CSCI B669 Scientific Data Management and Preservation
Indiana University
Syllabus

Meets: Wednesdays, 5:30-8:00, 919 E. 10th St. Informatics East 130
Instructor: Beth Plale
Office: 919 10th St., Informatics East 256
Email: plale@cs.indiana.edu
Office phone: 812 855 4373
Office hours: TBA or by appointment

Credits: 3

Course Overview: Environmental sensors, sequencing instruments, and the Internet all contribute to fundamental changes in the nature of scientific research, suggesting data-driven research as the **4th Paradigm of Science**. Digital data produced through computation is not a commodity that is consumed in a single use, but is an important and invaluable intellectual asset that can be used repeatedly to fuel new ideas and insights. Managing research data for the long-term, ensuring its continued access, has emerged as a major challenge. But as the well-known 2003 "Atkins report" states, "absent systemic archiving and curation of intermediate research results… data gathered at great expense will be lost". In this course we examine the full lifecycle of digital data with a focus on data generated and used in the course of advancing scholarship and science.

The course is divided into several sections:

I. **Motivating Applications**
II. **Data, Metadata and Semantics**
III. **Big Data: Analytics**
IV. **Big Data: Data Management**
V. **Data Preservation**

This is a reading intensive and hands-on course. It covers a large and broad body of material. Most of the readings and content of the course are provided below, though you should expect readings will be added. The schedule may change subject to student needs and interests and opportunities that may arise. While the majority of the students are expected to have a computer science background, not all will. Students should expect to feel more comfortable with some material and less so with others.

Learning Objectives:

- Develop an understanding of the big applications motivating science
- Gain working understanding of tools in use in big data data management
- Learn principles behind data management
- Gain understanding of challenges of data analysis at scale
- Develop appreciation for issues of data curation and preservation

Requirement and Expectations:

- Attendance in all course sessions is expected. If you have a conflict that will result in missing class, please notify me and complete work in advance. Only in extenuating circumstances should an extension be asked for.
- Complete required readings at the level of preparation thorough enough to discuss and critique readings for each week.
- Complete all assignments on time.

Evaluation:

Class project …… 25%

Examination(s) ……. 30%

Final project ……35%

Class participation …… 10%

Class Participation:

Evaluation of class participation will be based on attendance and completion of in-class activities. The latter will consist of things such as in-class writings, reports from in-class group work, etc.

Course readings: course readings will be taken from *Scientific Data Management: Challenges, Technology, and Deployment, A. Shoshani and D. Rotem, eds. CRC Press, 2010; and from relevant and current research papers.* Good motivating and background information for the course can be found in *4th Paradigm: Data Intensive Scientific Discovery, Tony Hey, Stewart Tansley, and Kristin Tolle eds.*

Assignments:

Class project: the student working as part of a small team, will conduct an evaluation of two comparable tools. The evaluation will include hands on experience with the tools. The result of the project will be an in depth tutorial delivered in the classroom. Evaluation will be based on presentation (10%) and quality and depth of presentation materials (90%).

Final project: the final project will be on a topic of the student's choosing. The instructor will provide projects from which the student can choose or the student can choose a project based on their research. The project will likely have a technology component, and possibly a programming component, but it need not if the student is not comfortable with programming and has a sufficiently interesting project thesis. Students can work in teams on the final project, but the amount of work needs to be equivalent to what one would do alone.

Course Schedule and Readings (reading list will be updated)

I. Motivating Applications

Week 1 (Jan 9): Motivation

Readings:

Jim Gray on eScience: A Transformed Scientific Method, *Jim Gray,* 4th Paradigm, pp. xix – xxxiii.

Got Data? A Guide to Data Preservation in the Information Age, *Francine Berman,* Communications of ACM, Dec 2008, 50(12) pp. 50-56.

Optional:

Week 2 (Jan 16): Motivating Applications

Readings:

Special Online Collection: Dealing with Data. In the 11 February 2011 issue, *Science* joins with colleagues from *Science Signaling*, *Science Translational Medicine*, and *Science* Careers to provide a broad look at the issues surrounding the increasingly huge influx of research data. This collection of articles highlights both the challenges posed by the data deluge and the opportunities that can be realized if we can better organize and access the data. *Science* is making access to this entire collection **FREE** (simple registration is required for non-subscribers). http://www.sciencemag.org/site/special/data/

Optional:

Gray's laws: database-centric computing in science, *Alexander S. Szalay, José A. Blakeley,* 4th Paradigm, pp. 5-12.

2020 Vision for Ocean Science, 4th Paradigm, pp. 27-38.

Instrumenting the earth: next-generation sensor networks and environmental science, *Michael Lehning, Nicholas Dawes, Mathias Bavay, Marc Parlange, Suman Nath, Feng Zhao*, from 4th Paradigm, pp. 45-51.

II. Data Interoperability, Metadata, and Semantics

Week 3 (Jan 23) : Interoperability and Metadata

Tutorial on Tools and Introduction of Class Project

Readings:

Interoperability and Data Integration in the Geosciences, *Michael Gertz, Carlos Rueda, and Jianting Zhang*, from Scientific Data Management, Ch. 10, pp. 369-398.

Scientific Process Automation and Workflow Management, *Bertram Ludäscher, Ilkay Altintas, Shawn Bowers, Julian Cummings, Terence Critchlow, Ewa Deelman, David De Roure, Juliana Freire, Carole Goble, Matthew Jones, Scott Klasky, Timothy McPhillips, Norbert Podhorszki, Claudio Silva, Ian Taylor, and Mladen Vouk*, Scientific Data Management, Ch 13.

Week 4 (Jan 30) : Semantics and Ontologies

Readings:

Linked Data - The Story So Far, *Christian Bizer, Tom Heath, Tim Berners-Lee*: Heath, T., Hepp, M., and Bizer, C. (eds.). Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS). Vol 5, Issue 3, pp. 1-22, 2009.

Ontologies : A contribution to the DL/DB debate*, Nadine Cullot, Christine Parent, Stefano Spaccapietra, and Christelle Vangenot,* The first International Workshop on Semantic Web and Databases, Sept 2003

Optional:

III. Big Data: Data Analytics

Week 5 (Feb 6) : Data Analytics Benchmarks

Readings:

MalStone: Towards a Benchmark for Analytics on Large Data Clouds, *Collin Bennett, Robert L. Grossman, David Locke, Jonathan Seidman and Steve Vejcik*, The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010), ACM, 2010.

Week 6 (Feb 13) : Tools Evaluations: Pig and R

Week 7 (Feb 20) : Tools Evaluations: Cassandra and MongoDB

Week 8 (Feb 27): Tools Evaluations: MPI I/O and netCDF

Week 9 (Mar 6): Data Analytics at Scale

Readings:

Discovering Emergent Behavior from Network Packet Data: Lessons From the Angle Project, *Robert L Grossman, Michal Sabala, Yunhong Gu, Anushka Anand, Matt Handley, Rajmonda Sulo and Lee Wilkinson*, in Next Generation Data Mining, edited by Hillol Kargupta, Jiawei Han, Philip S Yu, Rajeev Motwani and Vipin Kumar, CRC Press, Boca Raton, 2009, pages 243-260.

<Big Data: Data Analytics>

Week 10 (Mar 20) : Midterm Examination

IV. Big Data: Data Management

Week 11 (Mar 27) : noSQL Data Stores

Readings:

Optional:

Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications. In the *Proceedings of ACM SIGCOMM*, Aug 2001

Week 12 (Apr 3) : Data Driven Computing

Readings:

Parallel Data Storage and Access, *Robert Ross, Alok Choudhary, Garth Gibson, And Wei-keng Liao*, Scientific Data Management, CRC Press, 2010, pp. 35-72.

Memcached Design on High Performance RDMA Capable Interconnects, *Jithin Jose, Hari Subramoni, Miao Luo, Minjia Zhang, Jian Huang, Md. Wasi-ur-Rahman, Nusrat S. Islam, Xiangyong Ouyang, Hao Wang, Sayantan Sur, and Dhabaleswar K. Panda*, 2011 Int'l Conf on Parallel Processing, IEEE Computer Society, pp. 743 – 752.

Dyamic Storage Management, *Arie Shoshani, Flavia Donno, Junmin Gu, Jason Hick, Maarten Litmaath, and Alex Sim*, Scientific Data Management, CRC Press, 2010, pp. 73-114.

Week 13 (Apr 10): Provenance Capture and Actionable Data

Readings:

Metadata and Provenance Management, *Ewa Deelman, Bruce Berriman, Ann Chervenak, Oscar Corcho, Paul Groth, and Luc Moreau,* Scientific Data Management, Ch 12, pp. 433-466.

Provenance from Log Files: a BigData Problem, Devarshi Ghoshal and Beth Plale, under review 2013.

<actionable data>

Optional:

Open Provenance Model v1.1 http://eprints.ecs.soton.ac.uk/21449/

The foundations for provenance on the web. *L. Moreau*, Foundations and Trends in Web Science, 2(2-3) (2010) 99-241.

VI. Data Preservation

Week 14 (Apr 17): Data Lifecycle; Trident Workflows for Digital Libraries

Readings:

Hedstrom, M. (2001). Exploring the concept of temporal interoperability as a framework for digital preservation. *Third DELOS Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries (September 8-9, 2001; Darmstadt, Germany)* 6 pp.

Week 15 (Apr 24): Long term repositories; Institutional Repository as

Readings:

Management and preservation of research data with iRODS, Hedges, M., Hasan, A., & Blanke, T. (2007). *Proceedings of the ACM First Workshop on Cyberinfrastructure: Information Management in e-Science*. ACM, New York, pp. 17-22.

Week 16 (May 1): Final exam

Academic Integrity

Academic integrity requires that students take credit only for their own ideas and efforts. Misconduct, including cheating, fabrication, plagiarism, interference, or facilitating academic dishonesty, are prohibited because they undermine the bonds of trust and cooperation among members of this community and between us and those who may depend on our knowledge and integrity. Complete details are contained in the Indiana University *Code of Student Rights, Responsibilities, and Conduct.*