What Does Open Data Mean to Data Science?

Philip E. Bourne PhD, FACMI
Stephenson Chair of Data Science
Director, Data Science Institute
Professor of Biomedical Engineering
peb6a@virginia.edu
https://www.slideshare.net/pebourne
@pebourne
Let me answer the question with a story...

The case of the trauma surgeon...
Not convinced ... try this one...

- North Virginia Technology Council Announces a Hackathon
- Teams compete internally – undergrads led by Daniel Mietchen and Pete Alonso
- Team selected and competes against GMU, VT, VCU
- UVA wins!
The Problem

In 2014, the latest year available, more than 7,400 veterans took their own lives, accounting for 18 percent of all suicides in America.

Latest Estimate – 20 vets commit suicide every day

Source: U.S. Department of Veterans Affairs, US Census Bureau, data.gov
Veteran Demographics

Veteran Population

18.8 million  Total
17.2 million  Male
1.6 million  Female

Race Distribution

- White: 79%
- Other: 0%
- Hispanic: 1%
- Asian: 7%
- Amer. Indian/AN: 2%
- African-American: 11%
- 2+ races: 1%

Age Ranges

- 18-29: 5%
- 30-44: 14%
- 45-54: 15%
- 55-64: 23%
- 65+: 44%

Period Served

- Vietnam War: 33%
- Gulf War II: 11%
- Gulf War I: 12%
- World War II: 8%
- Korean War: 10%
- Peacetime: 25%
Veteran and General Suicide Victim Demographics

Suicide Rate by Gender

<table>
<thead>
<tr>
<th></th>
<th>General Population</th>
<th>Veteran Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>39.2</td>
<td>13.92</td>
</tr>
<tr>
<td>Male</td>
<td>41.8</td>
<td>21.83</td>
</tr>
<tr>
<td>Female</td>
<td>17.3</td>
<td>6.24</td>
</tr>
</tbody>
</table>

Suicide Rate by Age

<table>
<thead>
<tr>
<th></th>
<th>General Population</th>
<th>Veteran Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>15-34</td>
<td>15.3</td>
<td>70.4</td>
</tr>
<tr>
<td>35-54</td>
<td>18.5</td>
<td>47.7</td>
</tr>
<tr>
<td>55-74</td>
<td>17.0</td>
<td>30.4</td>
</tr>
<tr>
<td>75+</td>
<td>18.6</td>
<td>32.0</td>
</tr>
</tbody>
</table>

Suicide Rate by Race

Key

- General Population
- Veteran Population

<table>
<thead>
<tr>
<th>Race</th>
<th>General Population</th>
<th>Veteran Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian/Pacific Island</td>
<td>3%</td>
<td>1%</td>
</tr>
<tr>
<td>Am Indian/AK Nat</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>African-American</td>
<td>6%</td>
<td>5%</td>
</tr>
</tbody>
</table>

Source: 2012, 2014 data acquired from U.S. Department of Veterans Affairs, data.gov, 2014 Centers for Disease Control and Prevention reports

May 22, 2018
Correlation: Health Care by State

- ~2M Veterans lack health insurance
- 42% Unaware of VA benefits
- Complicated priority system (VA)
  - False PTSD diagnosis – est. 47,000 undiagnosed each year

Source: 2014 data acquired from Veterans Affairs and census.gov
Correlation: Social Isolation by State

Source: 2014 data acquired from Veterans Affairs, data.gov and census.gov
Correlation: Social Isolation as Measured by State

Percentage of Veterans in Rural and Urban Areas vs. Suicide Rate

Source: 2014 data acquired from Veterans Affairs, data.gov and census.gov
## Firearm Regulation

<table>
<thead>
<tr>
<th>State</th>
<th>Total Gun Laws</th>
<th>Ammunition Regulations</th>
<th>Background Checks</th>
<th>Buyer Regulations</th>
<th>Dealer Regulations</th>
<th>Gun Trafficking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arizona</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nevada</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>New Mexico</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Utah</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>All States</td>
<td>26.5</td>
<td>0.72</td>
<td>2.46</td>
<td>2.4</td>
<td>2.7</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Source: 2014 data acquired from Veterans Affairs, data.gov and census.gov
Firearm Access

Veteran Suicide

Firearm Regulation

Source: 2014 data acquired from U.S. Department of Veterans Affairs
**Data collected using various Python data mining methods and respective Social Media APIs**

May 22, 2018
Twitter Veteran Population Sentiment
(Hover for State Rating)
Recommendations

- Limit firearm possession based on mental health status
- Utilize social media data to increase outreach
- Have VA Resources complement private health care
- Community: Content Creation and Mental Health Training
Open data is driving data science which in turn is/will change the way we do everything...
Example - photography

Time

From a presentation to the Advisory Board to the NIH Director

Volume, Velocity, Variety

Digital camera invented by Kodak but shelved

Megapixels & quality improve slowly; Kodak slow to react

Digital media becomes bona fide form of communication

Instagram, Flickr become the value proposition

Democratization

Dematerialization

Demonetization

Disruption

Deception

Film market collapses; Kodak goes bankrupt

Phones replace cameras

“NDs” - Exponential Framework

Democratize

Dematerialize

Demonetize

Digitize
So what is the problem? ...

There is lots of data but it is hard to find and is not persistent ...
We are not FAIR

• Digital assets (objects) within that system are data, software, narrative, course materials etc.

• Assets are to varying degrees FAIR – Findable, Accessible, Interoperable and Reusable

FAIR: https://www.nature.com/articles/sdata201618
There is lots of data, but it gets lost quickly

• Big Data
  • Total data from NIH-funded research currently estimated at 650 PB*
  • 20 PB of that is in NCBI/NLM (3%) and it is expected to grow by 10 PB this year

• Dark Data
  • Only 12% of data described in published papers is in recognized archives – 88% is dark data^  

• Cost
  • 2007-2014: NIH spent ~$1.2Bn extramurally on maintaining data archives

* In 2012 Library of Congress was 3 PB
Funders and publishers come at this from a perspective of reproducibility ...
I Can't Reproduce My Own Work

It took several months to replicate this work.
The problem is more profound .. It inhibits doing the data science research in the first place ...
Stating the problem is easy..

What are some of the solutions?
Both funders and institutions see the need to move from pipes to platforms...
Example: NSF and NIH Approaches

The NIH Data Commons Pilot Phase is expected to span fiscal years 2017-2020, with an estimated total budget of approximately $55.5 Million, pending available funds.
What evidence is there that platforms work?

• Airbnb is a platform that supports a trusted relationship between consumer (renter) and supplier (host)

• The platform focuses on maximizing the exchange of services between supplier and consumer and maximizing the amount of trust associated with a given stakeholder

• It seems to be working:
  • 60 million users searching 2 million listings in 192 countries
  • Average of 500,000 stays per night.
  • Evaluation of US $25bn

Bonazzi & Bourne 2017 PLOS Biology 15(4) e2001818
Open Data Lab

Storage Layer
- Amazon Web Services S3
- GitHub

Development Layer
- R
- PyCharm
- Jupyter

HPC Layer
- Amazon Web Services EC2
- Local UVA
  - Rivanna
  - Ivy
  - Project X

Discovery Layer
- Zenodo
- The Dataverse
- WIKIDATA

Container Layer
- Docker
- Tectonic by CoreOS
- Rancher
Platforms support 4 of the 5 pillars of data science

Data Acquisition  Data Integration & Engineering  Machine Learning & Analytics  Visualization & Dissemination  Ethics, Law, Policy, Social Implications
In summary

• Open data defines much of the new economy and contributes to social good
• This may be incentive enough
• We are all part of this fourth paradigm
• To fully realize the potential of open data we must be FAIR
• We need to breakdown silos – platforms help
Since libraries are experienced with open knowledge for the public good they have a key role to play. But how?