Data Literacy for Data Science Professionals

Jian Qin
School of Information Studies
Syracuse University, NY, USA
Feeling the pressure of data deluge in the digital information world ...

...in our health care

http://ars.els-cdn.com/content/image/1-s2.0-S1053811905002508-gr4.jpg
in our neighborhood

http://www.redfin.com/homes-for-sale#!market=boston&region_id=112&region_type=1&v=8

Your Interests First
From open (house) to close, our agents are on your side.
In the business world...

(Feinleib, 2012)
...in science research

http://www.sciencemag.org/content/331/6018.cover-expansion
### Shift in Science Paradigms

<table>
<thead>
<tr>
<th>Period</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thousand years ago</td>
<td>Science was empirical, describing natural phenomena</td>
</tr>
<tr>
<td>A few hundred years ago</td>
<td>Theoretical branch using models, generalizations</td>
</tr>
<tr>
<td>A few decades ago</td>
<td>A computational approach simulating complex phenomena</td>
</tr>
<tr>
<td>Today</td>
<td>Data exploration (eScience) unify theory, experiment, and simulation</td>
</tr>
<tr>
<td></td>
<td>- Data captured by instruments or generated by simulator</td>
</tr>
<tr>
<td></td>
<td>- Processed by software</td>
</tr>
<tr>
<td></td>
<td>- Information/Knowledge stored in computer</td>
</tr>
<tr>
<td></td>
<td>- Scientist analyzes database/files using data management and statistics</td>
</tr>
</tbody>
</table>

Emerging concepts that are going to stay and matter to information professionals
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 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.
“Data scientists are the people who understand how to fish out answers to important business questions from today’s tsunami of unstructured information.”

Data literacy

“the ability to examine multiple measures and multiple levels of data, to consider the research, and to draw sound inferences”
(Love, 2004, p. 22)

“the ability and skills in collecting, processing, manipulating, evaluating, and using data for scientific inquiry.”
(Qin, 2006)
Not all kinds of data literacy are equal

- Bioinformatics
- Biomedical informatics
- Ecological informatics
- Geo-informatics
- Scientometrics
- Social informatics
- ...
But there is a core set of knowledge and skills across different types of data literacy...
Analytical skills

- 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
“Hacking” skills: from data sources to patterns, relationships, and trends

Data sources → "Hacking" → Analytical tools → Knowledge → Data products

Data lifecycle: Research lifecycle
Data management skills

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

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

- 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

- Operation systems
- Repository systems
- Database systems
- Programming languages
- Encoding languages
- Specialized programming

Communication skills

- Interviews
- “Ice breaking”
- Community building
- Institutionalization
- Stakeholder buy-in
CAS in Data Science program at

Core courses (3 credits each):
- Database
- Applied Data Science

Elective (any 3 courses):

**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
Data literacy skills: an information science version.
References


