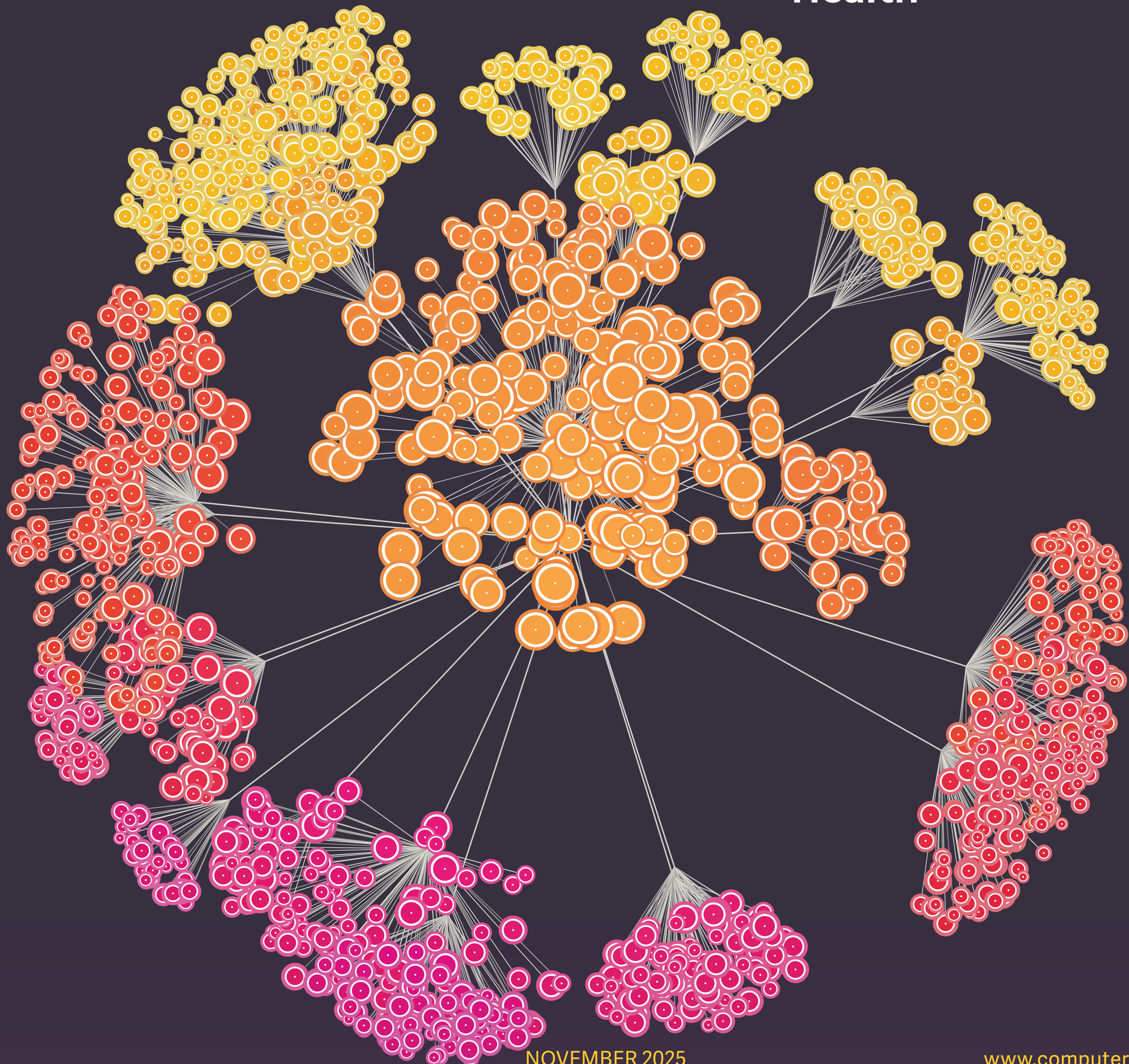
COMPUTING

edge

- Visualization
- Distributed Computing
- Generative AI
- Health



NOVEMBER 2025

www.computer.org

IEEE COMPUTER SOCIETY

IEEE

COMPUTING edge

IEEE COMPUTER SOCIETY computer.org

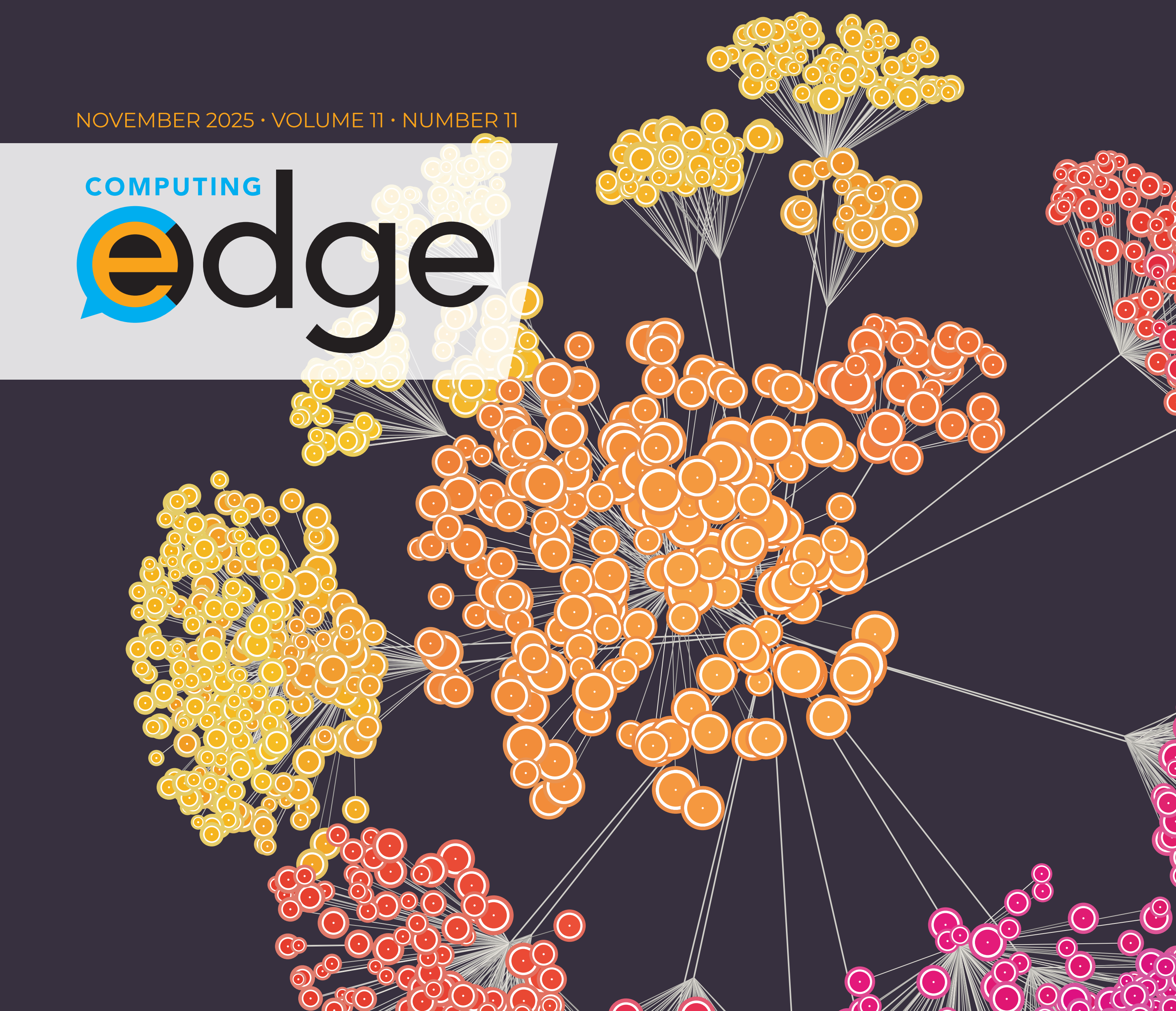
Printed with inks containing soy and/or vegetable oils

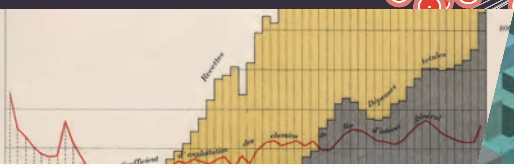
Circulation: *ComputingEdge* (ISSN 2469-7087) is published monthly by the IEEE Computer Society. IEEE Headquarters, Three Park Avenue, 17th Floor, New York, NY 10016-5997; IEEE Computer Society Publications Office, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720; voice +1 714 821 8380; fax +1 714 821 4010; IEEE Computer Society Headquarters, 2001 L Street NW, Suite 700, Washington, DC 20036.

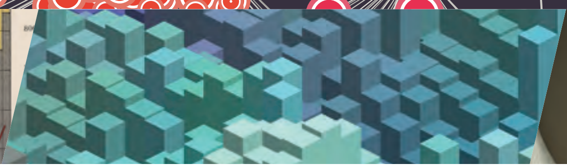
Postmaster: Send address changes to *ComputingEdge*-IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Printed in USA.

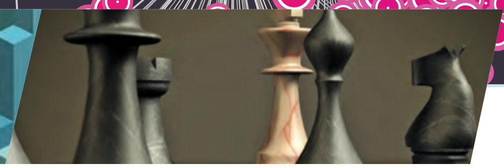
NOVEMBER 2025 · VOLUME 11 · NUMBER 11

COMPUTING edge









18
Business Data Visualization, Beyond the Boring

34
The Compute Continuum: Trends and Challenges

42
Generative AI and Health Care: Brief Survey

Visualization

8 **Preparedness for Visualization in the Next Pandemic**
FINTAN MCGEE, MUNA ABU SIN, MIN CHEN, DAVID EBERT, KAZUO MISUE, PANAGIOTIS D. RITSOS, AND ANTJE WULFF

18 **Business Data Visualization, Beyond the Boring**
ROBERT KOSARA

Distributed Computing

24 **Distributed Federated Deep Learning in Clustered Internet of Things Wireless Networks With Data Similarity-Based Client Participation**
EVANGELIA FRAGKOU, ELEFTHERIA CHINI, MARIA PAPADOPOULOU, DIMITRIOS PAPAKOSTAS, DIMITRIOS KATSAROS, AND SCHAHRAM DUSTDAR

34 **The Compute Continuum: Trends and Challenges**
MAURO TORTONESI

Generative AI

38 **Artificial Intelligence Adoption Is Easy for Gigs, Start-Ups and Small Companies—But Not Mid-Sized and Large Enterprises**
STEPHEN J. ANDRIOLE

42 **Generative AI and Health Care: Brief Survey**
ROY RADA

Health

48 **Economics of Agentic AI in the Health-Care Industry**
NIR KSHETRI

54 **Health Care 4.0 and Clinical Internet of Things**
JOANNA F. DeFRANCO

Departments

4 Magazine Roundup
7 Editor's Note: Envisioning New Horizons for Visualization
64 Conference Calendar

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

A Serious Game for Dena'ina Language and Heritage

This article, featured in the September 2025 issue of *Computer*, presents the research, design, and development of Dnigi, a serious game to facilitate learning Alaska Native heritage, focusing on Dena'ina language and history. The authors envision Dnigi as a learning aid, targeting an undergraduate audience as developers and users.

Computing in Science & Engineering

Do Research Software Engineers and Software Engineering Researchers Speak the Same Language?

Anecdotal evidence suggests that research software engineers (RSEs) and software engineering researchers (SERs) often use different terminologies for similar concepts, creating communication challenges. To better understand these divergences, the authors of this April–June 2025 *Computing in Science & Engineering* article investigated how software engineering fundamentals from the SER community are interpreted within the RSE community, identifying aligned concepts, knowledge gaps, and areas for potential adaptation.

IEEE Annals of the History of Computing

Twentieth Century Computer Product Proposals: A Wealth of Information About Information Technologies and Their Uses

This article, featured in the April–June 2025 issue of *IEEE Annals of the History of Computing*, describes the features of formal data processing proposals written in the twentieth century by computer vendors; in this case by IBM. Tens of thousands of such documents were carefully prepared but remain nearly invisible to members of the IT community and to its scholars, as these were considered highly confidential between customers and vendors. The features of this class of document are presented with two case studies.

IEEE Computer Graphics and Applications

MarsIPAN: Optimization and Negotiations in Mars Sample Return Scheduling Coordination

Resource allocation problems touch almost every aspect of modernity. The authors of this July/August 2025 *IEEE Computer Graphics and Applications* article examine communication bandwidth optimization and negotiation in NASA's early stage Mars Sample Return (MSR) mission, which places multiple robots into a single region on Mars. They present a design study conducted over two years at the NASA Jet Propulsion Laboratory with MSR, which characterizes the design and evaluation of the deployed MarsIPAN schedule browser.

IEEE Intelligent Systems

A Drone Early Warning System for Predicting Threatening Trajectories

Over the last few years, there has been increasing use of drones by terror groups and in armed conflict. Several technologies have been developed to detect drone flights. However, much less work has been done on the drone threat prediction problem (DTPP): Predicting which drone trajectories are threatening and which ones are not. The authors of this article, which was in the July/August 2025 issue of *IEEE Intelligent Systems*, propose the drone early warning

 2469-7087/25

system (DEWS), a framework to solve this problem.

IEEE Internet Computing

A Vision for Computational Decarbonization of Societal Infrastructure

Modern society is at a critical inflection point with rapidly accelerating demand for energy due to growth in domestic manufacturing, datacenters, artificial intelligence (AI), electric vehicles, and electric heat pumps. Sustaining this growth while also reducing society's carbon emissions will necessitate a shift beyond the long-standing focus on improving energy efficiency to optimizing carbon efficiency. This article from the March/April 2025 issue of *IEEE Internet Computing* lays out a vision for a new field of computational decarbonization, which focuses on optimizing and reducing the lifecycle carbon emissions of complex computing and societal infrastructure systems.

IEEE micro

Lunar Lake an Intel Mobile Processor: SoC Architecture Overview (2024)

This article, featured in the May/June 2025 issue of *IEEE Micro*, introduces Lunar Lake (LNL), a codename for the Core Ultra Series mobile processors designed by Intel, released in September 2024. LNL enhanced the partition of cores to performance and efficient clusters with the ability to contain software load to the desired hardware at runtime, it optimized performance cluster with single threaded core, revised the idle state management, reduced frequent CPU wakes, enhanced the memory subsystem power states, and added fine-grain power delivery. Overall, LNL's advanced features and robust design make it a versatile solution for next-generation client products, paving the way for future innovations in mobile computing.

IEEE MultiMedia

Bilateral Two-Dimensional Multiview Discriminant Analysis for Image Recognition

Multiview images are now ubiquitous in real-world applications. However, most multiview learning methods fail to exploit the spatial information hidden in them. To handle this, the author of this January–March 2025 *IEEE MultiMedia* article proposes bilateral 2-D multiview discriminant analysis, named B2DMvDA. B2DMvDA extracts features from both sides of the image, which can not only leverage spatial information like 2-D multiview discriminant analysis (2DMvDA) but also handle the general form of multiview data.

IEEE pervasive COMPUTING
MOBILE SYSTEMS | UBIQUITOUS COMPUTING | INTERNET OF THINGS

Immersive Learning: Evaluating Virtual Reality for Geological Compass Education

In the field of Earth sciences, an essential aspect of geology students' education is acquiring the skill of using a geological compass, but classroom training and limited field trips are often insufficient. The immersive and interactive nature of the virtual reality (VR) system provides a valuable resource for students and educators, offering new opportunities for immersive learning experiences. This April–June 2025 *IEEE Pervasive Computing* article presents an evaluation study of an innovative VR training system focused on improving practical skills in geological compass usage among geology students.

IEEE SECURITY & PRIVACY

Digital Forensics and Jurisdictional Challenges for the Industrial Internet of Things

The Industrial Internet of Things (IIoT) generates heterogeneous data, which are difficult to analyze. Legal aspects of permissibly handling IIoT data derived through privacy-constrained data sources also encumber investigation. In this article, featured in the May/June 2025 issue of *IEEE Security & Privacy*, the authors present a lateral analysis of digital forensic processes for IIoT.

IEEE Software

Quantum Approaches for Vehicle Routing Optimization on Noisy Intermediate Scale Quantum Platforms: Applications of QAOA and Quantum Annealing for Vehicle Routing Problems

Transportation significantly impacts urban energy consumption and environmental sustainability, making optimization critical. This article from the September/October 2025 issue of *IEEE Software* discusses how quantum computing provides a promising approach to improving transportation systems by tackling vehicle routing problem variations using the quadratic unconstrained binary optimization model.

IT Professional

Digital Twins in Industry 4.0: A Practitioner's Perspective

The rapid evolution of Industry 4.0 (I4.0) is transforming industrial processes, emphasizing the convergence of digital technologies with manufacturing. Central to this paradigm shift are digital twins, virtual counterparts that mirror real-world entities, processes, and systems. This article, featured in the July/August 2025 issue of *IT Professional*, explores the central role digital twins play in the I4.0 landscape. Building on the experiences the authors developed in the design and implementation of advanced digital twin solutions, the article adopts a practitioner's perspective to discuss practices, lessons learned, and design guidelines that characterize this emerging application field.

Envisioning New Horizons for Visualization

Visualization is one of humanity's oldest tools. It can be traced all the way back to cave paintings created by early humans thousands of years ago. Over the course of human history, visualization has been used in countless ways, from maps and geometry to written language and scientific drawings. Today, it continues to evolve, notably within the field of computer science research. This issue of *ComputingEdge* explores new and changing uses of graphic visualization, including its use in pandemic preparedness and its improvements for business applications. The articles also discuss emerging applications of distributed computing, artificial intelligence (AI), and machine learning (ML). The issue concludes by focusing on how AI is specifically affecting the health-care industry.

Visualization can benefit a variety of users, from researchers preparing for pandemics to business users seeking to comprehend data. In "Preparedness for Visualization in the Next Pandemic" from *IEEE Computer Graphics and Applications*, the authors examine visualization as a vital tool to facilitate a rapid and effective response to future pandemics. *IEEE Computer Graphics and Applications* article "Business Data Visualization, Beyond the Boring" analyzes business users' preferences and how to tailor business visualizations to better suit their wants and needs.

Topical applications of distributed computing include distributed federated deep learning and the compute continuum. The authors of "Distributed Federated Deep Learning in Clustered Internet of Things Wireless Networks With Data Similarity-Based Client Participation" from *IEEE Internet Computing*, present a protocol to perform distributed federated learning, fully addressing the requirements and peculiarities of the wireless ad hoc IoT environment. In "The Compute Continuum: Trends and Challenges" from *Computer*, the author explains the compute continuum, what it offers, and challenges associated with its implementation.

AI is already having a big impact on companies of all sizes as well as the health care industry. In *IT Professional* article "Artificial Intelligence Adoption Is Easy for Gigs, Start-Ups and Small Companies—But Not Mid-Sized and Large Enterprises," the author shows how smaller companies may have a competitive advantage to larger companies when it comes to adopting AI, ML, and generative AI. The author of "Generative AI and Health Care: Brief Survey," from *Computer*, assesses the future of AI in health care, particularly when it comes to advances in natural language processing, radiation oncology, and palliative care.

The last two articles continue the conversation on how modern technologies, such as AI and the Internet of Things, are being incorporated into the framework of health care. In *IT Professional* article "Economics of Agentic AI in the Health-Care Industry," the author illustrates how agentic AI can transform health-care systems by automating workflows, enhancing collaboration, and improving patient outcomes. *Computer* article "Health Care 4.0 and Clinical Internet of Things" highlights the success factors of Health Care 4.0 as well as the benefits and challenges of the Clinical Internet of Things.

EDITORS: Melanie Tory, m.tory@northeastern.edu
Daniel Keefe, dfk@umn.edu

This article originally appeared in
IEEE Computer Graphics and Applications
vol. 45, no. 3, 2025

DEPARTMENT: PEOPLE IN PRACTICE

Preparedness for Visualization in the Next Pandemic

Fintan McGee, *Luxembourg Institute of Science and Technology (LIST), L-4362, Esch-sur-Alzette, Luxembourg*
Muna Abu Sin, *Robert Koch Institute, 13353, Berlin, Germany*
Min Chen, *University of Oxford, OX1 3QG, Oxford, U.K.*
David Ebert, *University of Oklahoma, Norman, OK, 73071, USA*
Kazuo Misue, *University of Tsukuba, Tsukuba, 305-8573, Japan*
Panagiotis D. Ritsos, *Bangor University, LL57 2DG, Bangor, U.K.*
Antje Wulff, *Carl von Ossietzky Universität Oldenburg, 26129, Oldenburg, Germany*

This article discusses considerations on how visualization can be best positioned to help respond to future pandemics. We examine visualization, along with the corresponding and necessary enabling technologies and platforms, as a tool to facilitate a rapid and effective response to a forthcoming pandemic. We consider challenges in terms of an infrastructure supporting world-wide response, corresponding training and stakeholder engagement, integration of future technologies, and appraisal of such systems. Finally, we discuss how addressing these challenges also helps emergency response beyond infectious diseases.

Data visualization became an important element of our response to the COVID-19 pandemic. It was used to inform decision-making from governments and organizations, to frequently inform the public through different media, to assess the impact of implemented policies, and to facilitate epidemiological analysis. Such usage examples highlight the versatility of visualization to be applied in a plethora of applications, as well as indicate its multifaceted importance in responding to future pandemics.

In this article, we discuss how visualization can continue to play such an important role in future pandemic response, both in terms of how it can be applied in our preparation for them and considering how we can utilize it better during such events. We elaborate on six key challenges (see Figure 1) relating to the use of visualization during the COVID-19 pandemic, as identified from published works (e.g., the work of Chen et al.[3] and Dykes et al.[5]) and discussed during the Dagstuhl Seminar 24091. We map these challenges to corresponding actions that we propose to prepare for future pandemics and beyond.

In our analysis, we consider important visualization enablers, identified during the COVID-19 pandemic, such as: a) the technologies needed to support it, b) its importance for modeling and epidemiological analysis, c) insights and awareness from stakeholders that can drive it, and d) how to evaluate its effectiveness. We point to synergies with other domains, such as artificial intelligence (AI), human–computer interaction, and edge and distributed computing. We discuss infrastructure along with any necessary, novel, and existing technologies, methods, expertise, and attitudes needed for making visualization a "first-class citizen" in our future response to pandemics. Finally, we consider how visualization and infrastructure can be used to

Digital Object Identifier 10.1109/MCG.2025.3561619
Date of current version 18 July 2025.

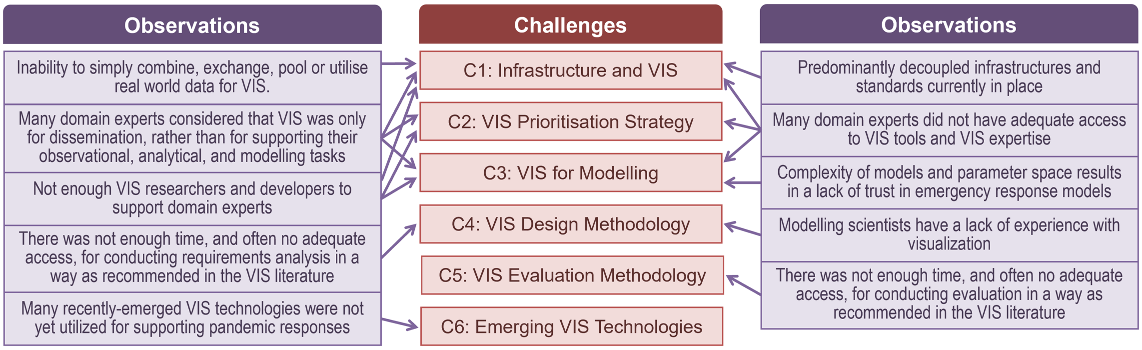


FIGURE 1. Overview of the observed problems and associated challenges.

support other emergency responses and even research in general beyond pandemics.

C1: INFRASTRUCTURE FOR DATA, MODELING, AND VISUALIZATION

As evidenced by the pandemic, addressing global issues necessitates collaboration on a global scale. Interdisciplinary teams must work together to implement data, visualization, and modeling technologies. More than ever, reliable digital infrastructures supporting the timely and accurate dissemination of data, information, and models in appropriate formats to the right users are essential. However, the predominantly decoupled infrastructures, standards, and visualizations are significant challenges. There are separate and nonintegrated solutions available for various fields (e.g., the collection and storage of medical and healthcare-related real-world data, the modeling of epidemiological scenarios, simulation efforts, or the diverse tasks performed by government organizations). Shared, interoperable, and accessible infrastructures for supporting cross-institutional visualization, simulation, prediction, and modeling teams are missing, thus, hindering global collaboration in pandemics.[10] Initially, modeling scientists struggled with a lack of data. Even when available, the vast and diverse array of data sources in the digital healthcare landscape must be effectively managed.[9] The inability to simply combine, exchange, pool, or utilize real-world data is a significant issue. Real-world data remain difficult to integrate due to heterogeneity, nonstandardization, and fragmented systems. Without a common semantic standard, even similar data cannot be easily shared. In addition, up-to-date information is crucial for visualization and modeling, making outdated data less useful. A significant challenge for decision-makers is dealing with data that are frequently potentially out of date, and the absence of data reflecting recent changes.

To address these challenges, overcoming decoupled infrastructures and standards becomes the critical action point. A readiness data platform optimized for pandemic visualizations with integrated visualization capabilities can help. It should be standards-based yet flexible, ensuring interoperability across systems, services, and teams while consolidating diverse data sources efficiently. Given that models generate a considerable amount of data that may not necessarily require persistent storage, it is also necessary to consider flexible approaches for intermediate storage and archiving. The platform must ensure data accuracy and quality, hinged on mutually agreed-upon metadata descriptions. This necessitates agreement on worldwide standards for the representation of the most relevant data valuable for pandemic visualizations. Semantically, rich data models require meticulous requirements engineering. To reach valuable data models, versatile questions must be answered, such as which data sources are relevant at which time, which data formats occur, which visualizations require which data, and which access rules should be followed. For visualization and analysis, another promising avenue lies in an agent-based flexible transformation approach, wherein simple agents dynamically translate data into a unified format. This concept, although in need of further validation, offers potential solutions to the multifaceted challenges of data integration.[3]

The infrastructure, hardware, and software of the platform should have readily deployable visualization capabilities, scaling visualizations expediently and, optimally, incorporating real-time processing features. Embedding of standard visualizations is an integral part. Rapidly created, customizable visualization dashboards are necessary to create effective workflows for different stakeholders.[6] In addition, a stringent data protection regime needs to be integrated, alongside data access controls to provide domain experts with appropriate access to data and visualization tools. The construction of this platform would necessitate intensive interdisciplinary collaboration and human resources. Ongoing research into automated visualization techniques (e.g., ontology and agents), also remains pivotal.[7]

C2: MAKING VISUALIZATION A FIRST CLASS CITIZEN

In a pandemic response, there are many factors to consider with respect to visualization, related to the amount of data, associated sources and users (and their tasks), as well as the modes and types of visualization. Data can be collected from sources (testing centers, hospitals, etc.), for different regions, at different levels (often aggregated), and with data results featuring future predictions and uncertainty. The data themselves are voluminous, dynamic, high-dimensional, and may arrive in numerous streams. The set of users is also large and diverse, for example, healthcare professionals (doctors, nurses, administrators, etc.), epidemiological modeling scientists, machine learning (ML) modeling scientists, governmental decision makers, and the general public, with many subgroups and subroles. There are different modes of visualization, with a wide range of response times, such as routine data monitoring (real time or in a few minutes), supporting decision-making (in tens or hundreds of minutes), in-depth data analysis and data mining (in days), and supporting model-development workflows (in weeks or months). Visualizing data for public communication and engagement can also be considered a separate mode. Finally, related to the modes are the different types, from simple dashboard visualizations to interactive visualization tools for data analysis and model development. Public engagement may require other types, such as storytelling visualizations. These modes and types need to be considered as early as possible in the process to mitigate their impact on any rapid response.

Among the large diverse set of users, many perceive that visualization is only for dissemination, overlooking the benefits of visualization supporting their observational, analytical, and modeling tasks (see also C4). This is also exacerbated by the lack of direct support of visualization scientists and practitioners in the many workflows for pandemic responses.

The complexity of the aforementioned factors can be addressed considerably by having visualization capabilities available at an infrastructural level. At the beginning of the COVID-19 pandemic, to the best of our knowledge, most countries (if not all) encountered challenges in providing visualization facilities due to the lack of adequate infrastructural support (e.g., regional dashboards only became available after many months, and epidemiological modeling workflows did not include adequate support for visualization). Attempts to develop visualization-related infrastructure during the pandemic often encountered many obstacles (e.g., limited availability of experienced developers or readily deployable software solutions).[7] Embedding visualization in the infrastructure is challenging (see also C1).

The two potential approaches are embedding visualization software in data infrastructures and close-coupling visualization infrastructure with data infrastructure. Regardless of which approach is taken, providing adequate and sustainable human resources in the development as well as deployment of visualization-enabled infrastructures is an important point that requires action. The more domain experts can access visualization capabilities from infrastructures and visualization software and the more they can receive direct support from visualization experts embedded in their teams or in collaboration with, the more users will appreciate the uses of observational, analytical, and model-developmental visualization, in addition to disseminative visualization. In this way, visualization will become a “first-class citizen” in data science. Meanwhile, it is necessary to make visualization experts as part of the human infrastructure in preparation for future pandemics (e.g., by attaching visualization practitioners to infrastructures and allowing them to become familiar with rapid methods for requirements analysis and evaluation), while providing support to a wide range of users.

C3: VISUALIZATION FOR MODELING AND ANALYSIS

The COVID-19 pandemic gave rise to an unprecedented level of complex epidemiological modeling activities around the world, while highlighting issues in epidemiological modeling workflows, where visualization can and should play a significant role in supporting modeling scientists (see Figure 2). In addition to huge data volume and complexity (see C1 and C2), there are the challenges of large parameter spaces and gaining trust in the model prediction. The huge heterogeneous dynamic data streams mean that it is

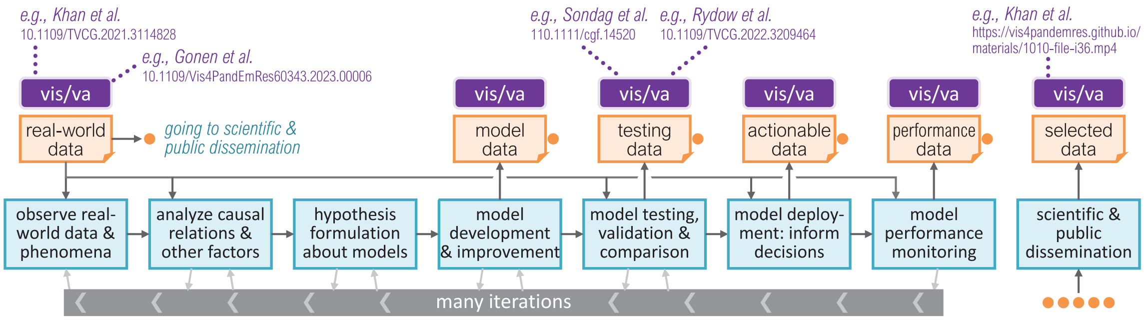


FIGURE 2. Detailed breakdown of the complexity of modeling processes and the associated opportunities for visualization.

time-consuming for scientists to gain useful information from a variety of data to improve a model. Automated or semiautomated visualization can enable scientists to observe data from different data sources, notice unexpected patterns, compare model prediction with newly arrived ground-truth data, and compare the performance of different models (including parameter variations). The process for parameter optimization often features a significant amount of unplanned trial-and-error exploration of a parameter space, while the total number of parameter sets that can be tested is very limited in comparison with the size of the parameter space. Ultimately, it is important that analysts trust the results, especially given the impact of bad outputs informing pandemic response.

Fundamentally, the model space, which includes all possible models due to variation of their structures and parameters, is huge. For a pandemic response, as with most applications, it is intractable to explore all models in the space in order to find the optimal model. Many approaches have been used to reduce the amount of brute-force exploration (e.g., uncertainty analysis for determining if further exploration is beneficial, sensitivity analysis for focusing the exploration on a subset of parameters, ensemble modeling for making use of a group of less-optimized models, and so on). Some (ML) approaches are often referred to as a black-box, as the model internals are not clear. Being able to explain how A leads to B and trusting each step on the way is even more important when an ML model is used. Engendering trust in the model and understanding why the model makes the predictions the way it does are important action points to be taken, especially in cases when the modeling approach is based on AI. Some efforts were made to develop models using ML during the COVID-19 pandemic, and there will no doubt be more such models being developed in the future. Many visualization techniques have been developed to support ML and can be used to further engender explainability and trust in these models, ensuring that they are accepted by their different stakeholders.[2] Visualization can be used to explore model sensitivity to small input changes, and also to optimize parameters.[12] There is also much to be offered in terms of helping users to understand the uncertainty of models. Related to this are ensemble visualization approaches, which convey information about model outputs over a range of parameters or runs.

C4: VISUALIZATION AWARENESS AND STAKEHOLDER DESIGN

Traditional methods of visualization development require close engagement with stakeholders and take time. Public health emergency responses are characterized, among other challenges, by time pressure and uncertainty while policies are implemented in parallel at a population level, with a quickly evolving and changing evidence-base.

It is challenging but necessary to define and agree on short-, medium-, and long-term perspective milestones and outcomes for the approaches chosen. These can be user- and/or purpose-led approaches, such as data analysis, and dashboards for dissemination of data with different levels of complexity/aggregation/stratification. The focus of the approach may be epidemiological, characterized by time, place, and person. As mentioned in other challenges, the turn-around time, amount, and dynamism of data- and evidence-generation during a pandemic differs considerably. In addition, in relation to these aspects, challenges around data validity, confidentially, relevance, and limitations exist and require epidemiological assessment and classification from the expert users.

Flexibility, agility, and continuous adaptation of surveillance, data generating, and monitoring systems are essential. To leverage the full potential of visualization for epidemiological data validation, exploration, analysis, reporting and dissemination, stakeholder engagement, and dedicated time for knowledge exchange to foster reciprocal understanding are a prerequisite but difficult to achieve in emergency situations. While visualization can provide a range of support to epidemiological modeling workflows, often modeling scientists are not accustomed to such support, perceiving visualization as only for dissemination of modeling results[3] (see also C2). During the COVID-19 pandemic, some visualization teams found it difficult to follow recommended practice, such as having frequent meetings with domain experts and conducting field observation. Many design decisions had to be made quickly as part of emergency responses. This challenge can be addressed by building awareness of the capabilities of visualization among domain experts and vice versa (by developing understanding of epidemiological measures and modeling). It is necessary to develop mutual understanding and build trust based on best practice examples in the interpandemic phase and engage in collaborative projects that make the case for visualization and have an impact on the initially agreed and well-defined purpose. This can be done, for example, by providing better documentation of past experience (e.g., through notebooks).[5] An important role is played by collaborative projects where domain and visualization experts codevelop approaches, models, and agile solutions to support epidemiological data analysis and dissemination during interpandemic phases. These are an opportunity to build the base for a more effective uptake and integration of visualization tools during the pandemic response where iterative processes are dominating. Ensuring a suitable composition of teams for this should be a consideration for any project proposal, and possibly even a prerequisite for project calls issued by funding agencies.

In addition, well-categorized and analyzed technical solutions in the visualization literature can be quickly adopted and adapted. For example, dashboards can be further enhanced for fast deployment[6] and easy customization, which must be clearly communicated for each use case (see C1 and C2).

Within the visualization community, this challenge can also be addressed by advancing theoretical understanding and developing requirements analysis methodologies that enable visualization researchers to identify requirements rapidly (e.g., the use of four-level visualization tasks in organizing visualization teams).[3,5] To encourage such progress, the visualization community should encourage research into alternative methods, while avoiding an assumption that the current status-quo for requirements analysis is the only acceptable approach.

C5: EVALUATION

In terms of building systems as we prepare for the next pandemic, current and emerging work on evaluating visualizations is, and will remain, immensely useful. However, the goal here is to develop visualization and demonstrate effectiveness for the user's task, rather than a rigorous evaluation to advance visualization research. It is necessary to develop new evaluation strategies that facilitate and do not hinder rapid responses during a pandemic. For example, during the COVID-19 pandemic, a key difficulty in evaluating visualizations was receiving unbiased feedback that reflects on the quality of visualization approaches and not other issues with the overall infrastructure that supports said visualizations. The close proximity of the visualization experts to end users may result in personal dynamic clouding an appraisal of approaches, even if the visualization is in some way successful. In addition, the domain experts may not be aware of alternative techniques,[5] and they may accept flaws without being aware of visualization best practices.

Furthermore, in terms of rapid response, traditional comparative evaluation of visualization using a null hypothesis approach is too time-consuming and low level. Crowds-sourced evaluations may be considered to yield fast results, yet the domain experts' insight must also be taken into account. This may be remedied by having a combination of evaluation strategies for experts and nonexperts. Nevertheless, the complexity, specificity, and intricacies of visualization may still require a level of visualization expertise from the evaluation participants. Pandemic response visualization deployment may be considered to have a lot in common with a design study approach to visualization (such as that described by Sedlmair et al.[13]), albeit with a very limited precondition phase. Reflection time is very limited, and writing time is even more so. Waiting until there is time to do serious reflection, as well as collection and analysis of evaluation data, unfortunately means that poor performance may have been part of an emergency response tool for quite some time.

Indeed, the agile improvement process during a pandemic would benefit from rapid feedback and reflection. Therefore, the development of new (and adaption of existing) evaluation methods is an action point. For example, with an instance-based evaluation method, as soon as an issue is identified in a rapid evaluation process, instead of waiting for statistically significant evaluation data, visualization experts can analyze the issue (symptom), determine the likely causes, identify possible remedies, and consider adverse side-effects[4] before selecting a solution. This would rely on visualization experts' broad knowledge about problem–solution space, and their ability to address an issue without introducing new problems by anticipating potential side-effects and applying best practices. Likewise, standardized questionnaires and quick feedback mechanisms, potentially integrated with the infrastructure itself, would be suitable for time constrained scenarios and may help rapid evaluation. Therefore, to prepare for the next pandemic, it is desirable for the visualization community to improve, collectively, its theoretical understanding, technical knowledge, and practical design skills in evaluation in a rapid response mode. This includes exploring and researching alternative methods for evaluation. Researchers in the next pandemic should be able to leverage existing theory and reduce time spent obtaining and responding to feedback.

C6: EMERGING TECHNOLOGIES

An important consideration in any future infrastructure is the integration of emerging and novel technologies, as these become available and pervasive. For example, the use of AI in modeling the COVID-19 pandemic (see C3). Similarly, wastewater monitoring had been used to monitor viral load. To facilitate continuous monitoring and rapid information dissemination, other tools were deployed via smartphone application, for reporting symptoms, or tracking and indicating exposure to known infection nodes. In this regard, any infrastructure that will be put in place for preparation and response to future pandemics should be able to accommodate the integration of emerging and future technologies into visualization solutions. For example, technologies from the domains of edge computing and distributed computing can enable such an infrastructure to combine high resolution, frequently updated localized information with coarser large-scale information as an epidemic or pandemic evolves. This can provide new opportunities for targeted fast interventions capitalizing on localized infrastructure and issues. Developments in ubiquitous visualization systems[8] indicate a future where visual analysis and data sense-making will take place away from our desktops in different locations, settings, and scales, e.g., in regional healthcare, hospitals, and urban and rural areas, as well as in our personal environments. Coupled with more capable data collection through smartphones, edge and distributed commuting nodes, ubiquitous computing infrastructures, and AI-enabled data processing, we will be able to provide faster public information and interventions. The quarantine restrictions of a pandemic also motivate the development of distributed command and control centers, where users each have their own remote set of displays, but integrated as part of a unified system.

In addition, emerging tech has been used in new ways, such as for training and emergency response, Existing approaches for first response training in augmented reality, such as visualizing of radiation threats in an environment,[11] may be adapted for a pandemic response training use case, to help understand risk of airborne transmission. Moreover, as such beyond-the-desktop visualization systems become easier to integrate with the current mainly web-driven data and visualization infrastructure, the opportunity for tighter integration with the infrastructures described in this work becomes significantly more viable. Automatic visualization has been a topic of interest for some time, and in 2005, Brodlie et al.[1] outlined the need for a visualization infrastructure with many autonomous facilities. However, currently the technical advancement in automatic visualization has not been fast enough to meet the aforementioned visualization needs at the scale encountered in the COVID-19 pandemic. However, the synergies between ML and visualization mean visualization should go hand in hand with any emerging AI- or ML-based approaches as part of any pandemic response. Visualization for explainable and trustworthy ML, is an important contemporary research topic[2] in the field of information visualization (see also C3). We anticipate that the synergies between visualization and AI will be instrumental in the development of future infrastructures. In visualization research leveraging emerging technologies, pandemic response should be considered a use case to allow flexibility in the choice of technologies used to address the next pandemic.

Finally, integrating emerging technologies in infrastructures, such as the ones described in C1, requires the development of standards that facilitate interoperability. In some themes, such as for ubiquitous visualization aforementioned, this is already a key motivational driver for new systems.[8] The same approach should be adopted for further emerging technology innovations.

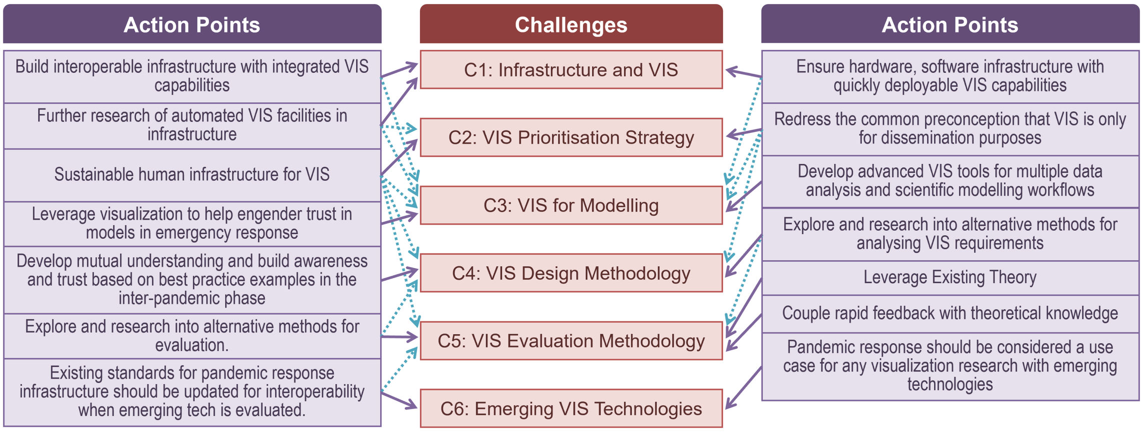


FIGURE 3. Action points and associated primary (solid arrows) and related (dotted arrows) challenges.

CONCLUSION

Addressing each of the challenges described will enable data visualization to be at its most effective in the next pandemic. We summarize action points for each challenge in Figure 3. These are not simple problems and are often major challenges outside the context of the pandemic. The lessons learned from the COVID-19 pandemic can prepare us for other emergencies, beyond those related to infectious diseases, as they often exhibit characteristics similar to pandemics, including suddenness and unpredictability, social disruption, and increasing demand for medical care. Some ways in addressing the challenges helps with general emergency response are as follows.

1) *Risk management and response planning:* Visualization for modeling and analysis to support (C3), and highly reliable digital infrastructure, for data and visualization (C1 and C2).
2) *Improve information sharing and communication:* Establish an infrastructure framework to improve interoperability and accessibility (C1), and take time to promote mutual understanding among stakeholders (C4).
3) *Strengthen medical and public health systems:* Establish reliable digital infrastructure (C1), develop integration/linkage between visualization and digital infrastructure (C2), and continue to ensure close coordination among stakeholders (C4), which can be done only in the interpandemic phase.

Rapid and highly specific, but extremely innovative, approaches can be transformed into sustainable solutions, which can be used beyond the context of pandemics. In Germany, the COVID-19 Data Exchange Platform[a] was initially established as a secure, expandable, and interoperable data platform for the provision of research data on COVID-19 that connects university hospitals nationwide. Although originally designed for the management of data related to the novel coronavirus, the associated frameworks, processes, and infrastructures are not restricted to this domain. It was recognized that the structures can also be operated and expanded beyond the use case in the sense of a generic Research Data Platform for the Network University Medicine[b] or the German Portal for Medical Research Data.[c] This is an example of how technologies can be kept alive, generating added value out of the emergency situation also for other domains and increasing the state of preparedness.

ACKNOWLEDGMENT

The work of Min Chen and Panagiotis D. Ritsos was supported by the UKRI/EPSRC project under Grant EP/V054236/1.

The authors would like to thank Schloss Dagstuhl and the participants of seminar #24091 "Reflections on Pandemic Visualization," which gathered researchers from the visualization community and domain experts to reflect on research work done in response to the COVID-19 pandemic.

[a][Online]. Available: https://num-codex.de/home
[b][Online]. Available: https://www.highmed.org/en/num-rdp
[c][Online]. Available: https://forschen-fuer-gesundheit.de/

REFERENCES

1. K. Brodlie et al., "Visual supercomputing: Technologies, applications and challenges," *Comput. Graphics Forum*, vol. 24, no. 2, pp. 217–245, 2005.
2. A. Chatzimparmpas, K. Kucher, and A. Kerren, "Visualization for trust in machine learning revisited: The state of the field in 2023," *IEEE Comput. Graphics Appl.*, vol. 44, no. 3, pp. 99–113, May/Jun. 2024.
3. M. Chen et al., "RAMPVIS: Answering the challenges of building visualisation capabilities for large-scale emergency responses," *Epidemics*, vol. 39, 2022, Art. no. 100569.
4. M. Chen and D. S. Ebert, "An ontological framework for supporting the design and evaluation of visual analytics systems," *Comput. Graphics Forum*, vol. 38, no. 3, pp. 131–144, 2019.
5. J. Dykes et al., "Visualization for epidemiological modelling: Challenges, solutions, reflections and recommendations," *Philos. Trans. Roy. Soc. A: Mathematical, Phys. Eng. Sci.*, vol. 380, no. 2233, 2022, Art. no. 20210299.
6. S. Khan et al., "Propagating visual designs to numerous plots and dashboards," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 86–95, Jan. 2022.
7. S. Khan, P. H. Nguyen, A. Abdul-Rahman, E. Freeman, C. Turkay, and M. Chen, "Rapid development of a data visualization service in an emergency response," *IEEE Trans. Serv. Comput.*, vol. 15, no. 3, pp. 1251–1264, May/Jun. 2022.
8. M. Marcel Borowski, P. W. S. Butcher, J. B. Kristensen, J. O. Petersen, P. D. Ritsos, and C. N. Klokmose, "DashSpace: A live collaborative platform for immersive and ubiquitous analytics," *IEEE Trans. Vis. Comput. Graphics*, early access, Feb. 03, 2025, doi: 10.1109/TVCG.2025.3537679.
9. F. Martin Sanchez, K. Gray, R. Bellazzi, and G. Lopez-Campos, "Exposome informatics: Considerations for the design of future biomedical research information systems," *J. Amer. Med. Informat. Assoc.*, vol. 21, no. 3, pp. 386–390, 2013.
10. L. Maxwell, P. Shreedhar, D. Dauga, P. McQuilton, R. Terry, and A. Denisiuk, "Fair, ethical, and coordinated data sharing for COVID-19 response: A scoping review and cross-sectional survey of COVID-19 data sharing platforms and registries," *Lancet Digit. Health*, vol. 5, no. 10, pp. e712–e736, 2023.
11. F. McGee, R. McCall, and J. Baixauli, "Comparison of spatial visualization techniques for radiation in augmented reality," in *Proc. ACM Conf. Hum. Factors Comput. Syst.*, 2024, pp. 1–15.
12. M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller, "Visual parameter space analysis: A conceptual framework," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 2161–2170, Dec. 2014.
13. M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012.

FINTAN MCGEE is a lead research and technology scientist at the Luxembourg Institute of Science and Technology, L-4362, Esch-sur-Alzette, Luxembourg. His research interests include visual analysis of complex networks and immersive data visualization in augmented reality. McGee received his Ph.D. degree in computer science from Trinity College, Dublin, Ireland. He is the corresponding author of this article. Contact him at fintan.mcgee@list.lu.

MUNA ABU SIN is currently seconded from the Robert Koch Institute, 13353, Berlin, Germany to the Federal Ministry of Health. She is trained as a medical doctor with a focus on infectious diseases and is passionate about infection prevention and control. She has more than 10 years of experience as a public health professional and epidemiologist in the field of antimicrobial resistance and use surveillance at the national and international levels. Contact her at abu-sinm@rki.de.

MIN CHEN is a professor of scientific visualization at Oxford University, OX1 3QG, Oxford, U.K., and a fellow of Pembroke College, CB2 1RF, Cambridge, U.K. His research interests include many aspects of data science in general, and visualization and visual analytics in particular. Contact him at min.chen@oerc.ox.ac.uk.

DAVID EBERT is an associate VP for research and partnerships, a Gallogly chair professor of electrical and computer engineering, and the director at the Data Institute for Societal Challenges, University of Oklahoma, 73019, Norman, U.K., and an adjunct professor of electrical and computer engineering at Purdue University, West Lafayette, IN, USA. His research interests include visual analytics, human-guided AI, and data analytics. He is a fellow of IEEE. Contact him at ebert@ou.edu.

KAZUO MISUE is a professor at the Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba, 305-8573, Japan. His research interests include information visualization and automatic graph drawing. Misue received the Ph.D. degree in engineering from the University of Tokyo, Tokyo, Japan. He is a member of ACM and IEEE. Contact him at misue@cs.tsukuba.ac.jp.

PANAGIOTIS D. RITSOS is a senior lecturer at the School of Computer Science and Engineering, Bangor University, LL57 2DG, Bangor, U.K. His research interests include mixed reality, information visualization, and visual analytics. Ritsos received his Ph.D. degree in electronic systems engineering from the University of Essex, Colchester, U.K. He is a member of IEEE Computer Society. Contact him at p.ritsos@bangor.ac.uk.

ANTJE WULFF is a professor for Big Data in medicine at the School of Medicine and Health Sciences, Carl von Ossietzky Universität Oldenburg, 26129, Oldenburg, Germany. Her research interests include health data platforms and modeling, interoperability standards, and clinical decision-support systems. Wulff received her Ph.D. degree in computer science from Technical University Braunschweig, Braunschweig, Germany. Contact her at antje.wulff@uol.de.

Contact department editor Melanie Tory at m.tory@northeastern.edu or department editor Daniel Keefe at dfk@umn.edu.

IEEE COMPUTER SOCIETY

www.computer.org

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: Contact ombudsman@computer.org.

CHAPTERS: Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.

PUBLICATIONS AND ACTIVITIES

***Computer*:** The flagship publication of the IEEE Computer Society, *Computer*, publishes peer-reviewed technical content that covers all aspects of computer science, computer engineering, technology, and applications.

Periodicals: The IEEE CS publishes 12 magazines, 18 journals

Conference Proceedings & Books: Conference Publishing Services publishes more than 275 titles every year.

Standards Working Groups: More than 150 groups produce IEEE standards used throughout the world.

Technical Communities: TCs provide professional interaction in more than 30 technical areas and directly influence computer engineering conferences and publications.

Conferences/Education: The IEEE CS holds more than 215 conferences each year and sponsors many educational activities, including computing science accreditation.

Certifications: The IEEE CS offers three software developer credentials.

AVAILABLE INFORMATION

To check membership status, report an address change, or obtain information, contact help@computer.org.

IEEE COMPUTER SOCIETY OFFICES

WASHINGTON, D.C.:
2001 L St., Ste. 700,
Washington, D.C. 20036-4928
Phone: +1 202 371 0101
Fax: +1 202 728 9614
Email: help@computer.org

LOS ALAMITOS:
10662 Los Vaqueros Cir.,
Los Alamitos, CA 90720
Phone: +1 714 821 8380
Email: help@computer.org

IEEE

revised 2 September 2025

EDITOR: Theresa-Marie Rhyne, theresamarierhyne@gmail.com

This article originally appeared in Computer Graphics and Applications vol. 44, no. 5, 2024

DEPARTMENT: VISUALIZATION VIEWPOINTS

Business Data Visualization, Beyond the Boring

Robert Kosara, *Observable, San Francisco, CA, 94114, USA*

Business uses of charts and visualizations, and by extension business users, are usually considered mundane and boring. But they, too, want to get their audience's attention, emphasize a point they are making, or simply break out of the monotony of the limited palette of common chart types. I believe that there is ample opportunity to develop new approaches and build better tools that go far beyond the current one-size-fits-all approach to creating charts—much more than is currently recognized in the visualization community. The first step is to reexamine our notions of who business users are, and what they actually want and need.

Data visualization is a visual medium, and yet the output of most visualization and business intelligence (BI) tools looks largely the same. Bar charts, line charts, and all sorts of variations of them, plus treemaps, scatterplots, and—grudgingly—pie and donut charts. It does not matter if the data are about money or people, whether the numbers are big or small, they all get squeezed into charts that are largely interchangeable were it not for some axis labels.

This visual monotony is a real problem when it comes to the common business use case of presenting numbers. How do you make something stand out? How is the audience expected to remember a particular chart if they all look virtually the same? Why is BI and charting software so limited in the palette of visualizations it can produce?

I believe that this is driven by ideas about who the users of visualization are, and what they want, that are largely inaccurate and based on certain preconceived notions of "expert" users. There is also an unfortunate reinforcement loop here where most users only have access to tools that provide them with a relatively small set of inflexible options, which is then assumed to be all they want and need.

0272-1716 © 2024 IEEE
Digital Object Identifier 10.1109/MCG.2024.3407489
Date of current version 25 October 2024.

This is really unfortunate, because it undersells the variety of interesting use cases that could inspire new visualization research, and at the same time keeps any advances in research from being able to improve the choices real users have.

In this article, I present a different perspective—my *viewpoint*, if you will—that will hopefully demonstrate some of the interesting questions that could be asked, and interesting research that could be performed, that would at the same time help a large number of real users with their tasks every single day.

THE SME MYTH

From the perspective of visualization researchers, it is not unreasonable to expect that visualizations will be the focus of people's work. We also like to think of users as some sort of scientists or subject-matter experts (SME), who spend the majority of their days in intense and well-defined data- and visualization-focused work.

Those users exist, but the majority of people out there using Excel, Tableau, PowerBI, etc., use data and visualization as a means to an end. Instead of being experts in a narrowly focused area, they work across many different areas depending on what their current task demands. Being what we might call *business domain experts* (BDEs) requires a broad set of skills and knowledge about a variety of different domains.

These users want to accomplish a task: track performance, brief a higher up, help make a decision, raise awareness, etc. This is an incredibly common use case for charts and visualizations. These users sometimes make those charts themselves, or they have them built for them by their IT departments, BI analysts, etc.

What do these people need? How much does the research world know about the chart creators, their end users, and the many people in between? This is not a small user base, quite the opposite. Tens of millions of people (possibly even an order of magnitude more) use charts like this on a regular basis.

This leads to practices that are often considered less than ideal by researchers, but that work in the right contexts. Perhaps adding clip art or stock photos to charts helps break up the monotony of a chart-heavy presentation and draws attention to important facts, perhaps using a pie chart is perfectly reasonable when there are only a few values to be drawn and they are part of a whole, perhaps a few large numbers are more informative than a complex chart when their exact values are what matters.

I have seen many such dashboards and presentations, and spoken to many such users, but their existence does not seem to have been noticed by many visualization researchers (with a few exceptions[5]). The questions they ask are often not answered by existing research: how many pie slices is too many? How do I make better tables? When is an image on a dashboard okay? What chart types can I use to break up a wall of bar charts?

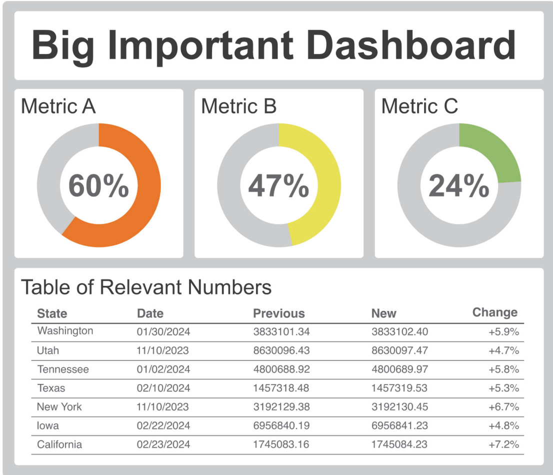


FIGURE 1. This mockup of a dashboard shows some of the elements commonly found on real dashboards: large numbers, donut charts, and a table of dates and numbers.

NUMBERS ARE IMPORTANT

In the world of visualization research, visual representations tend to replace numbers, not complement them. This is understandable, since we usually deal with too many numbers to represent, and care more about finding patterns of interest rather than specific numerical values. Researchers also often want to come up with techniques that work in many contexts, instead of being limited to one very particular use case, so the importance of particular values of those numbers is usually lost on us.

The user perspective is very different, however. Many business users (BDEs) are intimately familiar with the numbers they need to do their jobs, and have to see them. One reason is that they might have specific numeric targets they want, or have to hit. Another is that they expect numbers to change in a certain way, and want to see how far off the actual values might be. Finally, users want to feel confident that the visualization is showing them correct, up to date data.

When we build visualization tools to help people understand data, we implicitly assume that they are showing the data correctly and provide more value than the raw numbers. The people who are supposed to use these tools have to be able to trust them, but on what basis? A certain level of skepticism is healthy, and trust needs to be earned. This is an even bigger issue when your job and potentially lives depend on them being correct.

I know of a study, which was unfortunately never completed or published, where two engineers were trying to make the numbers they were seeing in an early version of Tableau match the ones they were used to from their existing tools. Even though they spent two hours working, tweaking their settings, trying different calculations, double-checking everything within the tool, etc., they never managed to make the numbers match. Understandably, they were skeptical of a tool that might be doing something wrong, or their abilities to use such a tool (it later turned out that they had been looking at an older snapshot of their data).

There is a reason so many dashboards show a few large numbers across the top (Figure 1): these numbers are important and need to be seen right away. If they are different from what is expected, immediate action might be necessary. If they are as expected however, the viewer can often move on to other tasks instead of having to spend any more time parsing the details of the dashboard.

In addition to large numbers, users also often ask for tables. BI tools can create tables, but they are usually an afterthought and provide little to no real functionality. While they can usually be sorted by different columns and in different directions, any more advanced kinds of functionality are usually missing:

can numbers be copied from the table (which is a common task for reporting)? Can a user perform simple calculations (like they are used to from Excel), perhaps even with numbers from different tables on the same dashboard? Can the table be grouped by a categorical column and calculations be performed on each partition? Can numbers be turned into little charts?

Imagine a dashboard showing sales for different branches of a corporation as absolute numbers, and using those for ranking. Would it not be reasonable to normalize these numbers by branch size? In a case I know of (which does not appear to have been published), the business user being interviewed had to copy the numbers into a spreadsheet, by hand, to perform this simple calculation—and do it again every month.

All this talk of numbers may seem out of place in an article about visualization, but to many users, visuals without numbers are meaningless. Visualization should integrate with numbers, not replace them. It should provide richer, more powerful ways of working with data, whether it be numbers or visuals. The challenge here is to extend visualization research to include numbers when they are meaningful to users, not just purely visual representations. The importance of tables has been recognized in research,[1] but there is still very little work addressing tables as first-class citizens of dashboards.

In addition to solving real users' problems, there is also a chance to reimagine the spreadsheet here: how can we make it easy to transform numbers, whether into other numbers or into visuals, in a unified way? Perhaps tables would not be so popular if we could treat the elements of charts as numbers and perform operations on them, like compute differences, turn them into percentages, etc.? We consider visualizations just another representation of numbers, so why are they often such a dead end?

VISUAL PREFERENCES

The pie chart has been widely used for many years, and perhaps for as many there has been criticism. In 1923, for example, Karsten wrote that, "[it] is worthless for study and research purposes." And yet, he spent seven pages explaining pie charts in his book and recommending them for uses that demand readers' attention. After all, "[a] circle is, of all geometrical patterns, the easiest resting spot for the eye."[7]

Even though treemaps came from the research community,[9] their use in BI tools today is usually quite different from how they were initially conceived: they rarely show hierarchy, and instead work as single-level part-to-whole charts for many values (though sometimes for even just a few). They are used as an alternative to bar and pie charts, presumably because they are more attractive than the former and less controversial than the latter.

In both cases, however, I believe that the reason they are used is one of visual preference. As Karsten observed 100 years ago, circles are attractive shapes, and while treemaps are made up of rectangles, they are more visually appealing than the jagged cliffs of bar charts. Many dashboards also include donut charts, often with a large number in their center (see Figure 1). They are used to break up the monotony of a dashboard, and provide a point of focus, on what would otherwise be a wall of largely identical-looking charts.

Similarly, visual embellishments on charts are usually discouraged, but what if instead they could be not only supported by tools, but even recommended? The use of semantically relevant colors has been shown to help people read charts faster and retain more information from them.[8] What could tools be doing, automatically, to find relevant color combinations for categorical data? What other visual channels could benefit from similar treatment?

Going further, the addition of images to charts is often derided as "chart junk," but can help with memory as well.[2] Providing the tools to incorporate these elements into charts and visualizations in a meaningful and subtle way would support people who need to present information to audiences and who currently struggle to do so.

Visual preferences are not usually studied in research papers when they compete with other reasons for choosing a chart type. But in practice, these decisions and tradeoffs are quite common. Just like users want or need to create interest in presentations, they do the same for dashboards and other displays of data. Research should take this into account and come up with means of incorporating visual interest and preferences into chart recommendations and formalisms.

VISUALLY RICH CHARTS

It may seem like a given that charts today have a certain minimal look to them, but that is also a large part of the problem of monotony described above. There is a stark contrast between the monotonous, anemic-looking charts BI and charting tools tend to produce and the rich and varied charts created by data journalists, or the beautiful charts often created by hand in the past (see Figure 2).

Why can software running on powerful computers not create charts that are at least as visually rich and interesting (or even more so!) as what people were able to draw by hand hundreds of years ago? This would even solve some of the common problems with popular chart designs that tend to draw criticism from visualization experts.

Take the dual-axis chart in Figure 2, for example. Note the grid inside the bars, which is only drawn where it is necessary. A common issue with grids is that they are too visually prominent, and while this design is still somewhat busy, it is a big step up from a grid filling the entire chart. The bars here are not stacked, but the yellow revenue bars sit behind the gray expenses bars. This allows direct comparison between them, and of course stacking expenses and revenue would make little sense. The choice of bars here is helpful though, to separate the two absolute data series from the red line showing the fraction of expenses to revenue. Note again the grid under the line being drawn only where necessary, and the way the little gap on the right separates the line from the axis that does not apply to it.

This would not be a terribly useful design if all charts in a presentation looked like it, but it does show the kind of care and thought that is so rare in mass-produced charts coming out of our tools today. The specific aesthetic choices here are also less important than the overall point, that this kind of more bespoke chart has all but disappeared from software-created charts.

But why? Why can software not offer more choices and customizations, beyond the garish and often useless embellishments like shadows, 3-D, etc.? Why can modern charts not look rich and exciting, and each have their own flavor, when they are created by software? Should software not allow for much more data-specific choices and variety than the manual tools of the early 1900s?

This might be a worthy use case for using the AI tools that are currently *en vogue* for all sorts of uses, to provide recommendations based on context and a library of interesting charts. Finding rich (and plentiful) enough training data would probably be a challenge, but this could be an interesting approach. Another would be to look for inspiration from charts like the one I discussed above, and add new degrees of freedom to charting and visualization tools. The simple act of hiding the grid where it is not necessary makes a large difference. There must be many more design touches that are simple and effective, but also just not available in tools today.

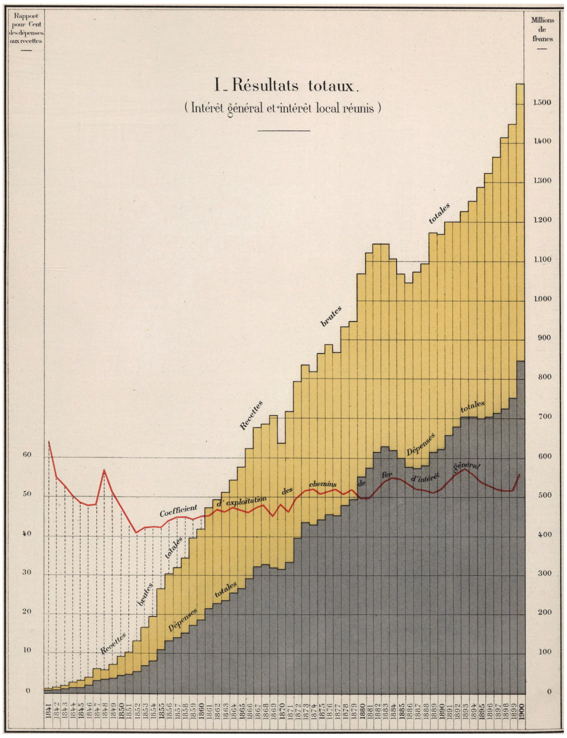


FIGURE 2. Dual-axis chart with data-dependent grid. The red line shows what percentage the gray bars (expenses) are of the yellow bars (revenue). Bars are not stacked, yellow bars are behind and always taller. Note the grid inside the bars is drawn only where necessary, and under the red line to connect it to the scale on the left. From Ministère des Travaux Publics, *Album de Statistique Graphique de 1900*, 1906. Image credit: David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries.

FROM LINTING TO RECOMMENDING

Many people who use charts see the kind of visualizations created by news outlets like the *New York Times* or the *Washington Post*, and want to be able to build their own that have similar impact. In presentations, they fight for their audience's attention, and they want to be able to emphasize important points. Or, in the words of one of the participants in our study,[3] they want to "show ordinary data in extraordinary ways."

One hurdle is that BI tools and chart makers are geared toward generally applicable charts. Bar charts, line charts, even treemaps are widely usable and useful. But that leads to monotony when most of the charts in a presentation end up being the same type. Even varying the color scheme or adding call-outs does not have the same impact as a more bespoke approach.

Research has recently developed the notion of visualization linting,[4] based on the *linting* tools provided in code editors and development environments. Linters apply rules to look for common mistakes, such as uninitialized variables, and flag them to the user.

Similarly, visualization linting is based on the assumption that the software has to watch for, and correct, the user's mistakes. Perhaps instead, or in addition, it could try to help users find new and interesting visual representations. This would help surface options for showing data that the user was not aware of, or that they would need to spend time trying out to see if they are suitable. Why does software not help with that?

In particular, this would help with techniques that are the opposite of general, and very often do not produce usable charts. An example of this odd class of techniques is the connected scatterplot,[6] which leads to unreadable "hairballs" for most datasets, but can be highly effective and interesting when it does produce a legible chart. Currently, the only way to find out is to create such a chart (which is not even possible in most chart chooser style applications). It is not difficult to imagine the development of metrics that would allow an application to recommend trying the connected scatterplot if the number of line crossings is below a certain threshold, however, and perhaps even offer it as a suggestion only in such a case. Similarly, unit charts and other unusual charts might be recommended when the data to be shown appears to result in a usable chart based on some metrics.

The pie chart is another example of this. It clearly cannot work in the general case, what if you end up with thousands of categories in your data? But that is not an unknowable fact. We can query databases about the number of categories. We can count the unique values in an array and then use that knowledge to weigh the suitability of each chart type. We can set a threshold for when to recommend the pie chart at all.

All the above depend not only on the dataset's overall properties, but the actual distribution and shape of the values. Research tends to favor methods that are universal and do not depend on the values as such (though even the bar chart suffers when the data are "noisy"[10]). To enable users to enjoy a broader range of chart types, software needs to take the specifics of a dataset into account so it can offer more unusual chart types that do not usually (or rarely ever) produce good charts.

CONCLUSION

The charts and visualizations used by millions of people every day are strangely impoverished and limited. Computer graphics has managed to create visuals that are virtually indistinguishable from reality, even when being generated in real time, yet visualizations are largely relegated to the most basic shapes and bare-bones customizations.

The users of data visualization deserve better. I hope to have pointed to some interesting use cases and needs in business contexts. Perhaps readers will agree that they are much more interesting than is often assumed, and open up many exciting research questions and directions. Research could have a huge impact by developing new ways of creating visualizations, and perhaps launching a new generation of data visualization tools. That ultimately would help millions of business users understand, present, and act on the numbers they depend on every day.

ACKNOWLEDGMENTS

The author would like to thank R. J. Andrews for his tireless research to surface the amazing charting work done in the past, and for pointing him to the example image used in this piece.

REFERENCES

1. L. Bartram, M. Correll, and M. Tory, "Untidy data: The unreasonable effectiveness of tables," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 686–696, Jan. 2022.
2. S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks, "Useful junk? The effects of visual embellishment on comprehension and memorability of charts," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2010, pp. 2573–2582.
3. M. Brehmer and R. Kosara, "From jam session to recital: Synchronous communication and collaboration around data in organizations," *Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 1139–1149, Jan. 2022.
4. Q. Chen, F. Sun, X. Xu, Z. Chen, J. Wang, and N. Cao, "Vizlinter: A linter and fixer framework for data visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 1, pp. 206–216, Jan. 2022.
5. E. Dimara, H. Zhang, M. Tory, and S. Franconeri, "The unmet data visualization needs of decision makers within organizations," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 12, pp. 4101–4112, Dec. 2022.

6. S. Haroz, R. Kosara, and S. L. Franconeri, "The connected scatterplot for presenting paired time series," *Trans. Vis. Comput. Graphics*, vol. 22, no. 9, pp. 2174–2186, Sep. 2016.
7. K. G. Karsten, *Charts and Graphs*, New York, NY, USA: Prentice-Hall, 1923.
8. V. Setlur and M. C. Stone, "A linguistic approach to categorical color assignment for data visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 698–707, Jan. 2016.
9. B. Shneiderman, "Tree visualization with tree-maps: 2-D space-filling approach," *ACM Trans. Graphics*, vol. 11, no. 1, pp. 92–99, 1992.
10. J. Talbot, V. Setlur, and A. Anand, "Four experiments on the perception of bar charts," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 12, pp. 2152–2160, Dec. 2014.

ROBERT KOSARA is senior data visualization developer at Observable, San Francisco, CA, 94114, USA. His research interests revolve around the uses of visualization for communication: how do we read charts, how do we use unusual chart types to capture attention and help a message stick in the viewer's memory, how do presentation and storytelling with data work in different contexts, etc. Contact him at rkosara@me.com.

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

EDITOR: Schahram Dustdar, dustdar@dsg.tuwien.ac.at

DEPARTMENT: INTERNET OF THINGS, PEOPLE, AND PROCESSES

This article originally appeared in IEEE Internet Computing vol. 28, no. 6, 2024

Distributed Federated Deep Learning in Clustered Internet of Things Wireless Networks With Data Similarity-Based Client Participation

Evangelia Fragkou, Eleftheria Chini, Maria Papadopoulou, Dimitrios Papakostas, and Dimitrios Katsaros, *University of Thessaly, 38334, Volvos, Greece*

Schahram Dustdar, *TU Wien, 1040, Vienna, Austria*

Federated deep learning is the method of choice for performing deep learning in environments where data sharing is not allowed due to privacy/security issues. However, all of the solutions based exclusively or substantially on the existence of a coordinating server are not a fit for wireless Internet of Things environments operating as ad hoc networks due to the excessive communication overhead. This article develops the DISCREETER distributed protocol to perform distributed federated learning, fully addressing the requirements and peculiarities of the wireless ad hoc IoT environment. The algorithm is based on a hierarchical organization of the underlying network along with a generic principled way to select the clients that will participate in the federation based on the similarity of the data they produce. We evaluate the proposed methodology on real data, targeting a time series prediction task, using a recurrent neural network in each node as the learning entity. The obtained results attest to the design merits of our approach.

Federated (deep) learning (F(D)L)[7] emerged to cope with data privacy/security issues that occurred in the traditional cloud-based (deep) learning. In the basic form of FDL, a set of nodes each performs rounds of local training of their deep learning model and then sends the resulting weights, i.e., their model, to a server that performs aggregation and sends back the averaged weights; these weights are then adopted by the nodes, which iterate the same procedure again until convergence.

Still, FDL falls short for servicing a lot of modern applications running over the Internet of Things (IoT).[6] Such applications are based on a large number of distributed and ad hoc interconnected devices, which generate private data that cannot be sent wirelessly over the network to the server due to privacy and network capacity issues. Moreover, their lack of resources does not allow them to perform massive wireless transmission of their local trained models for many obvious reasons—e.g., nodes residing on the paths leading to the server will rapidly deplete their energy and die out due to the excessive communication. To be adjusted to wireless ad hoc network environments, hybrid FDL schemes with device-to-device communication and a coordinating server have been proposed,[4] as have schemes based on only device-to-device communication.[11]

However, neither of the proposed approaches, as explained in the "Related Work" section, can cope with the grand challenge of developing FDL schemes for ad hoc IoT networks, namely, the purely distributed

Digital Object Identifier 10.1109/MIC.2024.3439068
Date of current version 24 January 2025.

nature of the architecture and communication efficiency to alleviate the *broadcast storm problem*[15] in resource-starved environments. Even some proposals that assumed network backbone construction based on clustering,[9] which is a classic technique to cope with broadcast storms, fail to fully address the FDL requirements in a wireless ad hoc environment.

In particular, a communication-efficient FDL scheme for ad hoc IoT networks must do the following:

- It requires a radically different approach than all previously proposed ad hoc network clustering schemes, which are based on node IDs. This is because it is very common in these protocols for the resulting cluster heads (CHs)—i.e., the nodes that will eventually perform the local averaging in the FDL scheme—to be each other's one-hop neighbors (contrast this with Figure 1 solution). This will negatively impact the communication efficiency of the scheme due to the interference since all of the intracluster communication will be carried out by the CH. Moreover, almost all ad hoc clustering algorithms produce clusters with very small cardinality; thus, local averaging/consensus is not very efficient with respect to the communication reduction.
- It should take into consideration that communication range restrictions (a range of a few hundred meters) result in the wireless nodes being frequently in very close proximity, thus producing very similar data, e.g., sensor measuring temperature.

Therefore, based on these challenges/opportunities, we contribute a purely distributed FDL scheme called DISCREETER:

- We develop an ad hoc network clustering scheme appropriate for performing local and global aggregations that reduces wireless interference.
- We develop a client participation scheme into the federation that is based on data similarity to reduce communication during local aggregation and reduce the overall data transmitted over the network in a principled way.
- We evaluate DISCREETER on convergence speed and communication reduction, proving its virtues.

RELATED WORK

The introduction of FL[7] was a major step in addressing privacy concerns in machine/deep learning. A plethora of variations and advances to the initial model were proposed afterward,[6] but since these models assume

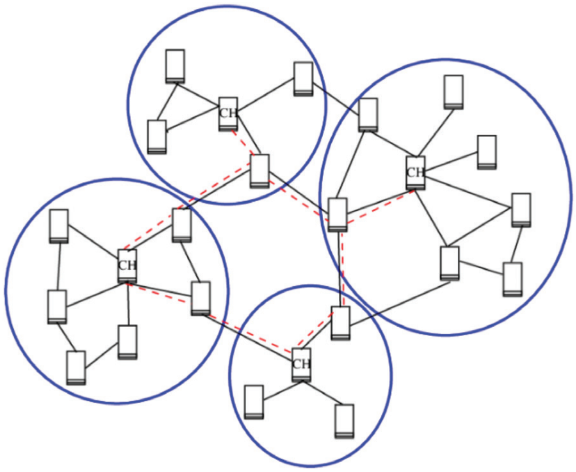


FIGURE 1. A clustered ad hoc network with CHs elected at the center of clusters. Black lines indicate wireless links. Dashed red lines indicate communication paths between CHs.

the existence of a central coordinating server or hybrids[4]—and therefore are not the focus of the article—they will not be surveyed.

The work of Wang et al.[15] has considered the purely distributed nature of wireless ad hoc networks, but it limits itself to determining which wireless links to activate to avoid interference. Similarly, a purely distributed model with only device-to-device communication has been described,[9,11] and other efforts can be found in work by Beltran et al.,[3] but these works produce excessive communication since they assume an all-to-all communication pattern. The work most close to ours is reported by Taya et al.,[13] but their method works in a function space, not in the weights/gradients space.

The literature on client participation in FL is quite rich: it has been addressed as a secretary selection problem,[8] as a sampling scheme,[5] and so on.[10]

We depart from all of the aforementioned works in that we develop an FL scheme for wireless ad hoc IoT environments 1) assuming a purely distributed scheme without a "central" coordinating server, thus avoiding excessive communication and fast energy depletion of nodes; 2) avoiding a completely flat wireless network, thus preventing an all-to-all communication pattern; and 3) deciding the client participation in the federation based on data similarity with its neighboring clients rather than adopting *data-agnostic* methodologies.

PROBLEM FORMULATION

We assume the existence of a wireless ad hoc IoT network comprising n nodes/devices connected to each other through wireless bidirectional links whose topology is described by a graph. Each node *i* generates a dataset D_i with $|D_i|$ data points. Each point (*x*, *y*) consists of a multidimensional feature vector $x \in R^m$ and a label $y \in R$. We let *f*(*x*, *y*, *w*) denote the loss associated with the point (*x*, *y*) based on learning model parameter vector $w \in R^m$. The local loss function $F_i(w)$ at node *i* is defined as

$$F_i(w) = \frac{1}{D_i} \sum_{(x,y) \in D_i} f(x, y, w).$$

A completely analogous loss function $F_C(w)$ is defined for each cluster. The total loss function is then defined as the average loss across the set of clusters:

$$F(w) = \frac{1}{\text{number of clusters}} \sum_{\forall C} F_C(w).$$

The goal of our FL model training is to find the optimal model parameters w^* for *F*:

$$w^* = \text{argmin}_{w \in R^m} F(w).$$

DATA SIMILARITY-BASED HIERARCHICAL DISTRIBUTED FL (DISCREETER)

The proposed FL method is purely distributed and strives for communication efficiency by leveraging two ideas:

- Imposition of a hierarchy over the ad hoc network so that the weight exchange is localized in neighborhoods (and not across the whole network). Moreover, this local communication is performed in a way that avoids (most of) the interference with nearby neighborhoods.
- Selection of the nodes that will participate in the federation based on how similar the data they generate are. Thus, handling of nonindependently identically distributed (non-IID) and sparse data becomes feasible in a natural way.

These two ideas define the two major components of the proposed methodology and are explained in the next paragraphs.

Creation of a Hierarchical IoT Topology

The world of ad hoc networking has developed many protocols for creating a one-level hierarchy over a flat ad hoc network; these are termed as backbone or clustering protocols. However, they fall short for the case of FL over ad hoc networks because 1) they are based on node IDs and, thus, make bad CH selections by selecting CHs not based on the topological position of the CH and 2) create a lot of small (in cardinality) clusters. In FL, we would need CHs to be in a proper position so as to communicate quickly with every cluster member and also to communicate quickly and with no interference with other CHs.

To address both of these challenges, we adopt and customize to the FL context a method for creating generalized dominating sets that will act as the backbone. We adopt the *max-min* d *cluster formation protocol*,[1] which creates dominating sets with the property that a dominated node (i.e., cluster member) can be at a maximum distance of *d* hops from its dominator (i.e., its CH). This algorithm operates *d* rounds of flood-max, where the largest node IDs travel and "prevail" (i.e., become CHs), and then *d* rounds of flood-min take place, giving the opportunity for some not-so-large node ID to "prevail," thus reducing the size of the clusters created by the flood-max rounds.

The main virtue of this algorithm is that it can significantly reduce the number of clusters for any $d \geq 2$; however, the CHs of nearby clusters can still lie at a one-hop distance from each other, causing severe interference effects. To alleviate this drawback, after the execution of a *convergecast* operation that allows each CH to become aware of its cluster members and of its cluster topology, the CH calculates the betweenness centrality (BC)[2] of its cluster members and of itself and assigns the CH role to the member with the highest BC value or retains this role if no member has a larger BC that itself. In this way, the CH becomes the node that can communicate with its members in the fastest possible way, and, at the same time, it is not located in the network "border." Thus, we avoid interference with nearby clusters and also facilitate intercluster communication for carrying out the global averaging task.

Data Similarity-Based Client Selection in FL

In the "Related Work" section, we briefly mentioned various approaches for selecting the nodes that will participate in the federation; these range from complete random to various "principled" sampling. We depart from all of these approaches and consider the necessity of node participation from the perspective of whether the node produces data distributed "differently" from the data of its nearby nodes. This situation occurs very frequently in scenarios where the nodes are sensors sampling temperature or moisture, capturing photos of the same scene, and so on. In all of these cases, data can be seen as a series of numbers; even a photo can be seen as a series of its pixel intensities. Due to geographical proximity, many sensors may produce similar data.

Thus, we are facing the question of how we can calculate the similarity among time series (data produced by nearby nodes) in a distributed and communication-efficient way when 1) the actual time series are not allowed to be shared among nodes (for privacy and security issues) and 2) the similarity calculation must be done with minimal communication among nodes.

We address these challenges at the same time by having each node calculate a *sketch* (i.e., a concise summary fitting in one or two wireless packets) of its time series and sending this sketch to its CH. Then, the CH computes the similarities among the sketches and decides which node(s) will participate in the federation for the next rounds. What is the nature of a sketch?

Since we are dealing with time series of multidimensional points *x*, each node computes the discrete Fourier transform (DFT) (in the experiments, we implemented with the fast Fourier transform) and saves the first few coefficients. This set of a handful of coefficients comprises the *sketch* that it sends to its CH to perform the similarity check among its cluster members. The DFT transforms a signal from the time domain to the energy domain. The next question is whether the similarity check on the energy domain performed by the CH is equivalent to the similarity check on the time domain. This is guaranteed by Parseval's theorem, which states that if $\vec{X}$ is the DFT of a *n*-point signal $\vec{x}$, then

$$\sum_{i=0}^{n-1} |x_i|^2 = \sum_{f=0}^{n-1} |X_f|^2$$

and, for any two signals *x* and *y*, it holds that

$$\text{Distance}(x, y) = \text{Distance}(X, Y)$$

i.e., the DFT preserves the Euclidean distance between the two signals.

Should each node choose to send the entire set of DFT coefficients, then we would gain nothing in terms of communication efficiency and privacy. Instead, each node sends the most significant (with the highest energy) DFT coefficients to its CH, and, thus, the CH performs an approximate similarity check among the data produced by its cluster members. The fewer the coefficients sent to the CH, the less accurate the similarity check, and, thus, fewer nodes are able to be excluded from the federation. When more coefficients are sent to the CH, the similarity check is more accurate, and, thus, more nodes are able to be excluded from the federation; on the other hand, in this case, we increase the communication efficiency and energy consumption as well, but we compromise privacy since reconstruction of the original data is possible.

Of course, we have silently assumed that, in our clustered architecture, it holds that "nodes feel less anxiety when sending information to acquaintances"[13] (i.e., a cluster member to its CH). The issue of privacy violation (reconstruction of the original data) is also a major concern/problem for traditional FL since even gradient sharing can cause privacy violation. In practice, in each training round, each node generates and transmits a sketch of around a dozen coefficients. This sketch-based approach can be applied in finding similarities among weights as well.[14]

The Federated Averaging Procedure

Each node *i* has access only to its own local dataset D_i, and all nodes share a common neural network model. In the proposed method, the nodes of each cluster train their local model and share their DFT coefficients with their CH, which selects from which node it collects all of the local updates, including its own, and implements the averaging. Afterward, the CH broadcasts, through the selected gateway nodes, the updated parameters to its related nodes for them to continue training. Periodically, global averaging is implemented among CHs by exchanging their parameters. Specifically, the parameters are computed and propagated via the backbone to every CH to start the next round of training (see Algorithm 1). The possible communication links between the devices of each cluster (blue circles) are represented with the solid lines. Within each cluster, devices communicate only with the CH. For the global averaging, there is a backbone network established by the clustering procedure that connects the CHs with each other, which is represented with the red dashed line. Notice that there is an alternative red path among CHs to avoid using the same paths for conserving energy.

Pseudocode for the DISCREETER protocol is illustrated in Algorithm 1.

Complexities and Convergence

The proposed methodology involves computational and communication steps for setting up the backbone (i.e., clustering) among nodes and also computational and communication steps for extracting and transmitting the FFT coefficients. The computational complexity of clustering is *O*(*n*), where *n* is the number of network nodes, and each node incurs $O(1)$ communication complexity—i.e., $O(n)$ in total—for establishing the backbone. Moreover, the step of deciding the final CH by calculating centralities incurs an $O(v \times e)$ cost, where v is the number of nodes and edges inside a cluster. Typically, v and e are of the order of $O(\sqrt{n})$.

The extraction of the Fourier coefficients using the FFT algorithm entails a computational complexity of $O(m\log_2 m)$, where m is the size of the data. Since the total size of Fourier coefficients sent to CHs is orders of magnitude smaller than the size of exchanged weights, and, additionally, the coefficients can be packed in the beacon messages by nodes, we do not count this overhead as communication overhead. This is one more contribution of the present work, i.e., detecting data similarity among nodes with practically zero communication cost.

Cluster-local and global weight convergence can be proven for all standard convexity assumptions on the loss function.

EXPERIMENTAL EVALUATION

Experimental Setup

We developed a simulation framework in Python that accommodates a wireless network topology, generated as described by Basaras et al.[2]; a reliable communication means for message exchange among nodes; a specific neural network architecture identical for all network nodes; and local storage. The generic workflow runs as follows: each node performs local training using owned data and then exchanges its local weights wirelessly (if it participates in the federation); then performs aggregation (as described in the "The Federated Averaging Procedure" section); then adopts the new weights; and continues the local training until the (local/global) convergence ratio reaches the percentage of a threshold, bigger than zero (in our case the threshold is defined as 5% of loss). Once the loss values coincide with the threshold, we stop the learning process to prevent the *overfitting* effect, which is often observed when the loss values become exactly equal to zero.

For measuring performance, we used the following:

- *Accuracy*, defined as the ratio of correct prediction to total predictions.
- *Mean square error* loss function values in every round of training for every node.
- *Communication overhead*, measured by the number of nodes participating in the federation.

We performed a large number of experiments with networks comprising 10, 20, and more than 50 nodes; we also experimented with various types of neural architecture running on the nodes depending on the type of dataset used and the objective learning task. Here, we present a small, representative sample of the experiments. We worked on a 20-node network clustered with the proposed variation of the max-min two-cluster formation algorithm (d = 2) and with time series data described in detail in the "Experimentation Dataset" section for a prediction task. Therefore, we used a 15-unit gated recurrent unit (GRU) trained with stochastic gradient descent with the Adam optimizer.

We again emphasize that our solution idea is adaptive to any deep learning model with any type/kind of

Algorithm 1: DISCREETER Protocol.

```
Perform network clustering with max-min d
    cluster formation algorithm
    //Elected CHs are aware of their cluster topology,
    so
Calculate their BC and the BC of their cluster
    members
Re-elect CHs based on BC
Announce CHs and their IDs

At EACH NODE:
Repeat until convergence:
  Perform local training and weight update
  For every k rounds:
      Compute Fourier coefficients
      Send Fourier to node's CH
      If node elected to participate in the
         federation:
         Send local weights to its CH
      Adopt average weights, received from
         its CH

At EACH CH:
Repeat until convergence:
    Perform local training and updating weights//
       Acts also as a normal node
    For every k rounds:
       Receive Fourier coefficients from its
          cluster members
       Compute data similarity among its
          members
       Notify the selected (for cooperation) nodes
       Receive node's local weights
       Perform aggregation of these weights
       Send cluster-local aggregated weights to
          other CHs
       Receive aggregated weights from all other
          CHs
       Average all received weights from other
          CHs
       Transmit the average weights back to the
          selected cluster nodes
```

parameters as long as the cooperative training takes place by weight (or gradient) exchange, with 262,940 samples in total, recording the temperatures of five Chinese cities for the period 2010–2015. Each node is assigned a unique part of the data (around 13,000 samples each). The assignment of data parts to nodes is done so that parts belonging to the same city temperatures are assigned to topologically neighboring nodes. Therefore, the data of the network nodes are not IID, and, moreover, there is probably some similarity among the data of neighboring nodes. Finally, each part is divided into 27 *chunks* of 500 data points each so that we have 27 rounds of training for each node that performs 20 epochs of local training in every round. We used 80% of the dataset for training and the rest for testing.

Competitors

To compare the efficiency of our proposed method, we used two baseline methods [and their Ridge regression-based (i.e., L2 regularization) variants], focusing on the communication reduction and prediction accuracy. The first one is called FEDLwa_ANP, in which all nodes participate in the federation process, meaning that all nodes communicate with their CH and send their parameters. This method exploits all available data, and we expect that it will incur the maximum communication cost and, probably, the best accuracy. The second one is called FEDLwa_RNP, in which a portion (we experimented with 25%, 50%, and 75%, respectively) of nodes that are randomly chosen participate in the federation process. In this case, there is obvious communication reduction by excluding nodes from the training, but the selection process is arbitrary, which may cause accuracy degradation of the model because of not selecting representative distribution data.

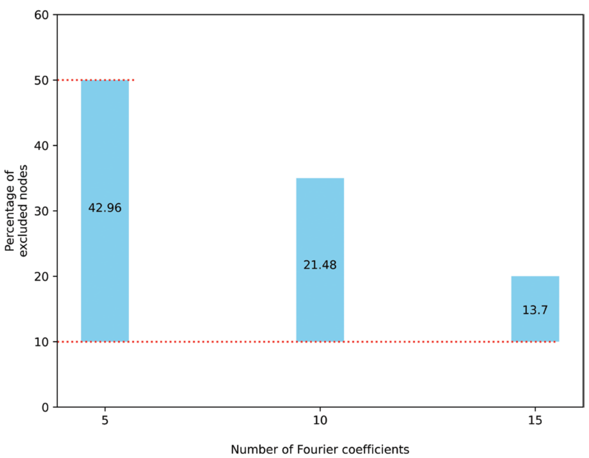


FIGURE 2. Percent reduction of cooperating nodes with respect to the coefficients.

Evaluation Results

First, we conducted experiments with 5, 10, and 15 Fourier coefficient variants, which resulted in 42.96% (varying from 10% to 50%) 21.48% (varying from 10% to 35%), and 13.7% (varying from 10% to 20%) average communication reduction in the network, respectively, preserving the high performance, as illustrated in Figure 2 and Table 1. The experimental evaluation shows that using 15 coefficients instead of 10 or five is a more reliable option for approximating our data since the more coefficients there are, the more accurate the representation of the original data (i.e., using more coefficients implies that we detect fewer nodes as similar from the perspective of their data similarity). This is experimentally justified since we have a percentage drop of almost 60% among the respective maximum communication reduction percentages, meaning that the number of excluded nodes lessens when switching from five coefficients to 15 coefficients.

Then, as depicted in Figures 3 and 4, DISCREETER not only retains the high performance, attained when all nodes globally train the neural network model (FEDLwa_ANP method: 96.59%), but it also outperforms this baseline, while it reduces the communication rounds

TABLE 1. Statistics of the representative node at the final round for the 20-node network configuration.

	Loss (%)		Mean Test Accuracy (%)
Algorithm	**10th Epoch**	**20th Epoch**	
FEDLwa_ANP	7.35	8.44	96.59
L2-FEDLwa_ANP	11.76	9.12	96.73
FEDLwa_RNP$_{25\%}$	29.67	32.83	81.47
FEDLwa_RNP$_{50\%}$	42.31	35.57	74.99
FEDLwa_RNP$_{75\%}$	50.708	46.64	72.38
L2-FEDLwa_RNP$_{25\%}$	24.99	27.22	85.63
L2-FEDLwa_RNP$_{50\%}$	48.95	38.61	74.72
L2-FEDLwa_RNP$_{75\%}$	45.17	41.52	76.22
DISCREETER$_{5coef}$	9.88	7.28	96.45
DISCREETER$_{10coef}$	10.31	10.47	97.26
DISCREETER$_{15coef}$	8.30	9.14	96.78
L2-DISCREETER$_{5coef}$	19.03	16.24	96.43
L2-DISCREETER$_{10coef}$	16.87	12.3	96.15

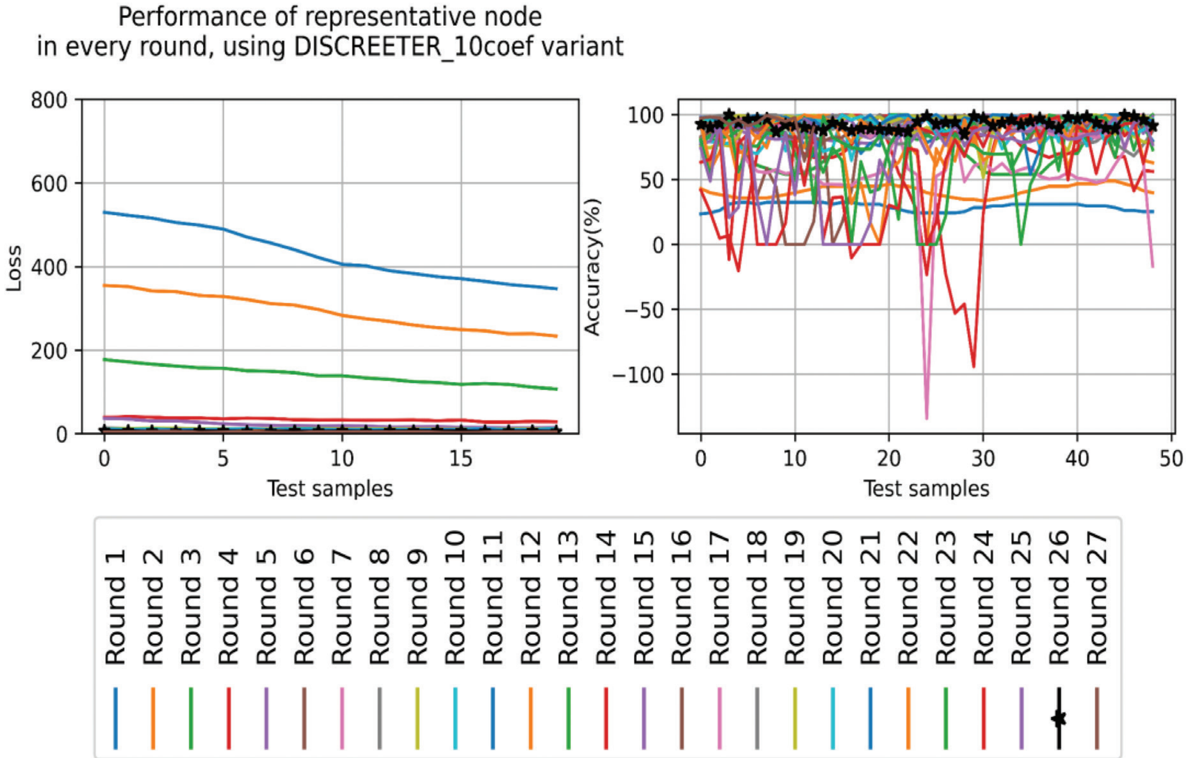


FIGURE 3. Progress of a representative node for DISCREETER with 10 Fourier coefficients with respect to the rounds. coef: coefficients.

of total training. More elaborately, DISCREETER's prediction accuracy (about 97.26%), in the case where 10 coefficients are selected for calculating the Euclidean distance exceeds, by about 0.7%, the levels of the corresponding FEDLwa_ANP method, with a significant average reduction of approximately 21.48% (varying from 10% to 35%) in the communication overhead among the participating nodes. It is shown that wiser selection of data leads to less noise diffusion in the network, which explains the reason why DISCREETER achieves better performance than the baseline method.

Compared to a heuristic of arbitrarily chosen participating nodes (FEDLwa_RNP), DISCREETER reaches up to a 34.3% increase in accuracy levels in some cases (see Table 1). The accuracy drop in FEDLwa_RNP is predictable since omitting significant data without a selection criterion may harm model generalization irredeemably. Especially in the case where 75% of nodes are excluded, we have an accuracy reduction of 11.15% comparatively with the FEDLwa_RNP$_{25\%}$ implementation.

However, the L2 regularization term mitigates the problem of exploding gradients; in our case, it decelerates the convergence of DISCREETER at a ratio equal to 55.1% since it transforms the loss function into a quadratic-like one. Hence, it might suppress the ability of a GRU to retain its memory and, thus, learn complex tasks. Conversely, when L2 is applied in FEDLwa_RNP, it accelerates its convergence speed by about 17%, especially in the case of the exclusion of 25% of the nodes, which is due to the many nodes and data being excluded from the federation. The aforementioned results are illustrated in Table 1 and in Figures 3 and 4.

CONCLUSION

In this work, we proposed a purely distributed FL methodology for small devices, connected via wireless ad hoc networks, with the aim of overcoming both privacy-preserving problems and excessive communication costs. Toward these goals, we developed an ad hoc network clustering scheme that creates relatively large clusters with the CH appropriately placed at the "center" of the cluster. Local averaging takes place inside each cluster, and a new node participation protocol into the federation was proposed, namely, DISCREETER. It is based on discovering the similarity of data from nodes that belong to the same cluster by calculating small summaries that are transmitted to the CH, which then decides which cluster members possess "different data" so as to enter the federation. The results show that the proposed heuristic reduces the communication overhead by up to 42%, while it preserves its high accuracy, up to 96%, compared to baseline heuristics (FEDLwa_ANP and FEDLwa_RNP).

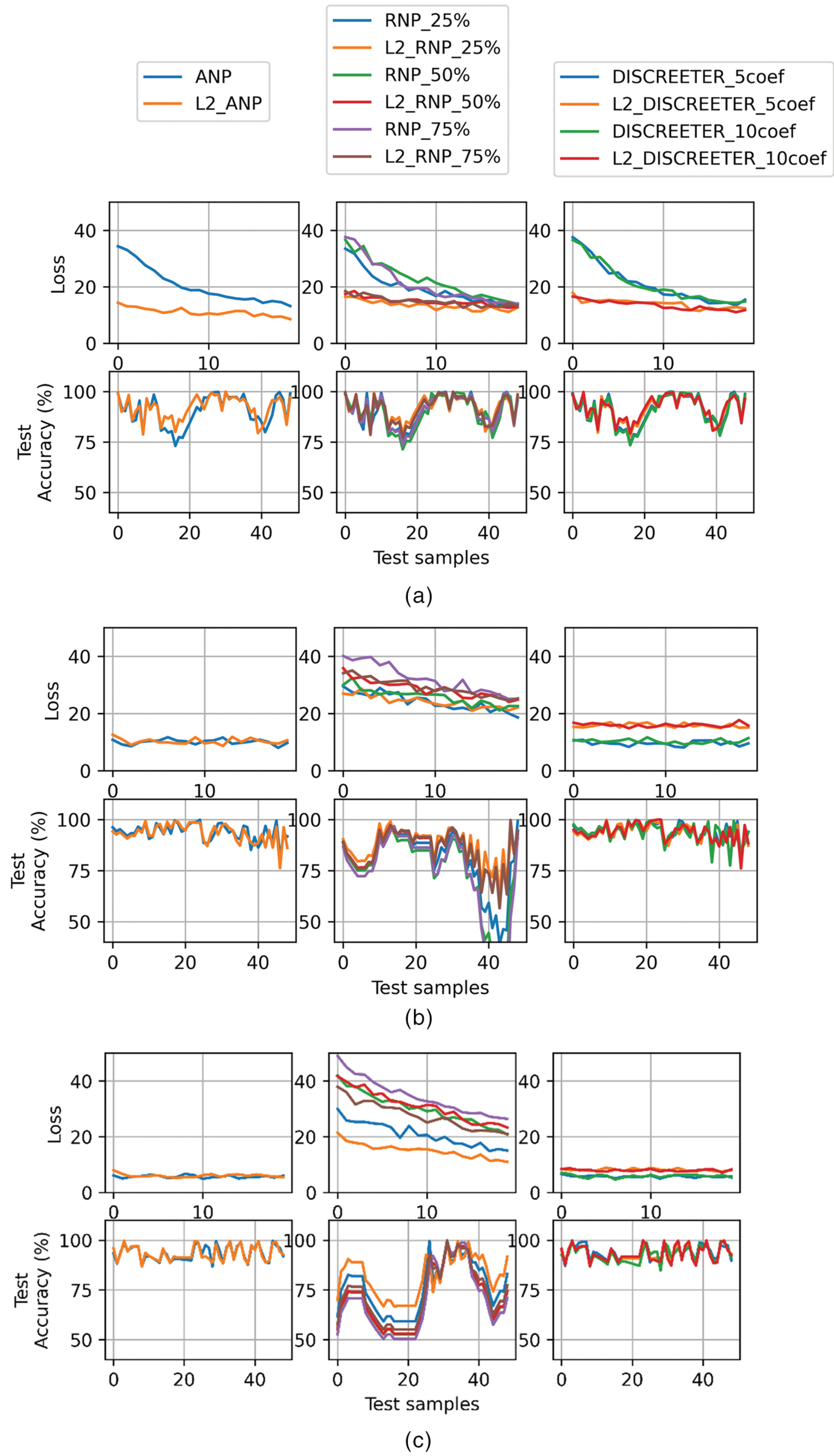


FIGURE 4. Representative node performance after the global averaging procedure: (a) Performance of a random node at round 5. (b) Performance of a random node at round 20. (c) Performance of a random node at round 26. ANP: all nodes participation; RNP: random nodes participation.

ACKNOWLEDGMENTS

Evangelia Fragkou's research work is supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI PhD Fellowships (Fellowship 5631).

REFERENCES

1. A. D. Amis, R. Prakash, T. H. P. Vuong, and D. T. Huynh, "Max-min d-cluster formation in wireless ad hoc networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun. 19th Annu. Joint Conf. IEEE Comput. Commun. Soc. (Cat. No.00CH37064)*, Tel Aviv, Israel, 2021, vol. 1, pp. 32–41, doi: 10.1109/INFCOM.2000.832171.
2. P. Basaras, G. Iosifidis, D. Katsaros, and L. Tassiulas, "Identifying influential spreaders in complex multilayer networks: A centrality perspective," *IEEE Trans. Netw. Sci. Eng.*, vol. 6, no. 1, pp. 31–45, Jan./Mar. 2019, doi: 10.1109/TNSE.2017.2775152.
3. E. T. M. Beltran et al., "Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 5, pp. 2983–3013, 4th Quart. 2023.
4. S. Hosseinalipour et al., "Multi-stage hybrid federated learning over large-scale D2D-enabled fog networks," *IEEE/ACM Trans. Netw.*, vol. 30, no. 4, pp. 1569–1584, Aug. 2022, doi: 10.1109/TNET.2022.3143495.
5. Y. Fraboni, R. Vidal, L. Kamenu, and M. Lorenzi, "A general theory for client sampling in federated learning," in *Proc. Int. Workshop Trustworthy Federated Learn.*, 2022, pp. 46–58.
6. Z. Li et al., "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3347–3366, Apr. 2023, doi: 10.1109/TKDE.2021.3124599.
7. B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, PMLR, 2017, pp. 1273–1282.
8. I. Mohammed et al., "Budgeted online selection of candidate IoT clients to participate in federated learning," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5938–5952, Apr. 2021, doi: 10.1109/JIOT.2020.3036157.
9. H. Ochiai et al., "Wireless ad hoc federated learning: A fully distributed cooperative machine learning," 2022, *arXiv:2205.11779*.
10. Q. Pan, H. Cao, Y. Zhu, J. Liu, and B. Li, "Contextual client selection for efficient federated learning over edge devices," *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 6538–6548, Jun. 2024, doi: 10.1109/TMC.2023.3323645.
11. S. Savazzi, M. Nicoli, and V. Rampa, "Federated learning with cooperating devices: A consensus approach for massive IoT networks," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4641–4654, May 2020, doi: 10.1109/JIOT.2020.2964162.
12. S. Chen. "Beijing, PM2.5." UCI Machine Learning Repository. Accessed: Jul. 7, 2024. [Online]. Available: https://archive.ics.uci.edu/dataset/381/beijing+pm2+5+data
13. A. Taya, T. Nishio, M. Morikura, and K. Yamamoto, "Decentralized and model-free federated learning: Consensus-based distillation in function space," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 8, pp. 799–814, 2022, doi: 10.1109/TSIPN.2022.3205549.
14. P. Tian, W. Liao, W. Yu, and E. Blasch, "WSCC: A weight-similarity-based client clustering approach for Non-IID federated learning," *IEEE Internet Things J.*, vol. 9, no. 20, pp. 20,243–20,256, Oct. 2022, doi: 10.1109/JIOT.2022.3175149.
15. J. Wang, A. K. Sahu, Z. Yang, G. Joshi, and S. Kar, "MATCHA: Speeding up decentralized SGD via matching decomposition sampling," in *Proc. 6th Indian Control Conf. (ICC)*, Hyderabad, India, 2019, pp. 299–300, doi: 10.1109/ICC47138.2019.9123209.

EVANGELIA FRAGKOU is a Ph.D. student in the Department of Electrical and Computer Engineering at the University of Thessaly, 38334, Volvos, Greece. Contact them at efragkou@uth.gr.

ELEFTHERIA CHINI is a B.Sc. graduate in the Department of Electrical and Computer Engineering at the University of Thessaly, 38334, Volvos, Greece. Contact them at echini@uth.gr.

MARIA PAPADOPOULOU is a B.Sc. graduate in the Department of Electrical and Computer Engineering at the University of Thessaly, 38334, Volvos, Greece. Contact them at marpapadopoulou@uth.gr.

DIMITRIOS PAPAKOSTAS is a postdoctoral researcher in the Department of Electrical and Computer Engineering at the University of Thessaly, 38334, Volvos, Greece. Contact them at papdimit@uth.gr.

DIMITRIOS KATSAROS is an associate professor in the Department of Electrical and Computer Engineering at the University of Thessaly, 38334, Volvos, Greece. Contact them at dkatsar@uth.gr.

SCHAHRAM DUSTDAR is a full professor of computer science heading the Research Division of Distributed Systems at TU Wien, 1040, Vienna, Austria. Contact them at dustdar@dsg.tuwien.ac.at.


Waseda University
IT Promotion
Research Institute
IEEE

EDITOR: Dimitrios Serpanos, Cti Diophantus and University of Patras, serpanos@computer.org

DEPARTMENT: CYBER-PHYSICAL SYSTEMS

This article originally appeared in Computer vol. 58, no. 3, 2025

The Compute Continuum: Trends and Challenges

Mauro Tortonesi, University of Ferrara

The compute continuum unifies computational, network, and storage resources across multiple layers, providing new opportunities to service developers and providers as well as challenges.

The compute continuum is an exciting concept, unifying computational, network, and storage resources across multiple layers, accessed through a largely homogeneous application programming interface (API).[1] This results in a distributed computing environment where resources can be seamlessly allocated from edge devices, including embedded and cyberphysical systems, to the cloud infrastructure. The strong focus on the homogeneous allocation and management of resources from the edge to the cloud, leveraging cloud-native architectures and modern technologies like containerization and orchestration, makes the compute continuum more suitable for the deployment and management of IT services with dynamic workloads and varying operational requirements than previous models, such as fog computing or cloudlets.

There are several enablers and drivers that contribute to the development of the compute continuum. First, containerization and the microservice paradigm have become a standard throughout the IT industry; standardized and lightweight service components are reasonably easy to manage from the deployment and migration perspectives. Second, the wide deployment of Kubernetes (K8s) has created a distributed system of K8s clusters, where we can assume to have ubiquitous orchestration functions in each cluster that can be accessed through a standard API; Kubernetes has become de facto standard cluster orchestrator in modern IT and has proved to work effectively from the edge to the cloud. The Kubernetes-driven homogenization of hosting platforms enables the realization of higher-level, system-wide orchestration functions that work across the compute continuum. Third, the increasing cost of cloud OPEX and the improvement of computing capabilities of embedded/edge hardware platforms make increasingly attractive the reallocation of computational tasks from the cloud to the edge. This trend is further accelerated by the increasing amount of Internet of Things (IoT)-generated data and the contextual computational and storage requirements of artificial intelligence/machine learning (AI/ML) tasks instantiated to process that data at the edge of the network. Finally, the increasing capabilities of modern mobile networks (5G and beyond) and the flexibility brought by network virtualization allows the creation of a high-speed, efficiently reconfigurable interservice connectivity fabric that meets the challenging demands of modern large-scale IT services.

Digital Object Identifier 10.1109/MC.2024.3520255
Date of current version: 26 February 2025

OPPORTUNITIES AND APPLICATION SCENARIOS

The compute continuum opens significant opportunities for service developers and providers, as it allows for the dynamic reallocation of services and continuous optimization across various metrics like revenue, user satisfaction, and sustainability. More specifically, the homogeneous management of resources at several layers enables: the seamless (re)allocation

of service components in the most convenient location along the edge-to-cloud spectrum; the continuous reconfiguration and reweaving of services to match high-frequency dynamics of workload and context; and the coherent optimization of the IT service fabric across the entire continuum. Thus, the compute continuum promises to improve operational efficiency and enable new applications in many domains.

For instance, Industry 5.0 applications deal with an increasing amount of machine-generated data that must be processed in a distributed fashion with dynamic workloads, and an increasing request for computationally intensive AI/ML tasks.[2] Leveraging the resources provided by the compute continuum, manufacturing companies can overcome the need to deploy dedicated and often proprietary platforms to support their data analysis requirements, with significant advantages in terms of CAPEX and time-to-market.

In the health-care domain, the compute continuum can be used to develop innovative solutions that improve patient care. For example, patient rehabilitation treatments—either in hospitals or at home—might require significant computational resources to process data in real time. The capability to dynamically allocate these tasks on the compute continuum, without having to purchase, deploy, and manage dedicated computational, network, and storage resources, might reduce costs and barriers to entry and speed up innovation in the critically important field of health-care-related services.

In field service management, the resources of the compute continuum could support planned maintenance in remote locations through the dedicated scheduled deployment of continuum resources (for example, mobile computer clusters, data, and AI models) in proximity of the location(s) of interest. This would allow operators to have a much broader and more capable range of support functions compared to the current state of the art, mostly based on platforms of relatively limited capabilities such as rugged tablets.

In vehicular networks, the compute continuum can be used to develop innovative solutions that improve the safety and efficiency of the transportation system. Going beyond the current street-level-focused approach based on offloading tasks to road side units (RSUs), the compute continuum could represent the perfect environment to develop a next generation of broader-focused and low-latency applications that dynamically allocate and leverage resources at different levels of aggregation (street, city, province, and so on) to support the vehicles in their tasks.

THE AI CONNECTION

AI is undergoing rapid development and has been attracting significant attention recently. This led to the development of many interesting methodologies and tools that represent a solid foundation for the development of service fabric management solutions for the compute continuum. Deep reinforcement learning (DRL)[3] has emerged as an interesting solution to train relatively robust and flexible agents that can efficiently manage the resource requirements across rather large portions of the compute continuum, in a wide range of conditions—and even adapt to significant context and environment changes. The adoption of offline DRL, compounded with digital twin approaches that enable agents to broaden their perspective and to validate their decision making through what-if scenario analysis,[4] holds even more interesting promises. And more approaches, such as active inference[5] and neuro-symbolic AI, are also being investigated. However, the correctness and quality of the decisions made by these solutions needs to be properly explored through trustworthy AI and explainable AI methodologies and tools, respectively.

In any case, the compute continuum will not only leverage AI but also actively enable it. We can expect that a large share of tasks that will be instantiated on

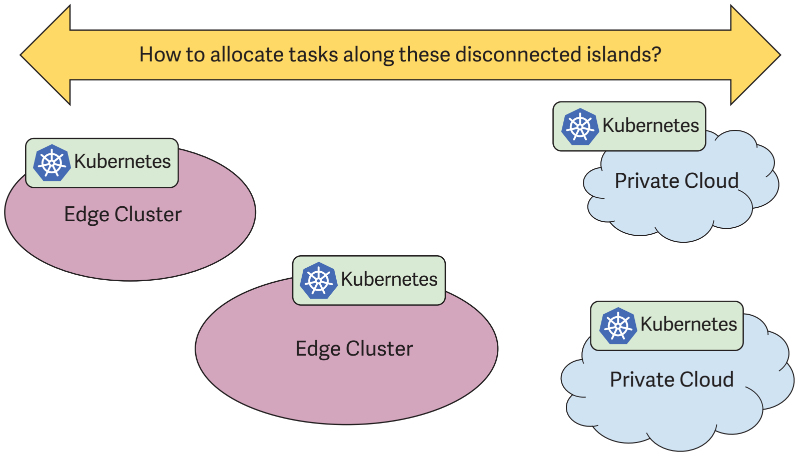


FIGURE 1. Multicluster orchestration still represents a challenge, despite Kubernetes's ubiquitous presence.

the compute continuum will heavily leverage AI/ML, and generative AI in particular. These AI-focused tasks will be based on ML models with complex and often application-specific lifecycles. ML models—either ad hoc/proprietary or foundational/open-weights—will very likely have to undergo multiple retrainings during their lifetime—either to exploit the knowledge acquired through new data sets or to address the issues of data and/or concept drifts.

As a result, there is the need to provide developers with functions that support the implementation and management of distributed AI services. We can thus envision the compute continuum to integrate a next generation of solutions that cater to AI service developers, automating the development of distributed inference and training of ML models and enabling the migration of AI services across the continuum. MLOps, federated and distributed learning, and continual learning represent very interesting functions that service developers could leverage on top of compute continuum.

CHALLENGES

The path to exploiting the potential of compute continuum faces several major challenges. At the federated/multicluster level, system level orchestration and intercluster coordination are still major challenges, as depicted in Figure 1, despite many attempts from both the research and the open source communities (KArmada, Submariner, Liqo,[6] and so on). Although we can rely on battle-tested scheduling and autoscaling solutions at the single cluster level, methods for reliable and automated scheduling, scaling, and service migration in a large-scale cluster federation have not reached production-level maturity yet. In fact, distributed decision-making in resource management across the compute continuum will need thorough validation and testing to avoid "ping-pong" effects, where cooperating agents in charge of decision making at the cluster level that could keep overturning each other's decisions—and, for example, endlessly migrating tasks across different clusters.

In addition, to reify the promise of a continuous computing environment, one needs to consider resource redeployment to bridge potential gaps in local or temporal network disconnections. This might require going beyond the adoption of disconnection resilience methodologies and tools, such as ghost devices from the IoT domain and circuit breakers from the microservices one, and to consider topology control solutions where clusters are (de)activated on demand or even moved between different locations, especially in the case of "high stakes" application scenarios.

At the market level, there is the significant question of what is the best business model for resource sharing across the compute continuum. Will we see the emergence of public continuum providers, which play a role similar to the one of Cloud providers, building, managing, and offering public access to small-scale data centers at the edge? Or will we instead see a community-driven approach, where people and companies will federate their servers and allow other members of the community to bid for the currently available resources on their platforms? Perhaps a combination of both? In any case, the high-level orchestration functions should be designed and tailored for the specific business models/commercial offers that will reach wide adoption.

Resource sharing also poses cybersecurity challenges. The capability of executing service components on a homogeneous—but shared—distributed environment brings significant advantages in time- and cost-to-market but also represents a disruptive model with respect to the single-tenant/silo model adopted by most edge- and IoT-focused applications. We will need to rethink threat models and adopt stricter cybersecurity policies in order to build a secure compute continuum.

Finally, the compute continuum will very likely include a small portion of clusters that do not use Kubernetes. This presents interoperability issues at different levels, such as the integration with non-Kubernetes-based orchestration APIs and the enforcement of deployment constraints on software components that can execute only on a restricted pool of compatible nodes—for example, legacy IT systems or high-performance computing-focused embedded and cyberphysical systems. These gaps in the continuum suggest developing higher-level abstractions that enable virtual resource modeling and management in an interoperable way, even across the least homogeneous part of the compute continuum.

The compute continuum is coming of age and calls for the rethinking of data and computational allocation and usage paradigms. The efficient resource management in this distributed environment requires advanced service fabric management solutions that enable the seamless deployment and orchestration of workflows from the edge to the cloud. This represents a major challenge that will demand a considerable amount of effort from both academia and industry in the years to come.

REFERENCES

1. S. Moreschini, F. Pecorelli, X. Li, S. Naz, D. Hästbacka and D. Taibi, "Cloud continuum: The definition," *IEEE Access*, vol. 10, pp. 131,876–131,886, 2022, doi: 10.1109/ACCESS.2022.3229185.
2. D. Serpanos, "Industrial Internet of Things: Trends and challenges," *Computer*, vol. 57, no. 1, pp. 124–128, Jan. 2024, doi: 10.1109/MC.2023.3331552.
3. J. Santos, E. Reppas, T. Wauters, B. Volckaert, and F. De Turck, "Gwydion: Efficient auto-scaling for complex containerized applications in Kubernetes through Reinforcement Learning," *J. Netw. Comput. Appl.*, vol. 234, Feb. 2025, Art. no. 104067, doi: 10.1016/j.jnca.2024.104067.
4. B. Sedlak, V. Casamayor Pujol, P. K. Donta, and S. Dustdar, "Equilibrium in the computing continuum through active inference," *Future Gener. Comput. Syst.*, vol. 160, pp. 92–108, Nov. 2024, doi: 10.1016/j. future.2024.05.056.
5. D. Borsatti et al., "KubeTwin: A digital twin framework for Kubernetes deployments at scale," *IEEE Trans. Netw. Service Manage.*, vol. 21, no. 4, pp. 3889–3903, Aug. 2024, doi: 10.1109/TNSM.2024.3405175.
6. M. Iorio, F. Risso, A. Palesandro, L. Camiciotti and A. Manzalini, "Computing without borders: The way towards liquid computing," *IEEE Trans. Cloud Comput.*, vol. 11, no. 3, pp. 2820–2838, Jul./Sep. 2023, doi: 10.1109/TCC.2022.3229163.

MAURO TORTONESI is an associate professor and the head of the Big Data and Compute Continuum Research Laboratory at the University of Ferrara, 44121, Ferrara, Italy. Contact him at mauro. tortonesi@unife.it.

EDITOR: Stephen J. Andriole, steve@andriole.com

COLUMN: LIFE IN THE C-SUITE

This article originally appeared in IT Professional vol. 27, no. 1, 2025

Artificial Intelligence Adoption Is Easy for Gigs, Start-Ups and Small Companies—But Not Mid-Sized and Large Enterprises

Stephen J. Andriole, *Villanova School of Business, Villanova, PA, 19085, USA*

AI, machine learning (ML) and generative AI (GenAI) platforms and tools provide an opportunity for smaller companies to solve business problems the same way large companies do—expect faster and cheaper. Gig workers, start-ups and small companies are in a strong position to exploit AI/ML/GenAI for competitive advantage—in many cases stronger positions than larger companies. C-suites in all-sized companies should be mindful of the opportunities and constraints around AI/ML/GenAI adoption.

Artificial intelligence (AI), machine learning (ML), and generative AI (GenAI) are the most important technologies in a generation. Every company on the planet is gearing up for the application of these technologies. But the process is often slow and expensive, especially in large companies. Expertise in AI/ML/GenAI is scarce and expensive. But the pressure to at least prototype some applications is relentless. Small or non-existent C-suites can move fast, but companies with large C-suites, federated organizational structures, and lots of corporate governance are always at a strategic disadvantage.

The stakes are high. The applications of AI/ML/GenAI to business models and processes are endless. Companies must examine AI/ML/GenAI methods, tools, techniques, platforms, and applications. They should especially focus on GenAI platforms and tools—like Gemini, Dall-E, ChatGPT, Claude, Copilot, Llama, Grok, Synthesia, and Midjourney—among hundreds of other tools at work as we speak, and develop the role that AI/ML/GenAI should play in enabling existing and new business models and processes across their companies: what do you want the technology to do? Admit applicants to college? Pick tomatoes with image recognition and robotics? Award (or deny) loans? Or maybe you want to develop a press release or a full-blown marketing plan for a new product or service? A business case for a new project? New code? Videos? Music? All well within the reach of the technology—today.

GenAI tools, platforms, and methodologies include large language models (LLMs), generalized pretrained transformers (GPTs), agentic AI, proprietary LLMs, custom GPTs, prompt engineering, data science, natural language processing, computer vision, and algorithms (especially neural networks). The major application areas include health care, transportation, finance, retail, gaming, manufacturing, marketing, customer service, human resources, agriculture, education, travel, entertainment, and computer security, among others that describe the range of possible applications, which is obviously enormous.

TECHNOLOGY ADOPTION

According to Gemini, "*technology adoption is the process of learning and using new technologies, and can involve individuals or organizations. It can also refer to the integration of new technologies into a business's operations and strategies.*" The "technology adoption curve" includes

- *Innovators*: The first to adopt a new technology, and are often the youngest, most social, and have the most financial resources.
- *Early adopters*: The second-fastest group to adopt a new technology, and are often younger,

Digital Object Identifier 10.1109/MITP.2025.3529859
Date of current version 20 February 2025.

have a higher social status, and have many financial resources.

- *Early majority*: Adopt a new technology after a varying amount of time, and make up the largest group of users.
- *Late majority*: Another large group of users that adopts a new technology.
- *Laggards*: The last group to adopt a new technology.

The technology adoption curve can be applied directly to AI/ML/GenAI, but there's a twist. 1) The complexity of the technology, 2) the speed with which it's developing, and 3) the breadth of its applied potential, makes this a special case that challenges companies of all shapes and sizes.

Gig workers, start-ups, and some small companies are in a strong position to exploit AI/ML/GenAI for competitive advantage. Remember that AI/ML/GenAI consists of platforms, tools, and methods. Companies do not have to develop anything to exploit AI/ML/GenAI for their customers and clients, which is different from traditional applications' development or "renting" enterprise applications in the cloud. Platforms like Gemini, Dall-E, ChatGPT, Claude, Copilot, Llama, Grok, Synthesia, and Midjourney—and many, many more—are always available. The payoff is not in the development of these application platforms but in their creative use. The opportunity lies with the creativity that companies bring to their customers and clients, and the paths to this creativity. Some companies have a lot of—although often crazy—creative paths. Others, the larger ones, have way too many crossroads that lead to nowhere.

SIZE MATTERS

We all know how addicted large companies are to processes, meetings, bureaucracies, vested interests, task forces, committees, governance, politics, lawyers, and long product development cycles. Public companies find it difficult to move at all, which often gets them into trouble. Mid-sized companies are schizophrenic: fast on Monday, painfully slow on Tuesday, and paralyzed for the rest of the week. Start-ups are chaotic, which makes them open to most anything. Gig workers pivot on a dime: they can adopt a new platform overnight, and then offer it to their clients. They can just as quickly discard a platform in which they've made only a minimal investment. "Fail fast/fail cheap" is the mantra.

So is there a special relationship that gig workers, start-ups, and small business can develop with AI/ML/GenAI? There absolutely is, and gig workers, start-ups, and small business should exploit this competitive advantage simply because larger companies cannot get out of their own way.

Many of these companies are really *small*, with usually fewer than 10 employees. In fact, the majority of small businesses in the United States have fewer than five employees.[1] They also have few standardized management processes. They often have no managerial bureaucracies at all. Due to the small number of employees, the nature and depth of corporate politics is much smaller than in medium-sized or large companies. Although product development is challenging, especially in large companies, gig workers, start-ups, and small companies can churn out ideas at an incredible pace. Although many of these ideas may make no sense, ideation is everywhere. Large companies struggle to generate ideas, and when they do, they're often pounded into the ground by those with powerful bureaucratic armies designed to kill ideas that might compete with theirs, or committees designed to think cautiously about releasing anything new, interesting, or disruptive into the market.[2]

> "Big businesses tend to be driven by bureaucracy and committees. If you're familiar with the expression 'design by committee,' you intuitively understand that when too many people are involved, the decision-making process is usually delayed and complicated. There's certainly truth to this; when decision-making is always a clunky process that involves dozens of people, you can't act quickly, which means you can't implement new ideas quickly."

Gig workers, start-ups, and small companies move much faster, have fewer vested interests to manage, and aren't afraid of cannibalizing excising product suites. As we all know, gig workers, start-ups, and small companies easily fuel more innovation than do large enterprises.

The technology curve maps onto the size of the companies that need to adopt AI/ML/GenAI.

- Gig workers and start-ups are the "innovators … the first to adopt a new technology, and are often the youngest, most social, and have the highest financial resources."
- Small companies … are "the second fastest group to adopt a new technology, and are often younger, have a higher social status, and have more financial resources."

- Mid-sized companies "adopt a new technology after a varying amount of time, and make up the largest group of users."
- Large companies are the "late majority, which is "another large group of users who adopt a new technology ... (along with) the laggards ... the last group to adopt a new technology."

HELPING THE HELPLESS

What took Marriott so long to respond to Airbnb? What about Kodak, Nokia, Segway, Blockbuster, Borders, Yahoo, Blackberry, Motorola, MySpace, Abercrombie & Fitch, and Toys"R"Us, among so many other companies that were ambushed by new entrants into their markets? These cases are well known and describe how gig workers, start-ups, and small companies can penetrate even the largest, most-established markets. Said a little differently, gig workers, start-ups, and small companies don't need much help adopting AI/ML/GenAI—it's what they do.

But medium-sized and large companies are dangerously slow to adopt technology for the reasons described earlier. Can big companies accelerate the AI/ML/GenAI adoption process? Maybe. It all depends on their C-suites. With that stipulation, here are five steps that might help:

1) *Step 1*: Education.
   Educate the company about AI/ML/GenAI, which starts in the C-suite—which should *loudly* announce the importance of understanding AI/ML/GenAI and its potential to impact the company's business processes and entire business model. Extend the educational program throughout the company and commit to ongoing education about AI/ML/GenAI. This step should also include an analysis of what its archrivals are doing with the technology.
2) *Step 2*: Acquisition.
   Constantly scan the market for companies to acquire, even if the acquisitions are "aqui-hires," *especially when they're aqui-hires of AI/ML/GenAI talent*. Many medium-sized and large companies do not have the talent to pursue AI/ML/GenAI initiatives, so they need help. They can try to upskill their teams, but it's much faster and efficient to just acquire companies and the people who understand the technology.
3) *Step 3*: Separation.
   Move the AI/ML/GenAI team far away from the corporate bureaucracy that constrains technology adoption and innovation. "Far away" means out of town, out of state, or across the country. It also means escape from the bureaucratic protocols that run medium-sized and large companies like task forces, committees, and long product development cycles.
4) *Step 4*: Prototype.
   The "away team" should fill the pipeline with as many demonstration prototypes as possible. AI/ML/GenAI prototypes should be the constant deliverables of the team and should be assessed according to some metrics that determine whether they can become minimum viable products (MVPs).
5) *Step 5*: MVPs.
   The objective is the development of products and services that can be tested in the marketplace, products, and services that can scale.

HOW BIG IS YOUR C-SUITE?

The adoption of AI/ML/GenAI can be fast, slow, or not at all. Gig workers can spend a month learning a platform like Synthesia and become capable of generating some incredibly creative solutions to a variety of problems. They can learn how to develop AI agents and custom GPTs. They can learn how to generate new code, music, or images. Start-ups can build small teams to do these and other things. Very importantly, gig workers and start-ups can deploy individual contributors to plow through the technology—where larger companies vastly prefer team efforts.

Small companies can also brand themselves around activities like text to video. Medium-sized and large companies can do some, but not all, of the same things if they take the aforementioned steps.

At the end of the day, all of this depends upon who's in the C-suite. The smaller the company, the smaller the C-suite, which makes it easier to adopt new technology. Large C-suites are complicated, political, and too often unwilling to take the plunge—into anything. But we already knew that, right?

REFERENCES

1. A. Hait, "The majority of U.S. businesses have fewer than five employees," *US Census Bureau*. Jan. 19, 2021. [Online]. Available: https://www.census.gov/library/stories/2021/01/what-is-a-small-business.html
2. S. Gibbons, "Why big and complex businesses find it difficult to innovate," *Forbes Magazine*. Mar. 1, 2022. [Online]. Available: https://www.forbes.com/sites/serenitygibbons/2022/03/01/why-big-and-complex-businesses-find-it-difficult-to-innovate/

STEPHEN J. ANDRIOLE is the Thomas G. Labrecque Professor of Business Technology in the Villanova School of Business, Villanova, PA, 19085, USA. Contact him at stephen.andriole@villanova.edu or https://andriole.com.


Computer
Trustworthy AI—Part II
Computer
TECHNOLOGY PREDICTIONS
Computer
Attributes of Quality
Computer
Venture Scientists and Technology Entrepreneurship
Computer
Dependable Computing
Computer
Computer
Ethics and Computing
Technologies
CS PRESIDENT'S MESSAGE 14
IEEE CS AWARDS 18
Computer
STATE OF TELEMEDICINE—PART I
IEEE

EDITOR: Hal Berghel, University of Nevada, Las Vegas, hlb@computer.org

DEPARTMENT: OUT OF BAND

This article originally appeared in
Computer
vol. 58, no. 7, 2025

Generative AI and Health Care: Brief Survey

Roy Rada, University of Maryland

AI is transforming health care. Big Data radiation oncology and high-touch palliative care illustrate different opportunities.

Artificial intelligence (AI) and health care have impacted one another since the beginning of AI. Traditionally, the branches of health care most dependent on digitized information benefited the most from AI. Do the latest advances in natural language processing impact the trajectory? Since large language models (LLMs) exploit free-form text, do they open opportunities for branches of health care traditionally reliant on human–human interaction? Or will the areas that get the most investment already be the ones that get the most investment in the future, as a continuation of the famous aphorism "the rich get richer, while the poor get poorer."

The history of computer applications in health care supports the famous "Rich get Richer" phenomenon. In 1968, Merton introduced the "Matthew effect"[1] as "This complex pattern of the misallocation of credit for scientific work must quite evidently be described as the Matthew effect … For unto everyone that hath shall be given." Evidence for the Matthew effect has appeared across occupations. A study of health care delivery in China concludes, "the Intercity Health Network generates new structural inequalities in health care access exhibiting a Matthew effect."[2] New investments in digital information systems in health care have typically gone into areas that most relied on digital information.[3] Historically, the first health care computer investments were in accounting and next, radiology and pathology. Health care and AI illustrate the Matthew effect. This essay next reviews the AI and health literature before focusing on two application areas, radiation oncology and palliative care, which sit at opposite extremes of the capital spectrum.

Digital Object Identifier 10.1109/MC.2025.3564446
Date of current version: 27 June 2025

AI AND HEALTH CARE

Deep learning has been successfully applied to clinical medicine for more than three decades.[4] Generative AI trains on multimedia libraries, while LLMs train on document libraries,[5] and both perform well on many health-care tasks.[6] While LLMs lack epistemic validity,[7] they create new opportunities for health care applications that converse with humans

A search of IEEE Xplore for ["artificial intelligence" AND ("health care" OR medicine) AND "systematic reviews" AND 2020-to-2025] returned 54 citations. Most citations were for highly focused topics, such as "multi-modal deep learning diagnosis of Parkinson's Disease,"[8] but nine were broad. One broad review shows that Big Data facilitates personalized medicine, predictive analytics, early illness diagnosis, precision diagnostics, and treatment optimization from electronic health records, genomes, wearables, and medical imaging.[9] In another broad review, adoption of AI applications depends on the factors of technical feasibility, ease of use, system quality, performance, usability, social influence, and trust.[10] The IEEE literature suggests that AI increases the effectiveness and efficiency of health care.[11]

RADIATION ONCOLOGY AND PALLIATIVE CARE

Radiation oncology is one of three oncology branches, the other two being medical and surgical oncology. In preparing a radiotherapy treatment plan, the doctors

acquire images, segment the target, and plan doses to kill cancer cells. Radiomics is a method to quantitatively analyze medical images to uncover tumoral patterns not appreciated by the naked eye and to support personalized therapy.[12] Doctors try to predict outcomes based on both radiomics and genomics. In the delivery of radiation, the team monitors patient changes and adapts treatment. Changing radiation based on patient motion occurs in real time under computer control. Radiation oncology relies on massive amounts of digital information and repetitive tasks[13]—"Big Data" incarnate.[14]

Radiation oncology is ripe for AI applications[15]: "AI … transformative applications in radiation oncology given … a heavy reliance on digital data processing and computer software." Radiation oncologists appreciate AI's impact and say that AI revolutionizes radiation therapy.[16] AI is transformative at every step of the radiation oncology process.

Palliative medicine differs sharply from radiation oncology. The World Health Organization states[17]: "Palliative care … improves the quality of life of patients and their families who are facing … life-threatening illness, through the prevention and relief of suffering … whether physical, psychosocial, or spiritual." Diseases benefiting from palliative medicine include cancer, heart failure, and stroke. Nevertheless, only 14% of people who need palliative care currently get it.[17] Health insurance underfunds palliative care, and hospitals, at a loss to themselves, provide more than half of the overall cost of palliative care.[18] A systematic review of challenges in palliative care concludes that the most significant barrier is a lack of resources.[19] If technology could reduce costs of delivering palliative care, then palliative care professionals have an obligation to explore using it.[20]

Palliative care uses AI less than radiation oncology. Grant and patent activity evidences the divide. With the U.S. National Institutes of Health (NIH) "Research Portfolio Online Reporting Tools," users search a repository of both intramural and extramural NIH-funded research projects and access patents resulting from NIH funding.[21] A search on this NIH repository for "radiation oncology" AND "artificial intelligence" returns 68 active projects linked to 60 patents. A search on "palliative care" AND "artificial intelligence" retrieves 10 active projects and 0 patents. Radiation oncology is "richer" than palliative medicine.

LARGE COMPANIES THAT DEVELOP ELECTRONIC HEALTH RECORDS SYSTEMS ARE WORKING WITH LARGE PROVIDER NETWORKS TO INTEGRATE LLMS INTO THE WORKFLOW.

How are LLMs special for palliative care? LLMs can forecast mortality or monitor pain. For predicting mortality, one project uses traditional machine learning on structured data sets,[22] while another project uses an LLM that reads unstructured medical records.[23] These two projects highlight the salient difference between LLM and non-LLM approaches. The LLM approach works on unstructured natural language. The LLM approach lends itself more readily to extensions and diffusion into electronic health records systems.

An LLM for patient mental health showed that the LLM system responds like a professional mental health counselor.[24] For the doctor, an LLM inferred wishes of mentally incapacitated patients based on prior records.[25] LLMs are well suited to palliative care.[26]

WORKFLOW

One challenge in the roll out of LLM applications is gaining the trust of the intended users. While this trust is a necessary condition for diffusion, another necessary condition is fitting into the workflow. The surge of interest in expert systems in the 1980s bares comparison to the surge of interest in LLMs. The 1980s MYCIN expert system was as good as expert physicians in its problem area of bacteremia, and it offered recency, accuracy, coherence, and transparency confirmed by clinical trials.[27] However, to use MYCIN, the physician needed to access a separate, stand-alone application and reenter patient information into that application which took enough time that the cost/benefit ratio was too high for the physician—in other words, MYCIN did not fit into the workflow.[3] Large companies that develop electronic health records systems are working with large provider networks to integrate LLMs into the workflow.[28] However, journals underreport these efforts.

Workflow systems rely on roles defined as functions with rules for passing messages among roles. LLMs can mediate between roles and between a role and the rules of that role.[29] In radiation oncology, automation of workflow with the support of LLMs is advanced,[30] but not in palliative care.

A hospice care system, which is the ultimate example of palliative care, suits LLM integration into the workflow.[31,32] The system would represent the roles of the interdisciplinary team and the roles of patients and their family, as "dignity therapy" illustrates. Dignity therapy helps patients reflect on end-of-life issues. The patient receives nine standard questions which guide a conversation. The dialog is recorded, transcribed, and edited into a legacy document for the patient to use in implementing the last steps of life.[33] However, dignity therapy is underutilized due to a lack of trained staff to help. With an LLM-automated social worker, patients could develop their dignity document.[34,35]

DISCUSSION

The IEEE literature highlights that AI with Big Data will improve health care. Radiation oncology relies on Big Data, attracts investment, and uses AI extensively. Palliative care relies on human–human interaction, attracts little capital, and uses AI sparingly. Radiation oncology is rich and getting richer. What is the future for palliative care?

Palliative care does not process vast amounts of information in real time, as radiation oncology does. Palliative care relies on natural language communication, and LLMs facilitate natural language communication. As death approaches, the roles of palliative care and natural language communication increase. LLMs are generalists[36] and could support palliative care applications both clinician-facing and patient-facing across every phase of the clinician and patient experience.

The betting AI entrepreneur invests in radiation oncology rather than palliative care. Radiation oncology uses LLMs, and as LLMs are commoditized, their use diffuses into poorer domains, such as palliative care. The radiation oncology patient cannot hope to understand the intricacies of radiation treatment, trusts the doctor, and hopes for a cure. For the hospice patient, by contrast, the doctor is less relevant, and the patient could benefit from LLMs but gets no help in using LLMs. Would the betting person see anything different here?

REFERENCES

1. R. K. Merton, "The Matthew effect in science: The reward and communication systems of science are considered," *Science*, vol. 159, no. 3810, pp. 56–63, 1968, doi: 10.1126/science.159.3810.56.
2. P. Zhao, J. Li, and M. Zhang, "Unequal roles of cities in the intercity healthcare system," *Nature Cities*, vol. 2, pp. 198–209, Jan. 2025.
3. R. Rada, *Information Systems and Healthcare Enterprises*. Hershey, PA, USA: IGI Global, 2007.
4. S. C. Lo, M. T. Freedman, J. S. Lin, and S. K. Mun, "Automatic lung nodule detection using profile matching and back-propagation neural network techniques," *J. Digit. Imag.*, vol. 6, no. 1, pp. 48–54, Feb. 1993, doi: 10.1007/bf03168418.
5. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nat. Med.*, vol. 29, no. 8, pp. 1930–1940, Aug. 2023, doi: 10.1038/s41591-023-02448-8.
6. X. Shi et al., "A bibliographic dataset of health artificial intelligence research," *Health Data Sci.*, vol. 4, Apr. 2024, Art. no. 0125, doi: 10.34133/hds.0125.
7. H. Berghel, "ChatGPT and AIChat epistemology," *Computer*, vol. 56, no. 5, pp. 130–137, May 2023, doi: 10.1109/MC.2023.3252379.
8. V. Skaramagkas, A. Pentari, Z. Kefalopoulou, and M. Tsiknakis, "Multi-modal deep learning diagnosis of Parkinson's disease—A systematic review," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 2399–2423, 2023, doi: 10.1109/TNSRE.2023.3277749.

9. A. Kumar and N. P. Singh, "A study on future data science and its impact in health care organizations," in *Proc. 15th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 1–7, doi: 10.1109/ICCCNT61001.2024.10724357.
10. G. B. Anugerah and A. N. Hidayanto, "What factors contribute to the acceptance of artificial intelligence on healthcare sector: A systematic review," in *Proc. 3rd Int. Conf. Intell. Cybern. Technol. Appl. (ICICyTA)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 477–482, doi: 10.1109/ICICyTA60173.2023.10428744.
11. A. Sharma and R. Kumar, "Artificial intelligence in health care sector and future scope," in *Proc. Int. Conf. Innovative Data Commun. Technol. Appl. (ICIDCA)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 210–214, doi: 10.1109/ICIDCA56705.2023.10100220.
12. A. K. Jha et al., "Implementation of big imaging data pipeline adhering to FAIR principles for federated machine learning in oncology," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 6, no. 2, pp. 207–213, Feb. 2022, doi: 10.1109/TRPMS.2021.3113860.
13. F. J. P. Hoebers et al., "Artificial intelligence research in radiation oncology: A practical guide for the clinician on concepts and methods," *BJR Open*, vol. 6, no. 1, Jan. 2024, Art. no. tzae039, doi: 10.1093/bjro/tzae039.
14. F. Abuhijla, I. Jaradat, A. Almousa, and R. Abuhijlih, "Impact of big data on radiation oncology practice," in *Proc. 1st Int. Conf. Cancer Care Inform. (CCI)*, 2018, p. 59, doi: 10.1109/CANCERCARE.2018.8618145.
15. E. Huynh et al., "Artificial intelligence in radiation oncology," *Nature Rev. Clin. Oncol.*, vol. 17, pp. 771–781, Aug. 2020, doi: 10.1038/s41571-020-0417-8.
16. M. Kawamura et al., "Revolutionizing radiation therapy: The role of AI in clinical practice," *J. Radiat. Res.*, vol. 65, no. 1, pp. 1–9, Jan. 2024, doi: 10.1093/jrr/rrad090.
17. S. Connor, "Global atlas of palliative care, 2nd edition," Worldwide Hospice Palliative Care Alliance, London, U. K., 2020, p. 120. [Online]. Available: https://www.iccp-portal.org/system/files/resources/WHPCA_Global_ Atlas_FINAL_DIGITAL.pdf
18. R. T. Sullender and S. A. Selenich, *Financial Considerations of Hospital-Based Palliative Care* (RTI Press Research Report Series). Research Triangle Park, NC, USA: RTI Press, 2016.
19. M. S. Amroud, P. Raeissi, S. M. Hashemi, N. Reisi, and S. A. Ahmadi, "Investigating the challenges and barriers of palliative care delivery in Iran and the world: A systematic review study," *J. Educ. Health Promotion*, vol. 10, no. 1, p. 246, 2021, doi: 10.4103/jehp.jehp_1325_20.
20. C. Barry, P. Paes, S. Noble, and A. Davies, "Challenges to delivering evidence-based palliative medicine," *Clin. Med.*, vol. 23, no. 2, pp. 182–184, Mar. 2023, doi: 10.7861/clinmed.2022-0336.
21. "RePORT: Research portfolio online reporting tools," Nat. Inst. of Health, Bethesda, MD, USA, 2025. [Online]. Available: https://report.nih.gov/
22. M. N. Patel et al., "Machine learning for targeted advance care planning in cancer patients: A quality improvement study," *J. Pain Symptom Manage.*, vol. 68, no. 6, pp. 539–547. e3, Dec. 2024, doi: 10.1016/j. jpainsymman.2024.08.036.
23. N. D. Agaronnik, J. Davis, C. R. Manz, J. A. Tulsky, and C. Lindvall, "Large language models to identify advance care planning in patients with advanced cancer," *J. Pain Symptom Manage.*, vol. 69, no. 3, pp. 243–250. e1, Nov. 2024, doi: 10.1016/j. jpainsymman.2024.11.016.
24. H. Qiu, A. Li, L. Ma, and Z. Lan, "PsyChat: A client-centric dialogue system for mental health support," in *Proc. 27th Int. Conf. Comput. Supported Cooperative Work Des. (CSCWD)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 2979–2984, doi: 10.1109/CSCWD61410.2024.10580641.
25. V. J. Nolan et al., "Incorporating patient values in large language model recommendations for surrogate and proxy decisions," *Crit Care Explorations*, vol. 6, no. 8, p. e1131, Aug. 2024, doi: 10.1097/cce.0000000000001131.
26. S. Tripathi, R. Sukumaran, and T. S. Cook, "Efficient healthcare with large language models: Optimizing clinical workflow and enhancing patient care," *J. Amer. Med. Inform. Assoc.*, vol. 31, no. 6, pp. 1436–1440, 2024, doi: 10.1093/jamia/ocad258.
27. V. L. Yu et al., "Evaluating the performance of a computer-based consultant," *Comput. Programs Biomed.*, vol. 9, no. 1, pp. 95–102, Jan. 1979, doi: 10.1016/0010-468x(79)90022-9.
28. H. Landi, "HIMSS24: How epic is building out AI, ambient technology for clinicians," Fierce Healthcare, Mar. 13, 2024. [Online]. Available: https://www.fiercehealthcare.com/ai-and-machine-learning/himss24-how-epic-building-out-ai-ambient-technology-clinicians#:~:text=When%20we%20 use%20the%20draft,that%20was%20 generated%20by%20AI.%22
29. K. J. Prabhod, "Integrating large language models for enhanced clinical decision support systems in modern healthcare," *J. Mach. Learn. Healthcare Decis.*

Support, vol. 3, no. 1, pp. 18–62, Jun. 2023. [Online]. Available: https://medlines.uk/index.php/JMLHDS/article/view/23.

30. R. Krishnamurthy, N. Mummudi, J. S. Goda, S. Chopra, B. Heijmen, and J. Swamidas, "Using artificial intelligence for optimization of the processes and resource utilization in radiotherapy," *JCO Global Oncol.*, vol. 8, Nov. 2022, Art. no. e2100393, doi: 10.1200/go.21.00393.
31. G. Demiris, D. Parker Oliver, and E. Wittenberg-Lyles, "Technologies to support end-of-life care," *Seminars Oncol. Nursing*, vol. 27, no. 3, pp. 211–217, Aug. 2011, doi: 10.1016/j.soncn.2011.04.006.
32. A. Suslow, C. Giehl, J. Hergesell, H. C. Vollmar, and I. Otte, "Impact of information and communication software on multiprofessional team collaboration in outpatient palliative care - A qualitative study on providers' perspectives," *BMC Palliative Care*, vol. 22, no. 1, Mar. 2023, Art. no. 19, doi: 10.1186/s12904-023-01141-4.
33. M. Martínez et al., "'Dignity therapy', a promising intervention in palliative care: A comprehensive systematic literature review," *Palliative Medicine*, vol. 31, no. 6, pp. 492–509, 2017, doi: 10.1177/0269216316665562.
34. Y. Lim, "Dignity and dignity therapy in end-of-life care," *J. Hospice Palliative Care*, vol. 26, no. 3, pp. 145–148, Sep. 2023, doi: 10.14475/jhpc.2023.26.3.145.
35. E. Wild, J. Weng, D. Pachman, C. Ingram, and M. Felt, "Feasibility and acceptability of virtual dignity therapy for palliative care patients with advanced cancer," *J. Pain Symptom Manage.*, vol. 67, no. 5, pp. e609–e610, 2024, doi: 10.1016/j.jpainsymman.2024.02.032.
36. M. Moor et al., "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, pp. 259–265, Apr. 2023, doi: 10.1038/s41586-023-05881-4.

ROY RADA is a professor emeritus at the Department of Information Systems, ITE Building, University of Maryland, Baltimore County, Baltimore, MD 21250 USA. Contact him at rada@umbc.edu.

EDITOR: Nir Kshetri, nbkshetr@uncg.edu

DEPARTMENT: IT ECONOMICS

This article originally appeared in IT Professional vol. 27, no. 1, 2025

Economics of Agentic AI in the Health-Care Industry

Nir Kshetri, University of North Carolina at Greensboro, Greensboro, NC, 27412, USA

This article explores the role of agentic AI in transforming health-care systems by automating workflows, enhancing collaboration, and improving patient outcomes. It discusses how agentic AI can address inefficiencies and fragmentation in global health care, leading to reduced costs and more personalized and higher quality care.

The global health-care sector is currently valued at $9 trillion, representing 11% of the world's gross domestic product.[1] It is argued that the traditional economic model of supply and demand—which involves buyers and sellers as the main participants, with buyers acting as informed judges of goods and services—faces challenges in the health-care market.[2] Prices serve as the key mechanism for coordinating market decisions in this framework. The "invisible hand" is believed to naturally lead to efficient resource allocation, assuming minimal interference. The health-care market differs significantly from the standard supply-and-demand model. Key differences include the involvement of third parties (insurers, governments), patients' inability to evaluate treatments, health-care providers being paid by insurers rather than patients, and insurers' rules determining resource allocation. These factors prevent the "invisible hand" from efficiently allocating resources, leading to potential inefficiencies in health-care delivery.[2]

As clinical practice becomes more collaborative, interdisciplinary teams work together to provide personalized care. However, the challenge arises in ensuring that all clinicians have real-time access to updated treatment plans and test results. While electronic health records (EHRs) are essential for business tasks like billing, they aren't designed to support the collaborative workflow needed in modern health care. This gap creates inefficiencies, as clinicians often lack the tools necessary for seamless communication and coordination (https://carealign.ai/what-is-care-orchestration/).

An upshot is that health-care delivery suffers from inefficiency, wasting billions annually due to unnecessary services, medication misuse, and overuse of emergency departments. Beyond clinical issues, inefficiencies also plague nonclinical processes, like scheduling, test result reporting, and prescription refills. Defined as using excessive resources to provide patient care, inefficiency often stems from unnecessary variations in both operational and clinical workflows, undermining overall system effectiveness.[3]

Agentic AI advancements prioritize efficiency by streamlining operations, automating complex tasks, optimizing decision making, and adapting to changing environments, addressing key challenges in industries like health care. Agentic AI-based solutions are being introduced by established medical technology companies, such as GE Healthcare, as well as by startup companies, such as Hippocratic AI, to enhance collaborative workflows in modern health care, addressing gaps that traditional medical IT systems like EHRs have struggled to fill. For instance, in response to the global nursing shortage, hospitals and clinics are increasingly adopting AI agents to perform nondiagnostic tasks, such as taking vital signs, scheduling, and maintaining patient records. Agentic AI health-care startup Hippocratic AI, for example, engages patients over the phone to manage tasks like preprocedure check-ins and posttreatment reminders, supporting health-care professionals while improving efficiency (https://www.fierceelectronics.com/ai/2025-mark-new-phase-agentic-ai). AI agents can thus help alleviate workload pressures in the health-care industry by reducing time burdens and enhancing doctor–patient

Digital Object Identifier 10.1109/MITP.2025.3529857
Date of current version 20 February 2025.

relationships through streamlined processes and improved efficiency.[4]

This article examines the challenges facing the global health-care sector, highlighting inefficiencies in care delivery and system fragmentation. It explores the potential of agentic AI to address these issues by enhancing collaboration, automating workflows, and improving patient outcomes while reducing health-care costs.

UNDERSTANDING AGENTIC AI AND ITS FUNCTIONALITY

In June 2024, Andrew Ng, founder of DeepLearning.AI, coined the term "agentic AI" to describe AI systems that autonomously handle complex tasks like data analysis and decision making with minimal or no human oversight.[5] Agentic AI uses independent AI agents (see "Exhibit 1: Key Steps in Building and Training AI Agents") that collaborate, reason, and solve complex problems, with large language models (LLMs) guiding decision making (https://www.weforum.org/stories/2024/12/agentic-ai-financial-services-autonomy-efficiency-and-inclusion/). These systems are capable of performing complex tasks to achieve specific goals and learning independently without human oversight.[6] Agentic AI can be viewed as the third wave of AI, which evolves from the foundations of predictive AI (first wave) and generative AI (second wave) (https://www.willowtreeapps.com/insights/agentic-ai-enhancing-workflows).

Generative AI tools, such as LLMs, excel at generating human-like text, enabling natural language interactions and creative tasks like content generation and summarization, while traditional programming provides structured, reliable solutions for precision and efficiency. Agentic AI combines these strengths, using LLMs for adaptive, flexible tasks and traditional programming for deterministic processes, like security and calculations. This hybrid approach allows agents to autonomously handle dynamic environments while ensuring critical functions remain precise and reliable, with various agent types working together to enhance overall performance (https://www.ibm.com/think/insights/agentic-ai).

Table 1 provides a comparison between generative AI and agentic AI, highlighting their key differences in terms of autonomy, decision making, adaptation, and environmental perception. Agentic AI systems leverage the creative capabilities of generative AI models like ChatGPT but differ in key aspects. They prioritize decision making and optimizing specific goals, such as sales, customer satisfaction, or supply-chain efficiency, rather than content creation. Unlike generative AI, they operate autonomously without human prompts and can independently carry out complex tasks, including searching databases or triggering workflows to complete activities (https://hbr.org/2024/12/what-is-agentic-ai-and-how-will-it-change-work).

Agentic AI operates through a four-step process to solve problems. First, it perceives by gathering and processing data from sources like sensors, databases, and digital interfaces, extracting relevant features and recognizing key entities. Next, it reasons using an LLM as a central engine to understand tasks, generate solutions, and coordinate specialized models, leveraging techniques like retrieval-augmented generation for accurate and relevant outputs. It then acts by integrating with external tools via application programming interfaces (APIs) to execute tasks, with built-in guardrails ensuring correct implementation, such as restricting authority for specific actions. Finally, it learns through a feedback loop, or "data flywheel," using interaction data to refine models and enhance performance, making it increasingly adaptive and effective over time.[4]

Having noted the above abilities, it is worth highlighting that while many organizations recognize the potential of autonomous agents, adoption remains in its early stages, with most demonstrations yet to reach full production readiness (https://www.coindesk.com/opinion/2024/11/19/ai-agents-can-help-crypto-become-the-currency-of-ai). AI agents greatly improve automation but don't fully replace it for complex tasks.[4] Moreover, while some systems exhibit agent-like behavior, defining truly "agentic" AI remains complex, as many technologies still rely on a mix of human input and machine actions.[5]

CHALLENGES IN THE CURRENT HEALTH-CARE SYSTEM

The health-care sector faces significant hurdles, including cognitive overload, fragmented care plans, and system inefficiencies, which collectively lead to delays and disjointed patient journeys.[7] Cognitive overload arises when an individual's working memory is overwhelmed by the need to process and manage more elements simultaneously than it can handle, hindering task performance.[8] The National Institutes of Health reports that medical knowledge doubles every 73 days, particularly in oncology, cardiology, and neurology. This rapid expansion creates a challenge for health-care professionals. For instance, an oncologist reviewing prostate-specific antigen (PSA) results faces the challenge of considering various factors, such as medications, therapies, procedures, symptoms, medical imaging, biopsies, and comorbidities, all spread across multiple documents and systems, often within a 15- to 30-min consultation window.[7]

Health system fragmentation involves the division of tasks and roles, like financing and service provision,

EXHIBIT 1: KEY STEPS IN BUILDING AND TRAINING AI AGENTS

Several key steps are involved in building and training AI agents.

First, it is crucial to identify the environment, tasks, and functions the AI agent will perform, such as within an app, website, or system, to ensure compatibility (https://www.uptech.team/blog/how-to-build-an-ai-agent). An AI agent, designed specifically for e-commerce brands, can deliver personalized, high-quality responses to address customer inquiries and resolve requests (https://docs.gorgias.com/en-US/articles/ai-agent-135134). It is also crucial to identify the target audience and their interaction needs. For example, medical AI agents must understand medical terminology (https://tinyurl.com/ydky3yyp).

Next, the development team will gather the data for training the AI agent (https://www.uptech.team/blog/how-to-build-an-ai-agent). Data must be accurate, relevant, and abundant data from internal sources, such as sales records and customer information, external sources like purchased or publicly available datasets, and user-generated content. Examples include social media posts and product reviews (https://www.uptech.team/blog/how-to-build-an-ai-agent), text transcripts from chat logs, support tickets, or e-mails, voice recordings for speech recognition, and interaction logs from similar systems to understand user behaviors and common queries (https://tinyurl.com/ydky3yyp).

The third step is selecting the appropriate tech stack and machine-learning model, based on the project's goals and deployment environment (https://tinyurl.com/ydky3yyp). This includes choosing programming languages, like Python or Java, and incorporating technologies, such as machine learning, natural language processing, computer vision, and robotic process automation (https://www.uptech.team/blog/how-to-build-an-ai-agent).

Step 4 involves designing the AI agent, which includes defining its architecture, data handling, and user experience. It is important to choose between modular or concurrent designs based on the agent's tasks, ensuring flexibility and efficiency. Additionally, focus on creating an intuitive user interface, incorporating accessibility features, and adding a feedback mechanism to improve the system.

Step 5 requires thoroughly testing and validating (https://tinyurl.com/ydky3yyp) the AI agent to identify glitches, biases, or unexpected behaviors. Key tests include unit, integration, functional, and usability testing, as well as optional edge case testing to evaluate how the agent handles extreme inputs. These tests ensure the agent operates effectively and provides a seamless user experience (https://www.uptech.team/blog/how-to-build-an-ai-agent).

The final step is to integrate the AI agent with existing systems and workflows, ensuring proper security measures for sensitive data. Regular monitoring is crucial to track key metrics like accuracy, response times, and resource usage. Additionally, gathering user feedback helps identify areas for improvement and ensures optimal performance (https://www.uptech.team/blog/how-to-build-an-ai-agent). It is also crucial to regularly monitor the AI agent's performance, checking query understanding, handling of complex conversations, and using tools for real-time insights on response times, success rates, and user satisfaction (https://tinyurl.com/ydky3yyp).

among different entities without sufficient coordination, often resulting in disconnected care and suboptimal outcomes.[9] These challenges in health care can be illustrated as a fragmented workflow, where multiple manual processes delay patient care. For instance, an oncologist manually checks PSA test results and requests additional tests without automation, leading to potential delays. Similarly, nurses and other medical professionals manually send requests for imaging or surgeries, and crucial patient information, like medical history, isn't automatically communicated, causing delays and risks. Each manual step adds time and complexity, reducing efficiency and potentially compromising patient safety.[7] These challenges hinder effective care orchestration, which relies on connecting individuals, data, systems, and workflows through technology to enhance coordination, foster collaboration, and improve health-care outcomes (https://carealign.ai/what-is-care-orchestration/).

The absence of collaborative workflow support in EHRs leads to inefficiencies for clinicians, stemming from inadequate communication and coordination tools. Clinicians often rely on workarounds, like paper or unsecured methods, to compensate for gaps in EHRs. These solutions can lead to inefficiencies, such as excessive time spent transcribing notes and a lack of real-time updates on patient conditions. In addition to wasting time, these workarounds can potentially harm patients by increasing the risk of errors and reducing the quality of care (https://carealign.ai/what-is-care-orchestration/).

These factors hinder the delivery of effective health-care services, resulting in poor health outcomes. For instance, delayed or missed cancer care paperwork increases the likelihood of patients postponing care or failing to follow through with appointments, tests, and prescriptions by nearly 50% (https://ldi.upenn.edu/our-work/research-updates/medical-red-tape-raises-the-odds-for-delayed-or-missed-cancer-care/). A Solutions Reach study found that cancer patients have a 25% rate of missed care, creating scheduling issues and backlogs, making it difficult to prioritize high-risk patients.

AGENTIC AI'S ROLE IN MAXIMIZING PATIENT VALUE AND ADVANCING HEALTH SYSTEM GOALS

To maximize patient value, health-care delivery should prioritize achieving high health outcomes relative to costs, focusing on efficiency and effectiveness in resource use.[10] These align with the World Health Organization's core objectives for health systems worldwide, which emphasize good health, responsiveness, and financial fairness as key to achieving optimal healthcare systems.[11]

Achieving Superior Health Outcomes and Fostering Overall Well-Being

Agentic AI can help achieve superior health outcomes by overcoming the various challenges and barriers discussed, such as inefficiencies in care delivery, data integration issues, and the complexity of decision making in health-care settings. AI agents can assist doctors by analyzing extensive medical and patient data, extracting critical insights to support better-informed care decisions.[4] Returning to the earlier example, an oncologist managing a patient with progressive prostate cancer can leverage agentic AI systems to optimize treatment planning by analyzing data, detecting abnormalities, and recommending additional tests. By automating scheduling and referrals, and providing real-time updates, the system ensures seamless communication among team members. This reduces manual workload, minimizes errors, and enables the oncologist to focus on decision making and patient care, improving efficiency and outcomes.[7]

In multiagent systems, AI agents collaborate to achieve complex objectives by distributing tasks among specialized subagents. These subagents

TABLE 1. Comparison of generative AI versus agentic AI (https://tinyurl.com/2s4va447).

	Generative AI	**Agentic AI**
Autonomy	Limited: Requires external prompts to produce responses and cannot function without human guidance.	High: Acts independently, handling tasks and making decisions without needing constant human input.
Behavior	Task-oriented and reactive: Responds to prompts to generate content like text or images, but lacks long-term goals and doesn't pursue overarching objectives, completing each task based solely on immediate input.	Goal-directed: Operates with a specific goal in mind, actively working towards it by taking purposeful actions, such as a self-driving car that makes decisions like steering and braking to ensure safe arrival at its destination.
Adaptation and learning	No: Operates based on learned patterns from training data and does not adapt or improve in real time without retraining using new data.	Yes: Learns from its experiences, adapting and improving over time, like a movie recommendation system that refines its suggestions based on user preferences.
Decision making	Basic: Makes decisions by selecting outputs based on learned patterns, such as predicting the next word in a sentence, but does not evaluate multiple alternatives or consider long-term consequences.	Complex: Evaluates multiple options and outcomes before making decisions, such as in stock-trading algorithms, where it analyzes data, predicts trends, and determines whether to buy or sell based on that information.
Environmental perception	No: Lacks environmental perception, working solely with data like text or images, without the ability to sense or interpret the physical world around it, reacting only to the input provided.	Yes: Makes decisions based on its understanding of the environment, using sensors or data inputs, such as cameras to detect obstacles, enabling it to navigate and adapt accordingly.

focus on specific workflow components, improving efficiency and coordination. Multiagent systems enhance problem-solving by enabling agents to interact, share information, and collectively address challenges beyond the capability of a single agent. This structure promotes scalability and adaptability, with each agent contributing to the system's overall functionality, ensuring more effective task completion through teamwork and specialization (https://www.multimodal.dev/post/ai-agentic-workflows). In the oncologist example above, when new clinical data (e.g., PSA levels, magnetic resonance imaging results, biopsy reports) is entered into the electronic medical record, a coordinating agent triggers workflows based on predefined rules. Specialized subagents, such as clinical data specialists, molecular test data agents, and biochemical data specialists, analyze various aspects of the data (e.g., clinical notes, genomic information, biochemical levels, imaging) to assess cancer aggressiveness, personalize treatment, and support decision making. These agents work together as part of a "virtual tumor board" to provide a comprehensive treatment approach.[7] The agents function autonomously within their defined scopes, retrieving additional data through APIs to assess disease progression. They collaborate to synthesize evaluations, which are processed by a coordinating agent. This agent triggers a recommendation generator that applies clinical decision support algorithms to propose treatment options. These recommendations are securely stored in the electronic medical record, with the oncologist notified for further action.[7]

With agentic AI, the challenges discussed earlier, such as delayed or missed cancer care paperwork leading to patients postponing care or failing to follow through with appointments and tests, can thus be minimized, ultimately reducing missed care for cancer patients. Additionally, AI agents offer 24/7 support to patients by providing information on medication usage, managing appointment schedules, and sending reminders to enhance adherence to treatment plans.[4] They also play a key role in remote patient monitoring, continuously analyzing data to deliver timely medical interventions, enhancing health-care delivery and improving patient outcomes through continuous care and quick responses to health changes.[6]

Due to their adaptability, emotional intelligence, and empathy, agentic AI systems excel in nonroutine, soft-skill roles, such as health care and caregiving. For instance, Hippocratic AI has developed AI agents for health care and social care, such as Sarah, who assists with assisted living, and Judy, who supports patients with preoperative tasks, including medication reminders and fasting advice (https://hbr.org/2024/12/what-is-agentic-ai-and-how-will-it-change-work).

Promoting Financial Fairness and Minimizing Costs

Agentic AI has the potential to transform health care by automating workflows, including care coordination, treatment planning,[6] and administrative tasks, like note-taking and claims processing.[4] This boosts cost efficiency, optimizes resource allocation, and enhances patient value through streamlined, autonomous systems.[6] In claims processing, AI agents autonomously review claims, verify documentation, and resolve discrepancies, reducing approval times by 30%.[4] Additionally, they analyze eligibility, identify bottlenecks, and expedite prior authorizations, decreasing manual review times by 40%.[4]

Agentic AI is thus likely to streamline processes, reduce approval times, and ultimately lower health-care costs by autonomously handling tasks, such as claims review, documentation verification, and discrepancy resolution. For instance, as of 2024, Hippocratic AI was reported to be developing AI health-care agents to perform nondiagnostic tasks typically handled by nurses, such as chronic care management and wellness coaching. These AI agents are more cost-effective, priced at $9/h compared to the $60.26/h for a registered nurse in California. The company partnered with NVIDIA to improve empathy inference for patient interactions and launched a staffing marketplace for AI agents in health care. It was also undergoing safety testing with health-care professionals and collaborating with over 40 provider groups.[5]

CONCLUSION

The health-care sector faces significant inefficiencies due to fragmentation, cognitive overload, and outdated IT systems, which hinder the delivery of high-quality care. Agentic AI has the potential to address these challenges by streamlining operations, enhancing decision making and improving collaboration among health-care professionals. By automating administrative tasks and providing real-time updates, AI can reduce errors, minimize delays, and optimize resource allocation, leading to better health outcomes. Ultimately, the adoption of agentic AI promises to promote financial fairness and maximize patient value, transforming the health-care system into a more efficient and effective model.

REFERENCES

1. Euronews, "How AI and Robotics are transforming the $9 trillion healthcare sector," *Euronews*, Jul. 17, 2024. [Online]. Available: https://www.euronews.com/business/2024/07/17/how-ai-and-robotics-are-transforming-the-9-trillion-healthcare-sector
2. N. G. Mankiw, "The economics of healthcare," Harvard University, Cambridge, MA, USA, 2017. Accessed: Dec. 29, 2024. [Online]. Available: https://scholar.harvard.edu/files/mankiw/files/economics_of_healthcare.pdf
3. K. A. James, S. E. Ross, B. Vance, R. Nath, M. I. Harrison, and D. R. West, "Inefficiency in primary care: Common causes and potential solutions," *Family Pract. Manage.*, vol. 22, no. 2, pp. 18–22, Mar./Apr. 2015.
4. E. Pounds, "What is agentic AI?" NVIDIA Blog, Oct. 22, 2024. [Online]. Available: https://blogs.nvidia.com/blog/what-is-agentic-ai/
5. A. Mathews, "Agentic AI is here and these 8 startups are leading the way in 2024," *AIM Research*, Sep. 10, 2024. Accessed: Dec. 28, 2024. [Online]. Available: https://aimresearch.co/ai-startups/agentic-ai-is-here-and-these-8-startups-are-leading-the-way-in-2024
6. "Gartner's top 10 tech trends of 2025: Agentic AI and beyond," *Productive Edge*, Nov. 19, 2024. [Online]. Available: https://www.productiveedge.com/blog/gartners-top-10-tech-trends-of-2025-agentic-ai-and-beyond
7. D. Sheeran and T. Kass-Hout, "How agentic AI systems can solve the three most pressing problems in healthcare today," *AWS Blog*, Dec. 4, 2024. [Online]. Available: https://aws.amazon.com/blogs/industries/how-agentic-ai-systems-can-solve-the-three-most-pressing-problems-in-healthcare-today/
8. M. Iskander, "Burnout, cognitive overload, and metacognition in medicine," *Med. Sci. Educator*, vol. 29, no. 2, pp. 325–328, 2019, doi: 10.1007/s40670-018-00654-5.
9. T. Bossert, N. Blanchet, S. Sheetz, D. Pinto, J. Cali, and R. P. Cuevas, "Comparative review of health system integration in selected countries in Latin America," Inter-American Development Bank, Washington, DC, USA, Tech. Note no. 585, 2014. [Online]. Available: https://publications.iadb.org/en/publication/11898/comparative-review-health-system-integration-selected-countries-latin-america
10. M. E. Porter, "What is value in health care?" *N. Engl. J. Med.*, vol. 363, no. 26, pp. 2477–2481, Dec. 2010, doi: 10.1056/NEJMp1011024.
11. World Health Organization, *The World Health Report 2000: Health Systems: Improving Performance*. Geneva, Switzerland: WHO, 2000.

NIR KSHETRI is a professor at the Bryan School of Business and Economics, the University of North Carolina at Greensboro, Greensboro, NC, 27412, USA, an *IT Professional* associate editor in chief, and IT Economics editor. Contact him at nbkshetr@uncg.edu.

EDITOR: Joanna F. DeFranco, The Pennsylvania State University, jfd104@psu.edu

DEPARTMENT: INTERNET OF THINGS

This article originally appeared in Computer vol. 57, no. 10, 2024

Health Care 4.0 and Clinical Internet of Things

Joanna F. DeFranco, *The Pennsylvania State University*

This article highlights the success factors of Health Care 4.0 as well as the benefits and challenges of the latest paradigm under the Internet of Medical Things: the Clinical Internet of Things.

Just as Industry 4.0 describes the transformation of manufacturing with a wide range of modern technologies, such as a "smart factory," Health Care 4.0 refers to state-of-the-art technologies that enhance patient outcomes and optimize health-care operations with real-time/data-driven insights.[1] The progression from Health Care 1.0 to 4.0 can be described as[2]:

- Health Care 1.0: patient/provider encounter
- Health Care 2.0: use technology to diagnosis, treat and monitor
- Health Care 3.0: electronic health records and integration with other units within a health-care system
- Health Care 4.0: smart, connected, and personalized treatment using data analytics and artificial intelligence (AI).

Technology paradigms are a great way to frame understanding with a set of practices. Paradigms shift with scientific discovery, rapid innovation, and advancement, especially when advancement has a social impact that influences societal behavior. With new medical paradigms, it can be crucial to differentiate between related terminology in that they encompass different aspects of the technology's implications and benefits.

Technologies, such as AI, Internet of Things (IoT), data analytics, robotics, 3D printing, and 5G cellular systems are leveraged in Health Care 4.0. Embracing Health Care 4.0 has also empowered patients by enabling management of chronic conditions.

This article will focus on the new Health Care 4.0 and one of its newest components: Clinical IoT. Clinical IoT is one of the dominant Health Care 4.0 components that is used along with the more familiar telemedicine and Internet of Medical Things (IoMT) components. Clinical IoT, telemedicine, and IoMT all fit into Health Care 4.0 to create a synergistic effect but also differ in their meaning, as shown in Table 1.

In general, these are interconnected concepts that differ in scope. The differences are that telemedicine focuses on remote delivery (that is, remote diagnosis and treatment), IoMT is a broad term related to the connection of medical devices and applications (for example, smart medical devices), and clinical IoT is a subset of IoMT focusing on enhancing patient care and clinical workflow in a clinical setting (for example, smart devices in a hospital setting).

These modern health-care technologies work together to transform and improve patient care and overall operational efficiencies. Clinical IoT is the most recent of the three. This term became more prevalent as the shift in the health-care industry recognized and adopted IoT technologies in clinical settings.

DISCLAIMER

The author is completely responsible for the content in this message. The opinions expressed here are her own.

Digital Object Identifier 10.1109/MC.2024.3426568
Date of current version: 23 September 2024

TABLE 1. Health Care 4.0 components meaning.

	Definition	Application	Examples
Telemedicine	Remotely delivering patient care[3]	Using technology to consult and treat patients	Robotic surgery, virtual consultations, remote monitoring
IoMT	A network of connected medical devices and applications that collect, exchange, and analyze health data[4]	Remote patient monitoring applications, smart medical devices, connected imaging systems, and health data analytics	Wearable monitors, closed loop insulin delivery system
Clinical IoT	A subset of IoMT – focused on the challenges and opportunities of integrating IoT technologies in clinical environments	Enhances clinical decision making, patient safety, and operational efficiency	Smart hospital beds that monitor patient movements and adjust for comfort and safety; IoT-enabled infusion pumps that ensure accurate medication delivery

TABLE 2. Technology advancement driving clinical IoT adoption.

Technology	Advancement	Application example
Sensors	Accuracy, decreased size, battery life, multiparameter sensor	Wearable sensors (that is, continuous glucose monitor), multiparameter monitor (for example, ECG, body temperature, respiratory rate, SpO_2, blood pressure), smart bed (for example, monitors patients and administer medicine based on recorded data)
Connectivity	5G networks (faster, more reliable data transmission)	Improve operational workflow environmental sensors feeding into a clinical support tool to monitor patient conditions and improve operational efficiency
Interoperability	Integration of IoMT devices with EHR	Wearable heart monitor: transmits data in real time to the health-care provider through integration with an EHR system
Data analytics and AI	Algorithm improvement	Closed-loop insulin pump uses real-time data processing improving medical decision making for a patient (for example, personalized treatment, and patient can adjust therapy in real time)

ECG: electrocardiogram; EHR: electronic health records; SpO_2: oxygen saturation.

HOW HAS CLINICAL IOT EVOLVED IN THE PAST FIVE YEARS?

Driven by technological advancements and increased adoptions, clinical IoT devices have evolved rapidly in the past five years. These devices have made a major impact on health-care delivery, making treatment more personalized and accessible. Some of the improvements in technology are shown in Table 2.

WHAT ARE THE CHALLENGES OF CLINICAL IOT?

The benefits, as shown in Table 2, are tremendous. However, this comes with several significant challenges. Some of these challenges (Table 3) span across technical, regulatory, ethical, and practical domains.

MOVING FORWARD

Clinical IoT is transforming health care but addressing these challenges is crucial for its successful adoption and implementation. The domains that are prime for cyberattacks include technology, education, and manufacturing and IoMT. However, IoMT (for example, and clinical IoT) is the most problematic.[6]

There are efforts to address some of these issues. For example, scientists at the National Institute of Standards and Technology have developed a lightweight,

TABLE 3. Clinical IoT challenges.

Challenge	Data security	Data privacy	Interoperability	Data overload	Connectivity
Examples	Ensuring the security of the sensitive data collected by IoT devices	Protecting confidentiality and privacy of patients at critical checkpoints (data are compiled and transferred)[5]; compliance with regulations: HIPAA (hiding sensitive diagnosis) and GDPR (right to erasure)	Systems using different standards and protocols challenging communication and integration	Need efficient data management to derive meaningful insights from the vast amount of data collected	Reliable network infrastructure is essential; issues in rural areas, dealing with outages, or disruptions

GDPR: General Data Protection Regulation; HIPPA: Health Insurance Portability and Accountability Act.

efficient, and flexible infrastructure solution for secure data sharing between organizations down at a granular level, promoting data security and privacy as well as system interoperability for the clinical data.[7] In addition, the IEEE recently published the Standard for Clinical Internet of Things (IoT) Data and Device Interoperability with TIPPSS (trust, identity, privacy, protection, safety, security) principles offering a comprehensive framework for designing secure, interoperable devices that protect individuals and maintain data integrity.[8] Advancements in technology will continue but the advancements need to address regulations and the quality requirements (for example, security, privacy, interoperability, etc.), as they play key role in overcoming the challenges to realize the full potential of clinical IoT.

There are also efforts to address the adoption of a patient-centric approach to increase the success of Health Care 4.0. A recent systematic literature review published by the National Institutes of Health highlighted 10 critical success factors that would influence the successful implementation of Health Care 4.0.[2] The results of the study highlighted factors that support culture, leadership, and employee skills required to move forward:

- digital integration and interconnectedness of the health care
- human-centric automation of health-care providers
- improve patient-centricity and patient experience
- use big data and analytics
- managing digital health-care supply chains
- strategies for promoting Health Care 4.0
- promote a culture for Health Care 4.0
- health-care leadership
- health-care employees' skills
- adoption of new business models.

Addressing both the technical challenges and implementing the critical success factors provide a path forward to facilitate an effective implementation of Health Care 4.0 and all of its components. This will enable society to take full advantage of smart, connected, and personalized treatment with minimal risk.

REFERENCES

1. M. Osama et al., "Internet of medical things and healthcare 4.0: Trends, requirements, challenges, and research directions," *Sensors (Basel)*, vol. 23, no. 17, Aug. 2023, Art. no. 7435, doi: 10.3390/s23177435.
2. M. Sony, J. Antony, and G. L. Tortorella, "Critical success factors for successful implementation of healthcare 4.0: A literature review and future research agenda," *Int. J. Environ. Res. Public Health*, vol. 20, no. 5, Mar. 2023, Art. no. 4669, doi: 10.3390/ijerph20054669.
3. J. DeFranco and M. Metro, "Internet of telemedicine," *Computer*, vol. 55, no. 4, pp. 56–59, Apr. 2022, doi: 10.1109/MC.2022.3143625.
4. H. Huang, J. Wang, S. Wang, and Y. Zhang, "Internet of medical things: A systematic review," *Neurocomputing*, vol. 557, Nov. 2023, Art. no. 126719, doi: 10.1016/j.neucom.2023.126719.
5. C. Dinh-Le, R. Chuang, S. Chokshi, and D. Mann, "Wearable health technology and electronic health record integration: Scoping review and future directions," *JMIR Mhealth Uhealth*, vol. 7, no. 9, Sep. 2019, Art. no. e12861, doi: 10.2196/12861.
6. T. Seals. "A look at the riskiest connected devices of 2024." Dark Reading. [Online]. Available: https://

www.darkreading.com/cyber-risk/riskiest-connected-devices-2024

7. J. DeFranco, J. Roberts, D. Ferraiolo, and D. C. Compton, "An infrastructure for secure data sharing: A clinical data implementation," *JAMIA Open*, vol. 7, no. 2, Jul. 2024, Art. no. ooae040, doi: 10.1093/jamiaopen/ooae040.
8. *IEEE Approved Draft Standard for Clinical Internet of Things (IoT) Data and Device Interoperability with TIPPSS - Trust, Identity, Privacy, Protection, Safety, Security*, IEEE SA - IEEE Standard 2933-2024.

JOANNA F. DEFRANCO is an associate professor of software engineering and the associate director of the Doctor of Engineering in Engineering Program at The Pennsylvania State University, University Park, PA 16802 USA, and an associate editor in chief of *Computer*. Contact her at jfd104@psu.edu.

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.

DECEMBER

1 December

- BDCAT (IEEE/ACM International Conf. on Big Data Computing, Applications, and Technologies), Nantes, France
- UCC (IEEE/ACM Int'l Conf. on Utility and Cloud Computing), Nantes, France

2 December

- APSEC (Asia-Pacific Software Eng. Conf.), Taipa, Macao
- RTSS (IEEE Real-Time Systems Symposium), Boston, USA

3 December

- SEC (IEEE/ACM Symposium on Edge Computing), Arlington, USA

4 December

- ICAIIHI (Int'l Conf. on Artificial Intelligence for Innovations in Healthcare Industries), Raipur, India

5 December

- ICA (IEEE Int'l Conf. on Agentic AI), Wuhan, China

8 December

- ACSAC (Annual Computer Security Applications Conf.), Honolulu, USA
- BigData (IEEE Int'l Conf. on Big Data), Macau, China

13 December

- iSES (IEEE Int'l Symposium on Smart Electronic Systems), Jaipur, India

14 December

- FOCS (IEEE Annual Symposium on Foundations of Computer Science), Sydney, Australia
- ICPADS (IEEE Int'l Conf. on Parallel and Distributed Systems), Hefei, China

15 December

- BIBM (IEEE Int'l Conf. on Bioinformatics and Biomedicine), Wuhan, China
- MCSoC (IEEE Int'l Symposium on Embedded Multicore/Many-core Systems-on-Chip), Singapore

17 December

- HiPC (IEEE Int'l Conf. on High Performance Computing, Data, and Analytics), Hyderabad, India

18 December

- ESAI (Int'l Conf. on Embedded Systems and Artificial Intelligence), Fez, Morocco

19 December

- ICVRV (Int'l Conf. on Virtual Reality and Visualization), Bogota, Colombia

2026

JANUARY

3 January

- VLSID (Int'l Conf. on VLSI Design & Int'l Conf. on Embedded Systems), Pune, India

14 January

- ICOIN (Int'l Conf. on Information Networking), Hanoi, Vietnam

26 January

- AIxVR (IEEE Int'l Conf. on Artificial Intelligence and eXtended and Virtual Reality), Osaka, Japan

31 January

- HPCA (IEEE Int'l Symposium on High Performance Computer Architecture), Sydney, Australia

FEBRUARY

2 February

- AIxDKE (Int'l Conf. on AI x Data and Knowledge Eng.), Laguna Hills, USA
- BigComp (IEEE Int'l Conf. on Big Data and Smart Computing), Guangzhou, China
- ICSC (Int'l Conf. on Semantic Computing), Laguna Hills, USA

16 February

- ICNC (Int'l Conf. on Computing, Networking and Communications), Maui, USA

MARCH

6 March

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

16 March
- PerCom (IEEE Int'l Conf. on Pervasive Computing and Communications), Pisa, Italy

17 March
- SANER (IEEE Int'l Conf. on Software Analysis, Evolution and Reengineering), Limassol, Cyprus
- SSIAI (IEEE Southwest Symposium on Image Analysis and Interpretation), Santa Fe, USA

20 March
- 3DV (Int'l Conf. on 3D Vision), Vancouver, Canada

21 March
- VR (IEEE Conf. on Virtual Reality and 3D User Interfaces), Daegu, Korea

22 March
- SSIAI (IEEE Southwest Symposium on Image Analysis and Interpretation), Santa Fe, USA

23 March
- SaTML (IEEE Conf. on Secure and Trustworthy Machine Learning), Munich, Germany

APRIL

12 April
- AST (IEEE/ACM Int'l Conf. on Automation of Software Test), Rio de Janeiro, Brazil
- FormaliSE (IEEE/ACM Int'l Conf. on Formal Methods in Software Eng.), Rio de Janeiro, Brazil
- ICSE (IEEE/ACM Int'l Conf. on Software Eng.), Rio de Janeiro, Brazil
- MOBILESoft (IEEE/ACM Int'l Conf. on Mobile Software Eng. and Systems), Rio de Janeiro, Brazil
- MSR (IEEE/ACM Int'l Conf. on Mining Software Repositories), Rio de Janeiro, Brazil

15 April
- COOL CHIPS (IEEE Symposium on Low-Power and High-Speed Chips and Systems), Tokyo, Japan

20 April
- PacificVis (IEEE Pacific Visualization Conf.), Sydney, Australia

MAY

4 May
- HOST (IEEE Int'l Symposium on Hardware Oriented Security and Trust), Washington, DC, USA

8 May
- BigDataSecurity (IEEE Conf. on Big Data Security on Cloud), New York City, USA
- CAI (IEEE Int'l Conf. on Artificial Intelligence), Granada, Spain
- HPSC (IEEE Int'l Conf. on High Performance and Smart Computing), New York City, USA
- IDS (IEEE Int'l Conf. on Intelligent Data and Security), New York City, USA
- SmartCloud (IEEE Int'l Conf. on Smart Cloud), New York City, USA

11 May
- SenSys (ACM/IEEE Int'l Conf. on Embedded Artificial Intelligence and Sensing Systems), Saint Malo, France

12 May
- RTAS (IEEE Real-Time and Embedded Technology and Applications Symposium), Saint Malo, France

13 May
- FCCM (IEEE Annual Int'l Symposium on Field-Programmable Custom Computing Machines), Atlanta, USA

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

19 May
- ICDE (IEEE Int'l Conf. on Data Eng.), Hong Kong, China