Response to Request for Input (RFI)-National Big Data R&D Initiative

Submitted by: San Diego Supercomputer Center (SDSC), University of California, San Diego,
9500 Gilman Dr., #0505, La Jolla, CA 92093
Contact: Michael Norman, PhD, Center Director
Tel (858) 822-5450, email mlnorman@sdsc.edu

Experience working with Big Data and role in Big Data innovation ecosystem: SDSC is a major academic supercomputing center funded by the NSF with a long history in Big Data innovation, shared community infrastructure and impactful, cross domain, Big Data projects ranging from WIFIRE to OpenTopo and the Neurosciences Information Framework (NIF), all underpinned by Big Data resources such as Gordon and soon, Comet.

Comments and Suggestions On The Initial Framework

What are the gaps that are not addressed in the Visions and Priority Actions document?

A strategic plan should recognize that Big Data systems are fundamentally different in nature and require specialized technological approaches and architectures. Key considerations:

- Big data applications are "end-to-end"— typically involving *pipelines* of processing with steps that include aggregation, cleaning, and annotation of large volumes of data; filtering, integration, fusion, subsetting, and compaction of data; and, subsequent analysis, including visualization, data mining, predictive analytics and, eventually, decision making.
- Big data applications are characterized by greater diversity in software packages and software stacks, as well as a rapidly evolving software ecosystem—users should be able to run the software that best suits their application. Easy-to-use portals, or *gateways*, that hide the details of the underlying infrastructure from the end user are essential to the success of big data and extreme computing.
- Big Data systems will be heterogeneous in nature. The trends in industry, as well as the architecture of systems like Gordon and SDSC's soon-to-be-deployed Comet, point towards a path forward where Big Data systems incorporate a range of capabilities matched up to the differing needs of different parts of a processing pipeline.
- Currently, Big Data systems in industry employ shared-nothing architecture with homogeneous, commodity components, to assist in manageability and scalability of the overall system. There is a separation between the active, stream-processing components of the system versus the analytics and *post facto* processing components. However, there is a desire to build "on-line everything" systems, where all processing could be done "inline," *with* the data stream rather than as a post processing step.

From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

There are a multitude of ideas at the frontiers of Big Data R&D with high societal impact. Investing in new technologies that support the ability to collect, analyze, and mine insights from extreme data sets has potentially high payoffs in climate science, cybersecurity, human genomics and personalized medicine, the "Internet of Things," and financial market stability.

What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem?

In order to do meaningful innovation, researchers and developers need access to Big Data systems comparable in scale to problems being targeted. Google, Facebook and other commercial entities have deployed extreme scale Big Data systems, but these systems are not generally available for research & development beyond their respective companies' proprietary needs. A game-changing investment would be one in extreme scale Big Data infrastructure (including the data sets) similar to the public investment in today's supercomputing centers.

How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

New fields of research will emerge from the analysis of real world problems. A specific example is the NSF-funded WIFIRE project that analyzes real time events for wildfire prediction or the work underway at UC San Diego to apply temporal data models and techniques in the social sciences. In the latter, SDSC is at the forefront of computational sociology – a field made more powerful by utilizing parallel, semi-structured databases that operate in a cloud environment. Both project examples manipulate data to construct events and map the interrelationships between entities. What results is the emergence of innovation potential in the dimension of situational analytics. Another relevant example for connecting academia, industry and government is SmartCity San Diego (SCSD). Fundamentally SCSD is about understanding energy and modeling for predictive analysis, but also contributes to a variety of related industry partners and calls for innovation in data science: measuring, modeling integrating and interpreting models that inform projects across domains and business sectors.   Funding cross-cutting projects analogous to SCSD would increase industry partnerships, lead to better informed city planning and have a direct impact on communities.

Rationale for inclusion in the strategic plan:

SDSC embodies a unique combination of computing at scale and deep data science expertise with a track record of innovation. Projects such as the NIH-funded NIF and IntegromeDB involve acquiring data, cleaning, transforming, creating ontologies and interlinking at a scale requiring SDSC's computing infrastructure. SDSC's innovation is in forming functional scientific applications and projects where Big Data tools and techniques can be effectively utilized. A specific example is NSF-funded OpenTopo, a massive store of LIDAR data in a parallel DB2 database with background analytics that can span machines enabling scientific analysis on a large-scale platform. SDSC has the computational capacity to serve the confluence of domain scientists and technologies, innovating at their junction, especially where they don't readily apply to each other – this is SDSC's strength.

SDSC has a track record of educating undergraduate and graduate students as well as anticipating post-graduate training needs. A recent example is the NIH-funded BD2K, "An Open Resource for Collaborative Biomedical Big Data Training" that provides a learning environment with domain relevant data and software connected to pedagogically informed teaching materials on 'best of breed' cloud platforms. SDSC's strength is in constructing such learning environments where many students can leverage a common, robust infrastructure.