

**Center for Data and Search Informatics
School of Informatics Indiana University
Research and Academic Vision
Spring 2009**

Research and Academic Vision

Digital imaging, sensors, analytical instrumentation and other technologies are becoming increasingly central to all areas of science. Increases in computing power drive advances in modeling and simulation that extend the reach of science and medicine. Improvements in networking increase access to information, instrumentation, and colleagues in the US and abroad. Digital data are the common thread linking these powerful trends in science and society.

What is the landscape of digital data research? A recent report by an interagency working group [1] on digital data casts the landscape into five roughly partitioned areas, which we find to be a useful guide for discussing our research and educational vision for the school:

- Data systems - scalability, systems integration, robustness, fault tolerance
- Data discovery and dissemination – capabilities for searching, understanding, visualizing and interacting with data; ethical use of data
- Data protection – protecting data security, privacy and confidentiality; intellectual property rights
- Data quality and disposition: validation, provenance, and long-term preservation; best practices for disposition decision making
- Data integration and interoperability – integration and interoperability of data and data tools; databases

DSI sees open research opportunities in these five areas when examined from the perspective of scale. Open challenges exist in looking at, making sense of, and tracking provenance of large amounts of data (*scale in volume*); in making meaningful data available to policymakers, non-domain experts through intelligent user interfaces that connect users and their goals (*scale in user diversity*); challenges in integration and interoperability across databases (*scale in heterogeneity*). Challenges exist in devices distributed across a population that must balance rich information, energy conservation, and bandwidth utilization (*scale in number and kind of devices*). Open challenges exist throughout in incentives and barriers to data sharing.

Success in the area includes the recent funding of the PTI Data to Insight Center by the Lilly Foundation as part of a larger multi-million dollar gift to fund PTI. Leake and Plale are collaborating on a \$450K NSF grant on data provenance. Plale and Groth are collaborating on an Eli Lilly funded effort. Plale lead IU in a \$20M data preservation proposal (NSF DataNet) with UIUC and UMichigan. Plale gave the keynote talk at the NSF Workshop on Instrumentation Needs of Computer and Information Science Engineering, Snowbird, Utah, July 2008. See also references [2, 3, 4, 5, 6, 7] for representative publications.

As to academics, a recent JISC report [8] identifies and distinguishes four data-related roles that will be needed in the future. These are as follows:

Data Creator - Researchers with domain expertise who produce data. These people may have a high level of expertise in handling, manipulating and using data.

Data Scientist - People who work where the research is carried out – or, in the case of data center personnel, in close collaboration with the creators of the data – and may be involved in creative enquiry and analysis, enabling others to work with digital data, and developments in data base technology.

Data Manager - Computer scientists, information technologists or information scientists and who

take responsibility for computing facilities, storage, continuing access and preservation of data.

Data Librarian - People originating from the library community, trained and specializing in the curation, preservation and archiving of data.

The educational needs of the data creator and data librarian we see as being addressed through academic programs in the domain discipline or in a traditional library school respectively. It is our view that the School of Informatics is uniquely positioned to lead the country in the education and production of *Data Scientists* and *Data Managers* at the undergraduate and graduate levels. The new specialization in data and search under development for the CS BS is a first step towards the data scientist program, however the curriculum will need continual refinement through additional and better-targeted courses (and faculty). The Informatics master's degree specializations could be a good fit for an Info MS Data Scientist or Info MS Data Manager.

New courses have been developed in this area (i.e., CS graduate data mining course (Fall 2009), and introductory topics in data and search informatics (Fall 2009)). Enrollments are good in all data-related courses. New and existing courses are beneficial to other research areas in the School. Security and health care applications depend on much better information integration. Beyond the School, curriculum partnerships could be developed. Data creators are not aware of the implications of the data they create for issues of ownership, custodianship, ethical data handling, etc. This knowledge is broadly beneficial and could be part of a minor for external students.

References

- [1] Harnessing the Power of Digital Data for Science and Society, Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, January 2009
- [2] N. Dexters, P.W. Purdom, and D. Van Gucht, Peak-Jumping Frequent Itemset Mining Algorithms, *Proceedings PKDD Int. Conf. Principles of Data Mining and Knowledge Discovery*, 2006
- [3] Kam Woods and Geoffrey Brown, Migration Performance for Legacy Data Access, *3rd Digital Curation Conference*, 2007
- [4] Yuqing Wu, Sofia Brenes, Dirk Van Gucht, and Pablo Santa Cruz, Trie Indexes for Efficient XML Query Evaluation. *WebDB*, 2008.
- [5] Kalpana Shankar, Order from chaos: the poetics and pragmatics of scientific recordkeeping', *Journal of the American Society for Information Science and Technology (JASIST)*, 2007, 58, 1457-1466
- [6] David B. Leake, Joseph Kendall-Morwick. Towards Case-Based Support for e-Science Workflow Generation by Mining Provenance. *ECCBR 2008*: 269-283
- [7] Yogesh L. Simmhan, Beth Plale, Dennis Gannon: Karma2: Provenance Management for Data-Driven Workflows *Int. J. Web Service Research* 5(2): 1-22, 2008.
- [8] Swan, Alma and Sheridan Brown, Skills, Role & Career Structure of Data Scientists & Curators: Assessment of Current Practice & Future Needs, JISC, 31 July 2008