Announcements 2/8

• **No Wednesday class** February 13.

• **Group topic summaries due at the beginning of class today** (one summary for the whole group).

• You should be working on your op-ed draft. If you have any questions about the topic, ask Fran in office hours.
<table>
<thead>
<tr>
<th>Wednesday Section</th>
<th>Friday Lecture (first half)</th>
<th>Second half of class</th>
<th>Assts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 9: NO CLASS</td>
<td>January 11: INTRO – DATA AND SOCIETY</td>
<td>Fran presentation demo</td>
<td></td>
</tr>
<tr>
<td>January 16: NO CLASS</td>
<td>January 18: BIG DATA 1; Topic groups / Topic materials information</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>January 23: Student presentations</td>
<td>January 25: BIG DATA 2</td>
<td>Student presentations</td>
<td>Op-Ed instructions</td>
</tr>
<tr>
<td>January 30: NO CLASS</td>
<td>February 1: DATA AND SCIENCE</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>February 6: NO CLASS</td>
<td>February 8: DATA STEWARDSHIP AND PRESERVATION</td>
<td>Student presentations</td>
<td>Group Topics due</td>
</tr>
<tr>
<td>February 13: NO CLASS</td>
<td>February 15: INTERNET OF THINGS</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>February 20: Student presentations</td>
<td>February 22: DATA AND PRIVACY / FOUNDATIONS</td>
<td>Student presentations</td>
<td>Op-Ed Drafts due</td>
</tr>
<tr>
<td>February 27: NO CLASS</td>
<td>March 1: DATA AND PRIVACY / POLICY AND REGULATION</td>
<td>Student presentations</td>
<td>Briefing instructions</td>
</tr>
<tr>
<td>March 6: Spring Break</td>
<td>March 8: Spring Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 13: Student presentations</td>
<td>March 15: DATA AND ENTERTAINMENT [ANDY MALTZ]</td>
<td>Student presentations</td>
<td>Op-Ed Drafts Returned</td>
</tr>
<tr>
<td>March 20: TOPICS PRESENTATIONS 1</td>
<td>March 22: DATA AND DATING</td>
<td>Student presentations</td>
<td>Topic Reports 1 due</td>
</tr>
<tr>
<td>March 27: Student presentations</td>
<td>March 29: DIGITAL RIGHTS 1</td>
<td>Student presentations</td>
<td>Op-Ed Finals due</td>
</tr>
<tr>
<td>April 3: NO CLASS</td>
<td>April 5: DIGITAL RIGHTS 2</td>
<td>Student presentations</td>
<td>Briefings due</td>
</tr>
<tr>
<td>April 10: Student presentations</td>
<td>April 12: DATA AND ETHICS</td>
<td>Student presentations</td>
<td>Op-Ed Finals returned,</td>
</tr>
<tr>
<td>April 17: Student presentations</td>
<td>April 19: CAREERS IN TECH [KATHY PHAM?]</td>
<td>Student presentations</td>
<td>Topic Reports 2 due</td>
</tr>
<tr>
<td>April 24: Student presentations</td>
<td>April 26: TOPICS PRESENTATIONS 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Stewardship, and Preservation
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.
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 ensure a stable “home” for data used for data-driven applications, innovation, commerce, research, etc.
  - “Homeless” data cease to exist ...
Stewardship and Preservation – why it matters ... (4:40 min)

http://youtu.be/N2zK3sAtr-4
“When” is data stewardship?

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

[Diagram showing the data life cycle stages: Acquire, Clean, Use/Reuse, Publish, Preserve/Destroy]
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 data should we preserve? Who decides?
- Who is responsible for data?
- How do we pay for stewardship and preservation infrastructure?
What should we preserve?

Data that is commonly of value ...

• Administrative / “business” data
  – Regulation and policy may mandate 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, reproducibility, 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.

Fran Berman, Data a

• 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

---

<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>

---

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”
http://searchdatamanagement.techtarget.com/news/article/0,289142,sid91_qci1186910,00.html

Table information partly based on “Data Retention – More Value, Less Filling”, John Murphy, http://www.tdan.com/view-articles/5222
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 (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” [grantee = institution]</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>

---

Fran Berman, Data and Society, CSCI 4370/6370
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
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**
Who is preserving your data? Common approaches to digital preservation

• **Personal data you want to keep**: You are preserving your data (on your own gear or via a service). You are responsible for ensuring that data is sustained over time (through fees, hardware migration, etc.)

• **Business data**: Companies determine what is valuable to them and include data preservation as part of their own infrastructure. Choices are made based on business priorities and regulation on what to retain and what to discard.

• **Government data**: The government is required to preserve many different kinds of data based on what is considered value (e.g. through NARA, the Library of Congress, GAO, agencies, etc.). You do not have access to all of it.

• **Research data**: Researchers preserve their data at their discretion if it is valuable, or required by funding sponsors, their institutions or publication. Where that data goes and who is responsible for it is often left up to the researcher.
Professionals in 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.
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
Data economics: 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

• In the private sector and public sector, alignment ensures preservation and access of valued data
  – Private sector: Retain data because it is critical for products and provides a competitive advantage
  – Public sector: Data retained for archival and historical purposes
Poor stakeholder alignment in academia: How to pay 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
Costs of data stewardship and preservation

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

Greater costs at the extremes (including “big” data) ...
Software, hardware, human, compliance costs

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
Data Stewardship and Preservation Today

- **Private data** – data still be retained / used / bought / sold / shared by companies for competitive advantage and to maximize profit
  - Terms and conditions often give them permission
  - Limiting legal barriers on data privacy highly dependent on where you are and what kind of data

- **Academic research data** – more recognition of value of preservation.
  - Business models still often insufficient
  - Value of libraries and community data repositories still not recognized by stakeholders as they should be as important players for the community
  - Researchers becoming better informed

- **Government data**
  - Good data preservation mechanisms still in place
  - Political winds may impact what data remains preserved and what doesn’t (e.g. EPA data)
Lecture 5 References (not already on slides)


• Digital Curation Center Data Life Cycle http://www.dcc.ac.uk/resources/curation-lifecycle-model


• OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf


• “Who will pay for Public Access to Research Data?”, Science Magazine, August 9 (on http://www.cs.rpi.edu/~bermaf/)
Break
Presentations
Presentation articles for February 15

• “Buoys, fisheries and aquaculture join the internet of things,” the Japan Times, https://www.japantimes.co.jp/life/2019/01/05/food/buoys-fisheries-aquaculture-join-internet-things/#.XDKUCVxKg2w [Yuming Z.]


Presentation Articles for February 20


• “This is how hackers can take down our critical energy systems through the Internet,” ZDNet, https://www.zdnet.com/article/this-is-how-hackers-can-take-down-our-core-water-energy-systems/ [Lexie Z.]


• “Home items are getting smarter and creepier, like it or not,” The Tribune, http://www.tribtown.com/2019/01/07/us-tec-gadget-show-smart-home-no-escape/ [Peter K.]
Presentation Articles for February 22


Presentation articles for Today


• “Inside the Wayback Machine, the internet’s time capsule”, the Hustle, https://thehustle.co/inside-wayback-machine-internet-archive  (Andrew L.)