Data Analytics Institute: A center of excellence for large-scale data inference and computation in business, engineering, science, and education

Background
Vast amounts of data currently inundate researchers across many fields and disciplines. A consistent, university-wide, large-scale resource is needed to provide high-performance, efficient, flexible data access and computation at Kansas State University (KSU). This resource would allow KSU to participate in future research opportunities while satisfying compliance and data governance requirements from federal funding agencies. Although large datastores are essential for educational and economic development, near-term shortfalls are projected of nearly 200,000 data scientists trained to advantageously utilize big data and convert it to $300B in economic growth.

Information is currently generated as massive, high-dimensional data sets with complex correlation structures and/or nontraditional formats. These data sets arrive with unprecedented velocity. Cutting-edge research in the social sciences, life sciences, physical sciences, and education generates petabytes of data that are transformatively collected, transmitted, stored, processed, and analyzed, revolutionizing how scientists, engineers, business people, and educators approach complex problems. High-dimensional data are generated in diverse fields, including agriculture, astronomy, climate science, ecology, energy, genetic analysis, geospatial sciences, and plant and animal health. These data are often generated in real time and require rapid analysis. Other web-based sources for current massive data sets provide new realms of data to explore, such as online searches, social networking activities, and financial transactions, with potential for improved business decisions and informed policy making.

Description
This initiative proposes development of an interdisciplinary Data Analytics Institute at Kansas State University-Manhattan, staffed primarily by data scientists from the Departments of Mathematics and Statistics in the College of Arts and Sciences and faculty associated with KSU’s Institute for Computational Resarch in Engineering and Sciences (ICRES) in the College of Engineering. Local HPC resources at ICRES uniquely provide a platform to train students and staff in cyberinfrastructure. ICRES staff and students deploy entire clusters and learn state-of-the-art high-performance computing and storage by contributing to ongoing research projects. KSU HPC-trained alumni now work at Google, Lawrence Livermore National Laboratory, Cerner, Garmin, and multiple other leading technology companies.

ICRES faculty, in collaboration with campus researchers, provide requisite skills for the design of big data studies, adaptation of algorithms for parallel computing, collection, storage, and retrieval of big data, modeling and analysis of such data, and interpretation of results. New tools for big data analytics will be developed and disseminated to the broader community. A fundamental goal of the Institute is the development of innovative curricula for undergraduate and graduate students to engage in large-scale data-driven science and engineering. The Institute could specifically contribute collective expertise to precision agriculture, bioinformatics, security, and enhancement of secondary education and provide significant advancements for federal and state initiatives on STEM workforce development.

Relevance
The establishment of a Data Analytics Institute aligns closely with K-State 2025 goals related to research, graduate, and undergraduate education, including research experiences for undergraduates. The Institute will focus on university strengths and critical needs, particularly in biosciences and animal health at KSU. Creating the Institute with cluster hires and/or joint appointments will strengthen and expand research funding opportunities throughout the university. In addition, the Institute will facilitate corporate partnerships with industry in the Kansas-Missouri Animal Health Corridor. The KSU Olathe campus offers a convenient venue for engagement between professional development and/or business related to big data.

KSU’s ICRES has consistently developed and influenced cyber-infrastructure for research and education. With existing collaborations between leading national and international research organizations (e.g., XSEDE) and the anticipated arrival of NBAR, ICRES has the potential to form alliances with and attract a multitude of cyber-enabled and bioinformatics companies to Manhattan. This project will be a catalyst for these endeavors, provide a vital research test bed, and establish a regional center to train the future cyber-enabled workforce.

Agency Contact Information
National Science Foundation
Michael Vogelius, Director of Mathematical Sciences, spantula@nsf.gov

Farnam Jahanian, PhD, Assistant Director for CISE, fjahania@nsf.gov
Summary: Vast amounts of data currently inundate researchers across many fields and disciplines. A consistent, university-wide, large-scale resource is needed to provide high-performance, efficient, flexible data access and computation at Kansas State University (KSU).

Opportunity: Cutting-edge research in social sciences, life sciences, physical sciences, and education generates potentially overwhelming amounts (many petabytes) of data that are transformatively collected, transmitted, stored, processed, and analyzed, thereby revolutionizing how scientists, engineers, business people, and educators approach complex problems.

Solution: The institute will work with researchers to design Big Data studies, as well as oversee the collection, storage, and retrieval of Big Data. The Institute will then model and analyze that data and interpret results. New tools for Big Data analytics will be developed. A fundamental goal of the Center will be the development of innovative curricula for undergraduate and graduate students to engage in large-scale, data-driven science and engineering.

Impact: Big Data and supercomputing are used to expand agricultural production (particularly genetic phenotyping to develop drought-resistant, pest-resistant strains of wheat), increase food security, investigate cyber attacks and consequently improve national cybersecurity, and develop bioinformatics. Development of a workforce trained in Big Data science is expected to help offset a predicted shortfall of 200,000 data scientists and generate $300 billion in economic growth.

Equipment & Expertise: Dr. Daniel Andresen – highly scalable distributed systems, supercomputing; Dr. Andrew Bennett & Dr. Gary Gadbury – high-dimensional mathematical data analysis; Beocat – KSU’s local supercomputer with over 3PB of available storage and 3,500 cores of computing power.