Introduction to the course, Introduction to digital data, Presentation model
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): https://www.cs.rpi.edu/~bermaf/Data%20Course%202019/Data.html
Announcements 1/11

• 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, you will need to come every time to reserve your place and sign the attendance sheet each time.

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

• No Wednesday class January 16.
Today (1/11/19)

• Why Data and Society?

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

• Lecture 1 – Data and Society

• Break

• How to create your article presentation (Fran)
Data is everywhere and impacts everyone
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 (but not comprehensive) 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 on the impacts of technology on society
  
  – The course should provide a complement to the data technologies material in the ITWS Data Science, Web Science, Data Analytics and other courses

• **The course will be structured to**
  
  – Develop your critical thinking around data and help you become a more data-literate professional
  
  – Evolve your oral and written communication and assessment skills

---

**Course foci:**
- Implications for data-driven commerce, science, society
- Data infrastructure, stewardship, policy, regulation
- Current and next-generation challenges and opportunities

**Guest Speakers this Semester:**
- Andy Maltz, March 15?
- Kathy Pham, April 19?
<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 16: NO CLASS</td>
<td>January 18: BIG DATA 1; Topic groups / Topic materials information</td>
<td>Student presentations</td>
<td>Assts.</td>
</tr>
<tr>
<td>January 23: Student presentations</td>
<td>January 25: BIG DATA 2</td>
<td>Student presentations</td>
<td>Assts.</td>
</tr>
<tr>
<td>January 30: NO CLASS</td>
<td>February 1: DATA AND SCIENCE</td>
<td>Student presentations</td>
<td>Assts.</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>Assts.</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>Student presentations</td>
<td>Assts.</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, Topic Reports 1 due</td>
</tr>
<tr>
<td>March 20: TOPICS PRESENTATIONS 1</td>
<td>March 22: DATA AND DATING</td>
<td>Student presentations</td>
<td>Assts.</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, Topic Reports 2 due</td>
</tr>
<tr>
<td>April 17: Student presentations</td>
<td>April 19: CAREERS IN TECH [KATHY PHAM?]</td>
<td>Student presentations</td>
<td>Assts.</td>
</tr>
<tr>
<td>April 24: Student presentations</td>
<td>April 26: TOPICS PRESENTATIONS 2</td>
<td>Student presentations</td>
<td>Assts.</td>
</tr>
</tbody>
</table>
Course Information
https://www.cs.rpi.edu/~bermaf/Data%20Course%202019/Data.html

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

• Presentation articles will be on the web for the week they will be given.
  – Embedded reference materials in the lecture will be given by URL. Lectures will be on the web.

• You are responsible for managing your assignments and participating in the class discussion. No late work.
How you’ll be graded

- Part. / Attend., 10
- Presentation 1, 15
- Presentation 2, 15
- Briefing, 10
- Op-Ed Final, 15
- Topic Report and Presentation (Group), 35
What you’ll be graded on 1
(additional grading specifics for Presentations later today)

• **Presentations (2, 15 points each, individual):**
  – Presentations on data-oriented current topics from the popular press.
  – **Students are responsible for scheduling their own Presentations and ensuring that both are done.** The expected Presentation template will be demonstrated at the end of today’s class.

• **Op-Ed (1, Final 15 points, individual):**
  – Op-Eds are a written “elevator pitch” designed to persuade the reader of a point of view.
  – Students will do a best effort “Draft Op-Ed” and receive feedback for their final Op-Ed. The **final op-ed should be a revision of the draft op-ed**, not a different topic / write-up. **The Final Op-Ed grade will be the one recorded for this assignment.** Op-Ed instructions will be given on January 25.
  – Students who are happy with their grades on the Draft Op-Ed can choose not to submit a Final Op-Ed. In that case, **please inform Fran** and the Draft Op-Ed grade will be recorded as the Final Op-ed grade.
What you’ll be graded on 2

• **Class participation (10 points based on attendance):**
  – Students are expected to attend 20+ out of 22 Wednesday and Friday class meetings for full Attendance credit.
    • You are responsible for making sure you sign the attendance sheet each time you are in class.
    • If you will miss class and your RPI school provides a written excuse, your absence will not be counted against your attendance grade
  – You are expected to participate in class discussions and be an attentive audience (and question asker) for speakers

• **Briefing (1, 10 points, individual)**
  – Short (2 pages undergrads, 3 pages grads) topic-focused paper
  – Instructions given March 1
What you’ll be graded on 3

• Topics Materials (Groups of 4, Group presentation / 15 points, group Topic Report / 15 points, Coordination / at most 5 points, based on individual assessments);
  – Get/read Bruce Schneier’s book (didn’t order it at the Bookstore, should be available through Amazon, etc.)
  – Fran will form Topic Groups on January 18.
  – Choose a topic from Schneier’s book and submit a one-pager describing your topic to Fran by February 8 (more detail later). Choose your Topic Report date preference (March 20 or April 26). (Fran may rebalance ...)
  – Assignment components:
    • Jointly written topic description / Feb. 8 – 0 points
    • Jointly written Topic Report (6-8 pages) – all must contribute – 15 points
    • Joint 15 minute talk (+ 5 min Q&A) (all must contribute) – 15 points
    • Individual assessment of group dynamics and coordination – used for 0-5 point coordination grade given to everyone
  – Topic reports will be provided on the web for the class to read.

Group grade (35 points):
• Joint Presentation: 15 points, usual rubric
• Joint written Topic Report: 15 points, group grade, information provided on Jan. 18
• Group coordination: at most 5 points for coordination based on materials and individual assessments

Fran Berman, Data and Society, CSCI 4370/6370
Grads and Undergrads

- There will be slightly different expectations for grad students and undergrads
  - Individual Briefing lengths and technical depth expectations in presentations and op-eds 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 or group’s own work. In cases where help outside project expectations was received, 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 of everything
  – 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 presentation articles as you can.

• **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)

**Skills you can improve if you spend time on the assignments and participating in class:**

• Persuasive writing
• Informative writing
• Presentation style
• Collaboration and communication skills
• Thinking on your feet (questions, discussion)
• Critical thinking / evidence-based reasoning
Why are you here? (take 5 min)

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.”
Digital data comes from everywhere

- Entertainment
- Health
- Education
- Commerce
- Physical Infrastructure and Smart Systems
- Communication / Community
- Research
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

Fra
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.
Data, data everywhere

- 7+ billion humans on earth
- 3.7 billion humans use the internet
- 2 billion active users on Facebook
- 9 billion email users

- **Every minute**
  - Snapchat users share 527,760 photos
  - Users watch 4,146,600 videos on YouTube
  - 456,000 tweets sent on Twitter
  - 156 million emails sent, 103 million of which are spam
  - The Weather Channel receives 18 million forecast requests

- **Every day**
  - Google processes 3.5 billion searches, 1.5 billion searches per day on other search engines
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

- **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)

Global “Datasphere”

- How much data is there?
- Where is it?
- How is it stored?
- How is it impacting individuals, society, commerce, life?

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
- There are currently 33 zetabytes of data in the global datasphere. By 2025, this will grow to 175 zetabytes.

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kilo</td>
<td>(10^3)</td>
<td></td>
</tr>
<tr>
<td>Mega</td>
<td>(10^6)</td>
<td></td>
</tr>
<tr>
<td>Giga</td>
<td>(10^9)</td>
<td></td>
</tr>
<tr>
<td>Tera</td>
<td>(10^{12})</td>
<td></td>
</tr>
<tr>
<td>Peta</td>
<td>(10^{15})</td>
<td></td>
</tr>
<tr>
<td>Exa</td>
<td>(10^{18})</td>
<td></td>
</tr>
<tr>
<td>Zetta</td>
<td>(10^{21})</td>
<td></td>
</tr>
<tr>
<td>Yotta</td>
<td>(10^{24})</td>
<td></td>
</tr>
</tbody>
</table>

Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Fran Berman, Data and Society, CSCI 4370/6370
How big is the Global Datashpere?

- 1 ZB = 1 trillion GBs

- Global DataspHERE in DVDs:
  - Stack pf DVDs between the earth and the moon 23 times
  - Stack of DVDs that would circle the earth 222 times
  - 1 download at 25 Mb/s (average connection speed across the US) would take 1.8 billion years for one person, 81 days for everyone on earth (without sleep ... 😊)
Where is data in the Global Datasphere located?

- **Core**: computing datacenters in the enterprise and the cloud
- **Edge**: enterprise hardened computers / appliances that are not in core datacenters (servers in the field, cell towers, gateways, regional datacenters, etc.)
- **Endpoint**: devices on the edge of the network including PCs, mobile phones, cameras, cars, wearables, sensors
Increase in data for critical uses (transportation, control systems, health devices, etc.) – greater need for security, fault-tolerance, robustness, etc.

**Figure 5.** Data Criticality Over Time

<table>
<thead>
<tr>
<th>Data Type</th>
<th>CAGR 2015 to 2025</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Data.</strong> Includes all data in the global datasphere.</td>
<td>30%</td>
</tr>
<tr>
<td><strong>Potentially critical.</strong> Data that may be necessary for the continued, convenient operation of users’ daily lives</td>
<td>37%</td>
</tr>
<tr>
<td><strong>Critical.</strong> Data known to be necessary for the expected continuity of users’ daily lives.</td>
<td>39%</td>
</tr>
<tr>
<td><strong>Hypercritical.</strong> 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)

**Devices connected to the Internet:**

- **1970 = 13**
- **1980 = 188**
- **1990 = 313,000**
- **2000 = 93,000,000**
- **2010 = 5,000,000,000**
- **2020 = 31,000,000,000+**

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

- **Does not require security**
  - 2010: 46%
  - 2015: 51%
  - 2020: 33%
  - 2025: 13%

- **Requires security protected**
  - 2010: 25%
  - 2015: 24%
  - 2020: 32%
  - 2025: 42%

- **Requires security unprotected**
  - 2010: 29%
  - 2015: 25%
  - 2020: 35%
  - 2025: 45%

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

Fran Berman, Data and Society, CSCI 4370/6370
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.
All data cannot be stored

- 2007 was the “crossover year”: Began to generate more digital data than storage to keep it
- By 2020, more than twice as much information will be created as storage available
Data infrastructure: 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
Need for data and IT-savvy professionals having tremendous impact on the workforce

From McKinsey Report on Big Data:
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

140,000–190,000
deeper 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.

Iron Bridge, Shropshire, England, first arch bridge to be made of cast iron

Iron Bridge

Savery Steam Engine, First industrially useful steam-powered engine

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?
Next two lectures

• Big data
  – How data is transforming business, public sector, research
  – What big data tells us, what big data doesn’t tell us
Lecture 1 Sources (not already in text)

- “The Digitization of the World: From Edge to Core” IDC Report

- “How much Data Do We Create Every Day?” Forbes


- “Data Age 2025: The Evolution of Data to Life Critical,”
Break
How you’ll be graded

- Presentation 1, 15
- Presentation 2, 15
- Op-Ed Final, 15
- Briefing, 10
- Part. / Attend., 10
- Topic Report and Presentation (Group), 35
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

- Q&A 2 pts
- Content 4 pts
- Talk 5 pts
- Visuals 4 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/questions/issues described of the article?

2. **What are the infrastructure 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. **Discussion**
   - What questions/issues arise from reading this 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?
- Is the font readable, are images used to help convey the points, etc.?

**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)
Cars Suck Up Data About You. Where Does It All Go?
Data and cars – who is listening and where is the data going?

• Cars are becoming “computers on wheels” – vehicles increasingly collect wide variety of data about trips, driving and drivers

• By 2020:
  – Ford vehicles will contain 100 million lines of code
  – 98% of new cars will be connected to cyberspace

• Wide variety of data about driving, trips and drivers collected through diagnostic systems, navigation systems, cellular connections, radar sensors, cameras, etc.

Editorial use only. Editorial credit: Radu Bercan / Shutterstock.com
What data can be tracked by cars?

- How fast you drive
- How hard you brake
- How much fuel your car uses
- What entertainment you prefer
- How often you wear a seatbelt
- Your eye movements
- Your weight
- Whether your hands are on the wheel
- Your phone calls and texts
- Your queries into websites
- Where you like to eat
- Etc.

Who owns the data?

- **Companies ...**
  - Few rules govern what data can be collected, what it can be used for, and how it can be used

- Typically, driver agrees to be tracked and monitored through the user agreement form needed to register a car’s in-dash system or a navigation app.

- If drivers do not agree, the systems/services may not be available

Tracking has both benefits and liabilities

• Benefits
  – Identifies issues, helps to promote maintenance, service, and safety
  – Shared information about traffic and navigation
  – Diagnostics and alerts
  – Lower rates for good drivers

• Liabilities
  – Higher rates for bad drivers
  – Invasion of privacy
  – No driver control over what information is used for or whether it is shared
Infrastructure: Cars are “computers that drive”. Data storage is complex in connected vehicles

• Vehicle systems handle driving, entertainment, biometrics, analysis, navigation, etc. in highly variable and largely unpredictable environments

• Much data from vehicles uploaded into the cloud (how much and when must be optimized).
  – Hitachi: 25 GB uploaded to the cloud per hour from connected cars

• Data challenges:
  – Variable capacity to upload: Cars move through areas of no connectivity and low data latency
  – Vulnerability to corruption: vehicles move between hot and cold temperatures
Automakers are multi-national, but laws are national

• Laws, policy and cultural norms about data privacy and autonomous systems vary widely by jurisdiction – challenging for multi-national companies.
  – Some companies (e.g. Otonomo) emerging to normalize, aggregate and coordinate automotive data from multiple sources and for multiple jurisdictions.

**U.S.:**

• Hodge-podge of federal and state laws about data privacy and autonomous systems
• No “guardrails” for the private sector: GM OnStar system privacy changes permitting data sharing led to numerous consumer complaints

**EU:**

• General Data Protection Regulation guarantees more privacy protections to European citizens
• Germany: guidelines for self-driving cars should protect data privacy of passengers

Fran Berman, Data and Society, CSCI 4370/6370
How private should your data be in a car?

• Should a car be a private space or a public space?
  – Who should be able to access data collected by / in your car and for what purposes?
  – Should your data be used to suggest improvements to your route, driving, destination choices?
  – Should data collected about you be used to market to you, shared with your insurance company, used against you in court?

• U.S. needs to take a stance on vehicle data privacy and develop appropriate standards, policy, legislation, infrastructure to that will enforce that stance.
References


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/questions/issues described of the article?

2. What are the infrastructure 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. Discussion
   • What questions/issues arise from reading this 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?
• Is the font readable, are images used to help convey the points, etc.?

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


Cars Suck Up Data About You. Where Does It All Go?  
Data and cars – who is listening and where is the data going?

• Cars are becoming “computers on wheels” – vehicles increasingly collect wide variety of data about trips, driving and drivers

• By 2020:
  – Ford vehicles will contain 100 million lines of code
  – 98% of new cars will be connected to cyberspace

• Wide variety of data about driving, trips and drivers collected through diagnostic systems, navigation systems, cellular connections, radar sensors, cameras, etc.
What data can be tracked by cars?

- How fast you drive
- How hard you brake
- How much fuel your car uses
- What entertainment you prefer
- How often you wear a seatbelt
- Your eye movements
- Your weight
- Whether your hands are on the wheel
- Your phone calls and texts
- Your queries into websites
- Where you like to eat
- Etc.
Who owns the data?

- **Companies ...**
  - Few rules govern what data can be collected, what it can be used for, and how it can be used

- Typically, driver agrees to be tracked and monitored through the user agreement form needed to register a car’s in-dash system or a navigation app.

- If drivers do not agree, the systems/services may not be available

---

Tracking has both benefits and liabilities

• **Benefits**
  – Identifies issues, helps to promote maintenance, service, and safety
  – Shared information about traffic and navigation
  – Diagnostics and alerts
  – Lower rates for good drivers

• **Liabilities**
  – Higher rates for bad drivers
  – Invasion of privacy
  – No driver control over what information is used for or whether it is shared
Infrastructure: Cars are “computers that drive”. Data storage is complex in connected vehicles

• Vehicle systems handle driving, entertainment, biometrics, analysis, navigation, etc. in highly variable and largely unpredictable environments

• Much data from vehicles uploaded into the cloud (how much and when must be optimized).
  – Hitachi: 25 GB uploaded to the cloud per hour from connected cars

• Data challenges:
  – Variable capacity to upload: Cars move through areas of no connectivity and low data latency
  – Vulnerability to corruption: vehicles move between hot and cold temperatures
Automakers are multi-national, but laws are national

- Laws, policy and cultural norms about data privacy and autonomous systems vary widely by jurisdiction – challenging for multi-national companies.
  - Some companies (e.g. Otonomo) emerging to normalize, aggregate and coordinate automotive data from multiple sources and for multiple jurisdictions.

**U.S.:**
- Hodge-podge of federal and state laws about data privacy and autonomous systems
- No “guardrails” for the private sector: GM OnStar system privacy changes permitting data sharing led to numerous consumer complaints

**EU:**
- General Data Protection Regulation guarantees more privacy protections to European citizens
- Germany: guidelines for self-driving cars should protect data privacy of passengers
How private should your data be in a car?

• Should a car be a private space or a public space?
  – Who should be able to access data collected by / in your car and for what purposes?
  – Should your data be used to suggest improvements to your route, driving, destination choices?
  – Should data collected about you be used to market to you, shared with your insurance company, used against you in court?

• U.S. needs to take a stance on vehicle data privacy and develop appropriate standards, policy, legislation, infrastructure to that will enforce that stance.
References


Your turn!
Presentation Articles for January 18


Presentation Articles for January 23

• “The Most Famous Person to Die in 2018, According to Data Science”, Huffington Post, https://www.huffingtonpost.com/entry/most-famous-celebrity-death-2018_us_5c26634ee4b08aaf7a903312 (Clarisse B.)


Presentation Articles for January 25


• “Electronic Voting was going to be the Future. Now paper’s making a comeback”, CNET, https://www.cnet.com/news/electronic-voting-machines-were-going-to-be-the-future-now-paper-ballots-make-a-comeback/ (Rufeng M.)
No class Wednesday. Class next Friday.

- **MAKE SURE YOU’VE SIGNED THE ATTENDANCE SHEET**  
  (enrolled and waiting list (*))

- **OFFICE HOURS TODAY 1-2, AE 218**

- **Next time (January 18):** Big Data 1; Presentations, Topic groups and instructions

- A pdf of these slides will be on the class website.