Big Data Analytics Institute: A Center of Excellence for Large-Scale Informatics in Business, Engineering, Science, and Education

Background
Vast amounts of data currently inundate researchers in all fields of study, but the data deluge especially impacts and challenges university researchers. A consistent, university-wide, large-scale resource able to provide high performance and efficient, flexible data access is therefore vital to those working to analyze and use this data. Attaining this type of resource and resulting data access will equip Kansas State University (KSU) for future research opportunities and satisfy compliance and data governance requirements for federal funding agencies. Although large datastores are essential for education and economic development, near-term shortfalls of nearly 200,000 trained data scientists able to utilize big data, and convert it to $300B in economic growth, are projected.

Information is currently generated as massive, high-dimensional, often heterogeneous data sets with complex correlation structures and/or nontraditional formats, typified by high volume, velocity, and variety. Big data refers to data sets characterized above and big data analytics are techniques for discovering patterns, unknown correlations, and large-scale inference methods for reliable variable selection and prediction. Cutting-edge research in social sciences, life sciences, physical sciences, and education generates petabytes of data that are transformatively collected, transmitted, stored, processed, and analyzed, thereby revolutionizing how scientists, engineers, businesspeople, 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. Web-based sources also provide new realms of massive data to explore: linked data, spatial data, and natural language text, with applications to search, business intelligence, social media, digital humanities, and digital arts. This offers potential for improved decision support and informed policy making.

Description
This initiative proposes development of an interdisciplinary Center for Data Analytics at KSU, 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. These faculty, in collaboration with campus researchers, will provide the requisite skills for designing big data studies; collection, storage, and retrieval of big data; modeling and analysis of such data; and interpretation of results. New tools for data modeling, integration, and mining, statistical machine learning, and information visualization will be developed and disseminated to the broader community. A fundamental goal of the Center will be development of innovative curricula for undergraduate and graduate students to engage in large-scale data science and computational applications such as data mining and visualization, recommendation, or pattern recognition. The Center will contribute collective expertise to applications of STEM such as precision agriculture, bioinformatics, health and medical informatics, security, and enhancement of secondary education. It will also provide significant advancements for federal and state initiatives on STEM workforce development.

Relevance
Establishing a Kansas Center for Data Analytics is well-aligned with KSU 2025 goals related to graduate and undergraduate education and research, including research experiences for undergraduates. The Center will focus on university strengths and critical needs, particularly biosciences, animal health, data mining, and informatics at KSU. In addition, utilizing cluster hires and/or joint appointments will strengthen and expand research funding opportunities university-wide. The Center will also facilitate corporate partnerships with industry in the Kansas-Missouri Animal Health Corridor. The KSU Olathe campus offers a convenient venue for business engagement and/or professional development 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 and the anticipated arrival of NBF, ICRES can form alliances with and attract many cyber-enabled bioscience and biotech companies. This project will stimulate these endeavors, provide a vital research test bed, and establish a regional center in which to train a future cyber-enabled workforce.

Agency Contact Information
Michael Vogelius, spantula@nsf.gov
Director, Math. Sciences, NSF
Farnam Jahanian, fjahanian@nsf.gov
Asst. Director, CISE, NSF
Summary: Vast amounts of data currently inundate researchers in all fields of study, specifically university researchers. We will develop new domain-specific techniques and tools for efficient, secure access to, synthesis of, and education about big data.

Opportunity: Cutting-edge research in social sciences, life sciences, physical sciences, and education generates petabytes of data that are transformatively collected, transmitted, stored, processed, and analyzed, thereby revolutionizing how scientists, engineers, business people, and educators approach complex problems. This presents urgent demand for raining and infrastructure for data science and applied computing (informatics) to support decision support, pattern recognition, adaptive systems, and similar applications.

Solution: The center will work with researchers to design big data studies; collect, store, and retrieve big data (currently petascale volume and velocity, with high variety); integrate, clean, model, analyze, and present such data; and interpret results. New tools for big data analytics will be developed. A fundamental goal of the Center will be development of applications in informatics and innovative curricula for undergraduate and graduate students to engage in large-scale data-driven science and engineering.

Impact: Big Data is used to improve agricultural production (particularly genetic phenotyping for development of more drought- and pest-resistant strains of wheat), increase food security, and investigate cyber attacks, thereby improving national cybersecurity and bioinformatics. Development of a workforce trained in data sciences and informatics and will compensate for a predicted shortfall of 200,000 data scientists and help generate a projected $300 billion in economic growth throughout the industry.

Expertise and Equipment: Dr. Daniel Andresen and Dr. William H. Hsu – highly-scalable distributed systems, supercomputing, informatics, machine learning, intelligent systems, visualization; Dr. Andrew Bennett and 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.