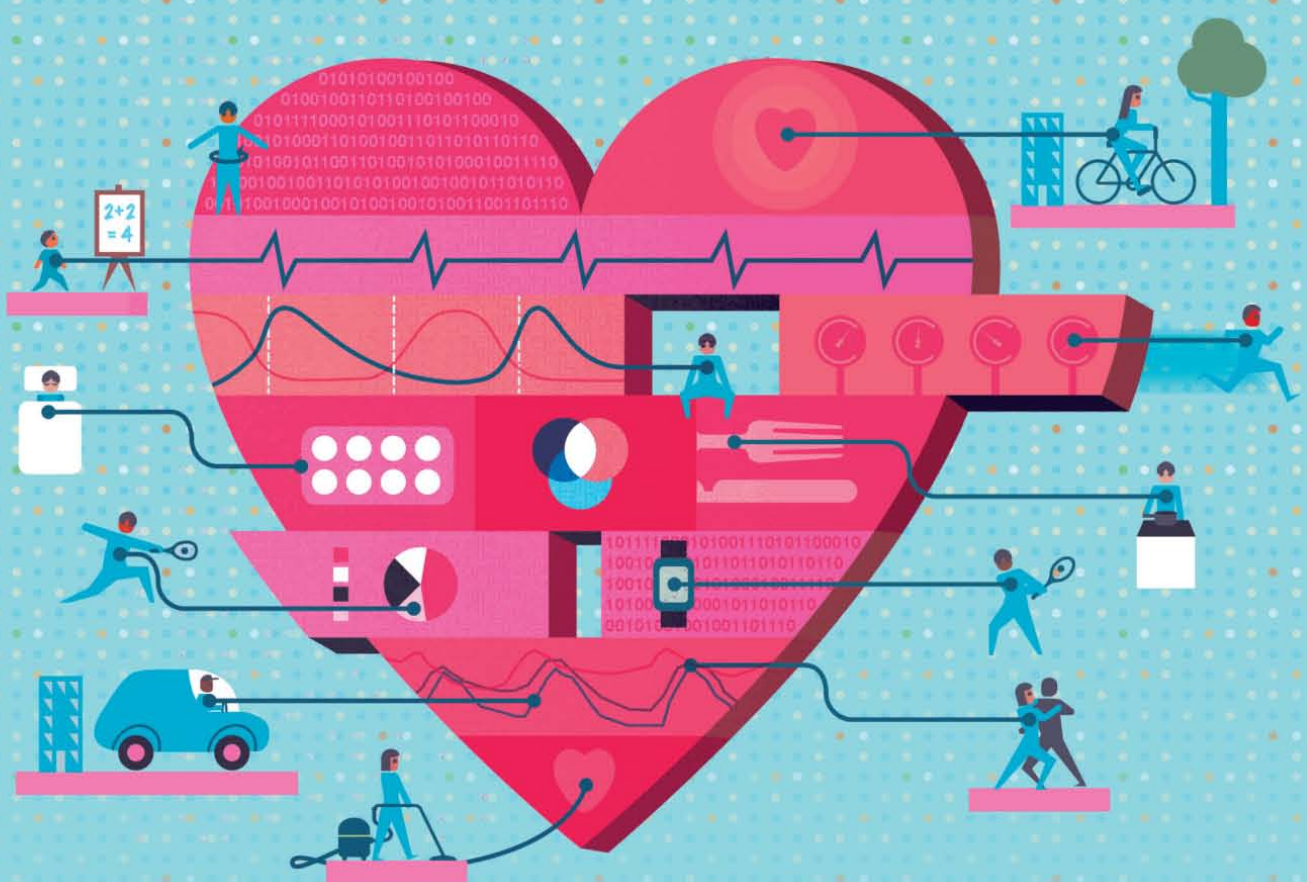
11.16

Computer

SMART HEALTH & WELL-BEING



IEEE

IEEE computer society

CELEBRATING 70 YEARS

www.computer.org/computer

myCS

HOME FOR YOUR CS SUBSCRIPTIONS

The Computer Society is excited to welcome you to a new digital publication experience! Beginning with the January 2017 issues, your digital subscriptions will be fulfilled through myCS, the interactive portal to our newly redesigned digital magazines. It just takes a moment to access this new platform that will be replacing the Qmags digital editions starting in January 2017. Please log in to myCS today using your CS member ID and password to access your subscriptions.

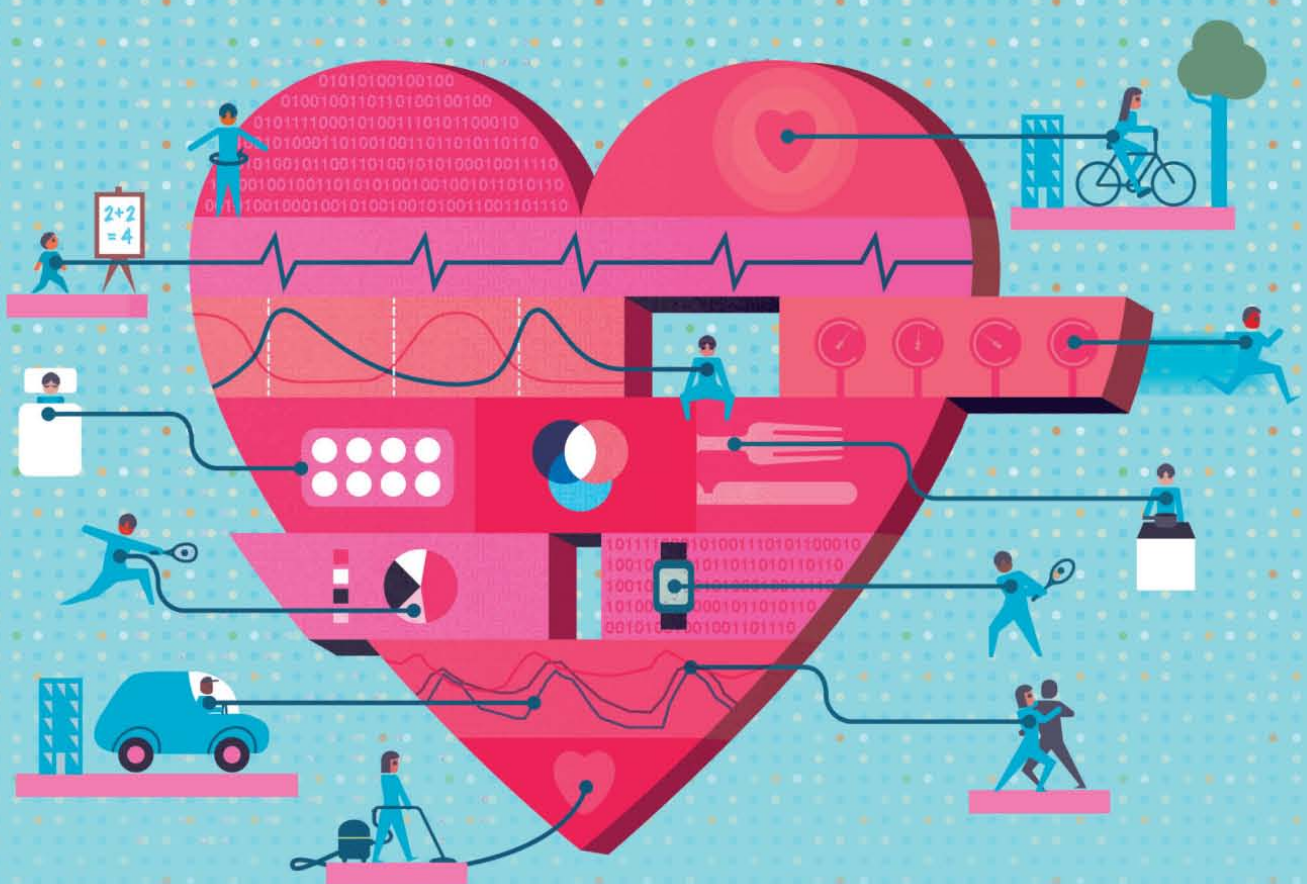
We have arranged with Qmags for your archival issues to remain available through 2017, accessible through the Qmags website using your existing email address and password. You are encouraged to download your existing Qmags archived issues by the end of this year to keep your personal library of past issues intact. myCS will archive your subscriptions going forward.

http://mycs.computer.org

11.16

Computer

SMART HEALTH & WELL-BEING



IEEE

IEEE computer society

CELEBRATING 70 YEARS

www.computer.org/computer

IEEE Computer Society Is Where You Choose the Resources that Fit Your Career

Find the membership that fits you best. IEEE Computer Society lets you choose your membership — and the benefits it provides — to meet your specific career needs. With four professional membership categories and one student package, you can select the precise industry resources, offered exclusively through the Computer Society, that will help you achieve your goals.

Select your membership	Preferred Plus		Training & Development		Research		Basic		Student
	$60 *IEEE Member*	**$126** *Affiliate Member*	**$55** *IEEE Member*	**$115** *Affiliate Member*	**$55** *IEEE Member*	**$115** *Affiliate Member*	**$40** *IEEE Member*	**$99** *Affiliate Member*	**$8** *Does not include IEEE membership*
Computer magazine (12 digital issues)*	✓		✓		✓		✓		✓
ComputingEdge magazine (12 issues)	✓		✓		✓		✓		✓
Members-only discounts on conferences and events	✓		✓		✓		✓		✓
Members-only webinars	✓		✓		✓		✓		✓
Unlimited access to *Computing Now*, computer.org, and the new mobile-ready myCS	✓		✓		✓		✓		✓
Local chapter membership	✓		✓		✓		✓		✓
Safari Books Online (600 titles and 50 training videos)	✓		✓						✓
Skillsoft online solutions (courses, certifications, practice exams, videos, mentoring)	✓		✓						✓
Two complimentary Computer Society magazine subscriptions	✓				✓				
myComputer mobile app	*30 tokens*				*30 tokens*				*30 tokens*
Computer Society Digital Library	*12 FREE downloads*		*Member pricing*		*12 FREE downloads*		*Member pricing*		*Included*
Training webinars	*3 FREE webinars*		*3 FREE webinars*		*Member pricing*		*Member pricing*		*Member pricing*
Priority registration to Computer Society events	✓								
Right to vote and hold office	✓		✓		✓		✓		
One-time 20% Computer Society online store discount	✓								

* *Print publications are available for an additional fee. See catalog for details.*

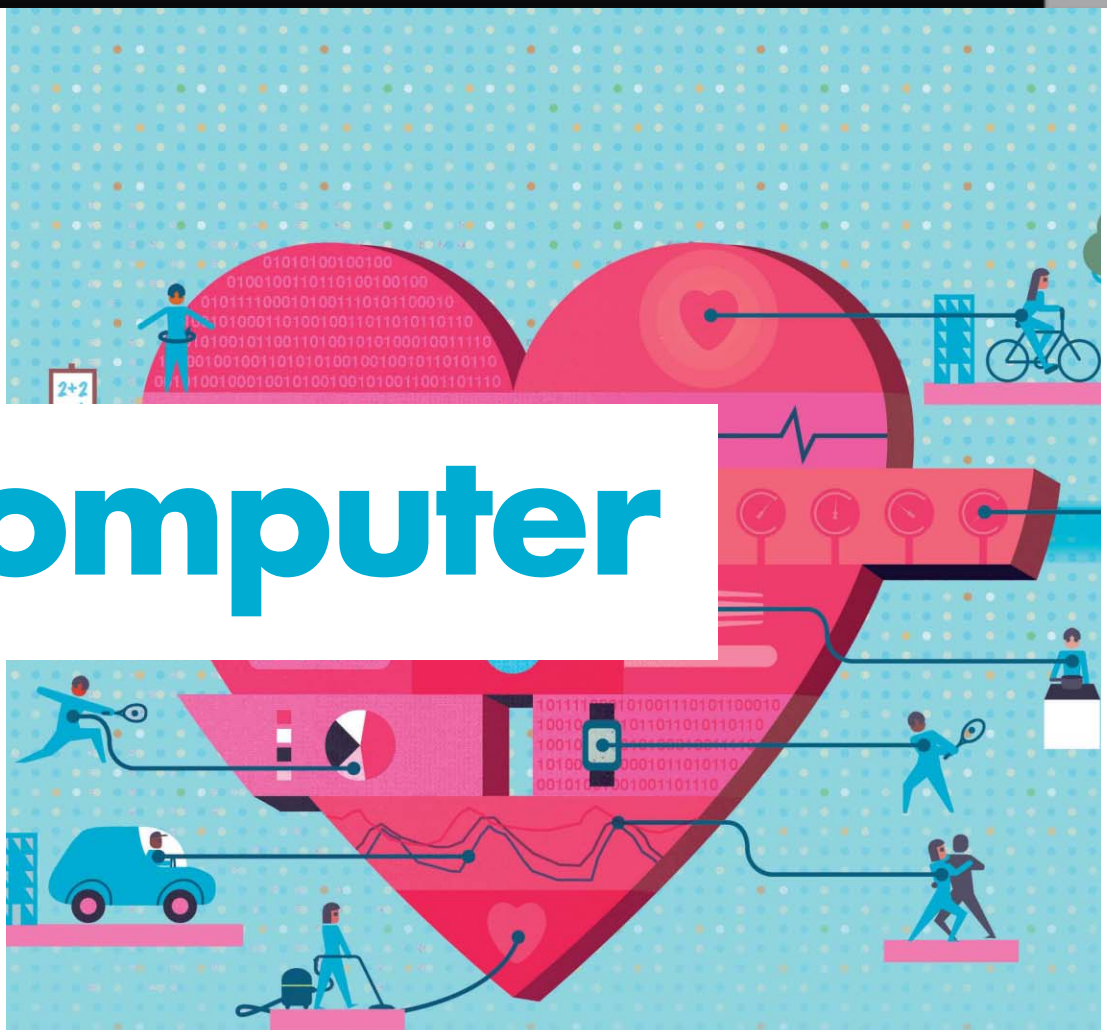
IEEE computer society

www.computer.org/membership

Computer

IEEE COMPUTER SOCIETY http://computer.org // +1 714 821 8380
COMPUTER http://computer.org/computer // computer@computer.org

Computer



GUEST EDITORS' INTRODUCTION

Smart Health and Well-Being

UPKAR VARSHNEY AND CARL K. CHANG

NOVEMBER 2016

FEATURES

14

The Future of Smart Health

S. JAY OLSHANSKY, BRUCE A. CARNES, YANG CLAIRE YANG, NORVELL MILLER, JANET ANDERSON, HIRAM BELTRÁN-SÁNCHEZ, AND KARL RICANEK JR.

22

Key Success Factors for Smart and Connected Health Software Solutions

NOEL CARROLL

29

Using Smart Homes to Detect and Analyze Health Events

GINA SPRINT, DIANE J. COOK, ROSCHELLE "SHELLY" FRITZ, AND MAUREEN SCHMITTER-EDGECOMBE

NOVEMBER 2016

CONTENTS

ABOUT THIS ISSUE

Advances in smart health technologies are giving a better quality of life to individuals and society.

FEATURES CONTINUED

38 Machine Learning in Cardiac Health Monitoring and Decision Support

SHUROUQ HIJAZI, ALEX PAGE, BURAK KANTARCI, AND TOLGA SOYATA

49 Privacy as a Service: Protecting the Individual in Healthcare Data Processing

XIANG SU, JARKKO HYYSALO, MIKA RAUTIAINEN, JUKKA RIEKKI, JAAKKO SAUVOLA, ALTTI ILARI MAARALA, HARRI HIRVONSALO, PINGJIANG LI, AND HARRI HONKO

COMPUTING PRACTICES

60 Architectural Approaches to Security: Four Case Studies

HUMBERTO CERVANTES, RICK KAZMAN, JUNGWOO RYOO, DUYOUNG CHOI, AND DUKSUNG JANG

See **www.computer.org/computer-multimedia** for multimedia content related to the features in this issue

RESEARCH FEATURE

68 Balanced Service Chaining in Software-Defined Networks with Network Function Virtualization

PO-CHING LIN, YING-DAR LIN, CHENG-YING WU, YUAN-CHENG LAI, AND YI-CHIH KAO

COLUMNS

6 32 & 16 YEARS AGO
Computer, November 1984 and 2000
NEVILLE HOLMES

77 COMPUTING EDUCATION
Engineering the New Boundaries of AI
AMIR BANIFATEMI AND JEAN-LUC GAUDIOT

80 AFTERSHOCK
Equity, Safety, and Privacy in the Autonomous Vehicle Era
VASANT DHAR

84 CYBERTRUST
The Fog of War in Cyberspace
ALEXANDER KOTT, ANANTHRAM SWAMI, AND BRUCE J. WEST

88 STUDENT DESIGN SHOWCASE
Tactile Digital Braille Display
GREG BYRD

91 OUT OF BAND
Chasing Elbridge's Ghost: The Digital Gerrymander
HAL BERGHEL

96 CLOUD COVER
Cloud Federation and the Evolution of Cloud Computing
DIMITRIOS G. KOGIAS, MICHAEL G. XEVGENIS, AND CHARALAMPOS Z. PATRIKAKIS

104 THE ERRANT HASHTAG
"I'm Not a Computer Scientist, but ..."
DAVID ALAN GRIER

Departments

8 Elsewhere in the CS LEE GARBER

10 Spotlight on Transactions MATTHEW B. DWYER

Membership News

5 70th Anniversary Milestones LORI CAMERON

83 IEEE Computer Society Information

100 Call and Calendar

Circulation: *Computer* (ISSN 0018-9162) is published monthly by the IEEE Computer Society. IEEE Headquarters, Three Park Avenue, 17th Floor, New York, NY 10016-5997; IEEE Computer Society Publications Office, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720; voice +1 714 821 8380; fax +1 714 821 4010; IEEE Computer Society Headquarters, 2001 L Street NW, Suite 700, Washington, DC 20036. IEEE Computer Society membership includes a subscription to *Computer* magazine.

Postmaster: Send undelivered copies and address changes to *Computer*, IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Canadian GST #125634188. Canada Post Corporation (Canadian distribution) publications mail agreement number 40013885. Return undeliverable Canadian addresses to PO Box 122, Niagara Falls, ON L2E 6S8 Canada. Printed in USA.

70TH ANNIVERSARY MILESTONES

IEEE Computer Society Technical Committees

For more than five decades, the IEEE Computer Society's Technical Committees (TCs) have served as strong communities centered on specific areas of computing. Currently, two technical councils address the broad areas of software engineering (TCSE) and test technology (TTTC), with 26 more keenly focused TCs, and one task force on the emergent area of rebooting computing. These groups range in size from 500 to 2,500 international members and have been instrumental in contributing to the professional development and advancement of CS members in current as well as unexplored technical areas. The TCs have long provided international forums for the exchange of ideas among practitioners and researchers in all aspects of computer research, development, methodology, and application. Serving as the focal point for society activities within a technical discipline, TCs directly organize and influence conferences, publications, standards development, educational activities, and awards.

Our TCs are leaders in today's hot topics of computer science, such as cybersecurity (TC on Security and Privacy), machine intelligence (TC on Pattern Analysis and Machine Intelligence), and semantic computing (TC on Semantic Computing), and they disseminate research information and opportunities via their conference activities. Most TCs have a flagship conference at which the field's most significant research advances are featured. TCSE, for example, sponsors more than 35 conferences annually, including its flagship, the International Conference on Software Engineering (ICSE). Another example comes from the smaller TC on Mathematical Foundations of Computing, which has sponsored the annual Symposium on Foundations of Computer Science (FOCS) for more than 50 years. The TC on Microprocessors and Microcomputers also organizes a set of popular conferences—Hot Chips, Hot Interconnects, Cool Chips, and Multi-Core SoC—with the first three nearing their third and fourth decades of activity.

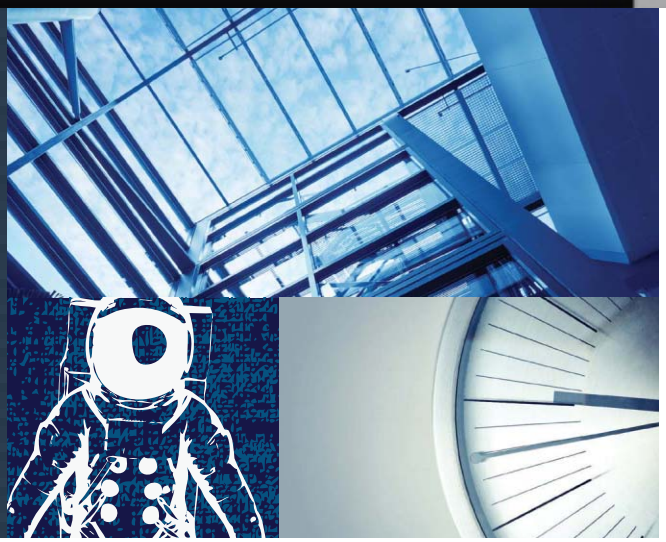
Publications also benefit from the efforts of TC members and volunteers, whether it is a direct tie to the editorial board membership, cross promotion at related conferences, special conference content publication, authorship, reviewers, and so on. TC membership is free and supports technologists at all career levels to expand their knowledge, professional network, and impact on the field.

—Lori Cameron

To learn more and join a TC, visit
www.computer.org/web/tandc/technical-committees.

IEEE computer society
CELEBRATING 70 YEARS

32 & 16 YEARS AGO



EDITOR **NEVILLE HOLMES**
holmeswn@yahoo.com.au

NOVEMBER 1984

www.computer.org/csdl/mags/co/1984/11/index.html

Letter (p. 5) "We can see that the term 'software maintenance' accurately describes the activities performed during this phase of the software life cycle. I propose that 'software maintenance engineers' should become standard nomenclature for those who perform the vital function of correcting, modifying, and maintaining the thousands of existing computer programs."

Graphical Programming (p. 7) "In this article, we describe our work on developing a programming methodology called Pict that permits humans to use their native intelligence in programming."

Interview (p. 42) "Four advances in technology—the microprocessor, the Winchester disk, the relational database, and a novel interprocessor communication scheme—formed the foundation for a startup company's approach to the $2 billion annual database market."

Software Development (p. 57) "To remain competitive in an international software field that includes system communications as well as programming, NEC set up an integrated, corporation-wide development and production project."

Software Management (p. 66) "This article presents the management perspective on the GTE software engineering culture methodology, environment, and tools as well as the problems of new technology transfer."

Workstation Networks (p. 74) "Our research had two primary goals. The first was to identify the factors behind the development of local area network–based systems of workstations. ... The second goal was to construct a profile of a workstation and its environment, to distinguish it from such products as word processors and intelligent terminals."

Packaging Workshop (p. 90) "This year's keynote session featured a panel that discussed wafer-scale integration [WSI], specifically the differences in Trilogy and Mosaic WSI, and the alternatives of silicon on silicon and other dense-packaging schemes."

Introspective Integration (p. 94) "Since inexpensive logic for testing and reconnecting modules eliminates the need to customize wafers, the concept of self-configurability is creating new interest in [WSI]. ... However, there seems to be little public debate about the merits of various approaches to WSI, and this note is an attempt to initiate a discussion and provide a personal answer."

Gentle Printing (p. 102) "A nonimpact office printer based on a patented, thin-film magnetic recording head has been introduced by Ferix Corporation. The Model 800 magnetic page printer ... can store images on its magnetic drum so that once an image has been created, it can be used indefinitely as a magnetic master in a duplicator mode to print 14 pages per minute without data retransmission."

Ombudsman (p. 113) "Computer Society [CS] membership has grown to almost 80,000, making it inevitable that some members would run afoul of the society business routine and face problems like lost magazines, inaccurate dues notices, and no answers to their complaints. To resolve such problems, the CS Governing Board created the post of ombudsman to help CS members cut through organizational red tape."

The Future (p. 114) "The personal expert system is a computer so small, yet so powerful, it can provide the problem-solving capabilities of a human expert anywhere they are needed."

Public Policy (p. 117) "The formation of a national commission on computer and communications systems has been recommended to Congress. The purpose is to examine the legal, economic, institutional, social, and technical aspects of safeguarding computerized resources, as well as to study the scope and nature of threats and vulnerabilities of computer/communications systems."

0018-9162/16/$33.00 © 2016 IEEE

NOVEMBER 2000

www.computer.org/csdl/mags/co/2000/11/index.html

Letter (p. 4) "I appreciated reading Neville Holmes's recent article on reforming the Olympic Games ('Olympic Games Reform: A Study in System Engineering,' Sept. 2000, pp. 91–93). I found the article rather appealing in both its content and its well-presented perspective of system engineering."

Mixed-Signal Chips (p. 12) "The technology's profile has increased dramatically with its growing use in intelligent handheld devices, which take advantage of the chips' powerful digital capabilities to leverage analog functionality."

Broadband Gaming (p. 16) "Now that video gaming has leaped onto the Internet, avid gamers—in their thirst for the fastest, most intense, and most realistic experiences—are adopting residential broadband technologies ... at a faster rate than the general public."

A Modern Museum (p. 22) "This article describes the issues we faced in the design and management of our Web-based image collection, the Thinker ImageBase, containing more than 75,000 images of artwork from the Fine Arts Museums of San Francisco."

Summary Automation (p. 29) "Although some summarizing tools are already available, with the increasing volume of online information, it is becoming harder to generate meaningful and timely summaries. [Available] tools ... are useful, but their application is limited to *extraction*—selecting original pieces from the source document and concatenating them to yield a shorter text. *Abstraction*, in contrast, paraphrases in more general terms what the text is about."

Compressed Search (p. 37) "In this article, we discuss the recent techniques that permit a fast and direct method for searching compressed text, and we explain how these new techniques can improve the overall efficiency of IR [information retrieval] systems."

Visual Search (p. 46) "My colleagues and I developed a prototype system called ImageScape ... to find visual media over intranets and the Web. The system integrates technologies such as vector quantization–based compression of the image database and k-d trees for fast searching over high-dimensional spaces."

Addressable Issues (p. 54) "In this article, we identify a set of issues organized within an overall framework that software developers must address for component-based systems (CBSs) to achieve their full potential."

Animated Networks (p. 63) "Nam, the network animator that we developed in our work at the VINT [Virtual InterNetwork Testbed] project, provides packet-level animation, protocol graphs, traditional time-event plots of protocol actions, and scenario editing capabilities. Nam benefits from a close relationship with ns, which can collect detailed protocol information from a simulation."

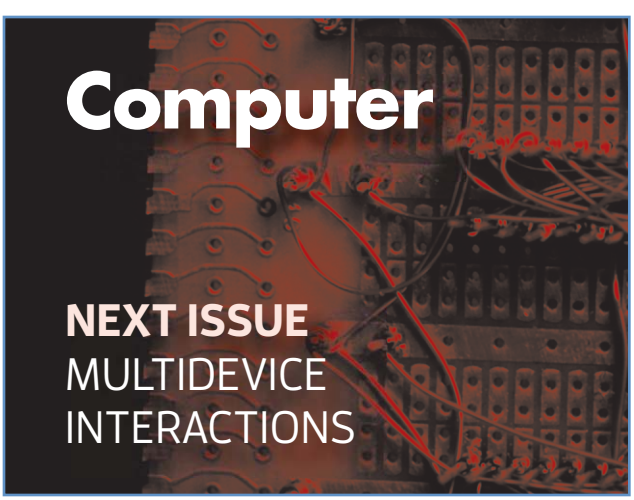
A Competition (p. 70) "The CS challenges undergraduates from around the world to put their learning to the test in the second annual CS International Design Competition [CSIDC]. Intended to replicate the type of project students might encounter in industry, the CSIDC is an extensive contest in which students work autonomously to design and build a working prototype of a device that could be used to solve a real-world problem."

Genetic Search (p. 118) "Genetic search algorithms enable intelligent and efficient Internet searches. They are especially useful when the search space is relatively large."

Clashing Models (p. 120) "The authors offer two techniques for avoiding the conflicting assumptions that often snare software projects in a costly and time-consuming spiderweb."

Background Computers (p. 123) "Making computers ubiquitous is not enough; we should also strive to make them invisible. But, in doing so, we will face many research challenges."

Program Coding (p. 128) "In our push to make programs more understandable, we have often overlooked the equally important goal of making them easier to code correctly."



ELSEWHERE IN THE CS

EDITOR **LEE GARBER**
lgarber@computer.org

Computer Highlights Society Magazines

The IEEE Computer Society's 13 peer-reviewed technical magazines cover cutting-edge computing topics including scientific applications, Internet computing, machine intelligence, pervasive computing, security and privacy, digital graphics, cloud computing, and computer history. Here, we highlight recent issues of other Computer Society magazines.

IEEE Software

Developers of **systems of systems** (SoSs) face numerous challenges. The authors of "Monitoring Requirements in Systems of Systems," from *IEEE Software*'s September/October 2016 issue, discuss their ReMinds tool, designed to meet these challenges by instrumenting SoS systems to extract events and data at runtime. ReMinds then defines requirements as constraints to check for expected behavior and properties.

IEEE Internet Computing

Measuring spam's cost for users and network operators and identifying who pays for it is difficult. In "Measuring, Characterizing, and Avoiding Spam Traffic Costs," from *IEEE Internet Computing*'s July/August 2016 issue, the authors provide a way to quantify those costs. They show that stub networks incur high spam traffic costs. However, they also found that some networks actually profit from spam traffic and might not want to filter it. They then present an algorithm to identify networks that would benefit from cooperating to filter such traffic.

Computing in Science & Engineering

As scientific data volumes continue to grow, researchers increasingly need a flexible computational infrastructure that can support the entire data science life cycle. The authors of "A Case for Data Commons: Toward Data Science as a Service," from *CiSE*'s September/October 2016 issue, explain their development of an **interoperable data commons infrastructure** that collocates data, storage, and computing resources with common analysis tools. Challenges remain, but development of such an infrastructure brings us one step closer to data science as a service for the scientific research community.

IEEE Security & Privacy

The gaming industry collects participants' data to generate marketing-related revenue and improve the playing experience. However, the need to protect player privacy complicates this process. "**Incorporating Privacy into Digital Game Platform Design:** The What, Why, and How," from *IEEE S&P*'s July/August 2016 issue, details an iterative approach that includes privacy-by-design principles in game development.

IEEE Cloud Computing

IEEE Cloud Computing's July/August 2016 special issue on **manufacturing and the cloud** includes articles on the economics and strategy of manufacturing and the cloud; cloud manufacturing's security, privacy, and forensic concerns; and a roadmap for using the Internet of Things and cloud computing in manufacturing.

IEEE Computer Graphics and Applications

"**Designing for Insight:** A Case Study from Tennis Player Analysis," which appears in *CG&A*'s July/August 2016 issue, describes a combinatory design process that uses incremental addition to generate increasingly complex data

0018-9162/16/$33.00 © 2016 IEEE

arrangements and thus create new ways to see the information and discover new insights.

IEEE Intelligent Systems

"Design of a **Multiagent System for Real-Time Traffic Control**," from *IEEE Intelligent Systems*' July/August 2016 issue, examines the various steps involved in analyzing and designing such a system for use at isolated street intersections. In the authors' model, the many agents designed for isolated intersections create, manage, and evolve their own traffic-signal plans.

IEEE MultiMedia

The authors of "**Multimedia Hashing and Networking**," from *IEEE MultiMedia*'s July–September 2016 issue, summarize shallow-learning–based and deep-learning–based hashing. State-of-the-art hashing techniques are widely used in high-efficiency multimedia storage, indexing, and retrieval. The authors also introduce multimedia information networks as a way to incorporate both visual and textual information to make deep learning practical in multimedia applications.

IEEE Annals of the History of Computing

"Two **Early Interactive Computer Network Experiments**," from *IEEE Annals*' July–September 2016 issue, looks at a couple of experiments that joined a System Development Corp. time-sharing computer with a system at the Stanford Research Institute in 1963 and with one at MIT Lincoln Laboratory in 1966 and 1967.

IEEE Pervasive Computing

The authors of "Displays as a Material: A Route to **Making Displays More Pervasive**," from *IEEE Pervasive Computing*'s July–September 2016 issue, advocate changing to an architecture that relies on autonomous pixels that independently sense input and convert it to a visual output. They also discuss their two display prototypes.

IT Professional

Rich visual information is becoming increasingly important in today's web. In "**Visual Information Retrieval**: The State of the Art," from *IT Pro*'s July/August 2016 issue, the author examines the process of searching for and retrieving images using a visual query, a process that is also called content-based image retrieval.

IEEE micro

"Ten Open Questions for Techno-Optimists," from *IEEE Micro*'s July/August 2016 issue, discusses some of the open questions regarding **productivity growth and economic gains** resulting from innovative IT.

SPOTLIGHT ON TRANSACTIONS

Finding Flaws in Natural Language Requirements

This installment of Computer's *series highlighting the work published in IEEE Computer Society journals comes from* IEEE Transactions on Software Engineering.

Matthew B. Dwyer, University of Nebraska–Lincoln

A software system begins in a user's mind. With some coaxing, the system's "idea" is expressed as natural language requirements. These requirements then drive the development of the software.

Existing approaches often structure natural language with templates that reduce ambiguity and normalize their format, making the requirements easier to understand and analyze.

A key challenge is determining whether the natural language sentences conform to the template format. Prior work allowed conformance checking, given a glossary defining the problem-domain keywords. However, practitioners often write requirements as a means of discovering and finalizing the glossary, which is difficult to do in advance.

In "Automated Checking of Conformance to Requirements Templates Using Natural Language Processing" (*IEEE Trans. Software Eng.*, vol. 41, no. 10, 2015, pp. 944–968), Chetan Arora and his colleagues explore whether modern approaches to natural language processing (NLP) can effectively find template conformance flaws in the kind of natural language requirements that practitioners create.

The authors describe the REquirements Template Analyzer (RETA), which uses NLP text-chunking techniques to overcome the need for a glossary. RETA, shown in Figure 1, integrates with existing requirements-authoring tools, enhancing them with the ability to report on conformance with different template structures, such as Chris Rupp's and EARS (Easy Approach to Requirements Engineering). RETA can also detect problematic constructs in requirements; for example, the requirement "The S&T module shall process the query data and send a confirmation to the database" might be ambiguous because "and" implies temporal ordering.

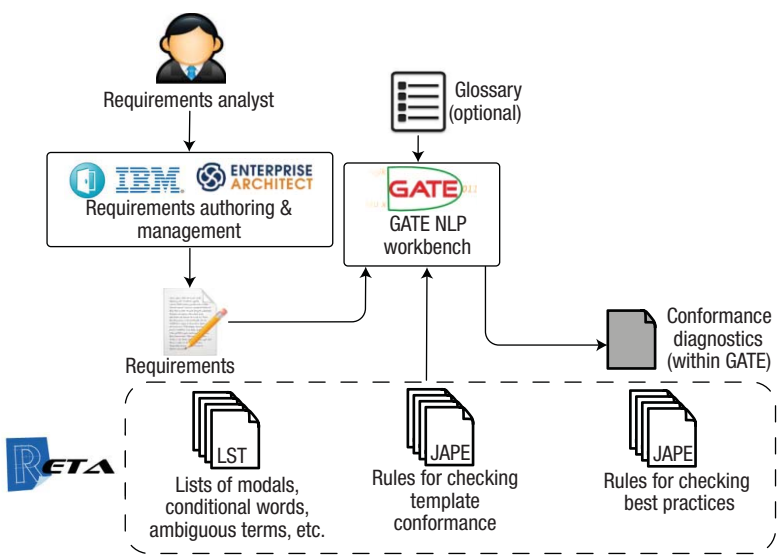


Figure 1. The REquirements Template Analyzer (RETA) allows requirements authoring tools to report on conformance with different template structures.

The paper evaluates the proposed method on four case studies reflecting the real-world variability found in industrial requirements specifications. This evaluation's breadth and depth effectively settles the question of how to check template conformance of natural language requirements for practicing engineers.

MATTHEW B. DWYER is a professor of computer science and engineering at the University of Nebraska–Lincoln. Contact him at dwyer@cse.unl.edu.

0018-9162/16/$33.00 © 2016 IEEE

COVER FEATURE
GUEST EDITORS' INTRODUCTION

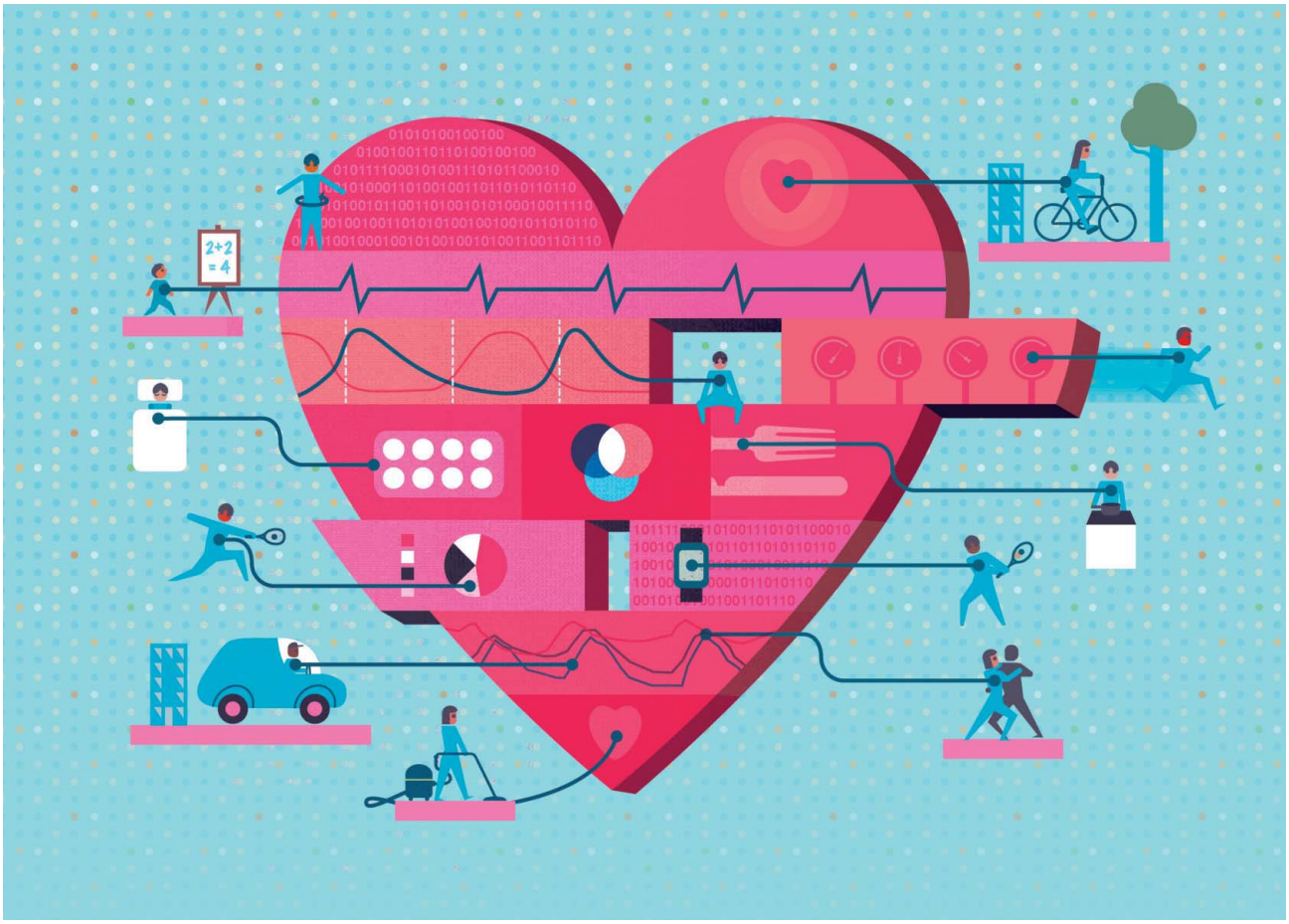
Smart Health and Well-Being

Upkar Varshney, Georgia State University
Carl K. Chang, Iowa State University

Advances in medicine and clinical care are increasingly tied to computing technologies. This special issue explores emerging trends in smart health and the benefits they bring to individual patients and society as a whole.

The $4 billion Human Genome Project successfully completed its mission to map the entire human genome 13 years ago. As hoped, the project galvanized development of personalized treatments and precision medicine as well as new capabilities for early disease detection and diagnosis. At the same time, the rapid expansion of big data analytics and cloud computing technologies has led to the creation of powerful new tools including virtualized tissue banks and disease-specific clinical-trials recruiting and selection databases. Such tools let researchers more nimbly leverage our growing knowledge of human biology to directly improve patients' health outcomes and quality of life.

GUEST EDITORS' INTRODUCTION



Despite these advances, many people still do not have access to affordable and effective healthcare, and significant challenges remain for many patient populations, including the elderly and those with chronic conditions or diseases. To address these needs, researchers are designing, implementing, and evaluating novel smart-health and wellness applications to better understand disease etiology and pathogenesis, reduce medical costs, customize care, and shift the focus from disease treatment to prevention. Key ongoing research activities include analyzing physiological and behavioral data from mobile and environmental sensors, improving telemedicine services, and exploiting emerging information sources such as social media and health data aggregators.

IN THIS ISSUE

For this special issue we received numerous high-quality submissions for consideration, and after rigorous assessment by many dedicated reviewers, we selected five articles showcasing several important facets of smart health. These include a vision of smart health's future, software innovations, behavioral aspects of health conditions, decision support for doctors, and privacy challenges in smart health.

In "The Future of Smart Health," S. Jay Olshansky, Bruce A. Carnes, Yang Claire Yang, Norvell Miller, Janet Anderson, Hiram Beltrán-Sánchez, and Karl Ricanek Jr. describe how longitudinal individual and crowdsourced data from wearable (and soon-to-be-implantable) sensors can be translated into empirically verifiable measures of risk that can be used by both patients and doctors to detect and treat disease, enhance quality of life, and increase longevity. Using several examples of applications their company is developing, the authors foresee the development of a health-data economy in which such technologies help engender new relationships among caregivers, patients, and industry.

In "Key Success Factors for Smart and Connected Health Software Solutions," Noel Carroll explains how smart and connected health (SCH) software innovations help providers and caregivers achieve desired coverage requirements and quality, reduce costs, and improve patient health outcomes and quality of life. Given the increasing reliance on software to support healthcare decisions, tools are needed to help evaluate the software's efficacy and impact. The author provides an overview of the key success factors for SCH software as it evolves and becomes more broadly adopted.

In "Using Smart Homes to Detect and Analyze Health Events," authors Gina Sprint, Diane J. Cook, Roschelle "Shelly" Fritz, and Maureen Schmitter-Edgecombe focus on how the application of behavior change detection (BCD) to patient activity data collected from environmental sensors can

improve our understanding of behavioral changes that accompany health events. The authors describe three case studies of older adults living in smart homes who experience major health events; they compared observed behaviors of the subjects with those described in the medical literature and found that the behaviors are consistent and thus can be automatically recognized and detected, highlighting the usefulness of sensor data mining for exploring the relationship between behavior and health.

In "Machine Learning in Cardiac Health Monitoring and Decision Support," Shurouq Hijazi, Alex Page, Burak Kantarci, and Tolga Soyata describe how portable medical devices can enhance and personalize healthcare. In particular, the authors introduce a system that identifies patterns in the flood of sensor data that can provide diagnostic decision support to clinicians. This system uses filtering and machine learning techniques to identify circadian patterns in long-term electrocardiograms of patients with certain genetic disorders. Based on this information, the patient's physician is provided with an estimate of health risk. In addition, the authors investigate the system's scalability, which—considering the explosive growth of medical data—is an important concern.

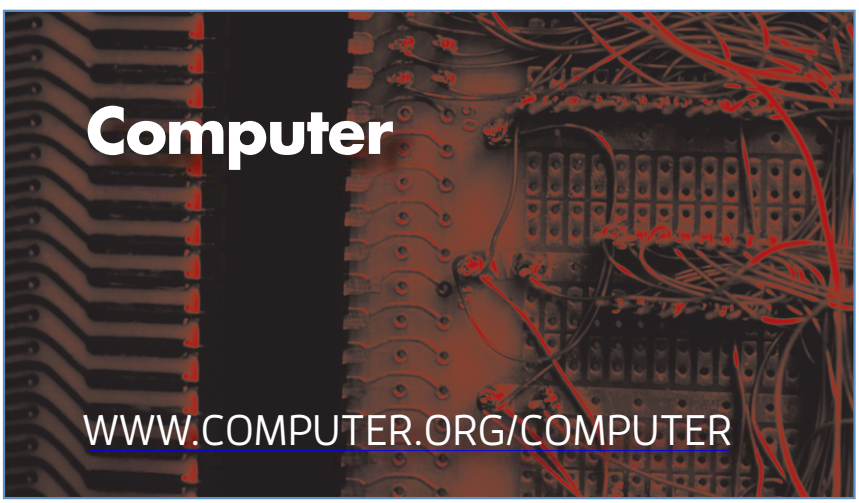
In "Privacy as a Service: Protecting the Individual in Healthcare Data Processing," Xiang Su, Jarkko Hyysalo, Mika Rautiainen, Jukka Riekki, Jaakko Sauvola, Altti Ilari Maarala, Harri Hirvonsalo, Pingjiang Li, and Harri Honko emphasize the continuing challenge and importance of privacy in e-healthcare delivery. The authors propose a privacy-focused architecture that provides tools for delivering user consent as a service. The architecture integrates data security and semantic descriptions into a trust-query framework that establishes the required interoperability and cooperation support for future smart-health services. This approach benefits all stakeholders through safer data management, cost and process savings, multiprovider participation, and innovative services based on emerging business models.

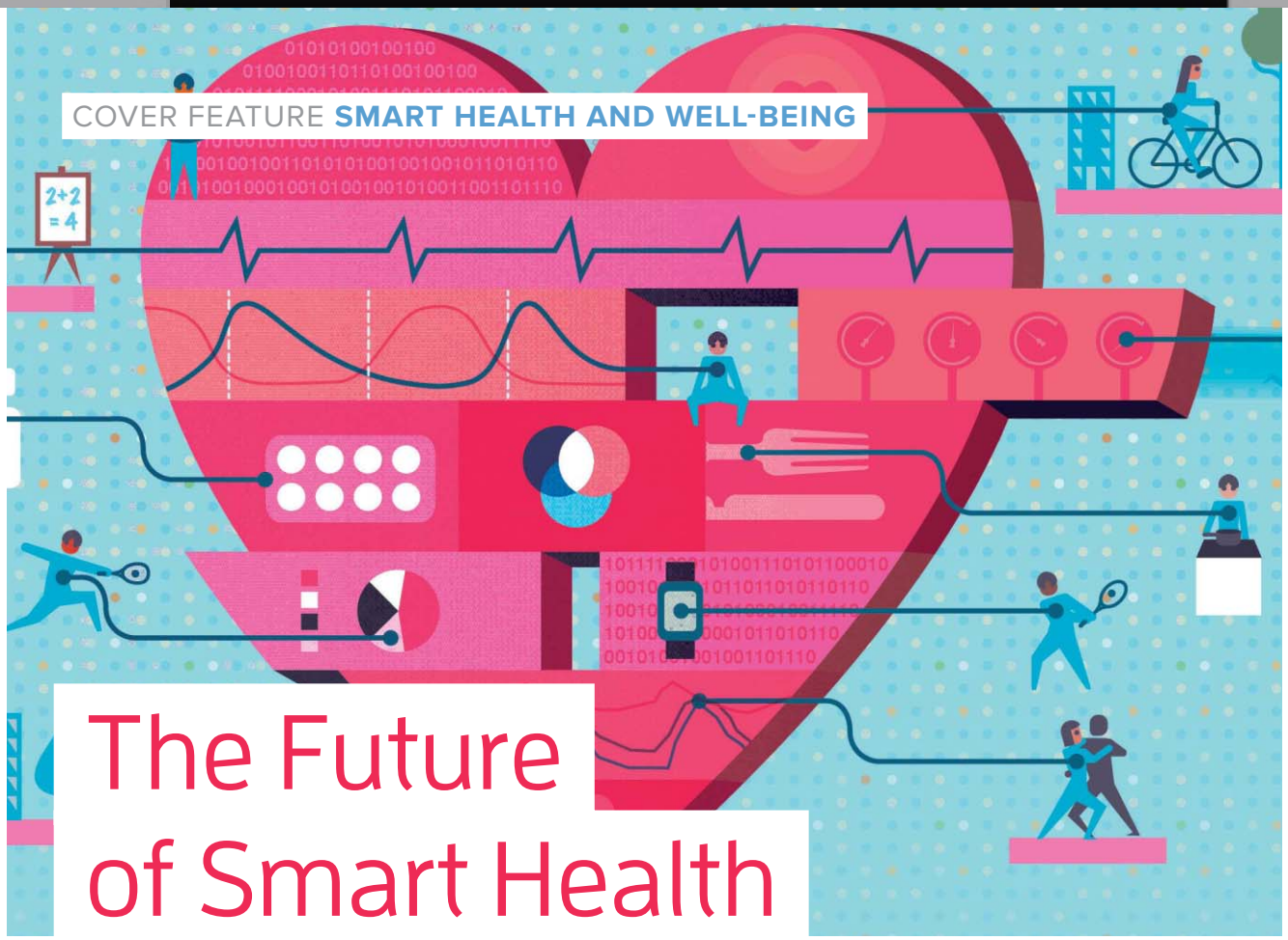
We thank the authors, reviewers, and *Computer*'s editor in chief and staff for their hard work in preparing this special issue. We hope it will help shed some light on some of the cutting-edge technologies in healthcare that are providing unprecedented benefits to individual patients and society as a whole.

ABOUT THE AUTHORS

UPKAR VARSHNEY is an associate professor of computer information systems in the J. Mack Robinson College of Business at Georgia State University. His research interests include mobile health, mobile computing, and health IT. Varshney received a PhD in telecommunications and computer networking from the University of Missouri. *Computer*'s area editor in health informatics, he is a member of IEEE, ACM, and the Association for Information Systems. Contact him at uvarshney@gsu.edu.

CARL K. CHANG is a professor of computer science as well as human–computer interaction and director of the Software Engineering Laboratory at Iowa State University. His research interests include requirements engineering, services computing, situational software engineering, and successful aging. Chang received a PhD in computer science from Northwestern University. He is a Fellow of IEEE and the American Association for the Advancement of Science, and a member of the European Academy of Sciences. Contact him at chang@iastate.edu.





COVER FEATURE **SMART HEALTH AND WELL-BEING**

The Future of Smart Health

S. Jay Olshansky, Bruce A. Carnes, Yang Claire Yang, Norvell Miller, Janet Anderson, Hiram Beltrán-Sánchez, and Karl Ricanek Jr., Lapetus Solutions

Longitudinal individual and crowdsourced health data from wearable sensors can be leveraged to increase the quality and length of life, giving rise to a new health data economy.

Human health has been tracked using various metrics since the advent of the demographic and actuarial sciences several centuries ago. Epidemiologists, sociologists, gerontologists, and public health experts have since become highly efficient at determining when in the lifespan diseases tend to emerge, what behavioral and genetic risk factors contribute to their onset and progression, how to treat them once discovered, and how basic biology influences both disease expression and length of life. In fact, during the 20th century, scientific, medical, and public health advances drove the first longevity revolution in the US: a more than 30-year increase in life expectancy at birth.

However, statistics do not capture what is actually happening inside our bodies. This became possible with advanced diagnostic and imaging technologies—including x-rays, positron emission and computed tomography, magnetic resonance imaging, electroencephalograms, electrocardiograms, and laparoscopic surgery—that enable physicians to not only observe disease after symptoms appear but increasingly to detect disease before symptoms are manifest, when intervention is most effective. For example, a routine colonoscopy can sharply reduce the risk of death from colon cancer, once a major cause of mortality in long-lived populations.[1] Yet, despite our newfound success in real-time disease tracking, there are still no physiological measures that can reliably and precisely determine biological age.[2]

THE EMERGENCE OF WEARABLE SENSORS

As interdisciplinary teams of medical researchers, biologists, and technologists search for the ideal biomarkers of aging, what has emerged in the interim is something that could not have been predicted even a few years ago—the ability to inexpensively and reliably monitor

0018-9162/16/$33.00 © 2016 IEEE

many bodily functions in real time using wearable devices. The wearable device industry has grown rapidly and is estimated to become a $31.6 billion market by 2020.[3] Instead of exclusively relying on invasive procedures such as extracting bodily fluids with syringes or using imaging devices that emanate potentially harmful radiation, we can now choose from a growing array of wearable technologies to keep close tabs on our bodies' inner workings. The immediate feedback that such technologies provide is somewhat akin to the biofeedback technology that was developed decades ago to help lower heart rate and control brain-wave functions.

Wearable technology currently has three primary applications in the healthcare domain.

The first, driven by profit and demand, is to monitor physiological attributes associated with a specific disease such as diabetes.[4,5] People with type I diabetes monitor their blood-sugar level frequently over the course of a day to determine when to introduce insulin. This is typically done with a device that draws blood multiple times during waking hours—an intrusive but necessary procedure. Newly developed wearable devices such as GlucoWise (gluco-wise.com) now make it possible to unobtrusively monitor blood-sugar level in real time. The benefits of continuous blood-sugar monitoring extend well beyond insulin control; the device can discover the body's unique and precisely measured insulin response to certain foods, and even evaluate personal insulin-based circadian rhythms to identify the best time of day to eat or avoid certain foods. Related uses of wearable technology include real-time collection of data on blood pressure (for hypertension control), resting and exercise-related heart rate, and sleeping patterns.

Another growing use of wearable devices is to enhance physical performance by tracking fitness activities in real time.[6] This not only creates a biofeedback loop informing users about which training methods yield the best results; it also enables documentation of "personal bests"—fastest time running, swimming, and so on—and the ability to set, measure, and exceed targeted levels of physical activity. Wearable devices that autonomously measure and track physical performance can be used to learn more about the body's unique functional attributes such as those associated with how certain foods or circadian rhythms influence performance. It is also popular to "gamify" fitness tracking devices to compare a user's physical achievements, such as number of steps taken in a given period of time, with those of friends, family, colleagues, or even strangers to foster competition and thereby improve performance.[7]

A third popular use of wearable devices is as a personal assistant or coach that constantly tracks some variable of interest. For example, the weight-loss industry discovered that keeping a food journal—a list of all foods consumed each day—can serve as a motivational tool and cheating deterrent for those trying to lose weight.[8] Instead of carrying around a notebook and writing down everything you eat, it is easier and faster to input data into a wearable device (or mobile application) that automatically keeps track of calories consumed or points used. Unlike the first two uses of wearable devices, which are passive data generators, this use of the technology can be either active or passive.

> THERE IS NO SHORTAGE OF INTERNAL BODILY FUNCTIONS TO MEASURE; THE CHALLENGE IS WHAT TO DO WITH ALL OF THE DATA.

Many other wearable applications are emerging that could extend the quality and length of life in ways never before thought possible.

TRANSLATING HEALTH DATA INTO EMPIRICALLY VERIFIED METRICS

Despite their enormous promise, wearable devices are a nascent technology that we are just beginning to understand. There is no shortage of internal bodily functions to measure; the challenge is what to do with all of the data. The thousands of wearable devices now flooding the market excel at generating information, but most devices are limited to tracking individual physiological parameters and motivating people to engage in more physical activity. However, getting a little buzz on your wrist or a "great job!" emoji from your fitness tracker after walking 10,000 steps doesn't seem like a

TABLE 1. Operational definitions of different physical activity levels.

Activity level	Definition	Examples
None	Sedentary	Sitting at a desk; conversing with friends; traveling in a car, bus, or train; reading; playing cards; watching TV; using a computer
Some	Activities two or more times per week involving at least 10 minutes of physical exercise that cause a slight increase in breathing or heart rate	Light walking
Moderate	Activities three or more times per week involving at least 15 minutes of physical exercise that cause light sweating and a slight to moderate increase in breathing or heart rate	Brisk walking, bicycling for pleasure, golf, gardening, dancing
Moderate/vigorous	Activities two or more times per week involving at least 20 minutes of continuous cardiovascular exercise that cause heavy sweating and/or large increases in breathing or heart rate	Running, lap swimming, aerobics classes, fast cycling
Vigorous	Activities three or more times per week involving at least 20 minutes of continuous cardiovascular exercise that cause heavy sweating and/or large increases in breathing or heart rate	Running, lap swimming, aerobics classes, fast cycling

sufficient reward—in time, a user is likely to get bored with the device and drop it in a drawer. What is needed is some means to aggregate and transmit the data to a third party that can leverage it in a more meaningful way.

Health data aggregation services are now available that collect millions of data points across time from wearable sensors—examples include Human API (humanapi.co), Welltok (welltok.com), and Validic (validic.com). Such crowdsourced data can be translated into empirically verified measures of risk to assist in disease prevention at both the individual and population level. For example, most physicians today rely on samples of bodily fluids taken before or during annual physicals to assess a patient's immediate risk of disease. This information could be complemented with extraordinarily rich data on bodily functions derived from wearable sensors and uploaded continuously between doctor visits. This would not only give the physician a more accurate picture of the patient's health status and risk factors, but could be aggregated with other patients' data to provide the equivalent of a personal longitudinal health survey. Researchers could also use this data, once aggregated, to track health trends at different levels of granularity in real time as well as apply data-mining solutions to develop new insights into the causes of diseases and ways to prevent them.

EXAMPLE APPLICATION

Lapetus Solutions has created several empirically based methods for translating data from wearable sensors into quantifiable measures of risk and benefit that can be used by individuals to measure and enhance their length and quality of life. Here we consider how a person's daily step count can be converted into a quantifiable measure of years of healthy life remaining.

Linking step count to health and longevity

Step count is a popular passive measure of physical activity provided by many wearable devices today, such as Fitbit. Although most researchers acknowledge that walking is a healthy behavior, to our knowledge no one before us had actually translated step count data from wearable devices into empirically verified measures of health benefits. Instead, a somewhat arbitrary threshold of 10,000 steps per day has been held up as a general standard.[9] However, the complex health benefits of walking cannot be captured by a generic step count applied equally to everyone. Children must walk more steps per day than the average adult to achieve optimum health, while many older people cannot realistically walk 10,000 steps per day. The number of calories burned and operationally defined levels of physical activity are influenced by surface grade—for example, walking uphill is far more taxing on the body. For healthy adults (with great heterogeneity across the age range), some highly active individuals with the lowest mortality have a step count greater than 12,500.[10]

Translating physical activity, including walking, into an empirically verified measure of an individual's health benefit has been well established in the scientific literature.[11] Moderate walking (30 minutes per day most days) reduces the risk of all-cause mortality by 27 percent (risk ratio = 0.73), and vigorous walking (20 minutes per day three times a week) lowers the risk of all-cause mortality by 32 percent (risk ratio = 0.68).[12] These results reveal that walking farther in a shorter period of time (greater exertion) correlates with lower mortality. This benefit, however, peaks at about 4 miles.

One way to translate daily step count from a wearable sensor into an empirically verified health benefit metric is to first translate the level of physical activity into walking speed, then convert walking speed into calories burned, and finally convert

TABLE 2. Relative risk of mortality for a 65-year-old US male who is 5'7" and weighs 175 pounds, as a function of step count and walking speed.

Activity level	Walking speed (mph)	Mile time (min)	Stride length	Steps/mile	Calories burned	Calories/step	Relative risk
Minimal	1.0	60	2.17' (26.0")	2,433	198	0.0814	1.00
	1.5	40	2.25' (27.0")	2,347	150	0.0639	1.00
Some	2.0	30	2.29' (27.5")	2,306	125	0.0542	0.81
	2.5	24	2.33' (28.0")	2,266	111	0.0490	0.81
Moderate	3.0	20	2.42' (29.0")	2,182	105	0.0481	0.73
	3.5	17	2.50' (30.0")	2,112	105	0.0497	0.73
Vigorous	4.0	15	2.58' (31.0")	2,046	109	0.0533	0.67

Table adapted from M.F. Leitzmann et al., "Physical Activity Recommendations and Decreased Risk of Mortality," *Archives of Internal Medicine*, vol. 167, no. 22, 2007, pp. 2453–2460.

calories burned into an operationally defined physical activity measure that can be scientifically linked to observed mortality risk.

Linking physical activity to walking speed

Physical activity can be categorized in many ways—for example, as Table 1 shows, according to frequency and intensity. Here we focus on walking speed, which typically ranges from 3.1 to 5.6 mph; beyond that speed is a transition into running. This range can be partitioned to define three levels of physical activity: *some*, 3.10–3.93 mph; *moderate*, 3.94–4.77 mph; and *vigorous*, 4.77–5.60 mph. It is also possible to generate fairly precise estimates of walking distance as a function of time traveled, step count, self-reported height, and correlated step length as a function of height. For example, a person who is 5'7" has a typical stride length of 27.5 inches; a person of this height would therefore take 2,307 steps to travel one mile, and 10,000 steps by this person would translate into a distance traveled of 4.33 miles.

Converting walking speed into calories burned

The number of calories burned (CB) while walking is a function of walking speed, weight, surface grade, and time spent walking (www.shapesense.com/fitness-exercise/calculators/walking-calorie-burn-calculator.shtml). Assuming a 0 percent surface grade,

$$CB = [0.0215 \times KPH^3 - 0.1765 \times KPH^2 + 0.8710 \times KPH + 1.4577] \times (WKG \times T),$$

where KPH is walking speed in kilometers per hour, WKG is weight in kilograms, and T is time in hours. The CB formula varies as a function of surface grade ranging from –5 percent to +5 percent, and another formula applies to grades ranging from +6 percent to +15 percent. Many step counters have the capacity to measure average grade of surface traveled during a day.

Converting calories burned into mortality risk

Health scientists have developed various mathematical models to associate mortality risk with physical activity. Using "life tables" that estimate how long a typical person of a given age and gender and with basic measurable physical characteristics (weight, height, and so on) will live based on actuarial and demographic data, they can help people improve their mortality risk profile by measuring and incorporating data related to exercise and other behaviors. These same quantitative models can be applied to wearable sensor data to determine the decreased risk of mortality from walking, running, and other activities.

To illustrate how a measured level of physical activity from sensor data can be converted into health and longevity metrics, consider the example of a 65-year-old US male who is 5'7" and weighs 175 pounds. As Table 2 shows, the medical literature indicates that a male of this age walking at a vigorous pace of 4 mph (2,046 steps/mile) would travel 1 mile in about 15 minutes and burn 109 calories, reducing his risk of mortality by 33 percent if done consistently over a period of time; if the same person walked at 3 mph (2,112 steps/mile) for 20 minutes, he would burn about 105 calories and reduce his risk of mortality by 27 percent.[12] The frame of reference in both cases is a walking speed of 1 mph. The key point is that step counts uploaded daily from a wearable device can be converted into reliable long-term assessments of an individual's health and longevity. Aggregating such data from millions of device users could lead to a gold standard for such metrics.

TOWARD A HEALTH DATA ECONOMY

The proliferation of passively generated, real-time wearable sensor data opens the door to development of a *health data economy* that will offer consumers novel benefits derived from such data. Here we discuss a few examples of Lapetus-initiated efforts in this area.

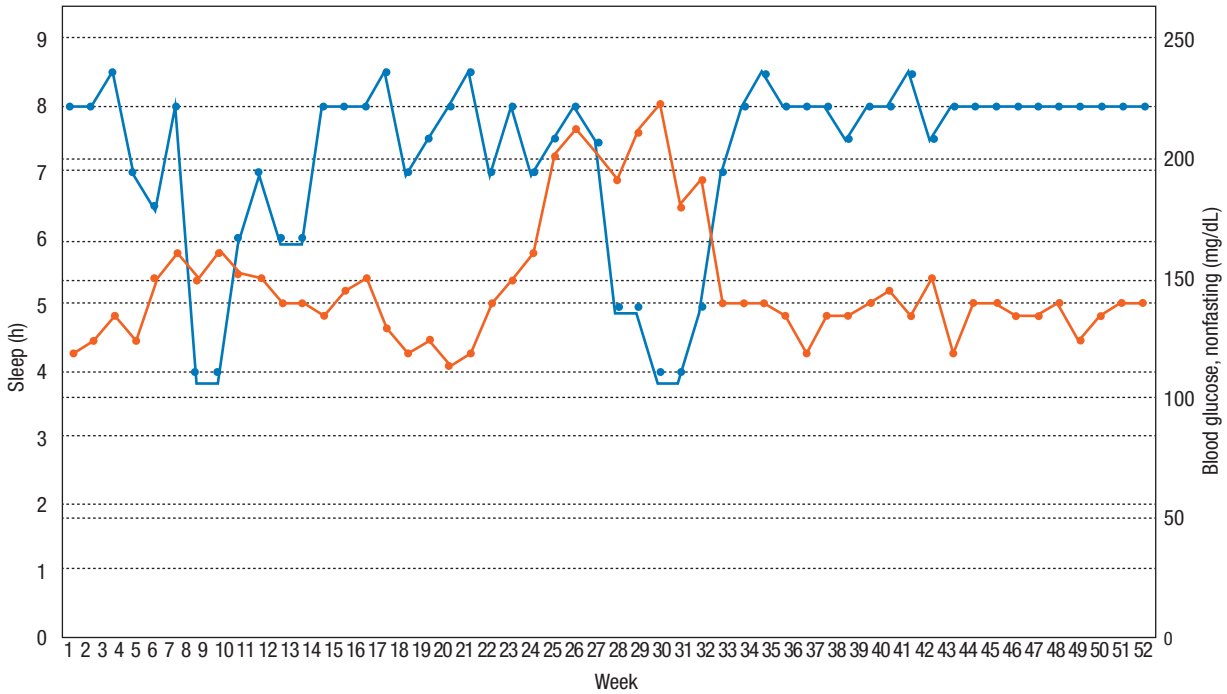


FIGURE 1. Hypothetical health chart showing sleep patterns and blood-glucose levels obtained from wearable sensors over a year. Such a chart can give a personal physician insight into hard-to-detect factors contributing to a patient's health problems.

Leveraging step-count data

Consider the use of a personal step counter as a self-motivation tool to improve one's health. Instead of simply monitoring the daily number of steps taken and congratulating you upon reaching some target, the device could provide more useful feedback—namely, specific health and longevity metrics directly associated with the steps taken. For instance, the 65-year-old male in the example above would not only receive daily updates of his step count but would also be informed that maintaining vigorous physical activity would reduce his risk of death by 33 percent and increase the probability of living an additional year by 25 percent. In addition to encouraging users to adopt a healthier lifestyle, such information could be used in planning for retirement and deciding whether to purchase long-term care insurance and an optimal age for doing so.

Longitudinal individual health data

Annual personal physicals, also known as physical health evaluations (PHEs), typically include the collection of blood and urine for analysis. These in-depth laboratory-screening procedures detect potentially serious medical conditions that patients can be unaware of, improve delivery of preventive services, and reduce patient worry. While extremely valuable, such assessments are only a snapshot of conditions within the body at a single moment. As such, they are ephemeral and impacted by behavioral and environmental conditions that existed at or just prior to the time the samples were collected.

What these tests fail to reveal is how patients have actually lived their life since the previous physical. Did they exercise regularly? Was their sleep pattern consistent and normal for someone their age? Did they experience periods of inactivity due to an illness or injury? Did their blood sugar or blood pressure ever fluctuate into an unhealthy range? By supplementing lab results with personalized longitudinal data obtained from wearable sensors, physicians would have a much fuller picture of their patients' current health status.

The process might work as follows. Your physician provides you with a wearable device, or you obtain one of many over-the-counter devices recommended for the same purpose, and register it online with the medical practice. The daily data generated by the device to help you monitor your health status or use as a motivational tool to enhance fitness is automatically uploaded to a longitudinal health database maintained by the practice for subsequent analysis.

Longitudinal health data is the gold standard for understanding and evaluating the relationship between various behavioral and environmental risk factors and health. Data from wearable devices can be used to generate charts that track all sorts of health barometers across time—sleep/wake patterns, blood pressure, blood sugar, physical activity, periods of sedentary behavior, and so on.

Figure 1 is a hypothetical example of a patient's health chart showing hours of sleep and blood-glucose levels over a year. Such a chart provides a previously unseen look at the patient's overall

health since the last annual physical and could prompt the physician to ask the patient during the personal health evaluation—for example, what caused an abnormal sleep pattern in February or why average blood-sugar levels rose to dangerous levels in June and July. With such data, the physician has a much better understanding of hard-to-detect factors contributing to potential health problems in the patient.

Longitudinal crowdsourced health data

Crowdsourced, or participatory, health studies[13] draw on data from social health networks such as PatientsLikeMe, smartphone health applications, wearable devices that collect health data, and companies that compile health data in different forms and for different purposes such as 23andMe, Quantified Self, and Genomera. The scientific vigor of crowdsourced health research varies widely, but it is clear that a new kind of research paradigm is emerging based on alternative sources of health data that citizen scientists and trained researchers alike are only now beginning to grasp.[14]

One way to ensure progress in this area is to create a longitudinal crowdsourced health database containing freely available anonymized data that meets well-defined standards. This would not only engage the public to accelerate health research, it would also create novel and innovative approaches to improving both individual and overall public health.

BLISS: Better Life and Income Scoring System

It is possible to unite personal health data with other numerically defined attributes, such as net worth and education level, to generate a more complete metric of an individual's relative risk than the FICO score and other credit risk metrics, which are limited to economic measures such as payment history and debt burden. Toward this end, Lapetus is launching the Better Life and Income Scoring System. BLISS weights and combines key financial indicators with longitudinal individual and crowdsourced health data, some of it derived from wearable sensors, to create a single, empirically verified metric—a BLISS score—ranging from 100 to 1,000. A higher BLISS score positively correlates with superior health *and* lower financial risk.

IT IS CLEAR THAT A NEW KIND OF RESEARCH PARADIGM IS EMERGING BASED ON ALTERNATIVE SOURCES OF HEALTH DATA.

A BLISS score would provide health and life insurance companies with a new, independently verified measure of risk to assess applicants as well as enable consumers to obtain the most favorable pricing for existing or future insurance—a win-win for the insurance industry. In addition, by yielding the most advanced estimates of health and longevity currently available, a BLISS score would give financial planners the ability to fine-tune a retirement plan to an individual's or family's unique attributes. Instead of relying on clients to determine how many years they expect to live in retirement—something most people simply cannot reliably answer—they could use the BLISS score to create a scientifically based plan. Long-term care insurance might not make sense for those with a high estimate of healthy life expectancy, while others' family history might suggest a high probability that such care will be needed.

We foresee the BLISS score becoming the basis of new forms of commerce in the coming health data economy. For example, health and life insurance companies could build different plans around various BLISS score ranges, while enterprises that aim to improve health, physical fitness, and financial well-being might target those with low or middle-range scores. In addition, companies might emerge to facilitate the collection of health data used to compute BLISS scores—for example, to pay consumers to register their wearable devices with them—as well as to develop novel types of sensors.

Wearable devices are being developed at breathtaking speed. The range of physiological attributes and bodily functions that such devices can monitor accurately and in real time is continually expanding, and it is only a matter of time before sensors are routinely implanted within our bodies—temporarily, or perhaps even permanently, becoming part of our identity. While it is impossible to

predict which devices, applications, and data aggregation companies will ultimately triumph in the marketplace, the emergence of a new health data economy driven by the flood of data from body area networks seems inevitable.

Sensors can now be plugged into a car to collect data on personal driving habits, enabling auto insurance companies to adjust premiums according to actual behind-the-wheel behavior instead of relying on actuarial data and rare events involving customers such as collisions and traffic citations. This technological advance benefits everyone: those who routinely drive safely are rewarded with lower premiums, auto insurance companies can create more equitable policies, and car manufacturers gain insights that can lead to vehicle safety and performance improvements.

We believe that smart health technologies will evolve along a similar path. People will learn more about their health and be educated on how to take better care of their own body; they will be motivated to do so not just through feel-good rewards but by meaningful financial incentives such as lower insurance premiums. Longitudinal data obtained from wearable and in-body sensors will likely lead to medical breakthroughs that improve individual as well as public health outcomes, reduce healthcare costs, extend longevity and enhance the quality of life, and engender new types of health data commerce. Issues of data privacy will no doubt arise with a health data economy: consumers should always be able to opt out and, more importantly, own their data—making it available to third parties only with permission.

The wearable device revolution is coming, and it will help launch a new era in public health and e-commerce that will benefit consumers and companies alike. It will be exciting to see how this revolution in health empowerment evolves over the coming years.

REFERENCES

1. B. Levin et al., "Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology," *CA: A Cancer J. for Clinicians*, vol. 58, no. 3, 2008, pp. 130–160.
2. S.S. Hall, "A Trial for the Ages," *Science*, vol. 349, no. 6254, 2015, pp. 274–1278.
3. "Wearable Technology Market Worth 31.27 Billion USD by 2020," press release, MarketsandMarkets, Dec. 2015; www.marketsandmarkets.com/PressReleases/wearable-electronics.asp.
4. Y. Hao and R. Foster, "Wireless Body Sensor Networks for Health-Monitoring Applications," *Physiological Measurement*, vol. 29, no. 11, 2008, pp. R27–R56.
5. S. Fox and M. Duggan, *The Diagnosis Difference, Part Three: Tracking for Health*, Pew Research Center, 26 Nov. 2013; www.pewinternet.org/2013/11/26/part-three-tracking-for-health.
6. R.T. Li et al., "Wearable Performance Devices in Sports Medicine," *Sports Health: A Multidisciplinary Approach*, vol. 8, no. 1, 2016, pp. 74–78.
7. Z. Zhao, S.A. Etemad, and A. Arya, "Gamification of Exercise and Fitness Using Wearable Activity Trackers," *Proc. 10th Int'l Symp. Computer Science in Sports* (ISCSS 16), 2016, pp. 233–240.
8. J.B. Wang et al., "Wearable Sensor/Device (Fitbit One) and SMS Text-Messaging Prompts to Increase Physical Activity in Overweight and Obese Adults: A Randomized Controlled Trial," *Telemedicine J. and e-Health*, vol. 21, no. 10, 2015, pp. 782–792.
9. Am. Heart Assoc., "Frequently Asked Questions about Physical Activity," updated 23 June 2015; www.heart.org/HEARTORG/Conditions/More/CardiacRehab/Frequently-Asked-Questions-About-Physical-Activity_UCM_307388_Article.jsp#.V9rbCfkrLRY.
10. C. Tudor-Locke and D.R. Bassett Jr., "How Many Steps/Day Are Enough? Preliminary Pedometer Indices for Public Health," *Sports Medicine*, vol. 34, no. 1, 2006; www.ncbi.nlm.nih.gov/pubmed/14715035.
11. D.E.R. Warburton, C.W. Nicol, and S.S.D. Bredin, "Health Benefits of Physical Activity: The Evidence," *CMAJ*, vol. 174, no. 6, 2006, pp. 801–809.
12. M.F. Leitzmann et al., "Physical Activity Recommendations and Decreased Risk of Mortality," *Archives of Internal Medicine*, vol. 167, no. 22, 2007, pp. 2453–2460.
13. M. Swan, "Scaling Crowdsourced Health Studies: The Emergence of a New Form of Contract Research Organization," *Personalized Medicine*, vol. 9, no. 2, 2012, pp. 223–234.
14. B.L. Ranard et al., "Crowdsourcing—Harnessing the Masses to Advance Health and Medicine, a Systematic Review," *J. General Internal Medicine*, vol. 29, no. 1, 2013, pp. 187–203.

ABOUT THE AUTHORS

S. JAY OLSHANSKY is the chief scientist of Lapetus Solutions and a professor in the School of Public Health at the University of Illinois at Chicago. A cofounder of the field of the biodemography of aging, he researches longevity forecasting and ways to slow human aging. Olshansky received a PhD in sociology from the University of Chicago. He is on the editorial board of numerous scientific journals and the board of directors of the American Federation for Aging Research, and is a Fellow of the Gerontological Society of America. Contact him at jay@lapetussolutions.com.

BRUCE A. CARNES is a senior scientist at Lapetus Solutions and an emeritus professor in the Donald W. Reynolds Department of Geriatric Medicine at the University of Oklahoma Health Sciences Center. A cofounder of the field of the biodemography of aging, he uses statistics to explore the factors that influence longevity as well as their relative importance, and to estimate upper limits for both the longevity of individuals and the life expectancy of populations. Carnes received an MS in population biology from the University of Houston and an MS in statistics and a PhD in theoretical ecology from the University of Kansas. An author of more than 150 academic papers and numerous book chapters and books, he is a member of the American Association for the Advancement of Science, the Gerontological Society of America, and Sigma Xi. Contact him at bruce@lapetussolutions.com.

YANG CLAIRE YANG is a senior scientist at Lapetus Solutions and a professor in the Department of Sociology and the Lineberger Comprehensive Cancer Center at the University of North Carolina at Chapel Hill (UNC). She specializes in the biodemography of aging, medical sociology, cancer epidemiology, social disparities in health, and quantitative methodology. Yang received an MS in statistics and a PhD in sociology from Duke University. She is a Faculty Fellow of UNC's Carolina Population Center and a member of the American Sociological Association, and has served on the editorial board of many social science journals as well as on the board of directors of the Population Association of America. Contact her at claire@lapetussolutions.com.

NORVELL MILLER is the president and chief business officer of Lapetus Solutions. A chartered financial analyst, he is the managing general partner of SEInteractive, a general partner of Covestco Seteura, and the managing director of EMS Financial. Miller currently serves or has served on the boards of several privately held companies including Allconnect, Arsenal Digital Solutions, BuildLinks, the MediaSpan Group, Pixel Magic Imaging, and VisionAIR. Previously, he cofounded and worked with numerous growth companies including DentalCare Partners, the Mobius Group, and Affordable Care. Miller received an MBA from Duke University. Contact him at norvell@lapetussolutions.com.

JANET ANDERSON is the chief marketing officer of Lapetus Solutions. She has more than 25 years of international sales and marketing experience in Europe, Latin America, and North America. Anderson spent most of her career with AEGON, one of the world's largest insurance companies, and most recently served with ReMark, a subsidiary of SCOR, as CEO of their North and Latin American business. Anderson received an executive MBA with an emphasis on international business from Loyola University Maryland. Contact her at janet@lapetussolutions.com.

HIRAM BELTRÁN-SÁNCHEZ is a senior scientist at Lapetus Solutions and an assistant professor in the Department of Community Health Sciences at the Fielding School of Public Health at the University of California, Los Angeles, where he is also a researcher at the California Center for Population Research. His research interests include the demography of health and aging, with a particular focus on Latin American countries; biodemographic patterns of health in adult populations; and the development and application of demographic methods to investigate health inequalities. Beltrán-Sánchez received a BS in actuarial sciences from Universidad Nacional Autónoma de México, an MS in mathematics from Northern Arizona University, and a PhD in demography from the University of Pennsylvania. He has published numerous articles on health and aging and collaborated with researchers and institutions around the world. Beltrán-Sánchez cofounded the Latin American Mortality Database, the largest data repository of mortality from 19 countries in Latin America. Contact him at hiram@lapetussolutions.com.

KARL RICANEK JR. is the CIO and chief data scientist of Lapetus Solutions and a professor in the Department of Computer Science at the University of North Carolina Wilmington (UNCW). A leading expert on facial analytics and facial recognition, he is director of the UNCW Institute for Interdisciplinary Identity Sciences (I3S) and founded its world-renowned Face Aging Group research lab. Ricanek has authored more than 80 scientific articles and multiple book chapters in the areas of machine learning, face recognition, and facial analytics; is affiliated with multiple international scientific working groups on these subjects; and holds numerous patents. He received a PhD in electrical engineering from North Carolina A&T State University. Ricanek is a Senior Member of IEEE and a member of ACM and Intelligence Advanced Research Projects Activity (IARPA). Contact him at karl@lapetussolutions.com.

COVER FEATURE **SMART HEALTH AND WELL-BEING**

Key Success Factors for Smart and Connected Health Software Solutions

Noel Carroll, University of Limerick

Healthcare delivery is being transformed by technology that personalizes, tracks, and manages patient information across devices. Smart and connected health will develop safer and more effective, efficient, equitable, and user-centered services through pervasive computing innovations.

Societal and demographic changes, coupled with economic challenges, are transforming how we deliver healthcare in our communities. Due to growing populations and medical advances, healthcare demands will inevitably outgrow medical professionals' capabilities to deliver safe, quality care in a timely manner. These factors have given rise to smart and connected health (SCH), a comprehensive sociotechnical model for managing healthcare through software solutions.

According to Gondy Leroy and her colleagues, technological advances in healthcare have encouraged the development of new technologies that drive connectivity across the healthcare sector—apps, gadgets, and systems that personalize, track, and manage care using just-in-time information exchanged through various patient and community connections.[1] This paradigm shift heavily emphasizes the process of software development in supporting SCH innovation. It has also contributed to a shift in healthcare practice, highlighting our growing reliance on and trust in software to support healthcare decisions. However, failure to correctly align software with healthcare needs can have serious—and potentially fatal—consequences.

MY RESEARCH

As an experienced researcher in applied SCH solutions and healthcare innovation, I'm often asked by industry professionals which reoccurring and emerging factors contribute to the success of SCH software innovations. To explore this question, I conducted a series of semi-structured interviews with 10 experts, including applied researchers, consultants, academics, and industry partners. Many of the participants have also been involved in national and international research projects and case studies. Drawing on these interviews and case study synopses, I synthesize here my findings regarding the key success factors (KSFs) for SCH software innovations.

Collaborating with healthcare institutions and companies over the years has allowed me to develop rich

0018-9162/16/$33.00 © 2016 IEEE

insights into and strong empathy for healthcare professionals and customers. These groups often cite the barriers and requirements involved in validating the problem–solution fit required for successful SCH innovations. This convinced me that I needed to identify the key ingredients of a practical yet innovative solution that maps healthcare and software needs while identifying technology's opportunities and mitigating its potential risks.[2]

My research, guided by design thinking,[3] yielded novel outcomes that successfully support the ability to innovatively bridge healthcare needs and software requirements. Design thinking is a human-centered, prototype-driven process that can be applied to any product, service, or business design.[4] It not only guided my identification of reemerging SCH success factors but also uncovered deeper insights on how to align healthcare needs with software requirements to address customers' pain points. In the context of SCH, a pain point is a real or perceived problem that causes discomfort, annoyance, embarrassment, or frustration and impacts a person's well-being.

To successfully launch a new SCH software solution, we must map the KSFs of healthcare software companies to guide their system analysts, designers, and software developers in addressing pain points.

WHY SMART AND CONNECTED HEALTH?

Over the past decade, software companies have increasingly focused on healthcare, yet medical errors are still inevitable because of the fragmented nature of medical information. However, there is growing interest in identifying methods that will transform healthcare from, for example, moving away from reactive care to proactive and preventive interventions and from clinic- to citizen-centered practices by empowering people through greater access to healthcare information.

SCH is supported by an interoperable digital infrastructure that facilitates the exchange of data and knowledge. It optimizes the use of rich and real-time data sources to support evidence-based health and wellness decisions. Thus, through the use of software tools and technologies, SCH extends healthcare services and processes beyond traditional healthcare boundaries, delivering more convenient and personalized healthcare services. SCH empowers individuals to manage their own health and treatment plans, often facilitated by integrated intelligent systems (for example, wearables) that monitor and support the maintenance of a desired health status.

SCH's ultimate goal is to develop safer and more effective, efficient, equitable, and user-centered health and wellness services through pervasive computing innovations. Thus, SCH software innovation contributes to the coverage and quality of healthcare services, improved health outcomes, reduced costs, and improved quality of life. Improved decision-making tools can increase the likelihood of saving lives, reducing costs, and ensuring a better quality of life before, during, and after treatment.

WHY KEY SUCCESS FACTORS?

CEOs often ask, "Within an SCH context, what worked well in the past and how do we incorporate this into our product/service design and development process?" Adopting KSFs requires a process with top-down support from management to champion new innovation and bottom-up participation from staff to develop healthcare software solutions.

KSFs are those functions, activities, or business practices—typically influenced by the market and defined and viewed by end users—that are critical to sustaining the technology provider–user relationship. Thus, KSFs can act as a scaffold to guide software development and result in value co-creation through software skills, processes, and systems. Organizations can develop KSFs in terms of the core competencies and capabilities that extend internal activities and improve practices and quality functions. Aligning these competencies with the KSFs increases the value and success of both the relationship and the healthcare software product/service value proposition—to both the organization's and users' benefit. In essence, adopting KSFs makes the software innovation lifecycle for SCH solutions less risky.

DESIGN THINKING

Design thinking is a formal process that captures pain points and influences the design and development of end products and services. This technique was extremely useful in guiding my research on the identification of healthcare needs and how they align with software requirements.[6] Such guidance is vital because healthcare software has much at stake, most notably patient safety. Design thinking moves beyond the gathering of software requirements and is not constrained by preconceptions of isolated software solutions.[5] As Figure 1 shows, design thinking has five key phases:

- empathizing with users to fully understand their experiences,

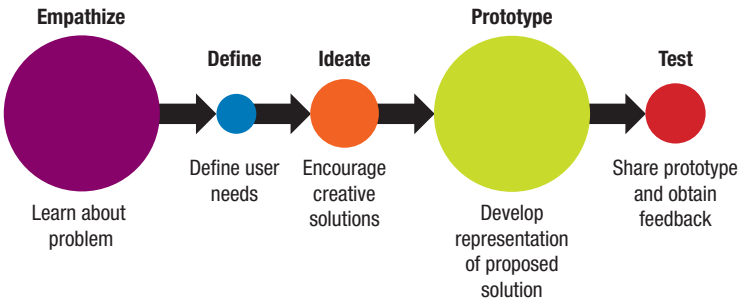


FIGURE 1. The five phases of design thinking.[6] This process helped identify factors leading to successful smart and connected health innovations and uncovered ways to better align healthcare needs with software requirements.

- defining a wide variety of possible SCH software solutions,
- ideating creative SCH solutions,
- prototyping ideas into tangible form, and
- testing to refine and examine the effect of SCH solutions.

This process fosters a learning lifecycle about solutions and carefully bridges our understanding of healthcare needs with the software design process. It also helps identify reoccurring themes that contribute to SCH software's success. Stanford University's d.school's (Institute of Design) design-thinking method, which is applied to SCH innovation in Table 1, is one simple, effective model.

The fascinating, sometimes meandering process of initiating a healthcare solution all the way through to testing the product and launching it into the market is influenced by many factors. Through the "stories of individuals," we can identify key lessons learned for guiding software innovation and use these to determine the KSFs.

MY RESULTS

Drawing on my research, I present the 10 KSFs that organizations should incorporate during SCH innovation development to ensure marketplace success.

Innovation champions

Improving current healthcare practices requires a sustainable healthcare innovation culture that truly empowers staff members to instigate changes. Such changes must have remarkable, rather than incremental, effects on healthcare to differentiate an organization from its competitors. Every project needs an innovation champion and a clinical champion, both of whom are bold, creative, and willing to push the boundaries to redefine healthcare. Therefore, one KSF is having innovation champions with the innate ability to identify SCH opportunities while understanding the impact of misaligning healthcare needs with software requirements.

Innovation champions play a key role during the empathize phase and are tasked with securing buy-in from the key decision makers who will execute and lead SCH solutions. Innovation champions are typically familiar with using empathy mapping to uncover what people think, feel, see, say, do, and hear and to identify core healthcare pain points. Innovation champions must also collaborate with clinical champions, usually consultants, who can implement and test proposed changes in a healthcare context. Innovation champions must adhere to feasible timeframes for delivering safe, high-quality software solutions. They are also key to influencing the political will to implement SCH solutions despite institutional and personnel resistance. Organizations should encourage idea generation among their staff, such as through an internal innovation program, to ensure that staff have the opportunity to share their ideas and champion change using SCH innovation.

Clear definition of success

To achieve healthcare goals using SCH solutions, organizations must communicate what "success" means (both internally and externally) and establish clear objectives. All staff members must understand what the end game is so that they can contribute to its achievement. This is a KSF because it encourages staff, particularly analysts and developers, to prioritize software requirements while targeting lucrative markets. In fact, the design thinking's empathize phase allows software developers to inform, reflect, understand, and decide on appropriate software development actions as co-defined by innovation champions and management. Typically, systems analysts would clearly define the healthcare problem of interest and then align their software processes with their definition of success. However, all employees must support an SCH innovation vision and the organization's short-, medium-, and long-term software-development strategies. One research participant, a software developer, explained that their organization uses the following—adapted from Robert Charette's[8] "why software fails" list—as a checklist of SCH "success requirements":

- realistic and articulated project goals;
- accurate estimates of needed resources;
- well-defined system requirements;
- good reporting of the project's status;
- managed risks;
- good communication among customers, developers, and users;
- use of mature technology;

TABLE 1. Stanford d.school's design-thinking process applied to smart and connected health (SCH) software solutions.

Design-thinking phase	Description	Contribution to SCH success
Empathize	Close collaboration with end users, patients, software developers, and management teams identifies the key healthcare problems and needs ("pain points"). This is achieved by observing, engaging with, and earning the trust of patients to learn about their experiences.	Supports innovation champions with ethnography skills, who can be sympathetic to patient needs and achieve rich insights by prompting deeper questions about patients' day-to-day experiences, for example, about taking a medication.
Define	Having gathered information on the core pain points, it is important to define these issues by contextualizing and synthesizing user needs; that is, clearly defining the problem and the associated needs.	Defines users' priorities in terms of healthcare needs, which influence the SCH software and system design process, and defines success (that is, addressing users' needs). For example, we might learn that patients experience social isolation, forgetfulness, and poor eyesight and need medication reminders.
Ideate	Software engineers generate new solution ideas by creating and prototyping low-resolution software prototypes through a sprint software-development cycle.	Takes action to develop and adopt human-centric software solutions that address patient needs; generates ideas around a theme, for example, a medication homecare assistant software solution.
Prototype	The development of software solutions with healthcare technology providers commences by ensuring software developers are clear on the product goal, identifying key factors to be tested, and monitoring patient feedback.	Develops an SCH solution within the sprint software-development cycle by collecting meaningful feedback in the testing phase. The feedback should provide a representative snapshot of the market reaction to the product. Examples include conceptualizing included features, such as machine learning, face recognition, personal healthcare assistant robots, and so on.
Test	The solution must be tested against the key healthcare needs by continuously observing and refining the prototype, learning about the user experience, and comparing multiple prototypes to reveal possible suppressed healthcare needs.	Refrains from explaining the SCH prototype to users, allowing them to interpret its functionality; tests the prototype's usability and usefulness, often prompting the need to address questions and concerns (for example, feedback might include that a personal assistant prototype offers a companion and greater sense of a security—which reduces social isolation—and provides friendly reminders that improve medicine adherence and quality of care).

- ability to handle the project's complexity;
- great development practices;
- good project management;
- management of stakeholder politics; and
- management of commercial pressures.

Identifying what could contribute to software failure can be incorporated into organizations' success definition and helps avoid the pitfalls associated with bringing SCH products to market.

Human-centric healthcare needs and software adoption

This KSF borrows techniques from design thinking's ideate phase and biodesign research's problem–need–solution approach.[7] It enables system analysts to immerse themselves in patients' lives and more deeply understand their healthcare needs. This removes the threat of technologists making assumptions about or guessing what patients need. It also supports organizations' software innovation developments and reduces the risks associated with developing products unfit for market.

Rather than one-size-fits-all technology solutions, particular focus should be placed on the use, usability, and usefulness of SCH and fit-for-purpose solutions in relation to the target market's healthcare needs. Therefore, SCH solutions must address real-world healthcare needs and be informed and validated through research rather than just assuming that the market will embrace the solutions. The empathize phase achieves this by investigating whether an SCH solution meets the end users' cultural needs and facilitates smoother software adoption. By addressing human-centric needs and software adoption, organizations can maximize their coverage and the quality of their healthcare software and services.

Knowledge of healthcare end users

This KSF targets the need to identify and understand those who are willing to pay for SCH software products and services, including healthcare institutions, healthcare professionals, caregivers, family members, and governments. The users ultimately co-create an SCH innovation's healthcare value. SCH users might interact with stakeholders who monitor wellness and contribute to patient treatment

by supporting the health workforce, tracking diseases, and monitoring public health through various software solutions. In addition, healthcare providers (HCPs) might play dual roles—as end users of the software solutions and as technology advocates recommending new software solutions to patients or healthcare institutions.

Therefore, certain core, intangible HCP views, attitudes, perceptions, and interactions (including cognitions and emotions) are critical to a healthcare software innovation's success. Adopting a multistakeholder engagement tool that ensures collaboration before, during, and after the software development lifecycle is extremely useful. Developing the software solution is only half the battle—marketing it to the right audience directly affects the organization's success and the solution's widespread adoption.

Software team and healthcare project management

Much of a solution's success hinges on the software team and coordinated healthcare project-management activities. For example, motivating a software team to deliver a high-quality solution on time is critical. This KSF thus encapsulates the transition from healthcare needs to innovative software solutions. Both the software product and the team supporting its development must be efficient and reliable to achieve the organization's vision of success.

A well-documented, transparent project management plan must be established. This plan must clearly identify key resources and software quality milestones to avoid loosely managed and unfocused development. A daily scrum meeting can ensure that the entire team is aware of operational developments and potential challenges within the organization. Specific governance protocols and adherence policies must be defined and documented to avoid deviating from what was originally agreed upon with the innovation champions and other key stakeholders. The development team must be compliant through timely reporting (regular, concise, and frequently reviewed) and regular training to ensure quality standards are maintained. Because poor management can have potentially devastating health consequences, SCH project management necessarily differs from software development for noncritical evolving systems.

SMART AND CONNECTED HEALTH IS A COMPREHENSIVE SOCIOTECHNICAL MODEL FOR MANAGING HEALTHCARE THROUGH SOFTWARE SOLUTIONS.

Evidence-based healthcare value co-creation

Value co-creation is critical to success and builds trust within the marketplace. This process involves testing and demonstrating the value derived from SCH software solutions through various sprint software developments. In most cases, value co-creation clearly and compellingly supported with data such as behavioral changes or clinical data is required for a software solution's value realization and wider adoption.

This process enables users to undertake specific tasks on multiple devices and experience single-point healthcare interactions. It also allows the organization to observe and understand how value is unlocked from the users' perspective. Much of this is achieved through extensive research that ideally proves the benefits of a specific SCH solution. More important, value is most often realized through improved health outcomes, reduced costs, and improved quality of life. Measuring value requires both quantitative (for example, a health value scorecard) and qualitative data (for example, patient satisfaction ratings). An organization can continue to improve software by creating new healthcare experiences and applying feedback throughout the innovation lifecycle.

Infrastructure and interoperability

Offering a safe and secure SCH software solution to users is a KSF that ensures that users avail of proper technical infrastructures, which is a key concern for HCPs and end users, particularly regarding data privacy and security. The testing phase informs systems analysts of the infrastructure and interoperability requirements for the proposed SCH solution. Interoperability requirements include whether systems need to be connected or are ready for emerging paradigms, such as the Internet of Things.

In addition, because most services are available online and generate real-time analytics, sufficient

attention must be paid to API scalability and tethering problems. If existing solutions are in place, the new SCH software solution must seamlessly integrate with these systems without jeopardizing the quality and safety of healthcare services. Quality and trust marks can be introduced to demonstrate that organizations meet high-quality software standards.

Software as a medical device (regulation and standards)

There are growing concerns among SCH organizations regarding intertwining healthcare, technology, and medical device regulations and standards. During the early prototype phase, organizations must consider their solutions' classification in terms of intended use and categorization as software as a medical device.[9] User and organizational concerns often focus on data security, ownership, and accuracy; user questions often include, "Where is my data?," "Who owns my data?," and "Can I trust this device?" Thus, assuring users that their data is secure and encrypted and can be safely transferred and stored is of vital importance, particularly in relation to healthcare and well-being decision-making tasks.

In addition, policies should govern how solution features are incorporated into the final product; that is, linking back to empathy and enabling decision makers to formalize new protocols and policies that ease implementation and adoption. Regulations and standards should not be viewed as hurdles that hamper SCH innovation. Rather, organizations should view regulations and standards as "rules of the game" that ensure their compliance so that they can focus on the innovation strategy. Using regulations and standards to guide the SCH innovation cycle also conserves resources by ensuring compliance at an early stage. In most cases, a product's compliance with medical device regulations gives the organization a competitive advantage in the crowded wellness marketplace.

From smart and connected health data analytics to knowledge

Healthcare data analytics has two core functions. First, from an organizational perspective, the technology provider must manage data to ensure that the software solution meets users' requirements, improves the quality and safety of care, and empowers people to manage their own care and well-being. Second, the data captured by various SCH software solutions must be accurate and calibrated to be meaningful to users. Failure to provide these two key functions will reduce the SCH software solutions' value proposition and customer retention rates.

User health self-management is an emerging—and necessary—feature of modern healthcare and wellness services. Therefore, it is important to clearly indicate who manages information—for example, whether a company will develop specific algorithms to mine data and generate diagnostic or predictive analytics—and how the data will be presented so as to be meaningful yet not cause information overload on smaller devices. Thus, as part of the much-hyped data-analytics field, data management and quality control are critical in SCH and the software market share. It is important for organizations to strike the right balance between capturing the required data and educating end users on their health status or progress. This also unlocks new knowledge and value from healthcare data, which strengthens cross-fertilization between research and industry to accelerate future innovations.

Software-driven educational experience

The final KSF is providing software-driven educational experiences. Internally, it is vital that organizations learn how their solution influences behavioral change in patients. Organizations must also encourage the key stakeholders involved in the software solution's execution to identify what worked well so that they can improve and evolve their software product/service development lifecycle and enhance their market opportunity analyses. Externally, it is critical that users learn how to empower themselves by using the SCH solution. Organizations must also try to ensure that users buy into future software products. Similar to social media, organizations need to identify methods that encourage users to embrace SCH as part of their everyday lives. The users' healthcare learning experience can take them on a journey of discovery to co-create value. This is critical if users are to manage their own healthcare and understand their progress with lifestyle interventions or treatments. Positive educational experiences also encourage users to promote the product in the marketplace.

Software innovation is poised to transform our access to and participation in our own healthcare and well-being. This challenge is being shaped by associated transformations to what it means to be healthy and to how we make informed healthcare choices. SCH software solutions are facilitating these transformations.

ABOUT THE AUTHOR

NOEL CARROLL is a research fellow at the Applied Research for Connected Health Technology Centre, University of Limerick. His research interests include healthcare software systems and innovation management, particularly techniques to assess and visualize technology's influence on healthcare service networks and novel methods to effectively communicate healthcare service process change. Carroll received a PhD in computer science from the University of Limerick. Contact him at noel.carroll@lero.ie.

However, healthcare is a complex field and, as most software practitioners are aware, software can fail!

The mismatch between healthcare needs and software solutions presents significant risks to healthcare as a result of "improper or unsafe use of technology."[10] My research helps address this by supporting SCH software development and evaluating the benefits of healthcare technology.[11] Organizations should incorporate the 10 KSFs to successfully address healthcare needs and introduce software solutions to the market. The clinical and commercial advantages of such an outcome-led design process turn directly into customer benefits.

This work has three clear implications for the SCH community:

- *Practical:* the 10 KSFs provide a roadmap or "scaffold" for software companies.
- *Theoretical:* this work bridges the interdisciplinary nature of SCH by explaining the benefits of adopting techniques such as design thinking.
- *Educational:* this research offers rich insight on how practitioners and academics can adopt a more informed view of SCH innovation throughout the software development process.

In the future, I plan to develop specific metrics for the KSFs as well as methods of incorporating them into the maturity phases of innovation. Doing so will allow organizations to assess their readiness to develop a new product, evaluate which skills need additional focus or training, and decide which elements of the SCH solution should be outsourced.

ACKNOWLEDGMENTS

Special thanks to the research participants for sharing their SCH insights and experiences. The Applied Research for Connected Health Technology Centre (www.arch.ie), an initiative jointly funded by Enterprise Ireland and the Industrial Development Agency, supported this work.

REFERENCES

1. G. Leroy, H. Chen, and T.C. Rindflesch, "Smart and Connected Health," *IEEE Intelligent Systems*, vol. 29, no. 3, 2014, pp. 2–5.
2. R. Owen et al., "A Framework for Responsible Innovation," *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, R. Owen, J. Bessant, M. Heintz, eds., Wiley, 2013, pp. 27–50.
3. T. Brown, "Design Thinking," *Harvard Business Rev.*, vol. 86, no. 6, 2008, pp. 84–91.
4. G.O. Matheson et al., "Leveraging Human-Centered Design in Chronic Disease Prevention," *American J. Preventive Medicine*, vol. 48, no. 4, 2015, pp. 472–479.
5. C. Giardino et al., "Key Challenges in Early-Stage Software Startups," *Proc. 16th Int'l Conf. Agile Software Development* (XP 15), 2015, pp. 52–63.
6. N. Carroll and I. Richardson, "Aligning Healthcare Innovation and Software Requirements through Design Thinking," *Proc. Int'l Workshop on Software Eng. in Healthcare Systems* (SEHS 16), 2016; doi:10.1145/2897683.2897687.
7. P.G. Yock et al., *Biodesign: The Process of Innovating Medical Technologies*, Cambridge Univ. Press, 2015.
8. R.N. Charette, "Why Software Fails," *IEEE Spectrum*, vol. 42, no. 9, 2005, pp. 42–49.
9. N. Carroll and I. Richardson, "Software-as-a-Medical Device: Demystifying Connected Health Regulations," *J. Systems and Information Technology*, vol. 18, no. 2, 2016; doi:10.1108/JSIT-07-2015-0061.
10. D.W. Meeks et al., "Exploring the Sociotechnical Intersection of Patient Safety and Electronic Health Record Implementation," *J. American Medical Informatics Assoc.*, vol. 21, no. e1, 2014, pp. e28–e34.
11. P. O'Leary et al., "Untangling the Complexity of Connected Health Evaluations," *Proc. Int'l Conf. Healthcare Informatics* (ICHI 15), 2015, pp. 272–281.

COVER FEATURE **SMART HEALTH AND WELL-BEING**

Using Smart Homes to Detect and Analyze Health Events

Gina Sprint and Diane J. Cook, Washington State University
Roschelle "Shelly" Fritz, Washington State University Vancouver
Maureen Schmitter-Edgecombe, Washington State University

Smart homes offer an unprecedented opportunity to unobtrusively monitor human behavior in everyday environments and to determine whether relationships exist between behavior and health changes. Behavior change detection (BCD) can be used to identify changes that accompany health events, which can potentially save lives.

Sensors have become ubiquitous: they are ambient, embedded in smartphones, and even worn on the body. Data collected from these sensors can form a time series in which each data point is paired with an associated timestamp. For human health, this time-series data is valuable when detecting and analyzing changes associated with seasonal variations, new lifestyle choices, new job situations, and so on. Furthermore, it can be useful when monitoring behavior changes that are related to health events such as cancer treatment, an injury, or onset of a chronic medical condition. Automatically tracking behavior changes from sensor data can help create a deeper understanding of the behavioral impact of these health events. Similarly, detecting these changes can alert individuals and their caregivers about potential health concerns.

Through behavior change detection (BCD), we can analyze the behavioral impact of health events using information about a resident's everyday behavior collected through smart-home sensors—without imposing any restrictions on the resident's routines. To test this method, we collected and analyzed data from ambient sensors placed in smart-home environments over multiple years and labeled the data with its corresponding activities using automated activity recognition (AR). To track changes in routine behavior, we quantitatively compared two or more time periods, or windows, of activity-labeled data. If the two time windows contained notably different activity information, this could indicate a significant behavior change.

We identified health events for three smart-home residents based on medical records and monthly interviews with the participants. We compared the

data surrounding the health event with their baseline normal data to determine if a significant behavior change had occurred, and then described the nature of the change. A clinician analyzed the corresponding behavior change to validate the change and explain the relationship between the health event and the behavior change.

> CLINICAL STUDIES SUPPORT A RELATIONSHIP BETWEEN DAILY BEHAVIOR AND COGNITIVE AND PHYSICAL HEALTH.

Clinical studies support a relationship between daily behavior and cognitive and physical health.[1] Most of the prior work in this area utilized wearable data to correlate home-based movement with health measures,[2] although smart-home data has been used to analyze mobility and time out of the home with respect to cognitive and physical health.[3,4] Our own earlier work showed that smart-home data can be analyzed over time to predict performance on cognitive health-assessment tests.[5] We hypothesize that the relationship between sensed behavior and health events can also be observed and analyzed using smart-home data, which has not yet been examined. Results from our case studies indicate that smart-home and machine-learning technologies can be used to understand the behavioral impacts of health events and to provide individuals with information that can point to possible health concerns.

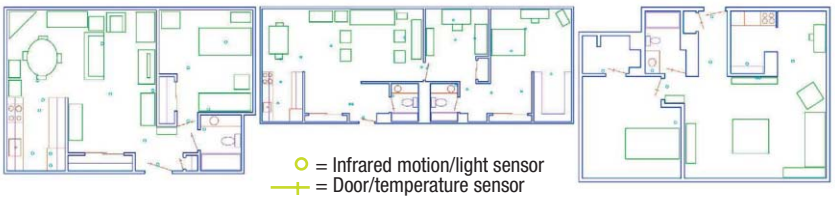
COLLECTING AND LABELING SMART-HOME DATA

For our case studies, we collected data in everyday home environments that we converted to smart homes using the CASAS "smart home in a box."[6] The three homes in this study were single-resident apartments, each with at least one bedroom, a kitchen, and a dining area. Figure 1 shows the floorplans, sensor positions, and sample labeled sensor data. Using CASAS, we equipped these homes with combination motion/light sensors on the ceilings and door/temperature sensors on cabinets and doors. The sensors continuously and unobtrusively monitored the residents' daily activities by sending text message–type updates, or sensor events, whenever they sensed a state change (for example, from "door closed" to "door open" or from "no motion" to "motion"). The CASAS middleware collected these sensor events and stored them in a relational database.

Once the sensor data was collected, we labeled each sensor event with the corresponding activity using the CASAS-AR algorithm.[7] Let $A = \{a_1, a_2, \ldots, a_T\}$ be the set of all activities. Given features $x \in \Re^d$ extracted from a sequence of sensor events ending at time t, AR's challenge is to map x onto a value $a \in A$, indicating the activity that occurred at time t. These labels provide a vocabulary for expressing and analyzing the sensed behavioral patterns. AR algorithms have been designed for wearable, phone, home, video, and other sensors using machine-learning techniques that range from naive Bayes classifiers and decision trees to more complex models including Gaussian mixture models and conditional random fields.[8,9]

AR is particularly well suited for this type of analysis because it does not require the sensor data to be pre-segmented into distinct activity sequences. Instead, it labels sensor events ("activities") in real time as the events occur. To do this, it moves a dynamic-size sliding window over the sensor events and extracts features x describing the current window of information. The features include the sensor event time of day, the size of the sliding window, the event count for each sensor within the window, the time elapsed for each sensor since its most recent event, the most recent event location and sensor identifier, and the sensor generating the most events in the previous two windows.

Training data for CASAS-AR was provided by external annotators who looked at one month of data and utilized both the house floorplan and resident information to generate corresponding ground-truth activity labels.[10] In addition, sensor identifiers were replaced by more general location-based descriptors, as shown in Figure 1. Using this method, CASAS-AR learns an activity model based on training data from multiple smart-home sites and can thus generalize for application to new smart homes with no training data. Although CASAS-AR has been tested with a number of classifiers including naive Bayes, decision trees, hidden



2014-06-15 03:38:28.094897 M009 ON	2014-06-15 03:38:28.094897 BedroomMotion ON Sleep
2014-06-15 03:38:29.213955 M009 OFF	2014-06-15 03:38:29.213955 BedroomMotion OFF Sleep
2014-06-15 03:38:17.814393 M015 ON →	2014-06-15 03:38:17.814393 BathroomMotion ON Bed-Toilet
2014-06-15 03:38:58.584179 M015 OFF	2014-06-15 03:38:22.584179 BathroomMotion OFF Bed-Toilet
2014-06-15 03:39:17.814393 M009 ON	2014-06-15 03:39:17.814393 BedroomMotion ON Sleep

FIGURE 1. Smart-home floorplan and sensor layout for three testbeds. Study participant 1's smart home (SH1) is shown at left, with three door/temperature sensors and 10 motion/light sensors; participant 2's smart home (SH2) is at center, with five door/temperature sensors and 23 motion/light sensors; and participant 3's smart home (SH3) is at right, with three door/temperature sensors and 10 motion/light sensors. Sample raw sensor data was converted to use generalized sensor identifiers and automatically labeled by the CASAS-AR algorithm with corresponding activity labels (bottom).

Markov models, and conditional random fields, a decision tree achieved the best performance. In this study, we analyzed the following activities: Hygiene, Sleep, Bed-Toilet, Eat/Drink, Enter/Leave home, Relax, and Work. For these activities in the three smart-home testbeds described in this article, CASAS-AR achieved a recognition accuracy of 98 percent using three-fold cross-validation.

DETECTING AND ANALYZING BEHAVIOR CHANGE

To determine whether a significant change in behavior occurred at the time of a health event and to analyze the nature of the behavior change, we introduced methods to quantify the amount of change in activity patterns between two windows of time-series activity data that were sampled by smart-home sensors and labeled by CASAS-AR. Let X denote a sample of time-series data where each day's data is expressed by extracted activity features, $X = \{x_1, x_2, x_k\}$, and let W be a window of n days such that $W \subseteq X$. For this study, activity features consist of the amount of time spent on each activity for a particular day and the sensor density of each activity (measured as the number of sensor events) for a particular day. We also collected the total amount of movement that occurred in the home for the day, expressed as the total distance traveled by the resident. These features were shown in earlier work to provide insight on behavior patterns that correlate with cognitive and physical health of smart-home residents.[10]

BCD compares two windows of data, W_i and W_j, within time series X. In this article, the windows are one week in length ($n = 7$) and BCD compares a baseline window ($i = 1$, the first week in our data subset representing normal behavior for the resident) with each subsequent window ($j = 2,3,\ldots$). We utilized three change-detection methods, each of which provided a slightly different perspective on the data comparison. Additionally, the more methods that detected a significant change, the greater the evidence for a behavior change.

Method 1: RuLSIF

The first method is a nonparametric approach that determines the amount of change between two time-series samples by comparing the probability distributions of the two samples. Instead of estimating the probability distributions (which is computationally costly), we directly estimated their ratio. Relative Unconstrained Least-Squares Importance Fitting (RuLSIF)[11] represents one such approach that estimates the ratio using the Pearson divergence dissimilarity measure.

RuLSIF does not explicitly provide a method to determine a cutoff threshold for Pearson divergence values that are considered significant change scores. To address this issue, we introduced a change significance test based on intrawindow variability and outlier detection. The proposed change significance test utilizes the day-to-day variability in human behavior patterns.[12] For a change between two windows to be significant, the magnitude of change (interwindow change) should exceed the day-to-day variability within each window (intrawindow change).

To compute the significance of the change score (CS) between two windows, we first generated a list of all possible daily change scores (DCSs) within each window—there are $2 \times$ Combination(n,2) such scores. Next, we applied boxplot-based outlier detection to see if the CS was an outlier when compared to the distribution of intrawindow DCSs. Here, an outlier is defined as an observation that appears to be inconsistent with other observations. To determine this, we computed the interquartile range (75th–25th percentile). CS values outside the 75th percentile + $1.5 \times$ interquartile range are considered outliers and thus significant. Advantages of this proposed significance test are that it is nonparametric and that it can be computed based on any window size.

Method 2: sw-PCAR

Our Permutation-based Change Detection in Activity Routine (PCAR) approach[5] was originally designed to analyze changes in longitudinal smart-home data. Here, we adapted the original approach to handle smaller windows of activity-labeled data. The resulting small-window

PCAR (sw-PCAR) algorithm breaks each day within the window into nonoverlapping hour-long time intervals. Each time interval has a corresponding probability distribution over the activities that occur at that time. For sw-PCAR, the days within two windows W_i and W_j are averaged to yield aggregate windows $\hat{W}_i$ and $\hat{W}_j$. Next, we computed the CS using the symmetric Kullback–Leibler (KL) divergence distance between the activity probability distributions in $\hat{W}_i$ and $\hat{W}_j$. Finally, we computed the significance of the distance value CS by concatenating data from windows $\hat{W}_i$ and $\hat{W}_j$ into one window W. All the time intervals within W were randomly shuffled and then split into two new subwindows, and the KL distance was computed for this permuted window pair.

This shuffling procedure was repeated N times to produce an N-length vector V of KL distances. If N is large enough, the corresponding set of KL distances forms an empirical distribution of the possible permutations of activity data for the two windows. sw-PCAR computed change significance by comparing CS to the permutation vector V using boxplot-based outlier detection, as we did with RuLSIF. If identified as an outlier of V, the CS is reported as significant.

Method 3: Virtual classifier

Our final method utilizes a binary classifier to detect and explain behavior change. This type of virtual classifier (VC) for change analysis was first proposed by Shohei Hido and his colleagues.[13] For this approach, feature vectors from window W_i were labeled with a positive class and feature vectors from window W_j were labeled with a negative class. VC trains a decision tree to learn a boundary between the virtual positive and negative classes. The resulting average prediction accuracy based on k-fold cross-validation is represented as p_{VC}. If a significant change exists between W_i and W_j, the average classification accuracy p_{VC} of the learner should be higher than the accuracy expected from random noise: p_{rand} = 0.5, the binomial maximum likelihood of two equal-length windows.

To determine the significance of the change between two windows, we used the inverse survival function of a binomial distribution to determine a critical value, $p_{critical}$, at which n Bernoulli trials were expected to exceed p_{rand} at α = 0.05 significance. If $p_{VC} > p_{critical}$, a significant change exists between windows W_i and W_j. If a change significance test concludes that the CS is significant, we would like to investigate possible reasons for the change. Typically, this requires computing features that summarize the data and provide a meaningful context for change and applying change tests for those specific features. One of the advantages of the VC, however, is that the decision-tree learner's output explains the source of change without relying on statistical tests. Upon detecting a significant change, the decision tree is retrained on the entire dataset and inspected to reveal the features that are the most valuable in discriminating between the two windows of data.

WE COLLECTED DATA FROM SMART HOMES WITH OLDER ADULT RESIDENTS USING SENSORS TO UNOBTRUSIVELY MONITOR THEIR DAILY BEHAVIOR.

ANALYZING BEHAVIORAL IMPACT OF HEALTH EVENTS

We collected data across multiple years from smart homes with older adult residents. For each study participant, we also recorded health events with their date and event type based on medical records and monthly interviews. Here, we describe three of these health events and use the case studies to illustrate BCD's use.

Case 1: Radiation treatment

Case 1 focuses on an 86-year-old female resident living in a smart-home testbed whom we refer to as SH1 (refer to Figure 1 for the smart-home floorplans). Three months into the data collection, the participant was diagnosed with lung cancer and started radiation treatment during W_{10}. We hypothesized that radiation treatment would have an observable and quantifiable impact on the participant's behavior. To validate this hypothesis, we used BCD to compare a one-week baseline of smart-home activity data (W_1) with two other weeks. The first comparison was with another pre-event week, namely the week immediately following the baseline (W_2). The second

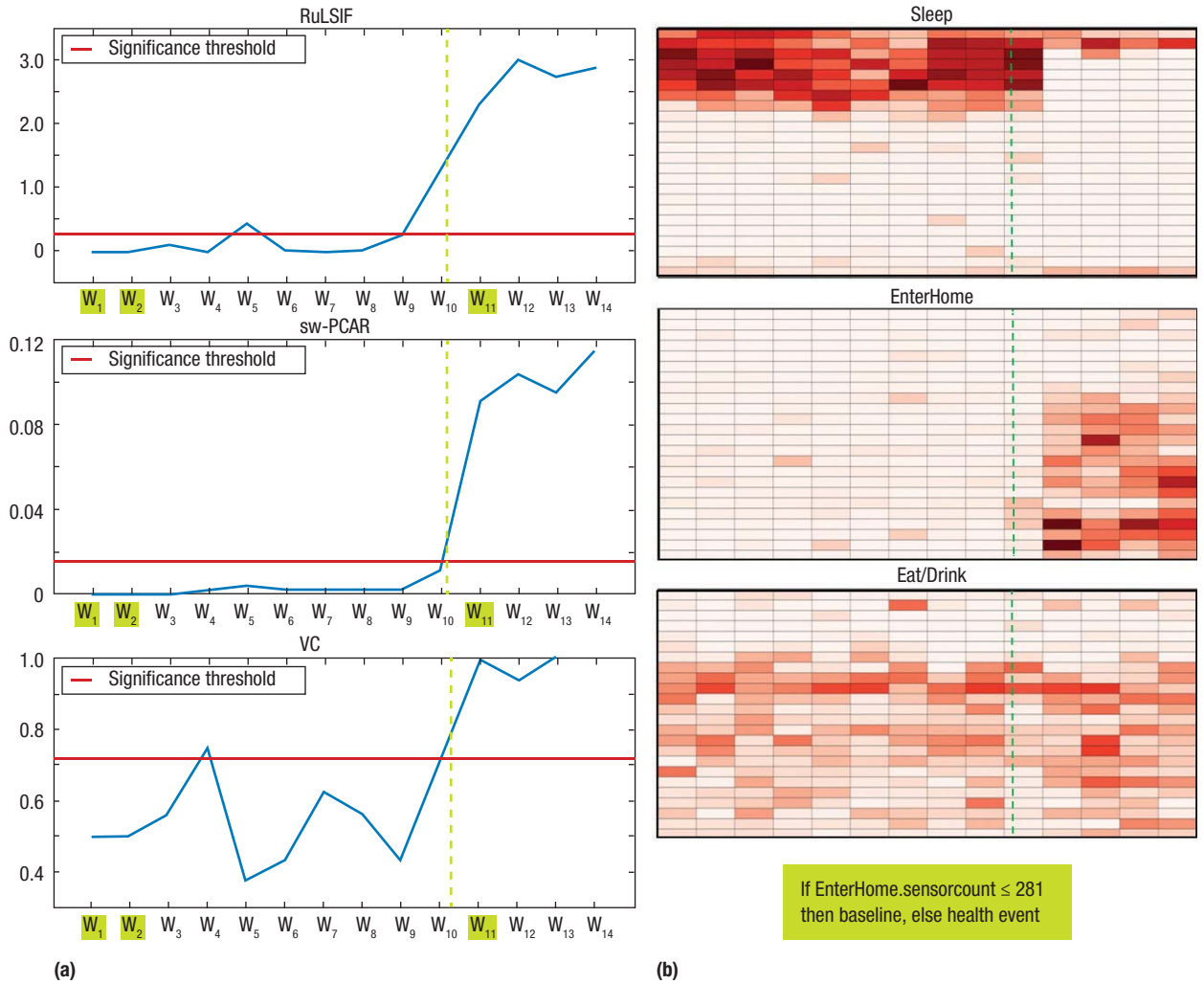


FIGURE 2. Results of SH1 health event analysis. (a) Overall change scores were plotted using Relative Unconstrained Least-Squares Importance Fitting (RuLSIF), small-window Permutation-based Change Detection in Activity Routine (sw-PCAR), and a virtual classifier (VC), comparing each week with the baseline week (W_1). Values above the red line show significant changes. (b) Density maps for selected activities Sleep, EnterHome, and Eat/Drink are plotted for the same time period. Darker colors in the density maps indicate more time spent on the activity during that hour of the day. The top-level rule generated using VC is highlighted in the box, indicating the activity feature that best discriminates the baseline week from the health event week. In each plot, the green dashed line indicates the occurrence of the health event.

comparison was with the first full week during which SH1 underwent radiation treatment (W_{11}).

Figure 2 illustrates results from applying each change-detection method, as well as associated activity density maps for SH1. Density maps were used in prior work to visualize levels of movement in the home.[4] Our activity density map is a heat map that visualizes the amount of time spent on a particular activity as a function of a 24-hour clock (*y*-axis), aggregated over one week (*x*-axis). The darker the color, the more time was spent on the activity during that particular hour of the day in the corresponding week.

As the density maps in Figure 2 show, the participant's level of sleep decreased once treatment started, and the number of times she left and returned home increased. Possible explanations for this are increased trips out of the home for appointments or visits from family and caregivers. Another impact of the treatment was an increased number of trips to the kitchen to eat or drink. These more frequent kitchen trips are consistent with the observation that radiation treatment increases thirst, resulting in a patient drinking more liquids throughout the day.[14]

Table 1 summarizes the CSs using the three BCD techniques described in this article. Scores were computed between two normal activity weeks (W_1 and W_2) and between a normal activity week and a week during the

TABLE 1. Change scores for smart-home residents.

Study participant (SH)	Method	W_1/W_2 (baseline)	W_1/W_{event} (health event)
SH1	Relative Unconstrained Least-Squares Importance Fitting (RuLSIF)	−0.017	2.298*
	Small-window Permutation-based Change Detection in Activity Routine (sw-PCAR)	0.001	0.091*
	Virtual classifier (VC)	0.500	1.000*
SH2	RuLSIF	0.010	3.315*
	sw-PCAR	0.004	0.042*
	VC	0.438	1.000*
SH3	RuLSIF	0.000	0.000
	sw-PCAR	0.000	0.001
	VC	0.500	0.750*

*Significant results

health event (W_1 and W_{11} for SH1 and SH2; W_1 and W_8 for SH3). For RuLSIF and sw-PCAR, larger values indicate greater change and values close to zero indicate no change. In the case of VC, values close to 0.5 indicate no change and values close to 1.0 indicate large change. Significant results are indicated with an asterisk.

For participant SH1, the behavior changes during radiation treatment are detected by each of the change-detection methods and the results are significant. The nature of the greatest change is highlighted by the decision tree that VC generates. As shown in Figure 2, the top-level feature is the number of sensor events that are related to an EnterHome activity. The number of times the participant (or a visitor) enters the home was larger during radiation treatment, with a great enough increase for this event to discriminate between baseline behavior and health event behavior.

Case 2: Insomnia

Case 2 is a 91-year-old female smart-home resident, referred to as SH2. During the data collection period, SH2 was diagnosed with insomnia. To measure the impact of this health event on her sleep and other routine activities, we used BCD to compare two weeks of normal behavior (weeks W_1 and W_2) and one week of baseline behavior with a week surrounding the insomnia diagnosis (weeks W_1 and W_{11}). The change scores are summarized in Table 1 and indicate that significant changes in overall routine are detected by all three methods.

Figure 3 shows that changes occurred not only during week W_{11} but also in the days leading up to the health event, and persisted for weeks following the insomnia diagnosis. The density maps in Figure 3 also show that the amount of sleep did decrease during this period. The behavior change also impacts relaxation, which is time spent in a favorite chair or couch with little movement and possibly napping. These relaxation periods occur during normal sleep hours but also throughout the day. In addition, the number of trips outside the home decreased during this time.

The VC actually found the corresponding decrease in EnterHome events to be the main discriminating feature between baseline and health event weeks. On the other hand, if we look slightly earlier at week W_{10}, VC again detected a significant change from the baseline week, and the main discriminating feature is the total movement in the home throughout the day (measured as distance traveled in the home). This factor could be considered when examining possible reasons for insomnia or the impact of decreased sleep.

Case 3: Fall

The third case is an 80-year-old female, referred to as SH3, living in a smart-home testbed. During the data collection period, SH3 fell in her home. She said her right leg hurt for several days afterward and "slowed her down." To analyze the impact of this health event, we compared data collected at baseline (W_1) with the following week, which also contained normal activity and no health events (W_2). We also compared W_1 with the week containing the health event (W_8).

As the results in Table 1 indicate, this health event had a subtler impact

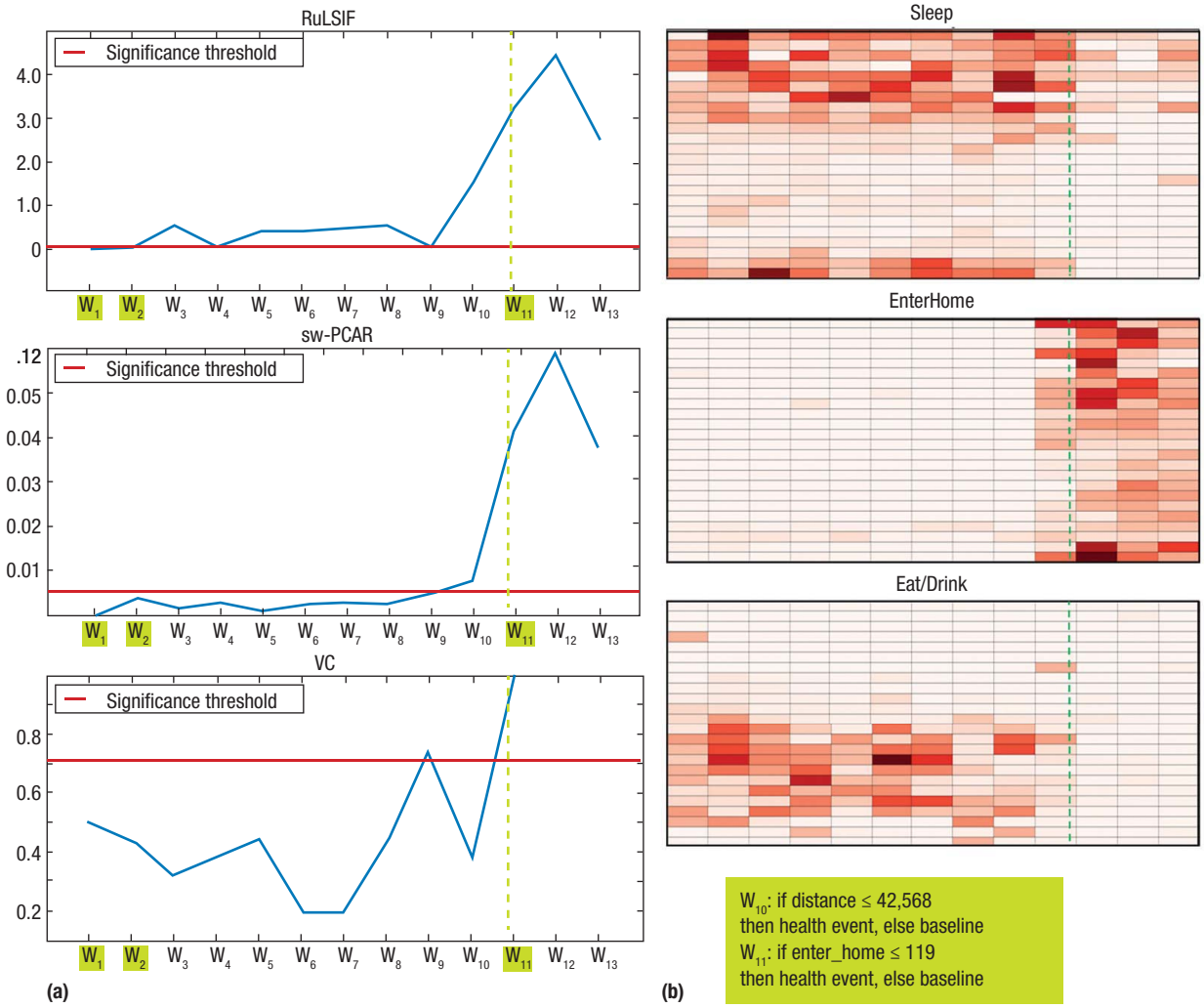


FIGURE 3. Results of SH2 health event analysis. (a) Overall change scores are plotted using RuLSIF, sw-PCAR, and VC for baseline week W_1 and health event week W_{11}. Values above the red line show significant changes. (b) Density maps for selected activities Sleep, Relax, and EnterHome for the same time period. The VC-generated rule is shown in the box. In each plot, the green dashed line indicates the occurrence of the health event.

on behaviors, at least those that could be detected by ambient smart-home sensors. RuLSIF and sw-PCAR detected almost no change between weeks W_1 and W_2 or between weeks W_1 and W_8. Only the VC method found the change during the health event week. As the VC-generated rule indicates, the difference is primarily detected based on the total distance that the individual traveled throughout the home on a daily basis. The decrease in movement is consistent with the observation that the injured leg caused SH3 to slow down. As the density plots indicate, there appears to be less impact on other routine activities such as sleep and bed/toilet transitions. There is an apparent slight decrease in trips out of the home, but not enough to be significant.

From these three case studies, we found that the ability to detect the behavioral impact of health events depends on the nature of the health event itself. Some events impact multiple activities including sleep, eating, and trips out of the home, whereas other events have a more localized impact. Detecting the actual health event occurrence (for example, a fall) and its impact might require additional, more sensitive sensors to be placed in the home or on the body. Systematically comparing different BCD window sizes might also provide insights on the typical duration of behavior changes that could be associated with different types of health events. Future work should analyze all of the BCD-detected changes to determine the broader spectrum of events that elicit changes, such as failed sensors or visitors in the home.

Detecting behavior changes associated with health events is valuable for researchers who want to better understand the relationship between health and behavior. These insights could also help care providers

SMART HEALTH AND WELL-BEING

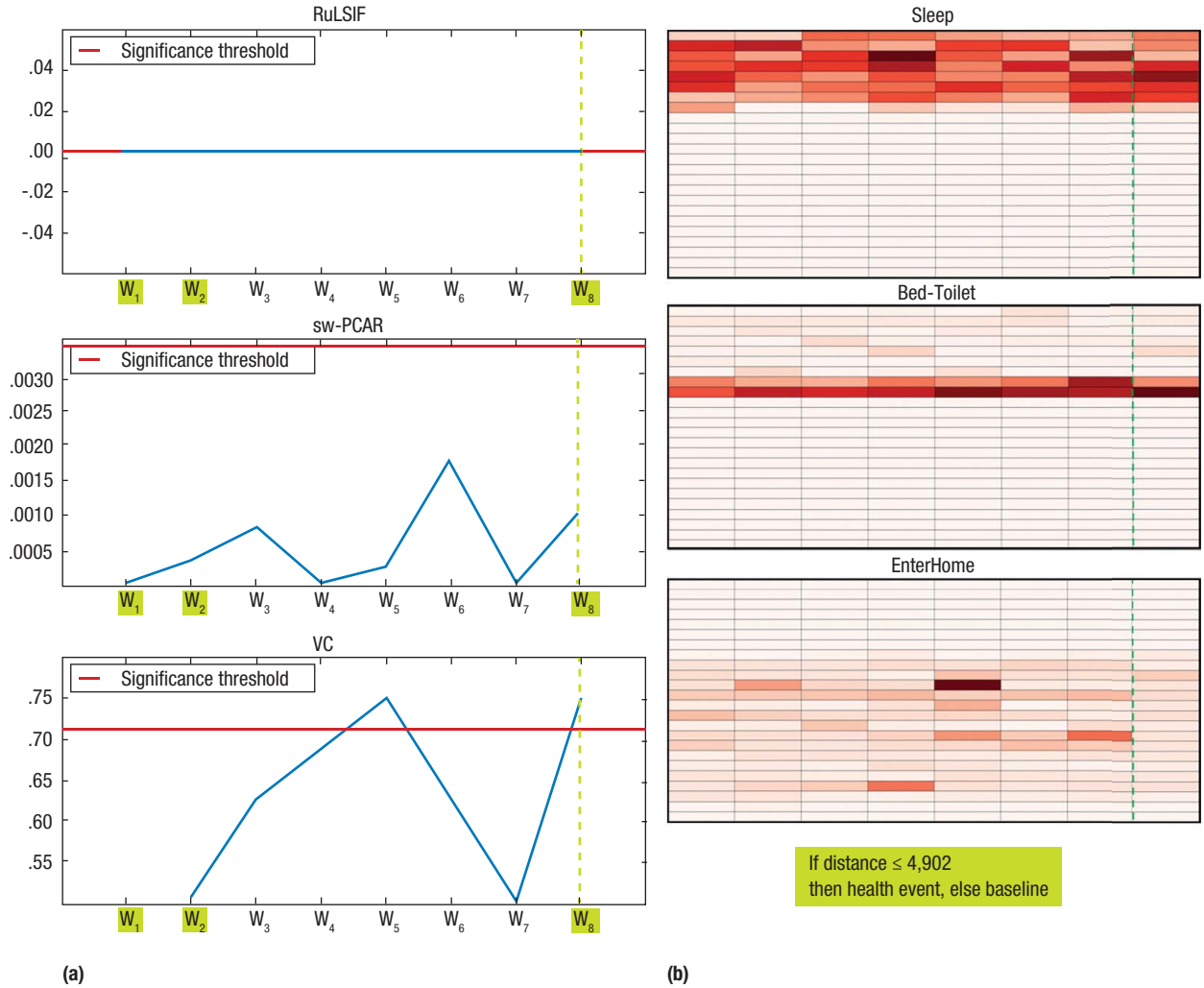


FIGURE 4. Results of SH3 health event analysis. (a) Overall change scores are plotted using RuLSIF, sw-PCAR, and VC for baseline week W_1 and health event week W_8. Values above the red line show significant changes. (b) Density maps for selected activities Sleep, BedToilet, and EnterHome are plotted for the same time period. The VC-generated rule is shown in the box. In each plot, the green dashed line indicates the occurrence of the health event.

respond to the needs of individuals who are experiencing health changes. An algorithm such as BCD can periodically look for changes in behavioral routine and alert the individual and his or her caregiver to them, as they can indicate changes in cognitive or physical health. Because BCD can analyze any type of sensor data, our continued research will adapt these methods to analyze smartphone and wearable data as well as data collected in smart homes.

ACKNOWLEDGMENTS

This work is funded in part by National Science Foundation grant 0900781 and National Institutes of Health grant R01EB015853.

REFERENCES

1. I.M. Lee et al., "Effect of Physical Inactivity on Major Non-Communicable Diseases Worldwide: An Analysis of Burden of Disease and Life Expectancy," *The Lancet*, vol. 380, no. 9838, 2012, pp. 219–229.
2. H.H. Dodge et al., "In-Home Walking Speeds and Variability Trajectories Associated with Mild Cognitive Impairment," *Neurology*, vol. 78, no. 24, 2012, pp. 1946–1952.
3. S. Robben, M. Pol, and B. Kröse, "Longitudinal Ambient Sensor Monitoring for Functional Health Assessments," *Proc. 2014 ACM Int'l Joint Conf. Pervasive and Ubiquitous Computing* (UbiComp 14), 2014, pp. 1209–1216.
4. S. Wang, M. Skubic, and Y. Zhu, "Activity Density Map Visualization and Dissimilarity Comparison for Eldercare Monitoring," *IEEE Trans. Information Technology in Biomedicine*, vol. 16, no. 4, 2012, pp. 607–614.
5. P.N. Dawadi, D. Cook, and M. Schmitter-Edgecombe, "Modeling Patterns of Activities Using Activity Curves," *Pervasive and Mobile Computing*, vol. 28, 2016, pp. 51–68.
6. D.J. Cook et al., "CASAS: A Smart

ABOUT THE AUTHORS

GINA SPRINT is a clinical assistant professor in the School of Electrical Engineering and Computer Science at Washington State University. Her research interests include wearable computing, machine learning, and technology applications for healthcare. Sprint received a PhD in computer science from Washington State University. Contact her at gsprint@eecs.wsu.edu.

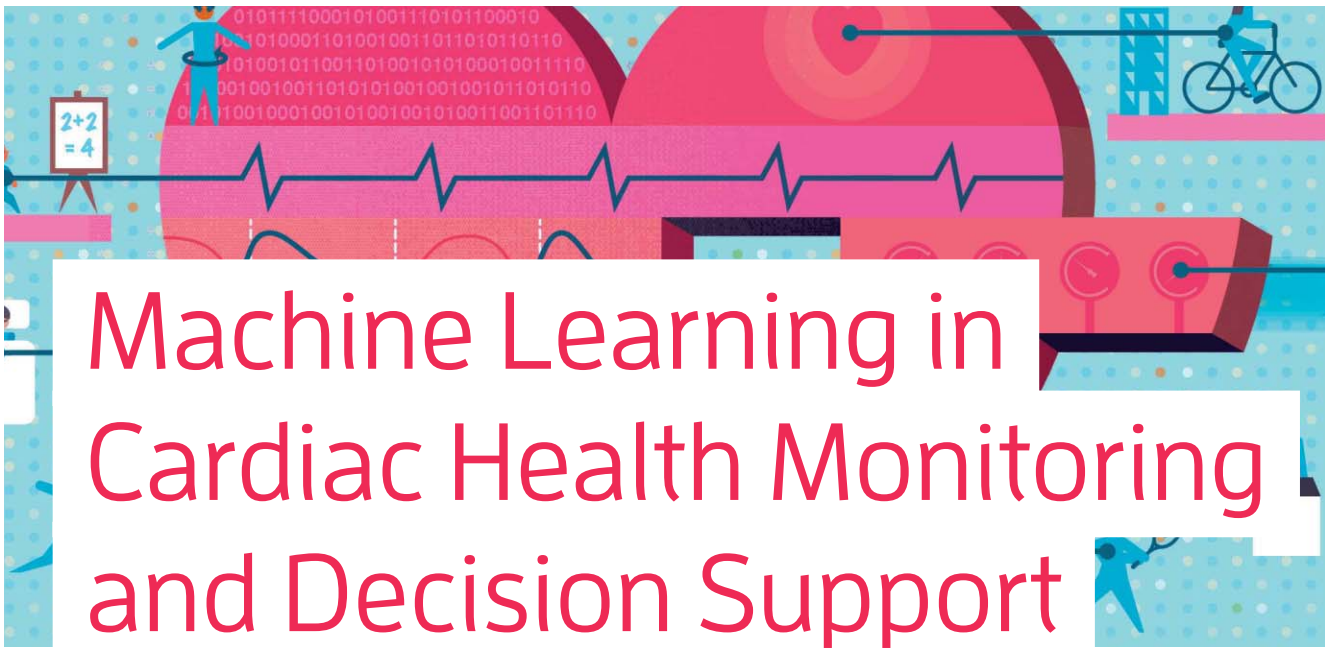
DIANE J. COOK is a Huie-Rogers Chair Professor in the School of Electrical Engineering and Computer Science at Washington State University. Her research interests include machine learning, data mining, and smart environments. Cook received a PhD in computer science from the University of Illinois. She is an IEEE Fellow. Contact her at cook@eecs.wsu.edu.

ROSCHELLE "SHELLY" FRITZ is an assistant professor in the College of Nursing at Washington State University Vancouver. Her research interests include the application of technology in the delivery of healthcare and human–computer interactions. Fritz received a PhD in nursing from Washington State University. Contact her at shelly.fritz@wsu.edu.

MAUREEN SCHMITTER-EDGECOMBE is a professor in the Department of Psychology at Washington State University. Her research interests include clinical and cognitive neuropsychology as well as assistive technologies for the aging population. Schmitter-Edgecombe received a PhD in clinical neuropsychology from the University of Memphis. Contact her at schmitter-e@wsu.edu.

Home in a Box," *Computer*, vol. 46, no. 7, 2013, pp. 62–69.
7. N.C. Krishnan and D.J. Cook, "Activity Recognition on Streaming Sensor Data," *Pervasive and Mobile Computing*, vol. 10, 2014, pp. 138–154.
8. A. Bulling, U. Blanke, and B. Schiele, "A Tutorial on Human Activity Recognition using Body-Worn Inertial Sensors," *ACM Computing Surveys*, vol. 46, no. 3, 2014, article no. 33.
9. L. Chen et al., "Sensor-Based Activity Recognition," *IEEE Trans. Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, 2012, pp. 790–808.
10. P. Dawadi, D.J. Cook, and M. Schmitter-Edgecombe, "Automated Clinical Assessment from Smart Home-Based Behavior Data," *IEEE J. Biomedical and Health Informatics*, vol. 20, no. 4, 2016, pp. 1188–1194.
11. S. Liu et al., "Change-Point Detection in Time-Series Data by Relative Density-Ratio Estimation," *Neural Networks*, vol. 43, 2013, pp. 72–83.
12. R. Refinetti, G.C. Lissen, and F. Halberg, "Procedures for Numerical Analysis of Circadian Rhythms," *Biological Rhythm Research*, vol. 38, no. 4, 2007, pp. 275–325.
13. S. Hido et al., "Unsupervised Change Analysis Using Supervised Learning," *Advances in Knowledge Discovery and Data Mining*, LNCS 5012, T. Washio et al., eds., Springer, 2008, pp. 148–159.
14. P. Beach et al., "Relationship Between Fatigue and Nutritional Status in Patients Receiving Radiation Therapy to Treat Lung Cancer," *Oncology Nursing Forum*, vol. 28, no. 6, 2001, pp. 1027–1031.

COVER FEATURE **SMART HEALTH AND WELL-BEING**



Machine Learning in Cardiac Health Monitoring and Decision Support

Shurouq Hijazi and Alex Page, University of Rochester
Burak Kantarci, University of Ottawa
Tolga Soyata, SUNY Albany

Portable medical devices generate volumes of data that could be useful in identifying health risks. The proposed method filters patients' electrocardiograms (ECGs) and applies machine-learning classifiers to identify cardiac health risks and estimate severity. The authors present the results of applying their method in a case study.

As personalized medicine becomes increasingly more sophisticated and affordable, portable medical devices have become ubiquitous and monitoring applications have begun to blend a range of functions. AliveCor, for example, offers an inexpensive smartphone electrocardiogram (ECG) attachment that can sample an individual's ECG, calculate real-time statistics, and share the recording with a physician.

In aggregate, portable medical devices generate data at a much higher rate than conventional systems, which can overwhelm medical personnel who must review reports for many patients. However, the data also presents scientists and engineers with the opportunity to create health-monitoring and decision-support systems that enhance and personalize healthcare. For example, decision-support systems based on machine learning (ML) can ease the review burden by filtering noise, errors, and irrelevant information so that the data reviewed contains only relevant clinical markers. ML algorithms learn patterns within the data, which serve as the basis for predictions about patient health. A machine can look through millions of reports and medical records to identify previously unknown drug interactions.[1] Such algorithms can significantly improve diagnostic accuracy, healthcare quality, and patients' quality of life.

ML algorithms have many potential applications in smart health. To explore one of these, we developed a method that filters data from long-term ECG recordings of patients with Long QT Syndrome (LQTS), uses ML to identify circadian patterns that signal risk of symptoms such as cardiac arrhythmia, and estimates the

0018-9162/16/$33.00 © 2016 IEEE

severity of that risk. LQTS is a disorder that primarily affects ion channels in heart muscle cells, allowing abnormal electrical activity to occur that can lead to sudden and dangerous arrhythmias. We tested our method using four classification methods against a database of 434 24-hour ECG recordings. We also explored how our method might scale with the volume of medical data. Scalability is rapidly becoming critical in medical studies. Analyzing massive amounts of data not only promotes a deeper understanding of the mechanisms that cause diseases but also makes personalized treatment possible. Such analyses can lead to breakthroughs in relating genes to diseases as well as providing the basis for treatment oriented to a particular patient's lifestyle and genetic makeup.

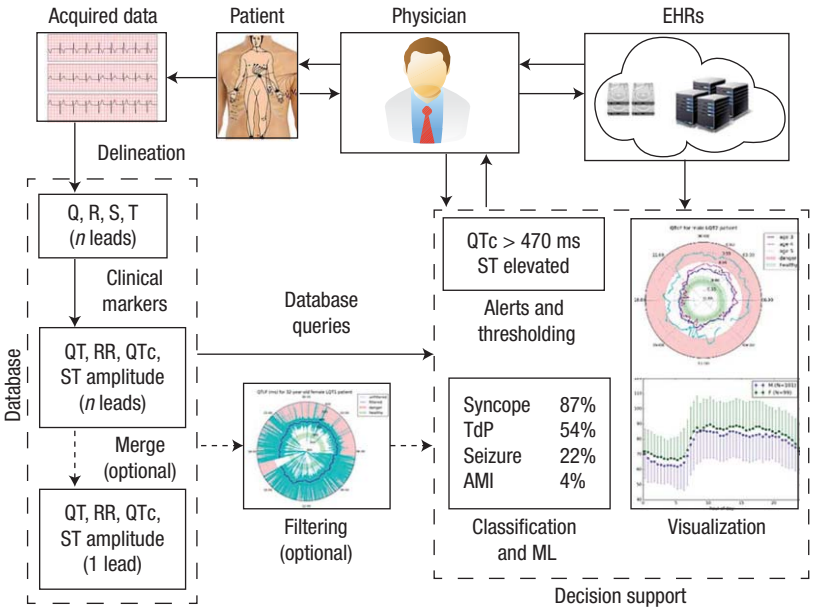


FIGURE 1. Conceptual workflow of a remote health-monitoring and decision-support system. Data from a patient at a remote location is acquired, preprocessed, and used to provide decision support in several forms. Visualization provides simple summaries to the physician or other clinician without any recommendations. Alerts are triggered by more urgent events, such as the violation of an established threshold. Classification of the patient's condition is based on the results of machine learning (ML), which involves comparing the patient to existing electronic health records (EHRs). AMI: acute myocardial infarction; Q, R, S, T: waves that indicate cardiac electrical state (on an electrocardiogram [ECG]); QT, RR, QTc, and ST: intervals in the cardiac electrical cycle, also measured on an ECG; TdP: torsades de pointes, a cardiac arrhythmia.

AN ML-BASED SYSTEM

We envision incorporating our method in a remote health-monitoring system that can provide feedback and decision support to a clinician. The system would use devices that acquire data through the Internet of Things and are connected to a cloud-based decision-support system.[2] The technological components of such a system are within reach, and advanced devices for acquiring medical data are becoming commercially available.[3] Sophisticated and powerful ML algorithms are already well understood and accessible.[4]

However, the human brain has unmatched reasoning abilities, so the physician is still the most important part of any medical decision-support system. Thus, the goal of our envisioned system is to provide physicians or other clinicians with concise, relevant information that can increase their diagnostic efficiency and accuracy.

Work flow

Figure 1 is a conceptual diagram of the workflow for a healthcare system that stores patient data electronically. After preprocessing and filtering patient data, the system stores it as an electronic health record (EHR). Each EHR gradually enriches the database, which will improve the accuracy of future ML results. A large database with many patients' records might not be as useful as a database with fewer patients but more information on each patient.

When many patients' records are aggregated and analyzed, steps must be taken to protect the individuals' privacy. Most protected health information (PHI), such as names and birthdays, can be removed from records in compliance with the Health Insurance Portability and Accountability Act (HIPAA) without detriment to the data mining process.[5] However, in some cases, it would be desirable to obtain more detailed information about certain patients from their physicians—an impossibility because it would violate HIPAA. Consequently, regulation, not just technology, can limit the acquisition of needed data.

Even after removing identifying information, what remains could be combined to statistically reveal a patient's identity. On one hand, PHI information such as age, gender, race, and genetic disorders, is critical to developing an effective decision-support system. On the other, including too much information on the wrong computer system risks violating HIPAA. Applications can also create privacy violations because some require explicitly protected information such as a patient's voice print[6] or city of residence. Researchers must keep these restrictions in mind during all stages of a study.

Decision support

An effective ML-based healthcare system capitalizes on the computer's vast computational capability and the physician's reasoning ability. Both machine and physician are looking for patterns, but the physician cannot analyze every heartbeat of every patient or be familiar with every disease's nuances. The machine can do all these tasks and then present its conclusions to the physician for confirmation.

Support types. As Figure 1 shows (see box at lower right), we envision three types of decision support: visualization, alerts, and classification.

Visualization puts long-term monitoring data in a concise and intuitive format,[7] which could significantly reduce the physician's data burden and enable timely and accurate decision making.

Alerts are alarms that activate when a value crosses an established threshold. The value can be simple to check, or the result of a more advanced algorithm. The threshold could be a clinical standard—for example, 480 ms for QTc, which is a measure of the ventricular depolarization and repolarization duration—or it could be tailored to a patient. For example, the physician might want to be notified only if a particular patient's QTc exceeds 600 ms.

Classification is the process of predicting the group a patient belongs in, such as people with a specific genotype or people at risk for certain cardiac events. Prediction of short-term outcomes is a primary goal. For example, the machine might predict that a patient is at high risk for a myocardial infarction in the next 12 hours.

The outputs from these support systems, such as plots and recommendations, would be attached to the typical ECG report that a cardiologist reviews. When real-time monitoring reveals an urgent issue, an alert would immediately be sent to both the patient and physician through SMS, pager, or an application.

Evolving symbiosis. The physician is still at the head of this process—ordering tests, analyzing records, adjusting prescriptions, and so on. The machine's visualizations and recommendations are simply additional decision-making tools. Over time, the database will expand, and the machine's classifications will be more accurate. But improvements will be symbiotic: as the machine's accuracy grows, the physician will develop an intuition for how and when the machine makes those accurate classifications and recognize when it might be fallible. For example, a patient might have an abnormal T-wave morphology that the algorithms did not process correctly, or a patient's heart rate had not reached the point at which problems would be identifiable. The physician can recognize the machine's limitations, and might opt for additional methods to measure risk such as prescribing a drug or exercise challenge or conducting a manual ECG analysis.

CASE STUDY PARAMETERS

In a case study to evaluate our method, we exploited ML's pattern-recognition abilities to classify risk in LQTS patients. The QT interval, which is typically used to measure the duration of ventricular repolarization (a clinical marker of the heart's electrical activity), can be abnormally long in some people who are taking certain medications or have certain genetic disorders.[8] A prolonged QT interval can trigger arrhythmias such as torsades de pointes (TdP), which are likely to cause serious symptoms such as seizures, fainting, or sudden death. It is therefore critical to monitor the QT interval in patients prone to this disorder using long-term ECG recordings. Data recordings of ambulatory patients over several hours or days are called Holter recordings or simply *Holters*.

Our study focused on congenital LQTS rather than the drug-induced form. In the database used for the study, we knew which recordings were from patients with symptoms, but we did not know if the symptoms came before or after the ECG recording. Consequently, we had no way to use ML to predict *when* symptoms would occur or to detect symptoms in real time. Instead, we attempted to identify when a recording came from a patient whom we knew had symptoms in the past or would have them in the future. In other words, we attempted to identify the patient's risk—an important concern for physicians, who must often prescribe medications and implantable devices on the basis of the perceived risk of symptoms. (Symptoms in this context are events, such as syncope, that are triggered by prolonged QT.)

Identifying high-risk patients who need extra prescriptions or monitoring *or* low-risk patients who would not benefit from those burdens would be highly valuable to both the physician and patient. Additionally, despite the limitations of this particular database, this study laid the groundwork for a future study at a time when a dataset with clinical outcomes becomes available.

Data preprocessing

Figure 2 illustrates the steps that transform raw ECG data into clinically useful measurements. Raw data

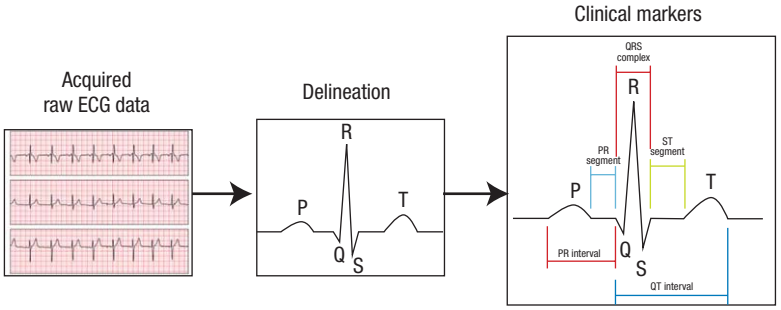


FIGURE 2. Detailed preprocessing steps using the workflow from Figure 1. The raw ECG signal contains too much data to feed into most ML algorithms, and the data has massive redundancy across leads (sensor locations) and heartbeats, as well as noise and errors. Preprocessing filters the raw ECG waveforms to extract only relevant clinical markers, such as the durations of the QRS complex and P and T waves.

contains massive redundancies as well as ectopic heartbeats and obvious noise and errors, which will not be useful in later processing. Preprocessing extracts only clinically relevant markers from the data, which reduces the data fed into the ML algorithm by multiple orders of magnitude while drastically improving classification accuracy and execution time. The delineated heartbeat (*Clinical markers* box) highlights several important measurements of the cardiac electrical cycle. Atrial and ventricular depolarization and repolarization are represented on the ECG as a series of waves: the P wave followed by the QRS complex and the T wave. The P wave occurs during atrial depolarization—when the atria contract. The QRS complex indicates ventricular depolarization—when the ventricles contract; at its end is the J point. The T wave occurs during ventricular repolarization—when the ventricles relax. Durations, amplitudes, and shapes at various parts of the ECG can be used to diagnose myriad illnesses.

Algorithm training

To identify patients at risk for LQTS symptoms, we trained ML algorithms using input variables extracted from raw ECG data. As Figure 2 shows, useful intervals and amplitudes are available after only two preprocessing steps: delineation and the computation of clinical markers.

We first annotated important markers in the ECG recording (such as the P, Q, R, S, and T peaks, onsets, and offsets) using delineation software to identify these points.[9,10]

The relevance of each clinical marker depends on the disease being studied. For example, STe, the elevation (voltage) during the ST segment, is of interest in heart-attack cases, and the shape of the P wave is of interest to diagnosing atrial enlargement. Many diseases affect the QT interval and its subintervals (QRS, the J point to T peak, and T peak to T end).

In our study, the QT and RR intervals were the most relevant markers. The QT interval alone is not enough information, because it will naturally lengthen and shorten in all individuals as their heart rate decreases or increases. The RR interval—the duration of a complete cardiac cycle—provides enough information to correct the QT for heart rate. We calculated the corrected QT (QTc) using the Fridericia equation:[11]

$$QTc = \frac{QT}{\sqrt[3]{\frac{RR}{s}}}$$

Reducing dimensionality

Preprocessing substantially reduces the data to be reviewed. The raw ECG data—sampled at 200 Hz, 16 bits per sample, on 3 leads, and over 24 hours—needed 100 Mbytes of storage. If only QTc interval values are required to detect LQTS symptoms, the storage requirement lowers to 1 Mbyte. Storage capacity alone is not a sufficient reason to have preprocessing; additional storage space is relatively inexpensive. Rather, preprocessing is necessary when using ML algorithms because the data reduction translates directly to a dimensionality reduction and faster processing in subsequent steps.

Reduced dimensionality is important because the "curse of dimensionality" remains a difficult problem in categorizing big data. That is, large volumes of data have a daunting number of features, with each feature having myriad possible values. There are so many dimensions to work with that it is too easy to separate the training data into the correct groups. The learned model becomes specific to the training set rather than generalizing to other data—a problem known as *overfitting*. Thus, an enormous amount of training data is required to ensure that there are several samples with each combination of values. With a fixed number of training samples, ML's predictive power can decrease as dimensionality increases.

Method selection

Among the ML methods for classification, supervised learning and clustering are the most popular. In our study, we focused on supervised learning. The alternative to supervised learning is clustering, also known as unsupervised learning, which generally tries to group data points into clusters according to their proximity to one another. A new data point can then be classified on the basis of the cluster into which it best fits.

Artificial neural networks, inspired by the neuron web in the human brain, can be used for both supervised and

TABLE 1. Pros and cons of four supervised learning classifiers.

Classifier	Advantages	Disadvantages
k-nearest neighbors	Simple to implement Easy to understand and interpret	Sensitive to noisy data and anomalies Computationally expensive for large datasets
Support vector machine	Flexible with nonlinear data Scales up with large sets of data Relatively resistant to the "curse of dimensionality"	Difficult to interpret feature importance Yields possibly unreliable confidence estimates
Random forest	Alleviates overfitting problem Easy to extract feature importance Resilient to missing data Scales to large datasets	Increases bias relative to single decision tree Different results possible in retraining on same data
AdaBoost	Automatically reduces dimensionality Relatively fast	Sensitive to noisy data and anomalies

unsupervised learning. Deep networks use many layers of artificial neurons to form input data abstractions, which lead to the formation of a final classification layer. In each layer, weights are applied to features of the previous layer to optimize performance.

The supervised learning methods that best fit our study were support vector machine (SVM), decision tree, and nearest neighbors. These algorithms classify previously unseen data points based on some function of the data points they have already seen. In an attempt to improve accuracy, some ML algorithms factor in the results of several classifiers in order to make their decision. Random forest and AdaBoost are among the most popular methods to use this ensemble learning technique.

Table 1 lists the pros and cons of the four methods we considered in our evaluation: *k*-nearest neighbors, SVM, random forest, and AdaBoost. The table is useful in identifying the best classifier for a given problem and set of computational constraints.

***k*-nearest neighbors.** The *k*-nearest neighbors algorithm finds the shortest distance between a new testing point and adjacent training points. It then classifies the testing point as the most common class among its *k*-nearest neighbors.

Support vector machine. The SVM method uses a training points subset to create hyperplanes that divide the data into classes, which it keeps as far apart as possible (thereby maximizing the distance between the hyperplane and different class samples). SVMs rely on a linear or nonlinear feature combination, depending on the kernel declared in the algorithm. We tested a linear kernel as well as a radial basis function (RBF) kernel.

Random forest. The random forest algorithm is an ensemble learning method that averages the results of several decision trees to classify its samples. As the name implies, each decision tree is trained on a random training data subset, perhaps using random features as well.

AdaBoost. AdaBoost, short for adaptive boosting, aggregates the results from many weak classifiers by iteratively retraining them to focus on fixing mistakes from the previous round. It then averages the results. In our experiments, AdaBoost always used decision trees as the weak classifier.

RUNNING THE CLASSIFIERS

We accessed a database containing 24-hour ECG recordings of 480 LQTS patients, including demographic information such as gender, age, and specific LQTS genotype.[12] We restricted our study to 434 recordings of patients with the most common LQTS genotypes (*LQT1* and *LQT2*) and the most complete demographic information (such as age and gender). The subjects' average age was 25 ± 18 years (newborns to senior citizens); 55 percent of the subjects were female, and 67 percent had the *LQT1* mutation. Our goal was to determine which of the patients would show symptoms of LQTS, such as seizures or syncope. That is, we were trying to identify ECG patterns that could reveal which genotype-positive patients will also be phenotype-positive. Given some measurements from an ECG, a classifier should simply tell us "symptoms expected" or "no symptoms expected," perhaps with a confidence value.

Algorithm implementation

We implemented all the classification algorithms—*k*-nearest neighbors, linear SVM and RBF SVM, random forest, and AdaBoost—using scikit-learn, an open source Python library built on SciPy and NumPy.[4] To assess a classifier's accuracy, we set aside 30 percent of the samples for testing, and trained only on the remaining 70 percent. Because some algorithms include inherent randomness in their operation and because the division between training and testing data is also random, we repeated the cycle of selecting training data, training, and testing 50 times for each classifier.

The average result from these trials—the Monte Carlo cross-validation

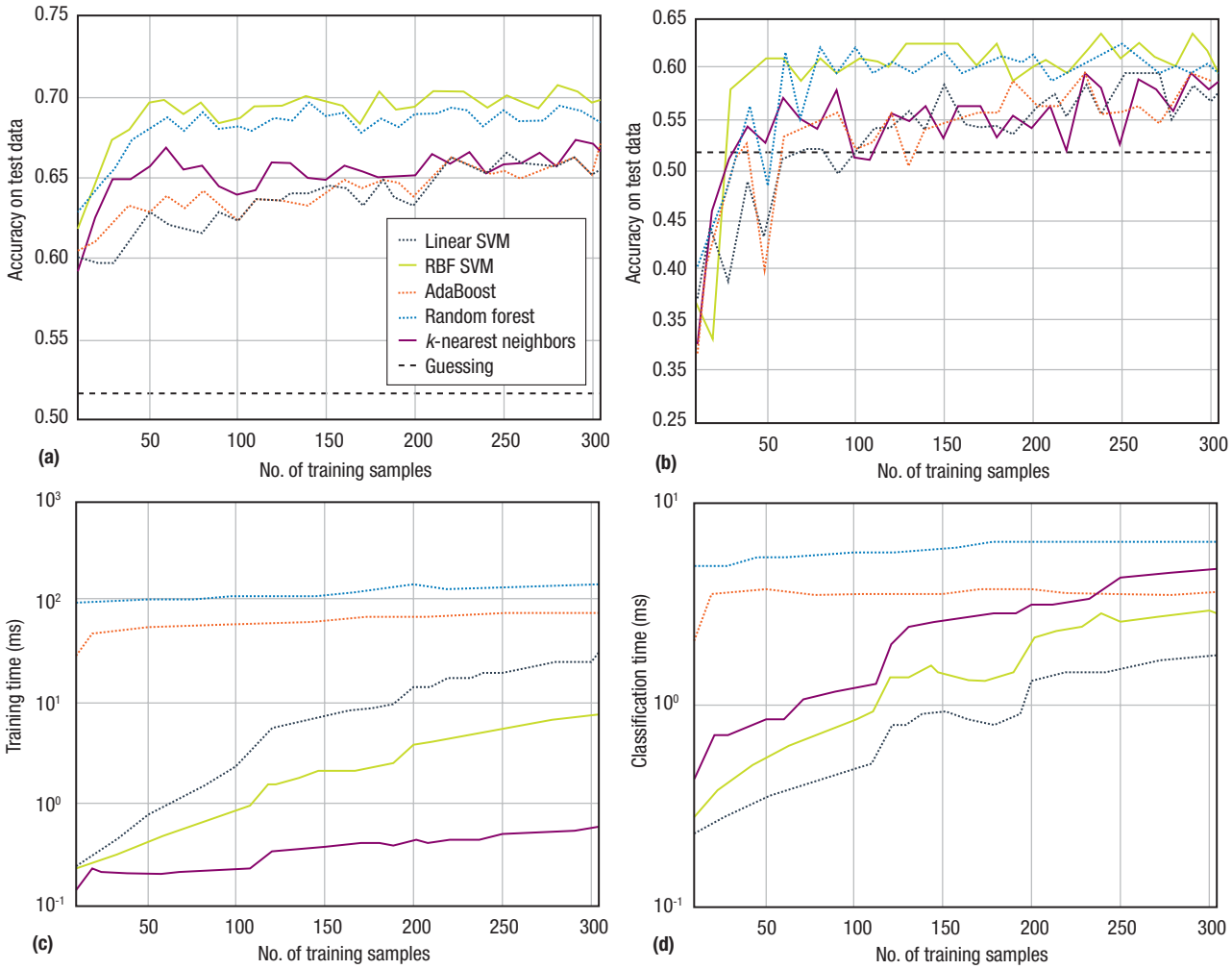


FIGURE 3. Classifier performance in the study. The input features were hourly QT and RR measurements for one day. Each data point is an average of 50 trials for which we randomly selected different training and testing data. Approximately 52 percent of the patients did not have symptoms, so the thin dashed line in (a) average accuracy and (b) minimum accuracy represents the performance achievable by simply guessing "no symptoms" every time. (c) Training time represents training conducted for increasingly larger subsets of the full training dataset of 304 samples, and (d) classification time represents how long validation took for the full test set of 130 samples. AdaBoost always used decision trees as the weak classifier.

score—told us how well a classifier would likely perform. The worst result showed how low a classifier's accuracy could be when the data was not evenly distributed across the random training/testing split (based on its underlying properties). For example, if most of the outliers or noisy recordings end up in the test set, or almost all of the asymptomatic patients end up in the training set, our trained model will not match up well with the data it's tested against. For classifiers that require input data to be normalized, we used scikit-learn's `StandardScaler()` function.

Feature selection

The four ML methods we used have inherent strengths and weaknesses, but their performance was also constrained by the data we provided them. We knew that QTc, and therefore QT and RR, are the measurements that cardiologists use most often to determine whether a LQTS patient is in danger. We also knew that people with different *LQT* genotypes tend to show more QTc prolongation at different times of the day.[13] We therefore decided to provide hourly QT and RR measurements as input to the ML classifiers. Each of our samples for training or classification consisted of 48 values (24 for QT and 24 for RR); increasing that number risked inflicting the curse of dimensionality.

To reduce dimensionality even more, we used chi-square (χ^2) tests to automatically select features that were likely to be the most useful. In general, the fewer the dimensions in the input, the fewer training samples are needed to achieve good performance, and the faster classifiers will run. Feature selection methods are also useful in the discovery of previously unknown

patterns and correlations between variables. For example, the machine might find that patients with a very specific genetic mutation are more at risk than others with seemingly similar mutations or that a measurement that is typically not used in the clinic actually carries significant information. Even if no new relationships are discovered, feature selection is useful to confirm the chosen model—to see if the machine picks the features expected.

RESULTS

We used the four classifiers to determine if the genotype-positive LQTS patients in our database had suffered or would suffer from any symptoms. Figure 3 illustrates how the performance of each classifier changes as we provided more training samples. We configured the *k*-nearest neighbors classifier to weight samples by distance rather than uniformly and composed the random forest with 100 trees rather than the default of 10. We set both random forest and the two SVMs to use balanced class weights. All other parameters were the scikit-learn defaults. Each of the classifiers' input samples contained 48 values.

We computed runtime results on an Intel i7-5930K and always used 30 percent of the 434 samples in the full dataset (training plus testing) for testing. However, we conducted training over increasingly larger subsets of the remaining 70 percent to determine how many samples would produce optimal results.

Average and minimum accuracy

Figure 3a shows average accuracy—the accuracies that we could expect from each classifier on the basis of 50 random training-data selections. Figure 3b shows minimum accuracy over the 50 trials—assuming we chose training data poorly, how well could each algorithm do? We found that in this worst-case scenario, more than 100 training samples could be required simply to break even—to exceed 52 percent (the guessing line). The highest scores, both minimum and average, came from random forest and the RBF SVM, which achieved 60 to 65 percent accuracy even with a poor selection of data and fewer than 100 training samples. The best classifier in our tests, the RBF SVM, averaged about 70 percent accuracy.

> IN FEATURE SELECTION, OUR AIM WAS TO REDUCE INPUT DIMENSIONALITY WHILE MINIMIZING RELEVANT INFORMATION LOSS.

Obviously, it is not desirable for the machine to make bad classifications. However, relatively low accuracy is manageable if we know when the computer was unsure of a result. We therefore tested the machine's average confidence in its responses. When the computer was incorrect, its average confidence was around 64 to 69 percent with the best classifiers, RBF SVM and random forest. When the computer was correct, its confidence was higher—around 68 to 74 percent. This test showed us the possibility of setting a threshold, below which the decision-support system could report "inconclusive" rather than marking a patient as having a high or low risk.

Scalability

Figures 3c and 3d show the results of measuring the runtime of the training and classification stages, which we used to estimate each classifier's scalability. Runtime was not a problem with this particular dataset, but it could be a limitation in other studies. Although the ensemble classifiers (AdaBoost and random forest) took longer than the others in both stages, adding training samples barely affected their runtimes. As we expected, *k*-nearest neighbors had essentially zero training time, but classification time increased with the number of samples because the algorithm had to compute distances to every point in the training set. In fact, at around 240 training samples, classification actually became slower than with AdaBoost. Because the two ensemble methods have very flat runtimes, SVM will also become slower than even random forest, given enough training samples.

All the classifiers except *k*-nearest neighbors (which does not really have a training stage) could not incorporate new data after training was complete. When a database grows, classifiers must be entirely retrained or the process of adding samples must use nontrivial techniques. The ability to add one or more training examples to a model without complete retraining, referred to as online ML, is an

important scalability feature. Consequently, we tested the perceptron online algorithm[14] and found that it achieved 68 percent accuracy, a level comparable to that of random forest.

Effects of feature reduction

We next investigated the impact of providing classifiers with QTc alone instead of QT and RR separately, which reduced the number of features from 48 to 24. Because QTc is designed to contain the LQTS-relevant information from QT and RR, we expected to see improved runtimes without lower accuracy. However, we found that accuracy decreased for random forest and AdaBoost (both of which are based on decision trees) as did the learning rate (more samples were required to reach peak performance). The feature reduction did not hurt RBF SVM's performance, which consistently yielded 70 percent accuracy when provided with enough training samples.

AdaBoost and random forest saw no improvement in runtime with fewer features. Predictably, *k*-nearest neighbors' classifications were faster, and all SVM runtimes improved. Because of these results, we revisited our feature choice, attempting to narrow the list as much as possible.

Searching for better features

Our use of QT, RR, and QTc was based on knowing what physicians measure in practice. However, we wanted to be sure that we did not overlook any other useful cardiac features, so we decided to evaluate 23 features at every hour of the day—a total of 552 measurements that included QT, RR, QRS, ST segment duration and elevation, QTp, JT, JTp, TpTe, and T-wave duration and amplitude. Additionally, we used several features from each patient's EHR: gender, age, *LQT* type (1 or 2), mutation type, and mutation location.

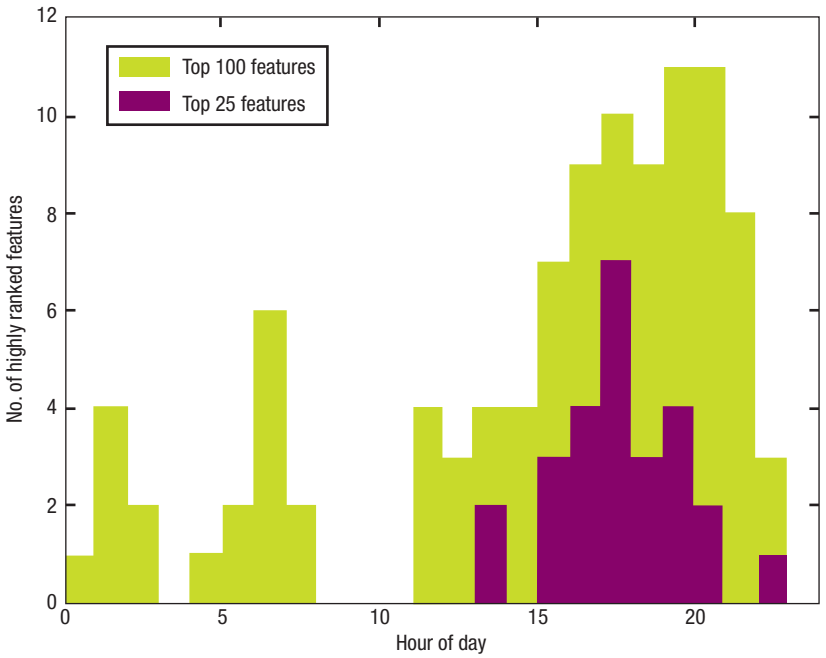


FIGURE 4. Feature importance versus time of day. Starting at 0, which represents midnight, the peaks indicate that late night, awakening, and the end of the workday are the best times to detect cardiac issues for this patient group.

Because we had only 434 training samples, we expected to have to reduce the new feature set's dimensionality to mitigate overfitting. Our aim was to reduce input dimensionality while minimizing information loss, which both the principal component analysis (PCA) and the χ^2 method satisfy. PCA projects the data to a lower dimensional space, while χ^2 selects the statistically best features. We measured classifier accuracy varying the number of preserved features from 1 to 512. Surprisingly, both feature selection methods allowed the classifiers to achieve 70 percent accuracy with only one feature or attribute. The most important features seemed to be QT-like measurements taken in the evening, where "QT-like" means QT, JT, QTp, JTp, or versions of these corrected for heart rate. Using the top 20 features with the random forest classifier yielded 72 percent accuracy (69 percent sensitivity and 75 percent specificity).

Figure 4 is a histogram of the times of day in which the top 25 and top 100 features appeared. As the figure shows, all the top 25 features are evident around 5 pm to 6 pm, which implies that perhaps fatigue at the end of the workday is unmasking cardiac issues. Expanding the search to the top 100 features begins to reveal other important times. One is first thing in the morning (6 am to 7 am), which is another stressful time of day.[15] Another is late night (1 am to 2 am), which makes sense for the *LQT2* subset of patients who tend to show more QTc prolongation during sleep.[13] Also important is the *lack* of highly ranked features around 8 am to 11 am, which implies that a clinical checkup in the morning might not be sufficient for the physician to accurately assess a patient's risk.

STUDY IMPLICATIONS

Our study had several implications for future analyses, including the effects of beta blockers, gender influence on classification, and measurement type.

Beta blocker effects

In one of our experiments, we found that separating QTc into QT and RR improved accuracy. However, many high-risk LQTS patients are on beta blockers—drugs prescribed to slow the heart rate—so it was possible that classifiers in that experiment were actually reacting to the increased RR that is characteristic of beta blockers, not to any novel pattern we had hoped to find. In particular, beta blockers might explain why the classifiers based on decision trees (AdaBoost and random forest) had higher accuracy when heart rate was available. To determine if this was the case, we conducted three evaluations.

The first was to train the classifiers on only beta blockers (BBs) or non-BB patients and then check accuracy. When the classifiers were trained on only BB patients, overall accuracy remained at approximately 70 percent. Accuracy within the BB group increased to approximately 90 percent, but accuracy in the non-BB group fell to approximately 60 percent. When we trained the classifier on only the non-BB group, we observed the opposite results. Although 90 percent accuracy is a marked improvement, the "guessing" line is higher in these subgroups: 67 percent of non-BB patients had no symptoms, and 62 percent of BB patients had symptoms. When combined, the line was 52 percent.

Our second experiment was to train the classifier on both groups as before, but to test the accuracy against each group separately. Results showed little difference: classification of only BB or only non-BB patients remained at 66 to 68 percent accuracy.

> THE ML SYSTEM WE ENVISION WILL PROVIDE A PERSONALIZED ASSESSMENT OF EACH PATIENT BASED ON A COMBINATION OF CRITICAL MARKERS.

Finally, in our experiment that used only QTc as input, we were not (ideally) providing any heart-rate information to the classifier. The classifier needed more training samples to reach peak accuracy, but that peak was still around 70 percent.

We concluded that the presence of BBs does not affect overall accuracy, but classifiers are more accurate when the groups are separated.

Gender

We expected gender to be a feature of significant importance in ML classification, as gender differences are known to influence many coronary heart diseases,[16] and males and females have distinct average QTc values and clinical prolongation thresholds. However, feature selection eliminated gender (along with age and mutation information) as very insignificant relative to most ECG measurements. We confirmed this by running RBF SVM and random forest with gender as an input, and found no difference in results and virtually no weight placed on that feature. It will be interesting to see under what conditions gender or the other static inputs become significant.

Measurement type

For our study, the feature of interest (QTc) has characteristics that fit well with hourly average measurements because QTC changes slowly and is corrected for heart rate. However, for other applications, even the extensive feature set we tested might not be enough. Heart rate, for example, can vary greatly during one hour. Perhaps a different measure such as heart rate variability would be more suitable. Future analyses might even include more exotic measurements, such as "ST elevation at 60 ms after J point during high heart rate" or "percentage of beats that T wave is inverted."

We have presented a workflow and a conceptual ML-based system for health monitoring that aims to analyze the ECGs of patients with an LQTS genetic disorder and to identify patients with increased risk of adverse cardiac events. The envisioned system will provide a personalized assessment of each patient by considering a combination of critical markers.

The results in Figure 3 provide insights into which classifier works best under constraints such as available training data or computational power. RBF SVM or random forest will yield the highest accuracy. RBF SVM is probably the better choice for experiments with relatively few training samples, but random forest's faster runtime will be preferable when training data grows to thousands of samples. As runtime

becomes more of a concern, feature selection gets more important.

In Figure 3a, all classifiers seem to be close to reaching a horizontal asymptote, meaning that their performance will not improve simply by adding more training samples. Instead, their inputs and parameters will need to be optimized. In previous work, we found that time of day was important in classifying patients as having the *LQT1* or *LQT2* genotype.[13,17] Both types show QT prolongation, but during different activities and different times of day. This finding is in large part why we structured our inputs as hourly data points in the study described. The dimensions also reduce well in this structure, as usually only a few hours during sleep are enough to differentiate LQTS types. However, other input structures and measurements should be investigated. In the selection of appropriate input features, the physician's knowledge and intuition remain critical.

The annotation algorithm we used had some trouble with noisy or abnormal ECGs; improved accuracy might require cleaner inputs and more accurate annotations. Additionally, we could construct more complex features such as T-wave symmetry measurements. Another possible approach is the use of a voting classifier, which attempts to aggregate the predictions of several other classifiers to reach a better result. However, our experience suggests that a voting classifier will be only slightly more accurate than the best individual classifier. Finally, a complete set of experiments will require trying other classification methods such as clustering and artificial neural networks.

The steps we used can be generalized to other types of medical data and illnesses. We expect that the refinement of our method and the growth of EHR databases will greatly improve the quality of care for patients with a variety of disorders.

ACKNOWLEDGMENTS

We thank Mehmet Aktas and Jean-Philippe Couderc from the University of Rochester's Department of Medicine for motivating this study and providing guidance about its clinical applications. This work was supported in part by National Science Foundation grant CNS-1239423.

ABOUT THE AUTHORS

SHUROUQ HIJAZI is a technology consultant at Ernst & Young. While conducting the research reported in this article, she was a research assistant in the machine-learning (ML) laboratory at the University of Rochester. Her research interests include ML techniques, cybersecurity, computer networks, the Internet of Things (IoT), and virtualization. Hijazi received a BS in electrical and computer engineering from the University of Rochester. She is a student member of IEEE. Contact her at shijazi@u.rochester.edu.

ALEX PAGE is a postdoctoral associate in the Heart Research Follow-up Program at the University of Rochester Medical Center. While conducting the research reported in this article, he was a PhD student in electrical engineering at the University of Rochester. His research interests include computer systems for analyzing medical data, such as databases, GPU acceleration, and ML techniques. Page received a PhD in electrical engineering from the University of Rochester. He is a student member of IEEE. Contact him at alex.page@rochester.edu.

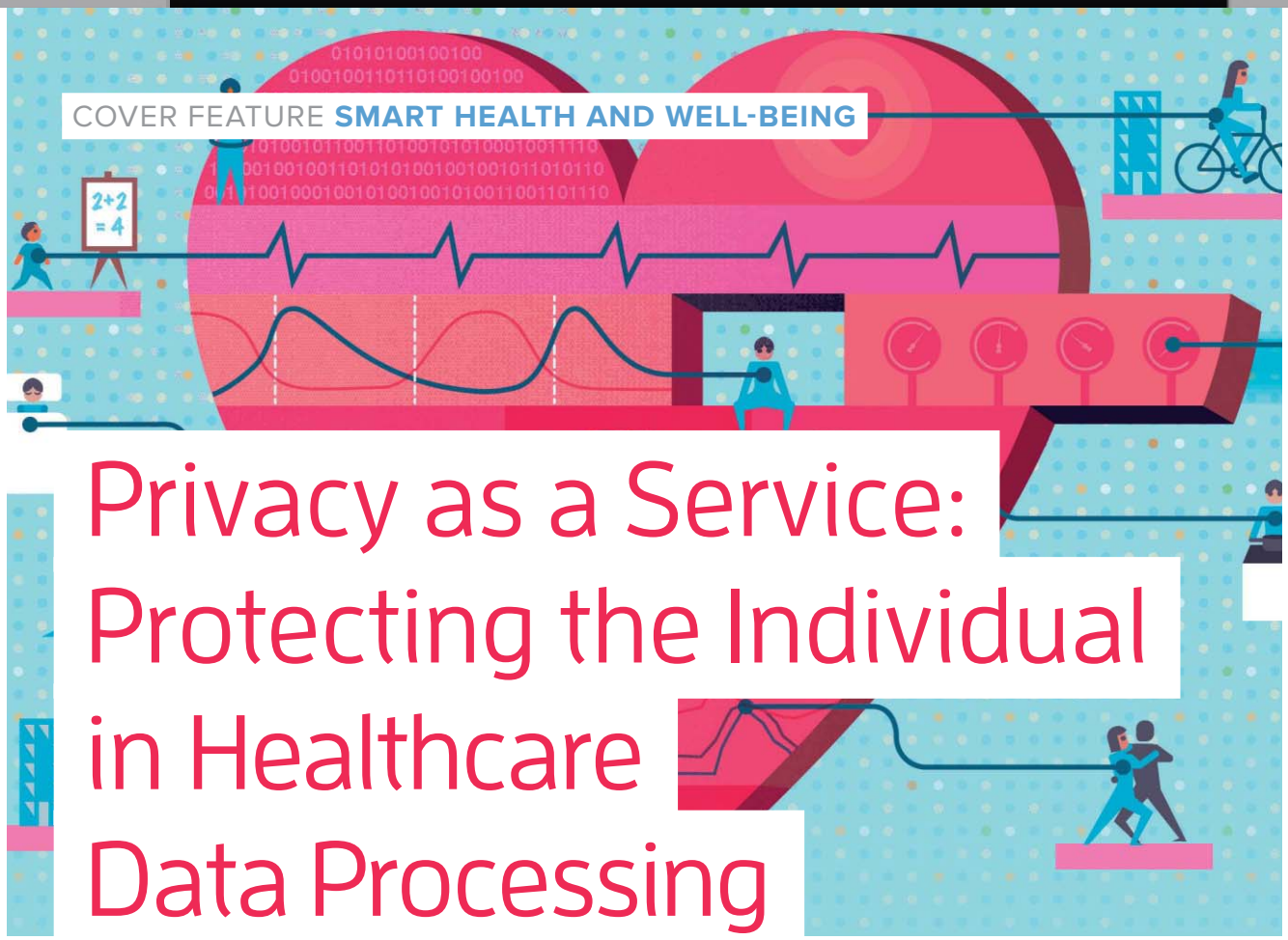
BURAK KANTARCI is an assistant professor in the School of Electrical Engineering and Computer Science at the University of Ottawa and a courtesy assistant professor in the Electrical and Computer Engineering Department at Clarkson University. His research interests include the IoT, big data in the network, crowdsensing and social networks, cloud networking, and digital health. Kantarci received a PhD in computer engineering from Istanbul Technical University. He is an editor of *IEEE Communications Surveys and Tutorials*, a Senior Member of IEEE, and a member of ACM. Contact him at burak.kantarci@uottawa.ca.

TOLGA SOYATA an associate professor in the Department of Electrical and Computer Engineering at SUNY Albany. His research interests include cyber-physical systems, digital health, and GPU-based high-performance computing. Soyata received a PhD in electrical and computer engineering from the University of Rochester. He is a Senior Member of IEEE and ACM. Contact him at tsoyata@albany.edu.

REFERENCES

1. T. Lorberbaum et al., "An Integrative Data Science Pipeline to Identify Novel Drug Interactions That Prolong the QT interval," *Drug Safety*, vol. 39, no. 5, 2016, pp. 433–441.
2. M. Hassanalieragh et al., "Health Monitoring and Management Using Internet-of-Things (IoT) Sensing with Cloud-Based Processing: Opportunities and Challenges," *Proc. IEEE Int'l Conf. Services Computing* (SCC 15), 2015, pp. 285–292.
3. D. Son et al., "Multifunctional Wearable Devices for Diagnosis and

Therapy of Movement Disorders," *Nature Nanotechnology*, vol. 9, 2014, pp. 397–404.
4. F. Pedregosa et al., "Scikit-Learn: Machine Learning in Python," *J. Machine Learning Research*, vol. 12, 2011, pp. 2825–2830.
5. "Summary of HIPAA Privacy Rule," US Department of Health and Human Services, 2006; www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations.
6. E.C. Larson et al., "Spirosmart: Using a Microphone to Measure Lung Function on a Mobile Phone," *Proc. ACM Conf. Ubiquitous Computing* (UbiComp 12), 2012, pp. 280–289.
7. A. Page et al., "An Open Source ECG Clock Generator for Visualization of Long-Term Cardiac Monitoring Data," *IEEE Access*, vol. 3, 2015, pp. 2704–2714.
8. C.E. Chiang, "Congenital and Acquired Long QT Syndrome: Current Concepts and Management," *Cardiology Rev.*, vol. 12, no. 4, 2004, pp. 222–234.
9. Y. Chesnokov, D. Nerukh, and R. Glen, "Individually Adaptable Automatic QT Detector," *Proc. IEEE Computers in Cardiology* (CinC 06), 2006, pp. 337–340.
10. A. Demski and M.L. Soria, "Ecg-Kit: A Matlab Toolbox for Cardiovascular Signal Processing," *J. Open Research Software*, vol. 4, no. 1, 2016; openresearchsoftware.metajnl.com/articles/10.5334/jors.86.
11. L.S. Fridericia, "The Duration of Systole in an Electrocardiogram in Normal Humans and in Patients with Heart Disease," vol. 53, 1920, pp. 469–486 (in German).
12. J. Couderc, "The Telemetric and Holter ECG Warehouse Initiative (THEW): A Data Repository for the Design, Implementation and Validation of ECG-Related Technologies," *Proc. IEEE Int'l Conf. Eng. Medicine and Biology Soc.* (EMBC 10), 2010, pp. 6252–6255.
13. A. Page et al., "QT Clock to Improve Detection of QT Prolongation in Long QT Syndrome Patients," *Heart Rhythm*, vol. 13, no. 1, 2016, pp. 190–198.
14. F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Rev.*, vol. 65, no. 6, 1958, pp. 386–408.
15. W.B. White, "Cardiovascular Risk and Therapeutic Intervention for the Early Morning Surge in Blood Pressure and Heart Rate," *Blood Pressure Monitoring*, vol. 6, no. 2, 2001, pp. 63–72.
16. A. Maas and Y. Appelman, "Gender Differences in Coronary Heart Disease," *Netherlands Heart J.*, vol. 18, no. 12, 2010, pp. 598–603.
17. A. Page et al., "Research Directions in Cloud-Based Decision Support Systems for Health Monitoring Using Internet-of-Things Driven Data Acquisition," *Int'l J. Services Computing*, vol. 4, no. 4, 2016, pp. 18–34.



COVER FEATURE **SMART HEALTH AND WELL-BEING**

Privacy as a Service: Protecting the Individual in Healthcare Data Processing

Xiang Su, Jarkko Hyysalo, Mika Rautiainen, Jukka Riekki, and Jaakko Sauvola, University of Oulu
Altti Ilari Maarala, Aalto University
Harri Hirvonsalo and Pingjiang Li, University of Oulu
Harri Honko, Tampere University of Technology

Health applications involve many data sources, individuals, and services that work against guarantees that an individual's personal data will not be used without consent. The proposed privacy-centered architecture integrates data security and semantic descriptions into a trust-query framework, enabling the provision of user consent as a service.

Healthcare's transition to the digital world has already reaped benefits such as more efficient processes and cost savings and has paved the way for new services and business models. However, the myriad organizations providing and consuming data sources and services have given rise to challenges, particularly with regard to how users can be assured that personal data is used only with their consent.

Recognizing privacy as a key obstacle to the full promise of digital healthcare, in 2012, the European Commission drafted the General Data Protection Regulation (GDPR), which became a regulatory directive for the EU in May 2015. EU member nations must incorporate the directive in their laws by May 2018. The GDPR recognizes that individuals need to control their own data, but it also states the need for trust to be built into personal data services through a combination of transparency, interchangeability, public governance, respectable companies, public awareness, and secure technology. Control is realized through consent that determines

GUIDING ARCHITECTURAL PRINCIPLES

We inherited our privacy-as-a-service (PRIAAS) architectural requirements from extended MyData principles augmented with the General Data Privacy Regulation (GDPR). We view these principles as foundational to human-centric data processing and personal information management.

- **Control.** Individuals have the right and practical means to manage their data and privacy according to the GDPR.
- **Access.** Data must be easy for the individual to access and use.
- **Translation.** There must be a way to convert data from single entities into a meaningful, machine-readable resource that can be used to create new services.
- **Interoperability.** To support an open business environment, the shared data infrastructure must enable the coordinated management of personal data, ensure interoperability, and facilitate the compliance of various entities to stricter data protection regulations.
- **Provisioning.** The infrastructure must allow individuals to change service providers and control their data management.

what data services can fetch and how it can be processed. Thus, the regulation has a twofold objective: restore control to individuals over the use of their personal data, and simplify the regulatory environment for business services. Specifically, the regulation calls for provisions to ensure user consent and to coordinate data services. According to the GDPR, "user consent" is an explicit indication of the data subject's wishes and "signifies agreement to the processing of the subject's personal data, either by statement or by clear affirmative action."[1]

To support this reform, we developed a privacy-driven architecture that provides tools for providing user consent as a service within the MyData infrastructure.[2] MyData is a procedural framework for describing personal data management that considers both the individual's digital rights and the healthcare organization's needs. It essentially acts as a bridge between multiorganizational data silos and fully decentralized web-based systems. Our architecture works within the MyData approach to incorporate privacy as a service (PRIAAS), which facilitates the management and reuse of private health information. PRIAAS is designed to accommodate a large number of data sources, individuals, and services—even when they are not known to the user. The architecture integrates data security and semantic descriptions into a trust-query framework to provide the interoperability and cooperation that health services will increasingly require. PRIAAS's benefits include safer data management, cost and process savings, and the ability to handle the multiprovider services that are often inherent in newer business models.

The sidebar "Guiding Architectural Principles" describes five principles that we followed in compliance with the GDPR and the MyData approach. PRIAAS is the first open solution that conforms to the GDPR, is poised for widespread use in Finland (an EU country), and is endorsed as part of the Finnish government's spearhead agenda.

CONSENT STANDARDS

Although personal information—whether a name, photograph, email address, bank details, or medical information—is routinely shared digitally across national borders, mechanisms remain organized around national boundaries, specific service provider rules, and legal frameworks.[3] Consent is typically through hardcopy signatures or static online interactions, such as filling out forms or clicking buttons and opt-in checkboxes. These static, actor-driven mechanisms are obviously ill suited to scaling and interoperability, and most fail to comply with requirements for distributed health services.

The GDPR was motivated in part by the need to move consent further into the digital realm, and other standards also address these limitations—notably the User Managed Access (UMA) protocol[4] and the Minimum Viable Consent Record (MVCR) specification.[5] Like GDPR, UMA and MVCR aim to give individuals unified control points for authorizing who and what can get access to their digital data, content, and services. All three are founded on simplicity, ease of use, user-centeredness, transparency, and standardization. GDPR sets the legal framework that calls for explicit, unambiguous and informed consent, transparency, and interoperability, whereas UMA and MVCR provide authorization and

consent technologies that address GDPR-based requirements.

User Managed Access

UMA is an access-management protocol that gives individuals control over their personal data, content, and services. The protocol, which is based on the OAuth 2.0 standard, focuses on connecting a service that provides an individual's personal data to another service that consumes the same data in a way that allows the individual to securely manage data access.

PRIAAS adopts several UMA protocol characteristics, including

- unified access control under a dedicated online service;
- application of the same policies across multiple sites;
- support for claims-based access policies, such as "over 18"; and
- access control that is easy for the individual to manage.

Minimum Viable Consent Record

MVCR describes requirements for creating a legitimate digital consent record, known as a consent receipt, aiming to minimize the information that individuals need to address to enable their explicit consent. The consent receipt serves as the individual's basis for communicating to organizations about consent details and how the receipt can be used to authorize data access. We adopted MVCR as part of our consent record because it is the best available solution for describing consent in a universal machine-readable record.

HUMAN-CENTRIC DESIGN REQUIREMENTS

As the EU model shifts to enable more flexible exchange of healthcare data yet keeps exchange control in the hands of individuals, so system architecture must shift to human-centric designs that are built around regulations such as the GDPR.

Figure 1 illustrates how the human-centric paradigm translates into an architectural design. To ensure trusted and fair use of data across organizations, the GDPR imposes user rights that force organizations to build tools enabling informed user consent for data management, delivery, and exchange. In Figure 1, the roman numerals represent these eight user rights:

I. the right to unambiguous consent;
II. the right that only relevant, necessary, accurate, and legitimate data is processed in a specific, fair, and transparent manner;
III. the right to access one's own personal data;
IV. the right to be properly informed when personal data is processed;
V. the right to rectification;
VI. the right to protection against the use of personal data for automated profiling;
VII. the right to be forgotten; and
VIII. the right to security measures.

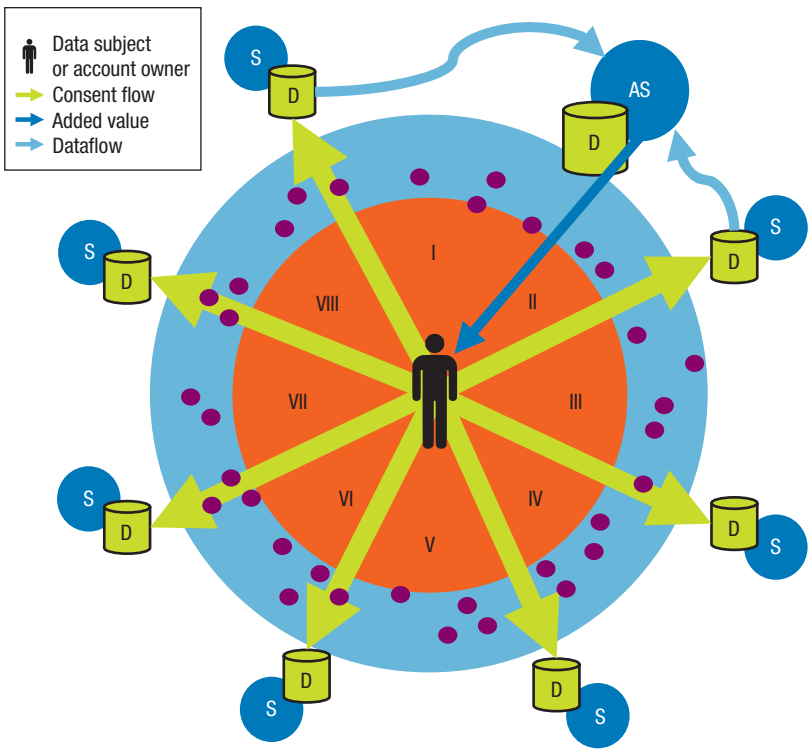


FIGURE 1. Aspects of a human-centric health services architecture. The individual produces data (D), which services (S) are designed to collect through a process imposed by an organizational entity. Applications (blue ring) present interfaces to users, who are increasingly involved in the organizational process of collecting and using their data. Aggregator services (AS) provide the individual with added value by correlating and analyzing data from a variety of organizational sources. The roman numerals in the red section represent enabling rights as set forth in the European General Data Protection Regulation (GDPR).

ROLES AND RESPONSIBILITIES

One of the central ideas in GDPR is that the control resides with the individual, who must be made aware of how personal data will be used before granting consent for its use. This requires centralized consent management in a distributed environment. For PRIAAS, this consent management is

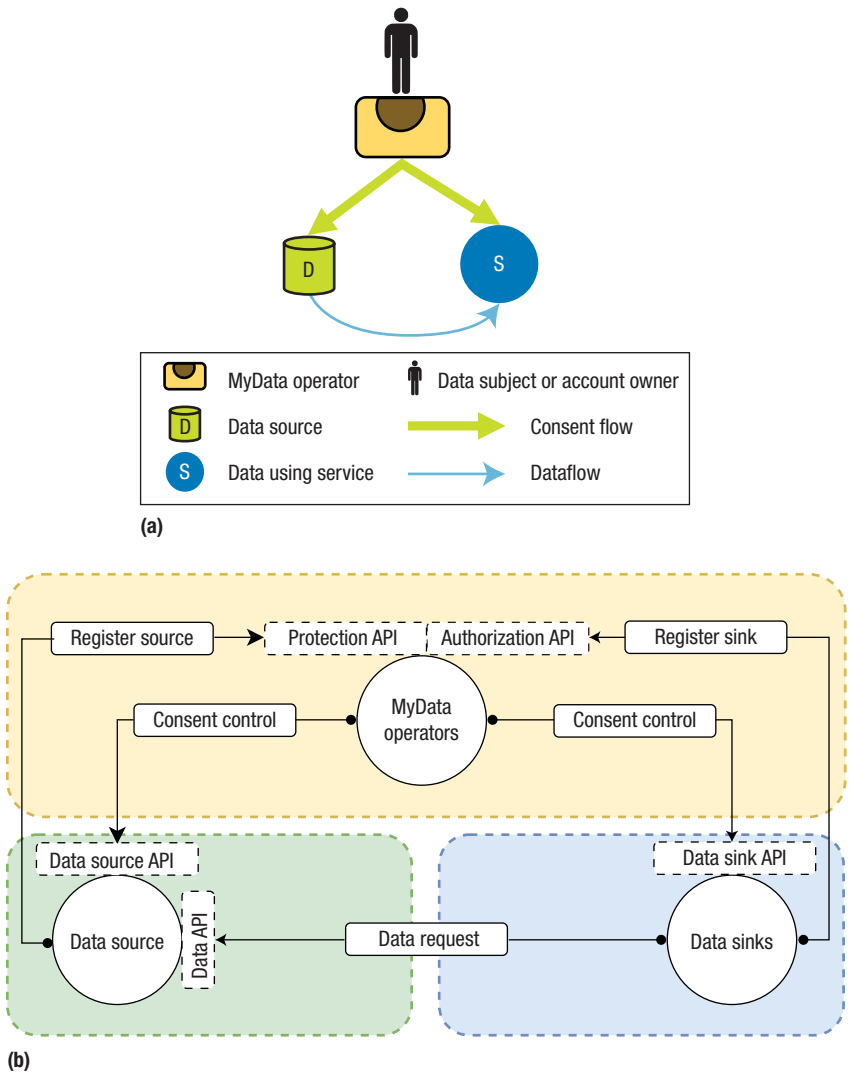


FIGURE 2. Consent management through MyData with our privacy-as-a-service (PRIAAS) system. (a) The MyData operator manages consent flow between the data source APIs (D) and the services consuming that data (S). (b) A more detailed diagram shows roles, interactions, and liabilities. In (b), the yellow screen represents the data owner. The MyData operators control consent and have protection and authorization APIs to both data sources and the services consuming that data (data sinks), the green and blue screens represent data sources and their APIs and data sinks and their APIs.

built on the MyData approach, which provides tools for tasks such as creating a service contract that honors rules for data exchange.

Managing consent

Central to the MyData approach is the MyData operator, a GDPR-compliant entity for service registration and consent management. Individuals use the operator to arrange and manage data exchange between sources and sinks, which are the entities that store, represent, and process data for user applications. MyData emphasizes portability and minimizes service provider lock-in, so individuals can choose MyData operators and migrate them. This account portability tends to increase trustworthiness because users have more control over processing.

As Figure 2a shows, the MyData operator manages consent through an individual's MyData account but the data itself is not necessarily streamed through the server hosting this account. The account maintains information on how the individual's personal data is connected to different services and the legal permissions and consent sources for data use. Figure 2b shows a more detailed representation. Data sources and services consuming data exchange information with the MyData account use MyData-compliant APIs. Individuals can grant access and give or cancel permissions for multiple data sources and services using this centralized interface. Any service provider can build a MyData API and enable their service to be connected with MyData accounts. Individuals can have multiple MyData operators and switch them as needed.

Consent management through MyData operators is a novel concept that lets users arrange and manage data exchange between sources and sinks. Through their MyData account, individuals can view, manage, and control consent easily and transparently through one operator's user interface. The resulting authorization process is simpler than UMA-based authorization, which requires multiple

interactions to enable authorized data transfer from source to sink.

User accounts held and managed by one or more trusted MyData operators also provide individuals with logical paths for controlling their personal data in complex environments of numerous data sources and consumers. Organizations acting on behalf of individuals can set up these accounts, or individuals can set up their own account services and provide them to organizations.

MyData operators provide web APIs that register sources and sinks through protection and authorization APIs. Services that implement the roles of sources and sinks must provide APIs for exchanging consent information, with the MyData operator acting as a broker. In practice, sources can use these APIs to inquire about the sinks' trustworthiness level before providing data access. Actual data exchange happens between sources and sinks without involving the MyData operator, which keeps the data architecture flexible and the MyData operator's role lightweight. Consequently, a variety of organizations can establish and maintain an operator service—which is important in accelerating the widespread adoption of a MyData ecosystem.

Creating a service contract

Figure 3 is a high-level sequence diagram of the process for creating a service contract, managing consent, and transferring data.

As (a) denotes in the figure, the process begins when users connect at least two services (source or sink) to their MyData account. Only connected services can receive consent.

In the second main step, (b), a sink is chosen, the user interface lists compatible sources and vice versa. The individual authorizes a sink to fetch data from a source and use this data according to rules that the user defines. The MyData operator records the parties involved in data use, the data being shared, and the rules for using in a pair of consent receipts that are stored in the user's MyData account and delivered to the sink and source. The operator constructs and cryptographically signs a token, which the sink uses to prove its authorization to the data specified in the consent receipt.

In the third step, (c), the sink makes a data request to the source presenting the token. Finally, the token and request description are delivered to the MyData operator, which uses previously stored information to determine whether the sink is authorized to access the requested data.

BENEFITS OF PRIAAS AND MYDATA

PRIAAS and MyData provide an open architecture with compelling benefits, such as ease of service expansion, flexible privacy monitoring, more efficient authorization relative to UMA, conformance with emerging regulations, cost savings, and improved health.

Ease of service construction

Because MyData is open source, developers can build access and services through public programming interfaces and libraries. Consent permissions—protection, authorization, and control—are managed through interfaces and programming instances that are separated according to purpose and usage rights. As a result, new consent operators, services, and applications naturally evolve.

> THE MYDATA OPERATOR NEVER STORES ANY GENERATED PERSONAL DATA BUT ACTS ONLY AS A TRUSTED CONSENT MANAGER.

Flexible privacy monitoring

Privacy is always protected because personal data moves between sources and sinks only with MyData operator permissions. In addition, MyData account management is separated from consent services and dataflows. The operator never stores any personal data generated by any source but acts only as a trusted consent manager and exposes the rights or limits to the use of an individual's data, on behalf of that individual. These flexible monitoring and governance mechanisms contrast sharply to existing models, in which consent is given on a service-by-service basis or through a single service operator.

To our knowledge, PRIAAS combined with MyData is the most comprehensive solution to date in providing informed consent management for healthcare data use. A 2015 literature review found no single solution that addressed even the majority of informed consent issues,[6] but it predates the development of the ForgeRock Identity Platform (www.forgerock.com

SMART HEALTH AND WELL-BEING

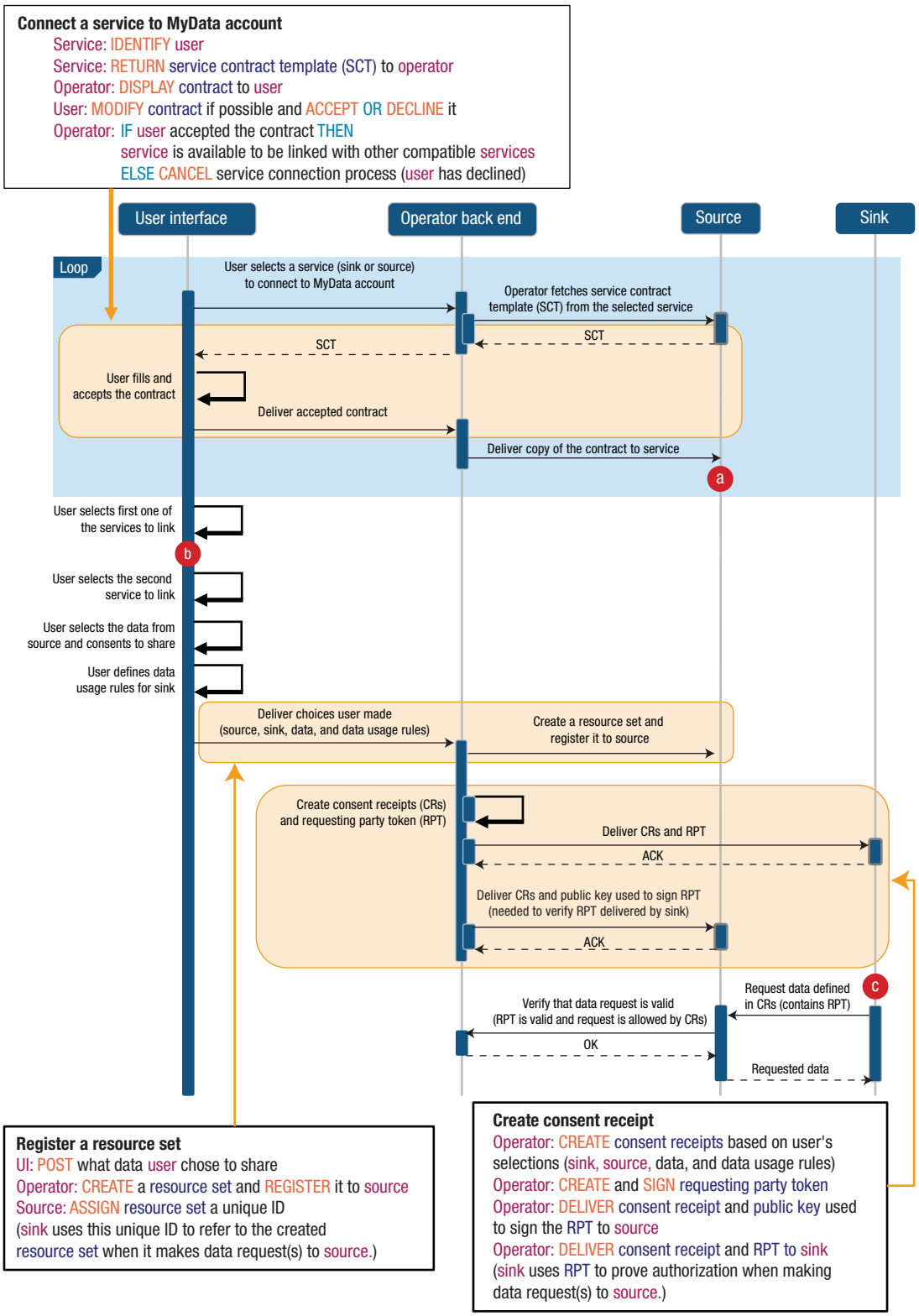


FIGURE 3. High-level sequence diagram of consent creation with examples from our sample implementation with multiple providers. (a) The user connects at least two services to his or her MyData account. Connected services are authorized to hold or release data. (b) If the first service selected is a sink, the user interface lists compatible sources; if it is a source, the interface lists compatible sinks. (c) Once consent is given, a sink can request data from a source.

/platform), which addresses evolving customer data privacy regulations based on the UMA protocol. We believe that this platform is the only comparable solution to PRIAAS and MyData in massively distributed private data environments.

More efficient authorization

PRIAAS borrows naming conventions from the UMA protocol, such as Protection API and Authorization API, and uses the concept of the resource set. However, PRIAAS does not conform to UMA protocol flow. Our authorization is based on a centralized authorization server similar to UMA, but because resource servers and clients are always discoverable and trusted, our authorization flow requires fewer messages compared to full UMA flow. For example, there is no longer a need to introduce the parties to each other in the beginning of authorization.

The authorization mechanism in PRIAAS is similar to OAuth 2.0 authorization code flow model because communication is expected to happen only between secure servers. However, PRIAAS differs in the way it defines the resource sets to be authorized: Instead of initiating registration by the resource server before the transaction, as in the OAuth 2.0 authorization flow, in PRIAAS, the resource owner initiates registration at the time of authorization transaction.

Conformance with emerging regulations

One major advantage of PRIAAS and MyData is conformance with regulations like the GDPR, which emphasize ease of use and interoperability. By brokering sources and sinks, the MyData operator enables trusted transfer of data and consent information with fewer messages, and the MyData account serves as a single hub for personal data management, allowing individuals to view, manage, and control their consents easily in a transparent and standardized way. Such standardization also facilitates interoperability.

Cost savings

Informed consent is considered valuable because it promotes trust in healthcare, which in turn ensures that people use healthcare more effectively.[7] Making consent safer and more efficient can reap even greater savings. Moreover, consent-management solutions like PRIAAS and MyData reduce the administrative time of medical personnel, who must otherwise process paper or online consent forms. Potential long-term profits stem from increased patient load and higher per-patient revenue or decreased per-patient cost.

The reduced learning curve is also a source of savings, as any new service investment has immediate costs in purchase, adaptation to the local organization, and staff training. MyData and PRIAAS are open source, which saves purchasing costs.

Finally, a comprehensive consent-management solution like MyData and PRIAAS promotes interoperability and reduces redundant documentation. Widespread adoption could move form system connectivity to interoperability among organizations and regions. The highest cost savings will come when a nation's health information systems are entirely interoperable. For example, complete interoperability within US health information systems has been estimated to yield savings of $77.8 billion annually.[8]

Improved health

Interoperability and cost savings are enablers that make personal data use rights a controlled, but ubiquitous, service. Greater trust in healthcare services leads to benefits that directly improve health, such as easier communication among patients with similar health problems and the discovery of clinical trials.[9]

SAMPLE IMPLEMENTATION

To validate the feasibility of PRIAAS, we developed a proof-of-concept implementation in which users authorize data sinks and sources to exchange data in a secure and trusted manner.

Data exchange

Data is exchanged only between sources and sinks through common data APIs, and consent is transmitted between operator and sink and operator and source. We divided the proof-of-concept implementation into separate parts: one each for MyData operators (one or more), sources, and sinks. Each part has distinct and interoperable roles and responsibilities in consent management that comply with regulations such as GDPR. The individual with a MyData account can complete all actions needed to establish consent for personal data to flow from a source to a sink. MyData-compliant sources provide data in a machine-readable format (JavaScript Object Notation) through RESTful interfaces.

Our implementation involved a health and wellness recommendation service that is based on multiple data sources. As shown in Figure 4a, the MyData operator manages user consent and data authorization by interacting with W2E, a general platform that aggregates wellness data

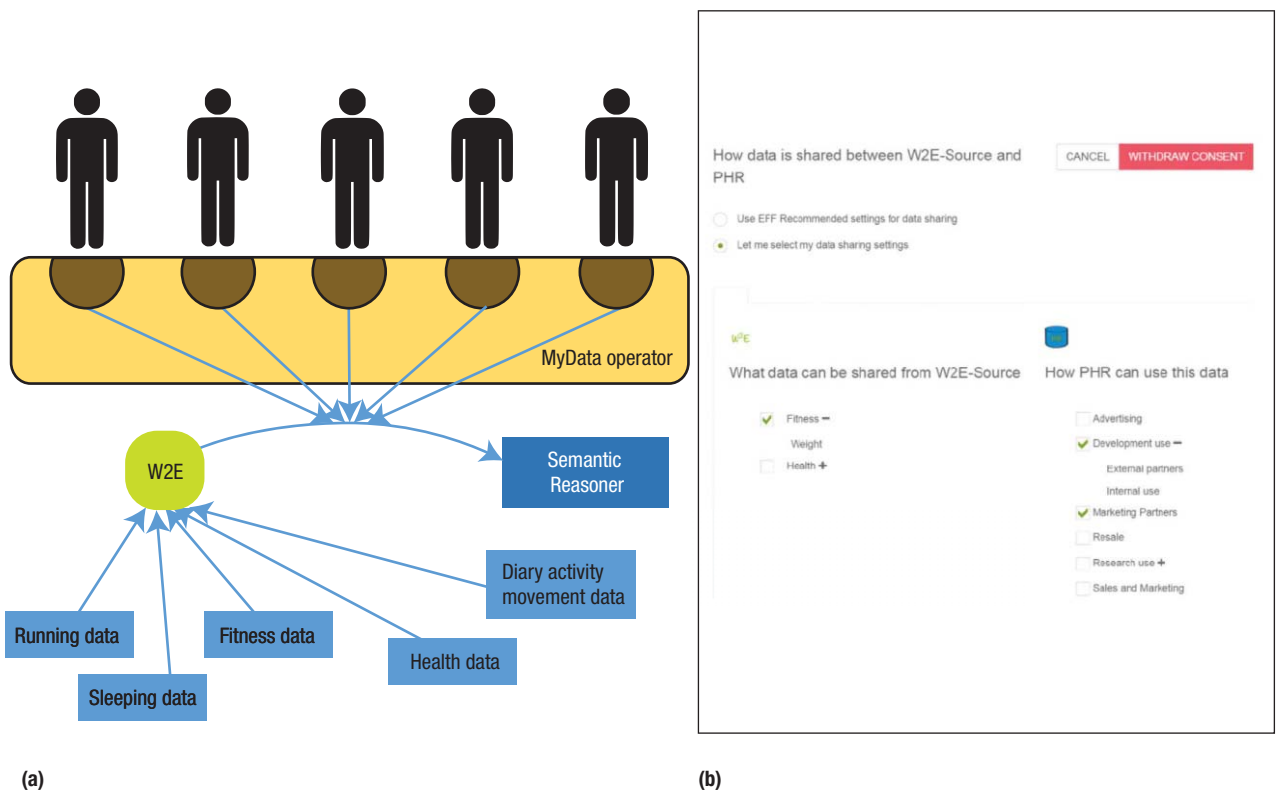


FIGURE 4. PRIAAS and MyData implementation that involves (a) multiple provider services and (b) an interface to manage consent for data use, in this case, to share fitness data with a personal health record (PHR).

(w2e.fi/frontpage), and components of the semantic reasoning service. Users must have an account at each of these services. The combination of W2E in MyData and the semantic reasoning service preserves privacy through pseudonyms and the separation of consent flow and dataflow. The W2E proxy accesses data from multiple sources that are not MyData compliant—mostly the back-end servers of health and wellness device manufacturers—and delivers the data to the semantic reasoning service. Data delivery is subject to the MyData operator, which manages the user's consent registry.

Figure 4b shows a screenshot of the consent-management interface, through which users connect sources and sinks and specify how the data sources can be used. In the screen portrayed, the user wants to share his fitness data from W2E with a personal health record (PHR). Hence, the PHR can fetch this data from W2E and apply it in a way that benefits users, such as giving their medical care providers more insight as well as supporting decisions relating to their health and healthcare.

Semantic reasoning

The reasoning service performs inference tasks based on ontologies and rules and makes health and wellness recommendations to user applications through an API. The recommendations result from a rule set, such as that in Table 1, which is drawn from publicly available Finnish healthcare guidelines. The rules infer a person's overall health, diabetes risk, and stress level from data fetched from multiple data sources. The semantic reasoning service can be maintained by officially authorized organizations that

TABLE 1. Sample rules for inferring health-related conditions in the semantic reasoning service.

Fact	Clause
TotalExercise	Exercise hasTimeStamp between(x,y) ∩ hasDuration ?d ➞TotalExercise hasDuration sum(?d) ∩ hasMeasurementDuration(y-x)
LowExerciseAmount	TotalExercise hasDuration ?d ∩ hasMeasurementDuration ?md ∩ ?d/?md < 0.04 ➞ LowExerciseAmount
EnoughIntenseExercise	Exercise rdf:type IntenseExercise ∩ hasTimeStamp between(x,y) ∩ sum(hasDuration)/hasMeasurementDuration > 0.0074 ➞ EnoughIntenseExercise
BMIIndex	Person hasWeight ?w ∩ hasHeight ?h ➞ Person hasBMI (?w/?h2)*703
Obesity	BodyMassIndex > 29.9 ➞ Obesity
EfficientSleep	SleepEfficiency > 84 ➞ EfficientSleep
OptimalBP	SystolicBloodPressure < 120 ∩ DiastolicBloodPressure < 80 ➞ OptimalBP
HypertensionDegree1	159 > SystolicBloodPressure > 140 ∩ 99 > DiastolicBloodPressure > 90 ➞ HypertensionDegree1
DiagnosedHypertension	(HypertensionDegree1 ∪ HypertensionDegree2 ∪ HypertensionDegree3) hasTimestamp between(x,y) ∩ avg(hasSystolic) > 140 ∩ avg(hasDiastolic) > 90 ➞ DiagnosedHypertension
UnhealthyDiet	Purchases hasTimestamp between(x,y) ∩ rdfs:subClassOf Fruits_Berries_Vegetables count+1 ∩ count < 2TimesPerWeek ➞ UnhealthyDiet
VeryHighType2DiabetesRisk	Age>64 ∩ Obesity ∩ DiagnoseHighBP ∩ NotEnoughIntenseExercise ∩ NotEnoughModerateExercise ∩ FamilyMember hasDiagnosedDiabetes ∩ HighBloodGlucose ∩ UnhealthyDiet ➞ VeryHighType2DiabetesRisk
OptimalHealth	NormalBMI ∩ (EnoughIntenseExercise ∪ EnoughModerateExercise) ∩ (NormalBP ∪ OptimalBP) ∩ EfficientSleep ➞ OptimalHealth
Stressed	(HypertensionDegree1 ∪ HypertensionDegree2) ∩ InefficientSleep hasTimestamp between(x, y) ➞ Stressed
ReduceTraining	Underweight ∩ HighExerciseAmount ∩ Stressed ➞ ReduceTraining
HealthyDiet	Overweight ∩ LowExerciseAmount ➞ Healthy Diet ∩ MoreTraining

guarantee the validity of a data-driven reasoning service for third-party applications that interact with the user, such as a FitBit for fitness data.

Figure 5 depicts the MyData core operations for establishing trust between the components that together realize the recommendation service. Both the health application and Semantic Reasoner have consent tokens to establish the trust required for data exchange. In the scenario in Figure 5, Semantic Reasoner separates user applications from the aggregated personal data and helps preserve the original data from exploitation by third-party applications.

As this implementation shows, PRIAAS and MyData facilitate and expedite the creation of new services, create data bindings between actors, and promote new business models to reuse-refine-reuse personal data within both the individual and group domains, with the goal of creating novel services and applications that enhance well-being.

Together, PRIAAS and MyData provide a holistic solution for consent delivery and management. Individuals have tools to manage their data as well as innovative services. Companies benefit from the new data-based business

SMART HEALTH AND WELL-BEING

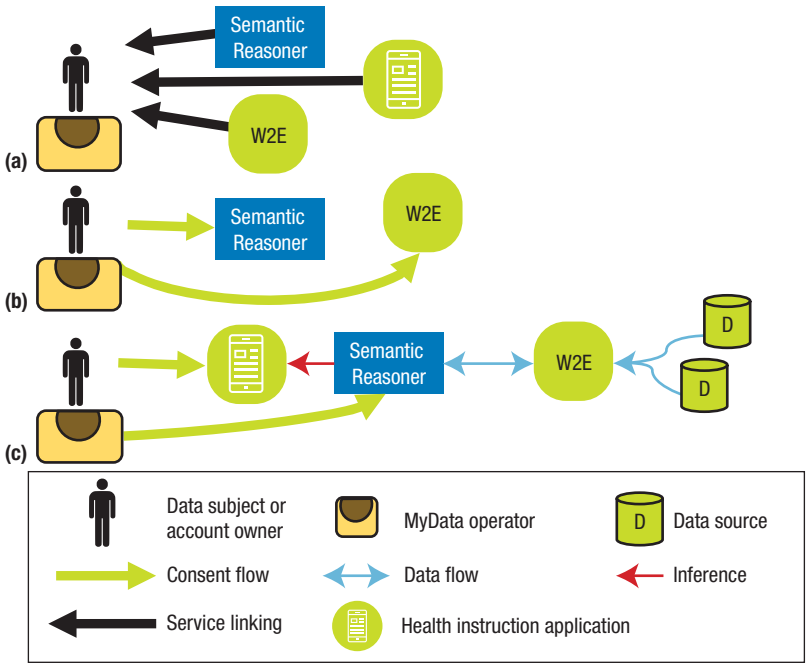


FIGURE 5. MyData operations for inferring health-related conditions from wellness data. The process to establish inference operations has three steps in which the user (a) links Semantic Reasoner (a health application) and the W2E aggregator service to his MyData operator account; (b) authorizes Semantic Reasoner to access his health data from the W2E aggregator service; and (c) authorizes the linked health application to use data from Semantic Reasoner for health-related guidance.

opportunities, and standardization enables interoperability and lowers the barrier for new companies and businesses to enter the healthcare-support market. Society benefits from the new services as well as standardized structures, processes, and policies that address individual rights to control data use.

PRIAAS and MyData focus on consent management for two reasons. First, consents are the backbone of any legislative framework that defines information processing from a human-centric perspective. Second, standardized consent that is both human and machine readable unites data management systems, legislative frameworks, and individual needs. These reasons suggest applications beyond healthcare. Indeed, with minor modifications, PRIAAS and MyData could be the basis for a consistent data collection and processing approach regardless of domain.

ACKNOWLEDGMENTS

This research has been supported by a grant from Tekes—the Finnish Funding Agency for Innovation as part of the Digital Health Revolution program.

REFERENCES

1. "Reform of EU Data Protection Rules," European Commission, 2 Aug. 2016; ec.europa.eu/justice/data-protection/reform/index_en.htm.
2. A. Poikola, K. Kuikkaniemi, and H. Honko, *MyData—A Nordic Model for Human-Centered Personal Data Management and Processing*, white paper, Open Knowledge Finland, Finnish Ministry of Transport and Communication, 2015; urn.fi/URN:ISBN:978-952-243-455-5.
3. J. Kaye et al., "Dynamic Consent: A Patient interface for Twenty-First Century Research Networks," *European J. Human Genetics*, vol. 23, no. 2, 2015, pp. 141–146.
4. T. Hardjono et al., *User-Managed Access (UMA) Profile of OAuth*, v. 2.0; specification by User-Managed Access Work Group, Kantara Initiative, 28 Dec. 2015; docs.kantarainitiative.org/uma/rec-uma-core.html.
5. *Kantara Initiative: Minimum Viable Consent Receipt Specification*, v. 0.7, 2015; github.com/KI-CISWG/MVCR.
6. R. Arnold, A. Hillebrand, and M. Waldburger, *Personal Data and Privacy*, Ofcom, 2015; stakeholders.ofcom.org.uk/binaries/internet/personal-data-and-privacy/Personal_Data_and_Privacy.pdf.
7. R. Roache, "Why Is Informed Consent Important?" *J. Medical Ethics*, vol. 40, no. 7, 2014, pp. 435–436.
8. J. Walker et al., "The Value of Healthcare Information Exchange and Interoperability," *Health Affairs*, vol. 24, no. 1, 2005; content.healthaffairs.org/content/early/2005/01/19/hlthaff.w5.10.
9. R. Steinbrook, "Personally Controlled Online Health Data—The Next Big Thing in Medical Care?," *New England J. Medicine*, vol. 358, no. 16, 2008, pp. 1653–1656.

ABOUT THE AUTHORS

XIANG SU is a postdoctoral researcher in computer science in the Center of Ubiquitous Computing at the University of Oulu. His research interests include semantic technologies, the Internet of Things (IoT), knowledge representations, and context modeling and reasoning. Su received a PhD in technology from the University of Oulu. He is a member of IEEE. Contact him at xiang.su@ee.oulu.fi.

JARKKO HYYSALO is a postdoctoral researcher in the Faculty of Information Technology and Electrical Engineering at the University of Oulu. His research interests include software architectures and processes and collaborative work. Hyysalo received a PhD in information processing science from the University of Oulu. Contact him at jarkko.hyysalo@oulu.fi.

MIKA RAUTIAINEN is a postdoctoral researcher in the Faculty of Electrical and Information Engineering at the University of Oulu. His research interests include content-based multimedia retrieval and management systems, scalable data-processing architectures, cognitive user experience, pattern recognition, digital image and video processing and understanding. Rautiainen received a PhD in technology in computer science from the University of Oulu. Contact him at mika@valossa.com.

JUKKA RIEKKI is a professor of software architectures for embedded systems and dean of the Faculty of Information Technology and Electrical Engineering at the University of Oulu, and a principal investigator in the university's Center of Ubiquitous Computing. His research interests include interactive, context-aware systems that support everyday tasks and edge computing driven by the IoT. Riekki received a PhD in technology from the University of Oulu. He is a member of IEEE. Contact him at jukka.riekki@oulu.fi.

JAAKKO SAUVOLA is a professor of data-intensive systems and advanced software at the University of Oulu and leader of Finland's High-Tech ICT Leverage from Long-Term Assetization (HILLA) Program. His research interests include mobility, system architectures, and data-intensive services and analytics. Sauvola received a PhD in technology from the University of Oulu. Contact him at jaakko.sauvola@oulu.fi.

ALTTI ILARI MAARALA is a doctoral student in the Department of Computer Science at Aalto University. His research interests include the IoT, big data, semantic technologies, knowledge representation, parallel algorithms, and computational genomics. Maarala received an MSc in computer science from the University of Oulu. Contact him at ilari.maarala@aalto.fi.

HARRI HIRVONSALO is an MSc student in the Faculty of Information Technology and Electrical Engineering at the University of Oulu. His research interests include software architectures, security and privacy, and personal data and identity management. Contact him at harri.hirvonsalo@oulu.fi.

PINGJIANG LI is an MSc student in the Center of Ubiquitous Computing at the University of Oulu. His research interests include ubiquitous computing and human–computer interaction. Li received a BS in computer science from the University of Jinan. Contact him at pingjiang.li@student.oulu.fi.

HARRI HONKO is a track lead in technology and regulation in the digital health revolution program and project manager in the Personal Health Informatics group at Tampere University of Technology. His research interests include identity, authorization, and consent management technologies; system architectures; and regulation for personal data management. Honko received an MSc in electronic engineering from Tampere University of Technology. He is a society affiliate of IEEE. Contact him at harri.honko@tut.fi.

COMPUTING PRACTICES

Architectural Approaches to Security: Four Case Studies

Humberto Cervantes, Metropolitan Autonomous University
Rick Kazman, University of Hawaii
Jungwoo Ryoo, Pennsylvania State University
Duyoung Choi, CodeOne
Duksung Jang, Keimyung University

An examination of the security approaches in industrial and open source projects shows that a strategic, systemwide architectural approach, implemented as a security framework or as a platform built using these frameworks, results in the highest security and lowest maintenance costs.

Software security is a complex multidimensional problem that touches coding, design, operation, and policy, yet security has not been a focus of software development. Typically, architects and developers focus on functionality first, and security is often applied as a band-aid solution after an application has been developed—the development team basically throws the code over the wall to a security team.

When development does take security seriously, most of the focus is on secure coding.[1] Although secure coding practices are clearly a critical step in producing secure software, these practices alone do not scale well for two reasons. The first is that security is a weakest-link phenomenon. Even if the vast majority of a system's code is secure, with only a tiny fraction containing vulnerabilities, the system is not secure. In other words, any exploitable vulnerability can compromise the entire system's security. Secure coding practices are also expensive, requiring careful attention not only to code, but also to extensive testing, inspections, and scanning.

For these reasons, we advocate an architectural approach to software security. It is true that architecture alone is insufficient, just as coding or process alone is insufficient to ensure secure software. However, we believe that architecture is central to defeating the problems of security being a "weakest link" phenomenon, and secure coding practices being a source of great expense in current security practices. Security is an architectural issue—its consequences are systemwide—and thus only a systemwide approach can cost-effectively address it.[2] Other security aspects, such as choosing strong passwords, keeping system software patches up to date, and training users and

0018-9162/16/$33.00 © 2016 IEEE

administrators, will be for naught if the underlying system is insecure.

To ensure that a system is secure, it must be properly designed with security in mind, which to our mind means using security frameworks early in the development project. Frameworks—whether commercial, open source, or homegrown—encapsulate best practices in design and coding in a reusable package. They are an efficient and economical way to realize architectural design intent[3] because they make it easier to approach a quality attribute concern, such as security, consistently across the system.

To support our assertion that frameworks are essential to a secure system, we studied three industrial projects and one open source project to gather empirical evidence about security practices. The studies include examples of web-based systems that were built both with and without frameworks, which provide insights into how security frameworks affect both system quality and development efficiency. Our main objectives were to understand how secure these systems were, how difficult and costly it was to create the security controls, and how much it would cost to maintain them in the future.

SECURITY DESIGN CONCEPTS

An architectural approach to software security relies on tactics, patterns, and frameworks, each of which can be employed during architectural design.

Tactics

Architectural tactics are techniques that an architect can apply to achieve quality attributes in a system—as such, they are foundational design decisions. Figure 1 shows security-related tactics from a catalog that lists the seven most important and widespread quality attributes.[2]

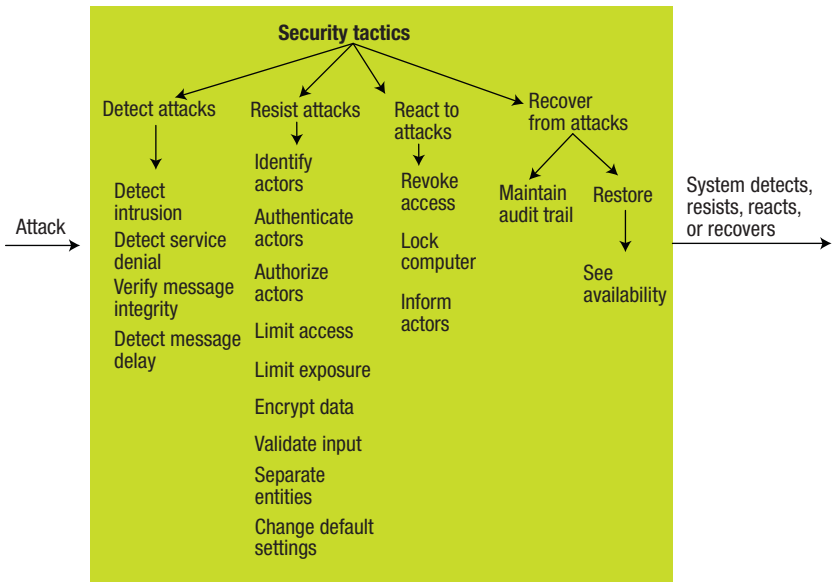


FIGURE 1. Security tactics categorization. This ontology describes the spectrum of architectural approaches to security, which can help software architects reason about incorporating security into system design early in development.

The security tactics listed are categorized into four groups, which represent the four major areas of concern for an architect aiming to design a secure system: how to detect that an attack is occurring, how to resist when you are under attack, how to react once you have been attacked, and how to recover from successful attacks. Within each of these categories are specific tactics. For example, tactics to detect attacks include

- *detect intrusion*, in which service request patterns are compared to a set of signatures of malicious behavior signatures;
- *detect service denial*, in which the pattern of incoming network traffic is compared to profiles of known denial-of-service (DoS) attacks;
- *verify message integrity*, in which checksums or hash values are used to verify the integrity of messages and files; and
- *detect message delay*, in which checks are performed on the time it takes to deliver a message, which can reveal any suspicious timing behavior.

The other three groups have similar tactics lists, the details of which are readily available.[2,4]

Tactics catalogs provide the architect with a conceptual framework

for reasoning about qualities such as security. These tactics can be used as a design or analysis checklist to ensure that different areas of security have been addressed. The catalogs do not describe how to implement the tactics, which is typically done through frameworks, patterns, or coding.[5]

Patterns

Design patterns are proven conceptual solutions to recurring design problems,[6] and myriad pattern varieties are described in catalogs. There are design patterns, architectural patterns, and anti-patterns. There are catalogs of patterns for specific quality attributes: availability, interoperability, performance, usability, security, and so on. Catalogs related to security are both technology-agnostic and technology-specific. One technology-agnostic pattern catalog describes 70 security patterns for concerns such as identity management, authentication, access control, secure process management, web services security, and cloud computing.[7] Another catalog, which is Java specific, describes 23 security patterns associated with different application layers.[8]

Patterns are a powerful tool for designing architectures, but their use can be problematic. The daunting number of available patterns makes it difficult to choose a pattern, for example. It can also be challenging to combine the chosen patterns into a single architecture, although the literature provides some methodological support to guide architects in choosing, tailoring, combining, and realizing patterns.[5–9]

A final challenge stems from the idea that patterns are underspecified. Consequently, developers often inadvertently undermine the pattern's intent.[10] Once a pattern has been selected, it must be instantiated, and, if the pattern is technology agnostic, it must be matched to a technology. These extra steps, which are needed to instantiate patterns, can require substantial design, development, and quality-assurance effort.[4,9]

Frameworks

A framework is a reusable software element providing generic functionality that addresses recurring concerns across a broad application range.[3] Frameworks generally implement several patterns and tactics.

Many frameworks in various languages address security concerns. Examples of Java frameworks that address concerns such as authentication and authorization include

- Spring Security (projects.spring.io/spring-security), an authentication and access control framework for securing Spring-based applications;
- Apache Shiro (shiro.apache.org), an authentication and authorization framework with cryptography and session management features; and
- JGuard (jguard.xwiki.com), an extension of Java Authentication and Authorization Service (JAAS) that provides more developer-friendly interfaces.

Other concerns, such as cryptography, are addressed by frameworks such as BouncyCastle (www.bouncycastle.org) and Jasypt (www.jasypt.org). The Java platform itself provides the JAAS security API (www.oracle.com/technetwork/java/javase/jaas/index.html). Other Java frameworks do not specialize in security, but they do address it nonetheless. An example is ZK (www.zkoss.org), a Web-based user interface framework that protects against cross-site scripting, denial-of-service, and cross-site request forgery attacks.[11]

We chose to describe Java security frameworks, but similar frameworks exist for other programming languages. Regardless of the language chosen, frameworks increase the programmer productivity by allowing programmers to focus on what they do best—coding business logic and creating end-user value—rather than worrying about the underlying technologies. There are, however, some downsides associated with framework use. Framework selection can be difficult, and the learning curve can be steep. Frameworks can be difficult to modify, and might give rise to lock-in concerns, as there is little control over how the framework evolves or whether it survives.

CASE STUDIES

Given this wealth of design concepts, we were interested in understanding how practicing architects approach security, and how well these design approaches secure the system and reduce the cost of creating and maintaining a secure architecture.

Objectives and procedure

For each of the four case studies, we attempted to assess the consequences of a strategic, architectural approach to security. Our goal was to explore the tradeoff space between the costs and benefits (effectiveness) of different architectural approaches to security and to determine if optimal project strategies exist for using them.

> WE WANTED TO UNDERSTAND HOW PRACTICING ARCHITECTS APPROACH SECURITY AND TO GAUGE THEIR METHODS' EFFECTIVENESS AND COST.

Our protocol was to interview the system's lead architect and to scan the system for vulnerabilities. Combined, these steps gave us information about software size, the estimated effort expended on security, and the number of vulnerabilities in the system.

Interviews. We interviewed architects about the approach to security, the code-base size, and the effort spent on security, asking five main questions:

- What were your primary drivers, that is, your system quality attributes, and how important is security among them?
- With respect to security, what approaches have you taken to address this quality attribute?
- How do you reason about tradeoffs?
- How did you ensure that your programmers conform to the security approaches? That is, what are your policies, and do you conduct code inspections and so on?
- What percentage of project effort do you estimate goes into security?

The interview questions were open-ended, but the discussion of security approaches was guided by the taxonomy of security tactics in Figure 1. We asked the last question regardless of whether the project had involved a security framework.

System scans. We used IBM's AppScan (www.ibm.com/software/products/us/en/appscan) to scan the system to identify its vulnerabilities. We chose AppScan in part because it can perform automated dynamic analyses, which probes the running application in a way similar to what a hacker would do. During our tests, we asked that existing firewalls be disabled to ensure that only the software of interest to us was tested. If a firewall were present, the requests from AppScan would potentially be blocked, which would prevent us from effectively analyzing the application's vulnerabilities.

Clearly, security can be measured in many ways, which motivated our use of a single scanner. AppScan, being fully automated, was an objective measurement tool and thus removed all subjectivity from this part of our assessment.

Projects

Our projects were a mix of locations, domains, and industrial and open source. However, all four systems we evaluated are web-based and follow a traditional three-tier client–server architecture.[2]

ACME. Based in Seoul, Korea, CodeOne (www.code1.co.kr) is a company specializing in software security. Its core business is to identify vulnerabilities in web applications, either by penetration testing or by using commercial vulnerability-assessment tools, and to help address any discovered vulnerabilities, primarily by using its homegrown proprietary security framework.

CodeOne's customers include retailers, banks, universities, and cities in the Republic of Korea. CodeOne periodically scans the websites of these organizations and removes vulnerabilities once they are detected. A typical website under Code One's scrutiny provides dynamic web content and consists of web front ends and database back ends. Many of the organizations outsource website development, and in CodeOne's experience these third-party developers are not particularly security conscious as a rule.

For the case study, we used data collected from scanning a website that represents the worst-case scenario but has a typical size and complexity—the kind CodeOne faces on a daily basis. When developing its site, the customer (which we call ACME, a fictitious name) initially paid no attention to security and later enlisted CodeOne to improve the architecture to address security.

The ACME web application was built with a combination of Java server pages and HTML. We were provided data for two key periods of the application's development: ACME Before and ACME After. ACME Before represents the application before the CodeOne security framework was introduced, and ACME After is the application with framework use. We interviewed CodeOne's chief software architect, who was closely involved in the vulnerability detection and removal processes for ACME.

Quarksoft. Based in Mexico City, Quarksoft (www.quarksoft.net) is a company that builds custom software for various domains, including government, banks, and insurance companies. Software developers use the Team Software Process, so the company routinely collects development data from its software engineers.

The web application analyzed for this case study is an internal tool for collecting development data and tracking project progress. We interviewed the tool's software architect. At the time of the study, the tool was being deployed for a pilot team in the organization. The tool was built in Java, and its security controls are implemented primarily through the

TABLE 1. Comparative data from the case studies.

Metric	Case study				
	ACME Before	ACME After	Quarksoft	OpenEMR	Web portal
Approach followed	No adoption	Partial adoption (CodeOne framework)	Full adoption (ZK and Spring frameworks)	No adoption	Full adoption (Extension of SWB core)
Primary development language	Java	Java	Java	PHP	Java
Size (KLOC)	7.93	8.55	16.56	255.6	105.6
Detected security issues (vulnerabilities)*	H: 154 M: 50 L: 99 I: 224 Total: 527	H: 0 M: 25 L: 99 I: 224 Total: 348	H: 0 M: 0 L: 0 I: 1 Total: 1	H: 8 M: 9 L: 475 I: 52 Total: 544	H: 0 M: 0 L: 370 I: 542 Total: 912
Analyzed URLs	756	756	24	3,497	413
Number of threat classes	9	8	1	8	9
Number of security tactics addressed	6/17	12/17	13/17	9/17	8/17
Number of security tactics addressed within application logic	5/17	5/17	0/17	6/17	2/17
Estimated effort for security (% of total project effort)	20	10	3	20	8

* High (H), medium (M), low (L), or informational (I) in severity

use of the Spring Security framework. The tool also uses the ZK framework in the presentation layer, which addresses several security concerns.

OpenEMR. OpenEMR (www.open-emr.org) is an open source electronic medical record (EMR) management system. It is used by many physicians and medical organizations that cannot afford OpenEMR's commercial counterparts. One of the lead architects of the OpenEMR project is a physician who ensures OpenEMR's relevance to its users. We have been participating in the OpenEMR project by providing recommendations on how to improve its software security.

The OpenEMR application is built using PHP and, at the time of the interview, no security framework was being employed. Only common library functions that address a limited set of security vulnerabilities, such as those susceptible to SQL injections, were used. For this case study, we interviewed OpenEMR's chief architect.

Web portal. For our fourth case study, we evaluated a publicly accessible web portal from an institution that requested anonymity, and we interviewed the portal architect. The portal was built using Semantic Web Builder (SWB; www.semanticwebbuilder.org.mx/swb/swb/SWB_Portal), a Java-based content management system (CMS) platform. The CMS platform handles security aspects that include authentication and authorization, as well as protection against SQL injection, cross-site scripting, and cross-site request forgery. The CMS is built using a number of Java frameworks, including Spring Security. The portal was developed as a set of extensions to the CMS' platform core.

COMPARATIVE RESULTS

Table 1 shows the results of our case studies. AppScan categorizes security vulnerabilities as having high (H), medium (M), low (L), or informational (I) severity. It also reports the number of analyzed URLs, the percentage that are vulnerable, and the number of threat classes (different vulnerability categories). Unfortunately, its criteria for classifying security vulnerabilities are proprietary, so we used the criterion established by CodeOne: an application is secure if it has zero high-severity vulnerabilities. In addition to the scan, we measured application size in KLOC.

Interview responses

Interviewees estimated the security effort, and described their use of the security tactics in Figure 1. We devoted most of the interview time to understanding the approaches the architect took with respect to security. We asked architects whether they considered each tactic and, if so, how it was implemented. We have room to give only a flavor of the responses (the interviews' full text and results are available at sites.psu.edu/interviewtactics).

For example, both CodeOne and QuarkSoft enforce tactics to detect and resist attacks, some of which include

- *Detect intrusion*. The CodeOne framework detects intrusion attempts through monitoring and provides visibility into attack attempts. The company includes real-time monitoring to report attack attempts. Quarksoft enforces intrusion detection primarily through firewall

appliances and uses Spring Security to guarantee that a session comes from a single source.
- *Verify message integrity.* The CodeOne framework supports this by associating all requests with a session. The company focuses on preventing session hijacking and cross-site scripting attacks.
- *Detect message delay.* In Quarksoft, this tactic is covered by ZK, which creates many short-lived objects when a session is initiated. Each object has a user ID, which the framework verifies, making it hard to replicate the IDs.
- *Authenticate actors.* This tactic is not covered by the CodeOne framework but it is implemented as part of ACME Before's application logic. Quarksoft enforces this tactic through Spring Security, which handles all URLs. Content transmission is ZK's responsibility.

Framework adoption

The case studies represent three approaches that differ in their degree of security framework adoption.

Full adoption. An organization adopts security frameworks from the beginning of their development process. Quarksoft falls into this category. The web portal also falls in this category since the CMS was selected from the beginning of their development, and one of the criteria used for selecting the CMS was its support for security.

As Table 1 shows, the full adoption approach provides the best results. For Quarksoft, AppScan identified 0 high-severity vulnerabilities, and cost was just 3 percent of the total project effort. For this case, 13 of the 17 tactics were covered, and none of them was addressed inside the application logic. The AppScan operators said that they were impressed by the level of security exhibited by this application.

AppScan also identified 0 high-severity vulnerabilities for the Web portal. The architect estimated that the effort dedicated to security was 8 percent of the total project effort, which is notably smaller than the efforts estimated for the no adoption and partial adoption case studies. In this case study, security was one of the primary drivers. Although the CMS handled part of the security, the portal architect described design decisions that supported security, such as the application's logical and physical structuring—decisions that represented a nontrivial effort. AppScan did identify many low and informational issues, but the architect determined that many of these issues stemmed from a single problem in a page section embedded by the CMS in most of the system's webpages. The problem proved easy to correct.

Partial adoption. In partial adoption, a security framework is introduced in the middle of the application's lifecycle. For ACME After, which falls in this category, security is excellent, with 0 high-severity vulnerabilities, but the cost of introducing security countermeasures is high. In this case, the cost was 10 percent of the total project effort. For this case, 12 of 17 tactics were employed but, despite the use of the CodeOne framework, only 5 of these tactics were implemented in the application logic. ACME After also has several medium- and low-severity vulnerabilities because, after conducting a risk analysis, they decided to remove only the most critical ones.

No adoption. A security framework is not used. ACME Before and OpenEMR belong in this category because neither employed any security framework but rather addressed security purely within the application logic using ad hoc coding solutions.

The data in Table 1 clearly shows that no adoption results in the worst security. ACME Before exhibited 154 high-severity vulnerabilities scattered across nine threat classes. Although OpenEMR had only 8 high-severity vulnerabilities, they were serious and persistent—the project has scanned its code since 2009, identifying security issues such as vulnerability to SQL injection and cross-site scripting, but it has been unable to eradicate these issues. OpenEMR implemented 9 of 17 tactics, but 6 of these were implemented in the application logic, not systematically.

For both ACME Before and OpenEMR, the cost of dealing with security issues has been considerable—an estimated 20 percent of total project effort.

VALIDITY OF RESULTS

Although our results are encouraging, we identified three threats to results validity.

Single scanner

We used only AppScan to assess security, and, although AppScan is the security industry's gold standard for vulnerability assessments, this choice could have unfairly biased our results. Also, because we could not establish what AppScan considered to be a secure application, we had to rely on CodeOne's criteria.

Effort estimation

There was no hard data to support the architects' effort estimates. However, judgments were consistent across the architects, which gives us more confidence in their estimates. The cases that did not use a framework-based approach to security estimated that they spent 20 percent of the total project effort on security, and the Quarksoft architect guessed that, without frameworks, they would have spent 30 percent. In the partial adoption case, the architect estimated that 10 percent of the project effort went to security. And in the two full adoption cases, the architects estimated that 3 to 8 percent was devoted to security.

Study number and type

Four case studies is a relatively small number, and all evaluations were of web-based systems. However, the studies involved different application domains, system sizes, and organizations, and projects were both industrial and open source. Consequently, we believe that the four studies represent a reasonable coverage of security issues faced by many real-world projects. Given that the results were mostly consistent, we feel confident in generalizing from them. However, our results might not apply to other system types, such as embedded systems.

WHY FRAMEWORKS?

We believe it is best to address security through frameworks for four main reasons:

- Application developers might be experts in their domains, but they are typically not security experts. Security framework developers, on the other hand, focus purely on information security.
- Even if developers have security experience, they should not write their own security controls.[12]
- Framework use increases the likelihood that security controls will be applied consistently across the application.
- Delegating security issues to frameworks allows developers to devote their energy to application logic, which increases overall productivity.

The last reason was most evident in the Quarksoft study. The architect was more motivated to address other quality attributes and hence did not spend much effort on security. Even so, he managed to create a system that was impressively secure because he delegated security considerations to the frameworks and ensured that they were applied consistently.

Although we continue to advocate the consistent use of security frameworks, they do have an important caveat: they must be maintained to keep pace with the ever-changing threat environment. Thus, the architect must pay continuous attention not only to application maintenance, but also to the maintenance of the security framework and its evolutionary path. Consequently, we recommend the full adoption of security frameworks combined with a maintenance program that focuses on updating the frameworks regularly.

> THESE CASE STUDIES PROVIDE CLEAR EVIDENCE OF THE SUPERIOR SECURITY OUTCOMES THAT RESULT FROM USING FRAMEWORKS.

The case studies we conducted provide clear evidence of the superior security outcomes that result from using security frameworks as an architectural approach, through either partial or full adoption. The effort required for partial adoption—introducing frameworks after the system is built—is significantly more than for full adoption—using frameworks or a platform built on frameworks from the beginning. Partial adoption is a suboptimal practice, but we have found that it is the most common way of adopting security frameworks. Most developers and architects worry about functionality first and security (and other qualities) later. The good news is that once the architecture has been refactored to use frameworks consistently, better security results are obtained. It is certainly better than no adoption, with security handled in an ad hoc fashion. This is typically an inefficient use of

programmers' time, and a system using this approach is, in our observations and experience, more vulnerable to security threats.

Thus, we recommend the use of security frameworks from the early phases of system construction. Not only will security be higher, but costs will be lower, as adopting a framework after the system has been built will undoubtedly take more resources than doing so from the very beginning.

REFERENCES

1. R. Seacord, *Secure Coding in C and C++*, 2nd ed., Addison-Wesley, 2013.
2. L. Bass, P. Clements, and R. Kazman, *Software Architecture in Practice*, 3rd ed., Addison-Wesley, 2012.
3. H. Cervantes, P. Velasco, and R. Kazman, "A Principled Way of Using Frameworks in Architectural Design," *IEEE Software*, vol. 30, no. 2, 2013, pp. 46–53.
4. J. Ryoo, R. Kazman, and P. Anand, "Architectural Analysis for Security," *IEEE Security and Privacy*, vol. 13, no. 6, 2015, pp. 52–59.
5. H. Cervantes and R. Kazman, *Designing Software Architectures: A Practical Approach*, Addison-Wesley, 2016.
6. F. Buschmann, K. Henney, and D. Schmidt, *Pattern-Oriented Software Architecture: A Pattern Language for Distributed Computing*, vol. 4, John Wiley & Sons, 2007.
7. E. Fernandez-Buglioni, *Security Patterns in Practice: Designing Secure Architectures Using Software Patterns*, Wiley, 2013.
8. C. Steel, R. Nagappan, and R. Lai, *Core Security Patterns: Best Practices and Strategies for J2EE, Web Services, and Identity Management*, Prentice Hall, 2012.
9. J. Ryoo, P. Laplante, and R. Kazman, "Revising a Security Tactics Hierarchy through Decomposition, Reclassification, and Derivation," *Proc. IEEE Int'l Conf. Software Security and Reliability* (SERE 12), 2012, pp. 85–91.
10. Y. Cai et al., "Introducing Tool-Supported Architecture Review into Software Design Education," *Proc. IEEE Conf. Software Engineering Education and Training* (CSEE&T 13), 2013, pp. 70–79.
11. "Security Features of ZK Framework," Potix Corp., 2013; www.zkoss.org/_w/images/e/ea/ZK_Security_Report.pdf.
12. M. Schwartz, "6 Ways to Strengthen Web App Security," *InformationWeek*, Sept. 2012; www.informationweek.com/security/application-security/6-ways-to-strengthen-web-app-security/240006962.

ABOUT THE AUTHORS

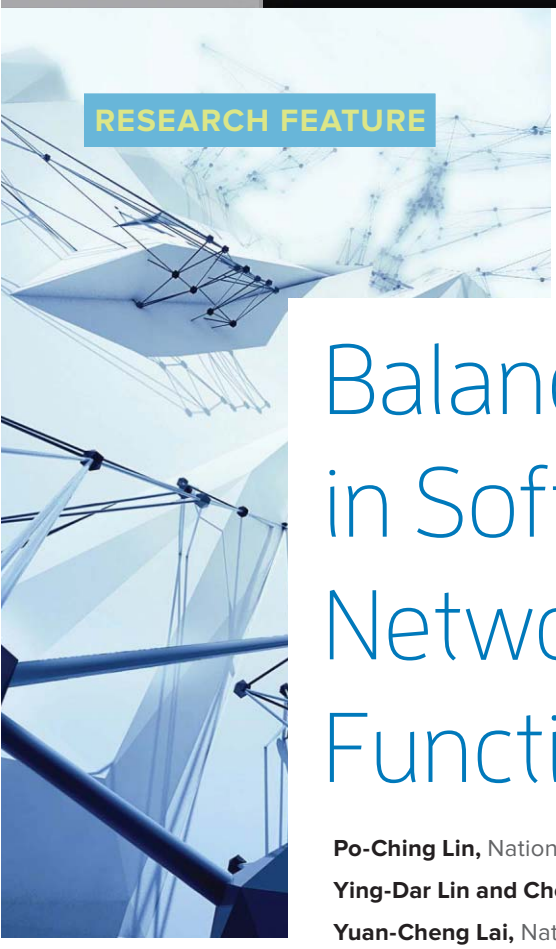
HUMBERTO CERVANTES is a professor of software engineering at Universidad Autónoma Metropolitana–Iztapalapa. His research interests include software architecture design methods and their adoption in industrial settings. Cervantes received a PhD in computer science from Université Joseph Fourier. He is a member of IEEE. Contact him at hcm@xanum.uam.mx.

RICK KAZMAN is a professor of information technology management at the University of Hawaii and a principal researcher in the Software Engineering Institute at Carnegie Mellon University. His research interests include software architecture, design and analysis tools, and software engineering economics. Kazman received a PhD in computational linguistics from Carnegie Mellon University. He is a member of IEEE and chair of the IEEE Technical Council on Software Engineering. Contact him at kazman@hawaii.edu.

JUNGWOO RYOO is a professor of information sciences and technology at the Pennsylvania State University. His research interests include information security and assurance, software engineering, and computer networking. Ryoo received a PhD in computer science from the University of Kansas. He is a member of IEEE. Contact him at jryoo@psu.edu.

DUYOUNG CHOI is director of research at CodeOne and a PhD student in the department of Computer Engineering at Keimyung University. His research interests include software security, natural language processing, and systems software. Choi received an MS in computer science from Daegu University. He is a member of the Korea Information Processing Society and the Korea Institute of Information Security and Cryptography. Contact him at dychoi@cit21.com.

DUKSUNG JANG is a professor in the Department of Computer Engineering at Keimyung University. His research interests include software engineering, natural language processing, and functional programming languages. Jang received a PhD in computer engineering from the Seoul National University. He is a member of the Korean Embedded Engineering Institute and the Korean Multimedia Engineering Society. Contact him at dsjang@kmu.ac.kr.



RESEARCH FEATURE

Balanced Service Chaining in Software-Defined Networks with Network Function Virtualization

Po-Ching Lin, National Chung Cheng University
Ying-Dar Lin and Cheng-Ying Wu, National Chiao Tung University
Yuan-Cheng Lai, National Taiwan University of Science and Technology
Yi-Chih Kao, National Chiao Tung University

Balanced Hash Tree (BHT) is a mechanism for service function (SF) chaining, service routing, and traffic steering that enables switches to select SF instances for load balancing without involving the controller. In an experimental evaluation, BHT decreased packet-in message-processing time by 92.5 percent and achieved near-perfect load-balancing performance.

Network architects often rely on a software-defined network (SDN) to decouple the control plane from the data plane for higher programmability.[1] They can then use network function virtualization (NFV) to extend the data plane to outsource service functions (SFs), which allows service providers to deploy virtual SFs on commodity servers instead of on specialized hardware. Virtualization facilitates the dynamic instantiation of SFs to meet resource requirements from input traffic, which effectively lowers deployment cost.

However, this strategy puts an extreme burden on the centralized controller in a large datacenter network, which manages a set of switches through a southbound interface such as that defined by the Open Networking Foundation's OpenFlow specification.[2] An SDN provides flexible connectivity between switches and hosts, but providing value-added services exclusively through the controller is not a scalable solution, and passing packets from the switches to the controller will incur excessive communication overhead.[3,4] The controller can chain SFs—order their sequence and bind them together in a group—by configuring the data plane to support NFV. However, load balancing with SF chains can rapidly overwhelm the controller in such a large network.

To address this issue, we developed Balanced Hash Tree (BHT), a load-balancing mechanism for SF chaining, service routing, and traffic steering. Rather than

0018-9162/16/$33.00 © 2016 IEEE

BHT BALANCES CHAINING BY REDIRECTING INCOMING FLOWS TO THE OPTIMAL SERVICE FUNCTION WHILE MITIGATING THE CONTROLLER'S WORKLOAD.

relying on the controller, BHT implements load balancing on the switches through the *select* group table, as described in the OpenFlow specification, which requires processing to be based on a switch-computed selection algorithm.[2] In BHT, switches use a hashing-based algorithm to determine the output port for load balancing among an SF's instances, and the switch uses the select group type to assign each flow to an action bucket. Flow distribution is similar to the well-known equal-cost multipath routing (ECMP) strategy, which balances loads using multiple equal-cost paths between two neighboring hops (en.wikipedia.org/wiki/Equal-cost_multi-path_routing). However, BHT is different because it also considers service chaining between SFs and can assign weights to different paths.

BHT has proven to be a successful alternative to load balancing through the controller—an approach taken by many existing load-balancing schemes. In a performance evaluation, BHT reduced packet-in message-processing time by 92.5 percent and its load-balancing performance was within 2.4 to approximately 5 percent of perfect.

SERVICE FUNCTION CHAINING AND LOAD BALANCING

A service chain, which is an ordered sequence of SFs, is typically used to build a required network service. For example, to configure web traffic to go through a firewall, an administrator can use SF chaining to combine and order intrusion-prevention and load-balancing functions; the controller can then instruct the switches to redirect traffic through the SFs in that chain.[5] The Internet Engineering Task Force (IETF) defines the architecture for SF chaining,[6] but in simple terms, for each SF chaining path, a tunnel is established between the roots of the trees of any two successive SFs—meaning that the entire SF chaining path consists of multiple tunnels.

Balancing through switches

OpenFlow switches process incoming traffic through a pipeline of one or more flow tables and determine the actions on a packet by matching the entries in the flow tables. If a switch cannot match any of the flow entries, it either drops the packet or sends a packet-in message to the controller to ask how to process the packet. In the latter case, the controller sends the switch a packet-out or a flow-modify message telling it how to process the packet and configure the flow table. The multipart message has the added benefit of allowing the controller to collect information from the switches, such as port statistics.

When a matched flow entry's actions specify that a packet should go to a specific group entry, the packet must be processed according to the actions specified in the group entry's action buckets, which are determined by the entry's group type. OpenFlow supports four group types: *all*, *indirect*, *fast failover*, and *select*. For the select type, one bucket in the group entry is executed, depending on the selection algorithm, but the algorithm's configurations and states are left unspecified.

Network function virtualization

NFV supports running virtualized SFs on commodity servers and makes it easier to deploy SF instances on the virtualization layer, increasing both scalability and flexibility.[7] The NFV infrastructure includes management and orchestration components, which determine the service chain for specific traffic. The controller is then programmed to enforce the orchestration.

When queried from the ingress switch that classifies incoming traffic, the controller runs service routing and then configures the switches to enforce the SF chain by traffic steering. Service routing is responsible for finding the optimal path through a particular service chain, given available resources in the NFV infrastructure. In traffic steering, the switches steer incoming traffic according to the flow entries set by the controller.

Network service header

The European Telecommunications Standards Institute (ETSI) specification on OpenFlow-based traffic steering[8] describes it as the process of matching five tuples in each incoming packet to set entries in the flow table. However, per-flow matching incurs a large number of flow entries because the same SF chaining path must be set multiple times for each group of five tuples.

To address that problem, the IETF has proposed using the network service header (NSH),[9] which enables traffic steering and service chaining to be carried out in the service plane.[10] The NSH carries two critical fields: the service path identifier (SPI), which identifies an SF chaining path; and the service index (SI), which identifies the SF's location in the path. The SI decreases by 1 each time a packet travels through an SF (after the SF is finished). The combination of SPI and SI determines the tunnels through which the packets will pass.

The NSH efficiently reduces the number of entries for traffic steering

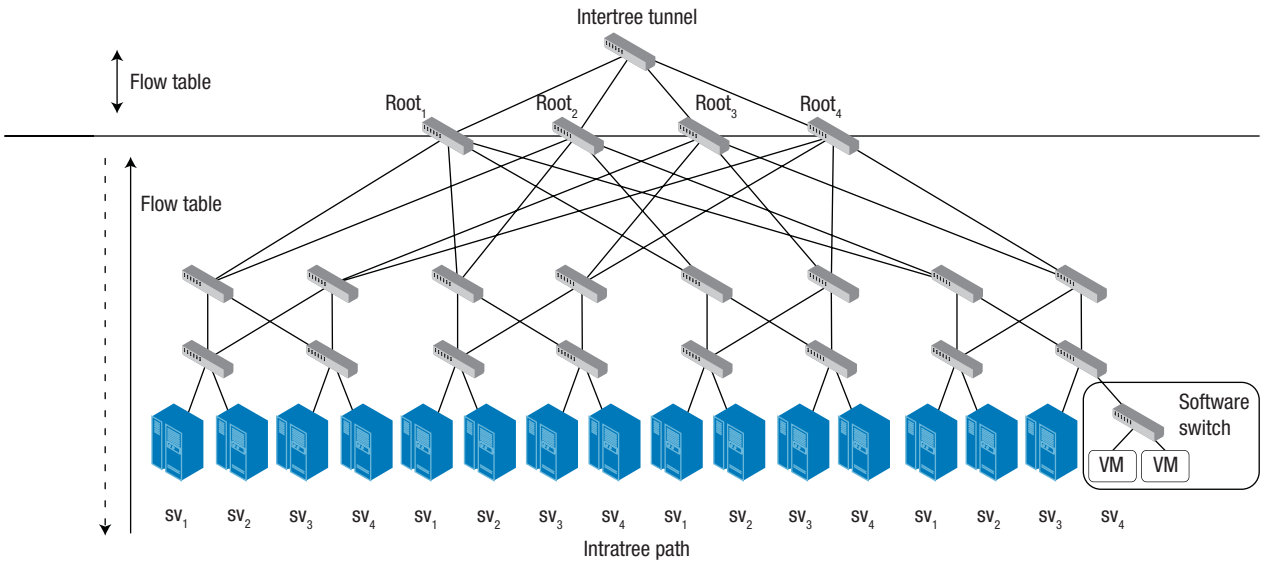


FIGURE 1. How Balanced Hash Tree (BHT) balances service functions (SFs). In this design, each physical server has one OpenFlow software switch, which runs two identical SF instances on two virtual machines (VMs), as shown in the enlarged box at far right. Multiple SFs make up an SF chain, represented as sv_1, sv_2, sv_3, ... sv_i. In the example, one server processes one chain. The SF path through the network is divided into an intertree tunnel, which interconnects the roots of successive SF trees through the flow entries (lines above the horizontal line), and the intratree path for distributing traffic (lines below the horizontal line). The arrows to the left represent packet-forwarding directions: up (solid), down (dashed), and both ways (double arrow).

because multiple SF chaining paths share the tunnels. The tunnel to be transformed for the next SF might consist of multiple equivalent instances for load balancing. The IETF draft does not specify the exact balancing method.

Controller-based balancing

Several proposed systems balance the load among SF instances, but all involve using the controller. One strategy implements multiple load-balancing algorithms on the controller, such as random, round-robin and load-based methods.[10] When it receives a packet-in message, the controller selects an SF instance according to the algorithms and then sets the flow entries. Another approach uses a dedicated controller for each service network to monitor SF instance loads for load balancing.[11] Yet another system monitors the server loads and the network status for processing packet-in messages when some flows request services.[12] A system for resource management and load balancing sets flow entries to divide incoming traffic.[13]

In all these systems, when the controller receives a packet-in message, it must track the loads of SF instances to select the appropriate one and then set the flow entries accordingly—tasks that greatly increase its workload.

HOW BALANCED HASH TREE WORKS

BHT has two main objectives: to balance SF chaining in a way that redirects incoming flows to the desired SFs, and to mitigate the controller's workload.

As Figure 1 shows, BHT establishes a tree of SFs and uses group entries to split incoming traffic among the SF instances. We assume that the datacenter network provides the NFV infrastructure, the network has a fat-tree topology, and the controller has information about both the network topology and the SF instances' locations. In this design, a service chain is mapped to a single service-chaining path that interconnects the roots of trees of SFs based on the NSH.

In Figure 1, SF locations are heterogeneous (different instances in the same pod), but BHT works equally well with homogenous locations (identical instances in the same pod). The intratree path is through the entries in the select group type, and the packets are redirected back to the root when an SF is finished. Thus, when the controller receives a packet-in message, it needs only to establish an intertree tunnel; it does not have to specify which instances will provide SFs.

Traffic steering

After classification to determine its traffic type (such as HTTP), a packet is mapped to the desired service chain according to the configured policy. It is then tagged with the NSH, which carries SPI and SI for traffic steering. SPI carries the service-chaining path identifier, and SI is initialized to the length of service chain.

Once SPI and SI determine the tunnels through which the packets will pass, BHT establishes the tunnels by imposing the NSH between the original packet and the Multi-Protocol Label Switching (MPLS) transport encapsulation in the outer network. We chose MPLS over a virtual extensible LAN (VXLAN) or generic routing encapsulation because its label format is simple and sufficient for our use. In an operation similar to that in MPLS, the switches in both ends of a tunnel will

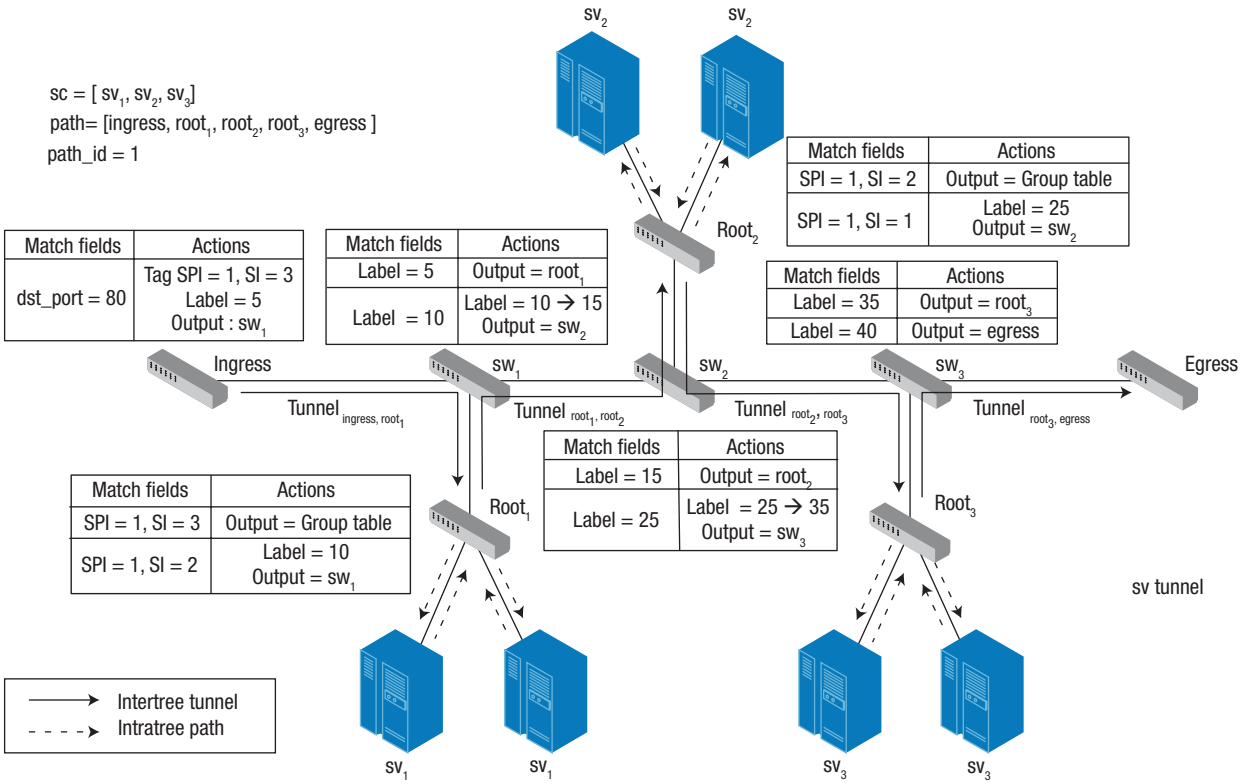


FIGURE 2. An example of traffic steering for an SF chain in which $Root_1$, $Root_2$, and $Root_3$ are the root switches for SFs sv_1, sv_2, and sv_3. The ingress switch first classifies incoming packets and initializes the service path identifier (SPI) and service index (SI) in the network service header as well as the label in the Multi-Protocol Label Switching (MPLS) transport encapsulation. The flow tables on the three root switches are in charge of changing the tunnel with a new label according to SPI and SI, while the other switches (sw_1, sw_2, and sw_3) perform label switching and forward packets. The egress switch removes the network service header and forwards the packets to their original destination.

distribute a locally unique label, and through repeated label switching the packets will move through the tunnel. After an SF processes a packet, the packet goes back to the root switch, which will change the tunnel and redirect the packet to the next SF depending on the SPI and SI.

The BHT module on the controller contains two functions related to traffic steering: `Packet-In_processing()` and `Service_chain_setting()`. We assume the controller knows the locations of the SF instances in terms of the attached switches and ports. It determines the group entries for each intermediate switch in the tree and then adds the group entries with the select group type (whether the SF locations are heterogeneous or homogeneous).

When the `Packet-In_processing()` function receives a packet-in message from the ingress switch, the controller assigns an SF chain to the flow specified in the message and checks whether the SF chain has been set. If it has, the controller sets a flow entry on the ingress switch to tag the flow's packets with the NSH. If it has not been set, the packet-in processing function calls the `Service_chain_setting()` function and assigns the new SF chain to the specified flow.

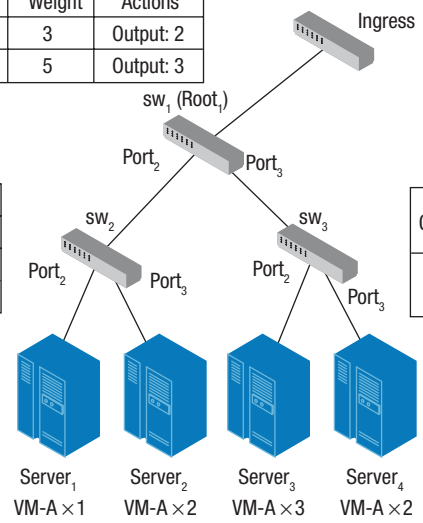
BHT assigns a unique path identifier to the SF chaining path and then checks whether every tunnel between any two successive SFs has been established. If it has, BHT uses SPI and SI as the match fields to set the entries on the root switch for changing the tunnel when the packets return to the root switch from a finished SF. Otherwise, it establishes a tunnel by distributing a locally unique label for the tunnel and then setting the entries for label switching.

Figure 2 shows a traffic-steering example for SF chain sc = [sv_1, sv_2, sv_3]. After an SF is finished, the corresponding root switch will change the tunnels (by changing the labels)—$Root_1$ for the tunnel from $Root_1$ to $Root_2$, $Root_2$ for the tunnel from $Root_2$ to $Root_3$, and $Root_3$ for the tunnel from $Root_3$ to the egress switch. The SF chain interconnects only the SF root switches, not the physical servers.

Load balancing

Because the controller knows the locations of SF instances, it can construct the tree for each SF and set the group entries according to the number of instances and their locations. In addition to the two traffic-steering functions, the BHT module on the controller contains the `Load_balancing()` function, which performs three tasks:

Group ID	Type	Action buckets		
		Index	Weight	Actions
1	select	1	3	Output: 2
		2	5	Output: 3



Group ID	Type	Action buckets		
		Index	Weight	Actions
1	select	1	1	Output: 2
		2	2	Output: 3

Group ID	Type	Action buckets		
		Index	Weight	Actions
1	select	1	3	Output: 2
		2	2	Output: 3

FIGURE 3. Load balancing in BHT. Physical servers can launch multiple VMs as SF instances, for example, SF-A in this case. After the tree has been established, the `Load_balancing()` function sets the group entries with multiple action buckets. The buckets consist of various forwarding ports, which are generated according to node branches. The corresponding weights are determined by the number of SF instances (VM–A × n) on the physical servers reachable from the output ports.

- For each SF, it calculates the paths from the switches to those SF instances attached to the ingress switch and leverages this information to find the SF's root switch (the switch nearest the ingress switch on the paths).
- It counts the number of instances to which an output port is attached for each intermediate node on the tree and records it as tree information.
- On the basis of the tree information obtained, it determines the actions and weights in the action buckets and sets the entries in the select group type.

Figure 3 shows an example of load balancing in BHT.

Selection mechanism of Open vSwitch

We assume that the network shown in Figure 2 uses Open vSwitch (openvswitch.org). The default selection mechanism on Open vSwitch is to calculate a score for each action bucket and have the switch select the action bucket with the highest score for the enforced action. When an incoming packet is assigned to an entry in the select group type, the first step is to retrieve the packet's destination media access control (MAC) address and then hash it to generate the basis value in the score calculation. The second step is, for each action bucket, hash its basis with index i and multiply the hashing result and the weight. The result is the action bucket's score. The score calculation is formulated as

```
basis = hash_mac(dst_mac)
score = (hash_int(i,basis) & 0xffff) *
        bucket→weight
```

The default selection calculation has a flaw, however. The hash functions are the same on every switch, so the packets with the same destination MAC address will get the same result with the same group entry, even on different switches. In other words, they will be all forwarded to the same port, and no packets will go through the other ports.

For that reason, BHT uses a modified calculation in which the destination MAC address and the switch's datapath identifier (dpid)—which allows the controller to manage switches—are hashed first to produce dpbasis, and dpbasis and the action bucket's index are hashed next. With this modification, selection varies with the switches even with the same group entry, but the selection for packets in the same flow is still uniform. The modification is formulated as

```
basis = hash_mac(dst_mac)
dpbasis = hash_int(dpid,basis) & 0xffff
score = (hash_int(i,dpbasis) & 0xffff) *
        bucket→weight
```

SAMPLE IMPLEMENTATION

To evaluate BHT's performance, we implemented it on the Ryu controller (osrg.github.io/ryu) using our modification of the Open vSwitch's default selection algorithm.

The BHT module on Ryu contained the `Load_balancing()`, `Packet_In_processing()`, and `Service_chain_setting()` functions. As implemented, the first function uses Dijkstra's algorithm[14] to calculate the SF trees and a list to store tree information. We applied `parser.OFPGroupMod()`—a function from

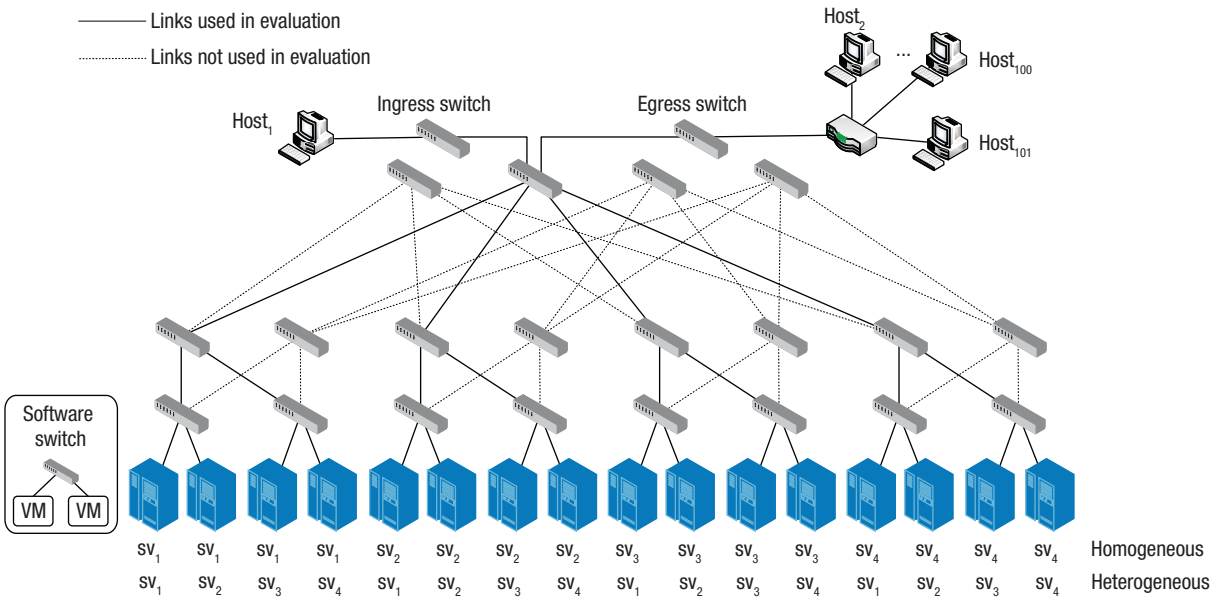


FIGURE 4. Network topology for our experiments to evaluate BHT's load-balancing ability. We employed Mininet 2.2 to emulate a virtual network. The VMs on Mininet emulate the SF instances. Mininet's iperf tool generated 100 TCP connections with different destination access control (MAC) addresses from $Host_1$ to $Host_{101}$ for emulating the flows that reach the desired SFs through SF chaining.

the Ryu APIs—to set group entries according to that information.

We had a packet-in message trigger the `Packet-In_processing()` as a thread, and used the `packet_in_handler()` function to process the message. If necessary, we could have the packet-in message trigger the `Service_chain_setting()` function, which relies on Dijkstra's algorithm to set a new SF chaining path.

We did not include the NSH in our implementation. Rather, we inserted VLAN tags to carry the SPI and SI fields, which simulated NSH use, and established tunnels in the transport encapsulation using the label value in the differentiated services code point (DSCP) field—a six-bit field in the IP header that specifies the per-hop behavior for a given packet flow.

EVALUATION RESULTS

We used two servers in our experiments. The first is an Intel Core i5-4590 running the Debian 7.7 OS at 3.30 GHz on a VMware workstation with the Ryu 3.15 controller and OpenFlow 1.3 switching protocol. The other server is an Intel Core i7-4790K also running Debian 7.7 but at 4.00 GHz and with Open vSwitch version 2.3. We chose the second server to emulate a datacenter network's tree topology by also running Mininet version 2.2 (mininet.org). Mininet creates a realistic virtual network that runs real kernel, switch, and application code on a single machine. Figure 4 shows the network configuration used in our experiments.

The SF types are irrelevant to our experimental results, and the locations of SFs can be either homogeneous or heterogeneous. The traffic was sent for 200 seconds, and the packet size was consistently 1,514 bytes. Once we sent all the traffic, we evaluated load-balancing performance in terms of the bytes received, which we considered the workload of an SF instance.

Our objectives in conducting the experiments were to evaluate packet-in message-processing time and load-balancing performance.

Packet-in message-processing time

To set a baseline for comparing packet-in message-processing times, we designed a mechanism that implements load balancing on the controller, as do many existing schemes. The load-balancing mechanism on the controller refers to the received bytes as the workload of an SF instance. When receiving a packet-in message, but before calculating the SF chaining path, the controller sends a multipart request message to get the port information of received bytes on the software switches to which the SF instances are attached. The fewer bytes a software switch has received, theoretically, the lighter the load of the SF instances attached to it will be.

After getting the received bytes in a time interval from the software switches, the controller selects the instances attached to a software switch with the lowest number of received bytes (the least loaded instances) to provide the SF for load balancing. Finally, the controller determines the SF chaining path and sets the flow entries.

In this experiment, we assumed the configuration in Figure 4: four SFs are in the datacenter network (sv_1 through sv_4), and each SF contains two instances on each of the four physical servers for that SF. We generated 100 TCP

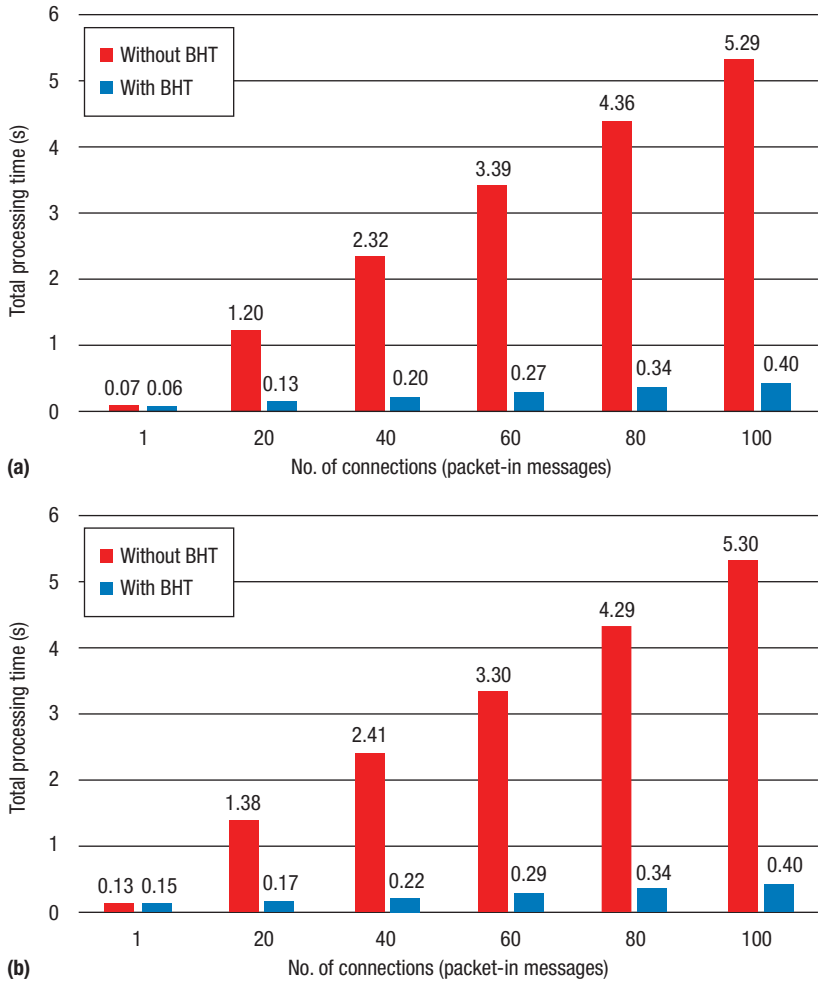


FIGURE 5. Processing times of packet-in messages with and without BHT. Times are in terms of connection number, which is synonymous with the number of packet-in messages, for (a) homogeneous and (b) heterogeneous SF locations. Time without BHT includes the latency from the queried switches reporting their load information back to the controller. With BHT, the controller outsources the selection of SF instances to the group table on the switches, which effectively mitigates the controller's workload. Total processing time without BHT was 5.3 seconds and, with BHT, 0.4 seconds—a reduction of approximately 92.5 percent. The result is similar for homogenous and heterogeneous SF allocation.

connections from $Host_1$ to enforce the same service chain over the four SFs.

Figure 5 shows processing times of packet-in messages with and without BHT. Packet-in processing time without BHT includes the latency from the queried switches reporting their load information back to the controller. We performed the queries in parallel by multithreading, and latencies were typically on the order of milliseconds. Considering the processing time of 5.3 seconds for 100 connections without BHT, almost all the processing time is still likely to be attributable to the controller's workload

Load-balancing performance

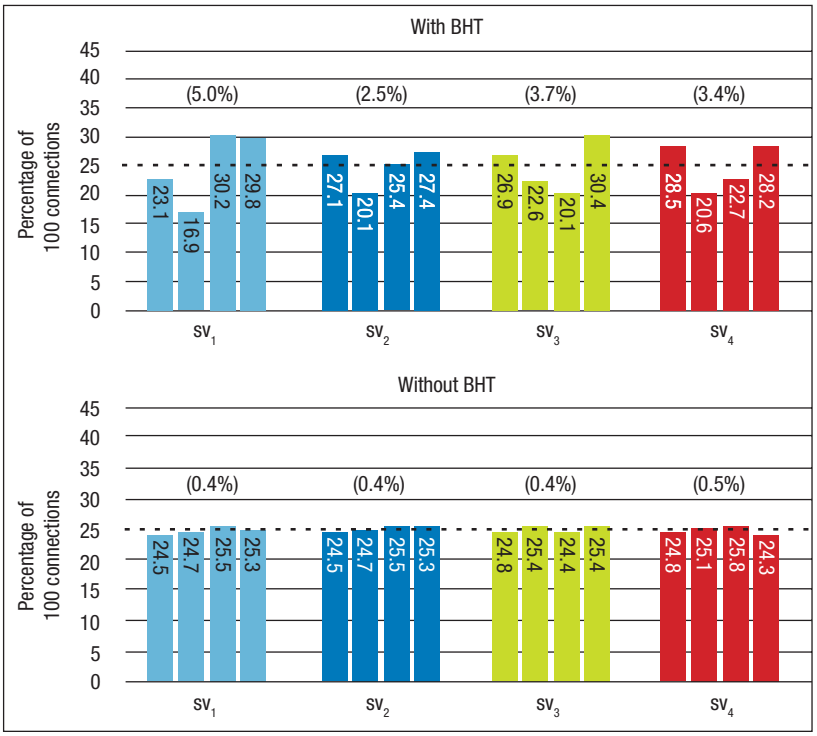
Figure 6 shows comparative load balancing for the same experiment. The comparison covers only load balancing among the four physical servers for each SF because the hypervisor handles balancing among the VMs inside the server. For each SF, the perfect load balancing is 25 percent on each of the four physical servers, so we evaluated load-balancing performance by calculating the average absolute difference between the actual load and the perfect load on each server.

Because the entry in the select group type is implemented by hashing, the loads are not as balanced as those with load balancing without BHT, but the absolute difference for BHT is within 2.4 to 5.0 percent of perfect performance. The load balancing with BHT is slightly better for heterogeneous SF allocation, but the difference is insignificant for any practical application. The results show that BHT efficiently balances loads among the servers, while simultaneously reducing the controller workload significantly.

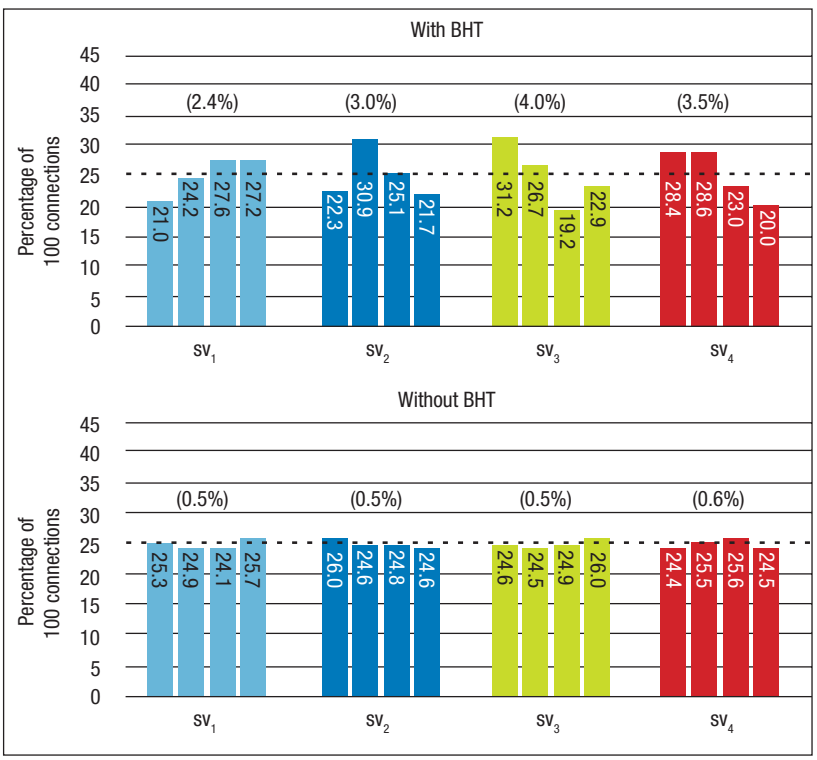
Initial evaluations show that BHT can be an effective-load balancing alternative to controller-based solutions, and we have already identified areas for extension. One is to accommodate multiple paths between adjacent SFs. In the current design, BHT considers only load balancing among the instances of the same SF. Extending BHT to cover load balancing among both the SF instances and paths simultaneously is an interesting issue for future work. Another open question for additional exploration is how to perform load balancing for diverse types of real-world traffic in a datacenter network.

ACKNOWLEDGMENTS

The work described in this article was supported in part by the Ministry of Science and Technology, Taiwan; Chunghwa Telecom and MediaTek; and the III Innovative and Prospective Technologies Project (1/1) of the Institute for Information Industry, which is subsidized by the Ministry of Economic Affairs, Taiwan.



(a)



(b)

FIGURE 6. Comparison of load-balancing performance. We compared load balancing on four physical servers with (a) homogenous and (b) heterogeneous locations with and without BHT. The horizontal dashed line represents perfect balancing at 25 percent. The numbers in parentheses above the bars denote the percentage departure of the average load balancing (average of the numbers over the four bars) from a perfect load balancing for that SF. In (b) with BHT, the sv_1 balancing was within 2.4 percent of the perfect balance.

REFERENCES

1. *Software-Defined Networking: The New Norm for Networks*, white paper, Open Networking Foundation, 2012; www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf.
2. *OpenFlow Switch Specification, Version 1.5.0*, ONF TS-020, Open Networking Foundation, 19 Dec. 2014; www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-switch-v1.5.0.noipr.pdf.
3. B. Nunes et al., "A Survey of Software-Defined Networking: Past, Present, Future of Programmable Networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, 2014, pp. 1617–1634.
4. Y.D. Lin et al., "An Extended SDN Architecture for Network Function Virtualization with a Case Study on Intrusion Prevention," *IEEE Network*, vol. 29, no. 3, 2015, pp. 48–53.
5. P. Quinn and T. Nadeau, *Problem Statement for Service Function Chaining*, IETF RFC 7498, Apr. 2015; tools.ietf.org/html/rfc7498015.
6. J. Halpern and C. Pignataro, *Service Function Chaining (SFC) Architecture*, IETF RFC 7665, Oct. 2015; tools.ietf.org/html/rfc7665.
7. R. Mijumbi et al., "Network Function Virtualization: State-of-the-Art and Research Challenge," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, 2016, pp. 236–262.
8. Group Specification: Network Functions Virtualisation (NFV); Virtual Network Functions Architecture, ETSI GS NFV-SWA 001 v1.1.1, European Telecommunications Standards Inst.; Dec. 2014; www.etsi.org/deliver/etsi_gs/NFV-SWA/001_099/001/01.01.01_60/gs_NFV-SWA001v010101p.pdf.

ABOUT THE AUTHORS

PO-CHING LIN is an associate professor in the Department of Computer and Information Science at National Chung Cheng University (CCU). His research interests include network security, network traffic analysis, and the evaluation of network systems performance. Lin received a PhD in computer science from National Chiao Tung University (NCTU). Contact him at pclin@cs.ccu.edu.tw.

YING-DAR LIN is a Distinguished Professor in the Department of Computer Science at NCTU and director of the Network Benchmarking Lab. His research interests include transforming networks into clouds, software-defined networks (SDNs), network function virtualization, 5G/mobile edge computing, and network security. Lin received a PhD in computer science from the University of California, Los Angeles. He is an IEEE Distinguished Lecturer, an Open Networking Foundation research associate, and coauthor of *Computer Networks: An Open Source Approach* (McGraw-Hill, 2011). Contact him at ydlin@cs.nctu.edu.tw.

CHENG-YING WU is an engineer at MediaTek. While conducting the research reported in this article he was a graduate student at NCTU. His research interests include network architecture, SDNs, and cloud computing. Wu received an MS in computer science from NCTU. Contact him at cywu.cs02g@nctu.edu.tw.

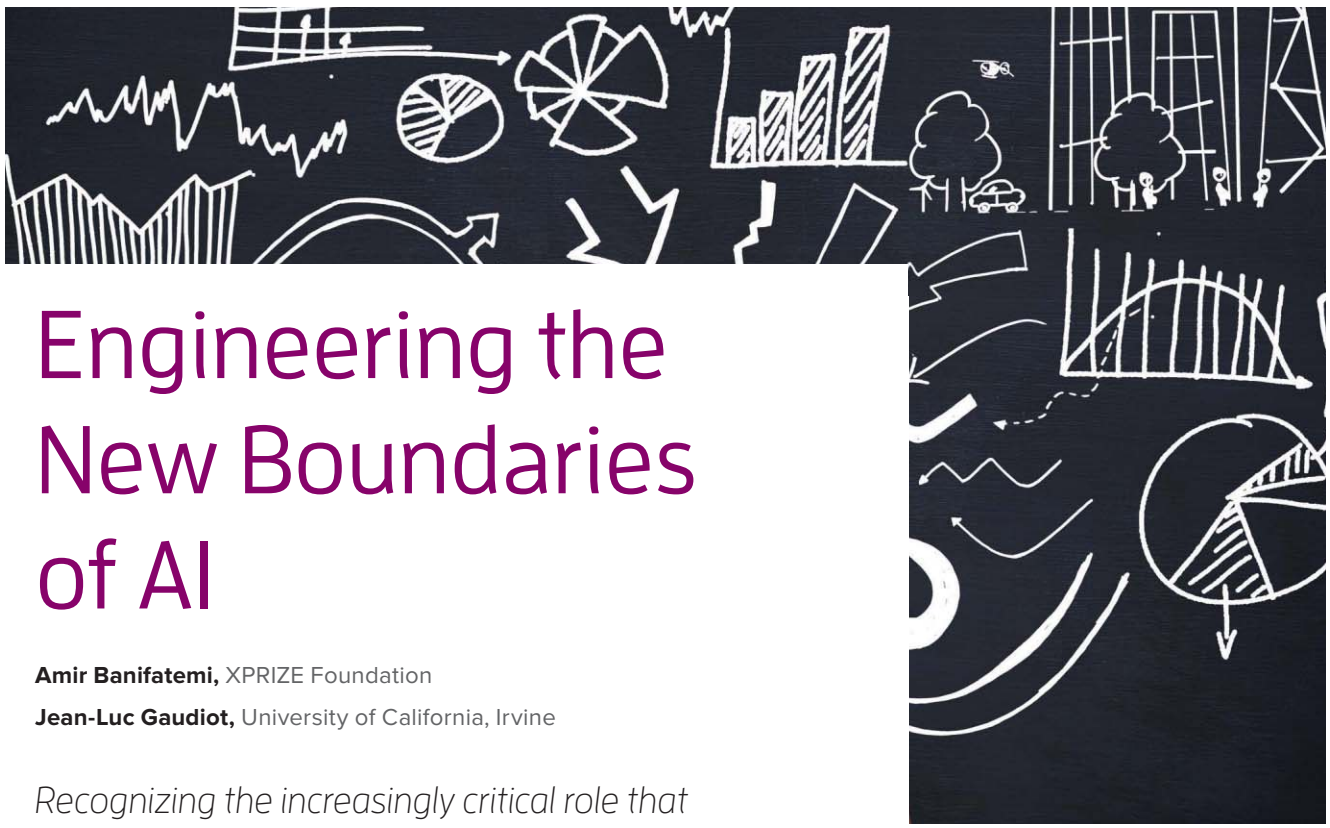
YUANG-CHENG LAI is a professor in the Department of Information Management at National Taiwan University of Science and Technology. His research interests include wireless networks, network performance evaluation, network security, and social networks. Lai received a PhD in computer science from NCTU. He is a member of IEEE. Contact him at laiyc@cs.ntust.edu.tw.

YI-CHIH KAO is director of the Network and System Division of the Information Technology Service Center at NCTU. His research interests include cyber forensics, network performance evaluation, SDNs, and service design. Kao received a PhD in industrial engineering and management from NCTU. He is a member of ACM and the Project Management Institute (PMI). Contact him at ykao@mail.nctu.edu.tw.

9. P. Quinn and U. Elzur, "Network Service Header," IETF Internet Draft, work in progress, 26 May 2016.
10. H. Uppal and D. Brandon, "OpenFlow-Based Load Balancing," cite seerx.ist.psu.edu/viewdoc/download?doi=10.1.1.168.5150&rep=rep1&type=pdf.
11. M. Koerner and O. Kao, "Multiple Service Load-balancing with OpenFlow," *Proc. IEEE 13th Int'l Conf. High-Performance Switching and Routing* (HPSR 12), 2012, pp. 210–214.
12. N. Handigol et al.,"Plug-n-Serve: Load-Balancing Web Traffic Using OpenFlow," *Proc. ACM SIGCOMM Conf. Data Communication* (SIGCOMM 09), 2009; conferences.sigcomm.org/sigcomm/2009/demos/sigcomm-pd-2009-final26.pdf.
13. Z. Qazi et al., "SIMPLE-fying Middlebox Policy Enforcement Using SDN," *Computer Communication Rev.*, vol. 43, no. 4, 2013, pp. 27–38.
14. E.W. Dijkstra, "A Note on Two Problems in Connexion with Graphs," *Numerische Mathematik*, vol. 1, 1959, pp. 269–271.

COMPUTING EDUCATION

EDITOR ANN E.K. SOBEL
Miami University; sobelae@muohio.edu



Engineering the New Boundaries of AI

Amir Banifatemi, XPRIZE Foundation
Jean-Luc Gaudiot, University of California, Irvine

Recognizing the increasingly critical role that AI plays in all aspects of modern society, the XPRIZE Foundation launched the IBM Watson AI XPRIZE. Interdisciplinary teams will advance current AI technologies as they compete for the grand prize.

AI surrounds us. It's in our search engines, automobiles, phones, video-streaming websites, and even the systems that run our financial markets. We interact with and rely on intelligent machines throughout our daily activities.

AI has made steady, linear progress toward adoption and integration over the past 50 years. It's now at an important inflection point—the adoption curve could soon transcend the linear growth pattern and take an exponential leap. With machine learning, AI technologies have the potential to make sense of vast, complex datasets. This begs the question: How can we best leverage both human intelligence and these powerful AI technologies to benefit humanity now and in the future?

ENTER XPRIZE

It's this question of the impact of AI on mankind that inspired the IBM Watson AI XPRIZE, the latest innovation challenge from the XPRIZE Foundation. IBM sponsored the competition with a prize purse of $5 million, including a grand prize of $3 million, a second-place prize of $1 million, and a third-place prize of $500,000. The goal of the competition is to accelerate the development of scalable AI solutions to address humanity's greatest challenges. The four-year competition will have milestone completions in 2017, 2018, and 2019, with the top three finalists competing for the grand prize in 2020.

Recognizing the competition's humanitarian importance and global reach, the IEEE Computer Society (CS) formed a partnership agreement in August 2016 with the XPRIZE Foundation. Select CS members will join the XPRIZE scientific advisory board, and the CS will draw from its global network of experts to help judge and advise the competitors. The CS will also encourage

XPRIZE ORIGIN STORY

The history of XPRIZE is unique. The first was inspired by the $25,000 Orteig Prize, which was awarded in 1927 to American aviator Charles Lindbergh for flying the *Spirit of St. Louis* on the first nonstop flight between New York and Paris.

The original XPRIZE, announced in 1996, was a $10 million prize to the first privately financed team that could build and fly a three-passenger vehicle 100 km into space twice within two weeks. The prize, later titled the Ansari XPRIZE for suborbital spaceflight, motivated 26 teams from 7 nations to invest more than $100 million in pursuit of the $10 million prize. On 4 October 2004, the prize was awarded to Mojave Aerospace Ventures, kicking off the personal spaceflight revolution.

The first competition created a funding model to stimulate broad investment in innovation that produces 10-fold return on the prize as well as a 100-fold increase in follow-on investment and social benefit.

The XPRIZE Foundation recognized that this extraordinary leverage could be beneficial for a range of global grand challenges in which these inducement prizes would lead to drastic and important improvements in quality of life.

Active competitions include the $30 million Google Lunar XPRIZE, the $20 million NRG COSIA Carbon XPRIZE, the $15 million Global Learning XPRIZE, the $10 million Qualcomm Tricorder XPRIZE, the $7 million Shell Ocean Discovery XPRIZE, the $7 million Barbara Bush Foundation Adult Literacy XPRIZE, and the $5 million IBM Watson AI XPRIZE.

its members to serve as mentors or to provide other technical expertise to entrants in support of improving their application or approach as well as in showcasing their team's progressive achievements.

The XPRIZE Foundation launched its first competition in 1996 in an effort to spur privately funded innovation and research for space exploration (learn more about XPRIZE's history in the "XPRIZE Origin Story" sidebar). With the success of this first competition, the foundation has attracted continued private funding for additional XPRIZE competitions in health, learning, energy, environment, global development and exploration.

These privately funded, prize-based scientific challenges are highly attractive to individuals, companies, and organizations, and they're clearly inspiring a solution-focused approach to grand challenges that is less research oriented than traditional academic or corporate settings. University- and corporation-based research and development traditionally have incentivized research advances through promotion- or salary-based structures. XPRIZE funding and publicity are untethered to any one corporation or institution, so they offer global recognition and opportunities for successful competitors.

XPRIZE competitions draw significant numbers of nontraditional innovators and entrepreneurs and enable them to focus completely on one area of innovation. With its goal to bring about "radical breakthroughs for the benefit of humanity" through incentivized competition, the XPRIZE approach doesn't impose budget requirements or other significant administrative overhead. Researchers, engineers, and entrepreneurs are free from constraints that might typically limit them with government- or consortia-based funding sources. By driving competitors to invest every intellectual and financial resource at their disposal to solve the problem, reach the goal, and win the prize, innovation can take center stage. A large number of groups can work in parallel toward the goal, and the resulting solution or benefit can be quickly adopted and realized by society.

INCENTIVE-DRIVEN COMPETITIONS SPUR INNOVATION

Challenge-based models offering substantial financial prizes have historically succeed in driving innovation. This departure from traditional research models enables competitors to focus on a solution's impact, rather than the particulars of the methods that are used to achieve it. Contributors are inspired to explore all possible and creative ways to meet that objectives and win.

Of course, prizes for innovation and technological achievements aren't new: The British government's simple and practical method to precisely determine a ship's longitude; Charles Lindbergh won the Orteig Prize for the first transatlantic flight from New York to Paris; and, more recently, Elbert "Burt" Rutan's team won the Ansari XPRIZE for developing the first reusable private spaceship.

With tech-focused prizes, awards are tied to achievements versus attempts, and teams are driven to leverage any and all resources and support, pulling together an impressive matrix of expertise, capital, collaboration support, and achievement-oriented structure. In addition, recent incentive-driven competitions have shown that teams invest their own resources in these competitions to the tune of 10 to 100 times the prize itself.

Competitions can come in all sizes—from the smaller Kaggle Competitions (kaggle.com/competition) to the larger $1 million Netflix Prize—thus there's more awareness and acceptance of such incentive prizes than ever. Of course, scale and size matter, and

multimillion-dollar competitions tend to have a greater impact on the market and innovation than smaller, more localized versions. However, their greater popularity has brought them increasingly into other domains—including the classroom. Educators have been enthusiastic to explore challenge-based learning, appreciating their potential for developing students' abilities to work in teams, to be inventive and resourceful, and to push beyond traditional lines of thought and institutional boundaries. Indeed, leveraging such challenges in the learning environment imparts some unique skillsets that are invaluable for students as they transition to the modern workplace.

THE AI XPRIZE

As the first "open" challenge in AI, the IBM Watson AI XPRIZE is unlike its predecessors. Rather than set a single universal goal for all teams, this competition invites teams to each create their own goal: an application of AI to a meaningful problem. Once goals are defined, the teams will use their own best available resources, along with assistance from selected experts in the professional AI and engineering communities, to develop their plans for the duration of the competition.

Each team must demonstrate progress and annual milestone achievements for their plan, which will ultimately qualify them to compete for the grand prize on the TED 2020 stage in front of a live in-person and online audience. The audience and a panel of expert judges will crown the grand-prize winner and rank the runners-up from the three finalists.

As AI encompasses a wide range of disciplines, teams can utilize several types of specialized technology in achieving their goals. Technologies that drive or support intelligent behaviors also integrate sensory inputs and cognitive capabilities. These systems support human–machine collaboration through their ability to possess domain understanding, learn from experience, reason toward specific goals, and interact naturally with human collaborators. Thus, AI applications benefit greatly from expertise in data science and machine intelligence as well as more humanities-focused disciplines.

To foster increasing democratization and availability in AI technologies and allow individuals from many disciplines to participate in technological and social AI innovation, the competition is structured to attract large, collaborative teams. In fact, institutions and organizations with computing resources, datasets, tools, and domain expertise are encouraged to join the competition as supporting partners to help teams with the development and execution of their plans. Similarly, individual engineers, researchers, and specific domain experts (healthcare, environment, education, art, policy, economics, and so on) can join teams as advisors or hands-on support—indeed, it's this cross-pollination of knowledge, techniques, social contexts, and methodologies that will contribute to accelerating education and skills development as it drives groups toward meeting their goals and winning the prize.

The XPRIZE's public visibility will have a tremendous impact on increasing our everyday awareness of AI, particularly the ways in which it benefits society. Through this kind of highly publicized, global, interdisciplinary, public/private competition, the XPRIZE Foundation also opens up a crucial forum for a variety of stakeholders to help define the guidelines and mechanisms by which AI could or should be used. The benefits of this kind of heterogeneous, team-based approach to address important humanitarian challenges is that many voices can be heard, allowing us to more knowledgably shape our future through technology.

Because this competition is, at its core, a large-scale collaboration employing interdisciplinary approaches to problem solving, it's a tremendous learning opportunity—indeed, the first real benefit is learning.

JOIN THE COMPETITION

Developers, engineers, entrepreneurs, and innovators throughout the world are encouraged to form or join an AI XPRIZE team. To read more or submit an entry, visit ai.xprize.org.

AMIR BANIFATEMI is the lead for the XPRIZE Foundation's IBM Watson AI XPRIZE. Contact him at amir.banifatemi@xprize.org.

JEAN-LUC GAUDIOT is a professor of electrical engineering and computer science at University of California, Irvine. He is IEEE Computer Society's president-elect. Contact him at gaudiot@uci.edu.

AFTERSHOCK

Equity, Safety, and Privacy in the Autonomous Vehicle Era

Vasant Dhar, New York University

Big data from onboard vehicular systems can be used to help determine liability in accidents, streamline insurance pricing, motivate better driving practices, and improve safety—all with potentially minimal impact on privacy.

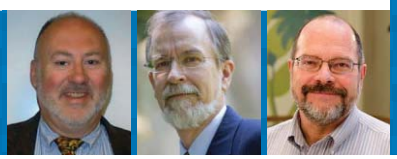
In 2015, an estimated 38,300 people were killed and 4.4 million injured in motor vehicle crashes in the US alone, with damage costs exceeding $400 billion.[1] In contrast, there were no commercial aviation fatalities and 136 civil aviation deaths in the US in 2015.[2] More than 90 percent of auto accidents result from human impairment such as drunk driving or road rage, errant pedestrians, or just plain bad driving. This has motivated the development of automotive safety systems that use sensors, cameras, lasers, and radar to monitor the vehicle and its surroundings and to prompt either the vehicle or driver to take corrective action to avoid an accident. These systems include forward-collision warning and automated braking, lane-departure warning, rear cross-traffic alerts, drowsiness detection, adaptive headlights, pedestrian detection, and automated parking assistance.[3]

Autonomous vehicles (AV) promise to further reduce collisions dramatically by removing the human element altogether. More than 30 tech companies and auto manufacturers are working on AVs and expect to start deploying them within the next three to five years.[4] Some safety experts have cautioned against allowing AVs on the road until collision-avoidance systems achieve perfection, though this wouldn't necessarily eliminate all accidents. However, even less-than-perfect systems could be compelling enough if they substantially reduce accidents, considering that every percentage point improvement equates to 400 lives saved and 40,000 injuries avoided annually.

It could be some time before AVs are commonplace. Although the physics associated with driving is well understood and sensor technology continues to improve, machines must learn human behavior and norms, which can be complex. In addition, the human interface in AVs is still evolving as manufacturers work to enhance the software and its security. Nevertheless, federal transportation officials appear receptive to self-driving

0018-9162/16/$33.00 © 2016 IEEE

EDITORS
HAL BERGHEL University of Nevada, Las Vegas; hlb@computer.org
ROBERT N. CHARETTE ITABHI Corp.; rncharette@ieee.org
JOHN L. KING University of Michigan; jlking@umich.edu



technology and recently published first-ever guidelines on safety best practices, existing and potential new regulatory tools, and the adoption of uniform practices across states (www.transportation.gov/AV).

Regardless of whether and when vehicles become fully autonomous, the data and analytics provided by current-day onboard systems are sufficient to determine fault in many accidents, and will improve with time. This capability, which already exists in AVs but is also installable on human-operated vehicles, demands that we revisit existing insurance laws and practices regarding liability and compensation, which have been in place for decades and are based on outdated assumptions about available evidence. Equally important, the data obtained from this technology can yield novel insights into how to prevent accidents and save lives, especially as human-operated and driverless vehicles begin to share the road. At the same time, we must ensure that vehicles' new data-gathering abilities don't lead to the violation of individual privacy, specifically the fundamental human right "to be left alone" as articulated by Samuel Warren and Louis Brandeis more than a century ago.[5]

AV DATA AS THE BASIS FOR INSURANCE EQUITY AND PRICING

Auto insurers rely on various forms of individual and demographic data—for example, your age, sex, and marital status; what type of vehicle you drive; where you live; how often and far you drive; your driving record; and your credit history—to assess customer risk factors and price insurance efficiently. However, when it comes to determining who's at fault in an accident, which insurer should pay for it, and what the amount should be, often there are no easy answers based on the available evidence. Data traditionally used to establish fault such as physical damage, police reports, and testimonies from victims and witnesses are often insufficient to achieve incontestable results.

Consequently, a complex patchwork of insurance laws has emerged across the US in which the determining factors for financial liability can vary considerably. Most states have "at-fault" or tort laws that let accident victims sue drivers at fault for medical expenses, material damages, lost wages, and sometimes pain and suffering, with different monetary limitations across states. Given that the data used to assign fault is often incomplete, resolving disputes about the degree of liability and the amount of compensation can be lengthy, costly, and frustrating for all parties. This problem led a dozen states to adopt "no-fault" laws that allow cases to bypass the litigation process by requiring the parties' insurers to cover their own policyholders' expenses regardless of who's ultimately held responsible. Such laws ensure quick payments (subject to monetary thresholds), but they don't address fairness, nor do they discourage bad driving behavior. They also lead to fraudulent claims by drivers in collusion with other unscrupulous agents, which can be difficult to disprove and raise premiums for law-abiding drivers.

Because fairness in insurance law has been elusive, economists and policymakers have focused largely on pragmatic steps to reduce the number and severity of accidents and ways to spread their cost equitably through society. However, quantifying the total costs of accidents, including preventive measures,[6] is difficult with limited data, and the larger issues of fairness and liability remain unaddressed.

The availability of big data obtained from onboard vehicular systems and roadway sensors changes the calculus of liability. Such data provides sufficient details about the circumstances of an accident—ranging from physical factors such as vehicle speed and angle of impact to environmental factors such as weather and lighting conditions to human factors such as driver awareness and attention—to easily, accurately, and reliably determine fault. This would nullify the litigation costs and delays associated with tort laws and eliminate the need for no-fault insurance.

If AV data can be used to establish fault in accidents with high precision, it isn't too much of a stretch to envision many liability decisions based on this data becoming largely automated.

If AV data can be used to establish fault in accidents with high degree of precision, it isn't too much of a stretch to envision many liability decisions based on this data becoming largely automated. This could significantly improve efficiency and produce more just outcomes, assuming accuracy is sufficiently high and the cost per error is sufficiently low.[7]

However, we need to think carefully about how the data are used to avoid some of the unintended consequences we might not like. For example, red-light cameras are already unpopular; most people probably wouldn't want computers to automatically issue other types of traffic citations, such as for speeding or tailgating, based on data data captured by onboard vehicular systems or roadway sensors. Likewise,

to avoid practices such as selling personal data or arbitrarily assigning a higher level of risk to policyholders to charge a higher premium, there would have to be explicit restrictions on the gathering and appropriate use of such data, including the granting of policyholder consent.

How can we enable the desirable uses of AV technology without intrusions on our privacy?

SAFETY AND PRIVACY TRADEOFFS

To understand the privacy risks associated with AV data, it's important to understand the nature of the data and what its analysis might yield. Onboard vehicular systems continuously record data, which can be translated into "instances" describing the vehicle's movement such as turning, accelerating, braking, changing lanes, and so on. Combining such instances can reveal unsafe behaviors by the vehicle operator—for example, weaving, following too closely, or "rolling through" stops. Combining instances across many drivers can reveal general patterns such as "tired drivers have more accidents," in which the degree of tiredness is estimated from the low-level sensor data.

Driving patterns can serve as useful coaching aids—for example, to inform a driver about the risks of driving late at night or changing lanes frequently, and ways to ameliorate those risks. However, if the instance-level data is in the hands of an insurer, auto manufacturer, or other entity, the driver might have little or no control over its "secondary use" or sale to third parties, which could have undesirable consequences. For example, if you collide with another driver, could that driver subpoena data from your insurer to support his claim that you appeared "tired" prior to the accident? What if your insurer sells data to a disreputable marketer or its servers are hacked?

Fundamental to addressing these concerns is where the data is stored and the analytics are conducted. If privacy is a top priority, one solution is to retain data on the onboard "black box" device and perform the analytics locally as well. In this case, to carry out risk assessment and premium pricing, insurers could transmit aggregate-level driving statistics to the device, which could then extract individual driver data for comparative analysis. This would ensure privacy and prevent unauthorized secondary use or sale of the data. In expectation of a lower premium, drivers could, however, elect to share onboard vehicular data with their insurer—preferably stripped of sensitive components such as location. The precedent for this type of privacy-preserving cryptographic protocol exists in several European countries, where drivers pay mileage-based road taxes without revealing personal data.[8] Of course, sharing instance-level data exposes the driver to potential data misuse.

Some scholars argue that a decision to keep this and other types of data private could be negatively viewed by stakeholders as a signal of high risk. For example, law professor Scott Pettet suggests that so-called voluntary data disclosure implicitly erodes privacy:[9]

> *In an economy with robust signaling, those with valuable credentials, clean medical records, and impressive credit scores will want to disclose those traits to receive preferential economic treatment. Others may then find that they must also disclose private information to avoid the negative inferences attached to staying silent. This unraveling effect creates new types of privacy harms, converting disclosure from a consensual to a more coerced decision.*

Although some policyholders such as new drivers might feel pressure to share their detailed driving data with insurers as proof of safe driving practices, aggregate statistics that are compared with a broader population along a few key dimensions should be sufficient for pricing the large majority of drivers fairly. For those who elect to share details, driving data is arguably less sensitive than, say, medical or financial data, which can be indicative of a chronic health condition or high credit risk, respectively. In contrast, drivers can readily alter their driving practices, and those who demonstrate safer driving behavior could probably limit data sharing and its associated risks with little economic penalty.

There will undoubtedly be significant trial and error before AVs are a common feature on the road. But regardless of the degree of autonomy vehicles are ultimately entrusted with, onboard systems will have increasingly powerful data-gathering capabilities that could be used to significantly improve the transportation ecosystem. Indeed, the analytics from today's fairly primitive systems already have the capacity to help determine liability in accidents, streamline insurance pricing, motivate better driving practices, and improve vehicle safety.

Despite the fine-grained nature of the data they collect, onboard vehicular systems need not expose drivers to serious privacy risks as long as we regulate the use of such data carefully. Technical and legal solutions are available to anonymize the data and prevent its unauthorized use. Just as wearable devices are enabling individuals to improve their health and fitness, vehicles could become yet another rich source of information for personal knowledge discovery by the "quantified self,"[10] while also making the world a safer place for everyone to navigate.

REFERENCES

1. S. Ziv, "2015 Brought Biggest Percent Increase in U.S. Traffic Deaths in 50 Years," *Newsweek*, 17 Feb. 2016; www.newsweek.com/2015-brought-biggest-us-traffic-death-increase-50-years-427759.

2. B. Jansen, "2015 One of the Safest on Record for Airliners," *USA Today*, 15 Feb. 2016; www.usatoday.com/story/news/2016/02/15/2015-another-safe-year-airliners/80398194.
3. J. Linkov, "Collision-Avoidance Systems Are Changing the Look of Car Safety," *Consumer Reports*, 17 Dec. 2015; www.consumerreports.org/car-safety/collision-avoidance-systems-are-changing-the-look-of-car-safety.
4. "33 Corporations Working on Autonomous Vehicles," blog, 11 Aug. 2016; www.cbinsights.com/blog/autonomous-driverless-vehicles-corporations-list.
5. S. Warren and L. Brandeis, "The Right to Privacy," *Harvard Law Rev.*, 15 Dec. 1890; groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html.
6. G. Calabresi, *The Cost of Accidents: A Legal and Economic Analysis*, Yale Univ. Press, 1970.
7. V. Dhar, "When to Trust Robots with Decisions and When Not To," *Harvard Business Rev.*, 17 May 2016; hbr.org/2016/05/when-to-trust-robots-with-decisions-and-when-not-to.
8. J. Balasch et al., "PrETP: Privacy-Preserving Electronic Toll Pricing," *Proc. 19th USENIX Security Symp.*, 2010, pp. 63–78.
9. S.R. Peppet, "Unraveling Privacy: The Personal Prospectus and the Threat of a Full Disclosure Future," *Northwestern Law Rev.*, vol. 105, no. 3, 2011, pp. 1153–1203.
10. T. Fawcett, "Mining the Quantified Self: Personal Knowledge Discovery as a Challenge for Data Science," *Big Data*, vol. 3, no. 4, 2015, pp. 249–266.

VASANT DHAR is a professor of information systems in the Leonard N. Stern School of Business and director of graduate studies at the Center for Data Science at New York University, as well as editor in chief of *Big Data*. Contact him at vdhar@stern.nyu.edu.

CYBERTRUST

The Fog of War in Cyberspace

Alexander Kott, Ananthram Swami, and Bruce J. West,
US Army Research Laboratory

A "cyberfog" security approach that splits data into numerous fragments and continually disperses them across multiple end-user devices could provide greater attack resiliency but also presents formidable technical challenges.

The great Napoleonic Age warfare theorist, Carl von Clausewitz, wrote about the fog of war as the fundamental uncertainty of information in a complex and adversarial world. More recently, the term "fog computing" has emerged to refer to the extension of cloud computing to the network edge. We see connections between these seemingly disparate notions. For example, it might be possible to improve the security of our networks and data by maximizing the "fogginess" of information as it appears to a cyberadversary. Even if partly compromised, this information would remain opaque to the adversary, while still being useful to us.

One way to achieve such opaqueness is to split data into numerous fragments and continually disperse them across multiple end-user devices. Many modern commercial databases employ data splitting, or sharding, for both security and scalability, but typically not for end-user devices. However, given the growing interest in fog computing and fog networks,[1] and the maturing of edge-network distributed databases such as GaianDB[2] as well as cyber-physical networks, it's time to explore the use of data splitting at the edge.

While potentially offering numerous benefits such as greater attack resiliency, this "cyberfog" approach also presents formidable challenges with respect to data and network management complexity; bandwidth, storage, and battery-power demands; data-reassembly latency; and intermittent connectivity. In a recent meeting at the US Army Research Laboratory, government scientists discussed these challenges with colleagues from academia and industry.

DATA DISPERSION AND REASSEMBLY

The database security community has demonstrated, through both research prototypes and successful products, the feasibility and value of data dispersion and, to a lesser extent, frequent repositioning of data shards.[3]

0018-9162/16/$33.00 © 2016 IEEE

EDITOR JEFFREY VOAS
National Institute of Standards and Technology;
j.voas@ieee.org

File confidentiality and integrity can be preserved, even when a cyberattack compromises a subset of the file servers.

Shamir's Secret Sharing scheme[4] can be seen as either a metaphor, or an actual component, of a cyberfog approach. Roughly, a Shamir-like data-dispersion scheme could enable information sharing in such a way that an adversary who succeeds in capturing a significant fraction of shards still won't be able to reconstruct any meaningful information from it. Such a scheme might help balance data-dispersion bandwidth requirements over time—for example, the bulk of data shards could be distributed during a lull in communications demands, whereas only the final and a few critical shards would be sent over the network during busy periods.

At the same time, there are significant obstacles to developing, validating, and implementing the complex mechanisms required to perform data dispersion. Increased diversification also creates new cyberattack surfaces and venues. In particular, a cyberfog approach could increase a network's vulnerability to availability attacks, even as it improves its resilience to confidentiality attacks. Consequently, the network might need to manage a complex tradeoff between availability and confidentiality in real time depending on users' tasks and circumstances. Achieving consistency would also be complicated.

Users eventually will request the dispersed data, which must be gathered and reassembled in a timely and efficient fashion. This could be helped by intelligent dispersion—putting data shards where they're more likely to be accessible when users are more likely to need them. While doing so, care must be taken not to introduce regularity into the dispersion scheme that would make it easier for adversaries to find that information. For example, CYRUS (Client-defined privacY-protected Reliable cloUd Service)[5] ensures user privacy and reliability by scattering files into smaller pieces across multiple clouds, so that no one cloud can read users' data.

To determine a user's data needs, there must be some means to automatically determine the relevance of information to the user. A cyberfog approach complicates this process: whereas in a conventional system two files in the same folder are likely relevant to the same issue, colocation of two data shards says nothing about their common relevance.

Timing issues in data dispersion and reassembly are also complex: the way a collection of information is dispersed—the data shards' size and distance from one another—depends on when and how rapidly the user will need these bundles of information, and the overhead for distributing and gathering each shard. The tradeoff between timeliness and security is dependent on the nature of the task: if maximum security need only be maintained for a short period of time, it might be acceptable that an adversary has a higher chance of obtaining the information after a given time interval. Researchers have explored placing data fragments and replicas so as to minimize latency in a dynamic disruption-tolerant network, taking into account users' social network structures.[6,7]

The network's topology, architecture, communication protocols, and other characteristics also influence the optimal means of data dispersion and reassembly. Fogging/defogging must take into account the size, density, complexity, and tempo of the network, the mobility and geographic proximity of users and nodes where data shards are stored, how soon sharded information will become stale, how soon stored information might be needed, and so on.

SITUATIONAL AWARENESS AND INFORMATION SEMANTICS

The ultimate goal of information accessibility is situational awareness (SA), and even timely and relevant information delivery doesn't guarantee high-quality SA. Not all data shards are equally valuable from the SA perspective: a given shard could be used to create multiple pictures or draw multiple conclusions, depending on how it's "glued" to other shards. SA thus presents a challenge with respect to discovering as well as gathering dispersed information.

A cyberfog approach will require novel methods of information fusion to achieve adequate SA, especially when data gathering is incomplete due to an adversary action or network failures. This entails knowledge of the semantic context of the information, which strongly influences how recipients understand it. Toward this end, semantic information theory[8] and perhaps sheaf theory[9] seem highly relevant to addressing cyberfog challenges.

With a cyberfog approach, the network might need to manage a complex tradeoff between availability and confidentiality in real time depending on users' tasks and circumstances.

Context is particularly important in protecting business tasks because an adversary might need very little information to disrupt a key element of a task. Consequently, the data-dispersion, data-gathering, and SA-formation processes must be designed and executed in such a way that information has high value for the users and low value for the adversary. This implies the need for a thorough model of the adversary's intent and prior knowledge.

RISK ASSESSMENT

Risk could serve as a comprehensive framework for characterizing cyberfog effectiveness. However, new risk models are needed to model poorly understood phenomena such as obfuscation that play an important role in this approach.

It's tempting to formulate the risk of failure in terms of data, such as the fraction of data captured by an adversary, but it should be analyzed in terms of the impact on a given task's objectives. This implies the need for an accurate model of the task, including its dependencies on network and computing assets—a highly complex modeling problem.

Other complexities arise in quantifying the impact of failure: the same failure can have very different consequences depending on its timing or how old the lost information is—the loss of dated information could be less important than that of recently obtained information. Additive properties of failures are important too—for example, knowing data item A and data item B might have high value, whereas knowing only one of the items would have zero value. The risk of a cyberfog approach also increases with the uncertainty of failure: if I know I lost data item A, I can do something about it; but if I'm uncertain, the approach's effectiveness is lessened.

Risk assessment in a cyberfog strategy would clearly benefit from a game-theoretic treatment. In this case, risk is highly dependent on the decisions and actions of the opponents, who are interdependent. This kind of game deviates strongly from the traditional zero-sum game because participants operate with partial information, bounded rationality, and so on. In fact, even the game's goals—the task's objectives—can be subject to change if some supporting assets fail or are captured by the adversary. Further, the game involves deception and obfuscation.

DECEPTION AND OBFUSCATION

Data dispersion presents adversaries with uncertainty as to where to find relevant information and how to reconstruct it from captured shards. A cyberfog approach also uses obfuscation and deception to increase uncertainty for the adversary. Obfuscation subjects information to multiple, equally possible interpretations, whereas deception aims to induce an incorrect interpretation that thwarts the adversary's goals. Obfuscation and deception can be achieved in many ways—for example, by providing a misleading view of the network's topology, traffic, and behavior.

Regardless of the means employed, effective obfuscation and deception can be difficult to implement. For example, creating believable fake business documents or network traffic is very challenging. The task is even harder if an adversary is able to observe network behavior and system use across both the physical and cyber dimensions.

Determining the fundamental limits of adversaries' ability to detect obfuscation and deception is also challenging. Because these are human fabrications, they're likely to be far less complex and rich in detail than real-world activities. As such, they might be vulnerable to sophisticated machine-learning techniques designed to detect anomalies. Thus, research is needed on ways to fool particular classifiers with particular inputs.[10] As AI systems become pervasive and increasingly sophisticated, understanding the difference between how machines and humans perceive obfuscation and deception will be critical to cyberfog success.

Given the extreme challenges and complexities inherent in a cyberfog environment, the use of formal methods could provide some assurance that the environment as well as the tools and activities we design for it exhibit certain properties. Unfortunately, formal methods are expensive to implement and can't yet eliminate the need for conventional testing. Nor are formal methods suitable for novel types of cyberattacks that contravene current models' assumptions.[11] Furthermore, it's unknown how well, if at all, formal methods apply to human factors such as the role of cognition in deception. Perhaps some of these difficulties could be mitigated by purposefully designing a cyberfog strategy that's more amenable to formal methods.

REFERENCES

1. M. Chiang, "Fog Networking: An Overview on Research Opportunities," Dec. 2015; arxiv.org/pdf/1601.00835.pdf.
2. G. Bent et al., *Network and Information Sciences International Technology Alliance*, US Army Research Lab/UK Ministry of Defence, 2016; nis-ita.org/Legacy/files/book/ITA%20eBook%20PDF.pdf.
3. A. Mei, L.V. Mancini, and S. Jajodia, "Secure Dynamic Fragment and Replica Allocation in Large-Scale Distributed File Systems," *IEEE Trans. Parallel and Distributed Systems*, vol. 14, no. 9, 2003, pp. 885–896.
4. A. Shamir, "How to Share a Secret," *Comm. ACM*, vol. 22, no. 11, 1979, pp. 612–613.
5. J.Y. Chung et al., "CYRUS: Towards Client-Defined Cloud Storage," *Proc. 10th European Conf. Computer Systems* (EuroSys 15), 2015; www.princeton.edu/~cjoe/CYRUS_EuroSys.pdf.

6. X. Zhuo et al., "Social-Based Cooperative Caching in DTNs: A Contact Duration Aware Approach," *Proc. IEEE 8th Int'l Conf. Mobile Ad-Hoc and Sensor Systems* (MASS 11), 2011, pp. 92–101.
7. W. Gao et al., "Cooperative Caching for Efficient Data Access in Disruption Tolerant Networks," *IEEE Trans. Mobile Computing*, vol. 13, no. 3, 2014, pp. 611–625.
8. P. Basu et al., "Preserving Quality of Information by Using Semantic Relationships," *Pervasive and Mobile Computing*, vol. 11, 2014, pp. 188–202.
9. G.E. Bredon, *Sheaf Theory*, 2nd ed., Springer, 1997.
10. P. McDaniel, N. Papernot, and Z.B. Celik, "Machine Learning in Adversarial Settings," *IEEE Security & Privacy*, vol. 14, no. 3, 2016, pp. 68–72.
11. K. Schaffer and J. Voas, "What Happened to Formal Methods for Security?," *Computer*, vol. 49, no. 8, 2016, pp. 70–79.

DISCLAIMER
This article doesn't reflect the positions or views of the authors' employers.

ALEXANDER KOTT is chief of the Network Science Division, US Army Research Laboratory. Contact him at alexander.kott1.civ@mail.mil.

ANANTHRAM SWAMI is senior research scientist for network science at the US Army Research Laboratory. Contact him at ananthram.swami.civ@mail.mil.

BRUCE J. WEST is senior scientist in mathematics at the Information Sciences Directorate, Army Research Office, US Army Research Laboratory. Contact him at bruce.j.west.civ@mail.mil.

STUDENT DESIGN SHOWCASE

VIDEO

Tactile Digital Braille Display

Greg Byrd, North Carolina State University

Students at North Carolina State University enhanced Polymer Braille's multiline braille display by adding new interactive features, additional rows of characters, and a mobile-device interface.

According to the World Health Organization, 39 million people worldwide are blind and another 246 million have low vision (www.who.int/mediacentre/factsheets/fs282). To access electronic content, these individuals have two options: text-to-audio conversion—which isn't always appropriate or desirable—and tactile displays. Tactile displays use pins that raise and lower to form braille characters. In addition to being nearly silent, such braille displays are more useful for conveying highly technical text and mathematical notations.

A braille tactile display provides one row of up to 80 characters. Each character consists of six or eight dots arranged in rows of two. The characters are refreshed by raising or lowering pins as the user navigates the computer screen. However, these one-line displays can be limiting for readers—especially compared with printed braille, which allows for 2.5D graphics that readers can use to naturally navigate to different parts of the page. However, multiline tactile displays have been difficult to design because of space constraints, the force required for mechanical actuators, and the actuation speed needed to make them readable and usable.

Polymer Braille is developing a 2.5D braille display using electroactive polymer (EAP) technology with piezoelectric actuators that raise and lower the braille dots. While the company has focused on the actuator technology, students in North Carolina State University's electrical and computer engineering senior design class have been developing system prototypes. For these proof-of-concept systems, they used LEDs instead of actuators to represent the braille output. The LEDs let sighted software design engineers use the prototypes for content development and feature testing.

The 2014–2015 senior design team built an initial prototype that controlled the LEDs to translate text to braille characters. This was an output-only device, so users couldn't send text or commands to the computer. The 2015–2016 team enhanced the initial prototype: they added Bluetooth wireless communication, two-way communication using a braille keyboard with navigation buttons, and an enclosure that supported a 12 × 30 character display.

See **www.computer.org/computer-multimedia** for multimedia content related to this article.

HIGH-LEVEL DESIGN

Figure 1 is a high-level block diagram of the system components. From the computer, text is translated into

0018-9162/16/$33.00 © 2016 IEEE

EDITOR **GREG BYRD**
North Carolina State University;
gbyrd@computer.org

braille by the NonVisual Data Access (NVDA) open source software package (www.nvaccess.org). The braille characters are transmitted via USB or Bluetooth to the braille display device, where they're displayed on the LED array. Information can also be transmitted from various input devices to the computer, in this case, an 8-key braille keyboard and a scroll wheel. The keyboard follows the Perkins Brailler standard: each key corresponds to one dot of the braille character.

The device has two microprocessors: a main controller and a peripheral controller. The main controller manages communication between the device and the computer. When a row of characters has been buffered, the main controller sends the braille data to the LED array, which displays them. The main controller also receives information (text and control signals) from the peripheral controller, which it then sends to the computer via the USB or Bluetooth channel.

HARDWARE CHALLENGES

The students made several changes to the previous year's prototype, including adding a keyboard and changing the display from 4 × 40 to 12 × 30 characters.

They designed a new motherboard to accommodate the keyboard, as shown in the lower part of Figure 2. The motherboard includes both microcontrollers. The keyboard buttons communicate with the peripheral controller, which interprets the signals and sends characters to the main controller. The different components communicate using the I^2C (Inter-Integrated Circuit) serial bus protocol.

The main controller also drives the LED cards for the display, communicating directly with each card over I^2C. This is an improvement over the previous version, which used dedicated controllers for a pair of rows. The reduction

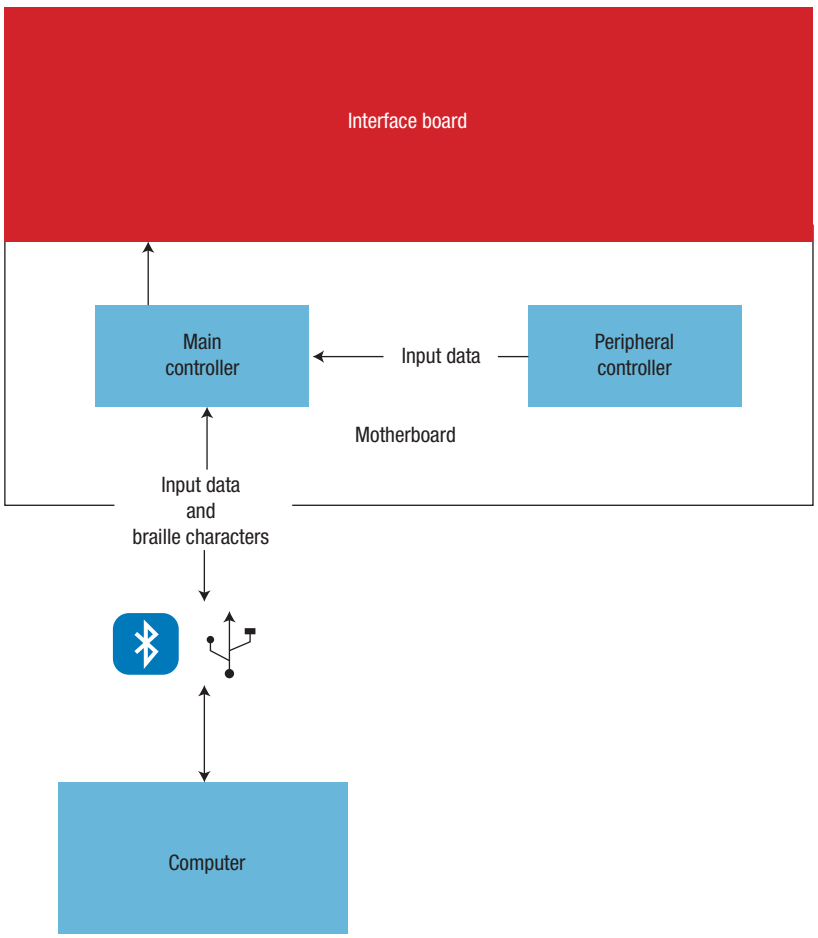


Figure 1. System components of the prototype 2.5D multiline braille display. Computer text is translated into braille characters that are transmitted via USB or Bluetooth to the braille display device and displayed on the LED array.

PROJECT DETAILS

- Title: Tactile Braille Display
- School: North Carolina State University, Raleigh, NC
- Students: Chris Baldrige, Zach Johnston, Daniel Mendoza, and Evan Wetherington (2015–2016); Richard Beauchemin, Kelina Farnham, Aasheesh Francis, and Nathan Hansen (2014–2015)
- Industry mentor: W. Shepherd Pitts (Polymer Braille, Raleigh, NC)
- Faculty mentors: Bobby Compton, Rachana Gupta, and Jean Shortley
- Support: National Science Foundation (SBIR Phase II grant IIP-1353625)

STUDENT DESIGN SHOWCASE

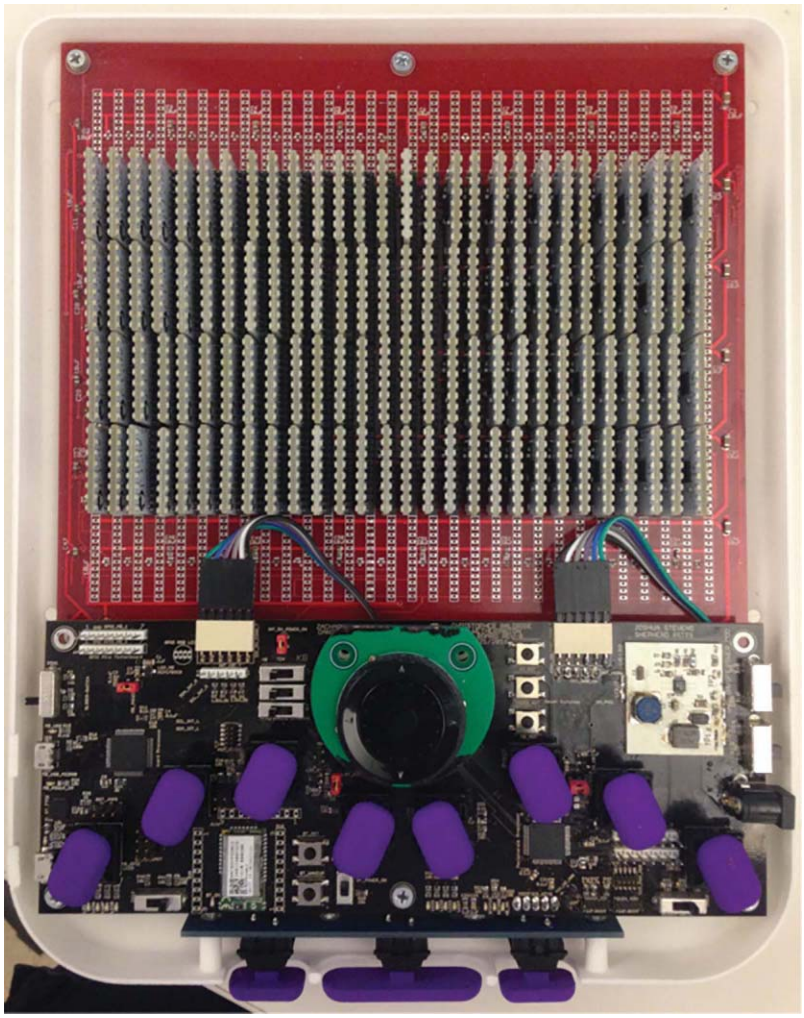


Figure 2. Final prototype internal components. The LED display is on top, with the microcontrollers and keyboard below.

in hardware allows the device to be thinner. As mentioned, the LEDs are temporary substitutes for the actuators that will raise and lower dots to form the braille characters. The actuators require relatively high voltage (hundreds of volts), which would require significant changes to the circuit board design. This requirement was determined to be out of this specific prototype's scope because of time limitations.

To implement the braille keyboard, the students connected key pads to linear switches that were connected to the peripheral controller. They chose a four-way navigational switch with a scroll wheel, rather than a joystick, because it felt more intuitive. Thumb-activated space, backspace, and enter keys were added along the bottom of the device.

The team also designed the outer enclosure by carefully modeling all of the internal printed circuit boards and components. They completed the modeling in SolidWorks (www.solidworks.com) and 3D-printed the enclosure with Shapeways (www.shapeways.com).

SOFTWARE CHALLENGES

This prototype's main software challenge, relative to the initial version, was the two-way communication between the device and the computer. The previous prototype was only a display, taking character data from the computer. This latest version requires the transmission of keyboard data from the device to the computer as well.

The NVDA software was used to translate text to braille. To handle input from the user, the student team implemented an NVDA plug-in that allows the software to respond to users' navigational and braille keystrokes. These events are translated into actions on the main computer screen.

In addition to input data from the computer, the main controller receives input from the peripheral controller. The peripheral controller detects the keystrokes and converts the braille to ASCII characters. It also detects when users have pressed the directional buttons on the scroll wheel or scrolled the wheel. This information is encoded and sent to the main controller over the I^2C bus.

The students successfully demonstrated their prototype at the university's May 2016 Design Day (watch an interview with the team at youtu.be/RjAXwlVyzNU?list=PL2Tk3txD5Sb1Q2bsJ7FRuFS4-9KvjsGD1). Moreover, they learned valuable lessons about communication and project management—although each member had clear responsibilities, they had to collaborate and adapt to the challenges posed by a complex project. Two of the students continue to work with Polymer Braille. The company anticipates bringing their product to market, offering new ways for people with visual impairments to interact with technology and more actively participate in the global digital community.

GREG BYRD is associate head of the Department of Electrical and Computer Engineering at North Carolina State University. Contact him at gbyrd@computer.org.

EDITOR HAL BERGHEL
University of Nevada, Las Vegas; hlb@computer.org

Chasing Elbridge's Ghost: The Digital Gerrymander

Hal Berghel, University of Nevada, Las Vegas

Gerrymandering has been with us for centuries, but now it's digital and out of control.

Elbridge Gerry's most lasting legacy may be the electoral abuse that bears his name: gerrymandering. A gerrymander is the manipulation of voting district boundaries for partisan effect. Although the term originated with Gerry's bid for the Massachusetts governor's office in the early 1800s to favor his Democratic-Republican party, the tactic it describes dates back at least as far as 1701.[1] Gerry's gerrymander failed to get him reelected, but its potential for manipulating election results was immediately recognized by the electoral-victory-at-any-cost partisans as a boon to oligarchy, and it's been used to great effect ever since.[2,3]

Recently, gerrymandering has been most successfully demonstrated in Republican redistricting efforts in selected states (including Arizona, Florida, Texas, North Carolina, Pennsylvania, New York, and Ohio) after the 2010 census. To illustrate the effect, in 2012, 1.4 million more Americans voted for Democrats than Republicans, and yet the Republicans won 33 more seats in the House of Representatives. As David Daley points out, "This was the first time since 1972 ... and only the second time since World War II that the party with the most votes did not also win the most seats."[2] And the consensus is that it's very likely to stay that way until 2022![3,4]

To some, including a few recent members of the Supreme Court, gerrymandering is politics as usual. But to others, it's a shift further away from our democratic ideals in that it violates the principle of one person, one vote—a proposition recognized by the Court since *Reynolds v. Sims* in 1964 that many of us hold dear. Regardless of political persuasion, all agree that gerrymandering is an effective way to manipulate election outcomes and circumvent the will of the majority. This is relevant to computer science, because computer programs are now available that can either make gerrymandering worse or mitigate against its harmful effects.

In our profession, we might think of gerrymandering as an insidious application of an exact cover algorithm: Given a state S consisting of k subsets (congressional districts), find a cover such that each element (voter) is in exactly 1 subset. Now, if you're a partisan tribalist who never really bought in to the one-person, one-vote business, you'd be tempted to find an exact cover such that your party receives the most seats even though it receives the least votes. That's where computer-based gerrymandering

comes in. There are computer programs that find such covers, and politicians that use them to redistrict to suit their parochial interests. There are also computer programs that can minimize the effects of gerrymandering that for the most part go unused. This column is about these programs and the political environment in which they arise.

THE LAW

Partisan politicians and special interests have always found the gerrymander intoxicating; it reduces the discomfort accompanying free and fair elections that can undo their efforts toward domination. In this regard, gerrymandering shares common cause with Plato's "noble lies" and Martin Heidegger's postmodern definition of truth as "that which makes a people certain, clear, and strong." These positions attempt to thwart the untidy side of democracy where majority rules (also known as "mob rule"). Our republic is in essence an outgrowth of the Founding Fathers' concern with this untidiness. Unfortunately, their preoccupation with the possibility of majority tyranny meant that they ignored other forms of oppression like oligarchy. That is, they addressed tyranny by majority, but not tyranny by minority.

The real question is: whose opinion matters in elections? Universal suffrage was never seriously considered by the Constitution's Framers, and no mention of it is made in the Constitution, therefore, the precise makeup of the electorate is fuzzy. So we're left with this loophole for sundry types of election manipulation, whether through voter disenfranchisement, voter suppression, vote nullification, or vote dilution.[5,6] Of these, the gerrymander is but one special type. The Constitution only insists that states should apportion seats based on the national census, but it's silent on how this should be accomplished. Not surprisingly this loophole translated into selective representation by ethnicity, gender, wealth, property ownership, taxes paid, and so on; this situation remained until 1964 when the Court ruled in *Reynolds* that the equal protection and due process clauses in the Bill of Rights demanded that each vote must be given roughly equal weight.[7] That can't happen when voting districts are dramatically malapportioned. Although *Reynolds* introduced the concept of one person, one vote into law, it did little to prevent gerrymandering because it failed to take into account the fact that American politicians are allowed to choose their electorate. As *Miami Herald* columnist Fred Grimm poignantly remarked, "[politicians] can't be trusted to put the public interests over their own job security" (www.miamiherald.com/news/local/news-columns-blogs/fred-grimm/article1978425.html). We'll return to this observation in the conclusion of this column.

The Supreme Court determined in *Davis v. Bandemer* (1986) that partisan gerrymandering was justiciable, but remained silent on what standards might be applied by the courts and legislatures. The only guidance the Court provided was the observation that "unconstitutional discrimination occurs only when the electoral system is arranged in a manner that will consistently degrade a voter's or a group of voters' influence on the political process as a whole."

> We might think of gerrymandering as an insidious application of an exact cover algorithm.

In 2004, judicial confusion increased when the Court ruled 5–4 to deny claims of gerrymandering in *Vieth v. Jubelirer* on the basis that while such claims aren't in fact justiciable without reasonable standards to determine unconstitutional electoral discrimination in place, it was possible that satisfactory standards might emerge in the future. In fact, it was in this case that four members of the Court sought to overturn *Davis* outright, but they failed to get a majority. However, in *League of United Latin American Citizens v. Perry* (2006), five justices expressed willingness to adopt a gerrymandering standard if a sound one could be found. In particular, they looked positively on the proposed standard of "partisan symmetry" proposed by Bernard Grofman, Gary King, and others.[8,9]

John Mackenzie, as well as Nicholas Stephanopoulos and Eric McGhee, further quantified partisan symmetry by calculating an "efficiency gap"; this measurement determines the efficiency with which votes are translated into seats—inefficient votes don't lead to victory.[10,11] The degree of inefficiency is the ratio of wasted votes to total votes in any election, and a vote is considered wasted when it's: cast for a winning candidate in excess of what was needed to win (accomplished through packing); cast for a losing candidate (cracking); or canceling out minority votes even though electoral outcomes are assured (stacking). In this way a gerrymander is essentially redistricting to force one party to waste more votes than another.[12] Plotting the efficiency gaps for both state and congressional districts from 1972 to 2012, Stephanopoulos and McGhee found that the Republican dominance since 2012 is due to "extreme gerrymandering," with the "highest levels recorded in the modern era."[11]

An update on the nebulous law principle of one person, one vote is in order. In 2016, the Court began to seriously consider judiciable gerrymandering standards in *Evenwel v. Abbott*, but even then the court waffled. *Evenwel* held that states can redistrict based on total population, but it didn't rule out other alternatives

such as voter population or number of citizens for state and local elections. This is an important issue because these elections determine the bias of congressional districts (read: the extent to which they are gerrymandered). However, the very standard that's used for measuring a district's "size" is highly partisan—because Republicans tend to be more successful at improving voter registration and turnout than Democrats (www.scotusblog.com/2016/04/opinion-analysis-leaving-a-constitutional-ideal-still-undefined), apportionment by numbers of registered voters rather than total population will produce a Republican advantage. This is even more critical since the Court overturned Section 4 of the Voting Rights Act of 1965 in *Shelby County v. Holder* (2013) (supreme.justia.com/cases/federal/us/570/12-96/#).

The practice of gerrymandering is widespread, although the effects are in some cases subtle. However, as can be seen in the references mentioned above, social scientists and legal scholars have made considerable progress on measuring whether and to what degree redistricting plans gerrymander. They might not be optimal, but they're reasonable.

ENTER ALGORITHMS

I've set a legal framework for the most significant gerrymander in history: the Republican State Leadership Committee (RSLC)'s REDMAP (Redistricting Majority Project), which was the Republican party's effort to dominate state governments and the House of Representatives in 2010.[3,4,13] Let there be no confusion over this—this was a national gerrymander to avoid as many competitive elections (that is, waste as many non-Republican votes) as possible. It worked. The party that had the most votes for House of Representatives candidates in both 2012 and 2014 lost control.

Because of this gerrymandering, FairVote claims that the majority party (Democrat) would have to win the 2016 national vote by more than 12 percent to earn a one-seat majority (www.fairvote.org/monopoly_politics). FairVote has successfully predicted congressional election outcomes two years prior to the elections for many years.

> Partisan analysts pack, crack, and stack their way to gerrymandered nirvana as they maximize their opponents' wasted votes.

What makes this possible? In a word, computers. That is, geographic information system (GIS) tools like Maptitude, Redistricter, iRedistrict, and MapInfo Pro. The tool most often associated with REDMAP is Maptitude (Caliper Corporation); of course, it isn't the tool per se that causes the gerrymandering, but partisan users who seek to subvert the will of the majority. GIS redistricting tools are the nuclear option for geopolitical mapping analysts. As Daley documented in his recent book, the capability exists for designing districts by race, ethnicity, gender, political affiliation, demographics, party registration, voter turnout, previous election behavior, and so forth.[3] This data is primarily, but not exclusively, driven by current census data. By combining and layering such data, answers to "what if?" questions become trivial. Partisan analysts pack, crack, and stack their way to gerrymandered nirvana as they maximize their opponents' wasted votes. In this manner, election outcomes are easily predicted. In 2012 for example, FairVote's Monopoly Politics project predicted 2014 congressional district outcomes with 99.7 percent accuracy—all due to the proliferation of noncompetitive districts in the US. Thus, FairVote concluded, "In the vast majority of cases, the particulars of candidates and campaigns have little impact on the end result. Uncompetitive races mean that outcomes are essentially predetermined, leaving voters without meaningful choices or a compelling reason to go to the polls" (www.fairvote.org/monopoly_politics).

This is how the notoriously distorted districts in Maryland, Pennsylvania, Ohio, New York, Wisconsin, North Carolina, Florida, Michigan, Iowa, Arizona, and Texas came to look the way they do.[3] By aligning the electorate into gerrymandered voting districts, the relationship between votes cast and number of seats taken might be severed for a decade at a time.

So although gerrymandering has been with us since the country's founding, 2010 marked the watershed in vote nullification and dilution by manipulating efficiency gaps with redistricting software. In the hands of partisan political operatives, this software puts an electoral twist on the social sorting that has characterized American life since the Europeans first arrived here.

What can be done? Once again, we look to computers. What's needed is serious software development in two key areas: the ability to recognize the gerrymander, and the ability to redistrict without the gerrymander. As a search of any computer science digital library will confirm, there have been a variety of attempts to automate redistricting over the past 40 years. However, they share a common weakness: they aren't tuned to recognize and prevent computer-based partisan gerrymandering, which only went viral in 2010. This is a new grand challenge for computing professionals (read more about software development in this area in the "Redistricting Research" sidebar).

Although there might be complexity in full automation of redistricting,[14] semi-automated programs should be able to *remove* partisan bias from redistricting because they've been used

REDISTRICTING RESEARCH

Research and development in redistricting tools has been ongoing since 2010. The following projects take interesting approaches in this area.

- Brian Olson's "optimally compact" equal-population system for impartial automatic redistricting. The site illustrates the results for all 50 states (http://bdistricting.com/2010).
- Kevin Baas's Auto-Redistrict program is based on genetic algorithms (http://autoredistrict.org).
- The Center for Range Voting's shortest splitline algorithm geographically splits state populations in halves, fourths, eighths, and so on, until a given number of districts are created (http://rangevoting.org).
- James Case's isoperimetric approach (www.siam.org/pdf/news/1237.pdf).
- Gregory B. Lush, Esteban Gamez, and Vladik Kreinovich's iterative clustering algorithm (www.cs.utep.edu/vladik/2007/tr07-51a.pdf).

for a decade to *inject* partisan bias into it. In addition, various voting systems reduce the impact of gerrymandering such as ranked choice voting with automatic runoffs (fairvote.org/rcv). There's no reason not to attack the gerrymander from two fronts. Although I've no solutions to recommend at this point, my hunch is that most if not all of the proposed solutions are better than what we have. If society will accept that optimal redistricting ensures outcomes consistent with the will of the majority, I'm confident that computer professionals could handle the rest.

There are many ways to characterize an election, from illegitimate or fraudulent at one extreme to transparent, fair, and open at the other. I'm not sure that we'll ever completely achieve the latter, but with modern computer technology we're in a position to move a lot closer by reducing the harmful effects of vote nullification and dilution through gerrymandering. All we lack is the will.

One major problem is that the public doesn't appreciate the extent to which gerrymandering corrupts government. Nor do they completely understand technology's role in the corruption. It doesn't help that journalists and scholars try to pin the problem on elusive algorithms. For example, a recent *InformationWeek* column led with "Why can't a simple formula replace the politically charged gerrymandering that's skewing our election processes?" (www.informationweek.com/government/open-government/wanted-honest-algorithms-for-voter-redistricting/a/d-id/1297859). A simple formula can be found, but this misses the point entirely.

The business of the gerrymander is to prevent competitive elections. The reason it persists is that it serves those in power and the special interests that put them there. In the 2010 Republican gerrymander, it served one party. But in the case of incumbents, it serves all parties. The fact is that those who hold "safe" (read: noncompetitive) seats—despite their public proclamations to the contrary—are willing supporters of gerrymandering, maybe more so than the wannabes. Where might we find one willing to lose a seat to uphold democratic principles? Politicians don't think that way, as Grimm so wisely noted.

Thus, any viable gerrymandering solution must include the proposition that elected officials are forever prevented from determining their own constituents. Further, it must prevent political operatives from manipulating independent redistricting commissions. Although these are admirable goals, realistically, any attempt to remove bias from politics defies experience. Therefore, the low-hanging fruit is to enlist committed computing professionals to the cause.

One straightforward approach is to build an academic consensus for the courts to consider that would spell out what optimal redistricting and gerrymandering identification algorithms might look like. Although current literature reveals a foundation, it doesn't convey any urgency. For that to happen, redistricting algorithms need to be drawn into mainstream computing curriculum and research. One of our grand challenges in computing should be to develop digital technology in support of free, open, and fair elections. It's a concept of global importance and application. Unfortunately, the art and science of digital gerrymandering currently is no more popular in computer science curricula than control fraud is in business schools—both topics appear resilient to rigorous study in the academy, and for much the same reasons.

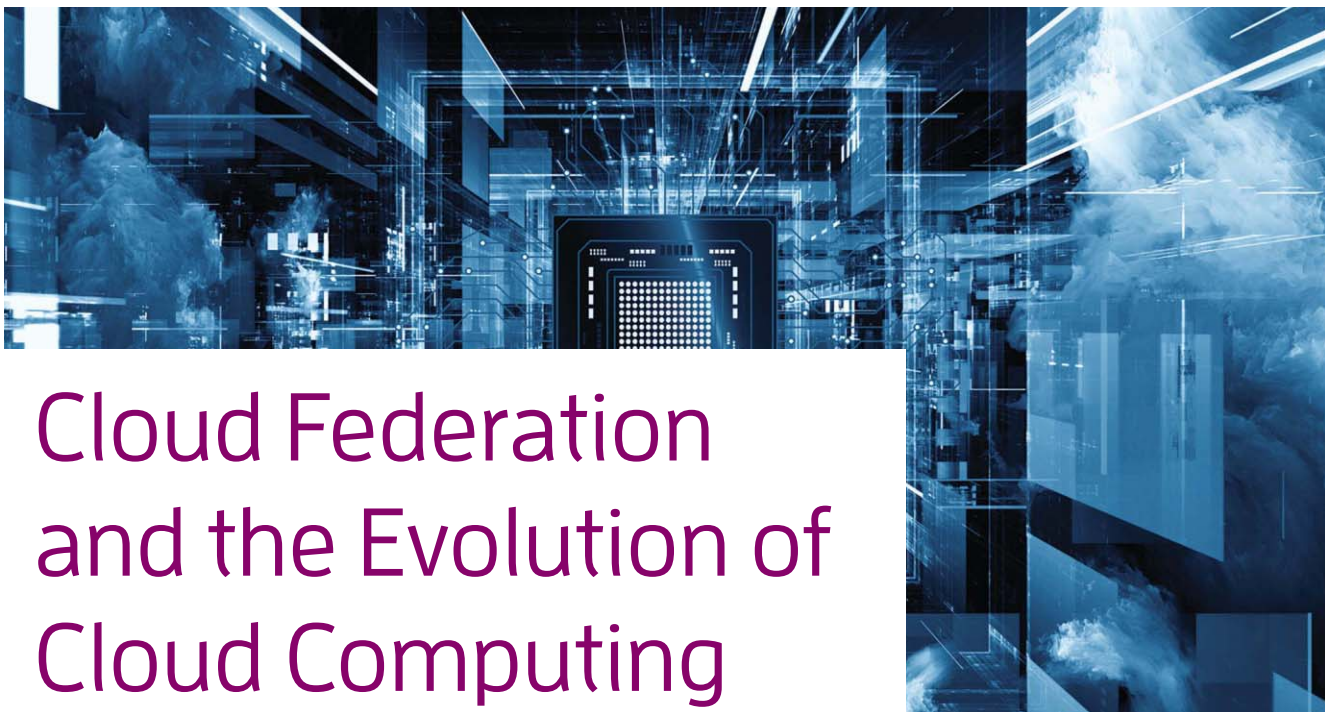
REFERENCES

1. E. Griffith, "The Rise and Development of the Gerrymander," PhD Dissertation, University of Chicago, published by Scott, Foresman and Company, 1907; ia802502.us.archive.org/29/items/risedevelopmento00grif/risedevelopmento00grif.pdf.
2. T. Wang, *The Politics of Voter Suppression*, Century Foundation, 2012.
3. D. Daley, *Ratf**ked*, Liverlight/Norton, 2016.
4. P. Pierson and J. Hacker, *Off Center*, Yale Univ. Press, 2006.
5. H. Berghel, "Digital Politics 2016," *Computer*, vol. 49, no. 1, 2016, pp. 75–79.
6. T. Campbell, *Deliver the Vote: A History of Election Fraud, an American*

Political Tradition—1742–2004, Carroll & Graf, 2004.
7. I. Somin, "Competing Interpretations of One Person, One Vote," *The Washington Post*, 3 Aug. 2015; www.washingtonpost.com/news/volokh-conspiracy/wp/2015/08/03/alternative-interpretations-of-one-person-one-vote/?utm_term=.f25ccaa43a3d.
8. B. Grofman and G. King, "The Future of Partisan Symmetry as a Judicial Test for Partisan Gerrymandering after *LULAC v. Perry*," *Election Law J.*, vol. 6, no. 1, 2007; gking.harvard.edu/files/jp.pdf.
9. R. Browning and G. King, "Seats, Votes, and Gerrymandering: Estimating Representation and Bias in State Legislative Redistricting," *Law and Policy*, vol. 9, no. 3, 1987; gking.harvard.edu/files/LP9.pdf.
10. J. Mackenzie, "Gerrymandering and Legislator Efficiency," Feb. 2010; udel.edu/johnmack/research/gerrymandering.pdf.
11. N. Stephanopoulos and E. McGhee, "Partisan Gerrymandering and the Efficiency Gap," *Univ. of Chicago Law Rev.*, 2015, pp. 831–900; lawreview.uchicago.edu/sites/lawreview.uchicago.edu/files/uploads/82_2/04%20Stephanopoulos_McGhee_ART.pdf.
12. R. Miller, "Gerrymandering Politics Out of the Redistricting Process: Toward a Planning Revolution in Redrawing Local Legislative Boundaries," The Urban Fringe blog, *Berkeley Planning Journal*, 2012; ced.berkeley.edu/bpj/2012/09/gerrymandering-politics-out-of-the-redistricting-process-toward-a-planning-revolution-in-redrawing-local-legislative-boundaries.
13. "The Redistricting Majority Project," Republican State Leadership Committee; redistrictingmajorityproject.com.
14. M. Altman and M. McDonald, "The Promise and Perils of Computers in Redistricting," *Duke J. of Constitutional Law and Public Policy*, vol. 5, no. 1, 2010, pp. 69-111; scholarship.law.duke.edu/djclpp/vol5/iss1/5.

HAL BERGHEL is an IEEE and ACM Fellow and a professor of computer science at the University of Nevada, Las Vegas. Contact him at hlb@computer.org.

CLOUD COVER



Cloud Federation and the Evolution of Cloud Computing

Dimitrios G. Kogias, Michael G. Xevgenis, and Charalampos Z. Patrikakis,
Piraeus University of Applied Sciences

To satisfy the demand for collective and collaborative cloud use, academia and industry want to interconnect heterogeneous clouds to form a federated system. This approach is promising but also faces significant challenges.

Cloud computing allows users to access computing services and resources on demand without having to buy their own infrastructures, and to pay only for what they use.[1] Many cloud companies—such as Amazon and Google—have developed their own platforms featuring proprietary interfaces, which isn't a problem as long as a single provider can fully satisfy its customers. However, the lack of standardization for interconnecting platforms makes it difficult for customers who need the combined services or resources of multiple providers. This often results in users being locked into specific providers and platforms.[2,3]

This issue has led to the idea of interconnected clouds, also known as *interclouds*.[2–5] Interclouds address single-provider approaches' limitations such as the lack of interoperability between platforms, limited resources being exhausted during times of peak customer demand, service interruptions, and quality-of-service (QoS) degradation.

INTERCLOUD

An intercloud is a cloud of clouds.[3] In essence, it's a large cloud comprising many smaller clouds, each having its own characteristics and serving different needs. An intercloud implementation could be any one or combination of

- *hybrid clouds*, in which private clouds access the resources of public clouds without the latter being aware of their participation;
- *multiclouds*, which utilize libraries from applications that enable the use of resources from multiple clouds, without any of them being aware of their participation;
- *sky computing*, an emerging model in which resources from multiple cloud service providers (CSPs) create a large, distributed, virtual infrastructure

0018-9162/16/$33.00 © 2016 IEEE

EDITOR SAN MURUGESAN
BRITE Professional Services; san@computer.org

able to support the establishment of trust between different clouds that might not be configured to trust or even recognize one another;[6]
- *multiclouds tournament*, an architecture comprising multiple clouds that utilizes a tournament model to balance resource offerings with users' consumption, thereby providing higer-quality services;[7] and
- *cloud federations*, an interconnected set of heterogeneous public and/or private clouds from voluntarily participating users and providers.[2,3]

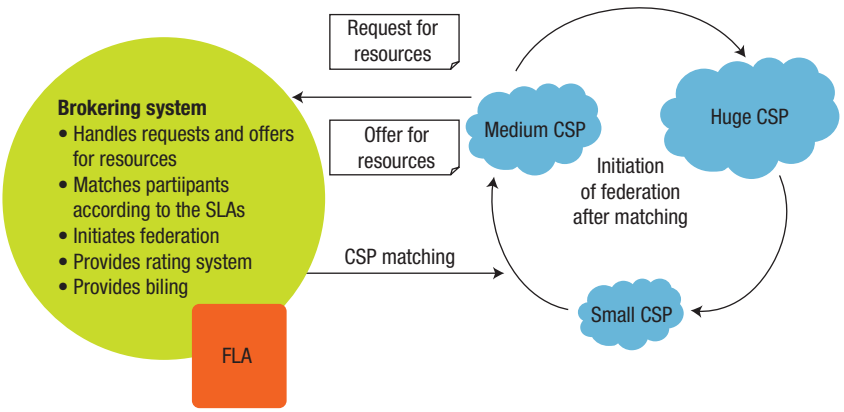


Figure 1. Cloud federation architecture. Users send requests for resources and cloud service providers (CSPs) send their responses to the broker (left), which matches users with providers based on billing, ratings, and service-level agreements (SLAs). This results in a federation (right), governed by a federal-level agreement (FLA).

CLOUD FEDERATION

Intercloud researchers have shown the most interest in cloud federations because it enables power-efficient, cost-effective, dynamic sharing of idle cloud resources and services. Federation members can sign service-level agreements (SLAs) to ensure QoS and availability.

The federation should

- have a defined marketing system that describes the cost of utilizing resources and services and that helps to valorize use,
- feature efficient geographic dispersion by allocating resources close to users to eliminate network problems that could interrupt service access, and
- follow rules in a federal-level agreement (FLA) describing the cooperation and relationship among participating clouds.

We disagree with the research literature's frequent interchangeable use of the terms "cloud federation" and "intercloud." In federations, cloud organizations participate voluntarily after signing an FLA. In an intercloud organization, no private or public cloud is necessarily aware of its participation. Also, interclouds are based on open standards that provide interfaces for interoperability. Cloud federations use a broker to translate and connect CSPs' own interfaces.

CLOUD FEDERATION ARCHITECTURE

For federations or interclouds to work properly, heterogeneous clouds must be able to interoperate. However, this can be difficult to achieve. For example, participating clouds might use different techniques to describe the services they offer. Users, however, need a mechanism to provide common access to available services. Thus, the cloud federation's architecture must employ interface standards, a service broker that translates between interfaces and provides updates on offered services and users' status changes, or a combination of the two.[3,8]

Cloud federations most often use brokerages. The common object request broker architecture (CORBA) and object request broker (ORB) middleware were initially the most popular approaches.[9] However, the advent of XML-based technologies such as SOAP has provided the ability to use the same language in the descriptions of all services, thereby avoiding the need for translation.

Figure 1 shows a cloud federation architecture with the broker playing a central role and the CSPs at the edges communicating mainly through the broker. The brokering system is in the cloud and matches the available federation resources with user demand, taking into consideration participants' SLAs. To achieve this, the broker must understand the various ways that each cloud describes its available resources and services[2,3] and then combine the gathered information seamlessly for the user. In some cases, the broker could provide users with resource and service pricing information, as well as bill them.

For the federation to function properly, all interested parties must sign an FLA that specifies interconnection rules and describes each participant's responsibilities and permissible behaviors, along with the financial, administrative, or other penalties for violating its terms. The parties can leave

the federation when they want, as long as they follow FLA procedures.

ADVANTAGES AND LIMITATIONS OF CLOUD FEDERATION

Cloud federations have pros and cons.

Advantages

Federation performance is guaranteed by the dynamic resource allocation—or *elasticity*—that lets clouds ask for other participants' idle resources or services when their own are exhausted. This achieves both uninterrupted service delivery and resource scalability, the latter being the result of the seamless, transparent operation between clouds for the delivery of an agreed-upon QoS level.

Federations also enable the geographic dispersion of resources, efficiently locating some near users[10] but also allowing participants to access more distant resources in case of local outages. This enables efficient commercialization of the offered resources and lower prices than single-cloud services can charge.[11]

And because the FLA clearly describes what each participant is offering, as well as the federation's rules, it ensures the commitment of the involved parties to the operation's performance.

Limitations

Although federation mechanisms can provide the agreed-upon performance, constant monitoring and increased security mechanisms are required to guard against accidents and malicious users.

Selecting which services a federation will offer is not trivial because they will have to come from multiple providers that have different cloud characteristics and that offer varying QoS levels. Thus, federation participations should deploy a service-selection mechanism, preferably automated, that uses a predefined set of criteria regarding the QoS that providers offer. Or they could dynamically negotiate SLAs to address user needs.

Federation members could also address the lack of a common repository for available services via peer-to-peer approaches using a distributed hash table overlay network for service discovery.[12] They could also utilize an intercloud root,[13] which produces an abstract view of a global catalog of federation services and resources offered in the connected clouds.

The mobility of virtual machines (VMs), which are common in cloud services, is important for providing uninterrupted performance and expected QoS levels. Hosts must meet requirements for factors such as memory use, state, status of running processes and applications, and LAN connectivity to be able to migrate a live VM from one physical node to another without disrupting network traffic. This is particularly critical in real-time services. In cloud environments, this migration could be challenging for VMs belonging to different clouds that have never shared resources and thus have no knowledge about each other's networking configurations. Thus, it's important to re-create the originating cloud's networking and communication environment in the destination cloud quickly enough to avoid excessive delays.

Federation participants must address data portability, focusing particularly on issues such as security and privacy, because services belonging to one CSP must frequently access data stored in another cloud.

> Cloud federations enable power-efficient, cost-effective, dynamic sharing of cloud providers' idle resources and services. This approach could promote more collaborative use of the cloud, but it also faces significant challenges.

LOOKING AHEAD

Early attempts at cloud federations haven't had all the characteristics that a true federation should possess. Instead, there have been multiclouds or hybrid clouds enhanced with some federation characteristics. However, these aren't as efficient as fully federated approaches.

True federations require brokering systems that can quickly communicate with cloud interfaces and find the right combination of resources and QoS to meet users' needs in the heterogeneous environment. In the process, the brokerages must keep in mind users' performance and cost requirements.

Content delivery networks (CDNs)—which have successfully provided high-quality data access for many users over the Internet—could serve as the framework for cloud-broker communication. But regardless of which approach is adopted, the CSPs' role is important, particularly for providing APIs that enable communication with brokers. Standards organizations such as IEEE could also play a major role in cloud-federation evolution by developing a reliable brokering system that is compatible with most cloud frameworks.

Federation participants must take special care in composing the terms of an FLA, which is the mechanism that ensures the system's integrity. A key concern is translating abstractly expressed requirements into concrete technical terms and functionalities.

Other issues include the establishment of trust among participants and the security of resource access and use, which is extremely important

in a dynamic environment such as a cloud federation.

IEEE's effort[14] to introduce a standard for a brokering-system is an important step toward the realization of true cloud federations. Researchers should also examine the characteristics proposed in different cloud technologies and architectures—such as fog computing's local hardware awareness[15]—that provide the technical capabilities that VMs could use to learn about cloud environments.

CALL FOR COLUMN CONTRIBUTIONS

For this column, we welcome short articles (1,500 to 2,000 words) discussing your ideas for advancing cloud computing or sharing your experiences in harnessing the cloud. We also solicit articles on recent developments, future trends, case studies, and topics such as cloud governance, cloud management and monitoring, risk management, disaster recovery, open source cloud computing, pricing, cloud economics, service-level agreements, standards, compliance, and legal issues. Please send proposals or submissions to San Murugesan at cloudcover@computer.org. For a list of previous Cloud Cover columns, visit http://tinyurl.com/computer-cloudcover.

REFERENCES

1. *Encyclopedia of Cloud Computing*, S. Murugesan and I. Bojanova, eds., Wiley-IEEE Press, 2016.
2. M.R.M. Assis and L.F. Bittencourt, "A Survey on Cloud Federation Architectures: Identifying Functional and Non-functional Properties," *J. Network and Computer Applications*, vol. 72, September 2016, pp. 51–71.
3. A.N. Toosi, R.N. Calheiros, and R. Buyya, "Interconnected Cloud Computing Environments: Challenges, Taxonomy, and Survey," *ACM Computing Surveys*, vol. 47, no. 1, 2014, pp. 7:1–7:47.
4. D. Bernstein and Y. Demchenko, "The IEEE Intercloud Testbed—Creating the Global Cloud of Clouds," *Proc. IEEE 5th Int'l Conf. Cloud Computing Technology and Science* (CloudCom 13), 2013, pp. 45–50.
5. B. Di Martino et al., "Towards an Ontology-Based Intercloud Resource Catalogue—The IEEE P2302 Intercloud Approach for a Semantic Resource Exchange," *Proc. IEEE Int'l Conf. Cloud Eng.* (IC2E 15), 2015, pp. 458–464.
6. K. Keahey et al., "Sky Computing," *IEEE Internet Computing*, vol. 13, no. 5, 2009, pp. 43–51.
7. M.R.M. Assis, L.F. Bittencourt, "Multiclouds Tournament Blueprint," *Proc. IEEE/ACM 8th Int'l Conf. Utility and Cloud Computing* (UCC 15), 2015, pp. 404–405.
8. G. Zangara et al., "A Cloud Federation Architecture," *Proc. 10th Int'l Conf. P2P, Parallel, Grid, Cloud, and Internet Computing* (3PGCIC 15), 2015, pp. 498–503.
9. M. Henning, "The Rise and Fall of CORBA," *ACM Queue*, vol. 4, no. 5, 2006, pp. 28–34.
10. L. Hongxing et al., "Virtual Machine Trading in a Federation of Clouds: Individual Profit and Social Welfare Maximization," *IEEE/ACM Trans. Networking*, vol. 24, no. 3, 2016, pp. 1827–1840.
11. J. Weinman, "Intercloudonomics: Quantifying the Value of the Intercloud," *IEEE Cloud Computing*, vol. 2, no. 5, 2015, pp. 40–47.
12. R. Ranjan and L. Zhao, "Peer-to-Peer Service Provisioning in Cloud Computing Environments," *J. Supercomputing*, vol. 65, no. 1, 2013, pp. 154–184.
13. D. Bernstein and D. Vij, "Intercloud Directory and Exchange Protocol Detail Using XMPP and RDF," *Proc. 6th World Congress on Services* (Services 10), 2010, pp. 431–438.
14. *IEEE P2302/D0.2 Draft Standard for Intercloud Interoperability and Federation (SIIF)*, IEEE, 2012; www.intercloudtestbed.org/uploads/2/1/3/9/21396364/intercloud_p2302_draft_0.2.pdf.
15. M. Zhanikeev, "A Cloud Visitation Platform to Facilitate Cloud Federation and Fog Computing," *Computer*, vol. 48, no. 5, 2015, pp. 80–83.

DIMITRIOS G. KOGIAS is an adjunct lecturer and a senior researcher in the Piraeus University of Applied Sciences' Department of Electronics Engineering. Contact him at dimikog@teipir.gr.

MICHAEL G. XEVGENIS is a junior researcher in Piraeus University of Applied Sciences' Department of Electronics Engineering. He is also a postgraduate student at Kingston University. Contact him at mxevgenis@teipir.gr.

CHARALAMPOS Z. PATRIKAKIS is an associate professor in the Piraeus University of Applied Sciences' Department of Electronics Engineering. Contact him at bpatr@teipir.gr.

CALL AND CALENDAR

CALLS FOR ARTICLES FOR *COMPUTER*

Computer plans an August 2017 special issue on collective computing.

Since the middle of the past decade, key cloud, crowdsourcing, Internet of Things, and wearable technologies—the cloud, the crowd, and the shroud—have emerged that together integrate machine intelligence with the collective intelligence of humans to deliver on-demand expertise to individuals and organizations. This collective-computing approach, which Georgia Tech professor Gregory Abowd proposed in an article in *Computer*'s January 2016 issue as a way to extend ubiquitous computing, has given rise to a new class of intelligent systems and applications.

This special issue will explore collective computing. Articles should address systems, applications, or evaluations of real-world deployments that reflect the combined use of the cloud, the crowd, and the shroud, as well as the challenges of creating collective-computing middleware, programming models, or methodologies.

Abstracts are due **20 December 2016** (mail to co-0817@computer.org), and articles are due **5 January 2017**. Visit www.computer.org/computer/cfp8 to view the complete call for papers.

Computer plans a September 2017 special issue on blockchain technology in finance.

Blockchain distributed-database technology is revolutionizing the finance industry. Key blockchain applications and technologies include cryptocurrencies such as bitcoin, as well as distributed ledgers, which are distributed databases that keep all transactions in a shared, replicated, synchronized bookkeeping record secured by cryptographic sealing. The technology is also used in smart contracts, which codify legal contracts, automatically execute contract terms, and codify laws and statutes.

Experts now recognize that blockchain technology has great potential in other areas including healthcare, pharmaceuticals, energy, insurance, and legal services.

For this special issue, *Computer* seeks submissions covering projects, associated research, in-depth surveys, reviews, and tutorials across blockchain-technology domains, particularly finance.

Abstracts are due **20 December 2016** (mail to co-0917@computer.org), and articles are due **1 February 2017**. (Guest editors will consider articles until this date even if authors didn't submit an abstract.) Visit www.computer.org/computer/cfp9 to view the complete call for papers.

CALLS FOR ARTICLES FOR OTHER IEEE CS PUBLICATIONS

IEEE Software plans a July/August 2017 special issue on reliability engineering for software.

As the number of smart, interconnected devices in our cars and homes grows, engineers must increasingly consider software reliability in the connectivity tools and platforms they create. Reliability engineering emphasizes dependability, whether at a critical moment or throughout the software's lifecycle.

This theme issue focuses on reliability challenges and successes in software engineering.

The guest editors seek articles featuring case studies, experience reports, practices, approaches, techniques, and guidelines, all involving practical software results.

Articles are due **1 December 2016**. Visit www.computer.org/software/cfp4 to view the complete call for papers.

IT Professional plans a July/August 2017 special issue on cognitive computing.

Cognitive computing refers to smart systems that learn at scale, reason with purpose, and interact with humans and other smart systems. Rather than being explicitly programmed, such systems learn and reason from their interactions with us and from experiences with their environment.

SUBMISSION INSTRUCTIONS

The Call and Calendar section lists conferences, symposia, and workshops that the IEEE Computer Society sponsors or cooperates in presenting.

Visit www.computer.org/conferences for instructions on how to submit conference or call listings as well as a more complete listing of upcoming computer-related conferences.

0018-9162/16/$33.00 © 2016 IEEE

This special issue seeks to provide readers with an overview of current cognitive-computing issues and practices, as well as a look into the future.

Articles are due **1 December 2016**. Visit www.computer.org/itpro/cfp4 to view the complete call for papers.

IEEE Transactions on Emerging Topics in Computing (TETC) plans the following special sections for its December 2017 issue, with articles due **1 December 2016**:

Emerging topics for disaster management: Visit www.computer.org/cms/computer.org/transactions/cfps/cfp_tetcsi_etdm.pdf for the complete call for papers.

Cyber social computing and cyber-enabled applications: Visit www.computer.org/cms/computer.org/transactions/cfps/cfp_tetcsi_csccea.pdf for the complete call for papers.

IEEE Transactions on Emerging Topics in Computing (TETC) and *IEEE Transactions on Learning Technologies (TLT)* plan a joint special section for their October–December 2017 issues, with articles due **1 December 2016**:

Innovation in technologies for educational computing: Visit www.computer.org/cms/computer.org/transactions/cfps/cfp_tetcsi_itec.pdf for the complete call for papers.

IEEE Internet Computing plans a September/October 2017 special issue on 5G technology.

The past two decades have witnessed phenomenal progress in wireless access to the Internet and telecommunications services.

Cellular network operators in several countries have deployed fourth-generation long-term evolution (4G LTE) radio technologies to keep up with the demand. Fifth generation—5G—cellular network technologies are slated for 2020.

This special issue will provide a comprehensive update of 5G radio and network technologies. The guest editors seek papers from industry and academia. Of particular interest are review and tutorial articles that summarize the subject area, and provide guidance to researchers and planners so that they can anticipate 5G's impact on networks and envision new mobile services.

Brief article descriptions are due **12 December 2016** (mail to ic5-2017@computer.org). Articles are due **12 January 2017**. Visit www.computer.org/web/computingnow/iccfp5 to view the complete call for papers.

IEEE Internet Computing plans a November/December 2017 special issue on agents for social media.

While most current social media is intended for sharing content, future social-media applications could offer models for other forms of interactions, including those involving business and government. Such models could make use of agents that form teams, partnerships, and communities; foster communication; and enable participants to collaborate to formulate policies and reach decisions. Realizing these capabilities will require researchers to address underlying computational challenges.

This special issue will address the questions, challenges, and opportunities that arise at the intersection of agents and social media. These contributions can include theoretical and applied research related to the modeling, design, and development of agents and multiagent systems for social media.

Brief article descriptions are due **13 January 2017**. Articles are due **13 February 2017**. Visit www.computer.org/web/computingnow/iccfp6 to view the complete call for papers.

SEEKING PAPERS ON COMPUTATIONAL SOCIAL SYSTEMS

IEEE Transactions on Computational Social Systems welcomes submissions on topics such as modeling, simulation, analysis, and the understanding of social systems from the quantitative and/or computational perspective. Learn more at www.ieeesmc.org/publications/transactions-on-computational-social-systems/call-for-papers-and-special-issues.

SEEKING PAPERS ON SOFTWARE ENGINEERING

Computing in Science & Engineering seeks submissions on scientific-software engineering. The magazine seeks to provide a venue for the publication of significant work in the field, recognizing that the development of scientific software differs significantly from that of other software. Learn more at www.computer.org/cms/computer.org/computingnow/docs/2016-software-engineering-track.pdf.

SECTION TITLE

EVENTS IN 2016 AND 2017

DECEMBER 2016

2–4 ..... T4E 2016
8–10 ..... CIT 2016
12–14 ..... WF-IOT 2016
19–22 ..... HiPC 2016

JANUARY 2017

22–25 ..... PRDC 2017
30 January–1 February ..... ICSC 2017

FEBRUARY 2017

4–8 ..... HPCA 2017
13–16 ..... BigComp 2017
20–24 ..... SANER 2017

MARCH 2017

13–18 ..... ICST 2017
18–22 ..... VR 2017
24–31 ..... WACV 2017
27–29 ..... AINA 2017

IEEE Security & Privacy plans a November/December 2017 special issue on digital forensics.

Modern societies are becoming increasingly dependent on open networks. However, these networks can attract cybercriminals.

In these cases, cybercrime detection and evidence collection can be difficult because clues are often buried in large data volumes. In addition, investigations of cybercriminal activities often span international borders and are subject to multiple jurisdictions and legal systems. These challenges require the use of digital forensics, which is thus becoming increasingly important.

The guest editors aim to collect information on the most relevant digital-forensics research.

Articles are due **1 March 2017**. Visit www.computer.org/web/computingnow/spcfp6 to view the complete call for papers.

IEEE Transactions on Emerging Topics in Computing (TETC) plans the following special section for its March 2018 issue, with articles due **1 March 2017**:

Reliability-aware design and analysis methods for digital systems, from gate to system level: Visit www.computer.org/cms/computer.org/transactions/cfps/cfp_tetcsi_rdamds.pdf to view the complete call for papers.

IEEE Transactions on Emerging Topics in Computing (TETC) plans the following special section for its June 2018 issue, with articles due **1 June 2017**:

Cybersecurity threats and defense advances: Visit www.computer.org/cms/computer.org/transactions/cfps/cfp_tetcsi_cstda.pdf to view the complete call for papers.

DECEMBER 2016

2–4 December: T4E 2016, 8th IEEE Int'l Conf. Tech. for Education, Mumbai, India; www.ask4research.info/t4e/2016

8–10 December: CIT 2016, 16th IEEE Int'l Conf. Computer and Info. Tech., Nadi, Fiji; nsclab.org/cit2016

12–14 December: WF-IOT 2016, IEEE 3rd World Forum on Internet of Things, Reston, Virginia; wfiot2016.ieee-wf-iot.org

19–22 December: HiPC 2016, 23rd IEEE Int'l Conf. High Performance Computing, Data, and Analytics, Hyderabad, India; www.hipc.org/hipc2016/index.php

JANUARY 2017

22–25 January: PRDC 2017, 22nd IEEE Pacific Rim Int'l Symp. Dependable Computing, Christchurch, New Zealand; prdc.dependability.org/PRDC2017

30 January–1 February: ICSC 2017, 11th IEEE Int'l Conf. Semantic Computing, San Diego; icsc.eecs.uci.edu/2017/index.html

FEBRUARY 2017

4–8 February: HPCA 2017, 23rd IEEE Int'l Symp. High Performance Computer Architecture, Austin, Texas; hpca2017.org

13–16 February: BigComp 2017, 4th Int'l Conf. Big Data and Smart Computing, Jeju Island, Korea; conf2017.bigcomputing.org

20–24 February: SANER 2017, 24th IEEE Int'l Conf. Software Analysis, Evolution, and Reengineering, Klagenfurt, Austria; saner.aau.at

MARCH 2017

13–18 March: ICST 2017, 10th IEEE Int'l Conf. Software Testing, Verification, and Validation, Tokyo; aster.or.jp/conference/icst2017/index.html

18–22 March: VR 2017, IEEE Virtual Reality 2017, Los Angeles; www.ieeevr.org/2017

24–31 March: WACV 2017, IEEE Winter Conf. Applications of Computer Vision, Santa Rosa, California; pamitc.org/wacv2017

27–29 March: AINA 2017, 31st IEEE/ACS Int'l Conf. Advanced Information Networking and Applications, Taipei; voyager.ce.fit.ac.jp/conf/aina/2017/index.html

VR 2017

IEEE Virtual Reality 2017 (VR 2017) is sponsored by the IEEE Computer Society.

VR 2017 is geared to a broad audience, including academics, researchers, industry employees, and VR enthusiasts.

The conference will focus on topics such as applications, input devices and graphics techniques for virtual, augmented, and mixed reality; advanced display technology; immersive projection technology; modeling and simulation; VR systems and toolkits; locomotion and navigation in virtual environments; teleoperation; and telepresence.

VR 2017 will take place 18–22 March in Los Angeles. Visit www.ieeevr.org/2017 for complete conference information.

THE ERRANT HASHTAG

EDITOR DAVID ALAN GRIER
George Washington University; grier@gwu.edu

AUDIO

"I'm Not a Computer Scientist, but …"

David Alan Grier, George Washington University

Computing is opaque to some scientists and science is opaque to some software engineers.

Graduate students can form communities in an instant. When they gather as strangers—at a conference, class, or party—they often demonstrate an ease in bonding with and trusting others. After exchanging bits of information, they quickly find that they have a common background and experience. "Your ways are my ways," they tell each other. "Your concerns are my concerns."

I saw such a community flourish at a recent summer school program as the participants discussed the impact of computer simulation on scientific research. For a week, 15 students laughed and smiled and gossiped while they discussed the ideas of theoretical physicist Fritz Rohrlich, who argued 25 years ago that we were "at the threshold of an era of new scientific methodology." He claimed that simulation was a "qualitatively new and different methodology for the physical sciences" because it fell between the traditional forms of theoretical and empirical research.

The members of this new community agreed that Rohrlich was probably correct, but they all found a unique reason to support their position and presented their ideas. Each student rejected the traditional criticism of scientific simulation—the claim that it fails to provide a clear, mathematical description of natural phenomena. Instead, they seemed comfortable with the common justification that it provides numerical answers in situations where mathematics can't.

As simulation pioneers J.M. Hammersley and D.C. Hanscomb once famously claimed, simulation exploits the "strength of theoretical mathematics while avoiding its associated weakness." However, most of the students ignored the strengths of simulation and focused on its weakness, arguing that simulation was "epistemologically opaque"—a term coined by philosopher Paul Humphreys. To a philosopher, simulation is epistemologically opaque because it can't be fully verified by a scientist. A scientist can verify that a simulation produces the same data that's produced by nature, but can't verify that the simulation produces those numbers the same way that nature does.

Each speaker offered a different reason for epistemological opacity, including the complexity of software, the division of programming labor, the use of software libraries, and even the age of many simulation languages. The most novel idea was based on a phrase that one student, Katerena, heard repeatedly while studying programming practices on a large simulation project: "I'm not a computer scientist, but … ."

Katerena observed that the phrase was being used to justify bad programming practices and to hide the code's internal operation. She said that her subjects, the researchers on the project, had learned programming from each other and hence were mediocre at it; they would recommend unsuitable computer languages, accept technical debt, code inefficient algorithms, design faulty file structures, or cling to awkward factorizations. In almost every case, the bad habit was introduced with the phrase in question.

From Katerena's research, it was a little too easy to conclude that scientists need better software engineering training. The "I'm not a computer scientist, but …" phrase not only tells us that computing is opaque to some scientists, but also reminds us that science is opaque to some software engineers. Software engineering permeates scientific research (one speaker characterized software engineering as "contaminating science"). For most projects, scientists need software to collect, simulate, and analyze data. More science is done on the desktop than in the lab or field.

Software engineering has the goal of making things that work, and scientific research aims to understand what's true, even when it can't make things that work. To one, a simulation is a program that has to work; to the other, it's a way of communicating the ideas we believe to be true. They have enough in common to be part of the same community, but they'll never be the same thing.

See **www.computer.org/computer-multimedia** for multimedia content related to this article.

DAVID ALAN GRIER is an associate professor at George Washington University. Contact him at grier@gwu.edu.

0018-9162/16/$33.00 © 2016 IEEE

Computer | Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page

Qmags THE WORLD'S NEWSSTAND®

CALL FOR SPECIAL ISSUE PROPOSALS

Computer solicits special issue proposals from lead experts. Proposed themes/issues should address timely, emerging topics that will be of broad interest to *Computer*'s readership. Special issues are an important component of *Computer*, as they deliver essential research insights and well-developed perspectives on new and established technologies and computing strategies.

We encourage submissions of high-quality proposals for the 2018 editorial calendar. The *Computer* editorial board will review proposals in early 2017.

Proposal guidelines are available at: **www.computer.org/web/computer/siGuide**

Deadline for proposal submission: **15 December 2016**

IEEE IEEE computer society