IST687 Applied Data Science

(To be offered in spring 2013)
Instructor: Jian Qin

Email: jqin@syr.edu
Office: 311 Hinds Hall
Phone: 315-443-5642

Time: on your schedule
Location: anywhere

Course Description

Applied Data Science includes three modules: 1) fundamentals about research data, 2) data management with case studies, and 3) use of research data and broader issues relating to tools for management, analysis, and visualization, as well as quality and publication practices. The first module provides an overview of research data and data management, including data fundamentals such as scales, types, levels, data structures and models, data formats, and technologies used to store, retrieve, and manage data. The second module uses case studies regarding data collection, processing, transformation, management, and analysis to provide students with hands-on experience. In the last module, students will be introduced to methods and tools for evaluating data quality and creating datasets from multiple data sources for analysis and visualization in various contexts. Students will work as an interdisciplinary team on a comprehensive research data project throughout the semester under the guidance of the instructor. The performance of each student is based on exercises, group reports, class participation, and the course project.

Learning Objectives

At the end of the course, students are expected to understand:

- Concepts and characteristics of research data and practices
- Principles and practices in data management and use
- Technologies used to manage, manipulate, and analyze data
- Procedures and methods of using data for inquiry

At the end of the course, students are expected to be able to:

- Identify the needs for organizing, reporting, and managing research data
- Represent datasets with metadata
- Create analysis-ready datasets to meet research needs
- Design and execute workflows for data management and analysis

What does it take to succeed in the course?

- An interest and passion in data science career in the corporate, academic, or government sector
- Curiosity on natural and social phenomena and basic computer skills, e.g., HTML coding, SQL scripting, and database design.
- Motivation to learn and excel in individual and team work.

Required Readings

There is no required textbook, but required readings will be available in Blackboard as electronic documents for reading and printing. Students are expected to read the assigned materials for discussions and coursework.

Contributions to Grade

The work for this class will involve a mixture of individual assignments, case study reports, and a final project.

- **Exercises** (4 x 8% = 32%) are designed for you to practice the necessary skills in carrying out data management and manipulation tasks.
- **Case study reports** (2 x 10% = 20%) are designed to maximize the usefulness of case study results and provide staged deliverables for the final project.
- **Final project** (35%): the project team will be formed early on in the semester and the team members will work together throughout the semester on many tasks.
- **Participation** (13%) includes your attendance and participation in class discussions and activities.

Academic Integrity

The academic community of Syracuse University and of the School of Information Studies requires the highest standards of professional ethics and personal integrity from all members of the community. Violations of these standards are violations of a mutual obligation characterized by trust, honesty, and personal honor. As a community, we commit ourselves to standards of academic conduct, impose sanctions against those who violate these standards, and keep appropriate records of violations. The academic integrity statement can be found at: http://www.ist.syr.edu/courses/advising/integrity.asp

Student with Disabilities

In compliance with section 504 of the Americans with Disabilities Act (ADA), Syracuse University is committed to ensure that “no otherwise qualified individual with a disability…shall, solely by reason of disability, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity…" If you feel that you are a student who may need academic accommodations due to a disability, you should immediately register with the Office of Disability Services (ODS) at 804 University Avenue, Room 308 3rd Floor, 315.443.4498 or 315.443.1371 (TTD only). ODS is the Syracuse University office that authorizes special accommodations for students with disabilities.

Schedule

Date	Topics
1/14 Week 1	Introduction to the course Overview of data science
1/21 Week 2	Fundamentals about data • Scales, types, and levels • Structures and models • Data formats and standards
1/28 Week 3	Data and research lifecycle • Stages of research lifecycle • Status of data at different stages of research lifecycle • Requirements for data management
2/4 Week 4	Data provenance • What is data provenance? Why does it matter? • Data provenance framework • Workflow management systems and data provenance
2/11 Week 5	Managing data with metadata • Metadata standards and tools for research data • Examples of metadata records for research datasets • When to create metadata for datasets? What metadata standards should I use?
2/18 Week 6	Managing data with repositories • What data repositories are out there? • Types and functions of data repositories • Technical architecture of data repositories
2/25 Week 7	Large-scale database systems (1) • NoSQL systems • Key-value stores • Tradeoffs of SQL and NoSQL
3/4 Week 8	Large-scale database systems (2) • Querying data in NoSQL • Lab
3/11 Week 9	Spring Break. No class.
3/18 Week 10	Preparing analysis-ready datasets (1) • Types of research problem and data needs

	• Methods of locating relevant datasets • Evaluation of the suitability of data
3/25 Week 11	Preparing analysis-ready datasets (2) • Data selection and transformation • Data mashing • Documentation of processing methods and parameters
4/1 Week 12	Developing data management project: • Data set characteristic analysis • Needs assessment • User roles (researchers, lab staff, IT staff, etc.) • Goals and Planning • Policy development • Curation and preservation • Enabling technologies • Quality control
4/8 Week 13	Interactive data products • Analysis-ready datasets • Interactivity of datasets • Tools for creating interactive datasets
4/15 Week 14	Data product creation and repurposing • What is a data product? • Legal and ethical issues in data product creation and repurposing • Design issues in data product creation
4/22 Week 15	Data quality, discovery, and publishing • Quality criteria • Data repositories and discovery • Directory services • Data publishing and citation
4/29 Week 16	Project presentations and discussions Wrap-up