White Paper for NSF Smart Cyberinfrastructure (CI) Workshop

Washington DC Feb. 25-27, 2020

Naveen Sharma, Rochester Institute of Technology, Rochester, NY

We live in the era where the confluence of "big data" and software offers unprecedented opportunities to accelerate progress across all aspects of human endeavor. Big data is already enabling dramatic transformations in science, engineering, medicine, and public policy [1]. Consequently, we are witnessing an explosion of research and technology development initiatives that employ various data-driven methods under the broad umbrella of machine learning (ML) and data-driven artificial intelligence (AI). While ML/AI have gained much popularity and have shown notable successes in business for some time, fostered by industries such as e-commerce, marketing, and process optimization significant efforts are underway in its applications to natural sciences, where algorithms are used to stimulate scientific discovery. For example, advances in image processing, big data, data visualization, coupled with domain-specific computational models for climate science are enabling climatologists to model climate changes from data such as high spatial resolution (HSR) remote sensing images of sea ice (sea ice acts as both an indicator and an amplifier of climate change) [2]. Similarly, applications of ML/AI along with advances in information extraction and big data processing are enabling: biologists to gain insights into how living systems adapt; health scientists to devise targeted treatments and interventions to optimize health outcomes; education researchers to personalize pedagogy to optimize learning outcomes; social scientists to study why organizations, societies, and cultures succeed or fail; urban planners to design for optimal traffic flow; material scientists to develop new materials with properties not seen before. Progress in many areas of human endeavor is increasingly enabled by our ability to acquire, share, integrate, and analyze disparate types and modalities of data (i.e. big data) and new methods and tools for data integration, analysis, modeling, and interpretation. Holistically, seamlessly integrated "big data" and "software enabled capabilities" with networking, security resources, tools and services, and people skills collectively enables these new capabilities [3]. We advocate that to exercise their full potential future CIs ("Smart CI") will be highly programmable by the domain experts. Also, CIs enabling solutions to urban challenges will need to seamlessly integrate with the community and its citizenry.

Current ML and AI technologies do not provide easy ways for domain experts who are not ML/AI experts to develop applications. The fundamental goal of ML is to induce or synthesize programs that are able to learn from data. However, in current practice the programs are reduced to a model or set of models mapping inputs to outputs. Learning is an optimization process driven by an objective function of a predefined form [4]. Also, despite advances in "big data" the task of developing ML/AI applications is largely manual and labor-intensive. The real promise of ML/AI is rather limited to organizations possessing advance skills in computing, algorithms, and statistics. Thus, advancing the capability and capacity for ML/AI use in predictions and data-driven decision making in science, engineering, and public policy requires extensive support and involvement of skilled data scientists. Domain experts working with data scientists develop computational abstractions for relevant domains and associated methods and tools for domain data analysis, simulation, visualization, and sharing, and integration. Currently, multiple NSF-funded research efforts across wide array of domains are underway that focus on building domain-specific frameworks or platforms. Generally, these systems intend to provide good toolboxes and libraries for ML/AI based on existing techniques and domain-specific data in one place. Examples for such CI span across computational physics, climate modeling, cardiovascular simulations, material discovery, plasma physics, hydrologic modeling, and many more. In general, the focus is on building CIs comprising proven domain-specific models and abstractions along with integrated domain-data and views from various disparate sources (e.g. open data sets). In most cases, ML models are the key component of these systems, but a typical solution involves multiple such models, along with significant levels of reasoning with the models' output and input. Current technologies do not make such techniques easy to use for domain experts who are not fluent in ML nor for ML experts who aim at testing ideas and models on real-world data in the context of the overall AI system. To realize full potential, we argue that the future CI systems will need to provide easy to use interface by being *highly programmable* – both to attract a larger user base as well as software evolution. Quickly, *variability* is the number of possible different evolutions of a system (CI) where as *programmability* is the capability of a system to change or to react to external stimuli (input) in order to alter its behavior [5]. A highly programmable CI system will raise the level of abstraction at which the user conceptualizes and develops ML/AI models. Lowering the technical barriers to access CI will enable easy and productive access to AI/ML tools for science as well as a large community of science stakeholders. On one hand, the complexity of ML/AI necessitates this capability for wider scale adoption; on the hand ML/AI techniques itself can help design this capability. Thus such a capability enables both *CI for ML/AI and ML/AI or CI* paradigms. A highly-programmable CI in a specific domain, e.g. hydrologic modeling or material discovery will most likely offer a domain-specific interface – i.e. a very high level and domain-specific programming language – conducive to that domain, it will enable synthesis of a set of programs implementing ML/AI models for the data presented. Users can pick and choose which models best fit their application. Such programmable interface will exercise CI in multitudes of innovative ways and far more than a canned interface.

By 2050 it is predicted that 66% of the global population will habitat in large- and mid-size cities. To date "smart city" solutions to issues of human development, while sometimes useful, do not get to the core of the urban issues, specifically at the community[a] and neighborhood levels. Having more high-quality data about the activities of citizens, households, and businesses in a community and having easy access to this data in a secured, timely and open fashion, can undoubtedly be of great help in designing service delivery systems or managing infrastructure efficiently. Community Based Participatory Research (CBPR) is widely used and has become a standard approach in a wide variety of domains, such as environmental and sustainability studies, criminology, community psychology, studies of race and gender, urban planning, community development and urban studies, migration studies and international development studies. "*CBPR emphasizes collaborative, equitable partnerships among researchers, stakeholders and community members throughout all phases of research. ... Communities are involved in decision-making throughout the research process, from developing research questions to disseminating research findings"* [6]. CBPR emphasizes the importance of identifying and validating the community's strengths and assets, avoiding an exclusive focus on problems. We argue that future CI systems focused on urban issues (or Urban CI) will need to seamlessly integrate with communities and neighborhoods empowering its citizenry. From driverless vehicles to software that runs subways systems to dynamic bus scheduling to smart grids, Urban CIs enable and help run a city more efficiently. However, over the long run, the parameters of these engineered systems and services, such as the quality of service in terms of use of time, comfort, economic cost, etc., must continuously adapt to societal issues. In fact, the Urban CIs will need to evolve as the community evolves. True integration with the community will enable community residents to use and contribute to Urban CI as part of evolution. In other words *Urban CIs shall be for the community and by the community!* Armed with data and access to analytics, communities can cut through ideological boundaries, focus on things that matter, and engage in conversations about challenges and opportunities. Such systems can engender processes with added benefit of creating a path of dialogue and inclusion to the urban poor and of responsive and knowledgeable government around practical issues for official organizations and private actors.

[a] A group of people living in the same defined area sharing the same basic values, organizations, interests, and sense of identity.

References

[1] Chen, H., Chiang, Roger H.L., and Storey Veda, C. “Business Intelligence and Analytics: From Big Data to Big Impact”

[2] Yang, Chaowei, Yu, Manzhu, Li, Yun, Hu, Fei, Jiang, Yongyao, Liu, Qian, Sha, Dexuan, Xu, Mengchao, and Gu, Juan. "Big Earth data analytics: a survey," Big Earth Data, v.3, 2019

[3] NSF’s Blueprint for a National Cyberinfrastructure Ecosystem “ Transforming Science Through Cyberinfrastructure”

[4] Parisa Kordjamshidia, Dan Roth, and Kristian Kerstingd “Declarative Learning-Based Programming as An Interface to AI Systems”

[5] Zenil, Hector “A Behavioural Foundation for Natural Computing and a Programmability Test”

[6] Collins, S. E. "Community-based participatory research (CBPR): Towards equitable involvement of community in psychology research community ," American Psychologist, vol. 73, pp. 884-898, 2018.