Data Science and What It Means to Library and Information Science

Jian Qin
School of Information Studies
Syracuse University

iSpeaker Series at Sungkyunkwan University
Seoul, Korea, December 8, 2015
Agenda

• What is data science?
• What is a data scientist?
• What areas of library work can benefit from data science?
What is data science?

The whole lifecycle of data from collection to analysis to preservation.

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

What is data science?

Gathering and massaging data to tell its story

“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.”

Science

A systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions.

Data Science

The study of the generalizable extraction of knowledge from data, which involves data and statistics or the systematic study of the organization, properties, and analysis of data and its role in inference, including our confidence in the inference.

Why is data science different from statistics and other existing disciplines?

- Raw material, the “data” part of data science, is increasingly heterogeneous and unstructured and often emanating from networks with complex relationships between the entities.

- Analysis of data requires integration, interpretation, and sense making that is increasingly derived through tools from computer science, linguistics, econometrics, sociology, and other disciplines.

- Data are increasingly generated by computer and for computer consumption, that is, computers increasingly do background work for each other and make decisions automatically.
The “scalability in decision making has become possible because of big data that serves as the raw material for the creation of knowledge.”

Main fields in data science

Methodologies:
- Computer science
- Statistics
- Visualization
- Geospatial data analysis and collection
- Data collection and digitization
- Interdisciplinary effort

Domain applications:
- Astronomy & physics
- Chemistry & material science
- Climate science
- Computational biosciences
- Digital humanities
- Ecoinformatics
- Energy & smart grid
- Neuroscience
- Social sciences
What is a data scientist?

• **Math skills:** Statistics and linear algebra

• **Computing skills:** programming and infrastructure design

• **Able to communicate:** ability to create narratives around their work

• **Ask the right questions:** involves domain knowledge and expertise, coupled with a keen ability to see the problem, see the available data, and match up the two.
Analysis of data problems: Story 1

- **Domain:** Global migration studies
- **What’s involved:** migrants, refugees, detention centers, refugee camps, Asylums, ...
- **Data types:** interview audio recordings, photos, articles, clippings, written notes, ...
- **Analysis software:** Atlas.ti, SPSS

---

**We’ve got a problem**

- **Researcher:** How to use Atlas.ti?
- **Data scientist:** What data do you have?
- **Data scientist:** How do you collect them?
- **Data scientist:** What do you do with the data?

---

**Bottleneck problem:**
- difficulty in finding the data by person, interview, and related artifacts and in transforming the data into analysis software
Analysis of data problems: story 2

- **Domain:** Thermochronology and tectonics
- **Data types:** Excel data files (lots of them), spectrum and microscopic images, annotations
- **Analysis:** modeling 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

What is involved: workflows in a research lifecycle

- Need tools and standards to streamline the project data flow to repository
- Data repository contains datasets that are
  - Verified
  - Organized
  - Formatted
  - Aggregated
  - Categorized
  - Linked
  - Searchable
  - Reusable

Rock samples

Test and observe samples by selected methods

Lab tests and analysis

Record test data
Annotate images
Annotate methods

Data files
Analysis of data problem: story 3

- **Domain**: collaboration networks in a data repository
- **What’s involved**: metadata describing DNA sequences
- **Data types**: semi-structured data in plain text format
- **Analysis**: identify entities and relationships, build the data into a database for querying and extraction
- **Bottleneck problems**:
  - Extremely large data sets with multiple entities, which makes manual processing impossible
  - Disambiguation of author names and correctly linking between entities
Analysis of data problems

Analysis of data problems is an analysis of domain data, requirements, and workflows that will lead to the development of solutions.
Skills required to perform analysis of domain data problems

- 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
Example 1: modeling research data for gravitational wave research

1. Understand research lifecycle
2. Workflows: steps and relationships
3. Data flows: what goes in and out at which step
4. Entities and attributes, relationships
5. Researcher’s practice and habits in documenting and managing data
Example 2: asking the right question in mining metadata

Metadata describing datasets is big data that can be used to study:

- Collaboration networks
- Scholarly communication patterns
- Research frontiers and trends
- Knowledge transfer
- Research impact assessment
What areas of library work can benefit from data science?
Data services and data-driven services

Library

Data services that support research, learning, and policy making (external)

Data discovery
Data consulting
Data literacy training
Data collection
Data mining
Data integration

Data-driven services that support library planning, management, and evaluation (internal)
Data-drive organization

- Consumer internet companies
  - Google, Amazon, Facebook, LinkedIn
- Brick-mortar companies:
  - Walmart, UPS, FedEx, GE

“A data-driven organization acquires, processes, and leverage data in a timely fashion to create efficiencies, iterate on and develop new products, and navigate the competitive landscape...”

Is your library (company, research center, etc.) a data-driven organization?

Data curation

“the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for reuse over time, and this new field includes authentication, archiving, management, preservation, retrieval, and representation.”

–UIUC GSLIS
Data collection

• Build data collections through
  • Institutional repositories
  • Community repositories
  • Developing tools for researchers to submit, manage, preserve, and discover data

• Develop data collections
  • Specialized
  • Analysis-ready
  • Reusable
  • Actionable
    • For library service planning, decision making, and evaluation
    • To support policy making, research, and learning
Data discovery

• Complex data landscape:
  • International, national, regional
  • Disciplinary, community
  • Open access vs. closed access

• Data sources for various purposes:
  • Utility data sources: open, reusable
  • Census data: open, but need additional processing/meshing to reach the analysis-ready state
  • Government data: open, reusable, but require additional processing
  • Disciplinary research data: access varies, require special knowledge to access and use

Data involving human subjects are under strict control by law and often follow additional compliance

12/8/2015 iSpeaker Series at Sungkyunkwan University, Seoul, Korea
Data consulting

- Search, locate, and verify data for particular research purposes
- Plan, design, and implement data curation and/or data analysis projects
- Provide training and consulting for statistical methods and tools
Data mining

• Using internal data:
  • Users, uses, expenses, collections, staff
  • Goal: improve efficiencies and service quality

• Using external data:
  • Trends and indicators in scholarly communication, technology, economy, and culture
  • Goal: adjust current services and plan for new services
Data integration

Data integration is the combination of technical and business processes used to combine data from disparate sources into meaningful and valuable information.


A process of understanding, cleansing, monitoring, transforming, and delivering data, which offers opportunities to develop data products as an infrastructure for research, learning, policymaking, and decision making.
A home buyer’s information integration

What houses for sale under $250K have at least 2 bathrooms, 2 bedrooms, a nearby school ranking in the upper third, in a neighborhood with below-average crime rate and diverse population?

Information integration

- Realtor
- School rankings
- Crime rate
- Demographics
Research data integration

Diabetes data and trends—Country level estimates:

Diabetes Data & Trends home page:
http://apps.nccd.cdc.gov/ddtstrs/default.aspx
Summary

• Data science is not a new discipline, but rather, a new way of utilizing data, methods, and tools to ask the right questions in solving problems.

• Practicing data science requires strong skills in math, computing, interpersonal communication, and asking the right questions.

• Libraries are at a strategic position in practicing data science. How to leverage this position relies on the
  • vision
  • courage of risk taking
  • knowledge of data science and related topics
  • careful planning
  • collaboration
Thank you!

Questions?