Creating intelligent cyberinfrastructure for democratizing AI

Dhabaleswar K. Panda[1] | **Vipin Chaudhary[2]** | **Eric Fosler-Lussier[1]** | **Raghu Machiraju[1]** | **Amit Majumdar[3]** | **Beth Plale[4]** | **Rajiv Ramnath[1]** | **Ponnuswamy Sadayappan[5]** | **Neelima Savardekar[1]** | **Karen Tomko[6]**

[1]The Ohio State University, Columbus, Ohio, USA

[2]Case Western Reserve University, Cleveland, Ohio, USA

[3]San Diego Supercomputer Center, Diego, California, USA

[4]Indiana University, Bloomington, Indiana, USA

[5]University of Utah, Salt Lake City, Utah, USA

[6]Ohio Supercomputer Center, Columbus, Ohio, USA

Correspondence
Dhabaleswar K. Panda, The Ohio State University, Columbus, OH, USA.
Email: panda@cse.ohio-state.edu

Abstract

Artificial intelligence (AI) has the potential for vast societal and economic gain; yet applications are developed in a largely ad hoc manner, lacking coherent, standardized, modular, and reusable infrastructures. The NSF-funded Intelligent CyberInfrastructure with Computational Learning in the Environment AI Institute ("ICICLE") aims to fundamentally advance *edge-to-center*, AI-as-a-Service, achieved through intelligent cyberinfrastructure (CI) that spans the edge-cloud-HPC *computing continuum*, *plug-and-play* next-generation AI and intelligent CI services, and a commitment to design for broad accessibility and widespread benefit. This design is foundational to the institute's commitment to democratizing AI. The institute's CI activities are informed by three high-impact domains: *animal ecology*, *digital agriculture*, and *smart foodsheds*. The institute's workforce development and broadening participation in computing efforts reinforce the institute's commitment to *democratizing AI*. ICICLE seeks to serve as *the national nexus for AI and intelligent CI*, and welcomes engagement across its wide set of programs.

INTRODUCING THE ICICLE AI INSTITUTE

Institute rationale: Artificial intelligence (AI) is transforming every sector of society, from transportation and autonomous vehicles to biodiversity and wildlife conservation to food production and smart foodsheds. It will be the driving force for the next information revolution. However, there is a massive gap between available AI techniques and end-user availability to various application domains. The ad hoc development of many AI applications lack coherent, standardized, modular, and reusable infrastructures.

Successful AI solutions for one use case rarely directly generalize to other use cases, or even the same use case with a slightly different context.

Infrastructure implementation necessitates proper AI abstractions. Concomitant with generalizability challenges, as the cyberinfrastructure (CI) grows increasingly complex, end users face bewildering choices when purposing AI toward insightful analytics, modeling complex systems, or developing automation. In environment-focused settings, scientists studying natural/managed ecosystems and resource flows necessarily deal with complex dynamics represented by a wide array of public and private data
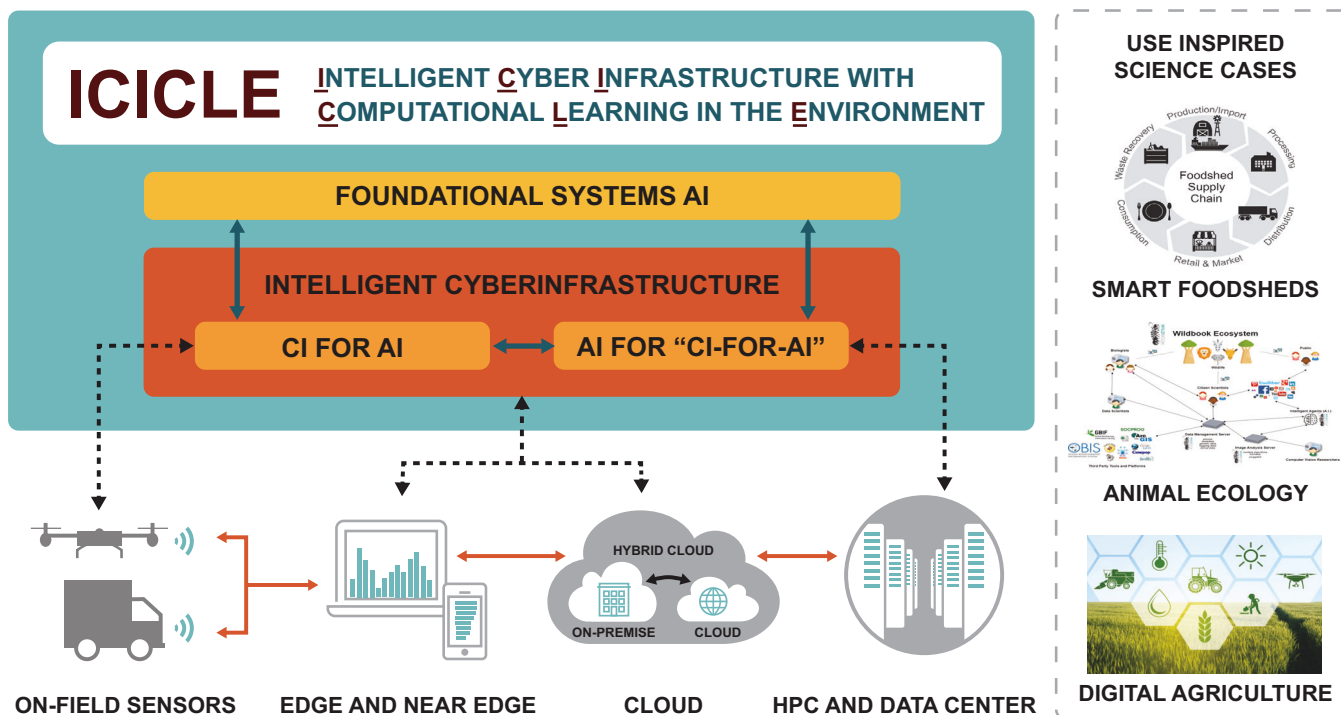
FIGURE 1   Overview and scope of the ICICLE project.

sources. Efficiently using AI infrastructures requires navigating model choice, cloud-or-local compute decisions, privacy and data protection, and heterogeneous architectures.

Addressing the challenge: The **ICICLE** (*Intelligent CyberInfrastructure with Computational Learning in the Environment*) AI Institute (ICICLE 2023) addresses these challenges by building the first and foremost *edge-to-center* AI-as-a-service enterprise, advancing foundational AI research and next-gen CI for AI *(CI4AI)*, while using AI techniques in building this CI *(AI4CI)*. This CI supports AI democratization, extending the accessibility and benefit of AI to a wider population (Figure 1) and across the computing continuum—from traditional large-scale HPC systems and data centers to newer activity in edge- and near-edge devices as well as cloud resources.

Development and deployment of this next-gen CI is driven by *use-inspired research* in three high-impact domains: *animal ecology*, *digital agriculture*, and *smart foodsheds*. ICICLE's design extracts both common and differentiated workflows across use cases, inspiring new CI for AI and enabling new discoveries in these domains. End users and AI application developers must be able to keep data access appropriately authenticated or private, and choose where computation occurs (e.g., data centers, the cloud, or edge devices). Thus, the "E" in ICICLE— environment—reflects two interacting themes within the institute: end use cases from real-world environments utilizing AI and in expanding computational environments.

The team: Realizing ICICLE's vision requires a highly cohesive and interdisciplinary team of scientists, engineers, and practitioners from thirteen universities and research institutes. Team expertise spans CI (high-performance computing, networking), AI (statistical machine learning, computer vision, knowledge graphs, and conversational AI), data privacy and trust, visualization, and use-inspired science in animal ecology, digital agriculture, and smart foodsheds. Additionally, ICICLE team members have expertise in the areas of broadening participation, workforce development (WFD), knowledge transfer, and collaboration. Under the umbrella of the ICICLE Institute, these diverse, individually talented, and individually focused researchers have formed a cohesive team capable of taking up the grand challenge of solving our proposed *plug-and-play* AI vision by tackling cross-cutting problems in CI, AI, and use-inspired science.

Building connections: The diverse expertise in ICICLE is being harnessed to grow the next generation of scientists, scholars, and workers—from kindergarten through professionals—using an inclusive plan that is synergistic with NSF's overall WFD portfolio. ICICLE's broadening participation in computing (BPC) plan leverages both the foundations of the NSF CISE BPC program and our own institutional resources to impact WFD, particularly reaching out to women, historically disadvantaged persons, and persons with disabilities.

As a national scale AI Institute, ICICLE also leverages NSF-funded infrastructure including *Tapis* from the Texas

Advanced Computing Center, *Jetstream* from Indiana University, and *Voyager* from the San Diego Supercomputer Center, thus bringing together multiple organizations to synergistically work together as a *nexus for this collaborative effort*.

VISION AND RESEARCH DIRECTIONS

ICICLE addresses a crucial reality: today, applying AI techniques to any specific use case is a nontrivial task for AI experts, let alone for users whose expertise lies in the domains outside AI. The vision of the ICICLE Institute is to empower users to seamlessly *plug-and-play* AI models in any given local computing environment across multiple use cases.

ICICLE aims to provide conversational and visualization-powered user interfaces where users are able to ask questions about the essential underlying context, available computational resources, and other information in a natural language much like speaking to agents such as Siri and Alexa. Next, ICICLE includes a novel realization of an AI model commons, that is, an ecosystem for producing, profiling, sharing, and distributing AI models. The commons will use existing models, exploit available unified semantic data representation provided by expansive knowledge graphs to create new models, and direct users to models most relevant to their needs. Finally, ICICLE is environment-aware: it manages AI holistically across the computing continuum from the edge and on-premise devices used for data collection to model training, curation, and storage at large HPC/cloud centers. Recognizing AI's inherent contextual sensitivity, adaptivity in ICICLE is an essential property for AI deployment with innovations for automated or human-in-the-loop adaptation to different application contexts (including tasks, environments, and user preferences) at the edge.

Research directions: All these foundational systems AI innovations will be powered by an intelligent CI that will provide high-performance model training, data management, edge intelligence, and wireless control and coordination for the computing continuum, which is in turn continuously improved by advanced AI techniques. Throughout, ICICLE includes cross-cutting design considerations to manage data privacy, accountability, and integrity, as well as visual analytics for explainability.

The institute itself provides an opportunity to study *AI ethics* because of its unique multimodal orientation on CI for AI, AI for CI, and on the democratization of AI. Through dialog, thought experiments, and a mapping exercise we asked "Where could issues of AI ethics arise in AI infused CI infrastructure for AI?" Through this exercise,

we categorized concerns as falling into six areas: democratization, fairness, accountability, trustworthiness, privacy, and the unknowns that arise from the use of the CI, (manuscript under review). In a step then towards operationalizing the project's statement of ethics, we developed a AI ethics framework. The framework, a living document, proposes a collection of guiding principles that are tailored to render ICICLE as a responsibly designed and managed CI. The guiding principles are a work in progress; they, however, must be complemented by outreach and WFD effort focused on the unique challenges of AI ethics in research infrastructure.

To empower plug-and-play AI, ICICLE must provide abstraction layers to shield users and AI and CI developers from the pervasive heterogeneities in the full AI development and application life-cycle. Such successful abstractions require the following advances in foundational systems AI: *knowledge graphs* as the knowledge backbone to provide a semantically rich abstraction for addressing data heterogeneity; a *model commons* to provide the abstraction for addressing model heterogeneity and improve discoverability, matchability, and interoperability of AI models; *adaptive AI* to enable in situ adaptation of AI models at the edge; *federated learning* to support edge-to-center across the computing continuum, decentralized, and privacy-preserving learning; and *conversational AI* to provide unified and user-friendly interfaces for human–machine interaction and improve the accessibility and usability of the entire ICICLE system. All these systems' AI research areas embrace and aim to address the full complexity and heterogeneity in the environment-bound, edge-to-center, and user-supporting scenarios targeted by ICICLE.

The ICICLE Institute brings under one roof a broad range of experts, institutions, and partners to develop next-generation AI and CI allowing for **democratized and plug-and-play AI modeling across the computing continuum** to rapidly impact salient use-inspired projects while training the next generation of AI-aware talent. Following the mantra "today's AI is tomorrow's CI," the ICICLE Institute is integrating tenets of foundational AI into CI in a virtuous cycle, whereby AI innovations in areas such as conversational AI and federated learning become integrated into the CI, while foundational CI research supports AI innovation by improving the ability to rapidly train, deploy, and secure models and data.

HIGHLIGHTED ACCOMPLISHMENTS

Even early in the life of the ICICLE Institute, we are starting to see synergistic accomplishments that advance our ambitious goals of combining foundational AI research
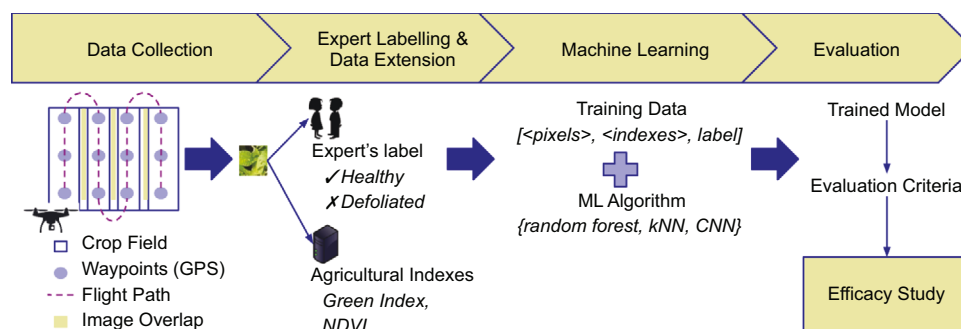
FIGURE 2 AI inspired agriculture.

with CI ecosystem advancements. ICICLE must be a sustainable, transformational nexus of collaborations among the AI, CI, and domain sciences while at the same time BPC and developing an AI-savvy and ethically aware workforce.

Institute research outputs, outcomes, and impacts take place across several dimensions but are oriented towards contributions to application domains (animal ecology, digital agriculture, smart foodsheds).

Animal ecology: Two guiding applications for this area are smart camera traps that can intelligently monitor animal behavior from a fixed (typically remote) location, and drone-based observations that can monitor behavior more dynamically. Early results include the release of benchmark datasets and models, as well as the release of a reference simulation application (camera traps). Research foci include knowledge graphs, resource profiling, model training, and model commons.

Early outreach and interactions include the Ohio Department of Natural Resources, and Division of Wildlife. Collaboration with the NSF-funded HDR DIRSE Imageomics Institute has resulted in contributions to the Experiential Field Course in Kenya, focusing on animal behavior inference from drone videos.

Digital agriculture: This application area enables democratized access of AI technologies to digital agriculture services. Its guiding applications are aerial crop scouting, crop health modeling, and autonomous in-field machinery; this requires HPC model training and inference, edge wireless, model commons (Figure 2). Early results include released data sets and AI models on soybean crop health (Ockerman et al. 2023). The digital agriculture group has initiated collaborative outreach to the Ohio Soybean Council, AgAID, AIFARMS, Tata Consulting Services and other industry partners. We have also establised the US-Indo research collaboration with Technology Innovation Hub on IoT (with IIT-Bombay).

Smart foodsheds: This application area envisions democratizing AI for food systems workers of the future

through access to computational workflows and tools. It strives to achieve this goal by means of a standardized ICICLE Integrated Knowledge and Learning Environment (IKLE) that serves as an entry point for food systems planners to explore their own ontologically encoded data sets and eventually mixed with public data (Tu, Wang, et al. 2023b). The group's early research results include the Persons-Projects-Organizations-Datasets ("PPOD") ontology, now publicly available on GitHub. It is developing a conversation agent and federated data capabilities. They have worked with the visual analytics research thrust on a prototype visualization platform for exploring KG and food flow data.

AI foundations: Within this thrust, the team is developing plug-and-play technology to allow construction of systems utilizing key AI infrastructure components, including models from model commons, knowledge graphs, federated learning and adaptive AI techniques. One example is a new holistic transfer learning methodology (Tu, Chen, et al. 2023a) that allows adaptation of pretrained models to new distributions without disrupting the original performance; when used in the smart camera traps setting, this empowers ecologists to adapt species classification models to work in new locations with few data. Similarly, information access in food system knowledge graphs will soon be driven by new adaptive conversational AI technology that provides effective few-shot in-context learning for knowledge-base question answering with large language models (Gu, Deng, and Su 2022).

CI4AI: Within the *CI4AI* thrust, early results include new top-k gradient sparsification methods and all-to-all sparse communication algorithms for high-performance model training that is 20% faster than SOTA method on BERT training with convergence (Hussain et al. 2022). The novel solution includes a new distributed scheduling algorithm for ensuring AI-adaptive, probabilistic per-packet real-time guarantees has the potential for improving network capacity by a factor of 5–20x. Finally, an intelligent resource provisioner with reinforcement learning has

reduced average wait time by 54.2% on three production GPU clusters (Ding et al. 2023).

AI4CI: Within the *AI4CI* thrust, early results include a hybrid analytical/ML model for searching functionally equivalent CUDA code (Xu et al. 2022), a family of CNN architectures for the detection of soybean defoliation (Ockerman et al. 2023), and improved predictive capability of the ML-based resource prediction models (Vallabhajosyula and Ramnath 2023). Within visual analytics, early results include reusable software components for exploring knowledge graph data that allow bi-directional interaction between multiple-coordinated views to gain a deeper understanding of the graph data. These components have been used to create an interactive knowledge and learning environment for smart foodsheds (as mentioned above), and also for visualizing loss landscapes of multiple training clients in a federated learning system (Tu, Wang, et al. 2023b).

PADI: Within the "privacy, accountability and data integrity" thrust (PADI), early results include novel privacy-preserving techniques applied to sequential data sharing (Jiang, Yilmaz, and Ayday 2023), the sharing of summary statistics from sensitive databases, and collaborative quality control for research databases (Dervishi et al. 2023). In ongoing research, we are leveraging model cards and model ontologies for accountability AI models and their use.

CI/SOFTWARE DEVELOPMENT AND RELEASE

The ICICLE team is also focused on translating research to publicly available CI/software components. It has embarked on a robust CI/software development, testing, and release plan thanks to strong cooperation between the software thrust members and the developers of various CI/software components. Many of these components are being integrated with the TAPIS framework (TAPIS 2023) to provide solutions in the computing continuum environment. Multiple rounds of releases with more than 20 components have been released so far (ICICLE-Software 2023). The team plans to continue periodic releases in realizing the edge-to-center AI-as-a-service enterprise.

OVERVIEW OF THE BROADER IMPACT ACTIVITIES

The ICICLE Institute is committed to an inclusive environment for all. Its code of conduct guides all members of the ICICLE community regardless of position or seniority and includes a mechanism for reporting and addressing issues that might arise. The institute further hosted a 3-h facilitated workshop on inclusion at its 2022 All Hands Meeting. Attendees were asked to make a personal pledge to take action to ensure inclusive academic spaces with follow-up surveys to evaluate effectiveness.

The strategy adopted for realizing the broader impacts goals of the institute is anchored in the Collective Impacts model where a single backbone group works collectively across WFD, BPC, and knowledge transfer on shared goals and shared measures.

The institute's overriding goal for WFD is an ethically aware AI workforce. We begun by defining an ICICLE Ethical AI Framework, before applying the framework to professional development of ICICLE team members themselves, to professional training opportunities, and to research experiences for undergrads, and K-12 student immersive experiences. The BPC goals are twofold: (i) making BPC part of the fabric of the institute, and (ii) building awareness and fostering actions leading to broadened participation.

Broader impacts are realized by ICICLE through deep interaction with stakeholder communities. The institute has diverse engagement vehicles including a student affiliates program, an academic affiliates program, an industry partners consortium for industry partners, and a stakeholder roundtable. The institute is flexible to new engagements and research interactions that we envision as it accelerates in its software availability and CI platform offerings.

Broader impacts programmatic activities are having early success in extending the impact of the institute. In particular, the ICICLE NexGens, an affiliate group composed of ICICLE students, is increasing student voices on institute decisions and increasing a students sense of belonging. The ICICLE Educational Fellows program (Indiana University 2023), which awarded its first cohort of five fellows in 2023, brings to the project an external perspective on stakeholder engagement and democratizing AI.

NEXUS OF COLLABORATION ACTIVITIES

The institute is actively becoming the nexus of multidisciplinary and multi-organizational collaborative activities for creating intelligent CI to democratize AI. Representative activities include (i) leading the CI resource concierge service for all AI institutes; (ii) leading a Special Interest Group (SIG) on AI Ethics across all AI institutes; (iii) collaboration between ICICLE and AI4Opt AI institute, optimizing food logistics for the USDA-LFPA program; (iv) Joint ICICLE-TIH-Bombay (India) Digital Agriculture activities; (v) diverse set of users using the released

CI/Software components; (vi) joint cross-cutting research and publications across groups/thrusts/organizations; and (vii) new research grants/proposals leveraging ICICLE activities.

FUTURE ACTIVITIES

Over the course of its first 5 years, ICICLE will establish itself as a leading institute for CI in the field of AI. Its impact will resonate within the domains of high-performance computing, AI and machine learning (ML), as well as in the application-inspired disciplines of animal ecology, digital agriculture, and smart foodsheds. There will be a notable influence on the foundational systems of AI and the interdisciplinary approaches that inherently incorporate AI for enhanced performance and robustness. Given ICICLE's emphasis on carefully selected projects, rubrics and assessment metrics are employed at the conclusion of every virtuous cycle to measure the scholarly impact on both foundational and translational domains. In addition, ICICLE is focused on training and engaging an AI informed and ethically aware workforce by means of efforts that are broadly inclusive. The broader impacts efforts complement technical developments to ensure that both research products and people associated with the institute enter the world accounting for and reflecting different perspectives.

CONCLUSIONS

ICICLE is positioned as the nexus for foundational and use-inspired AI for CI and CI for AI research among its collaborating institutions, domain scientists, and other NSF AI Institutes, along with external partners. Our overarching objectives are to (i) facilitate collaboration and synergy between CI researchers, AI researchers, and domain scientists; (ii) involve stakeholders throughout the research process; (iii) build and support a community of AI4CI and CI4AI activities; (iv) support the adoption of ICICLE technology, and (v) engage and partner with external organizations for technology transfer and commercialization. ICICLE aims to create a robust and effective ecosystem that accelerates both AI and CI research while maximizing their utility across various scientific disciplines and industries.

CONFLICT OF INTEREST STATEMENT
The authors declare that there is no conflict.

ORCID
Dhabaleswar K. Panda https://orcid.org/0000-0002-0356-1781

REFERENCES
ICICLE-Software. 2023. https://icicle.osu.edu/cyberinfrastructure/software/ (accessed July 4, 2023).

ICICLE. 2023. https://icicle.osu.edu/ (accessed July 4, 2023).

Dervishi, L., W. Li, A. Halimi, X. Jiang, J. Vaidya, and E. Ayday. 2023. "Privacy Preserving Population Stratification for Collaborative Genomic Research." In ISOC Network and Distributed System Security Symposium.

Ding, Q., P. Zheng, S. Kudari, S. Venkataraman, and Z. Zhang. 2023. "Mirage: Towards Low-interruption Services on Batch GPU Clusters with Reinforcement Learning. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '23)." Association for Computing Machinery, New York, NY, USA, Article 25, 1-13. https://doi.org/10.1145/3581784.3607042

Gu, Y., X. Deng, and Y. Su. 2022. "Don't Generate, Discriminate: A Proposal for Grounding Language Models to Real-World Environments." In Annual Meeting of the Association for Computational Linguistics. https://api.semanticscholar.org/CorpusID:254853929.

Hussain, M. T., G. S. Abhishek, A. Buluç, and A. Azad. 2022. "Parallel Algorithms for Adding A Collection of Sparse Matrices." In 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), 285–94. IEEE.

Indiana University. 2023. "ICICLE Educational Fellows." June 20, 2023. https://icicle.osu.edu/education-and-outreach/icicle-educational-fellows-program.

Jiang, Y., E. Yilmaz, and E. Ayday. 2023. "Robust Fingerprint of Location Trajectories Under Differential Privacy." In Proceedings on Privacy Enhancing Technologies Symposium, volume 2023, 5–20. NIH Public Access.

Ockerman, S., J. Wu, C. Stewart, and Z. Zhang. 2023. "A Reflection On AI Model Selection for Digital Agriculture Image Datasets." In Workshop on AI for Agriculture and Food Systems.

TAPIS. 2023. https://tapis-project.org/ (accessed July 4, 2023).

Tu, C.-H., H.-Y. Chen, J. Zhong, J. Zhong, Z. Mai, V. Pahuja, T. Berger-Wolf, S. Gao, C. Stewart, Y. Su, and W.-L. Chao. 2023. "Holistic Transfer: Towards Non-Disruptive Fine-Tuning with Partial Target Data." In Conference on Neural Information Processing Systems.

Tu, Y., X. Wang, R. Qiu, H.-W. Shen, M. Miller, J. Rao, S. Gao, et al. 2023. "An Interactive Knowledge and Learning Environment in Smart Foodsheds." *IEEE Computer Graphics and Applications* 43(3): 36–47. https://doi.org/10.1109/MCG.2023.3263960.

Vallabhajosyula, S., and R. Ramnath. 2023. "Towards Characterizing DNNs to Estimate Training Time using HARP (HPC Application Resource (runtime) Predictor)." In Practice and Experience in Advanced Research Computing (PEARC'23).

Xu, Y., Q. Yuan, E. C. Barton, R. Li, P. Sadayappan, and A. Sukumaran-Rajam. 2022. "Effective Performance Modeling and

Domain-Specific Compiler Optimization of CNNs for GPUs." In Proceedings of the International Conference on Parallel Architecture and Compilation Techniques, 252–64. https://doi.org/10.1145/3559009.3569674.

AUTHOR BIOGRAPHIES

Dhabaleswar K. Panda, ICICLE Director, is a Professor and University Distinguished Scholar of Computer Science and Engineering at The Ohio State University. His research interests are high-performance computing, exascale computing, deep/machine learning, big data, and cloud computing. He is an IEEE Fellow and the recipient of the 2022 IEEE Charles Babbage Award.

Vipin Chaudhary, ICICLE co-PI, is the Kevin J. Kranzusch Professor and Chair of Computer and Data Science at the Case Western Reserve University. His research interests include High Performance CI, AI, Data Science, Medical Computing, and Quantum Computing.

Eric Fosler-Lussier, ICICLE co-PI, is the John Makhoul Professor and Associate Chair in Computer Science and Engineering at the Ohio State University. He has research interests in Conversational AI, and is a Fellow of the IEEE and ICSA.

Raghu Machiraju, ICICLE co-PI, is a Professor of Biomedical Informatics, Pathology, and Computer Science and Engineering at OSU; he also serves as an Associate Chair for Growth in CSE and previously served as the Founding Faculty Lead of the Translational Data Analytics Institute. Research interests include domain-specific AI methods, particularly for clinical medicine, and systems AI.

Amit Majumdar is the Director of the Data Enabled Scientific Computing Division at the San Diego Supercomputer Center, UCSD and an Associate Professor in the Department of Radiation Medicine and Applied Sciences. Research interests include parallel computing, science gateways and CI resources, and software for HPC and AI.

Beth Plale, ICICLE co-PI, is the Michael A and Laurie Burns McRobbie Bicentennial Professor of Computer Engineering and Chair of the Department of Intelligent Systems Engineering at Indiana University Bloomington. Her research interests are in computational and data infrastructure, open science, AI ethics, and data accountability.

Rajiv Ramnath is a Professor of Practice in Computer Science and Engineering at OSU with extensive industry and interdisciplinary collaborations. His work covers data science, cyberinfrastructure, software engineering, enterprise strategy, and computing in education.

Ponnuswamy Sadayappan is a Professor in the Kahlert School of Computing at the University of Utah. His primary research interests are in compiler optimization and high-performance computing. He is a Fellow of the IEEE.

Neelima Savardekar is the Managing Director of ICICLE at The Ohio State University.

Karen Tomko is the Director of Research Software Applications. Her research interests include performance improvement for research software and cyberinfrastructure for AI.