Announcements 2/24

• NO class Wed March 1 and Friday March 3. Please work on your research paper draft.

• **Op-Ed final due today.** Hardcopy on Fran’s desk

• **Paper draft due on March 10.**

• Discussion article for Friday March 10. Please read:

<table>
<thead>
<tr>
<th>Wednesday Section</th>
<th>Friday lecture</th>
<th>First Half of Class</th>
<th>Second Half of Class</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 25: NO class</td>
<td>January 27</td>
<td>L2: Big data applications / Data and the election; Data and Target; Discussion</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>February 1: 6 presentations</td>
<td>February 3</td>
<td>L3: Data and Health / PDB, Precision Medicine; Discussion</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>February 8: NO class</td>
<td>February 10</td>
<td>L4: Data and Science / Earthquakes, LHC; Paper Instructions</td>
<td>4 Presentations</td>
<td>Op-Ed Draft Due</td>
</tr>
<tr>
<td>February 15: 6 presentations</td>
<td>February 17</td>
<td>L5: Data Cyberinfrastructure; Discussion</td>
<td>4 Presentations</td>
<td>Op-Ed Draft Back</td>
</tr>
<tr>
<td>February 22: 6 presentations</td>
<td>February 24</td>
<td>L6: Data Stewardship and Data Preservation; Discussion</td>
<td>4 presentations</td>
<td>Op-Ed Final Due</td>
</tr>
<tr>
<td>March 1: NO class</td>
<td>March 3</td>
<td>NO class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 8: 6 presentations</td>
<td>March 10</td>
<td>L7: Data Futures – Internet of Things; Discussion</td>
<td>4 presentations</td>
<td>Paper Draft Due</td>
</tr>
<tr>
<td>March 15: Spring Break</td>
<td>March 17</td>
<td>Spring Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 22: NO class</td>
<td>March 24</td>
<td>L8: Data rights and policy / U.S. and EU; Discussion</td>
<td>4 presentations</td>
<td></td>
</tr>
<tr>
<td>March 29: 6 presentations</td>
<td>March 31</td>
<td>Op-Ed Pecha-Kucha</td>
<td>Paper Draft Back</td>
<td></td>
</tr>
<tr>
<td>April 5: NO class</td>
<td>April 7</td>
<td>NO class</td>
<td></td>
<td></td>
</tr>
<tr>
<td>April 12: 4 presentations</td>
<td>April 14</td>
<td>Hilary Mason Guest Lecture</td>
<td>4 presentations</td>
<td>Final Paper Due</td>
</tr>
<tr>
<td>April 19: 4 presentations</td>
<td>April 21</td>
<td>L9: Data and Ethics; Discussion</td>
<td>4 presentations</td>
<td></td>
</tr>
<tr>
<td>April 26: 6 presentations</td>
<td>April 28</td>
<td>Paper Pecha-Kucha</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Today (2/24/17)

• Lecture 6: Data Stewardship and Preservation

• Discussion

• Break

• 4 Student Presentations
Lecture 6: Data Stewardship and Preservation
Why should we care about digital stewardship and preservation?

• In the data ecosystem, data provides a “natural resource”

• Critical to *accessing and utilizing* that resource is the *stewardship* of data today

• Critical to *sustaining* that resource is the *preservation* of data over time

• Data stewardship and preservation provides a stable “home” for data that is a pre-requisite for data-driven applications, innovation, commerce, research

  – “Homeless” data ceases to exist ...
Why data stewardship and preservation matter

http://youtu.be/N2zK3sAtr-4
What is data stewardship?

• *Environmental stewardship* refers to responsible use and protection of the natural environment through conservation and sustainable practices. *Wikipedia*

• We can define *data stewardship* similarly as the responsible use and protection of digital assets through management, infrastructure support, and sustainable practices.
“When” is data stewardship?

- Data stewardship and preservation important focus all throughout the “data life cycle”

{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

Acquire
Create, capture gather from:
- Lab
- Fieldwork
- Surveys
- Devices
- Simulations
- etc

Clean
- Organize
- Filter
- Annotate
- Clean

Use / Reuse
- Analyze
- Mine
- Model
- Derive ++data
- Visualize
- Decide
- Act
- Drive:
  - Devices
  - Instruments
  - Computers

Publish
- Share
  - Data
  - Code
  - Workflows
  - Disseminate
  - Aggregate
  - Collect
  - Create portals, databases, etc
  - Couple with literature

Preserve/Destroy
- Store to:
  - Preserve
  - Replicate
  - Ignore
  - Subset, compress
  - Index
  - Curate
  - Destroy

Fran Berman, Data and Society, CSCI 4370/6370
Data stewardship promotes access and use of digital data *today* and data preservation promotes the access and use of digital data *tomorrow*.

**Key Questions:**

- What should we preserve?
- Who should access it?
- How should we preserve it?
- Who is responsible for stewardship?
- How do we pay for it?
What should we preserve?

Data that is commonly of value ...

• Administrative / “business” data
  – Increasing business automation in 21st century. Stewardship part of the cost of doing business
  – Regulation and policy often mandates preservation

• Client / customer data
  – Competitive advantage in the private sector. Stewardship and preservation required to capitalize on this asset

• Research / public data
  – Access increasingly expected to support research, innovation, public information
Value is in the eye of the beholder …

Digital information we* want to keep over the long-term:

**We = “Society” / Public Sector**
- Official and historically valuable data (Census information, presidential emails, Shoah Collection, etc.)

**We = Private Sector**
- Administrative / business data (organizational utility); Client/customer data (competitive advantage)

**We = Research Community**
- Data important to projects, colleagues, domains, current and future efforts

**We = Me**
- My financial data, digital photos of my kids’ graduations, etc.

---

• **Many kinds of valued research data**
  - Data that is in demand by researchers for replication or reuse
  - Data that is mandated to be preserved by policy or regulation
  - Data that is expected to be preserved as part of good scholarly practice
  - Data that is highly cited
  - Data for which value accrues over time
  - Data that underlies assessment reports
  - Data that is costly to reproduce or cannot be reproduced
  - Data that is timely, costly or difficult to obtain, etc.
**Sarbanes-Oxley (Public Accounting Reform and Investor Protection Act of 2002)**

*Applies to all U.S. public company boards, management, and public accounting firms*

**Includes electronic records** (correspondence, work papers, memoranda, etc.) that are created, sent, or received in connection with an audit or a review.

1. “Don’t forget that email and instant messaging are business records …
4. Don’t assume that the retention requirement ...is ...7 years. ...most lawyers that understand information retention agree that business records need to be kept indefinitely.

*Kevin Beaver, “Thirteen Data Retention Mistakes to Avoid”*

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Retention Requirement</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarbanes-Oxley</td>
<td>Auditors must retain relevant data for at least 7 years</td>
<td>Fines to $5M and 20 years in prison</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Retain patient data for 6 years</td>
<td>$250K fine and up to 10 years in prison</td>
</tr>
<tr>
<td>Gramm-Leach-Baily</td>
<td>Ensure confidentiality of customer financial information</td>
<td>Up to $500K and 10 years in prison</td>
</tr>
<tr>
<td>SEC 17a</td>
<td>Broker data retention for 3-6 years. Some require longer retention</td>
<td>Variable based on violation</td>
</tr>
<tr>
<td>OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)</td>
<td>“a three year period is the minimum amount of time that research data should be kept by the grantee”</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>
What do we *have* to preserve?

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Retention Requirement</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarbanes-Oxley</td>
<td>Auditors must retain relevant data for at least 7 years</td>
<td>Fines to $5M and 20 years in prison</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Retain patient data for 6 years</td>
<td>$250K fine and up to 10 years in prison</td>
</tr>
<tr>
<td>Gramm-Leach-Baily</td>
<td>Ensure confidentiality of customer financial information</td>
<td>Up to $500K and 10 years in prison</td>
</tr>
<tr>
<td>SEC 17a</td>
<td>Broker data retention for 3-6 years. Some require longer retention</td>
<td>Variable based on violation</td>
</tr>
<tr>
<td>OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)</td>
<td>“a three year period is the minimum amount of time that research data should be kept by the grantee”</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>

**HIPAA (Health Insurance Portability and Accountability Act)**

- Applies to health information created or maintained by health care providers “who engage in certain *electronic transactions*, health plans, and health care clearinghouses” [www.hipaa.org]
- Title II: Requires HHS to create rules and standards for the use and dissemination of health care information
- Healthcare providers must retain healthcare records for a period of **not less than 6 years.**
What do we have to preserve?

- The U.S. Office of Management and Budget requires that **federally funded research data**, supporting documentation, scientific notebooks, financial records, etc. **be maintained by the grantee (typically institution) for 3+ years**

- Sponsored research grants: research data typically owned by the grantee (institution).

- Some federal and private sector contracts require that sponsors be granted ownership or some/all rights to data
  - Many institutions retain rights to data for research and education

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Retention Requirement</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarbanes-Oxley</td>
<td>Auditors must retain relevant data for at least 7 years</td>
<td>Fines to $5M and 20 years in prison</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Retain patient data for 6 years</td>
<td>$250K fine and up to 10 years in prison</td>
</tr>
<tr>
<td>Gramm-Leach-Baily</td>
<td>Ensure confidentiality of customer financial information</td>
<td>Up to $500K and 10 years in prison</td>
</tr>
<tr>
<td>SEC 17a</td>
<td>Broker data retention for 3-6 years. Some require longer retention</td>
<td>Variable based on violation</td>
</tr>
<tr>
<td>OMB Circular A-110 / CFR Part 215</td>
<td>“a three year period is the minimum amount of time that research data should be kept by the grantee” [grantee = institution]</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>
Public access expectations for sponsored research data and publications

• February 2013 OSTP Memo focus: Access to data and publications from federally funded research

• Federal R&D agencies asked to develop
  – Strategy for capitalizing on what exists and fostering public-private partnerships with scientific journals
  – Strategy for increasing / enhancing discoverability, access, dissemination, stewardship, preservation
  – Approach for measuring and enforcing compliance
Data stewardship promotes access and use of digital data *today* and data preservation promotes the access and use of digital data *tomorrow*.

Key Questions:

- What should we preserve?
- Who should access it?
- **How should we preserve it?**
- Who is responsible for stewardship?
- How do we pay for it?

Print media provides easy access for long periods of time but is hard to data-mine

Digital media is easier to data-mine but requires management of evolution of media and resource planning over time
Good Practice in Data Stewardship and Preservation

- **Replication** – make multiple copies of data and store some off-site
- **Refreshing** – transfer of data between “old” versions of the same storage to new versions of the same storage to reduce bitrot and alteration of data
- **Integrity assurance** – incorporate sufficient metadata, provenance information, checksums and other techniques to ensure the integrity of data systems, content, and context
- **Forward planning / migration** – pro-actively plan and transition data to ensure sustainability across multiple technology generations
- **Sustainable economic support** – develop business model to stably support data preservation efforts, technologies, and staffing over time
- **Compliance** – Ensure that preservation systems comply with current regulations, policies, and penalties that pertain to data
- **Security and disaster planning** – ensure appropriate levels of system security to demonstrate good practice and plan ahead for recovery from disaster scenarios

**Why are 3 copies used as best practice?**

- Approach comes from Lamport, Shostak, and Pease’s solution to the Byzantine General’s Problem
  - Method for agreement on a battle plan for a group of Byzantine generals communicating only by messenger
  - Analogous to reliable computer systems with malfunctioning components
- Solution: When generals can send unforgeable signed messages to one another, the minimum number required for agreement is 3.

*Paper in lecture references*
Key Players in research and public data stewardship and preservation: Librarians and Archivists

- **Archives** are the non-current records of individuals, groups, institutions, and governments that contain information of enduring value. The primary task of the **archivist** is to establish and maintain control, both physical and intellectual, over records of enduring value and ensure their content accessible for posterity.

- A **library** is an organized collection of sources of information and similar resources, made accessible to a defined community for reference or borrowing. The primary task of the **librarian** is to manage the information for discovery and use, and assist individuals in accessing and using library information.

- **Traditional professional skills expanded with key areas from information science:**
  - Knowledge of information architecture and information management systems
  - Markup languages, metadata formats, file types
  - Digitization, database management
  - Standards, policy and regulation
  - Data integrity, security, etc.
One size doesn’t fit all: Spectrum of research data libraries, archives, domain repositories

Fran Berman, Data and Society, CSCI 4370/6370
Data Stewardship and Preservation Glossary

- **Metadata** – Documentation relating to data content, structure, provenance (history), and context, “data about the data”
- **Identifier** – unique label used to reference an object or record
- **Curation** – maintaining and adding value to a trusted body of digital information for current and future use
- **Appraisal** – evaluation and selection of digital material for long-term curation and preservation
- **Authentication** – security measure designed to establish the validity of a transmission, message, or originator, or a means of verifying an individual’s authority
- **Ingest** – Controlled or secure transfer of material to an archive, repository, data center, or other custodial environment
- **Integrity** – Condition when data is unchanged from its source and has not been accidently or maliciously modified, altered, or destroyed
- **Digital Rights Management** – use of technologies to control how digital content is used and re-used
The Digital Curation Center’s Digital Data Life Cycle: Digital curation and preservation stages

Image: http://www.dcc.ac.uk/resources/curation-lifecycle-model
Archivist’s Perspective: Open Archival Information system (OAIS) Reference Model

- AIP: Archive Information Packages
- DIP: Dissemination Information Packages
- SIP: Submission Information Packages

Figure: OAIS functional entities, Wikipedia
More About OAIS

- OAIS is an archival approach that promotes preservation for a designated community.
  - “Open” in OAIS: model developed in open forum (not that all information in OAIS archive is unrestricted)

- Mandatory responsibilities for OAIS archives. **OAIS archives must**
  - Negotiate for and **accept appropriate information** from Information Producers
  - **Obtain sufficient control** of the information provided to the level needed to ensure long-term preservation
  - Determine which communities should become the “**Designated Community**” (who will understand the information provided)
  - Ensure that the information to be preserved is **independently understandable** to the Designated Community
  - Follow documented policies and procedures which **ensure that the information is preserved against all reasonable contingencies**, and which **enable the information to be disseminated** as authenticated copies of the original, or as traceable to the original
  - Make the preserved information **available** to the Designated Community

Fran Berman, Data and Society, CSCI 4370/6370
Data stewardship promotes access and use of digital data *today* and data preservation promotes the access and use of digital data *tomorrow*.

**Key Questions:**

- What should we preserve?
- Who should access it?
- How should we preserve it?
- **Who is responsible for stewardship?**
- How do we pay for it?
The Resource Gap – Data economics

Data infrastructure costs increase with usage, stewardship and access requirements, perceived value.

Greater costs at the extremes (including “big” data) ...

Fran Berman, Data and Society, CSCI 4370/6370
**What are we paying for?**

**Data infrastructure components may include**

- Maintenance and upkeep
- Software tools and packages
- Utilities (power, cooling)
- Space
- Networking
- Security and failover systems
- People (expertise, help, infrastructure management, development)
- Training, documentation
- Monitoring, auditing
- Reporting costs
- Costs of compliance with regulation, policy, etc. ...

**Resources and Resource Refresh**

**SDSC Data Storage Growth ‘97-’09**

- **Most valuable data replicated**
- **As research collections increase, storage capacity must stay ahead of demand**

Information courtesy of Richard Moore, SDSC
Current Economic Support Models for Digital Research Data

- **Subscription**
- **Fee for service**
- **Advertisement**
- **Crowd-sourcing, philanthropy**
- **Federal grants**

Fran Berman, Data and Society, CSCI 4370/6370
Who is responsible for data stewardship and preservation?

- The “Free Rider” non-solution:
  - (Economics) **Free rider** refers to someone who benefits from resources, goods, or services without paying for the cost of the benefit.
  - Free riding may be considered as a **free rider problem** when it leads to under-provision of goods or services, or when it leads to overuse or degradation of a common property resource. [Wikipedia]

- **Free Rider problem for data**: someone else (Google, the Govt., libraries, my institution, data creators, etc.) should pay for data stewardship / preservation, but not me
Aligning the Stakeholders

- Many Stakeholders in digital access and preservation
  - Stakeholders who benefit from use of the preserved asset
  - Stakeholders who select what to preserve
  - Stakeholders who own / have rights to the asset
  - Stakeholders who preserve the asset
  - Stakeholders who pay

- The greater the alignment between key stakeholder groups, the better the prospect for sustainable preservation
Sustainability

Economic sustainability for digital information requires

- **Recognition of the benefits** of long-term access and preservation
- **Incentives** for decision-makers to act
- **Means of selecting “valued” information** for long-term preservation
- **Mechanisms to support ongoing, efficient allocation of resources**
- **Appropriate organization and governance** of preservation and access activities

*From Blue Ribbon Task Force Interim Report*

**Digital sustainability** concentrates less on the solution and technology and more on building an infrastructure and approach that is flexible with an emphasis on interoperability, continued maintenance and continuous development.

**Digital sustainability** incorporates activities in the present that will facilitate access and availability in the future. [Wikipedia]
Lecture 2 Sources

- Digital Curation Center Data Life Cycle http://www.dcc.ac.uk/resources/curation-lifecycle-model
- OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf
- LOCKSS http://www.lockss.org/
- “Who will pay for Public Access to Research Data?”, Science Magazine, August 9 (on http://www.cs.rpi.edu/~bermaf/)

Fran Berman, Data and Society, CSCI 4370/6370
Discussion

• “Can the Internet be archived?”, The New Yorker, http://www.newyorker.com/magazine/2015/01/26/cobweb
Presentations for March 8

March 8


Presentations for March 10

- March 10
No class until March 8

• *March 10 Lecture:* Data Futures: Internet of Things; Discussion

• *Read for March 10 Discussion:*

Break
Presentations
Presentations for February 24

Feb 24 (data stewardship and preservation)


