Educating a New Breed of Data Scientists for Scientific Data Management

Jian Qin

School of Information Studies
Syracuse University

Microsoft eScience Workshop, Chicago, October 9, 2012
Talk points

› Data science (DS) and data scientists in the context of scientific data
› An iSchool version of the DS curriculum
› Findings and lessons from implementing the DS curriculum
› A new breed of data scientists: the iSchool approach
What is data science?

“An emerging area of work concerned with the collection, presentation, analysis, visualization, management, and preservation of large collections of information.”

Data science and scientific research

Plan, design, consult for, implement, and evaluate data management projects and services

Ingest, store, organize, merge, filter, and transform data and create analysis-ready data
What data scientists are expected to do: the job market

Scientific Data Management Specialist
- Design, develop, implement, and manage high-throughput automatic data processing infrastructure for large databases in a mature system
- Develop and improve the infrastructure supporting this system
- Interface with multiple data providers to design, build, and maintain their customized databases
- Clarify requirements, feature requests and bug reports for software developers and assist in testing code

Laboratory Data Management Specialist
- Administer operational database
- Assure the quality of data database content
- Interact closely with researchers, lab managers, and platform coordinators
- Track deliverables against budget and prepare data reports
- Collaborate closely with IT and bioinformatics colleagues
- Assist IT in gathering workflow requirements
- Test changes and updates in IT systems
- Create and maintain app documentation

Data Modeling/Management Specialist
- Working closely with the high performance computing and the IT manager
- Develop a data model for complex multi-scale rocks
- Design and organize a database and complex queries
- Integrate and manage multi-scale rocks subjected to large-scale scientific computing applications


10/16/2012
“We’re increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.”

What data scientists are expected to do: the difference from tradition

› Data scientists are more likely to be involved across the data lifecycle:
  - Acquiring new data sets: 33%
  - Parsing data sets: 29%
  - Filtering and organizing data: 40%
  - Mining data for patterns: 30%
  - Advanced algorithms to solve analytical problems: 29%
  - Representing data visually: 38%
  - Telling a story with data: 34%
  - Interacting with data dynamically: 37%
  - Making business decisions based on data: 40%

How should educational programs address the challenge?

A case of the CAS in Data Science program at Syracuse iSchool
Story 1: cognitive-demanding workflows and data management

- **Domain**: Thermochronology and tectonics

- **What’s involved**: rock samples from drilling and field observation, sliced and grained rock samples

- **Data types**: Excel data files (lots of them), spectrum and microscopic images, annotations

- **Analysis**: modeling and sensemaking by combining data from multiple data files with specialized software

- **Bottleneck problem**: manually matching/merging/filtering data is extremely cumbersome and the problem is compounded by the difficulty finding the right data files
Story 2: highly automated workflows

- **Domain:** Astrophysics: gravitational wave detection
- **What’s involved:** data ingestion from laser interferometers, raw data calibration and segmentation, workflow management, provenance

- **Data types:** streaming data from the laser interferometers, images
- **Analysis:** detection of “events”
- **Bottleneck problem:** tracking of data and processes and the relationships between them
What are expected of data scientists?

- Ability to use a wide variety of tools for documentation, analysis, and report of data
- Knowledge of a subject domain
- Data modeling, database and query design
- Collaboration, communication, and coordination
- OS, Programming languages
- Encoding languages
- Content and repository systems
Analytical skills: domain modeling

- Requirement analysis
- Workflow analysis
- Data modeling
- Data transformation needs analysis
- Data provenance needs analysis

- Interview skills, analysis and generalization skills
- Ability to capture components and sequences in workflows
- Ability to translate domain analysis into data models
- Ability to envision the data model within the larger system architecture
Analytical skills: from data sources to patterns, relationships, and trends

“Hacking”

Analytical tools

Knowledge

Data lifecycle

Research lifecycle

Data products
Data management skills: data lifecycle and infrastructural services

- Metadata standards
- Encoding language
- Semantic control
- Identify management

- Processed, transformed, derived, calculated, … data

- Scientific data formats
  - Common data format
  - Image formats
  - Matrix formats
  - Microarray file formats
  - Communication protocols

- Metaformats
  - DSV
  - CSV
  - XML

- Data structures
  - Tuple
  - Set
  - List
  - Array
  - Tree

- Physical data
  - Bits
  - Bytes
  - Characters
  - Strings

Infrastructural services

- Data source discovery
- Data curation
- Data preservation
- Data integration and mashup
- Data citation, publication, and distribution
- Data linking and interoperability
- …
Technology skills with excellent communication skills

<table>
<thead>
<tr>
<th>TECHNOLOGY SKILLS</th>
<th>COMMUNICATION SKILLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>› Operation systems</td>
<td>› Interviews</td>
</tr>
<tr>
<td>› Repository systems</td>
<td>› “Ice breaking”</td>
</tr>
<tr>
<td>› Database systems</td>
<td>› Community building</td>
</tr>
<tr>
<td>› Programming languages</td>
<td>› Institutionalization</td>
</tr>
<tr>
<td>› Encoding languages</td>
<td>› Stakeholder buy-in</td>
</tr>
<tr>
<td>› Specialized programming</td>
<td></td>
</tr>
</tbody>
</table>
No superman model for beginning data scientists

Data scientists
Core:
Applied data science
Databases

Data analytics
Data storage and management
General system management
Data visualization
The CAS in Data Science program at SU

› Required:
  – Data Administration Concepts and Database Management
  – Applied Data Science

› Elective:

**Data Analytics**
- Data Mining
- Basics of Information Retrieval Systems
- Natural Language Processing
- Advanced Information Analytics
- Research Methods
- Statistical Methods

**Data Storage and Management**
- Technologies for Web Content Management
- Foundations of Digital Data
- Creating, Managing, and Preserving Digital Assets
- Data Warehousing
- Advanced Database Management

**Data Visualization**
- Information Architecture for Internet Services
- Information Visualization

**General Systems Management**
- Enterprise Technologies
- Managing Information Systems Projects
- Information Systems Analysis
What we learned from the program development

› Data science is a moving target with multiple focal points
  – Versions from statistics, computer science, and library and information science

› Skills vs. theories
  – Students are anxious to learn skills but not so interested in theories
  – Theories help build visions

› Sufficient hands-on time for technologies and tools

› Authentic learning through real-world data management projects
Reconciliation of the two views of data science

“It’s an emerging area of work concerned with the collection, presentation, analysis, visualization, management, and preservation of large collections of information.”


“We’re increasingly finding data in the wild, and data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.”

The iSchool’s version of data science education

Eventually the iSchool data science program will build the foundation for super data scientists...
eScience Librarianship Curriculum Project:  
http://eslib.ischool.syr.edu/

Science Data Literacy Project:  
http://sdl.syr.edu/

CAS in Data Science:  
http://ischool.syr.edu/future/cas/data science.aspx