Big Data

Also in this issue:

> On the Impact of Being Open

> Ideas Ahead of Their Time: Digital Time Stamping

NOVEMBER 2015 www.computer.org

COMPUTING edge

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

Printed with inks containing soy and/or vegetable oils

IEEE computer society

STAFF

BPA WORLDWIDE

Circulation: ComputingEdge 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.

Postmaster: Send undelivered copies and address changes to IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Application to Mail at Periodicals Postage Prices is pending 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.

IEEE Computer Society Magazine Editors in Chief

Computer
Sumi Helal, *University of Florida*

IEEE Software
Diomidis Spinellis, *Athens University of Economics and Business*

IEEE Internet Computing
M. Brian Blake, *University of Miami*

IT Professional
San Murugesan, *BRITE Professional Services*

IEEE Security & Privacy
Shari Lawrence Pfleeger, *Dartmouth College*

IEEE Micro
Lieven Eeckhout, *Ghent University*

IEEE Computer Graphics and Applications
L. Miguel Encarnação, *ACT, Inc.*

IEEE Pervasive Computing
Maria Ebling, *IBM T.J. Watson Research Center*

Computing in Science & Engineering
George K. Thiruvathukal, *Loyola University Chicago*

IEEE Intelligent Systems
Daniel Zeng, *University of Arizona*

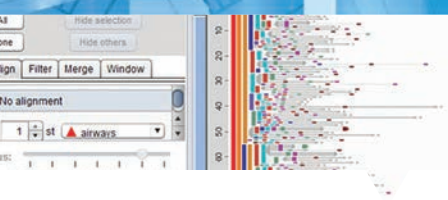
IEEE MultiMedia
Yong Rui, *Microsoft Research*

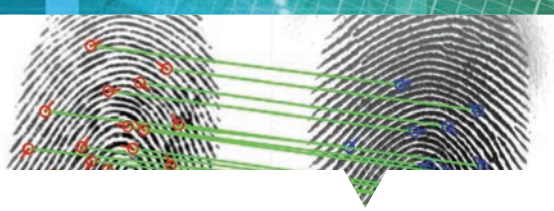
IEEE Annals of the History of Computing
Nathan Ensmenger, *Indiana University Bloomington*

IEEE Cloud Computing
Mazin Yousif, *T-Systems International*

NOVEMBER 2015 • VOLUME 1, NUMBER 11

COMPUTING edge





50

Ideas Ahead of Their Time: Digital Time Stamping

Departments

Spiking Neural Network Architecture

Paolo Montuschi, Polytechnic University of Turin

This installment highlighting the work published in IEEE Computer Society journals comes from IEEE Transactions on Computers.

ARM microprocessors are found in nearly every consumer device, from smartphones to gameboxes to e-readers and digital televisions. But did you know that, combined, these same ARM microprocessor cores can simulate the human brain?

The Spiking Neural Network Architecture (SpiNNaker), a massively parallel neurocomputer architecture, aims to use more than one million ARM microprocessor cores to model—in real biological time—nearly one billion spiking neurons.[1] The model comes from the University of Manchester's Advanced Processor Technologies Team under the guidance of Steve Furber, an IEEE Fellow and 2013 IEEE Computer Pioneer Award recipient (www.youtube.com/watch?v=x_H_6xG1TEs). Furber's vision is to apply computer engineering techniques to multidisciplinary research on information processing in the brain.

In 2013, *IEEE Transactions on Computers* (TC) published Furber and his colleagues' article on the SpiNNaker system's architecture and physical design.[1] Furber also prepared a video illustrating the paper's contributions (www.youtube.com/watch?v=EhPpxsK2Ia0). As editor in chief of *TC*, I invite you to read not only the original paper but also its 2015 follow-up.[2] In the most recent paper, Furber and his colleagues describe the innovative SpiNNaker software:

> *It possesses an architecture that is completely scalable to a limit of over a million cores, and the fundamental design principles disregard three of the central axioms of conventional machine design: the core-core message passing is non-deterministic …; there is no attempt to maintain state (memory) coherency across the system; and there is no attempt to synchronize timing over the system.*
>
> *Each of the million cores has … only a small quotient of physical resource. … The inter-core messages are small (≤72 bits) and the message passing itself is entirely hardware brokered, although the distributed routing system is controlled by specialized memory tables that are configured with software. The boundary between soft-, firm- and hardware is even more blurred than usual.*[2]

Each low-power ARM core has limited resources, so the SpiNNaker engine uses no more than 90 kilowatts of electrical power.

Furber and his team's research outlines a broader picture in which inexact computing—using less energy and power—could form the backbone of an emerging research and multidisciplinary application area. This topic is increasingly relevant in the context of sustainability and green computing. Their research also stimulates exploration of ways that the axioms of conventional machine design can be drastically changed to achieve different and very ambitious goals. Inspired by Furber and his colleagues, scientists might open themselves to new approaches in which creativity reshapes how computing systems are designed and implemented.

REFERENCES

1. S. Furber et al., "Overview of the SpiNNaker System Architecture," *IEEE Trans. Computers*, vol. 62, no. 12, 2013, pp. 2454–2467.
2. A.D. Brown et al., "SpiNNaker—Programming Model," *IEEE Trans. Computers*, vol. 64, no. 6, 2015, pp. 1769–1782.

PAOLO MONTUSCHI is a professor of computer engineering at the Polytechnic University of Turin. Contact him at pmo@computer.org and visit http://staff.polito.it/paolo.montuschi/news-from-EIC-TC.html.

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

Selected CS articles and columns are also available for free at **http://ComputingNow.computer.org**.

 Published by the IEEE Computer Society

Magazine Roundup

The IEEE Computer Society's lineup of 13 peer-reviewed technical magazines covers cutting-edge topics ranging from software design and computer graphics to Internet computing and security, from scientific applications and machine intelligence to cloud migration and microchip manufacturing. Here are highlights from recent issues.

Computer

By connecting the human body and the computer, we extend the boundaries of technology via physiological processes. *Computer*'s October 2015 special issue focuses on **physiological computing**.

IEEE Software

To survive and grow, software companies want to swiftly turn their ideas into products and are increasingly using **hackathons** to accomplish this. In a hackathon—the subject of "What Are Hackathons For?" in *IEEE Software*'s September/October 2015 issue—small groups work together to quickly produce software prototypes.

IEEE Internet Computing

Big, established players and small startups alike are racing to develop new devices, applications, and protocols for the wearable electronics market. *IEEE Internet Computing*'s September/October 2015 special issue focuses on the interdisciplinary design of efficient protocols and technologies to help implement the **small wearable Internet**.

Computing in Science & Engineering

A key step in supporting scientific progress is enabling **data-driven hypothesis management and predictive analytics** directly from simulation results. "Managing Scientific Hypotheses as Data with Support for Predictive Analytics," from *CiSE*'s September/October 2015 issue, introduces a new approach for doing this.

IEEE Security & Privacy

Organizations require effective **cybersecurity incident response teams** (CSIRTs) to protect their systems and data. However, researchers know relatively little about improving CSIRTs' performance because use of the teams is fairly new. The authors of "Improving Cybersecurity Incident Response Team

2376-113X/15/$31.00 © 2015 IEEE Published by the IEEE Computer Society November 2015

Effectiveness Using Teams-Based Research," from *IEEE S&P*'s July/August 2015 issue, identify steps to upgrade CSIRTs.

IEEE Cloud Computing

Cloud computing creates legal challenges related to, for example, data access, portability, and privacy; liability; and laws governing information use. *IEEE Cloud Computing*'s July/August 2015 special issue looks at **balancing privacy with legitimate surveillance and lawful data access in the cloud**.

IEEE Computer Graphics and Applications

IEEE CG&A's September/October 2015 special issue on **virtual reality (VR) software and technology** highlights recent advances in the field. The issue's five articles include state-of-the-art practical VR applications and suggest new research directions.

IEEE Intelligent Systems

IEEE Intelligent Systems' September/October 2015 special issue on **natural language processing** includes articles on the processing pipeline for extracting information from text; deep neural networks' use in machine translation; question-answering over knowledge bases, a critical task for search engines; and modeling machine translation.

IEEE MultiMedia

Ultra-high definition (UHD) video promises to significantly enhance the user experience with higher spatial resolutions, frame rates, and sample bit depths, as well as a wider color gamut. However, this requires increased bandwidth. "Manipulating Ultra-High Definition Video Traffic," from *IEEE MultiMedia*'s July–September 2015 issue, explores on-demand UHD video streaming using the latest video-compression and -delivery technologies.

IEEE Annals of the History of Computing

In the late 1960s, attorneys and programmers used the term "**embodying software**" to refer to a patent-drafting technique for software inventions. This strategy, sometimes successful, consisted of claiming a patent not for the application that was developed but for the type of computer for which the application served as the control system. "Embodied Software: Patents and the History of Software Development, 1946–1970," from *IEEE Annals*' July–September 2015 issue, argues that embodied software's history demonstrates that software patenting predated the software industry's birth, which is a different spin on the history of software patents.

IEEE Pervasive Computing

In "**Digitally Enhanced Learning?**" from *IEEE Pervasive Computing*'s July–September 2015 issue, the author explores whether pervasive technologies—such as smart boards, wireless student-response pads (called *clickers*), smart pens, tablets, and smartphones—are reaching their potential in the classroom.

IT Professional

Advances in mobile computing and communications, ambient intelligence, and ubiquitous sensors have driven **wearable computing**, the topic of *IT Pro*'s September/October 2015 special issue. The technology facilitates a new form of human-computer interaction via small, on-body devices that are always connected, hands-free, and less distracting than handhelds. This leads to a new form of synergy between humans and computers, offering consistency and multitasking capabilities.

IEEE Micro

In *IEEE Micro*'s July/August 2015 issue, the article "Achieving Exascale Capabilities through Heterogeneous Computing" provides an overview of Advanced Micro Devices' (AMD's) vision for exascale computing, focusing on heterogeneity's central role. **Exascale computing** requires chips to offer high performance while staying within power budgets. AMD sees high-volume GPU technology, working with CPUs, as the best way to achieve energy-efficient parallel computing.

Computing Now

The Computing Now website (http://computingnow.computer.org) features **up-to-the-minute computing news** and blogs, along with articles ranging from peer-reviewed research to opinion pieces by industry leaders.

Big Data: Opportunities and Concerns

Processors have gotten faster. Storage capacities have increased. Smartphones have proliferated. Social media has exploded. And an increasing number of devices connect to the Internet. These and other factors have led to a massive increase in the amount of information being generated for potential use by researchers, companies, and others.

Big data offers many potential benefits. For example, it enables scientists to make valuable discoveries and helps retailers figure out what their customers want. On the other hand, organizations struggle to access, store, process, and analyze vast quantities of data.

This issue of *ComputingEdge* looks at big data's significant opportunities and challenges.

IEEE Internet Computing's "Is Big Data a Transient Problem?" explores whether computing capabilities might "catch up" to the exploding amount of available information, making big data only a short-term challenge.

The authors of "Big Data: Next-Generation Machines for Big Science," from *Computing in Science & Engineering*, write about the next-generation supercomputers that will be necessary to solve the scientific grand challenges that the US Department of Energy's Office of Science has identified.

"Sharpening Analytic Focus to Cope with Big Data Volume and Variety," from *IEEE Computer Graphics and Applications*, looks at ways to make analytics more effective when working with huge amounts of data.

IEEE MultiMedia's "Multimedia Big Data" reports on important presentations made at the First IEEE International Conference on Multimedia Big Data, held in April 2015 in Beijing.

The authors of "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds," in *IEEE Cloud Computing*, highlight some critical big-data-related security and privacy issues.

ComputingEdge articles on other subjects include the following:

- In *Computer*'s "Anil Jain: 25 Years of Biometric Recognition," an expert in the field discusses biometric technology's evolution.
- In *IEEE Software*'s "On the Impact of Being Open," the authors analyze the similarities and differences among the open source movements they've participated in and present their expectations for the future.
- *IEEE Security & Privacy*'s "Ideas Ahead of Their Time: Digital Time Stamping" suggests that some old security concepts that were never widely implemented might now be ready for prime time.

2376-113X/15/$31.00 © 2015 IEEE Published by the IEEE Computer Society November 2015

Editor:
Theresa-Marie Rhyne

Sharpening Analytic Focus to Cope with Big Data Volume and Variety

Ben Shneiderman and Catherine Plaisant
University of Maryland

The growing volumes of data available from sensors, social media sources, Web logs, and medical histories present remarkable opportunities for researchers and policy analysts.[1] Although big data resources can provide valuable insights to help us understand complex systems and lead to better decisions for business, national security, cybersecurity, and healthcare, there are many challenges to dealing with the volume and variety of data.[2] Data cleaning and data wrangling[3] have received some attention with the development of application tools (such as OpenRefine, http://openrefine.org), but data focusing to sharpen the analytic focus remains a challenge. An admirable example of preprocessing strategies to clean and prepare data is the five- or six-step process used in many NASA remote sensing projects, such as the OMNI 2 dataset,[4] the Earth Observing System Data and Information System (EOSDIS) data tools (https://earthdata.nasa.gov/data/data-tools), and the Swift processing pipeline (http://swift.gsfc.nasa.gov/quicklook/swift_process_overview.html).

This analytic-focusing problem is also being addressed in familiar relational databases, large network (graph) databases, and elsewhere, but here we emphasize temporal event sequences in which streams of point and interval events are organized into records. For example, patient histories might include point events, such as diagnoses, tests, and surgeries, as well as interval events, such as medication episodes, dieting plans, or hospitalizations. Each patient history, with a large variable number of events, is considered just one record. In addition to events, patients have attributes, such as gender or age, and events may also have attributes, such as which physician ordered a medication or which hospital provided care.

Researchers and corporate innovators are building a growing number of visual analytics and statistical software tools to deal with temporal event sequences. These tools often have trouble dealing with two problems:

- *Volume of records*. The number of records may grow to billions, making it difficult to load or apply operations to the data.
- *Variety of patterns*. Within each record, there may also be thousands or millions of events, coded using thousands or tens of thousands of event categories. For example, medical histories can include diagnoses that come from the more than 90,000 ICD-9 (International Classification of Diseases, 9th revision) codes or medications that come from the over 31,000 medications listed in RxNorm. Most records are unique, and the variety makes it difficult to see global patterns such as relationships, clusters, or gaps as well as to identify errors or anomalies.

This problem exists in other domains as well. For example, for social media log analysis, a Twitter user may post thousands of times and retweet, reply, mention, or take other actions. Web log analysis for shopping sites may include thousands of website visits, often recorded by the types of products viewed and the purchases made.

To address these challenges, this article provides a taxonomy of analytic-focusing strategies for temporal event sequences. The 10 analytic-focusing strategies included here can help analysts cope with big data volume and variety. The strategies include extracting relevant records, event types, and key events; folding data to make cyclic patterns such as days or weeks clear; and pattern simplification strategies to simplify complex sequences of events.

Data Cleaning and Analysis

Although an overview of the variety in data can

 Published by the IEEE Computer Society

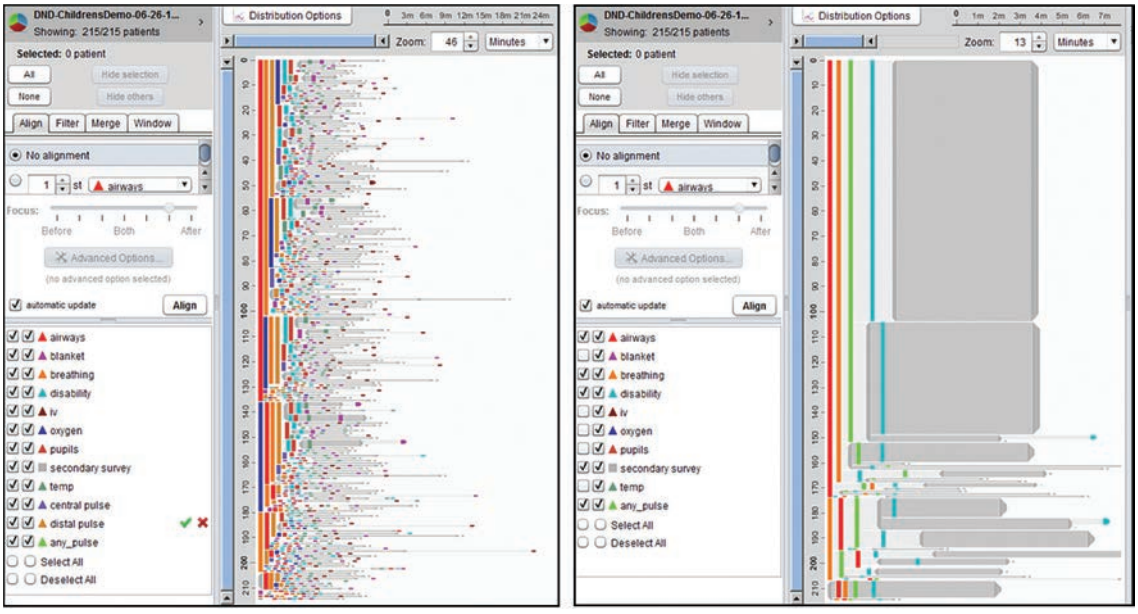


Figure 1. Analytic focusing: (a) before and (b) after. The goal was to find the patients who had received the correct treatment sequence (about half the patients, shown on the top of the sequence overview) and the ways that the required protocol was violated.

be useful, analysts also need ways to sharpen the analytic focus to enable useful visualizations of global patterns and anomalies of interest. Just as camera images need to focus on objects or faces of interest and telescopes are best when selected for spectral ranges (visual, ultraviolet, radio, x-ray, and so on), analytics tools will be most effective if users can focus their attention. The idea of an analytics pipeline or workflow is well-established in mature topics such as pharmaceutical drug discovery or NASA's remote sensing data analysis.

Cleaning the data is critical[3,5] and necessary before the sharpening of analytic focus can occur. Often the first stages are to detect omissions or duplications in the data; then data cleaning begins to deal with dubious values from incorrect processing, human data-entry error, failing sensors, and so on. This is often far more complex than analysts expect as they discover the vagaries of domain-specific data capture. One favorite example is the hospital that was trying to study the average length of emergency room stays, but it did not realize their data was faulty. The most extreme case was the patient who was admitted 14 times but discharged only twice. In discussions with hospital managers, it became clear that admission is tied to billable events so it is usually accurately collected, but discharges are occasionally skipped. A detailed set of technical reports from the Health and Social Care Information Center of the UK National Health Service (see www.hscic.gov.uk/article/1825/The-processing-cycle-and-HES-data-quality) describes their data cleaning process.

Once data analysts begin their work, there are many paths and questions, mostly driven by the analytic goal. In some mature application domains the accumulated experience of analysts has led to well-established strategies and workflows for analytic focusing, but in new domains, such as temporal event sequence analysis, we see a need for novel approaches. Our work in developing systematic, yet flexible strategies showed ways to structure analytic workflows,[6] especially for network data. T.D. Wang and his colleagues provided a starting point for developing systematic, yet flexible workflows for temporal event sequences,[7] which was pursued by Megan Monroe[8] and is refined here. Our application of these ideas in EventFlow (www.cs.umd.edu/hcil/eventflow) gave us dozens of case studies from real users who applied our tools for their own investigations.

Figure 1 provides one example of the dramatic difference between the initial "confetti" view to a focused version that fits the analytic goal. This example includes just a small dataset of 215 patient records, and the goal was to find the patients who had received the correct treatment sequence (about half the patients, shown on the top of the sequence overview) and the many ways that the required protocol was violated. The hospital managers are currently investigating the impact of these variations and training staff as needed.

Taxonomy of Analytic Focusing Strategies

This taxonomy of analytic focusing strategies for temporal event sequences is based on our experience

working with dozens of case studies using EventFlow[5] and other visual analytics tools for sequence analysis.[9–12] Most of the strategies can simultaneously lead to a reduction in volume and variety.

Extraction Strategies

The extraction strategies take only the relevant records, event types, or events from a large dataset, thereby making detection of meaningful patterns easier.

1. Goal-driven record extraction. For most projects dealing with large datasets, we find that the question at hand concerns only a fraction of the available records. For example, the US Army Pharmacovigilance Center has 15 million patient histories, but when the analytic goal was to determine how asthma medications had been prescribed in the past six years, they extracted a set of 182,000 records. At Washington Hospital Center, only 3,600 out of more than 1 million records had been given the treatment that was being studied. Traditional query and extraction tools (such as https://www.i2b2.org or http://btris.nih.gov) are needed before the focused analysis can begin.

Just as camera images need to focus on objects or faces of interest, analytics tools will be most effective if users can focus their attention.

2. Goal-driven event extraction. Some studies require a large portion of the records in a database, but only a small fraction of the events in those records. For example, in the US Army asthma study, only asthma-related medications and events were extracted.[13] In another study related to prostate cancer radiation treatment, the analytic goal was finding what durations and intensity of radiation produce the fewest bone fractures, yet still curtail the cancer. In that case, analysts removed events such as eye or dental exams or even the procedure's details to trim the dataset and greatly focus the analysis. We found that successful analysts start with just a few event types and then progressively add more as needed to refine the analysis.

3. Temporal windowing. In many cases, only a small window of time matters. The selection might be arbitrary (such as only recent data) or goal driven (for example, only the week before a purchase or a month before and two months after a surgery). Extraction tools are needed that limit the range of events, either using absolute dates (such as the holiday shopping period) or relative dates after alignment by a selected event (such as a surgery).

4. Random sampling of records. If the previous strategies fail to sufficiently reduce the data volume, random sampling of records (such as extracting every 10th record) may become a reasonable strategy. There is often some benefit in getting a rough indication of the prevalence of the patterns being sought.[14] On the other hand, random sampling of events within records does not seem useful.

Folding Strategy

The folding strategies replace a single long sequence with many shorter ones so that cyclic patterns, such as weekend days in a week, are easier to recognize.

5. Temporal folding. Some datasets have records that are long streams (for example, one person's entire Twitter history with thousands of events), which may be more successfully analyzed by folding (or splitting) each record into yearly, monthly, or weekly units. An Interpersonal Violence (IPV) study had detailed 90-day records of drugs and alcohol use as well as incidents of arguments, physical violence, sexual abuse, and so forth. Patterns were difficult to extract until the records were broken into weekly records, thereby revealing weekend conflicts and drug use.

Temporal folding may not address data volume issues as the number of records increases even though the number of events remains constant. However, once the temporal folding is done, the variety of patterns may be reduced and other strategies to reduce dataset size may become applicable.

Pattern Simplification Strategies

The central problem we faced with many datasets was that the variety of event types and the volume of events made it difficult to see meaningful patterns. Sometimes similar event types could be relabeled as a single event type, such as changing book purchase events that were labeled by book title to be a generic event of a book purchase. The goal was to have fewer diverse events, but in some cases we also had fewer total events.

6. Grouping event categories (aggregation). With the explosion of the number of event categories, aggregation becomes necessary. For example, there are

more than 400 types of lung cancer and over 200 types of bone cancer, making it impossible to see global patterns. Replacing all lung or bone cancers with two single event categories will reduce the variety of patterns. The number of events remains the same, but the simplification sharpens the analytic focus and allows analysts to determine that lung cancers often spread to bones, however bone cancers rarely spread to the lungs. Dynamic aggregation (undoing the grouping) is needed but certain combinations of data transformations may restrict the guarantee of reversibility.

7. Selecting sentinel events in a stream. In social media log analysis, such as the use of Twitter, sharpening the analytic focus might require thoughtful selection of sentinel events. A typical strategy is to keep only the dates of the first, 10th, 100th, and 1,000th tweets in each person's record. This dramatically reduces the clutter of tweets and makes the relationship to retweets, mentions, replies, and so on clearer. For the same reason, analysts might choose to retain only the dates of the first, 10th, 100th, and 1,000th followers. Similarly, in the medical domain, finding all new prescriptions following a gap of more than three months (and then removing others) may be what is needed to focus an analysis.

8. Converting multiple point events into a single interval event. When dealing with patient histories, a major simplification is to convert multiple point events, such as 25 normal blood pressure readings over 12 months, into a simpler, more meaningful single interval event that shows normal blood pressure for that 12-month period.

9. Converting multiple interval events in a single longer interval. The US Army asthma project raised this issue for patients who received repeated prescriptions for the same medication. Patients often refill asthma prescriptions early, which appears as an overlap of two intervals, or delay their refills, which appears as a gap between the intervals. Analysts simplified the patterns by merging intervals with overlaps of less than 15 days or gaps of less than 10 days resulting in long intervals indicating the drug "episode."

10. Identifying hidden complex events. In many application domains some high-level event, such as a heart attack or surgery, may consist of 20 to 100 events that all occurred within a given time period (blood tests, imaging, and so on) These component events may not be relevant to the analysis, so they can all be identified and replaced by a single event.

These pattern simplification strategies are either domain specific (event category aggregation can use domain ontologies) or goal driven (shaped by the specific question the analyst is trying to answer). Domain experts who visually inspect sample data will be able to tune parameters and see the effect of the simplifications.

Multistep Strategies

Extraction strategies can be performed in traditional database systems. After the extracted results are converted, they can be loaded in interactive visual analytics tools. Our experience indicates that extraction strategies continue to be useful inside the visual analytics tools as users eliminate errors and outliers or select groups of records of interest using interactive filtering widgets.

In time, the accumulated experience of analysts with particular application domains will lead to recommended or partially automated workflows.

Because visual inspection of the transformation results greatly improves the analytic focus, it needs to be conducted within the visual analytics tool, possibly on a sample of records at first. For example, the initial analysis of a data sample is usually extremely useful to identify data-cleaning strategies (for example, removing all records deactivated in the last month), new extraction strategies (such as identifying interesting event categories), or devising meaningful pattern simplification strategies (possibly selecting the first and fifth abnormal test).

In our experience pattern simplification strategies usually reduce data volume. We have implemented most of the strategies described here in EventFlow, but if the goal is to reduce the volume of records so that all needed records can be loaded in the visual analytics tool, those transformations will have to be executed in the source database or in a separate analytic-focusing tool. Progressive visual analytics techniques might also offer effective solutions.[15]

There are certainly other temporal analytics focusing strategies beyond extraction, folding, and pattern simplification. Some will be generalizable, while others will remain specific to event

sequence analysis. Some of those strategies might become an integral part of an established domain-specific workflow (such as for the analysis of drug usage patterns) or be task specific (such as exploratory analysis of the patterns leading to a fixed outcome). Here we emphasize user-driven strategies, but sequence mining methods might provide complementary strategies.[11] Understanding which strategies are relevant in each situation requires experience with the data and problem at hand. We believe that these analytic focusing strategies can be combined to sharpen analytic processes and enable users to deal with ever larger datasets. In time, the accumulated experience of analysts with particular application domains will lead to recommended or partially automated workflows.

Acknowledgments

We thank Theresa-Marie Rhyne for working closely with us, encouraging our work, and managing the review process. We also thank the anonymous reviewers and David Gotz, Gigi Lipori, Adam Perer, and Krist Wongsuphaswat for constructive and supportive comments on draft versions. This work is supported in part by Oracle. We gratefully acknowledge funding provided by the University of Maryland's Mpowering the State through the UMIACS (University of Maryland Institute for Advanced Computer Studies) Center for Health-Related Informatics and Bioimaging.

References

1. P.C. Wong and J. Thomas, "Visual Analytics," *IEEE Computer Graphics and Applications*, vol. 24, no. 5, 2004, pp. 20–21.
2. P.C. Wong et al., "The Top 10 Challenges in Extreme-Scale Visual Analytics," *IEEE Computer Graphics and Applications*, vol. 32, no. 4, 2012, pp. 63–67.
3. S. Kandel et al., "Research Directions in Data Wrangling: Visualizations and Transformations for Usable and Credible Data," *Information Visualization*, vol. 10, no. 4, 2011, pp. 271–288.
4. J. King and N. Papitashvili, "OMNI 2 Preparation," tech. report, NASA, Feb. 2013; http://omniweb.gsfc.nasa.gov/html/omni2_doc_old.html.
5. T. Gschwandtner et al., "A Taxonomy of Dirty Time-Oriented Data," *Multidisciplinary Research and Practice for Information Systems*, LNCS 7465, Springer, 2012, pp. 58–72.
6. A. Perer and B. Shneiderman, "Systematic Yet Flexible Discovery: Guiding Domain Experts during Exploratory Data Analysis," *Proc. ACM Conf. Intelligent User Interfaces* (IUI), 2008, pp. 109–118.
7. T.D. Wang et al., "Extracting Insights from Electronic Health Records: Case Studies, A Visual Analytics Process Model, and Design Recommendations," *J. Medical Systems*, vol. 35, no. 5, 2011, pp. 1135–1152.
8. M. Monroe, "Interactive Event Sequence Query and Transformation," doctoral dissertation, Dept. of Computer Science, Univ. of Maryland, June 2014; www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2014-17.
9. A. Rind et al., "Interactive Information Visualization for Exploring and Querying Electronic Health Records: A Systematic Review," *Foundations and Trends in Human-Computer Interaction*, vol. 5, no. 3, 2013, pp. 207–298.
10. D. Gotz and H. Stavropoulos, "DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data," *IEEE Trans. Visualization and Computer Graphics*, vol. 20, no. 12, 2014, pp. 1783–1792.
11. A. Perer and F. Wang, "Frequence: Interactive Mining and Visualization of Temporal Frequent Event Sequences," *Proc. ACM 19th Int'l Conf. Intelligent User Interfaces* (IUI), 2014, pp. 153–162.
12. K. Wongsuphasawat and J. Lin, "Using Visualizations to Monitor Changes and Harvest Insights from a Global-Scale Logging Infrastructure at Twitter," *Proc. IEEE Visual Analytics Science and Technology Conf.* (VAST), 2014, pp. 113–122.
13. C. Plaisant et al., "Interactive Visualization," *Big Data and Health Analytics*, K. Marconi and H. Lehman, eds., CRC Press, 2014, pp. 243–262.
14. D. Fisher et al., "Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster," *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems* (CHI), 2012, pp. 1673–1682.
15. C. Stolper, P. Perer, and D. Gotz, "Progressive Visual Analytics," *IEEE Trans. Visualization and Computer Graphics*, vol. 20, no. 12, 2014, pp. 1653–1662.

Ben Shneiderman *is a professor of computer science in the Institute for Advanced Computer Studies and the founding director of the Human-Computer Interaction Lab at the University of Maryland. Contact him at ben@cs.umd.edu.*

Catherine Plaisant *is a senior research scientist in the Institute for Advanced Studies and the associate director of research in the Human-Computer Interaction Lab at the University of Maryland. Contact her at plaisant@cs.umd.edu.*

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.

This article originally appeared in
IEEE Computer Graphics and Applications,
vol. 35, no. 3, 2015.

COMPSAC 2016

ATLANTA, GEORGIA, USA
JUNE 10-14

CONNECTED WORLD: NEW CHALLENGES FOR DATA, SYSTEMS & APPLICATIONS

COMPSAC is the IEEE Computer Society Signature Conference on Computers, Software, and Applications. It is a major international forum for academia, industry, and government to discuss research results and advancements, emerging problems, and future trends in computer and software technologies and applications. The technical program includes keynote addresses, research papers, industrial case studies, plenary and specialized panels, fast abstracts, a doctoral symposium, poster sessions, and a number of workshops and tutorials on emerging and important topics. The theme of the 40th COMPSAC is "Connected World: New Challenges for Data, Systems, and Applications". Our world becomes more and more connected every day, with billions of computer applications, devices, and services interacting globally to make our lives safer, convenient, and more enjoyable. Computations, as well as sharable data and applications, are becoming available everywhere. This explosive growth brings us closer together and requires innovative technical solutions. With the rapidly shrinking gap between cyber and physical domains, we face many new challenges and new opportunities for computers, software, and applications. COMPSAC 2016 will provide a platform for in-depth discussion of such challenges both in traditional and in emerging fields such as smart and connected health, wearable computing, the Internet of Things, cyber-physical systems, social networking, and the smart planet. COMPSAC 2016 will be organized as a tightly integrated union of several symposia, each of which focuses on a particular technical segment. Big data is a particularly highlighted theme in our symposia. Furthermore, COMPSAC 2016 will also include a student research symposium/competition, fast abstracts sessions, workshops, poster sessions, panels, and keynotes to facilitate discourse on and deepen the understanding of these challenges by furthering the fundamental contributions needed for advances in computing systems.

Authors are invited to submit original, unpublished research work, as well as industrial practice reports. Simultaneous submission to other publication venues is not permitted. In accordance with IEEE policy, submitted manuscripts will be checked for plagiarism; instances of alleged misconduct will be handled according to the IEEE Publication Services and Product Board Operations Manual. Detailed instructions for electronic paper submission, panel, workshop, and tutorial proposals, fast abstracts, industry papers, poster papers, doctoral symposium, and the review process are available at www.compsac.org.

IMPORTANT DATES

11/30/2015: Main conference abstracts due
12/11/2015: Main conference submissions due
2/29/2016: Paper notifications
3/6/2016: Workshop papers due
3/28/2016: Workshop paper notifications
4/7/2016: Camera ready and registration due

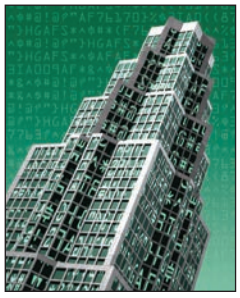
GENERAL INQUIRIES: For information, please contact Sorel Reisman, Chair of the Standing Committee at sreisman@calstate.edu, Sheikh Iqbal Ahamed, Chair of the Steering Committee at sheikh.ahamed@marquette.edu, or Ling Liu, General Chair at lingliu@cc.gatech.edu

2016 ORGANIZERS

Standing Committee Chair: Sorel Reisman, California State University, USA
Steering Committee Chair: Sheikh Iqbal Ahamed, Marquette University, USA
General Chairs: Ling Liu, Georgia Tech, USA and Dejan Milojicic, HP Labs, USA
Program Chairs: William Claycomb, Carnegie Mellon University, USA; Mihhail Matskin, KTH Royal Institute of Technology, Sweden; Hiroyuki Sato, University of Tokyo, Japan
Workshop Chairs: Motonori Nakamura, NII, Japan, Stelvio Cimato, University of Milan, Italy and Chung -Horng Lung, Carleton University, Canada

IEEE computer society

Editor: Jimmy Lin, jimmylin@uwaterloo.com



Is Big Data a Transient Problem?

Jimmy Lin

What's growing faster, Big Data or Moore's Law?

It's undeniable that the amount of data that organizations must store, process, organize, and analyze is growing rapidly. This requires increasingly larger clusters and datacenters, as well as increasingly complex software infrastructure to orchestrate the necessary computations. But is Big Data growing faster than Moore's Law is lowering the costs of computing capabilities to accomplish these tasks? For rhetorical convenience, I'm using Big Data to refer to all the things we want to *do* on massive collections of data, and Moore's Law to refer to exponential increases in computing capabilities for *doing it*. It's worth emphasizing that I don't literally mean the periodic doubling of transistors on a chip; I use Moore's Law as a convenient shorthand to refer to continued exponential advances in computing.

Logically, there are only three possibilities:

1. Big Data is growing faster than Moore's Law.
2. Big Data is growing at the same rate as Moore's Law.
3. Big Data is growing slower than Moore's Law.

The first two scenarios aren't particularly interesting: In the first case, what we can store will be bounded by Moore's Law and the rest of the data will need to be processed in real time (and then thrown away). The second case is essentially the status quo (hence, uninteresting). The third scenario, however, is intriguing: it suggests that computing capabilities are going to "catch up" to Big Data at some point. In other words, Big Data is a *transient problem*.

Defining the Question

What do I mean by a transient problem? Here's an analogy that might resonate with many: I remember when digital music first burst upon the scene about two decades ago. At the time, storing all those MP3s on my (gasp, desktop!) computer was a big deal. I distinctly remember my music collection consuming most of my hard drive, and having to sacrifice (delete) some files to make room for others. Over time, however, keeping MP3s around became less and less of a problem: storage technology improved many-fold, whereas the amount of music I could consume had a clear upper bound (24 hours in a day), and beyond a certain point, increased encoding quality didn't make a difference (at least to my ears). Today, all the music I could possibly want to listen to easily fits in my pocket (on my phone). In this sense, digital music storage was a transient problem that technology solved. Is Big Data the same way?

Of course, I'm assuming that Moore's Law will continue for some time, or more generally, exponential increases in computing technology will continue unabated. Obviously, there are physical limits,[1] but we're still pretty far from those. What do I mean by "for some time?" I have left this deliberately vague, because it depends on the particular prediction: when examining extrapolations of computing capabilities ("supply") with the demands of Big Data, it only matters if the pace of technological improvement will "hold up" until the anticipated crossover point. As to the broader question about continued technological progress (dire predictions about the end of Dennard Scaling notwithstanding), it depends on whether you're Cornucopian or Malthusian. This philosophical argument is beyond the scope of my piece, although I would note that Malthusians have essentially been wrong every time, because human civilization is still around and we don't (yet) live in a post-apocalyptic wasteland.

In this article, I only focus on *human-generated data* and leave aside data from scientific instruments (such as the Large Hadron Collider and the Square Kilometer Array), remote sensing (for

 Published by the IEEE Computer Society

Big Data Bites

"Big Data Bites" is a regular department in *IEEE Internet Computing* that aims to deliver thought-provoking and potentially controversial ideas about all aspects of Big Data. Interested in contributing? Drop me a line!

—*Jimmy Lin*

example, satellite imagery), surveillance (including traffic cameras), and related applications because the economics are quite different. Human-generated data benefit from what Jeff Bezos calls *the flywheel*: the virtuous cycle where insights from user-generated data are exploited to improve products and services, which lead to broader usage and even more user-generated data, thus closing the loop. Amazon, Google, Facebook, Uber, and countless other companies are all built on this powerful driver. In contrast, the economics of data not generated by humans are very different (and in some ways, less interesting).

Note that my definition of human-generated data is fairly expansive: it includes all forms of data generated by humans, including those in databases (for example, Amazon's vast product catalog is human generated in the sense that the products for sale are produced and consumed by humans), behavior logs, personal medical records, and even some aspects of the Internet of Things (the data generated by connected appliances are ultimately derived from human activity).

Data Bounds

The upper bound on human-generated data is the product of two terms: total human activity and the amount of data generated per unit time, or the *data density*. Let's examine the first term: by most accounts, the human population will stabilize sometime relatively soon. The "medium" scenario of Samir KC and Wolfgang Lutz[2] shows a continued world population increase, resulting in 9.17 billion in 2050, peaking around 9.4 billion in the 2070s, and declining somewhat to 9 billion by 2100. A competing analysis by Patrick Gerland and his colleagues[3] is somewhat more pessimistic, arguing that world population stabilization is unlikely this century. According to their models, there's an 80 percent probability that the world population will increase to between 9.6 and 12.3 billion in 2100.

Regardless, there appears to be a consensus that overall fertility rates are decreasing — the debate is mostly over how quickly — so the point remains that the human population on this planet won't grow indefinitely. This means, in turn, that there's a finite upper bound on human activity; after all, there are only 24 hours in a day. My analysis depends on the assumption that the human population won't grow without bound, which means that when we start colonizing the galaxy, all bets are off!

Let's look at the second term, the data density. As a specific case, consider the amount of human-generated textual data on the Web (for example, HTML pages): evidence suggests that it's growing slower than Moore's Law. Andrew Trotman and Jinglan Zhang present quite reasonable projections suggesting that by the middle of the next decade, "the storage capacity of a single hard drive will exceed the size of the index of the Web at that time," and that "within another decade it will be possible to store the entire searchable text on the same hard drive."[4] They explore the implications of this for the design of search engines, which is interesting but beyond the scope of the current discussion. You might quibble with the details of their projections, but the underlying point remains: when we talk about text, it's not growing as fast as we have room to store and index it.

The challenge in extending this argument to all human-generated media is that there's no upper bound to data density except for special cases like text, since we can arbitrarily improve sensor resolution (we'll eventually run into quantum limits, I suppose, but we're far from those). The argument with textual data "works" because text has a low and constant data density — which isn't the case with images and video, for example. What if we include all human-generated images and video on the Web? Imagine a dystopian future where all humanity does is create YouTube videos all day long: although the content's length in hours would be bounded, the data's size wouldn't, since the resolution could be made arbitrarily better. You might counter with the observation that, beyond a certain resolution, the human visual system can't tell, so the bandwidth of the human perceptual system might provide a natural upper bound. But what if I wanted to zoom in on a previously captured image or video? Then I'd want as high a resolution as physically possible. (Perhaps it wouldn't matter if nobody was watching!)

Why stop at video? What about a personal magnetic resonance image (MRI) scanner that continuously monitors and captures our physiological state? Or a swarm of nanobots living inside us that gathers detailed measurements of our molecular functionings? The limits imposed by physics aside, it's difficult to see an end to increased data density. However, it's important to remember that many of the technological trends that give rise to higher-resolution sensors are either directly or indirectly related to Moore's Law (for example, the increase in megapixels in digital cameras). Could it be that what Moore's Law giveth, data density taketh? In which case, the real question is: What's the growth in our ability to capture data at finer resolutions

compared to increases in computing power? The human population is simply the "constant" in the equation (albeit a fairly large constant), but if we're talking about exponential growth, the constant is basically irrelevant.

At this point, we might shift the argument to focus on useful data as opposed to all data. Suppose I went around capturing 4K-quality video of my every waking moment (technically possible today) — who cares and why would I possibly want that? Perhaps not now, but this is a failure of imagination: much of data science and Big Data analytics today is built on data we thought was useless two decades ago (in fact, some people call it *data exhaust*). One day, questioning the usefulness of certain data-collection activities could sound as quaint as asking: Why would we ever want to keep around click logs? What possible use could we have for them?

However, a more nuanced way to think about this issue is to compare the growth of Big Data with the extent to which we can exploit the data practically. Numerous studies have found roughly a log-linear relationship between the amount of data analyzed and its effectiveness in an application.[5,6] That is, achieving the next increment in effectiveness (for example, accuracy in a classification task) requires a multiple-fold increase in the amount of data. The relationship between Moore's Law and the slope of this effectiveness line is important. For example, if making an algorithm incrementally better requires four times more data, then one Moore's Law cycle (doubling capabilities) is insufficient to improve our algorithm. However, in rough terms, it does make the current problem half as difficult. In this case, we might say that practically exploitable Big Data is growing slower than Moore's Law. Yet, there's a hole in this argument, since it assumes that there won't be significant algorithmic improvements in the future. Perhaps some brilliant researcher will devise entirely new classes of algorithms that exploit Big Data much more efficiently?

Implications

So, is Big Data growing slower than Moore's Law? Hopefully, I've shown that it's plausible, at least in a suitably qualified or more restrictive form. Thus, it's worthwhile to consider some of the implications on future computing systems for Big Data.

The most important implication is what I call "the revenge of scale up." A nearly unquestioned assumption in the design of data processing systems today is the superiority of scaling "out" on a cluster of commodity machines as opposed to scaling "up" on a single "beefy" machine (more memory, more cores). Previously, scaling up simply wasn't an option because no single machine, no matter how powerful, was sufficient to handle the data-processing task at hand. Scaling out, however, incurs large costs in terms of synchronization, communication, and fault tolerance. If Big Data is indeed growing slower than Moore's Law, then we need to revisit the scale out versus scale up debate, because at some point, a single machine might become powerful enough to handle Big Data.

In fact, this debate is already under way. According to the analysis of Antony Rowstron and his colleagues,[7] at least two analytics production clusters (at Microsoft and Yahoo) have median job input sizes under 14 gigabytes and 90 percent of jobs on a Facebook cluster have input sizes under 100 gigabytes (in 2012). A study of enterprise Hadoop clusters at around the same time shows that the workloads are dominated by relatively small jobs.[8] So why are we still using distributed processing frameworks such as MapReduce or Spark when the data easily can be held in memory on a single machine? As my colleague Jens Dittrich puts it, why are we all obsessed with building a 1,000-horsepower supercar just to make a two-mile trip to the supermarket? Indeed, we're seeing a resurgence of interest in scale-up approaches, particularly from the academic community.[9-12]

So then, what's with all the petabytes that we're accumulating in our vast data warehouses? As it turns out, the process of extracting features (or "signals") from raw data is quite distinct from data mining and machine-learning algorithms for deriving insights from those features. In the first, we typically distill raw data into sparse feature vectors; during this process there's typically many orders of magnitude reduction in data size. The feature vectors then serve as input to machine-learning or data-mining algorithms. We still need large clusters for feature extraction, since the raw data are often immense and we need the aggregate throughput of disk spindles across many machines. However, the distilled feature vectors are quite manageable. For example, state-of-the-art large-scale machine learning today talks about billions of training examples with millions of parameters,[13] on the order of a trillion nonzero features in total (since the feature vectors are sparse). A trillion floating-point values occupy four terabytes of main memory: any day now, we'll purchase commodity machines with that much memory.

Similarly, consider a graph with a trillion edges: stored in the most naive manner as (source, destination) pairs, it would take eight terabytes. We'll purchase a commodity machine with that much memory soon enough (one Moore's Law cycle later, in fact). In general, graphs of human social relations are bounded by population size, which suggests that graph problems are progressively becoming easier with each generation of hardware. As a concrete example, Twitter's production graph recommendation service began with a scale-up approach, holding the entire follower graph in memory on a single machine (and exploiting replication for increased throughput).[14] Examples

of impressively fast machine learning on individual machines include Vowpal Wabbit (see https://github.com/JohnLangford/vowpal_wabbit), the lock-free "Hogwild" method for parallelizing parameter updates,[15] and recent work in matrix factorization.[16]

Decoupling feature extraction and machine learning suggests a heterogeneous architecture where we exploit clusters to munge the raw data, and then bring extracted features over to a single machine to perform the actual machine learning — in other words, scale out for data cleaning, feature extraction, and so on, and scale up for machine learning.

This architecture, however, raises two interesting questions: First, data scientists loathe multiple processing frameworks, which introduce impedance mismatches into their daily activities. Having one framework for feature extraction and another framework for machine learning introduces friction. Thus, it would be ideal to have a single framework that both scales up and out. Second, datacenter operations engineers prefer consolidated clusters with a homogeneous hardware configuration, from both the perspective of economics and management overhead. Modern cluster-management software such as Mesos[17] (and Google's equivalents) work best with homogeneous fleets of servers. This doesn't mean they're unable to handle workloads where certain jobs (that require lots of memory) can run only on certain machines (that have enough memory) — but it does add an element of complexity in scheduling and coordination.

Finally, if Big Data indeed is growing slower than Moore's Law, this means that the Big Data of today will fit in my pocket tomorrow — in the same way that my music collection, which occupied most of the disk on my desktop machine about 15 years ago, fits in my pocket easily today. How would information seeking change if we could store a cache of the Web in a mobile device we carry around all the time? We're already proceeding down this path: Ask yourself, when was the last time you searched Google just to go to Wikipedia? Or when you used a search engine as a bookmark to return to a page you've visited before (what the information retrieval community calls "refinding"[18])? In both cases, perhaps a local cache of the Web might do the job just as well, and has the additional advantages of freeing us from flaky connectivity and network latencies.

Already today, so-called low-power "wimpy" devices (such as mobile phones and tablets) are far more prevalent than traditional servers and PCs. The technology research firm Gartner forecasts that worldwide shipments of PCs in 2015 will total around 320 million units, compared to 2.3 billion mobile phones and tablets (www.gartner.com/newsroom/id/2791017). Thus, it's worthwhile to explore how infrastructure designed for "brawny" servers in a traditional datacenter might run in wimpy environments, and the implications of many thousands of wimpy devices within a relatively small area (say, in a Manhattan city block or sporting venue). Interesting work along these lines include deploying full-text search engines[19] and transactional databases[20] on mobile phones, and Web archiving infrastructure on Raspberry Pis.[21] In addition to scaling out and up, it's worthwhile to think about scaling "down" Big Data technology.

What does the future hold for Big Data? It could be the same qualitatively, just bigger and better, or there might be fundamentally disruptive forces that completely reshape the computing landscape. Trying to predict the future, of course, is a perilous exercise. At best, this discussion provides some deep insight on future developments in Big Data. At worst, it makes for an interesting cocktail conversation. Either way, it's worth the rumination.

Acknowledgments

I'd like to thank Andrew Trotman for various engaging discussions; and Charlie Clarke, Craig Murray, and Arnab Nandi for comments on previous drafts of this piece.

References

1. S. Lloyd, "Ultimate Physical Limits to Computation," *Nature*, vol. 406, 2000, pp. 1047–1054.
2. S. KC and W. Lutz, "The Human Core of the Shared Socioeconomic Pathways: Population Scenarios by Age, Sex, and Level of Education for All Countries to 2100," *Global Environmental Change*, 2014; doi:10.1016/j.gloenvcha.2014.06.004.
3. P. Gerland et al., "World Population Stabilization Unlikely This Century," *Science*, vol. 346, no. 6206, 2014, pp. 234–237.
4. A. Trotman and J. Zhang, "Future Web Growth and Its Consequences for Web Search Architectures," 2013; arXiv:1307.1179v1.
5. M. Banko and E. Brill, "Scaling to Very Very Large Corpora for Natural Language Disambiguation," *Proc. 39th Ann. Meeting of the Assoc. for Computational Linguistics*, 2001, pp. 26–33.
6. T. Brants et al., "Large Language Models in Machine Translation," *Proc. 2007 Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 858–867.
7. A. Rowstron et al., "Nobody Ever Got Fired for Using Hadoop on a Cluster," *Proc. 1st International Workshop on Hot Topics in Cloud Data Processing*, 2012, article no. 2.
8. Y. Chen, S. Alspaugh, and R. Katz, "Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads," *Proc. 38th Int'l Conf. Very Large Data Bases*, 2012, pp. 1802–1813.
9. R. Appuswamy et al., "Scale-Up vs Scale-Out for Hadoop: Time to Rethink?" *Proc. 4th ACM Symp. Cloud Computing*, 2013.
10. J. Shun and G.E. Blelloch, "Ligra: A Lightweight Graph Processing Framework for Shared Memory," *Proc. 18th ACM SIGPLAN Symp. Principles and Practice of Parallel Programming*, 2013, pp. 135–146.
11. K.A. Kumar et al., "Optimization Techniques for "Scaling Down" Hadoop on Multi-Core, Shared-Memory Systems," *Proc. 17th Int'l Conf. Extending Database Technology*, 2014, pp. 13–24.

12. F. Chen et al., "Palette: Enabling Scalable Analytics for Big-Memory, Multicore Machines," *Proc. 2014 ACM SIGMOD Int'l Conf. Management of Data*, 2014, pp. 705–708.
13. A. Agarwal et al., "A Reliable Effective Terascale Linear Learning System," *J. Machine Learning Research*, vol. 15, Mar. 2014, pp. 1111–1133.
14. P. Gupta et al., "WTF: The Who to Follow Service at Twitter," *Proc. 22nd Int'l World Wide Web Conf.*, 2013, pp. 505–514.
15. F. Niu et al., "Hogwild!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent," *Advances in Neural Information Processing Systems 24*, 2011, pp. 693–701.
16. Z. Liu, Y.-X. Wang, and A.J. Smola, "Fast Differentially Private Matrix Factorization," 2015; arXiv:1505.01419v2.
17. B. Hindman et al., "Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center," *Proc. 8th USENIX Symp. Networked Systems Design and Implementation*, 2011.
18. S.K. Tyler and J. Teevan, "Large-Scale Query Log Analysis of Re-Finding," *Proc. 3rd ACM Int'l Conf. Web Search and Data Mining*, 2010, pp. 191–200.
19. A. Balasubramanian et al., "FindAll: A Local Search Engine for Mobile Phones," *Proc. 8th Int'l Conf. Emerging Networking Experiments and Technologies*, 2012, pp. 277–288.
20. T. Mühlbauer et al., "One DBMS for All: The Brawny Few and the Wimpy Crowd," *Proc. 2014 ACM SIGMOD Int'l Conf. Management of Data*, 2014, pp. 697–700.
21. J. Lin, "Scaling Down Distributed Infrastructure on Wimpy Machines for Personal Web Archiving," *Proc. 24th Int'l World Wide Web Conf. Companion*, 2015, pp. 1351–1355.

Jimmy Lin holds the David R. Cheriton Chair in the David R. Cheriton School of Computer Science at the University of Waterloo. His research lies at the intersection of information retrieval and natural language processing, with a particular focus on Big Data and large-scale distributed infrastructure for text processing. Lin has a PhD in electrical engineering and computer science from MIT. Contact him at jimmylin@uwaterloo.ca.

This article originally appeared in IEEE Internet Computing, *vol. 19, no. 5, 2015.*

Editors: James J. Hack, jhack@ornl.gov | Michael E. Papka, papka@anl.gov

Next-Generation Machines for Big Science

James J. Hack | Oak Ridge National Laboratory
Michael E. Papka | Argonne National Laboratory
and Northern Illinois University

Addressing the scientific grand challenges identified by the US Department of Energy (DOE) Office of Science programs alone demands a total leadership-class computing capability of 150 to 400 Pflops by the end of this decade. The successors to DOE's three most powerful leadership-class machines are set to arrive in 2017 and 2018—the products of the Collaboration Oak Ridge Argonne Livermore (CORAL) initiative, a national laboratory–industry design/build approach to engineering next-generation petascale computers for grand challenge science. These mission-critical machines will enable discoveries in key scientific fields such as energy, biotechnology, nanotechnology, materials science, and high-performance computing and will serve as a milestone on the path to deploying exascale computing capabilities.

Meeting a Need

Since 2012, three US DOE supercomputers—two that support open scientific research and one focused primarily on nuclear security—have been among the top five fastest machines in the world. Each has followed a familiar life cycle: a debut on the international supercomputing scene followed by years of robust use before eventually meeting their limits in the face of increasing demand.

Such is the fate of machines whose capabilities represent the "highest domestic priority" with respect to DOE mission needs.

The US faces serious economic, environmental, and national security challenges based on its dependence on fossil fuels and the need to be energy independent. These machines, hosted in the DOE Office of Science's Leadership Computing Facility centers at Argonne National Laboratory and Oak

2376-113X/15/$31.00 © 2015 IEEE Published by the IEEE Computer Society

Ridge National Laboratory and at the DOE's National Nuclear Security Administration's laboratory, Lawrence Livermore National Laboratory, are the computational engines that are helping researchers across a broad range of disciplines look for new alternative energy sources and develop new energy technologies.

Teams of technicians, performance engineers, domain scientists, and computational scientists are needed to prepare these behemoths to effectively and efficiently execute the massive computations driving high-end simulation science and modeling. From day one, the petascale capabilities of these machines are maximized: jobs run round the clock throughout the year.

Research teams are given time (typically tens of millions of hours, or core-hours) to investigate many of the fundamental questions in science today—to capture new opportunities in combustion science, battery technology, materials science, and fusion energy. What these teams can look for and learn is limited only by the machine's physical capabilities, power constraints, and contention for a shared resource.

Maintaining US leadership in computational science and engineering requires the best resources—including a succession of computers with world-class speed and capability. All the major priorities highlighted in the DOE's 2014 strategic plan—preparing for climate change, securing our leadership in clean energy technologies, maintaining science and engineering as a cornerstone of our economic prosperity, and enhancing our nuclear security—depend critically on advanced computing and have sharply focused the mission need for leadership-computing facilities.

Increasing the supercomputing capability to hundreds of petaflops, a process now underway, is vital to the US achieving its goals—and to achieving exascale computing down the road. These pre-exascale systems will be many times more powerful than their predecessors, they'll be diverse (providing two distinct solutions), and they'll be deployed and operational within a 2017–2018 timeframe.

These new supercomputers are the outcome of the CORAL partnership, formed in early 2014 to simplify and streamline the processes to procure for each lab a radically more powerful and architecturally diverse pre-exascale system, with the long term goal of delivering exascale systems that will be 20 to 40 times faster than today's leading supercomputers.

CORAL aggregated the expertise of three national labs to define the ideal system requirements for their users—the scientific community—and then provided multiple large-machine awards as an incentive to industry to align their business plans to design and develop systems to meet these needs.

CORAL issued a single request for proposals for multiple laboratory acquisitions where the winning bids would both deliver and support the new architecture. In this way, vendors would have a large stake in the success of these pre-exascale machines, and the CORAL partners could encourage and fund applied research to find a specific solution to meet a specific need that could be hardened as a deliverable.

The CORAL partnership selected two multivendor bids: IBM/Nvidia/Mellanox and Intel/Cray. Oak Ridge and Lawrence Livermore announced their private-sector partners last fall. Computer manufacturer IBM, along with OpenPOWER Foundation partners Nvidia and Mellanox, would build two separate machines using a heterogeneous architecture, or one that couples central processing units (CPUs) with general-purpose graphics processing units (GPUs).

The Oak Ridge Leadership Computing Facility's (OLCF's) Summit, the IBM machine scheduled to arrive in 2017, will deliver more than five times the computational performance of Titan's 18,688 nodes, while using roughly 3,400 nodes. Summit's architecture will consist of nodes with multiple IBM POWER9 CPUs and Nvidia Volta GPU accelerators, using a coherent memory space that includes high-bandwidth memory on the GPUs and a high-speed NVLink interconnect between the POWER9 CPU and Volta GPUs. Summit's peak computational capability is expected to meet or exceed 150 Pflops.

This past spring, the Argonne Leadership Computing Facility (ALCF) announced that technology manufacturer Intel and computer manufacturing company Cray would deliver a massively parallel manycore system 18 times more powerful than Argonne's current system, Mira. It will be based on the third-generation Xeon Phi family of chips from Intel.

The new system, Aurora, is scheduled to arrive in 2018 and will be built using these advanced processors, second-generation Intel Omni-Path Architecture interconnect technology, a new memory architecture, and a Lustre-based parallel file system—all integrated by Cray's high-performance

computing software stack. Intel will coordinate the overall system and silicon design and integration, while Cray will provide software stack expertise and its large system manufacturing capabilities. Aurora's peak computational capability is expected to be at least 180 Pflops.

Ramping Up

To ensure that, starting from day one, the capability of a leadership machine is sufficient to meet the scientific challenges that it will eventually support, the Leadership Computing Facility centers run separate programs that provide training and code support to a small number of project teams and then turn them loose on the machines. These Early Science Program projects (science application codes) represent a large portion of the machines' current and projected computational workloads.

OLCF's Center for Accelerated Application Readiness (CAAR) is currently working with 13 partnership teams to refactor and port their codes for Summit. Teams consist of core application developers and OLCF staff and receive support from the IBM/Nvidia Center of Excellence at Oak Ridge. The CAAR application readiness phase is scheduled to continue through spring 2017 and involve intermediate versions of the Summit architecture while the full-scale system is being constructed.

The ALCF's Early Science Program will also award grants of preproduction time to project teams that will consist of application developers and facility staff, along with domain science experts. The Early Science Program for ALCF-3 will have two phases and two separate calls for participation. The first, Theta, targets an early production system based on Intel's second-generation Xeon Phi processor. The second, Aurora, targets the fully integrated system. (The call for participation for Aurora is expected in mid-2016.) The Early Science teams for both systems will receive technical support from the ALCF.

As Office of Science–designated scientific user facilities, OLCF and ALCF support ambitious scientific initiatives for a wide range of investigations through DOE's highly competitive user programs, INCITE (which covers a broad range of research campaigns) and ALCC (which targets DOE mission-specific investigations). These centers provide their user communities with unique opportunities and resources to explore their science at scales otherwise unavailable, thus allowing more rapid progress and yielding new insights that come from higher-fidelity investigations and simulations that incorporate more complete treatments of the phenomena being modeled.

Because many federal agencies rely on the Leadership Computing Facility to conduct important parts of their research, DOE will gather and take into account the high-performance computing needs of those agencies in this next phase of petascale computing. Even as these next-generation machines are being rolled out, DOE has started planning the evolution of its leadership-computing capabilities for the next decade. ■

Acknowledgments

The Argonne Leadership Computing Facility is a DOE Office of Science User Facility supported under contract DE-AC02-06CH11357. The Oak Ridge Leadership Computing Facility is a DOE Office of Science User Facility supported under contract DE-AC05-00OR22725. We gratefully acknowledge additional editorial assistance by Laura Wolf and Gail Pieper.

James J. Hack directs the National Center for Computational Sciences at Oak Ridge National Laboratory. His primary scientific interests include physical parameterization techniques, numerical methods and their implementation on high-performance computers, and diagnostic methods for evaluating simulation quality. Hack has a PhD in atmospheric dynamics from Colorado State University. He's actively involved in a number of national and international advisory and steering committees, including the DOE Office of Science and National Science Foundation appointments. Contact him at jhack@ornl.gov.

Michael E. Papka directs the Argonne Leadership Computing Facility at Argonne National Laboratory, where he also serves as Deputy Associate Laboratory Director for Computing, Environment, and Life Sciences. His research interests include the visualization and analysis of large data from simulation and experimental sources. Papka has a PhD in computer science from the University of Chicago. He's a Senior Fellow of the Computation Institute and an associate professor of computer science at Northern Illinois University. Contact him at papka@anl.gov.

This article originally appeared in Computing in Science & Engineering, *vol. 17, no. 4, 2015.*

Trustworthy Processing of Healthcare Big Data in Hybrid Clouds

Surya Nepal
Commonwealth Scientific and Industrial Research Organization

Rajiv Ranjan
Commonwealth Scientific and Industrial Research Organization

Kim-Kwang Raymond Choo
University of South Australia

As we delve deeper into the "Digital Age," we're witnessing an explosive growth in the volume, velocity, variety, veracity, and value (the 5Vs) of data produced over the Internet. According to recent Cisco[1] and IBM[2] reports, we now generate 2.5 quintillion bytes of data per day, and this is set to explode to 40 yottabyes by 2020[3]—that is, 5,200 gigabytes for every person on the earth. As noted in previous "Blue Skies" columns, data generated by Internet of Things (IoT) devices and sensors are part of the big data landscape.[4,5] IoT comprises billions of Internet-connected devices (ICDs) or "things," each of which can sense, communicate, compute, and potentially actuate, and can have intelligence, multimodal interfaces, physical/virtual identities, and attributes. ICDs can be mobile devices, sensors, medical imaging devices, individual archives, social networks, smart cameras, body sensors, automobile cosimulations, or software logs. In a nutshell, a large volume of veracity data is generated at high velocity from a variety of sources.

The amalgamation of ICDs with big data processing software frameworks and cloud-based hardware resources leads to the creation of novel big data applications in domains such as healthcare, traffic management, smart energy grids, and smart manufacturing. Managing large, heterogeneous, and rapidly increasing volumes of data, and extracting value out of such data, has long been a challenge. In the past, this was partially mitigated by fast processing technologies that exploited Moore's law. However, with a fundamental shift toward big data applications, data volumes are growing faster than they can be analyzed, regardless of increased CPU speeds or other performance improvements. Although the impetus for the remainder of our article comes from healthcare big data, the problems and solutions discussed are applicable to other application domains.

 Published by the IEEE Computer Society

Big Data in Healthcare

A 2015 Gartner report noted that data processing technologies haven't kept pace with the significant increase in the volume of digital healthcare data, and an integrated and trustworthy healthcare analytics solution can facilitate more effective decision making in patient care and risk management, improving quality of life, optimizing performance of services, and so on.[6] Medical professionals have made similar observations. For example, the chief information officer of Boston's Beth Israel Deaconess Medical Center explained that "working with big data in hospital systems is hugely challenging but at the same time holds tremendous promise in providing more meaningful information to help clinicians treat patients across the continuum of care."[7]

Consider, for example, the problem of managing petabytes of multimedia content produced by advanced medical devices in the healthcare or medical domain as exemplified by the following inventions and reports.

- In conjunction with traditional x-rays, medical imaging can now delve deeper into the human body, discovering and analyzing smaller and smaller details. A research team from Williams College at Harvard University has developed a new type of optical medical imaging device that captures high-resolution live video of human cells and molecules.[8]
- A report from AT&T reveals that medical content (x-rays, computed tomography, genetic data, and other pathology test reports) archives are increasing by 20–40 percent each year.[9] In 2012, there were 1 billion of above-mentioned content types in United States alone, accounting for one-third of global storage demand.
- According to another study, "In 2012, worldwide digital healthcare data was estimated to be equal to 500 petabyes and is expected to reach 25,000 petabyes in 2020."[10] Further, it's anticipated that in 2015, an average hospital will need to manage 665 terabytes of patient data, 80 percent of which will be unstructured medical imaging data.

The challenge is how to ensure data confidentiality and integrity when storing such data but still make it highly available, process it to extract actionable information for decision makers, including medical professionals, and share it with collaborators, while preserving the privacy of individual patients and giving them the full control of their data at all times. This challenge calls for a trustworthy big data processing platform.

> According to recent reports, we now generate 2.5 quintillion bytes of data per day, and this is set to explode to 40 yottabyes by 2020.

Private Clouds: What Are the Research Opportunities?

Existing technology deployments within a medical organization, including its internal, on-premise infrastructure (private clouds) for data storage and the image archiving and communication systems used by radiologists, radically limit efforts to harness the massive amount of medical imaging and other healthcare data. In other words, organizations face several limitations when using private clouds to process healthcare application data.

The first limitation is scalability. On-premise private cloud deployments might not consider future growth, resulting in limited scalability. This isn't surprising, as building highly scalable private clouds requires a large capital investment for procuring and installing computing and storage resources. However, the changing volume, velocity, and variety of data make it difficult to accurately plan private cloud capacity, and private clouds are often either under- or overprovisioned. To reduce capital investment, private clouds are always built with limited scalability.

Analytics is another possible limitation. Analytics models and software frameworks required to manage heterogeneous data might not be available in the private cloud because of higher operational costs. In general, as Editor-in-Chief Mazin Yousif notes, public clouds support the most commonly used analytics models and software frameworks because of their commercial interests, and private clouds deploy analytics models and software frameworks not available from public cloud providers or analytics models and software frameworks developed in-house.

A third limitation is data sharing. Data must be shared with collaborators who don't have access to private clouds

or who reside outside the perimeter defenses. For example, a medical practitioner from a hospital in a different jurisdiction might not be able to access the data stored in the private cloud because at present, healthcare providers are generally subject to exacting regulatory requirements to ensure the security and privacy of patient and other sensitive data.

Although private clouds are inherently trustworthy, these limitations hamper the use of private clouds for processing healthcare big data. We also note the evolution of externally hosted private clouds, which are managed by third parties but support strict security and privacy guarantees. For example, the Postgres Plus Cloud Database (PPCD) provides an externally hosted private cloud. This service includes strict security and auditing features that are in compliance with the Health Insurance Portability and Accountability Act (HIPAA).[11] PPCD's architecture ensures that database instances and data are hosted in complete isolation from other instances and data. However, this isn't possible with purely public clouds.

Further, an externally hosted private cloud model incurs higher leasing costs and offers fewer opportunities to optimize costs than public clouds (for example, leasing spot instances from the Amazon Elastic Compute Cloud is relatively cheaper than leasing standard instances). Although public cloud infrastructures offer the opportunity to optimize hosting costs, they're more prone to security and privacy attacks because of the multitenancy of virtual machines (VMs) and data.

On the other hand, public clouds support the scalability and easy sharing of data. Alan Sill, editor of the "Standards Now" column, also rightly pointed out that US-based cloud service providers must ensure that they meet HIPAA requirements and offer levels of service that provide privacy and are in compliance with various internationally recognized standards.

> Building trustworthy end-to-end big data processing platforms that exploit hybrid cloud infrastructures can be challenging.

The National Institute of Standards and Technology (NIST) defines the four stages of the big data lifecycle as collection, preparation, analysis, and action.[12] At different stages of the data processing, however, the data could be targeted by an attacker. For example, big data processing frameworks don't allow application orchestrators such as Apache Yarn (http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html) to control which physical server rack the mapper and reducer VMs are deployed on at runtime. Hence, these instances can be mapped to VMs from other applications because of multitenancy, exposing the data to different types of security and privacy attacks. Public cloud service providers and existing big data processing frameworks have no easy way of detecting or monitoring such data leakage. Therefore, data auditing,[13,14] data protection,[15] and privacy preservation[16] have emerged as salient areas of inquiry for researchers from industry and academia.

Another potential future research opportunity is bringing together the inherent features of public clouds (scalability) and private clouds (security) to build a trustworthy big data processing platform.

A Trustworthy Hybrid Cloud for Big Data Processing

There has been a paradigm shift toward hosting big data applications in hybrid infrastructures consisting of private and public clouds. However, building trustworthy end-to-end big data processing platforms that exploit hybrid cloud infrastructures can be challenging for several reasons.

First, existing big data ingestion frameworks (such as Apache Kafka and Amazon Kinesis), data storage frameworks (such as MongoDB, BigTable, MySQL, and Cassandra), parallel and distributed programming frameworks (such as Apache Hadoop and Apache Storm), scalable data mining frameworks (such as Apache Mahout and GraphLab), and distributed file systems (such as the Hadoop Distributed File System and Google File System) might not guarantee trustworthy (secure and privacy-preserving) data processing. Most of these frameworks can't support encryption of big data without compromising their inherent scalability and performance.

In addition, traditional data and distributed system security and privacy-preserving techniques can't be automatically adapted to operate efficiently in a hybrid cloud infrastructure de-

ployed with multiple big data processing frameworks to process big data with 5Vs characteristics. There are two core reasons for this. First, as noted in a previous column, most of the big data processing frameworks can only process data within a single private or public cloud datacenter.[5] Second, porting existing security and privacy-preserving techniques to multiple big data processing frameworks is a hard undertaking because they support diverse data programming abstractions (for example, MapReduce in Hadoop, continuous query operators in Storm, and transactional operators in MySQL and Cassandra) and perform computation on diverse dataflows (such as batch, streaming, and transactional).

Security and privacy controls in public cloud computing infrastructures support basic security features such as public key infrastructure (PKI)–based access control and authorization to VMs and binary storage resources. They might have limited capability to protect data and applications against security attacks such as denial-of-service (DoS) and Sybil attacks.

Data holders (such as healthcare providers) typically want to ensure that their data is protected from malicious insiders who might steal or exfiltrate the data for sale. Such data can then be used to derive direct clinical value or profit from possible insights.[7] Data movement and advanced encryption techniques (such as homomorphic encryption) can make it challenging to provide data holders full control of their data in hybrid clouds without affecting performance. This gets even more complicated when patients want full control of their own data.

Existing cryptographic schemes are unlikely to be suited to a hybrid cloud deployment because of computational efficiency limitations and other constraints. Attribute-based encryption (ABE), for example, is designed to provide the scalability and flexibility of real-time data sharing in computing environments, including the cloud.[17,18] However, in existing ABE schemes, user revocation remains challenging, particularly when there's a large number of users. In addition, existing ABE schemes require the cloud server to be fully trusted, and in the aftermath of Edward Snowden's revelations that the National Security Agency has been conducting wide-scale government surveillance, the requirement that all cloud servers in the deployment be trusted might be onerous. Therefore, it isn't surprising that cloud cryptography offering enhanced security without compromising usability and performance is an ongoing research topic.

In the event of a security breach, there might not be an easy way to conduct digital investigations, particularly across borders and between organizations, which would allow the victim to mitigate future risks and/or pursue the criminals through a criminal investigation or civil litigation. For example, would it even be possible to remotely collect evidence from a hybrid cloud in the event of a digital investigation or incident response[19]? In addition, as noted elsewhere, "it's currently unclear whether existing legislation, say in Australia, permits the use of such remote real-time evidence preservation and collection processes and tools to preserve evidential material stored or held overseas without a mutual assistance request."[20]

In the virtual laboratory approach, data is kept inside the private cloud.[21] The virtual laboratory hosts the data and supports a number of data processing algorithms. The output datasets are checked against all privacy rules before they're released. This approach isn't scalable because it's built for private cloud infrastructures. Furthermore, it doesn't support privacy-preserving computation over data from multiple, heterogeneous, and dynamic sources because the virtual laboratory is a trusted entity and resides within a defense perimeter.

> There's a need to balance the data's privacy and security against data sharing or performing scalable, efficient, near-real-time data analytics.

Therefore, there's a need to balance the data's privacy and security against data sharing or performing scalable, efficient, near-real-time data analytics. To this end, a data outsourcing approach has emerged.

The Data Outsourcing Approach and Encryption Techniques

Traditional access-control mechanisms have been successfully used for controlled data sharing in collaborative environments; however, the applicability of such mechanisms is limited in a hybrid cloud environment where some

data can reside outside the defense perimeter (that is, organizational boundaries). An alternative is to use the PKI infrastructure supported by public clouds. In addition to its data security limitations, this approach isn't scalable. Recently, several encryption techniques have been developed to address existing security concerns.

Proxy reencryption (PRE) enables data encrypted using one user's public key to be transformed in such a way that it can be decrypted with another user's private key.[22] The basic idea is that two parties publish a proxy key that allows an untrusted intermediary to convert ciphertexts encrypted for the first party directly into ciphertexts that can be decrypted by the second.

Identity-based encryption (IBE) allows any pair of users to communicate securely and to verify each other's signatures without exchanging private or public keys, keeping key directories, or using the services of a third party.[23] This scheme is ideal for sharing information among closed groups of people (for example, within an organization). The idea is based on the public key cryptosystem, but the public keys are generated using attributes (such as company name or IP address) and individual users have corresponding private keys.

Attribute-based encryption (ABE) aims to overcome one of the limitations of earlier IBE schemes—that is, their use of string-based attributes.[24] ABE is one of two applications of Fuzzy IBE, introduced by Amit Sahai and Brent Waters, which allows attributes to take value from a domain other than strings (the other application is IBE that uses biometric identities).[24] In an ABE system, a user's keys and ciphertexts are labelled with sets of descriptive attributes, and a particular key can decrypt a particular ciphertext only if there's a match between the attributes of the ciphertext and the user's key. Sahai and Waters' cryptosystem allows for decryption when a ciphertext and a private key share at least k attributes. Although this primitive was shown to be useful for error-tolerant encryption with biometrics, the lack of expressibility limits its applicability to larger systems.

> Data sharing approaches should be combined with data analytics approaches to support end-to-end trustworthy data sharing.

Existing solutions based on ABE and PRE introduce a heavy computation overhead on the data owner so don't scale well when fine-grained data access control is desired. To address this problem, a combination of ABE and PRE schemes have been proposed in the cloud security and cryptography literature to exploit the benefits of both schemes.

Moreover, existing data sharing techniques do not support the data analytics. A different branch of research has recently emerged in which the computation can be performed on encrypted data in the cloud. Craig Gentry introduced the first fully homomorphic encryption scheme in 2009.[25] This was a revolutionary cryptographic achievement, but the scheme was far too inefficient for any practical use, especially because of its computational complexity (running time). Since 2009, several works have improved upon Gentry's technique, leading to significant reductions in running time. Although many researchers have improved the processing time, homomorphic encryption has other limitations. For instance, it requires that all recipients have access to the same key to encrypt the inputs and decrypt the results, which might be difficult to arrange if they belong to different organizations. This also doesn't support computation over data from multiple sources. Furthermore, current fully homomorphic encryption solutions are limited to a small number of operations or their performance isn't suitable for real-time and complicated analysis. In addition to numerical operations, all data mining operations must be performed over encrypted data. An encrypted data versioning system is also needed. These challenges offer great opportunities for future research.

Data sharing approaches should therefore be combined with data analytics approaches to support end-to-end trustworthy data sharing and processing platforms in public clouds. This question leads to further research on secure multiparty computation. MPC takes private input data from multiple parties and carries out a joint computation on them while ensuring that the input data remains private to their owners during the computation process.

The focus so far has been on data privacy in a private or public cloud. Some applications require a hybrid cloud approach, in which privacy-sensitive data is kept in the private cloud and

de-identified data is kept in the public cloud. This approach works well in the health domain, where the de-identified data can be shared with collaborators and processed in the collaborators/public cloud environment. However, segregating private and public data, moving public data, and integrating results after processing are some of the challenging issues requiring further research.

To ensure the privacy of personally identifiable information (PII) and other sensitive healthcare data in a (hybrid) cloud environment (despite the varying legal requirements in different jurisdictions), it's necessary to ensure the security of the underlying cloud architecture or ecosystem—for example, through the use of cryptography and privacy-enhancing or preserving technologies. Therefore, we need efficient cloud cryptography as well as privacy-enhancing or preserving systems that can be deployed in healthcare settings. We must also ensure that the underlying cloud architecture or ecosystem is designed to facilitate the identification, preservation, and collection of evidential data in the investigation of a data breach incident.

Developing techniques and APIs that can guarantee data security and privacy and computation across a hybrid cloud ecosystem consisting of multiple private and public cloud datacenters remains an open and difficult research problem. Future efforts also need to focus on designing and developing computationally efficient privacy-preserving techniques that seamlessly scale across multiple big data processing frameworks by exploiting the elasticity of hybrid (multiple private and public) cloud infrastructures while adapting to uncertain data volume, data velocity, and data variety. This could be achieved by exploiting the inherent software-level configuration of big data processing frameworks for scaling existing security and privacy-preserving techniques.

In summary, efforts need to focus on the development of security and privacy techniques that can deal with changing volume, velocity, and variety of heterogeneous dataflow (batch, streaming, transactional); be ported to diverse big data programming frameworks (Apache Hadoop, Apache Storm, Apache Hive); deal with variable computational complexity due to heterogeneous VM, storage, and network configurations across multiple clouds; and be seamlessly implemented in multicloud orchestration APIs such as jclouds.

> Segregating private and public data, moving public data, and integrating results are some of the challenging issues requiring further research.

References

1. R. Pepper and J. Garrity, "The Internet of Everything: How the Network Unleashes the Benefits of Big Data," *Global Information Technology Report 2014*, Cisco Systems, 2014; http://blogs.cisco.com/wp-content/uploads/GITR-2014-Cisco-Chapter.pdf.
2. IBM, "Bringing Big Data to the Enterprise," www-01.ibm.com/software/in/data/bigdata.
3. J. Gantz et al., "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," IDC iView, 2012; www.emc.com/leadership/digital-universe/2012iview/index.htm.
4. R. Ranjan, "Streaming Big Data Processing in Datacenter Clouds," *IEEE Cloud Computing*, vol. 1, no. 1, 2014, pp. 78–83.
5. L. Wang and R. Ranjan, "Processing Distributed Internet of Things Data in Clouds," *IEEE Cloud Computing*, vol. 2, no. 1, 2015, pp. 76–80.
6. V. Shaffer, *Agenda Overview for Healthcare*, Gartner report G00270705, 2015; www.gartner.com/doc/2995217/agenda-overview-healthcare-.
7. O. Badawi et al., "Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference," *JMIR Medical Informatics*, vol. 2, no. 2, 2014, e(22); doi: 10.2196/medinform.3447.
8. D. Borghino, "New Medical Imaging Technique Delivers Streaming Video at Molecular Level," *Gizmag*, 7 Dec. 2010; www.gizmag.com/medical-imaging-tracking-molecules-live-tissue-video-rate/17202.
9. *Medical Imaging in the Cloud*, AT&T tech. report, 2012; www.corp.att.com/healthcare/docs/medical_imaging_cloud.pdf.
10. J. Sun and C. Reddy, "Big Data Analytics for Healthcare," *Proc. 19th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2013;

doi:10.1145/2487575.2506178; http://dmkd.cs.wayne.edu/TUTORIAL/Healthcare.

11. F. Dalyrimple, "Postgres Meets HIPAA in the Cloud," blog, 31 Mar. 2015; www.enterprisedb.com/postgres-plus-edb-blog/fred-dalrymple/postgres-meets-hipaa-cloud.
12. Nat'l Inst. of Standards and Technology, *DRAFT NIST Big Data Interoperability Framework: Volume 1, Definitions*, NIST, 2015; http://bigdatawg.nist.gov/_uploadfiles/BD_Vol1-Definitions_V1Draft_Pre-release.pdf.
13. C. Liu et al., "MuR-DPA: Top-Down Levelled Multi-replica Merkle Hash Tree Based Secure Public Auditing for Dynamic Big Data Storage on Cloud," *IEEE Trans. Computers*, preprint, doi: 10.1109/TC.2014.2375190.
14. C. Liu et al., "Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates," *IEEE Trans. Parallel and Distributed Systems*, preprint, doi: 10.1109/TPDS.2013.191.
15. J. Yao et al., "TrustStore: Making Amazon S3 Trustworthy with Services Composition," *Proc. 10th IEEE/ACM Int'l Conf. Cluster, Cloud and Grid Computing* (CCGRID 10), 2010, pp. 600–605.
16. X. Zhang et al., "A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving Intermediate Data Sets in Clouds," *IEEE Trans. Parallel and Distributed Systems*, vol. 24, no. 6, 2013, pp. 1192–1202.
17. V. Goyal et al., "Attribute-Based Encryption for Fine-Grained Access Control of Encrypted Data," *Proc. 13th ACM Conf. Computer and Comm. Security* (CCS 06), 2006, pp. 89–98.
18. J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-Policy Attribute-Based Encryption," *Proc. IEEE Symp. Security and Privacy* (SP 07), 2007, pp. 321–344.
19. D. Quick, B. Martini, and K.-K.R Choo, *Cloud Storage Forensics*, Syngress/Elsevier, 2014.
20. B. Martini and K.-K.R Choo, "Cloud Forensic Technical Challenges and Solutions: A Snapshot," *IEEE Cloud Computing*, vol. 1, no. 4, 2014, pp. 20–25.
21. C.M. O'Keefe et al., "Protecting Confidentiality in Statistical Analysis Outputs from a Virtual Data Centre," *Proc. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, 2013; http://www.unece.org/stats/documents/2013.10.confidentiality.html.
22. M. Blaze, G. Bleumer, and M. Strauss, "Divertible Protocols and Atomic Proxy Cryptography," *Advances in Cryptology—EUROCRYPT 98*, LNCS 1403, Springer, 1998, pp. 127–144.
23. R.L. Rivest, A. Shamir, and L. Adlemanm, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," *Comm. ACM*, vol. 21, no. 2, 1978, pp. 120–126.
24. A. Sahai and B. Waters, "Fuzzy Identity-Based Encryption," *Proc. 24th Ann. Int'l Conf. Theory and Applications of Cryptographic Techniques* (EUROCRYPT 05), 2005, pp. 457–473.
25. C. Gentry, "A Fully Homomorphic Encryption Scheme," PhD thesis, Stanford Univ., 2009; https://crypto.stanford.edu/craig/craig-thesis.pdf.

SURYA NEPAL *is a principal research scientist in the Digital Productivity Flagship at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. His research interests include cloud computing, big data, and cybersecurity. Nepal has a PhD in computer science from Royal Melbourne Institute of Technology, Australia. Contact him at surya.nepal@csiro.au.*

RAJIV RANJAN *is in the Digital Productivity Flagship at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, where he's also a senior research scientist, Julius Fellow, and project leader. At CSIRO, he leads research projects related to cloud computing, content delivery networks, and big data analytics for Internet of Things (IoT) and multimedia applications. Ranjan has a PhD in computer science and software engineering from the University of Melbourne. He has published more than 120 scientific papers. Contact him at rajiv.ranjan@csiro.au or http://rajivranjan.net.*

KIM-KWANG RAYMOND CHOO *is a senior lecturer in the School of Information Technology and Mathematical Science at the University of South Australia. His research interests include cyber and information security and digital forensics. Choo has a PhD in information security from Queensland University of Technology, Australia. Contact him at raymond.choo@fulbrightmail.org or https://sites.google.com/site/raymondchooau.*

This article originally appeared in IEEE Cloud Computing, *vol. 2, no. 2, 2015.*

Multimedia Big Data

Yonghong Tian
Peking University, China

Shu-Ching Chen
Florida International University

Mei-Ling Shyu
University of Miami

Tiejun Huang
Peking University, China

Phillip Sheu
University of California, Irvine

Alberto Del Bimbo
Università degli Studi di Firenze, Italy

The First IEEE International Conference on Multimedia Big Data (BigMM 2015) took place at the Chinese National Convention Center in Beijing, China, from 20–22 April 2015. This conference was jointly sponsored by the IEEE Technical Committee on Multimedia Computing (TCMC) and IEEE Technical Committee on Semantic Computing (TCSEM) and was hosted by Peking University. The motivation for organizing the BigMM conference is the proliferation of multimedia data and ever-growing requests for multimedia applications—including video-on-demand, interactive video systems, surveillance, social media, medicine, and healthcare—making multimedia the "biggest big data" and an important source of insights and information.

In a broader sense, multimedia big data is emerging as the next "must have" competency in our society. As an active and interdisciplinary research field, multimedia big data also presents challenges and opportunities for multimedia computing in the big data era. The BigMM conference thus aims to foster the growth of a new research community, acting as an international forum for researchers and practitioners in academia and industry to present research that advances the state of the art and practice of multimedia big data, identifies emerging research topics, and defines the future of the field.

The theme of BigMM 2015 was "Multimedia: The Biggest Big Data." The technical program consisted of invited talks, oral and poster presentations, panels, a summit, a grand challenge, and several featured workshops. More than 120 participants attended this conference.

Keynote Speeches

Three keynotes were presented in the morning of each day's session.

On the first day, Wen Gao of Peking University presented, "Video Big Data Compression and Analysis." Gao discussed his vision of the three grand challenges in video big data: ultra-efficient compression, object tracking and search in a large-scale surveillance network, and event recognition from real-world surveillance videos. He also demonstrated recent developments to tackle these challenges.

On the second day, Ramesh Jain of the University of California, Irvine, presented "Multimedia and Big Data." Jain introduced his vision of multimedia big data, combining personal and environmental situations. Toward this end, he proposed a situation-recognition framework using heterogeneous data streams to solve real-world challenges for multimedia big data.

Finally, Dick Bulterman of FXPAL presented, "Finding the Needle in the Haystack: Personalizing the Search Through Big Data." His talk focused on a more personalized view of big data: what if we want to search a large collection of data (the haystack) for a single set of important, person-bound items (the needle)? He also shared their experiences from various large-scale international projects in trying to address this problem, both in theory and in practice.

Main Conference Sessions

The conference received paper submissions from 23 countries. Due to the large amount of top-quality submissions, the regular paper acceptance was very competitive (with an acceptance rate of 22.47 percent). In addition to regular papers, 18 short papers and seven posters were accepted, all of which provided novel ideas, new results, and state-of-the-art techniques in the field. In the technical program, the papers were compiled into six oral sessions and one poster session, covering different aspects of multimedia big data such as content analysis, processing, retrieval, systems, and applications, as well as social big media analysis.

The success of BigMM 2015 is evidence that multimedia big data is becoming an active and inter-disciplinary research field in its own right. One major driving force is the amount of multimedia data, which has grown to the extent that the traditional multimedia processing and

2376-113X/15/$31.00 © 2015 IEEE Published by the IEEE Computer Society

Figure 1. Shu-ching Chen introduces the four panel experts: from left, Jitao Sang, Yong Rui, Wenwu Zhu, and Fei Wu.

analysis systems cannot handle the data effectively. As a consequence, several new problems were presented in BigMM 2015 papers, including multimedia big data harvesting, analysis, and retrieval; rare actions/event detection in surveillance big data; cloud-based image enhancements; and mobile crowd sensing. Some novel methods and techniques were also proposed to address the multimedia big data challenges, such as semisupervised multimodal clustering, temporal multiple correspondence analysis, cascaded filtering retrieval, and geometric consistent tree partitioning MinHash (the min-wise independent permutations locality sensitive hashing scheme).

The best paper prize, sponsored by China Security & Fire Technology, was awarded to Yafei Song, Xiaowu Chen, Xiaogang Wang, Yu Zhang, and Jia Li, of Beihang University, for their paper, "Fast Estimation of Relative Poses for 6-DOF Image Localization."

BigMM Summit and Panel

The BigMM Summit was a featured program at the conference, providing a premier forum for leading scholars to present their insightful opinions on the scientific and technological challenges of multimedia big data and a common vision on how to address them. The summit was jointly supported by IEEE-TCMC, IEEE-TCSEM, and ACM SIGMM China Chapter, and was sponsored by Cooperative Medianet Innovation Center at Peking University. Ten experts presented their vision talks in this summit, including Yong Rui from Microsoft Research Asia, Yonggang Wen from Nanyang Technological University, and Changsheng Xu from the Chinese Academy of Sciences.

The summit ended with a panel discussing "When Multimedia Computing Meets Big Data." Four experts, including Rui, Wenwu Zhu from Tsinghua University, Fei Wu from Zhejiang University, and Jitao Sang from the Chinese Academy of Sciences, were invited to discuss and debate their views and experience on this topic with each other and the audience (see Figure 1).

The BigMM Challenge

Each year, the BigMM organizing committee plans to organize an algorithmic competition to address one grand challenge in the field of multimedia big data, which is open to all tool vendors, academics, and corporations. The major aim is not to rank the participants but to recognize the most innovative, efficient, and methodologically advanced tools in the field.

This first BigMM 2015 Challenge was to address the problem of large-scale object tracking over a network of multiple cameras. In such a network, the observations of the same object should be visually different and widely separated in time and space. Moreover, the tracking system shouldn't require calibrated cameras or complete site models, which are not available in most situations. These requirements made the task very challenging but of practical importance.

Six teams participated in BigMM Challenge 2015. The winning team was from Wuhan University; the runner-up and third place teams were from Ningbo University and Nanjing University of Science and Technology, respectively. All these teams, together with other participant teams, presented their solutions in a separate session.

Workshops

Four workshops were held in conjunction with BigMM 2015. They covered several topics that were related to multimedia big data, including hyperspectral imaging, geometry and graphics, multimedia big data compression, and the emerging techniques on big surveillance data analysis. They were scheduled in the afternoon sessions of the first and third days. Overall, these workshops obtained a huge success in terms of the attendance and participation.

Related Special Issues

To further promote the research in multimedia big data, the BigMM 2015 organizing team is also organizing two special issues in *IEEE*

Transactions on Multimedia (T-MM) and the *International Journal of Multimedia Data Engineering and Management* (IJMDEM), with the same theme as the conference. The authors of regular papers were invited to submit the extended versions to these special issues, which are also open for submissions from the general public.

IEEE BigMM 2016 will be held in Taipei, Taiwan, ROC, in 20–22 April 2016. Visit www.bigmm.org for more conference information and the call for papers. **MM**

Yonghong Tian is a professor at Peking University, China. Contact him at yhtian@pku.edu.cn.

Shu-Ching Chen is an Eminent Scholar Chaired Professor at Florida International University. Contact him at chens@cs.fiu.edu.

Mei-Ling Shyu is a professor at the University of Miami. Contact her at shyu@miami.edu.

Tiejun Huang is a professor at Peking University, China. Contact him at tjhuang@pku.edu.cn.

Phillip Sheu is a professor at the University of California, Irvine. Contact him at psheu@uci.edu.

Alberto Del Bimbo is a professor at Università degli Studi di Firenze, Italy. Contact him at delbimboalberto@gmail.com.

This article originally appeared in IEEE MultiMedia, *vol. 22, no. 3, 2015.*

Toward a Hermeneutics of Data

Amelia Acker
University of Pittsburgh

Editor: Bradley Fidler

Recently I watched an all women panel on careers in data science hosted by the University of California, Berkeley's iSchool.[1] The panel members had a range of backgrounds and training, from advertising to educational research, statistics, and topic modeling. Some of the roundtable's experts had PhDs, and a few had MBAs. Each of the panelists worked at Bay Area startups and commerce sites in northern California (think Airbnb, Eventbrite, and Jawbone). These corporate data scientists represent a promising—and fast paced—new field of commerce, analytics, knowledge, and perhaps most importantly, technical change in the present world of networked computing. I was struck by the variety of different ways these information professionals approached the idea of "data" as they were speaking about the nature of their work. The engaging discussion on data science illustrated how data is not just a byproduct of computing technologies but an engine for dynamic change that drives society in different, fascinating directions.

Data science is the systematic process of creating, building, and organizing knowledge with data. It has recently become a "new" area of interest in computing sciences, bioinformatics (including public health), learning sciences, business and marketing, and the information sciences. Higher education institutions have begun to offer master's degrees in data science—few programs exist at the undergraduate or doctoral level, but many are soon to come.[2] The "newness" of data science has become all the rage of late, but for some, it's just a fresh coat of paint. As others have taken pains to point out, the discipline of data science simply appears to consolidate and leverage principles and techniques from a number of fields that already exist, such as statistics, machine learning, knowledge management, and information retrieval.[3,4] What's new is that data science aims to confront the massive volumes of data created and collected today. Looking closely at data now that it is big can inspire us to ask questions about how it has been handled, modified, managed, and circulated since people started leveraging data with information systems and computing machines.

New academic programs aren't the only place where we are seeing the impact of the "data deluge." Increasingly, we are seeing a public consciousness around personal data generation and collection by states and corporations. Data collection (telephony metadata, in particular) has come under intense, international political debate since the Snowden leaks in 2013. Earlier this spring, the US Circuit Court of Appeals for the Second Circuit found that the bulk collection of telephony metadata by the US National Security Agency (NSA) is not authorized by the USA PATRIOT Act, saying that the collection "exceeds the scope of what Congress has authorized."[5] Since the Snowden leaks, media coverage, online activism, and political pressure from around the world brought the normally banal term "metadata" to center stage despite the fact the collection of data about citizens is far from a recent development in surveillance states.

Consumers are increasingly aware that the online traces they create generate data that can be aggregated and turned into black gold. We're also seeing consumer backlash against the aggregation, collection, and data protection that has resulted in numerous security breaches to information systems that regularly put consumers, workers, and citizens at risk. Ethnographers, legal theorists, and communication scholars have suggested that new cryptocurrencies and data-obfuscation techniques in email encryption[6,7] have, in part, stemmed from this new consumer consciousness about how user data is aggregated and applied into new commercial products. From Home Depot and Target to the Office of Personnel Management hacks, social media users and ordinary citizens are facing security breaches that increasingly reveal the staggering amount of information that is collected through networked infrastructures about their behavior, preferences, relationships, and activities. Although citizens' concerns about data collection have existed for many decades,[8] and metadata and surveillance programs that leverage data into commercial applications and state governance are not new issues, I'm interested in asking how historians of computing are confronting new conceptions of data that circulate in society—in academic, commercial, and civic spheres—and what we might have to contribute to new scholarship about data.

In the last Think Piece article in the *Annals,* William Aspray presented information domains as a promising area for computing historians to consider as way of getting at the "larger meaning of information and information technology in society."[9] I want to take Aspray's argument a little further, building off of his notion of information domains such as data curation or archival science, and suggest that data—how it is being created, packaged, deployed, and understood—is fruitful for computing historians to consider as part of a larger trend over the last 50 years toward networked information systems. For information scholars, the difference between information and data is context. A piece of data without context is without meaning, but when

 Published by the IEEE Computer Society

data is put into context through practices such as aggregation, description, classification, organization, or application, it becomes meaningful information to people and machines. Data that has become information may also have multiple layers of context or acquire more contextual information over time. For example, in the US, birth records are connected to social security numbers that can be aggregated in the Social Security Death Index (SSDI) database of death records. A database of death dates also carries lifetimes of information, including legacy information systems (such as analog birth and death records) as well as evidence for other kinds of data (such as population statistics). The data aggregated and classified in the SSDI database acquire multiple layers of context and carry multiple ontologies about the categories of "life" and "death" depending on how the information is accessed and interpreted. This is just one example of how data can become information with different layers of context.

Here, I want to bring data to the fore, the ways that it figures into different realms, as a phenomena that increasingly seems to penetrate all information domains, including fields of scholarship and areas of society. I argue that historians of computing committed to documenting, charting, understanding, and explaining technological change can expand and shape a growing area of scholarly research, which is increasingly being called "data studies" and which has strong and direct links to the history of computing research agenda.

Historians of computing committed to documenting, charting, understanding, and explaining technological change can expand and shape a growing area of scholarly research.

Data as Computing History

Data, as traces of transmission, are becoming the fundamental organizing principle of emerging cultural records that represent vast swaths of data created as part of networked computing infrastructures.[5] In my work on new information objects created with mobile computing infrastructures, I am particularly drawn to the origins of how data, or traces of data, come to be in information systems. In my earlier work on the Short Message Service (SMS) text messaging protocol,[10] I was concerned with the metadata that encapsulated text messages as part of their transmission across wireless networks. The metadata of text messages (not the message content itself) are used to route network traffic information and to locate senders and recipients of text transmissions in wireless networks. The NSA surveillance programs that were uncovered in 2013 showed us that these context traces, the data about text messages, are useful and collected in all sorts of ways by network operators, handset manufactures, standards organizations, and surveillance programs.[11] These routing data are metadata, which represent their own kind of documentation, records of transactions between people, institutions, machines and cultures through time. The existence of new kinds of data, such as telephony metadata from text messages, points to a shift in the history of recorded information and the ways we communicate with mobile networks that is different from earlier, analog communication networks. We need metadata about transactions for the networked information infrastructure to work. Histories of data help us understand how these layers of context and meanings are acquired through their development, stabilization, and circulation.

Scholars have studied the emergence of the data collection, privacy, and the surveillance society as social constructions since the 1960s, and they can help us begin to make sense of this current data deluge. For example, JoAnne Yates,[12] Geoffrey Bowker and Susan Leigh Star,[13] and Christine Borgman[14] have examined the origins of new kinds of documents, formats, and information objects in information infrastructures in distinct eras and expert domains. Other information infrastructure scholars such as Michael Buckland,[15] David Ribes,[16] and Matthew Mayernik[17] have analyzed the stabilization of formats, systems, and standards and their influence on computing in cultures of information and documentation work. There has also been a spate of work that focuses on how these traces of transactions in the histories of networks has shifted

It is time for us to consider that data may become a central part of the history of computing, and it will need to be for the foreseeable future.

away from organizational and expert cultures to see how new data subjects are developing.[18–20] Still other scholars, influenced by Michel Foucault, come to the study of data traces in information infrastructures by way of privacy and data collection techniques under legislation, network architecture, and technical politics.[21–24] A final body of scholarship points to the ways that the Internet supports new modes of being, such as the "algorithmic self," where users create corpora of personal data traces across social media platforms.[25–27]

This is certainly not an exhaustive list of the work being done in data studies, but it is evidence that a growing number of social scientists, media scholars, and organizational theorists are engaging with data in the recent history of technology. Still, few examine the origins and stabilization of data as a focal point. If the rise of data science, legislation around data collection, and consumer consciousness toward data generation is part of everyday life, how can historians of computing help apprehend data and its growing centrality to the information domain of data scholarship in particular?

One way might be to analyze and describe how network infrastructures are created by the generation and implementation of data, which can provide a way for us to examine the development and design of networking architecture and technologies, in what Andrew Russell calls, "histories of networking."[28] Yet, data and infrastructures have always had an intertwined existence, and the entanglement has become tighter and harder to distinguish, describe, and interpret with new Internet technologies and next-generation wireless networks.[29] In a relatively short time (less than 25 years), mobile computing with handsets has become the primary way of communicating information in terms of volume, frequency, and penetration for much of the developed world.[30] Clearly, historians of computing must account for digital traces and new formats such as telephony metadata, but it is uncertain whether existing approaches that describe data can account for the complexities of today's networked infrastructures. To my mind, data's impact on society and studies of data have reached a point for which it is now time for historians of computing to historicize data directly. And there needs to be an equitable balance between studying the effect of data and studying context—the processes of its creation, stabilization, and transmission in information infrastructures. Given the possibilities of emerging data in contemporary society, it is time for us to consider that data may become a central part of the history of computing, and it will need to be for the foreseeable future.

I propose that one way to do this might be to look at data within different scales of infrastructure, as Paul N. Edwards has suggested.[31] Studying data at different scales produces different views of how technology develops as well as how specific technologies affect individual practices (such as recordkeeping or evidence building) and in organizational practices (like business communications), as James Cortada has extensively documented.[32,33] Building upon Thomas Misa's framework for scalar analysis,[34] Edwards describes the micro-, meso-, and macro-scales of society and how infrastructure can be approached in different ways at each scale. Micro refers to the individual or personal level, the day-to-day practices that make up our lives. The meso-scale is the organizational or institutional change that we see with groups of people across weeks and years. Finally, the macro-scale refers to infrastructure over long periods of time, decades or even centuries (what some have called the "long now"[35]).

The beauty of approaching data as it moves through information infrastructures at different scales of analysis is that scales of inquiry are adaptable, like a pocket telescope, extensible and collapsible with quick gestures. Although many studies of data in computing history are at the macro-scale, I'm particularly drawn to the meso-levels of infrastructures, where people create and rely upon new forms of data as information. It is at the meso-level where ethnographers who examine information systems (such as Peter Botticelli,[36] Kalpana Shankar,[37] and Susan Leigh

What will the future of data studies be? How can historians be faithful to particular information ages' data and distinguish them?

Star[38]) all find rapid change—in this messy, in-between area where groups of people are communicating with documents and where the stuff of data creation, stabilization, reception, and circulation actually happens.

The Future of Data Studies

We have benefited from the applications of information domains, theories of infrastructure, and histories of computer networks, but I believe that now we should turn toward studies of data at different scales of information infrastructure. Careful studies of data, and their interpretation and development in histories of networking can tell us more about context, change, and continuities over time when it comes to computing more broadly. There is a role for historians of computing to tell us more about how this moment of data science came to be by looking at information domains through data and the ways data acquires layers of context to become information. Aspray, and others have argued that histories of computing, and the Internet in particular, have been too focused and limiting.[28,39,40] A hermeneutics of data is needed at the micro-, meso-, and macro-scales of networked infrastructures. In asking the field to turn toward the "newness" of data in this moment, I am not arguing that we should jump on the "big data" bandwagon. Instead, I am asking for us to consider how and why data came to be viewed as new again, arguably one of the major cultural developments of the past decade across national boundaries and across fields of expertise and practice.

Young investigators at the nexus of computing history, information infrastructure studies, and communications have begun to examine data at different scales of infrastructure in some interesting and fascinating ways. For example, Brian Beaton has recently written about software that promotes new types of everyday data gathering with mobile devices, and he calls for specific groups of social actors to rework their social relations around continuous data exchange and to form themselves into new types of networked subjects (what he calls "crowdsourced selves").[41] Kevin Driscoll has written about the social history of database technologies, finding that data structures within collections can heavily influence the flavors of database populism that may arise and recordkeeping possibilities with systems such as in social media.[42] Megan Finn has examined analog data artifacts during the 1857 Tejon earthquake, showing that historical information infrastructures of the period before standardized timekeeping shaped popular understandings of natural disasters.[43]

Social histories such as these, of data moving through scales of information infrastructures, represent a valuable intervention into histories of networking but also into studies of information, system design, and communication technologies in different realms of society. What will the future of data studies be? How can historians be faithful to particular information ages' data and distinguish them? Some readers may find this call too concerned with the present, but with a turn toward the study of data in context, we can disinter the ways in which infrastructures of transmission shape recorded information, in this moment and over time. Computing historians are uniquely positioned to probe the entanglement of networked infrastructures, data, and cultures of computing in the recent past and near future. But the key requirement is to first elevate data, making it central to computing history, as it already is within society.

References and Notes

1. Berkeley iSchool, "Technology for the Greater Good: Careers in Data Science," video, 2014; www.ischool.berkeley.edu/events/20140903careersnindatascience/video2.
2. M. O'Neil, "As Data Proliferate, So Do Data-Related Graduate Programs," *Chronicle of Higher Education*, 3 Feb. 2014.
3. B. Darrow, "Data Science Is Still White Hot, But Nothing Lasts Forever," Future of Work blog, *Fortune*, 21 May 2015; http://fortune.com/2015/05/21/data-science-white-hot/.
4. J. Furner, "Information Science Is Neither," *Library Trends*, vol. 63, no. 377, 2015, pp. 362–363.

5. United States Court of Appeals for the Second Circuit, *ACLU v. Clapper*, 7 May 2015, p. 97.
6. B. Maurer, T.C. Nelms, and L. Swartz, "'When Perhaps the Real Problem Is Money Itself!' The Practical Materiality of Bitcoin," *Social Semiotics*, vol. 23, no. 2, 2013, pp. 261–277.
7. F. Brunton and H. Nissenbaum, "Vernacular Resistance to Data Collection and Analysis: A Political Theory of Obfuscation," *First Monday*, vol. 16, no. 5, May 2011; http://firstmonday.org/article/view/3493/2955.
8. S.E. Igo, "The Beginnings of the End of Privacy," *The Hedgehog Rev.*, vol. 17, no. 1, Spring 2015; www.iasc-culture.org/THR/THR_article_2015_Spring_Igo.php.
9. W. Aspray, "Information Society, Domains, and Culture," *IEEE Annals of the History of Computing*, vol. 37, no. 2, 2015, pp. 2–4.
10. A. Acker, "The Short Message Service: Standards, Infrastructure and Innovation," *Telematics and Informatics*, vol. 31, no. 4, 2014, pp. 559–568.
11. S. Landau, "Making Sense from Snowden: What's Significant in the NSA Surveillance Revelations," *IEEE Security & Privacy*, vol. 11, no. 4, 2013, pp. 54–63.
12. J. Yates, *Control Through Communication: The Rise of System in American Management*, JHU Press, 1993.
13. G.C. Bowker and S.L. Star, *Sorting Things Out: Classification and Its Consequences*, 1st ed., MIT Press, 1999.
14. C.L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, MIT Press, 2007.
15. M.K. Buckland and R.R. Larson, "Metadata as Infrastructure: What, Where, When and Who," *Proc. Am. Soc. Information Sciences and Technology*, vol. 42, no. 1, Jan. 2005.
16. D. Ribes, "Ethnography of Scaling, or, How to a Fit a National Research Infrastructure in the Room," *Proc. 17th ACM conf. Computer-Supported Cooperative Work & Social Computing*, 2014, pp. 158–170.
17. M.S. Mayernik, "Research Data and Metadata Curation as Institutional Issues," *J. Assoc. Information Science and Technology*, preprint, 30 Jan. 2015; http://onlinelibrary.wiley.com/doi/10.1002/asi.23425/abstract.
18. W.H.K. Chun, *Programmed Visions: Software and Memory*, MIT Press, 2011.
19. Rita Raley, "Dataveillance and Countervailance," *Raw Data Is an Oxymoron*, L. Gitelman, ed., MIT Press, 2013, pp. 121–146.
20. A.R. Galloway, *Protocol: How Control Exists after Decentralization*, MIT Press, 2006.
21. S.E. Landau, *Surveillance Or Security?: The Risks Posed by New Wiretapping Technologies*, MIT Press, 2010.
22. K. Shilton, "Participatory Personal Data: An Emerging Research Challenge for the Information Sciences," *J. Am. Soc. Information Science and Technology*, vol. 63, no. 10, 2012, pp. 1905–1915.
23. L. DeNardis, "Hidden Levers of Internet Control," *Information, Communication & Society*, vol. 15, no. 5, 2012, pp. 720–738.
24. K. Crawford, "When Big Data Marketing Becomes Stalking: Can Data Brokers Be Trusted to Regulate Themselves?" *Scientific Am.*, 28 Jan. 2004; www.scientificamerican.com/article/when-big-data-marketing-becomes-stalking/.
25. I. Gershon, *The Breakup 2.0: Disconnecting over New Media*, Cornell Univ. Press, 2011.
26. A.E. Marwick, *Status Update: Celebrity, Publicity, and Branding in the Social Media Age*, Yale Univ. Press, 2013.
27. A.N. Markham et al., "Algorithmic Identity: Networks, Data, and the Terrible Beauty of the Black Box," Selected Papers of Internet Research, 2014; http://spir.aoir.org/index.php/spir/article/view/891/466.
28. A.L. Russell, *Open Standards and the Digital Age*, Cambridge Univ. Press, 2014.
29. B. Fidler and A. Acker, "Metadata and Infrastructure in Internet History: Sockets in the Arpanet Host-Host Protocol," *Proc. 77th ASIS&T Annual Meeting*, vol. 51, no. 1, 2014.
30. Int'l Telecommunications Union, "World Telecommunication/ICT Indicators database," 18th ed., Dec. 2014; www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx.
31. P.N. Edwards, "Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems," *Modernity and Technology*, T.J. Misa, P. Brey, and A. Feenberg, eds., MIT Press, 2004, pp. 185–225.
32. J.W. Cortada, *The Digital Hand: Volume II: How Computers Changed the Work of American Financial, Telecommunications, Media, and Entertainment Industries*, Oxford Univ. Press, 2005.
33. J.W. Cortada, *Information and the Modern Corporation*, MIT Press, 2011.
34. T.J. Misa, "How Machines Make History, and How Historians (and Others) Help Them to Do So," *Science, Technology, & Human Values*, vol. 13, nos. 3–4, 1988, pp. 308–331.
35. D. Ribes and T.A. Finholt, "The Long Now of Technology Infrastructure: Articulating Tensions in Development," J. *Assoc. Information Systems*, vol. 10, no. 5, 2009, pp. 375–398; http://search.proquest.com/openview/12a0d1f3490378e6817f3fc4ed200341/1?pq-origsite=gscholar.
36. P. Botticelli, "Records Appraisal in Network Organizations," *Archivaria*, vol. 1, no. 49, Jan.

2000; http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12743/13929.
37. K. Shankar, "Ambiguity and Legitimate Peripheral Participation in the Creation of Scientific Documents," *J. Documentation*, vol. 65, no. 1, 2009, pp. 151–165.
38. S.L. Star, "The Politics of Formal Representations: Wizards, Gurus, and Organizational Complexity," *Ecologies of Knowledge: Work and Politics in Science and Technology*, S.L. Star, ed., SUNY Press, 1995, pp. 88–118.
39. W. Aspray and B.M. Hayes, *Everyday Information: The Evolution of Information Seeking in America*, MIT Press, 2011.
40. T. Haigh, A.L. Russell, and W.H. Dutton, "Histories of the Internet: Introducing a special issue of Information and Culture," *Information & Culture*, vol. 50, no. 2, 2015, pp. 143–159.
41. B. Beaton, "Safety as Net Work: 'Apps Against Abuse' and the Digital Labour of Sexual Assault Prevention," *MediaTropes*, vol. 5, no. 1, 2015, pp. 105–124.
42. K. Driscoll, "From Punched Cards to 'Big Data': A Social History of Database Populism," *communication +1*, vol. 1, no. 1, 2012, article no. 4.
43. M. Finn, "Information Infrastructure and Descriptions of the 1857 Fort Tejon Earthquake," *Information & Culture*, vol. 48, no. 2, 2013, pp. 194–221.

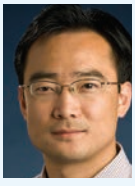
Amelia Acker *is an assistant professor in the School of Information Sciences at the University of Pittsburgh. Her research interests include information infrastructure studies, archival science, and data studies, specifically the material production and transmission of information objects in networked recordkeeping systems over time. Acker has a PhD in information studies from the University of California, Los Angeles. Contact her at aacker@pitt.edu.*

This article originally appeared in
IEEE Annals of the History of Computing,
vol. 37, no. 3, 2015.

Cross-Layer Cloud Resource Configuration Selection in the Big Data Era

Rajiv Ranjan
Commonwealth Scientific and Industrial Research Organization

Joanna Kołodziej
Cracow University of Technology



Lizhe Wang
China University of Geosciences

Albert Y. Zomaya
University of Sydney

The emergence of cloud computing has facilitated resource sharing beyond organizational boundaries and among various applications. This cloud resource sharing is primarily driven by resource virtualization and utility computing (the pay-as-you-go pricing model). The generic multilayered cloud service model is appealing to many parties—from small businesses looking for a low upfront infrastructure investment, to enterprises wanting to cut the cost of managing infrastructures, to research communities requiring large-scale data processing and computing power. In a cloud environment, computing resources (processors, storage devices, network bandwidth, and so on) and applications are provided as services over the Internet.

Fueled by an insatiable demand for new Internet services and a shift to cloud computing services that are largely hosted in commercial datacenters and in the large data farms operated by companies like Amazon, Apple, Google, Microsoft, and Facebook, discussions increasingly focus on the need to ensure application performance under various uncertainties. Through the infrastructure-as-a-service (IaaS) and platform-as-a-service (PaaS) concepts, datacenters virtualize their hardware and software resources and rent it on demand. In the cloud computing approach, multiple datacenter applications (such as content delivery networks, multitier Web, big data analytics, and large-scale scientific simulations) are hosted on a common set of servers. This allows for consolidation of application workloads on a smaller number of servers that can be better utilized, because different workloads might have different resource utilization footprints as well as temporal variations.

Multiple providers in the current cloud landscape offer IaaS and PaaS resources under heterogeneous configurations. Hence, application owners face a daunting task when trying to select cloud services that can meet their constraints. According to recent estimations, there are hundreds of IaaS providers around the world. Even within a particular provider there are different variations of services. For example, Amazon Web Services

 Published by the IEEE Computer Society 2376-113X/15/$31.00

(AWS) has 674 offerings differentiated by price, quality of service (QoS), and service-level agreement (SLA) features and location.[1] Further, every quarter they add about four new services, change business models (price and terms), and sometimes even add new locations. To select the best mix of IaaS and PaaS resource configurations from an abundance of possibilities, application owners must simultaneously consider and optimize application-level performance SLAs while dealing with complex software and hardware configuration dependencies; and managing heterogeneous sets of configurations (price, hardware features, software framework features, location, performance, and so on) remains a hard challenge.[2–4] For instance, it's not enough to just select an optimal cloud storage resource configuration for a content-delivery network application. Allocating corresponding CPU configurations (for example, type, speed, and cores) to content-delivery network media servers and encoding/decoding servers is essential to guaranteeing the ability to serve the content as fast as possible across a variety of devices.

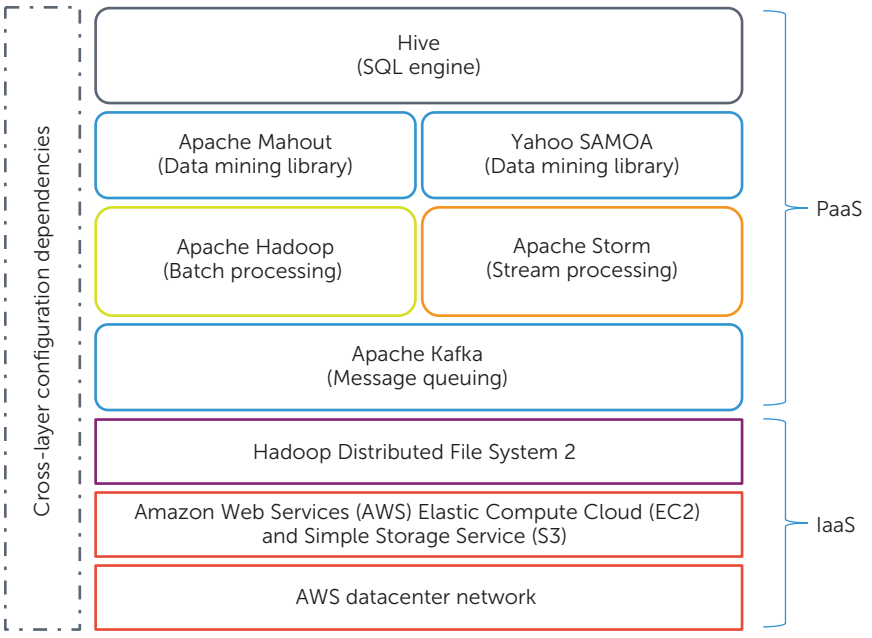


FIGURE 1. A stock recommendation application deployed over an Amazon Web Services datacenter.

Cross-Layer Resource Configuration Selection

To simplify understanding of the cross-layer resource configuration selection problem, consider a social-network-driven stock recommendation big data application deployed on an AWS datacenter, as illustrated in Figure 1.

This application needs to process both historical and real-time data, hence its application architecture consists of multiple and heterogeneous big data processing frameworks. Therefore, the application combines streaming free-form text data from the Twitter API with historical tweets (available via Twitter Firehose) stored in Amazon Simple Storage Service (S3) hardware resources. In the example in Figure 1,

- Apache Kafka is deployed as a high-throughput message-queuing framework;
- Apache Storm is deployed as a stream-processing framework that in turn exploits Yahoo Scalable Advanced Massive Online Analysis (SAMOA) as a data mining framework for classifying groups of tweets relevant to a particular stock;
- Apache Hadoop is deployed for processing historical tweets;
- Apache Mahout, which is hosted within the Apache Hadoop runtime environment, implements a Bayesian classifier algorithm for tweet grouping and classification; and
- the output of both batch and stream analytics subsystems is written to the Hadoop Distributed File System (HDFS).

To query the analytics result (for example, the top K most promising stock portfolios), Apache Hive is deployed to support search queries in Standard Query Language (SQL) format.

As Figure 1 shows, there are two application management layers in a big data application platform.[5,6] The first is a big data processing or PaaS framework (Apache Hadoop, Apache Storm, Apache Mahout, and so on) layer that implements software-based data processing primitives (for example, batch processing by Apache Hadoop or stream processing by Apache Storm). In the second IaaS layer, cloud-based hardware or IaaS resources (for example, CPU, storage, and network) provide hardware resource capacity to the higher-level PaaS frameworks.

The hard challenge is determining the optimal approach to automatically select IaaS resource and big data processing framework configurations

such that the anticipated application-level performance SLA constraints (for example, minimize event-detection and decision-making delays, maximize application and data availability, and maximize number of alerts sent per second) are consistently achieved, while maximizing cloud datacenter CPU utilization, CPU throughput, network throughput, storage throughput, and energy efficiency. For example, in the configuration in Figure 1, there's a need to optimally select configurations for Apache Hadoop (number of map and reduce tasks, map and reduce slots per CPU, max RAM per slot, and so on) and AWS CPU resource configurations (I/O capacity, RAM, CPU speed, local storage, cost, and so on) driven by application-level performance SLA constraints (for example, analyze 100 Gbytes of stock purchase tweets in x minutes subject to a maximum budget of y dollars). A similar cross-layer configuration challenge exists for other big data processing frameworks, such as Apache Storm and Apache Hive. In summary, managing such layered SLA dependencies across multiple big data processing frameworks is widely recognized as a challenging problem.[6–9]

Some cloud computing datacenter providers are also offering software-defined networks (SDNs) at the IaaS layer for deploying big data processing frameworks. In general, SDN deployments implement virtual switches and the software-level OpenFlow protocol to realize communication between the controller and forwarding devices. In SDN-enabled datacenters, the configuration selection and placement of big data processing frameworks in virtual CPUs needs to be coordinated with the selection of network routes between physical servers.

Existing big data application orchestration platforms (Apache YARN, Mesos, and Apache Spark) are designed for homogeneous clusters of IaaS resources. These platforms expect application administrators to determine the number and configuration of allocated IaaS resource types and provide appropriate configuration parameters for each IaaS resource type and big data processing framework for running their analytics tasks. Branded price calculators, available from public cloud providers such as AWS (http://calculator.s3.amazonaws.com/index.html) and Azure (http://www.windowsazure.com/en-us/pricing/calculator) and academic projects such as Cloudorado (www.cloudorado.com), allow comparison of IaaS resource leasing costs. However, these calculators can't recommend or compare configurations across big data processing frameworks and hardware resources while ensuring application-level SLAs (such as minimizing event-detection delay).

> Allocating IaaS-level cloud resources to PaaS-level big data processing frameworks is no longer a traditional time-minimization or resource-maximization problem.

Overview of Multicriteria Optimization and Decision-Making Approaches

The vast configuration diversity among the available cloud resources and big data processing frameworks makes it difficult for application administrators to select configurations or even determine a valid background for their decisions. Consequently, allocating IaaS-level cloud resources to PaaS-level big data processing frameworks is no longer a traditional time-minimization or resource-maximization problem but involves additional simultaneous objectives and configuration dependencies across multiple IaaS resources and big data processing frameworks. These include, but aren't limited to,

- maximizing classification accuracy for Apache Mahout,
- minimizing response time for map and reduce tasks in Apache Hadoop,
- minimizing stream processing latency in Apache Storm,
- maximizing network throughput for HDFS,
- maximizing CPU resource utilization, and
- minimizing energy consumption for the datacenter.

The ever-increasing heterogeneity of hardware being deployed in datacenters comprising multicore processors and coprocessors—general-purpose computing on graphics processing units (GPGPU), IntelMIC, and so on—further complicates the decision-making problem. Clearly, selecting configurations at the IaaS and PaaS layers for big data applications is a multiobjective optimization problem that doesn't have a single solution, but rather a set of tradeoff solutions (known as a Pareto front) corresponding to the SLA objective functions of each

big data processing framework (see Figure 1). Given that big data processing frameworks (such as Apache YARN and Apache Storm) share the same cluster of virtualized resources in a datacenter, workload contention among the frameworks further complicates computing the optimal configuration for meeting performance SLAs. For some big data workloads (high-volume batch processing of historical tweets by Hadoop in Figure 1), storage requirements dominate, whereas for others (transactional query processing by Apache Hive in Figure 1), computational requirements dominate, and for still others (real-time Twitter stream processing by Apache Storm in Figure 1), communication requirements dominate. Hence, in such complex application deployment scenarios, configuration selection techniques must have intelligence to help them determine which workloads should be combined on a common physical server (hosting multiple virtualized CPUs) to minimize resource contention due to workload interference. Such contention-related intelligence can be obtained via offline benchmarking as well as real-time SLA monitoring techniques—a very challenging problem.

A decision problem typically involves balancing multiple, and often conflicting, objective functions and constraints. Numerous research publications present different techniques to solve the configuration selection problem. Each technique is characterized by its search computational efficiency and flexibility in handling the diverse set of objectives and criteria. We classify these methods into three categories: constraint optimization, multicriteria evolutionary optimization (MCEO), and multicriteria decision analysis (MCDA).

Constraint Optimization

Constraint-optimization techniques such as linear programming can efficiently solve configuration selection problems where both the objective and the constraints are linear with respect to all decision variables.[10] Problem instances involving nonlinear objectives and constraints can be solved by applying techniques such as integer programming. However, in practice, neither of these techniques can handle the multiple conflicting objective functions (such as minimizing energy consumption while maximizing CPU utilization and HDFS network throughput).

To handle multiple objective functions, researchers have developed goal-programming techniques that transform the multiobjective problem into a single-objective optimization problem by assigning weights to objectives aggregated in an analytical function.[10] These techniques also support a combination of soft and hard (nongoal) constraints that can deviate, allowing for tradeoffs in achieving satisfactory solutions rather than focusing only on optimal solutions. Unfortunately, goal programming also has shortcomings; for example, the weights are problem-dependent and need to be decided a priori via methods such as application benchmarking, and the weights can lead to undesirable solutions if the relationships between the objective functions aren't clearly understood.

Multicriteria Evolutionary Optimization

MCEO techniques are capable of modeling and optimizing several objective functions simultaneously, and can find global optimal solutions.[11] This class of techniques includes various well-known biologically and physically inspired metaheuristics such as simulated annealing, genetic algorithms, particle swarm optimization, ant and bee colony optimization, and Tabu search. These techniques can handle search over an infinite number of feasible alternatives constrained by a finite number of quantitative configuration criteria. Unfortunately, because MCEO techniques are unconstrained, they can lead to high computational complexity. Constraining these techniques' running time requires integrating penalty functions with objective functions during the optimization process. With the evolution of parallel and distributed big data processing frameworks such as Apache Hadoop, it's possible to speed up these techniques via massive parallelization. However, novel research will be required in terms of modeling and implementation of a parallel version of these techniques that incorporates multiple cross-layered SLA objective functions (see Figure 1).

Multicriteria Decision Analysis

MCDA identifies combinations of configurations for frameworks and resources at the PaaS (such as number of map and reduce tasks instances in Apache Hadoop) and IaaS (such as CPU type and speed for hosting instances of map and reduce tasks in Apache Hadoop) layers to achieve application-level performance SLA objectives (such as minimizing data analytics delay). Formally, MCDA consists of "a collection of formal approaches which seek to take explicit account of multiple criteria in helping individuals or groups explore decisions that matter."[12] In general, the cross-layer configuration selection problem is MCDA, which can be briefly defined as a collection of techniques for providing the comparative analysis, ranking, and selection of the alternate combination of configurations[12] that best meets the application-level SLA objectives. Such a combination of configurations can be selected from a finite (known a priori) or very large/infinite (unknown a priori) set of all possible alternatives. A particular

alternative's "usefulness" is expressed by the application-level SLA objective function, the values of which depend on the payoff or decision matrix (static or dynamic) generated for the whole process.

MCDA techniques can be broadly classified as analytic hierarchy process (AHP), multiattribute utility theory (MAUT), and outranking methods.[1,12,13]

AHP is based on a pairwise comparison, with the criteria (cheapest CPU, cheapest storage, fastest map/reduce tasks, and so on) structured in a multilevel hierarchal relationship. The objective is defined at the top level, and the lower levels correspond to super-criteria, subcriteria, and so on. In the AHP tree, the selection process starts at the leaf criterion and progresses toward the top-level goal (final objective function). Each level represents the selection hierarchy corresponding to the weight or influence of different branches originating at that level. The analytic network process (ANP) is an extension of AHP that can be applied to solve decision-making problems that can't be structured hierarchically.

The analytic network process (ANP) is an extension of AHP that can be applied to solve decision-making problems that can't be structured hierarchically.

Unlike AHP, MAUT techniques are based on utility functions that quantify decision makers' preferences. MAUT aims to generate a means of associating a real number with each alternative (solution) to produce a preference order of alternatives consistent with the decision maker's opinion. The attribute values aren't fully determined in the alternative selection process, but can be influenced by some random factors. The consequences of IaaS and PaaS configuration selection should therefore be defined as probability vectors. MAUT techniques combine various preferences in the form of multiattribute utility functions for each criteria, which are combined with attribute weighted functions. The advantage of using MAUT is that the problem is constructed as a single objective function after successful assessment of the utility function. Thus, it becomes easy to ensure the achievement of the best compromise solution based on the higher-level objective function.

Outranking techniques are applied directly to partial preference functions, which are assumed to have been defined for each criterion. These preference functions could correspond to natural attributes on a cardinal scale, or could be constructed as ordinal scales. In this case, the preference functions must satisfy only the ordinal preferential independence condition. The key difference between MAUT and outranking techniques is that MAUT selects the best choice whereas outranking produces a list of alternatives.

Jose Figueira and his colleagues[13] apply AHP to evaluate and compare general features (security, performance, scalability, and so on) of cloud datacenter providers as defined by the Cloud Services Measurement Initiative Consortium.[14] Some authors have used AHP to develop approaches to select and rank SaaS applications such as salesforce automation products.[15] A hybrid decision-making technique proposed elsewhere combines multicriteria decision making (AHP) and evolutionary optimization techniques (genetic algorithms) for selecting the best CPU and webserver images relevant to public clouds (such as AWS).[2] Another approach[16] applies MAUT-based techniques to select SaaS applications driven by trustworthiness.[17] The selection criteria includes quality, cost, and reputation. Using the ELECTRE (elimination et choix traduisant la realité) outranking technique, other authors propose a cloud datacenter selection approach based on general features (not relevant to any application type).[18]

However, to the best of our knowledge, existing approaches based on constraint optimization, MCEO, and MCDA techniques can be used to select configurations of multiple big data processing frameworks and IaaS resources (CPU, storage, and SDN networks) simultaneously while handling cross-layer SLA dependencies.

Recent Efforts

Although we consider a stock recommendation application here, the challenges are relevant to other big data application types as well. These applications include natural hazard management, credit card fraud detection, remote healthcare, and smart energy grids.

Although interest in deploying big data applications on clouds is growing, the set of concepts needed to understand the decision-making problem across multiple layers is still emerging, rather than being well defined or understood. Re-

cent efforts have attempted to automate the configuration selection of Hadoop frameworks over heterogeneous cluster resources. Gunho Lee and his colleagues proposed dynamically allocating heterogeneous cluster resources to a Hadoop framework based on single-performance SLA constraints on storage size configuration.[19] Karthik Kambatla and his colleagues proposed selecting the optimal Hadoop configuration parameters over a given set of cluster resources by developing and profiling resource consumption statistics.[20] Similarly, others proposed selecting configurations of Hadoop frameworks and heterogeneous Amazon EC2 CPU resources under various what-if scenarios (number of map and reduce tasks, size, and distribution of input data).[21] In the Aroma system, the configuration of CPUs is specified, then Hadoop framework is implemented to meet data processing deadlines while minimizing CPU rental cost.[22]

Progress in optimized configuration selection for Hadoop frameworks is significant and sets a foundation for future research, which must focus on developing holistic decision-making frameworks that automate configuration selection across multiple IaaS resource types and big data processing frameworks to ensure application-level SLAs as required in many emerging application domains.

Developing cross-layer configuration selection techniques is difficult. The space of possible configurations for big data processing framework and IaaS resources grows exponentially with the increasing number of big data processing framework types and cloud datacenter IaaS resource types. Computing optimal solutions is time consuming, and is therefore intractable given current technology. The hard challenge will be to identify the most relevant configurations for each big data processing framework and its dependencies on lower-level IaaS resource configurations. More complexities exist in modeling the objectives and criteria for individual big data processing frameworks and simultaneously computing configuration alternatives at design and run time in response to changes in data volume, velocity, variety, and query types.

References

1. M. Whaiduzzaman et al., "Cloud Service Selection Using Multicriteria Decision Analysis," *Scientific World J.*, vol. 2014, 2014, article no. 459375; doi: 10.1155/2014/459375.
2. M. Menzel et al., "CloudGenius: A Hybrid Decision Support Method for Automating the Migration of Web Application Clusters to Public Clouds," *IEEE Trans. Computers*, vol. 64, no. 5, 2015, pp. 1336–1348.
3. M. Zhang et al., "A Cloud Infrastructure Service Recommendation Technique for Optimizing Real-Time QoS Provisioning Constraints," to be published in *IEEE Systems J.* in 2015.
4. R. Ranjan, "The Cloud Interoperability Challenge," *IEEE Cloud Computing*, vol. 1, no. 2, 2014, pp. 20–24.
5. T. Shah, F. Rabhi, and P. Ray, "Investigating an Ontology-Based Approach for Big Data Analysis of Inter-Dependent Medical and Oral Health Conditions," *J. Cluster Computing*, vol. 18, no. 1, 2015, pp. 351–367.
6. D. Abadi et al., "The Beckman Report on Database Research," *ACM SIGMOD Record*, vol. 43, no. 3, 2014, pp. 61–70.
7. R. Ranjan, "Streaming Big Data Processing in Datacenter Clouds," *IEEE Cloud Computing*, vol. 1, no. 1, 2014, pp. 73–83.
8. M. Kunjir, P. Kalmegh, and S. Babu, "Thoth: Towards Managing a Multi-System Cluster," *Proc. Very Large Databases*, vol. 7, no. 12, 2014, pp. 1689–1692.
9. R. Zhang et al., "Getting Your Big Data Priorities Straight: A Demonstration of Priority-Based QoS Using Social-Network-Driven Stock Recommendation," *Proc. Very Large Databases*, vol. 7, no. 12, 2014, pp. 1665–1668.
10. G.L. Nemhauser and L.A. Wolsey, *Integer and Combinatorial Optimization*, Wiley-Interscience, 1988.
11. M. Gendreau and J.-Y. Potvin, *Handbook of Metaheuristics*, Int'l Series in Operations Research & Management Science 146, Springer; doi: 10.1007/978-1-4419-1665-5.
12. U. Habiba and S. Asghar, "A Survey on Multi-Criteria Decision Making Approaches," *Proc. Int'l Conf. Emerging Technologies* (ICET 09), 2009, pp. 321–325.
13. J. Figueira, S. Greco, and M. Ehrgott, eds., *Multiple Criteria Decision Analysis: State of the Art Surveys*, Int'l Series in Operations Research & Management Science 78, Springer, 2005; doi:10.1007/b100605.
14. S.K. Garg, S. Versteeg, and R. Buyya, "A Framework for Ranking of Cloud Computing Services," *Future Generation Computer Systems*, vol. 29, no. 4, 2013, pp. 1012–1023.
15. J. Siegel and J. Perdue, "Cloud Services Measures for Global Use: The Service Management Index (SMI)," *Proc. Ann. SRII Global Conf.*, 2012, pp. 411–415.
16. C.W. Chen et al., "Conceptual Framework and Research Method for Personality Traits and Sales Force Automation Usage," *Scientific Research Essays*, vol. 6, no. 17, 2011, pp. 3784–3793.
17. N. Limam and R. Boutaba, "Assessing Software Service Quality and

This article originally appeared in IEEE Cloud Computing, *vol. 2, no. 3, 2015.*

Trustworthiness at Selection Time," *IEEE Trans. Software Eng.*, vol. 36, no. 4, 2010, pp. 559–574.
18. S. Silas, E.B. Rajsingh, and K. Ezra, "Efficient Service Selection Middleware Using ELECTRE Methodology for Cloud Environments," *Information Technology J.*, vol. 11, no. 7, 2012, pp. 868–875.
19. G. Lee, B. Chun, and H.K. Randy, "Heterogeneity-Aware Resource Allocation and Scheduling in the Cloud," *Proc. 3rd USENIX Conf. Hot Topics in Cloud Computing* (HotCloud 11), 2011, p. 4.
20. K. Kambatla et al., "Towards Optimising Hadoop Provisioning in the Cloud," *Proc. Conf. Hot Topics in Cloud Computing* (HotCloud 09), 2009, article 22.
21. H. Herodotou et al., "A What-if Engine for Cost-Based MapReduce Optimisation," *IEEE Data Eng. Bull.*, vol. 36, no. 1, 2013, pp. 5–14.
22. P. Lama and X. Zhou, "AROMA: Automated Resource Allocation and Configuration of MapReduce Environment in the Cloud," *Proc. 9th Int'l Conf. Autonomic Computing* (ICAC 12), 2012, pp. 63–72.

RAJIV RANJAN *is in the Digital Productivity Flagship at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, where he's also a senior research scientist, Julius Fellow, and project leader. At CSIRO, he leads research projects related to cloud computing, content delivery networks, and big data analytics for Internet of Things (IoT) and multimedia applications. Ranjan has a PhD in computer science and software engineering from the University of Melbourne. Contact him at rajiv.ranjan@csiro.au or http://rajivranjan.net.*

JOANNA KOŁODZIEJ *is an associate professor at Cracow University of Technology, Poland. Her research interests include data-intensive computing, grid and cloud computing, and artificial intelligence. Kołodziej has a habilitation in computer science from the Polish Academy of Sciences. She is a chair of the Polish chapter of the IEEE Computational Intelligence Society. She's a chair of the CHIPSET Cost Acrion project. Contact her at jokolodziej@pk.edu.pl.*

LIZHE WANG *is a ChuTian Chair Professor in the School of Computer Science at the China University of Geosciences (CUG). His research interests include high-performance computing, e-science, and spatial data processing. Wang has a doctor of engineering degree from the University of Karlsruhe, Germany. He's a fellow of the Institution of Engineering and Technology and the British Computer Society. Contact him at lizhewang@icloud.com.*

ALBERT Y. ZOMAYA *is the Chair Professor of High Performance Computing & Networking in the School of Information Technologies at the University of Sydney. His research interests include parallel and distributed computing, data intensive computing, and cloud computing. Zomaya has a PhD from the Department of Automatic Control and Systems Engineering, Sheffield University. He's a fellow of the American Association for the Advancement of Science, IEEE, and the Institution of Engineering and Technology. Contact him at albert.zomaya@sydney.edu.au.*

IEEE

IEEE computer society

intel
IEEE CS
Software
Professional Developer
Platinum
2016

Get the Recognition You Deserve

The 2016 Platinum IEEE Computer Society/Intel Software Developer Award recognizes the most talented software developers in the world.

3 WINNERS will be selected. The 3 developers with the highest exam scores will each receive the Award along with $3000. Winners are also invited to the IEEE CS Annual Awards Banquet in Atlanta, GA, where they will be honored by computing industry leaders.

Complete the Software Professional Developer Certification prior to May 1, 2016, and your test scores will automatically be entered to qualify you for these distinguished awards.

Validate your expertise. Gain global recognition. Advance your career.

FOR COMPLETE INFORMATION ON HOW TO ENTER, GO TO
www.computer.org/intel-award

Anil Jain: 25 Years of Biometric Recognition

Charles Severance, University of Michigan

Computer scientist Anil Jain discusses the evolution of the biometric recognition field.

Today, we don't think twice about swiping our finger to unlock a cell phone or walking into public areas where security cameras are performing real-time face recognition. If you turn on the television, you'll often see biometric technology such as fingerprint matching and facial recognition being used to solve crimes. The speed, memory, and sensors of today's computers make it feasible to use biometrics on a large scale. But it's taken decades of research to understand and build reliable and verifiable algorithms and techniques that underpin this high-stakes space.

I spoke with computer scientist and Michigan State University professor Anil Jain about the early days of biometrics and the field's future. You can see the entire interview at www.computer.org/computingconversations.

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

In the 1980s and 1990s, when mainframes were less powerful than today's wristwatches, there was a lot of focus on developing efficient algorithms for pattern recognition and image processing. For pattern recognition to evolve into biometrics, significantly more computing power was needed for real-time recognition:

> *It was serendipity in 1990 when professor Duncan Buell called me from Washington, DC, and said, "You do good image processing work. The NSA [National Security Agency] has funded the development of an FPGA [field-programmable gate array]—we can give you an FPGA board and some research money. Can you find a civilian application for this hardware?" The FPGA board was called Splash 2 and was an attached processor for Sun SPARCstation hosts.*

Although the FPGA's computational model (an array of Xilinx 4010 FPGAs; see Figure 1) was much simpler than that of general-purpose computers, it was well suited to many basic low-level image-processing algorithms like convolution, smoothing, edge detection, and local filtering as well

 Published by the IEEE Computer Society

as high-level operations such as point matching. But once the algorithms were ported on Splash, the question arose of which application areas to target:

> *One possible application of higher-level operations such as point matching is stereo matching in computer vision. There's a left image and a right image, and you find some landmarks in the two images and align them to obtain a depth map. As we were brainstorming other applications for point correspondence, my graduate student Nalini Ratha and I really liked the idea of implementing fingerprint recognition in FPGAs because fingerprint matching is essentially based on point-matching operations.*

The complete solution to fingerprint matching involved a point-matching algorithm combined with low-level filtering operations to enhance fingerprint images that are often of low quality and a bit blurred when captured.

For Anil, an interest in image processing evolved into a 25-year interest in fingerprint recognition. It's rewarding when your latest research finds its way into the mainstream media:

> *If you watch any crime show on TV these days, like* CSI: Crime Scene Investigation, *they'll show a computer extracting the minutia points from a fingerprint and doing the matching instantly.*

But if you watch those shows, you know that the current generation of fingerprint technology is never enough to solve the crime. There's always the next innovation—both on TV and in research:

> *For the past 100 years, fingerprint matching has been based on minutia points [see Figure 2]. But what happens if the fingerprint image doesn't have a sufficient number of points or the image quality is so poor that we can't extract enough reliable points? That's when you need to look at the image texture formed by the ridges and valleys that characterize the fingerprint.*

In 1998, Anil's graduate students, Salil Prabhakar and Sharath Pankanti, came up with a bank of filters that captured the texture characteristics of a fingerprint that could be used for fingerprint matching. Seventeen years later, these texture-matching algorithms are finding their purpose:

> *The sensors for fingerprint readers in mobile phones are only about 80 × 80 pixels in size. If you only capture a small part of the fingerprint, the number of minutia points isn't enough to establish a correspondence between two different impressions. This is where the texture information becomes especially useful for fingerprint comparison.*

Although many research results from the biometrics field are widely used in authentication systems—ranging from unlocking a mobile phone to large-scale national ID programs like Aadhaar in India (https://uidai.gov.in/aapka-aadhaar.html)—there are still many new areas to explore. As more sensors are embedded in mobile devices, novel approaches to continuously authenticate a device's owner become possible:

> *The traditional model of authentication is that you log in once and then just use your device. But that model revolved around sitting in front of a desktop computer. On a mobile phone, this notion of "authenticate once, use forever" is really not appropriate, which is why we have to keep unlocking our phones. The typical person might unlock his or her phone 40 or 50 times a day. So why doesn't the device learn who you are based on your behavioral patterns, how you swipe the screen, how you hold it, and its GPS location, or even turn the phone's camera on once in a while and capture an image of your face for recognition?*

Could we reissue a fingerprint representation similar to a credit card number that can be revoked and reissued?

Another important research area is the uniqueness of biometric traits like your fingerprint, face, or iris:

> *In principle, every fingerprint has a different friction ridge pattern. There are approximately seven billion people living on Earth right now, so there are about 70 billion fingers. We should be able to discriminate between these 70 billion fingerprints, but it doesn't quite work this way in practice because the pattern on the finger could be quite different from the two-dimensional image of the finger you use for recognition.*

Fingerprint recognition systems and other biometric recognition systems have small non-zero error rates that depend on the quality of the acquired biometric data.

A fundamental premise of any biometric trait is its persistence. Will a fingerprint or iris pattern change over time?

It's generally agreed that facial recognition systems become less reliable when the separation between two facial images of the same person exceeds about 10 years or so. But in the case of a fingerprint or iris, we have been led to believe that they last forever.

Because Anil has worked with law enforcement officials for more than 15 years on a wide range of research questions, they sometimes come to him with new questions, ideas, and data to analyze:

Just recently, along with my former student Soweon Yoon, I completed a study on the persistence of fingerprint recognition using data from the Michigan State Police. They gave us fingerprint records of about 16,000 individuals who had been arrested multiple times over a 12-year period. Using a multilevel statistical model, we showed that fingerprint recognition accuracy over this 12-year period doesn't degrade.

But what happens if somebody steals data that contains your biometric trait?

Today, the image or representation of your fingerprint is stored in your mobile phone or local bank. How do we secure it so that even if your data is stolen, it can't be used to impersonate you? This isn't as farfetched as one might think. The recent attack on the federal Office of Personnel Management resulted in the theft of fingerprint images of more than one million individuals. Although there's a need to collect fingerprints, we should avoid retaining the original versions in operational databases. Could we reissue a fingerprint representation similar to a credit card number that can be revoked and reissued?

Biometrics is a fascinating and continuously evolving application area for computing technology. When biometric data is used in critical situations like solving high-profile crimes or authenticating large financial transactions, it's important to have solid research that ensures the reliability and accuracy of these biometric recognition algorithms.

CHARLES SEVERANCE, Computing Conversations column editor and *Computer*'s multimedia editor, is a clinical associate professor and teaches in the School of Information at the University of Michigan. Follow him on Twitter @drchuck or contact him at csev@umich.edu.

This article originally appeared in Computer, *vol. 48, no. 8, 2015.*

Figure 1. The Splash 2 board consisting of an array of Xilinx 4010 field-programmable gate arrays. A sequential point-matching algorithm (assuming an average of 65 minutia points per fingerprint) executed on a Sun SPARCstation 20 runs at 100 matches per second. The same algorithm implemented on Splash 2 running at 1 MHz executes at 6,300 matches per second. (Source: Duncan Buell, University of South Carolina.)

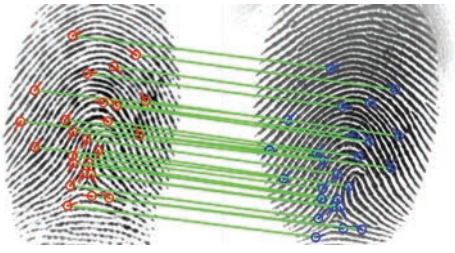


Figure 2. Two different fingerprint impressions (images) of the same finger, showing the corresponding minutia points. The number of paired minutiae is 25.

Keep up with the latest IEEE Computer Society publications and activities wherever you are.

Follow us on Twitter, Facebook, Linked In, and YouTube.

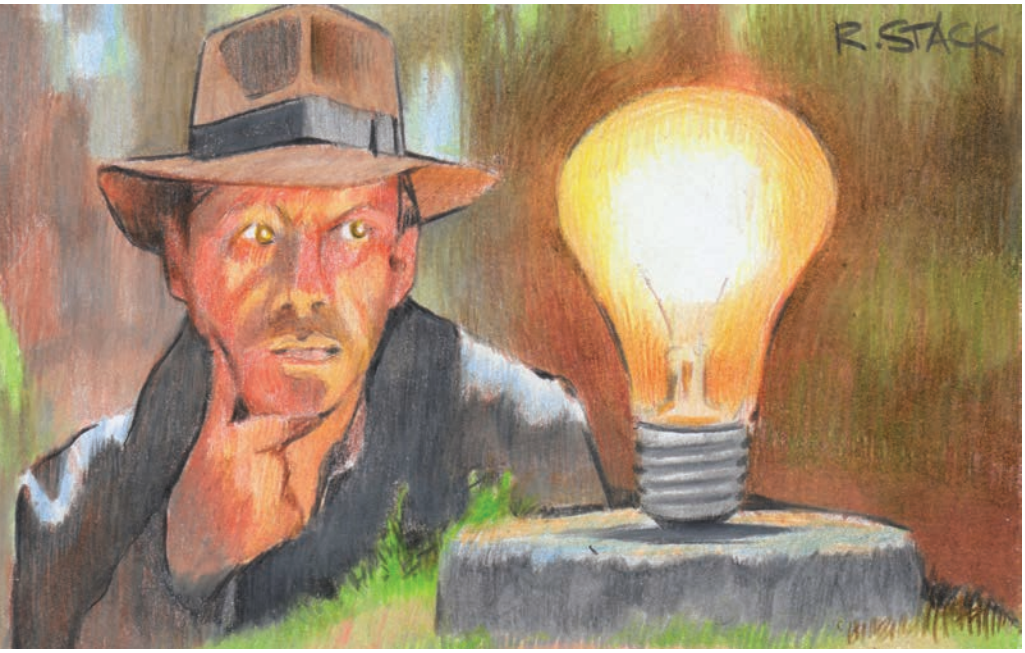
Twitter	@ComputerSociety, @ComputingNow
facebook	facebook.com/IEEEComputerSociety facebook.com/ComputingNow
LinkedIn	IEEE Computer Society, Computing Now
YouTube	youtube.com/ieeecomputersociety

IEEE computer society

Editors: Michael Lesk, lesk@acm.org | Jeffrey MacKie-Mason, jmm@umich.edu

Ideas Ahead of Their Time:
Digital Time Stamping

Michael Lesk | Rutgers University



Older ideas sometimes come back to life. Early ideas might be too far ahead of their time, be phrased in obsolete language, or suffer from bad luck in the exploitation process. Security applications face an additional barrier: the original idea might involve something secret, so others might be unaware of the whole strategy or even the actual idea. Thus, we ought to keep our eyes open for older ideas that might now be practical and useful.

The public tends to believe, perhaps from movies about Thomas Edison, that an inventor has a stroke of inspiration, and the idea moves immediately through the lab to production and worldwide acceptance. However, real technological change involves a great deal of uncertainty in development, production, and marketing. The people who take an idea and turn it into something with impact should get at least as much credit as those who originated it.

From Innovation to Acceptance

It takes longer than we think for innovations to move into practice. As a nontechnical example, consider the discovery of penicillin. Many people are familiar with some variant of the story that Alexander Fleming left a petri dish of bacteria in his lab and went on vacation, leaving the lab door open. A bit of mold blew into the room and killed the bacteria. The world owes a great deal to Fleming's failure to tidy up before he left the lab and his attention to the dish when he returned. This happened in 1928, and Fleming published a paper about it in 1929. Almost nobody paid attention.

It wasn't until 1938 that two biochemists, Howard Florey and Ernst Chain, followed up. Looking at the prospect of another world war, they decided to postpone basic research in favor of looking for something that would help treat infections. Chain went to the library to look for papers about anything that killed bacteria and spotted Fleming's paper. By luck, he discovered during a tea-time conversation that somebody at Oxford had a sample of the bacteria-killing mold. This sample was critical because Fleming—not a mold specialist—had misidentified the species of mold. Florey and Chain managed to get enough penicillin to treat a few sick mice in May 1940 and then a sick human in January 1941. The gap between Fleming's discovery and the attempt to exploit it was nine years.

Digital Time Stamping

What called this to my mind was a story in the *New York Times* about using Bitcoin technology to certify the existence of documents.[1] Paper documents can sometimes be

 Published by the IEEE Computer Society

analyzed chemically or microscopically to detect tampering, but establishing that a particular document existed on a particular date is much harder with bitstrings. Paper documents can be notarized, and some companies offer such an electronic service. But two Bellcore researchers, Stuart Haber and Scott Stornetta, had a new idea 25 years ago.

Suppose you have a good hash function. In this case, establishing that a document's hash existed is equivalent to knowing that the document existed, because you can't find an alternative string that will hash into the same result. The Haber-Stornetta insight was that if you had many documents and hashed each of them, and then hashed all the hashes together, you had one short string that depended on every original document. You could then publish that one string in a public place—originally the *New York Times* public notices section—and use it to confirm the existence of the documents you started with. This was patented as "Method for secure time-stamping of digital documents," US patent 5136647, filed in August 1990 and granted in 1992. The algorithm was also published in 1991 in a well-written paper beginning with a Shakespeare quote.[2]

Originally, the inventors attempted to market the idea as a certification service through a new company they founded, Surety. The idea received minimal public attention,[3] although the company ran a weekly public notice in the *New York Times* to publicly establish the combined hash value (see Figure 1).[4] I have no idea whether *Times* readers ever noticed these.

The time-stamping paper is cited in the original 2008 white paper, *Bitcoin: A Peer-to-Peer Electronic Cash System* by Satashi Nakamoto (believed to be a pseudonym).[5] The same idea is behind the "blockchain" that secures Bitcoin transactions. The original Bellcore patent would have expired in 2009, before Bitcoin became popular.

But in 2015, the *New York Times* noted that people had rediscovered the time-stamping idea as separate from Bitcoin and wanted to use it. So, there was an 18-year delay from invention to use in Bitcoin, and it was another seven years before somebody argued publicly and loudly for extracting the original algorithm and using it independently.

Public Notices 5100

3:34 pm 2/14/92 Bellcore TimeLink
Digital Time-Stamp Marker:
cb4b62bd61406b80bc53ef9e83b560
These markers prevent computer
document tampering. 445 South St.
Morristown, NJ, 07960, Rm. 2Q-346.

Figure 1. Publicly establishing the value of the hash of a chain of time-stamped documents in the *New York Times*.

Other Old Ideas

Many other examples exist. In the early 1980s, I built a system to give driving directions, but a combination of management and marketing problems and the lack of GPS technology meant that it went nowhere. Fifteen years later, I received a call from somebody asking if the old code could be dug out and made to work. By that time, it was too late—a lot of other people had done it better.

Ideas can also expire. Data compression was once a popular research topic but seemed of little practical importance until CDs were invented and getting as much data as possible onto a CD-ROM was significant for marketing. Then, distribution in physical form faded. If you look at a Google Ngram search for "data compression" you'll see interest in it peak from 1994 to 1996, then decline.

When ideas mature has as much to do with technology as marketing and fads. Figure 2 is a Google Trends plot of interest in Bitcoin matched with the price of bitcoins over the same time period. Correlation doesn't imply causation, of course, but one does wonder whether Bitcoin investors are driving the hype.

If old thoughts are a good source of research ideas, are there other good security-related ideas that could be mined from the literature? Almost certainly, yes. One of my early advisors was known for going to conferences in the 1980s and 1990s and claiming after each presentation that he'd thought of that idea in the 1960s.

Security History

Security is an area in which it's likely that good ideas exist but aren't widely known. All innovators run the risk that their work will be overlooked, perhaps because they don't publish prominently enough or because their papers don't happen to be read by anybody. The antibacterial effect of penicillin had been noticed by a French researcher in 1895, but he died young and nobody else carried on his work. Yet another obstacle with security research is that it's often done in secret. So mining old ideas for solutions to present-day problems is probably particularly relevant in the security area. There are many examples of hidden and older work that later became important.

For example, today we're aware that the concepts behind public-key cryptography were discovered by the British security research group Government Communications Headquarters a few years before the better-known announcements by Whitfield Diffie and Martin Hellman. However, the UK researchers didn't have the idea for digital signatures, nor did they anticipate many of the applications for asymmetric cryptography. The UK work wasn't disclosed until 1997, and work done in secret doesn't attract attention or exploitation.

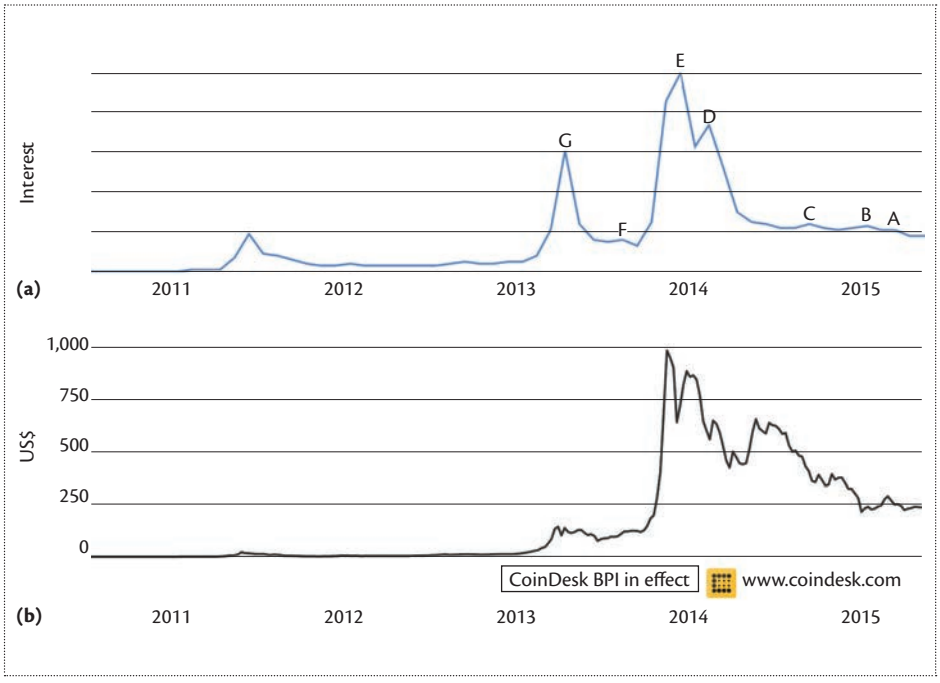


Figure 2. Interest in Bitcoin versus the price of bitcoins. (a) Google Trends plot of "Bitcoin" from 2010 to present. (b) Price of bitcoins from 2010 to present.

Similarly, work on code breaking during World War II inspired at least one thought of language translation. In 1947, Warren Weaver of the Rockefeller Foundation wrote a letter suggesting machine translation, saying "one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography."[6]

When I first read this, I of course didn't know that its author had been aware of computer decryption during the war. I had been taught that the first computers did numerical calculations, especially artillery firing tables. Not until 1974 was it generally known that the first electronic computer did a kind of text processing. Could research have moved faster if we had known of these precedents? Vannevar Bush's famous Memex essay inspired the whole field of information retrieval,[7] and Bush had been working in secret on code breaking with microfilm for years. If others had known of this work, perhaps somebody else might have been inspired to study this area.

Unearthing Security Innovations

So what might we look for in old security articles? Once upon a time, there was a lot of discussion about back doors into encrypted data and about key escrow systems to help the intelligence community—and these ideas are resurfacing. I can see the back door idea coming back, but not key escrow. Admittedly, the situation is much more confusing today, with national security issues sometimes tangled up with entertainment industry antipiracy crusades.

I had fun reading, for this article, a 1994 National Security Agency report (declassified in 2012) about the NSA's early activities.[8] The writer exults that almost none of the Eurocrypt 1992 papers were of any interest to the NSA efforts because the outside researchers seemed to have moved to theoretical topics, away from cryptanalysis or cryptosystem design. To me, it's strange for the NSA staff to be so dismissive of a thriving field of work, but it seems they were hoping that their ciphers wouldn't be at risk from outside efforts. And, in the spirit of looking for old ideas, that document reprints the recommendations for improving security that were made after the investigation of the Pearl Harbor attack. Many of these recommendations for improving security via better information exchange, more awareness of the need for alertness and proper management, and being aware of new risks would still be valid today.

At times, once we've solved a problem, we might want to look for other applications for the solution. For example, some early cybercrime articles discuss *phone phreaking* (the manipulation of the telephone system using tones that open a path to the switch controls to avoid being charged for calls and frustrating attempts to track calls—both of importance to bookies), which has been solved by moving telephone calls' metadata outside the voice path accessible to the sender. Why can't we do the same with email? It's been proposed many times, but its problems are administrative rather than technical, such as determining who would write its rules or validate the email systems. But I'm getting very tired of the daily phishing emails claiming to be from my Rutgers email service but easily observed to be from Africa or Asia.

Proving systems correct and building sandboxes were also popular topics at one time. Should these be resurrected? Provably correct code would be of use for not only security against attacks but also reliability in the face of accidents. Sandboxes could minimize the impact of flaws or break-ins. What about micropayments? Denial-of-service attacks rely on flooding a computer with a vast number of messages; if the bad guys had to pay for the messages, the attacks would be much harder to perform.

At one time, attackers engaged in *war dialing*—calling random numbers hoping to find modems

This article originally appeared in
IEEE Security & Privacy, *vol. 13, no. 4, 2015.*

attached to unprotected dial-in systems. That isn't an issue today, but random probing of computers on Wi-Fi networks is. There are now appropriate recommendations for dealing with this threat, but attackers have been placing receivers on roadsides to grab information from the electronic systems of passing cars. For example, Ishtiaq Rouf and his colleagues demonstrated how to use the RF signals from tire pressure monitoring to either track cars or falsely turn on a dashboard warning light.[9] Defenses to such attacks will be similar to what we already know. Rouf and his colleagues suggest several: software checking, encryption, and packet sequencing.

A common theme decades ago, and still important, is training: Do we educate enough people in cybersecurity defense? Similarly, management issues were important then and are now. As organizations decide to share information, they might find that their systems' least-defended part offers access to everything. Before 9/11, the US government partitioned the information on terrorist threats among its intelligence agencies; criticism of that approach led to greater information sharing, which then enabled a relatively low-level employee (Bradley Manning, now Chelsea Manning) to have access to, and betray, a wide range of information. Edward Snowden, by contrast, seems to have gained wider access via "social engineering" (that is, schmoozing); see previous sentences about management and training issues.

How might we enjoy the advantages of big data with fewer risks? Research in this area includes performing distributed database queries using limited access to each part of the database (viewing the data only through a specified kind of query—a sort of narrow hole into the data). Topics such as performing queries on encrypted data are regaining popularity.

A higher-level question is whether people would trade features for security. How many people using the Adobe PDF viewer know that it includes a 3D viewer? Would they give up this program for a simpler one with fewer vulnerabilities? In this case, users have a choice, because there are other PDF viewers. Operating system users' options are scarce. In addition, all users are at a disadvantage because evaluating security is difficult and there are no published ratings for the security of systems.

As with other ideas, new thoughts on improving security might require an appropriate technical environment, effective productization, suitable marketing, and good luck but might face the obstacle of confidentiality. Security suggestions are often made and tested in secret, and data regarding their effectiveness is often hard to come by. This is understandable; we don't want to publicize techniques to help hackers. But it also means that ideas aren't explored by a wider community and might be overlooked. Digital time stamping as a way to certify transaction dates, although not kept secret, was given a second life thanks to Bitcoin. Solutions to other current security problems might be hiding in old technologies; let's hope we can find some. ■

References

1. S. Ember, "Data Security Is Becoming the Sparkle in Bitcoin," *New York Times*, 2 Mar. 2015, p. B1.
2. S. Haber and W.S. Stornetta, "How to Time-Stamp a Digital Document," *Advances in Cryptology*, A.J. Menezes and S.A. Vanstone, eds., LNCS 537, Springer, 1991, pp. 437–455.
3. J. Markoff, "Technology: Experimenting with an Unbreachable Electronic Cipher," *New York Times*, 12 Jan. 1992.
4. *New York Times*, 16 Feb. 1992, p. L49.
5. S. Nakamoto, *Bitcoin: A Peer-to-Peer Electronic Cash System*, white paper, Bitcoin, 2008.
6. W. Weaver, Memorandum, Rockefeller Foundation, 1949; www.mt-archive.info/Weaver-1949.pdf.
7. V. Bush, "As We May Think," *Atlantic Monthly*, vol. 176, no. 7, 1945, pp. 101–108.
8. *Cryptolog*, National Security Agency, vol. 20, no. 1, 1994; https://www.nsa.gov/public_info/_files/cryptologs/cryptolog_126.pdf.
9. I. Rouf et al., "Security and Privacy Vulnerabilities of In-Car Wireless Networks: A Tire Pressure Monitoring System Case Study," *Proc. 19th USENIX Security Symp.*, 2010.

Michael Lesk is a professor at Rutgers University. Contact him at lesk@acm.org.

Editor: **Michiel van Genuchten**
VitalHealth Software
genuchten@ieee.org

Editor: **Les Hatton**
Oakwood Computing Associates
lesh@oakcomp.co.uk

On the Impact of Being Open

Robert Schuwer, Michiel van Genuchten, and Les Hatton

THERE'S MUCH DISCUSSION about being open, with topics such as open source software, open innovation, open research, and open education. Will the whole world be open, and, if so, what was all closed in the past? Many people credit software with starting the open movement in 1983 with Richard Stallman and the GNU Project. Others credit Linus Torvalds with starting the movement when he put the first version of Linux online in 1991. Here, we analyze the similarities and differences between the open movements we've been part of and come up with expectations for software's future.

Open Source Software

Software often starts as embedded software: it works only on a specific device and isn't separately charged. Next, it appears on the invoice as proprietary software but works only on the same company's hardware. IBM was the first company to enter this stage when it began selling software separately from its hardware in 1969. As software use expands, companies can't afford to develop all the required software themselves. This often leads to an industry or open standard, such as MS-DOS and Windows. The computer industry entered this stage when Compaq launched its PC in 1983, enabling startups such as Microsoft to sell their software on hardware from multiple vendors.

The next stage in software proliferation is open source.[1] Here, software is developed only once and then shared with the world. Some open source software is created by volunteers. This alternative is also attractive for companies that face increasing software development costs and aren't in the software business. Open source has changed software development in many industries. Whereas an engineer once might have started a project with requirements or an overall design, the first step now might be to look for open source components that do more or less what the requirements prescribe. Of course, the open source code's licensing conditions must be fully understood.

Today, many open source products are among the market leaders in their field, both visible and invisible to users. Examples include Firefox, Android, Linux, and Apache. *IEEE Software*'s Impact department has described four open source products: RealPlayer,[2] YAWL (Yet Another Workflow Language),[3] Bayesian networks,[4] and EYE.[5]

Open Education

In 2001, MIT announced its intention to make all its learning materials freely available through the Internet. A 2002 UNESCO conference in Paris coined such digital learning materials Open Educational Resources (OER). This development was also one of the drivers of Creative Commons—an organization that defines and maintains an open license framework for all creative expressions. According to Creative Commons,

2376-113X/15/$31.00 © 2015 IEEE Published by the IEEE Computer Society

the number of works published worldwide under that license has grown from 50 million in 2006 to 882 million in 2014.[6] Exactly how many of them are learning materials is unknown, but the number exceeds many millions.

In 2011, Sebastian Thrun and Peter Norvig made their residential Stanford course on artificial intelligence available for participants outside Stanford. This attracted 160,000 participants, of which 23,000 finished the course, earning a certificate of completion. This marked the start of the massive open online course (MOOC) movement. A MOOC provides a complete learning experience: content delivered in chunks by short videos, quizzes with immediate automated feedback, and an online examination. With an adequate performance, the learner earns a certificate. There are an estimated 3,000 MOOCs worldwide, often offered by renowned universities (see www.class-central.com). Companies such as SAP (openSAP) and organizations such as the International Olympic Committee also offer MOOCs.

David Wiley characterized the "Open" in OER as freely available and permitting the user "5R" rights: the right to retain (make a local copy), reuse (as is), revise (alter the learning materials), remix (mix the learning materials with other components), and redistribute.[7] Many MOOCs provide only free availability and keep the learning materials closed. However, learning materials' adaptability is important to make them fit the local educational context.

OER ultimately aims to provide higher-quality education. Its effect can be direct—for example, by making more digital learning materials accessible and creating more opportunities for new forms of pedagogy. Other direct effects are OER's improvement by peers and the opportunity to adapt OER to the specific educational context. An important indirect positive effect can be that teachers get inspired just by browsing OER materials. Another well-known effect is that the OER developers might pay more attention to quality because the materials are openly published and are reviewed by peers. Studies have verified this claim of higher quality.[8–10]

Open Research

Unquestionably, open source as a development model has revolutionized software development. It has become a true cooperative movement producing software often of the highest quality and without which the Internet simply wouldn't run in the same way. Perhaps lesser known but no less important has been its impact on scientific research.

Computation dominates the vast majority of scientific research. However, a fundamental principle of the scientific method, as promoted by the philosopher Karl Popper and others, is that the results must be independently repeatable. Results that aren't repeatable are, to put it bluntly, useless. For example, consider the infamous 1989 Fleischmann–Pons experiment reporting cold fusion.

The advent of large-scale computation has greatly complicated this simple but profound principle. In essence, as Darrel Ince and his colleagues discussed, the description of an algorithm is simply insufficient.[11] Instead, researchers must be able to access the complete means to reproduce the computational results. This includes the source code the original researchers produced and the source code of the support software they used, such as Perl or R, as well as the means to build it. (Even a different compiler switch can dramatically affect numerical accuracy in certain circumstances.)

With the advent of GNU/Linux and the many support systems written by enthusiastic volunteers worldwide, including statistics packages such as R, it's now possible to reproduce computational results. This will likely pave the way for the next advances in scientific research, as we learn which of the many results produced are sufficiently reproducible as to be relied upon. However, until this practice becomes mainstream, many scientific studies' computational results will likely be contaminated with unquantifiable errors. They simply aren't scientific in the classic sense of independent reproducibility.

Another important trend is the rise of open access journals. It started as an attempt to break the stronghold of some scientific publishers that had scientists writing for them for free while charging increasing subscription fees. With the rise of the Internet, it became possible to collect, review, and distribute papers digitally at much lower cost. Research has indicated that publishing in open access journals leads to more citations, which is one measure of quality.[12–14]

Will Being Open Take over the World?

We believe that, in the coming years, being open will take a larger place in the three fields we just discussed. The fact that we can now copy content for free and distribute it around the world for free allows the best content to travel to more consumers. Distribution has changed, and creation is changing as we speak. The worldwide distribution of software,

learning materials, and research products can be instantaneous and without cost.

However, this process still raises many economic and legal questions. The business models of being open and being free are unclear in many cases, and different license models don't allow the mixing and matching that some creators (such as software engineers and lecturers) would like. And, of course, software companies, educational institutions, and scientific publishers will do what they can to preserve their future. Some embrace the open future; others are trying to slow it down for as long as they can.

Sometimes valid reasons exist to keep things closed. For example, a company or institution needs to preserve and guarantee its product or architecture. A software company might have to guarantee that safety-critical software will work according to specifications and be willing to stand by its product in case of failure and any subsequent liability. A university will likely want to guarantee a consistent curriculum and offer a strong community of teachers and alumni. Some research journals have taken a century to achieve their fame and generate more citations than many open access journals combined.

A lesser-known effect is that being open will lead to more transparency, more competition at a larger scale, and less need for creation. We need great software, learning materials, and papers, but probably not that many because the winners will likely take it all. The main differences will probably be on the consumption side. Using software and software components shouldn't require much support. That is, users in many fields won't tolerate or even use software that requires them to take a course or read a manual.

In education, using open learning materials might remain less universal. A considerable amount of materials in fields such as law or history are still specifically national or local. Also, it helps to have a teacher with whom you can discuss course content and who can quiz you to assess what you understand and explain it again with additional examples, if necessary. So, open learning materials might impede the local learning process and not be accepted by teachers if they can't adapt them. Also, most educational institutions don't have to make a profit, so the need for global efficiencies might be less.

Nevertheless, some argue that being open will lead to better quality. Two mechanisms that help this process are reviews and the many users who help find problems early on. The evidence is being built continually, and although it's not quantifiable yet in all areas, being open has a strong argument in its favor if it's not only free but also better.

References

1. M. van Genuchten, "The Impact of Software Growth on the Electronics Industry," *Computer*, vol. 40, no. 1, 2007, pp. 106–108.
2. L. Bouchard, "Multimedia Software for Mobile Phones," *IEEE Software*, vol. 27, no. 3, 2010, pp. 8–10.
3. M. Adams, A.H.M. ter Hofstede, and M. La Rosa, "Open Source Software for Workflow Management: The Case of YAWL," *IEEE Software*, vol. 28, no. 3, 2011, pp. 16–19.
4. N.E. Fenton and M. Neil, "Decision Support Software for Probabilistic Risk Assessment Using Bayesian Networks," *IEEE Software*, vol. 31, no. 2, 2014, pp. 21–26.
5. R. Verborgh and J. De Roo, "Drawing Conclusions from Linked Data on the Web: The EYE Reasoner," *IEEE Software*, vol. 32, no. 3, 2015, pp. 13–17.
6. *State of the Commons*, Creative Commons; https://stateof.creativecommons.org/report.
7. D. Wiley, "Defining the 'Open' in Open Content," 2015; http://opencontent.org/definition.
8. J.L. Hilton III et al., "Cost-Savings Achieved in Two Semesters through the Adoption of Open Educational Resources," *Int'l Rev. Research in Open and Distance Learning*, vol. 15, no. 2, 2014, pp. 67–84.
9. T.J. Bliss et al., "An OER COUP: College Teacher and Student Perceptions of Open Educational Resources," *J. Interactive Media in Education*, 8 Feb. 2013; http://jime.open.ac.uk/articles/10.5334/2013-04.
10. B. de los Arcos et al., *Open Educational Resources (OER) Evidence Report 2013–2014*, OER Research Hub, 2014; http://oerresearchhub.org/about-2/reports.
11. D.C. Ince, L. Hatton, and J. Graham-Cumming, "The Case for Open Computer Programs," *Nature*, vol. 482, Feb. 2012, pp. 485–489.
12. Y. Gargouri et al., "Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research," *PLoS ONE*, vol. 5, no. 10, 2010; doi:10.1371/journal.pone.0013636.
13. J.M. Donovan, C.A. Watson, and C. Osborne, "The Open Access Advantage for American Law Reviews," to be published in *Edison: Law + Technology*; http://ssrn.com/abstract=2506913.
14. K. Wohlrabe and D. Birkmeier, "Do Open Access Articles in Economics Have a Citation Advantage?," Munich Personal RePEc Archive paper 56842, 2014; http://mpra.ub.uni-muenchen.de/56842/1/MPRA_paper_56842.pdf.

ROBERT SCHUWER is a lector (professor) in open educational resources at the Fontys University of Applied Sciences. Contact him at r.schuwer@fontys.nl.

MICHIEL VAN GENUCHTEN is the chief operating officer at VitalHealth. Contact him at genuchten@ieee.org.

LES HATTON is emeritus professor of forensic software engineering at Kingston University and the managing director of Oakwood Computing Associates. Contact him at lesh@oakcomp.co.uk.

This article originally appeared in IEEE Software, *vol. 32, no. 5, 2015.*

CLOUDERA, INC. is recruiting for our Palo Alto, CA office: Software Engineer: design & implement large distributed systems that scale well – to petabytes of data & 10s of 1000s of nodes. Mail resume w/job code #35905 to: Cloudera, Attn.: HR, 1001 Page Mill Rd., Bldg. 2, Palo Alto, CA 94304.

QATAR UNIVERSITY invites applications for research faculty positions at all levels. Candidates will cultivate and lead research projects at the KINDI Center for Computing Research in the area of Cyber Security. Qatar University offers competitive benefits package including a 3-year renewable contract, tax free salary, free furnished accommodation, and more. Apply by posting your application on the QU online recruitment system at careers.qu.edu.qa under "College of Engineering"

SENIOR SYSTEMS ANALYSTS Wesley Chapel/Tampa, FL area. Design & develop programs & systems. Work with MVS, z/OS, COBOL, DB2, VSAM, CLIST, REXX, JCL & HTML. Travel/reloc to various unanticipated locations as req'd. Send res to TechAffinity Consulting, Inc., 2255 Ashley Oaks Cir., Ste. 101, Wesley Chapel, FL 33544.

SOFTWARE TEST ENGINEER: system verification & validation testing; devel., maintain & implem. testing & QA methods. BS in CS or related + 2 yrs exp.; OR 4 yrs exp. Apply at http://prometric.submit4jobs.com/. Prometric Inc., 1501 S. Clinton St., Baltimore, MD 21224.

PURDUE UNIVERSITY. TENURE/TRACK FACULTY POSITIONS. The Department of Computer Science at Purdue University is in a phase of significant growth. Applications are solicited for seven tenure-track and tenured positions at the Assistant, Associate and Full Professor levels. Outstanding candidates in all areas of computer science will be considered. Review of applications and candidate interviews will begin early in October 2015, and will continue until the positions are filled. The Department of Computer Science offers a stimulating academic environment with research programs in most areas of computer science. Information about the department and a description of open positions are available athttp://www.cs.purdue.edu. Applicants should hold a PhD in Computer Science, or related discipline, be committed to excellence in teaching, and have demonstrated excellence in research. Successful candidates will be expected to conduct research in their fields of expertise, teach courses in computer science, and participate in other department and university activities. Salary and benefits are competitive, and Purdue is a dual career friendly employer. Applicants are strongly encouraged to apply online at https://hiring.science.purdue.edu. Alternatively, hardcopy applications can be sent to: Faculty Search Chair, Department of Computer Science, 305 N. University Street, Purdue University, West Lafayette, IN 47907. A background check will be required for employment. Purdue University is an EEO/AA employer fully committed to achieving a diverse workforce. All individuals, including minorities, women, individuals with

University of Nevada, Reno

UNIVERSITY OF NEVADA, RENO. The CSE Department invites applications for four tenure-track faculty positions. Two positions are at assistant professor level: the first position is in the area of big data with emphasis on security and privacy, and the second position is in the area of high performance computing with emphasis on parallel and distributed computing. The third position, in the area of cybersecurity, is at associate professor level, and will fill the role of Technical Director of the newly established Cybersecurity Center (CSC) at UNR. The fourth position is at assistant, associate or full professor level, and is open to all research areas, with preference given to candidates with expertise in embedded systems (Internet of Things, cyber-physical systems, VLSI design), machine learning (deep learning, data analytics, bioinformatics), computer graphics and visualization. Applicants must have a Ph.D. in Computer Science or Computer Engineering by July 1, 2016. Candidates must be strongly committed to excellence in research and teaching and should demonstrate potential for developing robust externally funded research programs. The department has several faculty with NSF Career awards and leaders in statewide and multi-state multi-million dollar NSF awards. Our research is supported by NSF, DoD, DHS, NASA, Google, Microsoft, Ford and AT&T. The department's annual research expenditures have exceeded $2M in recent years, while FY15 funding exceeds $3M. We offer B.S., M.S., and Ph.D. degrees and have strong research and education programs in Intelligent Systems, Computer and Network Systems, Software Systems, and Games and Simulations. In the last five years, the College of Engineering has witnessed an unprecedented growth in student enrollment and number of faculty positions. The College is positioned to further enhance the growth of its students, faculty, staff, and facilities as well as its research productivity and its graduate and undergraduate programs. UNR, Nevada's land grant University, has nearly 21,000 students. Reno is a half-hour drive to beautiful Lake Tahoe, an excellent area for a wide range of outdoor activities. San Francisco is within a four-hour drive. EEO/AA Women, under-represented groups, individuals with disabilities, and veterans are encouraged to apply. Apply online at:

https://www.unrsearch.com/postings/19013 (for the big data position)

https://www.unrsearch.com/postings/18990 (for the high performance computing position)

https://www.unrsearch.com/postings/19015 (for the cybersecurity position)

https://www.unrsearch.com/postings/19004 (for the open position)

Review of applications will begin on January 5, 2016 and will continue until the search closes on February 15, 2016. Inquiries should be directed to Ms. Lisa Cody, lacody@unr.edu.

 Published by the IEEE Computer Society

disabilities, and protected veterans are encouraged to apply.

VLOCITY is seeking a Mobile Software Engineer in San Francisco, CA to Dev. Vlocity iOS mobile apps. Ref Job ID: 9NSMQG & send res. To T. Dilley at hiring@vlocity.com.

IT PROFESSIONALS: Established IT firm (Edison, NJ) has multiple openings for IT Project Managers, Project Leads-JDE, Technical Project Leads, Technical Team Leads, Sr. Consultant/Software Engineers & Software Developers. Proj. Manager requires Master's or equiv. in Engg (any), CS, IS or related & 12 mos' relevant indus exp. Proj. & Team Leads require Master's or equiv. in Engg (any), CS, Computer Applications or related &12 mos' relevant indus exp. Project Lead-JDE also requires exp. configuring & customizing JD Edwards modules. For Proj. Managers & Proj. & Team Leads, we also will accept a Bachelor's or equiv. in the fields stated & 5 yrs' progressively responsible & relevant indus exp. Sr. Consultant/Software Engineers & Software Developers require a Bachelor's or equiv. in Engg (any), CS, IS or related & 24 mos' relevant indus exp. All positions based out of Edison, NJ HQ & subject to relocation to various unanticipated locations throughout the U.S. Qualified applicants mail resumes to: HR Manager, SYSTIME Computer Corporation (dba KPIT), 379 Thornall Street, Edison, NJ 08837.

MPHASIS CORP has multi openings at various levels for the follow'g positions at its office in NY, NY & unanticipated client sites thr/o the US 1. Info. Sys. Anyst* - Ana. & provide sys req & spec. 2. SW Dvlper* - Design, dvlp & modify SW sys. 3. Sys. Architect Dvlper* - Dvlp IT architecture 4. Graphic UI Desgr* - Design UI & perform UAT 5. N/W Infra Eng* - Maintain & TRBL n/w, design, dvlp, install n/w infra appl. 6. Business Operation Anyst* - Ana bus process thru app of s/w sol. 7. IT Mgr* - Plan & manage the delivery of IT proj. 8. Enterprise Svc Engagem't Mgr* - E2E sale of IT svc/prod. 9. Eng Engagem't Mgr* - Manage & direct business integration of proj activities. 10. Mkt Dvlpt Mgr* - Promote IT svc/prod. & impl bus plans. Must have a Bachelor/equiv and prior rel. exp, Master/equiv, or Master/equiv and prior rel. exp. Edu/exp req vary depending on position level/type. *Lead positions in this occupation must have Master/equiv+2yr or Bach/equiv+5yr progressive exp. Travel/relo req. Send resume & applied position to: recruitmentus@mphasis.com or 460 Park Ave. S., Ste# 1101, New York, NY 10016 Attn: Recruit.

SITECORE is seeking a Solution Engineer in Suas, CA to perform network modeling and analysis for company's CMS and module software products. Ref Job ID: 9ZTUYR & send res. J. Pillion at 2320 Marinship Way, Sausalito, CA 94965.

Fordham University
CIS Department
Tenure-track Assistant Professors

Fordham University invites applications for two tenure track Assistant Professor Positions in the CIS Department, to start in fall 2016. The two positions require a Ph.D. in Computer Science, Information Science or related fields, a commitment to teaching excellence, and an active program of research. One of the positions is in Cybersecurity and the other in Data Analytics.

These selected candidates are expected to teach graduate and undergraduate courses in Computer and Information Science, and conduct high-quality research.

For information about the department, visit http://www.cis.fordham.edu.

Applications can be electronically submitted to Interfolio Scholar Services:

For Cybersecurity Position: apply.interfolio.com/31854

For Data Analytics Position: apply.interfolio.com/31855

Include(1) Cover letter with qualifications,(2) Curriculum vitae,(3) Research Statement,(4) Teaching Statement,(5) Sample scholarship, and(6) At least three letters of recommendation. Applications will be accepted until the position is filled. Preference will be given to applications received by January 15, 2016.

For inquiries, contact: Palma Hutter at: hutter@fordham.edu.

Fordham University, an independent, Catholic University in the Jesuit tradition, is committed to excellence through diversity and welcomes candidates of all backgrounds. Fordham is an Equal Opportunity Employer.

NUS National University of Singapore

Multiple Tenure-Track or Tenured Faculty Positions in Computer Science

The Department of Computer Science at National University of Singapore (NUS) invites applications for several tenure-track or tenured faculty positions. We have positions dedicated to cyber security or big data analytics as well as positions open to all areas of computer science. While our main focus is on the assistant professor level, we also welcome exceptional candidates at the associate and full professor levels. For applications at the assistant professor level, candidates should demonstrate excellent research potential and a strong commitment to teaching. Candidates at more senior levels should have an established record of outstanding research achievements.

The Department of Computer Science at NUS is highly ranked internationally. It enjoys ample research funding, moderate teaching load, excellent facilities, and extensive international collaborations. The department covers all major research areas in computer science and boasts a thriving PhD program that attracts the brightest students from the region and beyond. More information is available at http://www.comp.nus.edu.sg/.

NUS offers highly competitive salaries and is situated in Singapore, an English-speaking cosmopolitan city and a meeting point of many cultures, both the east and the west. Singapore offers high-quality education, healthcare, and extremely low tax rates.

Interested candidates are invited to send, via electronic submission, the following materials to the Chair of the CS Faculty Search Committee, Prof. David Hsu, at csrec@comp.nus.edu.sg:

- A cover letter that clearly indicates main research interests
- Curriculum Vitae
- A teaching statement
- A research statement

Please arrange for

- at least 3 references

to be sent directly to the same e-mail address or provide the contact information.

Application review will commence on October 1, 2015 and continue until the positions are filled. To ensure maximal consideration, please submit your application by December 15, 2015.

JUNIPER NETWORKS

Juniper Networks is recruiting for our Sunnyvale, CA office:

Engineering Program Manager #30665: Lead cross functional teams in defining and managing project scope, delivery schedule, quality plan and risk management plan for complex software and hardware programs.

Software Engineer #26078: Develop and support ASIC drivers and toolkits; develop and support distributed networking operating system; and diagnose and provide solutions for product defects.

Software Engineer #26091: Design, develop, troubleshoot and debug software for enhancements and new products using Object Oriented/Functional Programming paradigms. Develop software in C, C++, Java and Python in a Linux environment.

Test Engineer #28997: Design, develop and implement testing methods and troubleshoot systems and equipment during all phases of product development. Develop comprehensive test plans based on changing and challenging product definitions, scaling targets and use case scenarios.

Software Engineer #10750: Design and develop fabric and interface software and drivers for next generation edge routers, including the development of device drivers for Company's proprietary Fabric ASICs and interface cards, toolkit for HSL2 interconnect, embedded software on FreeBSD and proprietary microkernel, diagnostic software to support field debugging and manufacturing.

QA Engineer #35064: Test NFV (Network Function Virtualization) functionality. Read and interpret standards documents, customer specifications, and functional specifications.

Software Engineer Staff #3986: Design, develop, troubleshoot and debug device drivers and networking software programs for enhancements and new Application-Specific Integration Circuits (ASICs). Develop software and tools used to implement distributed infrastructure software and platform software on embedded systems.

Technical Support Engineer #22016: Provide high-quality hardware and software technical support for switching products via email and telephone communication. Take ownership of high priority or sensitive customer issues and ensure prompt service restoration and resolution.

Technical Support Engineer #36483: Support, troubleshoot network impacting issues on company routing products (Internet backbone routers) to large Internet Service Providers and/or enterprise customers and deliver in-depth diagnostics and root-cause analysis. Understand customer network architecture, design, and layout and design and implement focused troubleshooting relevant to the network.

Technical Support Engineer #35717: Provide technical support to field engineers, technicians, product support personnel and other Technical Assistance Center (TAC) groups within the organization who are diagnosing, troubleshooting, repairing and debugging complex networked and/or wireless systems. Support Secured Routing products, working directly with our customers and partners when first-line product support has failed to isolate or fix problems in equipment or software.

Juniper Networks is recruiting for our Westford, MA office:

Technical Support Engineer #29712: Provide technical support to customers by resolving technical and non-technical problems related to router, protocols and network design.

Technical Support Engineer #35717: Provide technical support to field engineers, technicians, product support personnel and other Technical Assistance Center (TAC) groups within the organization who are diagnosing, troubleshooting, repairing and debugging complex networked and/or wireless systems. Support Secured Routing products, working directly with our customers and partners when first-line product support has failed to isolate or fix problems in equipment or software.

Mail single-sided resume with job code # to
Juniper Networks
Attn: MS A.8.429A
1133 Innovation Way
Sunnyvale, CA 94089

TU WIEN FAKULTÄT FÜR !NFORMATIK Faculty of Informatics

Announcement of an open position at the Faculty of Informatics, Vienna University of Technology, Austria

PROFESSOR of SECURITY

The Vienna University of Technology (TU Wien) invites applications for a professor position at the Faculty of Informatics. Excellent junior scientists are explicitly encouraged to apply.

The successful candidate will have an outstanding research record in the field of computer security. He/She should have a comprehensive research agenda, preferably spanning multiple subareas of the field. We expect the successful candidate to establish links to several of the main research areas of the Faculty of Informatics. Experience in raising funds and in managing scientific research projects is highly appreciated.

We offer excellent working conditions in an attractive research environment in a city with an exceptional quality of life.

For a more detailed announcement and information on how to apply, please go to:
www.informatik.tuwien.ac.at/vacancies

Application deadline: **December 31, 2015**

Université catholique de Louvain Full-time Academic Position in Computing Science

The ICTEAM Institute of the Université catholique de Louvain is now accepting applications for one tenure-track or tenured full-time position in Computing Science beginning September 2016. The Institute of Information and Communication Technologies, Electronics, and Applied Mathematics (ICTEAM) has more than 40 professors and 200 researchers, of which one fourth work primarily in computing science. The successful candidate will carry out research in areas relating to the general fields of data-intensive computing systems, software engineering, artificial intelligence, programming systems, software security, and bioinformatics. Other areas of competence will also be considered, since qualifications take precedence over specialization. The successful candidate will participate in undergraduate and graduate teaching within the curricula in Computing Science organized by the Louvain School of Engineering (in French and English). A Ph.D. in a computing-related field is required. Good knowledge of both spoken and written English and French is required, immediately or learned within two years. To apply, go to https://www.uclouvain.be/503094.html, click on "Accès public externe" and request the opening with Requisition Number 4611. The deadline for submitting an application is Nov. 23, 2015.

SYSTEMS PRODUCT CONSULTANT: Analyze & report on customer data for the sales of Tableau's visual analytics sw. Req Bach or foreign equiv in Comp Eng, Comp Sci, or rtd field, & 2 yrs exp in: consult w/ customers, define customer needs & reqmts, & create rpts & dashboards using analytical tools, incl ETL & Excel; design, devp & deploy Bus Intell & Visual Analytics sols util HTML, JavaScript, server-side scripting, & SQL; perform data analytics util RDBMS, incl SQL Server, MS ODBC & PowerPivot; extract & analyze large vols of complex data through data mining, machine learning, & inform retrieval techniques; & perform deep data driven analysis, stat analysis, metrics, predictive modeling, & data mining. Position at Tableau Software, Inc. in Seattle, WA. To apply, please e-mail resume to jobstableau@tableau.com.

THE OHIO STATE UNIVERSITY . The Computer Science and Engineering Department at The Ohio State University seeks to fill multiple tenure-track positions at the assistant professor level. We are particularly interested in recruiting in the following areas: cybersecurity, machine learning, distributed systems & cloud computing, and data management. The department is committed to enhancing faculty diversity; women, minorities, and individuals with disabilities are especially encouraged to apply. Some of these positions are partially funded by the university-wide Discovery Themes Initiative, a significant investment in key thematic areas, including the Data Analytics Collaborative which will establish a singular presence in data analytics at Ohio State. The university is also responsive to dual-career families and strongly promotes work-life balance through a suite of institutionalized policies. Applicants should hold or be completing a PhD in computer science & engineering or a closely related field, have a commitment to and demonstrated record of excellence in research, and a commitment to excellence in teaching. To apply, please submit your application via the online database. The link can be found at: https://web.cse.ohio-state.edu/cgi-bin/portal/fsearch/apply.cgi Review of applications will begin in December and will continue until the positions are filled. The Ohio State University is an Equal Opportunity/Affirmative Action Employer.

San Diego State University

Department of Computer Science

Chair of Computer Science

Department of Computer Science at SDSU seeks candidates for the Chair position with a PhD in Computer Science or a closely related field, and a sustained record of supported research. The Department is a dynamic and growing unit looking for a visionary Chair to lead it into its next phase of expansion.

We strive to build and sustain a welcoming environment for all. SDSU is seeking applicants with commitment to working effectively with individuals from diverse backgrounds and members of underrepresented groups.

For more details and application procedures, please apply via http://apply.interfolio.com/31841.

SDSU is a Title IX, equal opportunity employer. A full version of this ad can be found at: http://cs.sdsu.edu/

Cisco Systems, Inc. is accepting resumes for the following positions:

AUSTIN, TX: Program Manager (Ref#: AUS102): Coordinate and develop large engineering programs from concept to delivery. Telecommuting permitted.

BEAVERTON, OR: Software Engineer (Ref#: BEA1): Responsible for the definition, design, development, test, debugging, release, enhancement or maintenance of networking software.

BENTONVILLE, AR: Network Consulting Engineer (Ref#: BEN2): Responsible for the support and delivery of Advanced Services to company's major accounts.

BOXBOROUGH, MA: Technical Marketing Engineer (Ref.#: BOX18): Responsible for enlarging company's market and increasing revenue by marketing, supporting, and promoting company's technology to customers.

COLUMBIA, MD: Software/QA Engineer (Ref.# COLU2): Debug software products through the use of systematic tests to develop, apply, and maintain quality standards for company products.

CHICAGO, IL: Software Engineer (Ref#: CHI6): Responsible for the definition, design, development, test, debugging, release, enhancement or maintenance of networking software. Telecommuting permitted.

DENVER, CO: Product Manager (Ref.# DEN4): Own the complete product lifecycle including product vision & strategy, segmentation, solution, pricing, roadmap, planning and launch activities for the Cisco Policy Suite product line.

CLOVIS, CA: Network Consulting Engineer (Ref.# FRES1): Responsible for the support and delivery of Advanced Services to company's major accounts. Telecommuting permitted and travel may be required to various unanticipated locations throughout the United States.

MOORESTOWN, NJ: Systems Engineer (Ref.# MOO3): Provide business-level guidance to the account team or operation on technology trends and competitive threats, both at a technical and business level.

NORWALK, CT: Systems Engineer (Ref#: NOR1): Provide business-level guidance to the account team or operation on technology trends and competitive threats, both at a technical and business level. Travel may be required to various unanticipated locations throughout the United States.

RESEARCH TRIANGLE PARK, NC: Customer Support Engineer (Ref.# RTP302): Responsible for providing technical support regarding the company's proprietary systems and software. Telecommuting permitted. IT Engineer (Ref.# RTP13): Responsible for development, support and implementation of major system functionality of company's proprietary networking products.

RICHARDSON, TX: Test Engineer (Ref.# RIC4): Build test equipment and test diagnostics for new products based on manufacturing designs. Product Manager (Ref.# RIC621): Create high level marketing strategies and concepts for company solutions for markets and segments worldwide.

ROSEMONT, IL: Software Engineer (Ref.# ROSE14): Responsible for the definition, design, development, test, debugging, release, enhancement or maintenance of networking software. Telecommuting permitted.

SAN FRANCISCO, CA: CNG Staff (Ref.# SF13): Provide quality technical support for our client and partner base. Diagnose problems and troubleshoot company product line, including wireless access points, security appliances and switches.

SAN JOSE/MILPITAS/SANTA CLARA, CA: IT Manager (Ref.# SJ218): Manage the development and deployment of IT code that is pertinent to the company's Payroll, Stock, Travel, Webex Social and Mobile applications globally, rendering platform service and offerings. Customer Support Engineer (Ref.# SJ3): Responsible for providing technical support regarding the company's proprietary systems and software. Information Security Engineer (Ref.# SJ54): Responsible for ensuring the security of Cisco Security Cloud Operations network systems. Systems Administrator (Ref.# SJ12): Provide systems design and management function for business and/or engineering computer systems.

PLEASE MAIL RESUMES WITH REFERENCE NUMBER TO CISCO SYSTEMS, INC., ATTN: M51H, 170 W. Tasman Drive, Mail Stop: SJC 5/1/4, San Jose, CA 95134. No phone calls please. Must be legally authorized to work in the U.S. without sponsorship. EOE.

www.cisco.com

SOFTWARE ENGINEER needed f/t for design and development of high-level software applications with programming in HTML/CSS, Java, C++, etc; database management including schema design and SQL programming; profiling and deployment of various clustering technologies;analysis, design and implementation of applications for X.509 PKI certificate lifecycle management; some application testing duties (for quality assurance purposes).Requires Bachelor degree in Comp Sci, Comp Eng, Mathematics or related or foreign equiv. and 2 yrs exp. Background check req. Send resume and professional references to Marnie Euteneuer, Information Security Corp, 1011 Lake Street, Ste 425, Oak Park, IL 60301.

ERICSSON INC. has openings in Plano, TX for the position of: BUSINESS OPERATIONS ANALYST to design tools supporting service deliveries and identify new business opportunities for Ericsson in the given service area. Job ID#: 15-TX-289. CUSTOMER SOLUTIONS SALES MANAGER to lead and grow Cloud business by focusing on creating solutions for telecom operators, enterprises and channels. Frequent domestic travel required. Job ID#: 15-TX-1397. BUSINESS CONSULTANT to analyze complex customer requirements, identify scope of services and propose a range of solutions. Up to 35% Domestic Travel required. Job ID#: 15-TX-1392. APPLICATIONS DEVELOPER to develop, from a product (or group of products) perspective and end-to-end perspective, the best cost effective design and the adequate technology evolution. Job ID#: 15-TX-1885. SOLUTIONS ARCHITECT to build, lead and grow the services business within North America focusing on the area of Network Design and Optimization. Job ID#: 15-TX-2132. To apply please mail resume to Ericsson Inc. 6300 Legacy Drive, R1-C12 Plano, TX 75024 and indicate appropriate Job ID.

ERICSSON INC. has openings in San Jose, CA for the position of: Principal Consultant to plan, architect, and analyze needs for consulting for the Software Defined Network (SDN) practice area. Up to 70% travel required. Telecommuting is available for this position from anywhere in the United States. Job ID: 15-CA-2568. Engineer – Hardware to drive architecture and design work with teams focused on next generation SSR hardware. Job ID: 15-TX-2675. To apply please mail resume to Ericsson Inc. 6300 Legacy Drive, R1-C12 Plano, TX 75024 and indicate appropriate Job ID.

ENGINEER - SOFTWARE. Ericsson Inc. has an opening for the position of Engineer – Software in Piscataway, NJ to design, develop, test, plan and coordinate multiple features for a product suite. To apply please mail resume to Ericsson Inc. 6300 Legacy Drive, R1-C12 Plano, TX 75024 and indicate applying for 15-TX-3326.

SR. SVC CNSLTNT (NY, NY & unanticip client sites thrght US) Integrte CA App Dev prods w/other CA prods. Collab w/CA Svcs Directors & Sales/ Pre-Sales teams to scope potential proj

Help build the next generation of systems behind Facebook's products.

Facebook, Inc. currently has the following openings in **Menlo Park, CA (various levels/types)**:

Research Scientist (2956J) Research, design, & develop new optimization algorithms & techniques to improve the efficiency and performance of Facebook's platforms. **Data Scientist (5483J)** Perform research on available data using appropriate statistical techniques, including methodology & knowledge from the social sciences & large-scale data analysis techniques. **Security Engineer (5742J)** Design & build novel solutions to internal security challenges. Provide a secure computing environment through Enterprise Endpoints, IT Deployments, & Corp Extensions. **Software Engineer (SWE1015J)** Help build the next generation of systems behind Facebook's products, create web &/or mobile applications that reach over one billion people, & build high volume servers to support our content. **Data Scientist, Analytics (4425J)** Apply your expertise in quantitative analysis, data mining, & the presentation of data to see beyond the numbers & understand how our users interact with our core products. **BI Engineer (5825J)** Design & develop creative & innovative Business Intelligence/Analytic solutions from the data coming from various custom systems & databases. **Operations Program Manager, Tooling (5118J)** Define & document the end-to-end business processes used today & identify where & how tooling solutions will aid with streamlining & scaling the process. Document end-to-end business process (workflows) as they are defined throughout the program. **Production Engineer (2596J)** Participate in the design, implementation & ongoing management of major site applications & subsystems. **Application Engineer (Oracle) (3548J)** Develop & maintain integrated, scalable, corporate applications. Build solutions using Oracle technologies. **Data Engineer (4057J)** Architect, build, & launch new data models that provide intuitive analytics, & design, build, & launch extremely efficient & reliable data pipelines to move data (both large & small amounts) to serve insights & reporting.

Facebook, Inc. currently has the following openings in **Seattle, WA (various levels/types)**:

Production Engineer (327J) Participate in the design, implementation & ongoing management of major site applications and subsystems

Mail resume to: Facebook, Inc. Attn: SB-GIM, 1 Hacker Way, Menlo Park, CA 94025. Must reference job title & job# shown above, when applying.

& sales. Design, assess communicate how CA Srvc Virtualization & App test solutions solve business needs. REQS: Bach deg or for equiv in CS, Math, Engg (any) or a rel field +4 yrs of exp in job &/or a rel occup. Must have exp w/ Provdng IT cnsltng svcs in client envrmnt; Archtctng, scoping, assessing & implmntng App Test products; Java prgmmng, archtctre & in-mssng techs; C, C++, Java, VB.net, VB Script, Java Script & DOS; Intgrte CA App Perf Mgmt solution w/ CA Svc Vrtlztn & CA App Test sols to enable the feedback loop for DevOps process; Assessing, scoping & commnctng how CA Svc Vrtlztn & App Test sols can solve bus needs; Freq travel to unanticip client sites req; Wrk fr home anywhere in the US. Send resume to: Althea Wilson, CA Technologies, One CA Plaza, Islandia, NY 11749, Refer to Requisition #115538.

SR. RESEARCH ENGG (Santa Clara, CA) Dsgn a systm & autmtn wrkflw that deploys, upgrades & removes solutns. Dply systms on multi infrstrctre provdrs & prvnt & monitor unauth access, misuse, mod, or denial of a comp ntwrk & ntwrk-accessible resources. Execute agile tstng methods. Dply & config MySQL & Oracle DBs. REQS: 5 yrs exp in job &/or a rel occup. Must have exp w/ Decmpsng existing tech procses & implmntng them as autmtn wrkflws; Capturing & anlyzng tech reqs during proj implmntn in svc provdr envrmnts; Condct perf & load tstng, analysis & tuning of large-scale systms; Dsgng & implmntng highly avlble systms at web scale & preparing/ implmntng disaster recovery plans; Dplyng, using, maintng & upgrading Vrtlztn solutns such as VMware, kvm, Linux containers & LDOMs; Enterprise distrib & dplymnt of Red Hat Enterprise Linux; Dplyng & maintng virtual apps in VMware; Implmntng Lightweight directory Access Protocols; Send resume to: Althea Wilson, CA Technologies, One CA Plaza, Islandia, NY 11749, Refer to Requisition #115563.

SR. WEB CONTENT DVLPR (Santa Clara, CA) Archtct solution dsgns & engage in proj planning & execution. Engage in version cntrl, & prfm coding stndrd eval. Rev code, prvde tech guidnce & rslve tech issues. Visualize data on websites. Engage in unit test case frmwrk dvlpmnt & prep unit test cases. REQS: 5 yr exp in job &/or a rel occup. Must have exp w/Front end UI dvlpmnt; crtng web apps usng Angular JS Frmwrk; CoffeeScript, HAML, JavaScript, jQuery; Google Chart/Map libraries & Yeoman; CSS ; Flash/Flex, MXML, ActionScript; HTML5; Java. Send resume to: Althea Wilson, CA Technologies, One CA Plaza, Islandia, NY 11749, Refer to Requisition #115570.

ASSISTANT VICE PRESIDENT-DATA SCIENCE. Position available in Amherst, MA. Design and analyze variably sourced data sets using quantitative methodologies, computational frameworks, and systems. Develop machine learning algorithms and probabilistic models, create prototype systems, visualizations, and web applications and use SQL and NoSQL systems. Design experiments and analyze data using mathematical modelling package R. Disseminate findings to non-technical audiences through various media. Apply: B. O'Brien, Massachusetts Mutual Life Insurance Company, 1295 State Street, Springfield, MA 01111; Please Reference Job ID: 708201900.

CLASSIFIED LINE AD SUBMISSION DETAILS: Rates are $425.00 per column inch ($640 minimum). Eight lines per column inch and average five typeset words per line. Send copy at least one month prior to publication date to: Debbie Sims, Classified Advertising, *Computer* Magazine, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720; (714) 816-2138; fax (714) 821-4010. Email: dsims@computer.org.

In order to conform to the Age Discrimination in Employment Act and to discourage age discrimination, *Computer* may reject any advertisement containing any of these phrases or similar ones: "...recent college grads...," "...1–4 years maximum experience...," "...up to 5 years experience," or "...10 years maximum experience." *Computer* reserves the right to append to any advertisement without specific notice to the advertiser. Experience ranges are suggested minimum requirements, not maximums. *Computer* assumes that since advertisers have been notified of this policy in advance, they agree that any experience requirements, whether stated as ranges or otherwise, will be construed by the reader as minimum requirements only. *Computer* encourages employers to offer salaries that are competitive, but occasionally a salary may be offered that is significantly below currently acceptable levels. In such cases the reader may wish to inquire of the employer whether extenuating circumstances apply.

Apple Inc. has the following job opportunities in Cupertino, CA:

IS&T Technical Project Lead (REQ# 9REU8G). Deliver complex, global & cross-functional SAP pricing related projects. Prfrm analysis of complex bus data processing problems.

Hardware Development Engineer (REQ#9F7TCN). Dsgn camera HW for mobile platforms. Integrate optical, elctrcl, & mech components into camera sys. Travel required 20%.

ASIC Design Engineer (REQ# 99V24H). Dev tools & methodologies for semiconductor dsgn. Dev, maintain & enhance Phys Dsgn Verification & extraction flows.

Software Engineer Applications (REQ#9MBQ46). Des, lead and dlvr prods to improve Apple Online store.

ASIC Design Engineer (REQ# 9G3QNE). Perf reg STA runs, check qlty of constraints, Netlists & othr input colltrl.

Senior Software Development Engineer (REQ#9BXSY8). Respon for design & develop of embedded sw for cell/WiFi/NFC/Secure Element.

Hardware Development Engineer (REQ#9J2VNM). Evaluate and optimize organic light-emitting diode (OLED) materials for OLED devices. Travel req'd 20%.

Software Engineer Applications (REQ# 9X83JU) Support NoSQL db infrastructure through site reliability engg.

Systems Design Engineer (REQ# 9EZ5B7) Dvlp & optmze RF automation sys used on Apple's newest pdcts including iPhones, iPods, iPads, & others.

Mechanical Design Engineer (REQ# 9FYQY3). Conceive, des, and prod new mech components &assemblies to supp dev activities.

Hardware Development Engineer (REQ#9NELJH). Perform des & analysis of electr machines & related HW from an electr and mech des perspective.

Hardware Development Engineer (REQ#9FM2J3). Des, dev, & launch the next-gen display & panel des tech's for Apple prod's. Travel req'd 20%

Engineering Project Lead (REQ# 9F52UK). Rspnsbl for sys test process dvlpmnts, including upgrades to prdct feature testing reqs. Travel required 25%.

Software Engineer Applications (REQ#9UQT43). Bld Apple's next gen Employee Systems platform & suite of products. Travel req'd 25%.

Electronic Design Engineer (REQ# 99HUFZ) Dvlp & qualify advanced lithium ion battery tech for portable power consumer apps. Travel Req. 35%.

Software Development Engineer (REQ#9LL3QT) Perform Power Consumption Testing for Mobile Devices for Cellular modems.

ASIC Design Engineer (REQ# 9CVRYL) Res for converge of time analysis for high speed CPU designs.

Software Engineer Applications (REQ#9DNQ2V) Dsgn & dvlp mobile & web apps.

Software Development Engineer (REQ#9CYV24). Resp for SW arch & dev of image processing SW for embedded camera & media frameworks.

Firmware Engineer (REQ#9LKTWB) Dsgn & dvlp diagnostic SW for HW validation & factory testing.

Systems Design Engineer (REQ# 9FN48M) Desgn, dvlp, & optimize RF automation syss used on Apple's newest prdcts including iPhone, iPods, & iPads. Travel req. 25%.

Software Engineer Applications (REQ#9UGSXF). Provide a Java based Service-Oriented Arch pltfrm for various apps to interact w/each other.

Software Engineer Applications (REQ#9LA38S) Des & dev build & deploy sys for world wide point of sales sys.

Software Development Engineer (REQ#9G5UWX). Design, develop, & sup tools & process pipelines for analysis, mgmt., edit, & release of geospatial data for Apple Maps.

ASIC Design Engineer (REQ# 9FLQLT) Res for perf tuning, correlation, & verif for low-power high-perf microproc sys.

Hardware Development Engineer (REQ#9G5V5L). Dsgn & dev HW sys. Work on prototype bring-up & debugg'g & functional verfctn/mfg support.

Software Development Engineer (REQ#9PAMNX). Write & maintain SW for ingesting, transforming, & enriching data sets. Work w/ cont integration sys and SW eng lifecycles.

ASIC Design Engineer (REQ# 9GC4GF). Resp for Apple's Sensor dev high level performance using the FEM sim to optimize package deform & guide the BOM selection, package geo optim & process optim.

Software Development Engineer (REQ#9F4QJ6). Design & dev adv algorithms for mainly statistical signal & image processing using C & assembler programming.

Product Design Engineer [Multiple Positions Open] (REQ#9T2U9Q). Dev & practice analytical methods based on multiphysics modeling & simulation.

Software Development Engineer (REQ#9V4U64). Design and dev SW for real-time road traffic estimation and prediction.

Web Developer (REQ#9MY32P). Design, dev, validate, & deploy web based sys for managing Apple Mfg bus. processes.

Software Engineer Applications (REQ#9U2SR4). Analyze prod inventory, user interactions, & prefs to organize & prioritize content in stores.

Software Development Engineer (REQ#9E3QCK). Respon for design & develop of sw for WebKit web browser engine.

Systems Design Engineer (REQ# 9BX2L8) Dev & optmz RF autmtn sys used on Apple's newest prods incl: iPhones, iPods, iPads & othrs. Travel Req'd: 25%.

ASIC Design Engineer (REQ# 9E3TDZ). Devise timng mthds to perf timng anlys, cnvrgnce, closure, & final sign-off on submicron prcss tech.

Hardware Development Manager (REQ#9GVQAR) Des & dev snsng sys for consmr elctroncs dvcs.

Hardware Development Engineer (REQ#9NV3N6) Resp for acous

measmnts & des fdback of telecomm sys, multichan loudspkr & micr sys.

Software Development Engineer (REQ#9FK2X2) Prov securty guid across Apple's prod line (iOS, iOS X, web, and Windows technlgs).

Software Engineer Applications (REQ#9VYP32) Perf build, release, and sys integ on Apple prods.

Software Engineer Applications (REQ#9MC2H3) Des & dev SW for retail store wrkflw & plan'g syst.

Software Development Engineer (REQ#9DCT2A) Resp for tstng iOS-based cell dvcs & ensuring baseband & telphy feats comply w indus specs & w Apple prptry reqs.

ASIC Design Engineer (REQ# 9GD2KB) Des, dev, & integ PCIe dgtl HW for SoC ASICs.

Software Development Engineer (REQ#9ENTAX) Des, dev, implem & debug natl lang prcssng SW & tools.

ASIC Design Engineer (REQ#9FN5B8) Dsgn and dev multimedia IP's and subsys in SoC ASICs.

Systems Design Engineer (REQ#9ERNN5) Dev & optim RF autom sys used on Apple's newest prods incl: iPhones, iPods, iPads & othr wrlss techs. Travel Req'd: 23%.

ASIC Design Engineer (REQ#9G2N99) Des & dev hi-spd mem systs by perf sig & pwr integrity anlys.

ASIC Design Engineer (REQ#9FE3BZ) Implem cmplx, hi-perf & low-pwr microprocessor (CPU) units using gate-level logic des, P&R, & HDL synth.

Hardware Development Engineer (REQ#9H5UJD) Resp for the des & dev of new Apple prods. Travel Req'd: 20%.

Software Engineer Applications (REQ#9F534A) Des, dev, implem, maintn & oper lrg-scl distrb sys.

Mechanical Quality Engineer (REQ#9MJPHE). Supp new Apple prdct launches from a mechncl, tooling, engr'g, & prdct devt lvl. Travel Req'd: 35%.

Software Engineer Applications (REQ#9U3S4T). Analyze, des, prgrm, debug, & mod SW enhncmnts &/or new products used in local, ntwrkd, or Internet-rltd comp programs. Travel req'd 15%.

Systems Design Engineer (REQ# 9FZP9W) Eval the latest iPad, iPhone & iPod HW and SW sys. Travel req. 30%.

ASIC Design Engineer (REQ# 9CR39H) Resp for the bttm-line pwr & clock implmntation goals including dvlping & driving solutions.

Software Development Engineer (REQ#9F4T2L). Resp for grphcs drvr des, dev, debug & dplymnt.

Software Development Engineer (REQ#9DNPKZ). Test cell telephony function for iOS devices. Travel Req'd: 30%.

Operations Engineering Technical Project Manager (REQ#9N4UY9). Work under dir of tech program mgr, act as primary tech lead for product & mfg site. Expedite closure of tech oprtnl issues. Travel required 35%.

ASIC Design Engineer (REQ# 9CVRWE). Dev construct method for low power & high speed CPU designs.

ASIC Design Engineer (REQ# 9DFQ2F). Des, dev, implem & valdte embed SW for Apple's SoC bring up & tst.

Systems Design Engineer (REQ# 9FM4JC). Create & exec det'ld tst plns for antenna pasv & OTA perf.

Software Development Engineer (REQ#9J3TK8) Dvlp groundbreaking tech for lrge scale systms, spoken lang, big data, & artificial intelligence.

Business Systems Analyst (REQ# 9QK28C). Design and dev. SAP Retail IT solu. for Apple Retail logistics.

IST Technical Project Specialist (REQ#9QRPU8) Assist to dvlp and implmnt supply-chain systms.

Hardware Development Engineer (REQ#9DLTEB). Serve as a key eng in process dev for optical coatings, advanced touch panels & display panels.

Hardware Development Engineer (REQ#9LCPMN). Design, dev & characterize cutting edge circuits for display tech. Travel req'd: 20%.

Operations Advanced Manufacturing Engineer (REQ#9J8U9V). Des assemble & measure process for Apple prdcts, includ prdcts within Portbls, Dsktps, Wirels & Watch families. Travel req 30%

Engineering Project Coordinator (REQ# 9FLRW5). Coord iPhone serv projs & New Prod Intro Readiness.

ASIC Design Engineer (REQ# 9HTP4W) Des, imple & exe tests to valid & debug HW.

Firmware Engineer (REQ#9JXUM3). Des & dev embdd diagnos SW for Apple prods incl: iPhone, iPad & AppleTV. Travel Req'd: 20%.

Software Development Engineer (REQ#9SZVPU). Dev the transit data syst & web apps for facilitating operational workflows around gathering transit data for Apple Maps.

Apple Inc. has the following job opportunities in Jersey City, NJ:

Software Development Engineer (REQ#9F52NW) Excte test plans for test cell mobile dvcs on the Cell Radio Access Net. Travel req. 20%.

Apple Inc. has the following job opportunities in Maiden, NC:

Software Development Engineer (REQ#9S9VKG). Prvde Sr. lvl spprt for Bckup & Recov infrstrctre.

Refer to Req# & mail resume to Apple Inc., ATTN: L.J. 1 Infinite Loop 104-1GM Cupertino, CA 95014.
Apple is an EOE/AA m/f/ disability/vets.

SPLUNK INC. has the following job opportunities in San Francisco, CA: Senior iOS Engineer [REQ#9N4RXK]. Create, expand, & maintain highly scalable & fault tolerant mobile comp distr to customers & deployed on 300+ mil mobile devices. Software Engineer [REQ#9MWTLG]. Dev automation infrastructure & test cases, & write Co. apps. Software Engineer [REQ#9FP3RA]. Build & enhance Co. QA framework & architecture. Refer to Req# & mail resume to Splunk Inc., ATTN: J. Aldax, 250 Brannan Street, San Francisco CA 94107. Individuals seeking employment at Splunk are considered without regards to race, religion, color, national origin, ancestry, sex, gender, gender identity, gender expression, sexual orientation, marital status, age, physical or mental disability or medical condition (except where physical fitness is a valid occupational qualification), genetic information, veteran status, or any other consideration made unlawful by federal, state or local laws. To review US DOL's EEO is The Law notice please visit: https://careers.jobvite.com/Splunk/EEO_poster.pdf. To review Splunk's EEO Policy Statement please visit: http://careers.jobvite.com/Careers/Splunk/EEO-Policy-Statement.pdf. Pursuant to the San Francisco Fair Chance Ordinance, we will consider for employment qualified applicants with arrest and conviction records.

MISSISSIPPI STATE UNIVERSITY. PROFESSOR AND HEAD. DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING. Applications and nominations are being sought for the Professor and Head of the Department of Computer Science and Engineering (www.cse.msstate.edu) at Mississippi State University. The Head is responsible for the overall administration of the department and this is a 12-month tenured position. The successful Head will provide: uVision and leadership for nationally recognized computing education and research programs uExceptional academic and administrative skills uA strong commitment to faculty recruitment and development uA strong commitment to promoting diversity. Applicants must have a Ph.D. in computer science, software engineering, computer engineering, or a closely related field. The successful candidate must have earned national recognition by a distinguished record of accomplishments in computer science education and research. Demonstrated administrative experience is desired, as is teaching experience at both the undergraduate and graduate levels. The successful candidate must qualify for the rank of professor. Applicants must apply online at www.jobs.msstate.edu (PARF#9306) by completing the Personal Data Information Form and submitting a cover letter outlining your experience and vision for this position, a curriculum vitae, and the names and contact information of at least three professional references. Screening of candidates will begin January 15, 2016 and will continue until the position is filled. MSU is an equal opportunity employer, and all qualified applicants will receive consideration for employment without regard to race, color, religion, ethnicity, sex (including pregnancy and gender identity), national origin, disability status, age, sexual orientation, genetic information, protected veteran status, or any other characteristic protected by law. We always welcome nominations and applications from women, members of any minority group, and others who share our passion for building a diverse community that reflects the diversity in our student population. Please direct any questions to Dr. Jonathan Pote, Search Committee Chair (662) 325-3280 or jpote@abe.msstate.edu).

BAYLOR UNIVERSITY

Faculty Position

The Electrical and Computer Engineering Department of Baylor University seeks faculty applicants for a tenured/tenure-track Faculty Position at any level. Any area of expertise will be considered but applicants in computer engineering will be given special consideration. Applicants for assistant professor must demonstrate potential for sustained, funded scholarship and excellent teaching; applicants for associate or full professor must present evidence of achievement in research and teaching commensurate with the desired rank. The ECE department offers B.S., M.S., M.E. and Ph.D. degrees and is rapidly expanding its faculty size. Facilities include the *Baylor Research and Innovation Collaborative* (BRIC), a newly-established research park minutes from the main campus.

Chartered in 1845 by the Republic of Texas, Baylor University is the oldest university in Texas. Baylor has an enrollment of over 15,000 students and is a member of the Big XII Conference. Baylor's mission is to educate men and women for worldwide leadership and service by integrating academic excellence and Christian commitment within a caring community. The department seeks to hire faculty with an active Christian faith; applicants are encouraged to read about Baylor's vision for the integration of faith and learning at www.baylor.edu/profuturis/.

Applications will be considered on a rolling basis until the January 1, 2015 deadline. Applications must include:

1. **a letter of interest that identifies the applicant's anticipated rank,**
2. **a complete CV,**
3. **a concise statement of teaching and research interests,**
4. **the names and contact information for at least four professional references.**

Additional information is available at www.ecs.baylor.edu. Send materials via email to Dr. Keith Schubert at keith_schubert@baylor.edu. Please combine all submitted material into a single pdf file.

Baylor University is affiliated with the Baptist General Convention of Texas. As an Affirmative Action/Equal Employment Opportunity employer, Baylor encourages candidates of the Christian faith who are minorities, women, veterans, and persons with disabilities to apply.

INTERNATIONAL BUSINESS ALLIANCE USA, INC. seeks applicants for the following position at its worksite in Austin, Texas. Software Developer, Technical e-Configurator [Job Code TX-101]: Provides software technical support to e-configurator application. Analyze e-configurator requirements to determine implementation feasibility within configurator architecture. Participate in design of application architecture and modeling framework changes to support new configurator requirements. Collaborates with external teams to finalize design of interfaces and/or data communication formats. Reviews cost analysis of new design and makes recommendations. Develops design architecture and/or modeling framework changes using Trilogy/Versata Configuration Modeling Language (CML). Provides some technical guidance to developers when implementing new configurator requirements or supporting configurator production release. Conducts design and code reviews with other application developers to prevent errors and facilitate knowledge transfer. Works as technical team member of configurator project infrastructure and process changes, including regression testing, automated build, code generation, source control, to improve quality and efficiency of development process. Requirements: Bachelorís Degree in Computer Science, Engineering or related field plus two years of experience as an IT Professional using Trilogy/Versata SalesBUILDER Engine and Configuration Modeling Language (CML) in design and development of software applications, CML Extensions Model Packs, and SalesBUILDER Engine API and JNI. EOE. Apply by mail with cover letter referencing Job Code TX-101, and resume to: IBAUSA, Inc, Attn: SZ, 1092 Wilmington Avenue, San Jose, CA 95129.

UNIVERSITY OF ALABAMA

Tenured/Tenure-Track Faculty Positions, Computer Science

Openings exist for two Assistant/Associate/Full professors in software engineering with specific application to data analytics or computational modeling starting in Fall 2016. Outstanding candidates in all areas will be considered. At the time of appointment, candidates must have earned a Ph.D. in Computer Science or a related field. Candidates will be expected to form collaborations with the new NOAA Water Center on UA's campus (http://nws.noaa.gov/oh/nwc/). The CS department has twenty-four faculty members (15 tenured/tenure track faculty, seven of whom have interests in software engineering), over 600 undergraduates in an ABET accredited B.S. degree program, and 40 graduate students. The department also offers a Software Engineering Concentration for its undergraduates.

For additional details and to apply, visit http://se.cs.ua.edu/facultyjobs or contact Dr. Jeffrey Carver (carver@cs.ua.edu). Review of applications will begin immediately. The University of Alabama is an equal opportunity/affirmative action employer. Women and minority applicants are particularly encouraged to apply.

FIU FLORIDA INTERNATIONAL UNIVERSITY

Florida International University is a comprehensive university offering 340 majors in 188 degree programs in 23 colleges and schools, with innovative bachelor's, master's and doctoral programs across all disciplines including medicine, public health, law, journalism, hospitality, and architecture. FIU is Carnegie-designated as both a research university with high research activity and a community-engaged university. Located in the heart of the dynamic south Florida urban region, our multiple campuses serve over 55,000 students, placing FIU among the ten largest universities in the nation. Our annual research expenditures in excess of $100 million and our deep commitment to engagement have made FIU the go-to solutions center for issues ranging from local to global. FIU leads the nation in granting bachelor's degrees, including in the STEM fields, to minority students and is first in awarding STEM master's degrees to Hispanics. Our students, faculty, and staff reflect Miami's diverse population, earning FIU the designation of Hispanic-Serving Institution. At FIU, we are proud to be 'Worlds Ahead'! For more information about FIU, visit *fiu.edu*.

The School of Computing and Information Sciences (SCIS) seeks exceptionally qualified candidates for tenure-track and tenured faculty positions at all levels as well as non-tenure track faculty positions at the level of Instructor, including visiting instructor appointments. SCIS is a rapidly growing program of excellence at the University, with 30 tenure-track faculty members and over 2,000 students, including over 80 Ph.D. students. SCIS offers B.S., M.S., and Ph.D. degrees in Computer Science, an M.S. degree in Telecommunications and Networking, an M.S. degree in Cybersecurity, and B.S., B.A., and M.S. degrees in Information Technology. SCIS has received over $22M in the last four years in external research funding, has six research centers/clusters with first-class computing and support infrastructure, and enjoys broad and dynamic industry and international partnerships.

Open-Rank Tenure Track/Tenured Positions (Job ID# 508676)

SCIS seeks exceptionally qualified candidates for tenure-track and tenured faculty positions at all levels. We seek well-qualified candidates in all areas; researchers in the areas of computer systems, cybersecurity, cognitive computing, data science, health informatics, and networking are particularly encouraged to apply. Preference will be given to candidates who will enhance or complement our existing research strengths.

Ideal candidates for junior positions should have a record of exceptional research in their early careers. Candidates for senior positions must have an active and proven record of excellence in funded research, publications, and professional service, as well as a demonstrated ability to develop and lead collaborative research projects. In addition to developing or expanding a high-quality research program, all successful applicants must be committed to excellence in teaching at both the graduate and undergraduate levels. An earned Ph.D. in Computer Science or related disciplines is required.

Non-tenure track instructor positions (Job Opening 507474)

We seek well-qualified candidates in all areas of Computer Science and Information Technology. Ideal candidates must be committed to excellence in teaching a variety of courses at the undergraduate level. A graduate degree in Computer Science or related disciplines is required; significant prior teaching and industry experience and/or a Ph.D. in Computer Science is preferred.

HOW TO APPLY:

Qualified candidates for open- rank faculty positions are encouraged to apply to (Job Opening ID #508676); and candidates for instructor positions are encouraged to apply to (Job Opening ID# 507474). Submit applications at *facultycareers.fiu.edu* and attach cover letter, curriculum vitae, statement of teaching philosophy, research statement, etc as *individual attachments*. Candidates will be required to provide names and contact information for at least three references who will be contacted *as determined by the search committee*. To receive full consideration, applications and required materials should be received by December 31st, 2015. Review will continue until position is filled.

If you are interested in a visiting appointment please contact the department directly by emailing Dr. Mark Weiss at *Weiss@cis.fiu.edu*. All other applicants should apply by going to *facultycareers.fiu.edu*.

FIU is a member of the State University System of Florida and an Equal Opportunity, Equal Access Affirmative Action Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, national origin, disability status, protected veteran status, or any other characteristic protected by law.

COMPUTERS: TECH MAHINDRA (AMERICAS) INC. is seeking to fill numerous IT positions. Prgrm Mgrs to oversee & manage mult. IT projects, proj. planning, dvlpmnt, implementation, acct & delivery mgmt; Proj. mgrs to oversee & manage IT teams w/dvlpmnt of various sftwre apps. Sys. Analyst/Programmers/ Quality & Tech Architect/ Tech Solutions Architect /Quality Engg /Test Eng/ Sftwre Eng. / sftwre Developer / Systems Admin/DBAs/ DB Architect to analyze, design, dvlp, test & maintain comp software apps or databases through all phases of sftwre dvlpmt life cycle (Sftwre Eng. may also lead a team on various projects); Telecom solutions Architect / Application Architects to use various technologies to dsgn & dvlop telecom software apps. Mech. Eng (CAD/CAM) to design, develop, validate & perform structural calculations, product improvement, & provide tech. support to design teams at high levels utilizing specific mech. tools. Sales Eng/Bus. Analyst/ Mgmt Analyst for solutions/pre-sales activities w/relev. industry experience. IT Bus. Dev. Mgrs to create new business, negotiate contracts & dvlp proposals for customized IT solutions. Rel. Mgrs to manage/outsource commercial IT/ Eng. deals, monitor & maintain existing accts. All Tech/Mgrial. positions require a MS or BS degree or equiv. in CS, Comp Apps, CIS, IT, Eng, Bus mgmt/admin or closely related fields and relevant industry exp. Sales/Rel.Mgrs. require a MS or BS degree or equiv. in Bus. Admin, Eng. or closely related field & relev. Industry exp. Positions are based out of corp. HQ in 4965 Preston Park Blvd, # 500, Plano, TX 75093 & subject to travel & relocation to client sites located throughout the U.S. Mail resume & position applied for w/JOB CODE: 11IE15 to Visa Cell, Tech Mahindra (Americas) Inc., 1001 Durham Avenue, Suite 101, South Plainfield NJ 07080.

SR. RELAY ENGINEERS in Beaumont, TX area. Partic in scoping/estimating PM&C Capital projects. Research & determine transmission line & substation relay schemes & settings for PM&C/Grid Capital/Blanket projects. Occas travel req. Send res to JBT Electric LLC, 8876 Winzer Rd, Beaumont TX 77705.

University of Illinois at Urbana-Champaign – Positions in Computing

The Department of Electrical and Computer Engineering (ECE) at the University of Illinois at Urbana-Champaign invites applications for faculty positions at all levels and in all areas in computing, broadly defined, with particular emphasis on big data and its applications, including data analytics; data center and storage systems; parallel, high-performance, and energy-efficient computing; reliable and secure computing; distributed computing; bio-inspired computing; verification; wired/wireless networking; social networking; mobile, wearable sensing & applications; and computational genomics. From the transistor and the first computer implementation based on von Neumann's architecture to the Blue Waters petascale computer – the fastest computer on any university campus – ECE ILLINOIS faculty have always been at the forefront of computing research and innovation. Applications are encouraged from candidates whose research programs specialize in core as well as interdisciplinary areas of electrical and computer engineering. The department is engaged in exciting new and expanding programs for research, education, and professional development, with strong ties to industry. The ECE Department has recently settled into its new 235,000 sq. ft. net-zero energy design building, which is a major campus addition with maximum space and minimal carbon footprint.

Qualified senior candidates may also be considered for tenured full Professor positions as part of the Grainger Engineering Breakthroughs Initiative (http://graingerinitiative.engineering.illinois. edu), which is backed by a $100-million gift from the Grainger Foundation to support research in big data and bioengineering, broadly defined. In addition, the University of Illinois is home to Blue Waters petascale computer, which is supported by the National Science Foundation and developed and operated by the University of Illinois' National Center for Supercomputing Applications. Qualified candidates may be hired as Blue Waters Professors who will be provided substantial allocations on and expedited access to the supercomputer. To be considered as a Blue Waters Professor, candidates need to mention Blue Waters as one of their preferred research areas in their online application, and include a reference to Blue Waters in their cover letter.

Please visit http://jobs.illinois.edu to view the complete position announcement and application instructions. Full consideration will be given to applications received by December 15, 2015, but applications will continue to be accepted until all positions are filled.

The University of Illinois conducts criminal background checks on all job candidates upon acceptance of a contingent offer.

Illinois is an EEO Employer/Vet/Disabled www.inclusiveillinois.illinois.edu.

EXPEDIA, INC. currently has openings for the following opportunities in our Bellevue, WA office (various/levels/types): • Software Engineers: (728.SWE-AUG) Design, implement, and debug software for computers including algorithms and data structures. • Database Developers: (728.DBD-AUG) Coordinate changes to computer databases, test and implement the database applying knowledge of database management systems. • Reporting Analysts: (728.1557) Formulate and apply mathematical modeling and other optimizing methods to develop and interpret information that assists management with decision making. • Oracle Test Analysts: (728.1399) Create and execute test cases, and report and track test execution and defect metrics using test and defect management tools. • Managers, Engineering: (728.585) Responsible for architecture, design, construction, testing, and implementation of software. • Web Analysts: (728.1383) Formulate and apply mathematical modeling and other optimizing methods to develop and interpret information. • BI Developers: (728.1571) Participate in sourcing, organizing, maintaining, and standardizing large volumes of data through development of innovative tools, reporting dashboards, and well-organized. • Data Scientists: (728.1418) Apply advanced analytic techniques such as machine learning, data mining, and statistical modeling to design and implement mathematical models and algorithms to solve marketing problems. • Directors, Technology: (728.299) Build and manage a geographically distributed technology through subordinate managers and direct reports who support technology needs of core brand or function. 10% Travel to various unanticipated sites throughout the United States and internationally required. Send your resume to: Expedia Recruiting, 333 108th Avenue NE, Bellevue, WA 98004. Must reference position & Job ID# listed above.

EXPEDIA, INC. currently has openings for the following opportunities in our San Francisco, CA office (various/levels/types): • Software Engineers: (728.SWE-AUG-SF) Design, implement, and debug software for computers including algorithms and data structures. Send your resume to: Expedia Recruiting, 333 108th Avenue NE, Bellevue, WA 98004. Must reference position & Job ID# listed above.

SR. SOFTWARE ENGINEERS in San Francisco, CA to design database models & algorithms that can process massive volumes of data. Resume to: Job #3, Blueshift Labs, Inc., 88 First Street, Suite 500, San Francisco, CA 94105.

Intuit Inc.

currently has openings for the following positions in **Santa Clara County**, including **Mountain View, California** or any office within normal commuting distance:

Senior Business Data Analysts (Job code: I-58): Write SQL queries to pull sales data using Site Catalyst. Analyze web channel performance. **Software Engineers (Job code: I-398):** Apply software development practices to design, implement, and support individual software projects. Work on problems of moderate scope and complexity where analysis of situations or data requires a review of multiple factors of the overall product and service. Up to 20% travel may be required to work on projects at various, unanticipated sites throughout the United States. **Managers-Marketing Ops (Job code: I-29):** Complete technical digital marketing operations support such as tag management. Design financial and billing management systems for online advertising. **Senior Technical Data Analysts (Job code: I-260):** Serve as the key analytics partner for our Digital Marketing and Direct Response teams, partnering with them to uncover insights on the effectiveness of our marketing campaigns for Mint and Quicken, and to drive recommendations on marketing campaigns and tests.

Positions located in **San Diego, California**:

Senior Development Managers (Job code: I-290): Owns the strategy, goals and plans for the team's success and contributes to their organization's strategy and plans. Effectively implements SDLC and modifies to fit needs of the specific project. **Application Operations Engineers (Job code: I-186):** Apply master level knowledge of technology and operational best practices to drive the design, development and implementation of operational standards and capabilities for connected services that enable highly available, scalable and reliable customer experiences. **Staff Software Engineers in Quality (Job code: I-330):** Partner with cross-functional leaders and team members to deliver Intuit products, with greater efficiency and speed. Test software, including creating test cases, test plans, test data, and defect write ups.

Positions located in **Plano, Texas**:

Software Engineers in Quality (Job code: I-437): Apply best software engineering practices to ensure quality of products and services by designing and implementing test strategies, test automation, and quality tools and processes.

To apply, submit resume to Intuit Inc., Attn: Olivia Sawyer, J203-6, 2800 E. Commerce Center Place, Tucson, AZ 85706. You must include the job code on your resume/cover letter. Intuit supports workforce diversity.

Linkedin Corp.

Linkedin Corp. has openings in our **Mtn View, CA** location for:

Software Engineer (All Levels/Types) (SWE1015MV) Design, develop & integrate cutting-edge software technologies; **Database Engineering Manager (6597.289)** Contribute at a senior-level to the data warehouse design & data preparation batch processes by implementing a solid, robust, extensible design that supports key business flows; **CRM Application Administrator (6597.1176)** Support hundreds of internal users & millions of external users via scalable tools & technology; **Manager**, **Software Engineering (6597.954)** Responsible for leading & building a team to build the online serving platform; **Senior Data Scientist (6597.1413)** Collaborate with Product Development teams to build data-fueled products (utilizing programming languages such as Java, Python, & C/C++) that track & analyze data related to LinkedIn products; **Test Engineer (6597.1079)** Write & build automated suites, continuously design creative ways to break software, & identify potential bugs; **Site Reliability Engineer (6597.973)** Serve as a primary point responsible for the overall health, performance, & capacity of one or more internet-facing services; **Data Scientist (6597.1243)** Partner with Product Management, Engineering, Design, Marketing, & Business Operations to drive product strategy & data-driven product decisions.

LinkedIn Corp. has openings in our **Sunnyvale, CA** location for:

Software Engineer (All Levels/Types) (SWE1015SV) Design, develop & integrate cutting-edge software technologies; **Salesforce.com Developer (6597.958)** Develop, enhance, debugs, support, analyze, maintain & test new/existing functionality which supports internal business units or supporting functions; **Site Reliability Engineer (6597.610)** Serve as a primary point responsible for the overall health, performance, & capacity of one or more internet-facing services.

LinkedIn Corp. has openings in our **San Francisco, CA** location for:

Software Engineer (All Levels/Types) (SWE1015SF) Design, develop & integrate cutting-edge software technologies.

LinkedIn Corp. has openings in our **New York, NY** location for:

Software Engineer (All Levels/Types) (SWE1015NY) Design, develop & integrate cutting-edge software technologies.

Please email resume to: 6597@linkedin.com. Must ref. job code above when applying.

Focus on Your Job Search

IEEE Computer Society Jobs helps you easily find a new job in IT, software development, computer engineering, research, programming, architecture, cloud computing, consulting, databases, and many other computer-related areas.

New feature: Find jobs recommending or requiring the IEEE CS CSDA or CSDP certifications!

Visit www.computer.org/jobs to search technical job openings, plus internships, from employers worldwide.

http://www.computer.org/jobs

IEEE computer society | JOBS

The IEEE Computer Society is a partner in the AIP Career Network, a collection of online job sites for scientists, engineers, and computing professionals. Other partners include Physics Today, the American Association of Physicists in Medicine (AAPM), American Association of Physics Teachers (AAPT), American Physical Society (APS), AVS Science and Technology, and the Society of Physics Students (SPS) and Sigma Pi Sigma.

Call for Papers

ECSEE
European Conference
Software Engineering Education 2016

30 June and 1 July 2016, Seeon Monastery, Germany

Important Dates

Full Paper
Submission Deadline:
15 March 2016

Notification to authors:
11 April 2016

Camera ready version
6 May 2016

Author Registration and
Payment Deadline:
31 May 2016

Conference dates:
30 June and 1 July 2016

Conference Website

www.ecsee.eu

General chairs
Georg Hagel
Jürgen Mottok

Co-chairs
Hans Gruber
Dieter Landes
Irmgard Schroll-Decker

International Program Committee

Organizer

LESE
Lehren von Software Engineering

Learning of Software Engineering - Registered Association

Scope

Software Engineering (SE) is an important discipline and core part of almost all Computer Science curricula of universities.

Challenges in today's software development include increasing system complexity, short development cycles, shorter time to market, continuous change, and expected high quality of the software.

Software engineering education has to deal with these challenges. How can students or employees be prepared to master these challenges?

What are best practices to help them to work in different domains, ranging from app development for mobile devices to the development of really big applications for mainframe systems, from game development to working on highly secure systems?

How can we support students in their student life-cycle and how can we prepare them for lifelong learning?

How can we ensure that future software engineers meet industrial needs, with respect to technical as well as soft skills?

Important

At least one author of an accepted paper must register and attend the conference to present and discuss the paper. The paper will not be included in the conference proceedings if it is not presented at the conference.

Topics of Interest include, but are not limited to

- SE curriculum design
- Training, education, and certification of SE in adult education
- New methods, techniques, best practices, and experiences in SE education
- Pedagogical underpinning of SE education
- Illustrative examples to highlight SE topics in education
- Evaluation and assessment of students' SE-relates skills
- Assessment of different teaching models in SE
- Reflective learning
- Tools for SE education
- Support of lifelong learning in SE
- Social and cultural issues in SE education
- Games and social media in SE education
- Distance learning, online learning, E-learning, MOOCs
- Analysis / mining of educational data in SE

Best Paper Award

A Best Paper Award is appointed in *Educational Methods and Learning Mechanisms in Software Engineering Education*.

Submission

High quality contributions are accepted in the following categories:

- Research papers
- Experience reports from industry or universities
- Panel session

Submission guidelines can be found on the submission page. Papers must be submitted electronically.

Partners

IEEE

IEEE Education Society Austrian and German Chapters

Gesellschaft für Informatik e.V.

AK Fachdidaktik Software Engineering

Working Group Didactics of Software Engineering

IGIP

International Society for Engineering Education

iJET international journal of emerging technologies in learning

EVELIN
EXPERIMENTELLE VERBESSERUNG DES LERNENS VON SOFTWARE ENGINEERING

SPONSORED BY THE

Federal Ministry of Education and Research

KINNERET
Achi Racov Engineering School

Contact: ECSEE 2016 - Erika Antoni, Seybothstraße 2, 93053 Regensburg, Germany
erika.antoni@oth-regensburg.de

IEEE Symposium on Low-Power and High-Speed Chips

COOL Chips XIX

Yokohama Joho Bunka Center, Yokohama, Japan
(Yokohama Media & Communications Center, Yokohama, Japan)

April 20 - 22, 2016

CALL FOR CONTRIBUTIONS

COOL Chips is an International Symposium initiated in 1998 to present advancement of low-power and high-speed chips. The symposium covers leading-edge technologies in all areas of microprocessors and their applications. The COOL Chips XIX is to be held in Yokohama on April 20-22, 2016, and is targeted at the architecture, design and implementation of chips with special emphasis on the areas listed below. All papers will be published online via IEEE Xplore. Authors of best papers will be recommended to submit an extended version to a COOL Chips special issue of IEEE Micro.

Contributions are solicited in the following areas:

- **Low Power-High Performance Processors for - Multimedia, Digital Consumer Electronics, Mobile, Graphics, Encryption, Robotics, Automotive, Networking, Medical, Healthcare, and Biometrics.**
- **Novel Architectures and Schemes for - Single Core, Multi/Many-Core, NoC, Embedded Systems, Reconfigurable Computing, Grid, Ubiquitous, Dependable Computing, GALS and 3D Integration**
- **Cool Software including - Parallel Schedulers, Embedded Real-time Operating System, Binary Translations, Compiler Issues and Low Power Techniques.**

Proposals should consist of a title, an extended abstract (up to 3 pages) describing the product or topic to be presented and the name, job title, address, phone number, FAX number, and e-mail address of the presenter. The status of the product or topic should precisely be described. If this is a not-yet-announced product, and you would like to keep the submission confidential, please indicate it. We will do our best to maintain confidentiality. Proposals will be selected by the program committee's evaluation of interest to the audience.

Submission should be made through website.

Details instruction are in author's kit obtained from **http://www.coolchips.org/**

Author Schedule:	**February 8, 2016**	**Extended Abstract Submission (through website)**
	March 15, 2016	**Acceptance Notified (by e-mail)**
	April 1, 2016	**Final Manuscript Submission**

You are also invited to submit proposals for poster sessions. Submission should be made through website. Details instruction are in author's kit obtained from **http://www.coolchips.org/**

Author Schedule:	**March 22, 2016**	**Poster Abstract Submission (through website)**
	March 29, 2016	**Poster Acceptance Notified (by e-mail)**

For more information, please visit <**http://www.coolchips.org/**>.

For any questions, please contact the Secretariat <**cool_xix@coolchips.org**>.

Sponsored by the Technical Committees on Microprocessors and Microcomputers and Computer Architecture of the IEEE Computer Society. (approval pending) In cooperation with the IEICE Electronics Society and IPSJ.

#CES2016

THE STORY OF INNOVATION IS CONTINUALLY WRITTEN AND REWRITTEN AT CES.
DON'T MISS THE CHANCE TO ADD YOUR CHAPTER. REGISTER TODAY.

CES® 2016 JAN. 6-9, 2016

TECH EAST • TECH WEST • TECH SOUTH
LAS VEGAS, NV

REGISTER NOW at **CESweb.org**

CES
THE GLOBAL STAGE FOR INNOVATION

PRODUCED BY CEA