

COMPUTING  
**e**edge

- **Big Data**
- **Careers**
- **Artificial Intelligence**
- **Cloud Computing**

NOVEMBER 2023

[www.computer.org](http://www.computer.org)



IEEE COMPUTER SOCIETY D&I FUND

# Drive Diversity & Inclusion in Computing



*Supporting projects  
and programs that  
positively impact  
diversity, equity, and  
inclusion throughout  
the computing  
community.*

**DONATE TODAY!**



**IEEE Foundation**

## STAFF

### Editor

Cathy Martin

### Periodicals Portfolio Senior Managers

Carrie Clark and Kimberly Sperka

### Director, Periodicals and Special Projects

Robin Baldwin

### Production & Design Artist

Carmen Flores-Garvey

### Periodicals Operations Project Specialist

Christine Shaughnessy

### Senior Advertising Coordinator

Debbie Sims

**Circulation:** *ComputingEdge* (ISSN 2469-7087) 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 address changes to *ComputingEdge*-IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Printed in USA.

**Editorial:** Unless otherwise stated, bylined articles, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *ComputingEdge* does not necessarily constitute endorsement by the IEEE or the Computer Society. All submissions are subject to editing for style, clarity, and space.

**Reuse Rights and Reprint Permissions:** Educational or personal use of this material is permitted without fee, provided such use: 1) is not made for profit; 2) includes this notice and a full citation to the original work on the first page of the copy; and 3) does not imply IEEE endorsement of any third-party products or services. Authors and their companies are permitted to post the accepted version of IEEE-copyrighted material on their own Web servers without permission, provided that the IEEE copyright notice and a full citation to the original work appear on the first screen of the posted copy. An accepted manuscript is a version which has been revised by the author to incorporate review suggestions, but not the published version with copy-editing, proofreading, and formatting added by IEEE. For more information, please go to: [http://www.ieee.org/publications\\_standards/publications/rights/paperversionpolicy.html](http://www.ieee.org/publications_standards/publications/rights/paperversionpolicy.html). Permission to reprint/republish this material for commercial, advertising, or promotional purposes or for creating new collective works for resale or redistribution must be obtained from IEEE by writing to the IEEE Intellectual Property Rights Office, 445 Hoes Lane, Piscataway, NJ 08854-4141 or [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org). Copyright © 2023 IEEE. All rights reserved.

**Abstracting and Library Use:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

**Unsubscribe:** If you no longer wish to receive this *ComputingEdge* mailing, please email IEEE Computer Society Customer Service at [help@computer.org](mailto:help@computer.org) and type "unsubscribe *ComputingEdge*" in your subject line.

IEEE prohibits discrimination, harassment, and bullying. For more information, visit [www.ieee.org/web/aboutus/whatis/policies/p9-26.html](http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html).

## IEEE Computer Society Magazine Editors in Chief

### **Computer**

Jeff Voas, *NIST*

### **Computing in Science & Engineering**

Lorena A. Barba, *George Washington University*

### **IEEE Annals of the History of Computing**

David Hemmendinger, *Union College (Interim EIC)*

### **IEEE Computer Graphics and Applications**

André Stork, *Fraunhofer IGD and TU Darmstadt*

### **IEEE Intelligent Systems**

San Murugesan, *Western Sydney University*

### **IEEE Internet Computing**

George Pallis, *University of Cyprus*

### **IEEE Micro**

Lizy Kurian John, *University of Texas at Austin*

### **IEEE MultiMedia**

Balakrishnan Prabhakaran, *University of Texas at Dallas*

### **IEEE Pervasive Computing**

Fahim Kawsar, *Nokia Bell Labs and University of Glasgow*

### **IEEE Security & Privacy**

Sean Peisert, *Lawrence Berkeley National Laboratory and University of California, Davis*

### **IEEE Software**

Ipek Ozkaya, *Software Engineering Institute*

### **IT Professional**

Charalampos Z. Patrikakis, *University of West Attica*

COMPUTING  
**edge**



27

Data Science:  
Hype and  
Reality

39

Is It Live, or  
Is It Deepfake?

47

Serverless  
Computing for  
Scientific  
Applications

## Big Data

### 8 Challenges of Large-Scale Data Processing in the 1990s: The IPUMS Experience

DIANA L. MAGNUSON AND STEVEN RUGGLES

### 21 Technology Trends and Challenges for Large-Scale Scientific Visualization

JAMES AHRENS

## Careers

### 27 Data Science: Hype and Reality

NORITA AHMAD, AREEBA HAMID, AND VIAN AHMED

### 34 Careers in STEM: A Latina Perspective

ANDREA DELGADO, VERONICA G. MELESSE VERGARA, AND ANDREA SCHNEIBEL

## Artificial Intelligence

### 39 Is It Live, or Is It Deepfake?

NIR KSHETRI, JOANNA F. DEFRANCO, AND JEFFREY VOAS

### 42 The AI-Cybersecurity Nexus: The Good and the Evil

SAN MURUGESAN

## Cloud Computing

### 47 Serverless Computing for Scientific Applications

MACIEJ MALAWSKI AND BARTOSZ BALIS

### 53 Randy Shoup on Evolving Architecture and Organization at eBay

JEREMY JUNG

## Departments

- 4 Magazine Roundup
- 7 Editor's Note: Big Data: Then and Now
- 63 Conference Calendar

Subscribe to *ComputingEdge* for free at  
[www.computer.org/computingedge](http://www.computer.org/computingedge).



# Magazine Roundup

The IEEE Computer Society's lineup of 12 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 visualization and microchip design. Here are highlights from recent issues.

## Computer

### ***There Is Truth in Legends: Lessons for Team Performance From Hackathons***

---

In this article from the August 2023 issue of *Computer*, the authors present quantitative research of 281 online hackathon participants. They analyze the hackathon's mechanisms in depth and broaden an understanding of how its elements affect the participants' performance. They offer insights that may improve invention development methods in general.

## computing

in SCIENCE & ENGINEERING

### ***Science Gateways and the Humanities: An Exploratory Study of Their Rare Partnership***

---

Researchers and educators in humanities such as computational linguists, digital humanists, and those doing historical reconstructions are increasingly heavy users of computational and/or data resources. Many know about activities, working groups, and initiatives around the findable, accessible, interoperable, reusable (FAIR) principles and are a driving force

for improving the sharing of data and software. However, it seems that humanities researchers are less aware of the science gateways community and the end-to-end solutions that science gateways could provide and are therefore lacking a driving force for adoption of this technology. This article from the January/February 2023 issue of *Computing in Science & Engineering* clarifies some of the challenges and needs faced by computational researchers in the humanities that may explain their relatively low participation in the science gateways community.

## IEEE Annals

of the History of Computing

### ***Promoting Computing in the Postwar United States—The Case of UCLA***

---

In the mid-1940s, a differential analyzer and an electronic digital computer acquired by the University of California-Los Angeles (UCLA) made large-scale computing available to staff and students there and in surrounding industries. To serve these users, the university fostered both new professional organizations and pioneering curricula in what would later be called

computer science. Read more in this article from the April–June 2023 issue of *IEEE Annals of the History of Computing*.

## IEEE Computer Graphics and Applications

### ***Presenting Morphing Shape Illusion: Enhanced Sense of Morphing Virtual Object With Weight Shifting VR Controller by Computational Perception Model***

---

Haptic sensation is crucial for virtual reality, as it gives the presence of objects in a virtual world and thus gives a greater sense of immersion. To provide a sense of the shape of handheld objects, a haptic device that changes weight distribution is proposed. It is known that visual feedback enhances the haptic sensation of shape, and it is also known that it does for morphing shape as well. The authors' previous publication presented a perception model for the static shape of a virtual object. In this article from the July/August 2023 issue of *IEEE Computer Graphics and Applications*, they extend the model to produce a plausible sense of the morphing shape of handheld objects.



## IEEE Intelligent Systems

### ***DCAT: Combining Multisemantic Dual-Channel Attention Fusion for Text Classification***

---

Text classification is a fundamental and central position in natural language processing. There are many solutions to the text classification problem, but few use the semantic combination of multiple perspectives to improve the classification performance. This article from the *IEEE Intelligent Systems* July/August 2023 issue proposes a dual-channel attention network model called DCAT, which uses the complementarity between semantics to refine the understanding deficit. Specifically, DCAT first captures the logical semantics of the text through transductive learning and graph structure. Then, at the attention fusion layer (channel), logical semantics are used to perform joint semantic training on other semantics to correct the predictions of unlabeled test data incrementally.

## IEEE Internet Computing

### ***Serverless Vehicular Edge Computing for the Internet of Vehicles***

---

Rapid growth in the popularity of smart vehicles and increasing

demand for vehicle autonomy brings new opportunities for vehicular edge computing (VEC). VEC aims at offloading the time-sensitive computational load of connected vehicles to edge devices. However, VEC offloading raises complex resource management challenges and, thus, remains largely inaccessible to automotive companies. Recently, serverless computing emerged as a convenient approach to the execution of functions without the hassle of infrastructure management. In this article from *IEEE Internet Computing's* July/August 2023 issue, the authors propose the idea of serverless VEC as the execution paradigm for Internet of Vehicles applications.

## IEEE micro

### ***There's Always a Bigger Fish: A Clarifying Analysis of a Machine-Learning-Assisted Side-Channel Attack***

---

Machine learning has made it possible to mount powerful attacks through side channels that are otherwise challenging to exploit. However, due to the black-box nature of machine learning models, these attacks can be difficult to interpret correctly. Models that simply find correlations cannot be used to analyze the various sources of information leakage behind an attack. This article from the July/

August 2023 *IEEE Micro* issue highlights the limitations of relying on machine learning for side-channel attacks without completing a comprehensive security analysis.

## IEEE MultiMedia

### ***Edge-Assisted Virtual Viewpoint Generation for Immersive Light Field***

---

Light field (LF), which describes the light rays that emanate at each point in a scene, can be used as a six-degrees-of-freedom (6DOF) immersive media. Similar to the traditional multiview video, LF is also captured by an array of cameras, leading to a large data volume that needs to be streamed from a server to users. When a user wishes to watch the scene from a viewpoint that no camera has captured directly, a virtual viewpoint must be rendered in real time from the directly captured viewpoints. This places high requirements on both the computing and caching capabilities of the infrastructure. Edge computing (EC), which brings computation resources closer to users, can be a promising enabler for real-time LF viewpoint rendering. In this article from *IEEE MultiMedia's* April-June 2023 issue, the authors present a novel EC-assisted mobile LF delivery framework that can cache parts of LF viewpoints in advance

and render the requested virtual viewpoints on demand at the edge node or user's device. Numerical results demonstrate that the proposed framework can reduce the average service response latency by 45% and the energy consumption of user equipment by 60% at the cost of 55% additional caching consumption of edge nodes.



### ***Linguistic and Vocal Markers of Microbehaviors Between Team Members During Analog Space Exploration Missions***

The authors of this article from *IEEE Pervasive Computing's* April–June 2023 issue used machine learning classifiers and dialog state tracking models, combined with natural language processing techniques relying on lexicon-based methods and data-driven methods, to automatically detect positive and negative microbehaviors between team members in nine four-person teams in a simulated space habitat. Their findings indicate that the psycholinguistic markers extracted using the linguistic inquiry and word count, STRESS-net dictionaries, and acoustic features can achieve an f1-score up to 54.87% in a three-class classification problem. The findings suggest that modeling turns between the sender and target of microbehaviors is significantly more effective in detecting microbehavior than only modeling the sender's information. Finally, they demonstrate the effect of introducing

context for detection purposes. Dialog state tracking approaches that model the linguistic interaction between team members and incorporate contextual information about the task and sentiment of the conversation can further yield improved performance, depicting an f1-score of 57.73%.



### ***In Pursuit of Aviation Cybersecurity: Experiences and Lessons From a Competitive Approach***

The passive and independent localization of aircraft has been the subject of much cyberphysical security research. The authors of this *IEEE Security & Privacy* article from the July/August 2023 issue designed a multistage open competition focusing on the offline batch localization problem using opportunistic data sources. They discuss setup, results, and lessons learned.



### ***Toward a Free and Open Source-Driven Public Sector: An Italian Journey***

The authors of this article from the July/August 2023 issue of *IEEE Software* focus on the Italian free and open source software (FOSS) strategy, consisting of three key pillars: norms, tools, and community. The Italian approach seeks to develop a public sector that can not only use but also produce FOSS.



### ***Using Internet of Things Application for Energy-Efficient and Lightweight Internet of Drones Networks***

The Internet of Drones (IoD) is a rapidly growing technology with the potential to revolutionize various industries, but it also raises concerns about security and efficiency in communication between drones. Blockchain technology has the potential to solve these issues, but conventional blockchains face performance, computation, and scalability issues. This article from the July/August 2023 issue of *IT Professional* proposes using Internet of Things Application (IOTA) distributed ledger technology (DLT) to address these concerns in the IoD. IOTA offers a highly energy-efficient alternative to major DLTs such as Bitcoin and Ethereum while providing fast and secure solutions. The results show that IOTA outperforms not only Bitcoin but also low-energy DLTs like Ethereum by a huge margin. The use of IOTA DLTs in large-scale systems including the IoD can improve the efficiency and speed of tasks such as surveillance, delivery, and inspection. The article also discusses the potential challenges and future work for integrating IOTA DLTs in the IoD. 🌍







## Editor's Note

# Big Data: Then and Now

**D**ata isn't what it used to be. Over the years, the amount of data that humanity generates has increased exponentially. How did computer scientists grapple with increasing amounts of data decades ago? And what are the most difficult challenges in managing big data today? This *ComputingEdge* issue examines the fascinating past and present of big data.

"Challenges of Large-Scale Data Processing in the 1990s: The IPUMS Experience," from *IEEE Annals of the History of Computing*, describes a historically important project that improved dataset interoperability, accessibility, and preservation in the late 20th century. "Technology Trends and Challenges for Large-Scale Scientific Visualization," from *IEEE Computer Graphics and Applications*, discusses modern data storage, processing, and

visualization techniques with an emphasis on large and complex scientific datasets.

With the growth of big data, data scientists are now highly valued across industries. The authors of *Computer's* "Data Science: Hype and Reality" present the current state of data science as a career. In "Careers in STEM: A Latina Perspective," from *Computing in Science & Engineering*, three computing professionals recount their career paths and their efforts to broaden Latina participation in the field.

When used for good, artificial intelligence (AI) can be revolutionary; but in the wrong hands, it can be equally harmful. The authors of *Computer's* "Is It Live, or Is It Deepfake?" argue that we need reliable tools for detecting AI-produced deepfakes—in part so that we can fully realize legitimate use cases. *IT Professional's* "The AI-Cybersecurity Nexus: The Good and the

Evil" explains how AI can be utilized by both attackers and cybersecurity professionals.

This *ComputingEdge* issue closes with some cloud computing insights. The authors of *IEEE Internet Computing's* "Serverless Computing for Scientific Applications" propose an architecture for a popular cloud computing model. *IEEE Software's* "Randy Shoup on Evolving Architecture and Organization at eBay" features an interview with an expert who advises new companies to use a public cloud. 🌐



## DEPARTMENT: ANECDOTES

# Challenges of Large-Scale Data Processing in the 1990s: The IPUMS Experience

Diana L. Magnuson and Steven Ruggles, *Institute for Social Research and Data Innovation, University of Minnesota, Minneapolis, MN, 55455, USA*

*When it was launched in 1991, the Integrated Public Use Microdata Series (IPUMS) project faced a challenging environment and limited resources. Few datasets were interoperable and much data collected at great public expense was inaccessible to most researchers. Documentation of datasets was nonstandardized, incomplete, and inadequate for automated processing. With insufficient attention to preservation, valuable scientific data were disappearing (see Bogue et al., 1976). IPUMS was established to address these critical issues. At the outset, IPUMS faced daunting barriers of inadequate data processing, storage, and network capacity. This anecdote describes the improvised computational infrastructure developed in the decade from 1989 to 1999 to process, manage, and disseminate the world’s largest population datasets. We use a combination of archival sources, interviews, and our own memories to trace the development of the IPUMS computing environment during a period of explosive technical innovation. The development of IPUMS is part of a larger story of the development of social science infrastructure in the late 20th century and its contribution to democratizing data access.*

The U.S. Census Bureau played a key role in the development of computing technology. The two leading computer companies of the middle decades of the 20th century—IBM and Remington Rand’s UNIVAC division—had roots as data processing companies that built equipment for the Census Bureau, and the Bureau indirectly funded the development of the first commercial computer beginning in 1946 [26].

The 1960 census was fully computerized, and the Census Bureau began to publish data in machine-readable form. In 1962, the Census Bureau introduced the first machine-readable microdata file, a one-in-1000 sample of households drawn from the 1960 census. Following the 1970 census, the Census Bureau greatly expanded the quantity of microdata, providing

six 1% samples from the 1970 census—a 60-fold increase compared with 1960.

After the 1970 samples were released, the Census Bureau contracted to produce a new version of the 1960 microdata, expanded from 1-in-1000 sample density to 1-in-100. The 1960 work was carried out by the Data Use and Access Laboratories (DUALabs), another Census Bureau spin-off, which made the decision to adopt the coding schemes and record layout that had been used for the 1970 samples [24].

The availability of two compatible census years led to an explosion of research in the 1970s on social and economic changes over the course of the 1960s. The pair of microdata samples became essential tools of American social scientists. It was in this climate that two demographers proposed extending the data series backward by digitizing census microdata from earlier years. The enumeration manuscripts for the 1900 census were made accessible to scholars by the National Archives and Records Administration in 1972. Between 1976 and 1980, demographer Samuel Preston of the University of Washington directed a project funded by the National Science Foundation (NSF) to

create a one-in-750 sample of the 1900 census. Halliman Winsborough, a demographer at the University of Wisconsin, independently came up with the idea for creating samples of the 1940 and 1950 censuses. The 1940 and 1950 censuses were still protected by census confidentiality rules, so the samples were created within the Census Bureau between 1978 and 1984 with funding from an NSF grant. Preston moved to the University of Pennsylvania in 1979 and, with the release of the 1910 census manuscripts in 1982, he led a project between 1983 and 1989 to create a sample of the 1910 census with funding from both NSF and the National Institute of Child Health and Human Development (NICHD) [8], [10], [30], [34], [38], [39].

At the University of Minnesota, the use of computer technology to understand historical transformations through quantitative methods began in the decade between 1976 and 1985, the same period that the first historical microdata samples were being created. During this era, academic computing at the University was centralized on a mainframe computer in the nearby suburb of Lauderdale, Minnesota. Although punch cards remained the primary means of creating datasets, interactive computing through teletype machines became available in the mid-1970s [17], [36].

Historical demographers Robert McCaa (hired in 1974), Russell R. Menard (hired in 1976), and Steven Ruggles (hired in 1984) undertook collaborative work that contributed to the expansion and redefinition of social science data infrastructure. All three had been working with historical census microdata since the mid-1970s, using card punches to digitize historical enumeration manuscripts from Ramsey County, Minnesota (Menard) [4], Parral, Mexico (McCaa) [15], and Lancashire, England (Ruggles) [19].

By the mid-1980s, the limitations of the mainframe computing system for large-scale historical datasets were apparent. Shortly after he arrived at Minnesota, Ruggles began working on the Control Data Corporation CYBER mainframe computer using the newly created 1940 Public Use Microdata Sample. He used the Network Operating System "COP" command to transfer the data files from a temporary disk to a permanent one. The charge for executing that single command was \$500, using up half of his annual allocation of computer time.<sup>1</sup>

To avoid the high cost of the centralized computer system, the historians turned to microcomputers. McCaa had been a microcomputer enthusiast since they were introduced in the 1970s and by the early

1980s was using microcomputers for census data entry in Mexico [15], [16].<sup>2</sup> In 1985, Menard and Ruggles established the History Micro-Computer Lab—later known as the Social History Research Laboratory (SHRL)—with a grant from IBM. The lab initially had two IBM computers, but soon expanded to five, including the latest model, the IBM PC/AT. SHRL obtained and made available a wide range of historical datasets for teaching and research.

During the 1985–1986 academic year, we made available for classroom use: a large national sample from the 1900 federal population census; data from the population and agricultural census schedules for rural Ramsey County, Minnesota, for 1860 through 1880; biographical information on 1500 graduates of Yale College in the eighteenth century; and information on basic demographic parameters of early America that permits exploration of how migration and vital rates interacted to shape the growth of population. During 1986–87, we developed additional data sets, particularly census material from northern Mexico during the late eighteenth and early nineteenth centuries; European family reconstitution data from 1600 to 1800; and evidence on the family budgets of industrial workers in Germany and the United States during the 1890s; national samples of the federal census for the years 1910, 1960, 1970, and 1980; and census returns for the Red Lake, Minnesota Indian Reservation for 1900 and 1910 [35, pp. 70–71].

The 1900, 1910, 1960, 1970, and 1980 census samples made available in the SHRL were not the full samples, since the full datasets were far too large to be used on a microcomputer. Rather, they were tiny subsamples, with 2100 households (6000 to 10,000 persons) from each census. The earliest complete microdata sample was transferred to a microcomputer in 1987, when graduate student Ron Goeken was assigned the task of breaking the 1900 public use sample into pieces small enough to fit on floppy disks. Goeken then copied the files to Ruggles' microcomputer and reassembled them. The full sample was just under 10 MB, consuming half the computer's available disk space.<sup>3</sup>

<sup>1</sup>Steven Ruggles, interviewed by D. L. Magnuson, University of Minnesota, Jan. 9, 2014.

<sup>2</sup>Robert McCaa, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Dec. 6, 2021.

<sup>3</sup>Ronald Goeken, interviewed by D. L. Magnuson, University of Minnesota, Oct. 3, 2013; Ronald Goeken, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 6, 2021.

Ruggles had close connections to the Preston and Winsborough census projects. He was an undergraduate at the University of Wisconsin in 1978 when the 1940–1950 project was launched, and he was a postdoc there in 1984 when it was completed. In between, he was a graduate student at the University of Pennsylvania during the creation of the 1910 project, and he was among the earliest users of the 1900 microdata sample, which formed the basis for much of his dissertation research.

By 1989, Preston and Winsborough had both decided that they were done with census digitization projects, and they were happy to share their grant proposals. Borrowing the strongest parts of those, Ruggles and Menard sought NICHD funding to create a 1-in-500 public use microdata sample of the 1880 U.S. census [12], [27], [29].

The 1880 census was the first to collect data on marital status, family relationships, and parental birthplaces, making it especially useful for studies of household composition, fertility, and nuptiality [27]. Ruggles' research focused on long-run changes in family and household composition, including multi-generational households, boarding and lodging, single parenthood, and living alone. Even though the Census Office began collecting full information on family and household composition in 1880, it did not publish any statistics on the topic until 1940, and only a few rudimentary measures thereafter. The series of microdata files, therefore, offered unprecedented opportunities to examine changes in families and households for the country as a whole. Since the 1880 Census was the first to identify family relationships and marital status, it made a natural starting point for studies of long-run family change.

The 1880 proposal earned a priority score in the top 1%. The reviewers had one major reservation: They felt that the sample size was too small. Accordingly, the panel recommended funding only if the proposal were revised from the proposed 1-in-500 density to 1-in-100, with an appropriate adjustment of the budget. Ruggles and Menard enthusiastically agreed to make the change.

According to Ruggles, he and Menard did not know if “the first one [1880] would fly,” but once it did, they were confident that they could secure funding and “fill in the rest.”<sup>4</sup> Their confidence was not misplaced: Federal funding for the creation and dissemination of new samples for the 1850, 1860, 1870, 1900, 1910, 1920, and 1930 censuses came to the University of Minnesota between 1992 and 2002.

<sup>4</sup>Ruggles interview, Jan. 9, 2014.

## TECHNOLOGICAL ENVIRONMENT, 1989–1994

The University of Minnesota history department shared access to the Social Science Research Facilities Center (SSRFC) with other social science departments at the university. Located on the West Bank in the basement of Blegen Hall, SSRFC was organized in the 1964–1966 biennium and designed to facilitate and support research at the university [37]. SSRFC was designed to provide access to the mainframe computer in Lauderdale (approximately 5 mi/8 km from Blegen Hall), and housed keypunch machines, a card reader, a line printer, and a punch card sorting machine. As interactive computing became feasible in the later 1970s, SSRFC installed computer terminals; in the 1980s, SSRFC added a PDP-10 minicomputer and nine-track tape drive.<sup>5</sup>

SSRFC staff provided essential technical computing support for the 1880 census project (and subsequent census projects) and the Integrated Public Use Microdata Series (IPUMS) during the 1990s.<sup>6</sup> To complement the microcomputers in the History Department's SHRL, Ruggles and Menard began to assemble the computational infrastructure necessary to support the 1880 PUMS project.

The 1880 Public Use Microdata Sample grant (1989–1994) provided funding to purchase motorized Dukane 35-mm microfilm readers, data entry equipment, data processing equipment, and software. Computer Marketing Corporation 80386 microcomputers for data entry were selected through a competitive bidding process. Workstations were connected by a 2-Mb Ethernet to a Sun SPARCstation file server and a 9-track tape unit. Microcomputers configured as data entry terminals were directly networked with the Sun, which eliminated file transfers via modem or floppy disk. Graduate research assistant Bill Block owned a drill and came in one evening to drill holes through the concrete-block walls of four adjacent offices to connect the data entry terminals with the SPARCstation [11].<sup>7</sup>

Office space was provided by the History Department at the University of Minnesota, where the project was initially located. The 1880 project was

<sup>5</sup>Phil Voxland, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 21, 2021. John Easton, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 21, 2021.

<sup>6</sup>Phil Voxland, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 21, 2021. John Easton, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 21, 2021.

<sup>7</sup>William Block, interviewed by D. L. Magnuson, University of Minnesota, Dec. 19, 2013.



**FIGURE 1.** Graduate Research Assistant Office, 1993.

allocated a corner office to start. Graduate research assistants also made use of Ruggles' office, conveniently adjacent to the corner office. As the project work expanded in the early 1990s to include 1850 and 1920 PUMS and IPUMS, a second office was allocated by the department at the opposite corner of the same hallway; this space was used exclusively by graduate research assistants. The space was extremely tight (see Figure 1). The first SPARCstation server was named "legohead" because space was so crowded that graduate students should have had Legos on their heads so they could sit on top of one another [13], [14]. There was no funding for office furniture, so project personnel obtained old furnishings from the university's warehouse, with most items dating back to the 1940s or 1950s. Graduate research assistant Dan Kallgren remembered,

"We ended up getting stone-lined fireproof cabinets that were designed to hold index cards, but they were perfect for holding boxes of microfilm reels. The cabinets weighed hundreds of pounds apiece. We somehow manhandled them onto the back of my pickup truck and brought them over to the Social Science Tower... and we ended up with this weird collection of desks and stuff for the research assistants' office."<sup>8</sup>

<sup>8</sup>Daniel Kallgren, interviewed by D. L. Magnuson, University of Minnesota, Oct. 31, 2013.

In the 1990s era of computing technology, one could not separate the room from the technology.<sup>9</sup> The data entry operators (DEOs) were civil servants, and each had a designated desk with a desktop computer and microfilm reader. Because office space was tight, DEOs shared their workspace with graduate research assistants associated with the historical census projects. The workstations for graduate research assistants (RAs) were undesignated spaces that were used at all hours of the day and night for their work to verify the data and research procedural histories. It was understood that anyone could use any workstation, but in practice, grad RAs tended to make use of a favorite spot based on the time of day they were in the office. Shared office space created a productive synergy between and among DEOs and graduate research assistants (shown in Figure 2).

With a couple of exceptions, most graduate students arriving in the history department had limited computer experience. Lisa Dillon, Diana Magnuson, and Dave Ryden remember a supportive office culture of shared learning around computing skills and data construction. Graduate research assistants learned how to use the software from each other. Lisa Dillon recalled,

"I learned what I learned by sitting next to people and people showing me... it was [Matt]

<sup>9</sup>Lisa Dillon, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Feb. 4, 2022.



**FIGURE 2.** IPUMS Staff, 1996.

Sobek who showed me what a spreadsheet was and how it worked.”

Todd Gardner trained Diana Magnuson in the use of Norton Commander, a file-managing tool. Being “in the room where it happens” was critical to quickly skilling up and becoming a productive and creative member of the historical census projects and IPUMS. As evidence of this early era of computing technology, several members distinctly remember the very first electronic mail they saw and sent; an email exchange with Steve Ruggles, checking in on the grad RAs while he was on sabbatical in England in the spring of 1990.<sup>10</sup>

In addition to their data production tasks, graduate research assistants carried much of the administrative load in the early years of the 1880 project through the launch of IPUMS in the early 1990s. The history department was not accustomed to grant administration and did not have adequate capacity to meet the growing administrative needs of the projects. History department personnel were available for guidance and troubleshooting, but day-to-day administrative details were carried out by graduate research assistants in consultation with Ruggles and Menard. At the startup of the 1880 PUMS project, Dan Kallgren managed administrative duties that included coordinating purchasing most hardware and software assets for the project.

<sup>10</sup>Kallgren interview, Oct. 31, 2013; David Ryden, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Jan. 24, 2022; Catherine Fitch, interviewed by D. L. Magnuson, University of Minnesota, Jun. 16, 2015; Dillon interview, Feb. 4, 2022.

Additionally, he researched used commercial coffee makers, made his purchase from a restaurant supply company in St. Paul, and transported the large coffee maker, strapped to the back of his bicycle, back to the Social Science Tower on the West Bank. Coffee fueled DEOs, graduate students, and faculty alike, and the machine served as the locus of a community building and gathering space. David Ryden administered tasks related to the 1920 PUMS project, which began in 1993. Ryden also had a bicycle, and this came in handy for the many trips he made back and forth across campus to obtain signatures and drop off forms related to historical census project contracts. Beginning in the fall of 1995, Cathy Fitch gradually assumed a variety of IPUMS project administrative responsibilities. The procedural history was largely researched by graduate research assistants Diana Magnuson and Miriam King, a post-doc at the university’s Hubert H. Humphrey Center for Public Affairs. Thus, the computational and procedural infrastructure that developed in the 1990s was administered by graduate research assistants taking on a variety of responsibilities necessary to processing, managing, and disseminating what would become the world’s largest population data collection.<sup>11</sup>

Software development for the 1880 PUMS was the top priority as the project got underway. Microcomputers were used for data entry, post-entry data consistency checking, and verification. The data entry software was the Integrated System for Survey

<sup>11</sup>Kallgren interview, Oct. 31, 2013; Ryden interview, Jan. 24, 2022; Fitch interview, Jun. 16, 2015.

Analysis (ISSA), which had been developed by Westinghouse for processing the Demographic and Health Surveys [6]. Unlike competing data-entry software, all the individual-level information for a dwelling could be included on the screen simultaneously. This meant that the DEOs could see the individual-level information for the entire dwelling while entering the data for each individual. The software had extensive capabilities for real-time consistency checks between fields; for example, it could be set up to require that married, divorced, or widowed persons were of marriageable age.

The ISSA software had significant drawbacks. The system required extensive programming. It was awkward to navigate between data entry fields except in a fixed sequence. The 1880 project sought to capture long alphanumeric strings verbatim to ensure minimal information loss, and ISSA did not handle long strings well. There were few string handling functions, and forms with many long strings were unstable and frequently crashed [3], [6].

As work progressed, it became apparent that microcomputer memory capacity was insufficient to carry out the final recoding of string variables into numeric codes. Post-entry data processing was thus shifted to the Sun UNIX workstation, which was faster and had more memory and storage capacity [6]. This operation relied on FORTRAN-77 programs and data dictionaries containing all the alphabetic entries transcribed by DEOs and their corresponding numeric codes as classified by graduate research assistants [6].

As the 1880 project was underway, Ruggles and Menard obtained additional funding to launch projects to create samples of the 1850 census in 1992 and the newly released 1920 census in 1993 [23], [28]. The 1850 project continued to use the ISSA data-entry software, but ISSA could not accommodate the expanded length of the 1920 records. After an extensive search for appropriate commercial software, the 1920 investigators decided to develop customized data-entry software using Microsoft C and Liant Software's C-scape. This approach allowed design of screen forms that mirrored the manuscript enumeration form. In addition to developing an intuitive screen layout, interactive data consistency checking for string and numeric data was built into the program [6].

### FIRST IPUMS PROJECT 1992–1995

By 1991, 10 machine-readable public use microdata samples covering the decennial censuses from 1880

and 1990 (1850, 1880, 1900, 1910, 1940, 1950, 1960, 1970, 1980, and 1990) were publicly available or under development. Use of the PUMS as a time series was impeded by incompatibilities: They were created at different times by different investigative teams, resulting in different formats, different documentation, different record layouts, and different coding schemes. The only exception was the 1960 and 1970 samples, because the new version of the 1960 sample created in 1973 had been redesigned to share the same coding and layout as the 1970 samples.

To enable an analysis of long-run family change, between 1985 and 1989, Ruggles developed a set of FORTRAN programs that recoded selected variables into a common format across the available census samples, created subsets of the samples that were of manageable size, and pooled multiple censuses into a single file [33].<sup>12</sup> This first effort at making variables compatible across data sets had liabilities. Initially, the guiding principle when combining the samples was a "lowest common denominator" approach [24, p. 1405]. Only a small subset of variables was converted into the common codes. The result of this method was a significant loss of information, which meant that customization of the programs was usually necessary to meet the requirements of any specific research project [24].

Despite the limitations, there was considerable demand for custom-designed, consistent-format census microdata to meet the research and teaching needs of the university's graduate students and faculty, as well as a few researchers from other institutions. By 1991, SHRL had a server dedicated to common format extracts that ran nearly continuously to meet researcher demand. The source data were stored on a set of Write-Once Read Many disks read by an IBM 3363 optical drive. Extracts for a subpopulation across all eight existing census samples required up to a day to process, and the SHRL was preparing about 25 such extracts per month [24].<sup>13</sup>

In 1991, Ruggles proposed to the NSF to create a single integrated series that would "maximize comparability and minimize information loss" across the eight public use microdata samples [20], [21], [22], [33, p. 103]. Ruggles also anticipated adding data for the remaining publicly available census years (1860, 1870, and 1930), thus "constitut[ing] a resource of unprecedented power for the study of long-term social

<sup>12</sup>Todd Gardner, interviewed by D. L. Magnuson, University of Minnesota, Oct. 24, 2013; Matthew Sobek, interviewed by D. L. Magnuson, University of Minnesota, Oct. 10, 2013.

<sup>13</sup>Sobek interview, Oct. 10, 2013.

change” [22]. Ruggles remembers the moment he went into the History Department lounge on the sixth floor of the Social Science Tower and said, “IPUMS! Integrated Public Use Microdata Series! Isn’t that a great idea?” The response from the graduate research assistants was not enthusiastic.

“What a terrible name! You can’t call it that!” According to Ruggles, “It was universal; everyone thought it was just a horrible name. . . It wasn’t a bad idea to propose, just a terrible thing to call it.”<sup>14</sup>

The 1991 NSF IPUMS grant proposal outlined four steps for

“convert[ing] the public use sample series into a single coherent form: 1) planning and design of record layouts, coding schemes, and constructed variables that maximize comparability and minimize information loss; 2) software development for reformatting, recording, constructing new variables, consistency checking, and allocation of missing and inconsistent data; 3) data processing of approximately 65 million records; and 4) preparation of an integrated set of documentation for the entire series of datasets, including a general user’s guide, a volume of procedural histories, and a volume on technical characteristics and error estimation” [22].

IPUMS succeeded because of two key technical innovations of the early 1990s: 1) In 1991, IPUMS introduced the first structured metadata system for data integration; and 2) in 1995, IPUMS produced the first interactive web-based system for creating customized pooled datasets.

The IPUMS software was designed to be driven entirely by machine-processable metadata. The project adopted what is now known as a data warehousing approach, which transforms data from heterogeneous sources into a single schema. The key metadata element underlying the process was called a “translation table.” A separate translation table was prepared for each variable in the system; an excerpt of the translation table for the “Relationship” variable appears in Figure 3. The top of the translation table provides the input column locations for variables for each source dataset. For example, in the original 1880 dataset, the relationship variable was on the person record and ran

RELATE .TRN Relationship	1880	1900	1910	1940	1950	1960	1970	1980	1990	
##	21	23	79							
1880 P	09	11	28							
1900 P	14	16								
1920 P	11	14								
1940 P	11	14	97							
1950 P	16	20	63							
1960 P	01	02								
1970 P	01	02	105							
1980 P	02	04	140							
1990 P	09	10	184							
##										
#	IPUMS	1880	1900	1910	1940	1950	1960	1970	1980	1990
HEAD & RELATIVES (1-10):	01 01	100	100	100	019901999		0-	0-	000	00
Head/Householder	01 01						00	00		
Spouse	02 01	120	120	120	029902999		1-	1-	010	01
Husband, not Head	02 01	140	140							
2nd/3rd Wife (PG)	02 02	121	129							
Child	03 01	130	130	130	039903999		2-	2-	020	02
(1970 screw-ups)	03 01						20	20		
(1970 screw-ups)	03 01							22		
Adopted Child	03 01							26		
Stepchild	03 02	132	132	132						
Adopted, ns	03 03	131	131	131	049904999					03
Child-in-law	03 04			280						
Step Child-in-law	04 01	133	133	133	059905999		30	30	051	*
Parent	04 02	134	134							
Stepparent	05 01	210	210	210	079907999		32	32	040	05
	05 02	211	211	211						

FIGURE 3. Translation table.

from column 21 to column 23. The additional input field refers to the location of data quality flags in the original data, which provide information about missing or inconsistent responses.

The rest of the table has a row for each relationship code. On the left are the IPUMS labels for each category and the standardized IPUMS code. The IPUMS code comes in two parts: The first two digits are a lowest common denominator designed to be consistent across all samples, and the second two digits provide additional detail available in a subset of censuses. This composite coding system provides interoperability without losing information. The input codes are on the right of the table, with separate codes for each census year. For example, the Head/Householder category was coded 100 in the 1880–1910 censuses, 0199 in 1940, and 000 in 1980. In addition to the translation tables, other machine-processable metadata components specified the order in which variables were to be processed, the names and locations of input and output data files, and the record layout of output files [9].

Metadata-driven data integration was a major innovation. Prior data harmonization efforts used statistical packages or software code to recode multiple datasets into a common coding scheme. Such systems were cumbersome to program and difficult to maintain. The translation table provided a visual representation of recodes across all datasets simultaneously that greatly simplified the design of integrated codes. The system made it simple to develop tests of internal consistency

<sup>14</sup>Ruggles interview, Jan. 9, 2014.



of the recodes (e.g., each input code must correspond to just one output code) and to ensure that all values encountered in the data appear in the translation tables. Without structured machine processable meta-data, it would have been far more expensive to build the initial IPUMS or to expand it to cover additional data sources.

The second major IPUMS innovation was the interactive web-based data extract system. The extract system was made possible by the timing of the IPUMS project coinciding with the rapidly changing computing environment of the early 1990s. When IPUMS was proposed, few social scientists had access to the Internet, and the World Wide Web existed only on a single computer in Switzerland [5]. Data transfers in that period were accomplished by sending physical media through the mail, usually 10.5-in (26.67 cm) 9-track reel-to-reel tapes.

The original IPUMS proposal called for the purchase of 450 of these tapes. A third of the tapes were needed to hold one copy of the full set of census microdata in its original format. Another 150 tapes were needed to hold the transformed data in IPUMS format. Finally, the last 150 tapes were needed to send one copy of the IPUMS-format data to the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, which would handle the dissemination by sending copies on 9-track tapes to institutions around the country [22].

In the end, the IPUMS project purchased no tapes. In 1992 and 1993, data archivist Ann Gray at the Cornell Institute for Social and Economic Research made the Census Bureau microdata samples available for download through an anonymous file transfer protocol (FTP) server, a program that made it possible for outsiders to log on and download data. The IPUMS team was able to directly download the files over the Internet without using any tapes.

When it came time to disseminate the IPUMS data, the same technology was used. Instead of sending the data to ICPSR on 9-track tables, IPUMS set up an anonymous FTP site to disseminate the data [1], [33]; the first IPUMS dataset was downloaded on November 19, 1993.

On November 11, 1993, eight days before the first IPUMS Internet download, an undergraduate from the University of Illinois released the first successful web browser for a PC: NCSA Mosaic 1.0, thus opening broad access to the World Wide Web [18]. The first IPUMS website (see Figure 4) appeared 16 months later, one of the first 15,000 websites created. The initial IPUMS website offered the same functionality as the IPUMS FTP site: Users were able to

## Welcome to the Social History Research Lab!

### University of Minnesota

This web site currently contains the data and documentation for the **Integrated Public Use Microdata Series (IPUMS)**. The IPUMS is a database consisting of 23 samples of the U.S. Census from 1850 to 1990. The IPUMS assigns the different samples consistent codes and integrates their documentation. At present, you may only download compressed IPUMS data files. To obtain a DOS decompression program click [here](#).

If you need uncompressed data or have any other questions about the IPUMS database, contact us at [ipums@atlas.socsci.umn.edu](mailto:ipums@atlas.socsci.umn.edu)

If you have problems, questions, or suggestions about this page, send e-mail to [block@torgo.hist.umn.edu](mailto:block@torgo.hist.umn.edu)

The User's Guide for the IPUMS is available online in Word 6.0 format (see below). This document is 800 pages long, and is contained in 61 separate files. A printed version is available from us for \$30; contact [ipums@atlas.socsci.umn.edu](mailto:ipums@atlas.socsci.umn.edu)

#### IPUMS DATA FILES:

To download an IPUMS sample, just click on it:

[1850 sample \(6.9M\)](#)

[1880 sample \(21.1M\)](#)

[1900 sample \(3.6M\)](#)

[1910 sample \(14.0M\)](#)

[1920 sample \(14.4M\)](#)

[1940 sample \(62.4M\)](#)

[1950 sample \(72.8M\)](#)

[1960 sample \(77.4M\)](#)

[1970 5% state sample \(98.5M\)](#)

[1970 15% state sample \(98.5M\)](#)

[1980 B sample \(142.0M\)](#)

[1990 1% sample \(163.2M\)](#)

#### IPUMS USER'S GUIDE:

The following files taken together comprise the User's Guide. All files are in Word 6.0 format. As mentioned above, we will provide upon request a printed, double-sided version of the Guide for \$30. If you wish to print your own copy of the Guide, all of the following files should be downloaded (just click on each in succession). Each filename begins with "guide" followed by a letter, a numeral, and then a ".doc" extension. The letter and numeral describe the order of the file within the documentation (A7, A8, B1, B2, C1, etc). For a file describing the documentation files click [here](#).

**FIGURE 4.** First IPUMS website, March 1995.

download data and documentation files by clicking on them.

The largest file available in the 1993 version of IPUMS—the 1990 1% sample—was 163 MB. Files of that size were difficult to reliably transfer over the Internet and most social scientists found that scale of data challenging or impossible to store or process. To analyze changes over the entire period from 1850 to 1990—which was after all the point of the IPUMS—one would have to download the entire data series, 775 MB, and subset and merge the files to analyze change over time. Only investigators at major universities had the infrastructure to handle data on this scale [31], [32].

As noted, the SHRL had been creating customized harmonized data extracts for five years prior to the development of IPUMS. These extracts used a FORTRAN program to select specific variables and population subgroups needed to conduct a particular analysis and read the data from multiple files to produce a pooled extract with data from multiple censuses. Few users had the storage capacity or processing power needed to handle the very large public use microdata samples in their original form, so the process of

IPUMS Data Extract

Sample Selection

[Clear] [Submit]

**Household Variables**

What is your email address? <input type="text"/>	
<a href="#">Census Year</a>	<input type="checkbox"/> 1850 Sample <input type="checkbox"/> 1880 Sample <input type="checkbox"/> 1900 Sample <input type="checkbox"/> 1910 Sample <input type="checkbox"/> 1920 Sample <input type="checkbox"/> 1940 Sample <input type="checkbox"/> 1950 Sample <input type="checkbox"/> 1960 Sample <input type="checkbox"/> 1970 5% State Sample <input type="checkbox"/> 1970 15% State Sample <input type="checkbox"/> 1980 B Sample <input type="checkbox"/> 1990 1% Sample
<a href="#">Sample Density</a>	<input checked="" type="radio"/> Tiny <input type="radio"/> Small <input type="radio"/> Regular
<a href="#">File Type</a>	<input type="radio"/> Flat <input checked="" type="radio"/> Hierarchical
<a href="#">Data Quality Flags</a>	<input type="checkbox"/> Include all data quality flags



FIGURE 5. IPUMS data access system, 1995. Step 1.

IPUMS Data Extract Variable Selection

[Clear] [Submit]

**Household Variables**

<input type="checkbox"/> <i>Technical Variables</i>	
<input type="checkbox"/> RECTYP	<a href="#">Record type</a>
<input type="checkbox"/> YEAR	<a href="#">Census year</a>
<input type="checkbox"/> DATANUM	<a href="#">Data set number</a>
<input type="checkbox"/> SERIAL	<a href="#">Household serial number</a>
<input type="checkbox"/> NUMPREC	<a href="#">Number of person records following</a>
<input type="checkbox"/> SUBSAMP	<a href="#">Subsample number</a>
<input type="checkbox"/> HHWT	<a href="#">Household weight</a>
<input type="checkbox"/> NUMPERHH	<a href="#">Number of person in household</a>
<input type="checkbox"/> DWSIZE	<a href="#">Dwelling size</a>
<input type="checkbox"/> NMEMBERS	<a href="#">Number of members in sample unit</a>
<input type="checkbox"/> NUMHH	<a href="#">Number of households in dwelling</a>
<input type="checkbox"/> NUMHHTAK	<a href="#">Number of households sampled from dwelling</a>
<input type="checkbox"/> UNREL	<a href="#">Unrelated persons in household</a>
<input type="checkbox"/> SLPERNUM	<a href="#">Sample line person number</a>
<input type="checkbox"/> SELFWTHH	<a href="#">Self-weighting sample identifier</a>
<input type="checkbox"/> <i>Location Characteristics</i>	
<input type="checkbox"/> REGION	<a href="#">Census region and division</a> <input type="checkbox"/> range
<input type="checkbox"/> STATEICP	<a href="#">State (ICPSR code)</a> <input type="checkbox"/> range
<input type="checkbox"/> STATEFIP	<a href="#">State (FIPS code)</a> <input type="checkbox"/> range
<input type="checkbox"/> COUNTY	<a href="#">County</a>
<input type="checkbox"/> CIVDIV	<a href="#">Civil Division</a>
<input type="checkbox"/> COUNTY90	<a href="#">County, 1900 PUMS classification</a>
<input type="checkbox"/> CNTYGP97	<a href="#">County, group, 1970</a>
<input type="checkbox"/> CNTYGP98	<a href="#">County, group, 1980</a>
<input type="checkbox"/> PUMA	<a href="#">Public use microdata area</a>

FIGURE 6. IPUMS data access system, 1995. Step 2.

selecting variables, subsetting, and merging was essential to make the data usable.

The unlikely catalyst that broke this bottleneck was fantasy football. Graduate research assistant Todd Gardner, whose responsibilities were primarily coding and software development, was also commissioner of the graduate research assistants' office fantasy football league. Gardner was learning to use the Perl language to develop a Web application that produced box scores and fantasy football standings in a web output.<sup>15</sup> Unlike virtually all websites of the era, Gardner's fantasy football system was truly interactive: Users made a series of selections on a web form and the software then constructed a new customized web page based on those selections. When Gardner showed Ruggles his fantasy football website, they recognized the potential to adapt the system to create a web-based interactive front end that could operate the FORTRAN data extract program, allowing outside users to design their own customized datasets.

<sup>15</sup>Gardner interview, Oct. 24, 2013; Sobek interview, Oct. 10, 2013.

IPUMS Data Extract Case Selection

[Clear] [Submit]

**Household Variables**

<a href="#">Region</a>	<input type="checkbox"/> New England <input type="checkbox"/> Middle Atlantic <input type="checkbox"/> East North Central <input type="checkbox"/> West North Central <input type="checkbox"/> South Atlantic <input type="checkbox"/> East South Central <input type="checkbox"/> West South Central <input type="checkbox"/> Mountain <input type="checkbox"/> Pacific <input type="checkbox"/> Military/Military reservations <input type="checkbox"/> PUMS boundaries cross state lines <input type="checkbox"/> State not identified
<a href="#">State</a>	<input type="text" value="Alabama"/> <input type="text" value="Alaska"/> <input type="text" value="Arizona"/> <input type="text" value="Arkansas"/> <input type="text" value="California"/> <input type="text" value="Colorado"/> <input type="text" value="Connecticut"/> <input type="text" value="Delaware"/> <input type="text" value="District of Columbia"/> <input type="text" value="Florida"/>

**Person Variables**

<a href="#">Age</a>	from <input type="text" value="0"/> to <input type="text" value="All"/>
<a href="#">Sex</a>	<input type="checkbox"/> Male <input type="checkbox"/> Female
<a href="#">Race</a>	<input type="checkbox"/> White <input type="checkbox"/> Black/Negro <input type="checkbox"/> American Indian <input type="checkbox"/> Chinese <input type="checkbox"/> Japanese <input type="checkbox"/> Other <input type="checkbox"/> Other, nec
<a href="#">Marital Status</a>	<input type="checkbox"/> Married, spouse present <input type="checkbox"/> Married, spouse absent <input type="checkbox"/> Separated <input type="checkbox"/> Divorced <input type="checkbox"/> Widowed <input type="checkbox"/> Never married, single

FIGURE 7. IPUMS data access system, 1995. Step 3.

Gardner initially estimated that it would take him “a couple of weeks” to transform his fantasy football program into the front end of a data extract system, but the task was harder than he anticipated. It took Gardner approximately six months before the inaugural launch of the website in November 1995, and thereafter, the system was in a continuous cycle of refinement for the next few years [31]. Unlike the early FTP dissemination system and the early IPUMS website, which required a download of entire census samples, the data extraction system allowed researchers to customize their data request to meet their research questions and computing capacity. Researchers followed prompts to select only those subpopulations and variables needed for their analysis and the output was a single data file containing multiple census years with identical custom record layouts [7], [31], [32]. A key element of the extraction system was the user interface that was developed using Perl and JavaScript. The expertise of SSRFC staff member John Easton was the key to untangling some logistical snarls in the computing process. The major design goal was “to make extractions easy while retaining maximum flexibility for users.”<sup>16</sup> At the IPUMS landing page, users were guided through a step-by-step process to make selections about sample, variable, and case characteristics (see Figures 5–7). In 1996, the website was refined to accommodate fully automated online registration and dynamic pages generated from user selections and metadata. Between April 1995 (when the data extract system went live) and 1999, the system processed 10,000 data extract requests [32].

The convergence of increasingly fast and affordable personal computers; the expansion of access to UNIX workstations across universities; and the development of the Internet and the World Wide Web facilitated the dissemination of historical census data. The Internet allowed IPUMS to bypass physical dissemination of the data and the World Wide Web made it possible to customize and automate data extraction in an online environment.

A technological constraint parallel to the dissemination of historical census data was the creation and dissemination of integrated documentation. The sheer volume and inconsistency of the documentation across public use samples made its use unwieldy at best. Early IPUMS documentation, approximately 800 pages in length and contained in 61 separate files, was available in a downloadable Word for DOS 6.0 format.

<sup>16</sup>Todd Gardner, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 7, 2021; Easton interview, Oct. 21, 2021.

A printed version of the documentation was available for purchase at \$30 (see Figure 4). In 1996, Ruggles and Menard obtained funding to convert the documentation into hypertext format accessible on the World Wide Web. The hypertext documentation was integrated into the extract system, enabling researchers to make informed choices about the samples they were requesting [31].

## IPUMS 1996–1999

As the decade progressed, steadily expanding funding from both the NSF and the NICHD continued to support the creation of samples for additional censuses to extend IPUMS [24]. The growth of the census projects stretched the physical space and computing capacities where the data integration work took place.

During this period, the projects occupied three separate spaces on the West Bank of the University of Minnesota. As noted above, the Social History Research Lab began in space provided by the History Department. As the number of census project expanded, the department ran out of space. Desperate, and finding no on-campus solutions from the College of Liberal Arts, Ruggles recalled discovering that

“you can convert grants from on-campus to off-campus. . . If I had money, I could rent space. So, I arranged to rent space and then I just transformed three grants from being on-campus to being off-campus. . . Because we [had] just moved, we were able to shift money from indirect cost to direct costs. And so, we had more money available, and we were allowed to spend it on rent.”<sup>17</sup>

The projects thus rented space in the Cedar-Riverside People’s Center, a block from the campus. In addition to the census projects, the building housed a free medical clinic, a veterinary clinic, a gymnasium featuring nude volleyball, community theater instruction in theatrical combat, and a variety of “alternative” organizations. According to DEO Dianne Star, working at the People’s Center was “interesting, to say the least,” with “all sorts of people going in and out” of the building.<sup>18</sup> The project also found space in scattered on-campus locations. Data entry, data cleaning, processing, dissemination, and administrative functions were simultaneously carried out in buildings based on physical fit rather than function.

<sup>17</sup>Steven Ruggles, interviewed by D. L. Magnuson, University of Minnesota, Jan. 9, 2014.

<sup>18</sup>Dianne Star, interviewed by D. L. Magnuson, University of Minnesota, Sep. 26, 2013.

The People's Center was a particularly challenging work environment for technical rather than social reasons. Initially, even basic phone service was problematic [13].<sup>19</sup> Dial-up modems were used to transfer data, but they were frustratingly slow and unreliable. A thinnet network connection was explored, but the \$80,000 cost was prohibitive.<sup>20</sup> The solution was a wireless link connected by point-to-point microwave antennas. One antenna dish was located on the 12th floor of the Social Science Tower in the Political Science Department's computer lab; the other was positioned on the roof of the People's Center (0.2 mi/0.3 km away). Tom Lindsay, a graduate student liaison between SSRFC and the Social History Research Lab, had a table saw in his garage and built the frame to support the antenna mounted on the People's Center. A six-foot copper spike was buried next to the building with a copper wire run up the side of the building to ground the antenna. The Ethernet connection was threaded down through a skylight into the workspace.<sup>21</sup>

With the success and enthusiastic response by social scientists to the 1995 release of IPUMS, federal funding was obtained to launch a "comprehensive revision" of the database in 1998. Numerous variables were added, coding schemes were "improved and rationalized," and "virtually" all missing and illegible data were allocated through logical editing and "hot deck" probabilistic editing procedures. Furthermore, preliminary samples of 1860 and 1870 and an "expanded and improved" version of the 1920 sample were integrated into the IPUMS database. The extent of available documentation exploded from an 800-page guide to 3000 pages in a five-volume set [33]. As computing capacity increased and storage costs decreased across the decade of the 1990s, what once seemed out of reach was becoming normal.

## CONCLUSION

The IPUMS model proved highly scalable. In 1999, IPUMS expanded to incorporate data from censuses and surveys covering 102 countries, providing information on over a billion additional individuals. The North Atlantic Population Project provided access to international historical census data, including full-count census data from six countries. IPUMS now covers U.S. and

international labor force surveys, time-use surveys, health microdata, and spatial population data.

IPUMS is now the world's largest population database, providing information describing more than three billion persons as well as billions of data points describing the characteristics of places. The size of the data collection is currently about four terabytes. When the project was proposed in 1991, much disk storage would have cost about \$19.5 million in 2021 dollars. A decade earlier in 1981, the same storage would have cost some \$2.6 billion, based on the cheapest prices for hard disk space advertised in *Byte Magazine*. By 2001, the cost of 4 TB of the disk had declined to less than \$14,000, and today (in 2022) one can buy it for \$69.99.

A convergence of technological advances over the past four decades made IPUMS possible. Throughout the 1990s, IPUMS pushed the limits of what was possible with the available computing resources. This involved improvising workarounds of the constraints of data storage, processing, and networking. IPUMS developed key innovations around structured metadata for data integration and web-based data access software. These innovations made large-scale data integration feasible and allowed users with limited computer capabilities to take advantage of an enormous (for the time) data resource.

IPUMS was initially conceived primarily as a resource for historical demography. The audience is now much broader, with over 200,000 registered users from dozens of academic disciplines. The largest contingent of users, accounting for some 40% of the total, are economists, followed by sociologists, demographers, geographers, and public health researchers. The availability of historical and contemporary data in consistent format has contributed to a dramatic increase in qualitative historical research across multiple disciplines [25]. Google Scholar lists 26,700 books, articles, working papers, and dissertations citing IPUMS data, and a new paper appears approximately every 3 h. Beyond the academy, IPUMS has found substantial audiences in policy, planning, and news organizations. By democratizing data access, reducing redundant effort, and enabling investigators to construct customized demographic datasets, IPUMS has opened new research opportunities and stimulated new discoveries. 🧐

## BIBLIOGRAPHY

- [1] A. F. Anderson and L. Neidert, "Dissemination of data: Electronic produces," in *Encyclopedia of the U.S. Census: From the Constitution to the American Community Survey*, M. J. Anderson, C. F. Citro, and J. J. Salvo, Eds., Washington, D.C., USA: CQ Press, 2012, pp. 182–188.

<sup>19</sup>Dianne Star, interviewed by D. L. Magnuson, University of Minnesota, Sep. 26, 2013.

<sup>20</sup>Easton interview, Oct. 21, 2021; Thomas Lindsay, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Nov. 1, 2021.

<sup>21</sup>Lindsay interview, Nov. 1, 2021; Easton interview, Oct. 21, 2021; William Block, interviewed by D. L. Magnuson and S. Ruggles, University of Minnesota, Oct. 27, 2021.

- [2] A. G. Bogue, "The historian and social science data archives in the United States," *Hum. Factors*, vol. 44, no. 4, pp. 976–987, 1976.
- [3] J. Cushing and J. Ortuzar, "ISSA: An integrated system for survey analysis," *Population Index*, vol. 54, no. 3, p. 426, 1987.
- [4] K. Dillard, "Farming in the shadow of the cities: The not-so-rural history of rose township farmers, 1850-1900," *Ramsey County Hist.*, vol. 20, no. 3, pp. 3–19, 1985.
- [5] EFE, "Lab in Switzerland celebrates invention of the World Wide Web 30 years ago." Accessed: Aug. 1, 2022. [Online]. Available: [www.efe.com/efe/english/technology/lab-in-switzerland-celebrates-invention-of-the-world-wide-web-30-years-ago/50000267-3921488](http://www.efe.com/efe/english/technology/lab-in-switzerland-celebrates-invention-of-the-world-wide-web-30-years-ago/50000267-3921488)
- [6] T. Gardner, "Software development," *Historical Methods*, vol. 28, no. 1, pp. 59–62, Winter 1995.
- [7] T. Gardner, S. Ruggles, and M. Sobek, "IPUMS data extraction system," *Historical Methods*, vol. 32, no. 3, pp. 199–224, Summer 1999.
- [8] S. N. Graham, *1900 Public Use Sample User's Handbook*. Seattle, WA, USA: Center Stud. Demography Ecol., Univ. Washington, 1980.
- [9] P. K. Hall et al., "IPUMS metadata: Documenting 150 years of census microdata," *Historical Methods*, vol. 32, no. 3, pp. 111–119, Spring 1999.
- [10] R. Jenkins, *Procedural History of the 1940 Census of Population and Housing*. Drive Madison, WI, USA: Center Demography Ecol., Univ. Wisconsin-Madison, 1983.
- [11] D. Kallgren, "Project budget line-equipment record," ISRD Archive, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. 12/1/1989-5/18/1990, 1880.
- [12] D. L. Magnuson, "Curating our social science infrastructure: The MPC/IPUMS institutional history as a case study," presented at the *Social Sci. Hist. Assoc.*, Baltimore, MD, USA, Nov. 2015.
- [13] D. L. Magnuson, "The spatial evolution of the MPC," Oct. 28, 2016. [Online]. Available: [https://blog.popdata.org/the\\_spatial-evolution-of-the-mpc/](https://blog.popdata.org/the_spatial-evolution-of-the-mpc/)
- [14] D. L. Magnuson, "The evolution of our physical space," May 23, 2018. [Online]. Available: <https://blog.popdata.org/physicalspace/>
- [15] R. McCaa, "Calidad, clase, and endogamy in colonial Mexico: The case of Parral, 1788-1790," *Hispanic Amer. Historical Rev.*, vol. 64, no. 3, pp. 477–502, Aug. 1984.
- [16] R. McCaa, "Microcomputer software designs for historians: Word processing, filing and data entry programs," *Historical Methods*, vol. 17, no. 2, pp. 68–74, Spring 1984.
- [17] T. J. Misa, *Digital State: The Story of Minnesota's Computing Industry*. Minneapolis, MI, USA: Univ. Minnesota Press, 2013.
- [18] "Mosaic—The first global web browser." Accessed: Aug. 2, 2022. [Online]. Available: [www.livinginternet.com/w/wi\\_mosaic.htm](http://www.livinginternet.com/w/wi_mosaic.htm)
- [19] S. Ruggles, *Prolonged Connections: The Rise of the Extended Family in Nineteenth-Century England and America*. Madison, WI, USA: Univ. Wisconsin Press, 1987.
- [20] S. Ruggles, "Integration of the public use samples of the U.S. census," in *Proc. Amer. Stat. Assoc., Social Statist. Sect.*, 1991, pp. 365–370.
- [21] S. Ruggles, "The U.S. public use census microdata files as a source for the study of long-term social change," *IASSIST Quart.*, vol. 15, no. 2, Summer 1991, Art. no. 20, doi: 10.29173/iq703.
- [22] S. Ruggles, "Integrated public use microdata series," NSF, Alexandria, VA, USA, Tech. Rep. SES-9118299, 1992–1995.
- [23] S. Ruggles, "Public use microdata sample of the 1920 census," NICHD-DBSB, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. R01 HD29015, 1993–1998.
- [24] S. Ruggles, "The Minnesota population center data integration projects: Challenges of harmonizing census microdata across time and place," in *Proc. Amer. Stat. Assoc., Government Statist. Sect.*, 2005, pp. 1405–1415.
- [25] S. Ruggles, "The revival of quantification: Reflections on old new histories," *Social Sci. Hist.*, vol. 45, pp. 1–25, Spring 2021.
- [26] S. Ruggles and D. L. Magnuson, "Census technology, politics, and institutional change, 1790-2020," *J. Amer. Hist.*, vol. 107, no. 1, pp. 19–51, Jun. 2020.
- [27] S. Ruggles and R. R. Menard, "A public use sample of the 1880 U.S. census of population," *Historical Methods*, vol. 23, no. 3, pp. 104–124, Summer 1990.
- [28] S. Ruggles and R. Menard, "Public use microdata sample of the 1850 census," Sociol. Division, NSF, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. SBR9210903, 1992–1994.
- [29] S. Ruggles and R. R. Menard, "Public use sample of the 1880 census," NICHD-DBSB, Univ. Minnesota, Minneapolis, MN, USA, Tech. Rep. R01 HD25839, 1989–1993.
- [30] S. Ruggles et al., *Public Use Microdata Sample of the 1880 United States Census of Population: User's Guide and Technical Documentation*. Minneapolis, MO, USA: Social Sci. Hist. Res. Lab., Univ. Minnesota, 1994.
- [31] S. Ruggles, M. Sobek, and T. Gardner, "Distributing large historical census samples on the internet," *Hist. Comput.*, vol. 9, pp. 145–159, 1996.
- [32] S. Ruggles, M. Sobek, and T. Gardner, "Disseminating historical census data on the World Wide Web," *IASSIST Quart.*, vol. 20, no. 3, 1996, Art. no. 4, doi: 10.29173/iq68.

- [33] M. Sobek and S. Ruggles, "The IPUMS project: An update," *Historical Methods*, vol. 32, no. 3, pp. 102–110, Summer 1999.
- [34] M. A. Strong, S. H. Preston, and M. C. Hereward, "An introduction to the public use sample of the 1910 U.S. census of population," *Historical Methods*, vol. 30, no. 2, pp. 54–56, Apr. 1989.
- [35] University of Minnesota, *Minnesota Project Woksape: Innovation, Learning, Wisdom. IBM Advanced Education Project*. Minneapolis, MN, USA: Univ. Minnesota, 1989. [Online]. Available: <https://conservancy.umn.edu/handle/11299/207708>
- [36] University of Minnesota, "West bank computer center," *West Bank Buffer*, Nov. 1973. [Online]. Available: <https://conservancy.umn.edu/handle/11299/162893>
- [37] University of Minnesota, "The president's report, 1964-1966," 1966. [Online]. Available: <https://hdl.handle.net/11299/102308>
- [38] U.S. Bureau of the Census, *Census of Population, 1940: Public Use Sample Technical Documentation*. Washington, D.C., USA: U.S. GPO, 1984.
- [39] U.S. Bureau of the Census, *Census of Population, 1950: Public Use Sample Technical Documentation*. Washington, D.C., USA: U.S. GPO, 1984.



**PURPOSE:** Engaging professionals from all areas of computing, the IEEE Computer Society sets the standard for education and engagement that fuels global technological advancement. Through conferences, publications, and programs, IEEE CS empowers, guides, and shapes the future of its members, and the greater industry, enabling new opportunities to better serve our world.

**OMBUDSMAN:** Email [ombudsman@computer.org](mailto:ombudsman@computer.org)

**MEMBERSHIP & PUBLICATION ORDERS**  
**Phone:** +1 800 272 6657;  
**Fax:** +1 714 816 2121;  
**Email:** [help@computer.org](mailto:help@computer.org)

revised  
 25 July 2023



**EXECUTIVE COMMITTEE**

<b>President:</b>	Nita Patel
<b>President-Elect:</b>	Jyotika Athavale
<b>Past President:</b>	William D. Gropp
<b>First VP:</b>	Hironori Washizaki
<b>Second VP:</b>	Grace A. Lewis
<b>Secretary:</b>	Carolyn McGregor
<b>Treasurer:</b>	Michela Taufer
<b>VP, Membership &amp; Geographic Activities:</b>	Fernando Bouche
<b>VP, Professional &amp; Educational Activities:</b>	Deborah Silver
<b>Interim VP, Publications:</b>	Greg Byrd
<b>VP, Standards Activities:</b>	Annette Reilly
<b>VP, Technical &amp; Conference Activities:</b>	Grace A. Lewis
<b>2023–2024 IEEE Division VIII Director:</b>	Leila De Floriani
<b>2022–2023 IEEE Division V Director:</b>	Cecilia Metra
<b>2023 IEEE Division V Director-Elect:</b>	Christina M. Schober

**BOARD OF GOVERNORS**

**Term Expiring 2023:** Jyotika Athavale, Terry Benzel, Takako Hashimoto, Irene Pazos Viana, Annette Reilly, Deborah Silver  
**Term Expiring 2024:** Saurabh Bagchi, Charles (Chuck) Hansen, Carlos E. Jimenez-Gomez, Daniel S. Katz, Shixia Liu, Cyril Onwubiko  
**Term Expiring 2025:** İlkay Altıntaş, Nils Aschenbruck, Mike Hinchey, Joaquim Jorge, Rick Kazman, Carolyn McGregor

**EXECUTIVE STAFF**

<b>Executive Director:</b>	Melissa Russell
<b>Director, Governance &amp; Associate Executive Director:</b>	Anne Marie Kelly
<b>Director, Conference Operations:</b>	Silvia Ceballos
<b>Director, Information Technology &amp; Services:</b>	Sumit Kacker
<b>Director, Marketing &amp; Sales:</b>	Michelle Tubb
<b>Director, Membership Development:</b>	Eric Berkowitz
<b>Director, Periodicals &amp; Special Projects:</b>	Robin Baldwin

**IEEE BOARD OF DIRECTORS**

<b>President &amp; CEO:</b>	Saifur Rahman
<b>President-Elect:</b>	Thomas M. Coughlin
<b>Director &amp; Secretary:</b>	Forrest (Don) Wright
<b>Director &amp; Treasurer:</b>	Mary Ellen Randall
<b>Past President:</b>	K. J. Ray Liu
<b>Director &amp; VP, Educational Activities:</b>	Rabab Ward
<b>Director &amp; VP, Publication Services &amp; Products:</b>	Sergio Benedetto
<b>Director &amp; VP, Member &amp; Geographic Activities:</b>	Jill Gostin
<b>Director &amp; President, Standards Association:</b>	Yu Yuan
<b>Director &amp; VP, Technical Activities:</b>	John Verboncoeur
<b>Director &amp; President, IEEE-USA:</b>	Eduardo Palacio

# Technology Trends and Challenges for Large-Scale Scientific Visualization

James Ahrens, *Los Alamos National Laboratory, NM, 87545, USA*

*Scientific visualization is a key approach to understanding the growing massive streams of data from scientific simulations and experiments. In this article, I review technology trends including the positive effects of Moore's law on science, the significant gap between processing and data storage speeds, the emergence of hardware accelerators for ray-tracing, and the availability of robust machine learning techniques. These trends represent changes to the status quo and present the scientific visualization community with a new set of challenges. A major challenge involves extending our approaches to visualize the modern scientific process, which includes scientific verification and validation. Another key challenge to the community is the growing number, size, and complexity of scientific datasets. A final challenge is to take advantage of emerging technology trends in custom hardware and machine learning to significantly improve the large-scale data visualization process.*

Visually understanding the growing massive volume of scientific data from sensors and supercomputers is an ongoing challenge. In this Visualization Viewpoints article, I will discuss technology trends and challenges for large-scale scientific visualization. I was inspired to write the column by Chris Johnson's Visualization Viewpoints article titled "Top Scientific Visualization Research Problems."<sup>1</sup> I believe it is a good time for an evaluation of large-scale scientific visualization challenges due to the expected delivery of exascale supercomputers in the next few years and the explosion of high-resolution scientific sensors and instruments. As Johnson writes in his article, "it is important to refresh . . . regularly and to add new viewpoints. . . ." This article fundamentally differs from the list offered by Johnson in that this article does not attempt to make a complete list but focuses only on what I consider the most important key trends and challenges. I focus on trends, that is, software or hardware advances that originate from outside the field of large-scale visualization but have a significant impact on the field. Trends I address include the effects of Moore's

law on the scientific process, the order of magnitude differences between processor and storage speeds, and custom hardware and machine learning advances. I also discuss challenges, that is, problems to overcome. Note that challenges can be reframed as opportunities especially when one finds an exemplary solution.

## THE POSITIVE EFFECT OF MOORE'S LAW ON SCIENCE

The first technology trend is our ability to quickly generate and process massive scientific data due to Moore's law. Moore's law states that the number of transistors on a microchip doubles approximately every 18 to 24 months. This doubling effect has translated into faster processors that can simulate and process data extremely quickly. Moore's law has also driven a revolution in sensor technology, with sensors achieving higher resolution and better dynamic range over time.

The effect of our increasing computational and data capabilities on science is described in Hey *et al.*'s 2009 article, "Jim Gray on eScience."<sup>2</sup> Jim Gray identified four paradigms of science that evolved over time:

- 1) Empirical—thousands of years ago, science was mostly empirical, describing natural phenomena;

- 2) Theoretical—in the last few hundred years, science added theoretical approaches, including the use of models and generalizations;
- 3) Computational—in the last seven decades, science added computational simulations, to simulate complex phenomena;
- 4) Data exploration—recently, data exploration was added, which unifies (1–3) experiment, theory, and simulation.

Many scientific advances of the past decades can be linked to this revolution. Chen *et al.* describe some of these advances and identify challenges that can be addressed by the synergistic use of data intensive science and exascale computing.<sup>3</sup>

Gray's paradigms provide us a starting point for a better understanding of the modern scientific process. Figure 1 presents a hierarchy of scientific data representation levels that I believe provides a more complete picture. Relating this model to Gray's paradigms, a single experiment or simulation ensemble (represented by the data element level up to the ensemble level) utilizes the empirical, theoretical, and computational paradigms. The ensemble level up to the validation level encapsulates a unifying data exploration paradigm.

The modern scientific process is characterized by the increasing pace of scientific data generation via automated scientific experimental and simulation processes on larger and more capable experimental and supercomputing facilities. It involves multiple teams of people across the world. In my opinion, a corresponding modern scientific visualization approach needs to provide effective techniques for all the levels of the modern scientific process outlined in Figure 1. I would further assert that the visual understanding of the validation and verification process using experimental and simulation ensembles is central to the future of scientific visualization and scientific advances.

## CHALLENGES BASED ON THE MODERN SCIENTIFIC PROCESS

Our overall challenge is to create visualizations that quickly increase our understanding of scientific data for this modern scientific process. In this section, I will review the levels, describe how the state of the art in scientific visualization addresses these challenges, and identify new challenges.

*Data element level*—Starting at the base level, each data element in a simulation has a type such as scalar or vector. Significant effort has gone into successfully creating visualization approaches for scalar

and vector visualizations. Simulations are evolving to incorporate new more complex types of data elements. Challenges include visual representations for higher dimensional data element types including tensors and functions.













*Data structure and variable level*—The spatial data structure that uses data elements and represents both the structure and topology of the represented physical system such as a structured or unstructured meshing representation. Challenges include full direct visualization support for complex simulation data structures such as adaptive mesh refinement strategies and complex unstructured mesh types including higher order meshes. As an example, this means for adaptive mesh refinement strategies, visualizing the adaptive grid properties, such as grid resolution, knowing how the boundaries between grids of differing resolutions are handled (e.g., for correct isosurfacing) and how a data location is correctly sampled from the adaptive grid. Direct support for these representations is needed both to debug simulations as well as create accurate visualizations of these data structures. At this level, many variables needed for the simulation are represented using these data structures.

*System level*—All data that are needed to represent a single run of a simulation or experiment. Typically for a simulation, this consists of an input deck, associated input scientific databases, the specific version of the simulation code, and the resulting data structure level variable outputs over time. Note that a simulation can be thought of as a set of independent variables (the inputs) and the dependent variables (the outputs). Challenges include visually understanding the multivariate and temporal nature of results, including visually identifying what variables are most relevant and what relationships exist between input and output variables.

*Ensemble level*—A set of system level input/output results. Typically, such a set is selected using a statistical experimental design. Statistical experimental design is used to carefully choose input values to ensure coverage of the entire input/output space of simulations or experiments. Starting at this level and at all higher levels of abstraction, there is not currently a common data format to represent or visualize data at these levels. Without these data representations, it is difficult to build and share needed visualization methods for these higher level concepts. A challenge at this level is to develop effective ensemble visualization representations and methods.<sup>4</sup>

*Verification level*—Verification is the process of checking whether simulations or experiments meet their design requirements (e.g., Are they working



The Modern Scientific Process						
Validation level	Data: Overall scientific knowledge					
	Action: The simulation and experimental verification data is validated. Validation is the process of checking whether simulations match real-world experimental results. At this level the quality of overall scientific knowledge is assessed.					
Verification level	Data: Verified collection of simulations			Data: Verified collection of experiments		
	Action: Simulations are verified. Verification is the process of checking whether simulations meet their design requirements. The quality of knowledge is strengthened thru the use and comparison of a collection of ensembles of results from different simulations.			Action: Experiments are verified. Verification is the process of checking whether experiments meet their design requirements. The quality of knowledge is strengthened thru the use and comparison of a collection of ensembles of results from different experiments.		
Ensemble Level	Data: Ensembles of results from a set of independently developed simulations using different approaches and solvers.			Data: Ensembles of results from a set of independently developed experiments using different approaches and sensors.		
	Action: Simulation type 1, 2 to N are run with an ensemble of inputs.			Action: Experiment type 1, 2 to M are run with an ensemble of inputs.		
System Level						
	Data: Simulation type 1			Data: Experiment type 1		
Data Structure and Variable level	Action: Create all the information needed to represent a single simulation run such as the input parameters, associated scientific databases and simulation code that produces time-varying multi-variate outputs.			Action: Create all the information needed to represent a single experimental run such as descriptions of the scientific apparatus and sensors, associated sensor code and input parameters that produces time-varying multi-variate outputs.		
						
Data Element Level	Data: Scalar, Vector and/or Tensor or Function per element					
	Action: Create an element representation of the data					

**FIGURE 1.** A representation of the levels in the modern scientific process. A challenge for the visualization community is to explore the best visualization approaches for each of these levels as well as to be able to track, connect, and visualize relationships between levels. The levels increase in abstraction and complexity from the lowest level (data element) to the highest level (validation level). The images in each data row are symbolic. At the data structure and variable level, the images visually represent the heterogeneity of grid types for simulations, and differing types of sensors for experiments. At the system level, the images represent different simulation and experiment types that are ready to run. At the ensemble level, the images represent a collection of ensembles of results from the simulations and experiments. The different colors are meant to represent different simulation and experiment results.

correctly?). Debugging a simulation is a verification process. There is a diverse set of ensemble data at this level. The diversity derives from differences at the underlying levels (e.g., different simulations, data structures, or data elements). These collections form independent evidence for our scientific knowledge.

**Validation level**—At this level, there are collections of ensembles from simulations and experiments. These data taken collectively provide the current state of our understanding about the scientific area of study. Independently formulated simulations and experiments reduce the bias of any single one of these activities. Validation is the process of checking whether simulations match real-world experimental results.

Verification and validation challenges include the use of high-dimensional visualization techniques to visualize these simulation and experimental ensembles and the comparison between ensembles and ensemble members.<sup>5</sup> Research efforts on visualizing uncertainty, feature extraction, and comparative visualization provide a good starting point to address these challenges.

To deeply understand the reasons for differences at the validation level, a modern visualization tool should be able to seamlessly “move” between different levels of the scientific process, from the abstract ensemble levels at the top of the hierarchy to the spatial/temporal representations at the bottom of the hierarchy. For example, a high-dimensional information visualization technique may highlight a major difference between the output of two simulations in an ensemble. This difference would likely be best visualized via a spatially based scientific visualization technique. One idea for a potential solution to this challenge is a unified programming language that encapsulates both information and scientific visualization concepts. A good starting point could be Satyanarayan *et al.*'s<sup>6</sup> interactive programming language based on the grammar of graphics.

### OUR ABILITY TO STORE DATA IS ORDERS OF MAGNITUDE SLOWER THAN OUR ABILITY TO GENERATE DATA

A second technology trend is that since the 1980s, memory and storage devices improve at a much slower rate than processors. It takes orders of magnitude longer to store data than to produce it. This trend motivates reducing data from simulation or observational sensors as close to the source of data as possible. For sensors, this challenge has led to the development of “edge computing” solutions. A related challenge that arises as a result of the large volumes of data produced

by simulations and experiments is the need to store and process these results in a timely manner. One approach is to work to reduce the size of the data without reducing the data’s usefulness. Numerous approaches have been applied including compression, sampling, and multiresolution representations. For example, approximate solutions derived via sampling are described by Moritz *et al.*<sup>7</sup> and Biswas *et al.*<sup>8</sup>

An important area for consideration is efforts to quantify how these reduced representations will affect the visualizations that will be applied to this data. An additional challenge for scientific data is that for spatial datasets, preserving topological<sup>9</sup> and connectivity relationships can be as important as preserving data accuracy.

A promising solution for understanding large-scale scientific simulation data is to immediately visualize the data on the same supercomputing resource.<sup>10</sup> This approach is known as *in situ* visualization.<sup>11</sup> *In situ* visualization typically occurs without human intervention during batch runs. Key decisions are required about what variables to visualize, how to visualize them, and from what viewpoint. A challenge is to automate the *in situ* visualization process by selecting effective parameters. Zhu *et al.* provide a survey of automated techniques for information visualization.<sup>12</sup> Applying these approaches for scientific visualization would be a good starting point for work in this area. An additional challenge is to support visual exploration after the simulation has finished. Research efforts have focused on saving visualization extracts such as a collection of imagery for later exploration.<sup>13</sup>

### NEW TYPES OF CUSTOM HARDWARE FOR RAY-TRACING AND MACHINE LEARNING ARE BECOMING AVAILABLE

A third trend in response to the expected deceleration in Moore’s law is to optimize the transistors in a chip design to solve a specific problem, making such a custom chip more efficient than a general-purpose solution for the specific problem. Examples of this trend include the creation of ray-tracing and machine learning accelerators. Using these hardware accelerators for large-scale visualization and rendering poses significant challenges. For example, traditional visualization pipelines have typically generated geometry such as triangles that are passed to a renderer to render into an image. Geometry takes up memory space. One intriguing possibility is to avoid the creation of geometry and instead ray-trace imagery directly from the scientific data.<sup>14</sup>

## TAKING ADVANTAGE OF ADVANCES IN MACHINE LEARNING TO SOLVE SCIENTIFIC VISUALIZATION PROBLEMS

A fourth trend is the emergence of new machine learning techniques and their successful application to a broad variety of challenges.<sup>15</sup>

Two challenges arise from this trend. The first challenge is the use of visualization to explain the results of machine learning applications. This challenge is closely related to visualizing ensembles, since machine learning techniques are strongly data driven, and these machine learning data are typically ensembles themselves. Ideally, the strong interest in visualizing machine learning results will translate into more effective ensemble visualization approaches.

The second challenge is the application of machine learning techniques as part of the scientific visualization process. A significant success has been in the application of supervised learning techniques such as deep learning. Supervised learning techniques solve a classification or regression problem using an ensemble of data with well-defined inputs and outputs, and produce labeled or numerical outputs. Unsupervised learning techniques solve a clustering or dimensionality reduction problem using an ensemble of data with well-defined inputs, and produce a reduced representation for these inputs.

At a high level, we can think about machine learning being applied to the scientific data we are visualizing or the visualization process itself. When supervised machine learning is applied to the scientific data it can provide a “surrogate” version of the scientific data. The surrogate supports interpolation over the space of simulations at unknown input locations. Such a representation is useful because it supports sampling the scientific data space with “unknown” inputs in order to create approximate or reduced representations of the scientific data. For example, this approach might be used in ensemble analysis for verification and validation to help understand the space of all possible results from a small finite set of representative results. Surrogates could be used to create a shared basis for visually comparing ensembles.

When unsupervised machine learning is applied to the scientific data, it can help to identify similarities between elements on each level of the hierarchy (e.g., clustering similar data elements, data structures, and ensemble members). It can also help identify representative but reduced basis functions from the input variables through dimensionality reduction operations (such as using Principal Component Analysis to create a 2-D basis function for plotting data elements).

To apply machine learning techniques to automate the selection of visualization parameters for the visualization process, the space of all visualization processes needs to be defined, as well as a metric to evaluate the desired property of the visualization process. For example, if we want to use machine learning to help select an “optimal camera” for a visualization, the space is all possible camera positions (as defined by a set of camera parameters) and a metric for an optimal view could be a computable value such as image entropy<sup>16</sup> or defined by the results of a user study that evaluates the effectiveness of different camera positions. One reason for the success of machine learning in industry is that companies are able to gather significant amounts of user choice data via their web interface to their products. As visualization tools move to web-based interfaces and more users interact with visualization tools in this manner, the opportunity to gather more user visualization choice data will increase. This in turn will enable visualization researchers to use this data for machine learning to help automate visualization process choices.

## CONCLUSION

Best practices in large-scale scientific visualization include tracking technology trends and identifying how these trends will affect the scientific process. Ideally these trends can also be leveraged to help scientists cope with these changes. Examples include automating labor-intensive tasks such as running simulations and experiments, and curating and visualizing the generated scientific data to support scientific understanding and decision making. Technology trends such as extremely capable supercomputers and advances in machine learning offer a bright future for researchers and practitioners to creatively address the challenges highlighted in this Visualization Viewpoint. 🌐

## ACKNOWLEDGMENTS

The author would like to thank the anonymous reviewers for their helpful feedback. This work was supported by the Exascale Computing Project 17-SC-20-SC, a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration under Award number 14-017566 to the Los Alamos National Laboratory.

## REFERENCES

1. C. Johnson, “Top scientific visualization research problems,” *IEEE Comput. Graphics Appl.*, vol. 24, no. 4, pp. 13–17, Jul./Aug. 2004.

2. T. Hey, S. Tansley, and K. Tolle, "Jim Gray on eScience: A transformed scientific method," in *The Fourth Paradigm: Data-Intensive Scientific Discovery*. New York, NY, USA: The Fourth Paradigm, 2009.
3. J. Chen *et al.*, "Synergistic challenges in data-intensive science and exascale computing," in *Department of Energy Advanced Scientific Computing Advisory Committee Data Subcommittee Report*. WA, DC, USA: DOE Office of Science, 2013.
4. J. Wang, S. Hazarika, C. Li, and H. W. Shen, "Visualization and visual analysis of ensemble data: A survey," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 9, pp. 2853–2872, Sep. 2018.
5. E. Bertini, A. Tatu, and D. Keim, "Quality metrics in high-dimensional data visualization: An overview and systematization," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2203–2212, Dec. 2011.
6. A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer, "Vega-lite: A grammar of interactive graphics," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 341–350, Jan. 2016.
7. D. Moritz, D. Fisher, B. Ding, and C. Wang, "Trust, but verify: Optimistic visualizations of approximate queries for exploring big data," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 2904–2915.
8. A. Biswas, S. D. A., E. Lawrence, J. Patchett, J. Calhoun, and J. Ahrens, "Probabilistic data-driven sampling via multi-criteria importance analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 12, pp. 4439–4454, Dec. 2020.
9. J. Tierny, G. Favelier, J. Levine, and M. Michaux, "The topology toolkit," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 832–842, Jan. 2017.
10. U. Ayachit *et al.*, "Paraview catalyst: Enabling in situ data analysis and visualization," in *Proc. 1st Workshop Situ Infrastructures Enabling Extreme-Scale Anal. Vis.*, 2015, pp. 25–29.
11. H. Childs *et al.*, "A terminology for in situ visualization and analysis systems," *Int. J. High-Perform. Comput. Appl.*, vol. 34, no. 6, pp. 676–691, 2020.
12. S. Zhu, G. Sun, Q. Jiang, M. Zha, and R. Liang, "A survey on automatic infographics and visualization recommendations," *Vis. Inform.*, vol. 4, no. 3, pp. 24–40, 2020.
13. J. Ahrens, S. Jourdain, P. O'Leary, J. Patchett, D. Rogers, and M. Petersen, "An image-based approach to extreme scale in situ visualization and analysis," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, 2014, pp. 424–434.
14. I. Wald *et al.*, "OSPRay-a CPU ray tracing framework for scientific visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 931–940, Jan. 2016.
15. M. Raghu and E. Schmidt, "A survey of deep learning for scientific discovery," 2020, *arXiv:2003.11755*.
16. N. Marsaglia, Y. Kawakami, S. Schwartz, S. Fields, and H. Childs, "An entropy-based approach for identifying user-preferred camera positions," in *Proc. IEEE 11th Symp. Large Data Anal. Vis.*, 2021, pp. 73–83.

**JAMES AHRENS** is the director of the Information Science Technology Institute, Los Alamos National Laboratory, NM, 87545, USA. He is also the Department of Energy Exascale Computing Project (ECP) Data and Visualization lead for seven storage, data management, and visualization projects that will be a key part of a vibrant exascale supercomputing application and software ecosystem. His research interests include visualization, data science, and parallel computing. Ahrens received the Ph.D. degree in computer science from the University of Washington. He is a member of the IEEE and the IEEE Computer Society. Contact him at [ahrens@lanl.gov](mailto:ahrens@lanl.gov).

Contact department editor Theresa-Marie Rhyne at [theresamarierhyne@gmail.com](mailto:theresamarierhyne@gmail.com).

## DEPARTMENT: DATA

# Data Science: Hype and Reality

Norita Ahmad, Areeba Hamid, and Vian Ahmed, *American University of Sharjah*

*Data science is considered a young field by many. This column shares the growing trends of data science as one of the most sought-after career options and as an emerging discipline in almost every industry in the world.*

In 2012, the *Harvard Business Review* caused a stir by calling data scientist “the sexiest job of the 21st century.”<sup>1</sup> The denomination “data scientist” refers to a profession that makes sense of the vast amount of big data. However, scientists, statisticians, computer scientists, librarians, and other professions have been analyzing and “making sense” of data for ages. As such, the term “data science” is not new and can be traced back to 1962 when John W. Tukey published a book titled *The Future of Data Analysis*.<sup>2</sup> In his book, Tukey, one of the most influential statisticians of the 20th century, suggested that statistics is “pure mathematics,” but data analysis is “intrinsically an empirical science,” and, therefore, it should take the characteristics of science rather than mathematics.<sup>2</sup> As such, Tukey acknowledged that the two are related but are separate disciplines, and to make progress in data analysis, it is important to focus on the tools and attitudes.

Today we’re witnessing explosive growth in the field. This increase can be attributed to the growing amount of data generated by digital activities in our lives. According to the International Data Corporation, more than 59 zettabytes of data were captured in 2020, and the number is expected to increase with a five-year compound annual growth rate of 26%.<sup>3</sup> Data science enables companies not only to understand data from multiple sources but also to enhance decision making. As a result, data science is widely used in almost every industry, including health care, finance, marketing, banking, city planning, and more. With advances in technology, new approaches to the field, and people’s positive attitudes toward data science, there is every reason to believe that the field will continue to grow in the future.

## WHAT IS DATA SCIENCE?

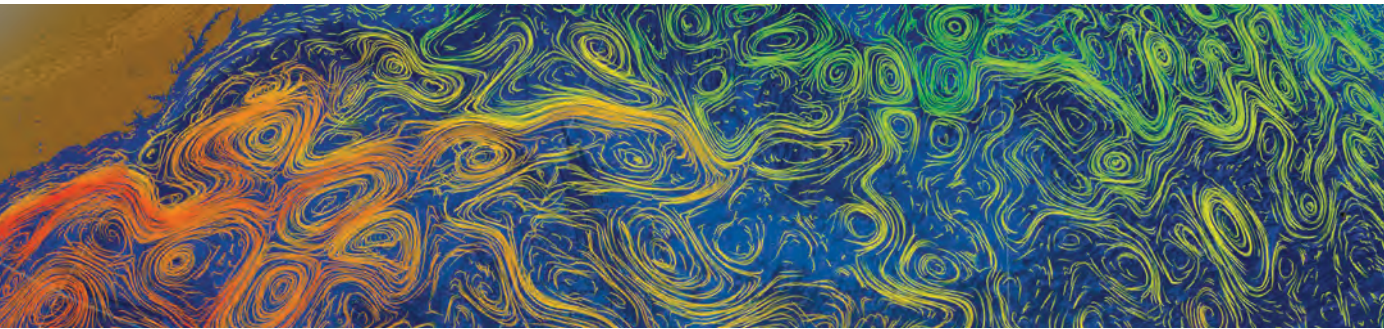
The term *data science* is so widely used today that its definition has become blurry. Some associate it with computer science and some with statistics; most frequently, it is linked to machine learning (ML) and data mining.<sup>4</sup> ML deals with algorithms for extracting patterns from data, while data mining pertains to the analysis of structured data. Data science takes both of these tasks into account in addition to other data-related tasks such as capturing, cleaning, and transforming unstructured data; the use of big data technologies; and handling data ethics and regulation.<sup>4</sup> Besides these engineering-oriented fields, data science can also be business oriented in the form of business intelligence and analytics. Moreover, the data science umbrella consists of back-end and front-end data science.<sup>5</sup> The back-end part is often referred to as *data engineering* and deals with hardware, computing, and data storage infrastructure. The front end focuses more on data analysis and ML.

Perhaps a better way of describing this vast topic is by describing what it is not.<sup>6</sup> Data science is not all about using data for prediction or merely about data analysis. It is not a discipline confined to science, technology, engineering, and mathematics fields, and, most importantly, it is not a single discipline at all. Like other sciences, it is best understood as a “collection of disciplines with complementary foundations, perspectives, approaches, and aims, but with a shared grand mission,” which is to use information and technologies to advance human society.<sup>6</sup>

---

Digital Object Identifier 10.1109/MC.2021.3130365

Date of current version: 14 February 2022



## THE EVOLUTION OF DATA SCIENCE

Tukey's work *The Future of Data Analysis* is known as a seminal moment in the history of data science. Tukey's impact on data science is immense. Besides Tukey, there are a few other prominent names and events that are worth examining.

For example, in 1974, Peter Naur, a well-known Danish computer scientist, published a book that was based on a survey of contemporary data processing methods in Sweden and the United States. Here, he defined data science as "the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated

---

*TUKEY ACKNOWLEDGED THAT THE TWO ARE RELATED BUT ARE SEPARATE DISCIPLINES, AND TO MAKE PROGRESS IN DATA ANALYSIS, IT IS IMPORTANT TO FOCUS ON THE TOOLS AND ATTITUDES.*

---

to other fields and sciences."<sup>7</sup> Following the success of his 1962 work, Tukey published a book in 1977 titled *Exploratory Data Analysis*, another important milestone in the field.<sup>8</sup> In the same year, the International Association for Statistical Computing was established as part of the International Statistical Institute, linking traditional statistical methods with computer technology in converting data into information and knowledge.

In 1996, the term data science was included for the first time in the title of the biennial conference of the International Federation of Classification Societies.<sup>9</sup> A year later, in his inaugural lecture for the H.C. Carver Chair in Statistics at the University of Michigan, Prof. Jeff Wu called for statistics to be renamed data science and the statisticians to be called data scientists.<sup>10</sup> In 2001, Prof. William Cleveland, a computer

scientist and professor of statistics at Purdue University, published an article titled "Data Science: An Action Plan for Expanding the Technical Areas of the Fields of Statistics." Just like Tukey, Cleveland was promoting the idea of merging various fields such as computer science and statistics.<sup>11</sup>

One of the most significant events in the field of data science was the launch of the *Data Science Journal* by the Committee on Data for Science and Technology of the International Council for Science in 2002.<sup>12</sup> The journal serves as a platform for everyone interested in data to present their work and exchange ideas. In 2005, the National Science Board published a report titled "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," which defines data scientists as "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection."<sup>13</sup>

In January 2009, in an interview with *McKinsey Quarterly*, Hal Varian, Google's chief economist, said that the occupation of statistician would be a sexy job in the next 10 years.<sup>14</sup> Then, in 2011, D.J. Patil, a renowned mathematician and computer scientist, wrote an article, "Building Data Science Teams," which many perceived to be the start of data science as the distinct professional specification known as data scientist today.<sup>15</sup> From there on, data science has been covered at length and started to gain more attention as businesses recognized the importance of data.

## THE EVOLUTION OF THE DATA SCIENTIST

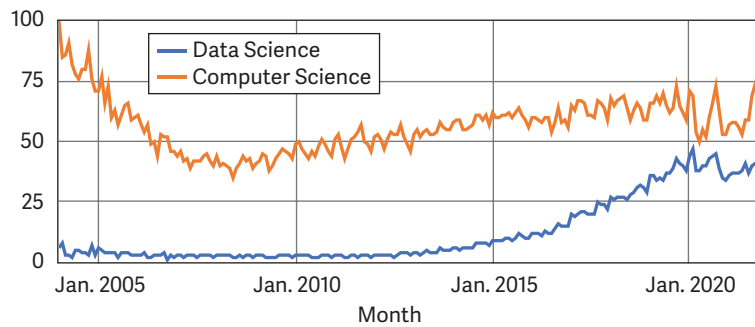
We can see that data science is not an entirely new field. It was derived from already existing topics. Statistics and computer science are the most closely related to what is now called *data science*. In fact, computer science is the top related query for data

science on Google Trends as of 2021. Over the past five years, the popularity of both fields has also risen similarly (see Figure 1). Computer science is also most relevant for data science software experts and developers. In early data science academic programs, most higher education institutions offered data science programs using existing courses with faculty from academic areas such as computer science and statistics, with some shifts in teaching and professional development.<sup>5,16</sup>

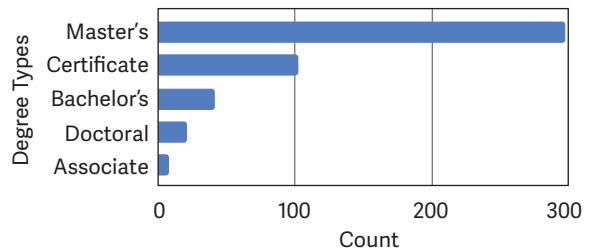
However, with the growing popularity and diversity of data science, institutions have created dedicated data science programs. In addition, many other fields also infuse data science across their disciplines to introduce data literacy and data analytics as important skills among their students.<sup>4,5</sup> One such discipline is business. The science of business statistics was virtually nonexistent until Silicon Valley tech giants started using big data.<sup>17</sup> The field of business analytics and intelligence has vastly grown ever since businesses started using data in the form of *insights data science* for strategic decision making, *product data science* for product testing and optimization, and *engineering data science* for business. Engineering data science includes ML engineers, data engineers, and analytical data engineers. These engineers are differentiated from other data scientists because they produce, rather than synthesize, data.<sup>18</sup>

## DATA SCIENCE EDUCATION AND CAREERS

Upon analyzing data from 465 U.S. data science programs,<sup>19</sup> the following conclusions can be drawn. Master's programs, at 63.6%, are the most common higher education programs for data science (see Figure 2). This conclusion also coincides with worldwide Google Trends for data science programs (see Figure 3), where a Master's degree program is more popular than any other type. Similarly, results from [datascienceprograms.org](http://datascienceprograms.org), which analyzed more than 500 universities in the United States, also found Master's degrees to be the most popular in the field.<sup>20</sup> The high



**FIGURE 1.** Google Trends for worldwide interest in data science versus computer science over time as of 29 October 2021. 100 designates the peak popularity of the term.

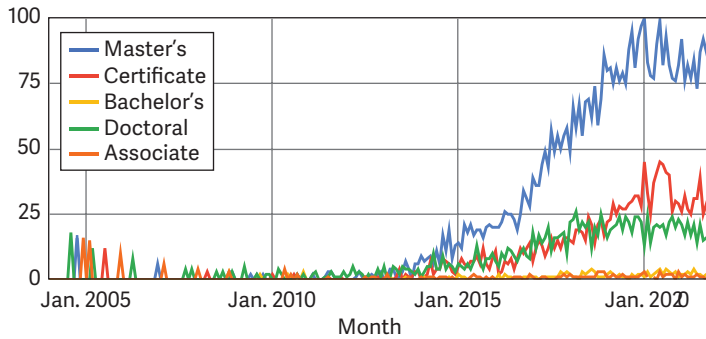


**FIGURE 2.** Counts of data science program degrees.

demand for Master's programs could be credited to the need for skills that other programs lack, the fact that Master's students already have relevant experience, and the growing popularity of data science professionals within the areas of business.<sup>5,21</sup>

Popularity within the business field can be seen through the high number of business intelligence and analytics programs offered, which is the most common program category in Rawlings-Goss's data (21.9%), as well as master of business administration programs (7.1%), most of which also have concentrations in business analytics (see Figure 4).<sup>19</sup> Another emerging field in data science education, as seen through these data, is health-care/biomedical informatics and information management. This is one of the fastest growing fields in the health-care sector today, requiring the acquisition, storage, retrieval, and use of medical data.<sup>22</sup> The data of Rawlings-Goss also show that fields such as computer science and statistics, from which data science heavily derives, continue to offer data science programs as well.<sup>19</sup>

Degrees in data science often target areas such as data wrangling and mining, data visualization, ML,



**FIGURE 3.** Google Trends for worldwide interest in data science programs over time as of 28 October 2021. 100 designates the peak popularity of the term.

programming (with R and Python being the most popular languages), and probability and statistics, whether offered as courses/certifications or dedicated degrees in these topic areas.<sup>20</sup> It should also be noted that, even though a variety of data science degrees is offered, there is no single organization that accredits data science programs. However, some organizations accredit related program areas such as the Accreditation Board for Engineering and Technology for computer technology, analysis, and engineering and the Association to Advance Collegiate Schools of Business for business analytics.<sup>20</sup> The top five universities for data science by region, based on research performance in data science from 14,160 universities in 183 countries, are listed in Table 1.<sup>23,24</sup>

There is a variety of educational approaches to data science, but many argue that theoretical education is not enough. To become a successful data scientist, one requires not only a knowledge

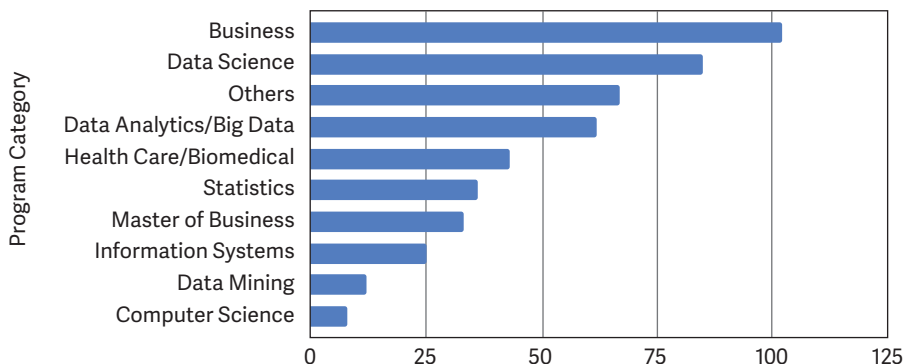
of theory and tools but also sufficient “real-world” experience to know how to reach, and when to trust, results.<sup>25</sup>

In general, data science jobs can be divided into either data generalists or specialists (see Figure 5 for details). Data generalists are commonly referred to as *data scientists* because they are broadly trained and “wear many hats” in the field of data science. While many organizations still look for generalists, the trend

toward data specialists is now increasing as companies use data science more specifically.<sup>19</sup> LinkedIn’s “Jobs on the Rise” reports from different regions around the world indicate that data analytics professionals, data analysts, and artificial intelligence (AI) professionals are the most in-demand roles.<sup>26</sup> Figure 5 also shows countries that listed each type of data science role among their top jobs of 2021. By 2026, it was reported that 11.5 million new data science jobs will be created in the United States alone.<sup>27</sup> The trends indicate that jobs are expected to continue to grow in data science.

### THE FUTURE OF DATA SCIENCE AND DATA SCIENTISTS

According to the U.S. Bureau of Labor Statistics, the average annual salary for a data scientist in 2021 was US\$103,930 in the United States.<sup>27</sup> A 2020 Burtch Works study also shows that the median salary for



**FIGURE 4.** Counts of categories of programs offered in data science in the United States.



data scientists is US\$95,000 at entry level, US\$130,000 at midlevel (US\$195,000 for managerial positions), and US\$165,000 for experienced data scientists (US\$250,000 for managerial positions).<sup>28</sup> Glassdoor also reports the average salary of a data scientist to be US\$117,212. This is higher than the average pay of US\$74,239 for programmers or US\$88,989 for statisticians, for example.<sup>29–31</sup> The high salary amount is attributed to the supply of data professionals still not catching up with the demand. Data scientists also require a high level of expertise and often acquire advanced degrees, as noted earlier with the popularity of Master's degrees.

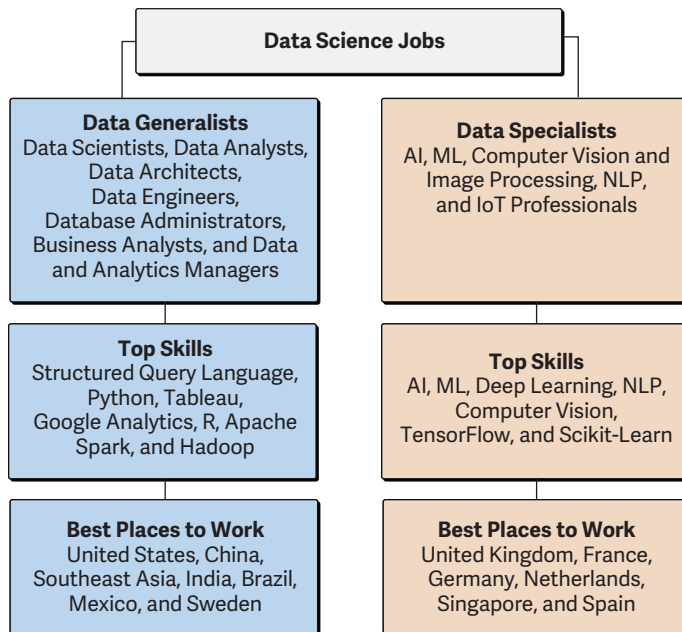
Given the hype around data science, the reality is that most companies still fail to use much of the data they collect and store during business activities. Gartner defines these unused data as "dark data."<sup>32</sup> According to Splunk's "State of Dark Data Report," 55% of an organization's data are dark.<sup>33</sup>

Yet, for a Fortune 1000 company, just a 10% increase in data accessibility will lead to more than US\$65 million additional net income.<sup>34</sup> The importance of understanding and utilizing data and the science behind it is becoming increasingly visible. Big data are now not an asset exclusive to big companies. The trend is heading to the "democratization of data," which essentially means the involvement of everyone, data scientist or not.<sup>35</sup> This will eventually lead to many trends in the future. An example includes the use of "small data" for analysis of the most vital data in situations where time, bandwidth, and energy expenditure are limited.<sup>36</sup> A closely linked concept is TinyML, which refers to ML algorithms that are designed to take little space and run on low-powered hardware, in embedded systems/the Internet of Things (IoT). Data democratization will also be driven by AutoML, which will make ML accessible to everyone.<sup>36</sup>

**TABLE 1.** The world's top undergraduate and graduate programs for data science, according to EduRank.<sup>23,24</sup>

Region	Top Five Universities
North America	<ol style="list-style-type: none"> <li>1. Stanford University, United States</li> <li>2. Harvard University, United States</li> <li>3. Massachusetts Institute of Technology, United States</li> <li>4. University of Illinois at Urbana-Champaign, United States</li> <li>5. University of Washington, Seattle, United States</li> </ol>
Latin America	<ol style="list-style-type: none"> <li>1. University of Sao Paulo, Brazil</li> <li>2. Federal University of Minas Gerais, Brazil</li> <li>3. National Polytechnic Institute, Mexico</li> <li>4. State University of Campinas, Brazil</li> <li>5. National Autonomous University of Mexico, Mexico</li> </ol>
Europe	<ol style="list-style-type: none"> <li>1. University of Oxford, United Kingdom</li> <li>2. University College London, United Kingdom</li> <li>3. University of Cambridge, United Kingdom</li> <li>4. Imperial College London, United Kingdom</li> <li>5. University of Manchester, United Kingdom</li> </ol>
Asia	<ol style="list-style-type: none"> <li>1. Tsinghua University, China</li> <li>2. National University of Singapore, Singapore</li> <li>3. Nanyang Technological University, Singapore</li> <li>4. Peking University, China</li> <li>5. Wuhan University, China</li> </ol>
Africa	<ol style="list-style-type: none"> <li>1. University of Pretoria, South Africa</li> <li>2. Ain Shams University, Egypt</li> <li>3. University of Stellenbosch, South Africa</li> <li>4. University of Cape Town, South Africa</li> <li>5. Mohammed V University, Morocco</li> </ol>
Oceania	<ol style="list-style-type: none"> <li>1. University of Melbourne, Australia</li> <li>2. University of New South Wales, Australia</li> <li>3. University of Sydney, Australia</li> <li>4. University of Technology Sydney, Australia</li> <li>5. University of Queensland, Australia</li> </ol>

Another emerging trend will be data-driven customer experience. There has been an increase in investment and innovation in online retail technology in recent years.<sup>35</sup> Interactions with businesses are also becoming more digital. So, it is realistic to expect less hassle in e-commerce, more user-friendly interfaces and front ends, quicker and smoother customer service, and greater levels of personalization in goods and services. Moreover, deep fakes, generative AI, and synthetic data are also expected to rise beyond the arts and entertainment industries.<sup>36</sup> Last, an increasing amount of data science will take place at the intersection of transformative technologies such as AI, the IoT, cloud computing, and superfast networks like 5G. These technologies will make new types of data transfer commonplace and increase automation to create smart homes, factories, and cities.<sup>36</sup> With data science's footprint in practically every aspect of our



**FIGURE 5.** Common job titles, top skills/areas of expertise, and best countries to work as data professionals. AI: artificial intelligence; NLP: natural language processing; IoT: Internet of Things.

everyday lives, there is always something new to learn. Technological change is perceived as fast, but data science, in particular, has been rising vigorously. 📈

## ACKNOWLEDGMENTS

We would like to thank Prof. Keith Miller and Prof. Joanna DeFranco for their valuable comments on previous versions.



## REFERENCES

1. T. H. Davenport and D. J. Patil, "Data scientist: The sexiest job of the 21st century," *Harvard Bus. Rev.*, vol. 90, no. 10, pp. 70–76, 2012.
2. J. W. Tukey, "The future of data analysis," *Ann. Math. Statist.*, vol. 33, no. 1, pp. 1–67, 1962, doi: 10.1214/aoms/1177704711.
3. M. Shirer and J. Rydning, "IDC's global datasphere forecast shows continued steady growth in the creation and consumption of data," IDC, Needham, MA, USA, May 8, 2020. [Online]. Available: <https://www.idc.com/getdoc.jsp?containerId=prUS46286020>
4. J. D. Kelleher and B. Tierney, *Data Science*. Cambridge, MA, USA: MIT Press, 2018.
5. R. A. Irizarry, "The role of academia in data science education," *Harvard Data Sci. Rev.*, vol. 2, no. 1, 2020, doi: 10.1162/99608f92.dd363929.
6. X.-L. Meng, "Data science: An artificial ecosystem," *Harvard Data Sci. Rev.*, vol. 1, no. 1, 2019, doi: 10.1162/99608f92.ba20f892.
7. P. Naur, *Concise Survey of Computer Methods*. New York, NY, USA: Petrocelli Books, 1974.
8. J. W. Tukey, *Exploratory Data Analysis*. vol. 2, 1977, pp. 131–160.
9. C. Weihs and K. Ickstadt, "Data science: The impact of statistics," *Int. J. Data Sci. Anal.*, vol. 6, no. 3, pp. 189–194, 2018, doi: 10.1007/s41060-018-0102-5.
10. D. Donoho, "50 years of data science," *J. Comput. Graphical Statist.*, vol. 26, no. 4, pp. 745–766, 2017, doi: 10.1080/10618600.2017.1384734.
11. W. S. Cleveland, "Data science: An action plan for expanding the technical areas of the field of statistics," *Int. Statist. Rev.*, vol. 69, no. 1, pp. 21–26, 2001. doi: 10.1111/j.1751-5823.2001.tb00477.x.
12. "About focus and scope," in *Data Science Journal*. London, U.K.: Ubiquity Press, 2021. [Online]. Available: <https://datascience.codata.org/about/>
13. D. Simberloff et al., "NSB-05-40, long-lived digital data collections enabling research and education in the 21st century," National Science Foundation, Alexandria, VA, USA, 2005. [Online]. Available: <https://www.nsf.gov/pubs/2005/nsb0540/>
14. "Hal Varian on how the web challenges managers," McKinsey & Company, New York, NY, USA, 2009. [Online]. Available: <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers>
15. D. J. Patil, *Building Data Science Teams*. Sebastopol, CA, USA: O'Reilly Media, Inc, 2011.
16. B. P. Ricca, B. E. Blaine, K. Donovan, and A. Geraci, "Changing paradigms: Faculty moving to data science from other disciplines," Mathematical and Computing Sciences Faculty/Staff Publications, May 17, 2019. [Online]. Available: <https://fisherpub.sjfc.edu/math>

- \_facpub/20/
17. T. H. Davenport and J. G. Harris, *Competing on Analytics: Updated, with a New Introduction: The New Science of Winning*. Cambridge, MA, USA: Harvard Univ. Press, 2017.
  18. G. Silvera, "Insights vs product vs engineering data science, and how each provides value to your business," LinkedIn, 2021. [Online]. Available: <https://www.linkedin.com/pulse/insights-vs-product-engineering-data-science-how-each-gordon-silvera/>
  19. R. Rawlings-Goss, *Data Science Careers, Training, and Hiring*. Cham, Switzerland: Springer International Publishing, 2019.
  20. "Guide for students looking for a degree in data science," Data Science Programs, 2021. [Online]. Available: <https://www.datascienceprograms.org/>
  21. "Why employers are looking for applicants with a Master's in data science", Cabrini University, Radnor, PA, USA. Accessed: Nov. 15, 2021. [Online]. Available: <https://www.cabrini.edu/graduate-degrees/programs/data-science/why-employers-are-looking-for-applicants-with-a-masters-in-data-science>
  22. T. Stobierski, "What is health informatics?" Northeastern Univ., Boston, MA, USA, 2021. [Online]. Available: <https://www.northeastern.edu/graduate/blog/what-is-health-informatics/>
  23. "World's 100 best data science universities" [2021 Rankings], EduRank, 2021. [Online]. Available: <https://edurank.org/cs/data-science/>
  24. "EduRank's university ranking methodology," EduRank. Accessed: Nov. 2, 2021. [Online]. Available: <https://edurank.org/methodology/>
  25. M. R. Berthold, "What does it take to be a successful data scientist?" *Harvard Data Sci. Rev.*, vol. 1, no. 2, 2019, doi: 10.1162/99608f92.e0eaabfc.
  26. "Jobs on the rise reports: The fastest-growing jobs in the world," LinkedIn. Accessed: Nov. 2, 2021. [Online]. Available: <https://business.linkedin.com/talent-solutions/emerging-jobs-report#all>
  27. "Occupational employment and wage statistics," U.S. Bureau of Labor Statistics, Washington, DC, USA, 2021. [Online]. Available: <https://www.bls.gov/oes/current/oes152098.htm>
  28. L. Burtch, "The Burtch works study salaries of data scientists & predictive analytics professionals," Burtch Works, 2020. [Online]. Available: [https://www.burtchworks.com/wp-content/uploads/2020/08/Burtch-Works-Study\\_DS-PAP-2020.pdf](https://www.burtchworks.com/wp-content/uploads/2020/08/Burtch-Works-Study_DS-PAP-2020.pdf)
  29. "Data scientist salaries," Glassdoor, 2021. [Online]. Available: [https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH\\_KO0,14.htm](https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm)
  30. "Programmer salaries," Glassdoor, 2021. [Online]. Available: [https://www.glassdoor.com/Salaries/programmer-salary-SRCH\\_KO0,10.htm](https://www.glassdoor.com/Salaries/programmer-salary-SRCH_KO0,10.htm)
  31. "Statistician salaries," Glassdoor, 2021. [Online]. Available: [https://www.glassdoor.com/Salaries/statistician-salary-SRCH\\_KO0,12.htm](https://www.glassdoor.com/Salaries/statistician-salary-SRCH_KO0,12.htm)
  32. "Dark data," Gartner. Accessed: Nov 15, 2021. [Online]. Available: <https://www.gartner.com/en/information-technology/glossary/dark-data>
  33. "Press release: Dark data research reveals widespread complacency in driving business results and career growth," Splunk, Apr. 30, 2019. [Online]. Available: [https://www.splunk.com/en\\_us/newsroom/press-releases/2019/dark-data-research-reveals-widespread-complacency-in-driving-business-results-and-career-growth.html](https://www.splunk.com/en_us/newsroom/press-releases/2019/dark-data-research-reveals-widespread-complacency-in-driving-business-results-and-career-growth.html)
  34. L. Myler, "Better data quality equals higher marketing ROI," *Forbes*, 2017. [Online]. Available: <https://www.forbes.com/sites/larrymyler/2017/07/11/better-data-quality-equals-higher-marketing-roi/?sh=750bc78c7b68>
  35. B. Marr, "What is data democratisation and why it is a business game-changer?" Bernard Marr & Co., 2021. [Online]. Available: <https://bernardmarr.com/what-is-data-democratisation-and-why-it-is-a-business-game-changer/>
  36. B. Marr, "The 5 biggest data science trends in 2022," *Forbes*, 2021. [Online]. Available: <https://www.forbes.com/sites/bernardmarr/2021/10/04/the-5-biggest-data-science-trends-in-2022/?sh=28d8692440d3>
- NORITA AHMAD** is a professor of management information systems at the American University of Sharjah, United Arab Emirates. She is a Member of IEEE. Contact her at [nahmad@aus.edu](mailto:nahmad@aus.edu).
- AREEBA HAMID** is a master of business administration student at the American University of Sharjah, United Arab Emirates. Contact her at [g00074541@aus.edu](mailto:g00074541@aus.edu).
- VIAN AHMED** is a professor of industrial engineering at the American University of Sharjah, United Arab Emirates. Contact her at [vahmed@aus.edu](mailto:vahmed@aus.edu).

## DEPARTMENT: DIVERSITY AND INCLUSION

# Careers in STEM: A Latina Perspective

Andrea Delgado , Veronica G. Melesse Vergara , and Andrea Schneibel, Oak Ridge National Laboratory, Oak Ridge, TN, 37830, USA

*Three Latina computing professionals at a large national laboratory reflect on the circumstances affecting the low representation of this segment of the population in STEM fields, and computing in particular. The authors share highlights of their path to STEM careers, and some of the efforts they are involved in for broadening participation in computing. They consider the roles of minority serving institutions, representation and mentoring, and advocacy.*

The U.S. Hispanic population has been growing steadily for decades. Last year, nearly 65 million Hispanic Americans made up 18.5% of the U.S. population (including Puerto Rico).<sup>1</sup> They were 18% of the overall workforce, but only 8.3% of the STEM professional labor pool.<sup>2</sup> Meanwhile, women make up 34% of all STEM occupations, and within the sciences, the proportion of women in computer science occupations was 26% in 2019, as reported by the National Science Foundation.<sup>3</sup>

While statistics describe the proportion of Hispanic Americans in STEM and women in STEM, the intersection between these two segments has not been systematically tracked. This highlights a significant challenge: quantifying the underrepresentation of Hispanic women in the workforce, particularly in STEM where fields are male dominated, including computer science.

A better understanding of this intersecting segment would be helpful in identifying specific obstacles that impact Latinas working in STEM fields in the US, which would be critical to identifying opportunities to improve

the number of Latinas in the field and increase retention by providing more prosperous career development paths.

In this article, we highlight some of the challenges Latinas in STEM face, including those specific to computing-related fields.

### LATINAS IN STEM

Although most of the Latinx population in the U.S. shares Spanish as their common language, different cultures, and traditions melt to create the Latinx<sup>a</sup> identity. Different backgrounds and circumstances will affect the individual journey into a STEM career and create a scientific identity. For example, some Latinas in STEM may be first-generation college graduates, while others may have traveled thousands of miles to pursue their education and jump-start their careers. In what follows, the authors reflect on their journeys as Latinas currently working in STEM.

### Three Paths to Stem Careers

One of the authors, now a technical staff member at Oak Ridge National Laboratory (ORNL), originally from Ecuador, moved to the U.S. as an exchange student. After being exposed to the education and opportunities available in the U.S., she decided to move permanently and complete her degree here due to the fact that career prospects in the computing field were more promising. She was able to pursue specialized education in computing and find a position that would apply those skills in the world of high-performance computing (HPC) at the U.S. Department of Energy national laboratory. Exposure to a wide range of career

---

Notice: This manuscript has been authored in part by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>).

---

1521-9615 © 2022 IEEE  
Digital Object Identifier 10.1109/MCSE.2022.3188195  
Date of current version 31 August 2022.

---

<sup>a</sup>The term Latinx refers to people of Latin American cultural or ethnic identity in the United States.

paths and industries in college and graduate school helped encourage her to remain on the STEM path.

A second author reflects on her journey as a Mexican–American woman born and raised in El Paso, Texas. Her path into a research scientist position at ORNL started with an internship opportunity at Texas A&M University, allowing her to experience research at one of the top nuclear physics facilities in the country. She also highlights the importance of representation and mentoring, two essential resources during her graduate degree studies in a male-dominated scientific field.

Born and educated in Venezuela, the third author is a science communicator supporting a research division in a national lab. Growing up in a low-income community meant that she had poor access to educational opportunities in STEM. She graduated college with a degree in communications and went on to pursue a career in science writing with the goal of helping build a bridge between research activity and scientific literacy in the general public. Her role has low employment prospects in the Venezuela, which made her explore other options. Although the U.S. has a strong job market for professional science communicators, bilingual ones are still scarce.

## THE LEAKY PIPELINE

Despite the different paths that bring Latinas to STEM, as a collective, they face similar challenges that are unique to this particular segment of the population. In addition to dealing with the challenges women in STEM face, including sexism, insufficient support from parents, and lack of opportunities for advancement, women of color also have to combat cultural stereotypes.

One common stereotype that Latinas have to counter is how they present themselves to others. Often, people expect Latinas to be extroverted and loud, and when a person does not fit that mold, they tend to face judgment. On the flip side, Latinas who are extroverted can often be labeled as too bold. In both cases, the result negatively impacts the individual's professional image, translating to serious setbacks for their career. As members of an underrepresented minority in the field, Latinas already struggle to develop a sense of belonging, but stereotypes like these compounded by the continuous efforts to present one's image in a way that minimizes judgment, can be detrimental to one's career.

Further, Latinas have to face stereotypes prevalent in popular culture that portray them as one-dimensional and in a narrow set of roles that rarely showcase them in science, technology, or leadership roles.

Breaking out of those molds can be extremely challenging and results in additional burden for Latinas working to advance their careers.

Several studies have shown that many women leave STEM fields in the first few years of their professional careers, for several reasons. Most organizations do not provide adequate policies and benefits to support a healthy work-life balance. This often leaves primary care parents with the tough choice of stepping away temporarily or switching to another field. Unfortunately, taking a leave for a prolonged period can make it extremely difficult for the individual to return to the workforce and it impacts their long-term earning potential. Traditionally, especially in Latino households, women have held that primary care role and, as a result, have a much harder time rejoining the workforce.

While no single solution can “plug” the pipeline, efforts are necessary at each educational level and career stage to encourage women of color to stay and thrive in STEM fields.

## PLUGGING THE PIPELINE AT EVERY LEVEL

Given that the number of Latinas continues to decrease as they advance in educational level and career stage, it is important to pursue mechanisms to “plug” the pipeline starting from K-12 but continuing all the way up to top-level leadership roles.

Many volunteers employed in academic, research, and government institutions, currently support organizations that promote STEM fields via after-school programs targeting K-12 students. However, participating in after-school programs can already be a challenge for Latinx students who face justifying the value of STEM programs to their parents. Providing informational sessions directly to parents can have a more significant impact in broadening participation. As part of community engagement goals, staff at research institutions, including ORNL, are encouraged to participate in local, national, and international events that promote careers at national laboratories and in STEM fields. Recently, ORNL participated as a mentor organization in the Winter Classic Invitational Student Cluster Competition.<sup>4</sup> As a result, ORNL hosted 12 teams from historically Black colleges and universities and Hispanic-serving institutions (HSIs) in a multiweek competition. Students were exposed to HPC and AI applications on different compute resources including Summit, the U.S. second fastest supercomputer (June 2022). This and similar competitions provide a unique venue to prepare students with skills related to careers in specialized computing fields. This type of outreach can

also be a great tool for students to connect with staff from underrepresented groups and begin growing their network, which is a key step toward retention.

### Role of Minority Serving Institutions

HSIs are defined in the Higher Education Act as not-for-profit institutions of higher learning with a full-time equivalent undergraduate student enrollment that is at least 25% Hispanic. This designation is the result of a grassroots effort in 1980 from a group of college students pushing for greater federal support for schools with a high Hispanic population. In 1986, 18 schools came together to form the Hispanic Association of Colleges and Universities,<sup>5</sup> which continues to advocate for Hispanic higher education.

HSIs are at the forefront of efforts to increase educational access and success for the nation's Hispanic citizens. They educate the majority of Latinx undergraduates in the U.S. Two out of three Latinx undergraduate students attend an HSI.<sup>6</sup> Furthermore, Latinx students at HSIs report higher graduation rates than the national average,<sup>7</sup> which emphasizes the importance of having a Latinx community as part of students' support systems.

Thanks to their designation as an HSI, colleges can receive grants from the U.S. Department of Education, enabling them to expand opportunities for Latinx students. These funds are awarded based on the HSI's student-support programs, community outreach programs, and efforts to increase the number of Latinx students. Federal funding also aims to increase the number of Latinx teachers in elementary and secondary education. In addition, HSIs offer several benefits for students. According to a 2020 report from *Excelencia in Education*,<sup>8</sup> some HSIs incorporate workforce preparation opportunities in their programs and emphasize learning from experience. Many HSIs have also partnered with local employers to help graduates transition into the workforce.

Furthermore, HSIs partner with local high schools to increase the number of first-generation Latinx college students. They also offer tutoring and mentoring opportunities, and integrate bilingual programs into their services. Thanks to their specialized programs and a supportive environment, HSIs help increase the number of Latinx college graduates.

## REPRESENTATION AND MENTORING

As Latinas advance in their STEM careers, they find fewer members of underrepresented minorities as their peers and even fewer in leadership roles. The

lack of diversity in top-level leadership positions can negatively impact the advancement of women of color in two ways. First, by discouraging women from pursuing specific roles due to the lack of representation observed, incorrectly reinforcing the perceived idea that they do not belong. Second, due to unconscious biases, people in leadership roles will tend to favor others most similar to them. This disproportionately impacts members from underrepresented groups and reduces their chances for career advancement.

In addition to the issues associated with educational barriers, such as lack of preparation and skills, lack of financial support and information on graduate school, another important aspect is the lack of role models. This is particularly important for Latinas in STEM. A recent study<sup>9</sup> reports on the findings from a survey involving nearly 1000 Black and Latinx students. Researchers found that thinking outside the stereotype of who the scientist is, including a scientist's interests and identity outside of their research, helps students feel more positive and encouraged in studying science themselves. Identifying with scientists in STEM can be incredibly important for young students, indicating that they can pursue and succeed in the fields that fascinate them.

One way to provide role models is by developing successful mentoring relationships. In the context of Latinas in STEM, mentorship can help combat the leaky pipeline. Mentorship relationships can have different forms, depending on the particular needs of the mentee at a given time in her career. Several studies have found that one of the essential factors in obtaining a doctoral degree was a positive mentoring experience,<sup>10</sup> and this statement is particularly true for early research experiences, such as summer internships. Similar studies have found that, regardless of the individual's background, the desire to pursue a graduate degree is influenced by a strong mentee-mentor relationship.<sup>11</sup> Mentoring has been linked to enhanced science identity, sense of belonging, and self efficacy,<sup>12</sup> which are critical aspects for improving degree attainment in the Latinx community. Thus, institutions need to allocate resources for mentoring programs to develop and target underrepresented groups. Furthermore, mentor training is crucial for the success of these programs.

## ADVOCACY: FROM BYSTANDER TO UPSTANDER

While individuals and organizations can contribute in many ways to broadening participation in STEM and computing, the focus varies depending on career stage, size and type of organization, and particular

areas of interest. Organizations and their leadership teams should ensure that diversity and inclusion goals are a core part of their mission and values. In addition to having clearly defined goals, organizations should commit funding to develop a diverse and inclusive culture. This includes providing and requiring training in diversity and inclusion topics, creating professional development programs to help amplify the voices of underrepresented minorities, and fostering a welcoming atmosphere where team members feel comfortable speaking up.

Efforts to broaden participation must include a multipronged approach: outreach to identify future candidates, recruiting at various organizations and regions, and retention by creating well-defined paths for career advancement.

We can identify two distinct roles at an individual level: members of underrepresented groups and allies. As a member of an underrepresented group, it is vital to participate in outreach activities both by attending recruitment events and advocating for participation at venues that may be unknown to the organization.

On the other hand, allies must be willing to become active participants in the conversation, be ready to speak out when they observe microaggressions, and educate themselves and others on diversity and inclusion topics continuously.

## CLOSING THOUGHTS

While overall the number of women in STEM fields is increasing, the gender STEM gap is still significant, and without appropriate and intentional action from organizations, leaders, and community members, it will fail to close at a fast-enough rate to match population growth. The factors that drive women away from STEM are exacerbated for women of color, who still represent a small fraction of professionals in STEM despite seeing higher numbers of K-12 students participating in STEM programs.

Institutions need to recognize the gap and take action to help improve recruitment practices and retention rates. Some specific steps that institutions can take include but are not limited to:

- › encourage staff and students to use professional societies to develop professionally and become members of the larger STEM community, which will help create a sense of belonging;
- › provide learning resources for becoming an effective mentor for a diverse student and staff population;
- › recognize the value of the diverse contributions from faculty/staff/students;

- › develop and implement a plan to recruit, retain, and acclimate members of underrepresented minorities;
- › build vertical and horizontal bridges to extend the resources and community of the department.

Leaders at institutions need to make conscious efforts and allocate financial and personnel resources to put inclusive practices in place. Developing an inclusive and welcoming atmosphere at an organization must become a priority and should be highlighted as a responsibility of all community members.

Finally, one of the most critical actions individuals can take to help “plug” the pipeline is to educate themselves on best practices to recognize and call out biases in themselves and others. Individuals can become allies in many ways, but the first step is to become knowledgeable about the challenges underrepresented minorities, including Latinx, face to determine how to best contribute to the organization’s diversity and inclusion goals. 🌍

## ACKNOWLEDGMENTS

This work used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Grant DE-AC05-00OR22725.

## REFERENCES

1. U. S. Census Bureau, “U.S. population estimates,” 2020. [Online]. Available: <https://www.census.gov/quickfacts/fact/table/US/PST045221>
2. “Labor force statistics from the current population survey,” 2022. [Online]. Available: <https://www.bls.gov/cps/cpsaat11.htm>
3. “The state of U. S. science and engineering,” 2022. [Online]. Available: <https://ncses.nsf.gov/pubs/nsb20221>
4. “2022 winter classic invitational student cluster competition.” [Online]. Available: <https://www.winterclassicinvitational.com/>
5. HACU, “Hispanic association of colleges and universities,” 2022. [Online]. Available: <https://www.hacu.net/hacu/default.asp>
6. H. A. of Colleges and Universities, “Hispanic higher education and Hispanic-serving institutions 2020 fact sheet,” 2020. [Online]. Available: [https://www.hacu.net/images/hacu/conf/2022CapForum/ResourcesMenu/2022\\_HSI\\_FactSheet.pdf](https://www.hacu.net/images/hacu/conf/2022CapForum/ResourcesMenu/2022_HSI_FactSheet.pdf)
7. G. A. Garcia and M. Taylor, “A closer look at Hispanic serving institutions,” *Higher Educ. Today*, 2017. [Online]. Available: <https://www.higheredtoday.org/2017/09/18/closer-look-hispanic-serving-institutions/>

8. J. Martinez and D. A. Santiago, "Tapping Latino talent: How HSIS are preparing Latino students for the workforce," *Excelencia Educ.*, 2020. [Online]. Available: <https://www.edexcelencia.org/research/publications/tapping-latino-talent>
9. U. Nguyen and C. Riegle-Crumb, "Who is a scientist? The relationship between counter-stereotypical beliefs about scientists and the STEM major intentions of Black and Latinx male and female students," *Int. J. STEM Educ.*, vol. 8, Apr. 2021, Art. no. 28.
10. D. G. Solorzano, "The road to the doctorate for California's Chicanas and Chicanos: A study of Ford Foundation minority fellows," CPS Rep., California Policy Seminar, Berkeley, CA, USA, 1993.
11. R. McGee and J. L. Keller, "Identifying future scientists: Predicting persistence into research training," *CBELife Sci. Educ.*, vol. 6, pp. 316–331, Dec. 2007.
12. H. Thiry, S. L. Laursen, and A.-B. Hunter, "What

experiences help students become scientists? A comparative study of research and other sources of personal and professional gains for STEM undergraduates," *J. Higher Educ.*, vol. 82, pp. 357–388, Jul. 2011.

**ANDREA DELGADO** is a research scientist in the Physics Division at Oak Ridge National Laboratory, Oak Ridge, TN, 37830, USA. Contact her at [delgadoa@ornl.gov](mailto:delgadoa@ornl.gov).

**VERONICA G. MELESSE VERGARA** is with Oak Ridge National Laboratory (ORNL), Oak Ridge, TN, 37830, USA. Contact her at [vergaravg@ornl.gov](mailto:vergaravg@ornl.gov).

**ANDREA SCHNEIBEL** is with Oak Ridge National Laboratory (ORNL), Oak Ridge, TN, 37830, USA. Schneibel received her Master of Scientific, Medical, and Environmental Communications degree in communication science from Universitat Pompeu Fabra, Barcelona, Spain. Contact her at [schneibelay@ornl.gov](mailto:schneibelay@ornl.gov).

## ADVERTISER INFORMATION

### Advertising Coordinator

Debbie Sims  
 Email: [dsims@computer.org](mailto:dsims@computer.org)  
 Phone: +1 714-816-2138 | Fax: +1 714-821-4010

### Advertising Sales Contacts

Mid-Atlantic US:  
 Dawn Scoda  
 Email: [dscoda@computer.org](mailto:dscoda@computer.org)  
 Phone: +1 732-772-0160  
 Cell: +1 732-685-6068 | Fax: +1 732-772-0164

Southwest US, California:  
 Mike Hughes  
 Email: [mikehughes@computer.org](mailto:mikehughes@computer.org)  
 Cell: +1 805-208-5882

Northeast, Europe, the Middle East and Africa:  
 David Schissler  
 Email: [d.schissler@computer.org](mailto:d.schissler@computer.org)  
 Phone: +1 508-394-4026

Central US, Northwest US, Southeast US, Asia/Pacific:  
 Eric Kincaid  
 Email: [e.kincaid@computer.org](mailto:e.kincaid@computer.org)  
 Phone: +1 214-553-8513 | Fax: +1 888-886-8599  
 Cell: +1 214-673-3742

Midwest US:  
 Dave Jones  
 Email: [djones@computer.org](mailto:djones@computer.org)  
 Phone: +1 708-442-5633 Fax: +1 888-886-8599  
 Cell: +1 708-624-9901

### Jobs Board (West Coast and Asia), Classified Line Ads

Heather Buonadies  
 Email: [hbuonadies@computer.org](mailto:hbuonadies@computer.org)  
 Phone: +1 623-233-6575

### Jobs Board (East Coast and Europe), SE Radio Podcast

Marie Thompson  
 Email: [marie.thompson@computer.org](mailto:marie.thompson@computer.org)  
 Phone: +1 714-813-5094



# Is It Live, or Is It Deepfake?

Nir Kshetri , University of North Carolina at Greensboro

Joanna F. DeFranco , The Pennsylvania State University

Jeffrey Voas , IEEE Fellow

*Let's build deepfake 'trust' so it can be put to good use.*

It's been four decades since society was in awe of the quality of recordings available from a cassette recorder tape. Today we have something new to be in awe of: deepfakes. Deepfakes include hyper-realistic videos that use artificial intelligence (AI) to create fake digital content that looks and sounds real. The word is a portmanteau of "deep learning" and "fake." Deepfakes are everywhere: from TV news to advertising, from national election campaigns to wars between states, and from cybercriminals' phishing campaigns to insurance claims that fraudsters file. And deepfakes come in all shapes and sizes—videos, pictures, audio, text, and any other digital material that can be manipulated with AI. One estimate suggests that deepfake content online is growing at the rate of 400 annually.<sup>1</sup>

There appear to be legitimate uses of deepfakes, such as in the medical industry to improve the diagnostic accuracy of AI algorithms in identifying periodontal disease<sup>2</sup> or to help medical professionals create artificial patients (from real patient data) to safely test new diagnoses and treatments or help physicians make medical decisions.<sup>3</sup> Deepfakes are also used to entertain, as seen recently on *America's Got Talent*,<sup>4</sup> and there may be future uses where deepfake could help teachers address the personal needs and preferences of specific students.<sup>5</sup>

Unfortunately, there is also the obvious downside, where the most visible examples represent malicious and illegitimate uses. Examples already exist.

Deepfakes also involve voice phishing, also known as *vishing*,<sup>6</sup> which has been among the most common techniques for cybercriminals. This technique involves

using cloned voices over the phone to exploit the victim's professional or personal relationships by impersonating trusted individuals. In March 2019, cybercriminals were able to use a deepfake to fool the CEO of a U.K.-based energy firm into making a US\$234,000 wire transfer. The British CEO who was victimized thought that the person speaking on the phone was the chief executive of the firm's German parent company. The deepfake caller asked him to transfer the funds to a

---

*IN MARCH 2019, CYBERCRIMINALS WERE ABLE TO USE A DEEPPFAKE TO FOOL THE CEO OF A U.K.-BASED ENERGY FIRM INTO MAKING A US\$234,000 WIRE TRANSFER.*

---

Hungarian supplier within an hour, emphasizing that the matter was extremely urgent. The fraudsters used AI-based software to successfully imitate the German executive's voice.<sup>7</sup>

In a more high-profile case, in January 2020, cybercriminals defrauded a bank in the United Arab Emirates of more than US\$35 million using deepfake voice technology. The technology was used to imitate a company director who was known to a bank branch manager. The manager authorized the transactions.<sup>8</sup>

---

Digital Object Identifier 10.1109/MC.2023.3252059

Date of current version: 26 June 2023



Deepfakes are also being used by nation states to spread misinformation and disinformation and pursue their national interests. For instance, in the Russo-Ukrainian war, both Russia and Ukraine have deployed deepfake videos against each other.<sup>9</sup>

What can be done to combat deepfakes? Could we create deepfake detectors? Or create laws or a code of conduct that probably would be ignored?

There are tools that can analyze the blood flow in a subject's face and then compare it to human blood flow activity to detect a fake.<sup>10</sup> Also, the European Union is working on addressing manipulative behaviors.<sup>11</sup>

**T**here are downsides to both categories of solutions, but clearly something needs to be done to build trust in this emerging and disruptive technology. The problem isn't going away. It is only increasing. 🤖

## DISCLAIMER

The authors are completely responsible for the content in this message. The opinions expressed here are their own.

## REFERENCES

1. Mordor Intelligence, "Deepfake content on the internet is growing at the rate of a whopping 400 year on year," *GlobeNewsWire*, Oct. 2022. Accessed: Feb. 25, 2023. [Online]. Available: <https://www.globenewswire.com/en/news-release/2022/10/27/2542944/0/en/Deepfake-content-on-the-internet-is-growing-at-the-rate-of-a-whopping-400-year-on-year.html>
2. "Synthetic medical imaging: How deepfakes could improve healthcare," *CISION PR Newswire*, Aug. 2022. Accessed: Feb. 25, 2023. [Online]. Available: <https://www.prnewswire.com/news-releases/synthetic-medical-imaging-how-deepfakes-could-improve-healthcare-301611438.html>
3. V. Gain, "How deepfakes and AI are being used to find new ways to treat diseases," *Silicon Republic*, Nov. 2022. Accessed: Feb. 25, 2023. [Online]. Available: <https://www.siliconrepublic.com/innovation/deepfake-ai-healthcare-diseases-insilico-medicine-pharma>
4. J. Kahn, "Deepfakes are stealing the show on 'America's Got Talent.' Will they soon steal a lot more too?" *Fortune Mag.*, Sep. 2022. Accessed: Feb. 25, 2023. [Online]. Available: <https://fortune.com/2022/09/06/deepfakes-america-got-talent-metaphysic-fraud-metaverse/>
5. "Deepfake teachers & technology: The future of K-12 public education?" Southwest Washington Educ., Washington, DC, USA, Sep. 2021. Accessed: Feb. 27, 2023. [Online]. Available: <https://swwededucation.org/deepfake-teachers-technology-the-future-of-k-12-public-education/>
6. J. Bateman, "Deepfakes and synthetic media in the financial system: Assessing threat scenarios," Carnegie Endowment for International Peace, Washington, DC, USA, Jul. 2020. Accessed: Feb. 25, 2023. [Online]. Available: <https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237>
7. N. Kshetri, *Cybersecurity Management: An Organizational and Strategic Approach*. Toronto, ON, Canada: Univ. of Toronto Press, 2021.
8. M. Anderson, "Deepfaked voice enabled 35 million bank heist in 2020," *Unite AI*, Oct. 2021. Accessed: Feb. 25, 2023. [Online]. Available: <https://www.unite.ai/deepfaked-voice-enabled-35-million-bank-heist-in-2020/>
9. B. Fowler, "Deepfakes pose a growing danger, new research says," *CNET*, Aug. 2022. Accessed: Feb. 25, 2023. [Online]. Available: <https://www.cnet.com/tech/services-and-software/deepfakes-pose-a-growing-danger-new-research-says/>
10. A. Nine, "Intel's AI can detect deepfakes with 96 percent accuracy," *Extremetech*, Nov. 2022. Accessed: Feb. 25, 2023. [Online]. Available: <https://www.extremetech.com/computing/340926-intels-ai-can-detect-deepfakes-with-96-percent-accuracy>

11. K. Collins, "EU strengthens disinformation rules to target deepfakes, bots, fake accounts," *CNET*, Jun. 2022. Accessed: Feb. 25, 2023. [Online]. Available: <https://www.cnet.com/news/politics/eu-strengthens-disinformation-rules-to-target-deepfakes-bots-fake-accounts/>

**NIR KSHETRI** is a professor of management in the Bryan School of Business and Economics, University of North Carolina at Greensboro, Greensboro, NC 27412 USA, and the "Computing's Economics" column editor for *Computer*. Contact him at [nbkshetr@uncg.edu](mailto:nbkshetr@uncg.edu).

**JOANNA F. DEFRANCO** is an associate professor of software engineering, associate director of the D.Eng. in Engineering program at The Pennsylvania State University, Malvern, PA 19355 USA, and an associate editor in chief of *Computer*. Contact her at [jfd104@psu.edu](mailto:jfd104@psu.edu).

**JEFFREY VOAS**, Gaithersburg, MD 20899 USA, is the editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at [j.voas@ieee.org](mailto:j.voas@ieee.org).



**IEEE COMPUTER SOCIETY**  
**Call for Papers**

Write for the IEEE Computer Society's authoritative computing publications and conferences.

**GET PUBLISHED**  
[www.computer.org/cfp](http://www.computer.org/cfp)

 IEEE COMPUTER SOCIETY

 IEEE

# The AI-Cybersecurity Nexus: The Good and the Evil

San Murugesan , *BRITE Professional Services, Sydney, NSW, 2152, Australia*

*AI is both good and bad for cybersecurity defenders—and for adversaries. This article outlines the nexus between AI and cybersecurity and explores how artificial intelligence (AI) can enhance information systems security. It discusses how threat actors also could use AI to create sophisticated attacks that evade detection, and how AI could become a victim of new sophisticated cyberattacks.*

Technological advancements are transforming what once seemed impossible into reality. But broader use of IT has brought—and will continue to bring—heightened security risks. Cyberattacks are more prevalent and harsher than ever and are no longer confined to businesses and governments as they used to be; they now stretch across every major sector—health,<sup>1</sup> manufacturing, logistics,<sup>2</sup> utilities,<sup>3</sup> and education,<sup>4</sup> for example. Last year was the most disruptive year with ransomware and other cyberattacks impacting businesses and governments—and critical infrastructure—as never before.<sup>5</sup>

The attack surface (threat landscape) is becoming broader and threat vectors continue to evolve due to the increasing use of the Internet of Things (IoT), cloud computing, data analytics, artificial intelligence (AI), robotics, and automation in a range of applications. For computer security professionals, businesses, and governments securing computing and information systems and protecting networked, computer-controlled critical infrastructure such as electricity generation, water treatment plants, and supply chain companies is a major challenge.

Going forward, security threats will grow in number, severity, and sophistication and cause even more severe impacts unless we implement effective proactive and reactive protective measures commensurate with ever-changing threats. Cybersecurity will remain a perpetual concern and challenge for everyone.

Protecting data, computer systems, and critical infrastructure calls for continuously updated strategies and new approaches as traditional approaches to monitoring, threat hunting, and incident response are labor-intensive and time-consuming, both leading to delayed remediation and exposure to cyberattacks. But there is

help: we can embrace artificial intelligence (AI) and machine learning (ML).

AI cybersecurity solutions, or “defensive AI,” can address problems that cannot be solved or effectively addressed by traditional cybersecurity solutions. By automating and improving core security functions, AI can transform security operations into streamlined, autonomous, continuous operations that speed remediation and offer better protection.<sup>6</sup> We can use AI for repetitive tasks to free up security staff for projects that require human ingenuity.

On the other hand, adversaries can—and do—use AI as an effective tool and aid to create intense, harmful cyberattacks that are more difficult to detect. This is referred to as “offensive AI” or “malicious AI.” Worse still, AI can be compromised by cyberattacks (adversarial attacks) impacting security and other applications of AI.

So, security professionals need to deeply examine the promise of AI to cybersecurity and security attacks on AI. Here, we briefly examine the AI-cybersecurity nexus, setting the scene for further exploration. We outline how AI can help protect information systems and networks against traditional cyberattacks and emergent threats. We also briefly outline two other key facets of the AI-cybersecurity nexus: how adversaries can exploit AI to launch sophisticated attacks and the need to protect AI from cyberattacks. Finally, we outline the benefits of using AI to power security solutions and offer recommendations for an AI-empowered secure digital future.

## CYBERSECURITY: A PERPETUAL CONCERN AND CHALLENGE

The cybersecurity landscape continues to widen accompanied by an incredible rise in security threats. Increased remote work and online activity; a rise in geopolitical tensions; greater use of cloud, IoT, wearables, and drones; an increase in network-connected devices for surveillance and smart home and building management; and expanding

networked critical infrastructures such as utilities have made our cyber and cyber-physical systems increasingly vulnerable to cyberattacks. Cyber adversaries are leveraging ongoing digital disruption and numerous applications that we now depend on to mount cyberattacks on individuals, organizations, governments, and critical infrastructure.

A recent report<sup>7</sup> by Splunk identifies over 50 top security threats and briefly outlines for each threat what we need to know, how the attack happens, and where the attack comes from. Key threats facing digital system assets, attack vectors, and threat agents are discussed in a 2019 publication<sup>8</sup> by the author. Various cybersecurity issues and considerations are indicated in detailed, interactive maps<sup>9</sup> available from The World Economic Forum.

To get an idea on the reality of cyberattacks, look at a few recent cybersecurity attacks. The last year's SolarWinds breach showcased how the supply chain could be compromised and the serious damage such an attack could cause globally.<sup>10</sup> Attacks last year on the Colonial Pipeline; a water facility in Oldsmar, Florida, USA;<sup>11</sup> and JBS, the world's largest meat supplier,<sup>12</sup> highlight today's critical infrastructure risks<sup>13</sup> and the need to better secure our manufacturing, supply chain, and utilities. Ransomware attacks have increased causing significant disruption to normal activities in hospitals, government offices, and businesses. New threats such as deepfakes<sup>14</sup> and others aided by AI and advanced attack tools are emerging.

The economic, business, and societal impact of cyberattacks is huge. If it were measured as a country, cybercrime—estimated to have inflicted damages totaling USD 6 trillion globally in 2021—would be the world's third-largest economy after the U.S. and China.<sup>15</sup> Cybercrime costs are expected to reach USD 10.5 trillion annually by 2025, which will be more than the global trade of all major illegal drugs combined. Cyberattack surface will be an order of magnitude greater in 2025 than it is today.

Another worrying trend is the increasing number of threat actors and cyber adversaries, ranging from amateur hackers and professional hacktivists to cybercriminals and state-sponsored actors, who are escalating alliances. These adversaries are widely disseminating insights and critical knowledge and collaborating to generate and benefit from coordinated, sophisticated cyberattacks.

Going forward, we must also address potential new security threats in the metaverse,<sup>16</sup> Web 3, autonomous vehicles, driverless cars, farm equipment, and spaceborne assets.<sup>17</sup>

Hence, cybersecurity will remain a perpetual concern and pose major challenges that we need to address satisfactorily. To address unprecedented security risks that threaten to undermine economic growth and public

trust, we need to adopt new cybersecurity strategies, approaches, and tools and constantly review and renew them. AI is a new powerful arsenal and ally in the fight against current and emergent cyberattacks; but it can also empower cyber adversaries.

## AI-CYBERSECURITY NEXUS

It is vital to examine the AI-cybersecurity nexus (see Figure 1) and explore its promise and pitfalls. AI and cybersecurity intersect and impact one another in three ways:

- 1) AI can help protect information and control systems and networks against traditional cyberattacks and emergent threats;
- 2) adversaries can exploit AI to launch sophisticated attacks;
- 3) AI is vulnerable to cyberattacks and we need to secure AI from security threats.

## AI-EMPOWERED CYBER DEFENSE

AI security solutions promise to have a significant impact on defensive operations, making security operations more effective and autonomous. They can be applied across each cybersecurity segment such as monitoring, detection, threat hunting, response, and proactive prevention and protection.

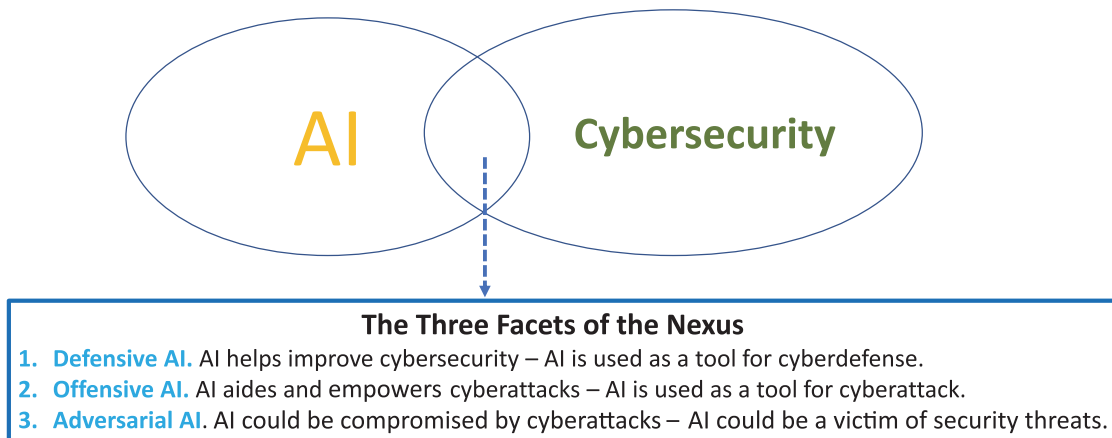
Operational security personnel are overwhelmed with security alerts and incidents that they must monitor and analyze. AI/ML technologies can help with this, searching large volumes of alerts and data concerning events with speed and a degree of accuracy that humans cannot match. They can also automate the detection of events that require human analysis, reducing the number of mundane tasks security personnel must perform.

AI can play key roles in and empower several cybersecurity operations, including<sup>18</sup>:

- 1) behavioral analytics;
- 2) threat intelligence;
- 3) ransomware attack detection;
- 4) smart identity governance;
- 5) strengthening cloud security;
- 6) online fraud detection;
- 7) defending against deepfakes; and
- 8) risk assessment.

AI can detect nuanced adversarial attacks, enhance data-driven decision-making during threat hunts, uncover previously undetectable tampering in devices, and quantify the risks associated with current IT system vulnerabilities. Security companies like Darktrace and SentinelOne use AI/ML to augment

## AI-Cybersecurity Nexus



**FIGURE 1.** AI and cybersecurity interact and impact one another along three dimensions.

threat detection and response and provide autonomous endpoint security.

### Benefits of Applying AI to Cyber Defense

The benefits of integrating AI into an organization's cybersecurity ecosystem are many and include the following<sup>19</sup>:

- › Improved cybersecurity protection and effectiveness due to AI's ability to detect nuanced attacks, heighten security, and enhance incident response.
- › Expedited detection and response cycle time, due to AI's ability to rapidly quantify risks and accelerate analyst decision making with data-driven mitigation measures.
- › Increased cost savings for organizations due to enhanced up-front timely protection, prevention, and mitigation of cybersecurity breaches and malicious attacks.
- › Reduced errors and inconsistencies than manual and semimanual processes.
- › Less time spent shifting through alerts and chasing false positives, freeing security personnel to focus on potential critical threats.
- › Improved workforce satisfaction because cybersecurity professionals can focus on higher level tasks instead of mundane time-consuming manual actions.
- › Enhanced customer satisfaction and brand reputation due to heightened cybersecurity protection and increased trust in the organization's security measures.

- › Better defense capabilities, reducing the scope of potential risks and strengthening security postures.

### MALICIOUS AI

Malicious AI, also known as "offensive AI," refers to the use of AI and ML to augment cyberattacks, enabling adversaries to launch highly targeted, sophisticated attacks more quickly and broadly than ever. AI can help attackers create stealthier, faster, more effective attacks that blend into normal background activity. This makes the attacks almost impossible to counter using traditional security controls.

AI-driven security threats are not only increasing, they are changing the character of threats.<sup>20,21</sup> They:

- › *Expand existing threats:* With AI tools, even amateurs can quickly and inexpensively generate attacks, expanding the ability to a broader range of threat actors.
- › *Introduce new threats:* AI systems can easily launch complex, intense, sophisticated threats that would be impractical for humans.
- › *Change the typical character of threats:* AI-powered attacks can be more effective, more finely targeted, and more difficult to attribute, easily exploiting vulnerabilities in complex systems.

To realize how the character of threats is changing consider, for example, spear-phishing campaigns. Unlike typical phishing campaigns, spear-phishing attacks<sup>22</sup> are created with a specific audience in mind, using reconnaissance from social media. Tailoring attacks to a specific victim produces, on average, 40

times the click-through rate of its boilerplate counterpart. AI attacks like these require AI-powered defenses that generate surgical responses to in-progress cyberattacks as soon as they appear.

## SECURING AI

AI is vulnerable to cyberattacks. The growing use of AI in cyberdefense and several other applications gives cyber adversaries greater incentives to attack AI systems and applications. AI and ML algorithms that are deployed to protect other systems and help in security operations can be attacked and controlled by an adversary. Such attacks could have serious consequences in areas such as industrial control, critical infrastructure, and autonomous driving.

AI models present new attack surfaces that current defenses do not protect against. Potential attacks on AI include data poisoning, data biasing, data or model theft, adversarial input attack, and inference attacks.<sup>23,24</sup> Consider the following adversarial examples and their impacts:<sup>25</sup>

- › *Computer vision*: Images can be modified by adding adversarial patches so as to fool image classification systems (e.g., those used by autonomous vehicles).
- › *Speech recognition*: Adding adversarial noise to a speech waveform may result in wrong textual translation.
- › *Social bot detection*: Similar to computer vision and automatic speech recognition, adversarial attacks can alter the features of social bots, without impacting their activity, thus allowing them to evade detection.
- › *Fake news detection*: Tampering with the textual content of an article, or even with its comments, may yield wrong article classifications.

These AI attacks are fundamentally different from traditional cyberattacks. Although AI's benefits largely outweigh its security and privacy risks, it is important to secure AI systems and applications.

## CYBERSECURITY: A CAT AND MOUSE GAME

Cyber threats will become more ubiquitous, more frequent, more severe, and more sophisticated. Like security professionals, bad actors will use AI and other advanced tools to intensify and complicate their attacks and the attack surface will broaden. Essentially, managing cybersecurity will continue to be a game of cat and mouse.

Organizations need to future-proof their systems and data by frequently updating their security strategies

and deploying new solutions powered by AI and other technologies. To do that, they will need people with knowledge and experience in applying AI and ML to cybersecurity, along with other cybersecurity skill sets.

Here are a few recommendations for security professionals<sup>18</sup>:

- › Research how AI has been and can be applied to secure cybersystems. Stay abreast of emerging trends at the nexus of cybersecurity and AI.
- › Advance your career by taking courses and earning certifications related to AI/ML and cybersecurity.
- › Develop an AI-enabled cybersecurity strategy and a roadmap and revisit these regularly. Identify hot spots, critical blind spots, and tasks that require significant human effort and find innovative ways to address them. AI cybersecurity strategy should augment and be aligned with an organization's overall security strategy—it should not be an independent, standalone initiative.
- › Address adoption barriers such as lack of personnel with strong AI cybersecurity skills, difficulty in acquiring and retaining such talent, data complexity, and initial implementation costs.

## CONCLUSION

AI is—and will be—a significant ally against intensifying cyber threats. Clearly, there is a compelling business case for using AI in cybersecurity. About 70% of executives surveyed globally said their organization cannot identify nor respond to cyber threats without AI.<sup>26</sup> AI's benefits to cybersecurity largely outweigh its security risks. Nevertheless, it is important to secure AI systems and applications.

Cybersecurity professionals and businesses should explore the entire spectrum of AI's potential security solutions and choose the AI capabilities that address the threats and best suit the application's needs. Of course, AI is not a silver bullet. Humans and AI must work together. But AI can support security personnel to maintain robust security operations around the clock.

We hope you are now inspired to examine further and adopt the new frontier in digital security, the AI cybersecurity nexus. 🤖

## REFERENCES

1. S. Murugesan, "Securing health data amid heightened threats and looming vulnerabilities," *Amplify*, vol. 35, no. 3, pp. 17–22, Mar. 2022.
2. D. Temple-Raston, *A 'Worst Nightmare' Cyberattack: The Untold Story of the SolarWinds Hack*. WA, DC, USA: NPR, 2021.


3. C. Grove, "Hard lessons from the oldsmar water facility cyberattack hack," *Secur. Boulevard*, Feb.–Sep. 2021. [Online]. Available: <https://securityboulevard.com/2021/02/hard-lessons-from-the-oldsmar-water-facility-cyberattack-hack/>
4. "Cybersecurity challenges facing higher education," *Reciprocity*, Aug./Feb. 2021. [Online]. Available: <https://reciprocity.com/resources/cybersecurity-challenges-facing-higher-education/>
5. D. Lohrmann, "Cyber review: The year ransomware disrupted infrastructure," *Government Technol.*, 2021. [Online]. Available: <https://www.govtech.com/blogs/lohmann-on-cybersecurity/2021-cyber-review-the-year-ransomware-disrupted-infrastructure>
6. J. B. Michael and T. C. Wingfield, "Defensive AI, the future is yesterday," *Computer*, vol. 54, no. 9, pp. 90–96, Sep. 2021.
7. "Top 50 security threats," *Splunk*, 2020. [Online]. Available: <https://tinyurl.com/50secthreats>
8. S. Murugesan, "The cybersecurity renaissance: Security threats, risks, and safeguards," *IEEE India Council Newslett.*, vol. 14, no. 1, Jan.–Mar. 2019. [Online]. Available: <https://tinyurl.com/sec-renaissance>
9. Cybersecurity: An Interactive Guide by WEF, 2022. [Online]. Available: <https://intelligence.weforum.org/topics/a1Gb00000015LbsEAE>
10. D. Temple-Raston, A 'Worst Nightmare' Cyberattack: The Untold Story of the SolarWinds Hack. WA,DC, USA: NPR, Apr. 16, 2021. [Online]. Available: <https://www.npr.org/2021/04/16/985439655/a-worst-nightmare-cyberattack-the-untold-story-of-the-solarwinds-hack>
11. C. Grove, *Hard Lessons from the Oldsmar Water Facility Cyberattack Hack*. Woodlawn, MD, USA: Security Boulevard, Feb. 9, 2021. [Online]. Available: <https://securityboulevard.com/2021/02/hard-lessons-from-the-oldsmar-water-facility-cyberattack-hack/>
12. A. Preis, *Five Lessons from the JBS Attack for Securing the Manufacturing Supply Chain*. Woodlawn, MD, USA: Security Boulevard, 2021.
13. C. Fradkin, *Cyberattacks in 2021 Highlighted Critical Infrastructure Risks*. Woodlawn, MD, USA: Security Boulevard, 2021.
14. "Deepfakes: New cybersecurity threats," *Cutter Advisor*. Accessed: Aug. 14, 2022. [Online]. Available: <https://www.cutter.com/article/new-enterprise-cybersecurity-threat-deep-fakes>
15. S. Morgan, *Cybercrime to Cost the World \$10.5 Trillion Annually By 2025*. London, U.K.: Cybercrime Magazine, 2021.
16. Security Staff, "9 security threats in the metaverse, security," Aug. 10, 2022. [Online]. Available: <https://www.securitymagazine.com/articles/98142-9-security-threats-in-the-metaverse>
17. V. Varadharajan and N. Suri, "Security challenges when space merges with cyberspace," 2022. [Online]. Available: <https://tinyurl.com/space-security>
18. S. Murugesan, *Cyber AI: Leveraging the AI-Cybersecurity Nexus for Heightened Protection*. Arlington, MA, USA: Cutter Business Technol. Journal, 2021.
19. "8 Benefits of using AI for cybersecurity," Cyber Management Alliance, May 4, 2021. [Online]. Available: <https://www.cm-alliance.com/cybersecurity-blog/8-benefits-of-using-ai-for-cybersecurity>
20. M. Brundage et al., "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *Electron. Frontier Found.*, Feb. 2018. [Online]. Available: [https://www.eff.org/files/2018/02/20/malicious\\_ai\\_report\\_final.pdf](https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf)
21. 3 ways AI will change the nature of cyberattacks, World Economic Forum, 2019. [Online]. Available: <https://www.weforum.org/agenda/2019/06/ai-is-powering-a-new-generation-of-cyberattack-its-also-our-best-defence/>
22. "Offensive AI: Surfacing truth in the age of digital fakes," *Wired*. Accessed: Dec. 2021. [Online]. Available: <https://www.wired.com/brandlab/2020/02/offensive-ai-surfacing-truth-age-digital-fakes/>
23. M. Campbell, "Protecting AI: We built the brains, but what about helmets?," *Computer*, vol. 54, no. 12, Dec. 2021. [Online]. Available: <https://www.computer.org/csdl/magazine/co/2021/12/09622289/1yEUqoMIgHS>
24. J. Wolff, *How to Improve Cybersecurity For Artificial Intelligence*. Washington DC, USA: The Brookings Institution, Jun. 9, 2020. [Online]. Available: <https://www.brookings.edu/research/how-to-improve-cybersecurity-for-artificial-intelligence/>
25. "Adversarial machine learning for protecting against online manipulation," *IEEE Internet Comput.*, Mar./Apr. 2022. [Online]. Available: <https://www.computer.org/csdl/magazine/ic/2022/02/09627787/1yQwEnBRGHC>
26. "Reinventing cybersecurity with artificial intelligence: The new frontier in digital security," *Capgemini Res. Inst.*, 2019. [Online]. Available: <https://tinyurl.com/cyberai-camgemini>

**SAN MURUGESAN** is the director of BRITE Professional Services, Sydney, NSW, 2152, Australia. He is a fellow of the Australian Computer Society, golden core and BoG member of IEEE Computer Society, life senior member of IEEE, and distinguished speaker of IEEE and ACM. Contact him at [san@computer.org](mailto:san@computer.org) or visit his Webpage [www.tinyurl.com/san1bio](http://www.tinyurl.com/san1bio).



## DEPARTMENT: VIEW FROM THE CLOUD

# Serverless Computing for Scientific Applications

Maciej Malawski , Sano Centre for Computational Medicine, 30-054, Krakow, Poland, and also AGH University of Science and Technology, 31-150, Krakow, Poland

Bartosz Balis , AGH University of Science and Technology, 31-150, Krakow, Poland

*Serverless computing has become an important model in cloud computing and influenced the design of many applications. In this article, we provide our perspective on what the recent landscape of serverless computing for scientific applications looks like. We discuss the advantages and problems with serverless computing for scientific applications and, based on the analysis of existing solutions and approaches, we propose a science-oriented architecture for a serverless computing framework that is based on the existing designs. Finally, we provide an outlook of current trends and future directions.*

Given the increasing role of simulations and data analysis in science today, researchers never have too much computing power. For this reason, various dedicated research computing infrastructures are built, including high performance computing (HPC) centers, large-scale computing clusters or grids, or smaller centers for research computing at universities and research institutes. On the other hand, distributed computing in the industry leverages large-scale datacenters, which provide computing resources based on the cloud computing model. Over the past ten years, these commercial offerings in the form of public clouds have been of interest to the scientific community, and the development of cloud solutions influenced the way traditional HPC hardware and software have evolved. These technological trends initially have included virtualization, containerization, on-demand access to resources, or object storage services. With increasing cloud adoption, advanced *cloud-native* technologies have emerged, with the Kubernetes container orchestration technology being their cornerstone.<sup>1</sup>

Recently, we observed a new trend in cloud technologies, which is generally called “serverless computing.” In general, serverless computing allows executing functions with minimum overhead in server management, combining developments in microservice-based architectures, containers, and the new cloud service models, such as

function-as-a-service (FaaS) and container-as-a-service (CaaS).<sup>2,3</sup>

The serverless computing model has not been designed to support scientific computing, rather it has targeted lightweight event-based applications. Still, as the research community is very open to exploring new and alternative ways of accessing computing resources and building scientific applications, we can see many attempts to evaluate the applicability of the serverless model for scientific applications and the desire to repurpose it to the requirements of the scientific community.

In this article, we provide examples of using serverless model for scientific applications, based on our experience in this area. This provides our perspective on how the recent landscape of serverless computing for scientific applications looks like. We discuss the advantages and problems with serverless computing for scientific applications and based on the analysis of existing solutions and approaches, we propose a science-oriented architecture for a serverless computing framework that is based on the existing designs. Finally, we provide an outlook of current trends and future directions.

## STRENGTHS AND WEAKNESSES OF SERVERLESS MODEL FOR SCIENTIFIC APPLICATIONS

There are several benefits of using serverless model or FaaS.

- › *Function as a useful abstraction:* Functions are the most fundamental abstractions in mathematics, which is called the language of science. Functions are also a very powerful abstraction in

computer science, with theoretical foundations in the lambda calculus and the functional programming paradigm. Functional programming has proven to be useful for distributed systems, with the examples of Erlang and Scala languages and actor systems implemented in them.<sup>4</sup> In scientific computing, modern programming languages, such as Julia, are also strongly influenced by the functional programming paradigm. For these reasons, FaaS offers a natural abstraction for scientific applications using distributed computing.

- › *Simplicity facilitating programming:* The general premise of serverless computing is to facilitate application programming by hiding the complexity of underlying infrastructure and relieving the programmer or user from managing the infrastructure. All the intertwined resource management issues, such as provisioning, scheduling, or autoscaling, are in principle handled by the provider in a much broader scope than in any other cloud service model or in any distributed programming environment.
- › *Highly elastic resource management model:* In serverless computing, the unit of resource allocation is a single function call, and the FaaS platforms are designed for serving large amounts of fine-grained requests very quickly. Our experiments have shown that it is possible to request thousands of concurrent function calls and they are invoked in parallel within seconds by a cloud provider, such as AWS or Google.<sup>5</sup> The overhead is low as compared to IaaS clouds (minutes) or in HPC systems (hours). This opens the possibility of better support for interactive and dynamic workloads, which are of interest for scientific applications.
- › *Deployment model using familiar programming languages and containers:* While FaaS was originally designed for Web or mobile applications based on JavaScript, Java, or Python, now it is possible to add custom language support or container images such as Docker. This is fundamental for scientific computing, where applications typically require diverse programming languages, libraries, and tools.

Another highly popular approach for scientific computing in the cloud is using a *container orchestration platform*, with Kubernetes being a de facto standard, supported by all major cloud providers. Kubernetes can be seen as a middle ground between plain infrastructure-as-a-service cloud and FaaS. On the one hand, Kubernetes hides the complexity of cluster management; on the other hand, application development in Kubernetes still

requires significant IT engineering skills. In our research, we have investigated various aspects of scientific workflow management in Kubernetes. This perspective lets us point out a few problems of FaaS in the context of scientific computing and contrast them with Kubernetes.

- › *Vendor lock-in:* Serverless code usually uses various cloud services through vendor APIs which increases the chance of vendor lock-in. Developing a portable solution in FaaS is harder than in Kubernetes.
- › *Observability:* With platform and infrastructure hidden behind APIs, system observability is out of control of the developer. However, observability is the cornerstone of experimental science. Diagnosing problems and getting important metrics related to performance, energy consumption, etc., can be more difficult in FaaS than in Kubernetes, where the developer has full control over the observability stack.
- › *Economics and performance:* The serverless model is very attractive for production systems that run 24/7 and need to handle variable workload. In such cases, a possibly higher per-cycle cost of FaaS can be mitigated by high elasticity. However, for a scientist who runs one-off batch workloads this is not necessarily the case. It has been shown that for data-intensive workloads, such as model training, FaaS is considerably more expensive and slower than IaaS.<sup>6</sup> *Performance isolation* is another challenge; For example, Malla and Christensen,<sup>7</sup> achieved the best performance on IaaS by allocating one CPU to one computational task. Such control over resource management, also possible in Kubernetes, is not available on FaaS, where the resources are managed by the underlying platform.<sup>8</sup>

As we can see, there are numerous potential benefits and challenges of using serverless computing for scientific applications. In the next section, we show examples of how these have been addressed in the specific solutions targeting scientific computing.

## EXAMPLES OF SCIENTIFIC APPLICATIONS AND FRAMEWORKS USING FAAS

From the beginning of the serverless model and from the first releases of FaaS services, they have been noticed as potential sources of computing for scientific applications. Here, we provide a set of selected examples, which we think nicely represent typical scenarios.

## PyWren

Perhaps the first framework for running compute-intensive workloads on serverless platforms that received wider popularity was PyWren.<sup>9</sup> As a simple, yet powerful, Python library, it allows running stateful functions in parallel, using shared cloud storage for input and output, similarly to a tuple space model proposed in Linda. An interesting technical solution in PyWren is to use Cloudpickle Python library, which allows one to serialize and execute remotely arbitrary Python code. Cloudpickle was developed by PiCloud.com, a startup company that offered simplified computing based on Python functions invoked on AWS cloud more than five years before FaaS model was proposed. PyWren has been applied to many embarrassingly parallel problems, such as MonteCarlo simulations and parameter sweeps, but also for MapReduce style data processing tasks, video encoding, and parameter optimization for distributed machine learning (ML) or distributed compilation.

## FaaSification of Scientific Applications

One of the first discussions of the potential for using FaaS for scientific computing was presented by Spillner *et al.*<sup>10</sup> They described four experiments that compare the performance and resource consumption, for example, benchmarks or applications: calculation of  $\pi$ , face detection, password cracking, and precipitation forecast. The experiments are run using AWS Lambda and a local testbed using Snafu tool developed by Spillner *et al.*<sup>10</sup> In addition to performance evaluation, there is a very interesting discussion about possible strategies for FaaSification (adaptation of existing software to FaaS) of monolithic applications, which can be done at varying levels of granularity: from whole functions to single lines of code. This discussion brings an important topic of the effort needed to make use of the serverless computing model to existing applications and the potential for using tools for automation.

## Serverless Scientific Workflows

FaaS can be a good fit for scientific workflows (graphs of tasks), in particular those with large numbers of relatively fine-grained tasks. HyperFlow, our workflow engine developed at AGH, was extended to FaaS platforms,<sup>11</sup> including Google Cloud Functions, AWS Lambda, and other functions based on HTTP request interface. In HyperFlow, stateless functions operate in a download–compute–upload sequence, using cloud storage for input and output. Initially we had to work around deployment problems by building custom-compiled binaries compatible with the operating systems of the FaaS providers, but recent support for the

Docker images solved this problem. Similarly, CaaS serverless platforms, such as AWS Fargate and Google Cloud Run, proved to be a viable solution for scientific workflows.<sup>12</sup> Other interesting examples include Triggerflow,<sup>13</sup> an event-driven workflow framework based on triggers, and abstract function choreography language (AFCL), which offers high-level notation for workflows with a rich set of control- and data-flow constructs.<sup>14</sup> In our opinion, support for serverless backends will become a natural evolution of scientific workflow engines.

## NumPyWren

Seemingly, it is hard to imagine using serverless platforms for dense linear algebra, a domain traditionally reserved for HPC. Nevertheless, as NumPyWren and LambdaPACK<sup>15</sup> tools show algorithms, such as matrix multiplication or decomposition, can run efficiently in the cloud. NumPyWren uses cloud object storage for communication, and while the latency of Amazon S3 is orders of magnitude higher compared to MPI-over-Infiniband, the aggregate bandwidth and its scalability allows efficiently decomposing the matrix calculations into basic operations on tiles of such size that the high latency is compensated by high bandwidth. While the experiments show that the performance achieved does not immediately beat the MPI implementation, the benefits of serverless approach are scalability, elasticity, and better resource utilization. Notably, the framework wisely combines various additional cloud services: SQS for task queue and Redis or DynamoDB as key/value store for managing the state.

## ROOT Lambda—Serverless Tools for High Energy Physics

The High Energy Physics community, having a long tradition of leveraging distributed computing infrastructures, shows a growing interest in exploring modern frameworks coming from the big data industry. A notable example includes Distributed RDataFrame, an extension to ROOT framework for data analysis adding the high-level functionality based on the Data Frame model. RDataFrame compiles operations, such as filters and aggregations, into a graph of tasks and supports multiple backends, including local multiprocessing, Apache Spark, Dask, and recently AWS Lambda.<sup>16</sup> When developing the AWS backend, we solved many technical problems, including containerized deployment of ROOT and remote access to storage at CERN. Despite the relatively large volume of data transfer between CERN and AWS, this approach scales to at least hundreds of parallel Lambda tasks, allowing interactive data analysis. This exemplifies

the potential of using serverless infrastructures as backend to domain-specific scientific tools by providing user friendly abstractions.

### FuncX

A custom-developed solution that aims specifically at supporting scientific applications using the serverless model is FuncX.<sup>17</sup> It supports running FaaS applications on federated resources ranging from local computers, via clusters and clouds, to supercomputers, with focus on applications with fine grained tasks. Examples of such applications are scalable metadata extraction, ML inference, crystallography, neuroscience, correlation spectroscopy, and high energy physics. The interesting feature of the approach is that it does not simply provide access to FaaS platforms, but brings together resources from multiple sources, including those dedicated to scientific computing and, e.g., equipped with specialized hardware, such as GPU. The example applications of FuncX show that scientific computing often relies on many tasks which do not necessarily require a traditional supercomputer.

### High-Throughput Biomedical Application Examples

High-throughput computing is often required in biomedical applications for screening of large space of molecular configurations. Examples in proteomics are replica exchange molecular dynamics (REMD), which have been successfully ported to serverless computing.<sup>18</sup> The usage of serverless architecture allows for more dynamic scaling of the workers (executed on FaaS), while it requires adding a communication layer using cloud object storage or Redis database. A similar approach was used in serverless implementation of Smith–Waterman dynamic programming algorithm for comparing protein sequences.<sup>19</sup> A recent survey<sup>20</sup> shows several examples of various applications of serverless computing to omics data analysis and integration, all representing high-throughput architecture with a scalable pool of resources obtained using the FaaS or CaaS model.

### Distributed ML

When discussing scientific or computational applications, one cannot exclude training and serving of ML models, which can also be a subject of porting to serverless architectures. One of the early examples<sup>21</sup> uses AWS Lambda for the inference of large neural network models. SIREN<sup>22</sup> is a distributed framework running compute-intensive batches of ML tasks on FaaS. Another example is FedLess,<sup>23</sup> which is a serverless

framework for secure training of ML models using a federated learning approach.

## LAYERED ECOSYSTEM OF APPLICATIONS AND GENERIC ARCHITECTURE OF THE FRAMEWORKS

Having studied examples of serverless scientific applications, we can observe an emerging layered architecture of their ecosystem, as shown in Figure 1. From the bottom-up, we have the basic layer of cloud storage and communication, which includes cloud object storage, queue systems, or caches. This layer provides state management for the stateless FaaS/CaaS layer. Next come various processing models—each of them relevant for scientific users and software engineers. Finally, the top layer includes ready to use frameworks for scientific applications, which typically provide high-level APIs or user interfaces.

Another general observation is that most of the frameworks have a very similar common architecture, as shown in Figure 2. The main component, called Execution Engine, is responsible for the task orchestration, and it uses some form of database, typically a key–value store, for managing the internal state of the application. The tasks are submitted to a task queue and then are processed using stateless FaaS or CaaS services, which use cloud storage for data exchange. There are possible variations of this architecture, including more distributed or decentralized orchestration, some frameworks do not use any queue, but directly invoke FaaS or CaaS functions using the public API, finally there are multiple options regarding database and cloud storage backends (see Figure 1). Nevertheless, this typical architecture can be considered as a standard blueprint for building computing frameworks using the serverless model.

## CONCLUSION

The examples presented here show that the concept of serverless computing can be applied in scientific computing, where traditional distributed processing or high-throughput approaches have been used. The new capabilities offered by serverless, including simplified programming model based on functions, highly elastic resource management, and convenient deployment model, allow not only for repurposing of existing applications and frameworks (parallel tasks, workflows, MapReduce, etc.), but can inspire new classes of scientific applications, which can be more event driven, interactive, and highly dynamic in resource usage, and also take advantage of the whole

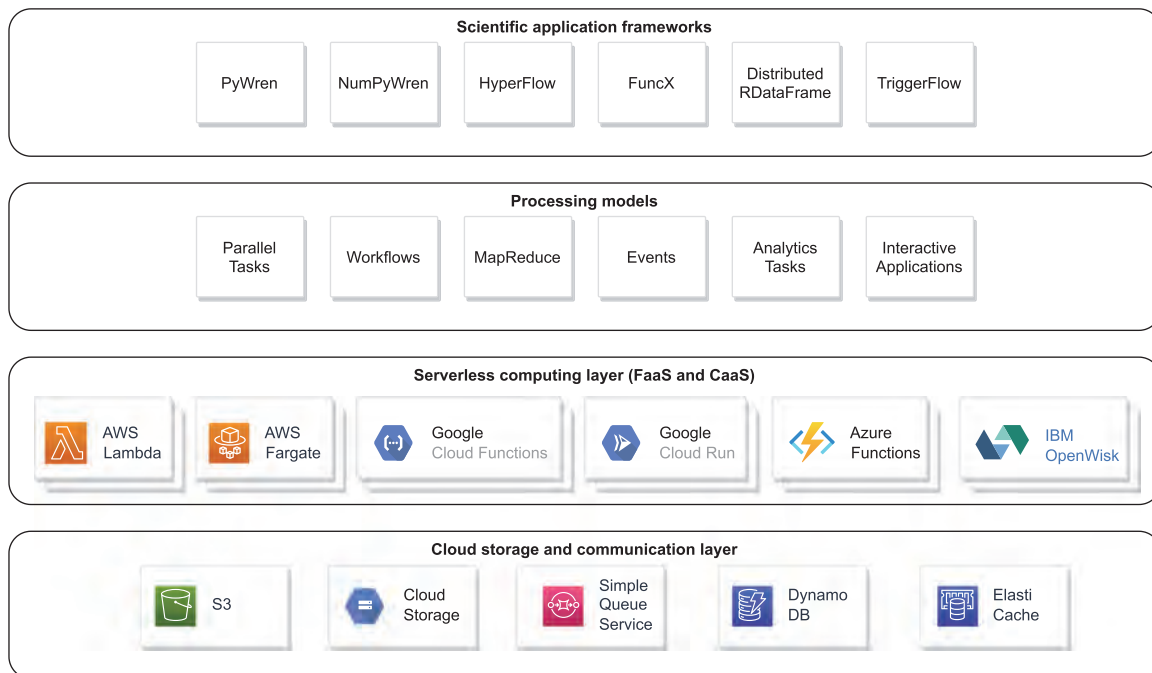


FIGURE 1. Layered ecosystem of serverless scientific applications.

continuum of resources from HPC, cloud, and other devices located at the edge.

There are of course limitations of serverless computing, such as vendor lock-in, observability issues, cost-performance tradeoffs, distributed state management, caching, and lack of tooling—these topics are now subject of active research.<sup>3</sup> Some trends, such as datacenter disaggregation,<sup>15</sup> convergence between HPC and cloud architectures, and increasingly elastic resource management in clouds, may suggest that some form of “serverless” computing will become prevalent. The future will show if a “serverless

supercomputer” may become an ultimate solution to scientific computing problems, but at least we are certain that the concepts presented here will influence the future developments in both compute infrastructures and the architectures of scientific applications. 🌐

## ACKNOWLEDGMENTS

This work was supported in part by the EU H2020 under Grant 857533, in part by the International Research Agendas Programme of the Foundation for Polish Science through Sano Project (<https://sano.science>), and in part by the European Regional Development Fund.

## REFERENCES

1. D. Gannon, R. Barga, and N. Sundaresan, “Cloud-Native applications,” *IEEE Cloud Comput.*, vol. 4, no. 5, pp. 16–21, Sep./Oct. 2017, doi: 10.1109/MCC.2017.4250939.
2. P. Castro, V. Ishakian, V. Muthusamy, and A. Slominski, “The rise of serverless computing,” *Commun. ACM*, vol. 62, no. 12, pp. 44–54, Nov. 2019, doi: 10.1145/3368454.
3. C. Abad, I. T. Foster, N. Herbst, and A. Iosup, “Serverless computing (Dagstuhl seminar 21201),” *Dagstuhl Rep.*, vol. 11, no. 4, pp. 34–93, 2021, doi: 10.4230/DagRep.11.4.34.
4. C. Hewitt, P. Bishop, and R. Steiger, “A universal modular ACTOR formalism for artificial intelligence,” in *Proc. 3rd Int. Joint Conf. Artif. Intell.*, 1973, pp. 235–245.

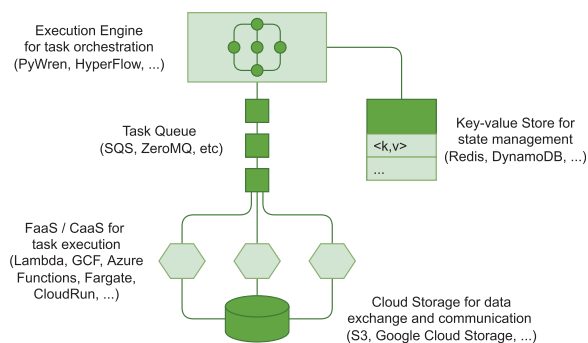


FIGURE 2. Generic architecture of a serverless execution framework for scientific applications.

5. K. Figiela, A. Gajek, A. Zima, B. Obrok, and M. Malawski, "Performance evaluation of heterogeneous cloud functions," *Concurrency Comput. Pract. Exp.*, vol. 30, no. 23, 2018, Art. no. e4792, doi: 10.1002/cpe.4792.
6. J. M. Hellerstein *et al.*, "Serverless computing: One step forward, two steps back," in *Proc. 9th Biennial Conf. Innovative Data Syst. Res.*, 2019. [Online]. Available: <https://sano.science>
7. S. Malla and K. Christensen, "HPC in the cloud: Performance comparison of function as a service (FaaS) vs infrastructure as a service (IaaS)," *Internet Technol. Lett.*, vol. 3, no. 1, 2020, Art. no. e137, doi: 10.1002/itl2.137.
8. E. van Eyk, A. Iosup, C. L. Abad, J. Grohmann, and S. Eismann, "A SPEC RG cloud group's vision on the performance challenges of FaaS cloud architectures," in *Proc. Companion 2018 ACM/SPEC Int. Conf. Perform. Eng.*, 2018, pp. 21–24, doi: 10.1145/3185768.3186308.
9. E. Jonas, Q. Pu, S. Venkataraman, I. Stoica, and B. Recht, "Occupy the cloud: Distributed computing for the 99%," in *Proc. Symp. Cloud Comput.*, 2017, pp. 445–451, doi: 10.1145/3127479.3128601.
10. J. Spillner, C. Mateos, and D. A. Monge, "FaaSter, better, cheaper: The prospect of serverless scientific computing and HPC," in *Proc. High Perform. Comput.*, 2018, pp. 154–168, doi: 10.1007/978-3-319-73353-1\_11.
11. M. Malawski, A. Gajek, A. Zima, B. Balis, and K. Figiela, "Serverless execution of scientific workflows: Experiments with hyperflow, AWS lambda and Google cloud functions," *Future Gener. Comput. Syst.*, vol. 110, pp. 502–514, Sep. 2020, doi: 10.1016/j.future.2017.10.029.
12. K. Burkat *et al.*, "Serverless containers – rising viable approach to scientific workflows," in *Proc. IEEE 17th Int. Conf. eScience*, 2021, pp. 40–49, doi: 10.1109/eScience51609.2021.00014.
13. A. Arjona, P. G. López, J. Sampé, A. Slominski, and L. Villard, "Triggerflow: Trigger-based orchestration of serverless workflows," *Future Gener. Comput. Syst.*, vol. 124, pp. 215–229, 2021, doi: 10.1016/j.future.2021.06.004.
14. S. Ristov, S. Pedratscher, and T. Fahringer, "AFCL: An abstract function choreography language for serverless workflow specification," *Future Gener. Comput. Syst.*, vol. 114, pp. 368–382, 2021, doi: 10.1016/j.future.2020.08.012.
15. V. Shankar *et al.*, "Serverless linear algebra," in *Proc. 11th ACM Symp. Cloud Comput.*, 2020, pp. 281–295, doi: 10.1145/3419111.3421287.
16. J. Kuśnierz *et al.*, "Serverless engine for high-energy physics distributed analysis," in *Proc. 22nd IEEE/ACM Int. Symp. Cluster, Cloud Internet Comput.*, 2022, pp. 13–16.
17. R. Chard *et al.*, "funcX: A federated function serving fabric for science," in *Proc. 29th Int. Symp. High-Perform. Parallel Distrib. Comput.*, 2020, pp. 65–76, doi: 10.1145/3369583.3392683.
18. M. E. Mirabelli, P. García-López, and G. Vernik, "Bringing scaling transparency to proteomics applications with serverless computing," in *Proc. 6th Int. Workshop Serverless Comput.*, 2020, pp. 55–60, doi: 10.1145/3429880.3430101.
19. X. Niu, D. Kumanov, L.-H. Hung, W. Lloyd, and K. Y. Yeung, "Leveraging serverless computing to improve performance for sequence comparison," in *Proc. 10th ACM Int. Conf. Bioinf., Comput. Biol. Health Inform.*, 2019, pp. 683–687, doi: 10.1145/3307339.3343465.
20. P. Grzesik, D. R. Augustyn, Ł. Wyciślik, and D. Mrozek, "Serverless computing in Omics data analysis and integration," *Brief. Bioinf.*, vol. 23, no. 1, Jan. 2022, Art. no. bbab349, doi: 10.1093/bib/bbab349.
21. V. Ishakian, V. Muthusamy, and A. Slominski, "Serving deep learning models in a serverless platform," in *Proc. IEEE Int. Conf. Cloud Eng.*, 2018, pp. 257–262, doi: 10.1109/IC2E.2018.00052.
22. H. Wang, D. Niu, and B. Li, "Distributed machine learning with a serverless architecture," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 1288–1296, doi: 10.1109/INFOCOM.2019.8737391.
23. A. Grafberger, M. Chadha, A. Jindal, J. Gu, and M. Gerndt, "FedLess: Secure and scalable federated learning using serverless computing," in *Proc. IEEE Int. Conf. Big Data*, 2021, pp. 164–173, doi: 10.1109/BigData52589.2021.9672067.

**MACIEJ MALAWSKI** is a research team leader on extreme scale data and computing at Sano Centre for Computational Medicine, 30-054, Kraków, Poland, and an associate professor at the Institute of Computer Science, AGH University of Science and Technology, 31-150, Krakow, Poland. His research interests include parallel and distributed computing, large-scale data analysis, cloud technologies, and scientific applications. Malawski received his Ph.D. degree in computer science. Contact him at [m.malawski@sanoscience.org](mailto:m.malawski@sanoscience.org).

**BARTOSZ BALIS** is an associate professor at the Institute of Computer Science, AGH University of Science and Technology, 31-150, Krakow, Poland. His research interests include environments for eScience, cloud computing, infrastructure automation, observability, and scientific workflows. Balis received his Ph.D. degree in computer science from the AGH University of Science and Technology. Contact him at [balis@agh.edu.pl](mailto:balis@agh.edu.pl).

DEPARTMENT:  
SOFTWARE ENGINEERING RADIO

# Randy Shoup on Evolving Architecture and Organization at eBay

Jeremy Jung 

## FROM THE EDITOR

In Episode 525 of “Software Engineering Radio,” Randy Shoup of eBay discusses the evolution of eBay’s tech stack with host Jeremy Jung. Topics include eBay’s origins, its five-year migration to multiple Java services, database sharing between old and new systems, building a distributed tracing system, advantages of cloud, why services should own their own data storage, effects of scale, rejoining a former company, choosing what to work on first, the book *Accelerate*, and improving delivery time. We provide summary excerpts in this article; to hear the full interview, visit <http://www.se-radio.net> or access our archives via RSS at <http://feeds.feedburner.com/se-radio>. —Robert Blumen

**Jeremy Jung: Fewer software-as-a-service products were available in 2000. What services are available today that weren’t available then?**

**Randy Shoup:** There was no cloud until 2006. There were a few vendors like Salesforce, but I couldn’t simply pay them to operate a technological software service. There were no monitoring vendors. Today, I would instrument everything with OpenTelemetry. For my back end, I would choose a distributed-tracing vendor. We also didn’t have distributed logging.

We built our own data centers, racked our own servers, and installed all the OSS in them. We still do all that because it’s cheaper for us at our scale, but the software developer in 2022 has this massive menu of options and can get a lot done through cloud vendors, software-service vendors, and so on.

In software, every year is better than the previous year. At that time, we were excited that we had all the tools and capabilities that we did have. The big companies rolled their own, and the most you could pay anybody else to do was rack your servers, but installing and operating software was your job.

**If eBay had started in the last 10 years, would it have made sense to start on a public cloud and then move to its own infrastructure later? Or did it make sense to start with your own infrastructure from the start?**

No one should ever start by building their own servers and their own cloud. You might outgrow the cloud vendors, but that doesn’t happen often. When it does, people write articles about it. Dropbox is a good example. By 2010–2012, the cloud had proven itself. Anybody who started since then should absolutely have started in the public cloud. Over time as the cloud bill grows, it could make sense to shift toward building and operating your own data centers, but it’s a big investment and takes years to develop the necessary internal capabilities. The more common migration is from proprietary data centers and colos into the public cloud.

---

Digital Object Identifier 10.1109/MS.2022.3210788  
Date of current version: 23 December 2022

## SOFTWARE ENGINEERING RADIO

Visit [www.se-radio.net](http://www.se-radio.net) to listen to these and other insightful hour-long podcasts.

### RECENT EPISODES

- » 532 — Peter Wyatt, CTO at PDF Association and project coleader of ISO 32000, and Duff Johnson, CEO at PDF Association and ISO Project coleader and US TAG chair for both ISO 32000 and ISO 14289 (PDF/UA), discuss the 30-year history of the PDF with host Gavin Henry.
- » 530 — Tanmai Gopal, CEO of Hasura.io, joins host Jeff Doolittle for a conversation about GraphQL.
- » 528 — Jonathan Shariat, coauthor of the book *Tragic Design*, discusses harmful software design with host Jeremy Jung.

### UPCOMING EPISODES

- » Eddie Aftandilan and host Priyanka Raghavan discuss Github Copilot.
- » Host Akshay Manchale talks to Andy Dang about AI/ML observability.
- » Dan Lorenc and host Robert Blumen discuss supply chain attacks.

---

*THE BIG COMPANIES ROLLED THEIR OWN, AND THE MOST YOU COULD PAY ANYBODY ELSE TO DO WAS RACK YOUR SERVERS, BUT INSTALLING AND OPERATING SOFTWARE WAS YOUR JOB.*

---

### What would you have done differently in your early days with eBay with the knowledge you have now, but the technology that existed then?

I would have moved us directly to what we would now call microservices—individual services that own their own data storage and that are only interacted with through the public interface. Amazon transitioned

from a monolith into microservices between 2000 and 2005. There's a famous Jeff Bezos memo from the early part of that that included the requirement that you never talk to anybody else's database, and you interact with other services only through their public interfaces. They didn't standardize around CORBA or JavaScript Object Notation (JSON) or GRPC, which didn't exist at the time, or around any particular interaction mechanism. But they did need to have this kind of microservice capability. That's modern terminology, where services own their own data, and nobody can talk in the back door. That is the one architectural thing that I wish in hindsight that I would have brought in, because that does help a lot. Amazon was pioneering in that approach, and a lot of people inside and outside Amazon didn't think it would work, but it did, famously.

### Microservice to you means having its own data store?

Several of the distinguishing characteristics are size and scope of the interface—the *micro* in *microservice*. You can have a service-oriented architecture with one big service, or some small number of large services. It may not have only one operation, but it doesn't have 1,000. And the handful or the several handfuls of operations are all about one particular thing. The other part of it that is critical is that the service must own its own data storage.

### When you started your new job at eBay, how did you figure out where the problems are or what to do next?

I lead the eBay velocity initiative, which is about delivering features and bug fixes more quickly to customers. We produced a value-stream map. That's a term from Lean where you look end to end at all the steps in a process and how long those steps take. Each step produces some value—a feature, some revenue, or something that helps the customer or the business.

Then you look for opportunities to improve—if a step takes five days, that is worth optimizing. We didn't talk with all 4,000 engineers or every team we had, but we sampled a few. After talking with three teams, we were already hearing some of the same things. We saw that software delivery was our current bottleneck—the amount of time it takes from an



engineer committing code to the code showing up as a feature on the site. Two years ago, before we started, the average was a week and a half. Now, for the teams that we've been working with, it's down to two days.

We used a book called *Accelerate* by Nicole Forsgren, Jez Humble, and Gene Kim (2018). If there's one book anybody should read about software engineering, it's that. It summarizes almost a decade of research from the State of DevOps reports. When the problem is software delivery, the book tells you all the continuous-delivery techniques, trunk-based development, and other things you can do to solve those problems.

Companies cluster. Organizations that are not good at deployment frequency and lead time are also not good at the quality metrics of mean time to restore and change failure rate. And the companies that are

excellent at deployment frequency and lead time are also excellent at lead time to recover and change failure rate. Companies or organizations are divided into low, medium, high, and elite performers. On average at the time and still on average, eBay is solidly in that medium-performer category. We've been able to move teams we worked with to the high category. So, we focused on moving the whole set of teams from that medium-performer category, where things are measured in weeks, to the high-performer category, where things are measured in days. 🙌



**JEREMY JUNG** is a technical lead in California, USA. Contact him at <https://www.jertype.com> or [contact@jeremyjung.com](mailto:contact@jeremyjung.com).

## IEEE Computer Society Has You Covered!

**WORLD-CLASS CONFERENCES** — Over 189 globally recognized conferences.

**DIGITAL LIBRARY** — Over 893k articles covering world-class peer-reviewed content.

**CALLS FOR PAPERS** — Write and present your ground-breaking accomplishments.

**EDUCATION** — Strengthen your resume with the IEEE Computer Society Course Catalog.

**ADVANCE YOUR CAREER** — Search new positions in the IEEE Computer Society Career Center.

**NETWORK** — Make connections in local Region, Section, and Chapter activities.

Explore all of the member benefits at [www.computer.org](http://www.computer.org) today!



# Get Published in the New *IEEE Open Journal of the Computer Society*

**Submit a paper to the new IEEE Open Journal of the Computer Society covering computing and informational technology.**

Your research will benefit from the IEEE marketing launch and 5 million unique monthly users of the IEEE *Xplore*® Digital Library. Plus, this journal is fully open and compliant with funder mandates, including Plan S.

**Submit your paper today!**

Visit [www.computer.org/oj](http://www.computer.org/oj) to learn more.



# Call for Articles

## IEEE Pervasive Computing

seeks accessible, useful papers on the latest peer-reviewed developments in pervasive, mobile, and ubiquitous computing. Topics include hardware technology, software infrastructure, real-world sensing and interaction, human-computer interaction, and systems considerations, including deployment, scalability, security, and privacy.

### Author guidelines:

[www.computer.org/mc/pervasive/author.htm](http://www.computer.org/mc/pervasive/author.htm)

### Further details:

[pervasive@computer.org](mailto:pervasive@computer.org)  
[www.computer.org/pervasive](http://www.computer.org/pervasive)



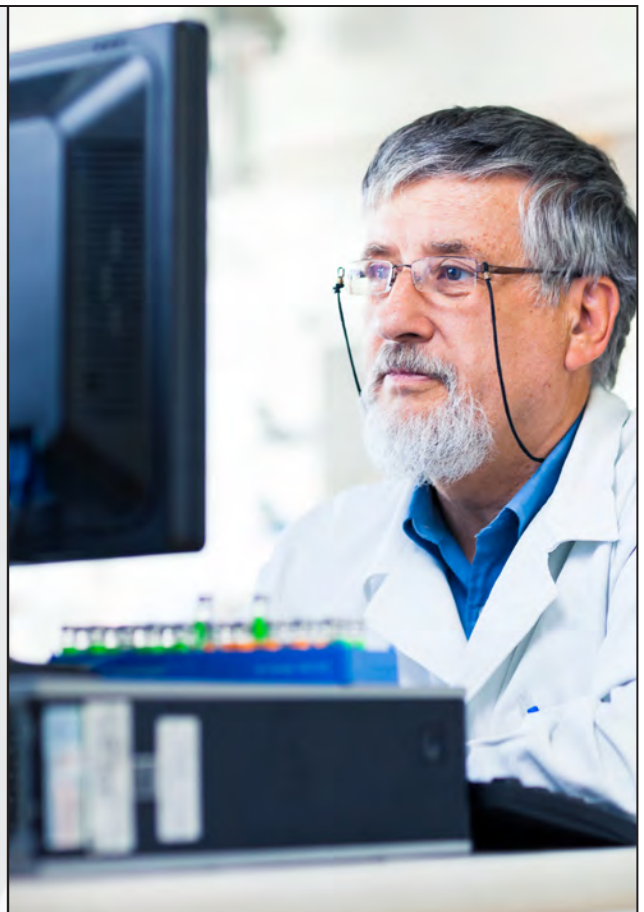
## Computing in Science & Engineering

The computational and data-centric problems faced by scientists and engineers transcend disciplines. There is a need to share knowledge of algorithms, software, and architectures, and to transmit lessons-learned to a broad scientific audience. *Computing in Science & Engineering (CiSE)* is a cross-disciplinary, international publication that meets this need by presenting contributions of high interest and educational value from a variety of fields, including physics, biology, chemistry, and astronomy. *CiSE* emphasizes innovative applications in cutting-edge techniques. *CiSE* publishes peer-reviewed research articles, as well as departments spanning news and analyses, topical reviews, tutorials, case studies, and more.

Read *CiSE* today! [www.computer.org/cise](http://www.computer.org/cise)



IEEE  
COMPUTER  
SOCIETY



# Publications Seek 2025 Editors in Chief

**Application Deadline: 1 March 2024**

The IEEE Computer Society seeks applicants for the position of editor in chief for the following publications:

- *Computer* magazine
- *IEEE/ACM Transactions on Computational Biology and Bioinformatics*
- *IEEE Computer Architecture Letters*
- *IEEE Intelligent Systems* magazine
- *IEEE Open Journal of the Computer Society*
- *IEEE Transactions on Big Data*
- *IEEE Transactions on Cloud Computing*
- *IEEE Transactions on Sustainable Computing*

Computer Society publications are the cornerstone of professional activities for our members and the community we serve. We seek candidates who are IEEE members in good standing, have strong familiarity with our publications, and possess an excellent understanding of the field as it relates to academic, industry, and governmental areas. Applicants must also have successful experience attracting and developing a diverse team of talented and respected individuals to serve key editorial board roles. Demonstrated managerial skills are also required, as they are necessary to ensure rich content and issue development, along with timely processing of submissions through the editorial cycle. Terms begin 1 January 2025.

**For complete information on how to apply, please go to [www.computer.org/press-room/seeking-2025-editors-in-chief](http://www.computer.org/press-room/seeking-2025-editors-in-chief)**



**Apply Today!**



IEEE

# SECURITY & PRIVACY

*IEEE Security & Privacy* is a bimonthly magazine communicating advances in security, privacy, and dependability in a way that is useful to a broad section of the professional community.

The magazine provides articles with both a practical and research bent by the top thinkers in the field of security and privacy, along with case studies, surveys, tutorials, columns, and in-depth interviews. Topics include:

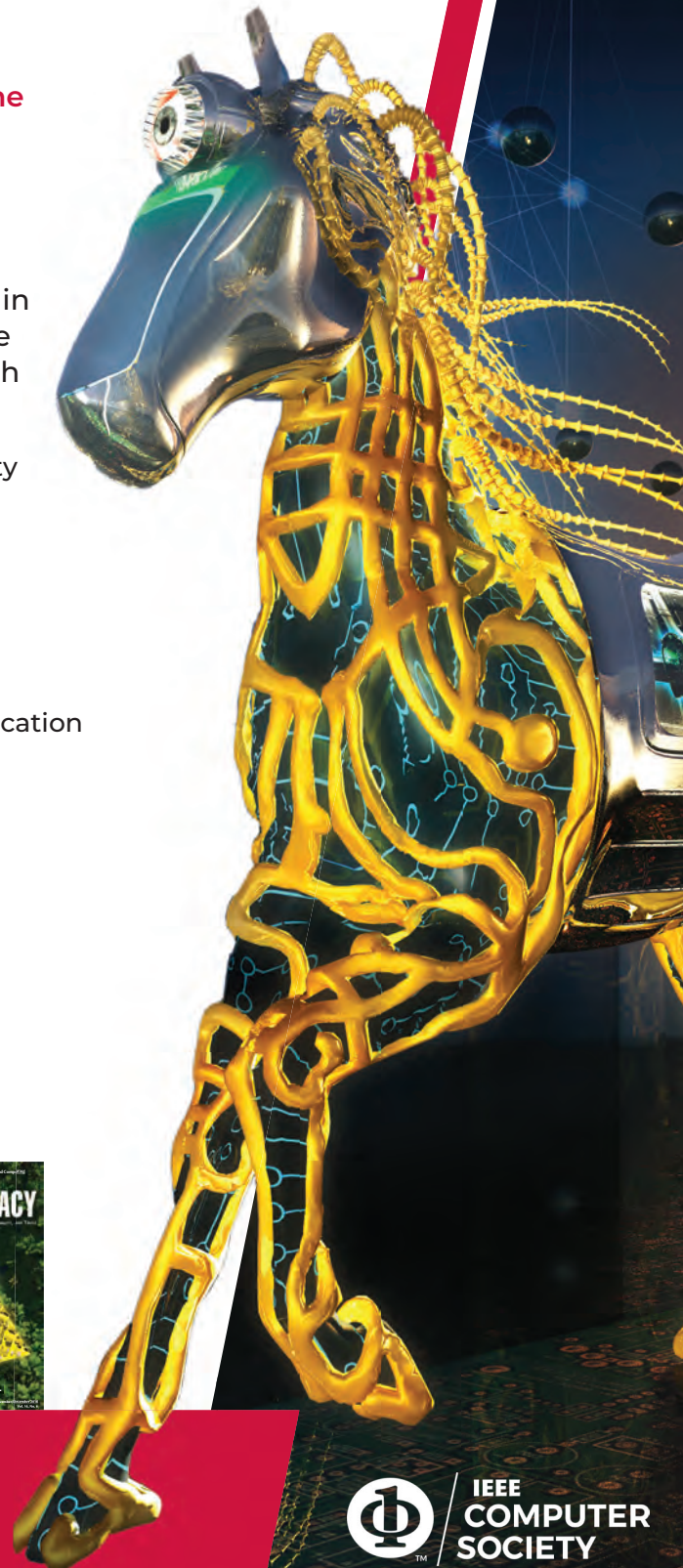
- Internet, software, hardware, and systems security
- Legal and ethical issues and privacy concerns
- Privacy-enhancing technologies
- Data analytics for security and privacy
- Usable security
- Integrated security design methods
- Security of critical infrastructures
- Pedagogical and curricular issues in security education
- Security issues in wireless and mobile networks
- Real-world cryptography
- Emerging technologies, operational resilience, and edge computing
- Cybercrime and forensics, and much more

[www.computer.org/security](http://www.computer.org/security)



Join the IEEE Computer Society  
for subscription discounts today!

[www.computer.org/product/magazines/security-and-privacy](http://www.computer.org/product/magazines/security-and-privacy)



# IEEE Internet Computing

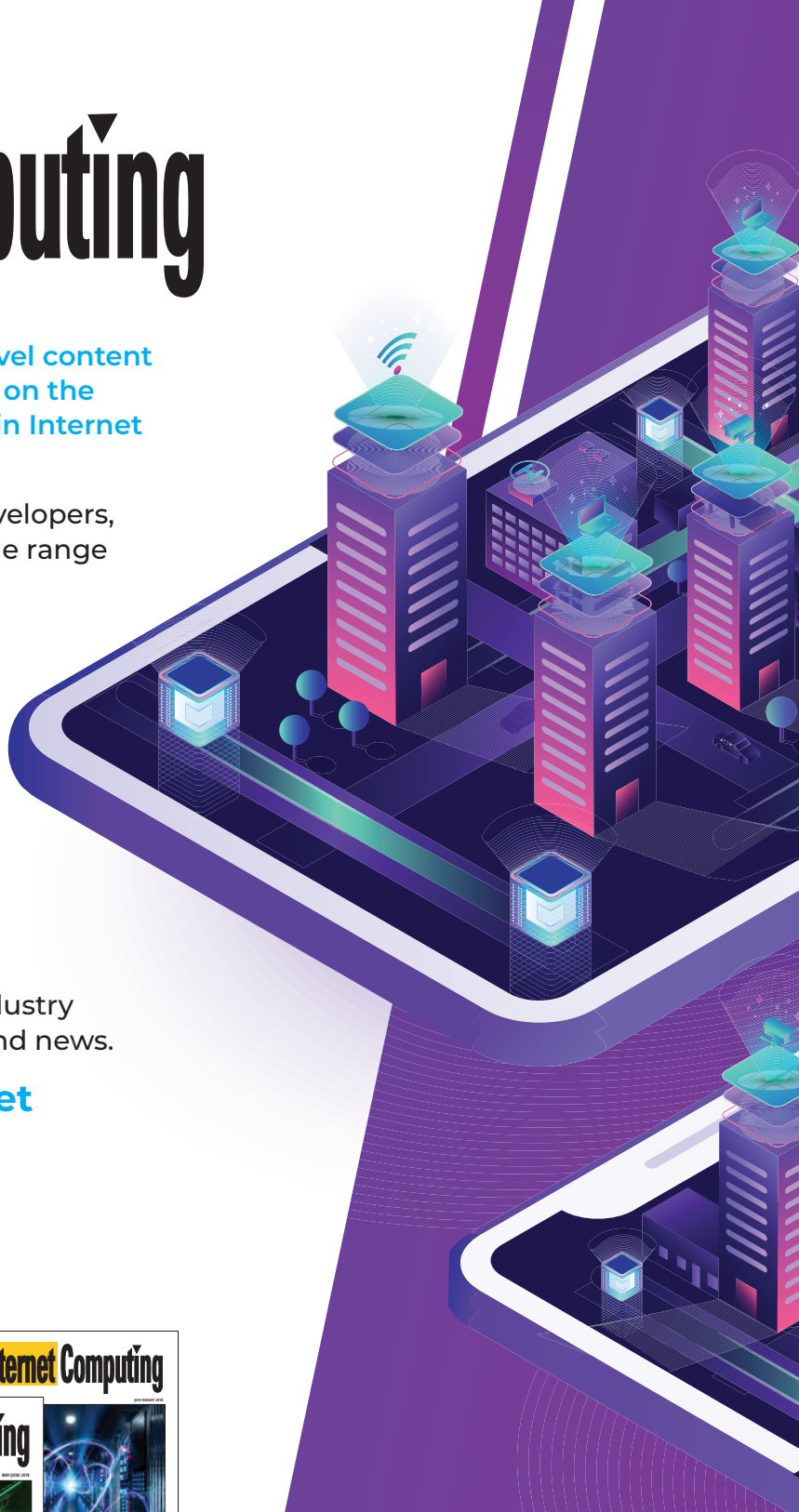
*IEEE Internet Computing* delivers novel content from academic and industry experts on the latest developments and key trends in Internet technologies and applications.

Written by and for both users and developers, the bimonthly magazine covers a wide range of topics, including:

- Applications
- Architectures
- Big data analytics
- Cloud and edge computing
- Information management
- Middleware
- Security and privacy
- Standards
- And much more

In addition to peer-reviewed articles, *IEEE Internet Computing* features industry reports, surveys, tutorials, columns, and news.

[www.computer.org/internet](http://www.computer.org/internet)



Join the IEEE Computer Society  
for subscription discounts today!

[www.computer.org/product/magazines/internet-computing](http://www.computer.org/product/magazines/internet-computing)





# IEEE Computer Society Volunteer Service Awards

*Nominations accepted throughout the year.*

## **T. Michael Elliott Distinguished Service Certificate**

Highest service award in recognition for distinguished service to the IEEE Computer Society at a level of dedication rarely demonstrated. i.e., initiating a Society program or conference, continuing officership, or long-term and active service on Society committees.

## **Meritorious Service Certificate**

Second highest level service certificate for meritorious service to an IEEE Computer Society-sponsored activity. i.e., significant as an editorship, committee, Computer Society officer, or conference general or program chair.

## **Outstanding Contribution Certificate**

Third highest level service certificate for a specific achievement of major value to the IEEE Computer Society, i.e., launching a major conference series, a specific publication, standards and model curricula.

## **Continuous Service Certificate**

Recognize and encourage ongoing involvement of volunteers in IEEE Computer Society programs. The initial certificate may be awarded after three years of continuous service.

## **Certificate of Appreciation**

Areas of contribution would include service with a conference organizing or program committee. May be given to subcommittee members in lieu of a letter of appreciation.



## **Nominations**

Submit your nomination at  
<http://bit.ly/computersocietyawards>

Contact us at  
[awards@computer.org](mailto:awards@computer.org)

IEEE

# COMPUTER ARCHITECTURE

# LETTERS

*IEEE Computer Architecture Letters* is a forum for fast publication of new, high-quality ideas in the form of short, critically refereed technical papers. Submissions are accepted on a continuing basis and letters will be published shortly after acceptance in IEEE Xplore and in the Computer Society Digital Library.

Submissions are welcomed on any topic in computer architecture, especially:

- Microprocessor and multiprocessor systems
- Microarchitecture and ILP processors
- Workload characterization
- Performance evaluation and simulation techniques
- Interactions with compilers and operating systems
- Interconnection network architectures
- Memory and cache systems
- Power and thermal issues at the architectural level
- I/O architectures and techniques
- Independent validation of previously published results
- Analysis of unsuccessful techniques
- Domain-specific processor architecture (embedded, graphics, network)
- High-availability architectures
- Reconfigurable computer architectures

[www.computer.org/cal](http://www.computer.org/cal)



Join the IEEE Computer Society  
for subscription discounts today!

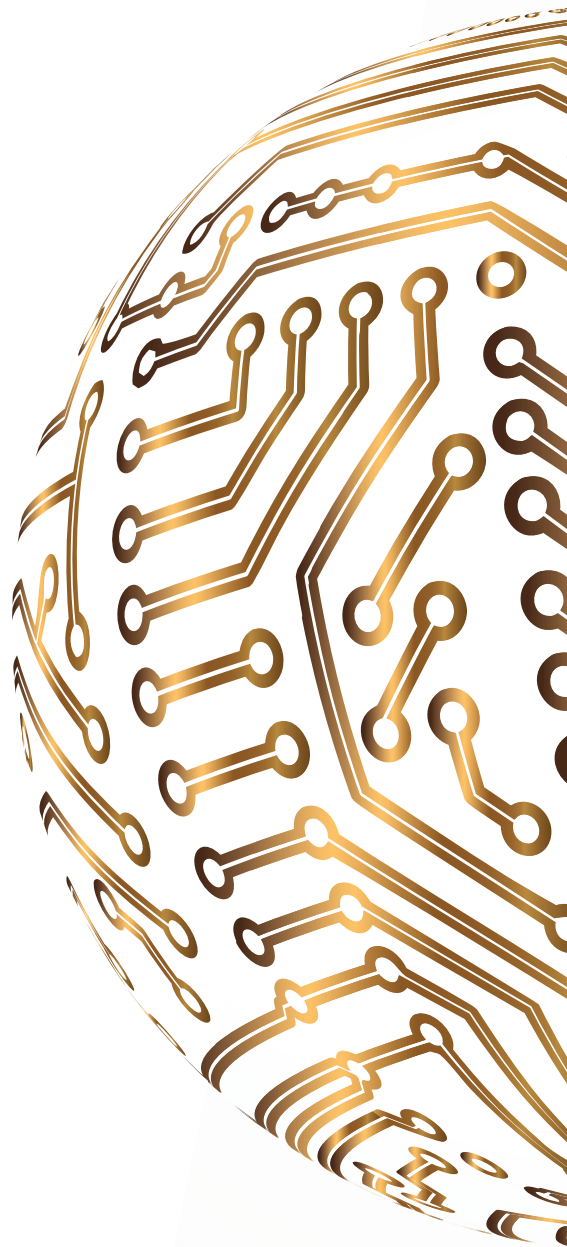
[www.computer.org/product/journals/cal](http://www.computer.org/product/journals/cal)



IEEE  
COMPUTER  
SOCIETY



IEEE





# Conference Calendar

IEEE Computer Society conferences are valuable forums for learning on broad and dynamically shifting topics from within the computing profession. With over 200 conferences featuring leading experts and thought leaders, we have an event that is right for you. Questions? Contact [conferences@computer.org](mailto:conferences@computer.org).

## DECEMBER

### 1 December

- ICDM (IEEE Int'l Conf. on Data Mining), Shanghai, China

### 4 December

- CloudCom (IEEE Int'l Conf. on Cloud Computing Technology and Science), Napoli, Italy
- CSDE (IEEE Asia-Pacific Conf. on Computer Science and Data Eng.), Nadi, Fiji
- ICA (IEEE Int'l Conf. on Agents), Kyoto, Japan
- UCC (IEEE/ACM Int'l Conf. on Utility and Cloud Computing), Taormina, Italy

### 5 December

- BIBM (IEEE Int'l Conf. on Bioinformatics and Biomedicine), Istanbul, Turkey
- ICRC (IEEE Int'l Conf. on Rebooting Computing), San Diego, CA, USA
- RTSS (IEEE Real-Time Systems Symposium), Taipei, Taiwan

### 6 December

- SEC (IEEE/ACM Symp. on Edge Computing), Wilmington, Delaware, USA

### 11 December

- ICAMLDL (Int'l Conf. on Advanced Machine Learning and Deep Learning), Raipur, India
- IRC (IEEE Int'l Conf. on Robotic

Computing), Laguna Hills, CA, USA

- ISM (IEEE Int'l Symposium on Multimedia), Laguna Hills, USA

### 14 December

- BCD (IEEE/ACIS Int'l Conf. on Big Data, Cloud Computing, and Data Science Eng.), Ho Chi Minh City, Vietnam

### 15 December

- BigData (IEEE Int'l Conf. on Big Data), Sorrento, Italy

### 18 December

- HiPC (IEEE Int'l Conf. on High-Performance Computing, Data, and Analytics), Goa, India
- iSES (IEEE Int'l Symposium on Smart Electronic Systems), Ahmedabad, India

## 2024

## JANUARY

### 3 January

- WACV (IEEE/CVF Winter Conf. on Applications of Computer Vision), Waikoloa, USA

### 17 January

- AIXVR (IEEE Int'l Conf. on Artificial Intelligence eXtended and Virtual Reality), Los Angeles, USA
- ICOIN (Int'l Conf. on Information

Networking), Ho Chi Minh City, Vietnam

### 27 January

- ASSIC (Int'l Conf. on Advancements in Smart, Secure and Intelligent Computing), Bhubaneswar, India

## FEBRUARY

### 1 February

- BICE (Black Issues in Computing Education Symposium), Santo Domingo, Dominican Republic

### 5 February

- AIMHC (IEEE Int'l Conf. on Artificial Intelligence for Medicine, Health and Care), Laguna Hills, USA
- CDKE (IEEE Int'l Conf. on Conversational Data & Knowledge Eng.), Laguna Hills, CA, USA
- ICSC (IEEE Int'l Conf. on Semantic Computing), Laguna Hills, USA

## MARCH

### 2 March

- CGO (IEEE/ACM Int'l Symposium on Code Generation and Optimization), Edinburgh, UK
- HPCA (IEEE Int'l Symposium on High-Performance Computer Architecture), Edinburgh, UK



#### 11 March

- PerCom (IEEE Int'l Conf. on Pervasive Computing and Communications), Biarritz, France

#### 12 March

- SANER (IEEE Int'l Conf. on Software Analysis, Evolution and Reengineering), Rovaniemi, Finland

#### 16 March

- VR (IEEE Conf. on Virtual Reality and 3D User Interfaces), Orlando, USA

#### 17 March

- SSIAl (IEEE Southwest Symposium on Image Analysis and Interpretation), Santa Fe, New Mexico, USA

---

### APRIL

#### 10 April

- SaTML (IEEE Conf. on Secure and Trustworthy Machine Learning), Toronto, Canada

#### 16 April

- ICDE (IEEE Int'l Conf. on Data Eng.), Utrecht, The Netherlands

#### 22 April

- VTS (IEEE VLSI Test Symposium), Tempe, Arizona, USA

#### 23 April

- PacificVIS (IEEE Pacific Visualization Symposium), Tokyo, Japan

#### 29 April

- DCOSS-IoT (Int'l Conf. on Distributed Computing in Smart Systems and the Internet of Things), Abu Dhabi, United Arab Emirates

---

### MAY

#### 6 May

- CCGRID (IEEE Int'l Symposium on Cluster, Cloud and Internet Computing), Philadelphia, USA
- HOST (IEEE Int'l Symposium on Hardware Oriented Security and Trust), Tysons Corner, Virginia, USA

#### 13 May

- ICCPS (ACM/IEEE Int'l Conf. on Cyber-Physical Systems), Hong Kong
- RTAS (IEEE Real-Time and Embedded Technology and Applications Symposium), Hong Kong

#### 20 May

- SP (IEEE Symposium on Security and Privacy), San Francisco, USA

#### 21 May

- ISORC (IEEE Int'l Symposium on Real-Time Distributed Computing), Tunis, Tunisia

#### 27 May

- FG (IEEE Int'l Conf. on Automatic Face and Gesture Recognition), Istanbul, Turkey
- ICST (IEEE Conf. on Software Testing, Verification and Validation), Toronto, Canada
- IPDPS (IEEE Int'l Parallel and Distributed Processing Symposium), San Francisco, USA

#### 28 May

- ISMVL (IEEE Int'l Symposium on Multiple-Valued Logic), Brno, Czech Republic

---

### JUNE

#### 4 June

- ICSA (IEEE Int'l Conf. on Software Architecture), Hyderabad, India

#### 16 June

- CVPR (IEEE/CVF Conf. on Computer Vision and Pattern Recognition), Seattle, USA

#### 19 June

- CHASE (IEEE/ACM Conf. on Connected Health: Applications, Systems and Eng. Technologies), Wilmington, Delaware, USA

#### 24 June

- DSN (IEEE/IFIP Int'l Conf. on Dependable Systems and Networks), Brisbane, Australia
- RE (IEEE Int'l Requirements Eng. Conf.), Reykjavik, Iceland

#### 25 June

- CAI (IEEE Conf. on Artificial Intelligence), Singapore

Learn more  
about IEEE  
Computer  
Society  
conferences

[computer.org/conferences](https://computer.org/conferences)

# Evolving Career Opportunities

*Explore new options—upload your resume today*

[careers.computer.org](https://careers.computer.org)

Changes in the marketplace shift demands for vital skills and talent. The **IEEE Computer Society Career Center** is a valuable resource tool to keep job seekers up to date on the dynamic career opportunities offered by employers.

Take advantage of these special resources for job seekers:



JOB ALERTS



TEMPLATES



WEBINARS



CAREER  
ADVICE



RESUMES VIEWED  
BY TOP EMPLOYERS

No matter what your career level, the IEEE Computer Society Career Center keeps you connected to workplace trends and exciting career prospects.



IEEE  
COMPUTER  
SOCIETY



IEEE