Introduction to the course, Introduction to digital data, Op-ed instructions, Presentation model
Announcements 1/19

• Welcome!

• Please sign attendance sheet each time you are here (your participation grade depends partly on attendance).
  – If you are on the waiting list and trying to get in the class, please do the same.

• If you decide to drop the class, please let me know (berm@rpi.edu) and I will let in someone on the waiting list.

• No Wednesday class January 24.

• Discussion article for next week (Friday, January 26). Please read:
Welcome to Data and Society, Low 3039
CSCI 6370 (Grads) / 4370 (Undergrads)
ITWS 6960 (Grads) / 4960 (Undergrads)

• Professor: Dr. Fran Berman

• Office: AE 218, 276-3794

• Office Hours: Friday 1-2 or by appointment (send email to bermaf@rpi.edu)

• Course website (linked off Fran’s RPI web page): http://www.cs.rpi.edu/~bermaf/Data%20Course%202018/Data.html
Today (1/19/17)

• Why Data and Society?

• Intro – about this course
  – Syllabus and grading expectations
  – Learning objectives and expectations
  – Why are you here?

• Lecture 1

• Break

• Op-Ed instructions, draft due February 9

• Data Presentation (Fran)
Data-driven innovation is a priority nationally, internationally, and in all sectors.
There’s more to the data story than technology

Policy and regulatory issues

Workforce evolution

New possibilities for innovation / new challenges for infrastructure

Privacy and rights

New modes of social and community interaction, organization

VR image: https://www.flickr.com/photos/nanpalmero/16237219524
Data and Society – about this course

• This course will provide a broad snapshot of the data-driven world
  – We’ll skim the sea of interesting data stuff, but we won’t / can’t include everything
  – We’ll focus more on societal issues than technical issues
  – The course should provide a complement to the material in the ITWS Data Science, Web Science, Data Analytics and other courses

• The course will be structured to
  – Increase your engagement with material
  – Evolve your professional communication and assessment skills
  – Help you develop as a “data-literate” professional

Course topics:
• Data Applications
• Data Infrastructure and Stewardship
• Data and Society

Guest Speaker this Semester:
• Leslie Mcintosh, Executive Director, Research Data Alliance / US
<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 17 : NO class</td>
<td>January 19</td>
<td>L1: CLASS INTRO AND LOGISTICS</td>
<td>Presentation Model / Op-Ed Instructions</td>
<td>Op-Ed instructions</td>
</tr>
<tr>
<td>January 24: NO class</td>
<td>January 26</td>
<td>L2: BIG DATA 1</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>January 31: NO class</td>
<td>February 2</td>
<td>L3: BIG DATA 2 - IoT</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>February 7: NO class</td>
<td>February 9</td>
<td>L4: DATA AND SCIENCE</td>
<td>4 Presentations</td>
<td>Op-Ed due Feb. 9</td>
</tr>
<tr>
<td>February 14: 5 Presentations</td>
<td>February 16</td>
<td>L5: DATA AND HEALTH / LESLIE McINTOSH GUEST SPEAKER</td>
<td>4 Presentations</td>
<td>Op-Ed drafts returned Feb. 21</td>
</tr>
<tr>
<td>February 21: 5 Presentations</td>
<td>February 23</td>
<td>L6: DATA STEWARDSHIP AND PRESERVATION</td>
<td>4 Presentations</td>
<td>Research Paper instructions</td>
</tr>
<tr>
<td>February 28: 5 Presentations</td>
<td>March 2</td>
<td>L7: DATA INFRASTRUCTURE</td>
<td>4 Presentations</td>
<td>Op-Ed Final due March 2</td>
</tr>
<tr>
<td>March 7 : 5 Presentations</td>
<td>March 9</td>
<td>NO CLASS / PAPER PREPARATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 14: Spring Break</td>
<td>March 16</td>
<td>SPRING BREAK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 21: NO class</td>
<td>March 23</td>
<td>NO CLASS / PAPER PREPARATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 28: 5 Presentations</td>
<td>March 30</td>
<td>L8: DATA RIGHTS, POLICY, REGULATION</td>
<td>4 Presentations</td>
<td>Research Paper due March 28</td>
</tr>
<tr>
<td>April 4: NO class</td>
<td>April 6</td>
<td>L9: DATA AND ETHICS</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>April 11: 5 Presentations</td>
<td>April 13</td>
<td>L10: DATA AND COMMUNICATION</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>April 18: 5 Presentations</td>
<td>April 20</td>
<td>L10: DATA FUTURES</td>
<td>4 Presentations</td>
<td></td>
</tr>
<tr>
<td>April 25: 5 Presentations</td>
<td>April 27</td>
<td>L11: HOT TOPICS / TBD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Course Information
http://www.cs.rpi.edu/~bermaf/Data%20Course%202018/Data.html

- Course website (above) will have all up-to-date information and materials
  - Syllabus may evolve slightly

- Discussion articles and presentation articles will be on the web
  - Embedded reference materials in the lecture will be given by URL. Lectures will be on the web.

- Discussion articles will be assigned prior to the Friday class in which they will be discussed. You are responsible for reading them and for participating in the discussion.
How You’ll be Graded

Grade distribution

- Research Paper: 25%
- Op-Ed Draft: 15%
- Op-Ed Final: 15%
- Presentation 1: 15%
- Presentation 2: 15%
- Participation: 15%

Fran Berman, Data and Society, CSCI 4370/6370
More about grading
(additional grading specifics for Presentations and Op-eds later today)

• **Presentations:**
  – Students are responsible for scheduling their Presentations and ensuring that both are done. Information about Presentations will be given at the end of today’s class.

• **Op-Eds, Papers:**
  – Information about Op-Eds will be given today. Information about the Paper Assignment will be given on February 23.

• **Class participation / attendance:**
  – **Students are expected to attend 18+ out of 20 Wednesday and Friday class meetings.** Attendance will be taken in class.
  – Participation grade: 10% attendance, 5% class participation (Discussions, Questions for speakers, etc.)

• **There will be a slightly different workload for grad students and undergrads**
  – Research paper lengths and technical depth expectations will be different.
  – In writing and presentations, each student will be assessed at a level appropriate to their educational level (undergrad or grad)
## Learning Objectives and Outcomes

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop greater data literacy</td>
<td>Be able to understand and explain the role that data plays as well as its limitations in various areas of research, commerce and modern life.</td>
</tr>
<tr>
<td>Develop critical thinking skills around data</td>
<td>Be able to read, understand, assess, and discuss data-oriented professional and popular publications and articles.</td>
</tr>
<tr>
<td>Develop communication skills around data</td>
<td>Be able to advance an evidence-based argument about data, data cyberinfrastructure and data-oriented efforts to both knowledgeable specialists within the field as well as non-specialists.</td>
</tr>
</tbody>
</table>
Academic Integrity

• Student-teacher relationships are built on trust. For example, students must trust that teachers have made appropriate decisions about the structure and content of the courses they teach, and teachers must trust that the assignments that students turn in are their own. Acts, which violate this trust, undermine the educational process. The Rensselaer Handbook of Student Rights and Responsibilities defines various forms of Academic Dishonesty and you should make yourself familiar with these.

• In this class, all assignments that are turned in for a grade must represent the student’s own work. In cases where help was received, or teamwork was allowed, a notation on the assignment should indicate your collaboration. If references or other materials are used, they should be cited. Submission of any assignment that is in violation of this policy will result in a penalty.

• If found in violation of the academic dishonesty policy, students may be subject to two types of penalties. The instructor administers an academic (grade) penalty, and the student may also enter the Institute judicial process and be subject to such additional sanctions as: warning, probation, suspension, expulsion, and alternative actions as defined in the current Handbook of Student Rights and Responsibilities. If you have any question concerning this policy before submitting an assignment, please ask for clarification.
You get out of the course what you put into it

• *Spend time on the writing and presentations.* Don’t do this at the last minute.
  – Do more than one draft
  – Have someone “red team” the assignments for feedback
  – Talk to Fran during office hours
  – Go to the RPI Center for Communication Practices

• Be prepared to *ask good questions and engage in class.* Read as many of the articles as you can (and *always read the discussion articles*).

• *You do the effort,* Fran = scorekeeper. Focus on building skills as an outcome of this class.

• *Do something hard* that stretches you and for which you can’t take shortcuts in this course (and otherwise)
Why are you here?

1. Name, major, school, grad or undergrad?
2. What do you expect to be doing after you finish your degree?
3. Why did you take this course?
4. What do you hope to get out of this course?
5. What areas / topics in the data landscape are of most interest to you?
6. What is the coolest recent thing you’ve heard about digital data?
Lecture 1: Introduction to Data
Lecture 1 Outline

• Some fundamentals
  – Data basics
  – How much data is there and where does it come from? What does the “global datasphere” look like?

• Data Transformation -- The Information Age
What is Digital Data?

• *Wikipedia*: “Digital data, in information theory and information systems, are discrete, discontinuous representations of information or works, as contrasted with continuous, or analog signals which behave in a continuous manner, or represent information using a continuous function.

• Although digital representations are the subject matter of discrete mathematics, the information represented can be either discrete, such as numbers and letters, or it can be continuous, such as sounds, images, and other measurements.”
Meaning and context increase the impact of data: Data, Information, Knowledge, Wisdom (DIKW)

- **Data** = Qualitative or quantitative values at the lowest level of abstraction
- **Information** = Data and its associated meaning
- **Knowledge** = Theoretical or practical understanding of information
- **Wisdom** = The quality of having experience, knowledge and good judgment

Considerable overlap and many definitions ...

Many articles use digital data and digital information interchangeably. For the most part, we will too.

... Where is the life we have lost in living?
Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information? ...

First recorded instance of DIKW in 1934 poem “Choruses from the Rock” by T.S. Eliot.

Digital data comes from everywhere

Entertainment

Education

Health

Commerce

Physical Infrastructure and Smart Systems

Research

Communication / Community
Transformative Potential of Data:
Massive-scale coordination, inclusion, access

Greater access → Greater participation, “democratization” possible

High quality, on-line education → On-line / on-site education solutions have the potential to transform higher education

Greater transparency, management, monitoring → More measurement, transparency, monitoring possible
Transformative Potential of Data: Emerging Technologies

Exascale computing → more compute and data at all tiers in the Branscomb Pyramid.
New breakthroughs in power and computer architectures required.

Smart Devices, Sensor Networks → More data-enabled devices and approaches drive crowd-sourced, real-time, and other aggregation applications

Information-Driven Analysis → X-informatics and X-analytics enable new targets for data-driven research and decision-making models

Fran Berman, Data and Society, CSCI 4370/6370
The “V’s”: All data are not alike

- **Volume**: amount of data, number of files
- **Velocity**: Rate at which data flows into an organization as well as speed of the “feedback loop” (can the data be where you want it when you want it)
- **Variety**: Diversity of data types and sources (“messiness” of using, combining, managing data)
- **Value**: importance of the data
- **Volatility**: how quickly data changes, how long the data is useful for

**Sources**:
- [http://strata.oreilly.com/2012/01/what-is-big-data.html](http://strata.oreilly.com/2012/01/what-is-big-data.html)

- **Validity**: legitimacy / accuracy of sources
- **Viscosity**: resistance to flow in the volume of data (improved infrastructure, management, and technologies can reduce viscosity)
- **Virality**: how quickly the data is dispersed and shared
- **Variability**: Extent to which data points differ from each other. (commonly used measures of variability: range, mean, variance and standard deviation)
The global datasphere: How much data is there?

- There won’t be an exaflop supercomputer until the end of the decade+.
- We have had exabytes of data for at least 10 years and hit a zettabyte in 2009-2010.
- By 2025, the global datasphere will grow to 163 zettabytes.

![Annual Size of the Global Datashere](image)

Source: IDC’s Data Age 2025 study, sponsored by Seagate, April 2017

Fran Berman, Data and Society, CSCI 4370/6370
Global datasphere doubles every two years

How big is ... (roughly)

- **A byte**: ~ 1 character
- **A megabyte**: ~ 1 small novel
- **A gigabyte**: 50 gigabytes → ~ 1 floor of books
- **A terabyte**: 10 Terabytes → ~ printed collection of the Library of Congress
- **A petabyte**: 2 Petabytes → ~ All US academic research libraries
- **An Exabyte**: 5 exabytes ~ all words ever spoken by human beings recorded in text
- **A Zettabyte**: 42 zettabytes → ~ all words ever spoken by human beings digitized as audio

Source: [http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html](http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html)
Where is data created in the network?

- **Core**: computing datacenters in the enterprise and the cloud
- **Edge**: enterprise hardened computers / appliances that are not in core datacenters (servers in the field, regional datacenters, etc.)
- **Endpoint**: devices on the edge of the network including PCs, phones, cameras, cars, wearables, sensors

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017
Confluence of growth trends in the global datasphere

- Growth of the Internet (> 1 billion users) and broadband availability
- Conversion of formerly analog information to digital
- Falling prices and increased performance for digital devices; ability to store more information and share it in standard formats
- Rise of automation, data-intensive, graphics-intensive, and “smart” applications
- Rise of data centers, cloud computing, social networks
- Increased computerization / roboticization of business, education, entertainment, etc.
- Growth in access, markets, prevalence of digital technologies
- Growth of the Internet of Things
- Increasing data from embedded systems
Increase in data for critical uses (transportation, control systems, health devices, etc.) – greater need for security, fault-tolerance, robustness, etc.

![Data Criticality Over Time](image)

<table>
<thead>
<tr>
<th>Data Type</th>
<th>CAGR 2015 to 2025</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Data. Includes all data in the global datasphere.</td>
<td>30%</td>
</tr>
<tr>
<td>Potentially critical. Data that may be necessary for the continued, convenient operation of users’ daily lives</td>
<td>37%</td>
</tr>
<tr>
<td>Critical. Data known to be necessary for the expected continuity of users’ daily lives.</td>
<td>39%</td>
</tr>
<tr>
<td>Hypercritical. Data with direct and immediate impact on the health and well-being of users. (Examples include commercial air travel, medical applications, control systems, and telemetry. This category is heavy in metadata and data from embedded systems.)</td>
<td>54%</td>
</tr>
</tbody>
</table>

Source: IDC’s Data Age 2025 study, sponsored by Seagate, April 2017

Fran Berman, Data and Society, CSCI 4370/6370
Huge growth projected from the Internet of Things (IoT)

IDC estimates that

In 2014, things in the digital universe approaching 200 billion, 10% (20 billion) of those wired and communicating with the Internet

In 2020, roughly 30 billion connected devices in the digital universe

Fran Berman, Data and Society, CSCI 4370/6370
Data security continues to be the Achilles Heel

Figure 16. Actual Status of Data Security

<table>
<thead>
<tr>
<th>Year</th>
<th>Does not require security</th>
<th>Requires security protected</th>
<th>Requires security unprotected</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>46%</td>
<td>25%</td>
<td>29%</td>
</tr>
<tr>
<td>2015</td>
<td>51%</td>
<td>24%</td>
<td>25%</td>
</tr>
<tr>
<td>2020</td>
<td>33%</td>
<td>32%</td>
<td>35%</td>
</tr>
<tr>
<td>2025</td>
<td>13%</td>
<td>42%</td>
<td>45%</td>
</tr>
</tbody>
</table>

Example
- Camera phone photos
- Digital video streaming
- Public website content
- Open source data

Example
- Corporate financial data
- Personally identifiable Information (PII)
- Medical records

Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017
Which data are useful?

- **Data is useful when we know something about it** — what it represents, where it was collected, what units are being used, etc.? **Metadata** a critical part of the data universe that makes data useful.

- IDC estimates tremendous growth in useful data from 2014 to 2020

Behind the scenes: the “Digital Shadow”

- Less than half of your digital footprint is related to individual actions – taking pictures, making VoIP calls, uploading files, etc.
- The rest of your digital footprint is “ambient” content and metadata related to you: surveillance images, banking records, medical records, information about your web searches and behavior in social networks, etc.
Who can collect, has rights to, and can use information about you, and under what conditions is the subject of national discussions world-wide.
Rising need for infrastructure: Increasing ubiquity of data spurs innovation in analysis, use, infrastructure, need for data skills in workforce
How we access, manage, use, store and preserve data varies widely

- **RETENTION TIMEFRAME:**
  - **Short-term** (few minutes, months, years) to **long-term** (decades, centuries, ...)

- **SIZE / SCALE:**
  - **Small-scale** (KBs, GBs, MBs) to **large-scale / “big”** (TBs, PBs, EBs)

- **PREPARATION:**
  - **Well-tended** (curated, sufficient metadata, cleaned and filtered) to **poorly tended** (flat files, insufficient metadata)

- **POLICY / REGULATION RESTRICTIONS:**
  - Subject to **more restrictive policy and regulation** (e.g. HIPAA) vs. subject to **less restrictive policy and regulation**

- **LIFE CYCLE PLANNING:**
  - Has a **data management and / or sustainability plan** vs. **ad hoc** approach

- **COMMUNITY ACCESSIBILITY:**
  - Shared with others in the community vs. **kept private**; Curated and organized using **community standards** vs. **ad hoc or home-grown** approaches
All digital data cannot be stored

- 2007 was the “crossover year”: Began to generate more digital data than storage to keep it.
- In 2013, current storage capacity could hold just 33% of the digital universe.

Source: IDC, 2008
2010 Update on the Storage Gap:
By 2020, more than twice as much information will be created as storage available

Figure 5: The Emerging Gap
Information Creation > Storage Available

Source: IDC Digital Universe Study, sponsored by EMC, May 2010
Need for data and IT-savvy professionals having tremendous impact on the workforce


140,000–190,000 more deep analytical talent positions, and 1.5 million more data-savvy managers needed to take full advantage of big data in the United States
Data Transformation: The Information Age
The Information Age

• “The Information Age (also known as the Computer Age, Digital Age, or New Media Age) is a period in human history characterized by the shift from traditional industry that the industrial revolution brought through industrialization, to an economy based on information computerization. The onset of the Information Age is associated with the Digital Revolution, just as the Industrial Revolution marked the onset of the Industrial Age.” Wikipedia
How did the Industrial Revolution Transform the World?

• Transition to new manufacturing processes in late 18th / early 19th century.
  – Hand production $\rightarrow$ machines, new chemical manufacturing, new iron production processes
  – Improved efficiency of water power and the increased use of steam power
  – Wood and bio-fuels $\rightarrow$ coal
  – England $\rightarrow$ Western Europe, US

• Major turning point, almost every aspect of daily life influenced in some way

Technological Innovation during the Industrial Revolution

• New technological capabilities and emerging needs had a transformative effect on
  
  – Work opportunities and workforce needs
  
  – National and international priorities
  
  – Economic, cultural, social, and political structures
  
  – Leading sectors (manufacturing, health, energy) and new areas for innovation and impact. Broad ripple effect from both.

Jacquard Loom – Precursor to the Programmable Computer

- **Jacquard loom** invented by Joseph Marie Jacquard and first demonstrated in 1801.

- Loom controlled by punch cards for the purpose of manufacturing textiles with complex patterns.
  - Rows of holes were punched on each card corresponded to one row of the design.

- Loom serves as an important conceptual precursor in the development of computer programming.
Social Innovation during the Industrial Revolution

- **Economic transformation**
  - Better standard of living
  - Better agricultural practices, housing, food supplies
  - Less expensive clothing and consumer goods

- **Urbanization**
  - Rise of factories and modern cities
  - Change in employment options

- **Social policy**
  - Child Labor laws
  - Growth in trade unions

*Cottonopolis* is a name given to the city of Manchester, in England. It denotes a metropolis of cotton and cotton mills, as inspired by Manchester's status as the international centre of the cotton and textile processing industries during this time.

Engraving by Edward Goodall (1795-1870), original title *Manchester, from Kersal Moor* after a painting of W. Wylde. Wikipedia (cropped from original)
Fast forward to the Information Age

- We are experiencing a transformation analogous to the Industrial revolution
- New technological capabilities and emerging needs again having a transformative effect on
  - Work opportunities and workforce needs
  - National and international priorities
  - Economic, cultural, social, and political structures
  - Leading sectors and new areas for innovation and impact. Broad ripple effect from both.
Social Impacts – adequate legal, regulatory, and policy underpinnings for data needed

• How do you maintain personal freedom and sufficient privacy / control over your information?

• What are your rights?
  – What do you own?
  – What can you distribute?
  – What can you charge for?

• What / whom do you trust?
  – Your data?
  – Your respondent?
  – Your hardware?
  – Your system / software?
What happens when digital data becomes the vehicle for progress in the Information Age?

• **APPLICATIONS:** How is digital data being used to drive new innovation? -- **How do we make the most out of data?**

• **INFRASTRUCTURE:** What kind of technological and human infrastructure is needed to support the access, management, use and re-use of digital data today and tomorrow? -- **How do we create a useful data ecosystem?**

• **SOCIAL IMPACT:** What social and community constructs are needed to realize data’s potential? – **What’s needed for a digitally responsible society?**
Next time: Big Data and Data Applications

• Read for next Friday’s discussion (article on the website):
Lecture 1 Sources (not already in text)


• “Data Age 2025: The Evolution of Data to Life Critical,”
Break
How You’ll be Graded

Grade distribution

- Research Paper: 25%
- Op-Ed Draft: 15%
- Op-Ed Final: 15%
- Presentation 1: 15%
- Presentation 2: 15%
- Participation: 15%

Fran Berman, Data and Society, CSCI 4370/6370
Op-Eds
Why is it good to know how to write an op-ed?

• Op-ed is a good example of a brief, persuasive communication to advance your point of view
  – Op-eds can have tremendous influence on community and stakeholders
  – Can establish you as an expert
  – Can be useful to your company, project or community
  – Can get your point of view into a more public discourse

• **Who is your audience:** General public

• **What is your purpose:** Persuasively get your point of view across
Op-Ed Detail -- Structure

Not all Op-Eds are like this, but many good Op-Eds have this structure:

• **Lede** – *Lead-in around a news hook or personal experience*

• **Thesis** – *your position (explicit or implied)*

• **Argument** – *should be based on evidence (stats, news, reports, expert quotes, scholarship, history, experience)*. Arguments often presented as a series of points.

• **Criticism pre-emption** – *take the lead in acknowledging the flaws in your argument and address potential counter-arguments*

• **Conclusion** – *circle back to lede?*
Op-Ed Tips

• Write in a way that smart people can relate to, even if they are not in your discipline. Don’t use buzzwords or talk “inside baseball” without explaining things.

• Pay attention to publication word count – op-eds are usually quite short

• *If you do this for real* (i.e. send it in to a publication rather than do it for class):
  – The final version may be reviewed and/or edited – what you send in may not be the final draft
  – Do your homework – everyone will read this
  – Be prepared for feedback – blogs, tweets, etc.

Fran Berman, Data and Society, CSCI 4370/6370
Grading Detail – Op-Ed
(Draft and Final each 15 points)

• Grade distribution for draft / final:
  – 7 points on editorial content: ideas, thesis, and support
  – 8 points on writing: does it work as an op-ed, is it compelling, does it make sense?

• Draft op-eds due February 9 at the beginning of class. **Bring a hard copy to class and turn it in at the beginning.**
  – Op-ed drafts will be returned with comments on **February 21.**
  – Final op-eds due on **March 2.**
  – Op-ed grades: Draft grade (15 points) + Final grade (15 points)

  – **Important note:** If your draft op-ed is strong, you may choose to not turn in the final op-ed and double your draft grade (op-ed = 2 X draft grade). **You must make this decision and let Fran know before February 21.**

• Op-eds should be in 12 pt. font and between 500 and 1000 words.
Grading Detail – Presentations

Do 2 of these, 15 points each

- You are responsible for ensuring that you sign up for 2 during the semester
- Note that you may do these on Friday Sessions or designated Wednesday Sessions
- Presentation articles will be given in class and can be found on the class website
- A print-out of your slides (please include your name) should be handed in by the beginning of the class in which you will present.

Point Distribution

- Content 4 pts
- Talk 5 pts
- Visuals 4 pts
- Q&A 2 pts

Fran Berman, Data and Society, CSCI 4370/6370
Grading Detail – Presentations

Presentation components
(10 minutes presentation + 5 minutes Q&A):

1. Summary and main points
   • What is the article about?
   • What are the main points of the article?

2. What are the technology and data issues?
   • How is data used to support the article’s point of view?
   • What data infrastructure, policy, practice, etc. is needed to be there for the data to play its role?

3. What is the broader context and future work?
   • What questions arise from reading this article?
   • What is the point of view presented in the article?

Note: You may need to read additional publications, websites for your presentations and reviews

Presentation Grading Metrics:

Talk (5 pts):
• Is the presentation compelling?
• Does the presentation tell an interesting story?
• Did the speaker use the timeframe effectively?

Visuals (4 pts):
• Are the slides well-organized and informative?
• Do the slides help tell the story?
• Are the slides visually interesting?

Content (4 pts):
• Does the speaker understand the topic?
• Has the speaker leveraged appropriate additional materials as needed to support their presentation?

Q&A (2 pts):
• Is the speaker well prepared for questions? Can they respond to them articulately?
Demonstration Presentation (Fran)
AI has found an 8-planet system like ours in Kepler Data
ScienceNews, December 14, 2017
https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data
AI discovers other solar systems like ours

- AI algorithm used Kepler telescope data to identify another solar system that hosts 8 planets
  - Our solar system was previously the only known solar system with 8 planets
- This is the first time an “exoplanet” has been discovered by an algorithm.

More specifics

- The finding was announced by NASA on 12/14/17

- The algorithm and exoplanet discovery is described in a paper accepted the The Astronomical Journal by Christopher Shallue (SW engineer, Google) and Andrew Vanderburg (astronomer, UT Austin)

- Kepler 90 is a sunlike star in the constellation Draco 2500 light-years from earth.

- AI algorithm discovered previously overlooked planet in the data: Kepler 90i.

---

What is a light year?
The light-year is a unit of length used to express astronomical distances. It is about 5.9 trillion miles. A light-year is the distance that light travels in a vacuum in 365.25 days. [Wikipedia]

What is an exoplanet? Planet beyond our solar system [NASA]

What is a planet? A planet is an astronomical body orbiting a star or stellar remnant that is massive enough to be rounded by its own gravity, is not massive enough to cause thermonuclear fusion, and has cleared its neighboring region of plantesimals. [Wikipedia]
The Kepler space telescope and Kepler 90i

• **Kepler 90i**
  - 30% larger than earth
  - 3rd planet from its sun
  - Rocky terrain
  - Orbits Kepler 90 once every 14 days
  - Estimated surface temp 800 degrees, about as hot as Mercury
  - Probably not habitable

• **Kepler Space Telescope**
  - The goal of the Kepler mission was to study Earth-sized planets orbiting Sun-like stars.
  - Kepler is a “Planet hunter” that has spotted more than 2500 confirmed exoplanets since its launch in 2009.
  - Focused on one patch of space for 4 years, now looking in other areas.
What was known before

• 7 previously known planets range from small rocky planets to gas giant

• All planets packed closer to their star than Earth is to the sun

• Kepler 90 may have even more planets as planets closer to their stars are easier to find.

Image: https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data
Backstory: Algorithm

• Previous algorithms used machine learning but this algorithm was the first to use a neural network approach to find overlooked exoplanets.

• Neural network algorithm searches for tiny dips in a star’s brightness when a planet passes in front of it.
  – Algorithm focuses on “the most promising” data, larger study to be done in future.

Research paper in the Astronomical Journal:
https://www.cfa.harvard.edu/~avanderb/kepler90i.pdf
First neural network algorithm used to find exoplanets

- Algorithm trained on more than 15,000 possible Kepler planetary signals that had been labeled by humans as exoplanet or non-exoplanet
- In the test set, the neural network correctly identified true planets and false positives 96% of the time.
- Neural network used test data to learn the characteristics of the light signal of an exoplanet
- Neural network then used on 670 star systems already known to host multiple planets were used in run. (Assumption was that multiple planet systems would be the best places to look for more exoplanets).
- Algorithm provided planet candidates which include some false positives and must be validated by other means.
- Among planet candidates are Kepler 90i and Kepler 80g around Kepler 80, which appear to be planets with high confidence.
Backstory: Data

- Focus of study was data gathered by Kepler between 2009 and 2013.
- Algorithm searches for tiny dips in data that indicate diminishing of a star’s brightness when a planet passes in front of it.
- Data hosted by
  - NASA’s Astrophysics Data System and NASA Exoplanet Archive
  - Mikulski Archive for Space Telescopes.

Image: https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data
What’s next?

• Algorithm to be used on all Kepler data (>150,000 stars) to identify new exoplanets.

• Model improvements:
  – Improved training set (add simulated or unlabeled data to increase size and performance)
  – Improved input representations (better method for flattening light curves, centroid information, robust means instead of medians)
  – View improvement (split local views, have secondary and tertiary views)
  – Additional features (stellar host features), etc.

• Continued focus on neural network algorithms for recognition and identification in broad spectrum of areas
  – “Low hanging fruit”: Astronomical bodies have no “privacy issues” but what about other targets?
References

• “AI has found an 8-planet system like ours in Kepler Data”, ScienceNews, December 14, 2017, https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data


Deconstructed Presentation

• You do not have to deconstruct the presentation you give. This is for demonstration purposes only.

• You do have to list references.
Grading Detail – Presentations

**Presentation components**
(10 minutes presentation + 5 minutes Q&A):

1. **Summary and main points**
   - What is the article about?
   - What are the main points of the article?

2. **What are the technology and data issues?**
   - How is data used to support the article’s point of view?
   - What data infrastructure, policy, practice, etc. is needed to be there for the data to play its role?

3. **What is the broader context and future work?**
   - What questions arise from reading this article?
   - What is the point of view presented in the article?

**Note:** You may need to read additional publications, websites for your presentations and reviews

**Presentation Grading Metrics:**

**Talk (5 pts):**
- Is the presentation compelling?
- Does the presentation tell an interesting story?
- Did the speaker use the timeframe effectively?

**Visuals (4 pts):**
- Are the slides well-organized and informative?
- Do the slides help tell the story?
- Are the slides visually interesting?

**Content (4 pts):**
- Does the speaker understand the topic?
- Has the speaker leveraged appropriate additional materials as needed to support their presentation?

**Q&A (2 pts):**
- Is the speaker well prepared for questions? Can they respond to them articulately?
References

• “AI has found an 8-planet system like ours in Kepler Data”, ScienceNews, December 14, 2017, https://www.scienecn.org/article/ai-has-found-8-planet-system-ours-kepler-data


AI has found an 8-planet system like ours in Kepler Data
ScienceNews, December 14, 2017
https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data
AI discovers other solar systems like ours

- AI algorithm used Kepler telescope data to identify another solar system that hosts 8 planets
  - Our solar system was previously the only known solar system with 8 planets
- This is the first time an “exoplanet” has been discovered by an algorithm.

Summary and main points

More specifics

- The finding was announced by NASA on 12/14/17.

- The algorithm and exoplanet discovery is described in a paper accepted the The Astronomical Journal by Christopher Shallue (SW engineer, Google) and Andrew Vanderburg (astronomer, UT Austin).

- Kepler 90 is a sunlike star in the constellation Draco 2500 light-years from earth.

- AI algorithm discovered previously overlooked planet in the data: Kepler 90i.

---

What is a light year? The light-year is a unit of length used to express astronomical distances. It is about 5.9 trillion miles. A light-year is the distance that light travels in a vacuum in 365.25 days. [Wikipedia]

What is an exoplanet? Planet beyond our solar system [NASA]

What is a planet? A planet is an astronomical body orbiting a star or stellar remnant that is massive enough to be rounded by its own gravity, is not massive enough to cause thermonuclear fusion, and has cleared its neighboring region of plantesimals. [Wikipedia]
The Kepler space telescope and Kepler 90i

- **Kepler 90i**
  - 30% larger than earth
  - 3\textsuperscript{rd} planet from its sun
  - Rocky terrain
  - Orbits Kepler 90 once every 14 days
  - Estimated surface temp 800 degrees, about as hot as Mercury
  - Probably not habitable

- **Kepler Space Telescope**
  - The goal of the Kepler mission was to study Earth-sized planets orbiting Sun-like stars.
  - "Planet hunter" that has spotted more than 2500 confirmed exoplanets since its launch in 2009.
  - Focused on one patch of space for 4 years, now looking in other areas.
What was known before

- 7 previously known planets range from small rocky planets to gas giant
- All planets packed closer to their star than Earth is to the sun
- Kepler 90 may have even more planets as planets closer to their stars are easier to find.

Image: https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data
Backstory: Algorithm

• Previous algorithms used machine learning but this algorithm was the first to use a neural network approach to find overlooked exoplanets.

  – Neural network algorithm searches for tiny dips in a star’s brightness when a planet passes in front of it.
    – Algorithm focuses on “the most promising” data, larger study to be done in future.

Research paper in the Astronomical Journal:
https://www.cfa.harvard.edu/~avanderb/kepler90i.pdf
First neural network algorithm used to find exoplanets

- Algorithm trained on more than 15,000 possible Kepler planetary signals that had been labeled by humans as exoplanet or non-exoplanet.

- In the test set, the neural network correctly identified true planets and false positives 96% of the time.

- Neural network used test data to learn the characteristics of the light signal of an exoplanet.

- Neural network then used on 670 star systems already known to host multiple planets were used in run. (Assumption was that multiple planet systems would be the best places to look for more exoplanets).

- Algorithm provided planet candidates which include some false positives and must be validated by other means.

- Among planet candidates are Kepler 90i and Kepler 80g around Kepler 80, which appear to be planets with high confidence.
Backstory: Data

- Focus of study was data gathered by Kepler between 2009 and 2013.

- Algorithm searches for tiny dips in data that indicate diminishing of a star’s brightness when a planet passes in front of it.

- Data hosted by
  - NASA’s Astrophysics Data System and NASA Exoplanet Archive
  - Mikulski Archive for Space Telescopes.

Image: https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data
What’s next?

• Algorithm to be used on all Kepler data (>150,000 stars) to identify new exoplanets.

• Model improvements:
  – Improved training set (add simulated or unlabeled data to increase size and performance)
  – Improved input representations (better method for flattening light curves, centroid information, robust means instead of medians)
  – View improvement (split local views, have secondary and tertiary views)
  – Additional features (stellar host features), etc.

• Continued focus on neural network algorithms for recognition and identification in broad spectrum of areas
  – “Low hanging fruit”: Astronomical bodies have no “privacy issues” but what about other targets?
References

• “AI has found an 8-planet system like ours in Kepler Data”, ScienceNews, December 14, 2017, https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data


Fran Berman, Data and Society, CSCI 4370/6370
Your Turn ...
Presentation Articles for January 26

• “We gave four good pollsters the same raw data. They had four different results.”, NY Times, http://nyti.ms/2cWlWL7 (Josh S)

• “Silicon Valley’s Mission to Save California Ag from Dying of Thirst,” Wired, https://www.wired.com/2017/05/silicon-valley-aims-save-californias-water/ (Alagu C)

• “Where you live can have a lot to say about your health,” Washington Post, https://www.washingtonpost.com/national/health-science/where-you-live-can-have-a-lot-to-say-about-your-health/2016/12/30/6d94c510-cc73-11e6-a747-d03044780a02_story.html?utm_term=.689dd5bb5168 (Nathalie P)

Presentation Articles for February 2

• “Rethinking Storage in the Age of Big Data”, ComputerWeekly.com,
  http://www.computerweekly.com/feature/Rethinking-storage-in-the-age-of-big-data (Jiyu H)

• “Google Flu Trends: The Limits of Big Data”, New York Times,

• “The Shazam Effect”, The Atlantic,

• “Giving Viewers What They Want,” NY Times,
No class Wednesday. Class next Friday.

• *Next time (January 26):* Big Data 1; Discussion; Presentations

• *Read for January 26 Discussion:*
  