Today (9/24/20)

• Personal Essay #2 **due by October 4, 12 a.m.**
  Information in 9/21/20 lecture

• Lecture – Differential Privacy

• Discussion

• Student Presentations
Reading for September 28

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Speaker</th>
<th>Date</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-31</td>
<td>Introduction</td>
<td>Fran</td>
<td>9-3</td>
<td>The Data-driven world</td>
<td>Fran</td>
</tr>
<tr>
<td>9-10</td>
<td>Data and COVID-19 - models</td>
<td>Fran</td>
<td>9-14</td>
<td>Data and COVID-19 – contact tracing</td>
<td>Fran</td>
</tr>
<tr>
<td>9-17</td>
<td>Data and the Opioid Crisis</td>
<td>Liz Chiarello</td>
<td>9-21</td>
<td>Data and Privacy - Intro</td>
<td>Fran</td>
</tr>
<tr>
<td>9-24</td>
<td>Data and Privacy – Differential Privacy and the Census</td>
<td>Fran</td>
<td>9-28</td>
<td>Data and Privacy – Anonymity</td>
<td>Fran</td>
</tr>
<tr>
<td>10-1</td>
<td>Data and Privacy - Law</td>
<td>Fran</td>
<td>10-5</td>
<td>Digital rights in the EU and China</td>
<td>Fran</td>
</tr>
<tr>
<td>10-8</td>
<td>Data and Elections</td>
<td>Fran</td>
<td>10-12</td>
<td>Columbus / Indigenous Peoples’ Day – NO CLASS</td>
<td>Fran</td>
</tr>
<tr>
<td>10-15</td>
<td>Data and Elections</td>
<td>Todd Rogers</td>
<td>10-19</td>
<td>Data and Elections – this election</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</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 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</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 Careers</td>
<td>Kathy Pham</td>
</tr>
<tr>
<td>12-10</td>
<td>Wrap-up</td>
<td>Fran</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lecture – Differential Privacy

- What is differential privacy?
- Differential Privacy and the Census
What is Differential Privacy and Why is it useful?

- What is Differential Privacy? NIST
- https://www.youtube.com/watch?v=JRURYTfBXQ
Differential Privacy [Dwork et al.]

• Differential Privacy is an approach which promotes privacy in statistical databases, i.e. information about the group but not specific individuals
  – Statistical database = set of data collected under the pledge of confidentiality for the purpose of producing statistics that do not compromise the privacy of the individuals that provided the data

• Privacy goal: Protect every individual while permitting statistical analysis of the DB as a whole. (“Nothing is learned” about an individual from before the query analysis to after).

• Differential privacy process introduces randomness to queries in a structured way that produces the same statistical results.
Differential Privacy Methodology

• Appropriately used, approach can be used to reasonably ensure privacy *no matter what other data is available*

• Basic idea:
  – *Introduce randomness into the results of queries* to a statistical DB of confidential data. (Promotes privacy wrt “query release”)
  – If done correctly, very accurate statistics of DB can be produced while ensuring high levels of privacy

• *Differential privacy is a condition of the release mechanism* (i.e. the trusted party releasing information about the dataset), rather than on the dataset itself.
  – If two datasets are similar, a differentially private algorithm should behave roughly the same on both datasets (i.e. the presence or absence of an individual will not affect the final output of the algorithm significantly)
How does it work?

• Procedure for achieving a randomized response wrt query “Do you have Property A?”
  1. Throw a coin
  2. If tails, then answer honestly
  3. If heads, then throw the coin again and answer YES if heads, NO if tails

• Why does this work?
  – Let $p =$ true proportion of people with Property A
  – By the procedure, we would expect to obtain $pX_{1/2} + 1/4$ YES answers.
  – so if we know the number $N$ of YES answers, we can figure out $p$ without being able to figure out individual entries ($p = 2N - 1/2$).
Increasing use in private and public sectors

- Adoption of differential privacy in real-world applications
  - 2008: U.S. Census Bureau, for showing commuting patterns.
  - 2014: Google's RAPPOR, for telemetry such as learning statistics about unwanted software hijacking users' settings.
  - 2015: Google, for sharing historical traffic statistics.
  - 2016: Apple announced its intention to use differential privacy in iOS 10 to improve its Intelligent personal assistant technology.
  - 2017: Microsoft, for telemetry in Windows.
  - 2019: Privitar Lens uses differential privacy API.
  - 2020: LinkedIn, for advertiser queries.
What are the limitations of differential privacy?

- Estimation from repeated queries -- with more and more queries, privacy is breached. (Exhausts the “privacy budget”)
- Data set must be “large enough” so that statistical variations will work and promote accuracy
- Method works in situation where queries about the group are of interest, not in the situation where queries about the individual are important.
- Requires substantial computation
- Trades-off privacy and accuracy
Differential Privacy and the Census

Discussion

• What do we use Census data for?
• Why does privacy of Census data matter?
• Why is differential privacy being used with Census data?
• What are the issues involved in the privacy vs. accuracy trade-off?
• How does differential privacy impact re-districting and other Census data-driven efforts?
Lecture 7 References (not already on slides)

• Differential Privacy, Wikipedia

• Differential Privacy, Simply Explained, YouTube, https://www.youtube.com/watch?v=gI0wk1CXlsQ

Presentations
Upcoming Presentations

• Presentations for September 28

• Presentations for October 1
Presentations for October 5


Need Volunteers – Presentations for October 8

• “Do Republicans or Democrats benefit from mail-in voting? It turns out neither.”, Science, https://www.sciencemag.org/news/2020/08/do-republicans-or-democrats-benefit-mail-voting-it-turns-out-neither (Richard)

Presentations for Today
