Data and Society

The Data-driven World – Lecture 2

9/3/20
Today (9/3/20)

• Personal Essay Assignment and Instructions
• Lecture 2
• Discussion
• Break
• Model Presentation
Reading for 10/10

• Read this article for the 10/10 class discussion.

• “Why It’s So Freaking Hard to Make a Good COVID-19 Model”, 538

<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, SLU</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</td>
<td>Fran</td>
<td>9-28</td>
<td>Data and Privacy – Anonymity, and Contextualization</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</td>
<td>Fran</td>
</tr>
<tr>
<td>10-8</td>
<td>Digital Rights in China</td>
<td>Fran</td>
<td>10-12</td>
<td>Data and Elections</td>
<td>Fran</td>
</tr>
<tr>
<td>10-15</td>
<td>Data and Elections</td>
<td>Todd Rogers, Harvard</td>
<td>10-19</td>
<td>Data and Elections</td>
<td>Fran</td>
</tr>
<tr>
<td>10-22</td>
<td>Data and Discrimination</td>
<td>Fran</td>
<td>10-26</td>
<td>Data and Discrimination</td>
<td>Fran</td>
</tr>
<tr>
<td>10-29</td>
<td>Data and Research</td>
<td>Josh Greenberg, Sloan</td>
<td>11-2</td>
<td>Data and Research</td>
<td>Fran</td>
</tr>
<tr>
<td>11-5</td>
<td>Data and the IoT</td>
<td>Fran</td>
<td>11-9</td>
<td>Data and the IoT</td>
<td>Fran</td>
</tr>
<tr>
<td>11-12</td>
<td>Data and Ethics</td>
<td>Fran</td>
<td>11-15</td>
<td>Data and Ethics</td>
<td>Fran</td>
</tr>
<tr>
<td>11-23</td>
<td>Cybersecurity</td>
<td>Bruce Schneier, Harvard</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11-30</td>
<td>Data and Infrastructure</td>
<td>Fran</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12-3</td>
<td>Data Science</td>
<td>Fran</td>
<td>12-7</td>
<td>Data Science Careers</td>
<td>Kathy Pham, Harvard</td>
</tr>
<tr>
<td>12-10</td>
<td>Wrap-up</td>
<td>Fran</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Writing Assignment – Personal Essay 1
Grading – Personal Essay
(4 of these, 8 points each)
Personal Essay: Your pandemic story

• 450-525 words / 11 point font / 8 points
• Send .docx to bermaf@rpi.edu by Wednesday, September 9 at 12:00 a.m.

• TOPIC: My Pandemic Story
  1. What happened to you when the pandemic hit?
  2. Pick 1:
     • What was a surprising benefit?
     or
     • What’s been the hardest part?
Writing Personal Narratives / Storytelling

• **GOAL:** Tell me (Fran/general public) an interesting (true) story on the assigned topic

• **PURPOSE:** Personal essays explore a *specific experience* and tell the story from your *point of view*. They may illustrate how a personal conflict or event left a lasting impression or how it changed your views or perspective.

• **TONE:** Can be more conversational than formal writing but should establish you as an *articulate and credible individual*.

• **FORMAT:**
  - **Introduction** -- Grab the reader and summarize your points
  - **Body** – main text that tells the story / provides information / explains and supports your points
  - **Conclusion** – may include a lesson, message, moral, take-away
Tips and Grading Rubric for Personal Narratives

TIPS

• Create an outline of the piece (don’t turn this in) before you write with the main points.
• Do more than one draft before turning your piece in.
• Spell and grammar check your piece
• Relevant statistics or facts should be cited and included as endnotes.
• Resources for writing personal essays:
  • https://www.thoughtco.com/write-the-perfect-personal-essay-3858745
  • https://www.indeed.com/career-advice/career-development/how-to-write-a-personal-essay

• GRADING RUBRIC
  (8 points total)
  
  • 4 points – content
    • Is the story compelling?
    • Does the content comply to the personal essay format?
  
  • 4 points – writing
    • Is there a clear tone and narrative?
    • Is it well-written (English, grammar, spelling, flow)?
Lecture 2 – The Data-driven World

Modeling the World on the Internet

- *Your world* represented by the Internet
  - Recommender systems
  - Filter bubbles and echo chambers

- *You represented on the Internet*
  - Profiles and collected data
  - Inferred data

- *Impacts of engaging in a modeled reality*
Everything is a representation on the Internet

• How does the Internet model your world?
  • What you see on the Internet depends on who it perceives you to be
  • What you see on the Internet is optimized, often to maximize your value as a product

• How does the Internet model you?
  • You on the Internet = your profile
  • Profiles developed from collected data, acquired data, inferred data
  • Information may be out of context, incorrect, irrelevant, not managed by humans
Recommender Systems

• **Recommender systems** provide users with personalized product and information offerings based on perceived user preferences
  • Custom suggestions based on user characteristics and past behavior
  • *Self-reinforcing*; can narrow exposure, exacerbate degenerative feedback loops
  • Recommender systems used by a variety of services: Netflix, Amazon, Spotify, YouTube, Facebook, on-line dating sites, etc.

• **Algorithmic issue:** How to tailor outcomes but broaden exposure?

• **Social issue:** How do user’s interests evolve in the presence of recommender systems?
  • User interests may degenerate based on internal dynamics, recommender systems can slow down or accelerate this process
Recommender system methods

• **Collaborative filtering methods**
  - New recommendations based solely on past interactions (no additional information needed)
  - Can be used to detect similar users and/or similar items
  - May or may not utilize a model (memory-based vs. model-based)
  - Drawback: “cold start problem”

• **Content-based filtering methods**
  - New recommendations based on past interactions and additional information
  - User information parametrizes a model of preferences
Recommender systems provide both convenience and a limiting “nudge” to the user

- **“Nudge”** (behavioral science) – use of positive reinforcement and indirect suggestions as ways to influence the behavior and decision making of groups and individuals (e.g. putting fruit at eye level)
  - Nudges alter people’s behavior in a predictable way without forbidding any options or changing their economic incentives

- **Parameters that can be varied in recommender systems**
  - Items shown to the user
  - Number of times an item is shown to the user
  - Size of pool of potential items
  - Accuracy / inaccuracy of predictions (amount of noise)

- **Ways to avoid degeneracy in recommender systems (results from DeepMind research simulations):**
  - Show items only finitely many times.
  - Grow the candidate pool of items shown to the user

- **Note that users typically want a good recommender system but not a quickly degenerating one....**
Netflix and data – a competitive advantage

- Netflix has a 90 second window to help users before they leave the platform for another service
- Algorithms produce $1B in revenue from customer retention
- 80% Netflix views come from recommendations

Recommendation infrastructure:
- 1300 clusters based on user’s viewing preferences
  - 2K taste groups
  - 7K shows and movies in the catalogue
  - Recommender systems use AI and machine learning to develop customized suggestions

Fran Berman, Data and Society, CSCI 4370/6370
Netflix Recommender System

- **Data Netflix collects/acquires to characterize each user**
  - Viewer interactions with Netflix services like viewer ratings, viewing history, etc.
  - Time of day, days of the week, location, device, etc. where a viewer watches (public info)
  - IP address (from device)
  - How long you’ve watched a show
  - Interactions with customer service
  - Web history (cookies, web beacons, advertising identifiers -- from browser)
  - Supplemental information such as demographic data, interest-based data, Internet browsing behavior (from third parties)
  - Name, email, payment method, telephone number, content rating, reviews (user provided)

- **Netflix uses your data**
  - To bucket you into a Taste Community
  - To parametrize its recommender system
  - To optimize audio and video encoding and adaptive bitrate selection
  - To figure out what artwork to show you
  - To parametrize user models for company purchases and creative projects
  - To guide advertising spend, advertising creative, and channel mix to identify new subscribers

- **Netflix shares data “for limited purposes”** with service providers, third-party companies tied to promotional offers with Netflix and law enforcement (upon request)

- **Netflix doesn’t sell member information**, ads to other companies or have third party developers providing applications on the platform
Ways the Internet limits your world – Filter bubbles and echo chambers

• **Echo chamber**: effect of a user’s interest being positively or negatively reinforced by repeated exposure to a certain item or category of items

• **Filter bubble**: when a user encounters only information and opinions that conform to and reinforce their own beliefs

• Filter bubbles and echo chambers limit the world represented to you on the Internet. **Do they limit your world view or conveniently target useful information?**
Ted Talk
Eli Pariser (9:04 minutes)
https://www.youtube.com/watch?v=B8ofWFx525s
How can we ensure that information expands our horizons, rather than limiting them?

• What should companies do?

• What should users do?

• What should government do?

• What’s the problem with having a targeted world?
Read more about it: Lecture 2 Resources

• “Introduction to Recommender Systems”,

• “Degenerate Feedback Loops in Recommender Systems”,


• “How Netflix’s Recommendation System works?”,

• “How to see everything Netflix knows about you,”
Model Presentation
The Amazing Disappearing Election
The Atlantic
What’s newsworthy?

- **Media chooses what you will see.** When there are less choices, we are seeing “agenda compression” and news homogeneity.

- **U.S. Media Collection** – percent of stories in 32 popular publications from Media Cloud mentioning “coronavirus or COVID or Wuhan”: 1/1/20-3/25/20
What bias does your news source have?


Stories mentioning “coronavirus or COVID or Wuhan”: 1/1/20-3/25/20

Fran Berman, Data and Society, CSCI 4370/6370
The diversity of stories in the news is falling in general; big stories contribute

<table>
<thead>
<tr>
<th>Focus</th>
<th>Before</th>
<th>Later</th>
</tr>
</thead>
<tbody>
<tr>
<td>The President</td>
<td>Obama presidency: 1 in 10 stories</td>
<td>Trump Presidency: 1 in 4 Stories</td>
</tr>
<tr>
<td>Sports</td>
<td>7% of the stories in 2014</td>
<td>5.2% of the stories in 2019</td>
</tr>
<tr>
<td>Protest, Black Lives Matter, uprising stories</td>
<td>Before Michael Brown’s killing: 7.4%</td>
<td>After Michael Brown’s killing: 11.3%</td>
</tr>
<tr>
<td></td>
<td>1 week after George Floyd’s killing: 21.5%</td>
<td>2 weeks after George Floyd’s killing: 32.7%</td>
</tr>
<tr>
<td></td>
<td>Just after Breonna Taylor’s killing: Less</td>
<td>After George Floyd’s death: More (currently roughly 2.5%)</td>
</tr>
</tbody>
</table>
How does the media decide what to cover?

- **Chartbeat** documents what readers will click on.
- Social media platforms play an important role:
  - Algorithms govern which stories Facebook and other social media platforms amplify.
- Provocative stories more likely to provoke comments than information stories (e.g. politics, personal health decisions).

*Chartbeat provides information on users and uptake.*

Fran Berman, Data and Society, CSCI 4370/6370
Other factors also influence what you see

• Tough economics:
  • fewer reporters producing less content; key local stories go unreported
• Consolidation of news sources
• High visibility media events shape subsequent media coverage

Image: https://billmoyers.com/story/media-consolidation-should-anyone-care/

Fran Berman, Data and Society, CSCI 4370/6370
Where does the data that publications use come from?

- **Media Cloud provides open source platform** with tools for audience analysis
  - *Explorer*: analysis of how digital news media covers your topic of interest
  - *Topic Mapper*: User topic-generator for a deeper dive into issues.
  - *Source Manager*: DB of print, broadcast and digital news collections
- **Data for this article** came from Media Cloud’s “U.S. Top Online News 2017” collection – set of 32 sources including
  - Websites for major newspapers (The New York Times, The Atlantic, etc.)
  - Cable and broadcast news networks (Fox News, etc.)
  - Magazines
  - Online-native publications (Breitbart News, etc.)
- Sources were “**the most popular U.S. new sites in 2017**” according to data from com.Score, Activate, and Alexa
Media Cloud: Structuring the Data

• Media Cloud indexes stories from tens of thousands of publications and adds metadata to each story to assist media researchers

• Platform organizes data by topic – 400+ topical bins

• Most popular stories trending to fewer bins
  
  • 2013: 37 most popular bins contained 50% of the stories
  
  • 2019: 24 most popular bins contained 50% of the stories
  
  • 2020: 14 most popular bins contained 50% of the stories (so far) [includes public health, disease]
Media compression and news diversity

• Is media agenda compression bad?
  • Harder for a novel stories to break through dominance that Trump and coronavirus have on media agenda in 2020
  • Stories that break through must be really big: Michael Brown’s killing in Ferguson, BLM, protests
  • Creating an echo chamber effect

• Promoting more diverse news
  • News outlets should regularly monitor the diversity of topics in their new coverage

• What should their objectives be?
  Focusing primarily on Chartbeat-type analysis prioritizes popularity and diminishes diversity

• Chicken and egg problem / shared responsibility?
  • News outlets have to find a way to maintain a viable business model with more diverse content
  • Audiences need to proactively read more widely
Questions?
References


• Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election, https://dash.harvard.edu/bitstream/handle/1/33759251/2017-08_electionReport_0.pdf

• Chartbeat website: https://chartbeat.com/products/dashboards/

• Media Cloud website: https://mediacloud.org/

Annotated Presentation
(Don’t present like this!)
References


• Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election, https://dash.harvard.edu/bitstream/handle/1/33759251/2017-08_electionReport_0.pdf

• Chartbeat website: https://chartbeat.com/products/dashboards/

• Media Cloud website: https://mediacloud.org/

The Amazing Disappearing Election

The Atlantic

OVER THE PAST SIX MONTHS, it has seemed like every news story is about the coronavirus, or President Donald Trump’s failed response to it. It’s been a challenge for every crucial issue like racial justice and police violence to break into the media agenda for more than a day or two.

Although the events of 2020 present unique challenges to the media—and to humanity at whole—the narrowing of the media agenda, a phenomenon I call “agenda compression,” dates back to the 2016 election. Former CBS President Les Moonves famously observed of Trump’s candidacy, “It may not be good for America, but it’s damn good for CBS.” News networks made the twists and turns of Trump’s candidacy to record ratings and ad sales, and the spotlight has not turned away from Trump for a second since he won the presidency.
What’s newsworthy?

• **Media chooses what you will see.** When there are less choices, we are seeing “agenda compression” and news homogeneity.

• **U.S. Media Collection** – percent of stories in 32 popular publications from Media Cloud mentioning “coronavirus or COVID or Wuhan”: 1/1/20-3/25/20
What bias does your news source have?


Stories mentioning “coronavirus or COVID or Wuhan”: 1/1/20-3/25/20

Fran Berman, Data and Society, CSCI 4370/6370
The diversity of stories in the news is falling in general; big stories contribute

<table>
<thead>
<tr>
<th>Focus</th>
<th>Before</th>
<th>Later</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The President</strong></td>
<td>Obama presidency: 1 in 10 stories</td>
<td>Trump Presidency: 1 in 4 Stories</td>
</tr>
<tr>
<td><strong>Sports</strong></td>
<td>7% of the stories in 2014</td>
<td>5.2% of the stories in 2019</td>
</tr>
<tr>
<td><strong>Protest, Black Lives Matter, uprising stories</strong></td>
<td>Before Michael Brown’s killing: 7.4%</td>
<td>After Michael Brown’s killing: 11.3%</td>
</tr>
<tr>
<td></td>
<td>1 week after George Floyd’s killing: 21.5%</td>
<td>2 weeks after George Floyd’s killing: 32.7%</td>
</tr>
<tr>
<td></td>
<td>Just after Breonna Taylor’s killing: Less</td>
<td>After George Floyd’s death: More (currently roughly 2.5%)</td>
</tr>
</tbody>
</table>
How does the media decide what to cover?

• **Chartbeat** documents what readers will click on

• Social media platforms play an important role
  • Algorithms govern which stories Facebook and other social media platforms amplify

• Provocative stories more likely to provoke comments than information stories (e.g. politics, personal health decisions)

*Chartbeat provides information on users and uptake*
Other factors also influence what you see

• Tough economics:
  • fewer reporters producing less content; key local stories go unreported
  • Consolidation of news sources

• High visibility media events shape subsequent media coverage

Image: https://billmoyers.com/story/media-consolidation-should-anyone-care/
Where does the data that publications use come from?

- Media Cloud provides open source platform with tools for audience analysis
  - Explorer: analysis of how digital news media covers your topic of interest
  - Topic Mapper: User topic-generator for a deeper dive into issues.
  - Source Manager: DB of print, broadcast and digital news collections

- Data for this article came from Media Cloud’s “U.S. Top Online News 2017” collection – set of 32 sources including
  - Websites for major newspapers (The New York Times, The Atlantic, etc.)
  - Cable and broadcast news networks (Fox News, etc.)
  - Magazines
  - Online-native publications (Breitbart News, etc.)

- Sources were “the most popular U.S. new sites in 2017” according to data from com.Score, Activate, and Alexa
Media Cloud: Structuring the Data

• Media Cloud indexes stories from tens of thousands of publications and adds metadata to each story to assist media researchers

• Platform organizes data by topic – 400+ topical bins

• Most popular stories trending to fewer bins
  • 2013: 37 most popular bins contained 50% of the stories
  • 2019: 24 most popular bins contained 50% of the stories
  • 2020: 14 most popular bins contained 50% of the stories (so far) [includes public health, disease]
Media compression and news diversity

- Is media agenda compression bad?
  - Harder for a novel stories to break through dominance that Trump and coronavirus have on media agenda in 2020
  - Stories that break through must be really big: Michael Brown’s killing in Ferguson, BLM, protests
  - Creating an echo chamber effect

- Promoting more diverse news
  - News outlets should regularly monitor the diversity of topics in their new coverage

- What should their objectives be?
  - Focusing primarily on Chartbeat-type analysis prioritizes popularity and diminishes diversity

- Chicken and egg problem / shared responsibility?
  - News outlets have to find a way to maintain a viable business model with more diverse content
  - Audiences need to proactively read more widely
Questions?
References


• Partisanship, Propaganda, and Disinformation: Online Media and the 2016 U.S. Presidential Election, https://dash.harvard.edu/bitstream/handle/1/33759251/2017-08_electionReport_0.pdf

• Chartbeat website: https://chartbeat.com/products/dashboards/

• Media Cloud website: https://mediacloud.org/

Presentation structure and grading
Presentation components and grading metrics

Presentation components (12 minutes presentation + 6 minutes Q&A):

- **Summary and main points**
  - What is the article about?
  - What are the main points/questions/issues described of the article?

- **What are the data issues?**
  - How is data used to support the article’s point of view?

- **Discussion**
  - What questions/issues arise from reading this article?

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

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?
More about presentations

Do 2 of these, 15 points each

• *You are responsible for ensuring that you sign up for 2 during the semester*
• Presentation articles will be given in class and can be found on the class website
• Send your slides as a .pdf attachment (please include your name) to *bermaf@rpi.edu* 15 minutes before the beginning of the class in which you will present.

**TIPS:**

• Practice your presentation. Use this as an opportunity to become a better public speaker
• **DO NOT SPEND MOST OF THE TALK READING YOUR NOTES**
• Talk so your audience can hear you. Engage with your audience
• Be prepared for questions and have an intelligent form of “I don’t know”
• Use visuals and text to make your slides interesting. Don’t make the font too small.
Your Turn!

- **Presentations for September 10**

- **Presentations for September 14**

- **Presentations for September 17**