

'Big Data' in the Social Sciences

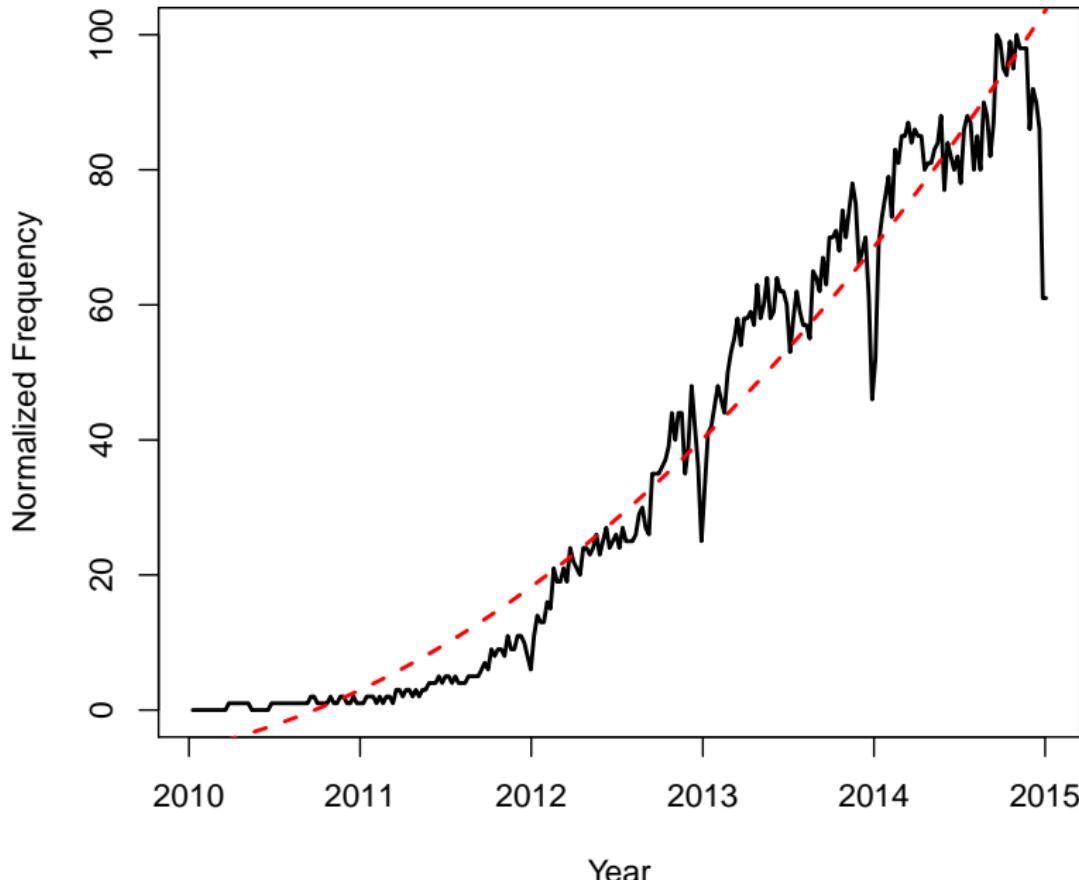
Christopher Zorn
`zorn@psu.edu`

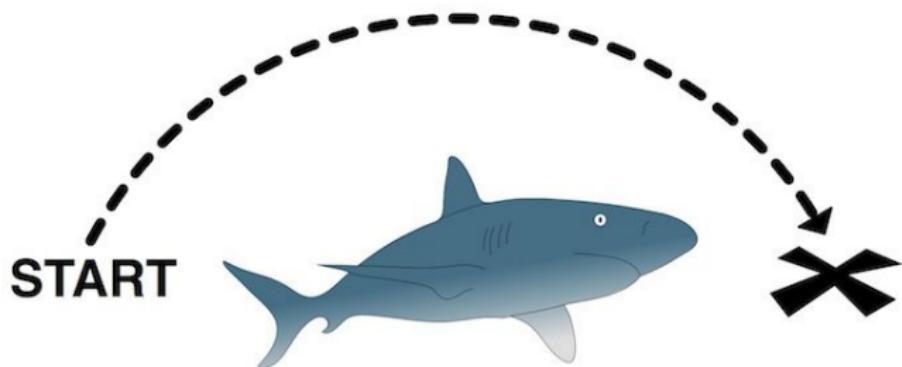
Workshop for Teachers of Quantitative
Methods for Social Scientists

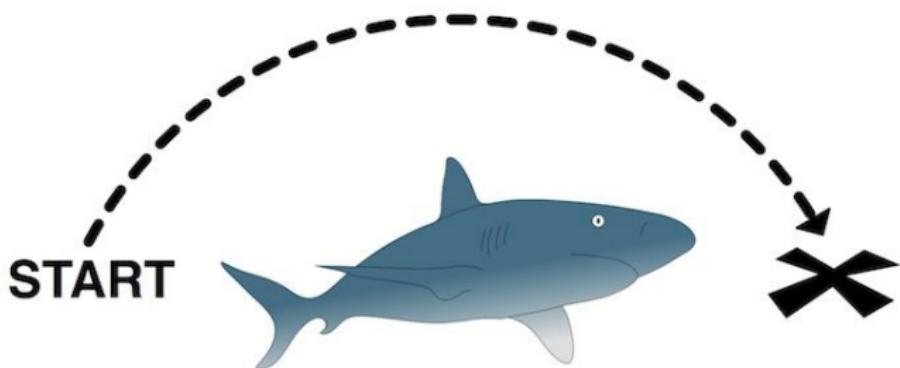
University of Oxford, January 8, 2015

Big

Google Trends Figures for 'big data'



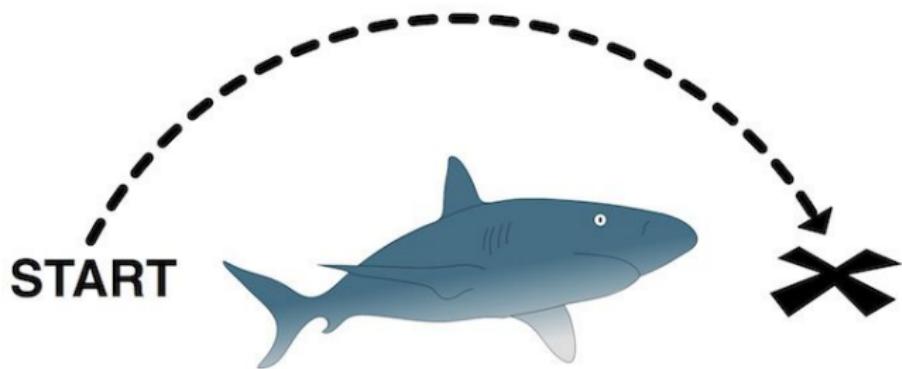




Google 0 🔍

[Web](#) [News](#) [Videos](#) [Images](#) [Shopping](#) [More ▾](#) [Search tools](#)

About 65,200 results (0.31 seconds)



Pictures of Big Data

Big Data is visualized in so many ways...all of them blue and with numbers and lens flare.

Three things:

- Data
- Tools
- Philosophy

1. Data

Data: Format

What we think of:

name	age	height
Chris	46	72
Zaryab	41	60
Evan	5	44

Data: Format

What we think of:

name	age	height
Chris	46	72
Zaryab	41	60
Evan	5	44

JSON:

```
[{"name": "Chris", "age": 46, "height": 72},  
 {"name": "Zaryab", "age": 41, "height": 60},  
 {"name": "Evan", "age": 5, "height": 44}]
```

Data: Format

What we think of:

name	age	height
Chris	46	72
Zaryab	41	60
Evan	5	44

XML:

```
<?xml version="1.0" encoding="UTF-8"?>
<rows>
    <row name age height="Chris 46 72" ></row>
    <row name age height="Zaryab 41 60" ></row>
    <row name age height="Evan 5 44" ></row>
</rows>
```

Data: Sources

- Open government /official data sources
- APIs / streaming data (Google, Twitter, etc.)
- MTurk / crowdsourcing
- “Data exhaust”

Official Data: E.g., Stata's wbopendata

The screenshot shows the World Bank Data homepage. At the top, there is a navigation bar with links for Home, About, Data, Research, Learning, News, Projects & Operations, Publications, Countries, and Topics. The 'Data' link is highlighted with a red background. Below the navigation bar is a red banner with the word 'Data'. Underneath the banner is a horizontal menu with links for By Country, By Topic, Indicators, Data Catalog, Microdata, Initiatives, What's New, Support, and Products. The main content area features a title 'Accessing World Bank Open Data in Stata' with a 'SHARE' button. A sidebar on the right is titled 'Recent News Posts' and lists two items: 'International Debt Statistics 2015 now available' (12/16/2014 - 12:47) and '2013 Global and Regional Estimates for Child Malnutrition Released' (11/19/2014 - 14:42). The main content area contains text about the wbopendata module and its benefits.

THE WORLD BANK
IBRD • IDA

Working for a World Free of Poverty

English Español Français フレンチ Русский 中文

Search

Home About Data Research Learning News Projects & Operations Publications Countries Topics

Data

By Country By Topic Indicators Data Catalog Microdata Initiatives What's New Support Products

This page in English | Español | Français | العربية | 中文

Accessing World Bank Open Data in Stata

Posted on 03/15/2012 - 09:54

Stata is a statistical computing package widely used in the business and academic worlds. We use it at the World Bank and it's great to see a new version of the `wbopendata`_module that gives Stata users direct access to much of the data on data.worldbank.org.

Academic institutions and hundreds of users are already taking advantage of it – why not give it a try?

Recent News Posts

International Debt Statistics 2015 now available
12/16/2014 - 12:47

2013 Global and Regional Estimates for Child Malnutrition Released
11/19/2014 - 14:42

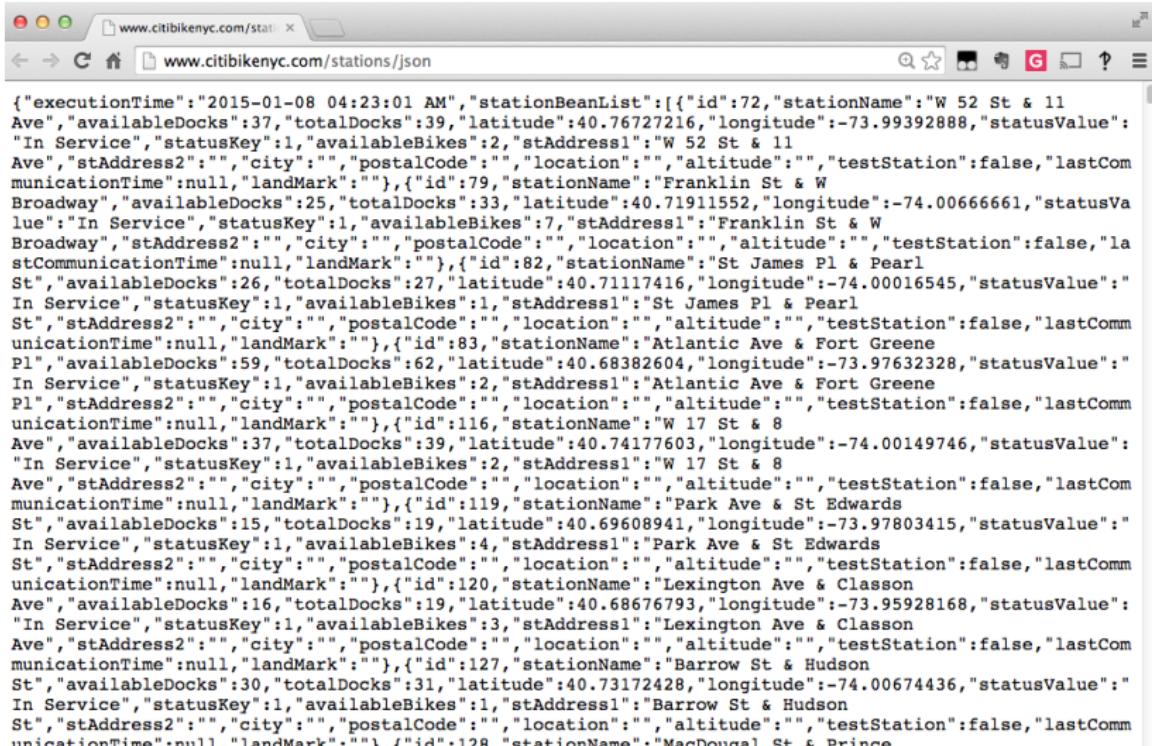
<http://data.worldbank.org/news/accessing-world-bank-open-data-in-stata>

Sources: APIs / Streaming Data

API = *Application programming interface*

- “...a set of programming instructions and standards for accessing a Web-based software application or Web tool.”
- Data portal / stream
- Examples:
 - United Nations Data
 - CitiBikeNYC API (real-time JSON data)
 - Others...

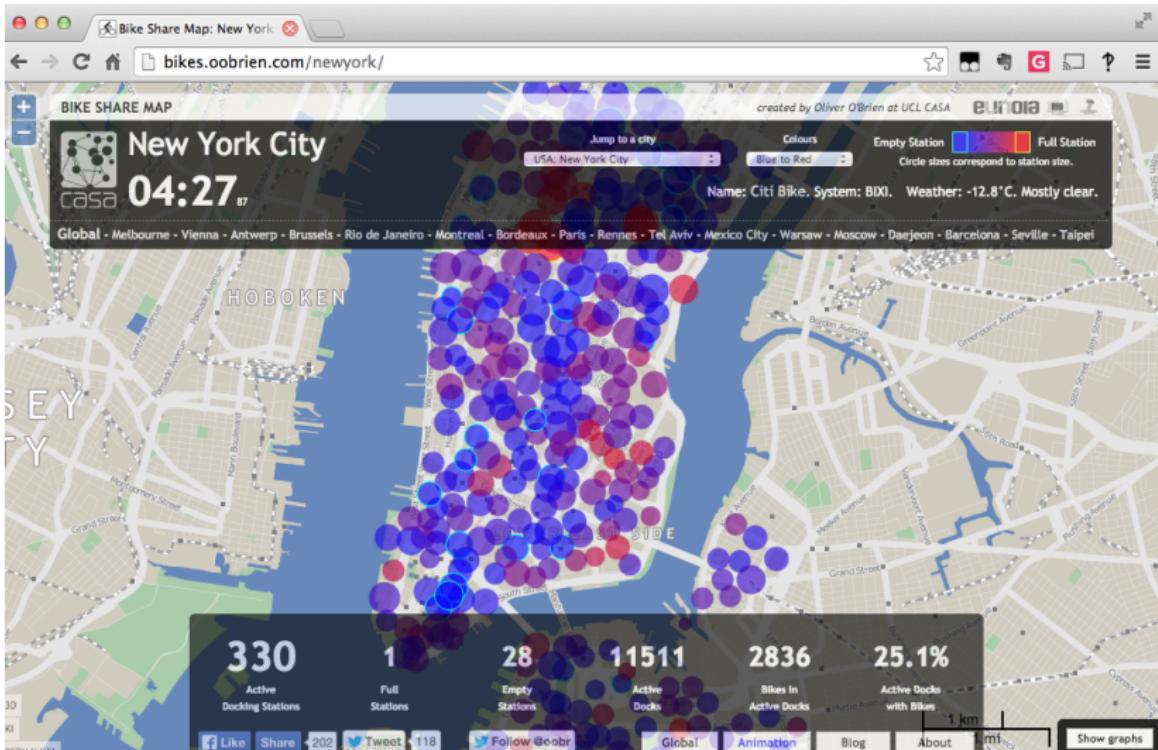
CitiBikeNYC Data



A screenshot of a Mac OS X desktop showing a web browser window. The address bar contains the URL www.citibikenyc.com/stations/json. The main content area of the browser displays a large block of JSON data representing CitiBike station information.

```
{"executionTime": "2015-01-08 04:23:01 AM", "stationBeanList": [{"id": 72, "stationName": "W 52 St & 11 Ave", "availableDocks": 37, "totalDocks": 39, "latitude": 40.76727216, "longitude": -73.99392888, "statusValue": "In Service", "statusKey": 1, "availableBikes": 2, "stAddress1": "W 52 St & 11 Ave", "stAddress2": "", "city": "", "postalCode": "", "location": "", "altitude": "", "testStation": false, "lastCommunicationTime": null, "landMark": ""}, {"id": 79, "stationName": "Franklin St & W Broadway", "availableDocks": 25, "totalDocks": 33, "latitude": 40.71911552, "longitude": -74.00666661, "statusValue": "In Service", "statusKey": 1, "availableBikes": 7, "stAddress1": "Franklin St & W Broadway", "stAddress2": "", "city": "", "postalCode": "", "location": "", "altitude": "", "testStation": false, "lastCommunicationTime": null, "landMark": ""}, {"id": 82, "stationName": "St James Pl & Pearl St", "availableDocks": 26, "totalDocks": 27, "latitude": 40.71117416, "longitude": -74.00016545, "statusValue": "In Service", "statusKey": 1, "availableBikes": 1, "stAddress1": "St James Pl & Pearl St", "stAddress2": "", "city": "", "postalCode": "", "location": "", "altitude": "", "testStation": false, "lastCommunicationTime": null, "landMark": ""}, {"id": 83, "stationName": "Atlantic Ave & Fort Greene Pl", "availableDocks": 59, "totalDocks": 62, "latitude": 40.68382604, "longitude": -73.97632328, "statusValue": "In Service", "statusKey": 1, "availableBikes": 2, "stAddress1": "Atlantic Ave & Fort Greene Pl", "stAddress2": "", "city": "", "postalCode": "", "location": "", "altitude": "", "testStation": false, "lastCommunicationTime": null, "landMark": ""}, {"id": 116, "stationName": "W 17 St & 8 Ave", "availableDocks": 37, "totalDocks": 39, "latitude": 40.74177603, "longitude": -74.00149746, "statusValue": "In Service", "statusKey": 1, "availableBikes": 2, "stAddress1": "W 17 St & 8 Ave", "stAddress2": "", "city": "", "postalCode": "", "location": "", "altitude": "", "testStation": false, "lastCommunicationTime": null, "landMark": ""}, {"id": 119, "stationName": "Park Ave & St Edwards St", "availableDocks": 15, "totalDocks": 19, "latitude": 40.69608941, "longitude": -73.97803415, "statusValue": "In Service", "statusKey": 1, "availableBikes": 4, "stAddress1": "Park Ave & St Edwards St", "stAddress2": "", "city": "", "postalCode": "", "location": "", "altitude": "", "testStation": false, "lastCommunicationTime": null, "landMark": ""}, {"id": 120, "stationName": "Lexington Ave & Classon Ave", "availableDocks": 16, "totalDocks": 19, "latitude": 40.68676793, "longitude": -73.95928168, "statusValue": "In Service", "statusKey": 1, "availableBikes": 3, "stAddress1": "Lexington Ave & Classon Ave", "stAddress2": "", "city": "", "postalCode": "", "location": "", "altitude": "", "testStation": false, "lastCommunicationTime": null, "landMark": ""}, {"id": 127, "stationName": "Barrow St & Hudson St", "availableDocks": 30, "totalDocks": 31, "latitude": 40.73172428, "longitude": -74.00674436, "statusValue": "In Service", "statusKey": 1, "availableBikes": 1, "stAddress1": "Barrow St & Hudson St", "stAddress2": "", "city": "", "postalCode": "", "location": "", "altitude": "", "testStation": false, "lastCommunicationTime": null, "landMark": ""}, {"id": 128, "stationName": "MacDougal St & Prince"}]
```

CitiBikeNYC: Data → Maps



Programmable Web (API clearinghouse)

The screenshot shows the ProgrammableWeb website. At the top, there's a navigation bar with links for API News, API Directory, For API Providers, For Developers, Listings, and Forum. Below the navigation is a promotional banner for a 'FREE Infrared Automation Guidebook'. The main content area features a large heading 'ProgrammableWeb: the world's largest API repository, GROWING DAILY'. Below this are search and filter options, and a table listing APIs. To the right, there's a sidebar for the API Directory Search.

Follow ProgrammableWeb to get API news and alerts as they break

+Follow Share Sign in/Sing Up

ProgrammableWeb API News API Directory For API Providers For Developers Listings Forum

FREE Infrared Automation Guidebook
The Ultimate IR Resource Guide for Automated Processes

Get the Guide

FLIR

ProgrammableWeb: the world's largest API repository, GROWING DAILY

Search Over 12,669 APIs

Search APIs

Filter APIs

By Category By Protocols/Formats Include Deprecated APIs

API Name	Description	Category	Updated
Google Maps	The Google Maps API allow for the embedding of Google Maps onto web pages of outside developers, using a simple JavaScript interface or a Flash interface. It is designed to work on both mobile	Mapping	12.05.2005

API Directory Search

Search over 12,669 APIs updated daily

Search APIs, mashups, developers

Browse by Category

Newest APIs

Latest Mashups

Add an API +

<http://www.programmableweb.com/apis/directory>

General packages:

- RCurl, httr
- JSON: jsonlite, rjson
- XML: XML, R4X

R Tools

General packages:

- RCurl, httr
- JSON: jsonlite, rjson
- XML: XML, R4X

Specific API data tools:

- Google: RGA, googleVis, translate, others
- twitteR (Twitter)
- Rook / Rfacebook (Facebook)
- tumblrR (Tumblr)
- Others: dvn (Dataverse), WDI (World Bank), etc. (see <http://cran.r-project.org/web/views/WebTechnologies.html>)

Simple API Example: Twitter

Setup:

```
library(httr)
library(twitteR)

reqURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
APIKey<-"ieDpmBzQ1khjVlbrR5RoAZL09"
APISecret<-"PnhfyeA3XbN0xMswN68z9hxThm5e5Z1AQ9dTYjvQllf0VGQQwk"
twitCred <- OAuthFactory$new(consumerKey=APIKey,
                               consumerSecret=APISecret,
                               requestURL=reqURL,
                               accessURL=accessURL,
                               authURL=authURL)

twitCred$handshake
```

Simple API Example: Twitter



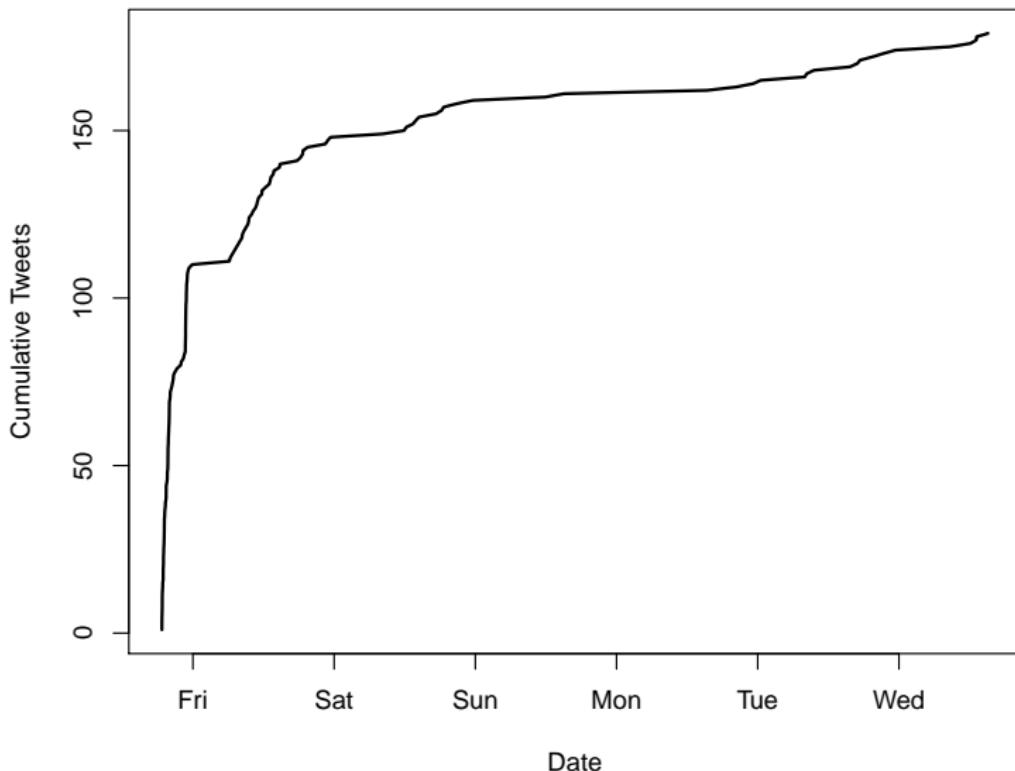
Simple API Example: Twitter

Pull “local” tweets:

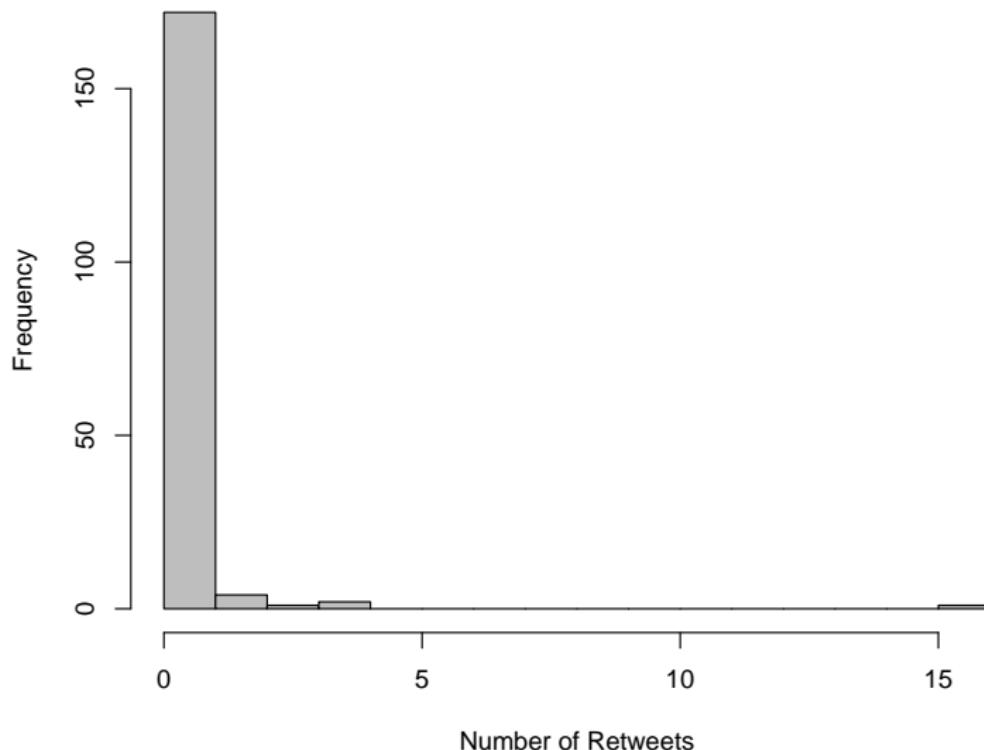
```
KaneTweets<-searchTwitter('Kane', geocode='51.603211,  
-0.066430, 10mi', since='2014-12-31',  
n=5000, retryOnRateLimit=1)
```

```
KaneTweetsData<-twListToDF(KaneTweets)
```

Cumulative Number of Tweets



Histogram: Number of Retweets



Harry Kane Word Cloud

<http://t.co/do7HdxNNvg>



Harry Kane Word Cloud

<http://t.co/do7HdxNNvg>



“Data Exhaust”



“Data Exhaust”

- Large public records



“Data Exhaust”

- Large public records
- Geotags / date stamps / etc.



“Data Exhaust”

- Large public records
- Geotags / date stamps / etc.
- Commercial / financial transactions



“Data Exhaust”

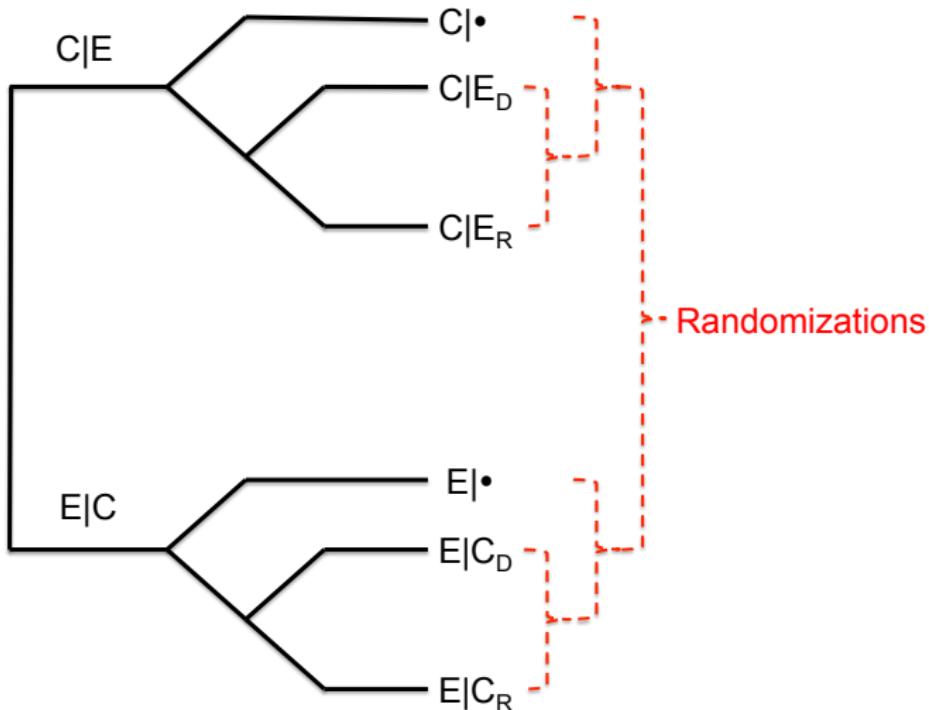
- Large public records
- Geotags / date stamps / etc.
- Commercial / financial transactions
- **Linkage** data (e.g., *your friends' locations*)



Crowdsourcing



Crowdsourcing

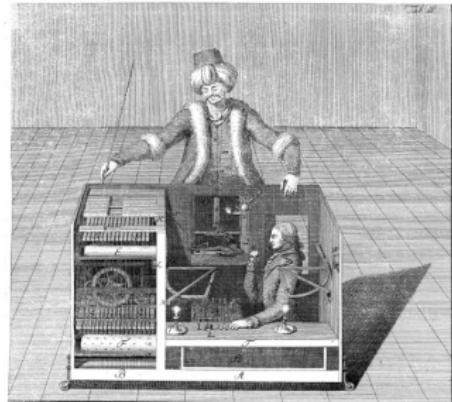


MTurk + Qualtrics

MTurk + Qualtrics

Details:

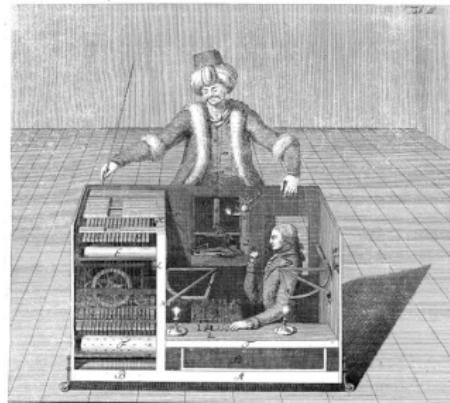
- 25 question, 10-minute survey
- 703 respondents
- *One week*



MTurk + Qualtrics

Details:

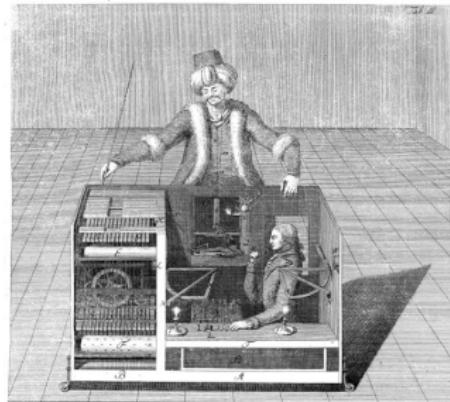
- 25 question, 10-minute survey
- 703 respondents
- *One week*
- **£100 (U.S. \$150)**



MTurk + Qualtrics

Details:

- 25 question, 10-minute survey
- 703 respondents
- *One week*
- **£100 (U.S. \$150)**

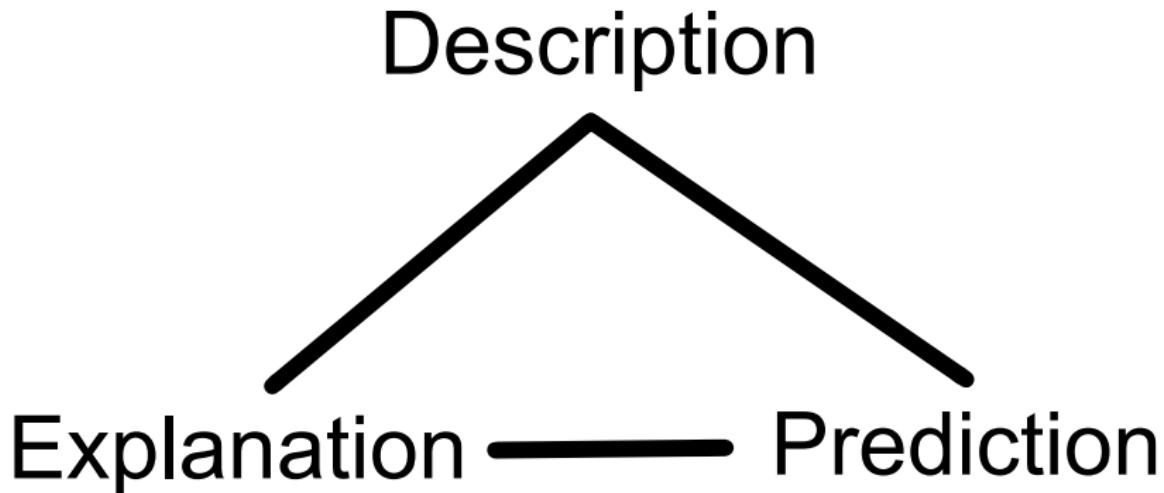


Other Advantages:

- Theory → Operationalization
→ Measurement → Analysis
- Hands-on Design Experience
- “Real” Challenges To Inference

2. Tools

Three Things



The Reality

Explanation

> Description

> Prediction

Descriptive Tools: Visualization

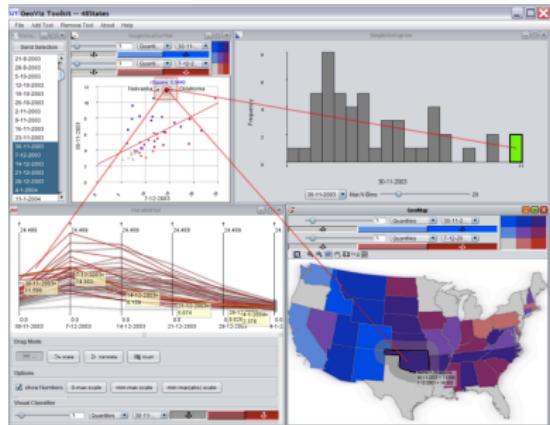
Descriptive Tools: Visualization

Some facts:

Descriptive Tools: Visualization

Some facts:

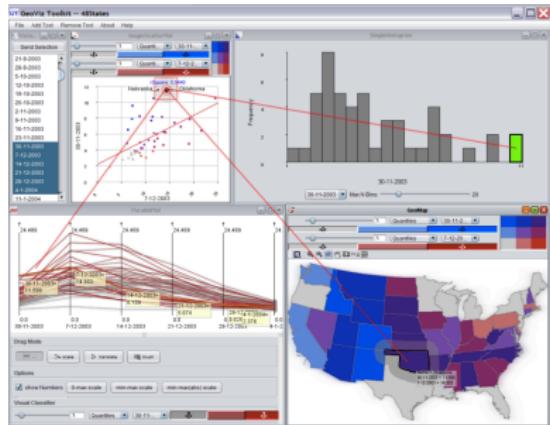
- People like pictures



Descriptive Tools: Visualization

Some facts:

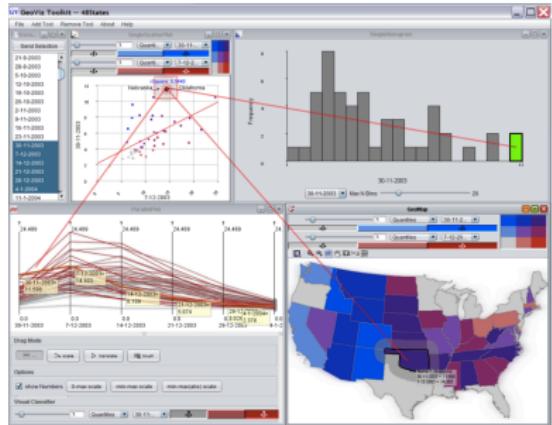
- People like pictures
- People love maps



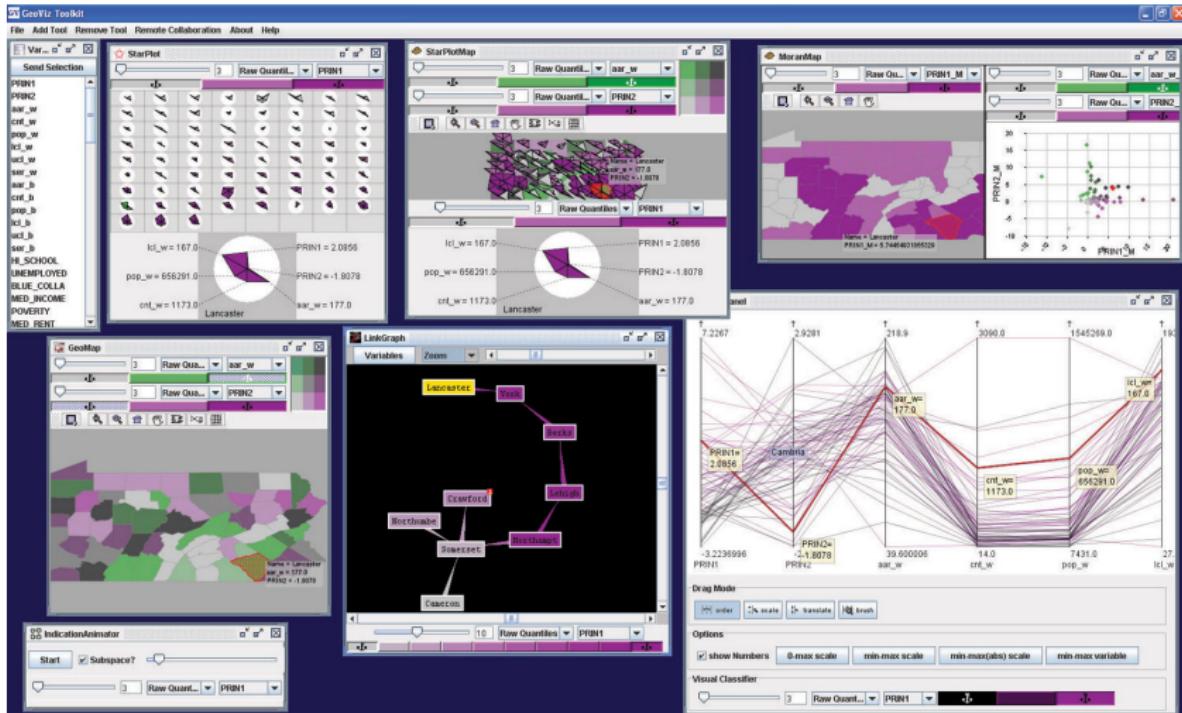
Descriptive Tools: Visualization

Some facts:

- People like pictures
- People love maps
- Making both is easier than ever

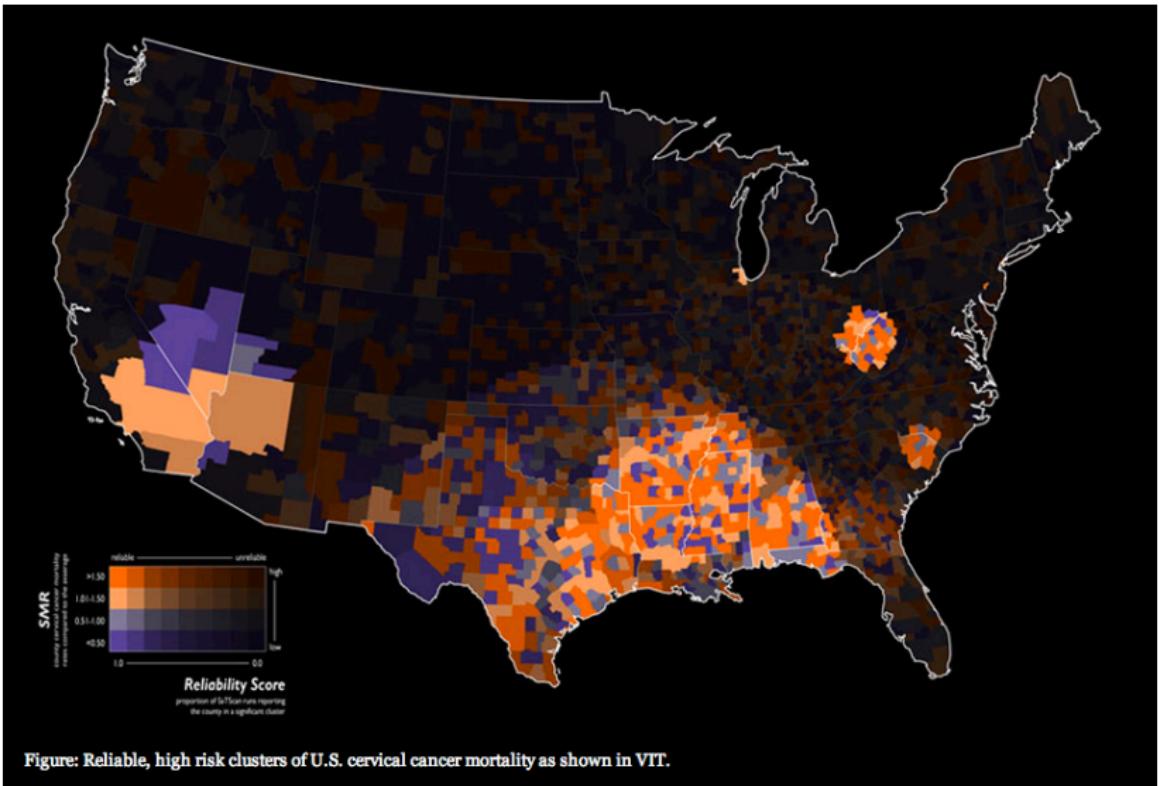


High-Dimensional Visualization



<http://www.geovista.psu.edu>

2D and 3D Mapping



Tools: Prediction

Tools: Prediction

Opportunities:

- Real-Time Data =
Real-Time
Out-Of-Sample
Predictions

Tools: Prediction

Opportunities:

- Real-Time Data =
Real-Time
Out-Of-Sample
Predictions
- Cross-Validation

Tools: Prediction

Opportunities:

- Real-Time Data =
Real-Time
Out-Of-Sample
Predictions
- Cross-Validation
- Kaggle, etc.

kaggle

Tools: Prediction

Opportunities:

- Real-Time Data =
Real-Time
Out-Of-Sample
Predictions
- Cross-Validation
- Kaggle, etc.

The logo for Kaggle, featuring the word "kaggle" in a lowercase, bold, sans-serif font. The letters are a vibrant blue color.

Challenges:

- Timing

Tools: Prediction

Opportunities:

- Real-Time Data =
Real-Time
Out-Of-Sample
Predictions
- Cross-Validation
- Kaggle, etc.

The Kaggle logo is displayed in a large, bold, blue sans-serif font. The word "kaggle" is written in lowercase, with each letter having a distinct vertical stroke. The letters are slightly rounded and have a modern, clean appearance.

Challenges:

- Timing
- Uncertainty

Tools: Prediction

Opportunities:

- Real-Time Data =
Real-Time
Out-Of-Sample
Predictions
- Cross-Validation
- Kaggle, etc.

kaggle

Challenges:

- Timing
- Uncertainty
- The “Arms Race”

3. Theory

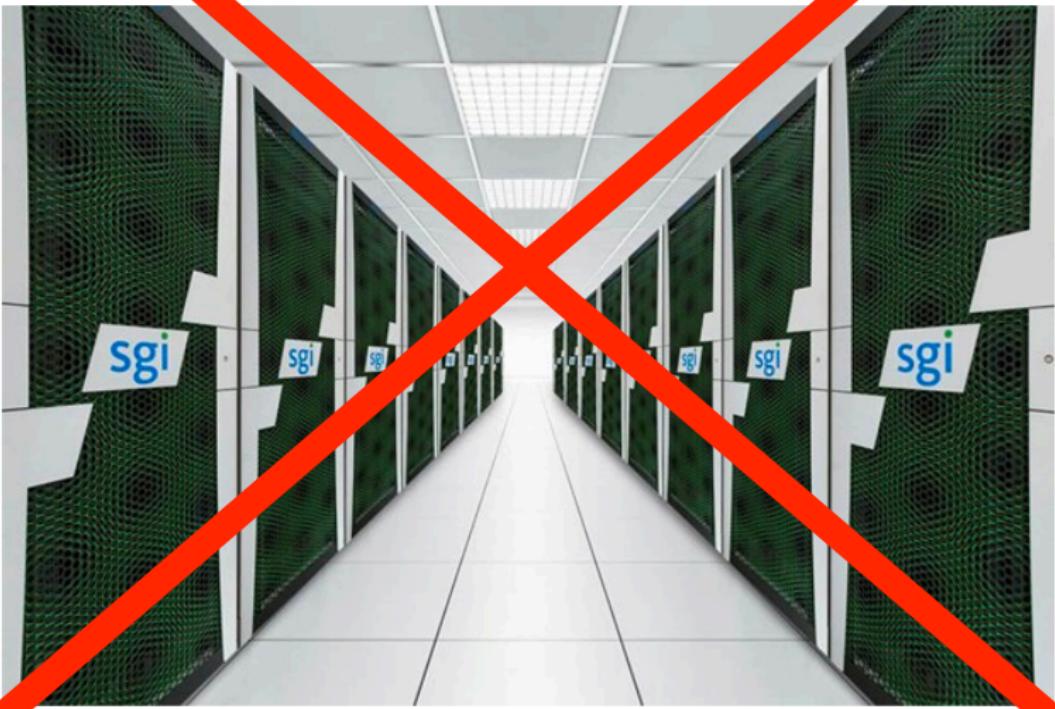
Big data and the death of the theorist

By Ian Steadman | 25 January 13



Big data and the death of the theorist

By Ian Steadman | 15 January 13



Whither (Wither?) Theory?

Whither (Wither?) Theory?

- The Counterargument: “It’s more important than ever!”

Whither (Wither?) Theory?

- The Counterargument: “It’s more important than ever!”
- The Power of Description

Whither (Wither?) Theory?

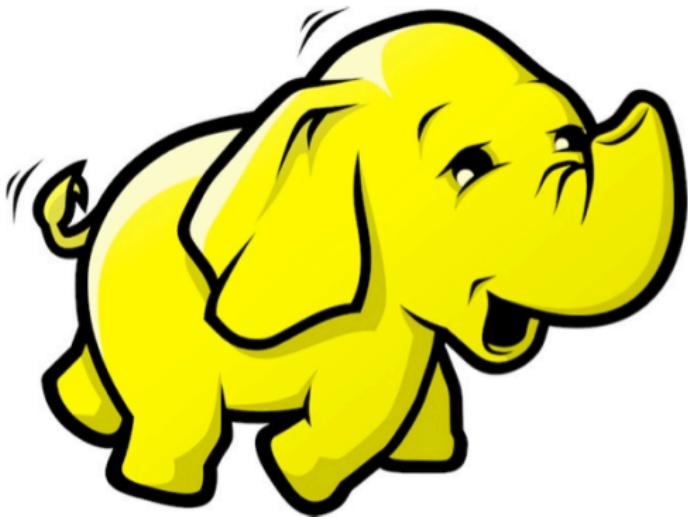
- The Counterargument: “It’s more important than ever!”
- The Power of Description
- Learn From Data

Whither (Wither?) Theory?

- The Counterargument: “It’s more important than ever!”
- The Power of Description
- Learn From Data
- Change your Borrowing Habits



Think different



Theorize
Think different
^

What Ought We Do (Differently)?

What Ought We Do (Differently)?

- Learn some (new?) things
 - Python/Perl/whatever
 - JSON and/or XML data formats

What Ought We Do (Differently)?

- Learn some (new?) things
 - Python/Perl/whatever
 - JSON and/or XML data formats
- Teach (a little) **programming**

What Ought We Do (Differently)?

- Learn some (new?) things
 - Python/Perl/whatever
 - JSON and/or XML data formats
- Teach (a little) **programming**
- Emphasis on **description / visualization** and **exploration**

What Ought We Do (Differently)?

- Learn some (new?) things
 - Python/Perl/whatever
 - JSON and/or XML data formats
- Teach (a little) **programming**
- Emphasis on **description / visualization** and **exploration**
- Greater attention to **prediction**

What Ought We Do (Differently)?

- Learn some (new?) things
 - Python/Perl/whatever
 - JSON and/or XML data formats
- Teach (a little) **programming**
- Emphasis on **description / visualization** and **exploration**
- Greater attention to **prediction**
- **Theorize different(ly)**: A little more induction, a little more humility

Thank you

<https://github.com/PrisonRodeo/QM2015>

