Today’s Class

• Assignment 4 due by 11:59 p.m. on 12/2
• Lecture
• Presentations
Assignment 4: Personal Essay 4/Op-Ed

Instructions

• Pick **ONE** of the following assignments. **LET FRAN KNOW WHICH ONE YOU ARE DOING.** Assignments are due by 11:59 p.m. on 12/2.

  – **Option 1: Personal Essay #4** -- You have 15 minutes with the President Elect. Tell him what you think the administration’s highest data-related priority should be and what specific thing the U.S. government should do. *(8 points, usual grading rubric)*

        OR

  – **Option 2: Op-Ed** -- Choose a data-related issue. Write an op-ed that persuades Fran of your thesis. *(14 points, usual grading rubric)*

• **Note:** Both options contribute up to the last 8 points of your grade. If you choose the Op-Ed option, up to 6 additional points of extra credit will be added to your total (depending on what you get)
<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Speaker</th>
<th>Date</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-31</td>
<td>Introduction</td>
<td>Fran</td>
<td>9-3</td>
<td>The Data-driven world</td>
<td>Fran</td>
</tr>
<tr>
<td>9-10</td>
<td>Data and COVID-19 - models</td>
<td>Fran</td>
<td>9-14</td>
<td>Data and COVID-19 – contact tracing</td>
<td>Fran</td>
</tr>
<tr>
<td>9-17</td>
<td>Data and the Opioid Crisis</td>
<td>Liz Chiarello</td>
<td>9-21</td>
<td>Data and Privacy - Intro</td>
<td>Fran</td>
</tr>
<tr>
<td>9-24</td>
<td>Data and Privacy – Differential Privacy and the Census</td>
<td>Fran</td>
<td>9-28</td>
<td>Data and Privacy – Anonymity</td>
<td>Fran</td>
</tr>
<tr>
<td>10-1</td>
<td>Data and Privacy - Law</td>
<td>Fran</td>
<td>10-5</td>
<td>Digital rights in the EU and China</td>
<td>Fran</td>
</tr>
<tr>
<td>10-8</td>
<td>Data and Elections 1</td>
<td>Fran</td>
<td>10-12</td>
<td>NO CLASS – Columbus / Indigenous Peoples’ Day</td>
<td>Fran</td>
</tr>
<tr>
<td>10-15</td>
<td>Data and Elections 2</td>
<td>Fran</td>
<td>10-19</td>
<td>Data and Elections 3</td>
<td>Fran</td>
</tr>
<tr>
<td>10-22</td>
<td>Data and Elections 4</td>
<td>Todd Rogers</td>
<td>10-26</td>
<td>Data and Research 1 / PERSONAL ESSAY 3</td>
<td>Fran</td>
</tr>
<tr>
<td>10-29</td>
<td>Data and Research 2</td>
<td>Josh Greenberg</td>
<td>11-2</td>
<td>Data and Discrimination 1 / PERSONAL ESSAY 3 DUE NOV 4</td>
<td>Fran</td>
</tr>
<tr>
<td>11-5</td>
<td>Data and Discrimination 2 / BRIEFING (TEAMS OF 2)</td>
<td>Fran</td>
<td>11-9</td>
<td>Data and the IoT 1</td>
<td>Fran</td>
</tr>
<tr>
<td>11-12</td>
<td>Data and the IoT 2</td>
<td>Fran</td>
<td>11-16</td>
<td>Data and Ethics</td>
<td>Fran</td>
</tr>
<tr>
<td>11-23</td>
<td>Cybersecurity / BRIEFING DUE 11/22 / PERSONAL ESSAY 4 or OP-ED</td>
<td>Bruce Schneier</td>
<td>12-3</td>
<td>Data Science / PERSONAL ESSAY OR OP-ED DUE 11/2</td>
<td>Fran</td>
</tr>
<tr>
<td>11-30</td>
<td>Data stewardship and preservation</td>
<td>Fran</td>
<td>12-10</td>
<td>Wrap-up</td>
<td>Fran</td>
</tr>
<tr>
<td>12-7</td>
<td>Data Careers</td>
<td>Kathy Pham</td>
<td></td>
<td></td>
<td>Fran</td>
</tr>
</tbody>
</table>
Reading for next time (12/3)

• What data scientists really do, according to 35 data scientists, Harvard Business Review

Lecture – Data stewardship and preservation

• Data Stewardship and Preservation
• The Internet Archive
Data stewardship promotes access and use of digital data *today* and data preservation promotes the access and use of digital data *tomorrow*.

- Which data should we preserve?
- Who maintains and preserves it?
- Who preserves the Internet?

Source: IDC, 2008
Who is preserving data?

- **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 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 valuable (e.g. through NARA, the Library of Congress, GAO, agencies, NSA, etc.). You do not have access to all of it.

- **Research data**: Researchers preserve their data at their discretion if it is valuable, or as required by funding sponsors, their institutions, or publications. Where that data goes and who is responsible for it is often left up to the researcher.
Some data must be preserved by law

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Type of data</th>
<th>Retention Requirement</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarbanes-Oxley</td>
<td><strong>Business data</strong> for U.S. public company boards, management, and public accounting firms</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><strong>Health data</strong> created or maintained by health care providers</td>
<td>Retain patient data for 6 years</td>
<td>$250K fine and up to 10 years in prison</td>
</tr>
<tr>
<td>OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)</td>
<td><strong>Federally funded research data</strong> – including supporting documentation, scientific notebooks, financial records, etc. be maintained by the grantee (typically institution)</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>
Data stewardship and preservation should be planned from the start

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

---

**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

---

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

---

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

Why data stewardship and preservation matter (4:40 min)

http://youtu.be/N2zK3sAtr-4
In research world, stewardship and preservation is uneven

The Arapadopsis Information Resource is a community database for plant biology. Originally funded by NSF for many years, the group launched a non-profit (Phoenix Bioinformatics) and now offers this DB and others through subscription.

Dataverse provides open source research data repository software, which can be used by researchers, journals, institutions and developers. Many institutions (e.g. Harvard) provide a local Dataverse used for stewardship and preservation of eligible data.
Best 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** – create 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
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 is not free

Costs of stewardship and preservation 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

Fran Berman, Data and Society, CSCI 4370, Information courtesy of Richard Moore, SDSC
Preserving the Internet: The Internet Archive

- Internet Archive is a digital library whose mission is “universal access to all knowledge”
  - Non-profit
  - Started by Brewster Kahle
- Free public access to collections of digitized materials, including websites.
- Internet Archive currently holds > 48 PB of materials including 20+M books and texts, 6+M movies and videos, 600K SW programs, 15M audio files and 480+B web pages in the Wayback Machine.
How does the Internet Archive preserve the web?

- Web crawlers work to preserve as much of the public web as possible. Webpages stored in the **Wayback Machine**.

- Users can view archived webpages.
  - Provides public access to code, images, source code from websites that may no longer exist or have been updated
  - About half of website hyperlinks included

- Not everything is crawled, only public Internet. Website owners can opt out.

- Frequency of website capture also varies per website, based on which crawl list(s) it’s on.
Challenges in preservation

• In October, 2020, the Internet Archive began to provide fact checks and context for Wayback Machine webpages
  – Idea is not to store misinformation without labeling it as such

We would like to acknowledge the hard work of the organizations we are building upon in order to provide context for archived web pages: FactCheck.org, Check Your Fact, Lead Stories, Politifact, Washington Post Fact-Checker, AP News Fact Check, USA Today Fact Check, Graphika, Stanford Internet Observatory, and Our.news.
Lecture Sources


• Internet Archive, https://en.wikipedia.org/wiki/Internet_Archive

• Internet Archive Website, https://archive.org/
Presentations
These are the last scheduled presentations. Make sure that you have signed up for two throughout the semester

• Presentations for 12/3
  – “Cooling our insatiable demand for data”, Science, https://science.sciencemag.org/content/370/6518/783

• Presentations for 12/7
• Presentations for Today
