ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

This response to the NITRD RFI is submitted on behalf of the RENCI-led National Consortium for Data Science (NCDS, http://data2discovery.org), an innovative initiative created to address the challenges of big data that offers an informed perspective on the issues outlined in the RFI. As a public-private partnership of academia, corporate partners, nongovernmental organizations, and governmental agencies, NCDS exemplifies an approach that spans sectors, domains, and organizations. NCDS identifies big data opportunities and research needs, potential solutions, and works to implement those solutions for the benefit of society.

Creating regional centers devoted to data science and solving the challenges posed by big data is central to a national-level big data strategy as outlined in the National Big Data R&D Initiative vision draft. The strategy of applying public funding to support long-term national research objectives, in this case data science research, has already proved successful. NIH used this approach when it funded Clinical and Translational Science Award centers and the NSF employed it to fund its Supercomputer Centers. We believe a similar investment for big data research, education and outreach centers will generate an equivalent return by advancing national big data R&D goals.

In our vision, these centers will be developed guided by a five-pronged vision. First, the focus of the centers should cross agency boundaries. Second, the centers should be focal points for innovation that leverage the best public-private collaborations to address leading-edge big data challenges through fundamental and applied research. Third, the centers should serve as knowledge hubs that capture best practices, best technologies, and related knowledge for dissemination and reuse. Fourth, the centers should work closely with the educational community, from high schools to graduate and professional schools, to develop curricula and programs that promote workforce development and maintain national competitiveness. Fifth, the centers should be regionally-oriented to facilitate local impact and local interactions. Finally, the centers should be created in a way that facilitates long-term sustainability, and outreach efforts to inform the general public and policy makers about the potential and limitations of Big Data should be an integral requirement of all centers.

The challenges associated with big data span sectors, agencies, and disciplines. On the supply side, federal agencies and the programs they fund generate terabytes, if not petabytes, of data each day. Where possible the centers should work to remove stovepiped cyberinfrastructure and barriers to using and sharing data. For example, in the Earth sciences each relevant agency, (NASA, NOAA, and NSF) develops its own data management and dissemination infrastructure. Centers that can examine ways to leverage common capabilities and approaches across agencies are essential. On the demand side, end users increasingly demand data synthesis that cuts across traditional disciplinary divides. Similarly, decision-makers will need to rely on increasingly sophisticated tools that integrate multiple levels of models and observational data from an array of sources not limited to one agency or program. To comprehensively address these types of challenges requires multi-agency approaches.

A set of nationally organized, regional big data centers, possibly implemented as regional innovation hubs, can promote coordinated research and development. Fundamental data science-related challenges require focused research and development in order to support a forward-looking, national-level big data strategy as outlined in the vision document. For example, there is an unexplored presumption that all data should be saved for all time. This assumption needs to be examined much more thoroughly. Do all data need to be saved? How is an assessment done? Sensors and cyberinfrastructure may be able to create and store data with 64-bit precision. However, from a data use standpoint, this level of precision may not always be necessary. At a minimum, additional research should be supported in the areas of networking, metadata, semantic knowledge creation, linked data, unstructured data models, middleware, and data security. A set of coordinated regional big data centers/innovation hubs would draw upon regional strengths to advance R&D while rapidly disseminating best approaches among them. Furthermore, these centers should support work that transfers basic research into the applied domain and operational environments. These fundamental research areas can help to address critical gaps and challenges faces by federal agencies as they seek to implement the various presidential and OSTP open data and open access policies.

Big data challenges include managing data that are significant in complexity, size, and velocity, and federal agencies may not have adequate resources to address these challenges. Data-intensive federal agencies, (e.g. Commerce, NASA, NOAA, NSF, USGS) will require a much better understanding of data life cycle issues related to big data and their core missions. While data growth is exponential, it is not clear agencies have the resources to save, let along manage, all these data. Data lifecycle also impacts data provenance, which is key element to ensuring the trustworthiness of data. Our envisioned big data centers should also identify and create best practices related to big data.

A central goal for these centers will also be facilitating the development of data science curriculum and workforce development programs. Numerous analyses show the demand for data-related jobs far outstrips the supply of qualified applicants. A network of data science centers should work to fill this void by producing curricula to support practical, applied, professional, and research education from high school through graduate school and professional training. Data as a science is in a position similar to computer science in the 1950s. There is growing recognition that data science should be treated as a specialization inside and outside of the academy.

Finally, these proposed centers should not operate on a traditional 1 – 3 year or piecemeal funding model. They should be formulated as a national-level investment in the future of U.S. scientific and economic competitiveness and as an essential element of a national security strategy. The initial levels of funding should be sufficient to ensure their long-term viability.

The NCDS is an exemplar that addresses all the data science challenges outlined above. The NCDS has implemented innovative educational efforts and a Data Fellow program. It is establishing a first-of-its-kind Data Observatory and Laboratory. The consortium marshals regional and national-level participants from both the public and private sectors to address these challenges. As NITRD's big data initiative moves forward, the NCDS welcomes the opportunity to serve a model example, to share lessons learned, and to operate as one of the inaugural big data centers.

Principle Point of Contact:

Stanley Ahalt, Ph.D., ahalt@renci.org, 919-445-9641
Chair, NCDS Steering Committee; Director, Renaissance Computing Institute
University of North Carolina at Chapel Hill.