

INTRODUCTION TO DATA SCIENCE

CSC 405/605



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

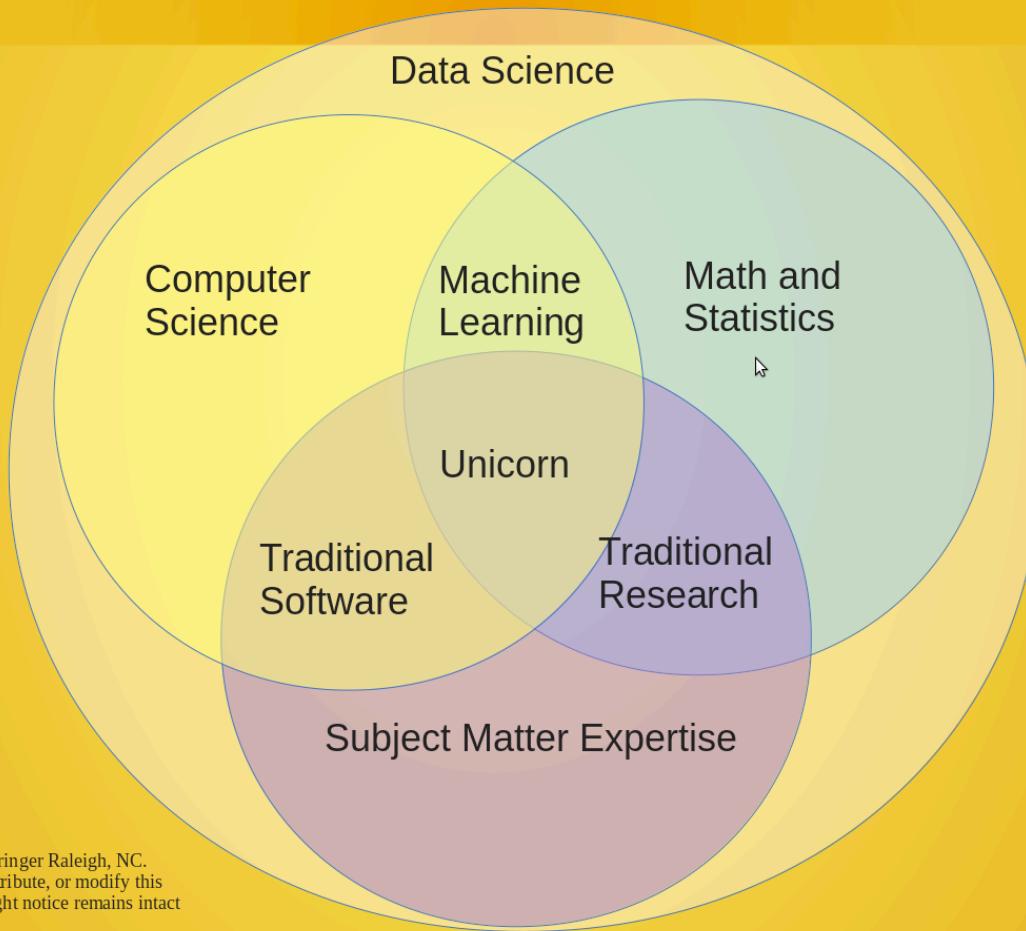
Goal of Data Science:

“Turn Data into Data Products”



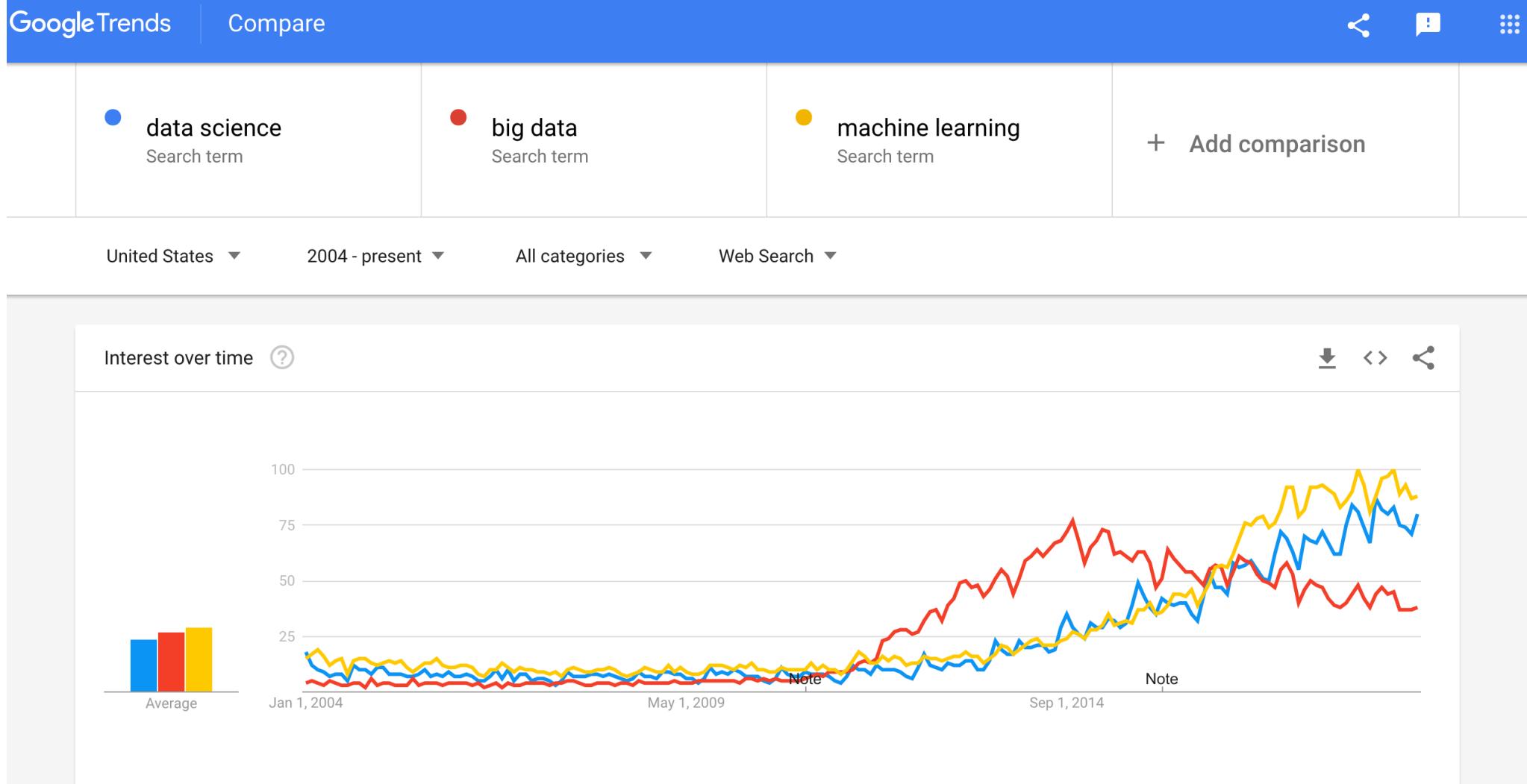
DATA SCIENCE

Data Science Venn Diagram v2.0



DATA SCIENCE

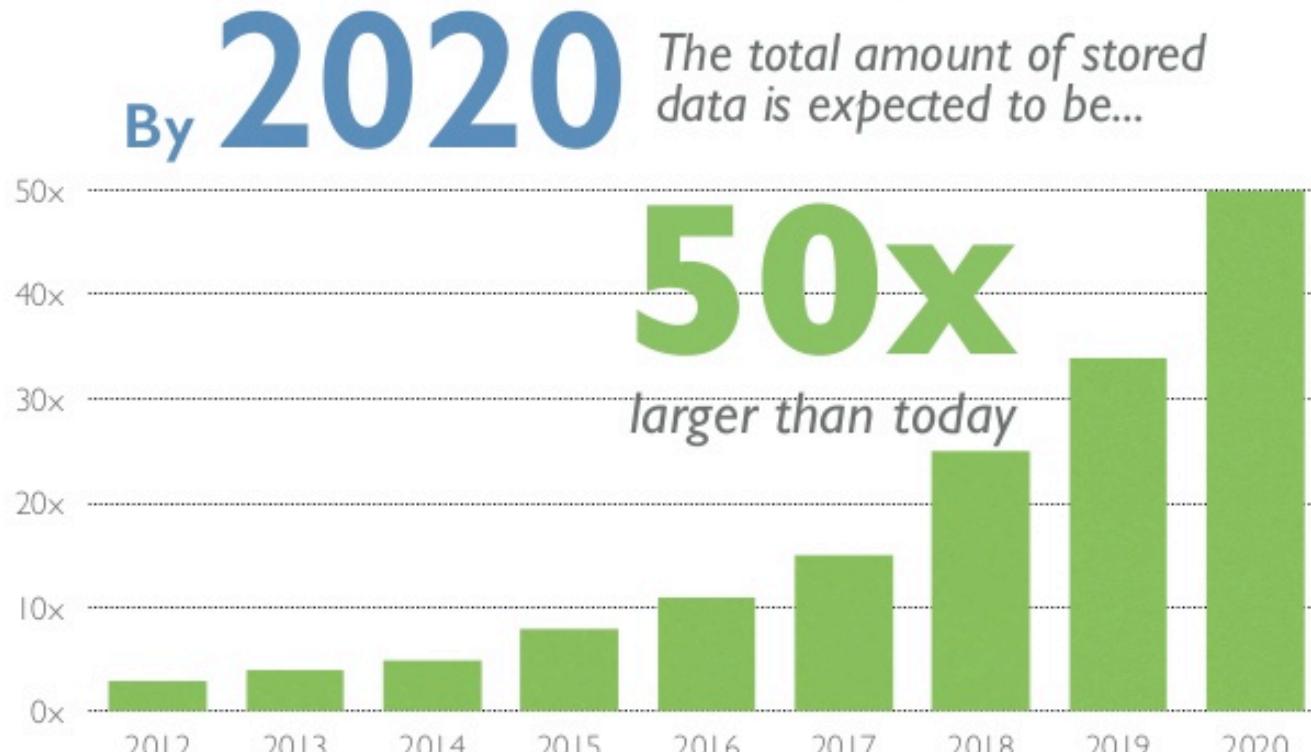
Why is Data Science exciting? – Google trends



DATA SCIENCE

Why is Data Science exciting? – Big Data

In 2014 the amount of information stored worldwide exceeded 5 ZetaBytes

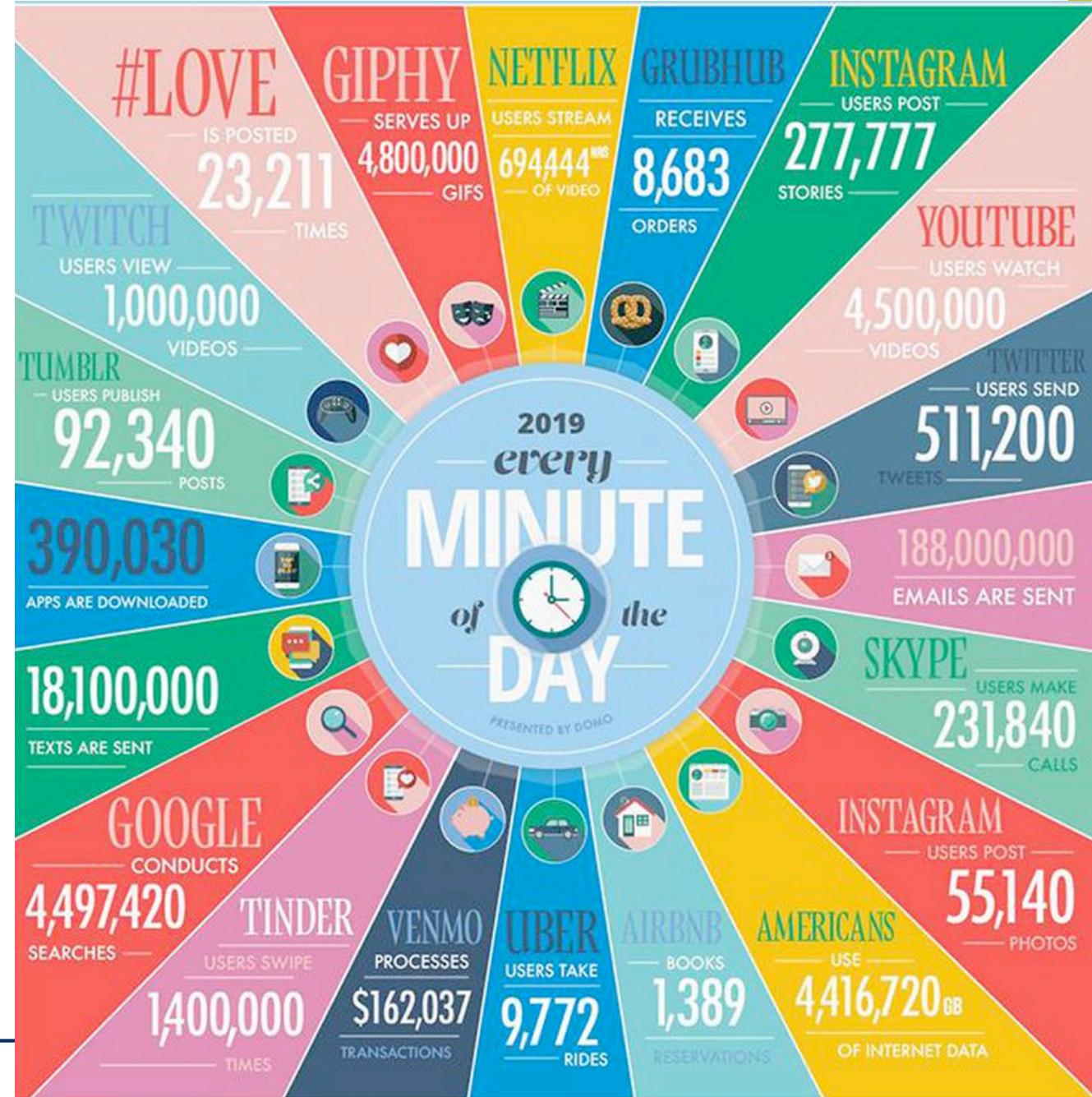


***Zettabyte** = 35,000,000,000,000,000,000 bytes



DATA SCIENCE

Why is Data Science
exciting? – Big Data



DATA SCIENCE - EXAMPLES

2012 Nate Silver's prediction of election

538 prediction



actual



For the Nate-haters, here's the 538 prediction and actual results side by side pic.twitter.com/jbny4pRX



Michael Cosentino

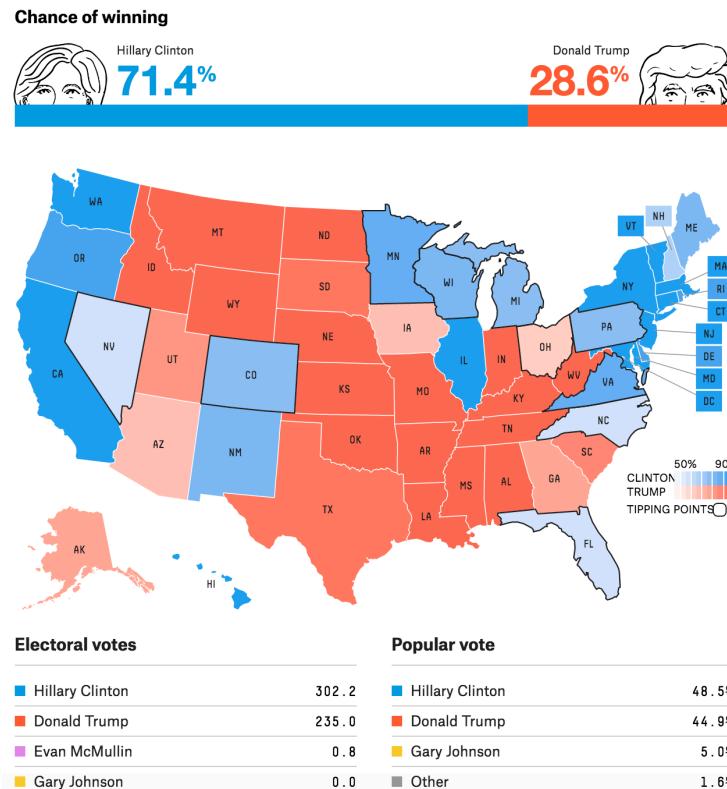
15 hours ago



DATA SCIENCE - EXAMPLES

2016 Nate Silver's prediction of election

<https://fivethirtyeight.com/features/how-i-acted-like-a-pundit-and-screwed-up-on-donald-trump/>



FiveThirtyEight

Politics Sports Science & Health Economics Culture



How I Acted Like A Pundit And Screwed Up On Donald Trump

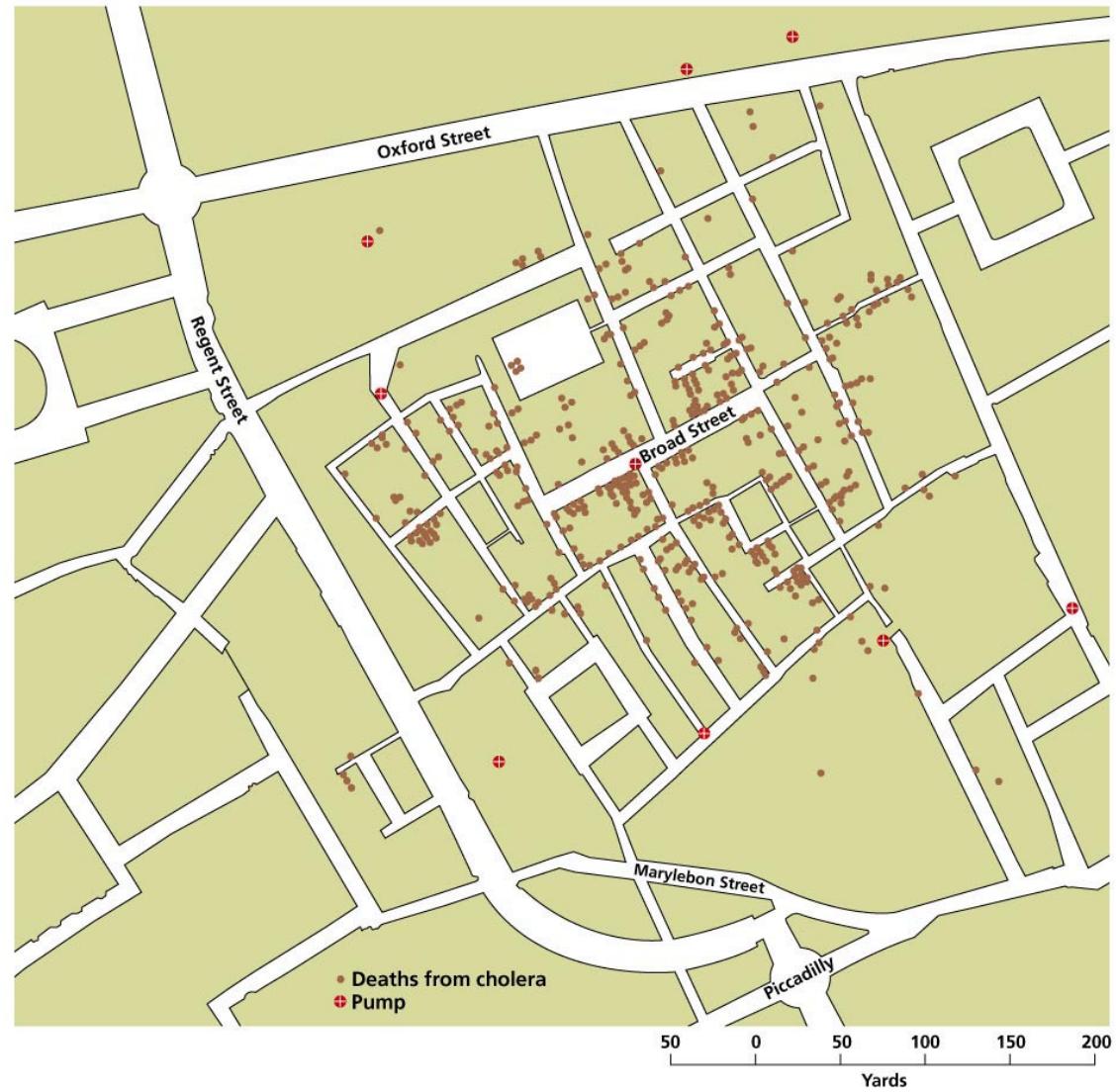
Trump's nomination shows the need for a more rigorous approach.

By Nate Silver
Filed under 2016 Election
Published May 18, 2016



DATA SCIENCE - EXAMPLES

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

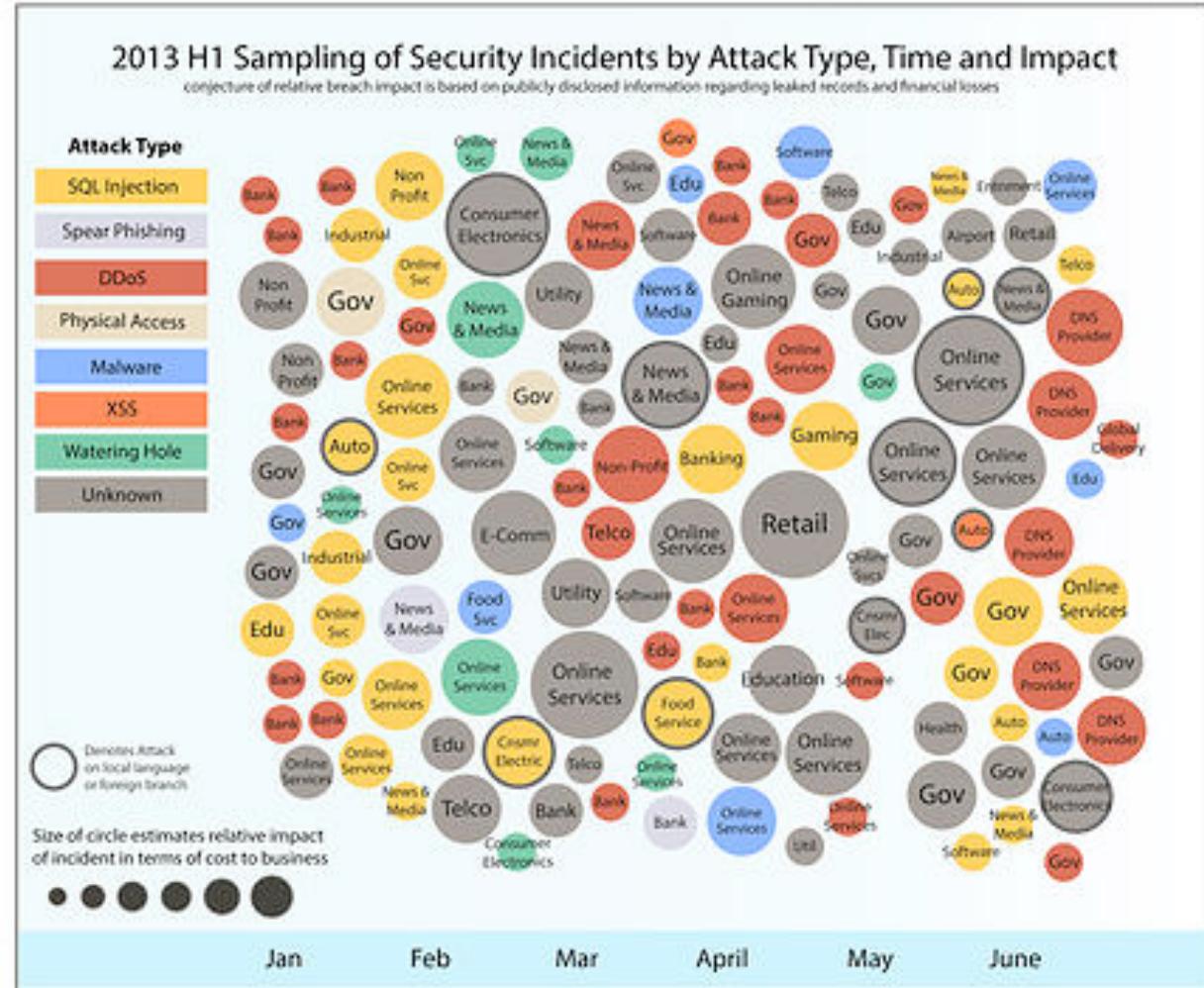


Copyright © 2006 Pearson Education, Inc., publishing as Benjamin Cummings.



DATA SCIENCE - EXAMPLES

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



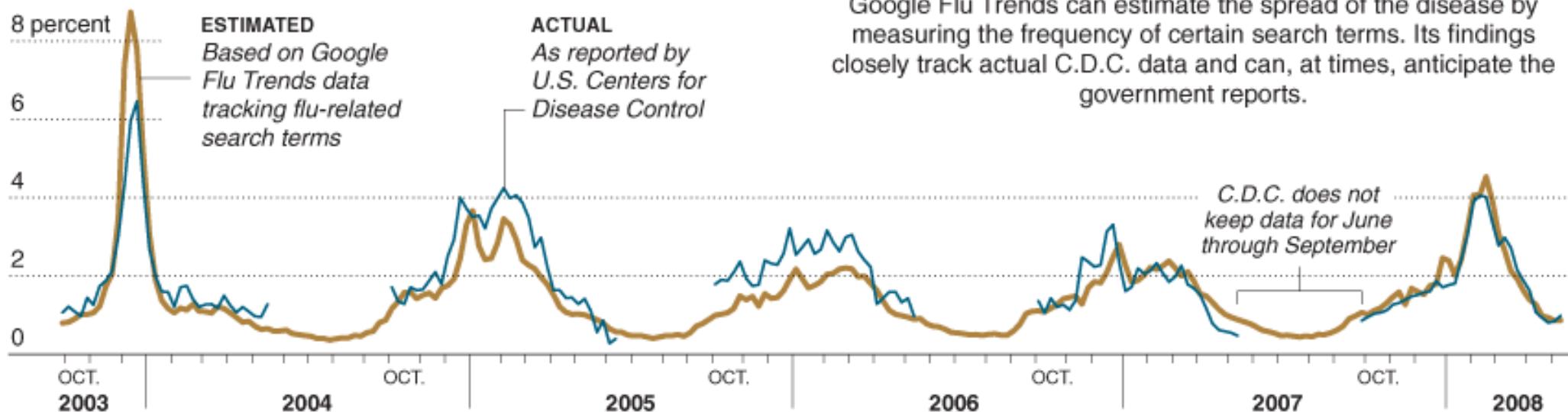
Source: IBM X-Force® Research and Development



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

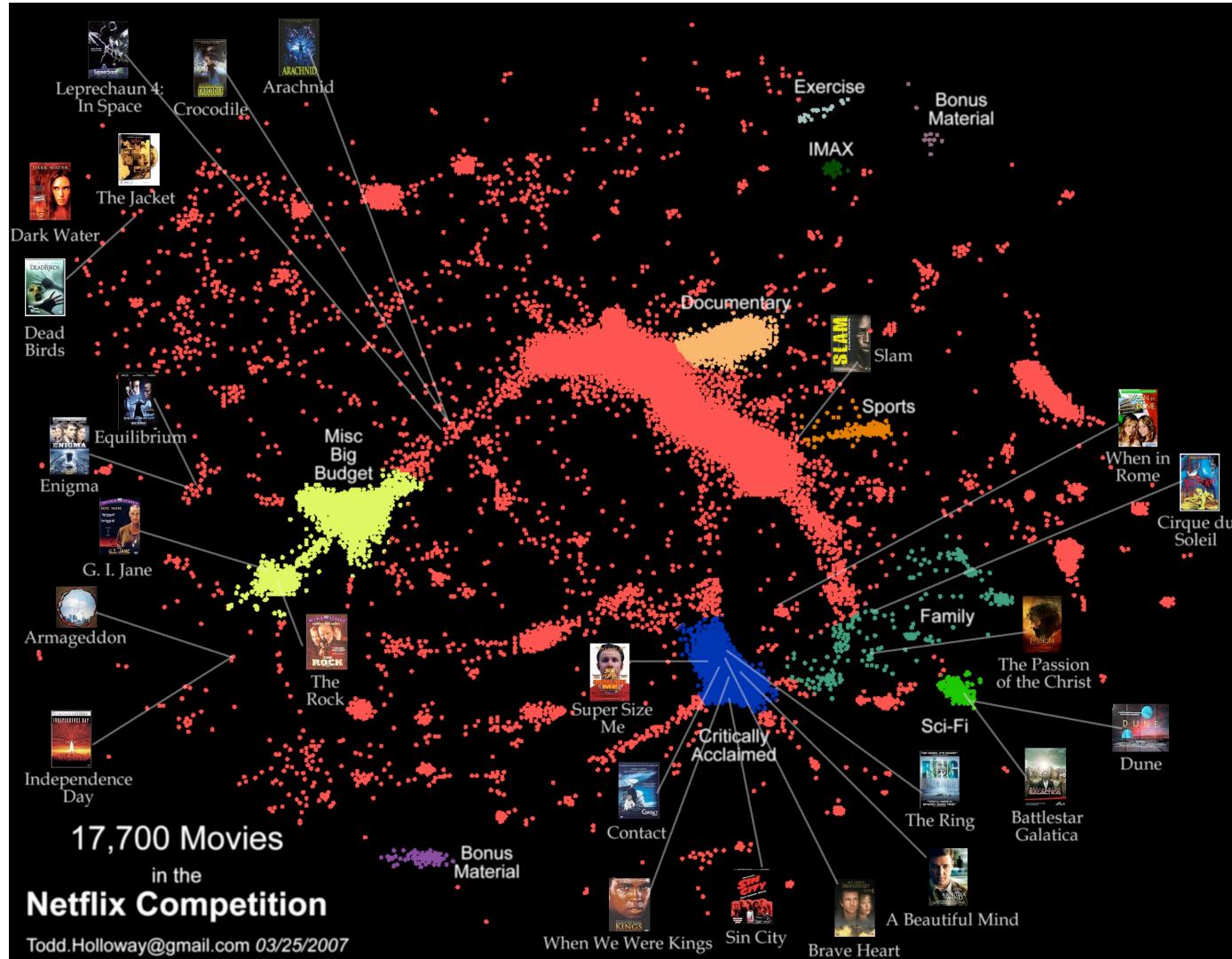
- **Netflix Challenge**

- How much someone is going to enjoy a movie
- Prior preferences of users
- $\langle \text{user}, \text{movie}, \text{date}, \text{grade} \rangle$
- 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	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:06:52

A yellow arrow points to the '% Improvement' column in the table, highlighting the 10.05% improvement achieved by BellKor's Pragmatic Chaos.



DATA SCIENCE - EXAMPLES



NOAA



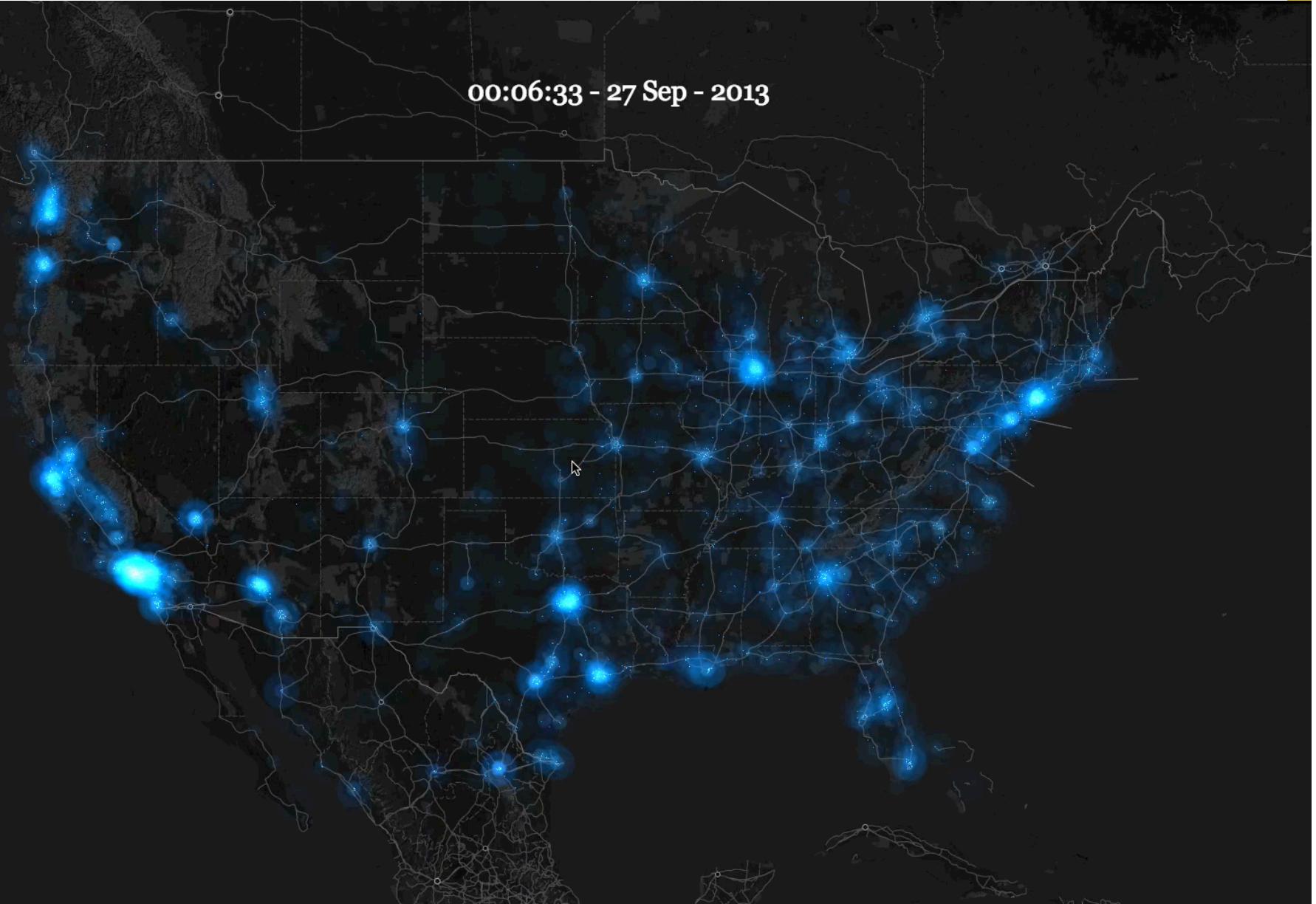
DATA SCIENCE - EXAMPLES



Twitter and Flickr



DATA SCIENCE - EXAMPLES



DATA SCIENCE - EXAMPLES



Data source: Facebook (2011)



DATA SCIENTIST

What is a Data Scientist job?

- **Data Scientist Job - “part analyst, part artist.” – IBM**
- **Data Driven Products**
- **\$120, 931 (data-scientist) versus \$66,537 (programmer)**
- **The fastest growing job segment**

 Economic Graph

Research

Resources

Blog

About

-
- **Skills Gaps | Demand for data scientists is off the charts –**
In 2015, there was a national surplus of people with data science skills. An employer in **Dallas** or **Atlanta** who wanted to hire data scientists had plenty of options; aside from in a few tech or finance-heavy cities like **San Francisco**, **New York City** and **Boston**, there weren't many local shortages. But today, 3 years later, the picture has changed markedly: data science skills shortages are present in almost every large U.S. city. Nationally, we have a shortage of 151,717 people with data science skills, with particularly acute shortages in **New York City** (34,032 people), the **San Francisco Bay Area** (31,798 people), and **Los Angeles** (12,251 people). As more industries rely on big data to make decisions, data science has become increasingly important across all industries, not just tech and finance. In that sense, it's a good proxy for how today's cutting-edge skills like AI & machine learning may spread to other industries and geographies in the future.



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 – Very Important***



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.....



DATASETS



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 (<https://pslcdatashop.web.cmu.edu/>)
- Academic Torrents (<http://academictorrents.com/>)

- **Private Companies**

- Data.World (<https://data.world/>)
- Quandl Financial Data (<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>)
 - PyCoders Weekly (<http://pycoders.com/>)

