Hi!

DS501: Data Gathering

Prof. Randy Paffenroth
rcpaffenroth@wpi.edu

Worcester Polytechnic Institute

Ok, try 2… hopefully with sound this time :-)

- https://www.youtube.com/watch?v=3_1reLdh5xw

Objectives for today

- To discuss “data gathering”
- We will go from the very general...
  - What kinds of data can you gather?
  - What are issues in data gathering?
- To the very specific…
  - How does one harness the Twitter stream (or perhaps more accurately the Twitter torrent!)
- By the end of class today you should be able to start processing Twitter to ask and answer interesting questions.

http://kathleendeery.com/wp-content/uploads/091615_1554_Drinkingfro1.jpg

Announcements

- Case study 1 is posted today!
- Case study 1 is due February 10 (**BEFORE THE START OF CLASS**)
  - Get started right away (like tomorrow), there is a lot of stuff you need to do!

What does a *high scoring* “case study” look like?

Here is an example from a previous class...

What is “data gathering”?

- Why do we start off the class with this topic?

need data to do analysis

Raw material

Quality of data Drives Results

Different sources

errors

Refining / cleaning

An example.

- Suppose you are a Data Scientist for a Presidential campaign.
- You are tasked with gathering data to decide where to spend the (very limited!) money the campaign has to improve your chances of winning.
- Where do you spend the money?
- https://datafloq.com/read/big-data-obama-campaign/516

historical Data

Stealing Data From Other Campaigns

Borrowing

Copying

Demographics

- Census

polling

ethnic, culture,

Channels

social network

Steal From Companies

Borrow

Search results

Geographically Labelled

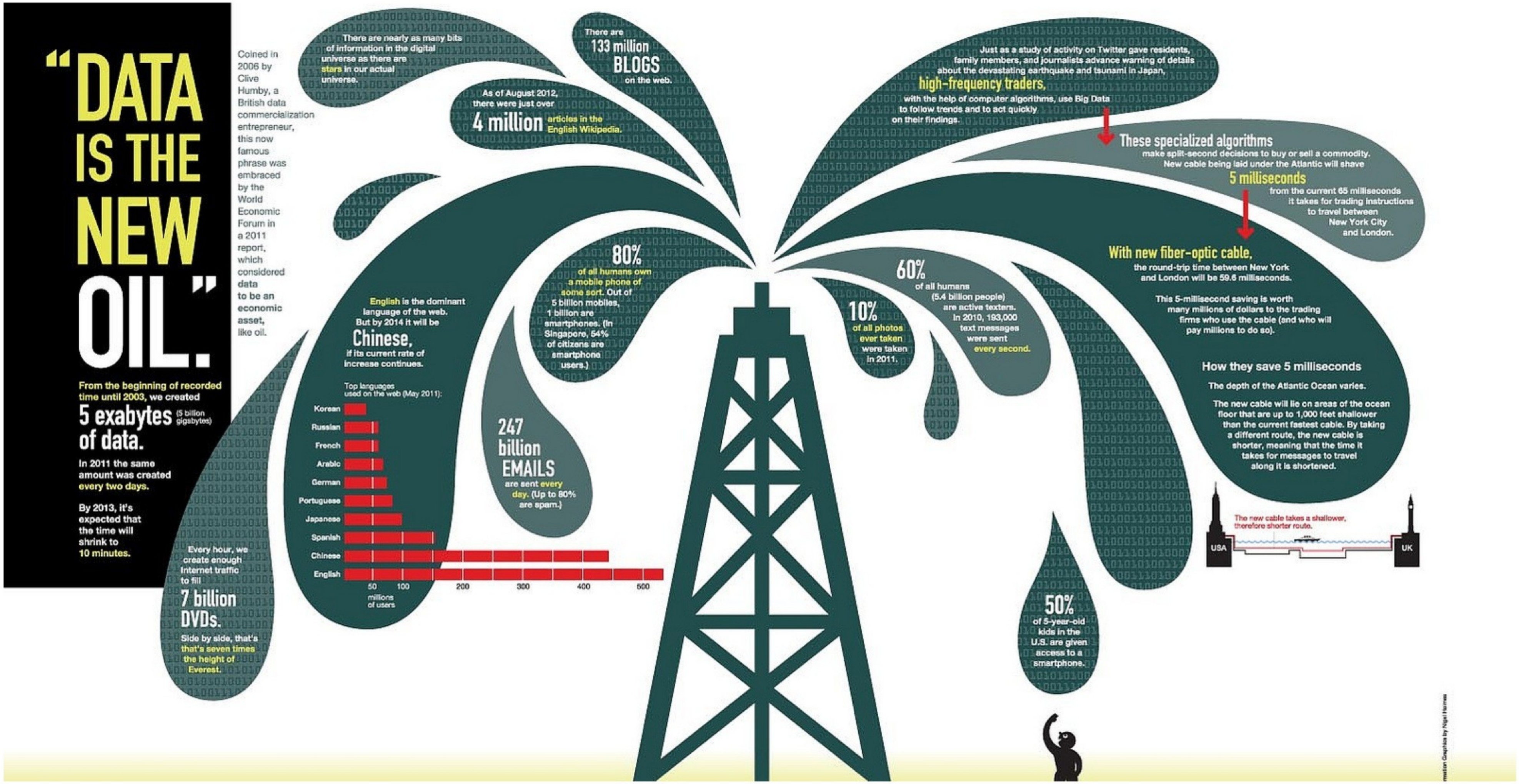
Contributions

weather → voter turnout

representatives

Location of voting

centers



“DATA IS THE NEW OIL.”

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.

From the beginning of recorded time until 2003, we created 5 exabytes (5 billion gigabytes) of data.

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Every hour, we create enough Internet traffic to fill 7 billion DVDs. Side by side, that's that's seven times the height of Everest.

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

There are 133 million BLOGS on the web.

As of August 2012, there were just over 4 million articles in the English Wikipedia.

English is the dominant language of the web. But by 2014 it will be Chinese, if its current rate of increase continues.

Top languages used on the web (May 2011):

80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)

247 billion EMAILS are sent every day. (Up to 80% are spam.)

10% of all photos ever taken were taken in 2011.

60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

50% of 5-year-old kids in the U.S. are given access to a smartphone.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, high-frequency traders, with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave 5 milliseconds from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 milliseconds

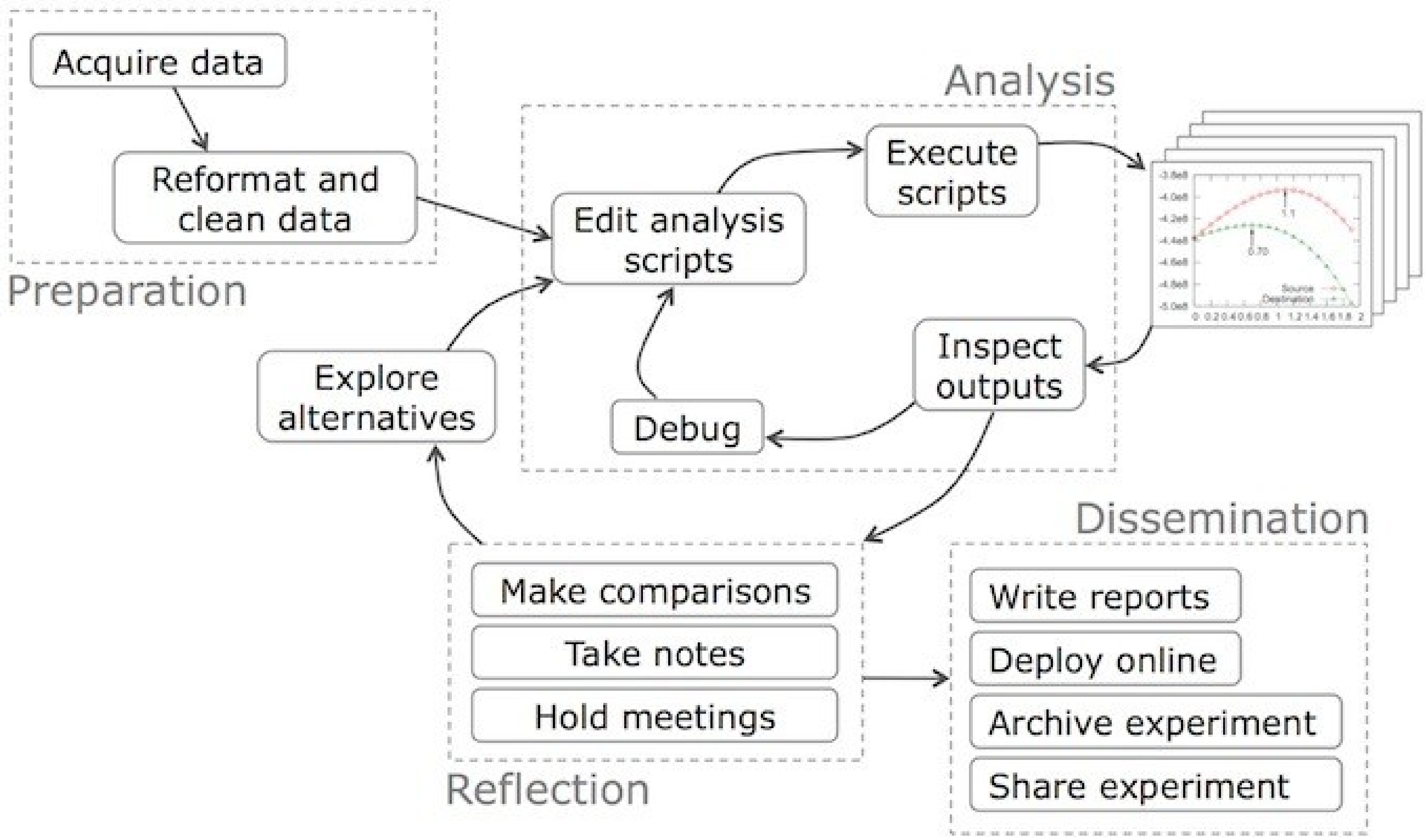
The depth of the Atlantic Ocean varies.

The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.

rmation Graphics by Nigel Holmes

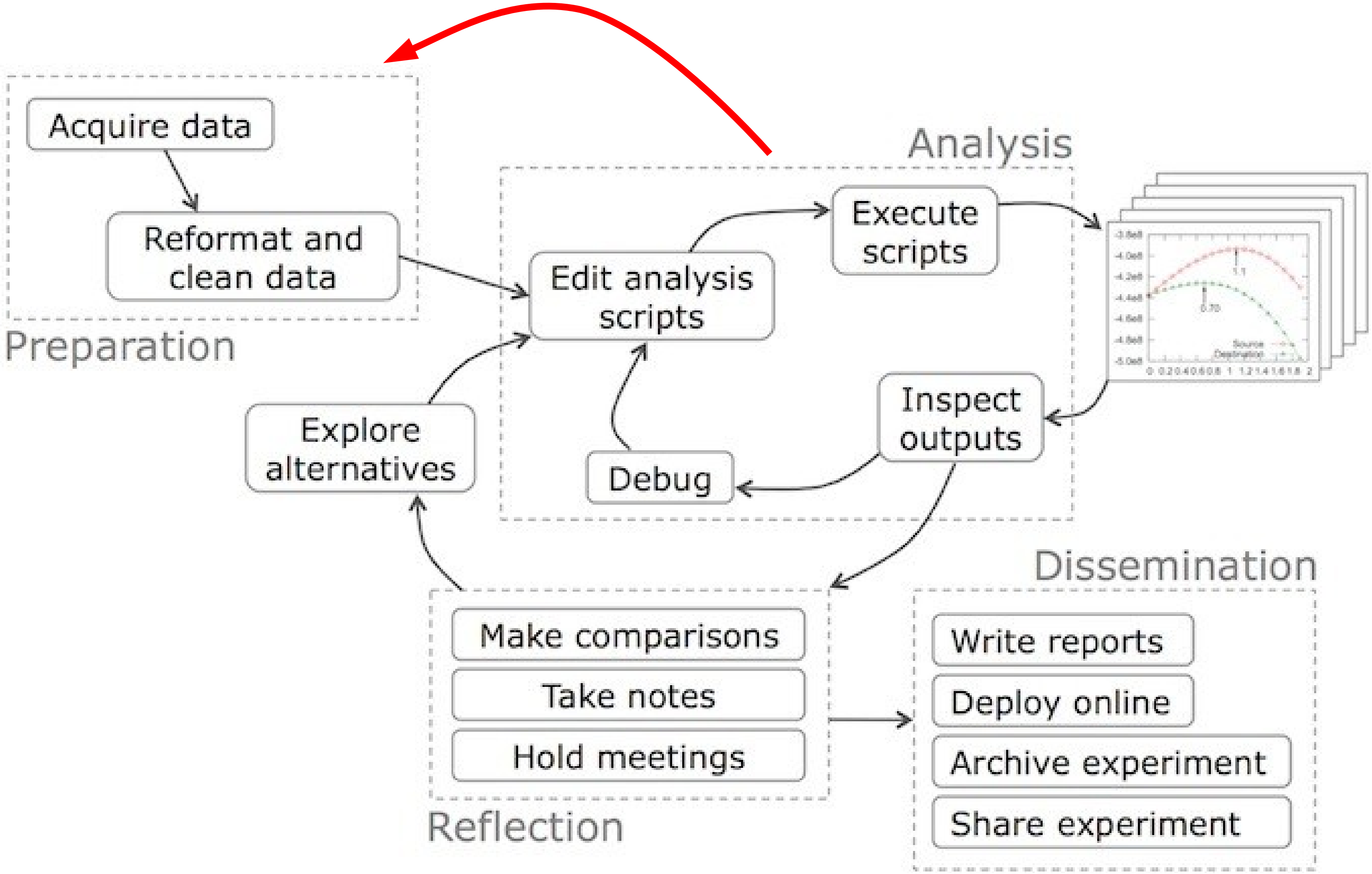
https://www.flickr.com/photos/seeminglee/8436405977

Data Science Work-flow



- http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

Data Science Work-flow



Preparation

Acquire data

Reformat and clean data

Analysis

Edit analysis scripts

Execute scripts

Inspect outputs

Debug

Explore alternatives

Reflection

Make comparisons

Take notes

Hold meetings

Dissemination

Write reports

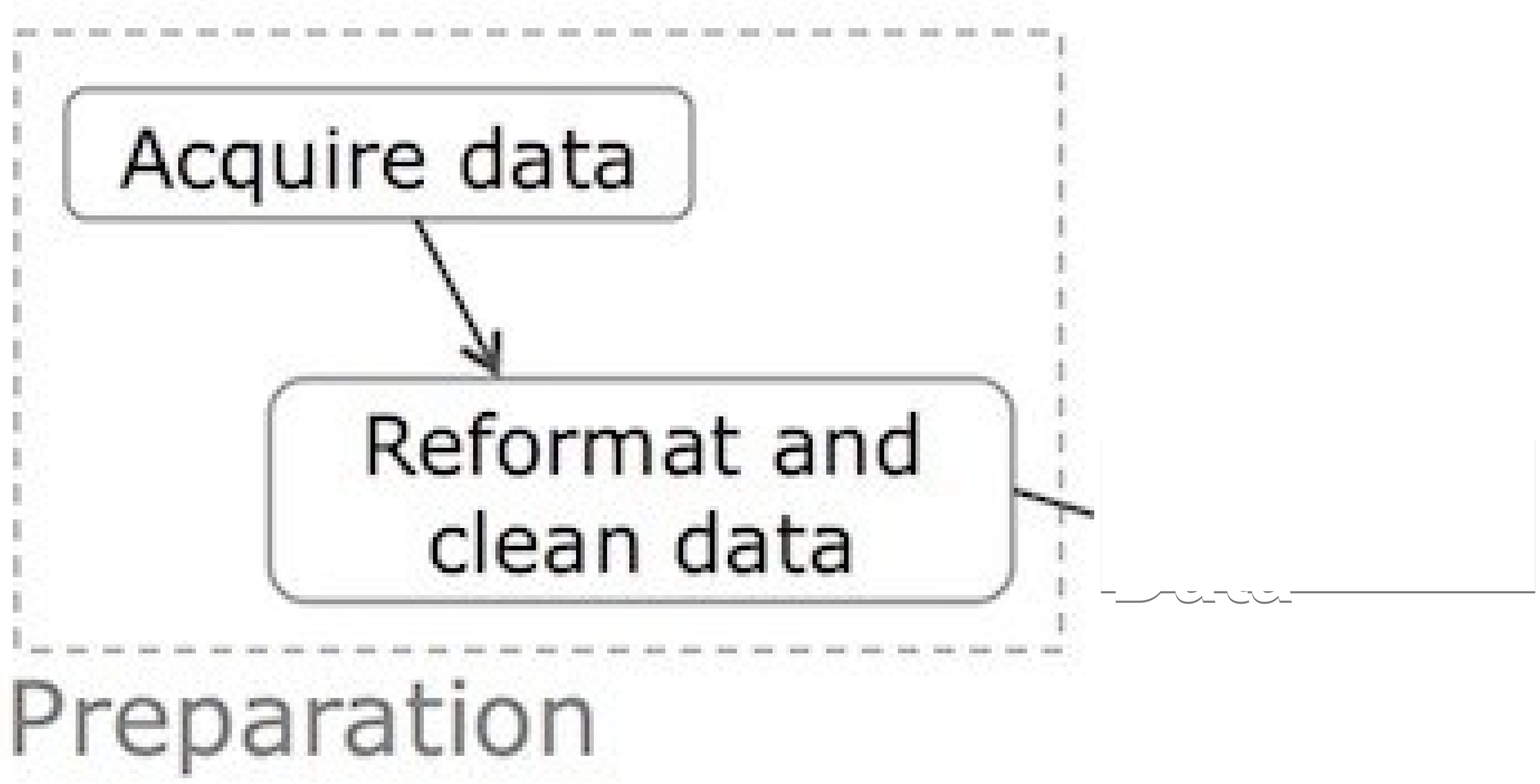
Deploy online

Archive experiment

Share experiment

- http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

First Step



Methods of data collection

- Surveys
  - Asking people what they think
- Experiments
  - Making your own little world and seeing what you can find our
- Observation
  - Looking at the world that is and seeing that is there

Observation

- others are hard

phone interviews

insurance

- surveys

Depends on domain

experiments are hard

expensive

How useful are suverys?

yes useful

Subjective

BIASeD

Potential Biases in Surveys

One of the main issues with sample surveys is that often the responses from the sample tend to favor some parts of the population over others. Then the results of the sample are not representative of the population and are said to be **biased**.

Example?

Boston - what is your Favorite Football team

Types of Biases

Bias can occur in a sample due to various reasons as follows :

1. **Sampling Bias**: As the term suggests, this kind of bias results from a flaw in the sampling method, most likely if the sample is **non-random**. Another way it can occur is due to **under-coverage** – having a sample that lacks representation from parts of the population. Responses by those not in the sample might be quite different from those in it, thus leading to misleading conclusions about the population.

Example: A telephone survey will not reach homeless people; incidentally, these groups of people may have very different views about life in general.

Random: Social Security NUMBER

Types of Biases

2. **Non-response bias**: This kind of bias results when some of the sampled subjects cannot be reached or refuse to participate. In fact, the subjects who are willing to participate may be different from the overall sample in some way, perhaps having strong views about the survey issues. The subjects who do participate may not respond to some questions, resulting in non-response bias due to missing data.

Types of Biases

3. **Response bias**: This kind of bias results from the actual responses. The responses of subjects may differ based on the particular manner ***the interviewer asks questions;*** subjects can often lie because they think that their responses may be socially unacceptable.

Types of Biases

There are many ways to change people's responses using subtle changes of wording for questions!

https://www.qualtrics.com/blog/writing-survey-questions/

Volunteer and Convenience Samples

Surveys are often carried out using easily obtainable samples. One such type of sample is the **convenience sample**.

- As the term suggests, this type of sample is easy and cheap for the interviewer to obtain.
- For example, an interviewer can stop people on the street or in front of a shopping mall to collect data from them.
- However, these kind of sampling schemes may result in serious **biases**.
- For example, working people may be under-represented if the interviews are conducted on workdays between 9 am-5 pm.

Volunteer and Convenience Samples

A common type of convenience sample in the **volunteer sample** where subjects volunteer to belong to the survey.

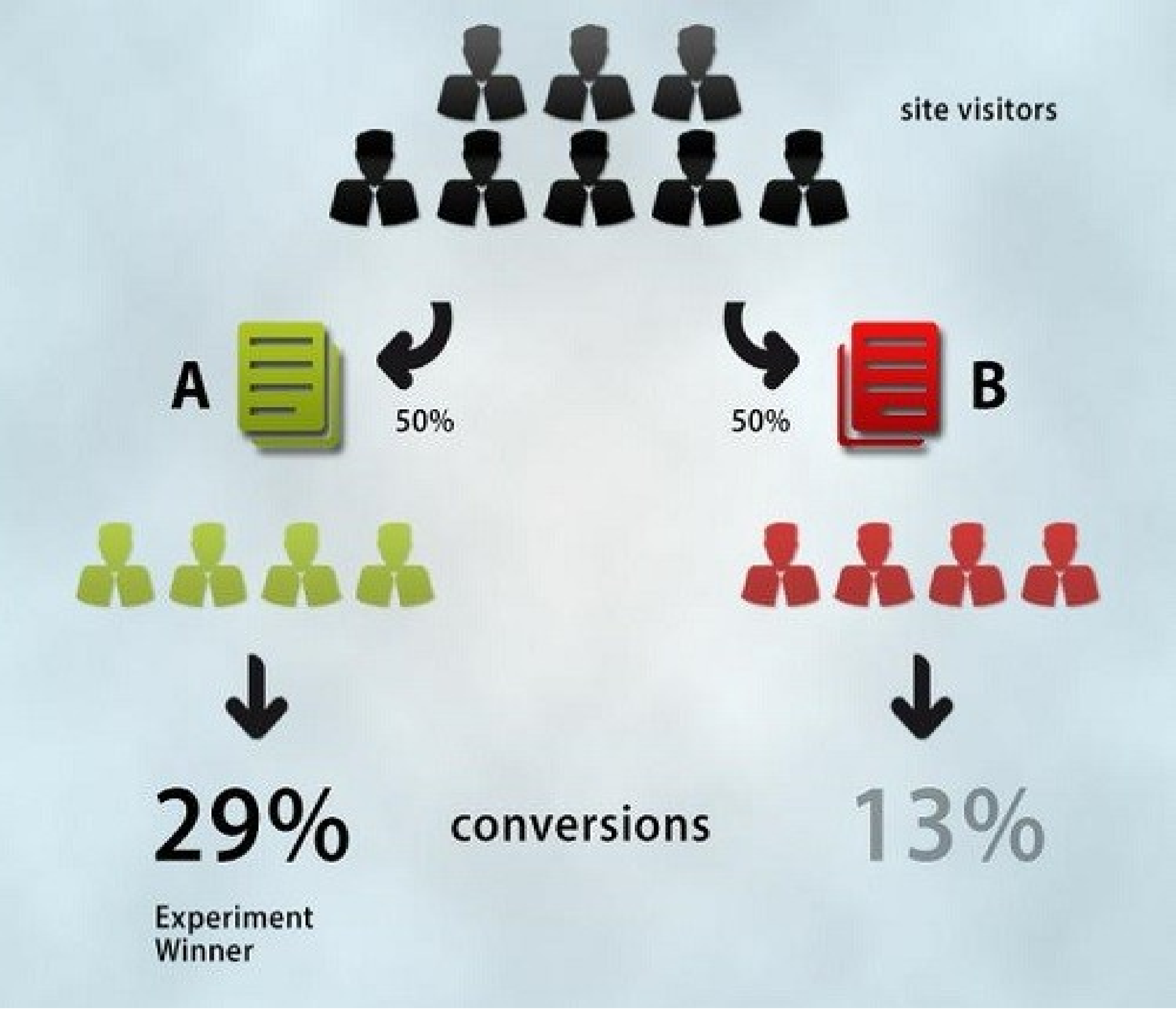
- A good example of volunteer samples are internet surveys where a person can voluntarily log in and can answer questions if he/she has access to the internet (E.g., *www.surveymonkey.com*).
- Unfortunately, volunteer sampling schemes inherently comes with biases.
- This is because, one segment of the population may be more likely to volunteer than other segments maybe because they have stronger opinions about a particular issue or a more likely to surf the web.

Volunteer and Convenience Samples

Wow, they sound pretty bad… do you actually do them in practice!?

How about experiments?

Inspirational example for experimental study: A/B testing



https://www.flickr.com/photos/rahulrodriguez/14683524180

Example of A/B testing

http://fortysevenmedia.com/blog/archives/google_a_b_testing_with_expressionengine_structure_freebie/

http://blog.hubspot.com/marketing/a-b-testing-experiments-examples

Experiments

- Design of statistical experiments is a broad and interesting topic in statistics.
- On which, graduate classes are taught!
- However, it will not be our focus here.

How about observations?

physical sensors

counting people

traffic

Groceries

amazon purchases

GPS traces

Getting data

- Pre-made datasets
- Making your own data

Note...

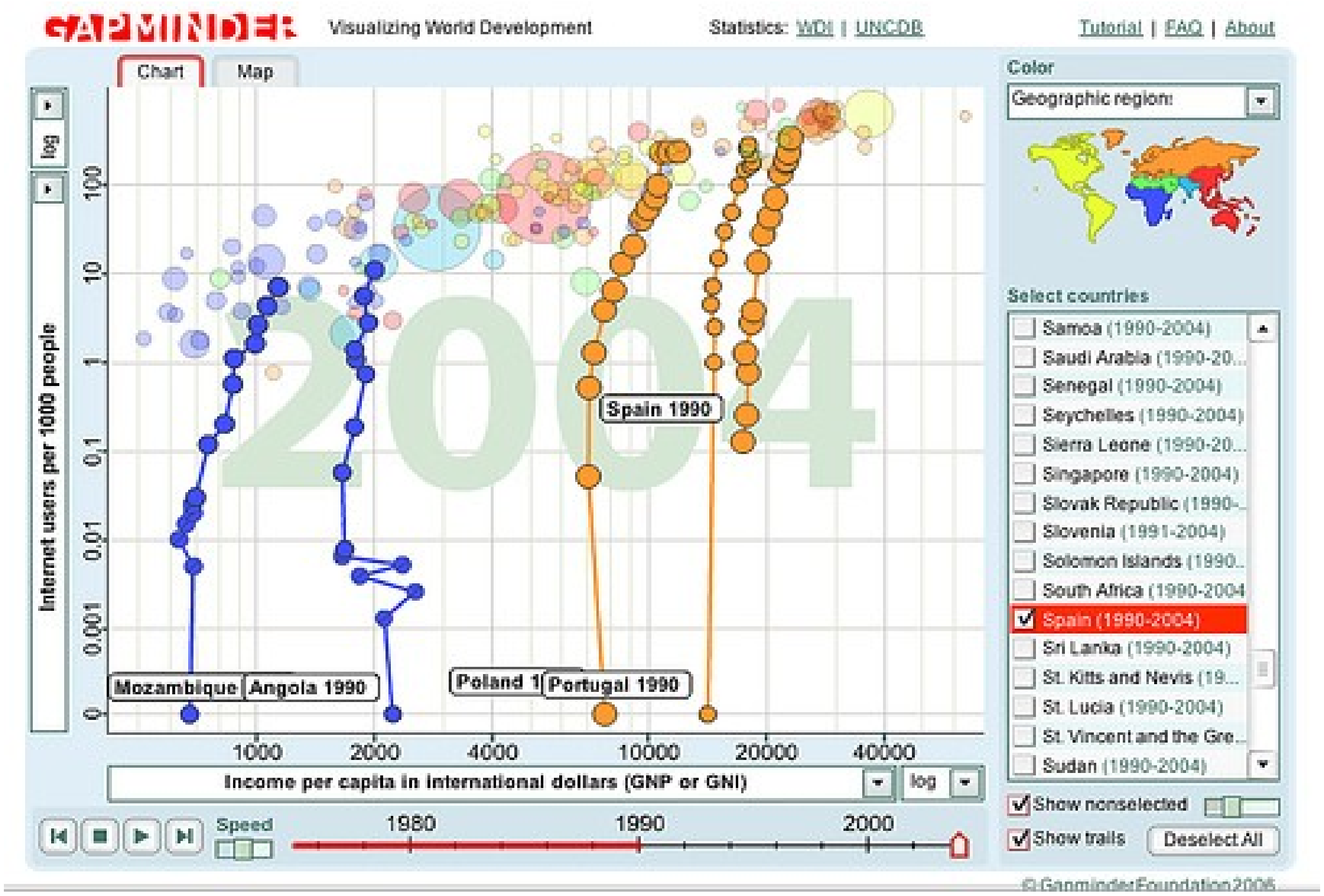
- Some examples shown at high level.
  - The details may be opaque but that is ok.
  - There are to wet your appetite for later work!
- Some examples are detailed
  - You are responsible for these!
- Though they all might help you
  - Get a job...
  - Write a paper...
  - Do research for your degree...

You have access to an **amazing** wealth of information

- http://archive.ics.uci.edu/ml/
- https://www.kaggle.com/
- http://www.data.gov/
- http://databank.worldbank.org/data/home.aspx
- http://www.transparency.org/
- And MANY, MANY more

One particular one that is quite interesting...

- http://www.gapminder.org



What About Data Quality?

- Generally, you have a problem if the data doesn’t mean what you think it does, or should
  - Data not up to spec : garbage in, glitches, etc.
  - You don’t understand the spec : complexity, lack of metadata.
- Data quality problems are expensive and pervasive
  - DQ problems cost hundreds of billion $$$ each year.
  - Resolving data quality problems is often the biggest effort in a data mining study.

Adapted from: “Data Quality and Data Cleaning: An Overview”
Theodore Johnson, SIGMOD 2003

Example

T.Das|97336o8327|24.95|Y|-|0.0|1000
Ted J.|973-360-8779|2000|N|M|NY|1000

- Can we interpret the data?
  - What do the fields mean?
  - What is the key? The measures?
- Data glitches
  - Typos, multiple formats, missing / default values
- Metadata and domain expertise
  - Field three is Revenue. In dollars or cents?
  - Field seven is Usage. Is it *censored*?
    - Field 4 is a censored flag. How to handle censored data?

Adapted from: “Data Quality and Data Cleaning: An Overview”
Theodore Johnson, SIGMOD 2003

Data Glitches

- Systemic changes to data which are external to the recorded process.
  - Changes in data layout / data types
    - Integer becomes string, fields swap positions, etc.
  - Changes in scale / format
    - Dollars vs. euros
  - Temporary reversion to defaults
    - Failure of a processing step
  - Missing and default values
    - Application programs do not handle NULL values well …
  - Gaps in time series
    - Especially when records represent incremental changes.

Adapted from: “Data Quality and Data Cleaning: An Overview”
Theodore Johnson, SIGMOD 2003

Conventional Definition of Data Quality

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
    - Special problems in federated data: time consistency.
- Consistency
  - The data agrees with itself.

Adapted from: “Data Quality and Data Cleaning: An Overview”
Theodore Johnson, SIGMOD 2003

One way to get data: Web page "crawling"

- HTML is all about how to display/show data, but not about giving you the data.
- Easy to download, but, hard to process
- Powerful, but can actually be quite complicated to get data from web sites.
- Rules: robots.txt

Example

http://finance.yahoo.com/q?s=ibm&ql=1

Crawling webpage and process HTML

- Screen scraping and web crawlers
  - https://commoncrawl.org/
  - http://www.crummy.com/software/BeautifulSoup/
  - http://scrapy.org/

Web crawling, example 1, Soccer!

http://nbviewer.jupyter.org/urls/dl.dropboxusercontent.com/u/8169386/FantasyLeague/FetchPremResultsFromBBC.ipynb

Web crawling, example 2

http://www-rohan.sdsu.edu/~gawron/python_for_ss/course_core/book_draft/web/web_crawling.html

How can we be more systematic?

Documented API

Other API for services you might have heard of

- LinkedIn API
- Google Plus API
- Facebook API
- GitHub API

Github

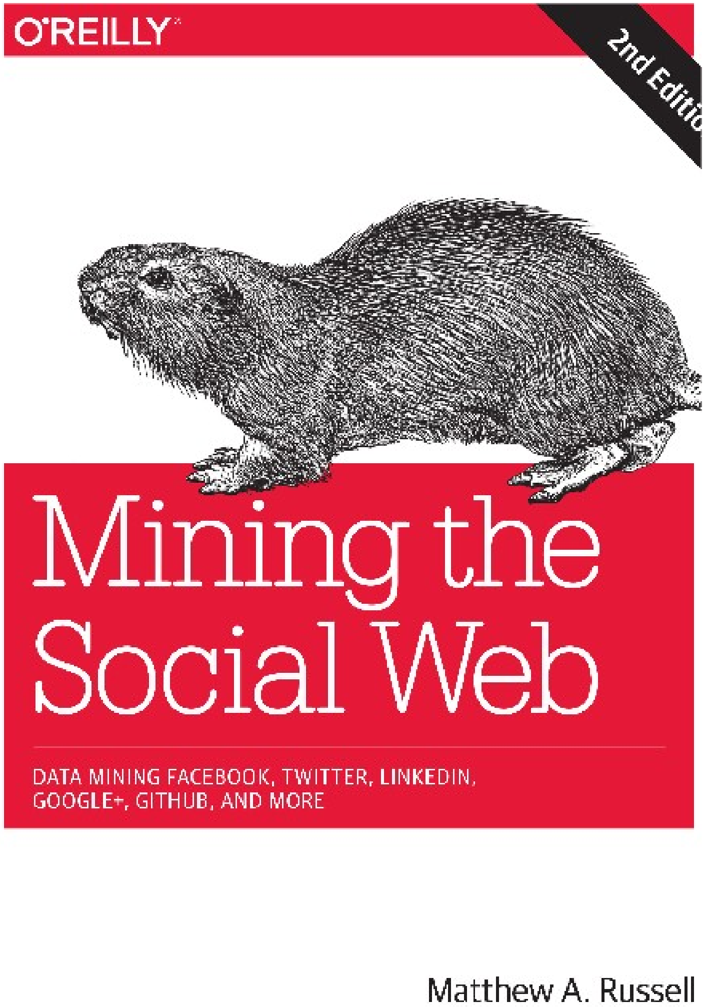
- Both an example of data gathering and a “pro tip”.
  - http://www.github.com
  - I **highly recommend** you use this for your case studies…
  - It make **collaboration on a team much easier**.

http://proquest.safaribooksonline.com.ezproxy.wpi.edu/book/web-applications-and-services/social-media/9781449368180/idot-a-guided-tour-of-the-social-web/ch07_html#X2ludGVybmFsX0h0bWxWaWV3P3htbGlkPTk3ODE0NDkzNjgxODAlMkZjaDA3X2h0bWwmcXVlcnk9

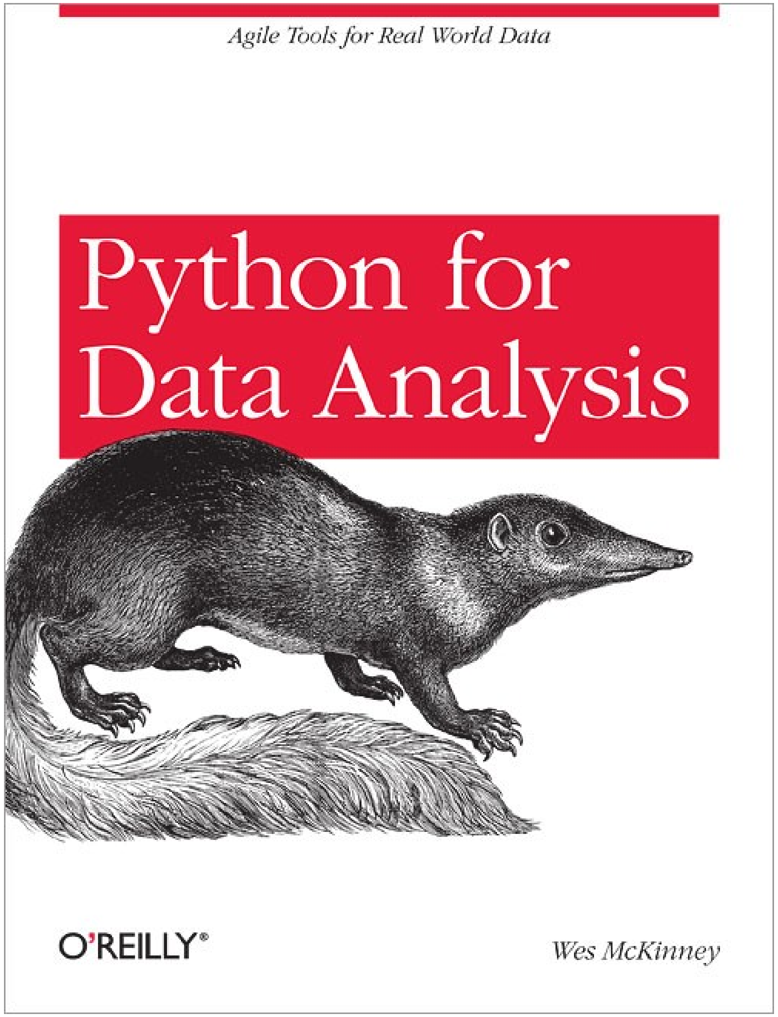
Streaming versus polling Twitter example

- https://dev.twitter.com/streaming/overview
- API – Application Program Interface
- REST - Representational state transfer

Great resource for Case Study 1 (i.e., Chapter 1 and 9 are ***required*** reading)


O'REILLY
2nd Editio
Mining the Social Web
DATA MINING FACEBOOK, TWITTER, LINKEDIN, GOOGLE+, GITHUB, AND MORE
Matthew A. Russell

Helpful book



Agile Tools for Real World Data

Python for Data Analysis

O'REILLY®

Wes McKinney

Why Twitter

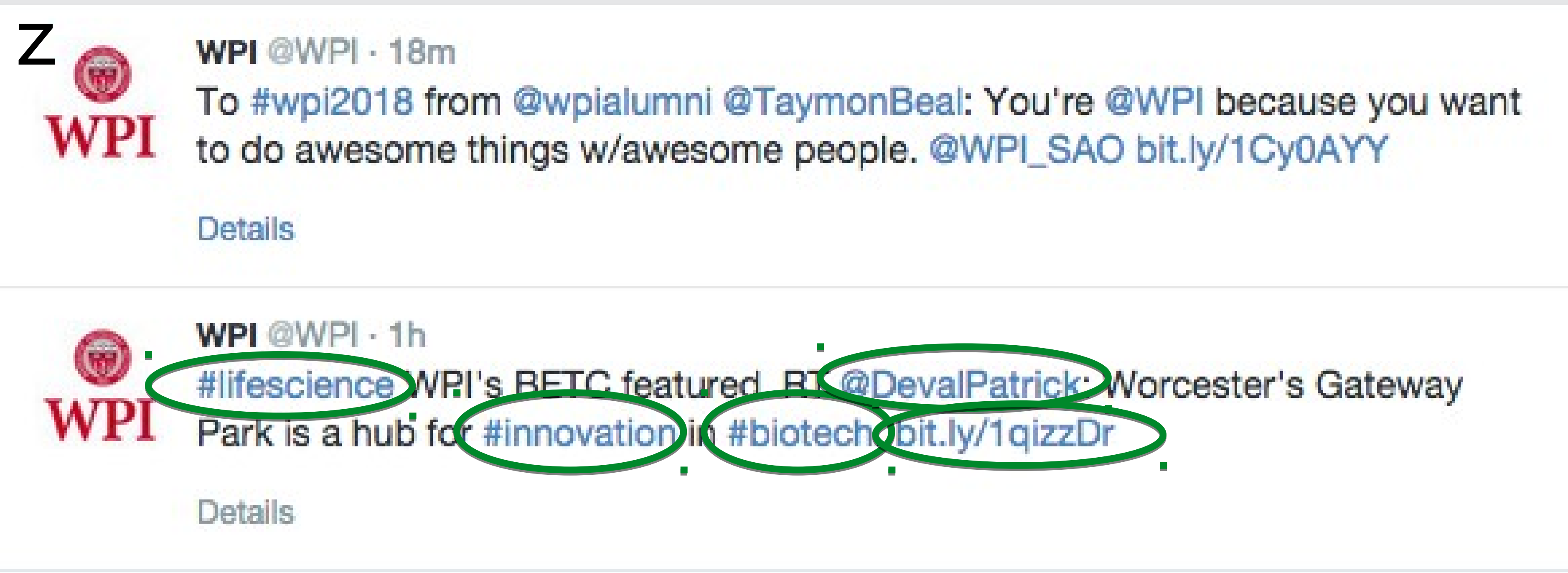
- Rich source of information
- Open for public consumption
- Well-documented API
- tweets happen at the “speed of thought” and are available in near real time.

Learn about the Data

https://support.twitter.com/articles/166337?lang=en#

- **Twitter Data**
  - Tweets: 140 characters (text + entities)

z



WPI @WPI · 18m

To #wpi2018 from @wpialumni @TaymonBeal: You're @WPI because you want to do awesome things w/awesome people. @WPI_SAO bit.ly/1Cy0AYY

Details

WPI @WPI · 1h

#lifescience WPI's BETC featured. RT @DevalPatrick: Worcester's Gateway Park is a hub for #innovation in #biotech bit.ly/1qizzDr

Details

https://twitter.com/search?q=wpi

Twitter as a Sensor Network

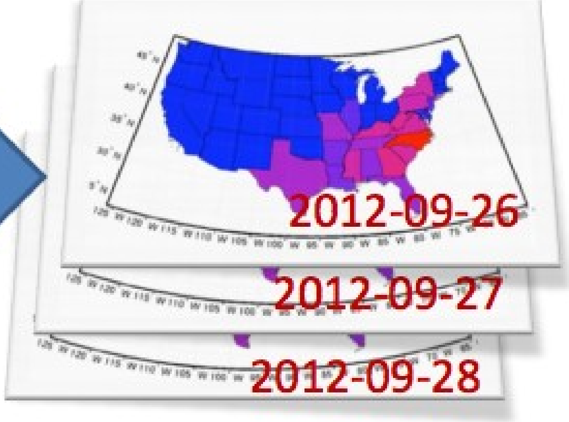
Humans as Sensors

Text Pre-process

Time	Location
2012-09-26 17:35:23	Wisconsin US
2012-09-27 12:17:52	N/A
2012-09-27 08:28:12	(-98.24, 23.22)
…	

Socio-scope

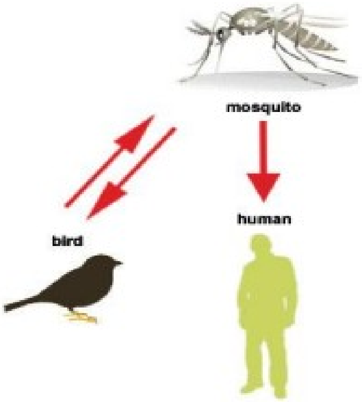
Spatio-temporal signal

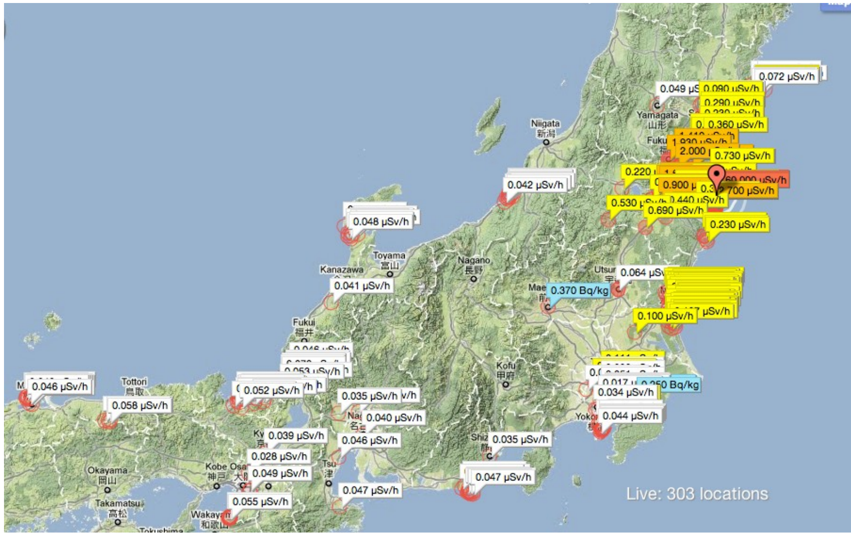


Transportation Safety

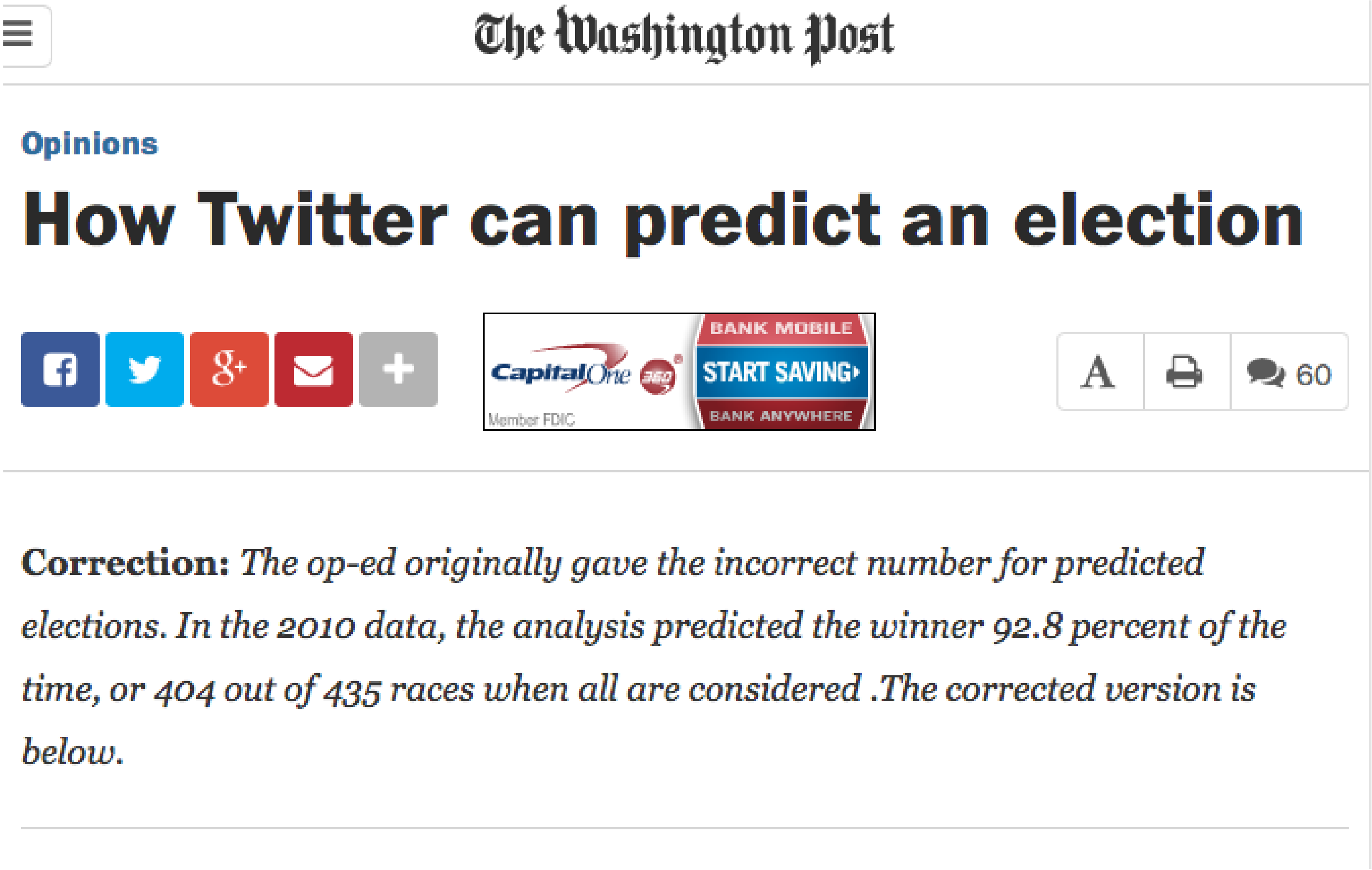
"16 deer got run over by cars in Wisconsin last month "

Public Health





Twitter predicts Election



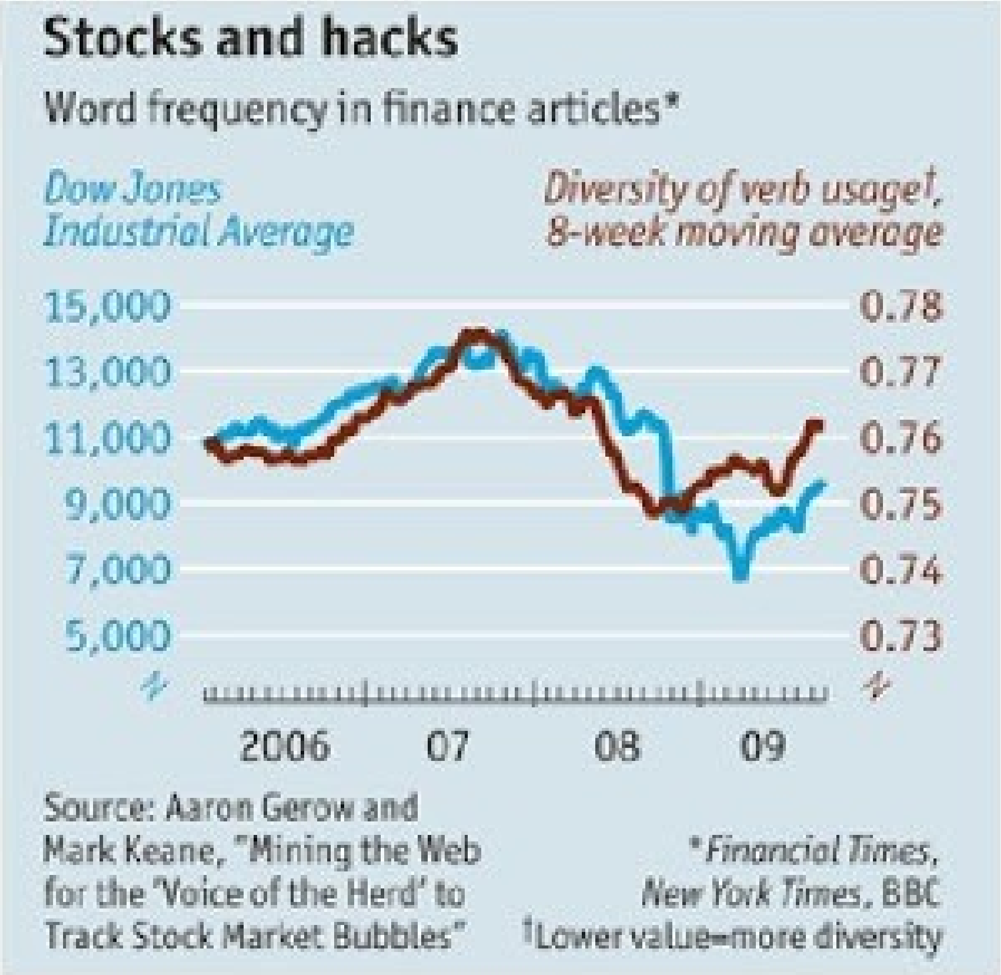
The Washington Post

Opinions

How Twitter can predict an election

Correction: *The op-ed originally gave the incorrect number for predicted elections. In the 2010 data, the analysis predicted the winner 92.8 percent of the time, or 404 out of 435 races when all are considered .The corrected version is below.*

Social Media predicts Stock Market



Stocks and hacks

Word frequency in finance articles*

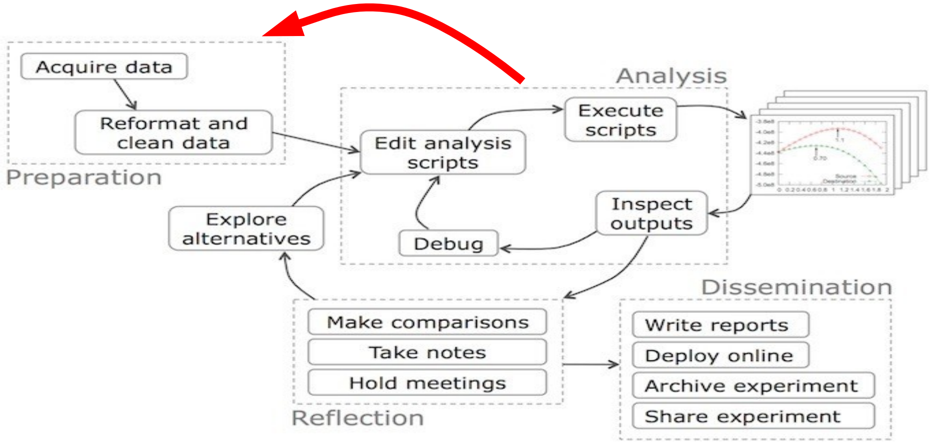
Source: Aaron Gerow and Mark Keane, "Mining the Web for the 'Voice of the Herd' to Track Stock Market Bubbles"

*Financial Times, New York Times, BBC

†Lower value=more diversity

Accessing Twitter Data from IPython Notebook: Workflow example

- **Get Connected:** Authorizing an application to access Twitter account data
- **Download Data:** Retrieving trends
- **Examine the Data:** Displaying API responses as pretty-printed JSON
- **Simple Analysis**
- **Collect more data**
- **More Analysis**



Preparation

Acquire data

Reformat and clean data

Analysis

Edit analysis scripts

Execute scripts

Inspect outputs

Debug

Explore alternatives

Reflection

Make comparisons

Take notes

Hold meetings

Dissemination

Write reports

Deploy online

Archive experiment

Share experiment

I am a brave professor...

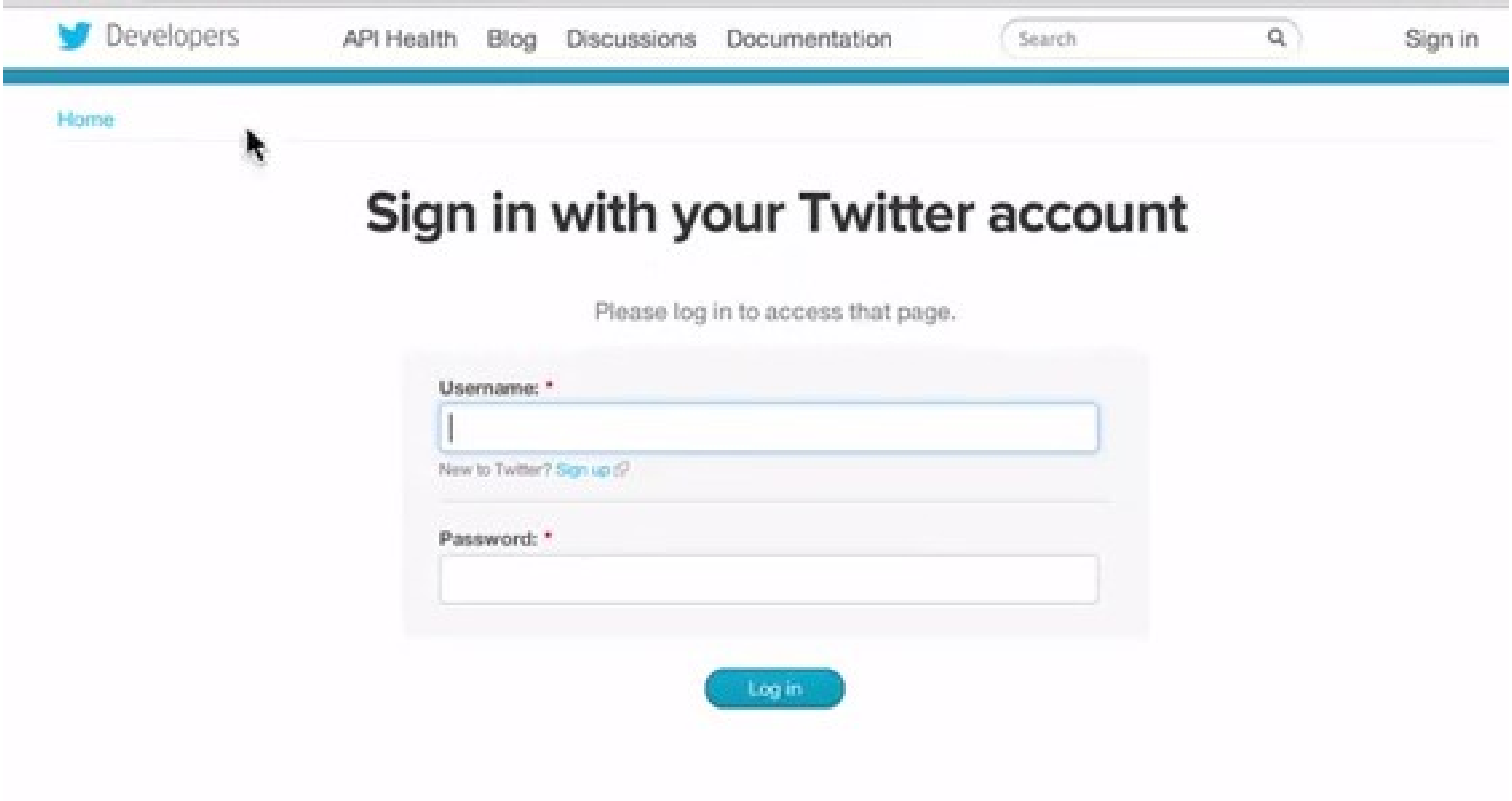
- We are going to do these demos live!
- I.e., we are going to analyze the Twitter stream as it is this very moment.
- Accordingly, any number of things can go wrong
  - I have a canned version as backup, but I think the live version is more fun :-)
- You are going to learn about data gathering like it really is.

To get code running

- Install "twitter" package in Canopy
- Generate app
    - Make sure phone number is in account
- Generate token
- Copy keys and tokens to code

Demo Example 1

Creating an Application



Sign in with your Twitter account

Please log in to access that page.

Username: *

New to Twitter? Sign up

Password: *

- https://dev.twitter.com/apps

OAuth

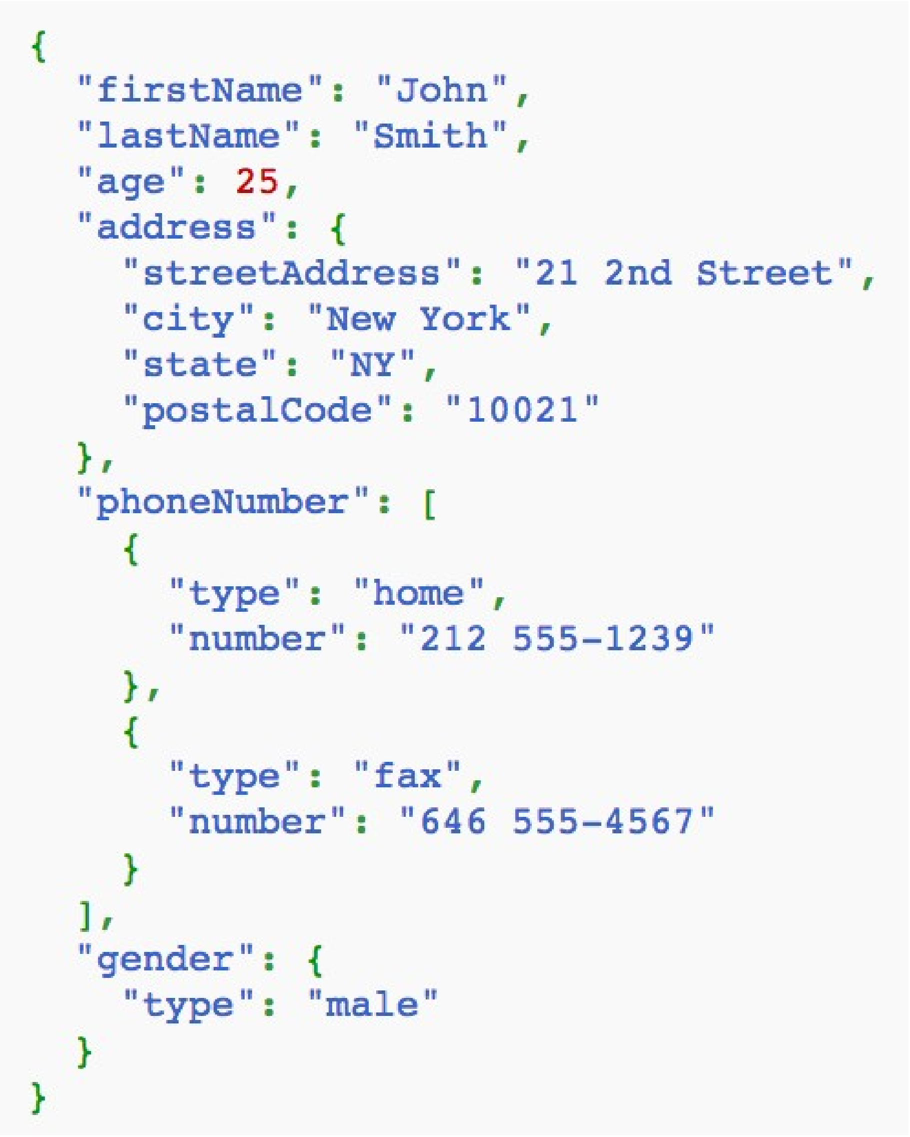
- OAuth is an open standard for authorization
- Short for **Open Authorization** (OAuth)
- A standard protocol for social web sites

See details: http://en.wikipedia.org/wiki/OAuth

Demo Example 1, 2

JSON

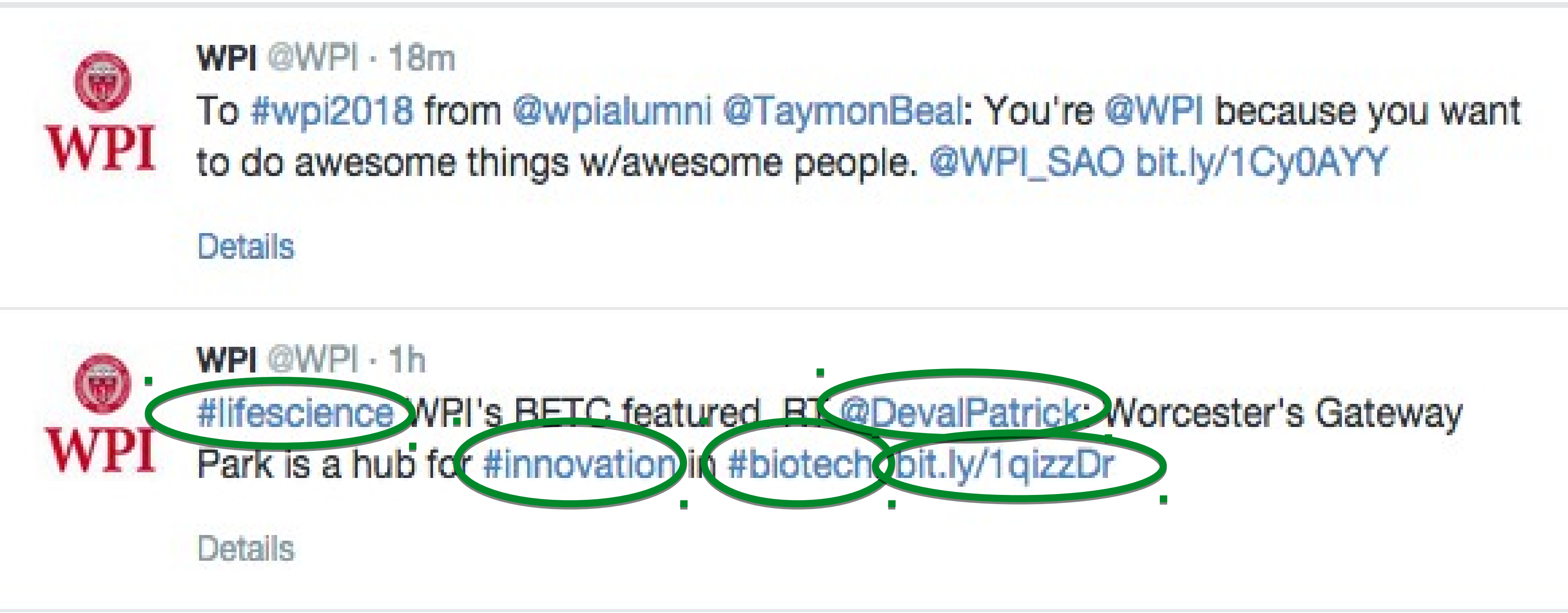
- **JavaScript Object Notation (JSON)**
- An open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs.
- A list of Dictionaries



```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    {
      "type": "home",
      "number": "212 555-1239"
    },
    {
      "type": "fax",
      "number": "646 555-4567"
    }
  ],
  "gender": {
    "type": "male"
  }
}
```

Demo Example 2, 3, 4, 5, 6

Extracting Tweet Entities



WPI @WPI · 18m

To #wpi2018 from @wpialumni @TaymonBeal: You're @WPI because you want to do awesome things w/awesome people. @WPI_SAO bit.ly/1Cy0AYY

Details

WPI @WPI · 1h

#lifescience WPI's BETC featured. RT @DevalPatrick: Worcester's Gateway Park is a hub for #innovation in #biotech bit.ly/1qizzDr

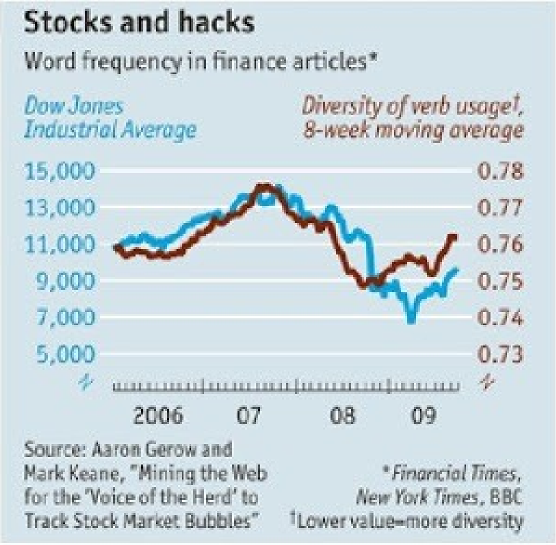
Details

https://twitter.com/search?q=wpi

Demo Example 6, 7, 8

Lexical Diversity of Tweets

- **Lexical Diversity = #different words used / # all words**



Stocks and hacks

Word frequency in finance articles*

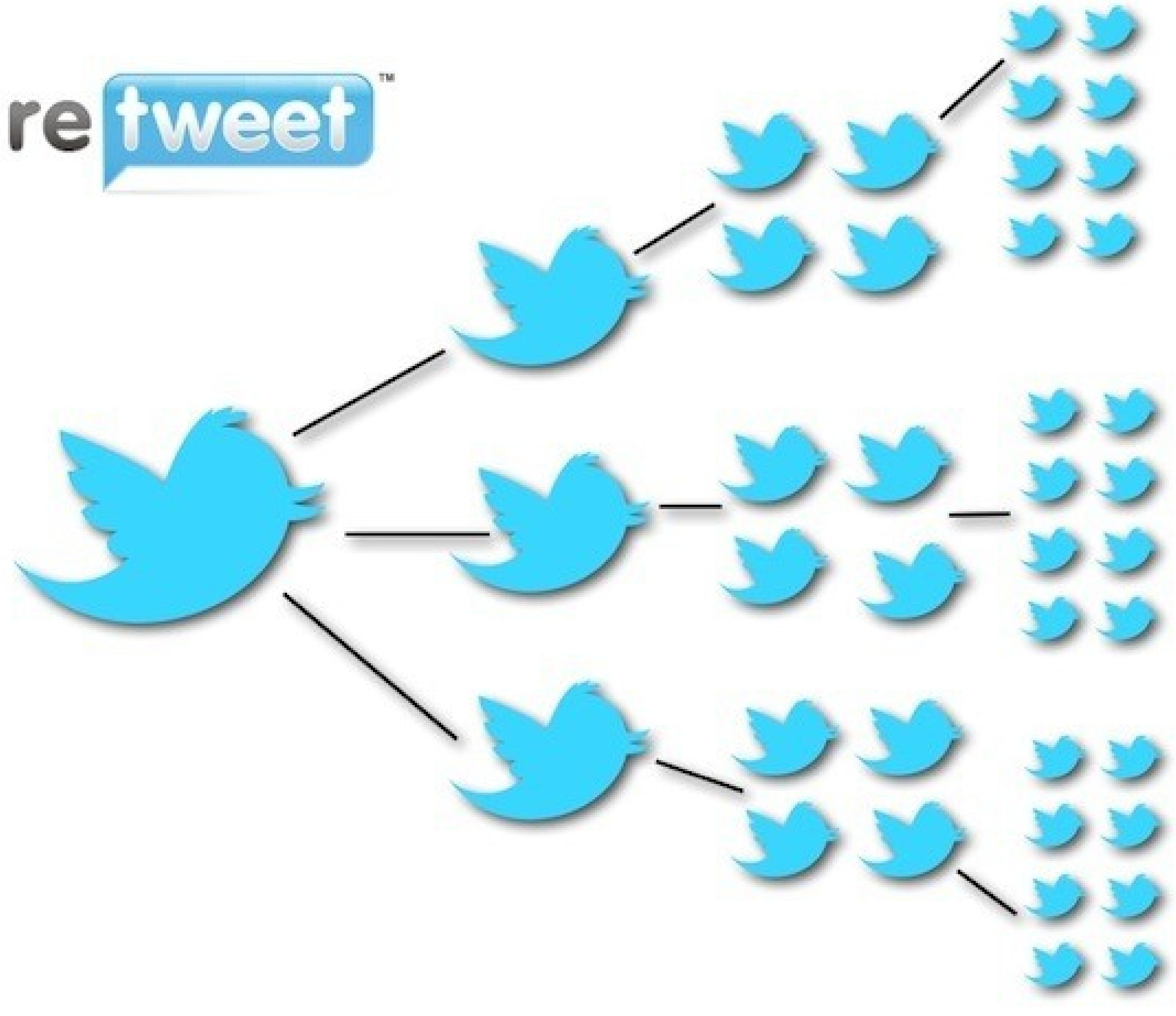
Source: Aaron Gerow and Mark Keane, "Mining the Web for the 'Voice of the Herd' to Track Stock Market Bubbles"

*Financial Times, New York Times, BBC

†Lower value=more diversity

Demo Example 9

Patterns in Retweets



Demo 10

Frequency of words...

- https://en.wikipedia.org/wiki/Zipf's_law

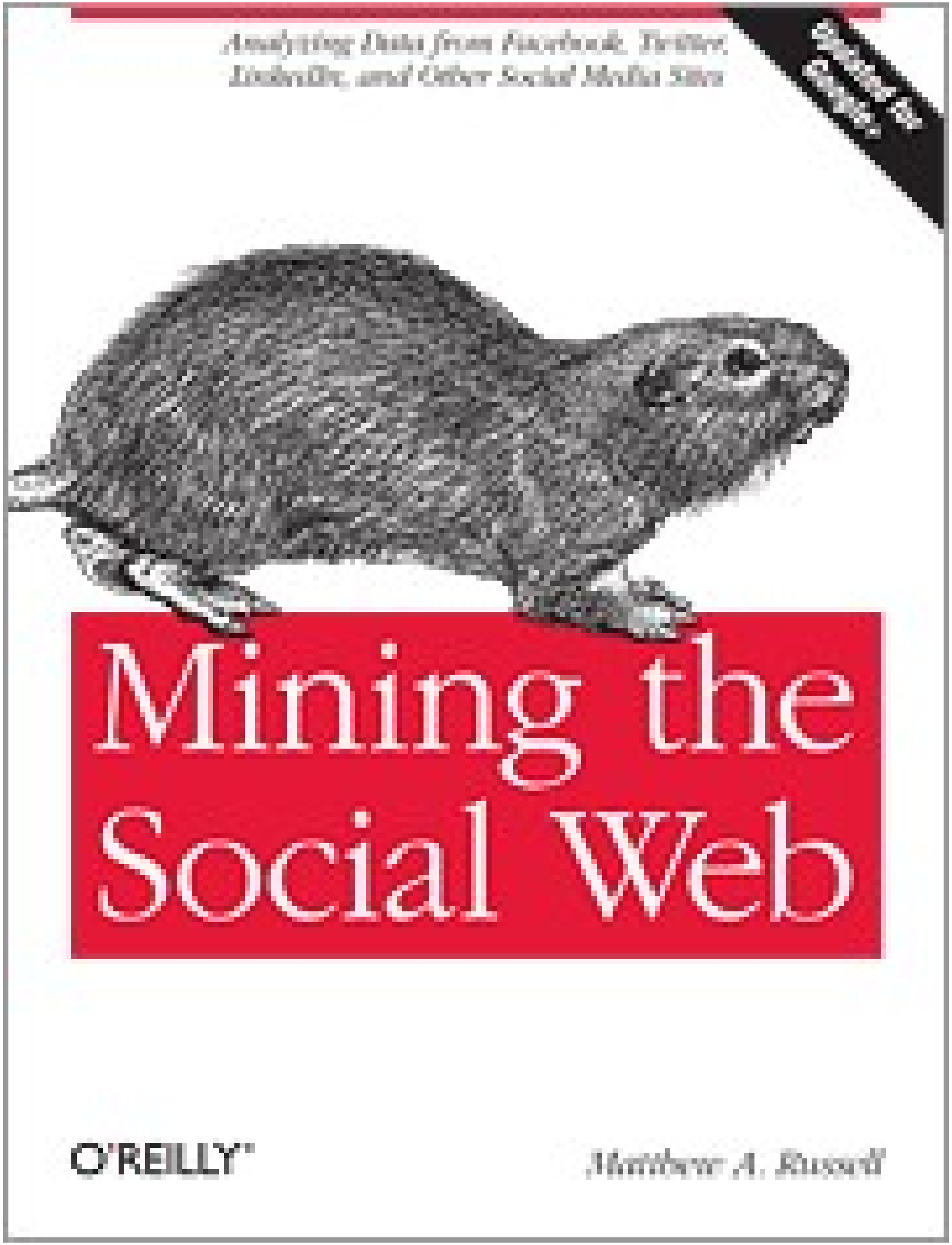
Demo 11

Online Course

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Getting and Cleaning Data
by Jeff Leek, PhD, Roger D. Peng, PhD, Brian Caffo, PhD

- https://class.coursera.org/getdata-007/



Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites

Updated for Google+

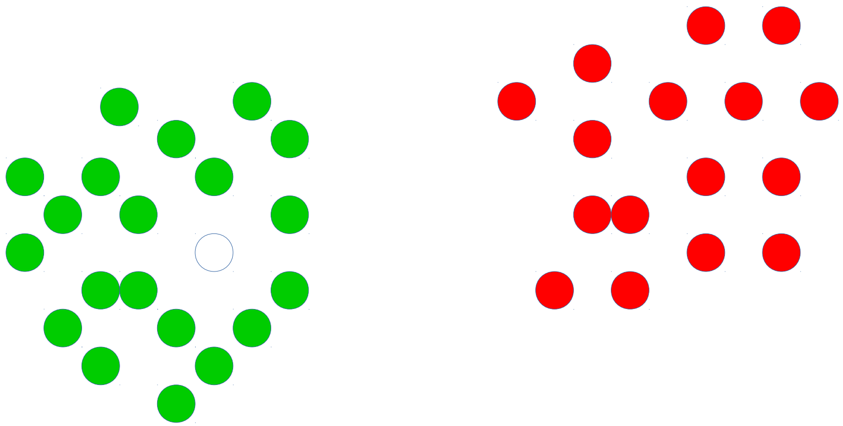
Mining the Social Web

O'REILLY®

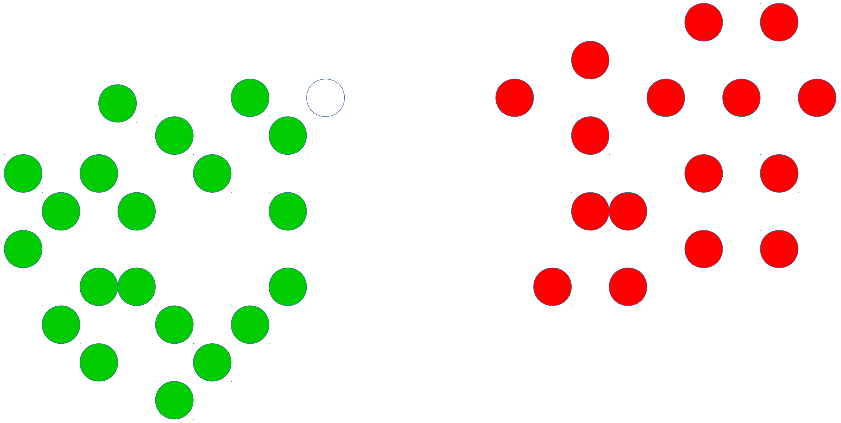
Matthew A. Russell

Backup

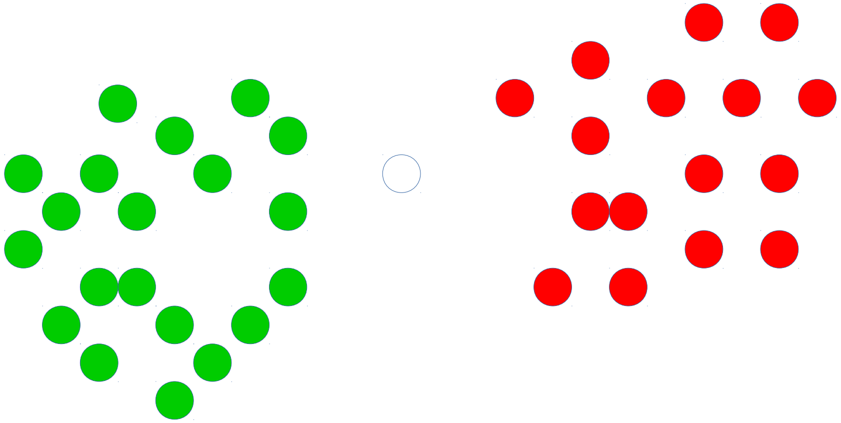
Problem three: Clustering



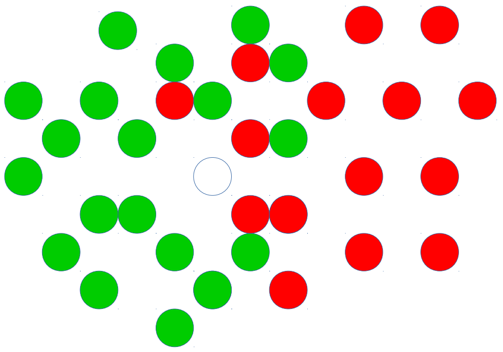
Problem three: Clustering



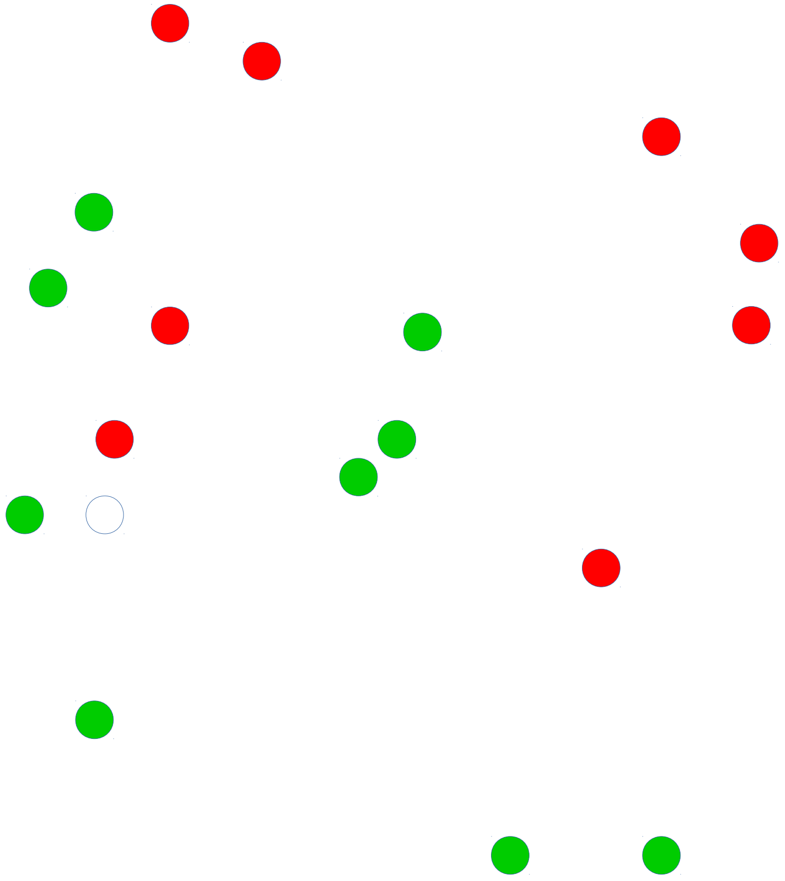
Problem three: Clustering



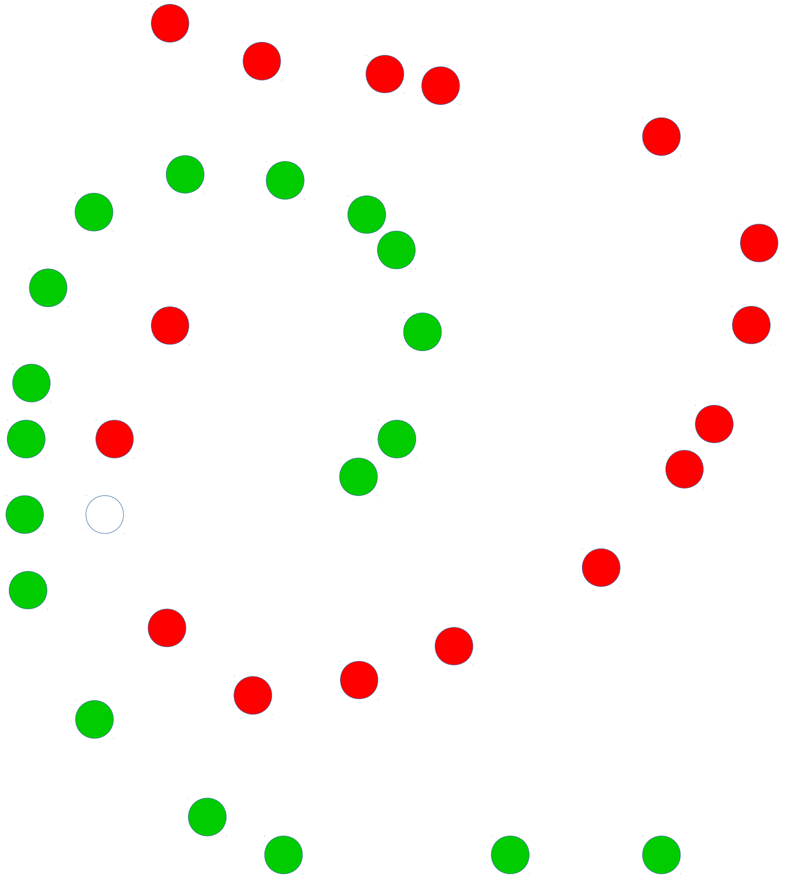
Problem three: Clustering



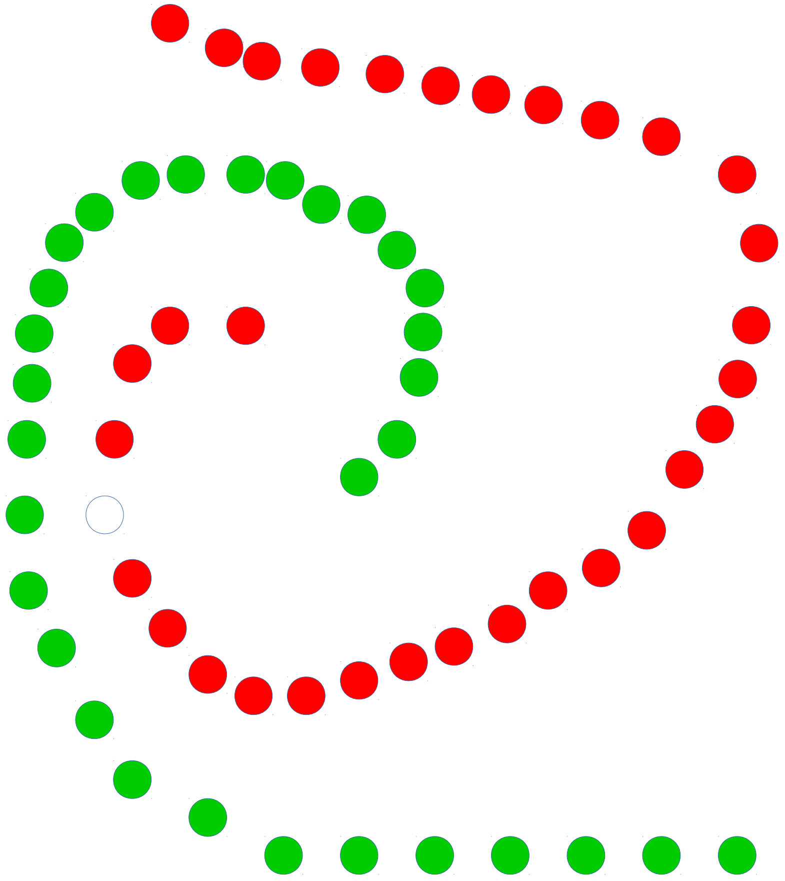
Problem four: Manifold learning



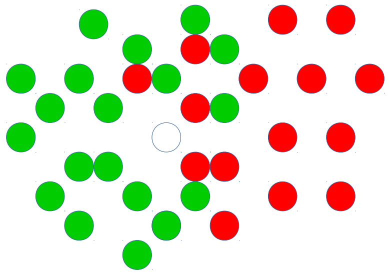
Problem three: Manifold learning

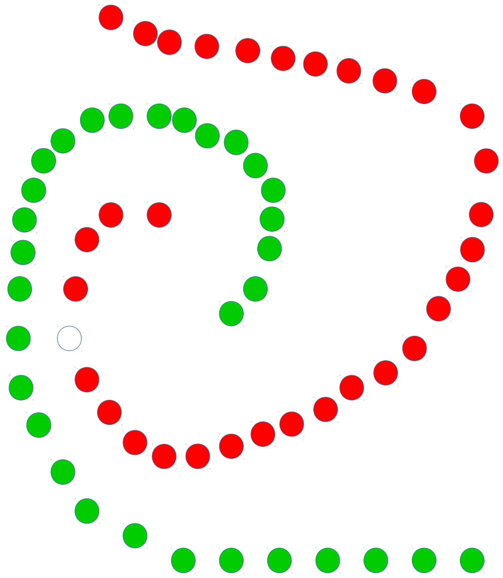


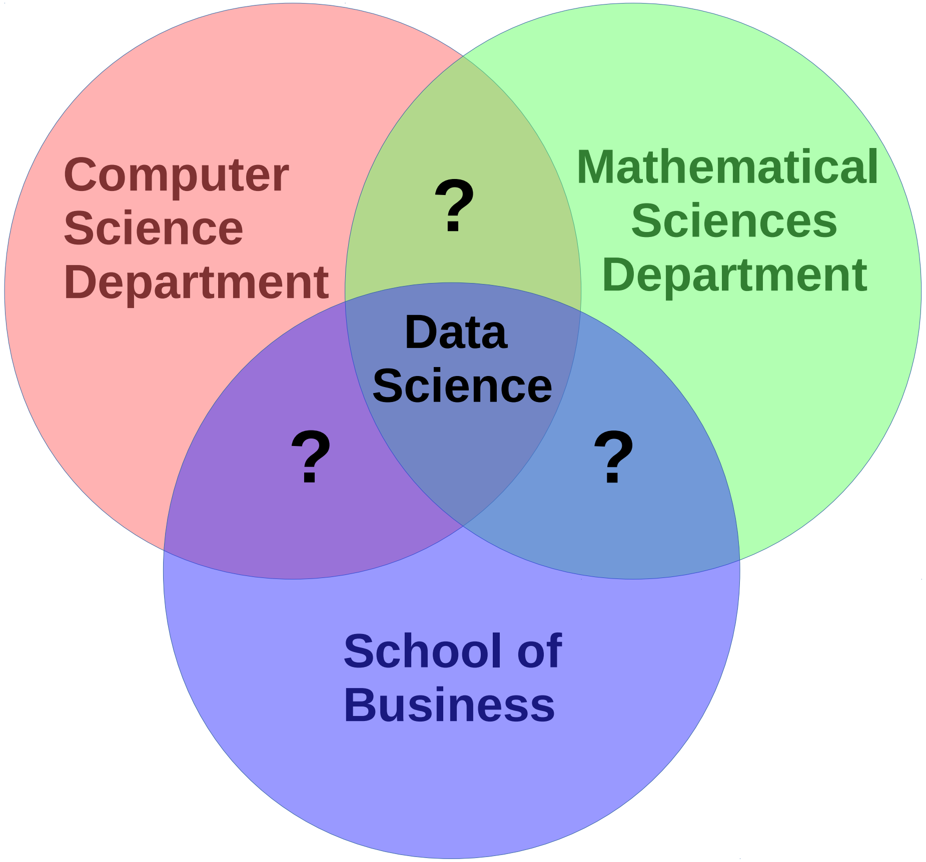
Problem three: Manifold learning



Questions? Comments? Jokes?







Computer Science Department

Mathematical Sciences Department

School of Business

Data Science

?

?

?