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

UNIVERSITY OF VIRGINIA DATA SCIENCE INSTITUTE

BIOMEDICAL ENGINEERING

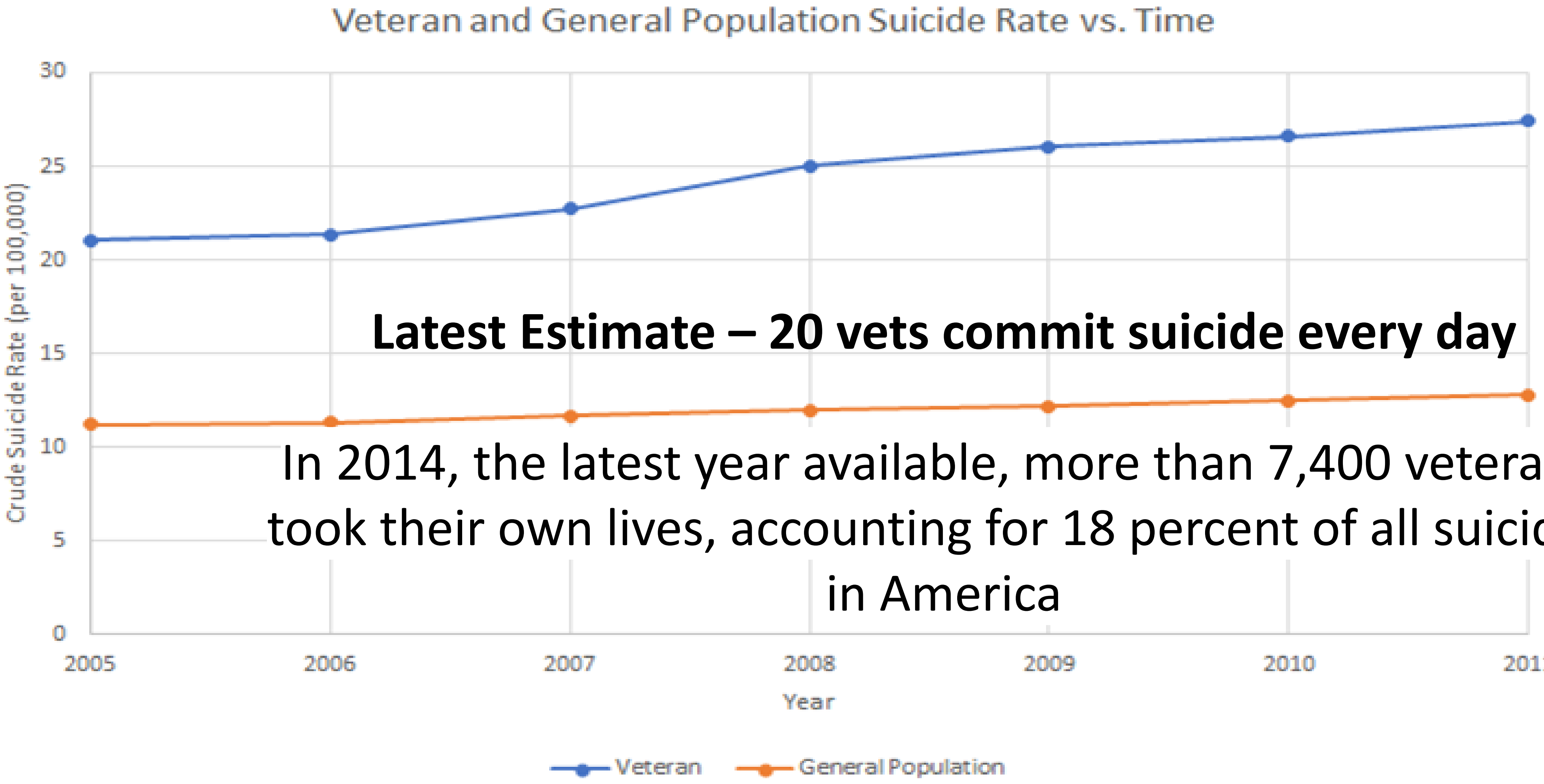
Let me answer the question with a story…

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



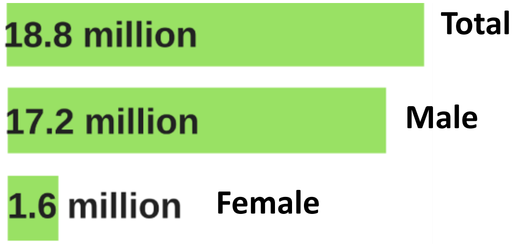
Latest Estimate – 20 vets commit suicide every day

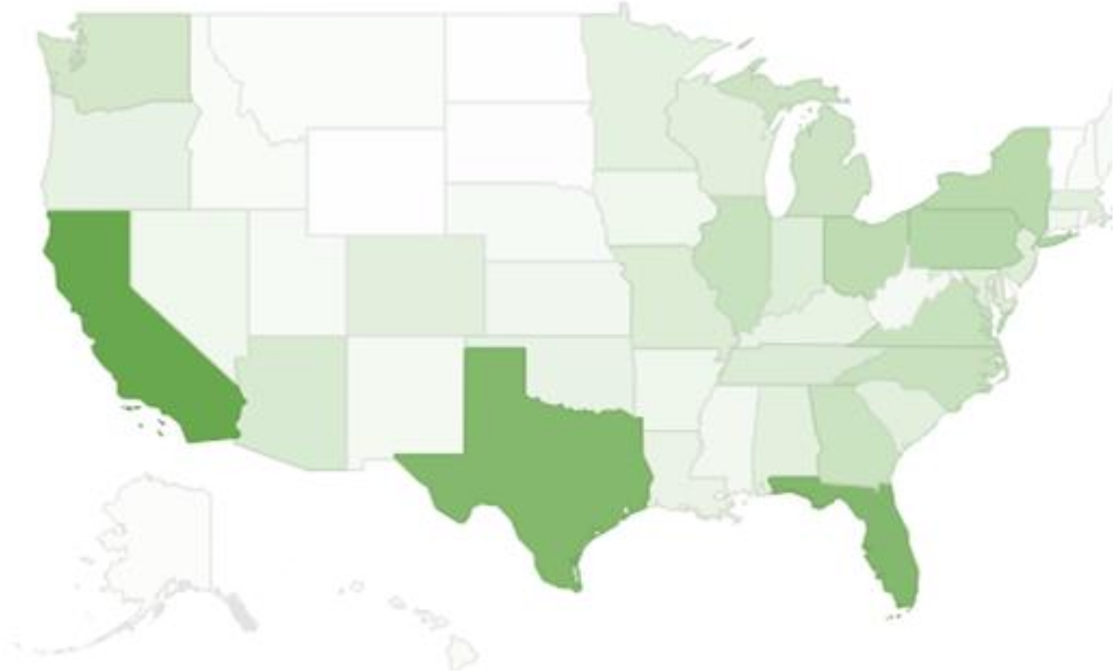
In 2014, the latest year available, more than 7,400 veterans took their own lives, accounting for 18 percent of all suicides in America

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

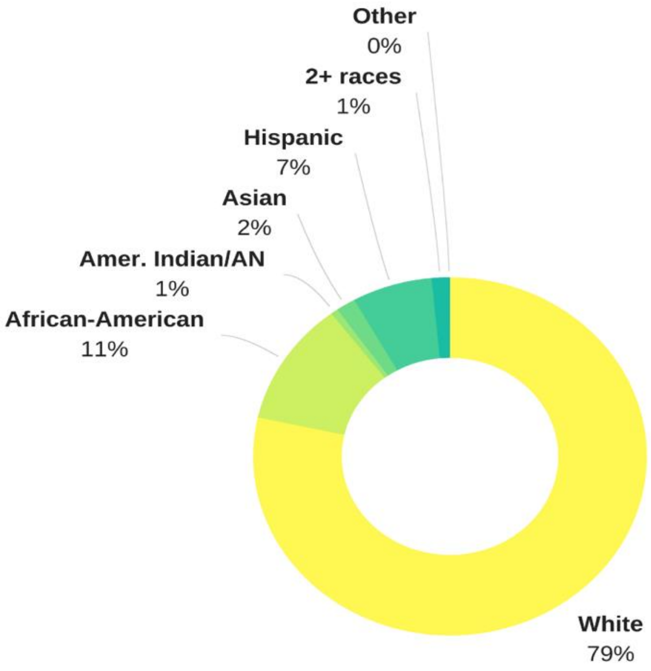
Veteran Demographics

Veteran Population

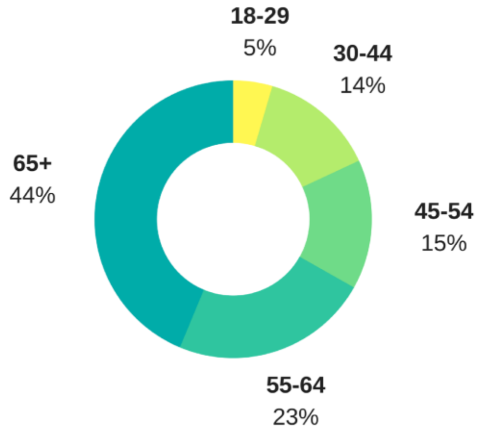




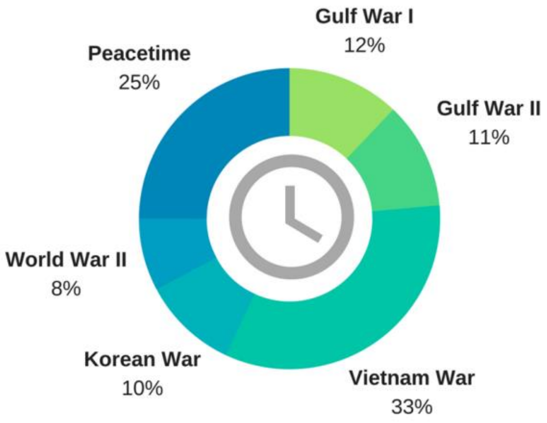
Race Distribution



Age Ranges



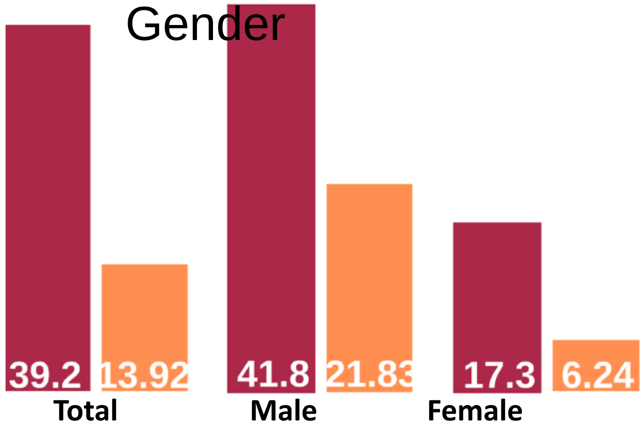
Period Served



Source: 2012, 2014, 2015 Data acquired from U.S. Department of Veterans Affairs, US Census Bureau, Department of Defense

Veteran and General Suicide Victim Demographics

Suicide Rate by Gender

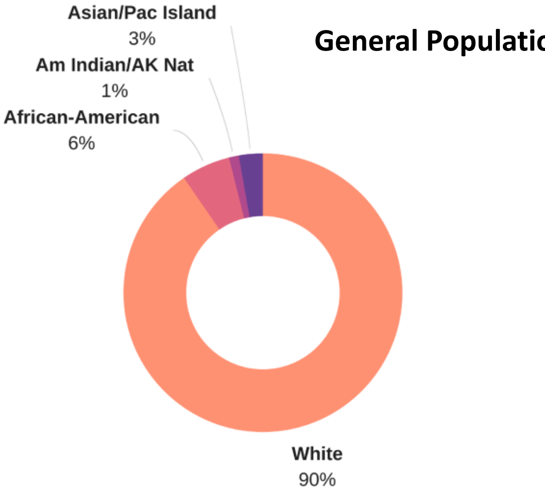


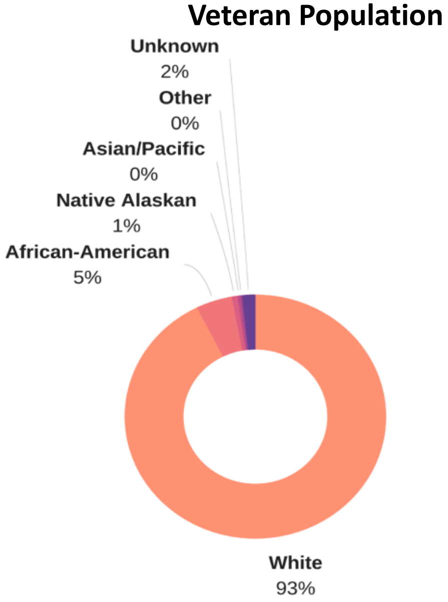
Key

🟧 General Population

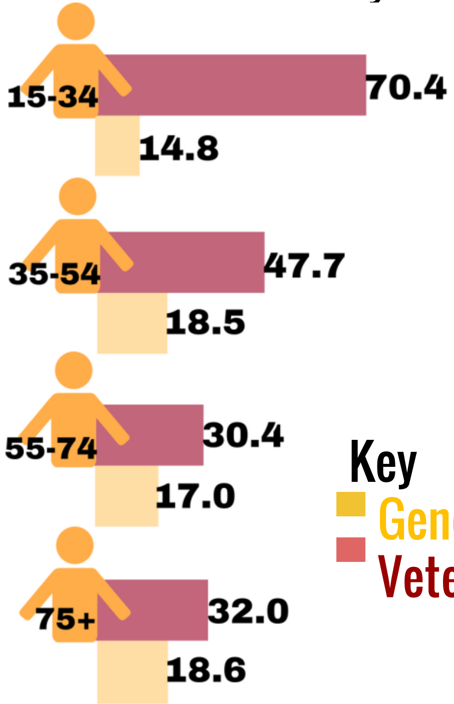
🟥 Veteran Population

Suicide by Race





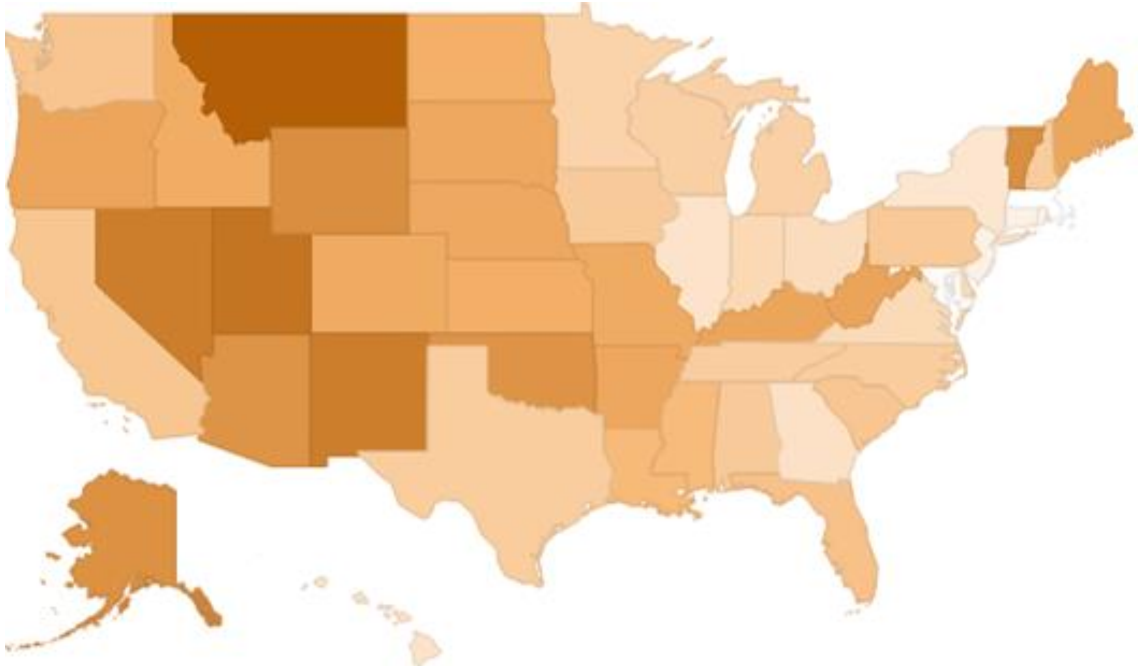
Suicide Rate by



Key

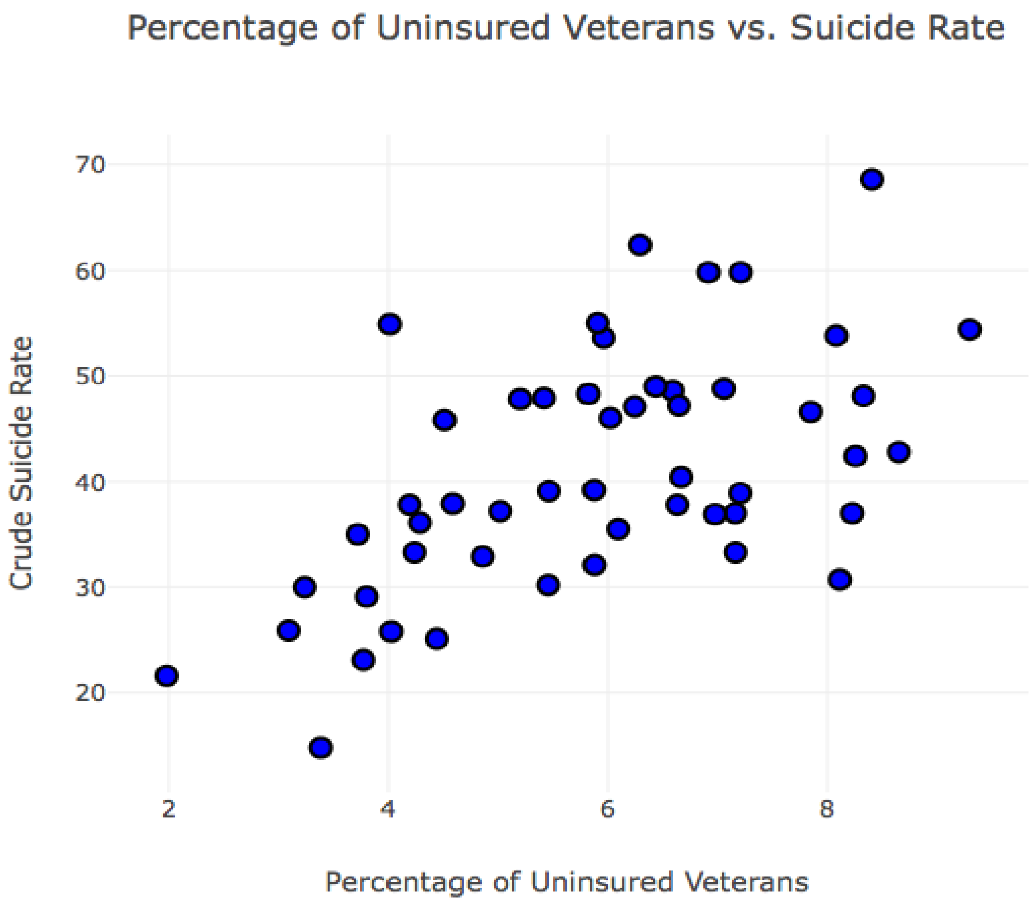
🟨 General Population

🟥 Veteran Population



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

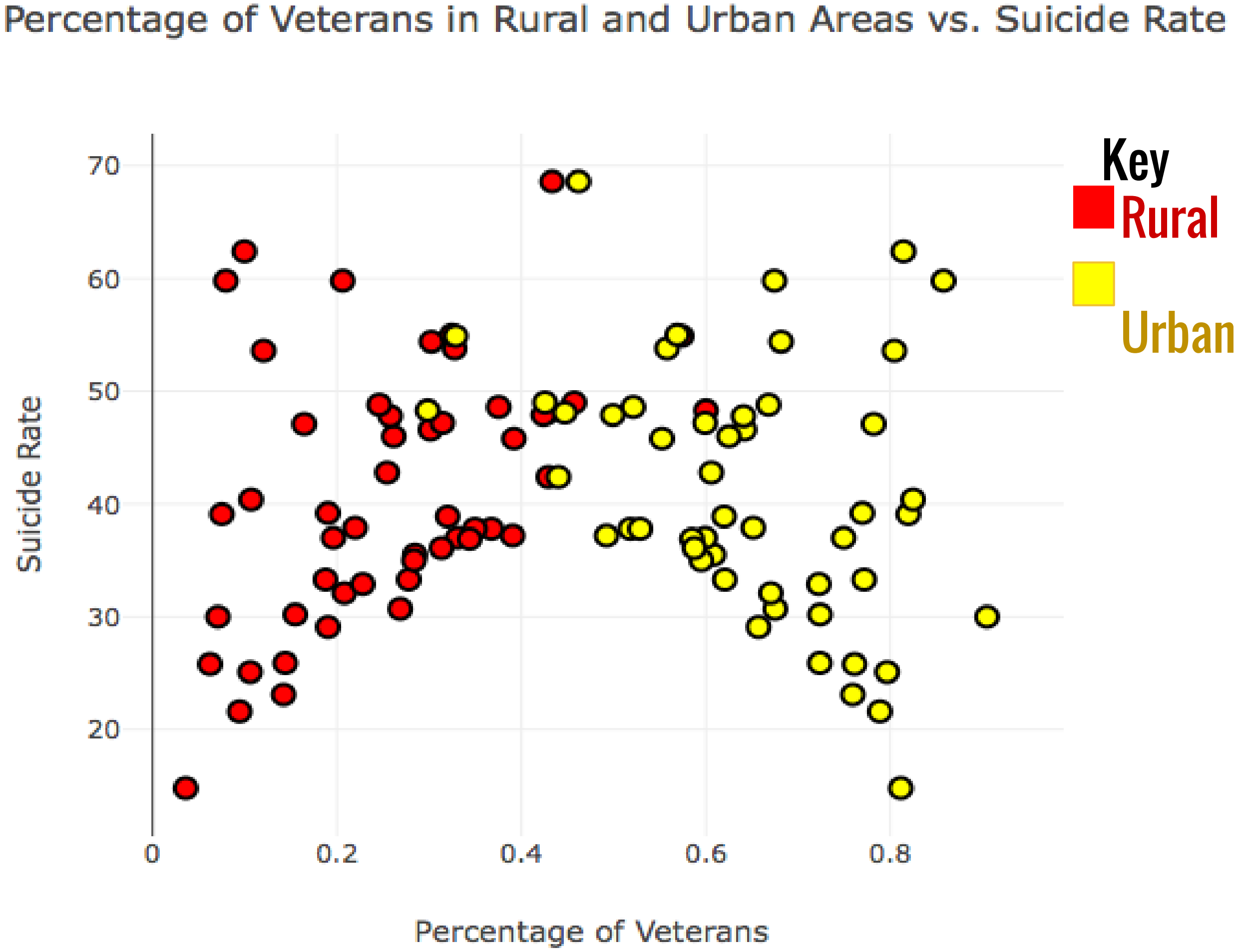
Correlation: Health Care by State



- ~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

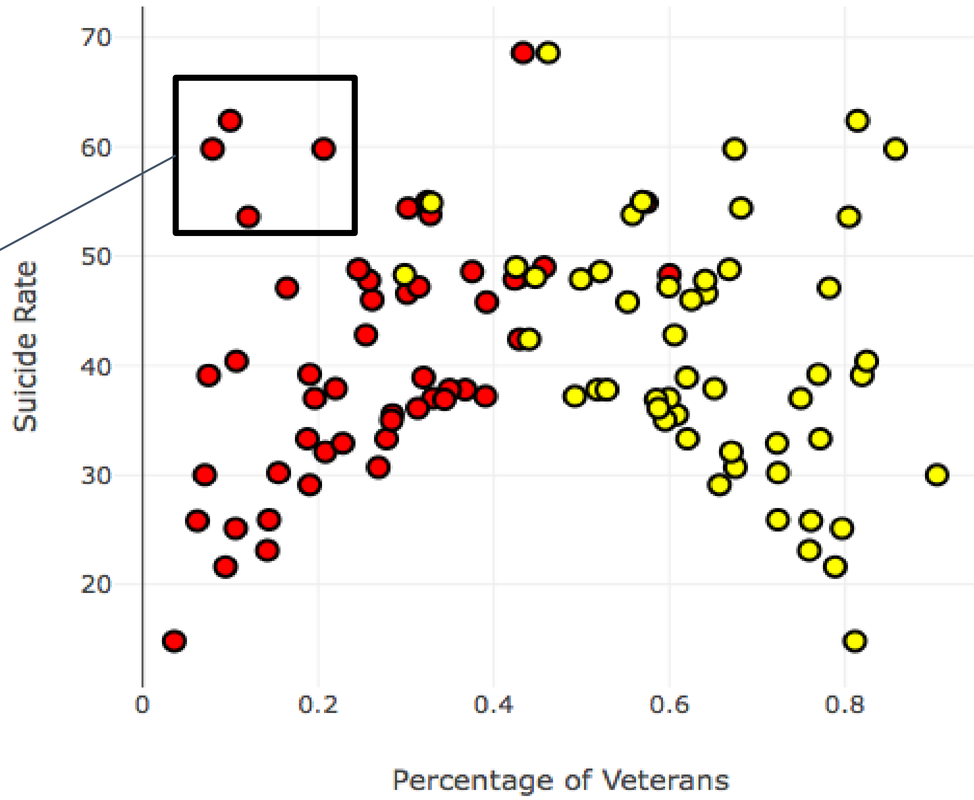


Percentage of Veterans in Rural and Urban Areas vs. Suicide Rate

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



Utah
Arizona
New Mexico
Nevada

Key

🟥 Rural

🟨 Urban

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

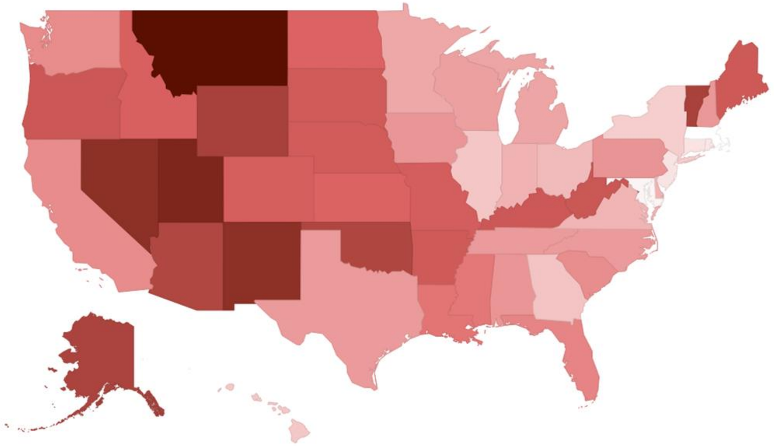
Firearm Regulation

State	Total Gun Laws	Ammunition Regulations	Background Checks	Buyer Regulations	Dealer Regulations	Gun Trafficking
Arizona	11	0	0	0	0	0
Nevada	11	1	0	0	0	0
New Mexico	10	0	0	0	0	0
Utah	11	0	0	0	0	2
All States	26.5	0.72	2.46	2.4	2.7	0.76

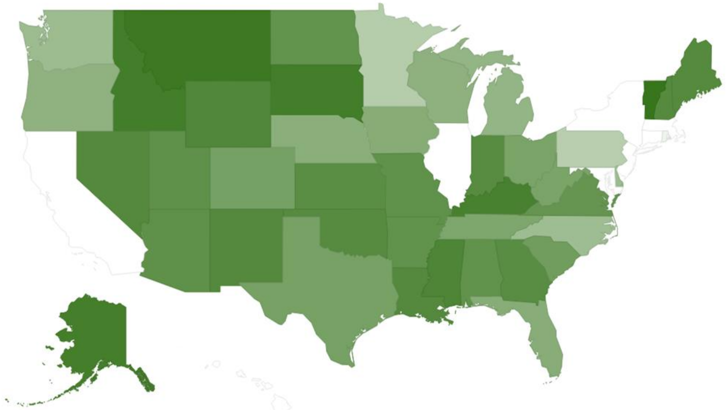
Source: 2014 data acquired from Veterans Affairs, data.gov and census.gov

Firearm Access

Veteran Suicide

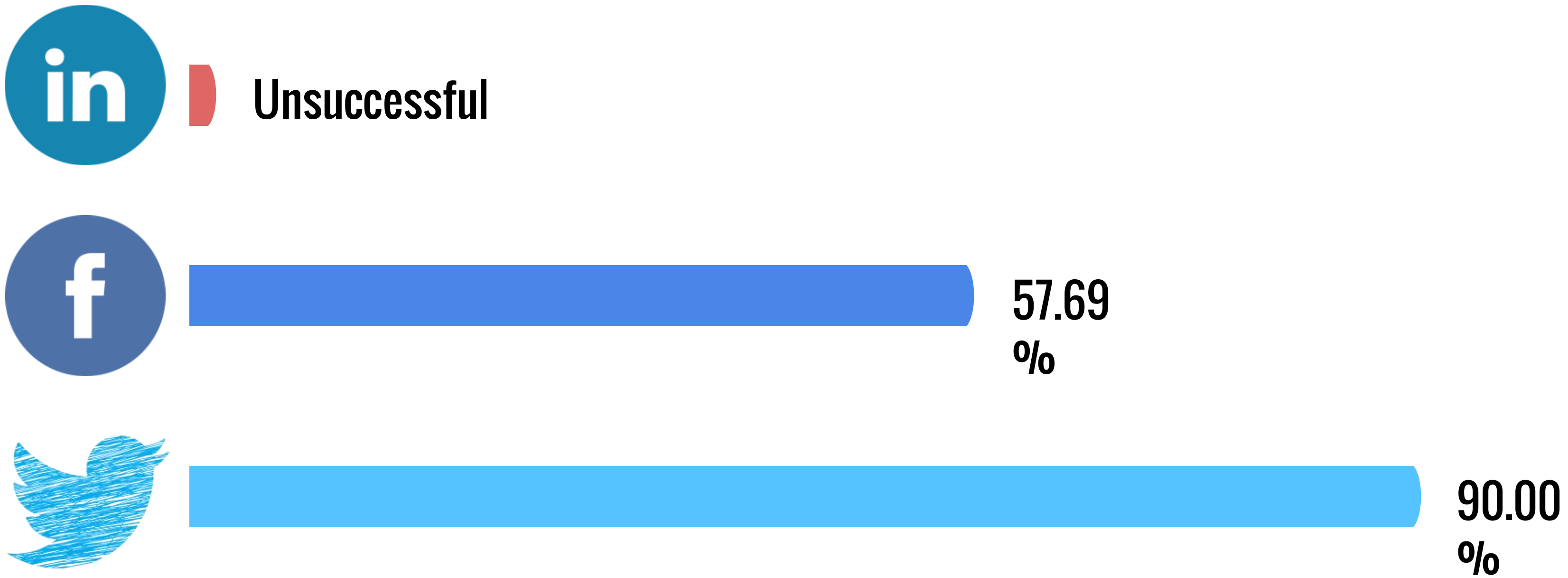


Firearm Regulation

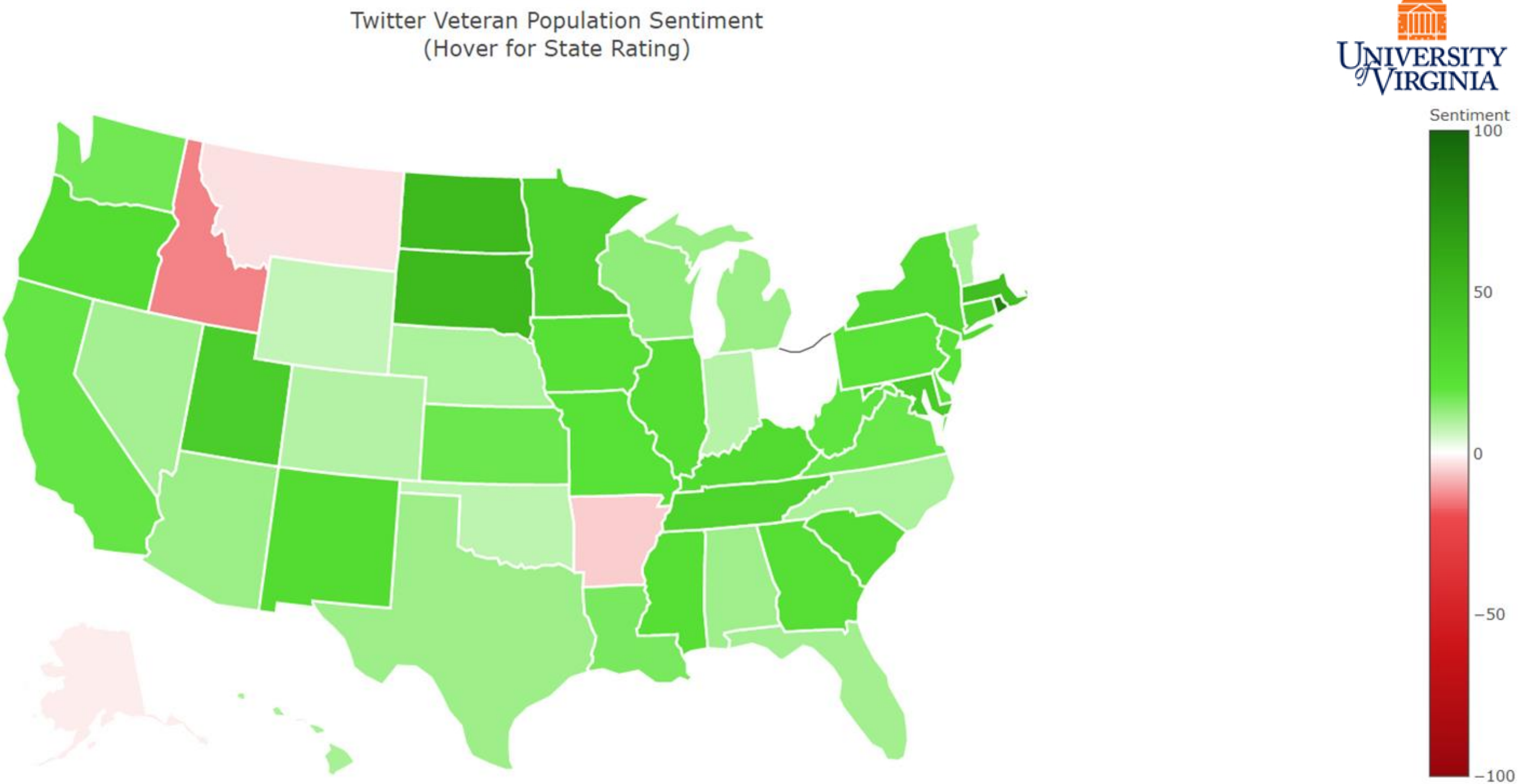


Source: 2014 data acquired from U.S. Department of Veterans Affairs

Social Media: Mining APIs and URLs



**Data collected using various Python data mining methods and respective Social Media APIs


Twitter Veteran Population Sentiment
(Hover for State Rating)
Sentiment
100
50
0
−50
−100

Recommendations

Limit firearm possession based on mental health status

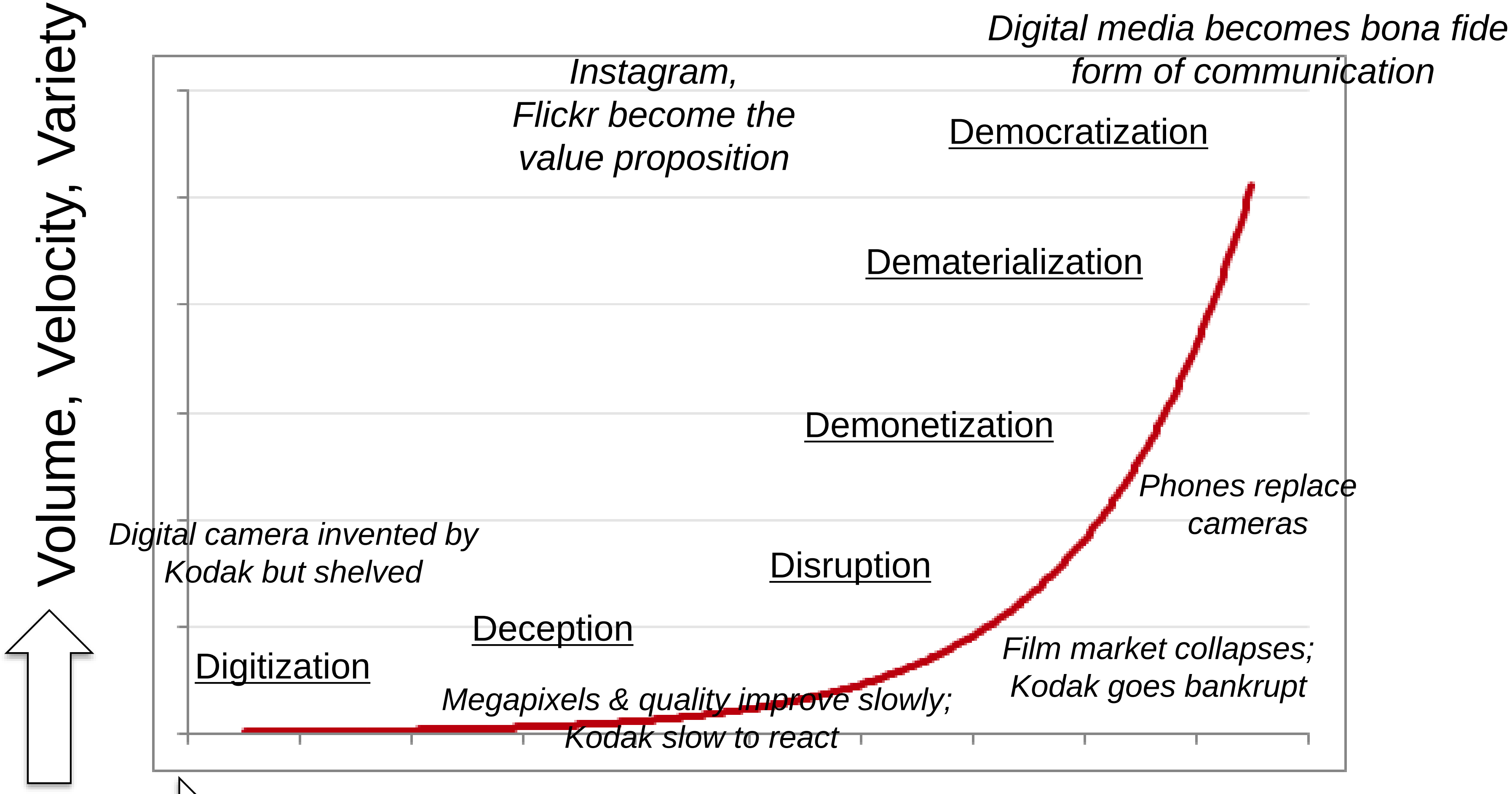
Have VA Resources complement private health care

Community: Content Creation and Mental Health Training

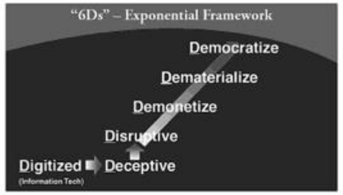
Utilize social media data to increase outreach

Open data is driving data science which in turn is/will change the way we do everything…

Example - photography



From a presentation to the Advisory Board to the NIH Director



The 6 Ds of Exponentials: Digitalization, Deception, Disruption, Demonetization, Dematerialization, and Democratization

Source: Peter H. Diamandis, www.abundancehub.com

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

https://www.workitdaily.com/job-search-solution/

There is lots of data, but it gets lost quickly

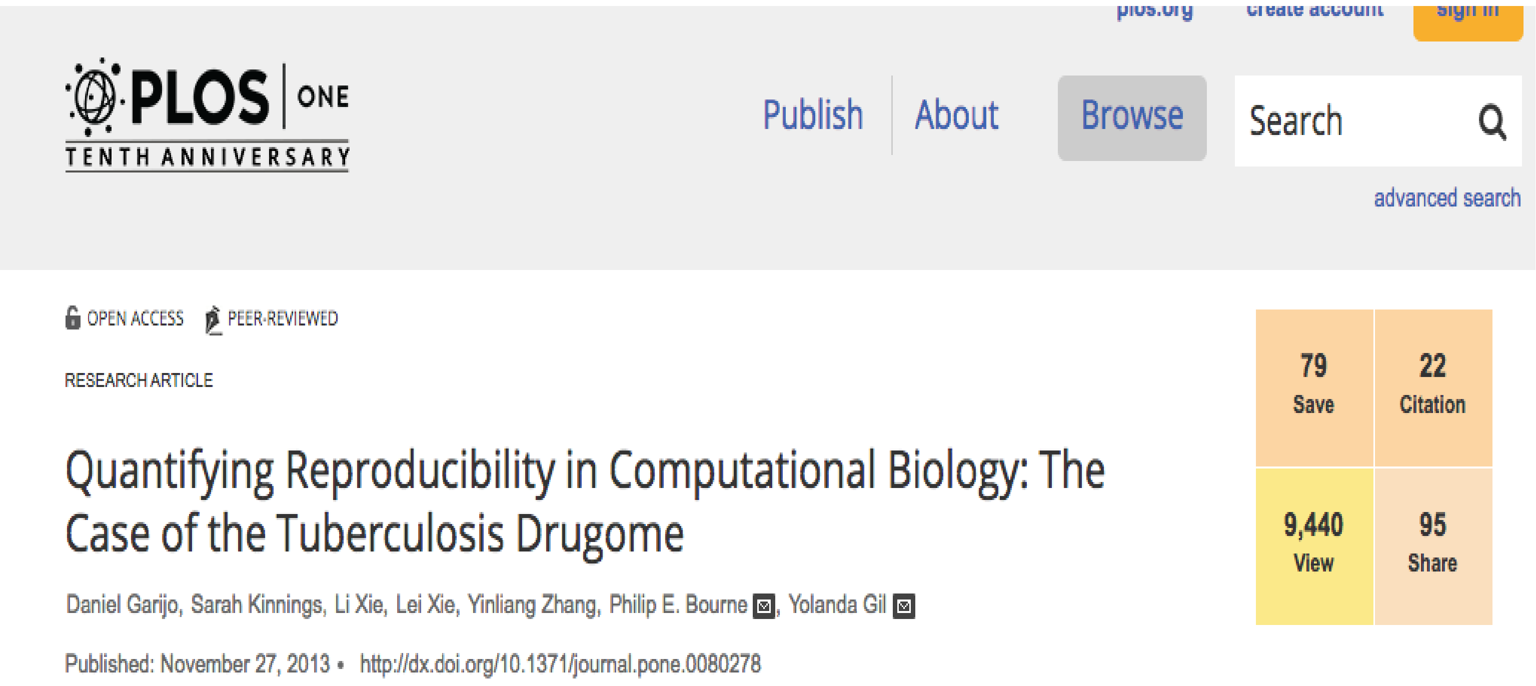
- Big Data
  - Total data from NIH-funded research currently estimated at 650 PB*
  - 20 PB of that is in NCBI/NLM (3%) and it is expected to grow by 10 PB this year
- Dark Data
  - Only 12% of data described in published papers is in recognized archives – 88% is dark data^
- Cost
  - 2007-2014: NIH spent ~$1.2Bn extramurally on maintaining data archives

* In 2012 Library of Congress was 3 PB

^ http://www.ncbi.nlm.nih.gov/pubmed/26207759

Funders and publishers come at this from a perspective of reproducibility …

I Cant Reproduce My Own Work



RESEARCH ARTICLE

Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome

Daniel Garijo, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne, Yolanda Gil

Published: November 27, 2013 • http://dx.doi.org/10.1371/journal.pone.0080278

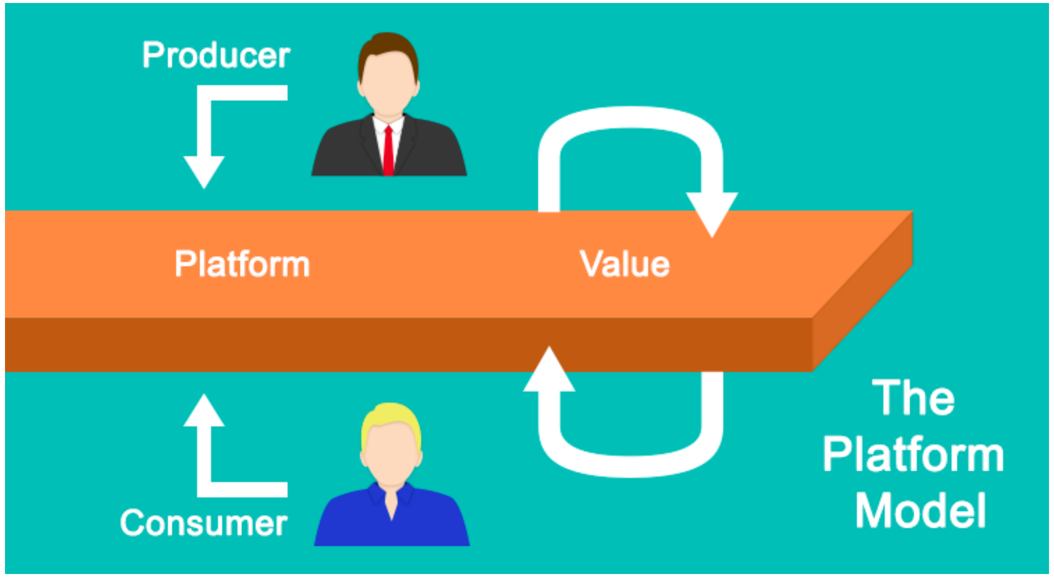
It took several months to replicate this work

The problem is more profound .. It inhibits doing the data science research in the first place …

Stating the problem is easy ..

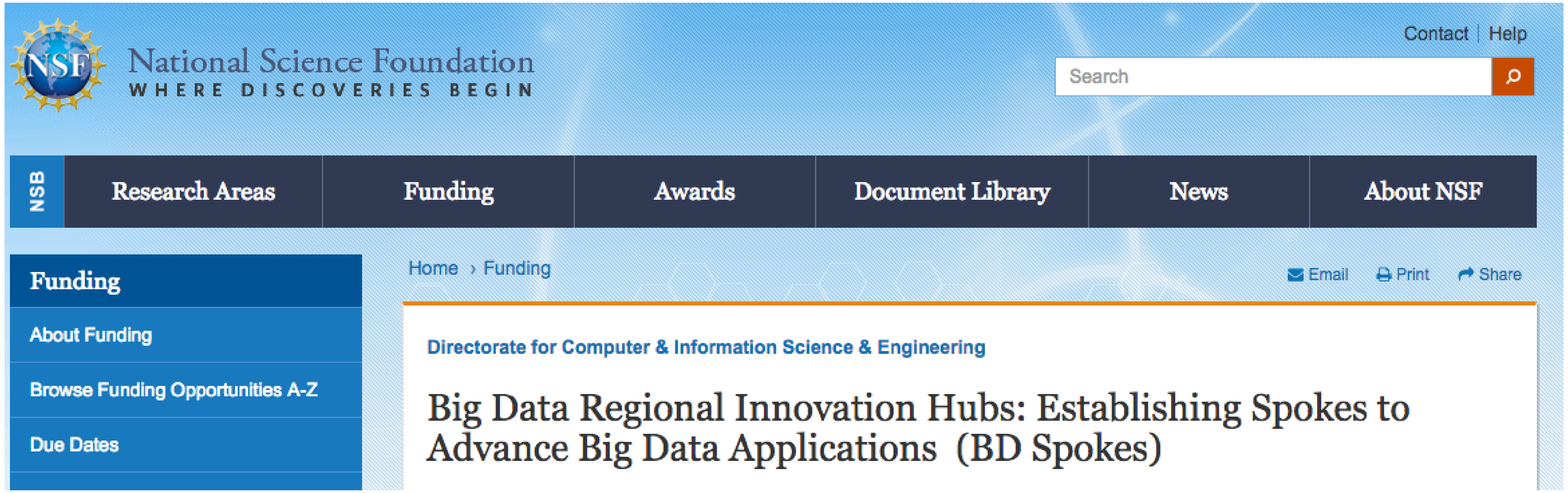
What are some of the solutions?

Both funders and institutions see the need to move from pipes to platforms…



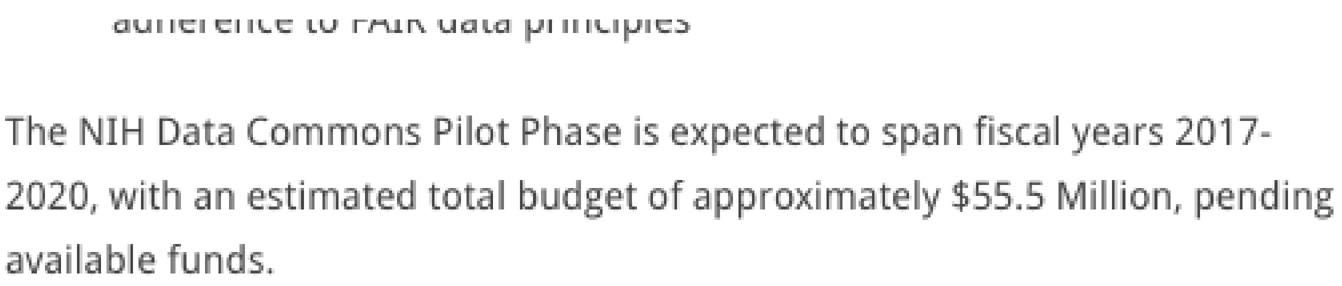
https://blog.lexicata.com/wp-content/uploads/2015/03/platform-model-750x410.png

Example: NSF and NIH Approaches



Directorate for Computer & Information Science & Engineering

Big Data Regional Innovation Hubs: Establishing Spokes to Advance Big Data Applications (BD Spokes)



adherence to FAIR data principles

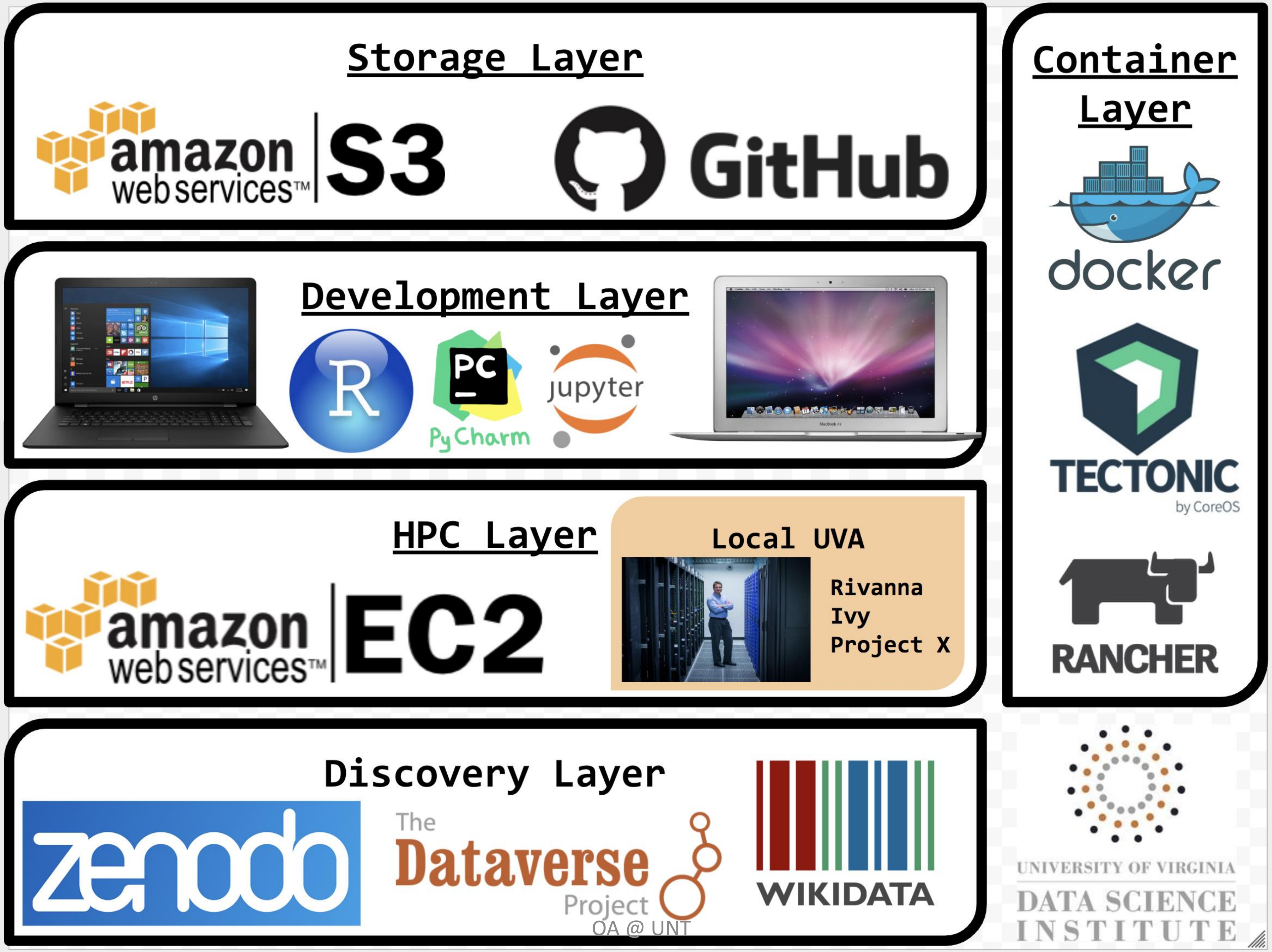
The NIH Data Commons Pilot Phase is expected to span fiscal years 2017-2020, with an estimated total budget of approximately $55.5 Million, pending available funds.

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

Development Layer

HPC Layer

Local UVA

Rivanna
Ivy
Project X

Discovery Layer

Container Layer

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?