Data Science for Social Impact
Examples, Opportunities, Challenges

Rayid Ghani
Center for Data Science & Public Policy
[Joint work with Sam Carton, Lauren Haynes, Jen Helsby, Kenny Joseph, Young-Soo Park, Joe Walsh – KDD 2016]
[Joint work with Sam Carton, Lauren Haynes, Jen Helsby, Kenny Joseph, Young-Soo Park, Joe Walsh – KDD 2016]
<table>
<thead>
<tr>
<th>EIS Status Summary For:</th>
<th>-- Freedom Division</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accidents</strong></td>
<td></td>
</tr>
<tr>
<td>Time Frame: 180 Days</td>
<td>Time Frame: 180 Days</td>
</tr>
<tr>
<td>No of Accidents: 0</td>
<td>No of Complaints: 2</td>
</tr>
<tr>
<td>Threshold: 2</td>
<td>Threshold: 3</td>
</tr>
<tr>
<td><strong>Injuries</strong></td>
<td></td>
</tr>
<tr>
<td>Time Frame: 180 Days</td>
<td>Time Frame: 90 Days</td>
</tr>
<tr>
<td>No of Injuries: 1</td>
<td>No of Uses of Force:</td>
</tr>
<tr>
<td>Threshold: 2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Pursuits</strong></td>
<td></td>
</tr>
<tr>
<td>Time Frame: 180 Days</td>
<td>Time Frame: 180 Days</td>
</tr>
<tr>
<td>No of Pursuits: 1</td>
<td>No of Events: 5</td>
</tr>
<tr>
<td>Threshold: 2</td>
<td>Threshold: 5</td>
</tr>
<tr>
<td><strong>Sick Leave/Days Off</strong></td>
<td></td>
</tr>
<tr>
<td>Time Frame: 30 Days</td>
<td></td>
</tr>
<tr>
<td>No of Events: 0</td>
<td></td>
</tr>
<tr>
<td><strong>Sick Leave/Vacation</strong></td>
<td></td>
</tr>
<tr>
<td>Time Frame: 90 Days</td>
<td></td>
</tr>
<tr>
<td>No of Events: 0</td>
<td></td>
</tr>
</tbody>
</table>

**Complaints**

| Time Frame: 180 Days |
| No of Complaints: 2 |
| Threshold: 3        |
[Joint work with Eric Potash, Joe Walsh], Joe Brew, Alex Loewi, Subho Majumdar, Andrew Reece KDD 2015]
Impaired Attention

Lack of Motor Skills

Hearing Loss

Learning Disability

Lower IQ

Memory Problems

[Joint work with Eric Potash, Joe Walsh, Joe Brew, Alex Loewi, Subho Majumdar, Andrew Reece KDD 2015]
WHO IS DROPPING OUT?

30% of high school students in the U.S. drop out.

Every year, 1.3 million students drop out of school.

7,000 every day of school or 1 every 26 seconds.

[Joint work with Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, KDD 2015]
Existing System can identify 45% of the at-risk students in 10th grade.
Operate 2.5x more toilets

Serve 45,000 more people
240,000 main breaks/yr nationwide
$13 billion in 2010 to repair
Expected $30 billion by 2040
180 breaks/yr in Syracuse, NY
How well can we predict main failures?

In a ranked list of the top 1% of blocks (50 blocks) with the highest risk of failure...

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>blocks identified by random selection</td>
</tr>
<tr>
<td>32</td>
<td>blocks identified by our model</td>
</tr>
</tbody>
</table>

I had a chance to look at your predictions versus water main breaks so far this year. So far, 7 of the top 50 have broken, and then at least another 5 are off by only a block. So, if we’re supposed to get to about 32 by the end of 2018, you all are on track!
46% of the Mexican Population lives in poverty
Our Results

2.85M

.78M

.21M

.84M

.77M

6.97M

Total people with imputed poverty indicators, who can now be better matched with social programs:

7,401,777
6,987 Jail Days
Data Science Problem Templates

• Can I detect who’s going to get lead poisoning early?

• Can I determine which home inspections to prioritize?

• Which policies do I modify to improve maternal mortality?

• How much impact is my after-school program having?

• Can I get data that helps me match employers with employees?
Data Science Problem Templates

- Can I detect who's going to get lead poisoning early?
- Can I determine which home inspections to prioritize?
- Which policies do I modify to improve maternal mortality?
- How much impact is my after-school program having?
- Can I get data that helps me match employers with employees?
Data Science Problem Templates

• Can I detect who's going to get lead poisoning early?  
  
• Can I determine which home inspections to prioritize?  
  
• Which policies do I modify to improve maternal mortality?

• How much impact is my after-school program having?

• Can I get data that helps me match employers with employees?
Data Science Problem Templates

- Can I detect who's going to get lead poisoning early?  
- Can I determine which home inspections to prioritize?  
- Which policies do I modify to improve maternal mortality?  
- How much impact is my after-school program having?  
- Can I get data that helps me match employers with employees?
Can I detect who's going to get lead poisoning early?

Can I determine which home inspections to prioritize?

Which policies do I modify to improve maternal mortality?

How much impact is my after-school program having?

Can I get data that helps me match employers with employees?
Data Science Problem Templates

• Can I detect who's going to get lead poisoning early?
  - 

• Can I determine which home inspections to prioritize?
  - 

• Which policies do I modify to improve maternal mortality?
  - 

• How much impact is my after-school program having?
  - 

• Can I get data that helps me match employers with employees?
Can I detect X early?
Common Challenges

- Privacy
- Security
- Interpretability
- Transparency
- Fairness and Ethics
What we need

Skills

- Computer Science, Programming, Databases
- Statistics & Machine Learning
- Econometrics & Social Science Methods
- Experimental Design
- Ethics & Legal Issues
- Communication
- Problem Formulation
Eric & Wendy Schmidt Data Science for Social Good Summer Fellowship

http://dssg.uchicago.edu

@datascifellows
Contact Information

Rayid Ghani
Center for Data Science & Public Policy
University of Chicago
rayid@uchicago.edu

Data Science for Social Good Summer Fellowship
http://dssg.uchicago.edu

Center for Data Science & Public Policy
http://dsapp.uchicago.edu

Code at github.com/dssg