can also help viewers understand the specific functional relationships between parts.<sup>10, 12</sup> Placing the arrows near contact points between parts that interact along the causal chain can help viewers better understand the causal relationships.

**(ii) Highlight causal chain step-by-step.** In both static and animated illustrations, highlighting each step in the causal chain of actions helps viewers mentally animate the assembly by explicitly indicating the sequence of interactions between parts. Static illustrations often depict the causal chain using a sequence of keyframes that correspond to the sequence of steps in the chain. Each keyframe highlights the transfer of movement between a set of touching parts, typically by rendering those parts in a different style from the rest of the

assembly. In the context of animated visualizations, researchers have shown that adding signaling cues that sequentially highlight the steps of the causal chain improves comprehension compared to animations without such cues.<sup>11,14</sup>

**(iii) Highlight important keyframes of motions.** The motions of most parts in mechanical assemblies are periodic. However, in some of these motions, the angular or linear velocity of a part may change during a single period. For example, the pistons in the assembly shown in Figure 8 move the cylinder up and down during a single period of motion. To depict such complex motions, static illustrations sometimes include keyframes that show the configuration of parts at the critical instances in time when the angular or linear velocity of a part changes. Inserting one additional keyframe between each pair of critical instances can help clarify how the parts move from one critical instance to the next.

### 3. SYSTEM OVERVIEW

We present an automated system for generating how-things-work visualizations that incorporate the visual techniques described in the previous section. The input to our system is a polygonal model of a mechanical assembly that has been partitioned into individual parts. Our system deletes hanging edges and vertices as necessary to make each part 2-manifold. We assume that touching parts are modeled correctly, with no self-intersections beyond a small tolerance. As a first step, we perform an automated motion and interaction analysis of the model geometry to determine the relevant motion parameters of each part, as well as the causal chain of interactions between parts. This step requires the user to specify the *driver* part for the assembly and the *direction* in which the driver moves. Using the results of the analysis, our system allows users to generate a variety of static and animated visualizations of the input assembly from any viewpoint. The next two sections present our analysis and visualization algorithms in detail.

### 4. MOTION + INTERACTION ANALYSIS

We analyze the input polyhedral model of an assembly to extract the degrees of freedom for each part, and also to understand how the parts move and interact with one another within the assembly. We encode the extracted information as an *interaction graph*  $G := (V, E)$  where, each node  $n_i \in V$  represents part  $P_i$  and each edge  $e_{ij} \in E$  represents a mechanical interaction between two touching parts ( $P_i, P_j$ ) (see Figure 2).

In order to construct this interaction graph, we rely on two high-level insights: first, the motion of many mechanical parts is related to their geometric properties, including self-similarity and symmetry; and second, the different types of mechanical interactions between parts are often characterized by the specific spatial relationships and geometric attributes of the relevant parts. Based on these insights, we propose a two-stage process to construct the interaction graph:

In the *part analysis* stage, we analyze each individual part to determine its type (e.g., spur gear, bevel gear, and axle)

and relevant parameters (e.g., rotation axis, side profile, and radius) using existing shape analysis algorithms. We store the extracted information in the corresponding nodes of graph  $G$ .

In the *interaction analysis* stage, we analyze each pair of touching parts and classify the type of mechanical interaction based on their spatial relationships and part parameters. We store the information in the corresponding edges of graph  $G$ . Our system handles a variety of part types and interactions as shown in Figure 3.

#### 4.1. Part analysis

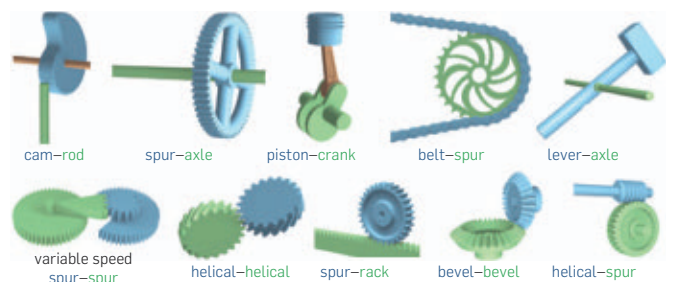
Our part analysis automatically classifies parts into the following common types: rotational gears (e.g., spur and bevel), helical gears, translational gears (i.e., racks in spur-rack mechanisms), axles, and fixed support structures (i.e., the stationary parts in an assembly that support and constrain the motions of other parts). The classifier is based on the geometric features of the parts. We rely on the user to manually classify parts that lack distinctive geometric characteristics such as cams, rods, cranks, pistons, levers, and belts. Figure 3 shows many of the moving parts handled by our system, and Figures 7a and 10b–c include some fixed support structures.

We also compute part parameters that inform the subsequent interaction analysis stage that determines how motion is transmitted across parts. For all gears and axles, we estimate the axis of rotational, helical, or translational motion. We compute teeth count and width for rotational and translational gears, and the pitch for helical gears. For rotational and helical gears, we also compute whether the part has a conical (e.g., bevel) or cylindrical (e.g., spur) side profile and its radii (e.g., inner, outer), as these properties influence the gear can interact with other gears. Since support structures often have housings or cutouts that constrain the rotational motion of other parts, we compute potential axes of rotation for these structures. Finally, we also compute potential rotation axes for user-classified cams, cranks, and levers.

To distinguish the different types of parts and estimate their parameters, we use the following shape analysis algorithms.

**Symmetry detection.** We assume that all gears and axles exhibit rotational, helical or translational symmetry and move based on their symmetry axes. We use a variant of the algorithm proposed by Mitra et al.<sup>22</sup> to detect such

**Figure 3. Typical part types and interactions encountered in mechanical assemblies and handled by our system. While we automatically detect most of these configurations, we require the user to manually classify cams, rod, cranks, pistons, levers, and belts.**



symmetries and infer part types and parameters based on the symmetry properties. If a part is rotationally symmetric, we mark it as either a rotational gear or axle, use the symmetry axis as the rotation axis, and use the order of symmetry to estimate teeth count and width. If a part is helically symmetric, we mark it as a helical gear, use the symmetry axis as the screw axis, and record the helix pitch. Finally, if a part has discrete translational symmetry, we mark it as a translational gear (i.e., rack), use its symmetry direction as the translation axis, and use the symmetry period to estimate the teeth count and width. Note that the symmetry detection method also handles partial symmetries that are present in parts like variable-speed gears (see Figure 3). If a part exhibits no symmetries and has not been classified by the user as a cam, rod, crank, piston, lever or belt, we assume it to be a fixed support structure.

**Cylinders versus cones.** Next, we analyze the side profiles of rotational and helical gears to determine whether they are cylindrical or conical, respectively. Specifically, we partition such gears into *cap* and *side* regions as follows. Let  $a_i$  denote the rotation/screw axis of part  $P_i$ .

We mark its  $j$ -th face as a cap face if its normal  $n_j$  is parallel to the part's rotational/screw axis, such that  $|n_j \cdot a_i| \approx 1$ , otherwise we mark it as a side face. We then build connected components of faces with the same labels, and discard components with only few faces as members (see Figure 4). Subsequently, we fit least squares cylinders and cones to the side regions and classify parts with low residual error as cylindrical or conical, respectively.

**Sharp edge loops.** Finally, we use *sharp edge loops*, which are 1D curves defined by sharp creases on a part, to determine additional part parameters for rotationally or helically symmetric parts, as well as cams, cranks, levers, and fixed support structures. We start by marking all mesh edges whose adjacent faces are close to orthogonal (i.e., dihedral angle in  $90^\circ \pm 30^\circ$  in our implementation) as *sharp* (see also Gal et al.<sup>7</sup> and Mehra et al.<sup>21</sup>). We then partition the mesh into segments separated by sharp edges, discard very small segments (less than 10 triangles in our tests), and label the boundary loops of the remaining segments as sharp edge loops. Next, we identify all the sharp edge loops that are (roughly) circular by fitting (in a least squares sense) circles to all the loops and selecting the ones with low residual errors. For rotationally and

helically symmetric parts, we use the minimum and maximum radii of the circular loops as estimates for the inner and outer radii of the parts (e.g., Figure 4-left shows the outer radius of a cylindrical gear).

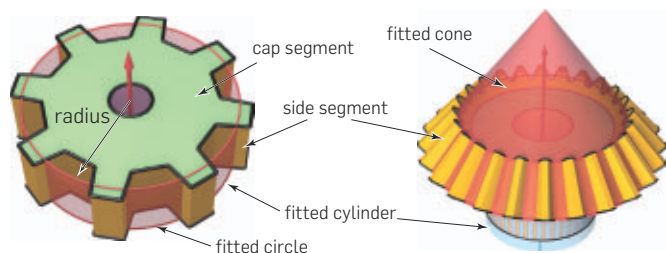
For fixed support structures, a group of circular loops with a consistent orientation often indicates a potential axis of rotation for a part that docks with the fixed structure. Such clusters of loops also indicate potential rotation axes for cams, cranks, and levers. We cluster circular loops in two stages (see Figure 5): For each loop we compute the normal of the plane that contains the fitted circle, which we call the *circle axis*, and cluster loops with similar circle axes. Then, we partition each cluster based on the projection of the circle centers along a representative circle axis for that cluster. The resulting clusters represent groups of circular loops with roughly parallel circle axes that are close to one another. We record the representative circle axis for each cluster as a potential rotation axis.

## 4.2. Interaction analysis

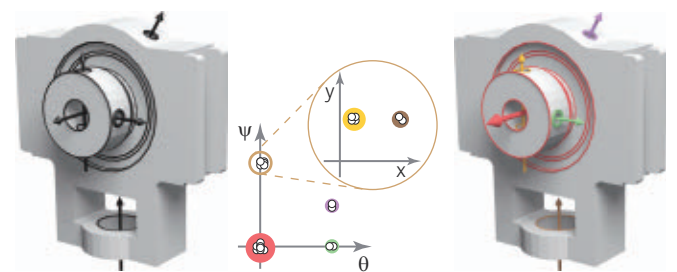
To build the edges of the interaction graph and estimate their parameters, we: (i) compute the topology of the interaction graph based on the contact relationships between part pairs; (ii) classify the type of mechanical interaction at each contact (i.e., how motion is transferred from one part to another); and (iii) compute the motion of the assembly. **(i) Contact detection.** We use the contact relationships between parts to determine the topology of the interaction graph. Following the approach of Agrawala et al.<sup>1</sup>, we consider each pair of parts ( $P_i, P_j$ ) in the assembly and compute the closest distance between them. If this distance is less than a threshold  $\alpha$ , we consider the parts to be in contact, and we add an edge  $e_{ij}$  between the nodes  $n_i$  and  $n_j$  in the interaction graph. We set  $\alpha$  to 0.1% of the diagonal of the assembly bounding box in our experiments.

As an assembly moves, its contact relationships evolve, that is, edges  $e_{ij}$  in the interaction graph may appear or disappear over time (see Figure 10c). We detect such contact changes using a space-time analysis. Suppose at time  $t$ , two parts  $P_i$  and  $P_j$  are in contact and we have identified their interaction type (see below). On the basis of this information, we estimate their relative motion parameters, compute their positions at subsequent times  $t + \Delta$ ,  $t + 2\Delta$ , etc., and

**Figure 4.** For rotationally and helically symmetric parts, we use their symmetry axes to partition the parts into cap- and side-regions. We fit cylinders or cones to the side regions to determine their profile types and extract sharp edge loops to estimate part attributes like radii.



**Figure 5.** We detect circular sharp edge feature loops on a part (left) and cluster the loops based on the orientations of their circle axes (middle) and the projections of the circle centers onto these axes (middle-inset). Such clusters represent groups of nearby loops with parallel axes, shown here in different colors (right).



compute the contact relationships at each time. We use a fixed sampling rate of  $\Delta = 0.1$  s with the default speed for the driver part set to an angular velocity of 0.1 radian/s or translational velocity of 0.1 unit/s, as applicable. Our method detects cases where two parts transition from touching to not touching over time. It also detects cases where two parts remain in contact but their contact region changes, which often corresponds to a change in the parameters of the mechanical interaction (e.g., the variable speed gear shown in Figure 3). For each set of detected contact relationships, we compute a new interaction graph. Note that we implicitly assume that part contacts change discretely over the motion cycle, which means that we cannot handle continuously evolving interactions, as in the case of elliptic gears. See original paper<sup>23</sup> for additional details.

**(ii) Interaction classification.** We classify the type of interaction for each pair of parts  $P_i$  and  $P_j$  that are in contact, using their relative spatial arrangement and the individual part attributes. Specifically, we classify interactions based on the positions and orientations of the part axes  $a_i$  and  $a_j$  along with the values of the relevant part parameters. For parts with multiple potential axes, we consider all pairs of axes.

*Parallel axes:* When the axes are nearly parallel, that is,  $|a_i \cdot a_j| \approx 1$ , we detect one of the following interactions: cylinder-on-cylinder (e.g., spur gears) or cylinder-in-cylinder (e.g., planetary gears). For cylinder-on-cylinder,  $r_i + r_j$  (roughly) equals the distance between the part axes. For cylinder-in-cylinder,  $|r_i - r_j|$  (roughly) equals the distance between the part axes. Note for cylinder-on-cylinder, the parts can rotate about their individual axes, while simultaneously one cylinder can rotate about the other one, for example, (subpart of) planetary configuration (see Figure 9).

*Coaxial:* When the axes are parallel and lie on a single line, we classify the interaction as coaxial (e.g., spur-axle and cam-axle).

*Orthogonal axes:* When the axes are nearly orthogonal, that is,  $a_i \cdot a_j \approx 0$ , we detect one of the following interactions: spur-rack, bevel-bevel, helical-helical, helical-spur. If one part is a rotational gear and the other is a translational gear with matching teeth widths, we detect a spur-rack interaction. If both parts are conical with cone angles summing up to  $90^\circ$ , we mark a bevel-bevel interaction. If both parts are cylindrical and helical, we mark a helical-helical interaction. If the parts are cylindrical but only one is helical, we mark a helical-spur interaction.

*Belt interactions:* Since belts do not have a single consistent axis of motion, we treat interactions with belts as a special case. If a cylindrical part touches a belt, we detect a cylinder-on-belt interaction.

These classification rules are carefully designed based on standard mechanical assemblies and successfully categorize most part interactions automatically. In our results, only the cam-rod and piston-crank interactions in the hammer (Figure 7a), piston engine (Figure 8), and the drum (Figure 10d) needed manual classification.

**(iii) Motion computation.** Mechanical assemblies are brought to life by an external force applied to a driver and propagated to other parts according to interaction types and part attributes. In our system, once the user indicates

the driver, motion is transferred to the other connected parts through a breadth-first graph traversal of the interaction graph  $G$ , starting with the driver-node as the root. We employ simple forward-kinematics to compute the relative speed at any node based on the interaction type with its parent<sup>5</sup>. For example, for a cylinder-on-cylinder interaction, if motion from a gear with radius  $r_i$  and angular velocity  $\omega_i$  is transmitted to another with radius  $r_j$ , then the imparted angular velocity  $\omega_j = \omega_i r_i / r_j$ . Our approach handles graphs with loops (e.g., planetary gears). Since we assume that our input models are consistent assemblies, even when multiple paths exist between a root node and another node, the final motion of the node does not depend on the graph traversal path. When we have an additional constraint at a node, for example, a part is fixed or restricted to translate only along an axis, we impose the constraint in the forward-kinematics computation. Note that since we assume that the input assembly is a valid one and does not self-penetrate during its motion cycle, we do not perform any collision detection in our system.

## 5. VISUALIZATION

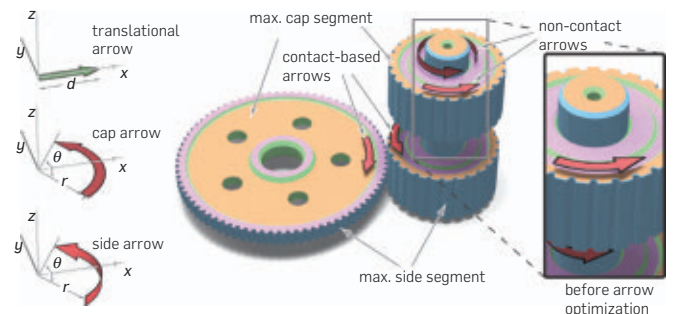
Using the computed interaction graph, our system automatically generates how-things-work visualizations based on the design guidelines discussed in Section 2. Here, we present algorithms for computing arrows, highlighting both the causal chain and important keyframes of motion, and generating exploded views.

### 5.1. Computing motion arrows

For static illustrations, our system automatically computes arrows from the user-specified viewpoint. We support three types of arrows (see Figure 6): *cap arrows*, *side arrows*, and *translational arrows* and generate them as follows: (i) determine how many arrows of each type to add; (ii) compute initial arrow placements; and (iii) refine arrow placements to improve their visibility.

For each non-coaxial part interaction encoded in the interaction graph, we create two arrows, one associated with each node connected by the graph edge. We refer to such arrows as contact-based arrows, as they highlight contact

**Figure 6. Translational, cap, and side arrows (left). Arrows are first added based on the interaction graph edges, and then to any moving parts without an arrow assignment. The initial arrow placement can suffer from occlusion (right-inset), which is fixed using a refinement step (center).**

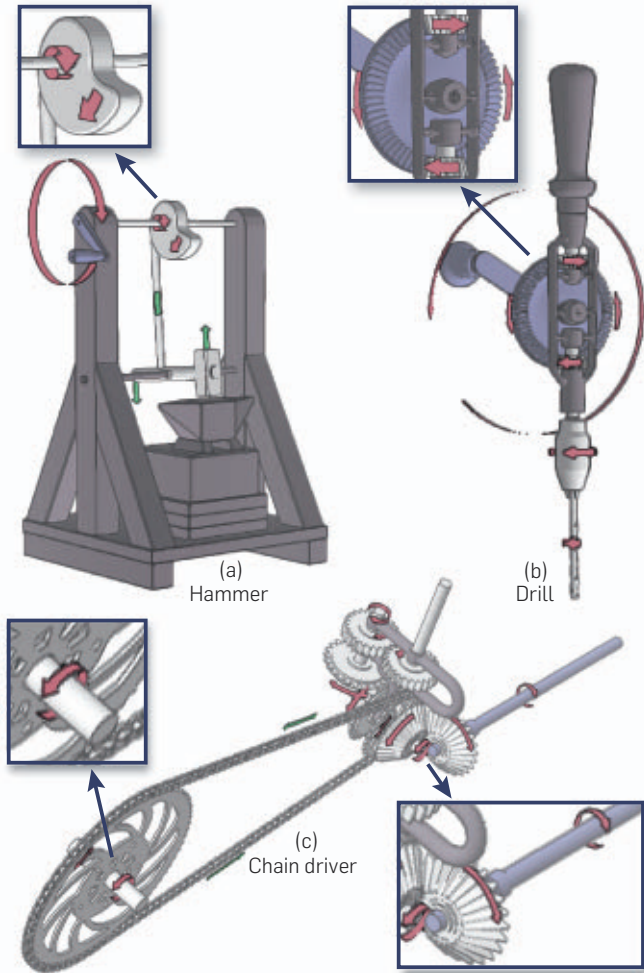




relations. We use the following rules to add contact arrows based on the type of part interaction:

- cylinder-on-cylinder*: add cap arrows on both parts;
- cylinder-in-cylinder*: add a cap arrow for the inner cylinder and a side arrow for the outer cylinder;

**Figure 7. Motion arrow results.** To convey how parts move, we automatically compute motion arrows from the user-specified viewpoint. Here, we manually specified the lever in the hammer model (a) and the belt in the chain driver model (c); our system automatically identifies the types of all the other parts.



- spur-rack*: add a cap arrow to spur and translational arrow to rack;
- bevel-bevel*: add side arrows on both (conical) parts;
- helical-helical*: add a cap arrow on both parts;
- helical-spur*: add a cap arrow on the cylinder and a side arrow on the helical part; and
- cylinder-on-belt*: add a cap arrow on the cylinder and a translational arrow on the belt;

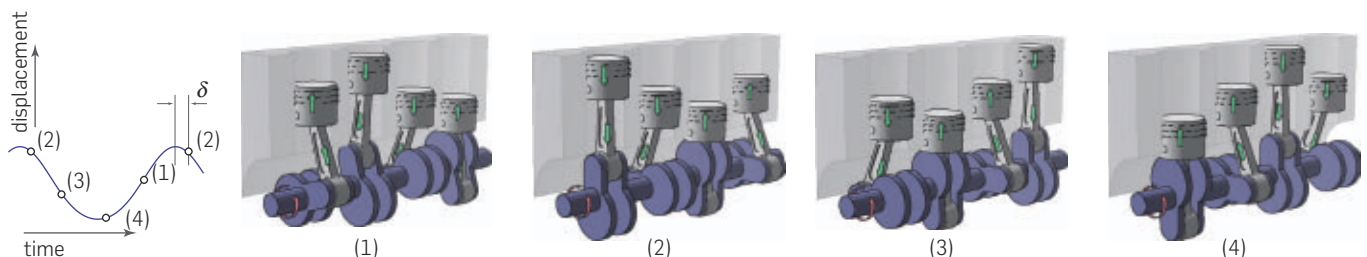
Note that these rules do not add arrows for certain types of part interactions (e.g., coaxial). For these interactions, we add a non-contact arrow to any part that does not already have an associated contact arrow. Furthermore, if a cylindrical part is long, a single arrow may not be sufficient to effectively convey the movement of the part. In this case we add an additional non-contact side arrow to the part. Note that a part may be assigned multiple arrows.

After choosing the number of arrows to add and associating a part with each one, we next compute their initial positions using the cap and side segments for each part (see Section 4). We use the z-buffer to identify the cap and side face segments with the largest visible areas after accounting for occlusion from other parts as well as self-occlusion. These segments serve as candidate locations for arrow placement: we place side arrows at the middle of the side segment with maximal score (computed as a combination of visibility and length of the side segment) and cap arrows right above the cap segment with maximal visibility. For contact-based side and cap arrows, we move the arrow within the chosen segment as close as possible to the corresponding contact point. Non-contact translational arrows are placed midway along the translational axis with arrow heads facing the viewer. The local coordinate frame of the arrows are determined based on the directional attributes of the corresponding parts, while the arrow widths are set to a default value. The remaining parameters of the arrows ( $d, r, \theta$  as in Figure 6) are derived in proportion to the part parameters like its axis, radius, and side/cap segment area. We position non-contact side arrows such that the viewer sees the arrow head face-on. Please refer to the original paper<sup>23</sup> for additional details.

## 5.2. Highlighting the causal chain

To emphasize the causal chain of actions, our system generates a sequence of frames that highlights the propagation of motions and interactions from the driver throughout

**Figure 8. Keyframes for depicting periodic motion of a piston engine.** For each of the pistons, we generate two keyframes based on its extremal positions (i.e., the top and bottom of its motion). We typically also add middle frames between these extrema-based keyframes, but due to the symmetry of the piston motion, the middle frames of each piston already exist as extrema-based keyframes of other pistons.

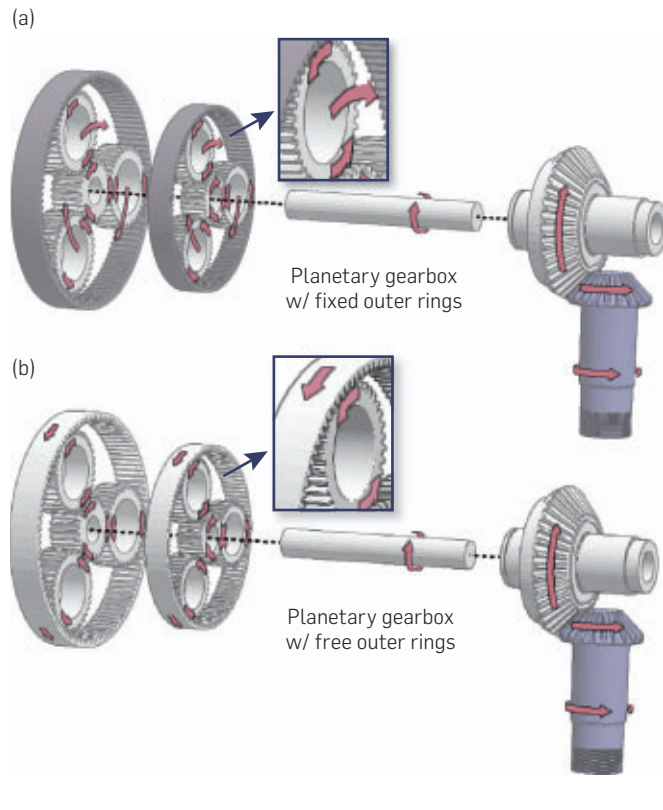


the rest of the assembly. Starting from the root of the interaction graph, we perform a breadth first traversal. At each traversal step, we compute a set of nodes  $S$  that includes the frontier of newly visited nodes, as well as any previously visited nodes that are in contact with this frontier. We then generate a frame that highlights  $S$  by rendering all other parts in a desaturated manner. To emphasize the motion of highlighted parts, each frame includes any non-contact arrow whose parent part is highlighted, as well as any contact-based arrow whose two associated parts are both highlighted. If a highlighted part only has contact arrows and none of them are included based on this rule, we add the part's longest contact arrow to the frame to ensure that every highlighted part has at least one arrow. In addition, arrows associated with previously visited parts are rendered in a desaturated manner. For animated visualizations, we allow the user to interactively step through the causal chain while the animation plays; at each step, we highlight parts and arrows as described above.

### 5.3. Highlighting keyframes of motion

As explained in Section 2, some assemblies contain parts that move in complex ways (e.g., the direction of motion changes periodically). Thus, static illustrations often include keyframes that help clarify such motions. We automatically

**Figure 9. Exploded view results.** Our system automatically generates exploded views that separate the assembly at coaxial part interactions to reduce occlusions. These two illustrations show two different configurations for the planetary gearbox: one with fixed outer rings (a), and one with free outer rings (b). The driver part is in blue, while fixed parts are in dark gray.



compute keyframes of motion by examining each translational part in the model: if the part changes direction, we add keyframes at the critical times when the part is at its extremal positions. However, since the instantaneous direction of motion for a part is undefined exactly at these critical times, we canonically freeze time  $\delta$  after the critical time instances to determine which direction the part is moving in (see Figure 8). Additionally, for each part, we also add middle frames between extrema-based keyframes to help the viewer easily establish correspondence between moving parts. However, if such frames already exist as the extrema-based keyframes of other parts, we do not add the additional frames (see Figure 8).

We also generate a single frame sequence that highlights both the causal chain and important keyframes of motion. As we traverse the interaction graph to construct the causal chain frame sequence, we check whether any newly highlighted part exhibits complex motion. If so, we insert keyframes to convey the motion of the part and then continue traversing the graph (see Figure 10c).

### 5.4. Exploded views

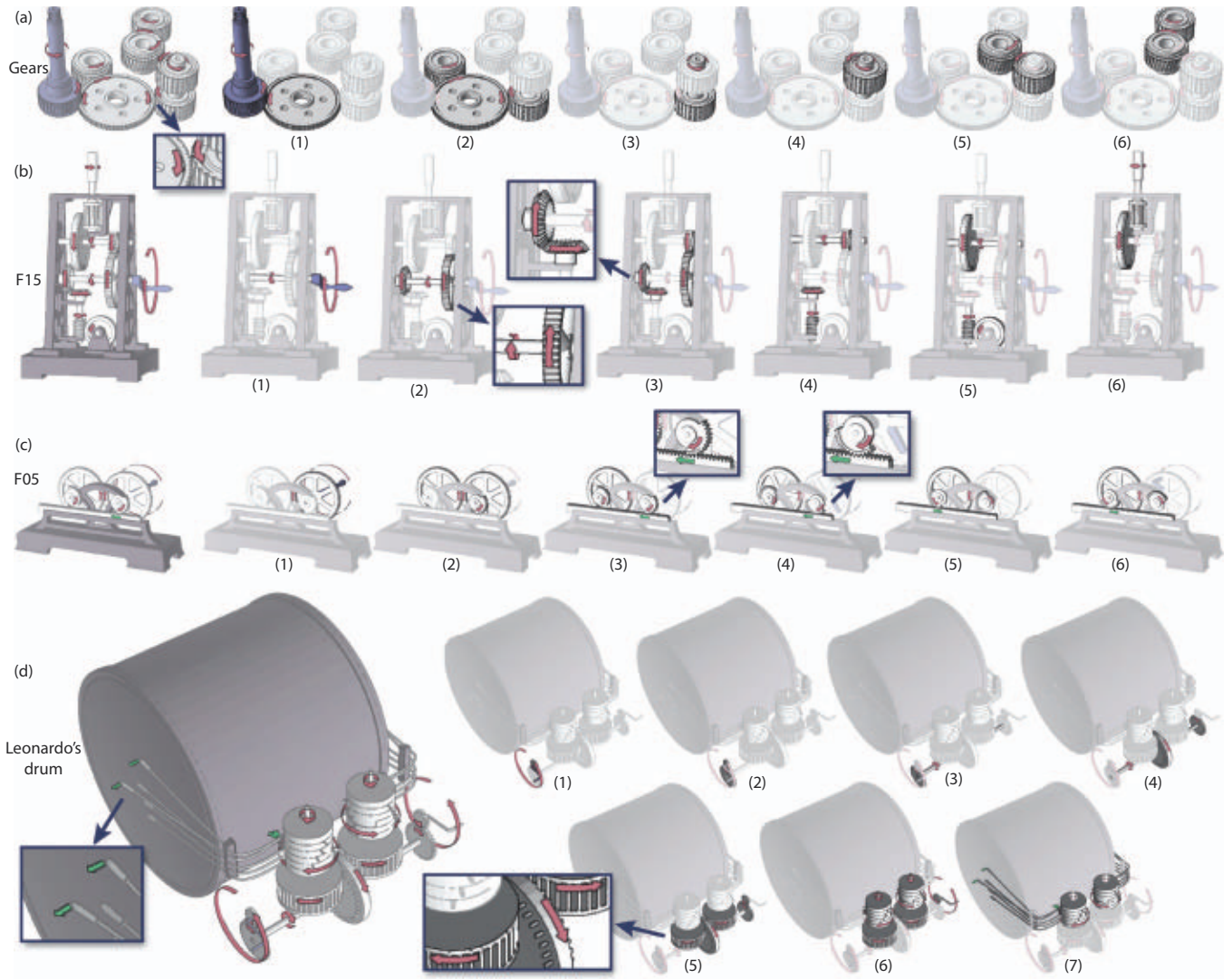
In some cases, occlusions between parts in the assembly make it difficult to see motion arrows and internal parts. To reduce occlusions, our system generates exploded views that separate portions of the assembly (see Figure 9). Typical exploded views separate all touching parts from one another to ensure that each part is visually isolated. However, using this approach in how-things-work illustrations can make it difficult for viewers to see which parts interact and how they move in relation to one another.

To address this problem, we only separate parts that are connected via a coaxial interaction; since such parts rotate rigidly around the same axis, we believe it is easier for viewers to understand their relative motion even when they are separated from one another. To implement this approach, our system first analyzes the interaction graph and then cuts coaxial edges. The connected components of the resulting graph correspond to sub-assemblies that can be separated from one another. We use the technique of Li et al.<sup>16</sup> to compute explosion directions and distances for these sub-assemblies.

## 6. RESULTS

We used our system to generate both static and animated *how-things-work* visualizations for ten different input models, each of which contains from 7 to 27 parts, collected from various sources. Figures 2, 7–10 show static illustrations of all ten models. Other than specifying the driver part and its direction of motion, no additional user assistance was required to compute the interaction graph for seven of the models. For the drum, hammer, and chain driver models, we manually specified lever, cam, rod and belt parts, respectively. We also specified the crank and piston parts in the piston model. In all of our results, we colored the driver blue, fixed support structures dark gray, and all other parts light gray. We render translation arrows in green, and side and cap arrows in red. In all the examples, analysis takes 1–2 s, while visualization runs at interactive rates.

**Figure 10. Illustration results.** We used our system to generate these how-things-work illustrations from 3D input models. For each model, we specified the driver part and its direction of motion. In addition, we manually specified the levers in the drum (c). From this input, our system automatically computes the motions and interactions of all assembly parts and generates motion arrows and frame sequences. We created the zoomed-in insets by hand.



Our results demonstrate how the visual techniques described in Section 2 help convey the causal chain of motions and interactions that characterize the operation of mechanical assemblies. For example, not only do the arrows in Figure 10a indicate the direction of rotation for each gear, but their placement near contact points also emphasizes the interactions between parts. The frame sequence in Figure 10b shows how the assembly transforms the rotation of the driving handle through a variety of gear configurations, while the sequence in Figure 10c conveys both the causal chain of interactions (frames 1–3) and the back-and-forth motion of the horizontal rack (frames 3–6) as it engages alternately with the two circular gears. Finally, our animated results (which can be found at: [http://vecg.cs.ucl.ac.uk/Projects/SmartGeometry/how\\_things\\_work/](http://vecg.cs.ucl.ac.uk/Projects/SmartGeometry/how_things_work/)) show how sequential highlighting of parts along the causal chain can help convey how motions and

interactions propagate from the driver throughout the assembly while the animation plays.

## 7. CONCLUSIONS AND FUTURE WORK

In this work, we presented an automated approach for generating *how-things-work* visualizations from 3D CAD models. Our results demonstrate that combining shape analysis techniques with visualization algorithms can produce effective depictions for a variety of mechanical assemblies. Thus, we believe our work has useful applications for the creation of both static and animated visualizations in technical documentation and educational materials.

We see several directions for extending our approach: (i) *Handling more complex models:* Analyzing and visualizing significantly more complex models (with hundreds or even thousands of parts) introduces additional challenges, including the possibility of excess visual clutter and large

numbers of occluded parts. (ii) *Handling fluids*: While the parts in most mechanical assemblies interact directly with one another via contact relationships, some assemblies use fluid interactions to transform a driving force into movement (e.g., pumps and hydraulic machines). One approach for supporting such assemblies would be to incorporate a fluid simulation into our analysis technique. (iii) *Visualizing forces*: In addition to visualizing motion, some *how-things-work* illustrations also depict the direction and magnitude of physical forces, such as friction, torque and pressure, that act on various parts within the assembly. Automatically depicting such forces is an open research challenge. □

**References**

1. Agrawala, M., Phan, D., Heiser, J., Haymaker, J., Klingner, J., Hanrahan, P., Tversky, B. Designing effective step-by-step assembly instructions. In *Proceedings of ACM SIGGRAPH* (2003), 828–837.
2. Amerongen, C.V. *The Way Things Work: An Illustrated Encyclopedia of Technology*, Simon and Schuster, 1967.
3. Brain, M. *How Stuff Works, Hungry Minds*, New York, 2001.
4. Burns, M., Finkelstein, A. Adaptive cutaways for comprehensible rendering of polygonal scenes. In *ACM TOG (SIGGRAPH Asia)* (2008), 1–7.
5. Davidson, J.K., Hunt, K.H. *Robots and Screw Theory: Applications of Kinematics and Statics to Robotics*, Oxford University Press, 2004.
6. Feiner, S., Seligmann, D. Cutaways and ghosting: Satisfying visibility constraints in dynamic 3D illustrations. *Vis. Comput.* 8, 5 (1992), 292–302.
7. Gal, R., Sorkine, O., Mitra, N.J., Cohen-Or, D. iWIRES: An analyze-and-edit approach to shape manipulation. *ACM TOG (SIGGRAPH)* 28, 3:#33 (2009), 1–10.
8. Goldman, D.B., Curless, B., Salesin, D., Seitz, S.M. Schematic storyboarding for video visualization and editing. *ACM TOG (SIGGRAPH)* 25, 3 (2006), 862–871.
9. Hegarty, M. Mental animation: Inferring motion from static displays of mechanical systems. *J. Exp. Psychol. Learn. Mem. Cognit.* 18, 5 (1992), 1084–1102.
10. Hegarty, M. Capacity limits in diagrammatic reasoning. In *Theory and Application of Diagrams* (2000), 335–348.
11. Hegarty, M., Kriz, S., Cate, C. The roles of mental animations and external animations in understanding mechanical systems. *Cognit. Instruct.* 21, 4 (2003), 325–360.
12. Heiser, J., Tversky, B. Arrows in comprehending and producing mechanical diagrams. *Cognit. Sci.* 30 (2006), 581–592.

13. Karpenko, O., Li, W., Mitra, N., Agrawala, M. Exploded view diagrams of mathematical surfaces. *IEEE Vis.* 16, 6 (2010), 1311–1318.
14. Kriz, S., Hegarty, M. Top-down and bottom-up influences on learning from animations. *Int. J. Hum. Comput. Stud.* 65, 11 (2007), 911–930.
15. Langone, J. *National Geographic's How Things Work: Everyday Technology Explained*, National Geographic, 1999.
16. Li, W., Agrawala, M., Curless, B., Salesin, D. Automated generation of interactive 3d exploded view diagrams. *ACM TOG (SIGGRAPH)*, 27, 3 (2008).
17. Li, W., Ritter, L., Agrawala, M., Curless, B., Salesin, D. Interactive cutaway illustrations of complex 3d models. *ACM TOG (SIGGRAPH)*, 26, 3:#31 (2007), 1–11.
18. Macaulay, D. *The New Way Things Work*, Houghton Mifflin Books for Children, 1998.
19. Mayer, R. *Multimedia Learning*, Cambridge University Press, 2001.
20. McGuffin, M.J., Tancau, L., Balakrishnan, R. Using deformations for browsing volumetric data. In *IEEE Visualization* (2003).
21. Mehra, R., Zhou, Q., Long, J., Sheffer, A., Gooch, A., Mitra, N.J. Abstraction of man-made shapes. In *Proceedings of ACM TOG (SIGGRAPH Asia)* (2009), 1–10.
22. Mitra, N.J., Guibas, L., Pauly, M. Partial and approximate symmetry detection for 3D geometry. *ACM TOG (SIGGRAPH)* 25, 3 (2006), 560–568.
23. Mitra, N.J., Yang, Y.L., Yan, D.M., Li, W., Agrawala, M. Illustrating how mechanical assemblies work. *ACM TOG (SIGGRAPH)*, 29 (2010), 58:1–58:12.
24. Narayanan, N., Hegarty, M. On designing comprehensible interactive hypermedia manuals. *Int. J. Hum. Comput. Stud.* 48, 2 (1998), 267–301.
25. Narayanan, N., Hegarty, M. Multimedia design for communication of dynamic information. *Int. J. Hum. Comput. Stud.* 57, 4 (2002), 279–315.
26. Nienhaus, M., Döllner, J. Depicting dynamics using principles of visual art and narrations. *IEEE Comput. Graph. Appl.* 25, 3 (2005), 40–51.
27. Seligmann, D., Feiner, S. Automated generation of intent-based 3D illustrations. In *Proceedings of ACM SIGGRAPH* (1991), ACM, 132.
28. Tversky, B., Morrison, J.B., Betrancourt, M. Animation: Can it facilitate? *Int. J. Hum. Comput. Stud.* 5 (2002), 247–262.
29. Viola, I., Kanitsar, A., Gröller, M.E. Importance-driven volume rendering. In *IEEE Visualization* (2004), 139–145.

**Niloy J. Mitra** University College London.

**Wilmot Li** Adobe.

**Yong-Liang Yang, Dong-Ming Yan** King Abdullah University of Science and Technology (KAUST).

**Maneesh Agrawala** University of California, Berkeley.

Text excerpt and illustrations from *The Way Things Work* by David Macaulay. Compilation copyright (c) 1988, 1998 Dorling Kindersley, Ltd., London. Text copyright (c) 1988, 1998 David Macaulay, Neil Ardley. Illustrations copyright (c) 1988, 1998 David Macaulay. Used by permission of Houghton Mifflin Harcourt Publishing Company. All rights reserved.

© 2013 ACM 0001-0782/13/01

# World-Renowned Journals from ACM

ACM publishes over 50 magazines and journals that cover an array of established as well as emerging areas of the computing field. IT professionals worldwide depend on ACM's publications to keep them abreast of the latest technological developments and industry news in a timely, comprehensive manner of the highest quality and integrity. For a complete listing of ACM's leading magazines & journals, including our renowned Transaction Series, please visit the ACM publications homepage: [www.acm.org/pubs](http://www.acm.org/pubs).

## ACM Transactions on Interactive Intelligent Systems



ACM Transactions on Interactive Intelligent Systems (TIIS). This quarterly journal publishes papers on research encompassing the design, realization, or evaluation of interactive systems incorporating some form of machine intelligence.

## ACM Transactions on Computation Theory



ACM Transactions on Computation Theory (ToCT). This quarterly peer-reviewed journal has an emphasis on computational complexity, foundations of cryptography and other computation-based topics in theoretical computer science.

PLEASE CONTACT ACM MEMBER SERVICES TO PLACE AN ORDER  
 Phone: 1.800.342.6626 (U.S. and Canada) +1.212.626.0500 (Global)  
 Fax: +1.212.944.1318 (Hours: 8:30am–4:30pm, Eastern Time)  
 Email: [acmhelp@acm.org](mailto:acmhelp@acm.org)  
 Mail: ACM Member Services  
 General Post Office  
 PO Box 30777  
 New York, NY 10087-0777 USA



Association for Computing Machinery

Advancing Computing as a Science & Profession

[www.acm.org/pubs](http://www.acm.org/pubs)

# Technical Perspective

## Finding People In Depth

By James M. Rehg

WHEN THE MICROSOFT Kinect for Xbox 360 was introduced in November 2010, it was an instant success. Via the Kinect, users can control their Xbox through natural body gestures and commands thanks to a *depth camera* that enables gesture recognition. In contrast to a conventional camera, which measures the color at each pixel location, a depth camera returns the distance to that point in the scene. Depth cameras make it easy to separate the Xbox user from the background of the room, and reduce the complexities caused by color variation, for example, in clothing.

While the role of the depth camera in the success of the Kinect is well-known, what is less well-known is the innovative computer vision technology that underlies the Kinect's gesture recognition capabilities. The following article by Shotton et al. describes a landmark computer vision system that takes a single depth image containing a person and automatically estimates the pose of the person's body in 3D. This novel method for pose estimation is the key to the Kinect's success.

Three important ideas define the Kinect architecture: tracking by detection, data-driven learning, and discriminative part models. These ideas have their origin in object recognition and tracking research from the computer vision community over the past 10 years. Their development in the Kinect has led to some exciting and innovative work on feature representations and training methods. The resulting system is a dramatic improvement over the previous state of the art.

In order to recognize a user's gesture, the Kinect must track the user's motion in a sequence of depth images. An important aspect of the Kinect architecture is that body poses are detected independently in each frame, without incorporating information


from previous frames. This *tracking by detection* approach has the potential for greater robustness because errors made over time are less likely to accumulate. It is enabled by an extremely efficient and reliable solution to the pose estimation problem.

The challenge of pose estimation, as in other vision problems, is to reliably measure the desired variables, while remaining unaffected by other sources of variability. Body pose is described by a vector of joint angles. When you bend your elbow, for example, you are changing one joint angle. However, the appearance of your elbow in a sequence of depth images is affected by many factors: your position and orientation with respect to the camera, the clothing you are wearing, whether your build is thin or stocky, and so forth. An additional challenge comes from the large number of pose variables. Around 30 joint angles are needed to describe the basic configurations of the human body. If each joint could assume only five positions, this would result in  $5^{30}$  possible poses. Fortunately, joints are coupled during coordinated movement, and many achievable poses, such as those found in yoga, are rarely encountered in general settings.

The authors employ *data-driven learning* to address the tremendous variability in pose and appearance. Motion capture data was used to characterize the space of possible poses: actors performed gestures used in gaming (for example, dancing or kicking) and their joint angles were measured, resulting in a dataset of 100,000 poses. Given a single pose, a simulated depth image can be produced by transferring the pose to a character model and rendering the clothing and hair. By varying body types and sizes, and by sampling different clothing and hairstyles, the authors automatically obtained a huge training dataset of depth images.

The final idea is the use of *discriminative part models* to represent the body pose. Parts are crucial. They decompose the problem of predicting the pose into a series of independent subproblems: given an input depth image, each pixel is labeled with its corresponding part, and the parts are grouped into hypotheses about joint locations. Each pixel can be processed independently in this approach, making it possible to leverage the Xbox GPU and obtain real-time performance. This efficiency is enhanced by a clever feature design.

The Kinect's impact has extended well beyond the gaming market. It has become a popular sensor in the robotics community, where its low cost and ability to support human-robot interaction are hugely appealing. A survey of the two main robotics conferences in 2012 (IROS and ICRA) reveals that among the more than 1,600 papers, 9% mentioned the Kinect. At Georgia Tech, we are using the Kinect to measure children's behavior, in order to support the research and treatment of autism, and other developmental and behavioral disorders.

In summary, the Kinect is a potent combination of innovative hardware and software design, informed by decades of computer vision research. The proliferation of depth camera technology in the coming years will enable new advances in vision-based sensing and support an increasingly diverse set of applications. 

**James M. Rehg** (rehg@gatech.edu) is a professor in the School of Interactive Computing at the Georgia Institute of Technology, Atlanta, where he directs the Center for Behavior Imaging and co-directs the Computational Perception Lab.

# Real-Time Human Pose Recognition in Parts from Single Depth Images

By Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore

## Abstract

We propose a new method to quickly and accurately predict human *pose*—the 3D positions of body joints—from a single depth image, without depending on information from preceding frames. Our approach is strongly rooted in current object recognition strategies. By designing an intermediate representation in terms of body parts, the difficult pose estimation problem is transformed into a simpler per-pixel classification problem, for which efficient machine learning techniques exist. By using computer graphics to synthesize a very large dataset of training image pairs, one can train a classifier that estimates body part labels from test images invariant to pose, body shape, clothing, and other irrelevances. Finally, we generate confidence-scored 3D proposals of several body joints by projecting the classification result and finding local modes.

The system runs in under 5ms on the Xbox 360. Our evaluation shows high accuracy on both synthetic and real test sets, and investigates the effect of several training parameters. We achieve state-of-the-art accuracy in our comparison with related work and demonstrate improved generalization over exact whole-skeleton nearest neighbor matching.

## 1. INTRODUCTION

Robust interactive human body tracking has applications including gaming, human–computer interaction, security, telepresence, and health care. Human pose estimation from video has generated a vast literature (surveyed in Moeslund et al.<sup>12</sup> and Poppe<sup>17</sup>). Early work used standard video cameras, but the task has recently been greatly simplified by the introduction of real-time depth cameras.

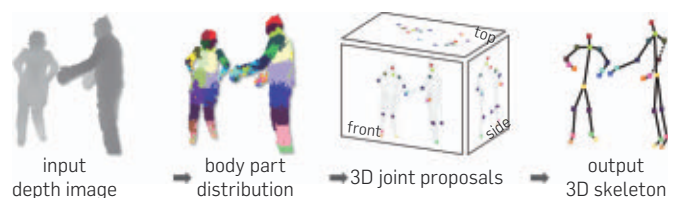
Depth imaging technology has advanced dramatically over the last few years, finally reaching a consumer price point with the launch of Kinect for Xbox 360. Pixels in a depth image record depth in the scene, rather than a measure of intensity or color. The Kinect camera gives a  $640 \times 480$  image at 30 frames per second with depth resolution of a few centimeters. Depth cameras offer several advantages over traditional intensity sensors, which are working in low light levels, giving a calibrated scale estimate, and being color and texture invariant. They also greatly simplify the task of background subtraction, which we assume in this work. Importantly for our approach, it is rather easier to use computer graphics to synthesize realistic depth images of people than to synthesize color images, and thus to build a large training dataset cheaply.

However, even the best existing depth-based systems for human pose estimation<sup>16,22</sup> still exhibit limitations. In particular, until the launch of Kinect for Xbox 360, of which the algorithm described in this paper is a key component, none ran at interactive rates on consumer hardware while handling a full range of human body shapes and sizes undergoing general body motions.

### 1.1. Problem overview

The overall problem we wish to solve is stated as follows. The input is a stream of depth images, that is, the image  $I_t$  at time  $t$  comprises a 2D array of  $N$  distance measurements from the camera to the scene. Specifically, an image  $I$  encodes a function  $d(x)$  which maps 2D coordinates  $x$  to the distance to the first opaque surface along the pixel's viewing direction. The output of the system is a stream of 3D *skeletons*, each skeleton being a vector of about 30 numbers representing the body configuration (e.g., joint angles) of each person in the corresponding input image. Denoting the output skeleton(s) at frame  $t$  by  $\theta_t$ , the goal is to define a function  $F$  such that  $\theta_t = F(I_t, I_{t-1}, \dots)$ . This is a standard formulation, in which the output at time  $t$  may depend on information from earlier images as well as from the current image. Our solution, illustrated in Figure 1, pipelines the function  $F$  into two

**Figure 1. System overview. From a single input depth image, a per-pixel body part distribution is inferred. (Colors indicate the most likely part labels at each pixel and correspond in the joint proposals.) Local modes of this signal are estimated to give high-quality proposals for the 3D locations of body joints, even for multiple users. Finally, the joint proposals are input to skeleton fitting, which outputs the 3D skeleton for each user.**



The original version of this paper appeared in the *Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition*, 129–1304.

intermediate representations: a *body parts image*  $C_t$  is first computed, which stores a vector of 31 probabilities at every pixel, indicating the likelihood that the world point under that pixel is each of 31 standard body parts. The second intermediate representation is a list of *joint hypotheses*  $J_t$ , which contains triples of (body part, 3D position, confidence) hypotheses, with say five hypotheses per body part. Finally, the joint hypotheses are searched for kinematically consistent skeletons. With these intermediate representations, pseudocode for  $F$  may be written as the steps

$$C_t = \text{ComputeBodyParts}(I_t) \quad (1)$$

$$J_t = \text{ComputeJointHypotheses}(C_t, I_t) \quad (2)$$

$$\theta_t = \text{FitSkeleton}(J_t, \theta_{t-1}) \quad (3)$$

An important attribute of our solution is that only the final stage uses information from previous frames, so that much of the image interpretation is performed independently on every frame. This greatly enhances the system's ability to recover from tracking errors, which are inimical to almost all existing solutions. In this paper, we focus on the first two stages of the pipeline: the computation of body part labels and joint hypotheses from a single depth image. As we will be dealing with each input frame independently, the  $t$  subscripts will be elided in the subsequent exposition.

## 1.2. Related work

Previous work on the problem has, as noted above, been largely focussed on pose estimation from conventional intensity images (the 2D problem), but of course many of the ideas transfer to the 3D case, and, as we show below, our approach may also be applied to the 2D problem. Thus, we discuss both 2D- and 3D-based work in this section. Many 2D systems make use of a known background and use only the human silhouette as a basis for estimation, thereby gaining invariance to clothing, at the cost of increased ambiguity in the solution.

Of particular interest are systems that perform “one-shot” estimation, from a single image. Such systems are valuable not just as a way of avoiding the error accumulation of tracking-based systems, but also because testing and evaluation of a one-shot system is much simpler than testing a tracking-based solution. Agarwal and Triggs<sup>1</sup> treated pose estimation as a nonlinear regression problem, estimating pose directly from silhouette images. Given a *training set*  $T$  of  $(I, \theta)$  pairs, and parameterizing the function  $F$  (writing it  $F_\Phi$ ), the parameters  $\Phi$  are found that optimize accuracy on the training set, typically a function of the form  $E(\Phi) = \sum_{(I, \theta) \in T} d(\theta, F_\Phi(I))$  for some accuracy measure  $d(\cdot, \cdot)$ . This approach forms the essence of the machine learning approach to pose estimation, which is also followed in our work. A large number of papers based on this approach improved the regression models, dealt with occlusion and ambiguity, and re-incorporated temporal information.<sup>13, 25</sup> Such models, however, by

classifying whole-body pose in one monolithic function, would require enormous amounts of training data in order to build a fully general purpose tracker as required in the Kinect system. A second difficulty with the regression approach is that the existing methods are quite computationally expensive.

One approach to reducing the demand for training data is to divide the body into parts and attempt to combine the per-part estimates to produce a single pose estimate. Ramanan and Forsyth,<sup>18</sup> for example, find candidate body segments as pairs of parallel lines, clustering appearances across frames. Sigal *et al.*<sup>24</sup> use eigen-appearance template detectors for head, upper arms and lower legs, and nonparametric belief propagation to infer whole body pose. Wang and Popović<sup>26</sup> track a hand clothed in a colored glove. Our system could be seen as automatically inferring the colors of a virtual colored suit from a depth image. However, relatively little work has looked at recognizing parts of the human body. Zhu and Fujimura<sup>28</sup> build heuristic detectors for coarse upper body parts (head, torso, arms) using a linear programming relaxation but require the user to stand in a “T” pose to initialize the model. Most similar to our approach, Plagemann *et al.*<sup>16</sup> build a 3D mesh to find geodesic extrema interest points, which are classified into three parts: head, hand, and foot. Their method provides both location and orientation estimate of these parts, but does not distinguish left from right and the use of interest points limits the choice of parts.

## 1.3. Approach

Our approach builds on recent advances in object recognition,<sup>7, 27</sup> which can identify object classes such as “sheep,” “building,” and “road” in general 2D images. In particular, the use of randomized decision forests allows recognition from a 20-class lexicon to be performed in real time.<sup>20</sup>

The adaptation to the pose estimation problem is relatively straightforward. Starting with a 3D surface model of a generic human body, the surface is divided into 31 distinct body parts (Section 3.1). An object recognition algorithm is trained to recognize these parts, so that at run time, a single input depth image is segmented into a dense probabilistic body parts labeling.

Reprojecting the inferred parts into world space, we localize spatial modes of each part distribution and thus generate (possibly several) confidence-weighted proposals for the 3D locations of each skeletal joint. A combination of these joint locations comprises the output pose  $\theta$ . Each proposal carries an inferred confidence value, which can be used by any downstream tracking algorithm Eq. (3) for robust initialization and recovery.

We treat the segmentation into body parts as a per-pixel classification task. (In contrast to classification tasks in object recognition or image segmentation, the machinery of Markov Random Fields has not proved necessary in our application.) Evaluating each pixel separately avoids a combinatorial search over the different body joints, although within a single part there are of course still

dramatic differences in the contextual appearance (see Figure 2). For training data, we generate realistic synthetic depth images of humans of many shapes and sizes in highly varied poses sampled from a large motion capture database. The classifier used is a deep randomized decision forest, which is well suited to our multi-class scenario and admits extremely high-speed implementation. The primary challenge imposed by this choice is the need for large amounts of training data, easily obtained given our use of synthetic imagery. The further challenge of building a distributed infrastructure for decision tree training was important to the success of our approach, but is beyond the scope of this paper.

An optimized implementation of our algorithm runs in under 5ms per frame on the Xbox 360 GPU, at least one order of magnitude faster than existing approaches. It works frame by frame across dramatically differing body shapes and sizes, and the learned discriminative approach naturally handles self-occlusions and poses cropped by the image frame. We evaluate both real and synthetic depth images, containing challenging poses of a varied set of subjects. Even without exploiting temporal or kinematic constraints, the 3D joint proposals are both accurate and stable. We investigate the effect of several training parameters and show how very deep trees can still avoid overfitting due to the large training set. Further, results on silhouette images suggest more general applicability of our approach.

#### 1.4. Contributions

Our main contribution is to treat pose estimation as object recognition using a novel intermediate body parts representation designed to spatially localize joints of interest at low computational cost and high accuracy. Our experiments also carry several insights: (i) synthetic depth training data is an excellent proxy for real data; (ii) scaling up the learning problem with varied synthetic data is important for high accuracy; and (iii) our parts-based approach generalizes better than even an oracular whole-image nearest neighbor algorithm.

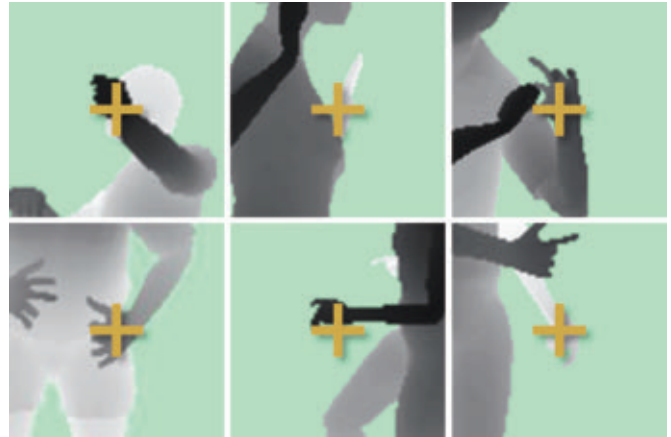
#### 1.5. Sensor characteristics

Before describing our algorithm in detail, we describe the process by which we generate training data for human pose estimation. In order to do so, we first describe the characteristics of the depth sensor we employ, as those characteristics must be replicated in the synthetic data generation.

As described above, the camera produces a  $640 \times 480$  array of depth values, with the following characteristics.

- Certain materials do not reflect infrared wavelengths of light effectively, and so ‘drop out’ pixels can be common. This particularly affects hair and shiny surfaces.
- In bright sunlight, the ambient infrared can swamp the active signal preventing any depth inference.
- The depth range is limited by the power of the emitter, and safety considerations result in a typical operating range of about 4 m.
- The depth noise level ranges from a few millimeters close up to a few centimeters for more distant pixels.

**Figure 2. Example renderings focusing on one hand, showing the range of appearances a single point on the body may exhibit.**



- The sensor operates on the principle of stereo matching between an emitter and camera, which must be offset by some baseline. Consequently, an occlusion shadow appears on one side of objects in the depth camera.
- The occluding contours of objects are not precisely delineated and can flicker between foreground and background.

## 2. TRAINING DATA

Pose estimation research has often focussed on techniques to overcome lack of training data,<sup>13</sup> because of two problems. First, generating realistic intensity images using computer graphics techniques<sup>14, 15, 19</sup> is hampered by the huge color and texture variability induced by clothing, hair, and skin, often meaning that the data is reduced to 2D silhouettes.<sup>1</sup> Although depth cameras significantly reduce this difficulty, considerable variation in body and clothing *shape* remains. The second limitation is that synthetic body pose images are of necessity fed by motion-capture (‘mocap’) data, which is expensive and time-consuming to obtain. Although techniques exist to simulate human motion (e.g., Sidenbladh et al.<sup>23</sup>), they do not yet produce the range of volitional motions of a human subject.

### 2.1. Motion capture data

The human body is capable of an enormous range of poses, which are difficult to simulate. Instead, we capture a large database of motion capture of human actions. Our aim was to span the wide variety of poses people would make in an entertainment scenario. The database consists of approximately 500,000 frames in a few hundred sequences including actions such as driving, dancing, kicking, running, and navigating menus.

We expect our semi-local body part classifier to *generalize* somewhat to unseen poses. In particular, we need not record all possible combinations of the different limbs; in practice, a wide range of poses prove sufficient. Further, we need not record mocap with variation in rotation about the



vertical axis, mirroring left–right, scene position, body shape and size, or camera pose, all of which can be added in post-hoc.

Since the classifier uses no temporal information, we are interested only in static *poses* and not motion. Often, changes in pose from one mocap frame to the next are so small as to be insignificant. We thus discard many similar, redundant poses from the initial mocap data using ‘furthest neighbor’ clustering<sup>10</sup> where the distance between poses  $\theta_1$  and  $\theta_2$  is defined as  $\max_j \|\theta_1^j - \theta_2^j\|_2$ , the maximum Euclidean distance over body joints  $j$ . We use a subset of 100,000 poses such that no two poses are closer than 5cm.

We have found it necessary to iterate the process of motion capture, sampling from our model, training the classifier, and testing joint prediction accuracy in order to refine the mocap database with regions of pose space that had been previously missed out. Our early experiments employed the CMU mocap database,<sup>5</sup> which gave acceptable results though covered far less of pose space.

## 2.2. Generating synthetic data

We have built a randomized rendering pipeline from which we can sample fully labeled training images. Our goals in building this pipeline were twofold: realism and variety. For the learned model to work well, the samples must closely resemble real camera images and contain good coverage of the appearance variations we hope to recognize at test time. While depth/scale and translation variations are handled explicitly in our features (see below), other invariances cannot be encoded efficiently. Instead, we learn invariances—to camera pose, body pose, and body size and shape—from the data.

The synthesis pipeline first randomly samples a pose from the mocap database, and then uses standard computer graphics techniques to render depth and (see below) body parts images from texture-mapped 3D meshes. The pose is retargeted to each of 15 base meshes (see Figure 3) spanning the range of body shapes and sizes. Further, slight random variation in height and weight gives extra coverage of body shapes. Other randomized parameters include camera pose, camera noise, clothing, and hairstyle. Figure 4 compares the varied output of the pipeline to hand-labeled real camera images.

In detail, the variations are as follows:

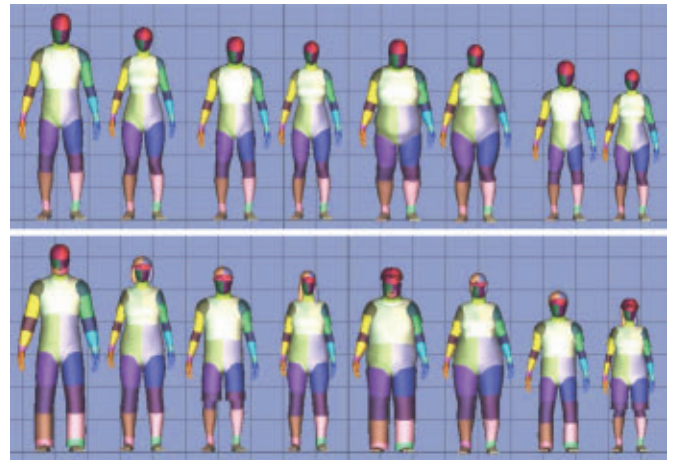
**Base character.** We use 3D models of 15 varied base characters, both male and female, from child to adult, short to tall, and thin to fat. Some examples are shown in Figure 3 (top row). A given render will pick uniformly at random from the characters.

**Pose.** Having discarded redundant poses from the mocap data, we retarget the remaining poses to each base character and choose uniformly at random. The pose is also mirrored left–right with probability  $\frac{1}{2}$  to prevent a left or right bias.

**Rotation and translation.** The character is rotated about the vertical axis and translated in the scene, uniformly at random.

**Hair and clothing.** We add mesh models of several hairstyles and items of clothing chosen at random; some examples are shown in Figure 3 (bottom row).

**Figure 3. Renders of several base character models. Top row: bare models. Bottom row: with random addition of hair and clothing.**



**Weight and height variation.** The base characters already have a wide variety of weights and heights. To add further variety, we add an extra variation in height (vertical scale  $\pm 10\%$ ) and weight (overall scale  $\pm 10\%$ ).

**Camera position and orientation.** The camera height, pitch, and roll are chosen uniformly at random within a range believed to be representative of an entertainment scenario in a home living room.

**Camera noise.** Real depth cameras exhibit noise. We distort the clean computer graphics renders with dropped out pixels, depth shadows, spot noise, and disparity quantization to match the camera output as closely as possible. In practice however, we found that this noise addition had little effect on accuracy, perhaps due to the quality of the cameras or the more important appearance variations due to other factors such as pose.

We use a standard graphics rendering pipeline to generate the scene, consisting of a depth image paired with its body parts label image. Examples are given in Figure 4.

## 3. BODY PART INFERENCE AND JOINT PROPOSALS

In this section, we describe our intermediate body parts representation, detail the discriminative depth image features, review decision forests and their application to body part recognition, and finally discuss how a mode finding algorithm is used to generate joint position proposals.

### 3.1. Body part labeling

A key innovation of this work is the form of our intermediate body parts representation. We define several localized body part labels that densely cover the body, as color-coded in Figure 4. The parts are defined by assigning a label to each triangle of the mesh used for rendering of the synthetic data. Because each model is in vertex-to-vertex correspondence, each triangle is associated with the same part of the body in each rendered image. The precise definitions of the body parts are somewhat arbitrary: the number of parts was chosen at 31 after some initial experimentation with smaller numbers, and it is

**Figure 4. Synthetic and real data. Pairs of depth image and ground truth body parts. Note wide variety in pose, shape, clothing, and crop.**



convenient to fit the label in 5 bits. The definitions of some of the parts are in terms of particular skeletal joints of interest, for example, ‘all triangles intersecting the sphere of radius 10 cm centered on the left hand.’ Other parts fill the gaps between these parts. Despite these apparently arbitrary choices, later attempts to optimize the parts distribution have not proved significantly better than the set described in this paper.

For the experiments in this paper, the parts used are named as follows: LU/RU/LW/RW head, neck, L/R shoulder, LU/RU/LW/RW arm, L/R elbow, L/R wrist, L/R hand, LU/RU/LW/RW torso, LU/RU/LW/RW leg, L/R knee, L/R ankle, L/R foot (Left, right, upper, lower). Distinct parts for left and right allow the classifier to disambiguate the left and right sides of the body. Even though this distinction may be ambiguous, the probabilistic label we output can usefully use even ambiguous labels.

Of course, the precise definition of these parts could be changed to suit a particular application. For example, in an upper body tracking scenario, all the lower body parts could be merged. Parts should be sufficiently small to accurately localize body joints, but not too numerous as to waste capacity of the classifier.

### 3.2. Depth image features

We employ simple depth comparison features, inspired by those in Lepetit et al.<sup>11</sup> At a given pixel with 2D coordinates  $\mathbf{x}$ , the features compute

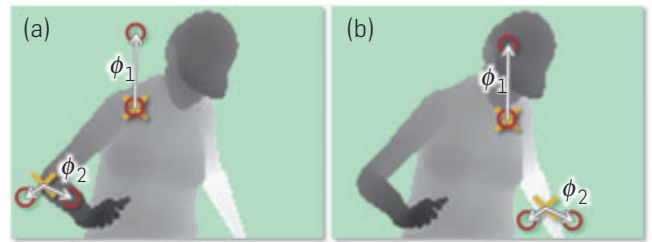
$$f_\phi(I, \mathbf{x}) = d_I\left(\mathbf{x} + \frac{\mathbf{u}}{d_I(\mathbf{x})}\right) - d_I\left(\mathbf{x} + \frac{\mathbf{v}}{d_I(\mathbf{x})}\right), \quad (4)$$

where  $d_I(\mathbf{x})$  is the depth at pixel  $\mathbf{x}$  in image  $I$ , and parameters  $\phi = (\mathbf{u}, \mathbf{v})$  describe offsets  $\mathbf{u}$  and  $\mathbf{v}$ . The normalization of the offsets by  $\frac{1}{d_I(\mathbf{x})}$  ensures that the features are depth invariant: at a given point on the body, a fixed *world space* offset will result whether the pixel is close or far from the camera. The features are thus 3D translation invariant (modulo perspective effects). If an offset pixel lies on the background or outside the bounds of the image, the depth probe  $d_I(\mathbf{x}')$  is given a large positive constant value.

Figure 5 illustrates two features at different pixel locations  $\mathbf{x}$ . Feature  $f_{\phi_1}$  looks upward: Eq. (4) will give a large positive response for pixels  $\mathbf{x}$  near the top of the body, but a value close to zero for pixels  $\mathbf{x}$  lower down the body.

Feature  $f_{\phi_2}$  may instead help find thin vertical structures such as the arm.

**Figure 5. Depth image features. The yellow crosses indicate the pixel  $\mathbf{x}$  being classified. The red circles indicate the offset pixels as defined in Eq. (4). (a) The two example features give a large depth difference response. (b) The same two features at new image locations give a much smaller response.**



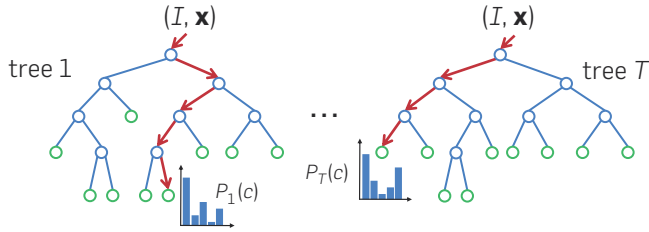
Individually, these features provide only a weak signal about which part of the body the pixel belongs to, but in combination in a decision forest they are sufficient to accurately disambiguate all trained parts. The design of these features was strongly motivated by their computational efficiency: no preprocessing is needed; each feature need read at most three image pixels and perform at most five arithmetic operations; and the features can be straightforwardly implemented on the GPU. Given a larger computational budget, one could employ potentially more powerful features based on, for example, depth integrals over regions, curvature, or local descriptors, for example, shape contexts.<sup>3</sup>

### 3.3. Randomized decision forests

Randomized decision trees and forests<sup>2,4</sup> have proven fast and effective multi-class classifiers for many tasks, and can be implemented efficiently on the GPU.<sup>20</sup> As illustrated in Figure 6, a forest is an ensemble of  $T$  decision trees, each consisting of split and leaf nodes. Each split node consists of a feature  $\phi$  and a threshold  $\tau$ . To classify pixel  $\mathbf{x}$  in image  $I$ , the current node is set to the root, and then Eq. (4) is evaluated. The current node is then updated to the left or right child according to the comparison  $f_\phi(I, \mathbf{x}) < \tau$ , and the process repeated until a leaf node is reached. At the leaf node reached in tree  $t$ , a learned distribution  $P_t(c|I, \mathbf{x})$  over body part labels  $c$  is stored. The distributions are averaged together for all trees in the forest to give the final classification

$$P(c|I, \mathbf{x}) = \frac{1}{T} \sum_{t=1}^T P_t(c|I, \mathbf{x}). \quad (5)$$

**Figure 6. Randomized Decision Forests.** A forest is an ensemble of trees. Each tree consists of split nodes (blue) and leaf nodes (green). The red arrows indicate the different paths that might be taken by different trees for a particular input.



**Training.** Each tree is trained on a different set of randomly synthesized images. A random subset of 2000 example pixels from each image is chosen to ensure a roughly even distribution across body parts. Each tree is trained using the algorithm in Lepetit et al.<sup>11</sup> To keep the training times down, we employ a distributed implementation. Training three trees to depth 20 from 1 million images takes about a day on a 1000 core cluster.

### 3.4. Joint position proposals

Body part recognition as described above infers per-pixel information. This information must now be pooled across pixels to generate reliable proposals for the positions of 3D skeletal joints. These proposals are the final output of our algorithm and could be used by a tracking algorithm to self-initialize and recover from failure.

A simple option is to accumulate the global 3D centers of probability mass for each part, using the known calibrated depth. However, outlying pixels severely degrade the quality of such a global estimate. We consider two algorithms: a fast algorithm based on simple bottom-up clustering and a more accurate algorithm based on mean shift, which shall now be described.

We employ a local mode-finding approach based on mean shift<sup>6</sup> with a weighted Gaussian kernel. We define a density estimator per body part as

$$f_c(\hat{\mathbf{x}}) \propto \sum_{i=1}^N w_{ic} \exp\left(-\left\|\frac{\hat{\mathbf{x}} - \hat{\mathbf{x}}_i}{b_c}\right\|^2\right), \quad (6)$$

where  $\hat{\mathbf{x}}$  is a coordinate in 3D space,  $N$  is the number of image pixels,  $w_{ic}$  is a pixel weighting,  $\hat{\mathbf{x}}_i$  is the reprojection of image pixel  $\mathbf{x}_i$  into world space given depth  $d_i(\mathbf{x}_i)$ , and  $b_c$  is a learned per-part bandwidth. The pixel weighting  $w_{ic}$  considers both the inferred body part probability at the pixel and the world surface area of the pixel:

$$w_{ic} = P(c|I, \mathbf{x}_i) \cdot d_i(\mathbf{x}_i)^2. \quad (7)$$

This ensures that density estimates are depth invariant and give a small but significant improvement in joint prediction accuracy. Depending on the definition of body parts, the posterior  $P(c|I, \mathbf{x})$  can be pre-accumulated over a small set of parts. For example, in our experiments the four body parts covering the head are merged to localize the head joint.

Mean shift is used to find modes in this density efficiently. All pixels above a learned probability threshold  $\lambda_c$  are used as starting points for part  $c$ . A final confidence estimate is given as a sum of the pixel weights reaching each mode. This proved more reliable than taking the modal density estimate.

The detected modes lie on the *surface* of the body. Each mode is therefore pushed back into the scene by a learned  $z$  offset  $\zeta_c$  to produce a final joint position proposal. This simple, efficient approach works well in practice. The bandwidths  $b_c$ , probability threshold  $\lambda_c$ , and surface-to-interior  $z$  offset  $\zeta_c$  are optimized per-part on a hold-out validation set of 5000 images by grid search. (As an indication, this resulted in mean bandwidth 0.065m, probability threshold 0.14, and  $z$  offset 0.039m).

## 4. EXPERIMENTS

In this section, we describe the experiments performed to evaluate our method. We show both qualitative and quantitative results on several challenging datasets and compare with both nearest-neighbor approaches and the state of the art.<sup>9</sup> We provide further results in the supplementary material. Unless otherwise specified, parameters below were set as 3 trees, 20 deep, 300k training images per tree, 2000 training example pixels per image, 2000 candidate features  $\phi$ , and 50 candidate thresholds  $\tau$  per feature.

**Test data.** We use challenging synthetic and real depth images to evaluate our approach. For our synthetic test set, we synthesize 5000 depth images, together with the ground truth body parts labels and joint positions. The original mocap *poses* used to generate these images are held out from the training data. Our real test set consists of 8808 frames of real depth images over 15 different subjects, hand-labeled with dense body parts and seven upper body joint positions. We also evaluate on the real depth data from Ganapathi et al.<sup>8</sup> The results suggest that effects seen on synthetic data are mirrored in the real data, and further that our synthetic test set is by far the ‘hardest’ due to the extreme variability in pose and body shape. For most experiments, we limit the rotation of the user to  $\pm 120^\circ$  in both training and synthetic test data, since the user faces the camera ( $0^\circ$ ) in our main entertainment scenario, though we also evaluate the full  $360^\circ$  scenario.

**Error metrics.** We quantify both classification and joint prediction accuracy. For classification, we report the average per-class accuracy: the average of the diagonal of the confusion matrix between the ground truth part label and the most likely inferred part label. This metric weights each body part equally despite their varying sizes, though mislabelings on the part boundaries reduce the absolute numbers.

For joint proposals, we generate recall-precision curves as a function of confidence threshold. We quantify accuracy as average precision per joint, or mean average precision (mAP) over all joints. The first joint proposal within  $D$  meters of the ground truth position is taken as a true positive, while other proposals also within  $D$  meters count as false positives. This penalizes multiple spurious detections near the correct position, which might slow a downstream

tracking algorithm. Any joint proposals outside  $D$  meters also count as false positives. Note that *all* proposals (not just the most confident) are counted in this metric. Joints invisible in the image are not penalized as false negatives. Although final applications may well require these joints, it is assumed that their prediction is more the task of the sequential tracker Eq. (3). We set  $D = 0.1m$  below, approximately the accuracy of the hand-labeled real test data ground truth. The strong correlation of classification and joint prediction accuracy (the blue curves in Figures 8(a) and 10(a)) suggests that the trends observed below for one also apply for the other.

#### 4.1. Qualitative results

Figure 7 shows example inferences of our algorithm. Note high accuracy of both classification and joint prediction across large variations in body and camera pose, depth in scene, cropping, and body size and shape (e.g., small child versus heavy adult). The bottom row shows some failure modes of the body part classification. The first example shows a failure to distinguish subtle changes in the depth image such as the crossed arms. Often (as with the second and third failure examples), the most likely body part is incorrect, but there is still sufficient correct probability mass in distribution  $P(c|I, \mathbf{x})$  that an accurate proposal can be generated. The fourth example shows a

failure to generalize well to an unseen pose, but the confidence gates bad proposals, maintaining high precision at the expense of recall.

Note that no temporal or kinematic constraints (other than those implicit in the training data) are used for any of our results. Despite this, per-frame results on video sequences in the supplementary material show almost every joint accurately predicted with remarkably little jitter.

#### 4.2. Classification accuracy

We investigate the effect of several training parameters on classification accuracy. The trends are highly correlated between the synthetic and real test sets, and the real test set appears consistently ‘easier’ than the synthetic test set, probably due to the less varied poses present.

**Number of training images.** In Figure 8(a), we show how test accuracy increases approximately logarithmically with the number of randomly generated training images, though starts to tail off around 100,000 images. As shown below, this saturation is likely due to the limited model capacity of a 3 tree, 20 deep decision forest.

**Silhouette images.** We also show in Figure 8(a) the quality of our approach on synthetic silhouette images, where the features in Eq. (4) are either given scale (as the mean depth) or not (a fixed constant depth). For the corresponding joint prediction using a 2D metric with a 10 pixel true positive threshold,

**Figure 7. Example inferences. Synthetic (top row), real (middle), and failure modes (bottom). Left column: ground truth for a neutral pose as a reference. In each example, we see the depth image, the inferred most likely body part labels, and the joint proposals shown as front, right, and top views (overlaid on a depth point cloud). Only the most confident proposal for each joint above a fixed, shared threshold is shown.**



**Figure 8. Training parameters versus classification accuracy. (a) Number of training images. (b) Depth of trees. (c) Maximum probe offset.**



we got 0.539mAP with scale and 0.465mAP without. While clearly a harder task due to depth ambiguities, these results suggest the applicability of our approach to other imaging modalities.

**Depth of trees.** Figure 8(b) shows how the depth of trees affects test accuracy using either 15k or 900k images. Of all the training parameters, depth appears to have the most significant effect as it directly impacts the model capacity of the classifier. Using only 15k images, we observe overfitting beginning around depth 17, but the enlarged 900k training set avoids this. The high accuracy gradient at depth 20 suggests that even better results can be achieved by training still deeper trees, at a small extra run-time computational cost and a large extra memory penalty. Of practical interest is that, until about depth 10, the training set size matters little, suggesting an efficient training strategy.

**Maximum probe offset.** The range of depth probe offsets allowed during training has a large effect on accuracy. We show this in Figure 8(c) for 5k training images, where ‘maximum probe offset’ means the max. absolute value proposed for both  $x$  and  $y$  coordinates of  $\mathbf{u}$  and  $\mathbf{v}$  in Eq. (4). The concentric boxes on the right show the five tested maximum offsets calibrated for a left shoulder pixel in that image; the largest offset covers almost all the body. (Recall that this maximum offset scales with world depth of the pixel.) As the maximum probe offset is increased, the classifier is able to use more spatial context to make its decisions, though without enough data it would eventually risk overfitting to this context. Accuracy increases with the maximum probe offset, though levels off around 129 pixel meters.

### 4.3. Joint prediction accuracy

In Figure 9, we show average precision results on the synthetic test set, achieving 0.731 mAP for the mean-shift clustering algorithm and 0.677mAP using the fast clustering algorithm. Combined with body part classification, the fast clustering runs in under 5 ms on the Xbox GPU, while mean shift takes 20 ms on a modern 8 core desktop CPU.

In order to get an idea of the maximum achievable mAP, we compare the mean shift algorithm to an idealized setup

that is given the *ground truth* body part labels. On the real test set, we have ground truth labels for head, shoulders, elbows, and hands. An mAP of 0.984 is achieved on those parts given the ground truth body part labels, while 0.914mAP is achieved using the inferred body parts. As expected, these numbers are considerably higher on this easier test set. While we do pay a small penalty for using our intermediate body parts representation, for many joints the inferred results are both highly accurate and close to this upper bound.

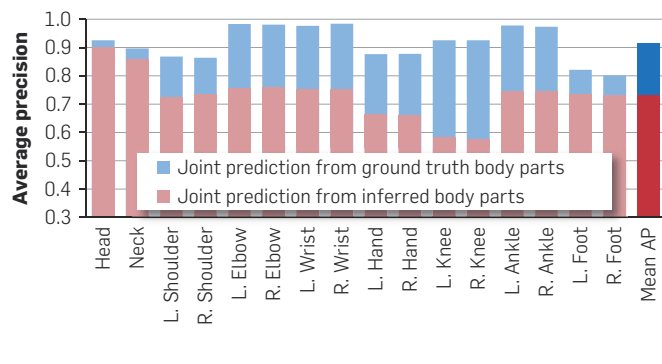
**Comparison with nearest neighbor.** To highlight the need to treat pose recognition in *parts*, and to calibrate the difficulty of our test set for the reader, we compare with two variants of exact nearest-neighbor whole-body matching in Figure 10(a). The first, idealized, variant matches the ground truth *test skeleton* to a set of training exemplar skeletons with optimal rigid translational alignment in 3D world space. Of course, in practice one has no access to the test skeleton. As an example of a realizable system, the second variant uses chamfer matching<sup>9</sup> to compare the test image to the training exemplars. This is computed using depth edges and 12 orientation bins. To make the chamfer task easier, we throw out any cropped training or test images. We aligned images using the 3D center of mass and found that further local rigid translation only reduced accuracy.

Our algorithm, recognizing in parts, generalizes better than even the idealized skeleton matching until about 150k training images are reached. As noted above, our results may get even better with deeper trees, but already we robustly infer 3D body joint positions and cope naturally with cropping and translation. The speed of nearest-neighbor chamfer matching is also considerably slower (2 fps) than our algorithm. While hierarchical matching<sup>9</sup> is faster, one would still need a massive exemplar set to achieve comparable accuracy.

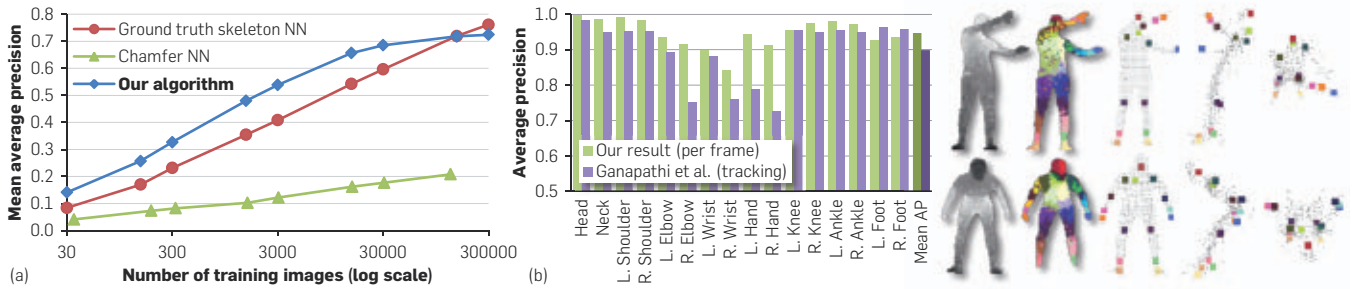
**Comparison with Ganapathi et al.**<sup>8</sup> Ganapathi et al. provided their test data and results for direct comparison. Their algorithm uses body part proposals from Plagemann et al.<sup>16</sup> and further tracks the skeleton with kinematic and temporal information. Their data comes from a time-of-flight depth camera with rather different noise characteristics to our structured light sensor. Without any changes to our training data or algorithm, Figure 10(b) shows considerably improved joint prediction average precision. Our algorithm also runs at least 10× faster.

**Full rotations and multiple people.** To evaluate the full 360° rotation scenario, we trained a forest on 900k images containing full rotations and tested on 5k synthetic full rotation images (with held out poses). Despite the massive increase in left–right ambiguity, our system was still able to achieve an mAP of 0.655, indicating that our classifier can accurately learn the subtle visual cues that distinguish front and back facing poses. Residual left–right uncertainty after classification can naturally be propagated to a tracking algorithm through multiple hypotheses. Our approach can propose joint positions for multiple people in the image, since the per-pixel classifier generalizes well even without explicit training for this scenario.

**Figure 9. Joint prediction accuracy.** We compare the actual performance of our system (red) with the best achievable result (blue), given the ground truth body part labels.



**Figure 10. Comparisons. (a) Comparison with nearest neighbor matching. (b) Comparison with Ganapathi et al.<sup>9</sup> Even without the kinematic and temporal constraints exploited by Ganapathi et al.,<sup>9</sup> our algorithm is able to more accurately localize body joints.**



**5. DISCUSSION**

We have seen how accurate proposals for the 3D locations of body joints can be estimated in super real-time from single depth images. We introduced body part recognition as an intermediate representation for human pose estimation. Use of a highly varied synthetic training set allowed us to train very deep decision forests using simple depth-invariant features without overfitting, learning invariance to both pose and shape. Detecting modes in a density function gives the final set of confidence-weighted 3D joint proposals. Our results show high correlation between real and synthetic data, and between the intermediate classification and the final joint proposal accuracy. We have highlighted the importance of breaking the whole skeleton into parts, and show state-of-the-art accuracy on a competitive test set.

As future work, we plan further study of the variability in the source mocap data, the properties of the generative model underlying the synthesis pipeline, and the particular part definitions. Whether a similarly efficient approach can directly regress joint positions is also an open question. Perhaps a global estimate of latent variables such as coarse person orientation could be used to condition the body part inference and remove ambiguities in local pose estimates.

**Acknowledgments**

We thank the many skilled engineers in Xbox, particularly Robert Craig, Matt Bronder, Craig Peeper, Momin Al-Ghosien, and Ryan Geiss, who built the Kinect tracking system on top of this research. We also thank John Winn, Duncan Robertson, Antonio Criminisi, Shahram Izadi, Ollie Williams, and Mihai Budiu for help and valuable discussions, and Varun Ganapathi and Christian Plagemann for providing their test data.

**References**

- Agarwal, A., Triggs, B. 3D human pose from silhouettes by relevance vector regression. In *Proceedings of CVPR* (2004).
- Amit, Y., Geman, D. Shape quantization and recognition with randomized trees. *Neural Computation*, 9, 7 (1997), 1545–1588.
- Belongie, S., Malik, J., Puzicha, J. Shape matching and object recognition using shape contexts. *IEEE Trans. PAMI* 24, 4 (2002), 509–522.
- Breiman, L. Random forests. *Mach. Learn.* 45, 1 (2001), 5–32.
- CMU Mocap Database. <http://mocap.cs.cmu.edu>.
- Comaniciu, D., Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. PAMI* 24, 5 (2002).
- Fergus, R., Perona, P., Zisserman, A. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of CVPR* (2003).
- Ganapathi, V., Plagemann, C., Koller, D., Thrun, S. Real time motion

capture using a single time-of-flight camera. In *Proceedings of CVPR* (2010).

- Gavrila, D. Pedestrian detection from a moving vehicle. In *Proceedings of ECCV* (June 2000).
- Gonzalez, T. Clustering to minimize the maximum intercluster distance. *Theor. Comp. Sci.* 38 (1985).
- Lepetit, V., Lagger, P., Fua, P. Randomized trees for real-time keypoint recognition. In *Proceedings of CVPR* (2005).
- Moeslund, T., Hilton, A., Krüger, V. A survey of advances in vision-based human motion capture and analysis. *CVIU* 104(2–3) (2006), 90–126.
- Navaratnam, R., Fitzgibbon, A.W., Cipolla, R. The joint manifold model for semi-supervised multi-valued regression. In *Proceedings of ICCV* (2007).
- Ning, H., Xu, W., Gong, Y., Huang, T.S. Discriminative learning of visual words for 3D human pose estimation. In *Proceedings of CVPR* (2008).
- Okada, R., Soatto, S. Relevant feature selection for human pose estimation and localization in cluttered images. In *Proceedings of ECCV* (2008).
- Plagemann, C., Ganapathi, V., Koller, D., Thrun, S. Real-time identification and localization of body parts from depth images. In *Proceedings of ICRA* (2010).
- Poppe, R. Vision-based human motion analysis: An overview. *CVIU* 108(1–2) (2007), 4–18.
- Ramanan, D., Forsyth, D. Finding and tracking people from the bottom up. In *Proceedings of CVPR* (2003).
- Shakhnarovich, G., Viola, P., Darrell, T. Fast pose estimation with parameter sensitive hashing. In *Proceedings of ICCV* (2003).
- Sharp, T. Implementing decision trees and forests on a GPU. In *Proceedings of ECCV* (2008).
- Shotton, J., Winn, J., Rother, C., Criminisi, A. *TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation*. In *Proceedings of ECCV* (2006).
- Siddiqui, M., Medioni, G. Human pose estimation from a single view point, real-time range sensor. In *IEEE International Workshop on Computer Vision for Computer Games* (2010).
- Sidenbladh, H., Black, M., Sigal, L. Sidenbladh, H., Black, M., Sigal, L. Implicit probabilistic models of human motion for synthesis and tracking. In *Proceedings of ECCV* (2002).
- Sigal, L., Bhatia, S., Roth, S., Black, M., Isard, M. Tracking loose-limbed people. In *Proceedings of CVPR* (2004).
- Urtasun, R., Darrell, T. Local probabilistic regression for activity-independent human pose inference. In *Proceedings of CVPR* (2008).
- Wang, R., Popović, J. Real-time hand-tracking with a color glove. In *Proceedings of ACM SIGGRAPH* (2009).
- Winn, J., Shotton, J. The layout consistent random field for recognizing and segmenting partially occluded objects. In *Proceedings of CVPR* (2006).
- Zhu, Y., Fujimura, K. Constrained optimization for human pose estimation from depth sequences. In *Proceedings of ACCV* (2007).

**Jamie Shotton, Toby Sharp, Andrew Fitzgibbon, Andrew Blake, and Mat Cook** ([jamiesho, tsharp, awf, ablake, and a-macook]@microsoft.com), Microsoft Research, Cambridge, UK.

**Alex Kipman and Mark Finocchio** ([akipman and markfi]@microsoft.com), Xbox Incubation.

**Richard Moore** ST-Ericsson.

## **American University** Assistant/Associate Professor

The Department of Computer Science in the College of Arts and Sciences at American University invites applications for two full-time, tenure-track positions in the broadly defined area of gaming. One position will be filled at the assistant professor level, the other at the rank of assistant or associate professor. Areas of the candidates' research expertise may include, but are not limited to: game programming, game engine development, artificial intelligence, graphics, multi-player online environments, human-computer interactions, or mobile game development. Experience with video game research involving educational, social or political issues is desirable as these faculty positions are part a long-term initiative at American University focused on interactive gaming strategies for more effective means of public engagement.

Responsibilities will include: teaching courses across the computer science curriculum from introductory courses to upper-level major electives; establishing an externally-funded research program that can involve student research participation; and service to the department and University. The Department of Computer Science offers BS degrees in computer science and computational sciences, as well as minors in computer science and computational mathematics. For more information about our program, see [www.american.edu/cas/cs/index.cfm](http://www.american.edu/cas/cs/index.cfm). To be considered, applicants must have a PhD in computer science or related discipline, post-doctoral and/or industry experience preferred. Applicants must submit a letter of application, curriculum vitae, teaching statement, research statement, and three letters of recommendation. Materials can be submitted online (highly preferred) at <http://academicjobsonline.org/ajo>, via email to [CS@american.edu](mailto:CS@american.edu), or in hard copy to:

Search Committee  
Department of Computer Science  
American University  
Washington, DC 20016-8050

The search committee will begin reviewing applications on January 7, 2013. American University is an EEO/AA institution, committed to a diverse faculty, staff, and student body. Women and minority candidates are strongly encouraged to apply.

## **Baylor University** Lecturer of Computer Science

The Department of Computer Science seeks a dedicated teacher and program advocate for a lecturer position beginning August, 2013. The ideal candidate will have a master's degree or Ph.D. in Computer Science or a related area, a commitment to undergraduate education, effective

communication and organization skills, and industry/academic experience in game development, especially with graphics and/or engine development. For position details and application information please visit: <http://www.baylor.edu/hr/index.php?id=81302>

Baylor, the world's largest Baptist university, holds a Carnegie classification as a "high-research" institution. Baylor's mission is to educate men and women for worldwide leadership and service by integrating academic excellence and Christian commitment within a caring community. Baylor is actively recruiting new faculty with a strong commitment to the classroom and an equally strong commitment to discovering new knowledge as Baylor aspires to become a top tier research university while reaffirming and deepening its distinctive Christian mission as described in Pro Futuris (<http://www.baylor.edu/profuturis/>).

*Baylor is a Baptist university affiliated with the Baptist General Convention of Texas. As an AA/EEO employer, Baylor encourages minorities, women, veterans, and persons with disabilities to apply.*

## **Baylor University** Assistant, Associate or Full Professor of Computer Science

The Department of Computer Science seeks a productive scholar and dedicated teacher for a tenured or tenure-track position beginning August, 2013. The ideal candidate will hold a terminal degree in Computer Science or a closely related field and demonstrate scholarly capability and an established and active independent research agenda in one of several core areas of interest, including, but not limited to, game design and development, software engineering, computational biology, machine learning and large-scale data mining. A successful candidate will also exhibit a passion for teaching and mentoring at the graduate and undergraduate level. For position details and application information please visit: <http://www.baylor.edu/hr/index.php?id=81302>

Baylor, the world's largest Baptist university, holds a Carnegie classification as a "high-research" institution. Baylor's mission is to educate men and women for worldwide leadership and service by integrating academic excellence and Christian commitment within a caring community. Baylor is actively recruiting new faculty with a strong commitment to the classroom and an equally strong commitment to discovering new knowledge as Baylor aspires to become a top tier research university while reaffirming and deepening its distinctive Christian mission as described in Pro Futuris (<http://www.baylor.edu/profuturis/>).

*Baylor is a Baptist university affiliated with the Baptist General Convention of Texas. As an AA/EEO employer, Baylor encourages minorities, women, veterans, and persons with disabilities to apply.*

## **Beihang University** Faculty Position Opening

International Research Center on Big Data (RCBD), Beihang University invites applications for a tenure-track professorate any rank with research interests in large-scale data analysis aka "big data".

Candidates irrelevant areas such as data management, distributed computing, machine learning, visualization and software science are especially encouraged to apply. Exceptional candidates in all areas will be considered. Candidates are expected to have a PhD degree and demonstrated ability in creating novel techniques for building large-scale data service systems and mining large-scale text, social media, and scientific data. Five-year academic experience is highly desirable. Our salary level is competitive and commensurate with qualifications and experience. The position is open until filled.

### Contact

Department of Human Resources, Beihang University  
No. 37, Xueyuan Road, Beijing, 100191, P.R. China  
tel: +86-10-82317779, 82316107;  
fax: +86-10-82317777  
e-mail: [rsrcrb@buaa.edu.cn](mailto:rsrcrb@buaa.edu.cn);  
web: <http://rsc.buaa.edu.cn/>

## **California State University, Fullerton** Department of Computer Science Assistant Professor

The Department of Computer Science invites applications for a tenure-track position at the Assistant Professor level starting fall 2013. For a complete description of the department, the position, desired specialization and other qualifications, please visit <http://diversity.fullerton.edu/>.

## **Carnegie Mellon University in Rwanda** Assistant Professor

Carnegie Mellon University's College of Engineering ([www.cit.cmu.edu](http://www.cit.cmu.edu)) has opened an exciting new location in East Africa, focusing on the development and applications of information and communication technology (ICT). This regional Centre of Excellence (CoE) in Kigali, Rwanda, offers first-class graduate education in a region of the world booming with opportunities for technology innovation. Striving to become the technology hub for East Africa, Rwanda is investing heavily in infrastructure and capacity building in the critical areas of ICT and engineering. With a history of excellence in higher education, Carnegie Mellon in Rwanda (CMU-Rwanda) is addressing the needs of Rwanda's and the region for highly skilled professionals by offering degree programs, the Master of Science Information

Technology (MSIT) and, beginning in 2015, the Master of Science in Electrical and Computer Engineering (MSECE). The first class of CMU-Rwanda MSIT students enrolled in August 2012.

Carnegie Mellon's College of Engineering is consistently ranked amongst the top ten in the USA and the world. The College includes seven academic departments granting BS, MS, and PhD degrees, two graduate-only degree granting departments (Information Networking Institute and Carnegie Mellon Silicon Valley Campus), several multidisciplinary research centers, and two Institutes (Carnegie Mellon CyLab, and Institute for Complex Engineered Systems).

Successful applicants for faculty positions with CMU-Rwanda will be expected to spend two academic semesters at Carnegie Mellon, Pittsburgh or Carnegie Mellon University Silicon Valley before assuming positions in Kigali.

Carnegie Mellon is seeking exceptional candidates who want to contribute to innovative, interdisciplinary graduate programs at CMU-Rwanda in the areas of networking, communications, cyber security and privacy, software engineering, mobile technology, cloud computing, energy systems, image and signal processing, embedded systems, entrepreneurship, and innovation and technology management.

Candidates must possess a PhD in CS, ECE or a related discipline, and an outstanding record in research, teaching and leadership. Applications should include a comprehensive resume, including a complete list of publications, a list of 3-5 professional references, statements of research and teaching interests (less than 2 pages each),

and copies of 2 research papers (journal or conference papers).

Applications should be submitted via email to the information provided below.

Contact Person: Professor Bruce Krogh,

Director of CMU-R

Apply URL: <http://www.cmu.edu/rwanda/jobs/faculty.html>

Email Address: [rwanda.coe@cit.cmu.edu](mailto:rwanda.coe@cit.cmu.edu)

---

### **Central Michigan University** **Department of Computer Science** **Three positions**

The Department of Computer Science at Central Michigan University wishes to fill three positions in the area of information technology: Mobile Development, Security and Informatics. To apply and receive more information please go to <http://www.jobs.cmich.edu>

---

### **Columbia University** **Department of Computer Science** **Faculty Positions**

The Department of Computer Science at Columbia University in New York City invites applications for tenured or tenure-track faculty positions. One or more appointments at the Assistant Professor, Associate Professor and Full Professor levels will be considered.

The successful candidate should contribute to the advancement of the Department by

developing an externally funded research program, being a thought leader in the profession, contributing to the undergraduate and graduate educational mission of the Department and providing active service to professional societies. The Department is especially interested in qualified candidates who can contribute, through their research, teaching, and/or service, to the diversity and excellence of the academic community.

Candidates are sought in all areas of Computer Science. Candidates are also sought whose research overlaps with the newly formed Institute for Data Sciences and Engineering at Columbia: <http://idse.columbia.edu>. Candidates must have a Ph.D. or its professional equivalent by the starting date of the appointment. Candidates at the Assistant Professor and Associate Professor levels without tenure must have the potential to do pioneering research and to teach effectively. Candidates at the tenured level (Associate or Full Professor) must have a demonstrated record of outstanding research accomplishments, excellent teaching credentials and established leadership in the field.

Candidates should apply online at <http://academicjobs.columbia.edu/applicants/Central?quickFind=56990>

and should submit electronically the following: curriculum vitae including a publication list, a statement of research interests, a statement of teaching interests, contact information for three people who can provide letters of recommendation, and up to three pre/reprints of scholarly work. The search will close no sooner than 12/31/2012, and will remain open until filled.



## **Yale School of Engineering & Applied Science** **Department of Electrical Engineering**

### **Junior Search in Communications and Networking at Yale University**

Yale University's Electrical Engineering Department invites applications from qualified individuals for a tenure-track, non-tenured faculty position in the area of communications and networking. Subfields of interest include wireless communications, networking, signal processing, network optimization, network economics, machine learning, and network science. All candidates should be strongly committed to both teaching and research and should be open to collaborative research. Candidates should have distinguished records of research accomplishments and should be willing and able to participate in shaping Yale's expanding program in electrical engineering. Yale University is an Affirmative Action/Equal Opportunity Employer. Yale values diversity among its students, staff, and faculty and strongly welcomes applications from women and under represented minorities. The review process will begin November 15, 2012. Applicants should include a CV, a research statement, a teaching statement and submit to <http://academicjobsonline.org/>.

### **Senior Search in Communications and Networking at Yale University**

Yale University's Electrical Engineering Department invites applications from qualified individuals for a tenured faculty position in the area of communications and networking. Subfields of interest include wireless communications, networking, signal processing, network optimization, network economics, machine learning, and network science. All candidates should be strongly committed to both teaching and research and should be open to collaborative research. Candidates should have distinguished records of research accomplishments and should be willing and able to take the lead in shaping Yale's expanding program in electrical engineering. Yale University is an Affirmative Action/Equal Opportunity Employer. Yale values diversity among its students, staff, and faculty and strongly welcomes applications from women and under represented minorities. The review process will begin November 15, 2012. Applicants should include a CV, a research statement, a teaching statement and submit to <http://academicjobsonline.org/>.

### **Senior Position in Computer Engineering at Yale University**

Yale University's Electrical Engineering Department invites applications from qualified individuals for a tenured faculty position in computer engineering. Subfields of interest include systems on a chip, embedded systems, VLSI, design automation, energy-efficient computing, low-power circuits, verification, networked systems, mobile computing, sensor networks, and biodevices. All candidates should be strongly committed to both teaching and research and should be open to collaborative research. Candidates should have distinguished records of research accomplishments and should be willing and able to take the lead in shaping Yale's expanding program in computer engineering. Yale University is an Affirmative Action/Equal Opportunity Employer. Yale values diversity among its students, staff, and faculty and strongly welcomes applications from women and under represented minorities. The review process will begin on November 15, 2012. Applicants should include a CV, a research statement, a teaching statement and submit to <http://academicjobsonline.org/>.



Candidates can consult <http://www.cs.columbia.edu> for more information about the Department.

Columbia is an affirmative action/equal opportunity employer with a strong commitment to the quality of faculty life.

---

### Connecticut College Bioinformatics Postdoctoral / Visiting Faculty

Connecticut College is seeking candidates with research interests in the use of data mining / machine learning for analyzing biological data. See details at [cs.conncoll.edu/bioinformatics.htm](http://cs.conncoll.edu/bioinformatics.htm) for this faculty position to begin in August.

---

### Connecticut College Tenure-Track CS Professor

Connecticut College is seeking candidates with research interests in visual aspects of computing, including graphics, visualization, human/computer interaction, virtual/augmented reality, animation, and other fields related to visual media. See details at [cs.conncoll.edu/tenure-track-position.htm](http://cs.conncoll.edu/tenure-track-position.htm) for this tenure-track position to begin in August.

---

### Eastern Washington University Assistant Professor – Computer Science

Assistant Professor – Computer Science - Eastern Washington University – September 2013.

Responsibilities include teaching a broad range of computer science courses such as object-oriented programming in Java, data structures, software engineering, finite state automata and concurrent programming. Other activities include: curriculum development, student advising, scholarly activities and service. Requires a Ph.D. in Computer Science or related, and potential for excellence in teaching and scholarship. Preference may be given to candidates with abilities in Computer Game Development, Mobile Application Development, or Parallel Programming. For complete information and application instructions, visit: <https://jobs.hr.ewu.edu> Eastern is an AA/EO employer.

---

### Max Planck Institute for Software Systems Tenure-track openings

Applications are invited for tenure-track and tenured faculty positions in all areas related to the study, design, and engineering of software systems. These areas include, but are not limited to, data and information management, programming systems, software verification, parallel, distributed and networked systems, and embedded systems, as well as cross-cutting areas like security, machine learning, usability, and social aspects of software systems. A doctoral degree in computer science or related areas and an outstanding research record are required. Successful candidates are expected to build a team and pursue a highly visible research agenda, both independently and

in collaboration with other groups. Senior candidates must have demonstrated leadership abilities and recognized international stature.

MPI-SWS, founded in 2005, is part of a network of eighty Max Planck Institutes, Germany's premier basic research facilities. MPIs have an established record of world-class, foundational research in the fields of medicine, biology, chemistry, physics, technology and humanities. Since 1948, MPI researchers have won 17 Nobel prizes. MPI-SWS aspires to meet the highest standards of excellence and international recognition with its research in software systems.

To this end, the institute offers a unique environment that combines the best aspects of a university department and a research laboratory:

a) Faculty receive generous base funding to build and lead a team of graduate students and post-docs. They have full academic freedom and publish their research results freely.

b) Faculty supervise doctoral theses, and have the opportunity to teach graduate and undergraduate courses.

c) Faculty are provided with outstanding technical and administrative support facilities as well as internationally competitive compensation packages.

MPI-SWS currently has 11 tenured and tenure-track faculty and 40 doctoral and post-doctoral researchers. The institute is funded to support 17 faculty and up to 100 doctoral and post-doctoral positions. Additional growth through outside funding is possible. We maintain an open, international and diverse work environment and seek applications from outstanding researchers



Florida International University is a multi-campus public research university located in Miami, a vibrant, international city. FIU is recognized as a Carnegie engaged university. Its colleges and schools offer more than 180 bachelor's, master's and doctoral programs in fields such as computer science, engineering, international relations, architecture, law, and medicine. As one of South Florida's anchor institutions, FIU is worlds ahead in its local and global engagement, finding solutions to the most challenging problems of our time. FIU emphasizes research as a major component of its mission and enrolls 48,000 students in two campus and three centers including FIU Downtown on Brickell and the Miami Beach Urban Studios. More than 160,000 alumni live and work in South Florida. For more information about FIU, visit <http://www.fiu.edu/>.

The School of Computing and Information Sciences seeks exceptionally qualified candidates for multiple tenure-track and tenured faculty positions at all levels as well as non-tenure track faculty positions at the level of Instructor.

#### Tenure track/tenured positions (Job ID # 505004)

We seek well-qualified candidates in all areas of Computer Science and researchers in the areas of programming languages, compilers, databases, information retrieval, computer architecture, scientific computing, big data, natural language processing, computational linguistics, health informatics, and robotics, are particularly encouraged to apply. Preference will be given to candidates who will enhance or complement our existing research strengths.

Ideal candidates for junior positions should have a record of exceptional research in their early careers. Candidates for senior positions must have an active and proven record of excellence in funded research, publications, and professional service, as well as a demonstrated ability to develop and lead collaborative research projects. In addition to developing or expanding a high-quality research program, all successful applicants must be committed to excellence in teaching at both graduate and undergraduate levels. An earned Ph.D. in Computer Science or related disciplines is required.

#### Non-tenure track instructor positions (Job ID# 505000)

We seek well-qualified candidates in all areas of Computer Science and Information Technology. Ideal candidates must be committed to excellence in teaching a variety of courses at the undergraduate level. A graduate degree in Computer Science or related disciplines is required; significant prior teaching and industry experience and/or a Ph.D. in Computer Science is preferred.

Florida International University (FIU), the state university of Florida in Miami, is ranked by the Carnegie Foundation as a comprehensive doctoral research university with high research activity. The School of Computing and Information Sciences (SCIS) is a rapidly growing program of excellence at the University, with 36 faculty members and over 1,500 students, including 75 Ph.D. students. SCIS offers B.S., M.S., and Ph.D. degrees in Computer Science, an M.S. degree in Telecommunications and Networking, and B.S., B.A., and M.S. degrees in Information Technology. SCIS has received approximately \$17.5M in the last four years in external research funding, has six research centers/clusters with first-class computing infrastructure and support, and enjoys broad and dynamic industry and international partnerships.

#### HOW TO APPLY:

Applications, including a letter of interest, contact information, curriculum vitae, academic transcript, and the names of at least three references, should be submitted directly to the FIU Careers website at <https://jobsearch.fiu.edu> refer to **Job ID# 505004** for tenure-track or tenured positions and to **Job ID# 505000** for instructor positions. The application review process will begin on January 7, 2013, and will continue until the position is filled. Further information can be obtained from the School website <http://www.cis.fiu.edu> or by e-mail to [recruit@cis.fiu.edu](mailto:recruit@cis.fiu.edu).

*FIU is a member of the State University System of Florida  
and is an Equal Opportunity,  
Equal Access Affirmative Action Employer.*

regardless of national origin or citizenship. The working language is English; knowledge of the German language is not required for a successful career at the institute.

The institute is located in Kaiserslautern and Saarbruecken, in the tri-border area of Germany, France and Luxembourg. The area offers a high standard of living, beautiful surroundings and easy access to major metropolitan areas in the center of Europe, as well as a stimulating, competitive and collaborative work environment. In immediate proximity are the MPI for Informatics, Saarland University, the Technical University of Kaiserslautern, the German Center for Artificial Intelligence (DFKI), and the Fraunhofer Institutes for Experimental Software Engineering and for Industrial Mathematics.

Qualified candidates should apply online at <http://www.mpi-sws.org/application>. The review of applications will begin on January 10, 2013, and applicants are strongly encouraged to apply by that date; however, applications will continue to be accepted through January 2013.

The institute is committed to increasing the representation of minorities, women and individuals with physical disabilities in Computer Science. We particularly encourage such individuals to apply.

**National Taiwan University**  
**Professor - Assistant, Professor - Associate, Professor**

The Department of Computer Science and Information Engineering has faculty openings at all

ranks beginning in August 2013. Highly qualified candidates in all areas of computer science/engineering are invited to apply. A Ph.D. or its equivalent is required. Applicants are expected to conduct outstanding research and be committed to teaching. Candidates should send curriculum vitae, three letters of reference, and supporting materials before February 28, 2013, to Prof Ai-Chun Pang, Department of Computer Science and Information Engineering, National Taiwan University, No 1, Sec 4, Roosevelt Rd., Taipei 106, Taiwan. Email Address: [faculty\\_search@csie.ntu.edu.tw](mailto:faculty_search@csie.ntu.edu.tw) Earlier submission is strongly encouraged.

**Northeastern University**  
**College of Engineering**  
**Professor and Chair**

Northeastern University College of Engineering invites applications and nominations for the position of Chair of the Electrical and Computer Engineering Department. The department is a large, successful and growing academic enterprise that includes 47 full-time faculty with over \$13 million in annual research funding, over 450 MS and Ph.D. students and more than 400 undergraduates. The department hosts several federally-funded research centers in diverse fields including sensing and imaging, energy transmission, microwave materials and information assurance. The department is committed to innovative curricular development at all levels and has a top-ranked cooperative education program.

**Qualifications:**

A doctoral degree in Electrical and Computer Engineering or a closely related field is required. Previous experience managing a federally-funded research group or center, or similar significant administrative experience, is desirable.


Applicants should submit a detailed Curriculum Vitae, three professional references and a strategic vision statement. To apply, visit: <http://apptrkr.com/298812>

For more information contact: Prof. Miriam Leeser; Email: [mel@coe.neu.edu](mailto:mel@coe.neu.edu); Phone: 617-373-3814.

**Princeton University**  
**Computer Science**  
**Postdoctoral Research Associate**

The Department of Computer Science at Princeton University is seeking applications for post-doctoral or more senior research positions in theoretical computer science. Candidates will be affiliated with the Center for Computational Intractability (CCI) or the Princeton Center for Theoretical Computer Science.

Candidates should have a PhD in Computer Science, a related field, or on track to finish by August 2013. Candidates affiliated with the CCI will have visiting privileges at partner institutions NYU, Rutgers University, and The Institute for Advanced Study. To ensure full consideration, we encourage candidates to complete their applications, (including letters of recommendation) by December 10, 2012. Applicants should submit

**MILWAUKEE SCHOOL OF ENGINEERING**

## SOFTWARE ENGINEERING FACULTY

The Milwaukee School of Engineering invites applications for a full-time faculty position in our Software Engineering program beginning in the fall of 2013. Rank will depend on qualifications and experience of the candidate. Applicants must have an earned doctorate degree in Software Engineering, Computer Engineering, Computer Science or closely related field, as well as relevant experience in software engineering practice.

The successful candidate must be able to contribute in several areas of software engineering process and practice while providing leadership in one of the following: computer security, networks, software architecture and design, or software requirements.


MSOE expects and rewards a strong primary commitment to excellence in teaching at the undergraduate level. Continued professional development is also expected.

Our ABET-accredited undergraduate software engineering program had its first graduates in Spring 2002. Founded in 1903, MSOE is a private, application-oriented university with programs in engineering, business, and nursing. MSOE's 15+ acre campus is located in downtown Milwaukee, in close proximity to the Theatre District and Lake Michigan. Please visit our website at [www.msoe.edu](http://www.msoe.edu).

**Submit all application material via email in pdf format to: [work@msoe.edu](mailto:work@msoe.edu)**

Applicants should include a letter of application, curriculum vitae, statement of teaching interests, and names (with email and physical addresses) of at least three references.

MSOE IS AN EQUAL OPPORTUNITY/AFFIRMATIVE ACTION EMPLOYER



**Tenure Track Faculty**

**CIS Department**  
**Temple University**

Applications are invited for tenure-track, open rank, faculty positions in the Department of Computer and Information Sciences at Temple University.

The junior position is in the software systems area, which includes

- Software Engineering and Applications,
- Database Systems, and
- Programming Languages.

The senior position for Associate or Full Professor is open to all areas of computer science/engineering. Applicants for the senior position are expected to have an outstanding track record. Please submit applications with all requested information online at <http://academicjobsonline.org>. For further information check <http://www.cis.temple.edu> or send email to search committee chair Dr. Eugene Kwatny at [gkwatny@temple.edu](mailto:gkwatny@temple.edu). Review of candidates will begin on January 2, 2013 and will continue until the positions are filled. Temple University is an equal opportunity, equal access, affirmative action employer.

a CV and research statement, and contact information for three references. Princeton University is an equal opportunity employer and complies with applicable EEO and affirmative action regulations. Apply to: <http://jobs.princeton.edu/req-uisition#1200777>

Apply URL: [https://jobs.princeton.edu/applicants/jsp/shared/position/JobDetails\\_css.jsp?postingId=192005](https://jobs.princeton.edu/applicants/jsp/shared/position/JobDetails_css.jsp?postingId=192005)

**Southern Methodist University Position #50049**  
**Department of Computer Science and Engineering**  
**Faculty Position in Computer Science and Engineering**

The Department of Computer Science and Engineering in the Lyle School of Engineering at Southern Methodist University invites applications for a faculty position in computer science and engineering beginning Fall 2013. Individuals with experience and research interests in all areas of computer science and engineering are encouraged to apply. Priority will be given to individuals with expertise and research interest in *data mining, informatics, computer systems and networking*, and related areas. The search is focused at the tenure-track assistant professor level. The successful candidates must have or expect to have a Ph.D. in computer science, computer engineering, or a closely related area by date of hire. Successful applicants will demonstrate a deep commitment to research activity in computer science

and engineering and a strong record of excellence in teaching.

The Dallas/Fort Worth area, one of the top three high-tech industrial centers in the country, has the largest concentration of telecommunications corporations in the US, providing abundant opportunities for industrial research cooperation and consulting. Dallas/Fort Worth is a multifaceted business and high-tech community, offering exceptional museums, diverse cultural attractions, and a vibrant economy.

The CSE Department resides within the Bobby B. Lyle School of Engineering and offers BS, MS, and Ph.D. degrees in Computer Engineering and Computer Science, the Doctor of Engineering in software engineering, and the MS in Security Engineering and Software Engineering. The department currently has 15 faculty members with research concentrations in security engineering, software engineering, computer networks, telecommunications, data mining, database systems, VLSI and digital systems, and computer arithmetic. Additional information may be found at: [www.lyle.smu.edu/cse](http://www.lyle.smu.edu/cse).

To receive full consideration, interested individuals should send a complete resume and names of three references, including a one-page statement of research interests and accomplishments by December 21, 2012 to:

[csesearch@lyle.smu.edu](mailto:csesearch@lyle.smu.edu)

or

CSE Faculty Search Position #50049  
Department of Computer Science  
and Engineering

SMU  
Dallas, TX 75275-0122

Review of applicants will begin immediately and will continue until the positions are filled. Hiring is contingent upon the satisfactory completion of a background check.

*SMU will not discriminate on the basis of race, color, religion, national origin, sex, age, disability, or veteran status. SMU is committed to nondiscrimination on the basis of sexual orientation.*

**State University of New York at Binghamton**  
**Department of Computer Science**  
**Two Tenure-track Assistant Professor positions**

Applications are invited for two tenure-track Assistant Professor positions beginning Fall 2013. Preferred specializations include embedded systems, energy-aware computing and security. The Department has about 800 majors, including 63 full-time PhD students. Junior faculty have a significantly reduced teaching load for at least the first three years. A new NSF supported industry-university collaborative research center on energy-smart electronic systems offers an added venue for research and funding. Apply online at: <http://binghamton.interviewexchange.com>

First consideration given to applications received by **January 31, 2013**.

We are an EE/AA employer.

**SYRACUSE UNIVERSITY**  
**Department of Electrical Engineering and Computer Science (EECS)**  
**Associate or Full Professor**

The EECS department invites applications for a tenure track/tenured Associate or Full Professor in the discipline of Computer Science for a position partially funded by JPMorganChase Corporation. The ideal candidate would be an accomplished researcher in Cybersecurity. Particular focus is in the area of mobile system security, requiring an understanding of the latest technological developments in mobile computing and mobile communications, the rapidly increasing penetration of world markets by mobile platforms, and the attempts by the financial industry to reach consumers through mobile platforms.

A doctorate in Computer Science or a closely related field is required at the time of employment. Start-up funds commensurate with the needs of the individual will be provided for the successful candidate.

Review of applications will begin immediately and continue until the position is filled. The starting date is expected to be in August 2013. For more information or to apply, please visit the following web site: <http://www.sujobops.com> (Job# 070191). For detailed information about the Department of EECS, please see the following web site: <http://www.eecs.syr.edu>.

The University and surrounding areas offer a vibrant intellectual and cultural atmosphere, a diverse ethnic community, great public education systems, affordable homes, and many other assets that make it a great place to live and work. Syracuse University is strongly committed to gender and ethnic diversity. Women and members of minorities are especially encouraged to apply.

Syracuse University is an Affirmative Action/Equal Opportunity Employer.



THE UNIVERSITY of  
NEW MEXICO

**CHAIR**  
**Dept. of Electrical & Computer Engineering**

The School of Engineering at the University of New Mexico invites applications for the position of Department Chair of the Electrical and Computer Engineering Department. The position will also include a tenured appointment as Professor of Electrical and Computer Engineering. Candidates must have earned a Ph.D. in Electrical or Computer Engineering or a closely related field. Experience commensurate with that of a full Professor or equivalent industry experience is required. Salary will be commensurate with qualifications and experience.

Preferred qualifications include demonstrated engineering leadership, a research record of marked distinction and significant impact, a strong commitment to undergraduate and graduate education, prior administrative experience, proven teaching skills, exceptional communication and interpersonal skills, and the ability to articulate a vision for furthering the national stature of the Electrical and Computer Engineering Department.

For best consideration, complete applications must be received by **January 21, 2013**. The position will remain open until filled. Each application must include a cover letter summarizing the applicant's perspectives on research, teaching and leadership, a detailed CV, and the names, mailing addresses, e-mails, and telephone numbers of three references (prior consent will be sought before contacting references). For complete job posting and how to apply, go to [www.ece.unm.edu](http://www.ece.unm.edu)

Applications should be submitted online through <https://unmjobs.unm.edu>, by referencing posting #0818004

Inquiries should be sent to: [ChairSearch@ece.unm.edu](mailto:ChairSearch@ece.unm.edu)

The University of New Mexico is an equal opportunity/affirmative action employer and educator. We especially encourage members of underrepresented groups to apply.

### Swarthmore College Tenure Track Assistant Professor

Swarthmore College has a strong institutional commitment to excellence through diversity in its educational program and employment practices and actively seeks and welcomes applications from candidates with exceptional qualifications, particularly those with demonstrable commitments to a more inclusive society and world.

**Applications are invited for a tenure track position at the assistant professor level beginning Fall semester 2013.** Swarthmore College is a small, selective, liberal arts college located 10 miles outside of Philadelphia. The Computer Science Department offers majors and minors at the undergraduate level. Applicants must have teaching experience and should be comfortable teaching a wide range of courses at the introductory and intermediate level. Candidates should additionally have a strong commitment to involving undergraduates in their research. A Ph.D. in CS by or near the time of appointment is required. We are particularly interested in applicants that add breadth to our department, including the areas of databases, networking, security, theory, compilers, and programming languages. Strong applicants in other areas will also be considered.

Priority will be given to applications received by December 15, but will be accepted until the position is filled. Applications should include a vita, teaching statement, research statement, and three letters of reference, at least two that speak to the candidate's teaching ability.

#### Apply for this Job:

Richard Wicentowski  
Email: [jobs2013@cs.swarthmore.edu](mailto:jobs2013@cs.swarthmore.edu)  
Fax: 610-328-8606  
Phone: 610-328-8272  
Apply URL: <http://goo.gl/LPYF2>

### Technological Institute of the Philippines

Cubao, Quezon City / Quiapo, Manila, Philippines

#### LOOKING FOR AN INTERESTING AND FULFILLING TEACHING CAREER?

Visit: [www.tip.edu.ph](http://www.tip.edu.ph)

\* Awards from the Commission on Higher Education, highest governing body for Higher Education Institutions

- ▶ AUTONOMOUS status for TIP Quezon City
- ▶ DEREGULATED status for TIP Manila
- ▶ Center of Development for Civil Engineering (TIP QC)
- ▶ Center of Development for Computer Engineering (TIP QC/Manila)
- ▶ Center of Development for Information Technology Education (TIP QC/Manila)

\* FAAP-PACUOCA

Accredited Programs

\* QUALITY MANAGEMENT SYSTEM CERTIFIED TO ISO 9001:2008

WE ARE LOOKING FOR FULL TIME FACULTY MEMBERS WITH  
PH.D IN COMPUTER SCIENCE

PH.D IN INFORMATION SYSTEMS  
PH.D IN INFORMATION TECHNOLOGY

Who can assist in building world-class Computer Science, Information Systems, and Information Technology programs that will address and manage the rapid developments in Computer Science, Information Systems, and Information Technology

Very competitive compensation package and fringe benefits await the qualified candidates

Send resume to [amelita.dolom@tip.edu.ph](mailto:amelita.dolom@tip.edu.ph) or [dolomillete@yahoo.com.ph](mailto:dolomillete@yahoo.com.ph)

### University College London (UCL)

Department of Computer Science  
Faculty Positions

The Department of Computer Science at University College London (UCL) invites applications for faculty positions in the areas of Networking, Systems, and Programming Languages. We seek world-class talent; candidates must have an outstanding research track record.

In Networking and Systems, our interests include operating systems, systems security, distributed systems, networking, and their intersection, with an emphasis on experimental system-building. Appointments in Networking and Systems will be made at the rank of Lecturer, Senior Lecturer, or Reader (equivalent to Assistant Professor, junior Associate

Professor, and senior Associate Professor, respectively, in the US system), commensurate with qualifications.



Cornell University

**The SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING AT CORNELL UNIVERSITY** in Ithaca, New York, invites applications for **tenure-track Faculty positions in electrical and computer engineering**. We are particularly interested in outstanding candidates that can make an impact on the following

two areas; however, candidates from all areas will receive full consideration.

**Advanced Nanoelectronic Devices** — We are looking for excellent candidates who will push the limits of the many avenues of nanodiscovery, and develop innovative post-CMOS devices/circuits from them. New approaches could involve electrons, photons, spins, quantum states, material phase transitions, etc. Such individuals will be able to leverage Cornell's strength in solid state physics, materials, and nanosciences, as well as our world-class Nanoscale Science & Technology Facility (CNF).

**Unconventional Imaging** — We are looking for excellent candidates with expertise that spans the broad areas of information processing, physical sensing, and computation. Ideal candidates would have experience developing new parallel sensing technologies such as 3-D sensing and imaging, light-field imaging, lensless microscopy, and compressive sampling with single sensor elements, as well as expertise developing associated algorithms and information processing techniques needed to interpret the complex, data-intensive outputs of such systems.

Successful applicants must hold a doctoral degree in an appropriate field before they can be appointed in a tenure-track position. Additionally, they must have demonstrated an ability to conduct outstanding research, and show promise for excellent teaching. We anticipate filling the position at the assistant professor level, but applications at all levels will be considered. Salary and rank will be commensurate with qualifications and experience.

Applicants should submit a curriculum vitae, a research statement, copies of two representative publications, a teaching statement, and complete contact information for at least three references. Personal statements summarizing teaching experience and interests, leadership efforts and contributions to diversity are encouraged.

Applications must be made on-line at <https://academicjobsonline.org/ajo/jobs/2280>. Applications received by **January 18, 2013** will receive full consideration. Applications will be evaluated on an ongoing basis until the positions are filled.

The School of Electrical and Computer Engineering, and the College of Engineering at Cornell embrace diversity and seek candidates who will create a climate that attracts students of all races, nationalities and genders. We strongly encourage women and underrepresented minorities to apply.

Cornell University 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.

Cornell University is an affirmative action, equal opportunity educator and employer.

### Northwestern University

#### Assistant Professor in Database Systems

The Department of Electrical Engineering and Computer Science at Northwestern University invites applications for a tenure-track assistant professor position in database systems to start in fall 2013. We are interested in exceptional candidates in all areas of database systems, but have a particular focus on areas such as large-scale data management, integration of structured and unstructured data, parallel and distributed data mining and analytics, stream databases, and database engines for scalable computing and emerging computer architectures.

A Ph.D. in Computer Science or Computer Engineering is required, as is a clear track record of success in database systems. Successful candidates will be expected to carry out world class research, collaborate with other faculty, and teach effectively at the undergraduate and graduate levels. Compensation and start-up packages are negotiable and will be competitive.

Northwestern EECS consists of over 50 faculty members of international prominence whose interests span a wide range. Northwestern University is located in Evanston, Illinois on the shores of Lake Michigan just north of Chicago. Further information about the Department and the University is available at <http://www.eecs.northwestern.edu> and <http://www.northwestern.edu>.

To ensure full consideration, applications should be received by **February 15, 2013**, but applications will be accepted until the position is filled.

To apply, first read full upload instructions at <http://eecs.northwestern.edu/academic-openings.html>. Applicants will submit (1) a cover letter, (2) a curriculum vitae, (3) statements of research and teaching interests, (4) three representative publications, and (5) at least three, but no more than five references. For general questions about the search or application assistance post submission, contact [db-search@eecs.northwestern.edu](mailto:db-search@eecs.northwestern.edu).

The aforementioned application materials may also be sent to: **Database Systems Faculty Search Committee, Department of Electrical Engineering and Computer Science, Technological Institute, L359, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA.**

*Northwestern University is an equal opportunity, affirmative action employer. Qualified women and minorities are encouraged to apply. It is the policy of Northwestern University not to discriminate against any individual on the basis of race, color, religion, national origin, gender, sexual orientation, marital status, age, disability, citizenship, veteran status, or other protected group status. Hiring is contingent upon eligibility to work in the United States.*

In Programming Languages and Verification, our interests span theory and practice, including semantics, program analysis, language design, compilation, and theorem proving. Appointments in Programming Languages and Verification will be made at the rank of Lecturer or, where warranted, Senior Lecturer (equivalent to Assistant Professor, and junior

Associate Professor, respectively, in the US system), commensurate with qualifications.

Candidates must hold an earned Ph.D. in Computer Science or a closely related field by the time they begin their appointment. They will be evaluated chiefly on the significance and novelty of their research to date, and their promise for leading a group in a fruitful program of research. They must also demonstrate a zest for innovative and challenging teaching at the graduate and undergraduate levels. A proven record of ability to manage time and evidence of ability to teach and to supervise academic work by undergraduates, masters, and doctoral students are desirable. Our department is a highly collaborative environment, and we seek future colleagues who enjoy working collaboratively. Candidates should further be committed to public communication, and to UCL's policy of equal opportunity, including working harmoniously with colleagues and students of all cultures and backgrounds.

Since 1973, when UCL CS became the first ARPAnet node outside the United States, the department has been a leading centre for networking research, as demonstrated by its long-standing strong presence in the SIGCOMM conference proceedings. UCL CS also is home to world-leading research in the theory and practice of program verification, including seminal contributions in separation logic and automatic termination proving.

UCL, the third oldest university in England, and the first to admit students without regard for their race or creed, sits a few blocks from the British Museum in central London, a vibrant metropolis of 8.1 million that offers an ever-renewing panoply of cultural and culinary choices, and is a budget airfare away from all of continental Europe.

Further details about UCL CS, the posts, and how to apply may be found

at <http://www.cs.ucl.ac.uk/vacancies>

All application materials must reach UCL by the 11th of January, 2013.

We particularly welcome female applicants and those from an ethnic minority, as they are under-represented within UCL at these levels.

UCL Taking Action For Equality.

---

### University of Alabama in Huntsville Computer Science Department Assistant Professor

The Computer Science Department of the University of Alabama in Huntsville (UAHuntsville) invites applicants for a tenure-track faculty position at the Assistant Professor level. A Ph.D. in Computer Science or closely related area is required. Our areas of interest are Data Science and Cyber Security & Information Assurance. We have a strong commitment to excellence in teaching, research and service. Successful applicants should have good communication and teaching abilities,

and will be expected to develop and maintain a strong externally funded research program.

UAHuntsville is strategically located along the Tennessee River in scenic North Alabama in a rapidly expanding high technology area. Huntsville has one of the largest Research Parks in the nation and is home to NASA's Marshall Space Flight Center, the Army's Redstone Arsenal, and many high-tech industries that present many opportunities to pursue collaborative research. The surrounding population of approximately 375,000 well-educated and highly technically skilled people has access to excellent public schools and inexpensive housing. The University has an enrollment of approximately 7700 students with more than ten research centers in diverse areas including information technology, modeling & simulation, Earth Science, and space science that provide further opportunities for collaborative projects.

The Computer Science Department has 12 full-time faculty and offers B.S., M.S., and Ph.D. degrees in Computer Science and M.S. degree in Software Engineering. There are approximately 254 undergraduate CS majors, 92 MS, and 34 Ph.D. candidates enrolled in our program. Current faculty research interests include Software Engineering, Visualization, Pattern Recognition and Image Processing, Distributed Systems, Graph Theory, Data Mining, Information Technology, Artificial Intelligence, Modeling and Simulation, Multimedia Systems, Information Assurance and Networking. According to recent NSF figures the CS department ranks 30th in the nation in federal research funding.

Please submit a detailed resume with references to include names of three professional references with contact information, including name, address, phone number and email address to: **Chair, Search Committee, Computer Science Department, The University of Alabama in Huntsville, Huntsville, AL, 35899**. Qualified women and minority candidates are encouraged to apply. Initial review of applicants will begin in November 2012 and will continue until a suitable candidate is found.

The University of Alabama in Huntsville is an equal Opportunity/Affirmative Action Institution.

---

### University of California, Riverside Tenure-track faculty position

The Department of Computer Science and Engineering, University of California, Riverside invites applications for **tenure-track** faculty position beginning the 2013/2014 academic year with research interests in **Systems**. Candidates in Networks, Operating/Distributed Systems and Cyber-security are especially encouraged to apply. Exceptional candidates in all areas will be considered. Positions require a Ph.D. in Computer Science (or in a closely related field) at the time of employment. While the department is primarily seeking to hire at the Assistant Professor level, exceptional senior candidates with outstanding research, teaching and graduate student mentorship records may be considered. Junior candidates must show outstanding research, teaching and graduate student mentorship potential. Salary level will be competitive and commensurate with qualifications and experience. Details and application materials can be found at [\[ucr.edu/facultysearch\]\(http://ucr.edu/facultysearch\). Full consideration will be given to applications received by January 1, 2013. We will continue to consider applications received until the position is filled. For inquiries and questions, please contact us at \[search@cs.ucr.edu\]\(mailto:search@cs.ucr.edu\). EEO/AA employer](http://www.engr.</a></p></div><div data-bbox=)

---

### University of Colorado Assistant Professor

The Department of Computer Science at the **University of Colorado Boulder** invites applications for two full-time tenure-track positions in the areas of cyber-physical systems (CPS) and machine learning (ML). While these positions are at the Assistant Professor level, exceptional candidates at a more senior level may be considered.

These positions will complement existing efforts in cyber-physical systems and machine learning in the department and across the campus. For the CPS position, we seek candidates with active research programs that integrate physical systems with relevant analytic techniques; application areas include, without limitation, ubiquitous and embedded systems, robotics, bio-medical systems, smart vehicles and network systems. ML candidates with research programs in the areas of numerical optimization, speech and language processing, and large-scale learning systems are of particular interest. Candidates with relevant research programs involving energy systems are strongly encouraged to apply for either position.

The University of Colorado is an Equal Opportunity/Affirmative Action employer.

Application materials are accepted electronically at:

For cyber-physical systems Jobs at CU posting quick link: [www.jobsatcu.com/applicants/Central?quickFind=71252](http://www.jobsatcu.com/applicants/Central?quickFind=71252)

For machine learning Jobs at CU posting quick link: [www.jobsatcu.com/applicants/Central?quickFind=71257](http://www.jobsatcu.com/applicants/Central?quickFind=71257)

---

### University of Kentucky Computer Science Department

The University of Kentucky Computer Science Department expects to hire an Assistant Professor to begin employment in August of 2013. Candidates must have earned a PhD in Computer Science or closely related field at the time employment begins. Applications are now being accepted.

Review of credentials will begin immediately and continue until the position is filled.

The department seeks to hire energetic researcher/educators who are interested in the application of advanced computing to challenging and relevant problems. We favor researchers who can collaborate to solve problems involving multiple disciplines. All areas of computing will be considered, but database/data mining and scientific computation are a focus area.

To apply, a University of Kentucky Academic Profile must be submitted at <https://ukjobs.uky.edu/applicants/Central?quickFind=237966>.

For more detailed information about this position, go to [www.cs.uky.edu/opportunities/faculty](http://www.cs.uky.edu/opportunities/faculty).

The University of Kentucky is an equal opportunity employer and especially encourages applications from minorities and women.

### University of Michigan-Dearborn Assistant/Associate Professor

The Department of Computer and Information Science (CIS) at the University of Michigan-Dearborn invites applications for a tenure-track faculty position in software engineering. Rank and salary will be commensurate with qualifications and experience. We offer competitive salaries and start-up packages.

Qualified candidates must have, or expect to have, a Ph.D. in computer science or a closely related discipline by the time of appointment and will be expected to do scholarly and sponsored research, as well as teaching at both the undergraduate and graduate levels. Candidates at the associate professor rank should already have an established funded research program. The CIS Department offers several BS and MS degrees, and participates in several interdisciplinary degree programs, including an MS program in software engineering and a Ph.D. program in information systems engineering. The current research areas in the department include computer graphics and geometric modeling, database systems, multimedia systems and gaming, networking, computer and network security, and software engineering. These areas of research are supported by several established labs and many of these areas are currently funded.

The University of Michigan-Dearborn is located in the southeastern Michigan area and offers excellent opportunities for faculty collaboration with many industries. We are one of three campuses forming the University of Michigan system and are a comprehensive university with over 8900 students. One of the university's strategic visions is to advance the future of manufacturing in a global environment.

The University of Michigan-Dearborn is dedicated to the goal of building a culturally-diverse and pluralistic faculty committed to teaching and working in a multicultural environment, and strongly encourages applications from minorities and women.

A cover letter, curriculum vitae including e-mail address, teaching statement, research statement, and three letters of reference should be sent to:

Dr. William Grosky, Chair  
Department of Computer and Information Science  
University of Michigan-Dearborn  
4901 Evergreen Road  
Dearborn, MI 48128-1491  
Email: wgrosky@umich.edu,  
Internet: <http://www.cis.umich.edu>  
Phone: 313.583.6424, Fax: 313.593.4256

The University of Michigan-Dearborn is an equal opportunity/affirmative action employer.

### University of Nevada, Las Vegas – UNLV Tenure/Tenure Track Professor Position in Computer

The Department of Computer Science at the University of Nevada, Las Vegas invites applications for a tenure track/tenured position in the area of software engineering and/or software development commencing Fall 2013. The rank is open. Candidates applying for Associate or Full Profes-

or ranks are expected to have established publication and grant records. Candidates for Assistant Professor must show potential for research and creative activities. The preference will be given to applicants with extensive software development backgrounds, and candidates are encouraged to provide information on their roles in past or present software engineering projects. All applicants must have a Ph.D. in Computer Science from an accredited college or university. Submit a letter of interest, a detailed resume listing qualifications and experience, and the names, addresses, and telephone numbers of at least three professional references who may be contacted. Applicants should fully describe their qualifications and experience, with specific reference to each of the minimum and preferred qualifications because this is the information on which the initial review of materials will be based. Although this position will remain open until filled, review of candidates' materials will begin on February 10, 2013, and best consideration will be gained for materials submitted prior to that date. Materials should be addressed to Prof Kazem Taghva, Search Committee Chair. For a complete position description and application details, please visit <http://jobs.unlv.edu> or call 702-895-2894. EEO/AA Employer.

### University of New Mexico Computer Engineering Faculty Position

The Department of Electrical and Computer Engineering (ECE) at the University of New Mexico invites applications for one or more full-time positions at the level of Assistant Professor. An Assistant Professor hire will be a probationary appointment leading to a tenure decision. Associate or Full Professor ranks will also be considered. Candidates must have earned a Ph.D., in Electrical or Computer Engineering or a closely related field by August 1st, 2013.

The preferred expertise areas are (i) bioengineering, (ii) hardware systems for Big Data, and (iii) computer networking and communications systems with an emphasis in applications such as biomedical image analysis and visualization, reconfigurable computing, and data analytics.

For complete job posting and how to apply, go to [www.ece.unm.edu](http://www.ece.unm.edu)

The University of New Mexico is an equal opportunity/affirmative action employer and educator. We especially encourage members of under-represented groups to apply.

### University of North Florida School of Computing Assistant Professor

The School of Computing at The University of North Florida is inviting applications for a tenure-track assistant professor position beginning August 2013. Interested applicants must have a terminal degree in computing or closely related field and at least one degree in an engineering field. Research interests should include security and/or cloud computing. Applicants must complete an online application at [www.unfjobs.org](http://www.unfjobs.org) and upload a letter of interest, a curriculum vitae, a list of three references with contact information, statement of purpose detailing teaching ex-

perience and a vision for research, and unofficial transcripts. Direct questions to Dr. Robert Roggio at [broggio@unf.edu](mailto:broggio@unf.edu). Application review will begin in October, 2012.

### The University of North Texas Department of Computer Science and Engineering Assistant Professor

The Department of Computer Science and Engineering at the University of North Texas (UNT) is seeking candidates for a tenure-track faculty position at the Assistant Professor level beginning August 15, 2013. The department plans to build on its existing strengths in Computer Security, including network security and intrusion detection, secure software systems, vulnerability analysis, and machine learning techniques applied to computer security. Candidates should have demonstrated the potential to excel in research in one or more of these areas and in teaching at all levels of instruction. A Ph.D. in Computer Science, Computer Engineering or closely related field is required at the time of appointment. An Applicant's record must include high quality publications.

The Computer Science and Engineering department is home to 812 bachelor students, 140 masters students and 82 Ph.D. students. The UNT Center for Information and Computer Security, housed in the department, has been recognized by the National Security Agency as a National Center for Academic Excellence in Information Assurance Research and Education and offers several certificate programs in computer security. Additional information about the department and center are available at the websites: [www.cse.unt.edu](http://www.cse.unt.edu) and [www.cics.unt.edu](http://www.cics.unt.edu), respectively.

#### Application Procedure:

All applicants must apply online to:  
<https://facultyjobs.unt.edu/applicants/Central?quickFind=51736>.

Submit nominations and questions regarding the position to Dr. Philip Sweany ([sweany@cse.unt.edu](mailto:sweany@cse.unt.edu)).

#### Application Deadline:

The committee will begin its review of applications on December 1, 2012 and continue until the position is closed.

#### The University:

With about 36,000 students, UNT is the nation's 33rd largest university. As the largest, most comprehensive university in Dallas-Fort Worth, UNT drives the North Texas region. UNT offers 97 bachelor's, 82 master's and 35 doctoral degree programs, many nationally and internationally recognized. A student-focused public research university, UNT is the flagship of the UNT System. UNT is strategically located in Denton, Texas, a vibrant city with a lively arts and music culture, at the northern end of the Dallas-Fort Worth metroplex. The DFW area has more than six million people, with significant economic growth, numerous industrial establishments, and excellent school districts.

The University of North Texas is an AA/ADA/EOE committed to diversity in its educational programs.

## University of Northern Iowa Assistant Professor of Computer Science

The Department of Computer Science at the University of Northern Iowa invites applications for a tenure-track assistant professor position to begin August 2013. Applicants must hold a Ph.D. in Computer Science or a closely-related discipline. The department seeks candidates able to participate widely in the CS curriculum and conduct a research program involving undergraduates.

Detailed information about the position and the department are available at <http://www.cs.uni.edu/>

To apply, visit <http://jobs.uni.edu/>. Applications received by January 15, 2013, will be given full consideration. EOE/AA. Pre-employment background checks are required. UNI is a smoke-free campus.

## University of Pittsburgh School of Information Sciences Assistant Professor

The School of Information Sciences (<http://www.ischool.pitt.edu>) at the University of Pittsburgh is seeking candidates for two tenure-stream assistant professorships to start in the fall term of 2013. The primary areas of interest include:

Information Assurance (Position #06441)

- ▶ Application and system security
- ▶ Digital forensics
- ▶ Trust, security, privacy

Web Science (Position #02336)

- ▶ Data-intensive scholarship
- ▶ Information visualization
- ▶ Data mining
- ▶ Semantic web
- ▶ Web engineering

This top-ranked information school (iSchool) offers degree programs in Information Science & Technology, Library & Information Science, and Telecommunications & Networking. Candidates are sought who have research and teaching interests in alignment with the School's signature strengths.

For a complete description, please visit <http://www.ischool.pitt.edu/news/facultyopenings.php>.

Contact Person: Search Committee

Email Address: [sissearch@sis.pitt.edu](mailto:sissearch@sis.pitt.edu)

Apply URL: <http://www.ischool.pitt.edu>

Phone: 412-624-5129

Fax: 412-624-5231

The University of Pittsburgh is an Equal Opportunity, Affirmative Action employer and strongly encourages women and candidates from under-represented minorities to apply.

## University of Rochester Faculty Positions in Computer Science: HCI and Big Data

The University of Rochester Department of Computer Science seeks applicants for multiple tenure track positions in **human-computer interaction (HCI)** and **big data research** (including

machine learning and data mining, cloud computing, e-science applications, and very large databases) Candidates must have a PhD in CS or a related discipline. Applicants for the big data position will also be considered for a faculty search in that area by our Department of Electrical & Computer Engineering. Additional information and online application instructions appear at <http://www.cs.rochester.edu/dept/recruit>.

The Department of Computer Science is a research-oriented department with a distinguished history of contributions in systems, theory, artificial intelligence, and HCI. We have a collaborative culture and strong ties to cognitive science, linguistics, and ECE. Over the past decade, a third of its PhD graduates have won tenure-track faculty positions, and its alumni include leaders at major research laboratories such as Google, Microsoft, and IBM.

The University of Rochester is a private, Tier I research institution located in western New York State. The University of Rochester consistently ranks among the top 30 institutions, both public and private, in federal funding for research and development. Teaching loads are light and classes are small. Half of its undergraduates go on to post-graduate or professional education. The university includes the Eastman School of Music, a premiere music conservatory, and the University of Rochester Medical Center, a major medical school, research center, and hospital system. The greater Rochester area is home to over a million people, including 80,000 students who attend its 8 colleges and universities.

The University of Rochester has a strong commitment to diversity and actively encourages applications from candidates from groups under-represented in higher education. The University is an Equal Opportunity Employer.

## The University of Texas at San Antonio Faculty Positions in Computer Science

The Department of Computer Science at **The University of Texas at San Antonio** invites applications for one or more tenure/tenure-track positions at the **Assistant, Associate or Professor** level, starting Fall 2013. We are particularly interested in candidates in software engineering, distributed systems, and operating systems. Other related areas of computer science will also be considered.

The Department of Computer Science currently has 22 faculty members and offers B.S., M.S., and Ph.D. degrees supporting a dynamic and growing program with 634 undergraduates and more than 160 graduate students, including 81 Ph.D. students. See <http://www.cs.utsa.edu/fsearch> for **application instructions and additional information on the Department of Computer Science**.

Screening of applications will begin on January 2, 2013 and will continue until the positions are filled or the search is closed. UTSA is an EO/AA Employer.

**Chair of Faculty Search Committee**

**Department of Computer Science**

**The University of Texas at San Antonio**

**One UTSA Circle**

**San Antonio, TX 78249-0667**

**Phone: 210-458-4436**

**Fax: 210-458-4437**

## University of Toronto Department of Computer Science and Donnelly Centre for Cellular and Biomolecular Research Assistant Professor

The Department of Computer Science and the Donnelly Centre for Cellular and Biomolecular Research at the University of Toronto invite applications for a tenure-stream position in Computational Biology or Bioinformatics. The appointment is at the rank of Assistant Professor and will begin on July 1, 2013.

We seek outstanding applicants with demonstrated excellence in research at the highest level and with the potential for excellence in undergraduate and graduate teaching. Although we expect candidates to have a PhD and postdoctoral training in the computational sciences (computer science, computational biology and quantitative biology), exceptional candidates with recent or imminently-expected PhDs will be also considered.

The Department of Computer Science is an international leader in research and teaching, with recognized strength in most areas of computer science. The Donnelly Centre is an interdisciplinary research institute at the University of Toronto with the mandate to create a research environment that encourages integration of biology, computer science, engineering and chemistry, and that spans leading areas of biomedical research. The successful candidate will have the opportunity to take advantage of the University's strengths in biology and bioinformatics--and computational, medical and biological sciences more broadly--and to facilitate further interaction with other units. To facilitate such interactions, the successful candidate will hold a joint appointment in the Department of Computer Science (51%) and in the Donnelly Centre (49%).

Salaries are competitive with our North American peers and will be determined according to experience and qualifications. Toronto is a vibrant and cosmopolitan city, one of the most desirable in the world in which to work and live, and a major centre for advanced computer, medical and biological technologies with strong ties to the University.

Applicants should apply online at <http://recruit.cs.toronto.edu/>, and include curriculum vitae, a list of publications, a research and teaching statement, and the names and email addresses of at least three references. Other supporting materials may also be included. We will not accept applications submitted by post. If you have any questions regarding this position, please contact Sara Burns at [recruit@cs.toronto.edu](mailto:recruit@cs.toronto.edu).

Review of applications will begin on January 7, 2013 and continue until the position is filled. To ensure full consideration applications, should be received by February 4, 2013.

For more information on the Department of Computer Science see [www.cs.toronto.edu](http://www.cs.toronto.edu) and for the Donnelly Centre for Cellular and Biomolecular Research see [www.thedonnelycentre.utoronto.ca](http://www.thedonnelycentre.utoronto.ca)

The University of Toronto is strongly committed to diversity within its community and especially welcomes applications from visible minority group members, women, Aboriginal persons, persons with disabilities, members of sexual minority groups, and others who may contribute to the further diversification of ideas. All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority.

**University of Toronto**  
**Dept. of Computer Science**  
**Assistant Professor**

The Department of Computer Science at the University of Toronto invites applications for a tenure-stream position in the area of Machine Learning. The appointment is at the rank of Assistant Professor and will begin on July 1, 2013.

Candidates should have (or be about to receive) a Ph.D. in Computer Science or a related field. We seek outstanding applicants with demonstrated excellence in research at the highest level and with potential for excellence in undergraduate and graduate teaching.

Salaries are competitive with our North American peers and will be determined according to experience and qualifications. Toronto is a vibrant and cosmopolitan city, one of the most desirable in the world in which to work and live, and a major centre for advanced computer technologies.

The Department of Computer Science is an international leader in research and teaching, with recognized strength in most areas of Computer Science. The department also has close interdisciplinary ties to other units within the University and strong interactions with the computer industry.

Applicants should apply online at <http://recruit.cs.toronto.edu>, and include curriculum vitae, a list of publications, a research and teaching statement, and the names and email addresses of at least three references. Other supporting materials may also be included. We will not accept applications submitted by post. If you have any questions regarding this position, please contact Sara Burns at [recruit@cs.toronto.edu](mailto:recruit@cs.toronto.edu).

Review of applications will commence on January 7, 2013 and continue until the position is filled. To ensure full consideration, applications should be received by February 4, 2013.

For more information on the Department of Computer Science, see [www.cs.toronto.edu](http://www.cs.toronto.edu).

The University of Toronto is strongly committed to diversity within its community and especially welcomes applications from visible minority group members, women, Aboriginal persons, persons with disabilities, members of sexual minority groups, and others who may contribute to the further diversification of ideas.

All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority.

**University of Toronto Mississauga**  
**Department of Mathematical &**  
**Computational Sciences**  
**Assistant Professor**

The Department of Mathematical & Computational Sciences, University of Toronto Mississauga and the Graduate Department of Computer Science, University of Toronto invite applications for a tenure-stream position in Computer Systems. The appointment is at the rank of Assistant Professor and will begin on July 1, 2013. Specific areas of interest include operating systems, networks, distributed systems, database systems, computer architecture, programming languages and software engineering.

The University of Toronto is an international

leader in computer science research and education, and the Department of Mathematical and Computational Sciences enjoys strong ties to other units within the University. The successful candidate will be expected to participate actively in the Graduate Department of Computer Science at the University of Toronto.

Candidates should have (or be about to receive) a Ph.D. in computer science or related field. We seek outstanding applicants with an ability to pursue innovative research at the highest level and with a strong commitment to undergraduate and graduate teaching. Evidence of excellence in teaching and research is required. Salaries are competitive with our North American peers and will be determined according to experience and qualifications.

Applicants should apply online at <http://recruit.cs.toronto.edu>, and include curriculum vitae, a list of publications, a research and teaching statement, and the names and email addresses of at least three references. Other supporting materials may also be included. We will not accept applications submitted by post.

Review of applications will begin on January 7, 2013 and continue until the position is filled. To ensure full consideration applications should be received by February 4, 2013.

For more information about the Department of Mathematical and Computational Sciences please visit our home page: <http://www.utm.toronto.ca/math-cs-stats/>.

The University of Toronto is strongly committed to diversity within its community and especially welcomes applications from visible minority group members, women, Aboriginal persons, persons with disabilities, members of sexual minority groups, and others who may contribute to the further diversification of ideas.

All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority.

**University of Wisconsin-Green Bay**  
**Assistant Professor or Lecturer**

The University of Wisconsin-Green Bay seeks candidates for an Assistant Professor or Lecturer position, depending on qualifications, in the Computer Science department. **For more complete information:** <http://www.uwgb.edu/hr/jobs/position808.html>

**University of Wisconsin – Milwaukee**  
**Assistant, Associate, Full Professor**

Faculty position in the Department of Health Informatics and Administration (HIA) to serve as Director of the newly established Center for Biomedical Data and Language Processing (BioDLP). The appointment will begin on or before August 19, 2013.

APPLY: <http://jobs.uwm.edu/postings/10878>

**US Air Force Academy**  
**Coleman-Richardson Endowed Chair**

U. S. AIR FORCE ACADEMY Department of Computer Science is accepting applications for our

2013-2014 Coleman-Richardson Endowed Chair position. See <http://www.usafa.edu/df/dfcs/dfcs/jobs.cfm> or call (719) 333-7474 for details. U.S. Citizenship required.

**Utah State University**  
**Assistant Professor of Computer Science**

Applications are invited for multiple open faculty positions at the Assistant Professor level, for employment beginning Spring or Fall semester, 2013.

USU offers competitive salaries and outstanding medical, retirement and professional benefits including an annual tax-free contribution to your retirement fund of 14.2% of your annual salary (see <http://www.usu.edu/hr/> for details). The department currently has approximately 300 undergraduate majors, 70 MS students and 25 PhD students. There are 14 full time faculty positions and one lecturer position in the department. The department's BS degrees are accredited by the Computing Accreditation Commission of ABET, Inc. UtahStateUniversity is a Carnegie Research Doctoral extensive University of over 20,000 students, nestled in a mountain valley 80 miles north of Salt Lake City, Utah. Opportunities for a wide range of outdoor activities are plentiful. Housing costs are at or below national averages, and the area provides a supportive environment for families and a balanced personal/professional life. Women, minority, veteran and candidates with disabilities are encouraged to apply. USU is sensitive to the needs of dual-career couples. UtahStateUniversity is an affirmative action/equal opportunity employer, with a National Science Foundation ADVANCE Gender Equity program, committed to increasing diversity among students, faculty, and all participants in university life.

Applicants must have completed a PhD in Computer Science by the time of appointment. The positions require demonstrated research success; a significant potential for attracting external research funding; excellence in teaching both undergraduate and graduate courses; the ability to supervise student research; a commitment to department, university, and community service; and excellent communication skills. The department is interested in strengthening its focus in the following areas: Cyber Security, Software Engineering, Databases, HCI, Parallel and Distributed Computing Systems and Networks, and Mobile Computing.

Applications must be submitted using USU's online job-opportunity system. To access this job opportunity directly and begin the application process, visit <https://jobs.usu.edu/applicants/Central?quickFind=52583>.

To be considered, an application must include a letter of interest, a current curriculum vita (statements of research experience and interests, proposals written and funded, publications, and teaching experience), letters from at least three references, and completion of the faculty interests survey at: <https://www.surveymonkey.com/s/USUCSFACSRCH>

(This is not a hot link and so you will need to copy and paste this link into your browser) The review of applications will begin 30 days after posting of this announcement on the USU Jobs web site and continue until the positions are filled.





From the intersection of computational science and technological speculation, with boundaries limited only by our ability to imagine what could be.

DOI:10.1145/2398356.2398382

Rudy Rucker

## Future Tense Share My Enlightenment

*I self-publish, and you get to sail my aether wave for free.*

I'VE ALWAYS THOUGHT of my novels as “beatnik science fiction,” not that anyone else uses those words. “Beatnik” is just something I like. I’m more what you’d call a kiqqie or a qrude. I live in a hole in the ground. I eat dirt. These are modern times.

At 36 I published my first novel, *Bad Brain*, about a brain in a jar that grows tentacles, rides a bicycle to the studio of a talk-radio station, and hollows out the head of an anti-beatnik broadcaster. Having preserved the original gray matter in its own jar, the bad brain takes up residence within the vacated skull and entertains the radio audience in offbeat but positive ways, tutoring the broadcaster’s brain all the while. At book’s end, the now-peace-loving broadcaster’s brain is restored to its place, and the bad brain rides his bicycle into the sunset in search of further ways to improve the world. *Bad Brain* appeared in paperback and as an aether wave. It was met with indifference, mutating to derision and scorn.

No matter. I developed a following. I won an award.

At night, alone in my burrow, I’d rub my feelers over the emerging good reviews. My quill would stiffen. My ink-sac would fill. I wrote more beatnik SF novels.

As I stand before you today, I’m 66, with a stack of beatnik SF novels to my credit. Meanwhile, my sales have turned anemic, with ever-smaller print-runs. The cretinous, slavering fans have become oblivious to my work. The reviewers jeer, and exhort me to stop.

As a comeback stratagem, I pub-



lished my autobiography, *Beatnik SF Writer*. My long-term publisher and I thought it might serve as a late-life mainstream break-out title. It bombed, and my long-term publisher dropped me.

**At this point  
my plan was  
to distribute  
my novel  
as malware.**

What next? I wrote another beatnik SF novel, *On The Nod*, about a Kentucky boy on a galactic roadtrip with a drug-addled alien cuttlefish searching for its soul, with the soul found in the gut of a microscopic cockroach in a you-tweak-it gene bar in Oakland, CA.

I found a small publisher for *On The Nod*. For reasons that were, I maintain, solely logistical, it bombed, too. The small publisher dropped me.

I began writing another beatnik SF novel. What else could I do? I should mention, by the way, that at all times I have had at least a few loyal followers, my cognoscenti. I dedicated my new novel to them. This one I called *Zip Zap*, about an allegedly insane man who befriends a possibly imaginary sea slug from the 10<sup>th</sup> [CONTINUED ON P. 135]



# IEEE 9<sup>th</sup> World Congress on Services (SERVICES 2013)

June 27—July 2, 2013, Santa Clara Marriott, CA, USA (center of Silicon Valley)  
Federation of 5 Theme Topic Conferences on Services from Different Angles  
(<http://www.servicescongress.org>)

## IEEE 6<sup>th</sup> International Conference on Cloud Computing (CLOUD 2013)



Cloud Computing is becoming a scalable services delivery and consumption platform in the field of Services Computing. The technical foundations of Cloud Computing include Service-Oriented Architecture and Virtualizations. Major topics cover Infrastructure Cloud, Software Cloud, Application Cloud, Social Cloud, & Business Cloud. Visit <http://thecloudcomputing.org>.

## IEEE 20<sup>th</sup> International Conference on Web Services (ICWS 2013)

ICWS 2013 will feature data-centric, web-based services modeling, design, development, publishing, discovery, composition, testing, QoS assurance, adaptation, and delivery technologies and standards. Visit <http://icws.org>.



## IEEE 10<sup>th</sup> International Conference on Services Computing (SCC 2013)

SCC 2013 will focus on services innovation lifecycle e.g., enterprise modeling, business consulting, solution creation, services orchestration, optimization, management, business process integration and management. Visit <http://conferences.computer.org/>



## IEEE 2<sup>nd</sup> International Conference on Mobile Services (MS 2013)

MS 2013 will feature all aspects of mobile services including modeling, construction, deployment, middleware, and user experience with a special emphasis on context-awareness in social settings. Visit <http://themobileservices.org/2013>.



## IEEE 2<sup>nd</sup> International Conference on Services Economics (SE-BigData 2013)

SE 2013 will focus on quantitative analysis of impact on service financial and enterprise operational outcome from BigData initiative, including outcome metrics identification, risk assessment and trade-off of IT solution selection and optimization of IT expense. Visit <http://ieeese.org/2013>.



Sponsored by IEEE Technical Committee on Services Computing (TC-SVC, <http://tab.computer.org/tscs>)

Conference proceedings are EI indexed. Extended versions of invited ICWS/SCC/CLOUD/MS/SE papers will be published in IEEE Transactions on Services (TSC, SCI & EI indexed), International Journal of Web Services Research (JWSR, SCI & EI indexed), International Journal of Business Process Integration and Management (IJBPIIM), and IEEE IT Pro (SCI & EI Indexed).



IBM Research



### Submission Deadlines

ICWS 2013: 1/28/2013  
CLOUD 2013: 1/28/2013  
SCC 2013: 2/11/2013  
MS 2013: 2/11/2013  
SE 2013: 3/1/2013  
SERVICES 2013: 3/1/2013

Contact: Liang-Jie Zhang (LJ)  
zhanglj@ieee.org  
(Steering Committee Chair)



IEEE TRANSACTIONS ON  
**SERVICES  
COMPUTING**



# Open Collaboration Research

Open Access  
Open Data  
Open Government  
**Research**

Free, Libre, and  
Open Source  
Software  
**Research**

Wikipedia  
**Research**

Open Collaboration:  
Wikis, Social Media, etc.  
**Research**



# OPENSYM 2013

Hong Kong, China  
Aug 5-7, 2013

## Call for Submissions

Research papers and posters, doctoral symposium  
Experience reports, demos, panels, tutorials, workshops

### Open Access, Data, and Government chair:

Anne Fitzgerald  
Queensland University of Technology

### Free, Libre, and Open Source chairs:

Jesus M. Gonzalez-Barahona  
Gregorio Robles  
Universidad Rey Juan Carlos

### Wikipedia chairs:

Heather Ford  
Mark Graham  
Oxford Internet Institute, University of Oxford

### Open Collaboration (Wikis, Social Media, etc.) chair:

Jude Yew  
National University of Singapore

### General chairs:

Ademar Aguiar  
Universidade do Porto

Dirk Riehle  
Friedrich-Alexander University  
Erlangen-Nürnberg

Waltraut Ritter  
City University of Hong Kong



<http://wikisym.org/cacm/>  
<http://opensym.org/cacm/>