RFI Response to National Big Data R&D Initiative

Srinivas Aluru, Professor
Georgia Institute of Technology

Personal Information:

Srinivas Aluru, Professor
College of Computing
Georgia Institute of Technology
1336 Klaus Advanced Computing Building
266 Ferst Drive, Atlanta GA 30332.
Ph: 404-385-1486
Email: aluru@cc.gatech.edu

Big Data Related Experience and Role in Big Data Innovation Ecosystem:

Georgia Tech (GT) is a top ranked public technology and science research institute. All of its computing and engineering departments are ranked in the top 10 by the U.S. News & World Report, with over half in top 5. We are the largest producer of engineering degrees awarded to women and underrepresented minorities. Research and education at GT is known for its real-world focus, and strong ties to industry. We are located in Atlanta, the eighth largest economy in the nation and a central hub for southeast industrial sectors. A large number of GT faculty have active research programs in big data, spanning applications and foundations, including collaborations with industry, government, and other Atlanta universities.

GT faculty projects have garnered support from all of the big data initiatives launched by major federal agencies (e.g. NSF Big Data, NIH BD2K, DARPA XDATA and ADAMS). I lead one of the eight NSF-NIH midscale big data awards made during the first round of big data federal funding. I have participated in the subsequent NITRD events including the first white house workshop on big data, and the Data to Knowledge to Action meeting. I co-chair the GT Faculty Council on Data Science and Engineering, which is charged with developing and executing a strategic plan for GT in big data. We are working towards a new Interdisciplinary Research Institute focused on data engineering and science, accompanied by an Innovation Hub for supporting industry engagement and entrepreneurship for translating GT big data research and expertise into economic and societal benefits. To support these activities, we are embarking on the construction of a 20 floor, 480,000 sq. ft. office tower augmented with 80,000 sq. ft. data center. The building will provide a long term base for GT efforts in high performance computing, big data, and computer modeling and simulation. The project will be a public-private partnership, with GT leasing about half the space and the rest made available for like-minded industry partners. Through this academia-industry co-location effort, we will drive economic development, creation of new jobs, technology engagement and transfer. We anticipate occupancy in 2017-2018.

Comments and Suggestions:

National Data Repositories: Access to data is one of the biggest impediments to big-data research. Many of the largest datasets originate in industry, government, or healthcare settings, and are not readily accessible due to issues such as privacy, loss of competitive advantage, etc. This is rightly identified as a priority issue in the Big Data National R&D Initiative. A similar problem existed in the high performance community decades ago, which was effectively solved

through national resources such as NSF TeraGrid, NSF XSEDE, shared DOE supercomputing facilities, etc. An effort of similar magnitude should be launched focused on data sharing, hosting community repositories, and supporting the compute-in-place paradigm. Where needed effective anonymization and identification/de-identification research should be supported to facilitate pervasive access to realistic datasets. Universities can act as nodal agencies to support such a national network. Georgia Tech is willing to be a partner in the area, and we are developing physical infrastructure capabilities to support such an endeavor.

Foundations on Big Data Analysis: Currently**,** much of the big data research is carried out in an adhoc manner, in the context of applications where data constitutes an overwhelming challenge. As the field progresses, unifying themes and common techniques with broad applicability are bound to emerge. Federal initiatives are already supporting research into core technologies. This should be further emphasized to encourage the fast development of broad foundations to support big-data research and development. Initiatives should target proposals that lead to development of paradigms accompanied by efforts to prove applicability in multiple domains. The analog of broadly applicable computability theory, and widely taught and used algorithmic paradigms, do not yet exist in the data domain.

Privacy and Ethical Considerations: Establishment of policies and guidelines for data sharing, along with policies for their responsible use, can remove the uncertainty in data sharing and accelerate big data research. Recent research and events such as the OSTP-MIT workshop on big data privacy explored privacy protection technologies, unsuspected breaches of privacy that can occur through seemingly de-identified big data, and rigorous computer science foundations for controlling access and ensuring privacy. Moving forward, it is important to establish uniform and legally acceptable protocols across all areas in the realm of big data, to remove impediments to research. At the same time, mandates for sharing big data where access to such data by the broadest possible community of researchers is vital for the public good, should be considered with appropriate safeguards. Specific examples that may fit this scenario include genomics driven medical care, and access to community health data that can predict impending outbreaks and environmental degradation that is responsible for illnesses.