Introduction to Machine Learning Applications

Spring 2021

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Agenda

- Class logistics
- Instructor
- Data and society
- Why am I excited about Data Science?
- What does it mean to be a data scientist today?
- What will we cover in course?

Class Logistics

When: Monday & Thursday 12:20 pm to 2:10 pm

Where: SAGE 3704

Website: https://spring2021introtoml.github.io/

Instructor	TA: Yuanyuan Liu
Office Hours: Tuesday 3 pm to 5 pm **Webex: https://rensselaer.webex.com/meet/manikl** **Email: manikl@rpi.edu**	**Email: liuy55@rpi.edu** **Webex: https://rensselaer.webex.com/meet/liuy55** **Office hours: Friday 11 am to 1 pm**

Lydia Manikonda (Ph.D. in CS)

- Assistant Professor
- Decision-making systems (social media)
  - focusing on public health,
  - Marketing,
  - Intelligent Systems
- Leveraging machine learning and artificial intelligence techniques

You https://forms.gle/A5PTpxVo4wdjxXmZ8

- Your major
- Programming background
- What you want out of the class

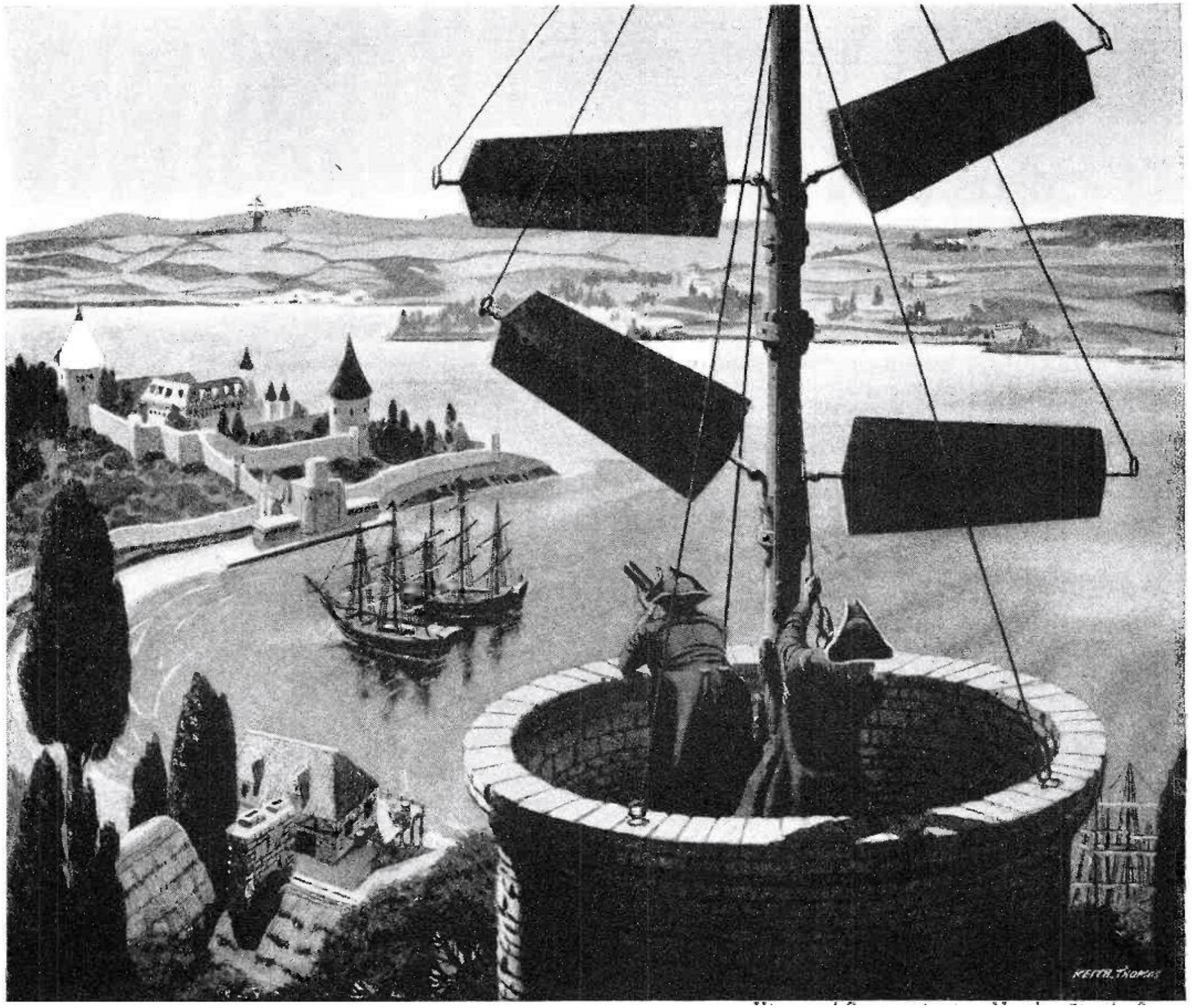
Data and Society

There have been profound changes in technology and the data/information processes define our society

Internet 0.1 Beta (18th Century)

Semaphore Telegraph

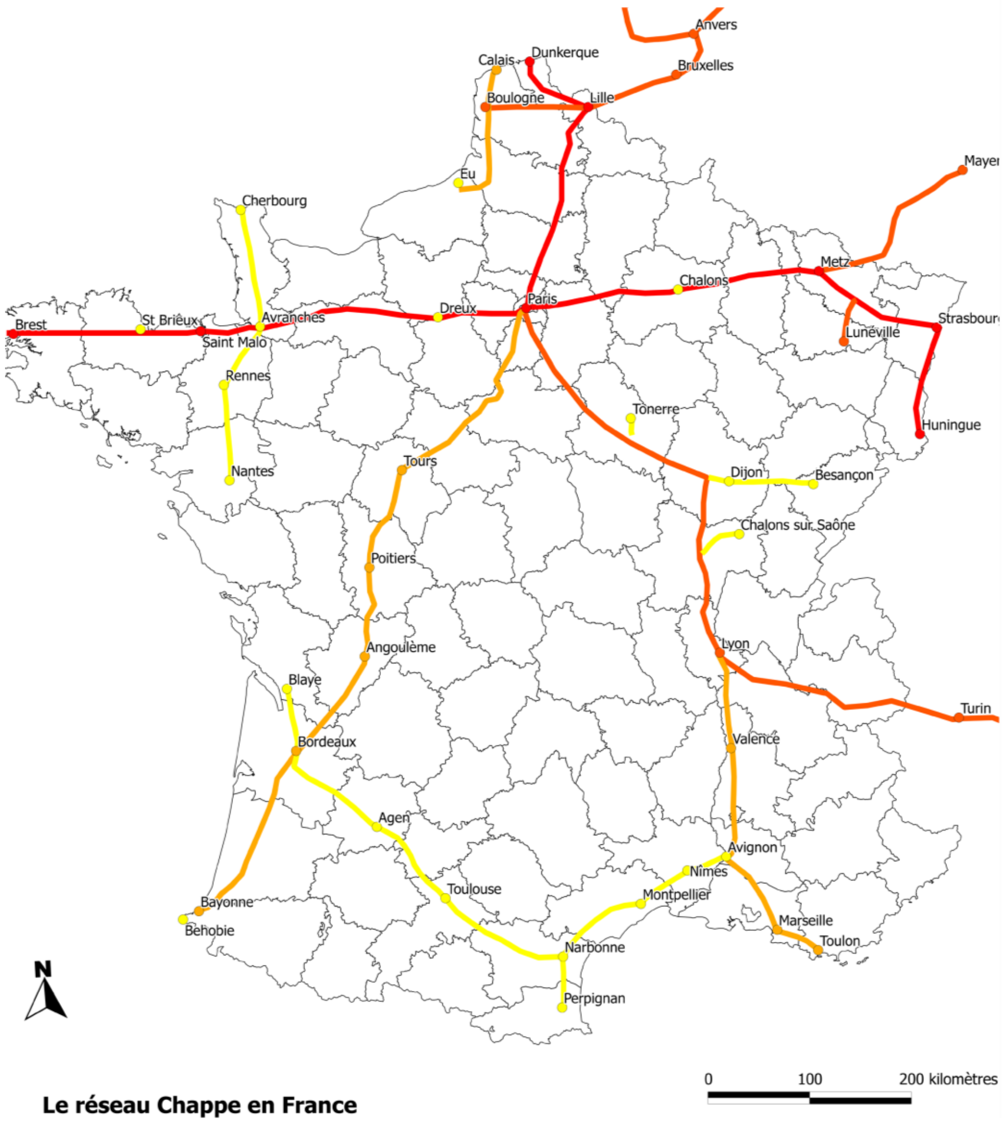
- Visual texting by position of the mechanical elements;



By The drawing is signed "Keith Thomas" in lower right corner [Public domain], via Wikimedia Commons

Internet 0.1 Beta (18th Century)

- Over 50 stations connecting France
- Shows the extent to which people will go to communicate



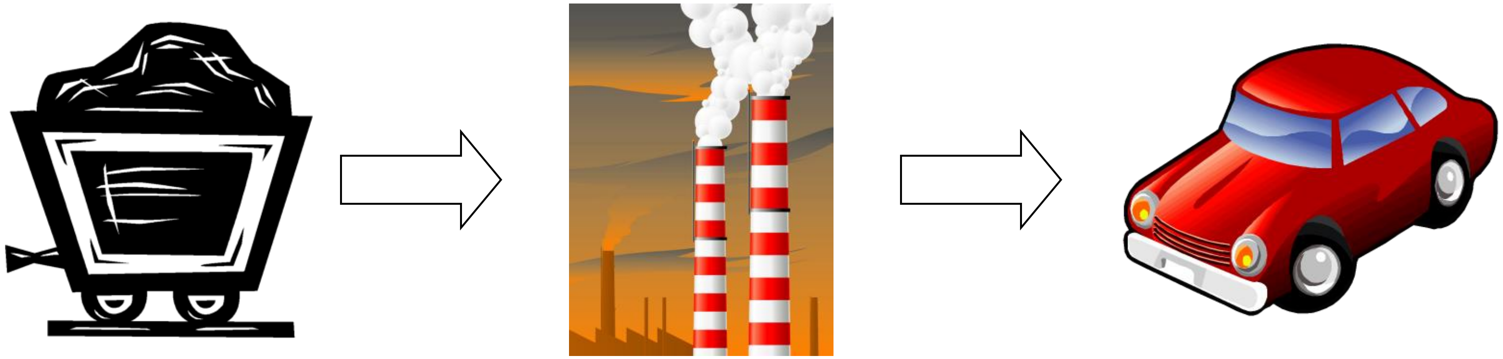
Le réseau Chappe en France

By Jeunamateur [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

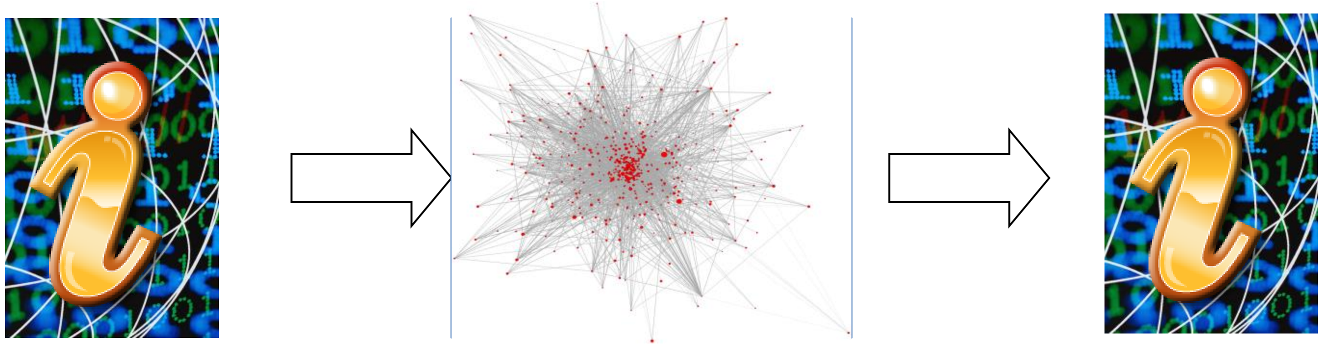
“We create as much information in two days now as we did from the dawn of man through 2003.”

-Eric Schmidt, Former CEO of Google

Information Economy

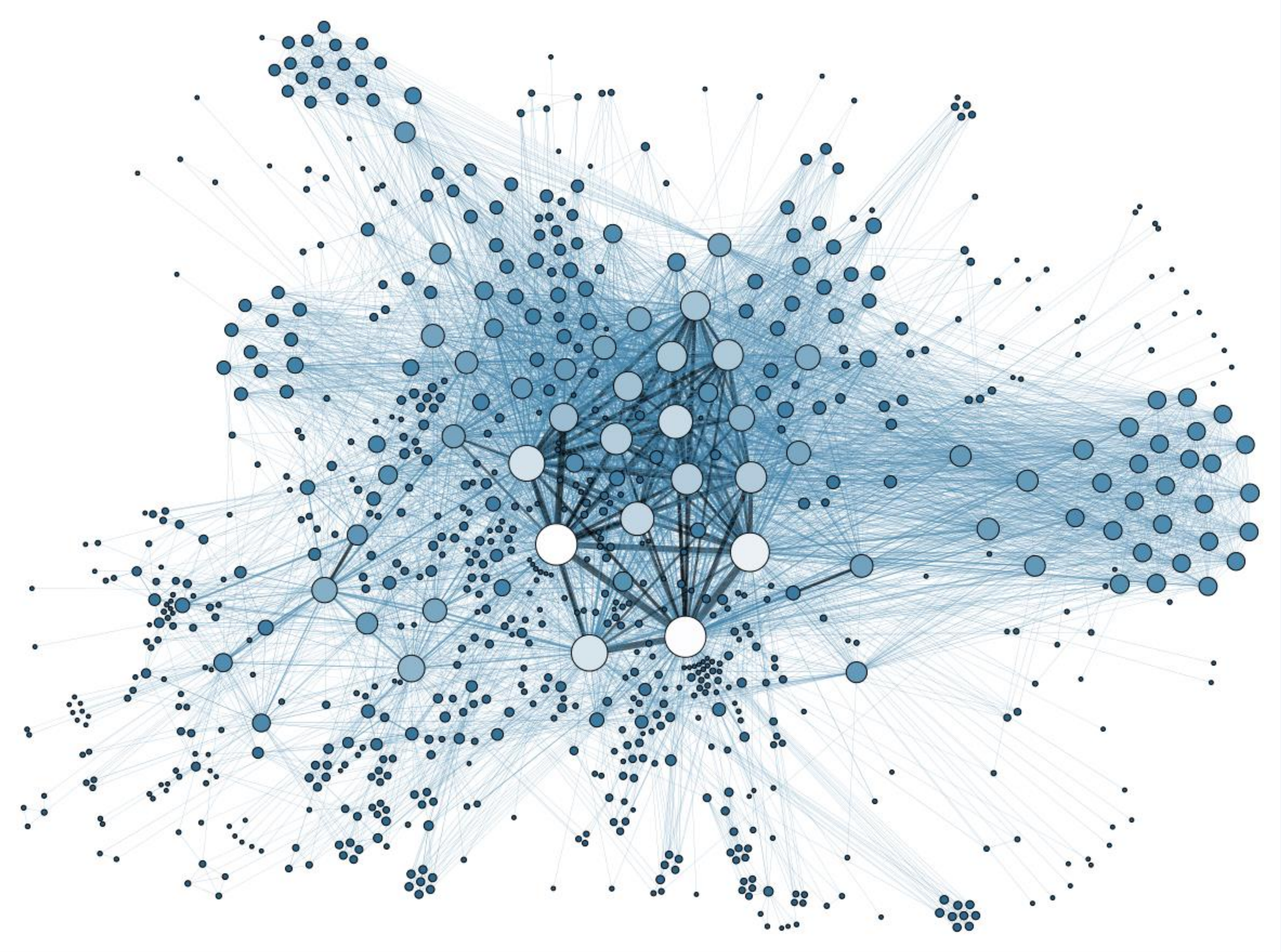


INFORMATION-BASED BUSINESS PROCESS



INFORMATION TECHNOLOGY

Why am I excited about Data Science?



Data, Analytics, and AI are Changing the World

Martin Grandjean - Own work : http://www.martingrandjean.ch/wp-content/uploads/2013/10/Graphe3.png

•CC BY-SA 3.0

“Analytics is the discovery and communication of meaningful patterns in data.”

-Wikipedia

More data. More analytics.

The Internet, the Original Big Data Problem

“PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.”

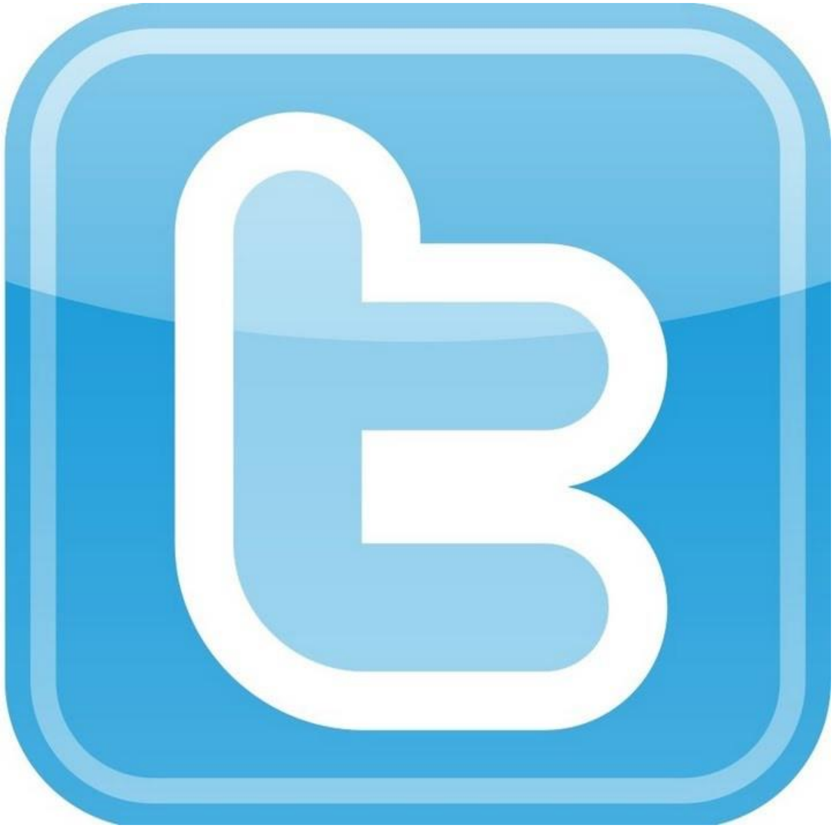
- From "Facts about Google and Competition” via Wikipedia [https://en.wikipedia.org/wiki/PageRank].

Internet of Things

“The internet of things (IoT) is the network of physical devices, vehicles, buildings and other items—embedded with electronics, software, sensors, actuators, and network connectivity that enable these objects to collect and exchange data.”

– Internet of Things Global Standards Initiative via Wikipedia.

Web 2.0 Social Networks



Real-world Examples

Disney

ROLE OF DATA: How many tickets did we sell?

Disney – Data Warehouse Stage

ROLE OF DATA: How much did our customers spend? How can we understand different customer types?

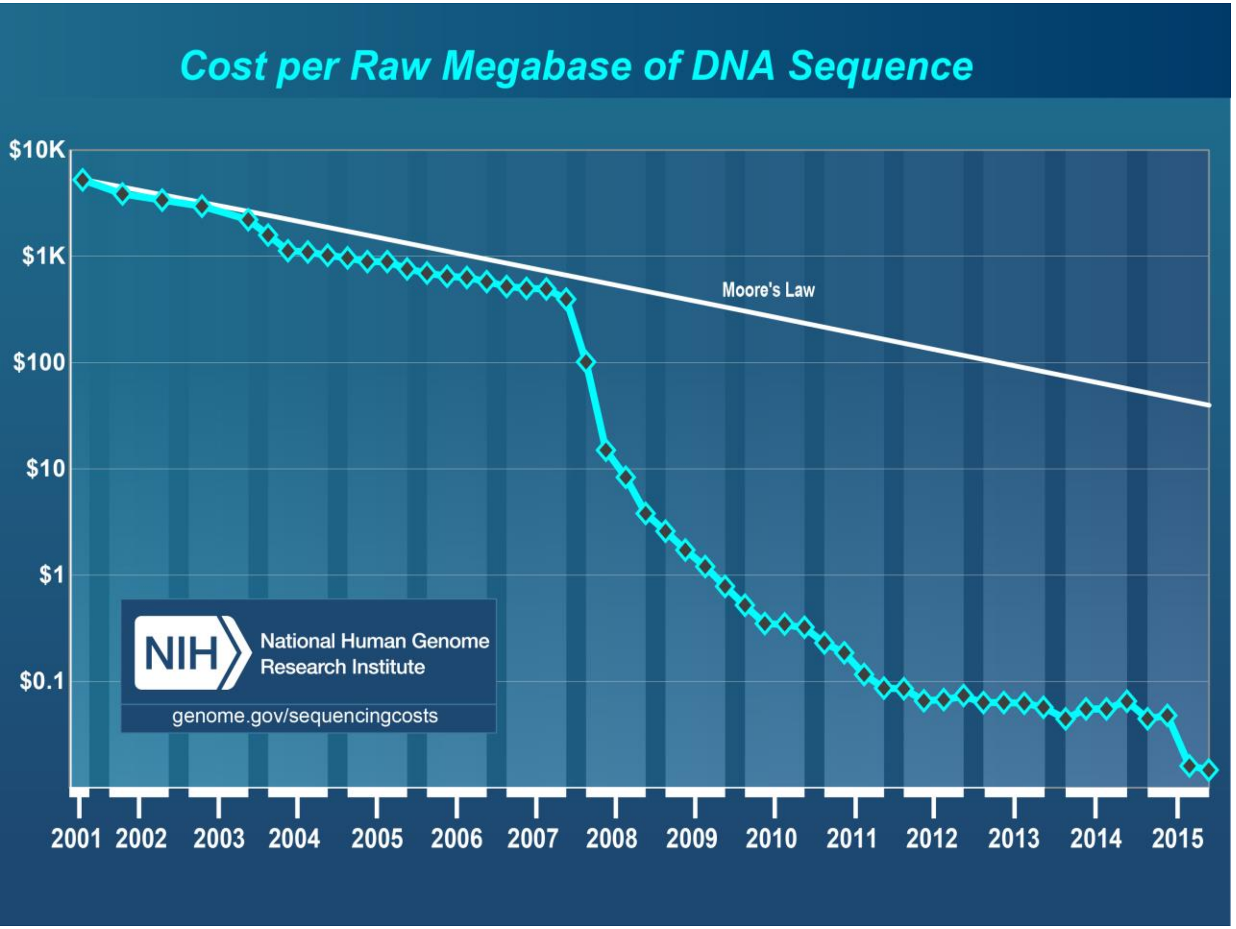
YOUR KEY TO THE WORLD (407) 934-3344

09/06/10 To 09/10/10 ADULT
TRICIA BALLAD
079007100910000078
DLX A02C03
ID#9915 4041 7603 0086 08

Disney – Big Data

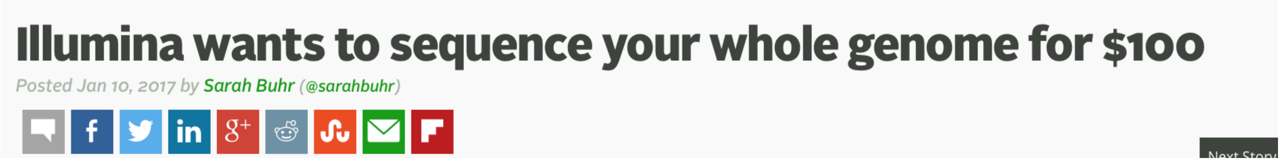
ROLE OF DATA: What path did customers take through the park, when did they leave? How long did they stand in line? When did they spend money on souvenirs and where? How often did they go to the bathroom and did they have to wait? How long did they spend at dinner in the Mexican pavilion compared with the German pavilion? How does the speed of entry correlate with tipping behavior?

Big Data and Bioinformatics



https://www.genome.gov/images/content/costpermb2015_4.jpg

Tremendous drop in cost of sequencing DNA



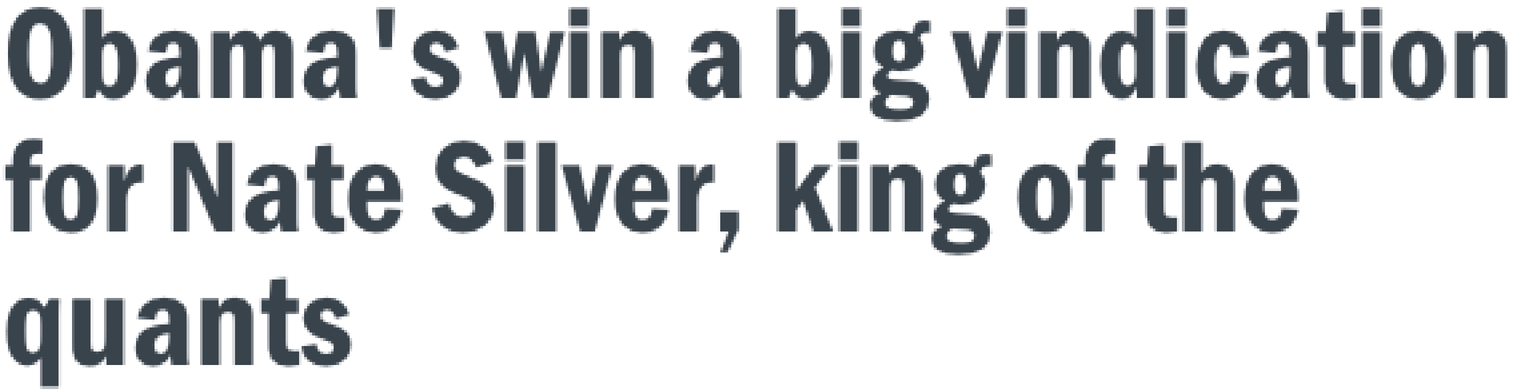
Illumina wants to sequence your whole genome for $100

Posted Jan 10, 2017 by Sarah Buhr (@sarahbuhr)

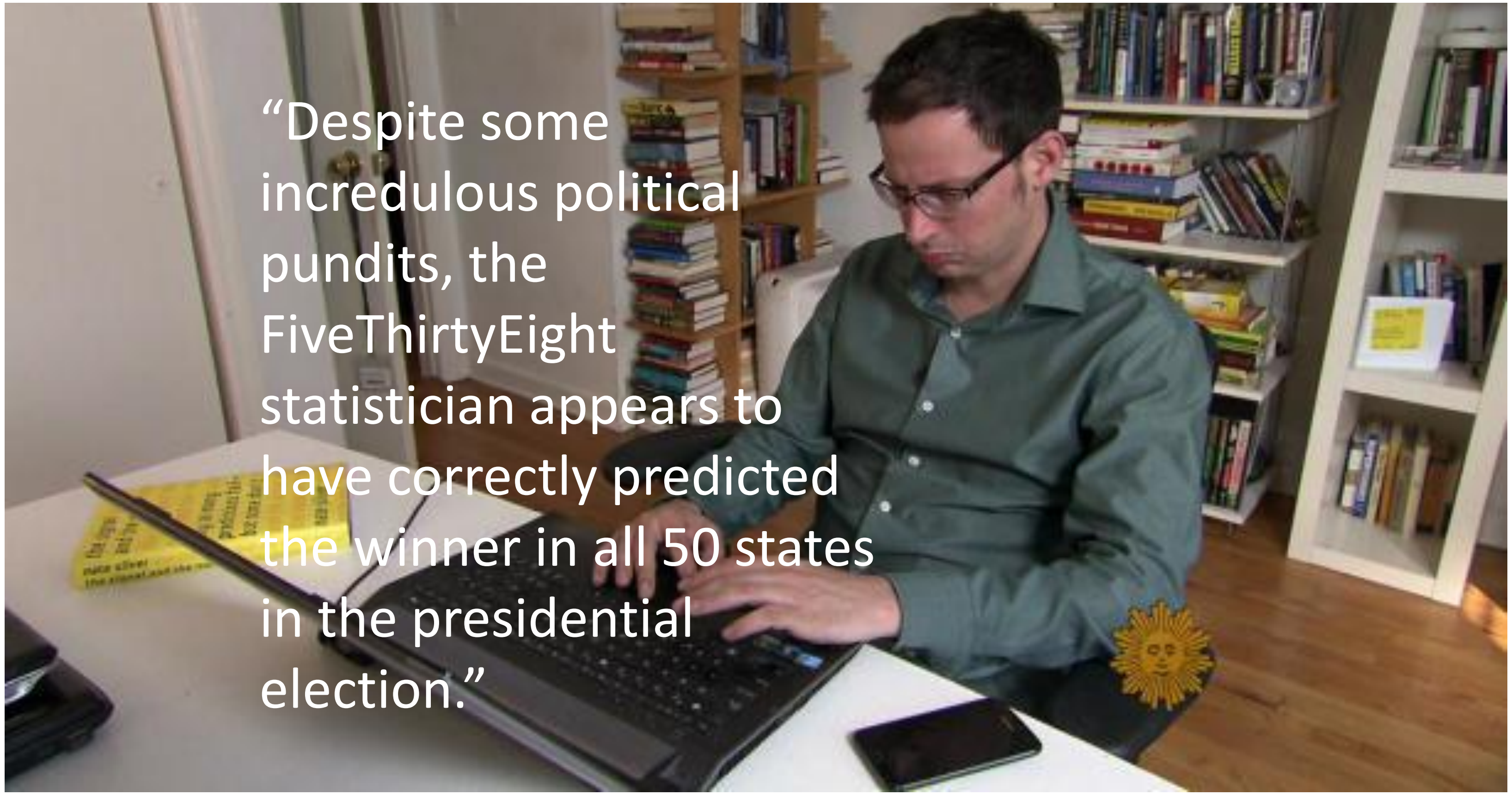
Big Data and Astronomy

"To store the Big Data the MWA produces, you'd need almost three 1 TB hard drives every two hours."

[Source: http://www.skatelescope.org/news/pawsey-centre/](http://www.skatelescope.org/news/pawsey-centre/)

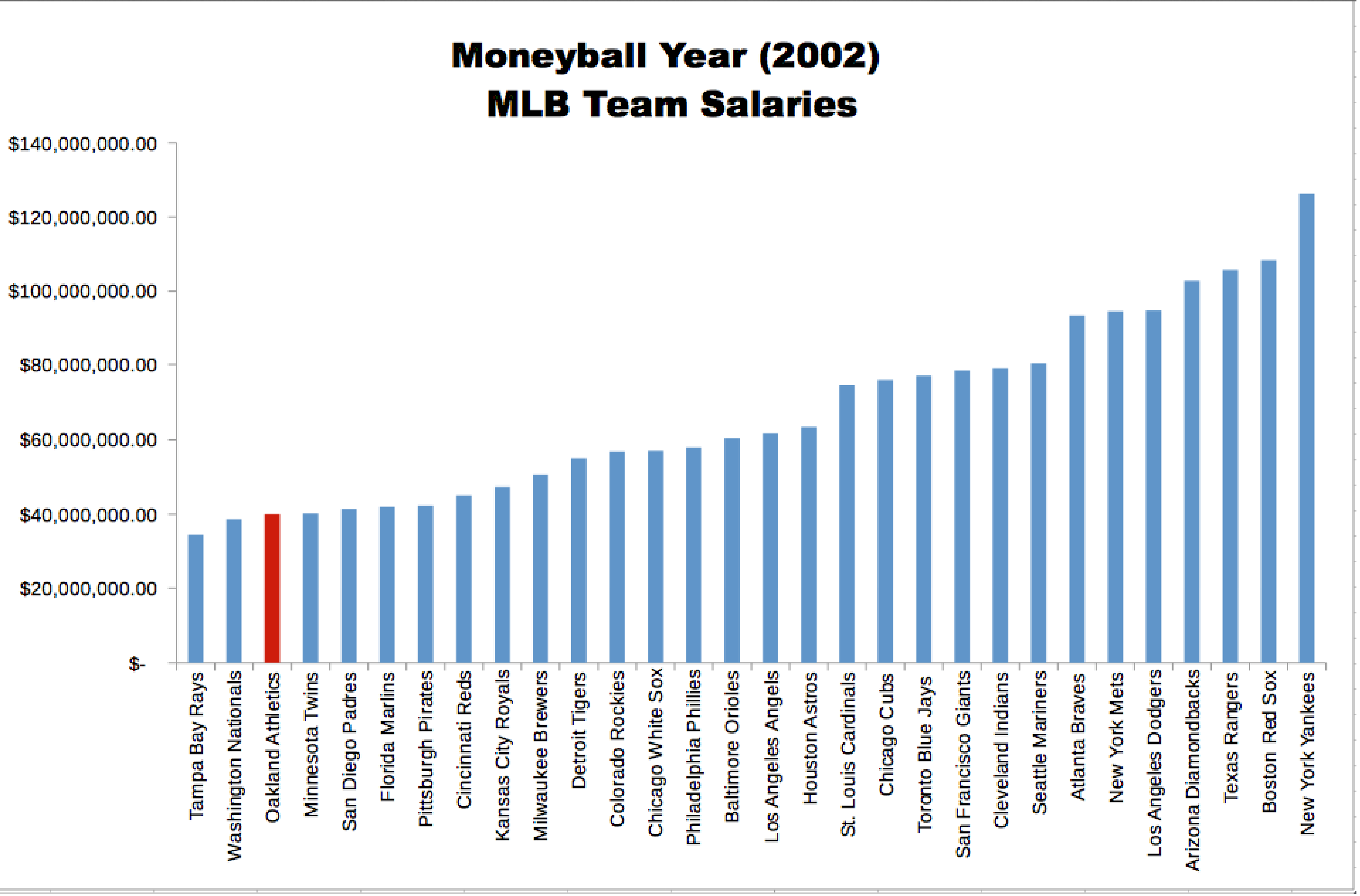


Obama's win a big vindication for Nate Silver, king of the quants



“Despite some incredulous political pundits, the FiveThirtyEight statistician appears to have correctly predicted the winner in all 50 states in the presidential election.”

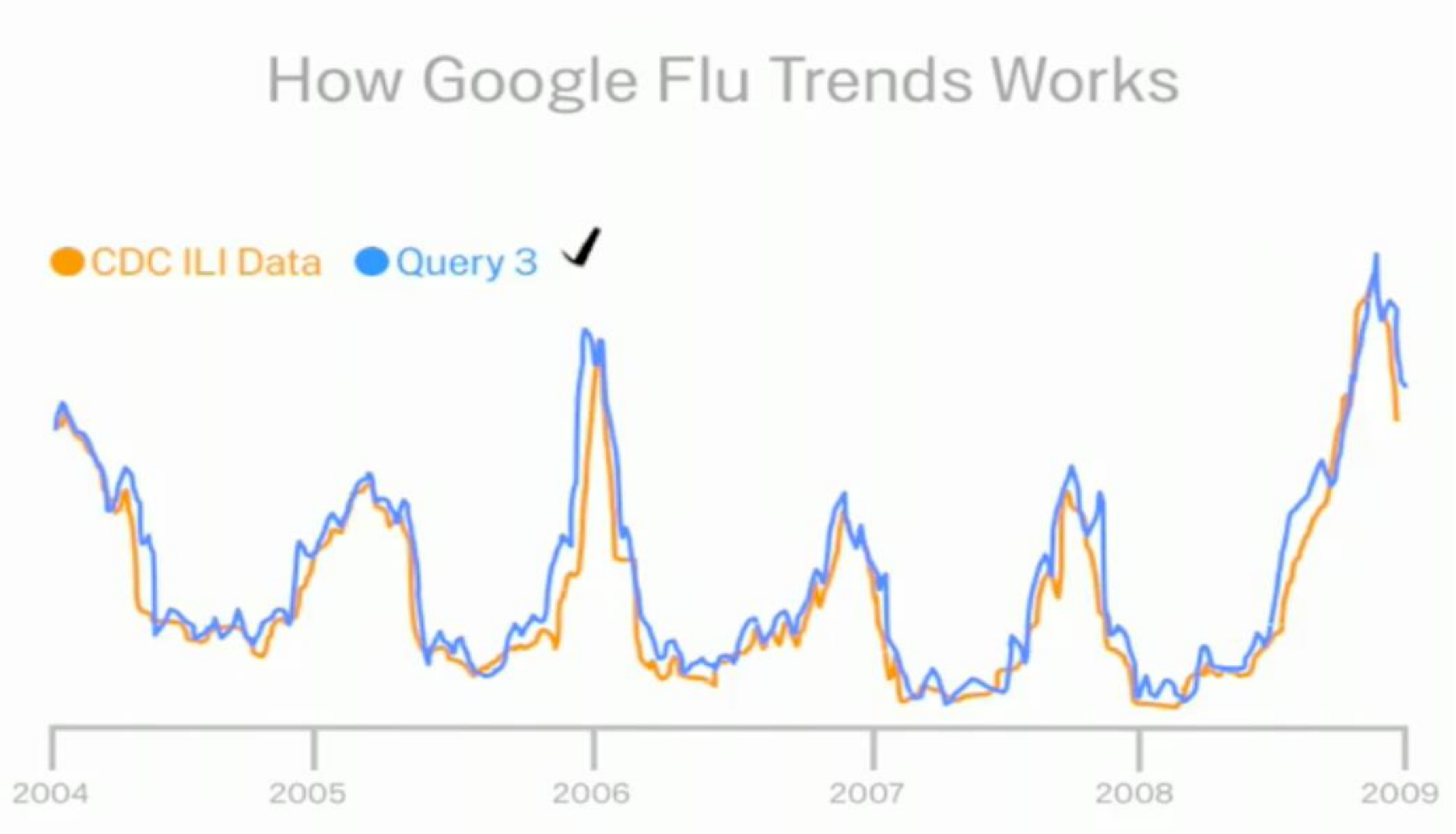
Source: http://news.cnet.com/8301-13510_3-57546161-21/obamas-win-a-big-vindication-for-nate-silver-king-of-the-quants/



http://www.youtube.com/watch?v=AiAHlZVgXjk

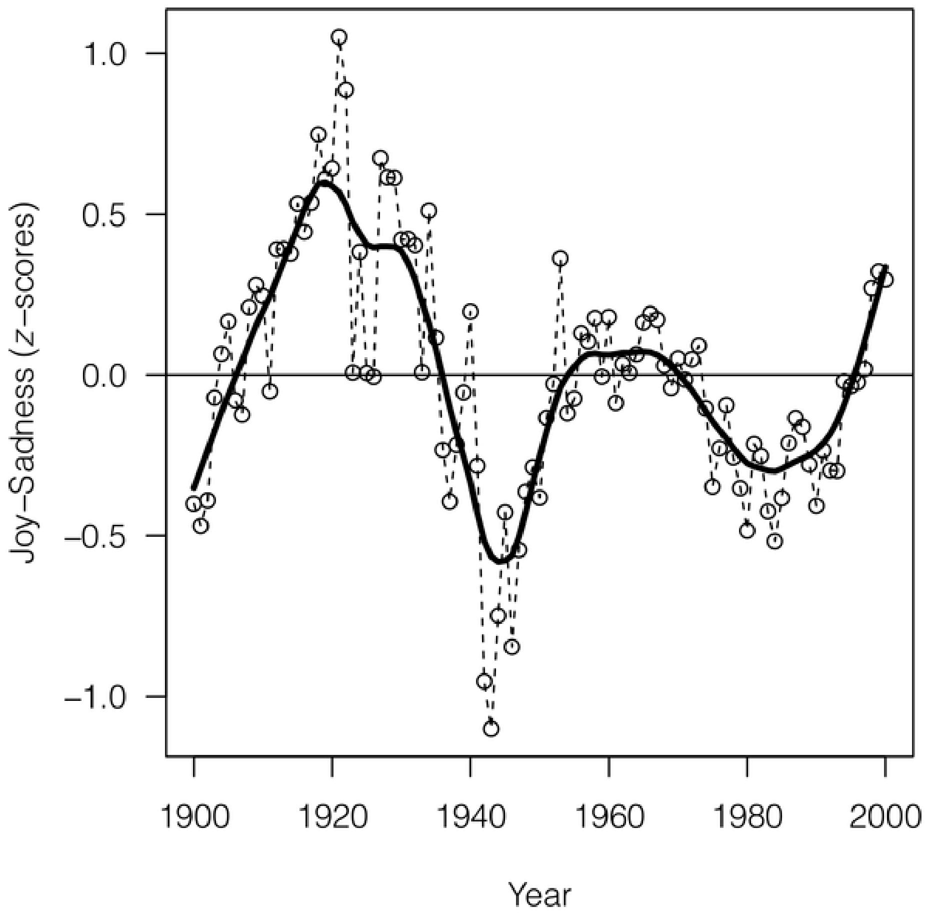
Darryl Leewood (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Google Flu Trends



http://www.google.org/flutrends/about/how.html

The Expression of Emotions in 20th Century Books

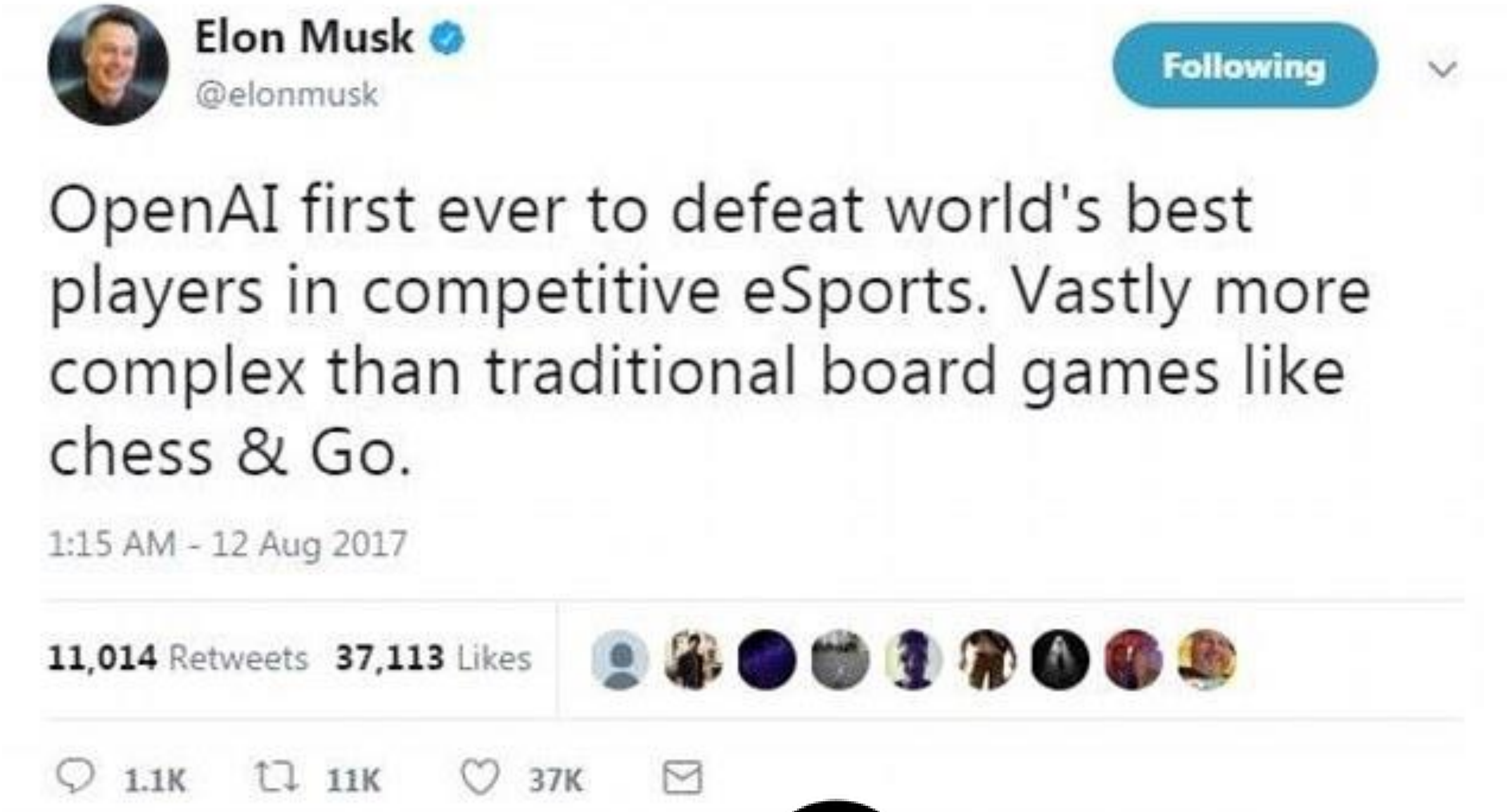


"using the data set provided by Google that includes word frequencies in roughly 4% of all books published up to the year 2008. We find evidence for distinct historical periods of positive and negative moods"

Source:
http://www.plosone.org/article/info:doi/10.1371/journal.pone.0059030

“AlphaGo is almost like the God of Go”

AlphaGo’s move 37 described as "So beautiful.”



Elon Musk
@elonmusk

OpenAI first ever to defeat world's best players in competitive eSports. Vastly more complex than traditional board games like chess & Go.

1:15 AM - 12 Aug 2017

11,014 Retweets 37,113 Likes

1.1K 11K 37K

© Twitter / @elonmusk

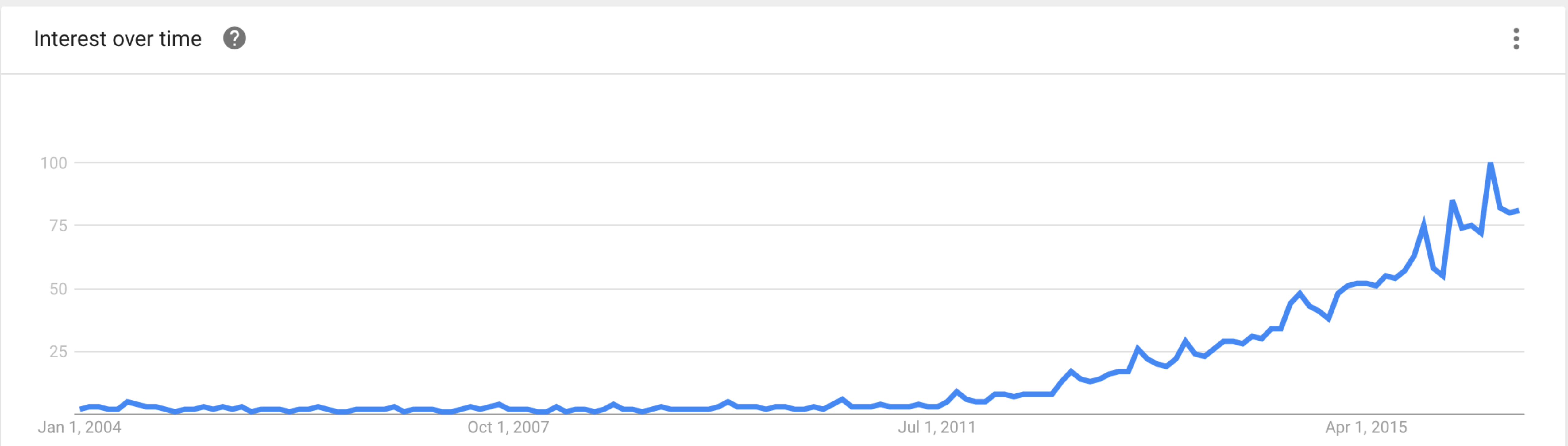
OpenAI

For example,

- Robots -- https://www.youtube.com/watch?v=uhND7Mvp3f4
- Games -- https://www.youtube.com/watch?v=8tq1C8spV_g
- Shopping -- https://www.youtube.com/watch?v=ZZ0qBLOqqyo
- ….

What does it mean to be a data scientist today?

What is a “Data Scientist”?



By Special Collections, University of Houston Libraries [CC0], via Wikimedia Commons

The data scientist has been described as the sexiest job of the 21st century, and people with the broad range of skills to truly be a data scientist have been called unicorns.

“There’s a joke running around on Twitter that the definition of a data scientist is ‘a data analyst who lives in California.’”

— Malcolm Chisholm

http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

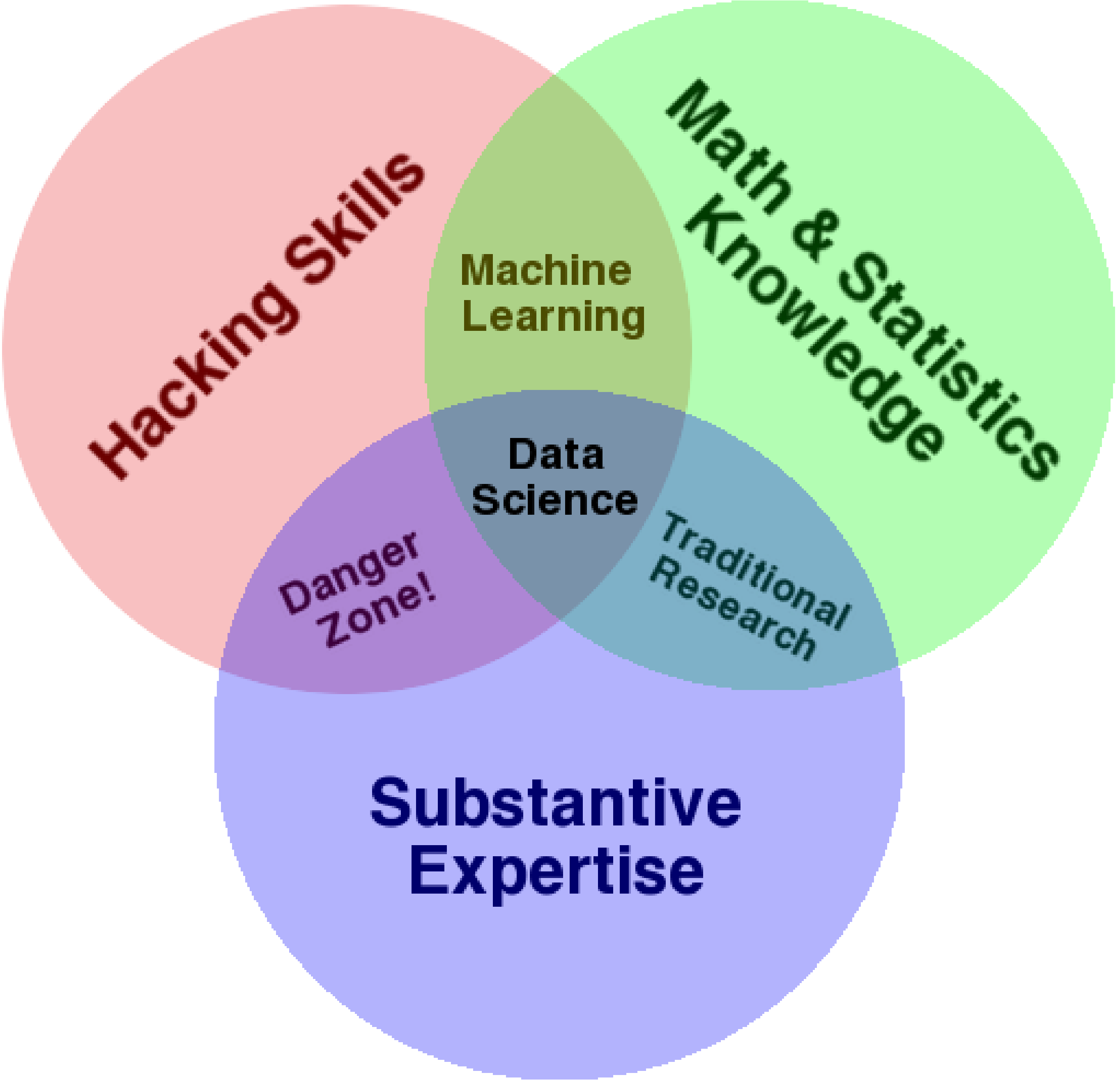
Data scientists are “analytically-minded, statistically and mathematically sophisticated data engineers who can infer insights into business and other complex systems out of large quantities of data.”

— Steve Hillion

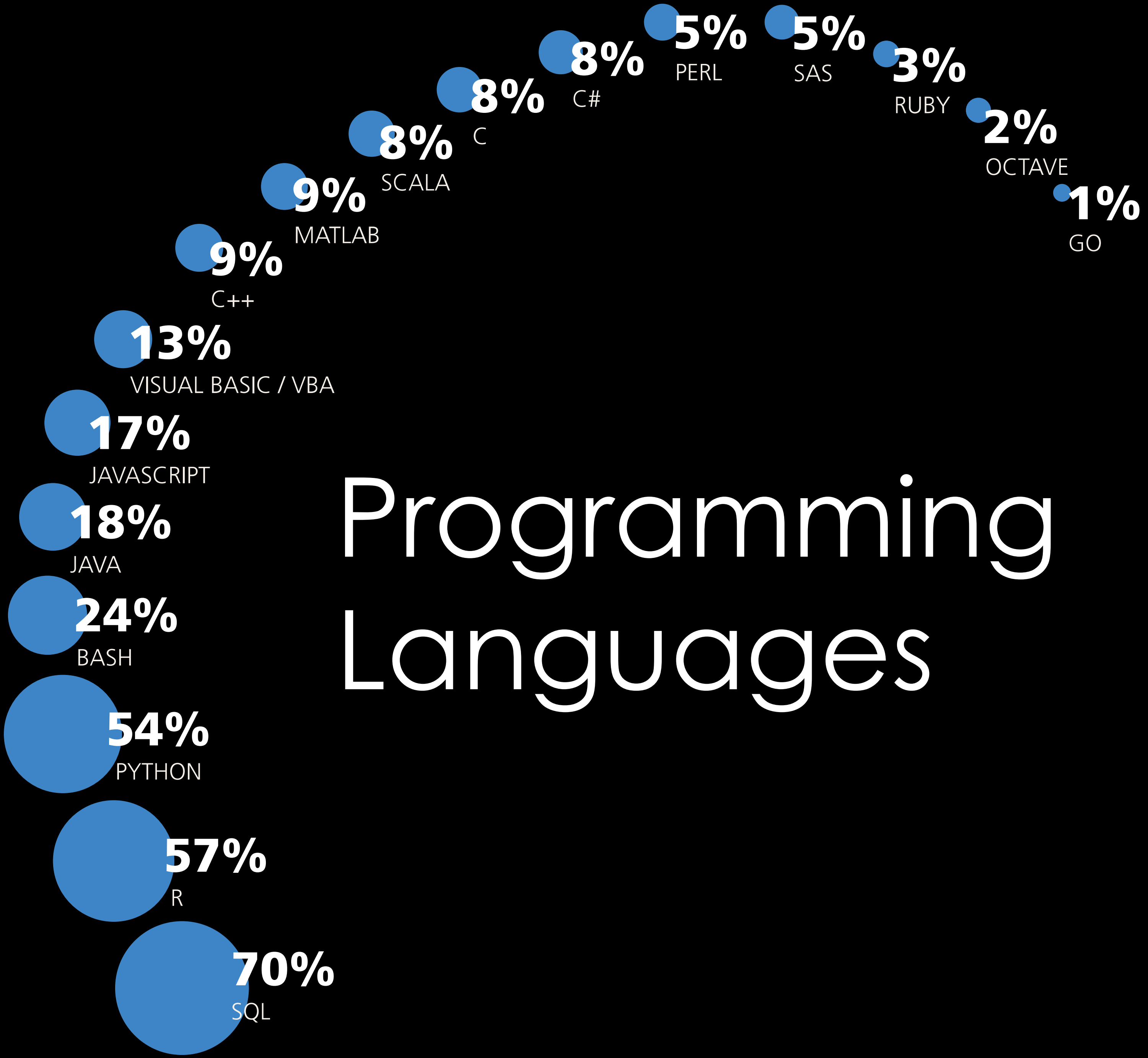
http://bigdata-madesimple.com/what-is-a-data-scientist-14-definitions-of-a-data-scientist/

Data Science

- Hacking Skills
- Math & Statistics Knowledge
- Substantive Expertise
- Ability to Learn



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



Programming Languages

SQL > R > Python

Cluster Analysis:

- ➢ Python > R: *data scientists*
- ➢ R > Python: *analysts*

Python users had higher salaries.

Highest Paid?

- ➢ Scala

Credit: DS100.org Class1 Presentation

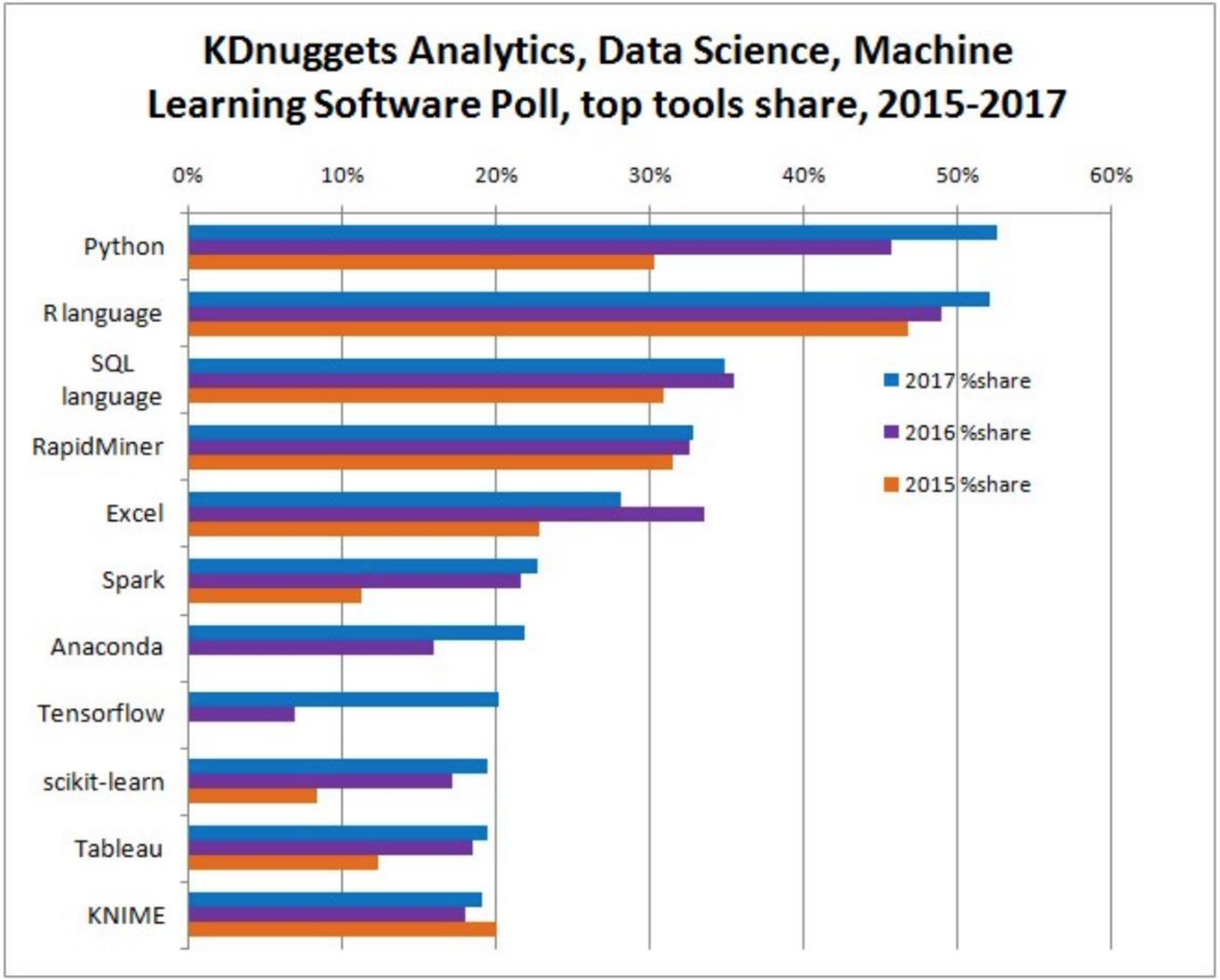


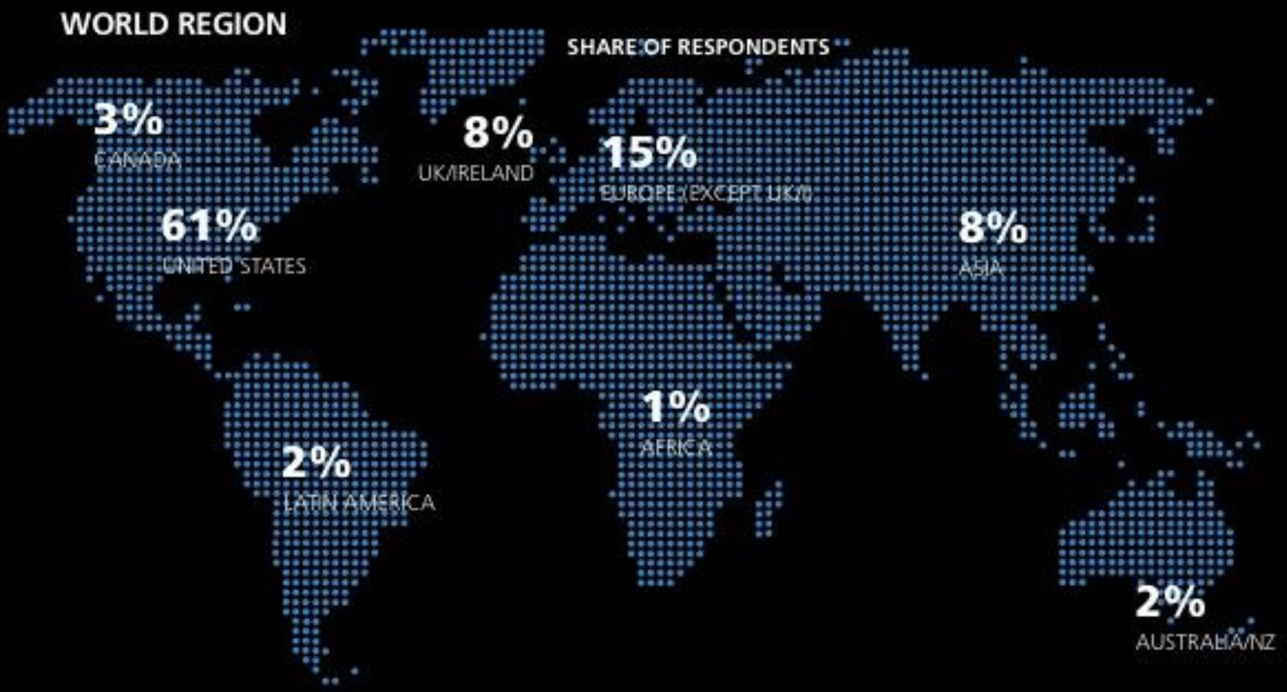
Fig 1: KDnuggets Analytics/Data Science 2017 Software Poll: top tools in 2017, and their usage

Source: http://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html

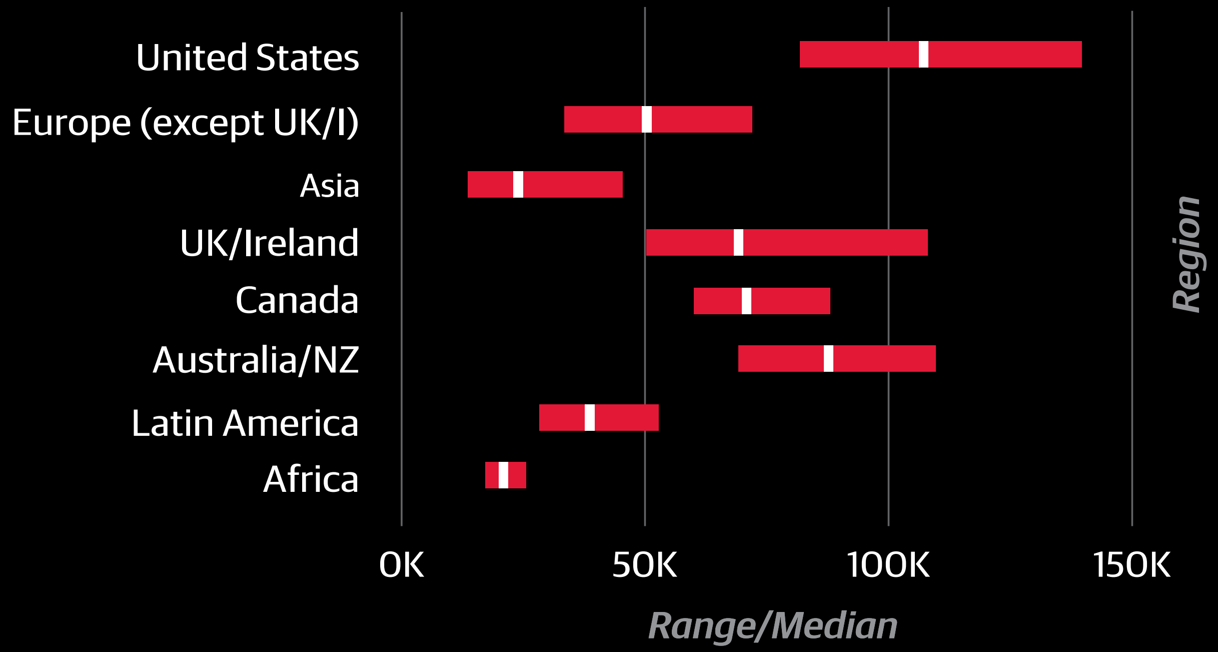
2017 Survey, Python is overtaking R.

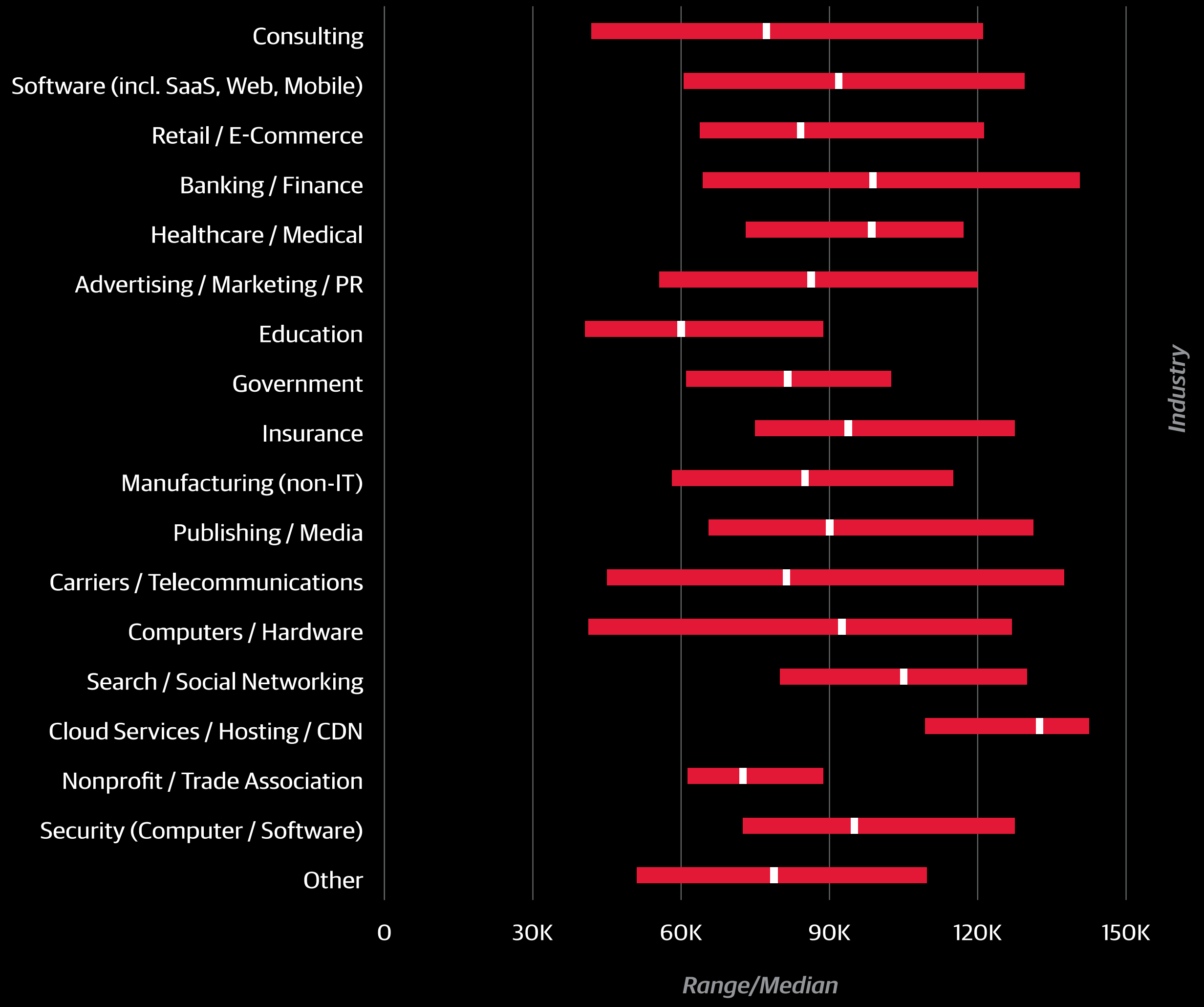
Python is "glue" that holds machine learning ecosystems together.

How much are they paid?



SALARY MEDIAN AND IQRC (US DOLLARS)





Salary Depends on Location and Industry

Elements that make this field exciting!

Helps answer questions such as:

1. Is this or that? A/B Testing
   - If we propose new discounts, will it be better for our business?
2. Clustering patterns
   - What are the buying habits, age, location, etc., patterns of my customers?
3. Predictions
   - Will a customer purchase the product again within a few months of time?
4. Anomalies
   - All of a sudden there was a shift in the purchase pattern behavior of a customer
5. Hypothesis testing
   - Is there a relationship between the purchase patterns of two customers buying similar products?

What will we cover in course?

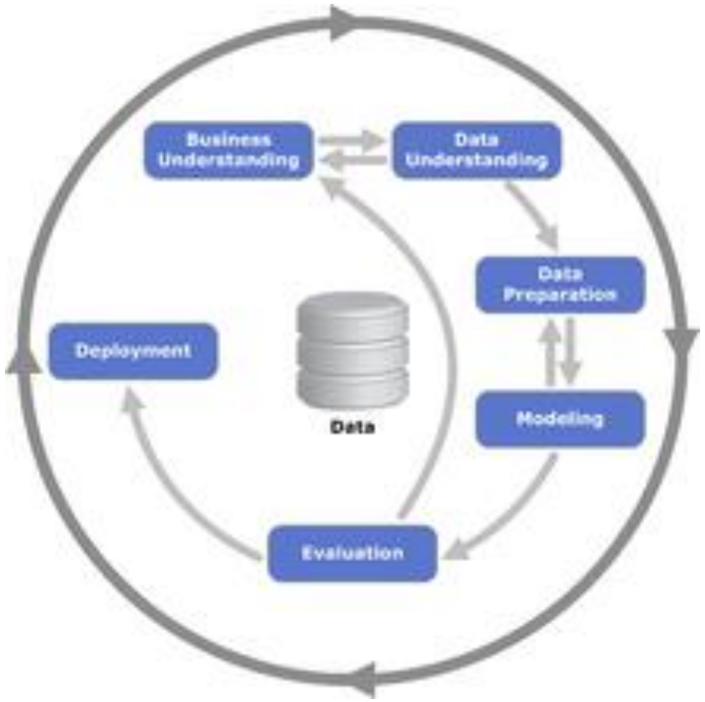
Class Goals

- **Prepare** for advanced courses in analytics from across the RPI campus.
- **Enable** you to gain skills necessary to begin careers as data scientists.
- **Empower** you to apply analytics to solve real world problems.

A few topics that we will cover in the class

- Python (basics, conditionals, loops, functions, etc)
- Exploratory data analysis (Data Visualization, creation, manipulation)
- Regression
- Unsupervised learning (Clustering)
- Supervised learning (Decision trees, Random forests, KNN, SVM, NeuralNets)
- Deep Learning (CNNs, RNNs)
- …

Class Overview



{JSON}

CSV

Spark

API

DATA MUNGING

Retrieve-Filter-Missing Data-Data Cleaning-Aggregate-Merge-Missing Values-Feature Creation-Text Tools (Lemmatization-Stemming-Corpus-Bag of Words-TFIDF))-Sampling-K-Fold Cross Validation

DATA FUNDAMENTALS

Variable-Vector- Matrix-Dataframe-CSV-JSON-For Loop-if/else- Function

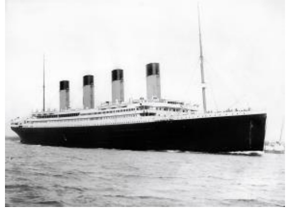
Modeling

R

python

GitHub

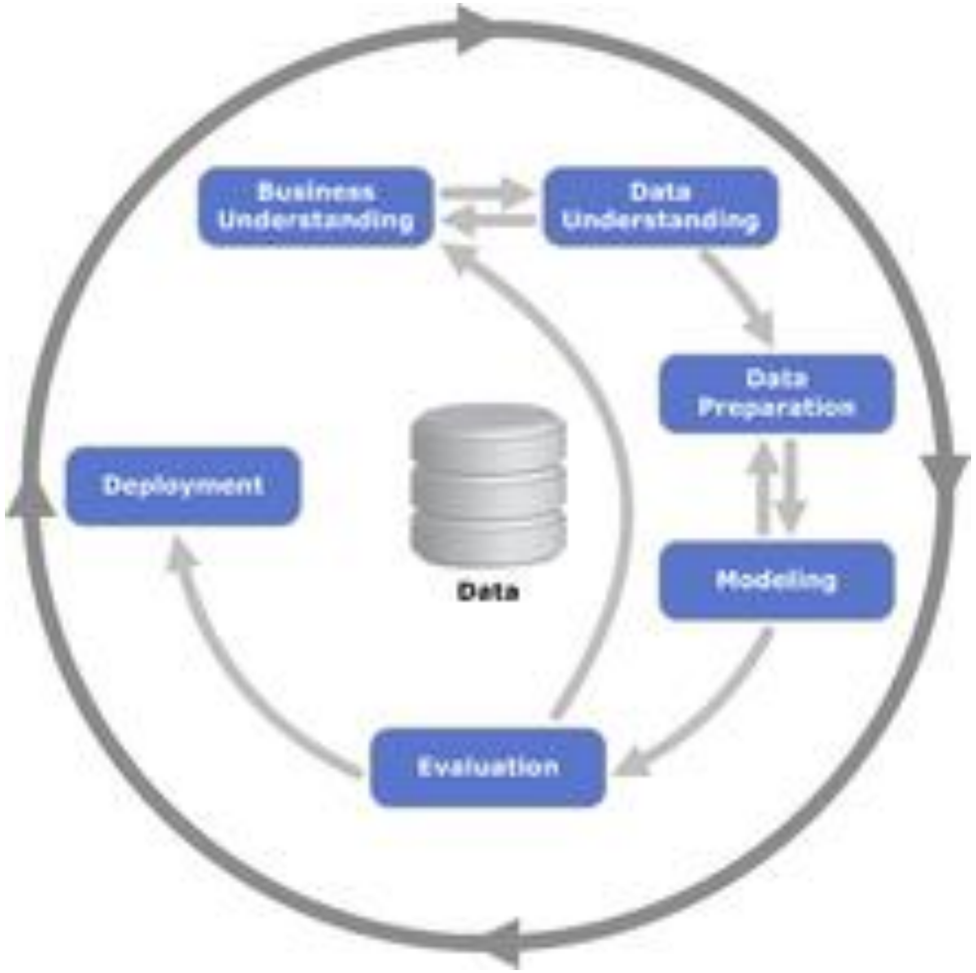
kaggle



Yummly

Basic Data Science Principles

- Defining the problem
- Data structures
- Missing data
- Exploratory data analysis
- Modelling
- Evaluation



Computer Science Principles in Data Science

- Software development and version control
- Relational data models
- Issues surrounding parallelism and big data

Statistics (&CS) Principles in Data Science

- Inference & prediction in modeling
- Model design and cross validation
- Feature extraction
- Processing of image and text data
- Issues surrounding parallelism and big data

Real Analytics on Real Data

Real data has issues

- Need to gain experience with issues like missing data, highly nested data
- 80% of the work a data scientist does is collecting, cleaning and organizing data

Real Tools for Analytics

- Both Python and R in Jupyter Notebooks
- Important packages (Pandas, Numpy, Seaborn)

Things will break:

  - Learn how to troubleshoot code
  - Get help through Piazza

It takes time to get good at data science.

1. Understand why you are doing something.
2. Read the error message.
3. Google the error message.
4. Consider other methods.
5. Ask for help.

Grading

Component	Weight
Assignments & Quizzes	15%
Research Translation Exercise	5%
Project*	25%
Midterm	25%
Final Exam	30%

Participation in the class matters!!

On Time Policy

- 5 days of “late” time for homework for sickness/deadline conflicts
- 20% per day for each late day
- Please let me know if having problems.

Collaboration Policy

- It is OK to work in the same location as someone and ask questions. It is not OK to share code.
- You should produce everything that is submitted.

Communications & Homework Submissions

- Communication (Announcements etc.) will be done through piazza.
    - Any questions related to the class should be posted on piazza.
    - Email communication with the professor/TA is only for questions that are case-based.
    - Participation in the class could help in curving the grades at the end of the semester.
        - For example, 2 students received same number of cumulative points at the end, participation is considered to break the tie
- Homework submission using blackboard.

Computing Environment

- Expected to eventually work on your own laptop environment.
- BETA – Google Colab provides a computing environment for Python which is robust and free.

Quizzes

- Surprise quizzes through the semester
    - To incentivize you to review the readings prior to class
- Please arrive on time
    - You will receive a 0 if you arrive late

Exams and Project

- Midterm exam
- Final exam

- Project – initial presentation, 2 main reports and a final presentation

Special Announcement – AAAI Workshop

- Scientific Discovery with AI

https://sites.google.com/view/ai4be/

When: **February 9, 2021** 9 am to 4 pm ET