
NSF
NSF's Convergence Accelerator
2021
PORTFOLIO
GUIDE

TABLE OF CONTENTS

NSF's Convergence Accelerator

CONVERGENCE ACCELERATOR OVERVIEW

Solutions for today's national-scale societal challenges are hard to solve within a single discipline. Instead, these challenges require convergence to merge ideas, approaches, and technologies from a wide range of diverse sectors, disciplines, and experts.

Launched in 2019, the National Science Foundation's Convergence Accelerator builds upon research and discovery to accelerate use-inspired convergence research into practical application. The program funds a cohort of teams to work interactively toward solving grand societal challenges that impact thousands of people positively.

Funded teams begin in phase 1; a fast-paced nine-month hands-on journey, which includes the program's innovation curriculum, formal pitch, and phase 2 proposal evaluation. The innovation curriculum includes user discovery, human-centered design, team science, communication skills, and pitching, assists teams in developing their solution, and preparing the teams for the next phase. The program's team-based approach creates a co-opetition environment, stimulating the sharing of innovative ideas toward solving complex challenges together, while in a competitive environment to try and progress to phase 2.

Teams are comprised of disciplines and expertise from academia, industry, government, non-profit, and other

communities of practice. Disciplines include all science and engineering fields, but also other disciplines such as law, healthcare, communications, and business management to accelerate the solutions forward. As teams apply the Program's convergence research fundamentals and innovation processes, the teams' pioneering ideas are transformed along the journey—moving it to a proof-of-concept, then prototype, and finally a solution. Teams also develop partnerships to support their solutions toward sustainability and transition to practice.

The Convergence Accelerator is a unique NSF program. While the program is focused on advancing research toward societal impact; the program is intentionally developed around four key components to provide the highest impact. The four components include a convergence research approach, strong multi-organization partnerships, high-impact deliverables, and track alignment.

- **Convergence Research**—Each research effort includes a multidisciplinary approach to accelerate use-inspired research into practice in ways that benefits society at scale.
- **Partnerships**—Funded teams must create partnerships with many types of organizations from academia, industry, government, non-profit, and other sectors, to ensure that the research efforts are use-inspired and have a clear path to transition to practice. Each partnership is different and may provide needed expertise, represent end-users; or provide resources, services, and infrastructure to advance the solution forward.
- **Deliverables**—Deliverables or solution outputs can take many forms; such as hardware, software, data, services, processes, protocols, standards, and more, but each solution must provide a positive impact on society at scale.
- **Track Alignment**—Each funded effort must align to the program's identified convergence research track topics and have the potential for strong integration with other efforts.

To date, the Convergence Accelerator program is comprised of two cohorts. The 2019 cohort, launched in September 2019, is developing solutions in Open Knowledge Networks (Track A) and the Future of Work (Track B). The cohort included 43 phase 1 teams, but after a down-select features eight phase 2 teams. The 2019 Cohort teams are halfway through the 24-month phase. All teams have prototypes solutions, strong partnerships, and are in the process of developing a sustainability plan to ensure the solutions' impact.

The 2020 cohort, awarded in September 2020, includes 29 phase 1 teams developing solutions in two transformative research areas—Quantum Technology (Track C) and AI-Driven Data Sharing & Modeling (Track D). Over the last nine months, the teams worked to build a proof-of-concept for their solution, developed strong partnerships, participated in the Program's innovation curriculum, and completed the formal pitch and phase 2 proposal evaluation. Selected teams for phase 2 will continue to apply program fundamentals to develop solution prototypes and to build a sustainability model to continue impact beyond NSF support.

The 2021 cohort, to be announced in Fall 2021, is in the final planning stages. Awarded phase 1 teams will develop low-fidelity prototypes focusing on convergence research topics; the Networked Blue Economy (Track E) and Trust & Authenticity in Communication Systems (Track F).

More information about the Convergence Accelerator can be found on the NSF's website: www.nsf.gov/od/oia/convergence-accelerator

TRACK A:

OPEN KNOWLEDGE NETWORKS

Vast amounts of data are produced every day, yet many organizations lack the accessibility to draw insights from these data and make data-driven decisions. Knowledge networks (or repositories) with massive amounts of world knowledge help to power the next wave of AI exploration, driving innovations from scientific research to the commercial sector. Knowledge networks/graphs provide a powerful approach for data discovery, integration, and reuse, but require an investment in their creation and maintenance. Today, only the biggest tech companies have the resources to develop and exploit significant knowledge graphs and networks.

To enable data to be freely accessible, especially to government, academia, small business, and nonprofits organizations, NSF's Convergence Accelerator is funding the creation of nonproprietary infrastructure for building Open Knowledge Networks (OKNs). Using artificial intelligence and machine learning, teams are building infrastructure, tools, and applications to identify data, link data points, describe relationships, and gather information at speed and scale—providing data to knowledge, knowledge to insights, insights to understanding.

The OKNs connect people, events, places, environments, health, and more, removing domain boundaries, linking data, its attributes, and relationships to other data to be accessible for decision-makers, analysts, researchers, and the American public to answer interesting and pressing questions. Currently, teams are focusing on urban flooding, judicial court records, biomedical health, geospatial information, and technology infrastructure for knowledge network creation and use.

The Open Knowledge Networks phase 2 efforts include:

AI and Machine Learning Infrastructure Tools and Applications

- **OKN Infrastructure**—Led by the University of Michigan/MIT, the OKN Infrastructure is building infrastructure for constructing novel OKNs and OKN-powered applications. This solution provides tools to make the creation and maintenance of high-quality datasets and apps more cost-effective and more widely accessible.
- **KnowWhereGraph (KWG)**— Led by the University of California, Santa Barbara, KWG provides knowledge graph and geo-enrichment services for environmental intelligence applications. The solution enriches data with pre-integrated custom-tailored knowledge about any locale of interest, thereby reducing the time to find, combine, and reuse data. The initial application areas are focused on decision support related to food systems, supply chains, and humanitarian aid, but can easily be expanded to other application areas as well.

Domain-based Open Knowledge Networks

- **SCALES**— Led by Northwestern University, the SCALES open knowledge network is designed to be a public resource to help provide insights based on judicial court records. SCALES is creating tools to decode court records and transform this data into actionable information that aids a variety of uses, including legal scholars, journalists, policymakers, judiciary, and citizens.
- **Urban Flooding Open Knowledge Network (UF OKN)**—Led by the University of Cincinnati, the UFOKN is addressing urban flooding impacts to assist decision-makers and urban planners in real-time response and long-term planning.
- **Biomedical Open Knowledge Network**—features SPOKE, an open knowledge network being developed by the University of California, San Francisco, connects millions of biomedical facts including molecules, pharmacological compounds, organs and diseases, food nutrients, and more. Centered around knowledge representation and reasoning, the team is developing applications using graph theory, advanced visualizations, and real-world clinical evidence to advance drug development and precision medicine.

Integrating the Knowledge Networks

Data2Knowledge Consortium—Knowledge graphs are rapidly emerging as key infrastructure to integrate the diverse information needed to solve complex societal challenges—from climate change and human health to capturing business value from the AI revolution. The Open Knowledge Network phase 2 teams are collaborating on “track integration” to create the Data2Knowledge Consortium to ensure that the outcome from the Convergence Accelerator Track is “greater than the sum of the parts”. Comprised initially of the current Open Knowledge Network phase 2 teams, the objective of the Data2Knowledge Consortium is to facilitate a thriving ecosystem for open knowledge graph development and use.

Biomedical Open Knowledge Network

Connecting the Dots to Advance Precision Medicine for Everyone

Lead PI: Sergio Baranzini
sergio.baranzini@ucsf.edu

Sui Huang
sui.huang@isbscience.org

Sharat Israni
Sharat.Israni@ucsf.edu

Overview

The human brain cannot possibly integrate the vast and rapidly growing amount of information modern societies have been able to amass. This hampers the generation of new knowledge, specifically in the biomedical sciences and its implications for human health, where the subject complexity is vast and stakes are high. Our knowledge network will incorporate billions of factual relationships among biomedical concepts, providing a discovery engine that will enable doctors, researchers, the pharmaceutical industry and the citizen scientist to explore biomedicine in its whole might.

Description

Human health has become so complex that even doctors turn to Google to understand difficult cases. Then, based on their extensive training, they can better interpret, diagnose and treat illnesses. Clearly, Google is neither based purely on accepted science nor is it specialized enough to handle rare or complex conditions, find a cure for a given disease, or discover the root of a biological process. For those cases, specific information needs to be integrated by skilled researchers into formulating the right hypothesis (consistent with previous evidence and maximizing utilization of current knowledge), which then needs to be tested experimentally.

The response to each of the previous scenarios requires navigating a deluge of complex data and information and connecting the dots in a meaningful way. Our knowledge network (SPOKE) integrates millions of biomedical concepts into a knowledge engine to enable doctors, drug developers and citizen scientists connect the dots and produce a biologically meaningful answer to these questions. Healthcare and related industries represent 1/5 of the entire US economy. We recently incorporated Mate Bioservices, a company that will commercialize applications of SPOKE. We anticipate extensive adoption of this platform will have a significant societal impact by reducing healthcare costs, health disparities and accelerating therapeutics, ultimately improving the quality of life for every American.

Differentiators

Even as high-throughput modern technology in biomedicine has facilitated the acquisition of vast amounts of data, it has only widened the chasm between its generation and its interpretation. Those approaching such a complex task based on individual strengths are bound to fail. While few efforts have been devoted to addressing this vacuum, we are pioneering the paradigm of Knowledge Networks in Biomedicine - a paradigm amply proven in Web Search - into a discipline that is inherently graph-theoretic. Our experience with systems biology, graph theory (LLNL), PI participation in NCATS Biomedical Translator, and the long track record of creating SPOKE are testaments to our vision and commitment to transforming data into knowledge. This has been acknowledged even by the NIH Office of Data Science Strategy.

Mate Bioservices aim to revolutionize the way we conduct research in the healthcare industry through: i) harmonizing biomedical research into a central knowledge network (KN); ii) designing the AI/ML tools needed to answer complex scientific questions; and iii) integrating KN and AI/ML tools into sophisticated user interfaces that are specifically designed for various customer profiles.

Road Map

As part of our phase 1 program, we have already developed and made available a fully functional biomedical KN. Our phase 2 proposal involves continuous development of the engine and 4 products powered by the KN throughout the 2-year period with a commercialization plan to ensure sustainability. Over the first year of phase 2, we have established a Governance Committee for SPOKE, expanded the KN with additional knowledge sources, developed a prototype of the open access network visualization tool, secured rights for commercialization of SPOKE-powered products, and completed a number of analyses that will validate the network's quality and utility. In the remaining year of phase 2, we will:

- Apply for non-dilutive funding. **(Q3, 2021)**
- Secure first contract with Pharma via Mate Bioservices **(Q4, 2021)**
- Release a web-accessible network visualization tool aimed at citizen scientists. **(Q1, 2022)**
- Produce a report outlining recommendations for mitigating risk associated with ethical, legal, and social implications of network use. **(Q3, 2021)**
- Obtain seed funding. **(Q1, 2022)**
- Deploy the first clinical decision support system (alpha) in the UCSF neurology practice with real patients **(Q2, 2022)**
- Develop and apply computationally intensive analyses to the network, enabling sophisticated validations of the network against real-world observations. **(Q2, 2022)**

Partnerships

Institute for Systems Biology will continue to work with us both in the scientific development of the KN and with resources to expand its utility and reach.

Lawrence Livermore National Laboratory will continue leveraging its extraordinary computing power and technical knowhow on graph theory and analytical approaches to guarantee an organic and balanced growth of the graph.

Indiana University: A partnership borne from the Convergence Accelerator phase 1 (Team B6656/7036), Katy Börner's team at Indiana University will lend their world-class expertise in complex graph visualizations and analysis to develop an open access network visualizer aimed at citizen scientists.

University of California, San Diego: Another partnership formed in phase 1 of the Program, Peter Rose from the San Diego Supercomputer Center will bring expertise in protein domain and structure, as well as geolocation information, to further expand the network.

University of California, San Francisco Innovation Ventures: Having already provided seed funding for the establishment of Mate Bioservices, they will continue to provide support for the handling of intellectual property matters.

National Center for Advancing Translational Sciences: Our KN is an active participant of the NCATS Translator Program as an Autonomous Relay Agent Team.

ELSI: In phase 2 we have partnered with Camille Nebeker (UCSD) and Erin Kenneally (Elchemy) to formally advise on ethical legal and social implications of our project.

NASA: We have secured a strategic partnership with NASA's GeneLab to process biological datasets acquired during spaceflight with SPOKE and will programmatically link the two platforms.

Intellectual Property

Inventions disclosures related to uses of the network (not the network itself) have been filed with UCSF Innovation Ventures. This will ensure proper documentation of the resulting IP from this project for those application areas.

KnowWhereGraph

A Knowledge Graph for Environmental Intelligence Applications

Lead PI: Krzysztof Janowicz
janowicz@ucsb.edu

Pascal Hitzler
hitzler@ksu.edu

Wenwen Li
wenwen@asu.edu

Dean Rehberger
rehberge@msu.edu

Mark Schildhauer
schild@nceas.ucsb.edu

Overview

We are developing a cross-domain knowledge graph for environmental intelligence applications. Our KnowWhereGraph supports data-driven analytics and decision making by answering questions such as “What is here” or “What happened here before?” anywhere on Earth. A series of geospatially-aware graph analytics tools and services, including GeoEnrichment services, graph integration and visualization tools are being developed to investigate applications such as disaster relief, agriculture and food supply chains, and enable rapid forecasting of environmental change and its potential health impacts across multiple scales.

Description

Geospatial data and the locations of places and events on Earth--are critical in understanding where and when things happen and, more important, why they happened or will happen. Our GeoEnrichment service is a powerful solution that provides on-demand access to area briefings at a high spatial and temporal resolution for any location on the Earth’s surface. Today, nearly 80% of the time invested in a data-intensive project is still spent on data discovery, retrieval, entry, cleaning, and apportionment. This significantly hinders the rate of data-driven decision making. Our KnowWhereGraph solution addresses this challenge by providing (1) an open knowledge graph that interlinks cross-domain decision data and (2) a set of geoenrichment services that enable ready access to well curated, location-aware graph data.

The team has made significant progress in Year 1 in developing the seed graph using a common schema that connects multi-source data in terms of disaster, air quality, climate hazards, crop history, experts and expertise, administrative boundaries etc. The statements in the graph have excitingly exceeded 300 million. We expect this number to continue growing as more automated graph generation and integration approaches are being developed.

Our team are also pioneers in developing spatially-explicit machine learning models to provide GeoAI-ready data to empower intelligent decision making. In Year 1 of the project, we have centered our efforts in supporting project verticals including the disaster relief subteam to assemble quickly needed datasets for rapid disaster response and evacuation after major devastating events, such as hurricanes, have occurred. We are also developing graph solutions for understanding and sustaining food supply chain resilience. As a technology-driven project, our goal is to demonstrate how novel geospatial solutions can inform downstream stakeholders from industry, nonprofits, and government agencies.

Differentiators

Spatial is special. Our team provides unique expert- ise in representing and integrating geospatial data using knowledge graph technologies and GeoAI-based services: We have contributed to international semantic standards; our partner Esri offers the market’s most advanced spatial analytics technology, while Oliver Wyman and partner companies provide expertise in applying remotely sensed

imagery and machine learning models to global food markets, supply chains, and farms. The academic team has a track record in technology transfer, the development of vocabularies, and lifting geospatial data to the graph, while our partners NCEAS, USGS, and USDA are among the largest providers and integrators of geospatial data. This combination enables us to rapidly develop our enrichment services, connect them to vast amounts of data, and apply them to real-world applications, including disaster relief and food supply chains.

Road Map

Year-1 milestones: (M1.1) We developed a knowledge graph with data at human-environment nexus, including remotely sensed imagery. (M1.2) Published semantically-enabled schema for data alignment and deduplication (M1.3). A developed a GeoEnrichment prototype (MVP) that integrates into Esri's ArcGIS. (M1.4) Prototype and graph tested with disaster relief, food systems and supply chain pilots. Existing GeoEnrichment tools will be used as a baseline to evaluate success. **Year-2 milestones:** (M2.1) Refined graph and schema with inputs from external partners. (M2.2) Enhanced end-user interface and visualization tools. (M2.3) Broadened GeoEnrichment beyond a GIS by developing APIs for dashboards and question answering. (M2.4) Deployment to broader applications and use cases. (M2.5) Final services and graph; established public-private partnership.

Partnerships

Academia Partners
University of California, Santa Barbara's Center for Spatial Studies, National Center for Ecological Analysis and Synthesis (NCEAS), Climate Hazards Center, KSU's Center for AI and Data Science, and ASU offer expertise in knowledge engineering, GeoAI, data synthesis, and environmental modeling. MSU contributes expertise in precision and digital agriculture and complements our data with a historic perspective on soils and slavery (Matrix). USC will contribute expertise in environmental economics with a focus on air pollution on labor.

Industry Sector Partners
Esri will provide expertise for developing and testing the knowledge graph-based GeoEnrichment services. Oliver Wyman and start-ups will test project capabilities with its customers in supply chain opti- mization and commodity markets.

Nonprofit Partners
Food Industry Association: will provide access to the food industry with a focus on sustainable agriculture. Direct Relief will apply project work to humanitarian aid supply chains.

Government Partners
USGS will provide expertise in lifting its National Map portal data to the graph, while USDA ARS & NRCS will provide data and expertise about sustainable agriculture and soil data.

Intellectual Property

We are committed to openness and will release products under a permissive BSD-3 license for the software, and CC0 or CC-BY for data. We will adhere to semantic technologies standards by W3C and OGC. While encouraging the least restrictive licensing, data or software licensed under more restrictive conditions will be accommodated to allow for broad participation from the industry.

OKN Infrastructure

Knowledge Network Infrastructure with Application to COVID-19 Science and Economics

Lead PI: Michael Cafarella
michjc@csail.mit.edu

Oren Etzioni
orene@allenai.org

Matthew Shapiro
shapiro@umich.edu

Overview

Knowledge Networks are a novel and potentially transformative form of data, but building applications on top of them is too difficult, time-consuming, and expensive. We are building a Knowledge Network Programming System that makes it far easier to build novel knowledge-powered applications, while also improving the knowledge resources themselves.

Description

Knowledge Networks like Wikidata are a compelling new type of data – akin to a "structured world wide web" – that have enabled new applications, such as structured web search and voice assistants. Unfortunately, only the most technically sophisticated organizations have had the resources to build these difficult-to-engineer applications. As a result, most of these next-generation applications never actually get built, and users cannot benefit from them. Our Knowledge Network Programming System will dramatically reduce the cost of building knowledge-powered applications by combining programming tools with recent advances in data management and machine learning. This will happen in two ways. First, it will make Knowledge Network data items easy to use by integrating them directly into a client programming language, much as hashtables and other standard data structures are built into every programming language today. Programmers will be able to assume that high-quality data is "built-in" rather than an additional load-and-integration step. Second, it will take the social curation methods pioneered by Wikidata and other knowledge networks and apply them to all kinds of data, not simply graph-structured information. This system will combine elements of Wikidata and a traditional database system. It makes standard data objects (tables, files, functions, and so on) amenable to Wiki-style social versioning, improvement, and sharing. With these two thrusts, the Knowledge Network Programming System makes data easier for developer to use and easier for developers to curate in the first place. Source code for our prototype system is available online.

We are testing the system using several Knowledge Networks developed as part of this project. The first is the CORD-19 network, which describes over 500K scientific papers on COVID-19 and related historical coronavirus research. This is already public and has been covered in the Wall Street Journal, the New York Times, and elsewhere.

The second is a network that describes macroeconomic statistics in the United States, including recent budgetary and economic responses to COVID-19. We believe these networks will be useful for our project at the same time they support national priorities.

Concretely, this project will yield new software, in the form of the programming system and toolset. It will also yield novel data resources, in the form of the above Knowledge Networks. Finally, it should yield novel applications that both illustrate the programming system and are useful on their own.

Differentiators

All Knowledge Network applications that we are aware of rely on traditional software engineering tools. We are unaware of any system that addresses application development per se, even though other data types (say, relational databases) have extensive dedicated tooling. The programming system is unusual in its application of data management and

machine learning methods to goals traditionally associated with programming languages.

The Knowledge Networks help with validating our system and with crucial social needs, but are especially notable for the agility with which we can create them. The CORD-19 Knowledge Network was initially released in March, 2020. In the first year of this program we have made several advances that make new Knowledge Networks faster and easier to develop, especially those extracted from document corpora.

The research team is unusual in its level of experience with shipping data development systems, knowledge network production, and relevant domain expertise. PI Michael Cafarella is a principal research scientist at MIT CSAIL. He has published on databases and is one of the co-creators of the Hadoop system. PI Oren Etzioni is the CEO of the Allen Institute for Artificial Intelligence. The Allen Institute is a nonprofit research organization that arguably employs the largest set of Knowledge Network engineers outside a major tech firm. PI Matthew Shapiro is the Lawrence R Klein Collegiate Professor of Economics at Michigan, and an expert in macroeconomics and public finance. He serves as chair of the Federal Statistics Advisory Committee.

Road Map

There are several core deliverables for this work: **(1)** the programming system, **(2)** the scientific and economics Knowledge Networks, **(3)** infrastructure used to produce those networks, and **(4)** application code built to demonstrate the programming system.

The programming system is under active development and a useable prototype is online. The Allen Institute has released regular updates to CORD-19 and continues to do so. The economics Knowledge Network is under development now. The infrastructure software is under development; we have front-loaded our budget to accelerate this work so it can be useful to other teams during the NSF program. We are translating the working data pipeline for CORD-19 to run entirely using the programming system, both to test the system and to exploit the new features it enables.

Partnerships

In the long term, we believe the system potentially has many users in the scientific, policymaking, and corporate worlds. We are collaborating with several large institutions, including financial and engineering firms, to test the system. We have spoken with several venture capital firms about ensuring the project and funding beyond this program.

Additional Senior Personnel from the University of California, Berkeley are working on front-end software for adding data to a Knowledge Network. Personnel at the University of Washington, Seattle are working on integrating the programming language with back-end data resources. Finally, we are also working with researchers in the social sciences to evaluate the system's ability to help in varied research settings.

Intellectual Property

Code and data created for this project has been and will be released into the public domain. The only exceptions will be when some purchased source datasets (say, containing certain economic statistics) have restrictions that prevent us from doing so.

SCALES

Transforming the Accessibility and Transparency of Federal Courts

Lead PI: Luis A. Nunes Amaral
amaral@northwestern.edu

Charlotte S. Alexander
calexander@gsu.edu

Rachel D. Mersey
rdmersey@northwestern.edu

Adam R. Pah
a-pah@kellogg.northwestern.edu

David L. Schwartz
david.schwartz@law.northwestern.edu

Overview

The U.S. court system produces millions of records per year. These records are supposed to be open to the public, but in practice they are trapped behind paywalls and dysfunctional, outdated software. Our project's goal is to enable a broad spectrum of public stakeholders to efficiently access, evaluate, engage with, and understand the work of the courts. Our mission is to create an Open Knowledge Network (OKN) that will serve as a foundation for advocates and researchers to analyze court data systematically. Our platform will also enable data gathering and integration that, in turn, supports intelligent analysis and meaning extraction so all citizens, entrepreneurs, journalists, lawyers, potential litigants, policy makers, scholars, and even the judiciary itself, can better understand and evaluate how the courts function.

Description

Data drives information and insight. Government agencies, central banks, health organizations, and law enforcement agencies all gather data in order to better understand and communicate the events, trends, and relationships between them that define our world. All of this is in service of the twin goals of understanding and transparency.

While the US court system collects similar data sets, those data are rarely used in support of the goals of understanding and transparency. This shortfall is the result of three features: availability of data, lack of data integration, and limited tools that support intelligent information analysis for non-technical users. Availability is hampered by the fact that much of the relevant court data resides behind a federal pay-for-use firewall. While individual case information is affordable, the data needed to do system level analysis would cost tens of millions of dollars to access. Integration is limited by data availability as well as the tools and skills required to support this work at scale. Understanding and insight, even if one had full access to the documents, is thus hindered by the lack of available tools for the journalists, legal scholars, and decision-makers who most need it.

With these issues in mind, we have brought together a team of computer and data scientists, legal scholars, journalists and policy experts, to develop a suite of tools to enable access to court records and analytics. The Systematic Content Analysis of Litigation EventS Open Knowledge Network (SCALES OKN) aims to provide access to both the data and the insights contained within them to a broad range of users with diverse technical skills. Our project encompasses five complementary elements. Data Access: We will make all of the data we incorporate into SCALES and the integration of such data freely accessible to the public. Data Integration: We will provide tools to upload relevant data sets (e.g., judicial, firm, and corporate profiles) and support data cleaning, normalization, and integration. These tools will include intelligent data interpretation supported by Natural Language Processing, Machine Leaning and crowdsourcing of instance tagging. The aim of this integration is the development of a dynamic knowledge graph that supports information access, analysis, and inference. Extraction of Insight: We will develop a public facing information access system that allows

users to explore the data to answer questions about trends, comparisons, and correlations by simply asking questions. Expanding on work in language processing, information goals, and intent guided analytics, the system provides access to not just the data, but the information and insight contained within. Education: We will develop materials aimed at providing users with an understanding of both the data, the integration, and analytics methods. Community: We will foster the nucleation and organic growth of a community of stakeholders that both add to and use the data and the system supporting it.

Differentiators

There are limited alternative sources of judicial records and none that enable systematic access and analysis for the public. Commercial legal services such as Westlaw, LexisNexis, and others have purchased many judicial records, but they limit access through their own fees and prohibit bulk downloads—foreclosing systematic analysis even for those who pay the hefty fees. A pioneering open alternative, The Free Law Project, maintains a user-generated free repository of court records, but it is not designed to support systematic analysis.

We are unique in that we are providing users with not just access to raw data but also the meaning that it supports through access to intelligent analytics learned from across the legal system.

Road Map

During **year one of phase 2** we have conducted extensive user tests and refined the user interface of the prototype based on that feedback. We implemented named entity recognition to further enrich the court record data with links to judge characteristics, nature of entity (i.e., business or government), and statutes—to aid in contextualizing the entities within these records. Further, we are developing an ontology of litigation events in order to systematically identify what occurs in a case.

In **year two**, we will expand the core technology and facilitate organic growth by helping external contributors to add data and annotations to SCALES. We will also expand on query/analytics capabilities following user needs.

Partnerships

During phase 2, we will engage partners in academia, law, industry, and journalism. We are forming an advisory board comprising prominent experts to evaluate technical, legal, and ethical aspects of the SCALES OKN. The Free Law Project will continue to contribute data and will serve on the SCALES advisory board. Our other partners (including the MacArthur Justice Center; American Bar Foundation; Jenner & Block; The Center on Wrongful Convictions; NYU Law, Technology, and Policy Clinic; and more) have contributed their time to help us define SCALES as a product and will continue to provide their time in testing SCALES as it is developed and expanded. We will also work with industry partners to identify corporate filings and collaborate on knowledge graphs that connect litigation with existing laws.

Intellectual Property

We plan to license for free all intellectual property we create under a standard open source license.

Urban Flooding OKN

Delivering Flood Information to AnyOne, AnyTime, AnyWhere™

UF-OKN

Lead PI: Lilit Yeghiazarian
lilit.yeghiazarian@uc.edu

Sankar Arumugam
sankar_arumugam@ncsu.edu

Ximing Cai
xmcai@illinois.edu

Torsten Hahmann
torsten.hahmann@maine.edu

Venkatesh Merwade
vmerwade@purdue.edu

Overview

On average flooding causes more than 100 billion dollars of economic loss and 500 deaths per year in the US. Some of this could be avoided if people had access to flood-related information such as flood inundation depth at any location or evacuation routes during a flood event. We will create a Google-like application that people can use to find out how flooding impacts them so they can take actions to mitigate it.

Description

Recent studies have shown that approximately 41 million people (~13% of the population) are at risk of severe flooding in the U.S. Despite such high risk of flood exposure, it is hard to find answers to simple questions such as "What is the total impact of flooding on a city?". This is because, while urban infrastructure is connected, data and models that describe them are not. We call this connected urban infrastructure the Urban Multiplex. It includes the power grid, transportation network, surface water and groundwater systems, storm water and sewage systems, drinking water systems, inland navigation and dams, all intertwined with the socioeconomic and public health sectors that form the fabric of modern cities. So, when one part of the Urban Multiplex fails from a flood, its cascading impacts across the city are unknown.

Our project aims to address the issue of quantifying flood impacts on an Urban Multiplex by developing a publicly accessible National-scale Urban Flooding Open Knowledge Network (UF-OKN). The proposed UF-OKN will connect multiple datasets, tools and models across the Urban Multiplex to ascertain and forecast the true impact of flooding. Product design, implementation and delivery are guided by user needs and partnerships with local, state and federal agencies and private industries. The UF-OKN can be used to answer questions such as: "which route can I take to work during a storm?" or "will my house lose power during a storm?" At an organizational level, a decision maker can ask questions such as: "Which neighborhoods and when to evacuate to minimize human loss"? or "How flood risk will change in my city in the next 20-30 years"?

We expect that the UF-OKN will directly and indirectly serve millions of people impacted due to flooding - by providing the necessary tools and resources to enable real-time response and long-term planning and decision making.

Differentiators

Currently, flood related information is available from different sources - and in different forms. The two most common ones are flood insurance rate maps (FIRMS), produced by FEMA and available as static GIS files or paper maps; and dynamic flood forecasts and water levels provided by the U.S. Geological Survey and the National Weather Service. All these datasets are available only through the respective agencies and require some knowledge of how to navigate their systems. Additionally, these datasets cannot be easily integrated to create a holistic view of a flood impact on an Urban Multiplex at different

temporal and spatial scales.

UF-OKN integrates flooding information with other related datasets in an Urban Multiplex so users ranging from an individual home owner to decision makers can get answers to their questions through a simple user interface or map on a mobile device or computer.

Road Map

Phase 2 will deliver key technologic and products to serve the needs of two User Archetypes representing emergency responders and federal planners involved in planning and coordination to mitigate flood impacts. Our Strategic Framework consists of five Planes: User Relations & Product Development; Technology Development; Research & Development; Transfer to Practice & Sustainability Model Development; and Urban Flooding Scientific Community Development. Activities and milestones achieved in each Plane inform decisions throughout the Strategic Framework.

Month 8 Milestones: partner data, models & results ingested into UF-OKN; live demo for focus groups with key partners; knowledge models running; demos ready for beta testing; initiate branding. Deliverable: Minimally Viable Product (MVP)

Month 12 Deliverables: potential user/buyer feedback; operational infrastructure in place; flood impact socioeconomic assessment.

Month 18 Milestones: draft business canvas; refined UF-OKN ready for deployment; partner training, UF-OKN testing, add new features/ functionalities; collect potential user/buyer feedback.

Month 24 Milestones: UF-OKN opens to general users; soft launch. Deliverables: Final business canvas.

Final Deliverable: Viable product/service, market evaluation.

Partnerships

Phase 1 partners (Cities of Wilmington, Raleigh (NC), NC Dept of Public Safety; two FL counties) contributed personnel time, data, model results, prototype testing, evaluation and feedback. All phase 1 partners will adopt UF-OKN for their operations, thus becoming our first users.

New phase 2 partners will contribute at the same level as phase 1 partners. They include U.S. Environmental Protection Agency (National repository of underground and above ground fuel storage tanks; National drinking water infrastructure data; Joint modeling of flood impacts in these critical facilities; real-time sensing data); U.S. Geological Survey (real-time sensing data; will participate in sustaining UF-OKN past phase 2); National Oceanic and Atmospheric Administration (personnel time); six additional FL counties.

Intellectual Property

We anticipate that significant intellectual contributions will result from this project. Given the collaborative and multi-institution nature of our team, IP created by this project may be jointly owned. Participating project team members have agreed to work together on the protection, maintenance, and commercialization of any jointly owned IP according to applicable laws and policies.

Data2Knowledge Consortium

Integrating Knowledge Across Domains to Address National Scale Challenges

Sergio Baranzini
sergio.baranzini@ucsf.edu

Michael Cafarella
michjc@csail.mit.edu

Krzysztof Janowicz
janowicz@ucsb.edu

Luis A. Nunes Amaral
amaral@northwestern.edu

Lilit Yeghiazarian
lilit.yeghiazarian@uc.edu

Overview

The Data2Knowledge Consortium—a combined activity of the five Open Knowledge Network projects in Track A of the NSF Convergence Accelerator—seeks to facilitate an open ecosystem for knowledge graph development, use, and maintenance. Knowledge graphs are rapidly emerging as key infrastructure to integrate the diverse information needed to solve complex societal challenges - from climate change and human health to capturing business value from the AI revolution. The goal of the Data2Knowledge Consortium is to bring the power of knowledge graphs into the hands of the public via an open system approach.

Description

Open knowledge networks are poised to fuel the next wave of AI exploration—enabling insights from massive amounts of world knowledge and driving innovations from scientific research to the commercial sector. The Open Knowledge Network (OKN) Track of the Convergence Accelerator—including the Data2Knowledge track integration activity—envisions an open, inclusive, community-oriented graph structure as a trustworthy knowledge infrastructure that facilitates and empowers a host of applications and opens new research avenues.

The recently released Final Report from the National Security Commission on AI calls for the establishment of "a National AI Research Infrastructure composed of cloud computing resources, test beds, large-scale open training data, and an open knowledge network that will broaden access to AI and support experimentation in new fields of science and engineering". The Data2Knowledge Consortium is a key building block for realizing this vision. Each of the OKN projects is bringing together data from existing specialized information sources, including government data platforms, and incorporating additional real-world knowledge and context. They are each adding unique value in their respective sectors: enabling new capabilities in biomedicine, urban infrastructure and flooding, understanding the federal judiciary; as well as developing new capabilities in representing geospatial information and context, and creating tools, technologies, and programming environments to support conversion of data into knowledge graphs.

Differentiators

While massive knowledge graphs have been created and are in use in proprietary applications, an open platform for heterogeneous data integration using knowledge graph technologies does not yet exist. This track integration activity brings together the complementary knowledge networks, tools and technologies being created by the five teams in Track A of the Convergence Accelerator to create the beginnings of an integrated Open Knowledge Network. User-centered design is also a major differentiator across the Convergence Accelerator projects. Projects are driven by use cases, rather than being technology-driven.

Road Map

The Data2Knowledge Consortium serves as a welcoming organization that employs an open system approach to help create a

growing, integrated, continuously updated OKN that hosts open, curated data to support development of AI applications; provide open transparent access to data whether the data are free or behind a paywall; empower the larger community to extend the OKN with additional information and knowledge from new sources and new domains; and provide training and, eventually, certification in the development and use of knowledge graph technologies.

The Consortium seeks ways to ensure the sustainability of this endeavor through a variety of partnerships, including public-private efforts, so that the OKN can persist well into the future, for as long as it is needed.

Partnerships

Each Track A team has its own extensive network of partners and collaborators who will be available to contribute to the Data2Knowledge Consortium. As an organization with an open approach, the Data2Knowledge consortium welcomes members and partners from all sectors interested in creating the Open Knowledge Network—from academia, industry, government, non-profit, and others.

Intellectual Property

Each entity participating in the consortium maintains all claims to existing IP, including proprietary data. Tools that are developed as part of the integration activities will be licensed to the consortium, while each team will retain IP developed as part of their individual efforts.

TRACK B:

FUTURE OF WORK

The world's technological advancements in AI, machine learning, and robotics are irrevocably shifting the future of work in unanticipated ways. NSF's Convergence Accelerator is focusing on solutions to train, reskill, upskill, and prepare the current and future workforce with industry needs and jobs of the future, as well as build a national talent ecosystem to stimulate the U.S. workforce and ensure continuing global competitiveness.

Teams comprised of academia, industry, non-profits, and end-user partners are converging together to develop disruptive future of work solutions that envision a positive national-sale societal impact whereby technology is utilized to create a STEM talent pipeline relevant to industry needs, keep workers safe and help them perform their jobs better, create new jobs, and facilitate accessibility and inclusivity. Solutions include developing the U.S talent pipeline through competency-based training intelligent tools to connect academic institutions with industry needs to prepare students for the workforce, improving workforce training and safety for emergency responders and manufacturing industry workers through human augmentation, and creating VR/AR tools to identify unique skills of neurodiverse individuals thus preparing them to thrive in the workforce.

The Future of Work funded phase 2 projects include:

- **SkillSync**— Industry 4.0 is changing the skills that workers need and companies require, leaving businesses vulnerable and colleges behind. SkillSync, led by Eduworks Corporation, uses AI and national skills data to help companies identify required skills, connect them with college continuing education departments, and enable colleges to respond with efficient, effective, and equitable reskilling programs.
- **NeuroAI@Work**—Led by Vanderbilt University, the NeuroAI@Work is a suite of AI-driven tools to support autistic individuals to successfully enter and contribute to the American workforce. The solution is a safe and effective tool for skill assessment, upskilling, independence, on-the-job support, and professional development. The team is comprised of engineers, psychologists, and business experts.
- **LEARNER**—Led by Texas A&M, LEARNER is an agile and adaptive Human Augmentation Technologies (HAT) integrated Emergency Response (ER) training platform that accelerates HAT adoption for safer and more efficient ER work, supports adaptive learning sensitive to ER workers' socio-technical opportunities and budgetary constraints, builds and retains skilled ER personnel, and accelerates next-gen workforce development.

Integrating the Future of Work Ecosystem

To guarantee the convergence research Future of Work track focus is greater than the "sum of its parts", phase 2 team solutions converge toward track integration, creating STEP UP the Skills-Based Talent Ecosystem Platform for Upskilling.

Comprised of the Future Work teams, STEP UP includes connecting the skills and talents of individual workers to the opportunities that most need them. By inclusively engaging America's human skill and talent, and the technologies that support, augment, and develop that talent, the group is ensuring every American may partake in the benefits of a thriving economy and the dignity of meaningful work.

B2: LEARNER

Learning Environments with Augmentation and Robotics for Next-gen Emergency Responders

Lead PI: Ranjana Mehta
rmehta@tamu.edu

E. Du
eric.du@essie.ufl.edu

J. Gabbard
jgabbard@vt.edu

A. Leonessa
aleoness@vt.edu

D. Srinivasan
sdivya1@vt.edu

Overview

The COVID-19 pandemic reinforces that Emergency Response (ER) workers do dangerous work while adapting to novel situations. Training institutes and organizations have reported a steep decline in ER trainings that are essential to the nation's critical infrastructure. A critical need exists to accelerate ER expertise development through adaptive, personalized learning platforms that deliver next-generation skills while integrating emerging human augmentation technologies (HATs). LEARNER, an agile and adaptive HAT-integrated ER training platform, will accelerate HAT adoption for safer and efficient ER work, support adaptive learning sensitive to ER workers' socio-technical opportunities and budgetary constraints, build and retain skilled ER personnel, and ultimately accelerate next-gen workforce development across other industry domains.

Description

LEARNER is a novel mixed-reality learning platform that has physical, augmented, and virtual reality components, where ER personnel will learn to work effectively with two HAT classes: powered exoskeletons (EXO) and head-worn AR interfaces (AR) for two ER skills curricula (e.g., Triage and Patient Handling). These HATs will showcase the modularity of LEARNER across physical and cognitive augmentation that have distinct learning requirements. We will develop, integrate, and assess EXO and AR learning modules into the LEARNER system across different access levels (home to field house to training centers). Our industry partner SARCOS Robotics has designed an upper-body EXO emulator interface integrated into the core LEARNER system. Concurrently, we remain engaged with our other industry partner Knowledge Based Systems, Inc. (KBSI), who are developing a working LEARNER prototype that utilizes a unique paradigm for learning by adapting to a variable set of learners' characteristics and contexts, through the incorporation of physiological, neural, and behavioral markers of learning into real-time AR/VR scenario delivery. Finally, we plan to test and evaluate the HAT-integrated LEARNER prototypes at a National ER Training Center (Texas A&M Engineering Extension Service; TEEX) in close guidance from our government partner (National Institute for Standards and Technology; NIST). Completion of customer/market needs assessment and determination of LEARNER business model will move the prototype towards achieving scale in the ER community. We will work with existing and future partners to submit non-dilutive grants and to license copyrights to the LEARNER curriculum across ER and broader industry domains. LEARNER will serve our responders by building a more capable and skilled ER workforce, safeguarding their health, improving their career longevity, and ensuring our nation's emergency preparedness.

Differentiators

Current VR based training platforms in ER

are focused on extensively simulating the environment. LEARNER will enable quick integration of emerging HATs (EXO and AR) into its platform enabling an individual or a group of users to learn and collaborate in tomorrow's human-technology ER teams. Access to training resources (i.e., facilities,

technology, budget, time away from duty) remains a critical barrier for effective and continued ER training. LEARNER provides multiple training delivery methods, from the home to embodied immersive training, that offer affordable and abundant opportunities for rapid repetition and skills refinement. ER workers are diverse in their makeup, in terms of their demographics, experiences, trust, and learning rates. In contrast, current ER training paradigms are static, with a one-size-fits-all approach. LEARNER utilizes personalized learning algorithms to reduce skills gaps across ER teams that enhance team operations. LEARNER is scalable across other work domains; affording HAT learning in, for e.g., manufacturing, thereby creating opportunities for broader industry adoption.

Road Map

Milestones: At **Month 8**, we completed the development of the curricula for two HAT-integrated ER courses: 1) Triage using AR tool (e.g., virtual triage tag); and 2) Patient Handling using EXO. By **Month 16**, we will have tested and embedded the adaptive learning algorithms into the LEARNER prototype. By **Month 17**, the courses will be embedded in the LEARNER prototype, and will be coupled with immersive VR environments with realistic scenarios and multisensory feedback (e.g., visual, haptic). By **Month 21**, LEARNER pilot test will be completed with ER workers in a high-fidelity ER benchmarking study at TEEX. Key inflection period will be during the pilot test evaluation (**Months 21-23**), followed by refining the prototype. Key deliverables: ER-based training curricula, personalized learning algorithm for ER training and AR/EXO learnings, AR/EXO-specific learning modules, and a functional LEARNER prototype.

Partnerships

We have assembled a team of academic researchers across three universities, industrial (SARCOS Robotics; EXO integration), government (NIST; training standards/ testing), and non-profit ER workforce development (TEEX; ER training and evaluation) partners. Our advisory board is composed of leaders from ER stakeholder organizations (International Association of Fire Chiefs, National Volunteer Fire Council, National Fire Protection Association; to advise on ER training needs/ constraints), government institutions (FEMA, US Naval Research Laboratory, Texas Division of Emergency Management; to guide on achieving scale in ER), and industry partners (Eksobionics, IHMC, Boeing, Ford, and ASTM; to share experiences in transition to practice, commercialization, and training and assessment in diverse industrial sectors).

Intellectual Property

The IP will include the LEARNER training platform, ER training scenarios, personalized learning algorithm, and EXO/ AR learning modules that would be protected in the form of copyrights and trademarks, registered through filings with USPTO to minimize infringement and illegal copying of materials. Licensees will be responsible for preventing infringement or illegal distribution of IP.

Visit us at www.projectlearner.com

NeuroAI@Work

Transforming Neurodiverse Employment Through Inclusion AI

NeuroAI@Work

Lead PI: Nilanjan Sarkar
nilanjan.sarkar@vanderbilt.edu

Susanne Bruyere
smb23@cornell.edu

James Rehg
james.rehg@cc.gatech.edu

Brian Scassellati
brian.scassellati@yale.edu

Zachary Warren
zachary.e.warren@vumc.org

Overview

We aim to enable the inclusion of adults with autism spectrum disorder (ASD) in the American workforce through novel, artificial intelligence (AI) tools and systems. Our primary users include working-age individuals with ASD; their support systems, including state vocational rehabilitation programs and advocacy organizations; and the many American companies seeking to hire people with ASD for their unique skills (e.g., EY, Microsoft, SAP). Our tools help these customers uniquely match job-seekers with job opportunities and then support them on the job—with novel skills assessments, social communications training in virtual reality, and robotic coaching supports—thereby empowering individuals with ASD to gain meaningful employment, and enabling employers to access the capabilities of a more neurodiverse 21st century workforce.

Description

Every year 70,000 children with ASD in the U.S. become work eligible adults, but currently more than 50,000 of them (80%) become unemployed or under-employed relative to their abilities. Their access to employment-related training or supports faces geographic, human resource, financial, or systems-level limitations. To address these challenges, our complementary suite of tools and technologies will bridge existing gaps between employers and job seekers with ASD.

Our suite of tools, depicted below, includes: A job skills assessment system based on visual AI; a virtual reality (VR) job-interview simulator and coaching system that detects user stress and engagement and provides real-time feedback; a collaborative VR simulator that assesses and supports communication and teamwork; a robotic coaching system that offers home-based training to build resilience to workplace distraction and to practice customer interaction; and a computer vision and machine learning-based system to assess non-verbal communication skills in the real world. Our team includes experts in machine learning, computer vision, artificial intelligence, assistive robotics, organizational and implementation science, and psychology, in collaboration with stakeholders representing individuals, employers, and systems. During and beyond phase 2, we will take a two-pronged approach to commercialization. First, we will develop commercialization plans with our expanding network of private- and public-sector partners, which has the potential to transform societal cost into great value for the nation.

Currently, not even 1% of the $175 billion annual cost to the U.S. to support unemployed autistic adults is for development of technologies that can be brought to market at scale. By increasing employment of these individuals, we not only reduce these massive costs, we position this large segment of the U.S. population to contribute substantially to the economy. Second, when needed, we will seek additional funding through formal SBIR/STTR grant mechanisms.

Differentiators

Isolated efforts to create technology-based solutions to support individuals with ASD have met with some technical success. However, there have not been any comprehensive, end-to-end solutions for barriers in the ASD employment pipeline—from skills assessment to training to coaching supports—nor solutions developed with vested interest from a range of stakeholders to optimize the chance of real-world acceptance and deployment.

Three core components of our plan uniquely position us to succeed. First, we have assembled a team with all the critical expertise and connections necessary to execute the work, including experts in cutting-edge AI-based technologies for people with ASD as well as critical convergent expertise in organizational, implementation, and clinical science. Second, our team represents organizations that allow access to people with ASD, which is necessary for human-centric participatory design research and represents a unique strength in scope and reach. Third, these types of technologies will never be transitioned to practice on a large scale unless there is buy-in from key stakeholders such as ours (employers, vocational rehabilitation groups), who have agreed to allow us into their workplaces and facilities to pilot systems, to gather user interaction data, and to successfully deploy our technologies through commercialization.

Road Map

Year 1 focused on prototype refinement toward MVP, delivery of MVPs to partners for testing, and data capture to understand barriers to uptake. Some of these activities have been delayed due to COVID-19, nonetheless we have made significant strides toward deployment and commercialization. For example, our partner The Precisionists, Inc., has executed a licensing agreement to deploy our technologies at their employment training center in Wilmington, DE. Similarly, our partner Amerigroup, a large regional insurance firm, is now executing a first licensing agreement for deployment of our technologies at their community-based clinics across TN, utilizing an insurance-plan reimbursement model to provide access to a large base of users.

At the same time, in **Year 2** we are also working with our current partners and new partners to expand the application of our tools toward even broader impact beyond phase 2. Our current work is focused on autism but has clear applicability to much larger underutilized segments of the workforce.

Partnerships

Our strong team of partners include:

1. Private-sector companies that seek cutting-edge tools for employing people with ASD: Auticon, EY, SAP, The Precisionists, Inc, and Amerigroup Insurance. They will provide access to employees with ASD and management staff, iterative input, and commit to commercial licensing agreements when appropriate, as well as act as testbeds.

2. Public-sector regional employment centers that provide job related support (Iowa, Washington, and California). They will provide access to their sites for deployment and provide user data.

3. Companies committing tools and distribution platforms (Floreo, SourceAmerica, and Microsoft): They will provide technical expertise and resources, including sharing VR and AI platforms to speed development and expand deployment.

Intellectual Property

Our private-sector partners have committed to execute licensing agreements for use of our tools as they become ready for commercialization. Our public-sector partners have committed to execute nonprofit agreements to deploy our technologies as scalable assessment and training opportunities for the individuals they support. Our digital-content partners have committed to work with us to embed our products within their platforms, potentially extending our reach to millions of current users beyond phase 2.

SkillSync

Reskilling Workers at the Speed of Business

Lead PI: Robby Robson
robby.robson@eduworks.com

Myk Garn
myk.garn@usg.edu

Ashok Goel
ashok.goel@cc.gatech.edu

Jeanne Kitchens
jkitchens@credentialengine.org

Matt Lisle
matt@c21u.gatech.edu

Elliot Robson
elliot.robson@eduworks.com

Overview

Industry 4.0 is causing rapid shifts in the skills workers and companies need, leaving companies vulnerable and workers with skill gaps. When companies look to college continuing education and professional development programs to fill these gaps, they find that colleges have not kept up with market demands and do not understand company needs. Communication is ad hoc and information is often hard to get. This leads to frustration, gets in the way of colleges serving their communities, and forces companies to spend more than desired on provisioning training. The ultimate effect is to reduce the number of reskilling opportunities for workers.

SkillSync is a web application that solves these problems. It uses AI and national skills data to help companies identify the skills their workers need. It then connects companies to college providers, facilitates communication between companies and colleges, and helps colleges create and offer programs that align with company needs. As Amazon, Uber, AirBnB, and similar apps have done in their market segments, SkillSync removes significant pain points in the $170B corporate training market by connecting reskilling consumers (companies) to reskilling providers (colleges) and enabling them to easily exchange just the right info.

Description

The SkillSync workflow starts with a company HR or talent professional. These users engage with SkillSync to identify and prioritize a set of skills needed for a new job or position or for employees to advance. SkillSync provides access to trending skills derived from national job data and can use AI to extract skill requirements from company job descriptions. Aided by a virtual assistant that helps them understand and make efficient use of SkillSync, company users build a detailed reskilling request and submit it to selected college training providers.

Upon receipt of a request, a college can use SkillSync to formulate a training proposal. To ensure that proposals target the right skills in an efficient manner, SkillSync uses AI to score the alignment between a set of training opportunities and a company's request. Training proposals can include a college's existing or custom offerings, parts of courses or whole courses, and externally available commercial and open educational resources. SkillSync's AI and skills management services create an "intelligent" user experience designed to meet expectations for efficiency, and accuracy.

Our AI research is focused on four operations needed to support the SkillSync workflow: (1) extracting and prioritizing knowledge, skills, and abilities (KSAs) from job descriptions and other unstructured sources, (2) removing unwanted, biased, or unallowable data from auto-generated KSAs, (3) assisting the user to formulate a comprehensive training request by suggesting KSAs for consideration, and (4) determining how well a set of courses or modules address a skills training request. SkillSync uses a transfer learning approach in which transformer- based, pre-trained language models are tuned through additional rounds of unsupervised learning to improve domain coverage. The resulting domain-specific models are then trained using labeled datasets to perform specific tasks, such as extracting KSAs from unstructured text and determining the degree of alignment between a set of courses and a training request.

Differentiators

SkillSync is first to implement a two-sided, digitally enabled process that helps companies and colleges engage with each other to obtain and provide reskilling. Differentiators include: (1) making nationally derived skills frameworks available to companies and colleges; (2) a novel AI-based alignment scoring tool that calculates the match between required skills and a college's existing instructional assets; (3) use of the Jill Watson intelligent agent developed by Georgia Tech to help users understand and use the SkillSync app; (4) the ability to extract skills from job postings, position descriptions, and college course information; and (5) bias controls discussed next section.

Bias Control and Research Results

Throughout the research process, we focused on documenting and mitigating undesired bias in the machine learning models used to create AI services. Large-scale language models have been criticized for encoding, or even amplifying, undesirable societal biases, and these biases have been particularly noted in relation to occupation and job-skill related terminology. For Skillsync AI services, we identified four potential points where undesirable biases may be introduced or mitigated: (1) large text corpora used to pre-train underlying language models may reflect biases; (2) labeled datasets used to fine-tune models to perform specific tasks may reflect biases; (3) labeled test sets used to measure performance of models during the research process may reflect biases; and (4) conscious or unconscious bias may be introduced by human raters used to label datasets (e.g., there is a strong correlation between gender of the human labeler and the frequency with which they detect soft skill KSAs in unstructured job descriptions).

SkillSync employs several techniques, including counterfactual data augmentation (CDA) and REG (a gender bias reduction technique) to mitigate undesirable biases in underlying datasets, and to address human rater bias we attempted to diversify the pool of human raters.

We were initially concerned that there might be tradeoffs when training models to minimize bias while maximizing performance, but we found that minimizing bias often helped, serving to prevent overfitting.

Participatory Design and Trials

During the first year of phase 2 development, Skillsync worked closely with corporate and college partners and team members such as the Business and Higher Education Forum, the Credential Engine, DXtera, the Open Syllabus Project, the National Association of State Workforce Agencies, CAEL, and UPCEA.

Using resources provided by these partners, Multiple focus groups were run from September 2020 through March 2021. A prototype was tested in May of 2021 by global corporations Halstead and Southwire, Georgia Tech, and the University of West Georgia. In these tests, SkillSync functioned correctly, and its users validated its value and features.

Road Map

SkillSync continues to be developed using a participatory and user-centric design approach. An MVP release due in **2022** will address input received from users; include more AI; more fully incorporate the Jill Watson intelligent agent; and add marketplace features. A further prototype trial is planned for **2021** and multiple full trials are planned for **2022**. In addition, the SkillSync team is studying sustainability options and will test a provisional business model in **2022**.

Intellectual Property

SkillSync skills management tools developed by Eduworks are, in part, open source. SkillSync AI tools are owned by or licensed to Eduworks. Partner data is licensed under data agreements. Jill Watson is owned by Georgia Tech with an agreement to license it as a service. Data produced by users belongs to the users.

For additional information and to contact the SkillSync team visit www.skillsync.com.

STEP UP

Skills-based Talent Ecosystem Platform for Upskilling

Ranjana Mehta
rmehta@tamu.edu

Nilanjan Sarkar
nilanjan.sarkar@vanderbilt.edu

Robby Robson
robby.robson@eduworks.com

Overview

STEP UP (Skills-based Talent Ecosystem Platform for Upskilling) connects workers, employers, and training providers to each other based on skills. Its goal is to provide diverse populations with equitable access to jobs and training and to help employers find, train, and hire workers in response to events that change the nature of work, ranging from the next pandemic to technological breakthroughs. STEP UP maps jobs and training to skills frameworks and offers skill gap analyses, training, and skills-based credentials to ensure that (a) workers are judged by the skills they have rather than formal educational achievements and previous job titles; (b) workers have (and are able to) access training that fills gaps in both the soft and technical skills they need for modern work environments; and (c) employers can accurately express the skills they need and identify workers that have them. The goal of STEP UP is to transition NSF Convergence Accelerator work to practice so that it has positive and lasting impact on American workers, especially on underserved groups and the missing millions in parts of the country that are often overlooked, as is so crucial for Future of Work.

Description

STEP UP! is a collaboration among three "Track B" (future of work at the human technology frontier) Convergence Accelerator projects : (1) LEARNER, which is providing first responders with new forms of training to prepare them for the use of technologies such as exoskeletons and Augmented Reality (AR), (2) NeuroAI@ Work (NeuroAI), which is creating innovative AI, Virtual reality (VR), and Robotics-based applications and training regimens that help neurodiverse individuals enter into and succeed in the workforce, and that help employers understand and make use of the skills these individuals offer; and (3) SkillSync, which helps companies understand and communicate the skills their workers need, and connects them to professional development and continuing education programs that can offer these skills, and facilitates the development of efficient, effective, and equitable reskilling programs. As individual projects, LEARNER meets employer needs for upskilling emergency responders; NeuroAI trains under-employed neurodiverse workers and connects them to potential employers, and SkillSync connects employers to training providers. STEP UP! combines the strengths of all three projects to create a virtuous cycle among workers, employers, and training. STEP UP will (a) allow employers to create job profiles that identify the skills needed for in-demand jobs; (b) enable individuals to create (and validate) skills profiles that identify the skills they have; (c) use AI to match skills profiles to job profiles and to identify skill gaps; (d) enable individuals to find and enroll in training to fill skills gaps; (e) issue skills-based credentials that validate newly acquired skills; and (f) enable individuals to provide existing or potential employers with those credentials. All of these functions are supported by AI that extracts skills from unstructured data and is used in matching and recommendation algorithms.

As a first step in STEP UP the Track B projects are joining forces to create a National Talent Ecosystem Council (NTEC) that will engage accomplished researchers, educators, practitioners, and policy makers who are working towards building a highly skilled and inclusive workforce. NTEC will provide oversight to the STEP UP project and additionally manage and disseminate the research, guidelines, standards, skills frameworks, and data produced by current and future Convergence Accelerator projects that address issues in education, training, and workforce development.

As a second step, STEP UP is creating guidelines for selecting the best training for acquiring a given set of skills under a variety of constraints, e.g., when physical access to a training center is not an option due to a pandemic or when an individual cannot engage in VR experiences due to motion sickness. These guidelines will help provide upskilling in response to unforeseen events. In parallel, STEP UP is developing skills frameworks and associated skills-based credentials for technical areas covered by the LEARNER, soft skills covered by NeuroAI, and for leadership training, which is a steppingstone to jobs that provide a living wage and an upward career path. These will be used in a proof-of-concept STEP UP portal scheduled for release by the end of 2022.

Differentiators

STEP UP is designed to address under-served populations and to enable the talent ecosystem to respond to new skills needs and unforeseen events. It applies machine-learning algorithms trained to avoid bias, has access to national job data so that it can detect emerging skills, provides innovative training to populations with non-standard needs and to populations that need to acquire newly emerging skills, and applies guidelines for selecting alternative modalities in response to worker circumstances. In addition, STEP UP is designed for a national talent ecosystem. Enterprise talent management and talent marketplace systems, in contrast, are used by single employers, while systems workers use to find jobs and employers use to recruit workers are different systems and do not facilitate access to training. STEP UP is a complete solution that closes the loop among workers, employers, and training providers and that is overseen by an organization (NTEC) dedicated to creating a robust equitable talent ecosystem.

Road Map

Work has started on forming NTEC and on setting up collaborations. NTEC will start as in an informal organization but will be incorporated in the future, likely as a 501©3 organization. It is anticipated that guidelines for training system selection and an initial set of skills frameworks, and related credentials will be in place in **Q1 of 2022** and that a proof-of-concept STEP UP web portal will be released by the end of 2022. This portal, which will be a cross between a web site and a full web application, is intended to show the "art of the possible" and inform the design and development of a full STEP UP platform. Thereafter the STEP UP project will seek the funding and required to develop a robust, fully featured, extensible, and scalable STEP UP platform. NTEC, for its part, is expected to continue as an important organization that otherwise supports national skills and talent pipeline initiatives and increases the impact and sustainability of Convergence Accelerator and other projects via symposia, publications, and networking.

Partnerships

Each Track B project has its own extensive network of team members and partners. All are available to contribute to NTEC and STEP UP Among these, key partners include the Texas A&M Engineering Extension Service (TEEX), Ernst & Young, the Boeing Corporation, Microsoft, the Frist Center for Autism and Innovation at Vanderbilt, the Credential Engine, the National Association of State Workforce Agencies, the Business Higher Education Forum.

Intellectual Property

Much of the IP initially contributed to STEP UP is owned by the Track B projects. This will be cross licensed for use in STEP UP Longer term, the IP in STEP UP will be licensed to NTEC by contributors or owned by NTEC. Data provided by STEP UP users will belong to the users and will be protected in accordance with privacy and data rights laws.

TRACK C:

QUANTUM TECHNOLOGY

Improving the U.S. industrial base, maintaining an edge in emerging areas of technology, creating jobs, and providing significant progress toward economic and societal needs are vital. Teams within the NSF's Convergence Accelerator Quantum Technology track are developing quantum sensors, devices, hardware, interconnects, networks, and simulations to deploy new technologies in a variety of applications, such as autonomous vehicles and healthcare. They are also creating an innovative curriculum by leveraging strong industry-university partnerships that are diverse and inclusive.

Quantum Technology funded phase 1 teams include:

Quantum Sensors

- **QuPID**—Led by the University of Chicago, the QuPID enables frequent and inexpensive measurements of thousands of proteins resulting in earlier interventions that increase the quality of health care and decrease costs.
- **Quantum Sensors**—Led by the University of Arizona, the Quantum Sensors is developing an entanglement-enhanced sensing architecture to benefit multitudes of domains, including secure inertial navigation, space and planetary terrestrial control, and healthcare monitoring.
- **PEAQUE**—Led by the University of Washington, the PEAQUE is addressing quantum computing scalability by innovating a chip-scale, multi-beam optical control system that empowers cold-atom quantum computing with 1,000s of qubits.

Quantum Devices, Hardware, and Interconnects

- **Hi-LINQS**—Led by the University of California-Los Angeles, the Hi-LINQS resolves the key quantum interconnect challenge in scaling up quantum systems by integrating different quantum computing platforms to unleash the power of quantum computers and systems.
- **TOPOLOGICAL QUBIT**—Led by Massachusetts Institute of Technology, , the Topological Qubit is addressing the complex challenge of quantum error correction in quantum computing by developing new quantum hardware for quantum computing.
- **AQS**—Led by Columbia University, the AQS is developing a novel atomic quantum simulator by combining analog quantum simulation and digital quantum computing in the same device.

Quantum Networks/Simulations

- **QuaNeCQT**—Led by the University of Maryland the QuaNeCQT is developing hardware to transform the internet into a quantum internet, which is essential to connecting the anticipated rapid expansion of the use of quantum computers.
- **SQAI**—Led by Pennsylvania State University, the SQAI is accelerating drug discovery using quantum computing and AI.

Workforce/Education

- **National Quantum Literacy Network AI-Powered Microcredentialing**—Led by Morgan State University, the NQLN AI-Powered Microcredentialing is developing a convergent solution to rapidly deploy a diverse quantum workforce.
- **QuSTEAM**—Led by the Ohio State University, the QuSTEAM is a transformational undergraduate curriculum aimed at addressing critical workforce needs in quantum information science and engineering.

AQS

Cloud-Accessible Atomic Quantum Simulator

Lead PI: Sebastian Will
sw3151@columbia.edu

Gabriella Carini
carini@bnl.gov

Alexander Gaeta
alg2207@columbia.edu

Layla Hormozi
hormozi@bnl.gov

Nanfang Yu
ny2214@columbia.edu

Overview

Quantum computing holds great promise for the future of computing. However, so far, available quantum computers have not been able to solve relevant real-world problems better than classical computers. We will build a quantum simulator based on programmable arrays of atoms that offers a realistic prospect to solve relevant computational problems faster than available classical and quantum computers. Expanding usability beyond individual users, our system will be cloud-accessible for broad use in academia, industry, and government.

Description

Quantum devices using atoms as quantum bits (qubits) open a path to scalability. Atoms are identical, highly reproducible in large quantities, and can be prepared, manipulated, and observed with the utmost precision. Atoms are nature's own qubits. In this highly convergent effort, connecting physicists, computer scientists, and engineers, we are realizing a novel integrated atomic quantum simulator that combines features of analog quantum simulation and digital quantum computing in the same device, utilizing over 1,000 atoms as precisely controlled qubits. This becomes possible through the integration of leading technology from atomic physics, nanophotonics, signal processing, and software engineering in close collaboration with potential users of the platform.

This project will make critical advances both in terms of novel quantum applications that it will enable and technological innovations through the construction of the device. With our platform, it will become possible to elucidate currently intractable problems in materials research, quantum chemistry, and graph problems relevant for optimization tasks. In particular, our system will be a powerful solver for the maximum-independent set problem, which relates to practical problems, such as achieving optimal coverage in 5G mobile data networks, effective deployment of national defense systems, and efficient routing for delivery and logistics, which will save natural resources, reduce humanity's carbon dioxide footprint, reduce cost and environmental impact. By integrating photonic nanotechnology on a broad frontier, we will also make a major leap towards the construction of quantum devices that will be more compact, robust, and reliable. This will pave the way towards an era of field-deployable atomic quantum devices for quantum simulation, quantum sensing, and quantum communication, which can be brought to market.

The quantum simulator will be professionally hosted as a user-facility at Brookhaven National Laboratory and will be accessible to users via a secure cloud-based web interface.

Differentiators

Distinct from currently available platforms for quantum computing, our system hybridizes elements of analog quantum simulation and digital quantum computing. As a result, our platform can operate both as a special-purpose quantum simulator and a gate-based universal quantum computer. This duality inspires new paradigms of quantum computing outside the beaten tracks; we will be able to efficiently

implement quantum algorithms with shallow circuit depth, reducing the need for complex gate-based quantum circuits that are prone to errors and decoherence.

Our system is also advantageous in terms of the quantity and quality of qubits, the fidelity and the variety of entangling gates. The largest numbers of coherent qubits reported so far are less than 100 for superconducting qubits and about 250 for a similar atomic device. Our system is scalable to more than 1000 qubits. In addition, massively parallel optical control will enable multi-qubit entangling gates (such as the Toffoli gate) with high fidelity to greatly enhance the circuit efficiency in ways that are not available in current qubit systems.

Road Map

We purse a vertically-integrated development process, broken down into five closely collaborating focus areas: (1) atomic platform, (2) lasers and photonics, (3) timing and control, (4) quantum compiler and user interface, and (5) quantum applications and algorithms. We are developing the system in a quantum co-design approach: The design of our quantum simulator is directly informed by relevant applications, allowing us to utilize the computational power of the platform with maximum efficiency. In phase 1, we have developed prototypes for a novel atomic source, chip-based lasers, a high-speed timing system, and a demo version of the user interface. The milestones for phase 2 are:

Year 1: (M.1.1) Demonstration of laser cooling of strontium atoms with chip-based lasers. (M1.2) Demonstration of programmable 1D arrays of atoms. (M.1.3) Connecting the timing and control system with the quantum compiler and cloud-based user interface. (M.1.4) Develop concepts for quantum speed-up exploiting the hybrid digital-analog architecture of our system.

Year 2: (M.2.1) Integration of the timing system for precision control of quantum operations on atoms. (M.2.2) Trapping of 1,000 individual atoms in 2D arrays. (M.2.3) Testing, benchmarking, and validating system performance. (M.2.4) Professional hosting as a user-facility within the Brookhaven National Lab infrastructure. (M.2.5) Go live with the cloud-based user interface for public user access.

Partnerships

Our team brings together world-leading expertise in all focus areas of the project. **Columbia University** contributes pioneering expertise in atomic quantum simulators, silicon-based nanophotonics, and holographic metamaterials. **Brookhaven National Lab** contributes unique experience with the deployment of high-tech user facilities and game-changing know-how in instrumentation that is not available on the free market. **Bloomberg LP** dedicates a substantial software engineering effort to develop the user interface. Recognized experts in quantum physics, quantum algorithms, and computer science from **Flatiron Institute, Yale University, New York University, Penn State University, Rice University,** and **JP Morgan Chase & Co.** support our co-design approach. Together with established quantum industry partner **IBM** and rising start-ups **Atom Computing** and **QuEra**, our team is well-positioned to make a "quantum leap" in the design of a paradigm-shifting atomic quantum simulator.

Intellectual Property

Access to our quantum simulation platform will be free for noncommercial users; we anticipate a service contract model for business and industry users. Hardware advances will be broadly shared with the research community. Software developed for the user interface will be made available under an appropriate open-source license.

Hi-LINQS

High-Coherency Chiral Links for Networking Quantum Systems

Lead PI: Kang Wang
wang@ee.ucla.edu

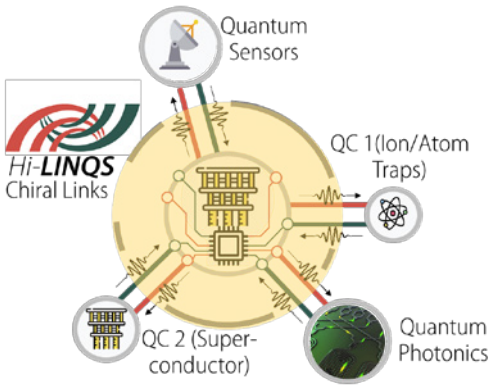
Clarice Aiello
Ajey Jacob
Jonathan DuBois
Thomas Ohki

Overview

LINQS aim to resolve one of the key technical challenges in scaling quantum computing (QC) - the lack of quantum interconnects and interconversion devices for networking. Current interconnects and converters, i.e., microwave cables, optic fibers, etc. are noisy, bulky, and suboptimal for QC applications. Our high-coherency, low-noise and scalable chiral quantum interlink platform will help integrate different QC platforms and unleash the power of quantum computers and systems.

Description

An ideal quantum interlink platform including interconnects and quantum converters must preserve the integrity of quantum information when networking different quantum systems. To date, QCs use conventional interlink technologies. As the industry ambitiously scales up the number of qubits, the need to improve interlinks and networking performance is becoming increasingly urgent among different quantum platforms (e.g., superconducting, ion/atom trap qubits, quantum sensors, and alike) operating in different frequency domains. To accelerate this scaling, high coherency interlinks that can transfer quantum states coherently between different quantum platforms are critically needed.



Our overarching goal is to accelerate integration of QCs and scaling. Our Solution: Providing innovative Chiral (i.e., directional) and hierarchical interconnects, and quantum converters with chip-level integration for networking different scaled quantum systems.

Thrust 1 - Intra-system chiral interconnect.
Within a platform or intra-system level, we focus on delivering miniaturized chiral microwave interconnects for e.g., superconductor systems. The approach uses novel chiral topological materials and structures to achieve chiral (one-way) microwave propagation in an on-chip microwave quantum circulator. In phase 1, prototypical devices were fabricated, and non-reciprocity was demonstrated.

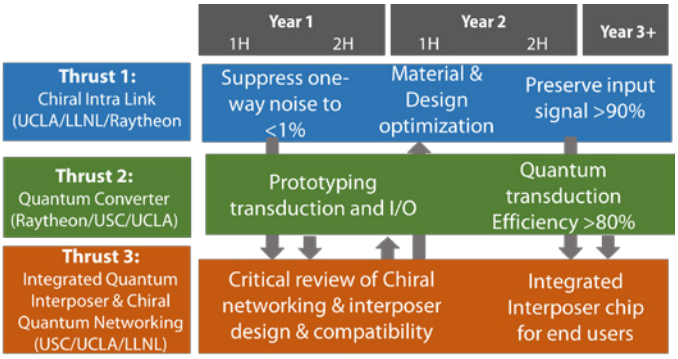
Thrust 2 - Inter-system quantum converters.
For connecting different QCs and sensors in a network, it is necessary to convert microwave to optical photons, which are preferred for networking and communications. To enable such distributed and fused QCs and quantum sensors (amongst SC, atom/ion trap, photonics, etc.), and leverage existing telecom infrastructures, high- coherency inter-band quantum converters will be designed and fabricated. The initial design has been completed and components have been obtained.

Thrust 3 -Integrated quantum interposer and chiral network.
Monolithic integration of quantum multichip modules is challenging at cryogenic temperatures. The benefits of chiral networking were initially modeled to guide our interposer designs developed in phase 1. Phase 2 will provide detailed assessment of chiral networking and develop quantum interposer technologies, that can integrate and preserve quantum features in heterogeneous integration. The quantum interposer chips are aimed at providing the benefits of scaling (size, speed, and power) as well as reliability (mechanical and thermal robustness at the hetero interfaces). Design considerations, partners and users have been identified.

Differentiators

Today's conventional interconnects are not scalable for networking QCs and sensor systems because they both cause unwanted noise and de-coherence, while also occupying a large footprint in QCs. Our solution is to suppress the noise, preserve high coherency, reduce the physical footprint, and improve the scalability with a roadmap similar to that of semiconductor technology. By separating inbound and outbound signal paths, noises from the environment and peripheral circuits are blocked, thereby minimizing the crosstalk. The decoherence from interlinks, interconnects and converters, may be reduced by over 100x (20 dB) compared to current devices. The quantum interposer chips will reduce the size by more than 1000x in volume. This is possible due to the fundamental knowledge of new materials, structures, and physics in academia and the expertise and infrastructures of Lawrence Livermore National Laboratory (LLNL) and Raytheon BBN to realize the accelerated development in integrating different frequency domains.

Road Map



Milestones. Year 1: demonstrate on-chip circulators with isolation > 20 dB; Design high Q integrated AlN resonator converters coupled microwaves and photons; Verify integration compatibility for inter- and intra-system links. Modeling of the design benefits for chiral links and networking.
Year 2: Improve insertion loss and decoherence of on-chip quantum circulators to < 0.5 dB; Improve microwave-to-photon conversion noise and efficiency > 80%; Multichip integration on a quantum interposer.

Year 3+: Demonstrate a complete product development pipeline for end-user groups. Establish a sustainable path for the project via partnership with industry, government and established centers.

Partnerships

Our team is constructed with complementary expertise, capabilities and resources to achieve our objectives and roadmap. The Core Partners and their responsibilities are University of California, Los Angeles (UCLA) (chiral material and structures, and Cryo-RF), Raytheon BBN (SC qubits, MW-optical converter), Lawrence Livermore National Laboratory (SC qubits, Cryo-RF). University of Southern California, Information Sciences Institute (MW-optical converter, interposer). As learned from phase 1, we will include new members from UCLA (modeling of chiral networking) and University of Michigan (new electro-optic material), and Tyndall (integration). We will also collaborate with the following NSF Convergence Accelerator members to broaden our impact, and assure the track success in quantum networking; C-695 and C527 (trapped ions), C-575 and C-520 (quantum sensors), and C-614 (quantum literacy for the workforce). We have also established an Industry Advisory Panel to focus our goal. Our long-term strategy is to engage with the stakeholders (through QED-C) and to commercialize our Hi-LINQS with industrial end-users and partners.

Intellectual Property

Our team will establish different agreements per the two distinct constituents. The first agreement ("Partnership Agreement") will establish procedures for managing intellectual property among the Core Partners of the program while respecting the existing obligations of each Partner. The second agreement ("Industry Affiliates Agreement") will establish the procedures for managing the IP between the core team and the industry affiliate members.

NQLN AI-Powered Microcredentialing

Rapid Microcredential Certification and QUAINT Search Engine

Lead PI: Timothy A. Akers
timothy.akers@morgan.edu

Denise Baken

Lily Milliner

Roxanne Hughes

Kevin Peters

Overview

The United States educational system broadly and the Department of Defense specifically are confronting one of the nation's deadliest threats– the need for a Quantum Literate workforce as a national and economic security imperative. Geopolitical competitors and potential adversaries, namely, China, Russia, and other nations, are outspending, outperforming, and rapidly educating their populations and military to occupy this new terrain. To address this deficit, the National Quantum Literacy Network (NQLN), a consortium of Historically Black Colleges and Universities, minority business enterprises, national laboratories, government agencies, and nonprofits, have been established to address hyper-disparities of historically underrepresented groups in the quantum literacy workforce. The NQLN serves to combat this deficit by educating ROTC students and retooling the military defense educational systems to implement Quantum Literacy as a national defense and economic strategy. Our solution is twofold: 1) to develop a Rapid Micro-Credential Certification program in Quantum Literacy and 2) create a QUAINT algorithm (Quantum Artificial Intelligence for Nascent Taxonomies) that rapidly identifies emerging quantum applications that impact the workforce development needs of the industry, defense, and government.

Description

Currently, data suggests that the skilled diverse workforce is not meeting the demand of industries in developing quantum technologies. This gap is projected to increase for at least the next decade. For example, historically underrepresented individuals and groups constitute less than 1 percent of the current quantum workforce; however, they are projected to comprise close to 50% of the nation's workforce by 2050. To date, the quantum industry is expected to generate over $2 billion within the next 7 years. However, because of these extremely low diversity estimates in the workforce, our NQLN team has recognized what we call as a hyper-disparity that is impacting historically under-represented workers and industries in this emerging scientific and economic quantum enterprise. If this, "hyper-disparity" remains or grows, the U.S. will experience a major tipping point in the diversity, equity, and inclusion gap in the quantum workforce. Because of these low numbers of diverse workers, the cost estimates have not been projected in this emerging industry. Therefore, the nation needs to understand and develop estimators to address and project the potential impact and risk to global stability and national security.

The NQLN, for example, has been developing two complementary innovations: 1) prototype modules to rapidly develop and implement micro-credential certifications designed to develop skills in quantum literacy across sectors; and 2) create a QUAINT search engine that employs artificial intelligence (A.I.) and Natural Language Processing (NLP). QUAINT will extract new and emerging quantum information from the web to build a database of taxonomy terms, definitions, and applications that link back to the micro-credential certification for rapid deployment to industry, government, and military.

Combined, the rapid micro-credential certification program and the QUAINT search engine can be initially a vital resource to military commanders, educators, and trainers. More importantly, these innovations will serve as spin-offs to industries such as healthcare, finance, utilities, telecommunication, and academia, among others. The NQLN prototypes will help to ensure a quantum-ready workforce for mission critical and mission ready needs, industrial workforce, or military command and control units.

Differentiators

The NQLN will be the first integrated quantum education program focusing exclusively on quantum literacy that targets historically underrepresented groups. More specifically, the NQLN team is engaging ROTC, military service academies, and non-service academy training institutes as prototype user groups

to test the micro-credential certification program and QUAINT search engine. The rationale for the NQLN differentiator is because 1) the U.S. military disproportionately recruits underrepresented groups and 2) has a workforce deficit in quantum literacy. For example, African Americans make up 19% of the military's active duty enlisted members, but only 9% of active duty officers, according to the Pew Research Center. Ideally, our goal is to close this gap by strategically introducing our prototype as a solution to narrow these hyper-disparities across the quantum enterprise.

Training in quantum literacy is unique across every quantum application because of the diversity of technologies, taxonomies, methods, and learning outcomes. Our prototypes will be adaptive to a variety of learning styles of potential candidates and iterative to accommodate a broadening field of careers that may not require advanced degrees. We focus our end-users to the military because it is an institution that provides advanced technical training opportunities and technologies to a wide variety of historically underrepresented groups. The ROTC cadets serve as a logical starting point for this prototype since they are open to explore various career pathways.

For example, ROTC students, with many coming from historically underrepresented groups, receive advanced training prior to their future career choices (e.g., industry, defense, government, or academia). Given that the DOD also has a workforce deficit in quantum technicians, the NQLN quantum micro-credential certification serves to address this deficit by focusing on historically underrepresented groups in high school and college ROTC programs to fill the DOD's urgent quantum workforce gap. In addition, our micro-credential certification will have broader applications to other industries and sectors.

Road Map

The following table represents the two-year outcomes for **phase 2** deliverables. As mentioned, the NQLN project focuses on rapid micro-credential certification in Quantum Literacy and is a complement to the QUAINT search engine that will serve to link quantum applications to quantum taxonomic terms, curricula, and training.

Year 1 Outcomes	Year 2 Outcomes
Micro-Credential Certification Programs for ROTC and Military	
• Develop Micro-Credential Certification Modules: testing and evaluation. • Develop bridge program between junior ROTC and college ROTC programs. • Modules include: Cohort, Quantum Application (Subject Matter Content), Taxonomy, Definitions, Visualizations, and Equations, Metrics	• Implement Micro-Credential Certification Modules to targeted ROTC programs nationally. • Implement junior ROTC bridge programs at targeted to ROTCs at HBCUs
QUAINT Search Engine Software	
• Develop QUAINT search engine software for beta-testing by employing Artificial Intelligence and Natural Language Process. • Develop a working database for quantum taxonomies across selected quantum applications	• Link QUAINT to standardized Micro-Credential program • Populate a database with quantum taxonomic terms, definitions, learning levels, applications, visualizations, and equations • Develop Micro-Credential Certification Templates
Quantum Literacy Convergence Exchange Program	
• Teams Collaborate to Exchanging Ideas across HBCUs and non-HBCUs	• Teams Collaborate to Integrate Ideas Learned through Team Projects for Adaptive Micro-Credential Certification Prototypes in Quantum Literacy
Broader Impact	
• K-12, Vocational, Community College, Undergraduate, Graduate, Industry, Government, Military	

Partnerships

The selected partnerships identified, among others, are critical in providing expertise in the areas of quantum science and its application to quantum literacy. The collaborators will work together to rapidly accelerate the development of 1) micro-credential certification modules and 2) quantum artificial intelligence for nascent taxonomies that extract quantum terms and applications from the internet in order to populate the micro-credential certification modules for curricula and training

Intellectual Property

During phase 2, the Intellectual Property and Technology Transfer of the NQLN products will be initially designed for end-users of the military. Once the technologies are tested the NQLN will scale up to ensure IP is protected through copyrights and patents. This will apply to the micro-credential certification modules and QUAINT software.

PEAQUE-Photonic Engine for Quantum Computing

Enabling Scalable Cold Atom Quantum Computing with Integrated Photonics

PEAQUE

Lead PI: Mo Li
moli96@uw.edu

Ben Bloom

Arka Majumdar

Adam Kauffman

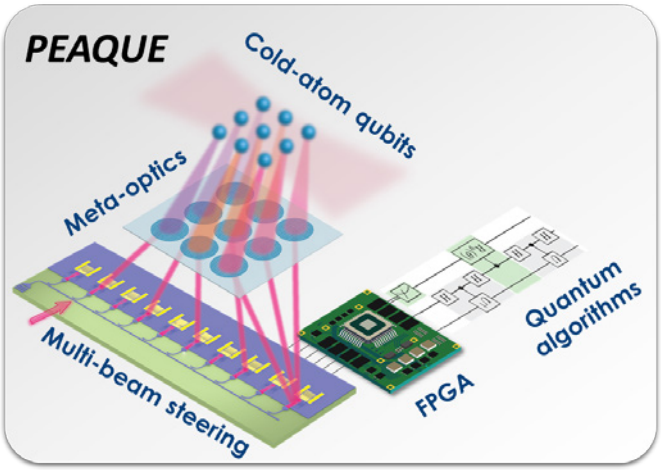
Birgitta Whaley

Overview

Quantum computing promises to solve critical problems that are intractable by classical computers, but currently it faces many scaling bottlenecks. We are developing a cold atom quantum control system in a miniaturized package, programmed with quantum software to empower quantum computing with 1000s of qubits and metrology with entangled atomic clocks. Dissemination of our system to the quantum technology industry will accelerate the realization of full-scale quantum computers capable of solving challenging problems from optimization algorithms to computational chemistry for drug discovery.

Description

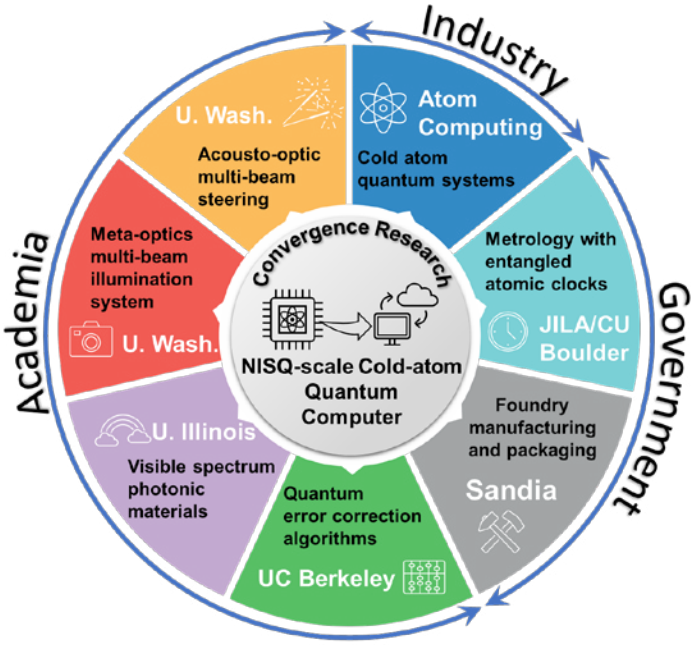
The PEAQUE project addresses quantum computing scalability by developing a powerful optical control engine that interfaces cold atom qubits with quantum software. The core of this engine is a chip-scale Multi-Beam Illumination and Steering (MBIS) system, which leverages the latest advances in nanophotonics and optical materials to generate a multitude of laser spots and precisely focus them on a dense atom array. Each MBIS module includes a pixel array of 10 lithographically patterned devices each capable of emitting 10 individually steerable laser beams, creating a module with 100 beams from a package sized only ~10 cm3, a miniaturization by 3 orders of magnitude over existing technology. Multiplexed modules in an MBIS engine will be able to perform high-speed, parallel gate operations on large 2D or 3D lattices of cold-atom qubits, which is impossible with existing technology. The power of MBIS will enable execution of quantum error correction (QEC) codes specially designed for cold atom qubits toward fault-tolerant computing. The photonic engine will also power entangled atomic clocks in a large lattice to advance quantum metrology. The complete PEAQUE solution package will accelerate building of a cold-atom noisy intermediate-scale quantum (NISQ) computer with capable of solving optimization problems and performing quantum simulations.



Our solution package includes a full stack of hardware and software to enable scalable cold-atom quantum computing and metrology.

Differentiators

Cold atom qubits, compared with other qubits (e.g., superconducting circuits, trapped ions, solid-state spins, or photons), have the decisive advantage of the ease in creating a large number of naturally identical qubits. To use them for quantum computing, optical control of each qubit in a large array is indispensable but has been a challenge to the current technology. PEAQUE will overcome this bottleneck with MBIS's unprecedented multibeam control and modulation capability. The PEAQUE hardware will be mass produced at wafer-scale using



industrial-scale foundries (Sandia National Laboratories (SNL)), and disseminated, along with electronics and software, in a solution package to the quantum community. To achieve the use-inspired research goals, our team consists of multidisciplinary experts in atomic physics, integrated photonics, MEMS, materials science, and quantum software

Road Map

In phase 1, we finalized the system design, developed the fabrication processes, and established a partnership with Sandia for foundry processes.
In phase 2, key milestones and deliverables include prototype MBIS modules, scalable manufacturing at foundries, control electronics systems, quantum software, and final atomic system demonstrations, in a timeline as listed below (starting from 09/2021):
Q2/Y1: Prototypes of single pixel MBIS
Q2/Y1: Field-programmable gate array (FPGA) control system with customized algorithm
Q3/Y1: MBIS integration with cold atom quantum system
Q4/Y1: MBIS integration with atomic clocks
Q3-4/Y1: Prototype of 10 pixel MBIS
Q4/Y1: Foundry process integration completed
Q1/Y2: Fault-tolerate color code algorithms developed
Q3/Y2: Error correction executed in cold atom quantum system
Q4/Y2: Lattice atomic clocks developed
Q4/Y2: 8-inch wafer process completed
Q4/Y2: Test kit/package ready to deliver

Partnerships

Our key partners include researchers across three universities (University of Washington: acousto-optics, nanophotonics, nanofabrication; University of Illinois Urbana-Champaign: wide bandgap materials, UC Berkeley: quantum error correction), two national labs (SNL: foundry services for wafer-scale MBIS fabrication; JILA: metrology and atomic clocks), and one key industrial partner (Atom Computing: cold atom systems for quantum computing and simulation).

Intellectual Property

Our team has established a comprehensive IP management plan that delineates the treatment of background IP, new IP generated through this project, licensing negotiation options, and confidentiality. With our experience in creating start-up companies, we will explore possible spin-offs for the PEAQUE technology.

QuaNeCQT

Quantum Networks to Connect Quantum Technology

Lead PI: Edo Waks
edowaks@umd.edu

Dirk Englund
englund@mit.edu

Saikat Guha
saikat@optics.arizona.edu

Norbert Linke
linke@umd.edu

Tripti Sinha
tsinha@umd.edu

Overview

The internet fundamentally changed every aspect of our lives by enabling computers to communicate with each other over long distances. But the current internet is not compatible with emerging quantum computers that store and process quantum information. QuaNeCQT will develop hardware that will enable the Internet to transmit quantum information over a vast network infrastructure, an essential requirement for the emerging quantum ecosystem.

Description

The current internet cannot transmit quantum information. This limitation relegates emerging quantum technologies, which are currently limited in their computational power, to be stand-alone systems that cannot be expanded or reach a broad user base. A quantum internet that can transmit quantum information would significantly boost quantum computing power by connecting multiple small quantum computers into powerful distributed quantum computers that can solve problems with major societal impact. It would revolutionize numerous industries that take advantage of quantum computing including banking, chemistry, medicine, and data analytics. A quantum network would also greatly increase the user base for quantum computers by providing secure access to end users (blind quantum computing) as well as certifying the legitimacy of quantum computing providers (quantum verification).

QuaNeCQT will enable internet service providers and networking companies to immediately transform their classical networks into the next generation quantum internet that delivers unprecedented security, data rates, and performance. We will do so by developing a comprehensive hardware solution composed of two modules, the qFC module and the qROADM module.

The qFC and qROADM modules convert a classical fiber network it into a fully-functional quantum network that can transmit quantum information and connect quantum computers. They transmit, route, and process both the quantum signals and classical signals required to run a quantum network. They will be fully integrated and equipped with a user friendly software interface for ease of operability. By connecting them to various peripherals (quantum computers, quantum sensors, detectors, etc ...) we can build quantum networks that directly connect emerging quantum technology and can be easily expanded. The qFC and qROADM modules will allow the quantum industry to immediately take advantage of the vast existing infrastructure that is our current internet. We will deploy and test these modules in the MARQI network, UMD and the DC area's local quantum network footprint which we have established in phase 1 of the project.

Differentiators

Currently, a quantum internet that connects quantum computers does not exist. Emerging quantum networks have focused almost entirely on secure point-to-point communication using quantum key distribution. These networks exchange classical information with security guaranteed by quantum physics. But they cannot transmit quantum

information between quantum computers. They therefore cannot interconnect them to scale computation power. A quantum internet that transmits quantum information would provide this essential and currently missing component. Our hardware solution will establish the first interconnection of quantum computers over the internet. It will allow quantum computers to work collectively to greatly increase their processing power. It will also enable end-users to access existing quantum computers remotely without divulging their intellectual property (blind quantum computing), while simultaneously protecting them from fraud (quantum verification). We will achieve this unprecedented objective by combining the most viable quantum computing architectures with advanced reconfigurable quantum photonic devices that can convert and route quantum signals in a network.

Road Map

In phase 2 we will develop compact packaged quantum interconnect hardware and deploy it in the MARQI network.

Y1-Q1- Begin development of the qFC and qROADM modules.

Y1-Q2 - Test hardware components and integrate hardware delivered from our industry partners

Y1-Q3 - Package and integrate hardware modules.

Y1-Q4- Develop software interface for qFC and qROADM modules. Establish plan for expansion and future connectivity of the MARQI network.

Y2-Q1 - Deploy hardware modules into the MARQI network. Install ion traps in MARQI end-nodes.

Y2-Q2 - Establish a connected network of ion traps over the MARQI network

Y2-Q3 - Distribute entanglement over end-nodes

Y2-Q4 - Demonstrate quantum communication between ion trap quantum computers with reconfigurable connectivity.

Partnerships

QuanNeCT started out with core partners that will continue through phase 2 and beyond: (1) **Cisco** is helping to develop hardware packaging and integration for a potential future product line; (2) **IonQ** is ensuring compatibility with their quantum computer and will serve as a node for the MARQI network; (3) **Brain Holding Ventures** will continue guide the commercialization and use case scenarios; (4) **Shinkuro's** CEO serves as the Chair of the MARQI advisory board with expertise in the development of the internet(5) **U.S. Army Research Laboratory** (Department of Defense) provides one of the central nodes of the MARQI network; (6) **Qrypt** provided secure networking expertise.

In phase 2, we added the following partners: (1) **Juniper Networks** will have an executive serve on the MARQI Advisory Board and advise on commercialization (2) **Fermilab** (Department of Energy) will work to deploy hardware on their QCNAST network testbed (3) **ColdQuanta** develop compact deployable ion traps; (4) **Sandia National Laboratories** (Department of Energy) and **Honeywell** will provide the ion chip traps; (5) **NTT electronics** and **ADVR** will provide customized quantum frequency conversion crystals, (7) **AIM** and **LionIx** will fabricate photonic integrated circuits.

Intellectual Property

QuanNeCQT developed and is in the process of patenting multiple key hardware components of a quantum network, including the qROADM and qFC modules. We have established IP agreements with our partners and the university entities.

Quantum Sensors

Quantum-Enhanced Inertial Measurement Unit

Lead PI: Zheshen Zhang
zsz@arizona.edu

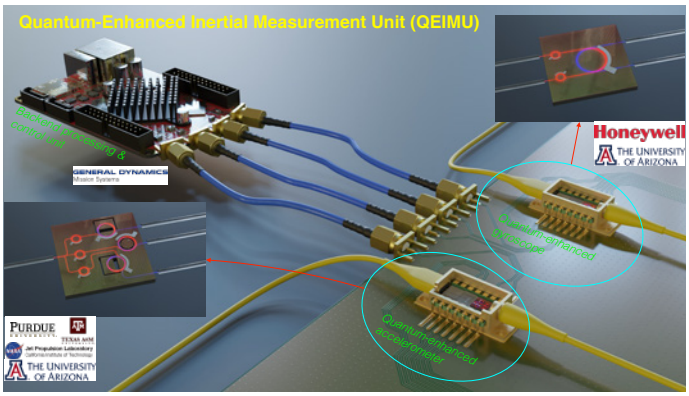
William Clark
william.clark@gd-ms.com

Nan Yu
nan.yu@jpl.nasa.gov

Jon Pratt
jon.pratt@nist.gov

Jianfeng Wu
Jianfeng.Wu@Honeywell.com

Overview



In this Track C Convergence Accelerator phase 2 project, our highly interdisciplinary and cross-sectoral team will develop a quantum-enhanced inertial measurement unit (QEIMU) for positioning and navigation at a performance level well beyond the current state-of-the-art inertial sensors. Phase 2 will build upon the phase 1 results, inputs from user interviews, and the concepts conceived with partners through prototyping meetings. The expected outcome of this project will be a landmark example of how quantum technologies can yield near-term societal impacts within a 5 to 10-year timeframe in diverse realms, such as aerospace navigation, self-driving cars, and space exploration.

Description

Phase 2 core team members are the University of Arizona (UA), General Dynamics Mission Systems (GDMS), Honeywell, NASA Jet Propulsion Laboratory (JPL), National Institute of Standards and Technology (NIST), Purdue University, and Texas A&M University (TAMU). The QEIMU prototype will comprise three principal pillars: 1) a quantum-enhanced gyroscope for angular velocity sensing (Honeywell, UA); 2) a quantum-enhanced accelerometer for linear acceleration sensing (Purdue, TAMU, UA); and 3) a backend processing unit for central control (GDMS). JPL will integrate the QEIMU components. Honeywell, JPL, and NIST will verify the QEIMU in their state-of-the-art calibration and environmental test facilities. The projected QEIMU performance for sensitivity, angle random walk, and bias is one-to-two orders-of-magnitude superior to the state-of-the-art classical inertial sensors. Therefore, QEIMU will enable unprecedented capabilities, including 1) spacecraft control and planetary terrestrial applications without a GPS-like system; 2) secure navigation for self-driving cars; and 3) precise measurements with entangled arrayed-weak force sensors for gravity, gravitational waves, and dark matter and energy, which have previously been scientifically unmeasurable. Since precise navigation and sensing are widely desirable and affect the daily lives of the general populace, we anticipate QEIMU will create a $2.5B market by 2035 and impact 700M people.

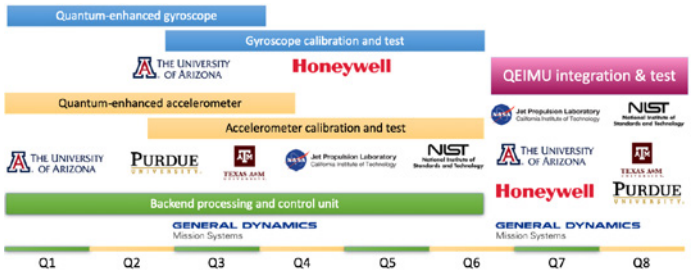
Differentiators

The high cost and large size, weight, and power (SWaP) of commercial optical gyroscopes and accelerometers prevent their market penetration for self-driving vehicles, autonomous robots, and various small-position and navigation platforms. While immense efforts have been dedicated to develop on-chip gyroscopes and accelerometers with reduced SWaP and production costs, due to weak signal-to-noise ratios their performance remains inferior to the navigation grade. The

performance improvement from fabricating larger photonic components to strengthen the inertial sensing signal is saturated by material properties constraints. The noise floor is deemed a barrier dictated by the fundamental quantum noise.

This phase 2 project will harness the revolutionary quantum sensing technology to break this noise barrier. We will miniaturize our quantum-sensing platform into a chip scale and integrate squeezed and entangled light sources with gyroscopes and accelerometers. Such differentiation will lead to the QEIMU prototype and a paradigm shift in inertial navigation and sensing.

Road Map



The two-year phase 2 timeline comprises parallel development of three key project components: 1) quantum-enhanced gyroscope **(end-Q3 deliverable)**; 2) quantum-enhanced accelerometer **(mid-Q4 deliverable)**; and 3) a backend processing and control unit **(end-Q6 deliverable)**. The calibration and testing deliverables for the quantum-enhanced gyroscopes and accelerometers are both **end Q6**, after which we will converge all three components into the integrated and verified QEIMU prototype as the final, **end-of-project deliverable**.

Partnerships

The successful phase 1 outcomes underpin phase 2. In particular, UA's entanglement-enhanced optomechanical sensing proof-of-concept laid the foundation for the QEIMU prototype. In parallel, the phase 1 team developed on-chip quantum-light sources (UA), wafer-scale optomechanical sensor arrays (Purdue), a test and calibration platform (TAMU and NIST), a backend processing and control unit (GDMS), and a modeling tool (UA). Phase 2 will readily integrate these functional modules to demonstrate QEIMU. Specifically, UA and Honeywell will develop a multilayer silicon-nitride platform to assemble QEIMU components. Honeywell will use its world-leading navigation sensor expertise to build quantum-enhanced gyroscope while Purdue and TAMU will collaborate to make triaxial accelerometers. GDMS, a certified electronics manufacturer for defense applications, will deliver QEIMU's backend processing and control unit. JPL will then integrate the individual components into the QEIMU prototype. The QEIMU calibration at NIST will be followed by environment tests at JPL to evaluate performance in future space-borne and aerospace applications.

Intellectual Property

UA and GDMS jointly own IP for an on-chip squeezed-light generation method and entangled radio-frequency photonic sensors. UA own IP for on-chip large-scale entanglement generation and has filed other quantum patent applications. Honeywell owns IP for integrated photonics and gyroscopes. We are coordinating with the tech transfer offices of each party to facilitate the success of the project.

Contribution to Quantum Ecosystem

Our project will help build a quantum ecosystem by distributing turnkey quantum-source modules to Convergence Accelerator teams and the broader community. Our contributions will include collaborating with C520, C695, C737, and other labs on customized quantum photonics, working with C614 and C581 on quantum education, and organizing workshops to engage stakeholders.

QuPID

Quantum Proteomics Insight Device

Lead PI: Peter Maurer
pmaurer@uchicago.edu

Ania Bleszynszki Jayich

Jason Cleveland
jcleveland@somalogic.com

Larry Gold

Karoly Holczer

Overview

Frequent monitoring of a person's proteome is proving to be the ultimate early diagnostic method for human disease. Our project brings together academia and industry to use quantum sensing to overcome a long-standing hurdle, enabling a proteomics device that permits routine testing at the scale of billions of people annually. Such technology will enable inexpensive and reusable proteomics devices which will allow monitoring of personal health trajectories to drive earlier interventions, thus improving the quality of healthcare while also decreasing costs.

Description

Single-molecule magnetic resonance measurements based on quantum biosensing offer a first-time-ever route to routinely analyze massive volumes of patient samples. Enabled by quantum technology we aim to shrink the physical size of currently available proteomics tests from the lab to a chip-scale, point-of-care technology to create QuPID (Quantum Proteomics Insight Device). Our approach relies on the unique high sensitivity of engineered atomic defects in diamonds – nitrogen-vacancy (NV) centers– as powerful quantum sensors capable of detecting individual proteins. Misidentification of proteins is a common problem in biosensing with mitigation requiring complex and laborious workflows. Our innovative technology combines our NV quantum sensors with a customized version of SomaLogic's protein capture molecules (SOMAmers) to allow us to correctly identify each tested protein molecule in a sample as complex as human blood, vastly simplifying the measurement process. Scaling up these quantum sensors to arrays of many millions of simultaneous measuring sites will enable automated measurement of most of the 20,000 proteins in a person's full proteome.

Disease manifests itself as a change in the quantities of proteins the body is producing. Therefore, unlike the genome, measuring the proteome on a regular basis has immense diagnostic value. SomaLogic has completed hundreds of thousands of tests that quantify thousands of proteins (7,000 currently, 10,000 by 2022) in blood and urine, but the tests require a full molecular biology lab with highly trained technicians. The miniaturized format based on our quantum sensor-enabled biochip will bring this test to the scale of hundreds of millions or billions of measurements per year, and will make it economical enough that every human could be measured once a year or more. This device will minimize costs by operating without any moving parts, by utilizing label-free detection techniques to eliminate any chemical processing, and by being reusable many times.

The NSF Convergence Accelerator program has allowed us to bring together leading academic experts in diamond fabrication, surface and biomolecular chemistry, magnetic resonance measurements, and quantum metrology and instrumentation, with SomaLogic, an industry leader in proteomics. Our continued success will enable a commercial pathway for a novel, broad-reaching medical diagnostic device with performance comparable or better than the current lab-based tests.

Differentiators

Routine blood tests already measure a handful of proteins. For example, the blood protein hemoglobin is commonly measured as a

test for anemia. Several existing commercial offerings allow measurement of larger protein panels (typically comprising tens to a few hundred proteins at most). Our approach is to bring measurement of 10,000 or more proteins into a healthcare and home setting, thus liberating patients from the limits of their local healthcare system. By providing a routine but comprehensive "proteome health check", conditions can be identified before symptoms arise to guide further testing and earlier intervention. Further, existing tests compare protein levels to average population levels, but every individual is different. Our device will be economical enough that people can be tested several times per year, creating a personalized baseline that their trajectory can be continuously compared against to flag changes in health driven by aging, disease, and other factors.

Road Map

Our project is based on four thrusts which will converge to a working prototype. **(1)** Demonstrate working surface chemistry for our protein binding measurement on glass **(Q2)**, transfer this protocol to our diamond sensor **(Q6)**, and integrate to an array on a diamond chip **(Q8). (2)** Demonstrate single **(Q4)** and then multiple **(Q6)** quantum sensor amplifiers to ultimately boost binding signal strength in our prototype **(Q8)**. **(3)** Build custom SOMAmers using two methods **(Q4, Q6)** and validate the signals they create when binding proteins using a separate method **(Q6),** and transfer them to our working prototype **(Q8). (4)** Fabricate **(Q2)** shaped diamond surfaces to demonstrate a 10-fold boost in signal strength upon protein binding **(Q4)** and integrate them into our prototype **(Q8)**.

Partnerships

Our Convergence Accelerator project is engaged across all technology sectors involved in the creation, production and use of this proteomic diagnostic device. This includes industry giants such as General Electric, Element 6 (de Beers) and Bruker BioSpin, as well as various start-ups. The world-leading producer of engineered diamond crystals, Element 6, is supplying special materials for us and keenly follows the development of our project. Bruker, a leading producer of magnetic resonance instrumentation, is providing us design and engineering resources to turn the quantum sensing technology we are developing into practical instrumentation. NVision, QDTI and Dust Identity are just a few of the growing number of start-ups using diamond quantum technology for various applications with whom we are engaged in the regular exchange of information. On the application side we are privileged to access SomaLogic's user base of their existing product, SomaScan, which ranges from large pharmaceutical companies such as Novartis and Amgen to individual health care providers.

Intellectual Property

An efficient research collaboration of four academic institutions and Somalogic was achieved by putting in place Material Transfer Agreements during phase 1 of the program. We have an IP management plan and will retain outside counsel to oversee the development of a co-owned strong IP portfolio and its coherent licensing strategy to assure the project's success after the Convergence Accelerator program.

QuSTEAM

Transforming Education for the Second Quantum Revolution

QuSTEAM Initiative

Lead PI: Ezekiel Johnston-Halperin
Johnston-Halperin.1@osu.edu

David Awschalom
awsch@uchicago.edu

Russel Ceballos
rcebal@uchicago.edu

Andrew Heckler
heckler.6@osu.edu

Angela Wilson
akwilson@msu.edu

Overview

We propose to develop QuSTEAM (Quantum Information Science, Technology, Engineering, Arts and Mathematics), a transformational undergraduate curriculum that will provide a national educational model for the emerging field of quantum information science and engineering (QISE). The development of QuSTEAM will rely on collaborative and research-based educational strategies to build a convergent and inclusive curriculum for a diverse community of future scientists and engineers.

Description

The development of QuSTEAM will rely on research-based educational practices to provide a convergent and inclusive curriculum to a diverse community of future scientists and engineers. The severe human-resource shortage in all areas of quantum science and engineering is projected to significantly slow the societal impact of the second quantum revolution. To address this need and accelerate the NSF Quantum Leap, a comparable leap in education strategy is required. The QuSTEAM curriculum will have a modular format with in-person, online, and hybrid delivery modalities to meet the educational needs of diverse stakeholders, including future quantum professionals and members of the current industrial workforce, community colleges, minority-serving institutions, and other bachelors and doctoral degree-granting institutions. We will draw on the extensive expertise of the core participants in both quantum research and STEAM education to create a new curriculum with multiple implementations at the module, class, minor, and certificate level. The curriculum will be piloted at the participating institutions, leveraging partnerships with industry and national labs while demonstrating the ability to train a quantum smart workforce at the scale necessary to support economic development.

Differentiators

Our paradigm shifts away from the hierarchal model of most undergraduate STEM programs in the U.S., where the most engaging, enjoyable content (i.e. content focused on field-leading innovations and societal impact) is traditionally delayed until the later years and students are first introduced to basic skill-building exercises, contributing to student attrition. Further, traditional pedagogy isolates STEM from broader Arts and Sciences engagement. In contrast, the QuSTEAM curriculum will seamlessly blend fundamental skill building with engaging, innovation-focused content from the outset, resulting in an inclusive and student-centered convergent educational experience in line with both modern pedagogy and the workforce needs of the rapid expansion of the community working in quantum information science and technology.

We will continue to directly engage the STEAM education research community to employ evidence-based practices in our curriculum development and will prioritize maximizing opportunities for diversity, equity and inclusion through targeted curriculum, instructor professional development, and independent external evaluation. By drawing from expertise at multiple institutions, students will have access to world-leading experts in QuSTEAM relevant disciplines - far broader access

than is possible within a single university or college. This access will manifest in convergent course material and instruction, for example enabling hybrid in-person and virtual on-line environments that blend experts and students from multiple institutions, providing a unique teaching and learning ecosystem.

Road Map

During **phase 1** needs finding and prototyping we have identified the establishment of a common template for an undergraduate minor and associated certificate programs as the key near term target for workforce development. During phase 2 we will move forward with building out these degree and certification programs, including initial offerings of the critical classes and modules at our respective universities, while continuing with needs-finding and assessment to provide dynamic feedback on evolving workforce needs. We plan to offer our introductory class for the first time in Spring 2022 and the full slate of core classes for our minor in Fall 2022/Spring 2023. By the end of **phase 2**, we anticipate having certificate and minor programs at all 5 of our R1 partners either approved or in-process as is locally appropriate. Our development of elective classes will extend throughout phase 2 and beyond, with the goal of having at least 2 electives defined and delivered at least once by Spring 2023. This education ecosystem is designed to be scalable to the national level, and will provide a template for a novel approach to STEM education more generally.

Partnerships

Our team consists of academic, national lab, and industrial partners. The backbone of our 20-institution academic team consists of 5 R1 institutional partners that have committed to teaching QuSTEAM classes and developing degree programs (Michigan State Univ., Ohio State Univ. Univ. of Chicago, Univ. of Illinois, and Univ. of Michigan) and the IBM-HBCU Quantum Center which is coordinating faculty from 10 of their 23 member institutions led by North Carolina A&T University. In addition, each of the R1 universities have identified one or more partners with whom they have an existing transfer pipeline to support engagement with student populations beyond traditional STEM demographics for a total of 66 faculty with a mixture of STEAM subject matter expertise and discipline-based STEM education research. This academic team is supported by a network of 20 collaborators including academic centers such as the with NSF Quantum Leap teams (all 3 QLCI: CIQC, HQAN, and Q-SEnSE; QII-TAQS), connections to the Convergence Accelerator Phase 1 Track C Cohort (Teams C-520, C-575, C-614, and C-702), and a DOE National Quantum Initiative center (Q-NEXT), as well as 13 industrial partners with interest in quantum workforce development, either as employers or as developers of educational content (or both): Applied Materials, GE Research, Honda, HRL, IBM, JPMorgan Chase, qBraid, Quantum Design, QED-C, Qubit by Qubit, Qutools, SRI International, TOPTICA, and the Unitary Fund.

Intellectual Property

The IP generated by our program will be primarily in the form of copyrighted course materials and educational software in the form of simple simulators. We plan to make all of these materials publicly available using a standard Creative Commons agreement with attribution shared among contributing participants.

SQAI

Scalable Quantum Artificial Intelligence for Drug Discovery

Lead PI: Swaroop Ghosh
Szg212@psu.edu

Mahmut Kandemir
Mtk2@psu.edu

Nitin Samarth
nxs16@psu.edu

Sean Hallgren
sjh26@psu.edu

Nikolay Dokholyan
nxd338@psu.edu

Overview

Drug discovery is time consuming, expensive, and challenging. Extensive time is needed to find ligands that would show promise as therapeutic drugs, and excessive cost is primarily attributed to inferior quality drug candidates that fail in clinical trials. Research scientists engaged in drug discovery will use SQAI to efficiently identify successful drug candidates against target binding sites, therefore reducing time of drug discovery and the costs of late failure, compared to current screening tools.

Description

The discovery of new drugs involves multifaceted challenges and opportunities. First, the size of the search space is enormous, equivalent to the total number of atoms in the universe. Second, we currently lack the tools to efficiently search the chemical space, resulting in a long and expensive drug discovery process. Discovering better methods for identifying possible drug candidates presents an unprecedented opportunity to treat emerging diseases and make healthcare affordable.

To address the search problem, classical approaches rely on machine learning which learns features from example drug molecules to find new drug molecules. However, this approach requires millions of training parameters and still fails to access certain regions of chemical space. We propose quantum machine learning (QML) to characterize the target region of chemical space using entangled qubits and generate high-quality drug molecules with less likelihood of late failure saving cost. It will also screen a large database of potential drug molecules to reduce the number of expensive experimental validation. QML can model classically impossible kernels to explore inaccessible regions of chemical space, thus generating diverse and novel molecules. We will optimize the QML circuits to improve their chances of success and execution cost on noisy and expensive quantum computers. Since noise can limit the search space, we will develop a noise resilient qubit. These activities will generate various quantum concepts and examples that will be used to prepare workshop material for K-12 education and a quantum-ready workforce to address Quantum Initiative Act.

SQAI aims cut down drug discovery time and cost by at least 10X via finding rich library of ligands that can address cancer and other diseases. Our resilient qubit and software toolchains will enable new applications e.g., material synthesis.

Differentiators

Conventional drug discovery uses Generative Adversarial Network (GAN) to generate synthetic molecules using a machine learning model. The model is trained iteratively by comparing the synthetic and real molecules until they become very similar. Although powerful, GAN is cumbersome to train even for generating small molecules. As a result the molecules may show poor affinity to receptor binding site leading to inferior/ infeasible drug. SQAI will develop QML models e.g., quantum GAN to exploit properties e.g., strong expressive power and noise of quantum computers to learn complex molecular distributions faster and generate

high-quality drug molecules. High-dimensional molecular dataset will be handled with small quantum hardware by working on feature space extracted from the data classically or quantumly. SQAI will be proven in real hardware from IBM/AWS. Our team includes, (i) QML expert from UMD to exploit the full potential of quantum computers, (ii) error mitigation expert from GaTech to increase success probability of QML workload, and (iii) drug discovery expert from UNC to experimentally validate our ligand library.

Road Map

Phase 1 has already produced low-fidelity prototypes for our deliverable (Table 1). By **Q3 of phase 2**, we will tune the low-fidelity prototypes of our deliverables to meet our refined approach. Medium-fidelity prototypes of the deliverables will be produced by Q5-Q6 which will be fine-tuned to high-fidelity prototypes by Q8. Nvidia, ApexQubit, IBM and ORNL will test thrusts 1 &2 prototypes, Q-NEXT Center will test thrust-3 prototype, Q-12 and academic partners will test education/ outreach prototypes in Q5. After Yr2, we will seek funding from DOE ASCR, NIH NIGMS, and NSF Bio/CNS programs to sustain our drug discovery efforts. By tuning our ligand library for relevant targets e.g., Ras from UNC, we plan to start a company by seeking funding from venture capitalists and STTR/ SBIR programs in Q10. SQAI framework and toolchains are also useful for domains e.g., material discovery for which we will seek funding from DOE ASCR, NSF-DMR, and DARPA BTO programs. We will also start start-ups in these domains by teaming up with our industrial partners in Q13.

Thrust	Deliverables
1	QML framework (D1); ligand library (D2)
2	Compiler (D3) and scheduling toolchain (D4)
3	π-periodic qubit (D5)
4	K12 outreach and professional development; UG curriculum (D6)

Partnerships

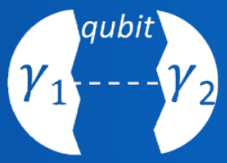
We will leverage, (i) computing resources at Penn State and Google for simulations and training, (ii) quantum hardware from IBM/AWS, (iii) Apexqubit and UNC's drug validation facility, (iv) Q-NEXT's expertise on qubit characterization, (v) BFTP for commercialization. These resources will pave path for conceptualization all the way to realization to achieve success. Our partners from industry, national labs and academia will guide us through technical obstructions in QML-based drug discovery. Collaboration with the Quantum Literacy Network (team C-614) will connect our discovery application to more equitable preparation of a quantum workforce. Cross-team sharing indicates that our deliverables will support many Track C and D teams.

Intellectual Property

SQAI will produce following IP, (i) drug discovery approaches; (ii) library of ligands for particular targets; (iii) π-periodic qubit. We are engaged with Penn State Office of Technology Management.

Topological Qubit

s-TQC: Topological Qubit for Robust Quantum Computing



Lead PI: Jagadeesh Moodera
moodera@mit.edu

Liang Fu
liangfu@mit.edu

Patrick Lee
palee@mit.edu

William Oliver
wi18222@mit.edu

Andrew Potter
acpotter@utexas.edu

Peng Wei
peng.wei@ucr.edu

Overview

The mission of the s-TQC program is to address the complex challenge of quantum error generation in quantum computing (QC) from the fundamental hardware level. Quantum error has been the main issue preventing the realization of full-scale quantum computers that have been pursued by a broad range of research institutes, major industries and startups. Despite the tremendous efforts in the algorithm-based quantum error correction, progress made at the hardware level will serve as the backbone for future quantum information sciences.

Description

In conventional quantum computing, whose basic element is known as a qubit, the computation is by performing operations with a group of qubits using logic gates. Maintaining high fidelity in qubit gate operations, i.e. controlling the quantum state of the qubit coherently and error-free, is a key challenge. Current quantum technology has an error rate of 1 part in 1000, i.e., 7 orders of magnitude higher than in classical computers. Furthermore, existing technology is only capable of fabricating a quantum chip consisting of several tens of qubits, far below for implementing algorithm-based quantum error corrections that need at least 1000 qubits. We hence plan to develop the robust next generation qubit that enables tolerating quantum error physically and intrinsically, thereby addressing this challenge from the hardware level. To do this, we propose to create a highly non-local qubit, which can be envisioned by splitting one conventional qubit into two coherent coupled sub-particles, each of which is known as a Majorana zero mode (MZM, $\gamma 1$ or $\gamma 2$) as shown in our logo. A MZM is a quasi-particle who is its own antiparticle and can be found in a new material known as topological superconductor. A pair of MZMs is equivalent to an electron (or a hole). Physically far separated the pair is yet long-range entangled: can serve to build an exotic non-local qubit that simultaneously dwell at two different locations in space. Such a qubit is immune to decoherence (or error) since any source of decoherence must simultaneously act on both MZMs in the spatially separated pair. Moreover, being an antiparticle of itself, a MZM do not decay and thus is intrinsically fault tolerant. Our recent work has demonstrated the signature of MZM pair on gold surface. In phase 2, we plan to build practical non-local qubits and their circuitry for robust QC.

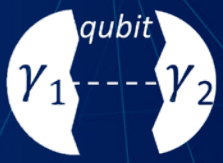
Differentiators

Compared to a conventional qubit, a MZM qubit or topological qubit (T-qubit) has the following advantages:

1. Efficiency: we expect that a T-qubit does not need extensive algorithm-based error corrections. Therefore, one T-qubit is capable of competing with 1000 conventional qubits;

2. Scalability: operating a T-qubit is much simpler than operating a conventional qubit, for example in superconductor-based Josephson junction qubits, and the coupling among T-qubits is easier than that in conventional qubits. This simplifies the design of a quantum chip, greatly enhancing the scalability.

3. Robustness: a T-qubit is based on a new superconductor material (gold) that can



operate at temperatures ten times higher than that for conventional superconducting qubits, thereby allowing the realization of more practical QC schemes. Furthermore, compared to other material platforms that could potentially host MZMs, ours, using common metals such as gold, is robust, stable and readily accessible. This makes our platform highly scalable, which inherently allows building complex quantum circuitry involving multiple T-qubits and is more compatible with industrial needs.

4. Sustainability: With active participation of industrial QC leader IBM in our qubit study, we anticipate a path beyond the phase 2 stage.

Road Map

There will be three milestones for phase 2. **(1-9 months)**: an experiment-theory collaboration work will be carried out to confirm the non-local nature of a pair of MZMs and demonstrate teleportation. This will serve to answer a fundamental question in science community regarding MZM property. We will deliver a nano scale device carefully designed for this experiment and carry out measurements at low temperatures. Once successful, the milestone work will result in high profile publications, which will have a strong influence in the field. **(6-24 months)**: we will deliver the logic element for topological QC i.e. a prototype T-qubit and achieve its operations. We have a pending patent describing how to build a T-qubit. Further, an elemental nano device structure will be delivered for the read-out of the T-qubit quantum state. The device will be tested, and the data will be analyzed by our theory team members to extract the information reflecting the quantum state of the T-qubit. If successful, this milestone work would open new avenues in both science and quantum engineering research directions. In parallel, we will pursue the milestone for theory development on T-qubit. **(1-9 months)**: modeling and simulation will be carried out to support material and measurement design. **(9-24 months)**: we will investigate new schemes for T-qubit error correction and quantum algorithm. This milestone work could enable new industrial startups and push QC industry.

Partnerships

Through discussions in phase 1, we have established the needed partnerships for phase 2 program. **IBM:** will closely partner in the development of T-qubit devices and measurements; **University of British Columbia:** will be involved in T-qubit algorithms and error correction; **University of Waterloo, Institute for Quantum Computing (IQC):** will participate in the development of robust materials for MZMs; **Lincoln Laboratory:** will collaborate in the development of T-qubit circuitry and quantum architectures; **Morgan State University:** will be involved in our Q-STEM outreach for quantum literacy.

Intellectual Property

One patent “Majorana Pair based Qubits for Fault Tolerant Quantum Computing Architecture using Superconducting Gold Surface States”, application number US20200356887A1 is currently pending

TRACK D:

AI-DRIVEN DATA SHARING & MODELING

AI research and development requires access to high-quality datasets and environments, and testing and training resources. The NSF's Convergence Accelerator is funding solutions to address data and model- sharing challenges through tool and platform development to enable easy and efficient data matching and sharing, and privacy protection tools and processes to ensure secure access to sensitive data.

AI-Driven Innovation via Data and Model Sharing funded phase 1 teams include:

American Sign Language

- **AI/ML-based Facial Analytics—**Led by Rutgers University, AI/ML-based Facial Analytics uses AI and machine learning to offer privacy for American Sign Language videos; help students learn to produce these expressions; advance understanding of co-speech gestures; and more.

Civil/Build Infrastructure

- **AI-Grid—**Led by Stony Brook University, AI-Grid is an AI-enabled solution for resilient networked microgrids.
- **Infrastructure Safety Monitoring—**Led by Howard University, Infrastructure Safety Monitoring informs maintenance, repair, and replacement decisions. The ISM solution supports widespread monitoring of real-world structures; ranging in complexity from simple vertical steel utility poles to agricultural structures to more complex structures such as buildings and bridges.
- **InstaTwin—**Led by Oregon State University, InstaTwin is an AI-based

technology that automatically segments, classifies, and extracts real-world features creating intelligent models for building design, energy management, renovation, and emergency planning.

Cybersecurity

- **aiShare—**Led by Carnegie-Mellon University, aiShare helps organizations detect cyber attacks faster. It offers two notable features such as assisting enterprises to generate and share high-fidelity synthetic data to enable the collective training of better attack detection models.

Environment

- **BurnPro3D—**Led by the University of California, San Diego, BurnPro3D, is a platform for public sector collaboration to proactively reduce the risk of devastating megafires. Leveraging the WIFIRE Commons data sharing and AI framework, BurnPro3D uses next-generation fire science in prescribed burns for vegetation treatments at an unprecedented scale.
- **Computing the Biome—**Led by Vanderbilt University, Computing the Biome is creating a data and AI platform for monitoring and predicting biothreats in a major U.S. city, and to drive economic sustainability by empowering businesses and advanced science missions to deliver valuable consumer apps and breakthroughs.
- **CRIPT—**Led by Massachusetts Institute of Technology, CRIPT is an AI-enabled cloud application and database, enabling polymer scientists to easily find and interact with complex data.
- **HydroGEN—**Led by the University of Arizona, HydroGEN is a web-based machine learning (ML) platform, generating custom hydrologic scenarios on demand.
- **Pisces ClimatePro—**Led by Columbia University, Pisces ClimatePro is a cloud-based, AI and machine learning platform that enables any utility or business to reduce their water-related climate risk and improve their resilience to future shocks.
- **Precision Epidemiology—**Led by the University of California, Davis, Precision Epidemiology is an online platform that converges data, AI models, and expertise across the livestock production and health space for animal health management.

Healthcare

- **ImagiQ—**Led by the University of Iowa, ImagiQ is a collaborative AI model for medical imaging. Through novel asynchronous and decentralized federated learning researchers will be able to develop AI models on large extramural cohorts of patient data without regulatory, administrative, and technical impediments.
- **LEARNER—**Led by Duke University, LEARNER is a new privacy-preserving AI model sharing and learning platform used for collaborative big health data mining, while preserving patient personal information.
- **MetaMatchMaker (M3)—**Led by the Research Triangle Institute, M3 is based on extensively trained transfer learning models making finding, accessing, and integrating datasets easier, cheaper, and faster.
- **Model Exchange—**Led by Princeton University, Model Exchange is based on extensively trained transfer learning models making finding, accessing, and integrating datasets easier, cheaper, and faster. The solution increases the pace and reduce the cost of discoveries and can be adapted to any field.
- **STRAIT Consortium—**Led by Vanderbilt University, the STRAIT Consortium is streamlining validation from inception to surveillance of medical imaging AI.

Tech Infrastructure

- **AI Maker—**Led by the University of California, San Diego, AI Maker allows users to input domain-specific keywords and/or sampl e data which locates the most suitable advanced models from a large collection of models.
- **Data Station—**Led by the University of Chicago, Data Station is a new data platform designed to democratize data sharing by lowering technical and human barriers.

AI-Grid

AI-Enabled Programmable Networked Microgrids

AI-Grid

Lead PI: Peng Zhang
p.zhang@stonybrook.edu

Scott Smolka
sas@cs.stonybrook.edu

Scott Stoller
stoller@cs.stonybrook.edu

Xin Wang
x.wang@stonybrook.edu

Yifan Zhou
yifan.zhou.1@stonybrook.edu

Overview

Coordinated networked microgrids (NMs) promise to significantly enhance power grid reliability. Three main challenges prevent their wide adoption: 1) Lack of understanding of NM dynamics; 2) Big data but limited/unscalable analytics; 3) Cyber-infrastructure bottlenecks. This project aims to develop AI-Grid: AI-enabled, provably resilient NMs. Key innovations are a programmable platform integrating reliable modeling under uncertainty, reachability analysis, formal control, high-assurance software architectures, and cybersecurity technologies to enable scalable, autonomic, and ultra-resilient microgrids and NMs.

Description

Microgrids are a promising new paradigm for electricity resilience. In August 2017, for example, multiple microgrids kept critical community services running in Houston despite utility grid outages caused by Hurricane Harvey. Coordinated networked microgrids (NMs), which allow microgrids to coordinate to support various smart city functions, are expected to provide increased electricity resilience during extreme events. As anticipated by the U.S. DOE, R&D of NMs will lead to the next wave of smart-grid research, which will help achieve the vision of a highly resilient grid. NMs are also expected to empower our nation's digital economic engine – the swiftly growing data centers. Recently, global Internet traffic surged by 40% between February and mid-April 2020 during the height of the Covid-19 containment measures, as a result of social activities moving online. This growth, coming on top of an exponential growth in demand over the past decade for data and digital services, makes uninterrupted data services of paramount importance.

Three main challenges have prevented NMs from serving as dependable resilient power resources and thus prohibited their wide adoption: 1) Lack of understanding of NM dynamics under frequent changes in status, ubiquitous uncertainties, fast ramping, low inertia, and non-synchronism; 2) Big data but limited and unscalable analytics as current technologies are unable to handle the volume of dynamic data needed for real-time decision making; and 3) Bottleneck in cyber-infrastructure due to delays, congestion, failures, cyberattacks, and the ever-increasing pace of functional/structural changes which can catastrophically plague microgrid cyber-networks.

To address these challenges, this project aims to develop **AI-Grid:** AI-enabled, provably resilient NMs. The key innovation is a programmable platform that integrates reliable modeling and prediction of system states under uncertainty, reachability analysis, formal control, high-assurance software architectures, and cybersecurity technologies to enable scalable, self-protecting, autonomic and ultra-resilient microgrids and NMs capable of coordinating ultra-scale distributed energy systems and cultivating America's smart communities and cities.

Differentiators

AI-Grid is a hardware-independent, software-defined platform that will enable previously unseen low CAPEX/OPEX and improved social

welfare for communities. It optimizes the use of real-time modeling and analysis to provide low power and energy costs with guaranteed high reliability, resiliency, and cybersecurity. It achieves AI-enabled microgrid operations, learning-based microgrid modeling, and a neural Simplex architecture for runtime safety and security assurance.

Road Map

A functional AI-Grid prototype platform will be fully tested and verified by **2021/11**. Demonstration of AI-Grid will be completed on ComEd's Networked Microgrids in Chicago by **2022/08**. AI-Grid will be implemented in the Energy Management Systems for operating EIP and The Plant by **2023/05**. Finally, AI-Grid models, data, and training materials will be fully accessible to US communities by **2023/08**.

Partnerships

Strong, cross-cutting partnerships are pivotal to AI-Grid's successes. The AI-Grid team has established 29 partnerships with America's leaders of all relevant sectors. The team's end-user partners include *Energy and Innovation Park*, a $1B data center microgrid project; Epic Institute, a global climate solutions organization which will use AI-Grid to manage *The Plant*—an old coal power plant being redeveloped into a global climate exhibition and convention center in NYC; and *Commonwealth Edison's* world famous networked microgrids in Chicago City. The team is developing an open-access programmable AI-grid platform with industry partners Hitachi, Grid Singularity and Schneider Electric. Major power utilities, including ComEd, ISO New England, PSEG Long Island, National Grid, and Eversource will provide data, guidance on grid integration and risk management, and assistance in the evaluation. Industrial partners include prominent companies in the power industry: RLC Engineering, RTDS, SEL, Schneider Electric, Quanta Technology, and Bloom Energy. These partners will provide equipment, data, dynamic models, and technical support. Connecticut Center for Advanced Technology (CCAT) will coordinate engagement between the academic and industrial partners. Brookhaven National Laboratory, with its New York Center for Grid Innovation, will collaborate on research, evaluation, and dissemination of AI-Grid.

Intellectual Property

The intellectual property used in the AI-Grid platform is being developed by the PIs' research groups. The AI-Grid platform will be publicly released to promote broader adoption and impact. IP agreements covering shared data, models, etc., will be negotiated with each relevant partner.

AI Maker

Intelligent and Context-aware Platform for Sharing, Search, and Matching of AI Models and Datasets

Lead PI: Jingbo Shang
jshang@ucsd.edu

Luca Bonomi
Rajesh Gupta
Dezhi Hong
Arun Kumar
Lucila Ohno-Machado
Giorgio Quer

Overview

AI modeling - that is, to discover underlying patterns from massive, domain-specific data – has become demanding in many domains (from smart homes to manufacturing, transportation to grid, health to sociology), e.g., AI modeling for COVID-19 imaging and contact tracing datasets. However, datasets and AI models are often siloed within each domain, posing significant barriers to advancing AI-driven research. We propose a match-making platform, AIMaker, which allows users to input domain-specific keywords and/or sample data and then locates the most suitable advanced models from a large collection of indexed models.

Description

Finding the most suitable advanced AI models for domain-specific datasets is one of the most significant challenges today, according to our extensive user interviews in phase 1. Domain experts are eager to explore beyond classical AI models; however, the lack of AI expertise in model selection and model refinement becomes the key obstacle.

To close this gap, we propose AIMaker, a "match-making" platform that connects domain users to suitable advanced models and/or datasets. AIMaker achieves this by transforming models and datasets into "computational resources" such that model-dataset pairs can be effectively searched and matched based on their semantics and contextual information, i.e., metadata. Sample data from users, when uploaded, will be processed by our privacy-preserving techniques to protect private information while extracting useful information for search, matching, and navigation.

As an illustrative example, clinicians working on COVID-19-related diagnoses based on patient imaging data can issue a simple query such as "Coronavirus hazard assessment from chest CT", and then they would be able to locate desired models and accompanying datasets amongst a myriad of other relevant models and datasets (e.g., pulmonary disease data).

AIMaker will unleash and facilitate novel AI-driven scientific and practical applications in various domains. It would also serve as a sharing portal of models and datasets (without storing the actual data). In phase 1, we have focused on two domains initially – biomedical and energy research. In phase 2, we will expand our pilot study to more domain experts from ten other Track D teams.

Differentiators

Our team provides unique interdisciplinary expertise in representing and integrating AI models and datasets from various domains in a search and matching service: We have contributed to international metadata standards in the IoT world; our partners Snowflake, IEEE DataPort, and OpenML offer the market's largest collection of models and cloud infrastructure. The academic team has a track record in technology transfer and development of standards. Our partner Google is the largest search engine while Amazon and Snowflake are among the largest cloud service providers and stakeholders of AI products. This combination enables us to rapidly develop our platform, connect to vast amounts of models and datasets, and test via pilot studies engaging multiple peer Track D teams focused on domain-specific AI research. This also makes AIMaker unique as a horizontal infrastructure platform that unifies domain expertise from other vertical teams. Our innovative components for extracting, expressing,

and organizing metadata about datasets and AI models in a standardized format are also useful to other teams.

Road Map

Q4, 2021: Refined metadata generation method and privacy protection technique by including more types of information describing datasets and models; assess the performance improvement (e.g., usability and privacy metrics).
Q1, 2022: Decide with external partners on design of APIs to provide to them for AI model metadata extraction and schema alignment.
Q2, 2022: A public, beta version of our platform for domain experts and participating vertical teams to use and test; define key metrics for evaluating search results based on pilot teams' input.
Q3, 2022: Refined matching function based on the public users' feedback and click logs.
Q1, 2023: An expanded collection of datasets and models indexed in our platform by incorporating more cutting-edge models from public sources and external collaborators.
Q2, 2023: Refined key components to ensure their effectiveness on emerging datasets and models.

Partnerships

IEEE DataPort will provide the dataset resources that they have and offer the opportunity to test and possibly deploy AIMaker on their platform to serve the entire IEEE community. **Amazon AWS, Google Tensorflow, Snowflake, Databricks, and Tempus** are interested in our model and dataset schema inference modules. **OpenML** will provide support for model fine-tuning after users are matched with the right models. We will also have the opportunity to integrate our matching service on the OpenML website to improve their search and matching quality.

We will have **various Track D Teams** as vertical pilot users of our system, including **D462, D521, D588** ((bio)medical research), **D532** (federated learning for medical imaging data sharing), **D542** (hydrologic forecasting), **D636** (polymer model sharing), **D655** (civil infrastructure), **D676** (wildfire detection and monitoring), and **D680** (veterinary research and data sharing), **D682** (unified model format). They will provide feedback upon using AIMaker and leverage AIMaker for their model and dataset sharing.

Intellectual Property

Algorithms, prototype systems and other artifacts developed in this project will be open-sourced, allowing maximum dissemination and use of the created materials and non-sensitive data. To ensure the sustainability and broaden the impact, these outcomes will be further integrated into a soon-to-be-launched community platform at UCSD for easy sharing and search of research datasets, AI models, and other project artifacts developed based on the open-source tools Dataverse and MLFlow.

AI/ML-Based Facial Analytics

AI/ML-based Facial Analytics for Natural Language

Facial Analytics for Communication & Education

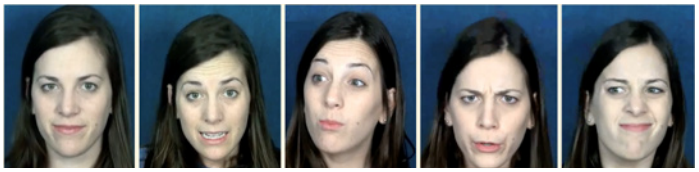
Lead PI: Dimitris Metaxas
dnm@rutgers.edu

Matt Huenerfauth
matt.huenerfauth@rit.edu

Carol Neidle
carol@bu.edu

Overview

This proposal involves development of **sustainable robust AI methods for facial analytics,** potentially applicable across domains but targeted here to new applications that address important problems related to use of facial expressions and head gestures in natural language. In *sign language*, critical linguistic information of many kinds is conveyed exclusively by **facial expressions and head gestures.**



Neutral | Yes-No Question | "If" clause | Wh-question | Negation

The fact that the face carries critical linguistic information poses major challenges for Deaf signers and for students of ASL as a non-native language.

Problem #1: The *>500,000* US **ASL signers** have no way to communicate anonymously through videos in their native language, e.g., about sensitive topics (such as medical issues). This is perceived to be a significant problem by the Deaf community. It also means, for example, that signed submissions to scholarly journals cannot be reviewed anonymously.

Problem #2: **2nd-language learners of ASL** (the 3rd most studied "foreign" language, with US college enrollments *>107,000* as of 2016) have difficulty learning to produce these essential expressions, in part because they don't see their own face when signing. In spoken language, these expressions also play an important role, but they function differently.

Problem #3: The role of co-speech gestures, including facial expressions and head movements, is not well understood because of inadequacies in current analytic tools. This has held back applications that rely on such correlations, such as the development of realistic speaking avatars and robots; technology for the elderly and those with disabilities; and detection of, e.g., deception and intent.

To address these problems, we are creating tools (1) to enable **ASL signers** to share videos anonymously by disguising their face without loss of linguistic information; (2) to help ***ASL learners*** produce these expressions correctly; and (3) to help ***speech scientists*** study co-speech gestures.

Description

The foundation for the deliverables designed to address the problems just described is provided by development of new AI methods for continuous multi-frame video analysis that ensure real-time, robust, and fair AI algorithm performance. The application design is also guided by **user studies.**

1) The ***Privacy Tool*** will allow signers to anonymize their own videos, replacing their face while retaining all the essential linguistic information.

2) The ***Educational Application*** will help learners

produce nonmanual expression and assess progress by enabling them to record themselves signing along with target videos that incorporate grammatical markings. Feedback will be generated automatically based on the computational analysis of the students' production in relation to the target.

3) The ***Research Toolkit*** includes: (a) a Web-based tool to provide *computerbased 3D analyses of nonmanual expressions from videos* uploaded by the user; and (b) extensions to *Sign-Stream®* (our tool for *semi-automated* annotation of ASL videos, incorporating computer-generated analyses of nonmanual gestures) to accommodate speech, and to our *Web platform* for sharing files in the new format. These software and data resources will facilitate

research in many areas, thereby advancing the field in ways that will have important societal and scientific impact.

Differentiators

(1) The proposed AI approach to analysis of facial expressions and head gestures—combining 3D modeling, Machine Learning (ML), and linguistic knowledge derived from our annotated video corpora—overcomes limitations of prior research. It is distinctive in its ability to capture subtle facial expressions, even with significant head rotations. This will have other applications, too, e.g. for sanitizing other data involving video of human faces, medical applications, security, driving safety, and the arts.

(2) The applications themselves are distinctive: nothing like our proposed deliverables exists.

Road Map

By the end of **Year 1**, we will have prototypes, with a partial set of features implemented, for the above deliverables. We will also have conducted user studies to improve our design and validate our tool performance. In **Year 2**, we will continue with user studies and incorporate the feedback, further extend the technologies, and complete the implementation of the deliverables. We will put in place appropriate mechanisms for short- and long-term dissemination/sharing of the software and data.

Partnerships and Sustainability

Gallaudet U. is a partner and will be involved in all components of the proposal. Anonymization is of particular interest for their *Deaf Studies Digital Journal,* as it will solve a problem with peer review of signed video submissions, enabling reviews to be "double blind." Anonymization (as well as the SignStream® enhancements) will be useful for a new gesture database they are building, examining how the Deaf communicate with hearing people. And the educational app will be tested in classes there.

We also work closely with Deaf-owned companies, including the largest distributor of ASL curricular materials in the US and Canada, *DawnSignPress (DSP).* DSP has supported our work in phase 1 and is interested in disseminating our pedagogical application along with its own curricular materials. *DSP* will also be able to anonymize the entries in its all-ASL dictionary (first of its kind) now under development. *Convo Communications*—the only Deaf-owned Video Relay Service provider—is interested in the prospect of using our anonymization tool for its interpreting and messaging services.

For SignStream® enhancements, our collaborators include a group of internationally renowned speech scientists, who will be providing feedback and testing out the software as development proceeds.

We have a longstanding partnership with the Rutgers Laboratory for Computer Science Research (LCSR). LCSR has been instrumental in the development of our annotation tools and Web interfaces for data sharing and will continue to support the products of this project beyond the funding period.

Intellectual Property

- *SignStream®,* our annotation tool for visual language data (to be extended for speech in phase 2), is distributed with an MIT License.
- The annotated video corpora we share on the Web are freely available for education and research.
- The new products will also be broadly accessible.

Additional Information

The convergent research team includes, as PIs:

- **Metaxas** - Distinguished Professor of CS; Director of the Center for Computational Biomedicine, Imaging & Modeling at Rutgers; expertise in ML, 3D modeling, human motion tracking, ASL analytics.
- **Huenerfauth** - RIT Professor and Director of the School of Information and Center for Accessibility & Inclusion Research (which evaluates tools for ASL students and deaf users, as well as linguistic technologies, including facial animations).
- **Neidle** - Linguistics Professor at Boston U. & Director of the American Sign Language Linguistic Research Project; specialist in syntax, ASL linguistics.

This team has been collaborating productively for decades on AI approaches to sign language recognition from video and development of related software tools and annotated video corpora.

aiShare

Improving Malicious Actor Detection with Collaborative Data and Model Sharing

Lead PI: Giulia Fanti
gfanti@andrew.cmu.edu

Nick Feamster
feamster@uchicago.edu

Michael Reiter
michael.reiter@duke.edu

Vyas Sekar
vsekar@cmu.edu

Lior Strahilevitz
lior@uchicago.edu

Overview

Cybercriminals and fraudsters pose a massive threat to American consumers and enterprises. However, detecting these malicious actors is expensive, time-consuming, and unreliable. aiShare is a software platform for enterprises and third-party providers of threat detection software to improve their threat detection capabilities by privately and proactively sharing threat intelligence with other relevant enterprises.

Description

In 2019, payment fraud losses exceeded $27 billion; for instance, retail fraud alone cost American enterprises over 1.8% of their annual revenue. Despite these staggering costs—which are ultimately borne by the end consumer—existing infrastructure for identifying fraud and cyberattacks is costly and ineffective. Most enterprises tackle this problem by applying detection models (both AI and traditional rule-based models) to enterprise data streams (e.g. payments, network traffic patterns, login attempts) to automatically identify malicious activity automatically. These models are typically either: (a) trained locally on private enterprise data, or (b) pre-trained by third-party vendors. Once these AI models identify possible malicious activity, they raise an alert, which must be processed manually by a team of analysts at each enterprise.

Unfortunately, a significant number of these alerts are false alarms; for example, the cybersecurity company LastLine estimates that on average, 50% of alerts in endpoint protection systems today are false positives. Network analysts therefore waste substantial time and money investigating alerts that do not correspond to attacks. Helping operators cope with this overload is our central problem: despite the massive damage caused by malicious activity (both fraud and cyberattacks), existing solutions are prohibitively ineffective and costly. In the words of one of our phase 1 interviewees, *"We are dealing with a fire hose. We can scale up everything except our team [of analysts]."*

The main reason so many analysts are needed to deal with alerts is that existing detection models cannot distinguish between benign anomalies and malicious ones. This happens for two reasons: (1) models are trained on insufficient or irrelevant data, (2) existing techniques for evaluating the confidence of an alert are primitive. aiShare tackles these problems by helping enterprises collaboratively share data about attacks perpetrated by fraudsters and cybercriminals. The aiShare software platform enables enterprises to cooperate in two new ways:

1) Enterprises can generate synthetic data, or fake data that mimics important patterns in the real data without without disclosing proprietary information. This synthetic data can be shared with other organizations to help them develop better AI models of both benign and malicious traffic. For example, if Enterprise X sees lots of traffic from a new botnet and shares synthetic network traces with Enterprise Y, Enterprise Y can learn to detect this botnet's traffic prior to being attacked.

2) Data sharing alone is not a panacea; organizations must also understand how to use shared data. aiShare incorporates privacy-preserving AI to allow organizations to adaptively learn which of their collaborators to trust on

different classes of data. Doing so is essential for building confidence in alerts. For example, if Enterprise X is an expert in classifying payment fraud from Russian IP addresses, aiShare helps other organizations learn to trust X for Russian IPs, and execute encrypted comparisons of alerts with the appropriate parties.

Differentiators

Today, the research community is exploring approaches to data sharing that, by design, hide data from participants and thus limit the detection models organizations can deploy. The utility of these approaches in cybersecurity and fraud detection is hampered, because in practice, analysts must inspect data and models.

By sharing synthetic data, aiShare allows organizations to learn from each other without sacrificing the privacy of their training data. Moreover, organizations can dynamically change their models and easily evaluate the quality or relevance of data from a provider. aiShare also uses encrypted computation only to process final model outputs, rather for the full AI learning pipeline. This enables different organizations to collaborate even if they locally use models with completely different rule sets, architectures, or input features.

Road Map

Phase 2 will involve prototype development, evaluation, and basic research to support the prototype. On the prototyping and evaluation front, our first year will have three parts: **(1)** developing the synthetic data backend by December 2021; **(2)** developing the adaptive model selection backend by April 2022; **(3)** developing a front end user interface by July 2022; **(4)** running preliminary user tests by September 2022. In the second year, we will run longitudinal studies with each industry partner to evaluate our prototype on their internal use cases. On the research front, we will study fundamental questions related to the prototype, including **(a)** developing synthetic data models with better privacy-utility tradeoffs, and **(b)** studying techniques to enable organizations to determine when and how to utilize shared data. This includes accounting for malicious behavior from participants and learning dynamic policies for data selection from diverse data sources. After two years, we will have a fully-functional prototype that incorporates our latest research and has been evaluated by four key industry partners for technical proficiency and business and legal practicality.

Partnerships

Malicious activity affects different sectors differently. We are therefore engaging industry partners from different sectors, and secondary partners in industry, academia, and nonprofits. Our primary industry partners are J.P. Morgan Chase (finance), Siemens (energy), and Bosch (consumer electronics), with secondary partners including VMware and Datavisor. During phase 1, these partners shaped our understanding of when data sharing can be used in the context of cybersecurity and fraud detection. In phase 2, these partners will help us deploy a prototype of aiShare on internal use cases that have been crafted specially for each of these companies. Because our partners represent different economic sectors with different concerns, each of them will evaluate different use cases. Collectively, they will demonstrate the feasibility of these ideas in a rich portfolio of scenarios.

Intellectual Property

The technologies have been (and will continue to be) released under a permissive open-source license.

BurnPro^{3D}

Fighting Wildfires Proactively with WIFIRE Commons

Lead PI: Ilkay Altintas
ialtintas@ucsd.edu

Yolanda Gil
gil@isi.edu

Kevin Hiers
jkhiers@talltimbers.org

Rod Linn
rrl@lanl.gov

Overview

A century of suppressing wildfires has created a dangerous accumulation of flammable vegetation on landscapes, contributing to megafires that risk human life and property, and permanently destroy ecosystems. Small controllable fires can dramatically reduce the risk of large fires that are uncontrollable. BurnPro^{3D} is a decision support platform to help the fire response and mitigation community understand risks and tradeoffs quickly and accurately to more effectively manage wildfires or conduct controlled burns.

Description

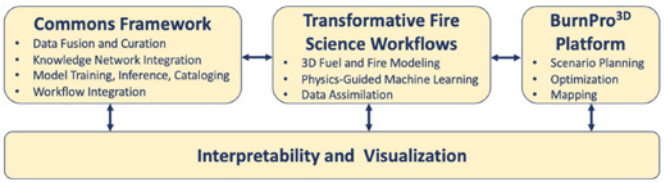
In 2020, wildfires swept across 10 million acres in the western US, killing dozens, destroying 10,000 structures, and causing $15 billion in property damage. Tens of thousands of firefighters risked their lives to fight the fires. The 2019-2020 fire season in Australia was a warning that the problem can worsen worldwide. Almost 50 million acres burned, driving some species to extinction, and emitting 300 million tons of CO_2.

Over the last decade, the WIFIRE team developed the first digital infrastructure to accelerate fire science and management through data, computing, and artificial intelligence (AI). Our initial efforts transformed the way fire response is being managed in California during the first six hours of a fire. However, even the most sophisticated approaches to fighting fires are *reactive* and cannot always control fire under extreme weather conditions. There is an urgent need to turn to *proactive* approaches to reducing the risk of severe fires by removing dangerous accumulations of vegetation.

Our platform, BurnPro^{3D}, is powered by the next- generation WIFIRE Data and Model Commons. It provides diverse users a common ground for understanding risks and tradeoffs related to controlled burns and wildfire management. BurnPro^{3D} can be used in fire mitigation when land managers prioritize land for treatment and burn bosses conduct controlled burns. It can also be used in fire response to evaluate unplanned fires for opportunities to work with incident commanders to manage wildfires to benefit ecosystems, instead of focusing solely on suppression. BurnPro^{3D} is the only space for active collaboration among these users, providing 3D, high-resolution models to increase the proactive use of fire to end devastating megafires. In both fire mitigation and fire response, BurnPro^{3D} can also support users in communicating risks and tradeoffs to regulators and the public.

Differentiators

Currently, wildfire related data, models and interfaces exist in silos. WIFIRE Commons



uniquely *brings these resources together*. As an example, with our partners, we generated the first 3D fuel datasets at the continental scale at 1m resolution to ingest into next-generation fire models and connect to the BurnPro^{3D} platform. The Commons enables the AI innovations necessary for users to optimize controlled burns and improve wildfire management using *fire model predictions made accurate* by 3D weather and vegetation data at a 30X higher resolution than previously available.

To achieve this vision, we are developing specific AI innovations to: (i) Use knowledge management

techniques to fuse data coming from diverse sources and prepare it for fire modeling; (ii) Conduct physics-based machine learning within next-generation fire models to use deep learning to understand complex processes that drive fire behavior; (iii) Apply constraint optimization methods to address complex tradeoffs in the decision process for the placement and timing of controlled burns; (iv) Employ explainable AI to increase the interpretability of data and models by diverse users all along the decision-making chain.

Road Map

Over the next two years, we will work closely with a small group of fire mitigation and response test users. In the first year, we will focus on creating proof of concept - showing that our users can collaborate more effectively via the BurnPro^{3D} platform. We will spend the first three months fully scoping their requirements, followed by six months building out the proof of concept for the platform, ending the year with three months of user testing. The second year will be used to develop BurnPro^{3D} into a minimum viable product and work with our collaborators to disseminate for broad use. In parallel, we will continue to develop the WIFIRE Commons framework that BurnPro^{3D} is built on, creating a space for the scientific user community to collaborate on AI-enabled fire science. We will also host sessions with potential future users who would benefit from platforms developed to solve related challenges, e.g., making decisions about how to manage power lines to prevent fires or when it is safe to send mutual aid during a fire. Last, with our NGO partner, the Climate and Wildfire Institute, we will solidify a model for long-term sustainability.

Partnerships

As exemplified by our 50+ collaborators and partner institutions, our multi-disciplinary team is supported by a consortium of influential advisors and users from 12 academic, ten for-profit, 16 government, and eight non-profit entities, in addition to seven other NSF Convergence Accelerator teams. Through these partnerships, we are building a culture of creating public/private partnerships as a vehicle to extend reach and use, while ensuring integration into existing systems for fire response and mitigation.

Initial partners contributed expertise, data, models, model products, prototype testing, evaluation and feedback. These core partners will adopt WIFIRE Commons for their operations and serve as the beta users of the BurnPro^{3D} platform. They will be joined by an expanded list of partners focused on transitioning our research to practical use, including various USFS Stations, U.S. Geological Survey, California Governor's Office of Emergency Services, Orange County Fire Authority, Los Alamos National Laboratory, Sandia National Laboratory, National Oceanic and Atmospheric Administration, NASA. Our convergence research effort focuses on accelerating solutions by transferring technology to agencies in ways that complement existing focus and strategic direction.

Intellectual Property

The WIFIRE Commons team is committed to open-source development and embraces the FAIR principles. The data and models in the Commons will be made available for use through open license for researchers. We will create restrictive data and software distribution and use agreements when necessary. All other data and models will be publicly available. Commercial applications built on top of the Commons by our external partners and others will be encouraged.

Visit us at https://wifire.ucsd.edu/commons

Computing the Biome

Sensing and Predicting Biothreats with AI

Lead PI: Janos Sztipanovits
janos.sztipanovits@vanderbilt.edu

Ethan Jackson
ejackson@microsoft.com

Overview

Individuals, industries, societies, and governments want to stay healthy. They need cost-effective systems to detect biological threats and predict future disease outbreaks as early as possible. COVID-19 acutely and painfully demonstrated the impacts of the unpredicted. The goals of this program, *Computing the Biome*, are twofold: (1) demonstrate an extensible data and AI platform that continuously monitors and predicts biothreats in a major U.S. city, and (2) create a framework for economic sustainability and global scalability of these results, by empowering businesses and advanced science missions to consume predictions and produce valuable consumer apps and breakthroughs.

Description

Predicting biological threats is hard. Earth's *biome* is home to hundreds of millions to possibly a billion species ranging from nanometer-sized viruses to kilometer-sized forests. These species are interconnected, co-evolving, and *moving* at breathtaking scales and speeds. As a result, biological threats such as emerging diseases, invasive species, and agricultural pathogens can appear unexpectedly and quickly harm our societies and ecosystems. They already cause hundreds of billions of dollars per year in economic damages.

Predicting these will require: (1) *continuous data* streams not yet available today, (2) *detailed models* harnessing expertise from across the science domains, and (3) modern *AI platforms* that use data and models to *compute the biome* in real-time – just as we continuously compute weather forecasts using real-time data streams and models. Fortunately, revolutions in sensing technology, AI, and consumer demand are about to transform how we compute the biome and predict threats. First, this team will produce and interconnect novel data streams ranging from kilometer-scale *hyper-local* weather, to *autonomously identified* disease transmitting insects (only millimeters in size), to *genomically recognized* known and novel viruses (only nanometers in size) – demonstrating that cross-cutting continuous data streams for biothreat detection and prediction can be rapidly unlocked.

Next, the team will combine their expertise in ecology, epidemiology, and virology to design new predictive models and anomaly detectors. Our team will develop the first of these high-impact AIs focused on predicting mosquito-borne diseases, which are difficult to control and impact over 600 million people per year. More broadly, the resulting data platform will empower development of new foundational methods for use by the AI community – based on real-world data and grounded in *the* societal challenges of our age.

Finally, economic sustainability will depend on a vibrant ecosystem where businesses and global missions can consume state-of-the-art models and produce applications and insights that people want to use. Even before COVID-19, the U.S. spent >$1 billion per year on biothreat mitigation. We want to deliver solutions that benefit these critical efforts.

Differentiators

Our main premise is that only a modern sensor network – that continuously monitors species at geographic scales across environments – will be capable of predicting complex biothreats early enough to manage risks. This perspective is based on the successes of existing

sensor networks and AI models to monitor and predict other complex phenomena (e.g. weather systems, smart power grids, and transportation systems).

Today, outbreaks of human disease are usually detected through clinical case data, news reports, and other digital data. WHO's GOARN system is a global aggregator of many of these data sources. It has successfully detected outbreaks early, but generally not early enough to stop their spread.

On the other hand, efforts like USAID's PREDICT program preemptively sampled the environment to look for future novel threats, even sampling coronaviruses in bats in China prior to COVID-19. However, these programs rely on manual sampling. We believe new platforms and AI could make these search efforts more efficient and cost-effective.

Road Map

Our first user is Harris County, Texas – home to the city of Houston and 4.7 million people. *First six months (foundations)*: (1) a unified data platform housing new biome data streams and tools for simulating biomes, (2) an equitable AI that uses simulations to design fair sensor networks – to be released as a global health planning tool, (3) an announcement and hackathon coinciding with WHO's *World Health Day.*

First year *(protect against known):* (1) socially equitable deployment of a sensor network into urban areas with high risk at *West Nile Virus (WNV)*, (2) streaming of biome data into continuous predictions, and (3) release of public health and clinical risk tools to protect communities.

Year 1.5 *(detect unknown)*: (1) development of biome baselines for detecting anomalies such as insecticide resistance and invasive species, (2) release of upgraded tools that guide insecticide use to reduce dangerous resistance and maximize sustainability, (3) recommend discussion of these results at the *World Economic Forum*, where human health and environmental sustainability are likely to be major intertwined topics.

Year 2 *(sustainability):* (1) real-time biome models going beyond WNV to other threat classes such as emerging human and agricultural pathogens, (2) AI-based biodiversity models, (3) creation of a non-profit to manage infrastructure and support business and science access.

Partnerships

(1) *Microsoft:* sensor nodes, species recognizers, models, and industry leadership. (2) Tomorrow.io: hyperlocal local climactic models for habitat suitability including newly launched satellite-based weather radars. (3) *Harris County Public Health:* equitable deployment and management of systems over the 1,800 mi2 of Harris County, Texas. (4) *Vanderbilt University:* open-source data platforms and application design studios for the wider community, and academic leadership. (5) *Johns Hopkins University:* AI-ready disease control policies and coordination with global health missions. (6) *University of Pittsburgh:* genomic data analytics for microbial threat detection and liaison with biotech stakeholders. (7) *University of Washington:* AI-enable epidemiological models and forecasts built on top of the above capabilities

Intellectual Property

Open platforms will be utilized, and arrangements have been made for data and code releases under open data and code licenses.

CRIPT

The Community Resource for Innovation in Polymer Technology (CRIPT)

Lead PI: Bradley Olsen
bdolsen@mit.edu

K. Aou | D. Audus | R. Barzilay | C. Borg

J. de Pablo | T. Jaakkola | S. Jegelka | K. Jensen | K. Kroenlein

Overview

Polymer materials, ranging from clothing and personal protective equipment to construction materials and food packaging, are fundamental to providing for our basic needs for food, shelter, health, and transportation. However, developing new polymers for next-generation products takes decades, and we must move faster to remain competitive. To accelerate this process, we are developing CRIPT, a polymer data ecosystem consisting of a web-based application and cloud database that allow polymer scientists to easily find, archive, and interact with complex polymer data. AI-driven chemistry tools and data-driven workflows within CRIPT will reduce the development time for polymer materials by an order of magnitude, creating a transformative impact on both the producers and buyers of the nearly $600 billion of polymers sold each year.

Description

The diversity of polymers and their properties has enabled them to fill critical roles in nearly every sector of our modern economy. However, this diversity also yields an incredibly large chemical design space, making it extremely challenging to navigate. To make the development process tractable, scientists often search locally, starting from known solutions and exploring new chemical designs around proven polymers. Being able to quickly review existing material designs and identify those having optimal properties is critical to accelerating polymer development.

Currently, searching among existing polymers is a daunting task because polymer data exists as small, disparate sets, making the navigation a complex process combining the harmonization of different data formats and the reconciliation of metadata, both of which currently require expert intervention. CRIPT offers a cloud database based on a new polymer-specific data model that simultaneously provides interoperability across different domains of polymer science and engineering, while retaining critical metadata that allows domain experts to correlate information across many independent records. A series of chemically-inspired AI innovations, including a chemistry-based query language, a graph-based schema preserving temporal structure in data, algorithms for automatic data validation, AI-human cooperative tools for data ingestion, and the integration of machines into the data ecosystem are also provided to add FAIR principles, trust in data, and ease of use to the system.

We anticipate extensive adoption of CRIPT will have a significant economic impact by leading to a more than $1B reduction in R&D costs for new material innovation and a societal impact by accelerating the speed of discovery twofold of the next generation of medical therapies, sustainable packaging, lightweight transportation materials, recycling technologies, and advanced textiles directly improving the quality of life for all Americans.

Differentiators

Current digital data solutions for polymers are fragmented and presented without adequate context to make data findable, interoperable, or reusable. Polymer data is provided as disparate, specialized sets each containing, at best, hundreds of polymer samples. In most cases, metadata associated with the reported properties, such as how physical or chemical measurements were carried out, as well as how each material sample was made, are not explicitly provided, limiting the data's utility.

CRIPT resolves these issues by structuring data so that connections can be easily drawn between polymers, processes, and properties. This data structure documents and indexes the heretofore lost history associated with the making

of a material. This feature alone provides a key advantage over existing polymer data solutions, wherein scientists must rely on rare domain expertise and multiple data sources to obtain the same information. As materials history provides a rich context for how a material is made and characterized, this allows CRIPT to assimilate data across sources with highly different standards, reconciling data and providing universal access.

By implementing new advances in AI, CRIPT's ecosystem enables an intuitive workflow: a new chemical structure query language provides for intuitive data search, natural language processing is used to translate users written descriptions into the standard data format, and validation tools help users to trust the data that they are encountering within the ecosystem. Coupling these features with visualization, analysis, and data set construction tools enables teams to share, collaborate, and communicate like never before to accelerate their innovation.

Road Map

Q3-Q4 2021. Partner-centered design phase: Refine CRIPT data model, ingestion tools, and functionality working with our select early-adopter partners. Milestone: Onboard all 13 of our early-adopter partners and their hardware.

Q1-2 2022. Community adoption phase: Design informative workflows and datasets with community users. Develop security infrastructure for housing open datasets. Milestone: Release CRIPT as a public tool with ingestion and visualization features for the community.

Q3-4 2022. Industrial expansion phase: Develop data privacy infrastructure for working with industrial partners and ability to use tools across disparate data collections. Milestone: Develop private instances of CRIPT and integrate tools from Track D partners for data privacy and federated learning.

Q1-2 2023. Incorporation phase: Establish corporate structure for start-up overseeing continual development of CRIPT. Milestone: Spin CRIPT into an independent non-profit that continues to maintain and improve the ecosystem.

Partnerships

Our multidisciplinary program brings together experts across academia (MIT and University of Chicago: polymer domain experts & computer scientists in AI/ML), industry (Citrine Informatics: database specialists & software development; Dow: industrial polymer experts), and government (National Institute of Standards and Technology: expert in informatics and standards). Our partnership with 6 academia research laboratories, 4 multi-institute research centers, 3 government labs, 5 materials manufacturers, and 3 industrial polymer consumers spans the full spectrum of stakeholders. We will also directly collaborate with 4 other teams within Track D, both leveraging their tools for searching data sets and models, federated learning, data privacy, and 3-dimensional modeling developed in the NSF Convergence Accelerator program within our own ecosystem. Conversely, our tools for validation, schemas for organizing and integrating small data, search tools, and data extraction algorithms will make valuable contributions to their efforts. To help catalyze further collaboration and community engagement, we will host several symposia and short courses for Track D and the polymer community.

Intellectual Property

CRIPT's AI innovations will be freely distributed. The data models and conceptual designs will be openly licensed; the associated code will be distributed open source under the MIT license. The code to operate the CRIPT app will be proprietary to CRIPT. Data will be FAIR compliant.

Data Station

Combining Data, Compute, and Market Forces

Lead PI: Raul Castro Fernandez
raulcf@uchicago.edu

Kyle Chard
chard@uchicago.edu

Ian Foster
foster@uchicago.edu

Michael Franklin
mjfranklin@uchicago.edu

Overview

Whenever data and models are shared, transformation ensues. Breaking down data silos unleashes value that makes companies and researchers more competitive. Entire disciplines change when researchers share benchmarks and models. However, three barriers prevent effective sharing: easy access to sensitive data, data discovery and integration, and data governance and compliance. Each of these challenges has both technical and human components. Data Stations are an entirely new approach to data management that directly addresses these challenges. Data Stations facilitate myriad data sharing scenarios that will democratize access to data and models, and ultimately unleash the value of data.

Description

The lifecycle of data-driven discovery centers around discovering relevant datasets, combining them, and accessing them. Delays in this lifecycle limit the ability of organizations of all kinds to extract value from data. While existing software addresses these barriers individually, no existing solution is able to tackle them collectively. In most cases, a solution to one challenge conflicts with the solution to another. For example, it is harder to discover relevant datasets when access is restricted, and to govern data when underlying datasets are not well integrated.

Tackling these challenges collectively requires a radically new data architecture to address both the *technical* and the *human* problem. Such an architecture must change how people access, share, and use data. To achieve these goals, we are developing a new architecture we call Data Stations.

Differentiators

In the Data Station architecture, both data and derived data products-such as ML models, query results, and reports are sealed and cannot be directly seen, accessed, or downloaded by anyone. The key idea is that interactions are inverted from the traditional model: instead of delivering data to users, users bring questions to data. For example, rather than download a dataset to train a ML model, a user presents the Data Station with examples of their desired results—sample data along with the variable to predict---and the Data Station identifies a suitable data and model combination, trains and optimizes various models on the data, and makes the best trained model available for inference (according to user-defined metrics of “best”). This inversion of compute and data mitigates many security risks of sharing sensitive data and democratizes the ability to ask questions of, and derive value from, data.

In today’s data platforms, data users must know what data are available before they can write a query to extract value from those data. In Data Stations, users cannot see the data a priori, so users must instead describe the goal they want to achieve and rely on the Data Station to compute a result, addressing the *technical* challenge. With Data Stations, data and compute are centralized in the platform, and this opens up opportunities to design and deploy incentive mechanisms, that encourage people to document data better, making it

more usable, more shareable, and increasing its value. In Data Stations, incentive mechanisms and market forces help address the *human challenge.*

Road Map

In phase 1 we have developed two low-fidelity prototypes, established strong feedback loops with a host of collaborators, and used the lessons learned to build a first prototype of Data Stations that we are working to deploy in our collaborators' infrastructure. Phase 2 focuses on execution in two directions: technical and convergence.

Q4 2021. Complete beta version of Data Station platform.

Q1 2022. Start pilot test of augmentation layer (Data Stations running alongside existing infrastructure). First focused meeting with fellow cohort teams.

Q2 2022. 2nd Workshop on Data Stations with industrial and academic collaborators. Completing augmentation layer.

Q3 2022. Host collaborator models and datasets in publicly available Data Station prototype.

Q4 2022. Second focused session with cohort teams and other collaborators. Tech transfer of research in incentive mechanisms to Data Station platform.

Q1 2023. Announcement of publicly available open-source version with a sustainability plan.

Q2 2023. 3rd Workshop on Data Stations with industrial and academic collaborators.

Q3 2023. Complete tech transfer. Launch of data-sharing consortia with collaborators.

Q4 2023. Open-source sustainability plan, onboarding materials, online presence for each data-sharing consortium.

Partnerships

NSF Convergence Accelerator Team, CRIPT (D636): We will collaborate to identify incentive mechanisms to encourage public/private organizations to share their data and deploy Data Station platform for the CRIPT.

BASF (largest chemical producer in the world): With a large data integration/discovery team, we have partnered with BASF and deployed a prototype in their environment. We will use this experience to refine the Data Stations' architecture and interface.

State Farm (insurance company): Cataloging, compliance and governance are some of the key issues Data Stations will address. We have partnered with State Farm to deploy Data Stations in-house, to refine and improve features in real applications.

Snowflake: As a processing layer and integration with existing data infrastructure, Snowflake and Data Stations complement each other.

Intellectual Property

Our sustainability plan revolves around an open-source community, including the software components, training materials, and tutorials.

Visit us at https://github.com/TheDataStation/

HydroGEN

A Machine Learning Platform for Hydrologic Scenario Generation

Lead PI: Laura Condon
lecondon@arizona.edu

Reed Maxwell
reedmaxwell@princeton.edu

Peter Melchior
melchior@astro.princeton.edu

Nirav Merchant
nirav@arizona.edu

Overview

HydroGEN is a web-based machine learning (ML) platform to generate custom hydrologic scenarios on demand. We combine powerful physics-based simulations with ML and observations to provide customizable scenarios from the bedrock through the treetops. Without any prior modeling experience, water managers and planners can directly manipulate state-of-the-art tools to explore scenarios that matter to them.

Description

Water is the driving force behind extreme events like floods, droughts and wildfires. These events have cost the US $234.3B in damages just in the past three years, and this figure is projected to increase. Recent events like the record setting wildfires in California and the mega drought on the Colorado river are merely the latest illustrations. Historical data are no longer a reliable guide for the risks we will face in the future. This uncertainty poses a huge challenge for decision makers.

The scientific community has developed models that can simulate complex changing systems. However, they are too computationally expensive for non-modelers to develop and use. As a result, the tools used for decision making lag behind the science and are often severely limited in their ability to predict evolving systems.

HydroGEN places sophisticated models in the hands of planners and decision makers. We train ML emulators on advanced physically based simulations and the observations, letting our users build customizable scenarios without any prior ML experience. Our platform goes beyond streamflow and is designed to provide spatially distributed simulations of complete watersheds.

Early adopters include water management agencies and resource managers interested in wildfire risk. Our External Advisory Board of second-generation users include water utilities, management districts, consultants, state officials and non-profits.

Figure 1: Illustration of the proposed HydroGEN platform

Differentiators

We know that data-driven models are not well equipped to predict out-of-sample behavior. What sets our team apart is our ability to train ML models for extreme events. HydroGEN builds off the first and only physics-based high-resolution groundwater surface water model in the US. This gives us a unique ability to train ML models using state of the science tools with a proven ability to capture watershed changes in both the surface and subsurface for events that have not yet happened.

Another major barrier to entry for advanced simulations are compute and data requirements. We have designed a scalable approach that allows our users to bring their own compute allocations and apply them directly in workflows without requiring any expertise in cloud

computing. Thus, we can rapidly grow our user base with minimal hardware requirements.

Finally, we have integrated two early adopters directly into our team and provided funding for them to closely engage in our design process. This means that in addition to our broader pilot studies, we will have immediate and direct input on our development from early adopters with projects that are already reaching millions of Americans.

Road Map

Year 1 is focused on our minimum viable product.

- We will release our first watershed prototype by the end of **Q1**. This will launch our first pilot user experience.
- The beta web interface will provide a platform for our first user experience tests (**Q2**).
- By the end of Year 1, we will have a functional national platform that can execute our end-to-end workflow using internal and external compute resources.

Year 2 is dedicated to improving performance, refining design and building our user base for sustainable phase 3 operations.

- We will launch additional pilot studies using the first national platform release (**Q1**).
- Our second release (**Q2**) will include improved designs as well as performance improvements for our ML architecture.
- The third release (**Q3**) will launch our pilot subscription-based model for phase 3. The second half of Year 2 will focus on broadening our community of users for phase 3.

After **Year 2**, we envision that phase 3 operations rooted in a subscription-based services.

Partnerships

Our core development team includes interdisciplinary experts from university, industry and government agencies. Software development is led by CyVerse; an NSF cyberinfrastructure project specializing in data and a workflow management with more than 84,000 users from thousands of institutions. Additionally, we have partnered with ViQi, a software company specializing in large-scale ML and data processing with interactive cloud-based visualization. Our modeling approach is built from the HydroFrame platform – the only high-resolution physically-based groundwater surface water model available for the US. Our ML team is led by the Princeton Center for Machine Learning with ML experts at three academic intuitions.

Our two early adopters are: (1) the Bureau of Reclamation, which is the nation's largest wholesale water supplier providing water to more than 31 million people and 10 million cres of farmland, and (2) WIFIRE, an operational fire response system for the State of California with more than 800,000 users. In addition, we have assembled an External Advisory Board that will guide our pilot studies and expanded user engagement. Our board includes federal agencies, regional and state-level water managers across five states, environmental consultants, national water programs and environmental groups (e.g., USGS, Water Now Alliance, America water, The Nature Conservancy).

Intellectual Property

All of the core tools of our platform are open source. We intend to continue an open-source development model for the tools developed in phase 2. Our sustainability plan follows a Software as Service model – we intend to implement a subscription-based model for access to the HydroGEN platform. Any commercial licensing will be supported by Tech Launch Arizona, the technology transfer arm within the University of Arizona.

IMAGIQ

Asynchronous and Decentralized Federated Learning for Medical Imaging

Lead PI: Stephen Baek
stephen-baek@uiowa.edu

Paul Chang
pchang@radiology.bsd.uchicago.edu

Daniel Rubin
dlrubin@stanford.edu

Nick Street
nick-street@uiowa.edu

Xiaodong Wu
xiaodong-wu@uiowa.edu

Overview

Medical imaging researchers routinely become frustrated due to numerous impediments related to data collection and sharing. *ImagiQ* is a peer-to-peer (P2P) network for distributed (federated) machine learning, connecting medical imaging researchers across the world for collaborative development and exchange of AI models, without the barriers of patient data sharing.

Description

Experts expect that AI will be at the core of clinical practice soon. A critical bottleneck towards this future, however, is the inherent *'data-hungry'* nature of AI models. In fact, development and validation of AI models require many annotated samples. However, such annotated data are incredibly expensive in medical imaging due to multitudes of reasons such as extensive physician time, strong ethics/privacy regulations set around medical images, etc., making extramural collaborations absolutely crucial. Unfortunately, in the current practice, data sharing between institutions is notoriously difficult due to regulatory, administrative, and technical impediments of patient data sharing. The key idea of federated learning is to avoid these practical challenges of data sharing by *not sharing data at all*. Instead, federated learning shares an AI model across different data sources, gets the model trained on locally available data, and aggregates the learning results to update the model. As a result, medical imaging researchers can train AI models on large and diverse cohorts of data, validate these AI models more thoroughly across different data sources, and hence, benefit patients with better predictions ultimately.

Differentiators

Unlike the existing federated learning solutions, where the learning process is governed and controlled by a centralized model server, ImagiQ is based on a novel technology called 'Peer-Adaptive Ensemble,' which democratizes and makes it infinitely scalable to a large network of hospitals. ImagiQ is a dynamically growing eco-system of AI models that can travel around different hospitals. The AI models learn to make better decisions as they see more diverse patient cases while visiting different hospitals. These traveling AI models form an ensemble, or a 'committee of AI models' at each hospital, in order to provide the most reliable and the most trustworthy decisions to the clinicians and patients. The committee members (AI models) are selected based on how well they perform locally on the hospital they are in. This tailors the AI committee to be optimized for the hospitals specific imaging devices/protocols as well as their unique patient demographics, which can vary from hospital to hospital. Such features make ImagiQ standing out of its competitors, by enabling more reliable and trustworthy predictions, as well as covering more diverse patient demographics and imaging protocols through the peer-adaptive ensemble.

Road Map

The first and foremost milestone would be the release of ImagiQ Version 1.0 by the **4th quarter of 2022.** ImagiQ Beta is available at https://github.com/stephenbaek/imagiqfl as

part of our **phase 1** deliverable. Our **phase 2** efforts will include activities to build a community around this beta version software such that more users can contribute by sharing their feedback (e.g. bug report) and posting feature requests, as well as by contributing lines of codes to this open source software. To this end, we will also aim to establish an advisory committee among developers of ImagiQ for the sustainable development and maintenance of the software, beyond the scope of this convergence accelerator project. Another major milestone is to accomplish clinical validations on more diverse medical imaging tasks, by the end of phase 2. We acknowledge that through validation of performance and usability in the context of the real clinical workflow is critical for the successful dissemination of the technology. Finally, we are enthusiastic about the translation of research outcomes into a profitable business model. To this end, we will finish conversion of the current provisional patent under preparation to a full utility patent application. Furthermore, we will navigate opportunities of either creating our own venture or licensing the technology to industry partners.

Partnerships

The core members for phase 1 included five academic institutions across the United States, including the University of Iowa, Stanford University, University of Chicago, Harvard University, and Yale University, as well as industry AI/medical imaging leaders such as NVIDIA, Lunit, IDx, and Imagoworks. During phase 1, these members have exchanged their ideas and visions through monthly meetings, interviews, and other pop-up meetings. In phase 2, we will expand our circle by including an overseas partner (Yonsei University) to increase the data diversity and to pose more realistic, practical challenges in federated learning. Furthermore, Inference Analytics, Inc. will newly participate in this project from phase 2, to facilitate the development of data discovery module.

Intellectual Property

The team plans to put together a provisional patent application to the United States Patent and Trademark Office (USPTO) that covers IP claims around the core concepts and potential embodiments of peer-adaptive ensemble learning. The source code implementation will be open to the public free of charge for both commercial and non-commercial uses, details as specified under Apache 2.0 License.

Contribution to Track Success

The team is dedicated towards the success of the Convergence Accelerator program, Track D. We engaged in a lot of productive cross-team discussions other than the ones assigned by the NSF. The strength of our team is the scalability and generalizability of our technology beyond medical imaging. For example, team Precision Epidemiology (D680) is developing an animal disease portal, which will produce unique collaborative opportunities for both teams. The team D680 will benefit from our federated learning solution such that they can expand their animal disease network to those who cannot publicly share their data. Meanwhile, our team will benefit from the experience and the expertise of D680 for the development of a shared data discovery portal. Similarly, we will collaborate with aiShare (D675) to understand the security aspects of federated learning as well as STRAIT Consortium (D462) and AI Maker (D727) who will provide expertise and knowhows on model exchange protocols and data/model brokerage (matchmaking), respectively.

Infrastructure Saftey Monitoring

An AI-driven Platform for Diagnosing Infrastructure Health

Lead PI: Claudia Marin
cmarin@Howard.edu

Anuj Karpatne
karpatne@vt.edu

Jale Tezcan
jale@siu.edu

Overview

After years of underinvesting in the Nation's infrastructure, federal and state departments of transportation, the U.S. Army Corps of Engineers, electric utility companies, and other decision-makers need accurate and timely information about an infrastructure's health to prioritize investment decisions. Unfortunately, there are currently no broadly applicable automated tools to assess the degradation of infrastructure health. Traditional monitoring methods, such as visual inspections, are subjective; thus, they are not reliable for critical resource allocation decisions. To improve the accuracy of infrastructure health, our team is developing an Artificial Intelligence (AI)-driven Infrastructure Safety Monitoring (ISM) platform to inform maintenance, repair, and replacement decisions. The ISM project will lead to widespread monitoring of real-world structures ranging in complexity from simple vertical steel utility poles and agricultural structures to more complex structures such as buildings and bridges.

Description

Infrastructure safety and resilience are critical to national security and socio-economic well-being. Deteriorating and deficient infrastructure has a cascading effect on the quality of life—it costs lives, jobs, disposable income and makes our national defense system vulnerable. According to the American Society of Civil Engineers, if current under-investment levels persist, the cumulative cost to the U.S. economy will reach 23.3 trillion dollars. By 2039, it is estimated that over 3 million U.S. jobs will be lost, and every household will lose 4.2% of its annual income. With budget shortfalls, it is critical to optimize the allocation of funds available for infrastructure maintenance and replacement, and address the absence of tools to support prioritization decisions.

The ISM platform aims to remedy the current lack of reliable structural health monitoring methods and tools by leveraging deep learning to extract essential information about structural health from video recordings of structures. Our preliminary investigations have shown that the ISM technology can successfully detect and locate damage, and quantify damage severity by tracking structural vibrations. Distinguishing features of the ISM platform are: (1) the use of video cameras as non-contact sensors; and (2) incorporation of physical laws and constraints into the machine learning algorithms (Physics Guided Machine Learning) to reduce the need for extensive data for training of models and to ensure the compatibility of the predictions with physical principles.

We are implementing the ISM platform on benchmark structures, including a pedestrian bridge, a water-flow control structure, vertical steel poles supporting 5G antennas, agricultural structures, a school building, and high-voltage equipment from electric transmission substations. We are field monitoring and creating validated models to demonstrate the platform's applicability to real-world infrastructure. Throughout phase 2, we will work with industry partners to explore technology transfer opportunities. Existing industry partners have expressed interest in adopting the ISM approach for phase 2 and beyond. The ISM team has social scientists and insurance industry supporting outreach to decision-makers and defining a strategy to address regulatory and societal challenges to the widespread adoption of the ISM platform.

Differentiators

The key aspects differentiating ISM from traditional approaches in infrastructure health monitoring are the method of data collection and the novelty of the AI-driven damage-detection algorithms.

Traditional monitoring approaches rely on a network of wired or wireless sensors for data collection. The high cost of installation, protection and maintenance of sensors limits the number of sensors, which in turn limits the damage detection capability.
The ISM approach extracts structural vibration data

from video recordings. This allows for tracking the movement of any point within the camera's field of view, rather than being limited to a set of discrete points defined by the sensor locations. Further, video tracking of structures in hazardous environments is easier, safer, more economical, and provides more accurate information than its sensor-based counterparts.

Existing damage detection algorithms are typically derived using simple laboratory models or numerical simulations. Their applicability rarely translates to real, in-service, structures. The machine learning models used in the ISM platform are validated and calibrated using field data from real structures.

The ISM team is highly multidisciplinary, with expertise that spans structural engineering, computer vision, AI, cybersecurity, and social science, from academia, government and industry. The team is leveraging high-quality field data from various sites to create, evaluate, and validate each component of the ISM platform and developing strategies to facilitate its widespread implementation, considering various scientific, regulatory, and societal issues.

Road Map

As part of our phase 1 program, we have identified six types of benchmark structures for full implementation of the ISM platform. Accordingly, the team has started collecting data on a dual purpose bridge/river control structure, a chimney tower, and utility poles. We developed the proof-of-concept of the video and machine learning modules, and in the process, refined our research goals and the implementation plan for phase 2.

Our plan for phase 2 involves the following tasks:

- Instrumentation of the benchmark structures for data collection **(Year 1)**.
- Finalization of the video tracking module for extracting structural displacements (**Year 1:** Initial module. **Year 2:** Refined module).
- Development, validation, and field-calibration of the machine learning models for the benchmark structures. (**Year 1:** Model development. **Year 2:** Model validation and calibration).
- Develop an optimal cybersecurity plan for the platform. (**Year 1:** Vulnerability assessment. **Year 2:** Development of a cybersecurity strategy.)
- Integration of video and machine learning modules developed for each benchmark structure into the ISM platform (**Year 2**).

Partnerships

Phase 1 Partners (all will continue in phase 2):

- U.S. Army Engineer Research and Development Center (ERDC): Industry perspective, data sharing, and providing access to structures for monitoring.
- District Department of Transportation (DDOT): data sharing and providing access to structures to monitor.
- Microsoft: Cloud computing, and advice in computer vision and machine learning.
- Natural Hazards Engineering Research Infrastructure (NHERI): DesignSafe and SimCenter data sharing platforms, data dissemination, and computational resources.

Additional phase 2 Partners:

- National Institute of Standards and Technology (NIST): Technology, data sharing, providing a testbed to apply the ISM platform.
- Additional Support: Supporting data sharing, connections with decision-makers, disseminating results. HMV Engineers (electrical infrastructure). Walter P Moore. Insurance Institute for Business & Home Safety. Berkshire Hathaway Specialty Insurance, Structural Integrity Associates, Smart Structures Technology Lab. Wall of Wind Florida International University, Natural Hazards Lab at the University of Florida, Michael Baker International, and Development Corporation of Columbia Heights.

Intellectual Property

Intellectual property (IP), including patents, copyrights, licensing, and know-how, is managed through the Howard University's IP office. The IP plan described in the proposal will be monitored and updated periodically.

InstaTwin

Digitizing the Built Environment with the Power of AI

INSTATWIN

Lead PI: Yelda Turkan
yelda.turkan@oregonstate.edu

Roger B. Chen

Fuxin Li

Yong K. Cho

Michael J. Olsen

Overview

Architects, engineers, contractors, and owners need reliable 3D digital models to plan, design, construct, and manage the Built Environment. Sensor technologies such as laser scanners (i.e., lidar) and 360° cameras can collect rich 3D data, but it is tedious and expensive to transform data into useful 3D digital information models. Our AI-based *InstaTwin* technology automatically segments, classifies, and extracts real-world features from 3D data to create digital representations of reality. The Artificial Intelligence (AI) components of *InstaTwin* have broad applicability and far-reaching impact on equity, energy, and sustainability.

Description

Buildings account for 40% of the world's energy consumption, primarily because older buildings are inefficient. To achieve global sustainability, we must transition old building stock to energy-efficient buildings. Digital representations of buildings can help us to assess, analyze, monitor, and renovate buildings and to alter their behavior in real time. Deep Reality's *InstaTwin* creates 3D models of the Built Environment. It converts 3D data collected from sensors such as lidar and 360° cameras and converts these data into digital representations of 3D objects. The resulting 3D model is called a "digital twin" or building information model (BIM). To understand the digitization problem, consider the analogous process of flatbed scanning, where a scanner captures a pixelized image of a physical document. The captured image provides a digital record (a 2D collection of pixels), but the true value of the scan resides in the words within the image. It is time-consuming to transcribe these words manually, but technologies such as Optical Character Recognition (OCR) significantly increased the accessibility to and utility of the information in the scanned document.

Scanning a building is analogous to scanning a document. Instead of a flatbed, a 3D laser scanner spins on a tripod and measures the contents of a building. The captured image provides a digital record (a 3D collection of points called a "point cloud), but the true value of the scan resides in the objects and building elements within the point cloud. It is even more time consuming to transcribe these objects than to transcribe words, and as of today, there exists no equivalent to OCR for creating BIMs.

Differentiators

We introduce *InstaTwin*, a platform that automatically generates 3D BIMs from point clouds. *InstaTwin* harnesses the power of AI to recognize objects within the 3D point cloud, just as OCR recognizes words within the 2D collection of pixels. *InstaTwin* empowers asset owners, architects, engineers, and facility managers to make better, faster, and more efficient decisions about the Built Environment.

InstaTwin is a data repository and processing framework that lidar and BIM users can access online. Collaborators and users submit data, test algorithms, process data into 3D models, view results, and download BIMs. Since our framework can benefit the industry immediately, it incentivizes users to supply data, train models, and support its development. Our diverse and convergent team spans academia, business, and industry. Several team members have worked in this space for 20 or more years and have spent their professional careers bringing this technology to where it is today. Industry leaders are helping us to develop and commercialize *InstaTwin* because they recognize its significant value to society.

Road Map

We have developed a two-year plan to create deep learning algorithms on a cloud-based platform and to sustain the platform after phase 2. In **Year 1**, we will develop and deliver novel deep neural network models, as well as a robust, fully-featured platform to store and distribute community-sourced, annotated 3D BIMs and point clouds of the Built Environment. We will establish an advisory board, develop a strategic plan, and obtain commitments from architectural, engineering, and construction (AEC) industry and standards groups. In **Year 2**, we will implement and refine the AI algorithms and the *InstaTwin* platform based on feedback from human-centered design workshops. We will then publicly launch the platform. We will also develop a business plan to secure funding to sustain the project and platform. We anticipate diverse and significant impacts because *InstaTwin* will address a large and growing market for building lifecycle management tools. This market desperately needs efficient digital solutions. Finally, *InstaTwin* will catalyze energy efficiency by encouraging the AEC industry to adopt smart buildings and efficient construction and maintenance practices.

Partnerships

We have assembled a team of academic researchers (Oregon State University, Georgia Institute of Technology, the University of Hawaii at Manoa, and Stony Brook University), industry (Allvision, Inc. and MPN Components), government (Oregon Department of Administrative Services), and non-profit/foundation (Castle, Alpha Tau Foundation, Inc., Design Building Lab) partners. Our stakeholder advisors include leaders from surveying and engineering firms (Sam O. Hirota, Inc., Lanmar Services, Feldman Surveyors, Maser Consulting, Direct Dimensions, Inc., Hypar, Inc., Plant Construction Company, Zachry Corporation, Skanska, and Precision Point) who collect 3D point cloud data and process them into 2D CAD drawings and 3D BIMs. They will beta test *InstaTwin* with data from their projects and provide feedback. The Open Geospatial Consortium will provide feedback on the data formats and schemas for the *InstaTwin* platform.

Intellectual Property

The IP will include the cloud-based *InstaTwin* platform and algorithms to extract detailed representations of physical reality from sensor data. Copyrights, patents, and trademarks will protect the platform and its supporting segmentation and S2BNet algorithms that synthesize point cloud data for incomplete point clouds, recognize and model scenes, and correct topological errors. We will distribute select deep learning algorithms of broad scientific value as open source material.

Visit us at research.engr.oregonstate.edu/deep-reality

LEARNER

A New Health AI Model Sharing and Learning Platform

Lead PI: Hai "Helen" Li
hai.li@duke.edu

Overview

A whole new world of health care insight is at our fingertips, thanks to quickly evolving machine learning (ML) techniques that can analyze vast amounts of data. But how can we effectively leverage these valuable data while ensuring that patient personal information remains private? We are developing a new privacy-preserving artificial intelligence (AI) model sharing and learning platform for health data, namely LEARNER, to enable collaborative big health data mining among doctors, biostatistics, and neurobiologists, etc. by integrating federated learning algorithms, trustworthy AI techniques, large-scale distributed optimization methods, and effective software tools.

Description

Recent advances in ML and AI have significantly impacted many data science applications over the past decade and led to breakthrough scientific advances across multiple disciplines. With the burgeoning adoption of ML and AI techniques and the rapid availability of massive health data, the healthcare sector has become one of the data science frontiers. However, the critical challenges to fully exploiting such big data lie in the complexity of ML model design and optimization and the data privacy concern as health data is intrinsically sensitive and identifiable.

To address these challenges, we will develop a novel trusted AI platform, namely LEARNER, to create model sharing for health science applications. It is a cloud-based platform that provides a suite of services in support of state-of-the-art privacy-preserving ML models and computational algorithms. Meanwhile, to pretrain ML models with optimal parameters for different applications and facilitate data sharing, LEARNER will also provide the health data repository to collect/share data and metadata and address FAIR (Findable, Accessible, Interoperable, Reusable) data principles.

LEARNER will be the first infrastructure that includes a suite of collaborative data analysis and privacy-preserving mechanisms in support of various types of health data analytics. Our project will fundamentally advance AI-driven health innovations and accelerate use-inspired convergence research in health data science through intensive infrastructure development and framework deployment. The proposed federated learning and trustworthy AI techniques are highly efficient and scalable and meet the critical needs of big data analysis and secure data mining. Such unique capabilities will enable new computational applications in many research areas. It advances and extends the relationship between engineering innovation and computational analysis.

In the long term, we believe the platform potentially has a large number of users in the scientific, policymaking, and corporate worlds. We collaborate with large medical centers, healthcare providers, and industrial partners to ensure the project has users and funding beyond this program.

Differentiators

We have assembled an experienced cross-disciplinary team and have brought together the required disciplines to carry out this project:

- ML: *Huang, Henao, Li, Lu, Zhan*
- Trustworthy AI: *Huang, Li*
- Biomedical data science: *Chen, Ding, Henao, Huang, D. Page, G. Page, Zhan*
- Computing systems and infrastructure: *Li*
- Research provenance and data infrastructure: *Infinia ML*

Both Duke and Pitt teams have been working on federated learning techniques for years and demonstrated their use in various application domains, including medical data analysis. For phase 2, RTI, UMN, and Infinia ML will join us, contributing to data harmonization, model interpretability, and software development, respectively. The integration of expertise enables computational thinking to address the critical problems in ML algorithms, theoretical foundations, and its application to biomedical data science. Our team also draws expertise from other projects in the Convergence Accelerator, thus capitalizing on existing talent to ensure overall track success.

Road Map

The main deliverable of this project is LEARNER, a cloud-based platform featuring (1) an easy-to-use web portal, (2) a suite of state-of-the-art privacy-preserving ML algorithms for health data analysis, (3) a scalable computation and storage physical platform, (4) the implementation of data collection, management, and analysis modules.

Year-1 milestones (M1-M4): The team will complete the asynchronous federated learning algorithms development for model sharing. The team starts working on the biomedical imaging data harmonization and phenotype standardization, and these data will be used to train the FL algorithms.

Year-1 milestones (M5-M8): The trustworthy AI techniques will be integrated with the developed federated learning algorithms. The genomic data harmonization strategies will be developed.

Year-1 milestones (M9-M12): A cloud computing environment (Amazon Web Service) will be set up and tested. The LEARNER platform starts to run for testing with the partners, collaborators, and users.

Year-2 milestones (M1-M4): We will complete the data harmonization strategies for electronic health record data. The privacy-preserving data sharing and publishing repository will be created. The hypothesis-sharing web portal will be built.

Year-2 milestones (M5-M8): The fair and interpretable health AI models will be pretrained and shared. The feedback from industry partners and pilot users will be used to improve our platform.

Year-2 milestones (M9-M12): The team will keep collecting feedback to improve the system and work with industry partners to achieve the model and data sharing research convergence.

Partnerships

ML/AI/statistics, biostatistics/data science experts from Duke University, University of Pittsburgh, University of Minnesota, and RTI International. form the **core technology development team,** and Infinia ML provides **software development** service.

We partner with **20+ medical schools/centers,** who will serve as pilot users for the LEARNER platform, contribute their domain expertise, and provide feedback.

In addition, our **industrial partners** will provide assists in the following aspects:

- Amazon: Model sharing and learning methods and cloud computing service
- Amazon & Accenture: Cybersecurity and trust-worthy AI techniques
- Microsoft: software development and evaluation
- Intuitive Surgical: health applications

External **Advisory Board** consists of members from all national stakeholders, including academic experts, medical centers, healthcare industry, AI companies, and non-profit institutes. All the team members will routinely communicate with the advisory board to prioritize and refine the research activities to best accommodate users' needs.

Intellectual Property

Inventions disclosures are being filed with Duke Ventures. This will ensure proper documentation of the existing IP and resulting IP from this project.

MetaMatchMaker

Discovery and Harmonization Through Transfer Learning

Lead PI: Grier P. Page
gpage@rti.org

Xiangqin Cui
xiangqin.cui@emory.edu

Ricardo Henao
ricardo.henao@infiniaml.com

Hai Li
Hai.li@duke.edu

Beatriz Martinez Lopez
beamartinezlopez@ucdavis.edu

Overview

Despite a massive global effort, we still cannot predict or effectively treat the adverse health events from Covid-19 infection. Mountains of data are being collected but they are heterogenous, in different formats, and are siloed across many databases. Bringing these data together is time consuming, expensive, labor intensive, and it slows down the pace of discovery. This issue is not unique to Covid-19 research. Between 35% and 80% of all data science budgets are spent on data integration before any analysis can be done. We are developing machine learning tools that are trained on thousands of datasets and millions of metadata terms to enable researchers and data scientists to locate and harmonize data quickly so they can do what they do best – gain insight and make discoveries. The approaches and tools we present here are appropriate for any field of study, including health, economics, environment, and business.

Description

"Everyone wants to do the model work, not the data work"

We present MetaMatchMaker (M3), an easy to use software suite based on a machine learning neural network model which makes finding, accessing, and integrating datasets easier, cheaper, and faster. Using a key AI innovation - transfer learning – we can adapt the neural net to address new scientific domains quickly. Think of this like Google translate but for scientific data.

The adoption of the M3 tool suite will significantly reduce the time and cost of preparing data. This means that studies will happen faster, data can be shared more easily, and the greatest barrier to entry for understaffed projects is removed. Ultimately this means faster insights, leading to improved health and quality of life.

Use Case - The Environmental Children's health Outcome (ECHO) study aims to find the causes of adverse health outcomes in children by combining studies from over 85 cohorts comprised of tens of thousands of mothers and newborns. It has taken 4.5 years and over $100 million to integrate data that is being used in hundreds of studies a year. Phase 1 prototype testing of M3 showed a 300% increase in speed of data integration. If ECHO had used M3 up front, analysis could have occurred 2 years earlier, and critical insights into diseases such as treatments for neonatal opioid withdrawal syndrome could have been discovered sooner.

Differentiators

Our team is deeply experienced with constructing common data elements (PhenX, Connects, NCATS Biomedical Translator), leading multi-center data coordinating centers (ECHO, NRN, PFDN), and serving as data stewards (NIDKK repository, LungMap, and BioDataCatalyst). We have partnered with world-class experts in machine learning to develop a novel approach to assemble datasets using a deeply trained transfer learning toolset.

In phase 1 we deployed two prototypes using this framework. First, a data discovery tool (metamatchmaker.com) lets researchers find existing public data that can be integrated into their own studies. Second, a metadata linker

tool lets researchers match datasets in hand. We are not aware of any other commercial product which accomplishes these tasks as accurately nor as efficiently as M3.
In phase 2 expertise from other Track D teams in the Convergence Accelerator will expand our reach into new scientific domains as well as secure access of data. Thus, we are capitalizing on existing talent to ensure overall convergent track success and expand the utility of M3.

Road Map

In phase 2 we will continue to train the core neural network, further develop the two phase 1 prototypes, and create and refine 3 additional tools.

- **Expand the neural network** to include 20+ new data repositories. (Q1, 2022)
- Develop our **metadata linker** into a user-friendly commercial product. (Q3, 2022)
- Train the neural network on **Electronic Medical Record** data and commercialize an EMR data extraction and formatter tool. (Q1, 2023)
- Train the network on data from experimental model organisms, animal and veterinary health to support **One Health research** in conjunction with Team D680 (Q2, 2023)
- **Facilitate access to protected data** in two ways. First, implement secure data passports (e.g., GA4GH AAI). Second, integrate methods in secure federated learning of AI models developed by team D588. (Q3, 2023).

Partnerships

Phase 1 Partners (continuing in phase 2)
InfiniaML provides world class machine learning expertise and engineering support. They will work with data harmonization experts at RTI International to develop and evaluate performance of the machine learning model.

Foundation for Atlanta Veterans Education and Research (FAVER) will continue to provide expertise in EMR data extraction, connect with researchers struggling with using EMR in their studies, and advise on ways to engage with the EMR market.

Additional phase 2 Partners
UC Davis: A new partner from the Convergence Accelerator Track D (D680), Dr. Martinez-Lopez's team will lend their expertise in animal and veterinary health to expand the scope of M3 into model organism and veterinary data, supporting biosurveillance research.

Duke/U. of Pittsburgh: A new partner from the Convergence Accelerator Track D (D588), Drs. Li and Huang will provide access to EMR data from Duke and UPMC, as well as provide their federated learning framework for AI model training on protected data.

Collaborators & Users
Agreements are in place with multiple data repositories, data coordinating centers, and institutions to evaluate the utility of M3 with real world data integration challenges: National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK repository), PhenX toolkit of common data elements, the NIH network of covid-19 therapeutic strategies (CONNECTS), a precision medicine genomics resource BioData Catalyst, and the VA phenome library (CIPHER).

Intellectual Property

MetaMatchMaker is the intellectual property of RTI International with minority ownership rights to Emory and Columbia Universities. Plans for copyrighting, branding, patens, and/or licensing agreements are being finalized.

Model Exchange

A Standardized Model Description Format (MDF) for Comutational Modeling

Lead PI: Jonathan D. Cohen
jdc@princeton.edu

Abhishek Bhattachargee
abhishek.bhattacharjee@yale.edu

Sharon Crook
sharon.crook@asu.edu

Padraig Gleeson
p.gleeson@ucl.ac.uk

Pranav Gokhale
pranav@super.tech

Terry Stewart
terry.stewart@gmail.com

Ted Willkesharon
ted.willke@intel.com

Tal Yarkoni
tyarkoni@gmail.com

Overview

The Model Exchange and Convergence Initiative (ModECI) is focused on developing a standard Model Description Format (MDF) that will support the exchange of computational models across diverse software platforms and domains of scientific research and technology development. This will leverage the explosive growth of computational modeling efforts in a wide range of scientific disciplines and spheres of technology development. By allowing models and execution/analysis tools developed in one domain to be ported transparently to others, ModECI will promote interdisciplinary exchange of theoretical and algorithmic developments, advances in efficiency of computation and, perhaps most importantly, facilitate the construction of more sophisticated models that integrate components developed in different environments or disciplines.

Description

As science takes on increasingly complex phenomena, such as climate change, the biome, and the human brain, it is relying increasingly on computational modeling for analysis and discovery. Similarly, as engineering has sought to build more sophisticated systems and machines, from structural monitoring and efficiency regulation to AI-driven medical image analysis and autonomous vehicles, computational models and algorithms have become an essential element.

These developments in both science and technology have produced a proliferation of software tools to assist in the development, execution and analysis of computational models that, in turn, has led to a "Tower of Babel" problem, producing barriers to the exchange of theoretical and algorithmic advances across disciplines, making it harder to replicate and validate existing results, impeding the creation of more sophisticated models that integrate components developed by researchers across disciplines, and introducing barriers to portability of models across software platforms and development environments. The impact this is having on the economy is inestimable. For example, the market for AI alone is exploding, from $20B in 2019 to estimates in excess of $100B over the next five years. What is not reported is how much of that will be wasted by missed opportunities or "reinventing the wheel" in different domains, but this is likely to be astronomical. The goal of ModECI is to overcome this balkanization of computational research and technology by developing an MDF that will permit transparent exchange of models. This can be thought of as the equivalent, for computational models, of the ubiquitous PDF format for documents, and its success will herald a new era of opportunity for coordination, integration and convergence in computational modeling efforts.

The MDF consists of two core components: a) a fully specified standard for the description of model structure and execution in the form of a computational graph, that can be serialized in text and binary formats; and b) a Python-based library of tools for importing, exporting, inspecting, and validating models in the format. These will be maintained in a publicly accessible GitHub repository, building on our prototype implementation, that will be used to review and accept appropriate community contributions to its maintenance and continued development.

Differentiators

To our knowledge, the ModECI effort is unique in scope. There are examples of standardization efforts within some disciplines, such as NeuroML in neuroscience and ONNX (Open Neural Network Exchange) in machine learning. However, these are tailored to their domains and lack critical functionality that will allow models to be exchanged across those domains and to others (for example, exchangeable expressions of process control). WebGME (Web-based Generic Modeling Environment) is a standard and associated set of tools for designing new modeling languages and applications, but does not unify the

expression of the models themselves, and thus requires either containerized execution or manual translation into a supported format. MDF is being designed in collaboration with the developers of these formats, as well as developers of widely used computing environments (such as PyTorch), to ensure interoperability and exchange. This will not only yield a more general exchange format, but will also provide access to the large ecosystem of tools and environments that each partner platform individually supports.

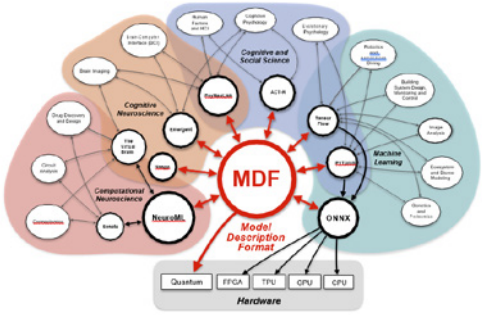
Critically, the MDF is a "horizontal" and not a "vertical" development effort. It is not meant as a new programming language or development environment for any particular domain or application. Rather, it will complement such efforts by facilitating exchange among existing (and any newly developed) environments, and provide a common target for efforts focused on increasing efficiency of model implementation.

It will also be used to establish a novel and unique pathway for mapping computational models to quantum platforms, through a partnership with Super.tech (a leading developer of quantum software) and collaborations established within Track C of the Convergence Accelerator Program. This promises to open up entirely new domains of application for quantum computing, while providing a path for optimization of AI and other computational applications requiring real-time constraint satisfaction (e.g., in decision making and problem solving), that have not been well served by the advances in deep learning and optimization on conventional hardware (e.g., CPUs and GPUs), but are well suited to implementation on quantum hardware. Thus, the MDF produced by ModECI will occupy a unique and valuable niche, simultaneously enhancing interoperability with and complementing existing standards.

Road Map

Our initial goal for phase 2 will be to fully develop interfaces to strategically selected "anchor" software platforms that are widely used in each target community, as the initial "spokes" of our hub-and-spoke approach.

These will help establish feasibility, and a critical mass of users to engage widespread use and



community support. Once that is accomplished, we will then focus on broadening our reach by supporting secondary software platforms within our target communities, and primary ones in a broader range of relevant scientific communities, and formalizing mechanisms for sustainability (e.g, gaining certification/affiliated status with organizations that promote and support open source development). Accordingly, **Year 1** will focus on finalizing deployment interfaces to the anchor environments begun during phase 1 (neuroscience: NeuroML; cognitive science: ACT-R; machine learning: PyTorch, ONNX; hardware: Super.tech quantum platform). In **Year 2** we will engage with secondary platforms (e.g., The Virtual Brain, SONATA, Nengo) and organizations that will provide mechanisms for further outreach and sustainability (e.g., COS, NumFocus, INCF).

Partnerships

Our partnerships fall broadly into two categories: a) developers of existing software platforms and/or disciplinary exchange formats with whom we have collaborated, and will continue to collaborate in the design and implementation of the the MDF; b) non- profit organizations dedicated to the evaluation, certification and dissemination of standards and platforms for open science. These include:

Modeling Environments: ACT-R, Emergent, PyTorch, ONNX, NeuroML, Nengo, PsyNeuLink, Allen Institute / SONATA, Super.tech, The Virtual Brain (TVB), and WebGME (General Modeling Environment)

Open Science: Brain Imaging Data Structure (BIDS), Center for Open Science (COS), International Neuroinformatics Coordinating Facility (INCF)

Intellectual Property

MDF will be developed as open source software under the oversight of ModECI using an Apache License.

Pisces ClimatePro

Quantifying America's Water & Climate Risks Using AI and Machine Learning

Lead PI: Upmanu Lall
ula2@columbia.edu

Casey Brown
casey@engin.umass.edu

Kenneth Kunkel
kekunkel@ncsu.edu

Scott Steinschneider
ss3378@cornell.edu

Naresh Devineni
ndevineni@ccny.cuny.edu

Overview

America's most critical resource, our water supply, is ill-prepared for climate and societal change, putting people and businesses at risk of costly disruptions. Water managers and corporate risk managers urgently need tools to understand and manage these risks. Pisces ClimatePro enables any utility or business to identify their climate risk and improve their resilience. The cloud-based, AI/ML analysis capability increases the preparedness of water systems and businesses for future shocks, reducing their risk of disruption in both the immediate and long-term future. Eager partners in corporate water risk management and financial risk hedging extend the accessible market.

Description

The nation's water is supplied by 50,000 independent water utilities, each unique yet all facing major risks—from climate variability and change and disruptions from catastrophic events. Cloud-computing, remote sensing capabilities, and advanced climate models enable incredible forecasting capability. Yet, most utilities across the US serve smaller communities and lack the modeling and analysis capabilities to utilize this technology. Even the largest water utilities lack the technical capacity to fully leverage these capabilities that could be their best defense against a variable and changing future. phase 1 target market discovery shows strong value for Pisces across sectors including: 1) small, medium and large utilities for reducing the disruptions from climate variability and better insights into how climate change could impact their operations and planning; 2) corporate risk managers expressed value in understanding, managing and disclosing climate risks; and 3) insurance providers see value in reaching new markets with products linked to water supply risk. The Resilience Navigator matches cloud-based AI/ML analytical capability and the proliferation of large data centers to these climate change problems and opportunities. The platform provides temporally consistent AI-based hydroclimate forecasts at any US location for the next 3-9 months with scenarios up to the next 80 years, providing utilities the planning and operational foresight they currently lack. In addition, AI-based water demand scenarios for the next 3 months to 80 years provide agile on-demand planning capabilities previously only available via costly specialty consultants. In response to demand from utilities in less drought-prone areas, the platform capabilities will also include water quality forecasts and flood forecasting capabilities, which are priorities in humid regions. Finally, transfer learning capability enables data pooling across sites to enhance model skill while maintaining confidentiality of individual utility data. Together these capabilities provide unprecedented foresight for water utilities to reduce climate disruptions and others to manage water related risks. Given estimated adoption rates, the Resilience Navigator will reduce the impact of climate change on 1,000-5,000 utilities and 100-300 million people, in addition to reduced business losses.

Differentiators

The Pisces ClimatePro is the first general

platform that provides water supply forecasts for any utility, at both short-term and long-range timescales, and integrates them with utility specific analysis capability. The platform will outperform non-scalable bespoke products developed by local engineering consultants, increasing capabilities and reducing costs by 10x. For the first time, water supply and water demand predictions will leverage pooled water utility data for unprecedented forecasting skill. Existing climate service providers focus on projects and products based on government climate models that do not include local system data, focus on distant future climates and do not target water utilities, creating an untapped market segment. In addition, because safe and reliable water supply is an input to many businesses while also an essential element of municipal population growth, water utility service is a risk for private sector interests as well. Our growing unique database of water utility risk data will underpin products and services for corporate risk managers, insurance and financial risk hedging products, and real estate investors, further providing each sector original information and business opportunities.

Road Map

2021 Q4: Onboard Development Team

2022 Q1: Develop Branding. **Q2:** Prototype 1 Complete. **Q3:** Customer Validation. **Q4:** Prototype 2 Complete.

2023 Q1: Finalize Business Model, Finalize Marketing Strategy. **Q2:** Prototype 3 Complete, Customer Validation. **Q3:** Market Ready Proof of Concept Launch. Pisces Incorporates. **Q4:** Series A funding round.

Partnerships

Over 50 utilities, insurers, and investors have recognized the unique value being offered by the Pisces Resilience Navigator and have united to support its launch.

Phase 1 established customer segments and value propositions for each segment through interviews, a subset of which are continuing as phase 2 co-development partners or product testers. We have identified the water utility sector as the beachhead customer and within that sector three segments: Large and Medium public utilities and regulated private water utilities that provide access to many small systems via a single partner. **Water Utility co-developers:** Association of Metro Water Agencies, SFPUC, Santa Cruz, Central States Water Resources, American Water, EPA (Michael Deane), MA and CT State Revolving Funds. **Insurance co-developers:** Global Parametrics, Swiss Re, Sciens Water. **Investment co-developers:** Adapting Markets, LLC, Skopos Labs. **Technology co-developers:** Track D Environmental Modeling Working Group.

Intellectual Property

Expected to register as a Digital Public Good with Open Source License, e.g., AGPL. Primary IP components:

- AI-based seasonal to decadal climate prediction at any location
- AI-based hydrologic prediction at any location
- AI-based water quality prediction at any location
- AI-based water system modeling at any location

Precision Epidemiology

Data-driven Disease Prevention and Control in Animal Health

pEPIC THE PRECISION EPIDEMIOLOGY CONSORTIUM

Lead PI: Beatriz Martínez López
beamartinezlopez@ucdavis.edu

Maria Jose Clavijo
mclavijo@iastate.edu

Xin Liu
xinliu@ucdavis.edu

Scott Stoller
stoller@cs.stonybrook.edu

Kun Zhang
kunz1@cmu.edu

Overview

Sustainable livestock production systems are needed to feed a growing world while protecting the planet. Livestock farming systems, in particular pork production, play a significant role in addressing this global challenge. However, sustainability of the livestock industry hinges on the maintenance of elevated animal health and wellbeing, and high production efficiency. Therefore, there is a need for proactive and refined best management practices that enhance animal health at the farm level. Our precision epidemiology consortium converges data, AI models and expertise across the livestock production and health space, providing an online user-friendly platform, called Disease BioPortal (https://bioportal.ucdavis.edu), for prediction and effective management of animal health problems.

Description

The US livestock industry has an enormous socio-economic impact, producing $195 billions in sales annually and generating >5.5 million jobs. In this project we focus on the swine industry, which is currently facing important global challenges with both endemic and emerging issues (e.g., swine influenza, antimicrobial resistance, African swine fever pandemic). However, our models and outcomes have been designed to be adapted to other disease and species by minor modifications in the data structure to strengthen the resilience of the livestock industry.

In swine production, disease outbreaks represent a significant economic loss for the industry, reducing productivity, compromising animal welfare and leading to an overdependence on antimicrobial drugs. Furthermore, the introduction of foreign animal diseases, such as African swine fever, could have cost $50 billion in losses and would worsen the current protein gap. Therefore, early detection and mitigation of animal health issues become crucial not only to maximize production efficiency and food safety and security but also to mitigate catastrophic health and economic consequences.

Phase 1 interviews of animal health data users and experts revealed three main barriers to improve animal health: 1) lack of data availability, integration, sharing, and use 2) poor data efficiency, data governance and privacy, and 3) absence of explainable prediction models and accessible user-friendly visualization tools. Our Disease BioPortal platform integrates multilevel animal health data with advanced prediction models, providing end-users (including farmers, veterinarians, pharmaceutical companies, researchers and policy makers) an easy and secure access to data and models through a simple user interface to support animal health decisions.

Differentiators

Current practices in the swine industry are usually reactive as most of the testing/interventions take place after having an outbreak or observing unusual clinical symptoms in animals. Ideally, we shall have a predictive early warning system that enables prevention, earlier detection and faster control of problems both at animal and farm level. Our

pEPIC approach and Disease BioPortal platform offers four unique value propositions: (1) a user-centered design informed by interviews with more than 40 organizations from pig companies to veterinary clinics and diagnostic labs, (2) data standardization, integration, secure sharing and communication capabilities, (3) innovative AI-based prediction models, cutting edge visualization tools, and domain-specific applications of data and models, and (4) a convergent team and academic-industry partnerships that brings together leading organizations, data providers, end-users and experts in computer and data science, visualization and computer-human interaction, causal discovery, software engineering, bioinformatics, epidemiology and animal health, diagnostics and management.

Road Map

Year 1:

- Continue with data collection, standardization, integration and curation to generate AI-ready datasets and facilitate data usage/sharing.
- Develop AI models (including explainable machine learning models, causal discovery and reasoning and topic models) and integrate data and models into user-friendly dashboards within the Disease BioPortal to facilitate their use/visualization by stakeholders.
- Expand our working groups with selected industry partners and end users to incorporate new data sources, increase our user cases, receive feedback and enhance user experience during the beta testing of Disease BioPortal new capabilities.
- Test the mobile version of Disease BioPortal, a key differentiator requested by many end users to be able to use it “on-the-go”.

Year 2

- Develop a federated learning approach to collaboratively share models without sharing the data.
- Publish AI models on GitHub
- Release of the new version of Disease BioPortal (new data, AI and visualization capabilities available for all end users).
- Organize Data challenges using large datasets to solve real-world problems in animal health.
- Develop and implement new services as well as the outreach, training and certification programs.

Partnerships

We have assembled a convergent research-industry team that will be able to gather all crucial data, develop the necessary methods and test them in real-farm applications. Through existing close collaborations and phase 1 expansions, our team consists of the top veterinarian schools in the US, the largest swine veterinary diagnostic laboratory (VDL at Iowa State University conducts 50% of all diagnostics in the US), the largest animal genetics company (PIC), the largest veterinary clinic in the US (Pipestone), and several of the top 10 largest swine producers (Seaboard, Pipestone, Hanor, Iowa Select, Tosh farms) and the National Pork Board. We have also extended our partnership with other companies such as GlobalVetLINK, the US leader in digital animal records and data aggregation; pharmaceutical companies such as Merck & Co. (MSD), which is the world’s seventh largest pharmaceutical company by market capitalization and revenue and; AWS research team to improve our data architecture. They all are key data providers and end-users.

Intellectual Property

Our Disease BioPortal code is already protected in the form of copyright, and we plan to seek other other types of IP protection and licensing in the future such as trademarks or patent applications for AI algorithms, processes, visualizations and the mobile App. We envision long term sustainability of our program using a subscription-base model that has already been approved by our University. Our tiered pricing model (by month or year) are based on the type of customer/end-user segment, the volume of data and the tools needed. For academia, research and students access to basic data and functionality is free of charge. Industry, enterprise or government organizations can opt for individual or pro tiers depending on the number of users per organization and the volume of data/tools needed.

STRAIT Consortium

Scalable, Traceable AI for Imaging Translation

Lead PI: Bennett A. Landman
Bennett.Landman@Vanderbilt.edu

Cheryl Carey
ccarey@SIIM.org

Yuankai Huo
Yuankai.Huo@Vanderbilt.edu

Steve Langer
Steve.Langer@FlowSigma.com

Ipek Oguz
Ipek.Oguz@Vanderbilt.edu

Overview

Reproducible science requires reproducible review. AI is poised to transform the medical imaging process from diagnosis through intervention, but these technologies are not reaching patients due to complex, non-scalable, and non-verifiable validation. STRAIT will streamline validation from inception to surveillance. This will allow for rapid development, assessment, and dissemination of AI models in medical imaging to provide innovators a clear path for commercialization

Description

Consider the current plight of an innovator with a groundbreaking diagnostic imaging AI model for detecting cardiovascular stenosis assessment on low dose computed tomography (CT). Current validation approaches would involve reimplementation, multi-site trials, and user-workflow studies. While the model may be highly innovative, stenosis is only one of 50+ potential findings on a typical chest CT. Clinical value is realized only when AI is integrated with existing informatics and clinical AI ecosystems, yet, clinical integration is costly. So, the economics for each algorithm quickly become untenable for most AI models, and their potential value is lost. We envisioned a new and different ecosystem, called STRAIT, where AI validation is performed conterminously with development, AI learn from each other in a trusted manner, and a model commons provides a systematic framework for AI discovery, comparison, and interaction. STRAIT will allow AI models to be validated and translated in a plug-and-play fashion.The **broader impact** of this project will develop a collaborative platform to sort AI models by organ and disease – reducing silos and translation barriers by dramatically decreasing the time from idea to test to implementation for the highly heterogeneous medical imaging industry (the 50 largest companies capture just 29% of the market). STRAIT's success will accelerate the convergence of AI model-centric medical imaging, impacting 80 million annual radiology examinations across 6,500 facilities in the US, and ultimately benefit healthcare industries, representing one fifth of the US economy.

STRAIT is a first-of-its-kind effort to create a community of academics, vendors and societies, crafting solutions to accelerate trusted, scalable validation of AI models and translation of them to healthcare practice. Our **phase 1 work** attracted substantial industry investment in its international Kaggle challenge, led a first-ever call for AI Models with crowd sourced expert peer review at the SIIM Annual Meeting, received extensive feedback on Model Zoo design, and demonstrated integration of model results to the SIIM virtual hospital using healthcare relevant APIs. In addition, new partners to enhance clinical translation, AI data training quality, and AI interoperability were added.

In phase 2, the primary deliverable will be clinical research testing of collaborative model-centric AI platform to meet the urgent needs of scalable validation and translation of model-centric AI in medical imaging. **The intellectual merit of our work** is to bridge the gap between AI developers, technical private sectors, and healthcare providers, representing a fundamental rethinking of how model-centric AI can be validated and translated in medical imaging,

algorithm design, and medical science. Through iterative interactions with stakeholders, the STRAIT open-source and integrated suite will facilitate assessment of AI models, lighten the burden of public and private entities, and bring the latest science into the hands of end-users in healthcare industry. We will host challenges and workshops to catalyze new research and grow this new ecosystem.

Differentiators

STRAIT offers three unique value propositions: (1) a community-level ecosystem informed by major players in medical imaging, (2) innovative AI-based model validation core technologies (e.g., federated validation, uncertainty estimation, explainability) with cutting edge visualization platform, and (3) unique capabilities for establishing AI trust.

Road Map

Q2. Complete back-end infrastructure.

Q3. Strategic planning for sustainability models

Q4. Public soft-launch and enter iterative testing / updates with partners and societies.

Q5. Public release in partnership with early adopters.

Q7. Deploy at international conferences.

Q8. Transition to sustainable community model.

Partnerships

The foundation of the team is a partnership between **core project team members** (Vanderbilt University, SIIM, FlowSigma, MD.ai, Kitware) and industry consortium, technology partners, and societies:

Industry consortium: MONAI, MLCommons
Societies: IEEE, MICCAI, SPIE, RSNA, ACR.

To leverage and contribute to publicly available Big Data, we have collaborated with Kaggle in phase 1 and will extend such collaboration in phase 2. To promote track success, we have teamed extensively with NSF Convergence Accelerator Track D innovators (who are advancing AI model creation and AI applications) to apply STRAIT innovations to new markets while incorporating developments in data management and model design.

Intellectual Property

The STRAIT will release core infrastructure, data and models with permissive licenses. Anyone will be able to use this platform, even if the users need their IP to be held confidentially. The contributions will transform the open-source community (with multiple avenues of long-term industry support), enable deep interactions with technical societies (with sustainable growth), and construct a core validation community around a 501(c) structure. Identified industry partners will use the techniques to grow both startup business models and established markets, providing for sustainability.

NSF's Convergence Accelerator

www.nsf.gov/od/oia/convergence-accelerator