Received: 15 June 2021 | Accepted: 12 January 2022

DOI: 10.1002/aaai.12042

SPECIAL TOPIC ARTICLE

Enabling AI innovation via data and model sharing: An overview of the NSF Convergence Accelerator Track D

Chaitanya Baru[1] | **Michael Pozmantier**[1] | **Ilkay Altintas**[2] | **Stephen Baek**[3] | **Jonathan Cohen**[4] | **Laura Condon**[5] | **Giulia Fanti**[6] | **Raul Castro Fernandez**[7] | **Ethan Jackson**[8] | **Upmanu Lall**[9] | **Bennett Landman**[10] | **Hai Helen Li**[11] | **Claudia Marin**[12] | **Beatriz Martinez Lopez**[13] | **Dimitris Metaxas**[14] | **Bradley Olsen**[15] | **Grier Page**[16] | **Jingbo Shang**[2] | **Yelda Turkan**[17] | **Peng Zhang**[18]

[1]National Science Foundation

[2]University of California San Diego

[3]University of Virginia

[4]Princeton University

[5]University of Arizona

[6]Carnegie Mellon University

[7]University of Chicago

[8]Microsoft Corporation

[9]Columbia University

[10]Vanderbilt University

[11]Duke University

[12]Howard University

[13]University of California Davis

[14]Rutgers University

[15]MIT

[16]RTI

[17]Oregon State University

[18]Stony Brook University

Correspondence
Chaitanya Baru, San Diego Supercomputer Center, University of California San Diego, 9300 Gilman Drive, La Joll.
Email: cbaru@ucsd.edu

Present address
Chaitanya Baru, University of California San Diego

Abstract

This article provides a brief overview of 18 projects funded in Track D—Data and Model Sharing to Enable AI Innovation—of the 2020 Cohort of the National Science Foundation's (NSF) Convergence Accelerator (CA) program. The NSF CA is focused on transitioning research to practice for societal impact. The projects described here were funded for one year in phase I of the program, beginning September 2020. Their focus is on delivering tools, technologies, and techniques

to assist in sharing data as well as data-driven models to enable AI innovation. A broad range of domain areas is covered by the funded efforts, spanning across healthcare and medicine, to climate change and disaster, and civil/built infrastructure. The projects are addressing sharing of open as well as sensitive/private data. In September 2021, six of the eighteen projects described here were selected for phase II of the program, as noted in this article.

INTRODUCTION

In September 2020, the National Science Foundation's (NSF) Convergence Accelerator (CA) funded 18 projects in phase I of its Track D on *AI-Driven Innovation via Data and Model Sharing* (NSF 2020). This paper provides a brief overview of these eighteen projects focused on transitioning research to practice in data and model sharing, for sharing of open as well as sensitive data/models with privacy concerns. From September 2020 to May 2021, the projects participated in the CA *innovation curriculum* (see Baru et al. in this Special Issue) and worked toward proof-of-concept prototypes to test and illustrate their ideas. Proposals for phase II were submitted to NSF in May 2021. Six of the 18 projects were selected for phase II, as noted below in their respective project descriptions.

NSF's Harnessing the Data Revolution Big Idea expressed the need for a *ModelCommons* for sharing data and data-driven models, particularly for ML/AI (NSF 2018). These ideas were further expanded upon at the *ACM KDD 2018 Workshop on Common Model Infrastructure*, August 2018 (CMI 2018). The *National AI R&D Strategic Plan* update released in June 2019 (NSTC 2019) emphasized related issues including (i) the urgency in developing and making accessible a wide variety of datasets to meet the needs of a diverse spectrum of AI interests and applications while dealing with associated challenges of standardization, privacy, etc., (ii) making training and testing resources responsive to commercial and public interests, and (iii) developing open-source software libraries and toolkits to enable data sharing and access. The efforts undertaken in Track D would help contribute to the creation of a *national AI research infrastructure*, as recommended by the National Security Commission on Artificial Intelligence (NSCAI 2021).

The Track D phase I projects span a wide variety of application areas addressing a range of issues related to sharing open as well as sensitive data and models, including privacy concerns and access control considerations. The eighteen projects can be classified into the following set of broad areas based on common issues and challenges, (1) Healthcare/Medicine, Neuroscience, American Sign Language (ASL), Meta-analysis, Veterinary Science; (2) Water, Climate Change, Wildfires/Hazards, Biome/Ecosystems; (3) Civil/Built Infrastructure; (4) Cybersecurity; (5) Polymer Data; and (6) Technology Infrastructure. The rest of this paper provides a summary of each project, organized by these themes, followed by a brief conclusion.

HEALTHCARE/MEDICINE, NEUROSCIENCE, AMERICAN SIGN LANGUAGE, META-ANALYSIS, VETERINARY SCIENCE

Seven projects in Track D are working on topics related to sensitive data and models derived from heterogeneous image and other types of data related to medicine, healthcare, neuroscience, animal care, and the American Sign Language (ASL). An overview of each project is provided below.

Scalable, TRaceable AI for Imaging Translation: Innovation to Implementation for Accelerated Impact (STRAIT I3)

PI: Bennet Landman, Vanderbilt University (Landman 2020). Currently, literally thousands of AI models are being published in the scientific literature each year for medical imaging which are unable to be used in actual clinical applications due to the lack of a consistent system for validation, access, and use of these models. The goal of the STRAIT I3 consortium is to integrate learning, validation, and evaluation within a common framework. By removing technical barriers to technology interchange, imaging AI methods can be developed, characterized, and deployed more quickly and at a lower cost. The project is implementing a *Model Zoo* to address this challenge by providing a validation/peer review process to make models more accessible and useable (Banalagay et al. 2014; Harrigan et al. 2016; Cai et al. 2020). The system incorporates a data provenance and annotation interface, and the validation process includes various checks, for example, for irreproducible implementations and overtraining of models. The system is being implemented with open-source

software on commodity hardware to facilitate easy reuse of the validated models. The team consisting of Vanderbilt University, Vanderbilt Medical Center, MD.ai, Kaggle, and the Society for Imaging Informatics in Medicine is working initially with open datasets and models for radiological assessment of COVID-19 pneumonia and will expand and scale up this approach to other medical imaging modalities including dermatology and ophthalmology.

ImagiQ: Asynchronous and decentralized federated learning for medical imaging

PI: Stephen Baek, University of Iowa (Baek 2020). A key challenge in using AI/ML methods in medicine is the ability to gather/create large collections of data to feed data-hungry AI/ML models. Sensitive medical data are distributed across many different, independent clinics/institutions. Data sharing is often hindered by a variety of regulatory, administrative, and technical impediments. This project is developing *ImagiQ*, a distributed ML platform which employs a novel approach called "Peer-Adaptive Ensemble" to support a dynamically growing eco-system of AI models which "travel" to different sites to be trained on the data at those sites (Xu et al. 2021). Models are also vetted against extramural data sets collected from different hospitals to assure generalizability and reliability. AI models in *ImagIQ* learn to make better decisions as they see data from more diverse patient cases at different hospitals. This ensemble approach of AI models can lead to more reliable and trustworthy decision making by clinicians and their patients (https://bit.ly/imagiq). The team at the University of Iowa is collaborating with industry partner NVIDIA and other partners in the medical AI industry to develop and test this approach.

A trusted integrative model and data sharing platform for accelerating AI-driven health innovation

PI: Hai Helen Li, Duke University (Li 2020). Similar to *ImagIQ*, this project is also implementing federated ML methods involving sensitive data. As mentioned, while massive collections of data are now available in the healthcare sector, the critical challenge to fully exploiting these big data is the complexity of the corresponding ML models in the presence of sensitive data and data privacy concerns. The project is implementing a system called LEARNER, which includes a cloud-based health data repository with data from multiple, independent sites. Privacy-preserving, federated learning over these sensitive data from multiple sites, is supported using an efficient method for communication during the federated learning phase (Li et al. 2020). The project is a collaboration among large medical centers, viz., Duke University School of Medicine and University of Pittsburgh Medical Center, and industrial partners, viz., Accenture, IBM, and Infinia ML. The set of platform services being developed as part of LEARNER have potential for use in almost any sector involving sensitive data, including scientific research, policymaking, and broadly in enterprise applications.

A standardized model description format for accelerating convergence in neuroscience, cognitive science, machine learning, and beyond

PI: Jonathan Cohen, Princeton University (Cohen 2020). This project is addressing the challenge of sharing machine learning models as well as cognitive science models by creating a *Model Description Format* (MDF, https://github.com/ModECI/MDF) whose adoption would be promoted by a Model Exchange and Convergence Initiative (ModECI, https://modeci.org) to enable sharing of data as well as theories and models in an objective, transparent, and reproducible way. The MDF represents models as computational graphs specifying both the architecture and control flow, coupled with a library of open-source tools for importing, exporting, and validating the models. MDF is being designed to be compatible with existing disciplinary standards, that is, ONNX in ML (https://onnx.ai) and NeuroML in neuroscience (https://neuroml.org/), and serve as a bridge between the two. The description format would support not only model dissemination but also validation and reproducibility; migration of models across domains, for example, across neuroscience and machine learning; use of models of brain function in machine learning applications; and integration of models at different levels of analysis, from neural models to models of cognitive function and, eventually, population-level models.

Application of sequential inductive transfer learning for experimental metadata normalization to enable rapid integrative analysis

PI: Page Grier, Research Triangle Park (Page 2020). The challenge addressed here is the current difficulty in performing integrative analyses and meta-analyses across multiple, distinct databases. Tools are needed to deal with data silos, even in the same domain, that use different terminologies and measurement schemes. Pretrained

Learning Models (PLMs) are being developed to combine methods used in Natural Language Processing (NLP) and in transfer learning to allow data-driven models built in one domain to be applied in another without the time and expense of having to develop large training datasets. A multidisciplinary team of researchers and experts in statistics, epidemiology, data harmonization, machine leaning, ethics, databases, imaging, and software engineering from RTI and collaborating organizations is developing a prototype PLM from a large existing manually trained dataset of PhenX (https://www.phenx.org/)—dbGAP (https://www.ncbi.nlm.nih.gov/gap) metadata linkage to link metadata across four diverse biomedical databases. This approach has the potential of being applicable to databases in many different domains in addition to just biomedicine.

Data and AI methods for modeling facial expressions in language with applications to privacy for the deaf, ASL education, and linguistic research

PI: Dimitri Metaxas, Rutgers University (Metaxas 2020). Sign languages require important linguistic information to be conveyed via facial expressions and head gestures. There are currently no tools available to enable deaf signers—and the over 500,000 users of ASL—to communicate anonymously through videos due to this visual cue requirement. This project is designing a privacy tool to enable ASL signers to share anonymized videos by disguising their faces without loss of linguistic information which could also be useful in a range of applications such as, say, sharing of sensitive information online, anonymous peer review of submissions to academic journals, and the like. The project is developing sustainable, fair, robust AI methods for facial analytics targeted to new applications addressing important problems related to the use of facial expressions and head gestures in signed and spoken language (Metaxas, Zhao, and Peng 2020). The approach combines 3D modeling, ML, and linguistic knowledge derived from annotated video corpora to overcome limitations of prior methods due to aspects like large head rotations, image blurring, and occlusions. Applications could include anonymization of sign language videos while retaining the essential linguistic information conveyed via expressions (Lee et al. 2021), tools to help sign language learners produce these grammatically meaningful expressions correctly, and tools for annotation of multimedia language data to advance the understanding of how gestures are used in conjunction with speech (Neidle, Sclaroff, and Athitsos 2001; Neidle 2020).

Data-driven disease control and prevention in veterinary health

PI: Beatriz Martinez-Lopez, UC Davis (Martinez-Lopez 2020). Maintaining good swine health is essential for achieving high productivity and efficiency in the large-scale global pork market. Effective use of AI algorithms to address veterinary health challenges requires integration of the vast amounts of data being collected in all steps of primary production in the pork industry from *farrowing* and *weaning* to *slaughter*. The data across these stages are often incomplete, inconsistent, and scattered across producers, diagnostic labs, and veterinary clinics. By expanding upon the *Disease BioPortal* platform (http://www.cadms.ucdavis.edu/) this project will facilitate data sharing and usage by AI applications as well as data/information visualization by a range of stakeholders including veterinarians, producers, the general public, and others. Data pipelines are being develop for effective multilevel data connection and integration and cost-aware adaptive sampling and explainable machine learning models to solve key problems in the swine industry including antimicrobial resistance (AMR) and swine influenza infections. Outcomes from this effort will help push the frontiers of *precision epidemiology* with the potential for improving animal health and welfare and securing the sustainability of US agriculture and food systems by providing data-driven decision tools. This project has been selected for phase II.

WATER, CLIMATE CHANGE, WILDFIRES/HAZARDS, BIOME/ECOSYSTEMS

Four Track D projects are addressing issues related to water management, impact of climate change, wildfires and other hazards, and monitoring the biome.

Hidden water and hydrologic extremes: A groundwater data platform for machine learning and water management

PI: Laura Condon, Univ of Arizona (Condon 2020). Understanding how human operations and groundwater interact especially as a result of extreme events such as droughts and floods requires combining state-of the-art groundwater science with operational water management tools. This is needed for improved hydrologic forecasting as well as to inform large-scale water management practices. The project is developing *HydroGEN*, a web-based ML platform that is able to generate custom hydrologic

scenarios on-demand by combining state-of the-art groundwater models, powerful physics-based simulations, and observations from operational management tools and machine learning. The tool is able to provide customizable scenarios from the bedrock through the treetops for a variety of users, including water managers and planners who may not have prior modeling experience but are, nonetheless, able to work with state-of-the-art tools to explore scenarios of interest. The integrated approach addresses biases in the current risk-assessment approaches that do not consider groundwater and the potential to improve long-term sustainability by more actively managing groundwater and accounting for groundwater–surface water interactions in making future estimates. The project partnership includes the Bureau of Reclamation, the largest wholesale water provider in the country providing water to more than 31 million people and 10 million acres of farmland, who are driving use case design and the metrics for evaluating success in phase I. The project also includes an educational component with hands-on activities and challenges for undergraduate students that provide them with real-world experiences in ML and data science. This project has been selected for phase II.

Water system data pooling for climate vulnerability assessment and warning system

PI: Upmanu Lall, Columbia University (Lall 2020). A major gap in the resiliency of America's water supply is the resiliency to climate variability and change, especially for the thousands of smaller utilities in the United States that typically lack the financial wherewithal and technical capacity to analyze these risks and assess their impact on operations. Physics-based hydroclimate models are unable to reproduce the basic statistics of observed phenomena and have poor forecast skills beyond short periods. Hybrid AI/ML approaches that are informed by causal variable structures and the topology of hydroclimatic networks are providing a novel way to overcome these barriers, including for extrapolation to extreme conditions at the space and time scales of direct interest for water system risk management applications. This project is developing a cloud-based, multiscale AI-enabled modeling platform to help quantify America's water supply risk at the level of water utilities and their regulatory state and federal agencies. Model outputs are generated based on the needs of identified users and will become a community data and modeling resource. The models being developed will assist in the strategic planning and operations of water systems in the face of an increasing frequency of floods and droughts under climate change and aging infrastructure conditions—factors that constitute significant risks to the nation's safe supply of water. The project incorporates collaborations with industry partners via the Water Initiative at the Columbia University Water Center (http://water.columbia.edu/research-themes/americas-water/) and the Water Innovation Network for Sustainable Small Systems at the University of Massachusetts (https://www.umass.edu/winsss/) and brings together experts in water systems, climate science, AI technologies, emulation models and software development to provide multiscale modeling for feature identification, spatiotemporal modeling and forecasting, functional dependence, inverse problems and transfer learning.

Artificial intelligence and community driven wildland fire innovation via a WIFIRE Commons infrastructure for data and model sharing

PI: Ilkay Altintas, UC San Diego (Altintas 2020). Longstanding efforts in fire suppression over many decades have created an accumulation of flammable vegetation on landscapes, contributing to megafires that risk human life and property and permanently destroy ecosystems. However, small controllable fires could reduce the risk of uncontrollable large fires. Building upon work in the WIFIRE Commons project (Moore 2019; Jain et al. 2020; Linn et al. 2020; Rowell et al. 2020), this effort is developing *BurnPro3D*, a decision support platform to help the fire response and mitigation community understand risks and tradeoffs quickly and accurately to manage wildfires and conduct controlled burns more effectively. The project is developing knowledge management techniques for fusing and preparing diverse data for use in fire modeling; physics-based ML for use in fire models and use of deep learning to understand complex fire behavior processes; constraint optimization methods addressing complex tradeoffs in the decision process for the placement and timing of controlled burns; and explainable AI for better interpretability of data and models by diverse users including fire managers, municipal leaders, and educators. The AI/ML capabilities and the WIFIRE Commons developed here are open resources for use in other domains as well. This project has been selected for phase II.

Deep monitoring of the biome will converge life sciences, policy, and engineering

PI: Janos Sztipanovits, Vanderbilt University (Sztipanovits 2020). Essential to convergence research is the

convergence of extremely heterogeneous and varied data from a wide spectrum of disciplines. This project is tackling the challenges in unifying such data focusing initially on making unified biome and ecological data sets available to a wide range of end users from researchers to policy makers and industries. The power of this unified dataset is demonstrated by the development of unified agent-based models for predicting mosquito populations. Mosquito-borne diseases already account for over 600 million cases of human disease each year, with a disproportionately large impact on disadvantaged communities in sub-Saharan Africa. The new science and technology ecosystem that would emerge from the converged data has the potential for wide impact on a range of issues from human health to agriculture, national security, ecology, the life sciences, engineering, and policy domains. The project seeks to create a new generation of scientists able to develop predictive AI models of the biome along with science-based methods and tools for shaping policies and delivering policy-aware tools to solve important societal-scale challenges. This project has been selected for phase II.

CIVIL/BUILT INFRASTRUCTURE

Intelligent surveillance platform for damage detection and localization of civil infrastructure

PI: Claudia Marin, Howard University (Marin 2020). Reliable structural health monitoring tools are needed for prioritizing maintenance and repair decisions regarding the nation's aging infrastructure to reduce the societal and economic impacts of aging, deterioration, and extreme events on civil infrastructure. Achieving this objective requires diverse perspectives from multiple disciplines, and partnerships crossing organizational, institutional, and disciplinary boundaries. This project is integrating advances in computer vision, ML, and pattern recognition with physics-based reasoning to develop a novel, accurate, field-calibrated computational platform for in situ monitoring of civil infrastructure. Key aspects of the platform are the use of video cameras instead of sensors, and incorporation of physics laws and constraints into ML algorithms. High-quality field data from various structures is being used to create, evaluate, and validate each component of the platform and develop strategies to facilitate widespread implementation, considering various scientific, regulatory, and societal issues. Phase I of the effort focuses on selection of benchmark structures, collection of data, and the development of a prototype of the video and ML modules. The overall objective is to deliver a field-calibrated computational platform that integrates the ML models with a video analytics module, to be implemented on selected benchmark structures, along with user manuals and educational materials for end-users.

Rapid development of intelligent, built environment Geo-databases using AI and data-driven models

PI: Yelda Turkan, Oregon State (Turkan 2020). Designers and planners increasingly require detailed 3D geospatial databases of the built environment to better meet the needs of citizens and emerging technologies such as autonomous vehicles. Traditionally, assembling and maintaining such databases has been expensive due to the costs of manual feature extraction. This project is creating a cloud-based service for ingesting scanned 3D point cloud models of *horizontal* (highway, bridges, etc.) and *vertical* (buildings, factory facilities) built infrastructure, called *Scan-to-BIM*, into Building Information Management systems to help the architectural, engineering, and construction (AEC) community reduce the costs of construction and renovation of public infrastructure. A sizable collection of benchmark datasets with annotated point cloud scans and corresponding BIM models are being used to develop the *Scan-to-BIM* framework for rapidly and reliably generating BIM models from scan data and validate the process with metrics related to parameters of interest to stakeholders, such as accuracy of modeled door widths (which are important to ADA compliance assessment), or fitness of the models for urban renewal and redevelopment projects. A *Scan-to-BIM* challenge is also being designed. The project is engaged in workforce development efforts in computer vision and geomatics to close the large gap between employment needs and workforce readiness, particularly for underrepresented minorities.

AI-enabled provably resilient networked microgrids

PI: Peng Zhang, Stony Brook University (Zhang 2020). Electric power grids need to be resilient to a variety of situations including cyberattacks, system faults, and other disaster events. The use of AI techniques for system management of networked microgrids has the potential to transform today's community power infrastructures into tomorrow's autonomic microgrids incorporating high levels of renewable energy and providing resilience. To understand system dynamics, this project is developing a neural-ordinary-differential-equations-net (ODE-Net)-enabled reachability theory which is able to achieve

online dynamic model discovery and formal stability verification of microgrids. Next, this ODE-Net is enhanced with Variational Stochastic Differential Networks (VSDNs) to accurately model the continuous dynamics of microgrids under missing data and system uncertainty. Further, to provide runtime safety assurance despite possible flaws and vulnerabilities in virtualized controllers, a Neural Simplex Architecture is being developed. If a safety violation is imminent, the decision module switches control from the AI-based advanced controller to a verified-safe baseline controller and incrementally retrains the AI-based controller (Jiang et al. 2021; Tang et al. 2021; Wang et al. 2021; Zhang 2021; Zhou and Zhang 2021a, Zhou and Zhang 2021b, Zhou, Zhang, and Yue 2021). The project is partnering with a $1B microgrid project at the Energy & Innovation Park (EIP) in New Britain, Connecticut, which aims to transform the traditional manufacturing base of that city to a new digital economy supporting the rapidly growing Data Center sector. Monitoring data, control models, and lessons learned from test beds, as well as data from the operational microgrids operated by collaborators, will be shared with the research and user communities. This project has been selected for phase II.

CYBERSECURITY

AI-enabled, privacy-preserving information sharing for securing network infrastructure

PI: Guilia Fanti, Carnegie Mellon University (Fanti 2020). Cyberattacks perpetrated by fraudsters and cybercriminals are resulting in huge losses to enterprises—exceeding $27 billion in 2019 alone. Today, these activities are largely detected through a combination of automated models that raise alerts on anomalous traffic, and human analysts who determine whether alerts are legitimate. However, current automated and human-based checking methods are falling behind due to (a) insufficient or irrelevant data for training automated models, leading to inability in distinguishing between benign versus malicious anomalies, and (b) lack of good techniques for evaluating the confidence of an alert. This project is developing the *aiShare* platform to generate synthetic data that mimics important patterns in the real data, without having to share the real data. The project team is interdisciplinary in composition, spanning AI/ML, security, privacy, networked systems, law, and policy. The team is addressing fundamental tradeoffs among privacy, utility, and efficiency along three key thrusts: (1) design and implementation of novel generative adversarial networks (GANs) to enable an enterprise to model its network data to inform anomaly detection by others (Yoon, Jarrett, and van der Schaar 2019), (2) design and implementation of new cryptographic protocols and systems workflows for efficiently comparing hypotheses, for example, suspicious identifiers, such as domain names, IP subnets, and program hashes, across enterprises to inform policy deployments, and (3) development of new legal and policy analyses on the implications of sharing such synthetic data, ML models, and hypotheses.

POLYMER DATA

A community resource for innovation in polymer materials

PI: Bradley Olsen, Massachusetts Institute of Technology (Olsen 2020). A vast amount of valuable information on polymers and other soft materials is currently locked across small, disparate data sets, creating a pressing need for a community-wide effort for easier data sharing. This project is developing a sharing infrastructure for data and models for such materials at multiple scales, from chemical bonds to molecular interactions. Text-based extraction schemes will be used for data extraction, polymer chemical structures will be extracted from images using an optical chemical structure recognition system, and new ML methods will be developed for data curation to enable integration of a wider range of data to overcome data sparsity and diversity. In developing these capabilities the project will address the current inability to deal with molecular distributions and stochastic reaction networks; characterization/data generation challenges for stochastic chemistry; challenges with nomenclature and molecular representation; and polymer properties determined on multiple scales from the chemical bond to the molecule to collective molecular interactions. The initial testbed will be in the area of *polyurethanes*—representing a large polymer market with diverse chemistry, substantial data availability in the patent and journal literature, and structure-processing-property relationships that are a playground for continued material innovation. The project will evaluate paradigms that mix embargoed and open data and explore models for ownership and credit that enable wider sharing of data across the community. All tools and standards will be made publicly available using open-source development projects. The project team includes members from industry (Dow, Citrine), academia (MIT), and government laboratories (NIST). This project has been selected for phase II.

TECHNOLOGY INFRASTRUCTURE

Two Track D projects are focusing on developing generic capabilities and computing infrastructure for data and

model sharing regardless of the particular application domain.

Intelligent sharing and search for AI models and datasets

PI: Jingbo Shang, UC San Diego (Shang 2020). A major challenge currently in using AI in practice is the effort needed to find AI-ready data and AI models that could potentially be reused for a different application/purpose. This project is building a hub and portal platform to provide search and matching capability for AI data and models, while also enforcing privacy constraints. The portal enables systematic incorporation of domain knowledge for substantially improved quality of results. The planned system deploys four novel techniques: (1) a fine-grained privacy control technique with adaptive descriptive statistics that strike a balance between the privacy needs of data owners and application-driven usability of models. Privacy-controlled access is observed throughout the system; (2) an automated metadata generation method that exploits information about AI models and datasets (e.g., data values, model parameters, auxiliary descriptions) to incorporate domain logic into semantics. This metadata, together with the models and datasets, will be organized as a text-rich network that can be systematically updated and navigated by the users; (3) a representation learning method that transforms information in the text-rich network into a latent space, where datasets/models with similar semantics would be close to each other. This learning over multimodal data enables comprehensive understanding about models and datasets; and (4) a learning-to-match model that bridges datasets and models under domain and user constraints. The constraints are also induced from schema alignment between models and datasets that can also filter obvious noncompatible model and dataset choices, thus significantly expediting the search and matching process. The project partners include the IEEE DataPort, OpenML, and a few other technology companies.

Data station: Orchestrating data and models

PIs: Raul Castro Fernandez, Ian Foster, University of Chicago (Foster 2020). Sharing of sensitive data with privacy concerns among independent parties in a distributed setting remains a challenge. There is need for data/software infrastructure to support this capability. This project is designing and implementing a computing environment called, *DataStation*, which supports enforcement of any arbitrarily specified data access and governance policies in such an environment (Fernandez et al. 2020). *DataStation* is a data platform that permits data owners to pool their data together and jointly extract the benefits of the combined data, such as better machine learning models and more precise answers. Once shared, data are sealed and cannot be accessed by anyone without explicit permission. Users employ novel data-blind query interfaces to submit their queries without seeing the data first, and results to the queries are made available to the users only if owners agree. The *DataStation* enables trustworthy data sharing by using novel authorization schemas, hardware, and protocols. It introduces a novel "data-blind" query interface that enable users to specify data-driven tasks, such as traditional data queries and machine learning model training, without the user requiring direct access to the data itself by following the paradigm of bringing computations to the data. Another novel feature is that task capsules can include user-defined metrics to determine which results are useful from a user's perspective, as well as which trust constraints need to be met for valid provenance of input datasets. The *DataStation* captures metadata each time derived data products are created and provides mechanisms to implement data governance and data access policies that adhere to specifications provided by data contributors. Only authorized users can access the data based on a novel *access-token* model that permits fine-grained yet scalable access control. The project is testing the effectiveness of this approach with a diverse set of use cases in materials science, biomedicine, and in corporate/enterprise scenarios.

SUMMARY

As indicated in the summaries above, six projects have been selected for phase II, from the 18 phase I projects, beginning in September 2021. Details of how the CA program activities are structured in phase I and phase II, along with the respective review processes and review criteria, are described in the article by Baru et al. in this Special Issue.

ACKNOWLEDGMENTS

The work performed by all the projects described in this article was supported by grants from the Convergence Accelerator program of the National Science Foundation under different NSF award numbers which may be obtained via the link to each project abstract provided in the References section.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Altintas, I. 2020. Artificial Intelligence and Community Driven Wildland Fire Innovation via a WIFIRE Commons Infrastructure for Data and Model Sharing. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040676&HistoricalAwards=false. Accessed: Feb 28, 2022.

Baek, S. 2020. ImagiQ: Asynchronous and Decentralized Federated Learning for Medical Imaging. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040532&HistoricalAwards=false. Accessed: Feb 28, 2022.

Banalagay, R., K. Covington, D. Mitch Wilkes, and B. Landman. 2014. "Resource Estimation in High Performance Medical Image Computing." *Neuroinformatics* 12(4): 563–73.

Cai, L., Yang, Q., Hansen, C., Nath, V., Ramadass, K., Johnson, G., Conrad, B., Boyd, B., Begnoche, J., Beason-Held, L., Shafer, A., Resnick, S., Taylor, W., Price, G., Morgan, V., Rogers, B., Schilling, K., and Landman, B. 2020. "PreQual: An Automated Pipeline for Integrated Preprocessing and Quality Assurance of Diffusion Weighted MRI Images." *Magnetic Resonance in Medicine* 86(1): 456–70.

CMI 2018. Workshop on Common Model Infrastructure at ACM KDD Conference 2018, London, UK, August 20, 2018. https://cmi2018.sdsc.edu/

Cohen, J. 2020. A Standardized Model Description Format for Accelerating Convergence in Neuroscience, Cognitive Science, Machine Learning, and Beyond. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040682&HistoricalAwards=false. Accessed: Feb 28, 2022.

Condon, L. 2020. Hidden Water and Hydrologic Extremes: A Groundwater Data Platform for Machine Learning and Water Management. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040542&HistoricalAwards=false. Accessed: Feb 28, 2022.

Fanti, G. 2020. AI-Enabled, Privacy-Preserving Information Sharing for Securing Network Infrastructure. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040675&HistoricalAwards=false. Accessed: Feb 28, 2022.

Fernandez, R., Chard, K., Blaiszik, B., Krishnan, S., Elmore, A, Obermeyer, Z., Risley, J., Mullainathan, S., Franklin, M. & Ian Foster, I. 2020. The Data Station: Combining Data, Compute, and Market Forces, arXiv:2009.00035 [cs.DB].

Foster, I. 2020. The Data Hypervisor: Orchestrating Data and Models. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040718&HistoricalAwards=false. Accessed: Feb 28, 2022.

Harrigan, R., Yvernault, B., Boyd, B., Damon, S., Gibney, K., Conrad, B., Phillips, N., Rogers, B., Gao, Y., and Landman, B. 2016. "Vanderbilt University Institute of Imaging Science Center for Computational Imaging XNAT: A Multimodal Data Archive and Processing Environment." *Article in Neuroimage* 124(Pt B): 1097–101.

Jain, P., Coogan, S., Subramanian, S., Crowley, M., Taylor, S., and Flannigan, M. 2020. "A Review of Machine Learning Applications in Wildfire Science and Management." *Article in Environmental Reviews*. 28(4): 478–505. https://doi.org/10.1139/er-2020-0019

Jiang, Z., Tang, Z., Zhang, P., and Qin, Y. 2021. "Programmable Adaptive Security Scanning for Networked Microgrids." *Article in Engineering* 7(8): 1087–100.

Lall, U. 2020. America's Water Risk: Water System Data Pooling for Climate Vulnerability Assessment and Warning System. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040613&HistoricalAwards=false. Accessed: Feb 28, 2022.

Landman, B. 2020. Scalable, TRaceable Ai for Imaging Translation: Innovation to Implementation for Accelerated Impact (STRAIT I3). NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040462&HistoricalAwards=false. Accessed: Feb 28, 2022.

Lee, S., Glasser, A., Dingman, B., Xia, Z., Metaxas, D., Neidle, C. & Huenerfauth, M. 2021. "American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users." In 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'21), October 18-22, 2021, New York, NY, USA: Association for Computing Machinery.

Li, H. 2020. A Trusted Integrative Model and Data Sharing Platform for Accelerating AI-Driven Health Innovation. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040588&HistoricalAwards=false. Accessed: Feb 28, 2022.

Li, A., Sun, J., Wang, B., Duan, L., Li, S., Chen, Y. & Hai Li 2020. LotteryFL: Personalized and Communication-Efficient Federated Learning with Lottery Ticket Hypothesis on Non-iiD Datasets, arXiv:2008.03371 [cs.LG].

Linn, R., Goodrick, S., Brambilla, S., Brown, M., Middleton, R., O'Brien, J., and Hiers, J. 2020. "QUIC-Fire: A Fast-Running Simulation Tool for Prescribed Fire Planning." *Environmental Modelling Software* 125:104616. (Online article). https://www.sciencedirect.com/science/article/pii/S1364815219307388

Martinez-Lopez, B. 2020. Data-Driven Disease Control and Prevention in Veterinary Health. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040680&HistoricalAwards=false. Accessed: Feb 28, 2022.

Marin, C. 2020. Intelligent Surveillance Platform for Damage Detection and Localization of Civil Infrastructure. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040665&HistoricalAwards=false. Accessed: Feb 28, 2022.

Metaxas, D. 2020. Data & AI Methods for Modeling Facial Expressions in Language with Applications to Privacy for the Deaf, ASL Education & Linguistic Research. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040638&HistoricalAwards=false. Accessed: Feb 28, 2022.

Metaxas, D., Zhao, L., and Peng, X. 2021. Disentangled Representation Learning and Its Application to Face Analytics. In *Deep Learning-Based Face Analytics*, edited by N. K. Ratha, V. M. Patel, and R. Chellappa. Springer, Cham. https://doi.org/10.1007/978-3-030-74697-1_3

Moore 2019. Report from the Workshop on Fire Immediate Response System, April 24-26, 2019, Moore Foundation. https://www.moore.org/docs/default-source/default-document-library/2019-firs-workshop-report.pdf

Neidle, C., Sclaroff, S., and Athitsos, V. 2001. "SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data." *Behavior Research Methods, Instruments, and Computers* 33(3): 311–20.

Neidle, C. 2020. *What's New in SignStream® 3.3.0?* Report No. 17, Boston, MA: American Sign Language Linguistic Research Project, Boston University.

NSCAI. 2021. Final Report of the National Security Commission on AI, April 2021. https://www.nscai.gov/2021-final-report/

NSF 2018. Harnessing the Data Revolution (HDR) at NSF. https://www.nsf.gov/cise/harnessingdata/. Accessed: Feb 28, 2022.

NSF 2020. Track D: AI-Driven Innovation via Data and Model Sharing. https://beta.nsf.gov/funding/initiatives/convergence-accelerator/portfolio. Accessed: Feb 28, 2022.

NSTC. 2019. The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update, National Select Committee on AI, National Science and Technology Council, June 2019. https://www.nitrd.gov/news/National-AI-RD-Strategy-2019.aspx

Olsen, B. 2020. A Community Resource for Innovation in Polymer Materials, Led by Massachusetts Institute of Technology. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040636&HistoricalAwards=false. Accessed: Feb 28, 2022.

Page, G. 2020. Application of Sequential Inductive Transfer Learning for Experimental Metadata Normalization to Enable Rapid Integrative Analysis. http://www.nsf.gov/awardsearch/showAward?AWD_ID=2040521&HistoricalAwards=false. Accessed: Feb 28, 2022.

Rowell, E., Loudermilk, E., Hawley, C., Pokswinski, S., Seielstad, S., Queen, L., O'Brien, J., Hudak, A., Goodrick, S., and Hiers, J. 2020. "Coupling Terrestrial Laser Scanning with 3d Fuel Biomass Sampling for Advancing Wildland Fuels Characterization." *Forest Ecology and Management* 462:117945. https://doi.org/10.1016/j.foreco.2020.117945

Shang, J. 2020. Towards Intelligent Sharing and Search for AI Models and Data Sets. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040727&HistoricalAwards=false. Accessed: Feb 28, 2022.

Sztipanovits, J. 2020. Deep Monitoring of the Biome Will Converge Life Sciences, policy, and Engineering. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040688&HistoricalAwards=false. Accessed: Feb 28, 2022.

Tang, Z., Qin, Y., Jiang, Z., Krawec, W., and Zhang, P. 2021. "Quantum-Secure Microgrid." *IEEE Transactions on Power Systems* 36(2): 1250–63.

Turkan, Y. 2020. Rapid Development of Intelligent, Built Environment Geo-Databases Using AI and Data-Driven Models. NSF Award abstract. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040735&HistoricalAwards=false. Accessed: Feb 28, 2022.

Wang, L., Zhou, Y., Wan, W., Ye, H., and Zhang, P. 2021. "Eigen-analysis of Delayed Networked Microgrids." *IEEE Transactions on Power Systems* 36(5): 4860–3.

Xu, J., Glicksberg, B., Su, C., Walker, P., Bian, J., and Wang, F. 2021. "Federated Learning for Healthcare Informatics." *Journal of Healthcare Informatics Research* 5(1): 1–19.

Yoon, J., Jarrett, D. & van der Schaar, M. 2019. "Time-Series Generative Adversarial Networks." Advances in Neural Information Processing Systems, NEURIPS2019, Vol.32, 2019, pp.5508-5519. Editors: Wallach, H., Larochelle, H., Beygelzimer, A., Alch'-Buc, F., Fox, E., and Garnett, R.

Zhang, P. 2020. AI-Enabled Provably Resilient Networked Microgrids. https://www.nsf.gov/awardsearch/showAward?AWD_ID=2040599&HistoricalAwards=false. Accessed: Feb 28, 2022.

Zhang, P. 2021. *Networked Microgrids.* Cambridge University Press, Cambridge, England.

Zhou, Y., and Zhang, P. 2021a. "Neuro-Reachability of Networked Microgrids." *IEEE Transactions on Power Systems*. 37(1): 142–55. DOI: 10.1109/TPWRS.2021.3085706

Zhou, Y., and Zhang, P. 2021b. "Reachable Power Flow: Theory to Practice." *IEEE Transactions on Power Systems* 36(3): 2532–41.

Zhou, Y., Zhang, O., and Yue, M. 2021. "Reachable Dynamics of Networked Microgrids with Large Disturbances." *IEEE Transactions on Power Systems* 36(3): 2416–27.

AUTHOR BIOGRAPHIES

Chaitanya Baru is Distinguished Scientist, San Diego Supercomputer Center, UC San Diego. From 2014 to 2018, he was Senior Advisor for Data Science at the National Science Foundation where he provided leadership for several NSF data programs including BIGDATA, Big Data Hubs, TRIPODS, Data Science Corps. From 2019 to 2021, he was Senior Advisor for the Convergence Accelerator and a member of the team that established the program.

Michael Pozmantier is a Program Director in the NSF Convergence Accelerator. He has worked in the private sector and government, including leading the Transition to Practice Program in the Science and Technology Directorate at the Department of Homeland Security, focusing on the commercialization of federally-funded cybersecurity research.

Ilkay Altintas is Chief Data Science Officer at the San Diego Supercomputer Center, University of California San Diego, founder and director of Workflows for Data Science Center of Excellence and WIFIRE Lab, and Founding Fellow of Halıcıoğlu Data Science Institute. Her passion is leading collaborative multi-disciplinary teams to make computational data science work more reusable, programmable, scalable, and reproducible.

Stephen Baek is currently Associate Professor of Data Science at the University of Virginia with research interests in geometric data analysis. After receiving a PhD from Seoul National University, Korea in 2013 he served as a presidential postdoctoral fellow of Korea, and subsequently joined the University of Iowa as Assistant Professor of Industrial and Systems Engineering, from 2015 to 2021.

Jonathan Cohen is Professor of Neuroscience and Psychology, and Co-Director of the Princeton Neuroscience Institute at Princeton University. His research focuses on the neural mechanisms underlying the human capacity for cognitive control, and its role in natural intelligence, using neural network modeling and laboratory studies, with implications for both machine learning and disorders of brain function in neuropsychiatric disorders.

Laura Condon is an Assistant Professor at the University of Arizona. She studies large-scale water sustainability and watershed dynamics. Her work combines physically based numerical modeling with machine learning and other statistical techniques. She is a lead developer of a national modeling framework.

Giulia Fanti is an Assistant Professor of electrical and computer engineering at Carnegie Mellon University. Her research focuses on the security and privacy implications of data sharing, explored through the lens of machine learning, distributed systems, and networking.

Raul Castro Fernandez is Assistant Professor of Computer Science at the University of Chicago interested in the economics and value of data and on understanding how to make best possible use of data, including building systems that provide the capability to share, discover, prepare, integrate, and process data. His research employs techniques from data management, statistics, and machine learning.

Ethan Jackson is Senior Director and Principal Researcher, Microsoft Healthcare. His research focuses on intelligent systems that make people, and the environments around them, healthier. He directs the Microsoft Premonition project, aimed at detecting movement of potential pathogens in the environment before they cause outbreaks in humans. He joined Microsoft in 2007 after receiving a PhD in Computer Science from Vanderbilt University.

Upmanu Lall is the Silberstein Professor of Engineering at Columbia University, and the Director of the Columbia Water Center. His research focuses on spatio-temporal hydroclimatic prediction at multiple scales through novel statistical and data science algorithms in the context of the dynamic risk assessment and mitigation for water, energy, agriculture, urban, and financial systems.

Bennett Landman is Professor and Department Chair of Electrical and Computer Engineering at Vanderbilt University. His research concentrates on applying image-processing technologies to leverage large-scale imaging studies to improve understanding of individual anatomy and personalize medicine.

Hai "Helen" Li is Clare Boothe Luce Professor and Associate Chair of the Electrical and Computer Engineering Department at Duke University. Her current research focuses on neuromorphic computing systems, deep-learning acceleration and security, memory design and architecture, and high-performance and energy-efficient computing systems. She is a Fellow of ACM and IEEE.

Claudia Marin is Professor and Graduate Student Director in the Department of Civil and Environmental Engineering, Howard University, Washington, DC. Her research interests include computational mechanics, experimental evaluation of large-scale structural and nonstructural components, and structural health monitoring. She has worked on developing protective systems to improve performance of bridges, buildings, and their contents under seismic and other extreme loads.

Beatriz Martínez-López is Professor of infectious disease epidemiology and Director of the Center for Animal Disease Modeling and Surveillance (CADMS) at the School of Veterinary Medicine at the University of California, Davis. She focuses on the application of disease modeling, risk assessment, geostatistical methods, or network analysis to unravel complex animal health problems and study diseases at the wild-domestic–human interface.

Dimitris Metaxas is Distinguished Professor of Computer and Information Sciences at Rutgers University and Director, Center for Computational Biomedicine, Imaging and Modeling (CBIM). His research focuses on novel theories for segmentation, dynamic object tracking and recognition, model-based learning, sparsity, physics-based and deformable object modeling, human behavior and movement analytics, computer animation of fluid phenomena, scalable solutions to multidimensional, and distributed learning problems.

Bradley Olsen is the Alexander and I. Michael (1960) Kasser Professor of Chemical Engineering at MIT. His research interests include sustainable polymers, polymer informatics, polymer physics, and natural materials. He holds a bachelor's degree in Chemical Engineering from MIT in 2003, PhD in Chemical Engineering from UC Berkeley in 2007, and was a postdoc at Caltech from 2008–2009.

Grier Page is a senior fellow and senior director of statistical genetics at Research Triangle Institute International. His current research applies genomics to transfusion medicine as well as pre-, peri-, and postnatal diseases. Additionally, he is developing artificial intelligence and natural language processing methods to enable data integration and harmonization.

Jingbo Shang is Assistant Professor in the Department of Computer Science and Engineering and the Halicioglu Data Science Institute, University of California San Diego. His research areas are in data mining, NLP, and ML. He has been recognized for his work on constructing structured knowledge from massive text corpora with minimum human effort, with a Grand Prize from Yelp Dataset Challenge and a Google PhD Fellowship.

Yelda Turkan is an Associate Professor in the School of Civil and Construction Engineering at Oregon State University. Dr. Turkan's research focuses on remote sensing and information technology applications for construction engineering and management. She leverages tools such as BIM, lidar, and augmented reality to innovate planning, monitoring, and controlling construction operations, and improve decision making in the Built Environment.

Peng Zhang is a Professor of Electrical and Computer Engineering, and a SUNY Empire Innovation Professor at Stony Brook University, New York. He has a joint appointment at Brookhaven National Laboratory as a Staff Scientist. His research interests include AI-enabled smart grids, quantum-engineered power grids, networked microgrids, power system stability and control, cybersecurity, and formal methods and reachability analysis.

How to cite this article: Baru, C., Pozmantier, M., Altintas, I., Baek, S., Cohen, J., Condon, L., Fanti, G., Fernandez, R. C., Jackson, E., Lall, U., Landman, B., Li, H. H., Marin, C., Lopez, B. M., Metaxas, D., Olsen, B., Page, G., Shang, J., Turkan, Y., and Zhang, P. 2022. "Enabling AI innovation via data and model sharing: An overview of the NSF Convergence Accelerator Track D." *AI Magazine* 43: 93–104. https://doi.org/10.1002/aaai.12042