Announcements 1/25

• How to have an argument ... 😊

• No Wednesday class January 30.

• Op-Ed instructions today after the lecture.

• Group topic summaries due at the beginning of class on February 8 (one summary for the whole group).

• Summary format (answer these questions in one page):
  – Who is in the group?
  – What problem will your topic report be on? (Provide some detail about this, including where it is described in the book).
  – What role will each of the group members play in developing the topic report and giving the presentation?
<table>
<thead>
<tr>
<th>Wednesday Section</th>
<th>Friday Lecture (first half)</th>
<th>Second half of class</th>
<th>Assts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 9: NO CLASS</td>
<td>January 11: INTRO – DATA AND SOCIETY</td>
<td>Fran presentation demo</td>
<td></td>
</tr>
<tr>
<td>January 16: NO CLASS</td>
<td>January 18: BIG DATA 1; Topic groups / Topic materials information</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>January 23: Student presentations</td>
<td>January 25: BIG DATA 2</td>
<td>Student presentations</td>
<td>Op-Ed instructions</td>
</tr>
<tr>
<td>January 30: NO CLASS</td>
<td>February 1: DATA AND SCIENCE</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>February 6: NO CLASS</td>
<td>February 8: DATA STEWARDSHIP AND PRESERVATION</td>
<td>Student presentations</td>
<td>Group Topics due</td>
</tr>
<tr>
<td>February 13: NO CLASS</td>
<td>February 15: INTERNET OF THINGS</td>
<td>Student presentations</td>
<td></td>
</tr>
<tr>
<td>February 20: Student presentations</td>
<td>February 22: DATA AND PRIVACY / FOUNDATIONS</td>
<td>Student presentations</td>
<td>Op-Ed Drafts due</td>
</tr>
<tr>
<td>February 27: NO CLASS</td>
<td>March 1: DATA AND PRIVACY / POLICY AND REGULATION</td>
<td>Student presentations</td>
<td>Briefing instructions</td>
</tr>
<tr>
<td>March 6: Spring Break</td>
<td>March 8: Spring Break</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 13: Student presentations</td>
<td>March 15: DATA AND ENTERTAINMENT [ANDY MALTZ?]</td>
<td>Student presentations</td>
<td>Op-Ed Drafts Returned 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></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></td>
</tr>
<tr>
<td>April 24: Student presentations</td>
<td>April 26: TOPICS PRESENTATIONS 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lecture 3: Big Data 2

- Big Data and Astronomy
- Big Data and the Public Sector
Big Data and Astronomy

Hubble mosaic of the crab nebula (public domain)

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

• Astronomy seeks to improve our understanding of celestial objects and phenomena.

• Data important in many areas of Astronomy:
  – **Observation** – data acquired from instruments (large-scale telescopes, radio astronomy arrays, satellites, etc.) stored and analyzed for research and applications.
  – **Theoretical** – computer models and simulations developed to describe the properties and characteristics of astronomical objects and phenomena (often called astrophysics)
Many open questions in Astronomy

A few of them [Wikipedia/Astronomy]:

• What is the origin of the stellar mass spectrum? That is, why do astronomers observe the same distribution of stellar masses—the initial mass function—apparently regardless of the initial conditions?
• Is there other life in the Universe? Especially, is there other intelligent life?
• Is the Solar System normal or atypical?
• What is the nature of dark matter and dark energy?
• What will be the ultimate fate of the universe?
• How did the first galaxies form? How did supermassive black holes form?
• What is creating the ultra-high-energy cosmic rays?
• Why is the abundance of lithium in the cosmos four times lower than predicted by the standard Big Bang model?
• Etc.
Telescopes are data generators

- Modern telescopes image enormous portions of the sky with very sensitive detectors.
  - **Large Synoptic Survey Telescope** has 3B pixel digital camera
  - Largest telescopes producing **petabytes of data**.

- Different telescopes focus on different parts of the sky, different kinds of bodies and phenomena, different radio-frequencies, etc.
  - Community metadata agreement important so that various surveys and collections can be combined.

- **Astronomical data doubles roughly every year**.
  - Community has created searchable archives and portals that allow data to be analyzed and “sampled” for individual studies.
Making the most of the data

• **Community and open science focus** optimizes usefulness of data through coordination and interoperability of different projects

• **Information technologies critical for astronomy data:**
  
  – **Common metadata** needed to identify distance, color, brightness, size, survey parameters, etc. of data and objects.
  
  – **Tools** needed for efficient analysis of massive data sets, search and down-sampling to manageable sets for researchers.
  
  – **Stewardship and preservation** needed to make datasets available for open community research.

State of the Art: The Large Synaptic Survey Telescope (LSST)

- **New telescope (LSST) being built** to photograph the entire available sky every few nights.
  - Innovations include novel 3 mirror design that will deliver sharp images over a wide diameter field of view
  - Images will be recorded by a 3.2 gigapixel CCD imaging camera – largest digital camera ever built
  - **Telescope will reside in northern Chile.** Full operations anticipated in 2022.

- **Science goals** include:
  - Studying dark energy and dark matter
  - Mapping small objects in the Solar System
  - Detecting transient optical events (novae, supernovae, etc.)
  - Mapping the Milky Way, etc.

Image: Todd Mason, Mason Productions Inc. / LSST Corporation - https://www.lsst.org/sites/default/files/LSST-Print5-high.jpg
Data, infrastructure, and the Large Synaptic Survey Telescope (LSST)

- LSST committed to making all data public as soon as it is taken.
  - LSST camera expected to take \(~1.28\) petabytes of data per year.
  - Computational and data infrastructure to deal with the data expected to include 250+TF computers and 15 PB of storage (10% of computing power and disk space to be used for user-generated products)
  - Data available on-line as prompt data (provided within 60 seconds of observation), daily data and annual data (most curated)

- LSST joins other large-telescopes and optical sky surveys (Sloan Digital Sky Survey, Hubble Telescope, Palomar Observatory Sky Survey, etc.) to provide broad spectrum of observed objects, events, characteristics of the universe.
Large-scale telescopes complex to build and fund (https://www.lsst.org/about/project-status)

LSST Project Schedule – 8.5 Months Contingency
How do modelers use astronomical data?

https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data
AI algorithm and astronomical data used to discover 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

- Interesting result: First time an “exoplanet” has been discovered by an algorithm.

Finding announced by NASA on 12/14/17

- The algorithm and exoplanet discovery described in a paper in 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 planetesimals. [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
Algorithm: Neural network algorithm used to find 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

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

Research paper in the Astronomical Journal:
https://www.cfa.harvard.edu/~avanderb/kepler90i.pdf

Fran Berman, Data and Society, CSCI 4370/6370
AI approach

- **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

- Trained neural network **used on 670 star systems already known to host multiple planets**. (Assumption was that multiple planet systems would be the best places to look for more exoplanets).
  - Algorithm-provided planet candidates included false positives and needed to be validated by other means.
  - Among planet candidates were Kepler 90i (orbiting around Kepler 90) and Kepler 80g (orbiting around Kepler 80), which appear to be planets with high confidence.
Kepler data

• Study utilized **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.

• **Stewardship:** 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](https://www.sciencenews.org/article/ai-has-found-8-planet-system-ours-kepler-data)
Ongoing work

• 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?
Data-driven Astronomy / (Science) Predictions

• How do we know if predictions are “right”?
  – Observations / data validate predicted phenomena and characteristics
  – Cross-calibration of different sources of observations and data validate one another
  – Models and data seem to explain new observed phenomena

• What happens if data-driven astronomy predictions are wrong?
  – Researchers may work on improving the models
  – Researchers may gather more data and/or ensure that the data is accurate
  – Researchers may look for other explanations ...
  – Applications based on poor predictions likely to be ineffective ...
    (garbage in / garbage out ...)

Fran Berman, Data and Society, CSCI 4370/6370
Big Data and the 2016 Election – Failure of Big Data?

FiveThirtyEight
2016 Election Forecast

Who will win the presidency?

Chance of winning

Hillary Clinton
71.4%

Donald Trump
28.6%

Electoral votes

Popular vote

Hillary Clinton
302.2

Donald Trump
233.0

Gary Johnson
0.9

Evan McMullin
0.8

Other
0.0

Fran Berman, Data and Society, CSCI 4370/6370
What Happened? Why were most predictions inaccurate?

- Were the models wrong?
- Were the interpretations wrong?
- Was the data faulty?
- Was the sampling / polls wrong?
- Was voter behavior just one of the low probability outcomes?

Map from http://www.270towin.com/2016_Election/interactive_map
How people voted: Exit Polls and Election Results

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Breakdown(s)</th>
</tr>
</thead>
</table>
| **Age**             | • 18-29: 55% for Clinton, 37% for Trump  
• 30-44: 50% for Clinton, 42% for Trump  
• 45+: 53% for Trump  
• 45-64: 44% for Clinton  
• 65+: 45% for Clinton |
| **Gender**          | • 54% of women voted for Clinton; 42% of women voted for Trump  
• 53% of men voted for Trump; 41% of men voted for Clinton |
| **Ethnicity**       | • White voters: 58% for Trump, 37% for Clinton  
• Black voters: 88% for Clinton; 8% for Trump  
• Hispanic and Asian voters: 65% for Clinton; 29% for Trump |
| **Education**       | • College grads: 49% for Clinton  
• Postgrads: 58% for Clinton  
• High school or less: 51% for Trump  
• Some college / Associate degree: 52% for Trump |
| **Religion**        | • Catholic: 52% for Trump  
• Protestants / Christians: 58% for Trump  
• Jewish: 71% for Clinton  
• Other: 62% for Clinton  
• No religion: 68% for Clinton |
| **Income**          | • Under $30K: 53% for Clinton  
• $30K-$49.99K: 51% for Clinton  
• $50K - $99.99K: 50% for Trump  
• $100K-$199.99K: 48% for Trump  
• $250K+: 48% for Trump |
| **Locale (Urban vs. Rural)** | • Cities with > 50K residents: 59% for Clinton, 35% for Trump  
• Rural areas: 62% for Trump, 34% for Clinton  
• Suburbs: 50% for Trump, 45% for Clinton |
Who can / did vote?

All Americans
320,000,000+

Voting age population
251,107,404 (78.5%)

Eligible voters
231,556,622 (72.4%)

Registered voters
~200,000,000 (62.5%)

Voters
138,884,643
(43.4% of all citizens, 60% of eligible voters)

Statistics from:
Election models: Model and Interpretation

Accuracy

Many challenges in modeling and interpretation:

- Raw polling data supplemented by estimates on how many people will vote and what undecided voters will do.

- Historical inferences about past patterns of turnout, demographics, economic conditions and party loyalty may not be accurate for present day.

- If polls shows that candidate “wins” by a small margin within the margin of error, it is risky to interpret this as a “win.”

From: https://projects.fivethirtyeight.com/2016-election-forecast/

Data Integrity – Was poll data accurate?

• Many **suspected that people lied** about voting for Trump

• Trafalgar Group’s approach to improving data accuracy -- Adjust numbers to account for people’s hesitance to admit a Trump vote
  – Used robotic calls for which Trump voters seemed more comfortable
  – Added a “neighbor” question -- Who do you think your neighbors will vote for? – and checked to see if the numbers were different
  – Created a demographic of people who had not voted in 6+ years but planned to vote for Trump

• Trafalgar predicted Trump win in Pennsylvania and Michigan (but not all states)
Sampling Accuracy

Key sampling questions

• How representative is the sample of population?
• How big is the sample / what is the margin of error?
• How biased are the sampling vehicles – land lines, human interviews, tweets, non-self screening respondents, etc.?
• How representative is the sample of turnout? For eligible voters? For eligible voters who actually vote?
• How accurate is the data (are people lying)?

Figure from http://www.forbes.com/sites/startswithabang/2016/11/09/the-science-of-error-how-polling-botched-the-2016-election/#75748a437da8
Who made correct predictions?

- **Investor’s Business Daily (IBD/TIPP poll)**
  
  **Predicted:** Trump would win by 1.6%

  **Approach:**
  
  - Start with random sample from public, adjust for census statistics and age, gender, religion, look at registered and likely voters, adjust for party registrations, **enthusiasm**
  
    - Poll made more calls to **smartphones** than landlines.
  
    - People represented wide range of people in the country (including a representative sample of types of phones used)
  
    - Poll questioned respondents about enthusiasm and factored this into results

- **USC / LA Times poll (USC economics prof Arie Kapteyn)**
  
  **Predicted:** Trump would win by 3%

  **Approach:**
  
  - Pollsters sought to **balance both big groups** (e.g. men and women) and **smaller groups** (e.g. young minority voters)
  
    - **Weighting of responses** in polls used to make them more fully representative. Sample includes representation of demographic statistics including race, gender, age.
Who made correct predictions?

- **Primary Model.com / Helmut Norpoth (Stonybrook political science prof)**
  
  **Predicted:** Trump would win against Clinton with 87% certainty.
  
  - Predicted last 5 presidential elections correctly; predicted the results of every presidential election except 1 in last 104 years.

  **Approach:**
  
  - Uses **primaries** rather than polls to predict outcomes.
  - Takes “swing of the electoral pendulum” into consideration (Republicans favored after two democratic terms)

- **Alan Lichtman / American University historian**
  
  **Predicted:** Trump wins

  **Approach:**
  
  - Developed 13 **T/F keys** that predict election outcome. True favors incumbent party. If 6+ are false, change is predicted.
  
  - Has worked in every election for the last 30 years.

  **Lichtman’s Keys:**
  
  1. **Party Mandate:** After the midterm elections, the incumbent party holds more seats in the U.S. House of Representatives than after the previous midterm elections.
  2. **Contest:** There is no serious contest for the incumbent party nomination.
  3. **Incumbency:** The incumbent party candidate is the sitting president.
  4. **Third party:** There is no significant third party or independent campaign.
  5. **Short-term economy:** The economy is not in recession during the election campaign.
  6. **Long-term economy:** Real per capita economic growth during the term equals or exceeds mean growth during the previous two terms.
  7. **Policy change:** The incumbent administration effects major changes in national policy.
  8. **Social unrest:** There is no sustained social unrest during the term.
  9. **Scandal:** The incumbent administration is untainted by major scandal.
  10. **Foreign/military failure:** The incumbent administration suffers no major failure in foreign or military affairs.
  11. **Foreign/military success:** The incumbent administration achieves a major success in foreign or military affairs.
  12. **Incumbent charisma:** The incumbent party candidate is charismatic or a national hero.
  13. **Challenger charisma:** The challenging party candidate is not charismatic or a national hero.
Using big data

• Big data is a great tool but it must be used with perspective, integrity, fairness, and validation
  – Big data is not a magic bullet ...

• Predictions and estimations (big data or otherwise) are not a guarantee of outcomes
  – Bias can be built into the system at many places: data collection, data sampling, models, interpretation, etc.
  – Garbage in, garbage out
Lecture 3 sources 1 (not on slides)


Lecture 3 Sources 2


• “The trouble is not with polling but with the limits to human interpretation of data,” Quartz, http://qz.com/832908/confirmation-bias-is-why-we-couldnt-predict-a-trump-victory/

• “There are Many Ways to Map election Results. We’ve Tried Most of Them.”, NY Times, http://www.nytimes.com/interactive/2016/11/01/upshot/many-ways-to-map-election-results.html?_r=0

• “Trump’s win isn’t the death of data – it was flawed all along,” Wired, https://www.wired.com/2016/11/trumps-win-isnt-death-data-flawed-along/

Lecture 3 Sources 3


Op-Ed Instructions
Why is it good to know how to write an op-ed?

• Op-Eds are a written “elevator pitch” designed to persuade the reader of a point of view.

• **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?*

**Lede Options**

- Current news
- Dramatic or personal anecdote
- Reference to popular culture or twist on conventional wisdom
- Anniversary of an event
- Major new study
Op-Ed Topics

- Op-eds are persuasive pieces, so the topic should be a little controversial. (If everyone agrees on something, there is no reason to convince them).
- Make sure your op-ed doesn’t just inform (provide information neutrally) but persuades (makes an evidence-based persuasive argument or advances an evidence-based opinion).
- Remember, there needs to be a data angle to your op-ed. (This is Data & Society ...).

<table>
<thead>
<tr>
<th>Largely uncontroversial topics</th>
<th>Good topics for op-eds</th>
</tr>
</thead>
<tbody>
<tr>
<td>World peace is good.</td>
<td>We are heading towards a global war in the middle east.</td>
</tr>
<tr>
<td>Recycling is good for the environment.</td>
<td>We need to pass legislation to reduce the number of plastic bags.</td>
</tr>
<tr>
<td>Technology has transformed music.</td>
<td>Copyright laws are killing hip hop.</td>
</tr>
</tbody>
</table>
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):
  – Timing is critical. A provocative thesis and lede can make a huge difference in terms of what gets published and when.
  – 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. Fact check, fact check, fact check.
  – Be prepared for feedback – blogs, tweets, etc.
Grading Detail – Op-Ed (Final is 15 points)

- Grade distribution for draft / final:
  - 7 points on editorial content: ideas, thesis, and supportive arguments
  - 8 points on writing: does it work as an op-ed (addresses an issue, presents a well-evidenced opinion, is it more than an information piece ...), does it tell a compelling story, is it well-written?

- Draft op-eds due February 22 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 by March 15.
  - Final op-eds due on March 29.
  - Op-ed grade given for Final Op-Ed only, draft op-eds will be graded but do not count for your grade UNLESS (see below)
  - Important note: If your draft op-ed is strong, you may choose to have the draft op-ed count as your final op-ed. In this case, you do not have to turn in a final op-ed. Let Fran know before March 29.

- Op-eds should be in 12 pt. font, double-spaced and between 500 and 1000 words. Op-eds should include references rather than links.
Break
Presentations
Presentation Articles for February 1


Presentation articles for February 8


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


• “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.)