

# **INTRODUCTION TO DATA SCIENCE**

**CSC 495/663**



THE UNIVERSITY *of* NORTH CAROLINA  
**GREENSBORO**

# DATA SCIENCE

## What Data Science?

- Tools, techniques, science to solve Big Data / Data problems
  - Big data is a broad term for *data-sets* that are large and/or complex.
- Challenges include analysis, capture, data-curation, search, machine learning, statistics, sharing, storage, transfer, visualization, and information privacy.

## How is it different from Data Analytics

- Data Analytics is a sub-set of Data Science
- Data Analytics is mostly about using Statistical / Computational methods and enterprise software to gain insights in the data
- Does not solve all components of data problems



# DATA SCIENCE

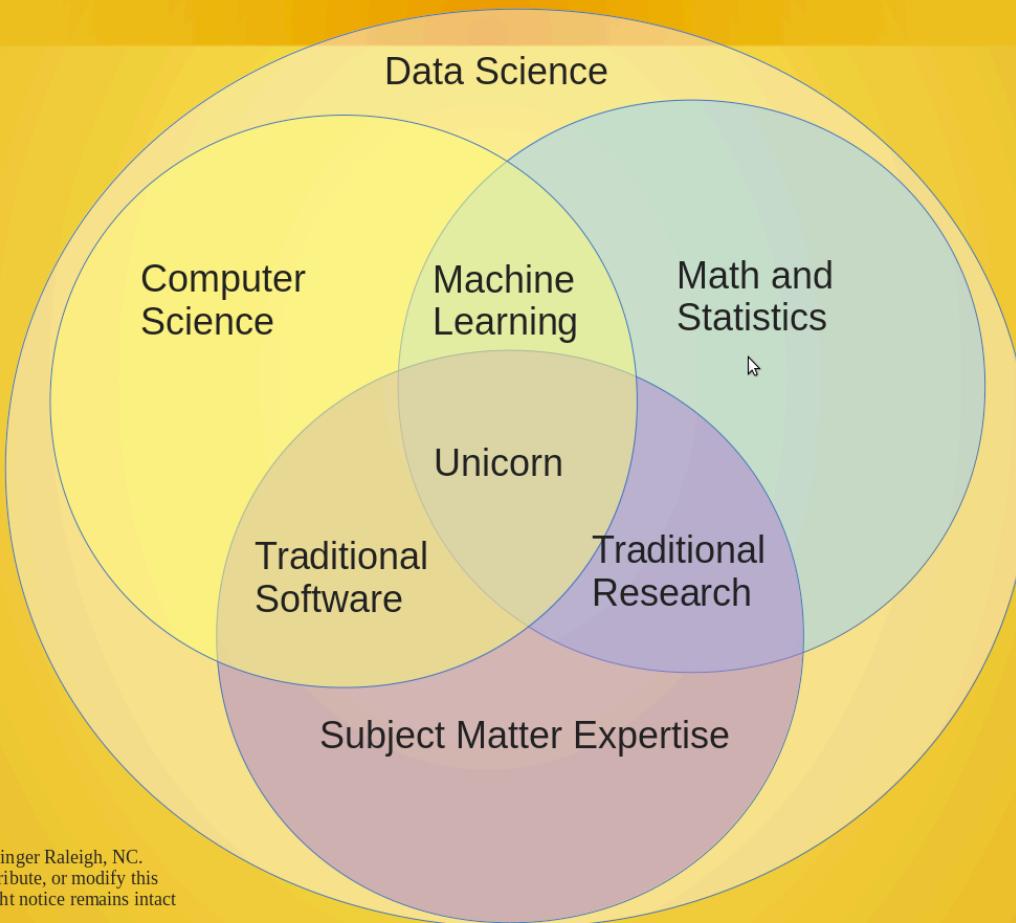
**Goal of Data Science:**

**“Turn Data into Data Products”**



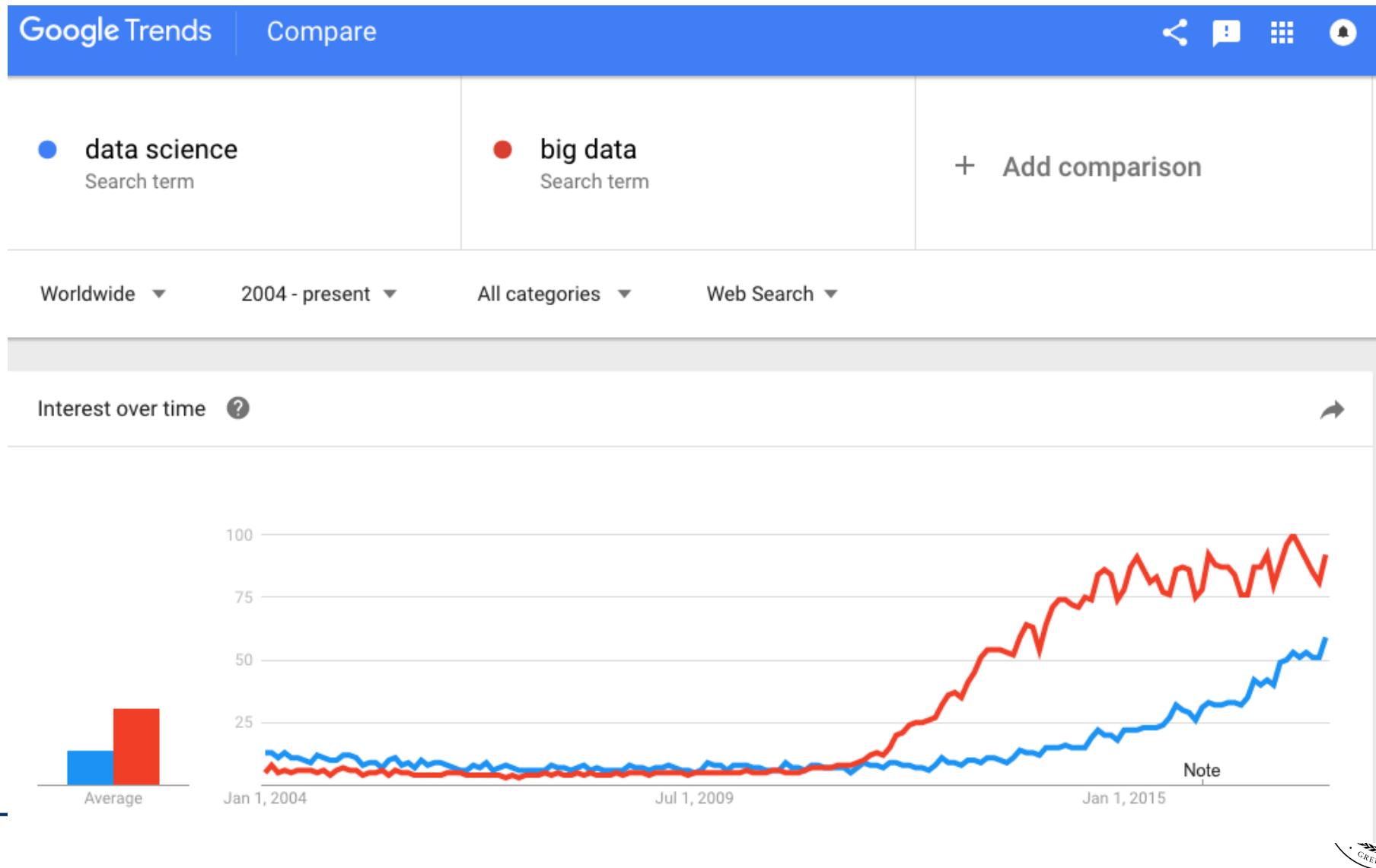
# DATA SCIENCE

## Data Science Venn Diagram v2.0



# DATA SCIENCE

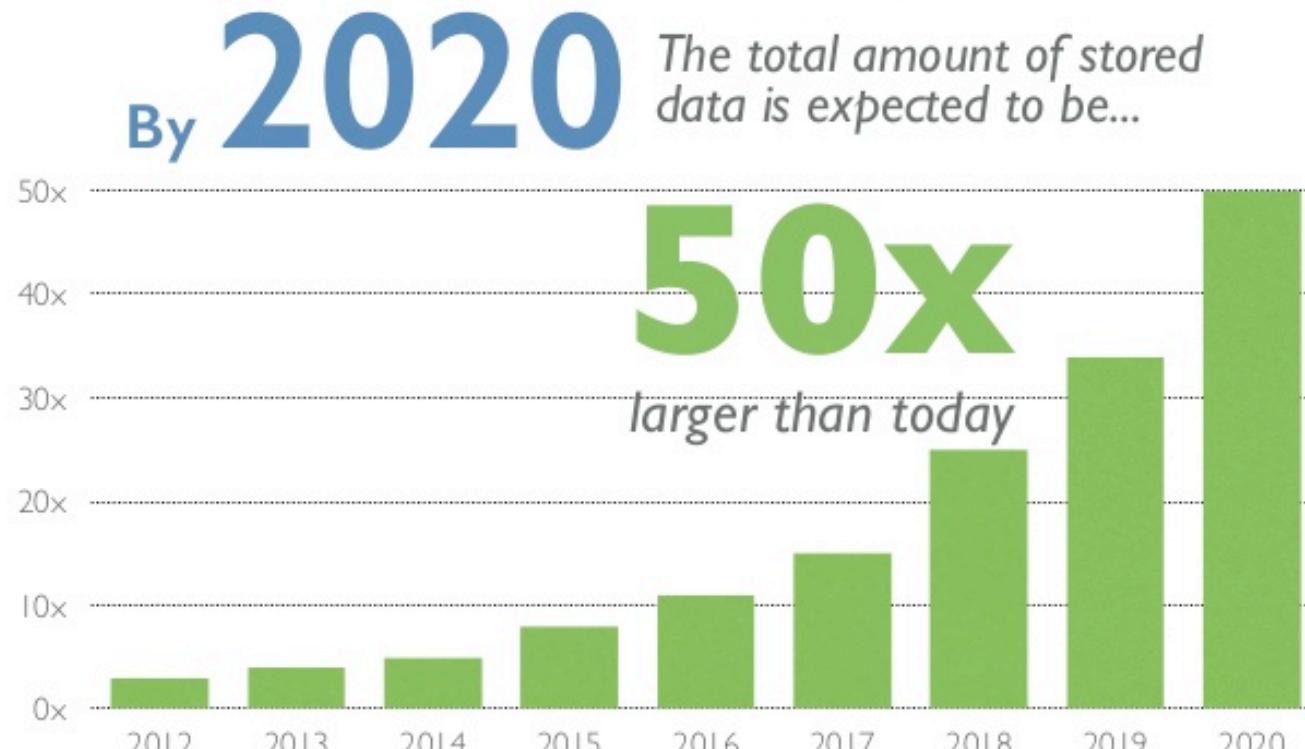
## Why is Data Science exciting? – Google trends



# DATA SCIENCE

## Why is Data Science exciting? – Big Data

In 2014 the amount of information stored worldwide exceeded 5 ZetaBytes

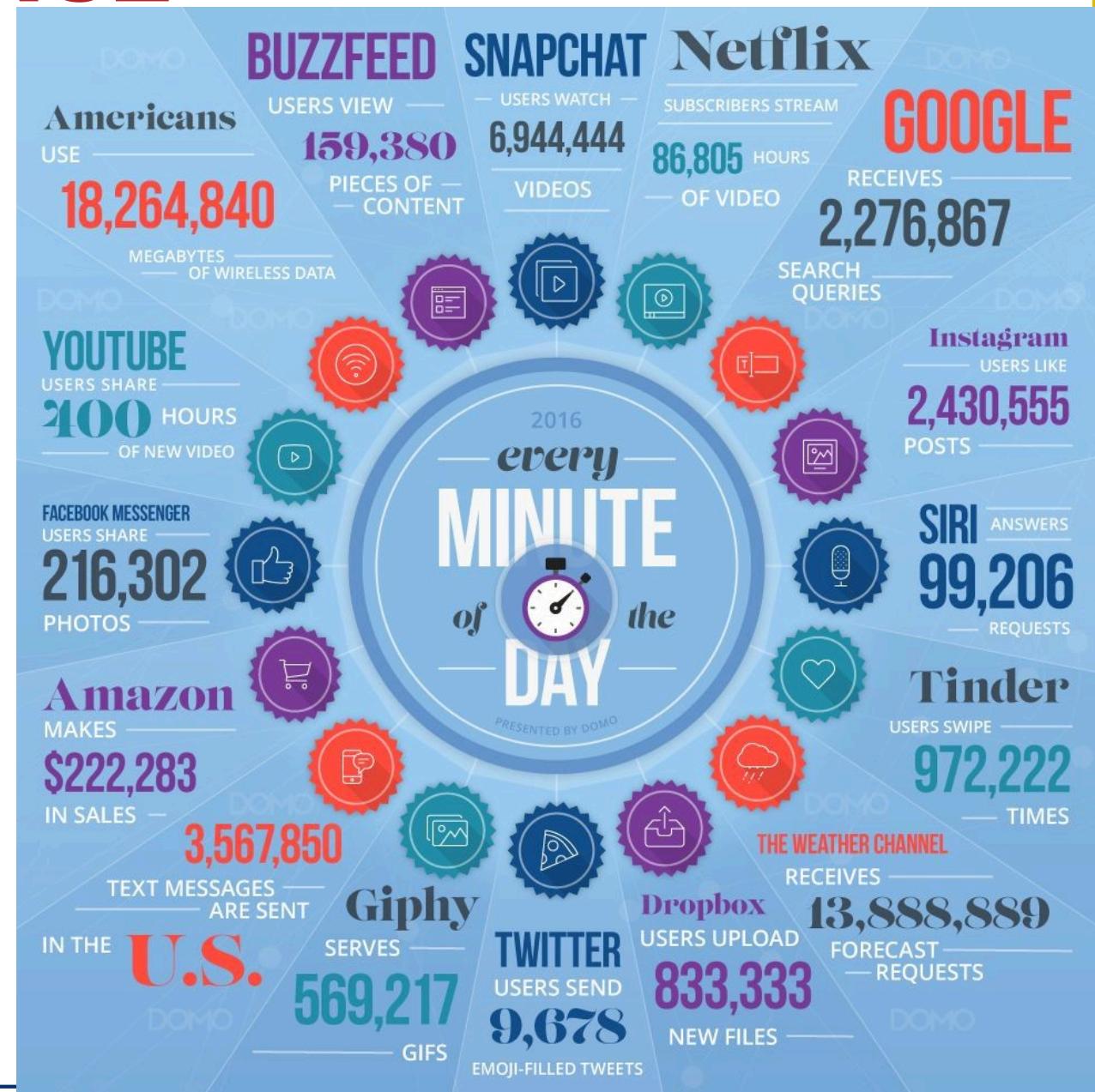


**\*Zettabyte** = 35,000,000,000,000,000,000 bytes



# DATA SCIENCE

Why is Data Science  
exciting? – Big Data



# DATA SCIENCE - EXAMPLES

# 2012 Nate Silver's prediction of election



For the Nate-haters, here's the 538 prediction and actual results side by side [pic.twitter.com/jbny4pRX](https://pic.twitter.com/jbny4pRX)



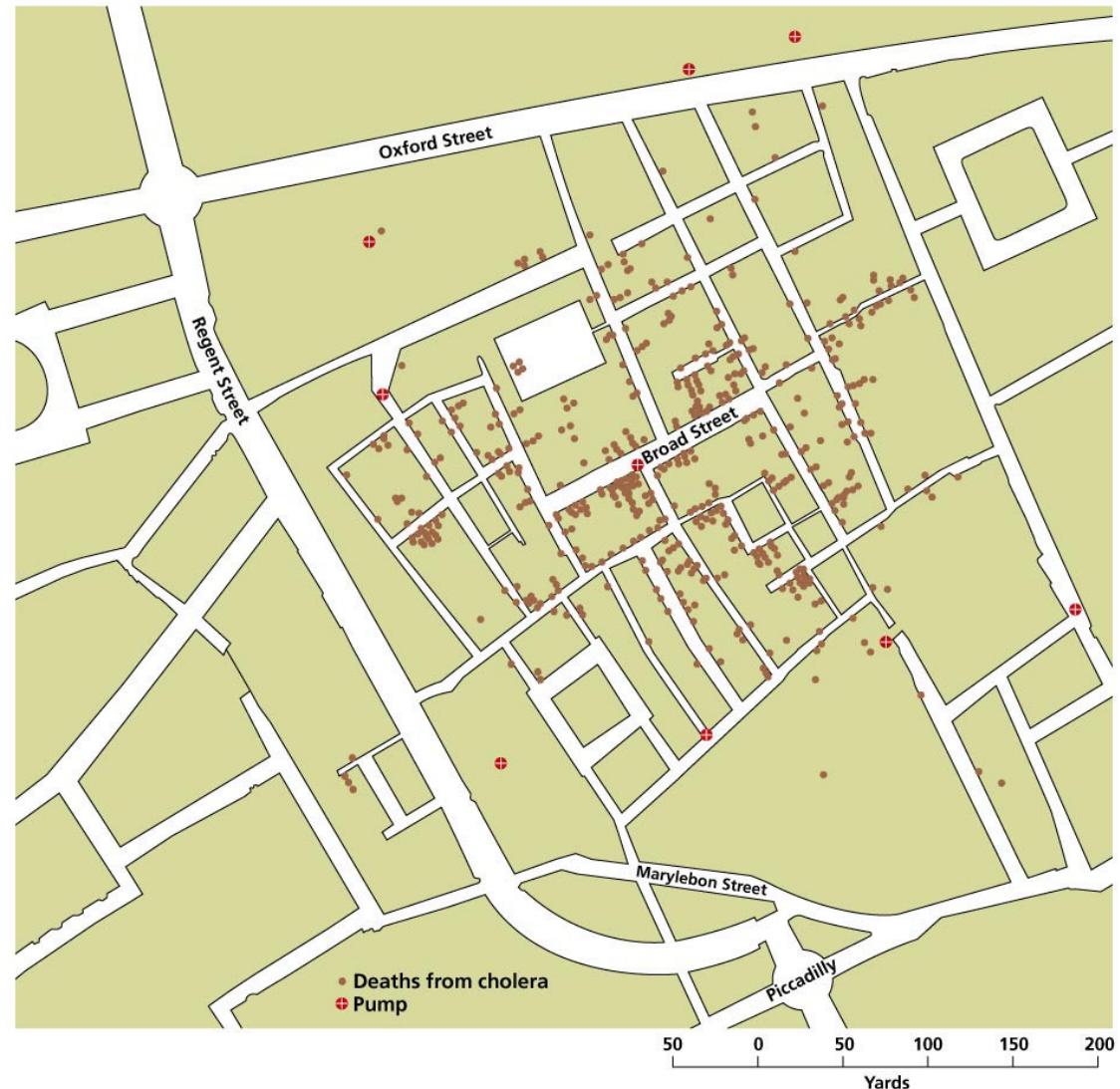
Michael Cosentino

15 hours ago



# DATA SCIENCE - EXAMPLES

- **1854 Cholera Outbreak**
  - London
  - Dr. John Snow
  - Water contamination
  - Removed pump handles

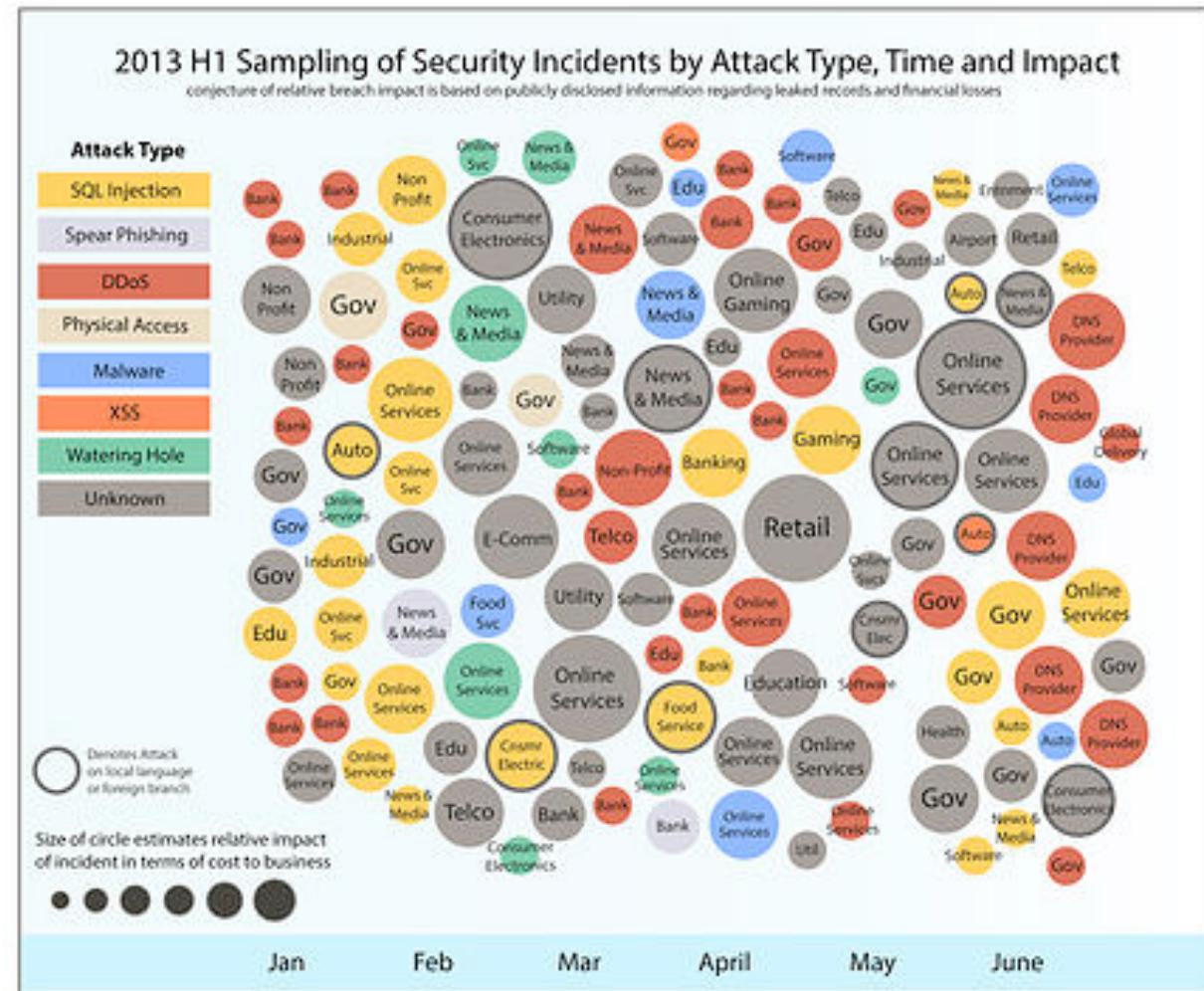


Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.



# DATA SCIENCE - EXAMPLES

- **Cyber-Security**
    - Fraudulent credit-card and transaction activity
    - Predicting criminal activity
    - Detecting Malware
    - Critical Infrastructure Protection



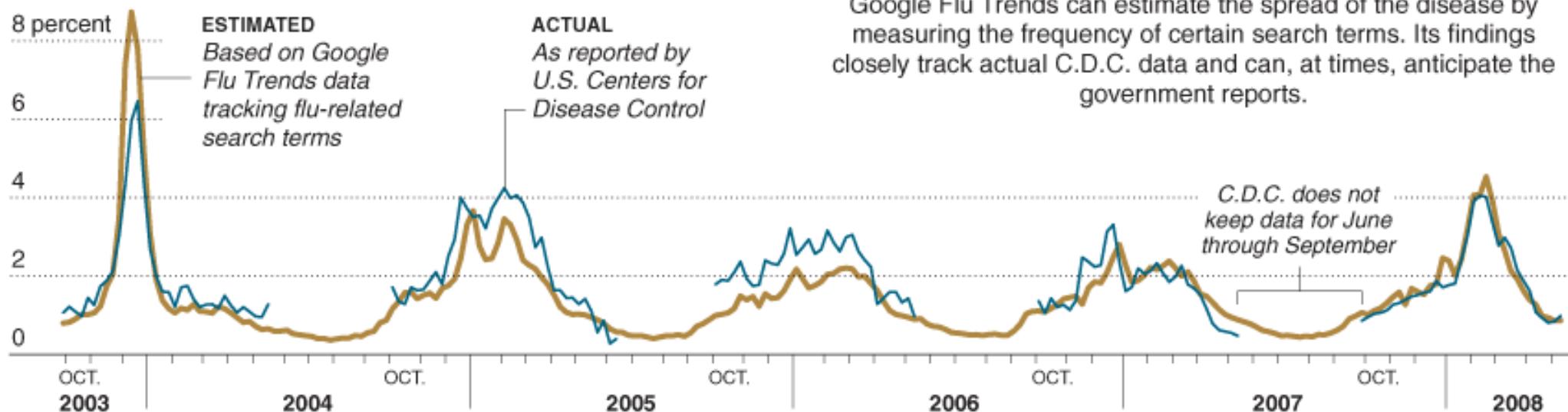
Source: IBM X-Force® Research and Development



# DATA SCIENCE - EXAMPLES

## Why is Data Science exciting?

PERCENT OF HEALTH VISITS FOR FLU-LIKE SYMPTOMS *Mid-Atlantic region*



Sources: Google; Centers for Disease Control

## Using Google to Monitor the Flu

Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.

THE NEW YORK TIMES



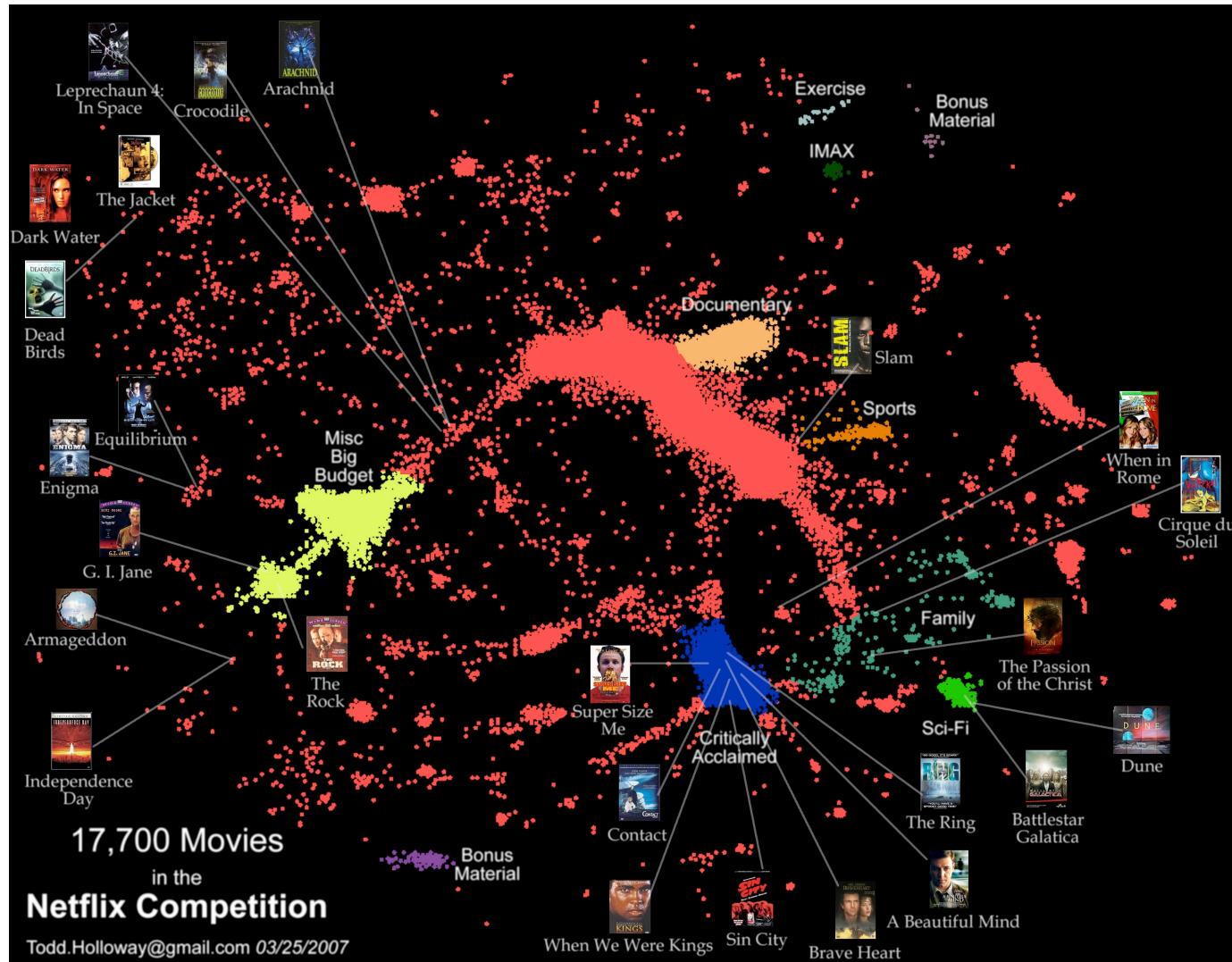
# DATA SCIENCE - EXAMPLES

- **Netflix Challenge**
  - How much someone is going to like a movie
  - Prior preferences of users
  - <user, movie, date, grade>
  - Training set 100 M
  - Qualifying set 2 M
  - Baseline *Cinematch*
  - \$1 Million prize awarded



# DATA SCIENCE - EXAMPLES

- Netflix Challenge



# DATA SCIENCE - EXAMPLES

- Netflix Challenge

**Netflix Prize**

Home Rules Leaderboard Register Update Submit Download

**Leaderboard 10.05%** Display top  leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	<a href="#">BellKor's Pragmatic Chaos</a>	0.8558	10.05	2009-06-26 18:42:37
<b>Grand Prize - RMSE &lt;= 0.8563</b>				
2	<a href="#">PragmaticTheory</a>	0.8582	9.80	2009-06-25 22:15:51
3	<a href="#">BellKor in BigChaos</a>	0.8590	9.71	2009-05-13 08:14:09
4	<a href="#">Grand Prize Team</a>	0.8593	9.68	2009-06-12 08:20:24
5	<a href="#">Dace</a>	0.8604	9.56	2009-04-22 05:57:03
6	<a href="#">BigChaos</a>	0.8613	9.47	2009-06-23 23:06:52

A yellow arrow points to the '% Improvement' column in the leaderboard table, highlighting the value '10.05' for the winning team.

2009
DATE 09.21.09
PAY TO THE ORDER OF: BellKor's Pragmatic Chaos
AMOUNT: ONE MILLION 00/100
FOR The Netflix Prize
Reed Hastings



# DATA SCIENCE - EXAMPLES



NOAA



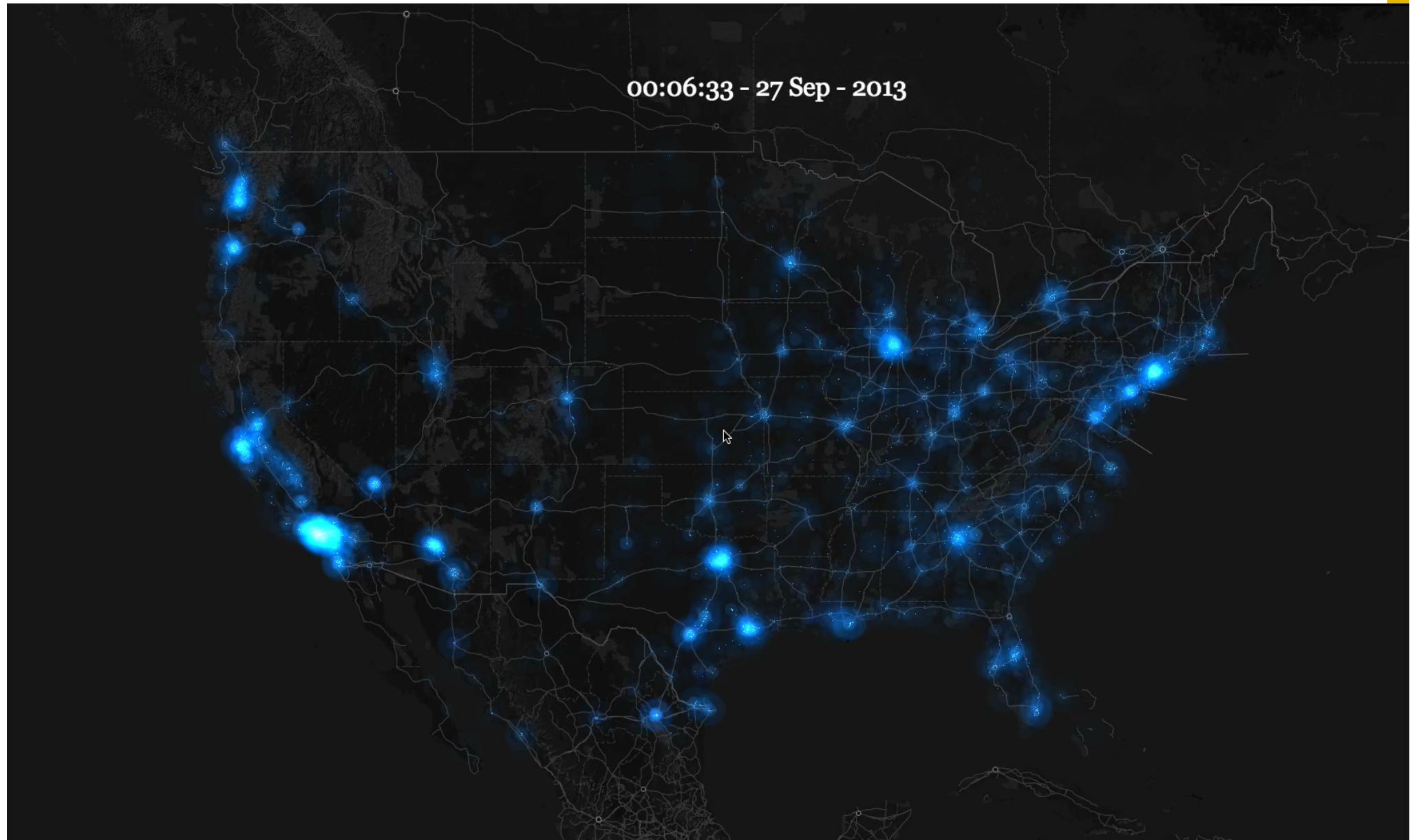
# DATA SCIENCE - EXAMPLES



Twitter and Flickr



# DATA SCIENCE - EXAMPLES



Twitter



# DATA SCIENCE - EXAMPLES



Data source: Facebook (2011)



# DATA SCIENTIST

## What is a Data Scientist job?

- **Data Scientist Job** - “part analyst, part artist.” – IBM
- **Learn and utilize a large number of tools and methods**
- **Data Driven Products**
- **\$118,709 (data-scientist) versus \$64,537 (programmer)**
- **A shortage of 140,000 to 190,000 jobs by 2018**
- **The fastest growing job segment**
- **Why are they in demand?**
  - Need of expertise in data science is far more than available



# DATA SCIENTIST

## How to be a Data Scientist?

- **Learn a lot of tools, methods and techniques**
- **Let the data speak for itself**
  - Traditional approach
    - Have a question / hypothesis, use the data to prove it
  - New approach
    - Dive into the data and figure out what it says
    - Ask why at every step
- **Be creative with the data**
  - Data munging or wrangling – convert to different formats and analyze
  - Visualize to understand
  - Learn from the data
- **Keep learning**



# DATA SCIENTIST

**How does the course help?**

- **New methods and techniques**
  - Access
  - Storage
  - Analysis
  - Machine Learning
  - Visualization
- **Change the old way of thinking**
- **Creative programming**

**The course is not going to teach you everything ☺**

**But will get you started on the path.....**



# **CLASS PROJECT**

**CSC 495/663**



THE UNIVERSITY *of* NORTH CAROLINA  
**GREENSBORO**

# DATASETS

- **Academic Datasets**

- UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/>)
- Stanford Large Network Dataset Collection (<http://snap.stanford.edu/data/>)
- Inter-university Consortium for Political and Social Research (<http://www.icpsr.umich.edu/>)
- Pittsburgh Science of Learning Center's DataShop (<http://www.learnlab.org/technologies/datasshop/>)
- Academic Torrents (<http://academictorrents.com/>)

- **Private Companies**

- Quandl (<https://www.quandl.com/>)
- Amazon Web Services Public Data Sets (<http://aws.amazon.com/datasets/>)
- Kaggle (<http://www.kaggle.com/>)
- Nytimes (<http://developer.nytimes.com/docs>)



# DATASETS

- **Gov. and NGO's**

- Data.gov (<https://www.data.gov/>)
- NYC Open Data (<https://nycopendata.socrata.com/>)
- DC Open Data Catalog (<http://data.dc.gov/>)
- OpenDataDC (<http://www.opendatadc.org/>)
- DataLA (<https://data.lacity.org/>)
- Project Open Data Dashboard  
(<http://data.civicagency.org/>)
- data.gov.uk (<http://data.gov.uk/>)
- US Census Bureau](<http://www.census.gov/>)
- World Bank Open Data (<http://data.worldbank.org/>)
- Humanitarian Data Exchange (<http://docs.hdx.rwlabs.org/>)
- Sunlight Foundation (<http://sunlightfoundation.com/api/>)
- ProPublica Data Store  
(<https://projects.propublica.org/data-store/>)



# DATASETS

- Other resources
  - 20 Big Data Sources  
(<http://www.smartdatacollective.com/bernardmarr/235366/big-data-20-free-big-data-sources-everyone-should-know>)
  - Center for Data Innovation  
(<http://www.datainnovation.org/category/publications/data-set-blog/>)
  - Data Science Central  
(<http://www.datasciencecentral.com/>)
  - Python API's (<http://www.pythonforbeginners.com/api/list-of-python-apis>)



# PROJECT IDEAS

## Twitter Human Mobility

- Mapping movement of Twitter users on temporal domain
- Utilize the geo-located tweets
- Twitter API  
(<https://dev.twitter.com/overview/documentation>)
- Google Maps  
(<https://developers.google.com/maps/?hl=en>)

## Publication Topic modeling and Network Analysis

- Topic modeling of publications to visualize interest in areas
- Connectivity of topics across areas
- Topic Modeling  
(<https://www.cs.princeton.edu/~blei/topicmodeling.html>)
- Elsevier API (<http://dev.elsevier.com/>)
- MIT scholarly resources (<http://libguides.mit.edu/apis>)



# PROJECT IDEAS

## Wiki Topic modeling and Network analysis

- Wikipedia topic modeling and connectivity of topics
- Interests based on edits across temporal domain
- Wikipedia API  
([https://www.mediawiki.org/wiki/API:Main\\_page](https://www.mediawiki.org/wiki/API:Main_page))
- DBpedia (<http://wiki.dbpedia.org/>)

## IMDB data analysis

- Graph analysis of IMDB dataset
  - Actors, genres, producers, directors connectivity
- Analysis of important features for movie success
- Predicting movie rating
- IMDB API (<http://www.omdbapi.com/>)



# PROJECT IDEAS

## Sentiment models - Twitter, IMDB, Amazon reviews

- Categorizing Polarity of tweets, IMDB movie reviews, amazon reviews
- Plotting heat maps for twitter
- Clustering popular products in amazon
- Predicting movie success
- Twitter API (<https://dev.twitter.com/overview/documentation>)
- 35 million Amazon reviews  
(<https://snap.stanford.edu/data/web-Amazon.html>)

## Million Song Dataset Analysis

- Popularity vs genre vs time
- Lyrics ngram analysis
- Clustering of songs
- API (<http://labrosa.ee.columbia.edu/millionsong/>)



# PROJECT IDEAS

## Facebook Network Analysis

- Clustering of Friends
- Influential Networks
- Post analysis
- Graph API (<https://developers.facebook.com/docs/graph-api>)

## Campus Safety Dataset Analysis

## Articles NY Times LDA over time

## Twitter Flu Tracker



# TEAMS

- **Teams Consist of 4-5 members**
- **Collaborative code repository**
  - Tracking via commit logs
- **Implementation via python**
- **Each team member must understand the methodology**
- **Visualizations for naïve user**
- **Effort vs Result**

