Technology Fundamentals of Business Analytics

Jason Kuruzovich

Agenda

- Intro/Classroom Rules
- Information and society
- Let’s get excited about analytics
- What do we mean by being a data scientist?
- Syllabus/Course Website
- Analytics Overview
  - Data Munging, Visualization, Statistics, and Machine Learning
- Technology Platforms
  - Docker
  - Azure
  - Git

Me

- Director of the Severino Center for Technological Entrepreneurship
- Associate Professor of Business Analytics
- Research on marketing, multichannel retailing, most recently entrepreneurship

Next Startup Tech Valley Event is September 7 at 5:30 at Brown's Revolution Hall
See more at
www.startuptechvalley.org

Classroom Rules

- Laptops/phones down when lecture/discussion
- Laptops up for hands on analytics exercises

DON'T WORRY, WE WILL GET A GOOD MIX OF EACH!

See what influenced me:
https://medium.com/@cshirky/why-i-just-asked-my-students-to-put-their-laptops-away-7f5f7c50f368#.h6b8fie71

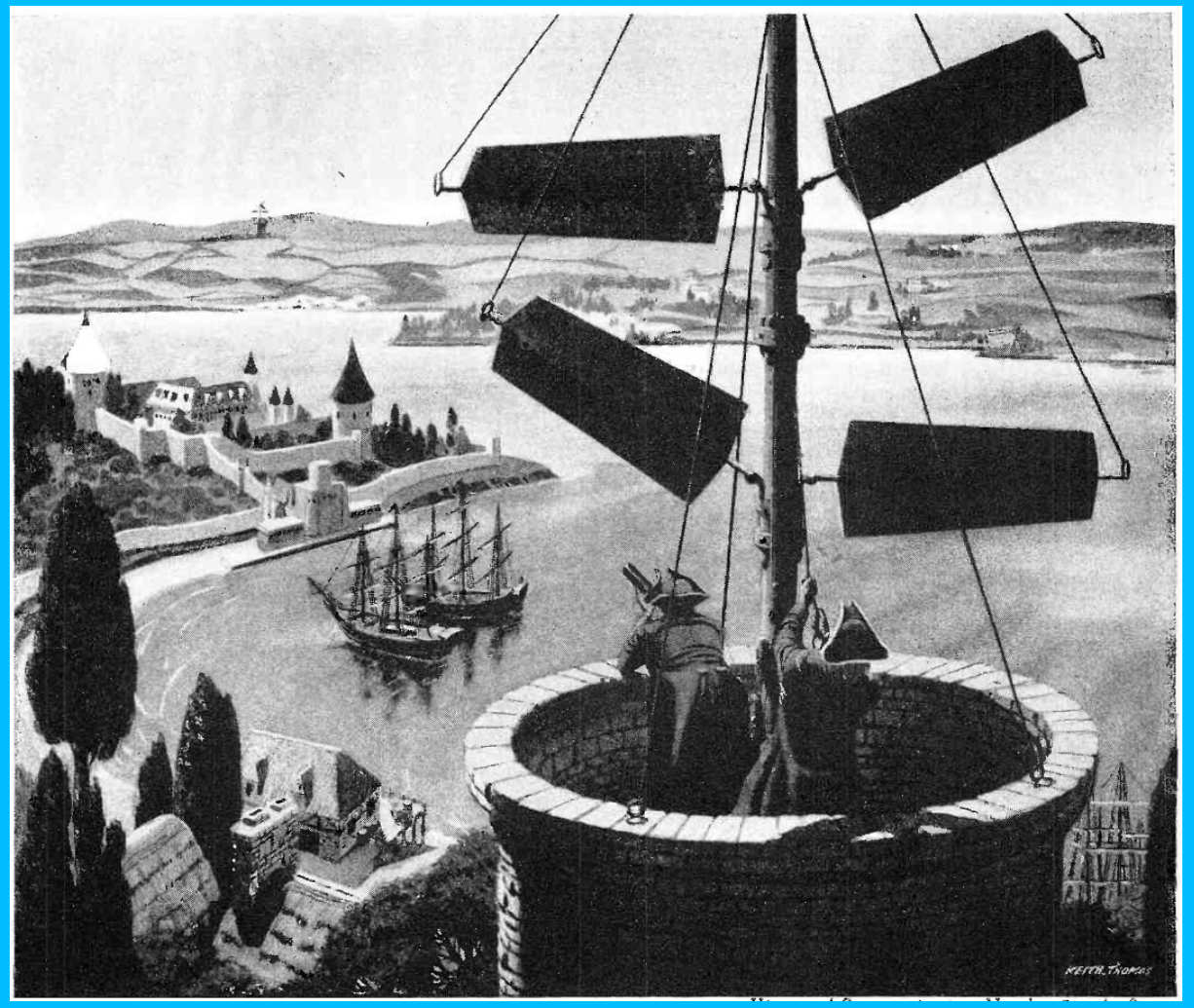
Information and society.

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

Internet 0.1 Beta (18th Century)

Semaphore Telegraph

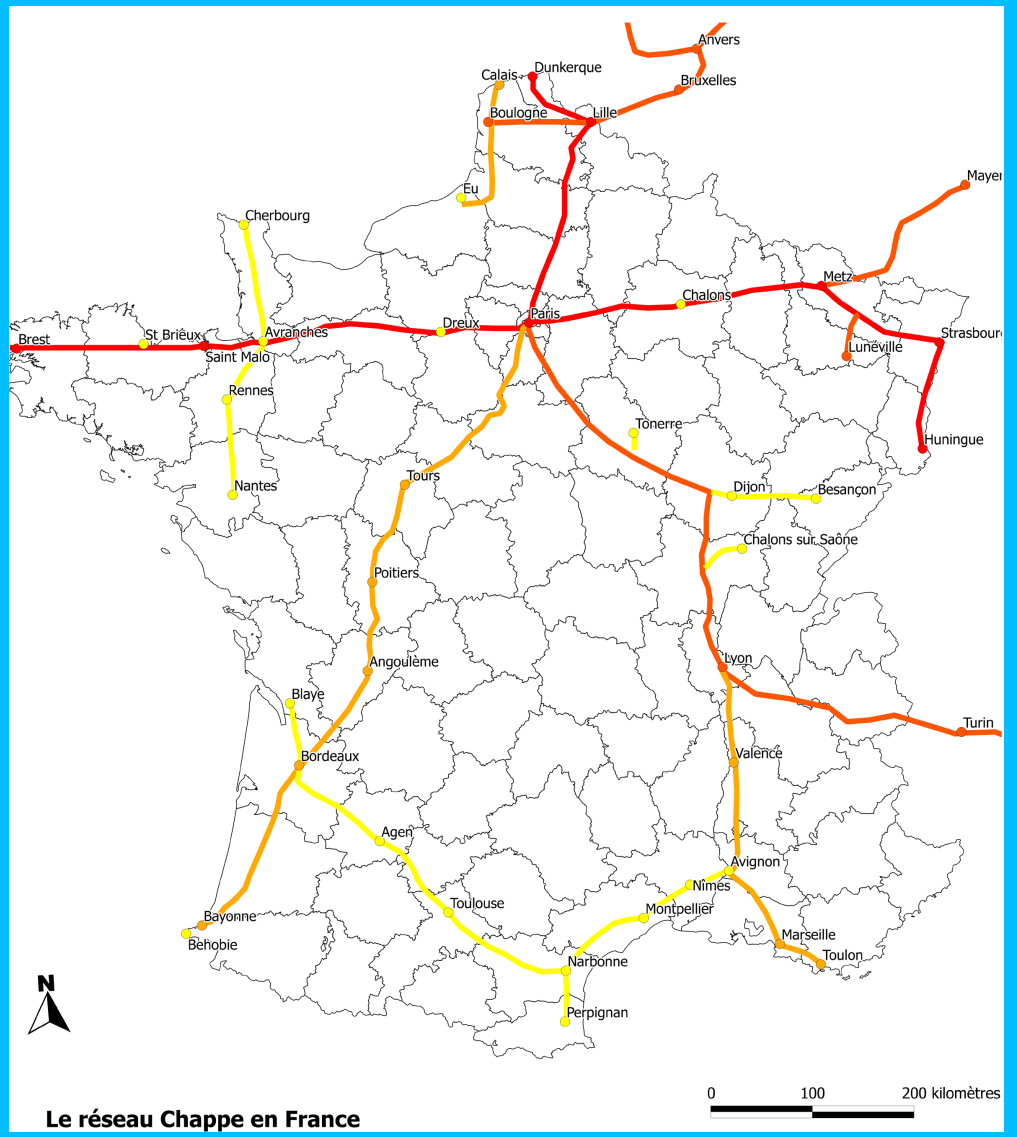
- Visual texting by position of the mechanical elements;



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

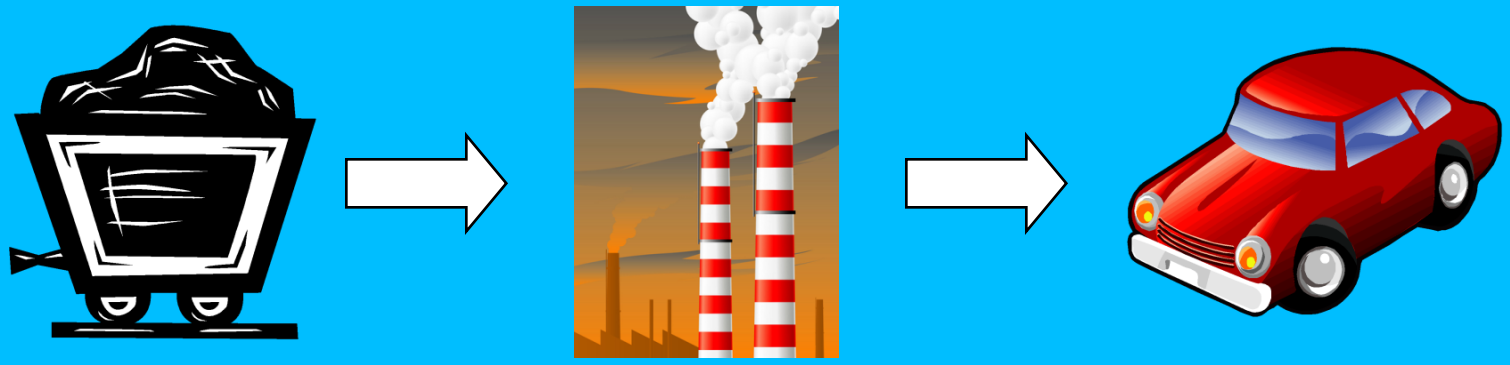


“We create as much information in two days now as we did from the dawn of man through 2003.”

-Eric Schmidt, Former CEO of Google

Information Economy

TRADITIONAL PRODUCTION PROCESS



INFORMATION BASED BUSINESS PROCESS

INFORMATION TECHNOLOGY

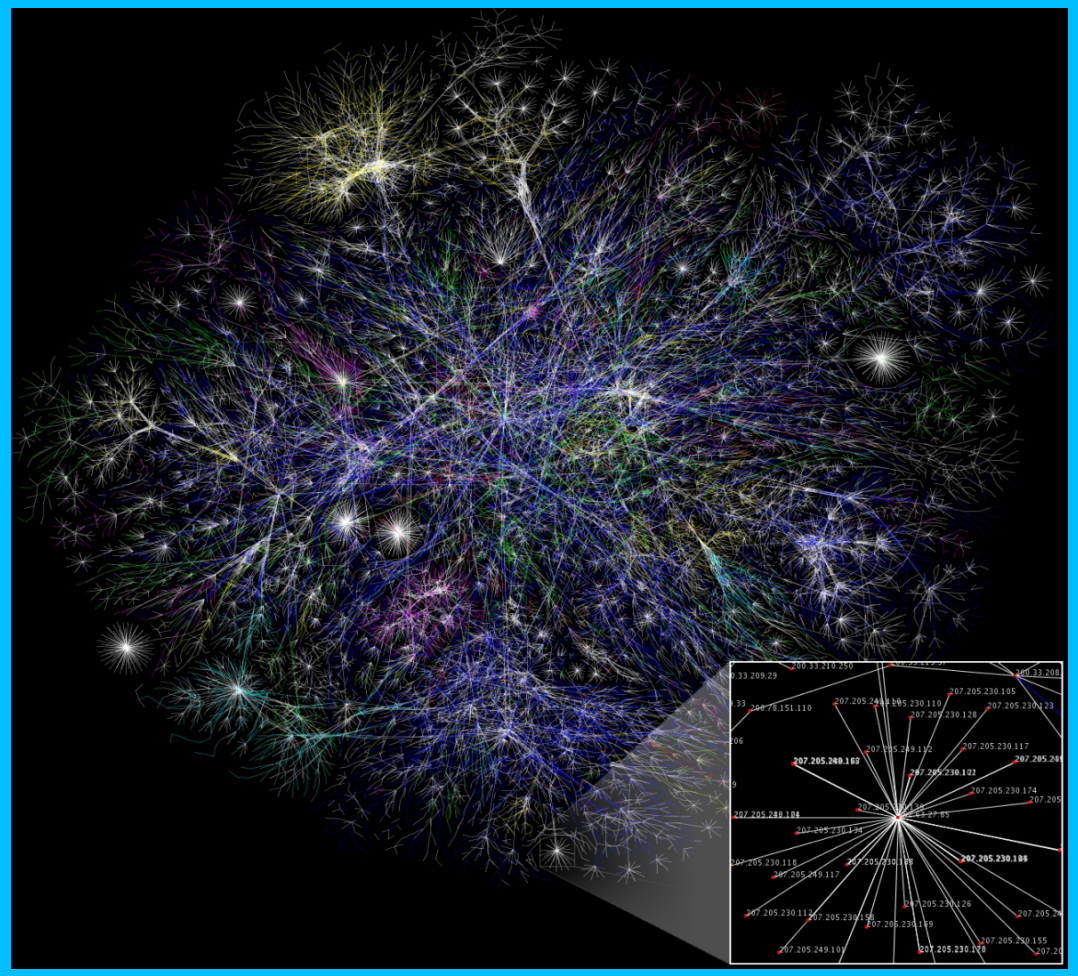
Let’s get excited about analytics.

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

-Wikipedia

More data. More analytics.

Web as Information Source



http://wiki.answers.com/Q/How_large_is_the_Internet

"One estimate puts the internet easily at a exabyte or 1024 petabytes or 1024^2 Terabytes. To think of it easier; if you filled a room that was 8' X 10' X 8' (celling) you could fit about 450 or so hard drives in there. Assuming you used even 2 TB hard drives you would still need over 1000 of those rooms filed to 'download the internet'".

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."

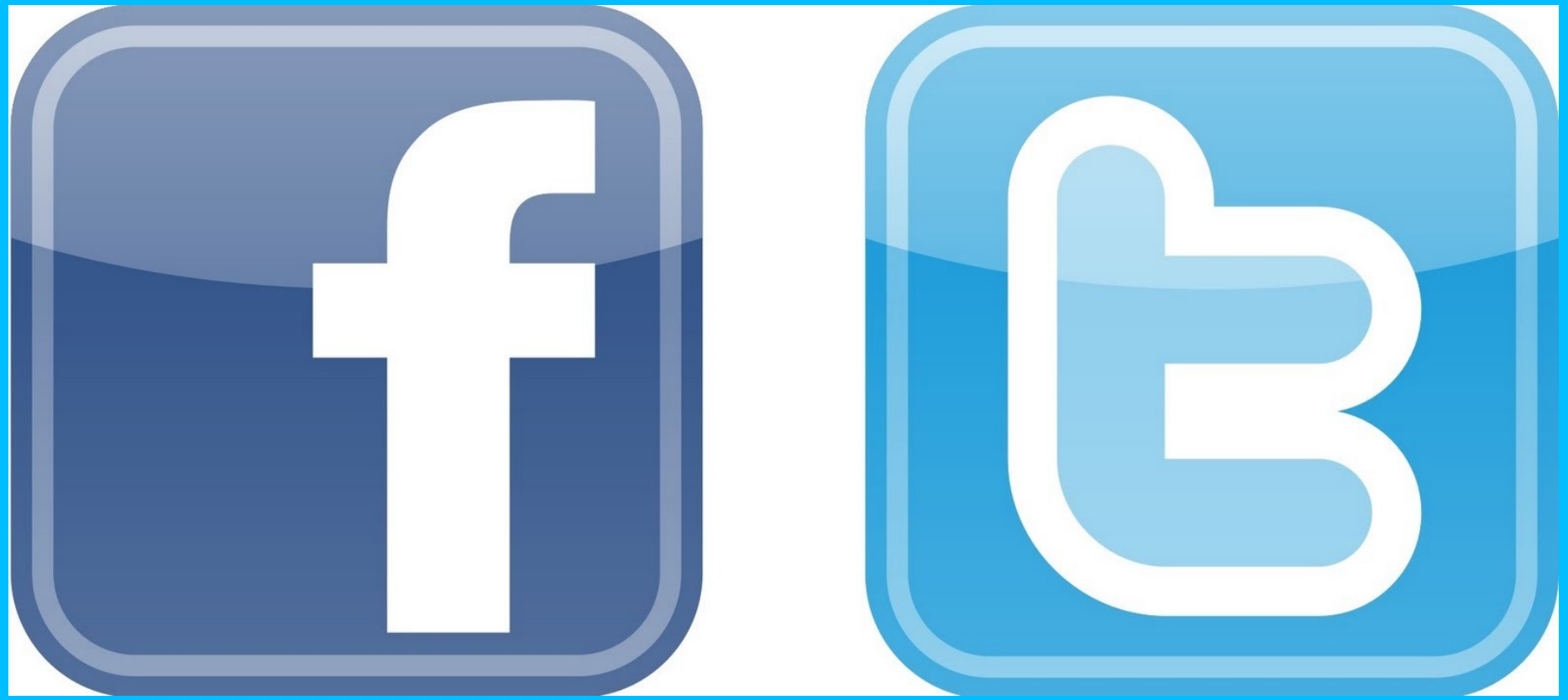
- From "Facts about Google and Competition" via Wikipedia [https://en.wikipedia.org/wiki/PageRank].

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



Disney

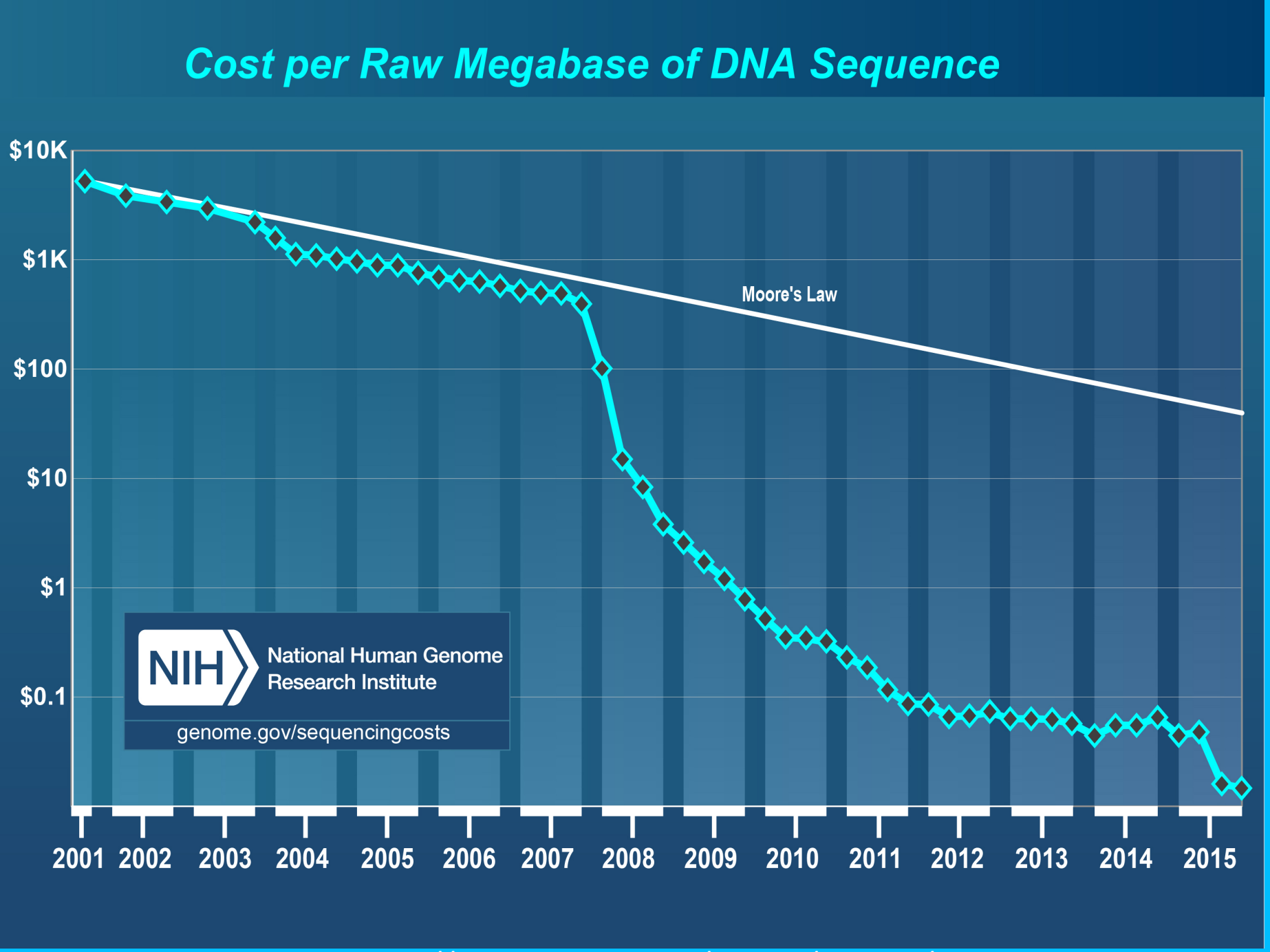
ROLE OF DATA: How many tickets did we sell?

Disney – Data Warehouse Stage

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

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?**



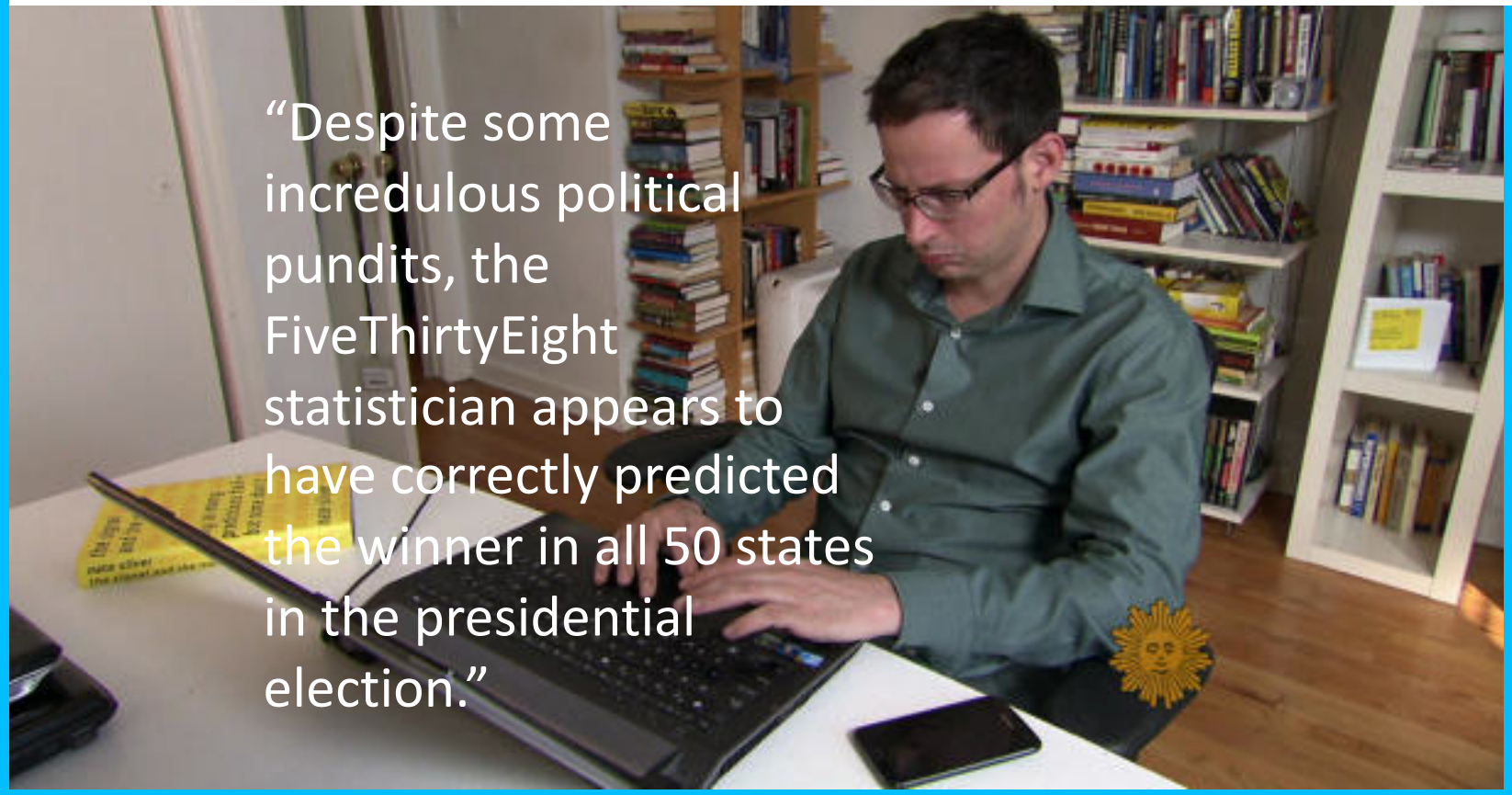
https://www.genome.gov/images/content/costpermb2015_4.jpg

Big Data and Astronomy

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

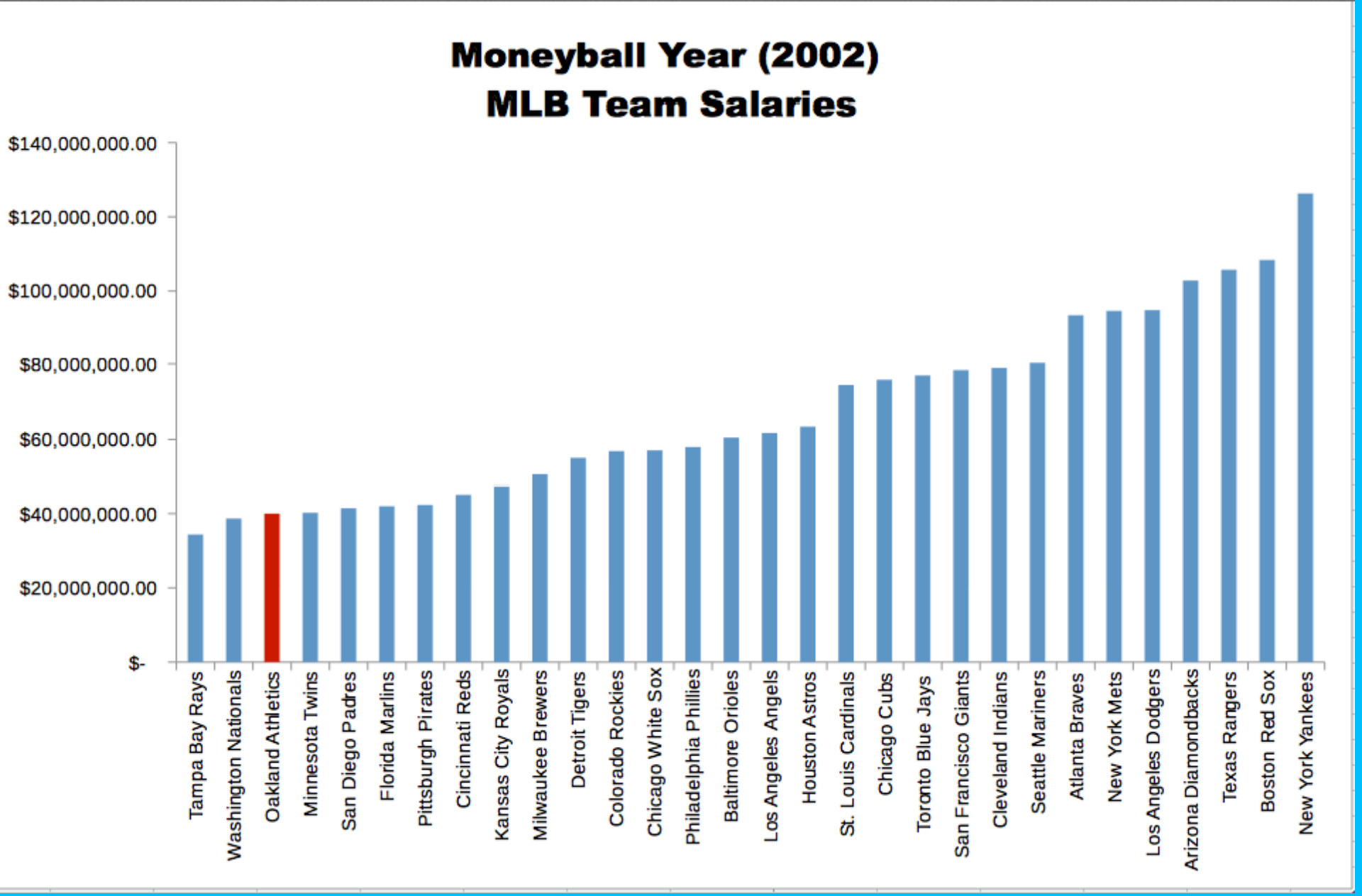
Source: http://www.skatelescope.org/news/pawsey-centre/

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



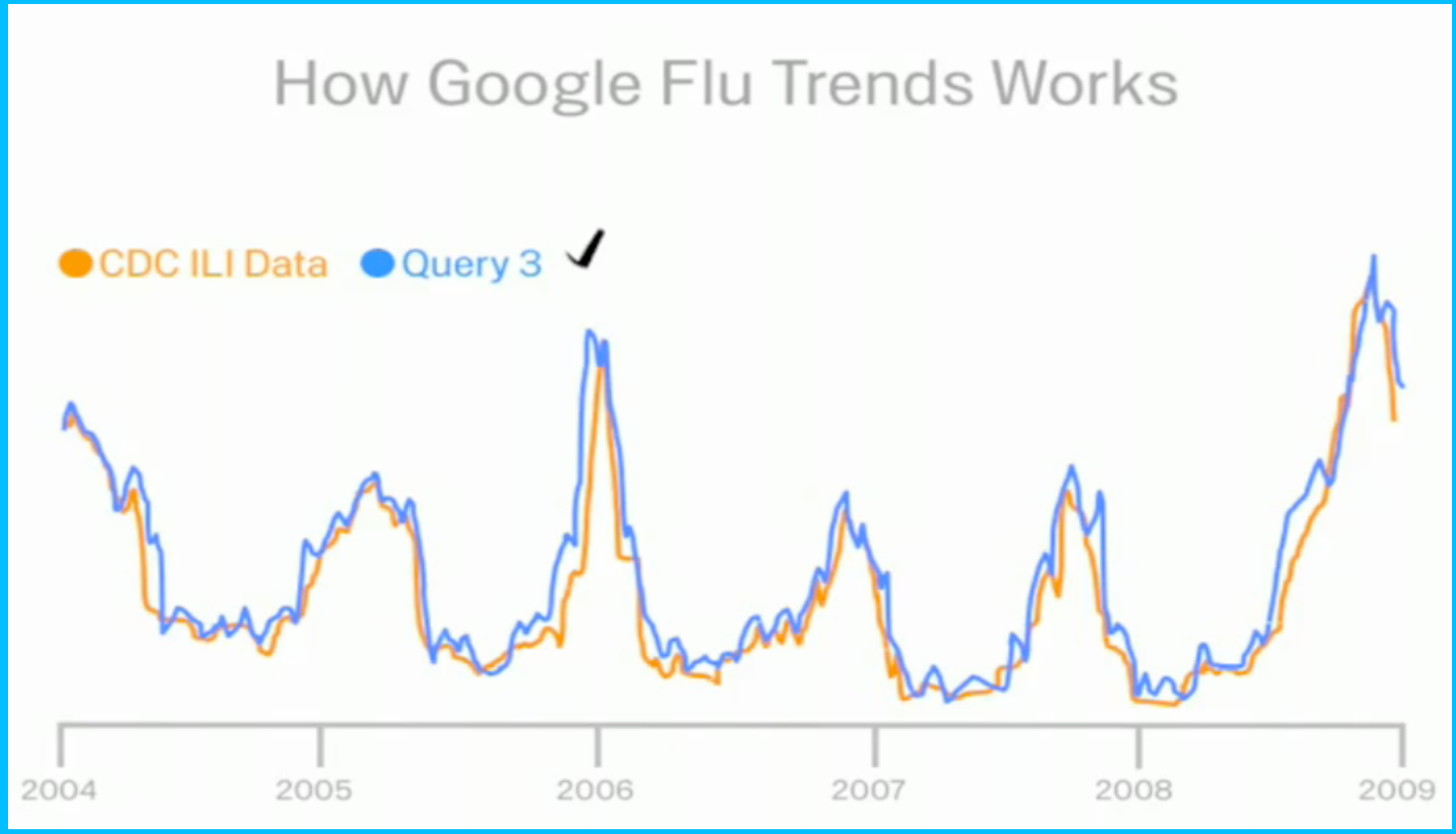
Source: http://news.cnet.com/8301-13510_3-57546161-21/obamas-win-a-big-vindication-for-nate-silver-king-of-the-quants/

https://datafloq.com/big-data-open-source-tools/os-home/



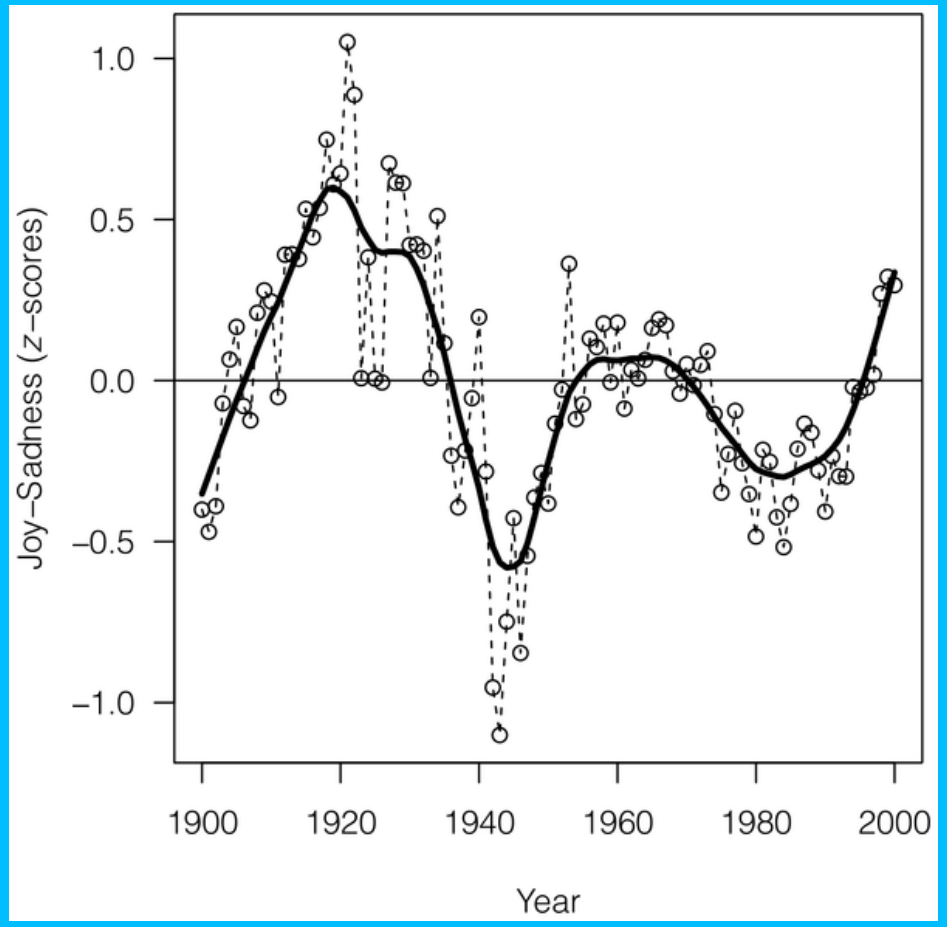
Trailer http://www.youtube.com/watch?v=AiAHlZVgXjk

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

What are other Examples of Analytics Changing the World?

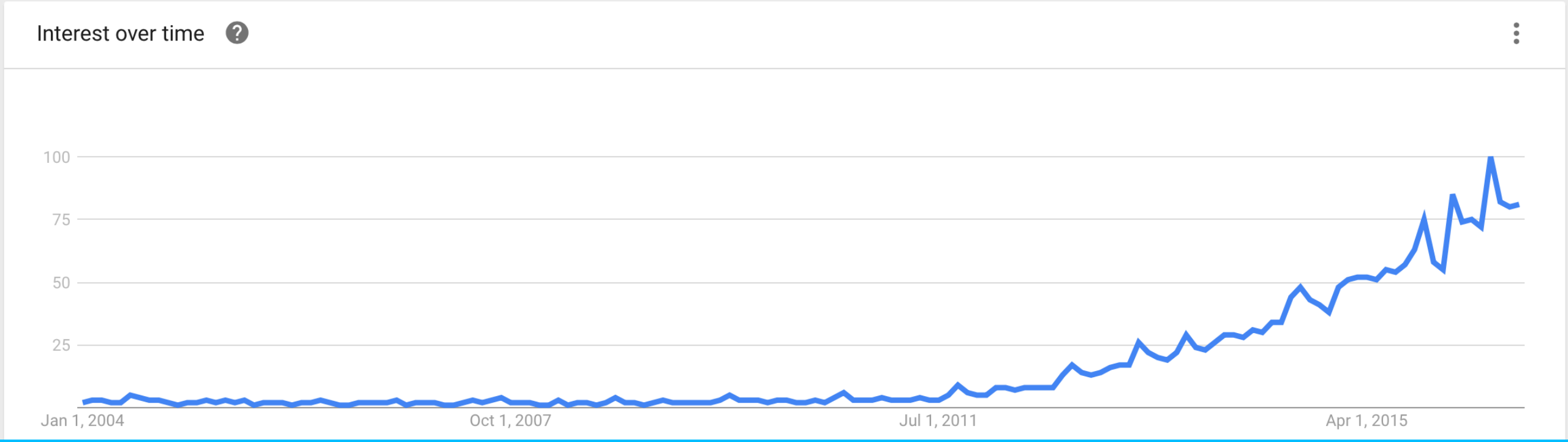
- Take 5 minutes find 1 example
- Start by telling the person next to you
- Be prepared to tell the class

What do we mean by being a data scientist?

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

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

What is a “Data Scientist”?



Accessed from Google Trends on 8/26/2016

“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

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

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

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

What skills are needed as a data scientist?

Data science is hard.

Key Tools of the Data Scientist

- **Basic Tools (R & PYTHON & SQL)**
- **Basic Statistics**
- **Machine Learning**
- **Multivariable Calculus and Linear Algebra**
- **Data Munging**
- **Data Visualization & Communication**
- **Software Engineering**
- **Thinking Like a Data Scientist**

From: http://blog.udacity.com/2014/11/data-science-job-skills.html

Data Science Venn Diagram

- Hacking Skills
- Math & Statistics Knowledge
- Substantive Expertise

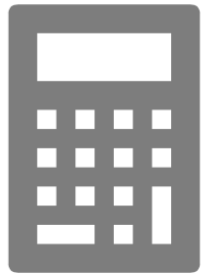
Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

TOP SOFTWARE TOOLS FOR DATA SCIENCE

Tool	2016	
	% share	% change
R	49%	4.50%
Python	45.80%	**51%**
SQL	35.50%	15%
Excel	33.60%	47%
RapidMiner	32.60%	3.50%
Hadoop	22.10%	20%
Spark	21.60%	**91%**
Tableau	18.50%	49%
KNIME	18.00%	-10%
scikit-learn	17.20%	**107%**

KD Nuggets http://bit.ly/kd2016top

Python



Statistics

Tableau

Databases

R

Spark

Github

Linked Data

Abstractions vs Tools

- Abstractions of data science
  - Matrices and linear algebra
  - Relations and relational algebra
  - MapReduce
  - Feature selection in Visualization

- Tools
  - Python/R
  - SQL/MongoDB
  - Spark/Hadoop (MapReduce)
  - Tableau (Visualization)

Syllabus and Course

https://jkuruzovich.github.io/tech-fundamentals-analytics/

Class Overview

{JSON}

CSV

Spark

API

DATA MUNGING

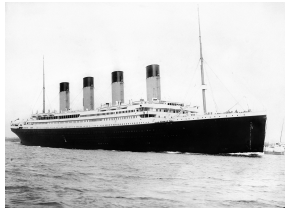
Retrieve-Filter-Missing Data-Data Cleaning-Aggregate-Merge-Missing Values-Feature Creation-Text Tools (Lemmatization-Stemming-Corpus-Bag of Words-TFIDF))-Sampling-K-Fold Cross Validation

DATA FUNDAMENTALS

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

Modeling

kaggle



R

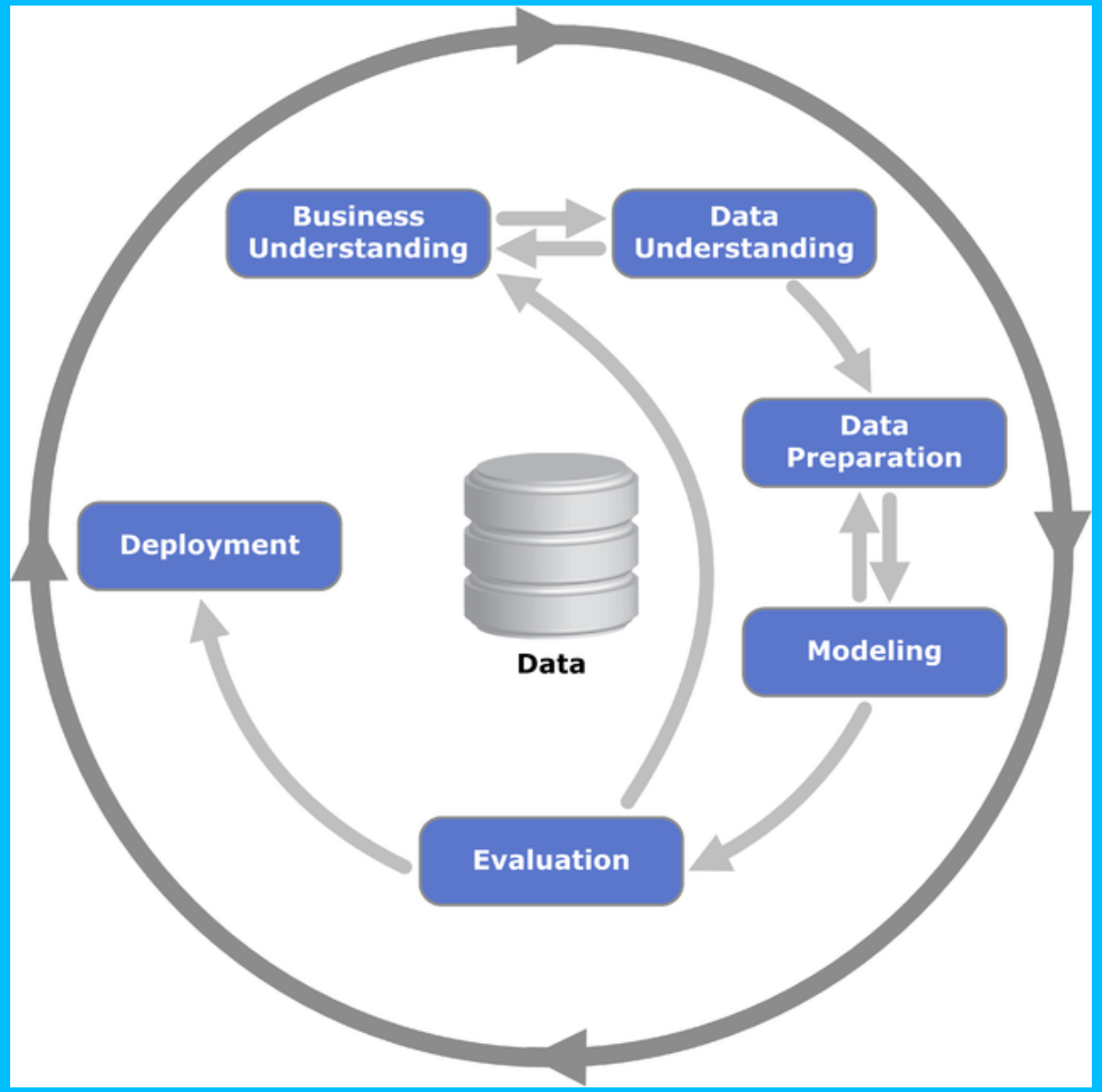
python

GitHub

Yummly

CRISP-DM

"Cross Industry Standard Process for Data Mining, commonly known by its acronym CRISP-DM, was a data mining process model that describes commonly used approaches that data mining experts use to tackle problems."
-Wikipedia



Credit: By Kenneth Jensen (Own work) [CC BY-SA 3.0], via Wikimedia Commons

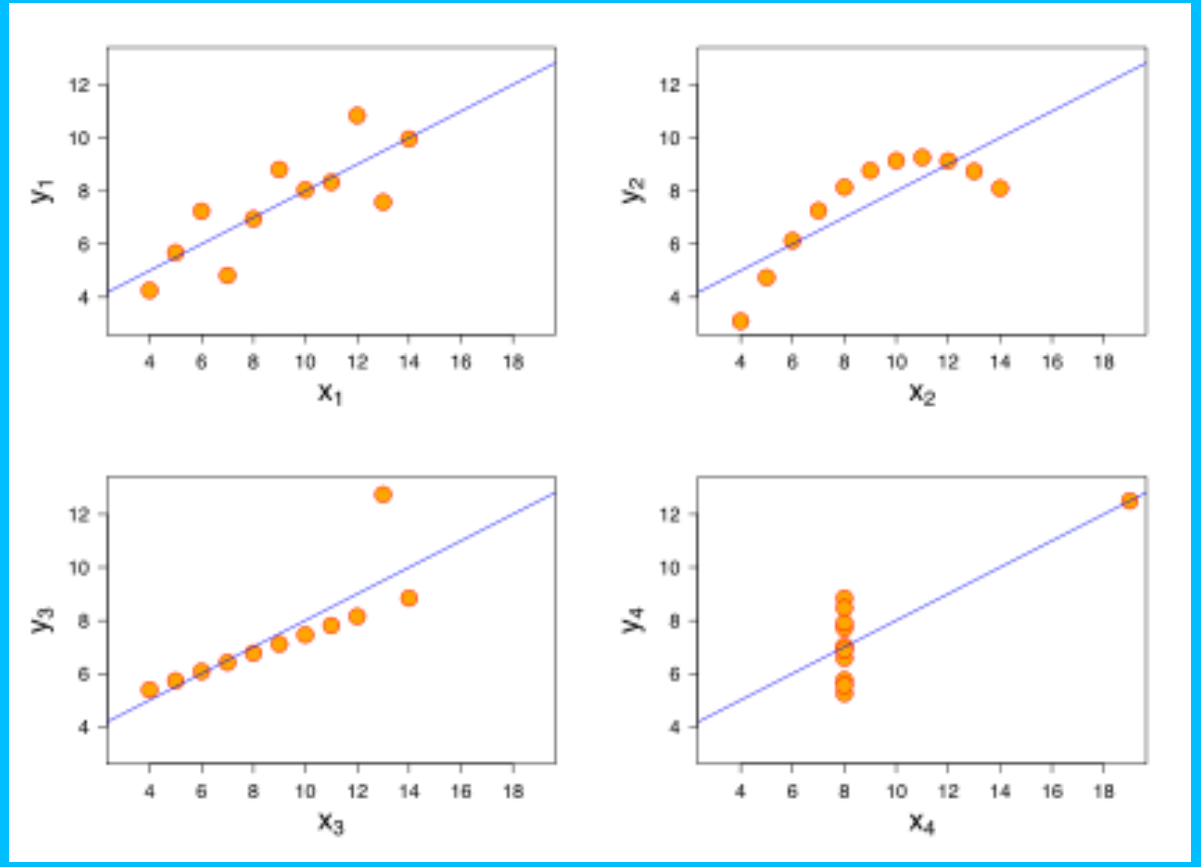
We can discover meaning through data munging visualization, statistics, and machine learning

Data comes in normal and *big data*

80% of the work a data scientist does is collecting, cleaning and organizing data.

Why Visualize?

All 4 sets for variables have the same mean, standard deviation, and correlation, and regression line



	Goals
Statistics	EXPLAIN the role of specific constructs
Machine Learning	CALCULATE an ACCURATE PREDICTION

Statistical Models

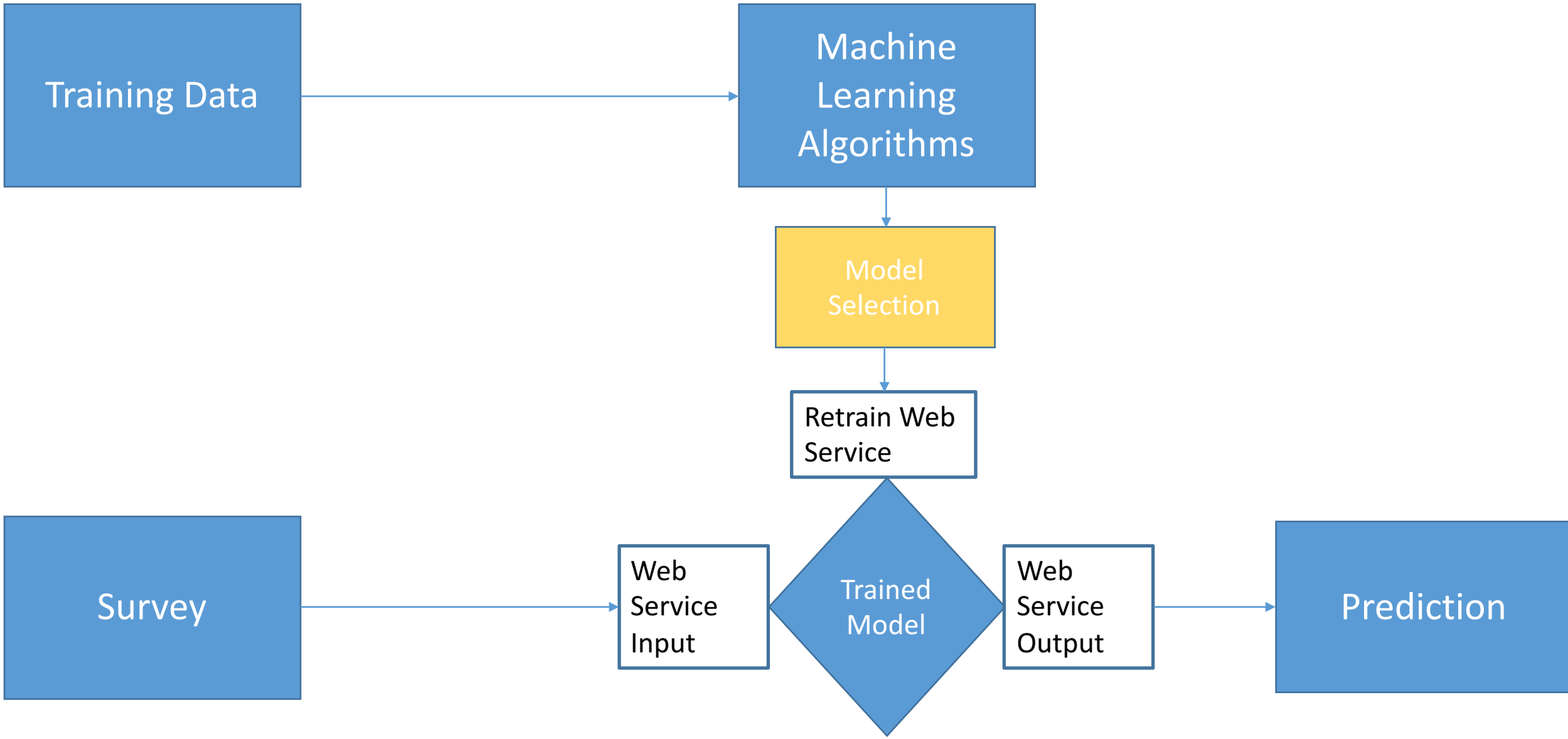
- Provide explanations for things
  - What is the role of gender in school performance
  - Which groups perform better
  - What foods reduce your likelihood of dying
- Estimate magnitude of effect
  - How much does an increase in police reduce crime?
  - How much does spending on education influence student performance?

Correlation does not imply causation.

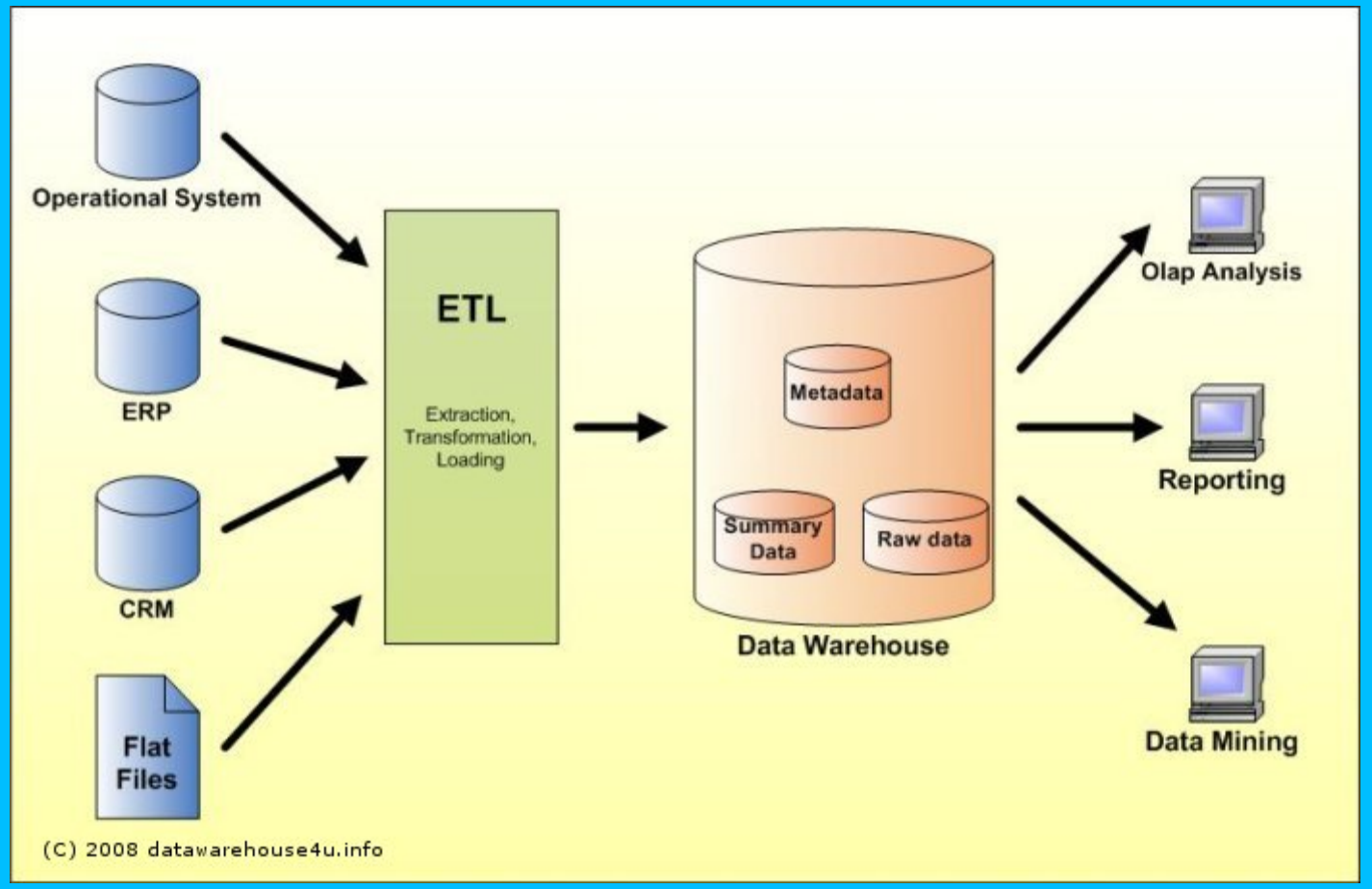
A good final note for statistics…next on to machine learning…

"Field of study that gives computers the ability to learn without being explicitly programmed”

-Arthur Samuel (on machine learning)


Training Data
Machine Learning Algorithms
Model Selection
Retrain Web Service
Survey
Web Service Input
Trained Model
Web Service Output
Prediction

Big Data 1.0


Operational System
ERP
CRM
Flat Files
ETL
Extraction, Transformation, Loading
Metadata
Summary Data
Raw data
Data Warehouse
Olap Analysis
Reporting
Data Mining
(C) 2008 datawarehouse4u.info

Big Data 1.0

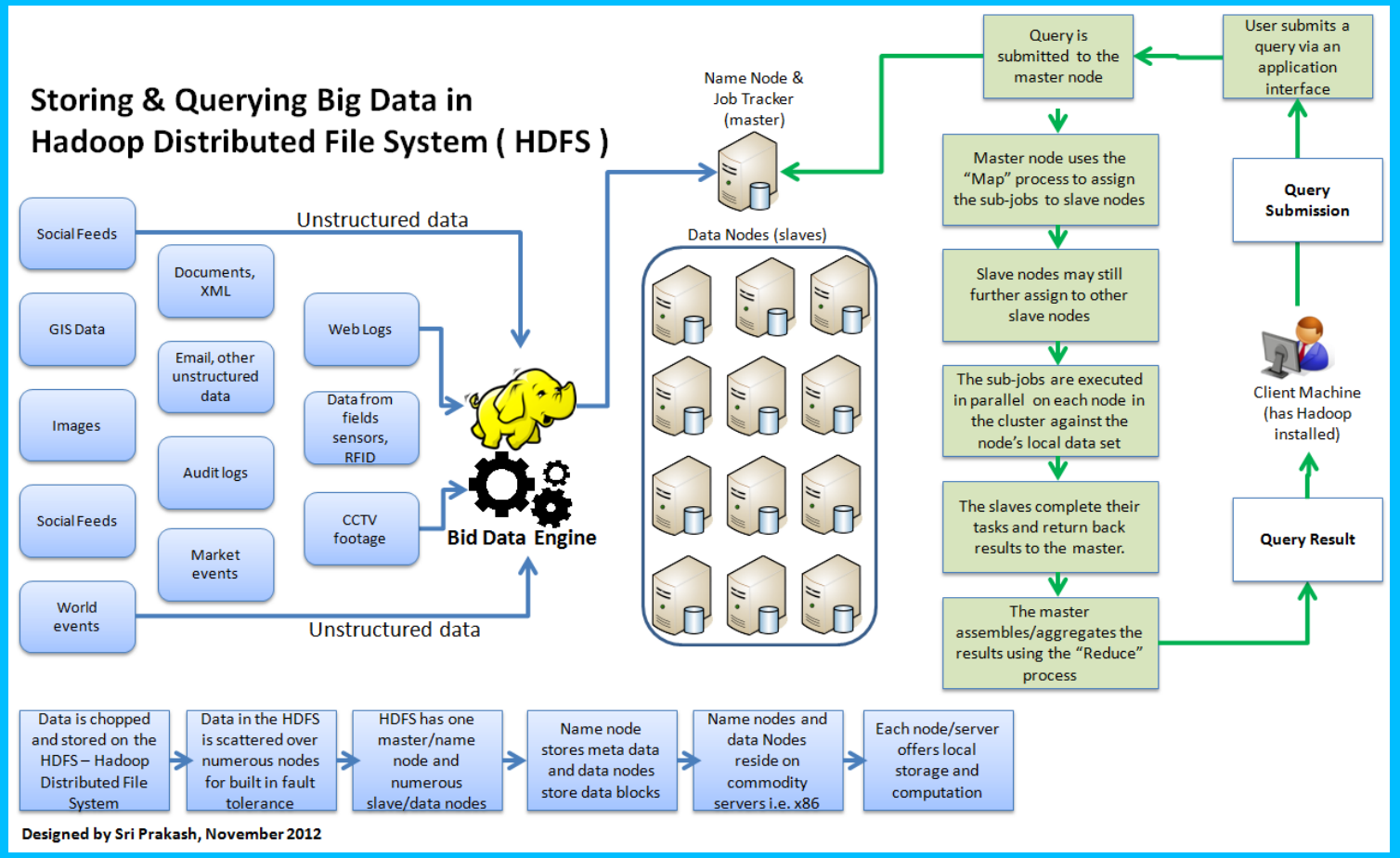
- Online Transaction Processing Systems (OLTP)
  - CRM, ERP, etc
- Extract, Transform, Load (ETL)
  - Extract data from all OLTP systems and put into a Data Warehouse
- Data Warehouse
  - System for creating reports, exploring existing data

Data Warehouse Limitations

- Not setup for extremely large datasets
  - Weblogs capturing all data
  - Data can be too large to hold in memory for a single location
  - Data from non-transactional sources (Twitter, logs, email) doesn’t fit as well with

Big Data 2.0

HADOOP and HDFS



It takes time to get good at data science.

Tips for Learning New Skills

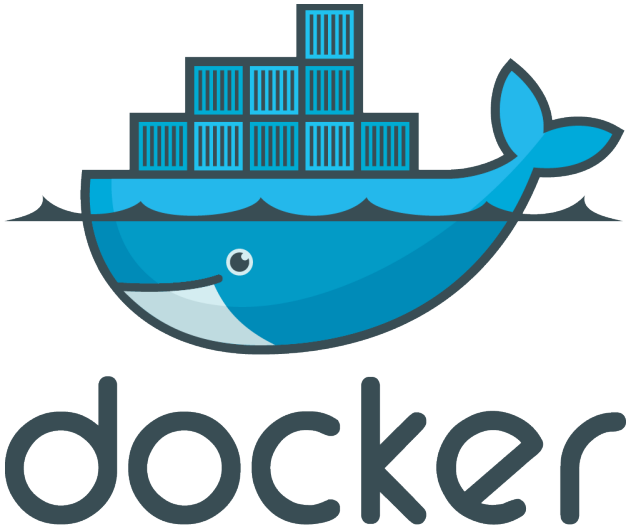
- Clarify terminology, mental models for what is being done conceptually
- Watch videos and lectures
- Practice – DO IT! Set aside time for practice
- Kaggle Competitions

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.

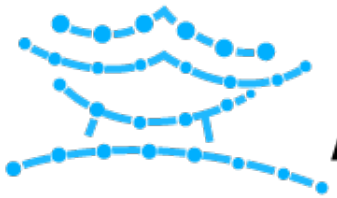
When R and when Python?
(Or why do I need both?)

Technology Platforms

kaggle



Azure
Machine Learning



Analytics dojo

jupyter Welcome to Python (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Python 3

Markdown CellToolbar

jupyter rackspace

Welcome to the Temporary Notebook (tmpnb) service!

This Notebook Server was **launched just for you**. It's a temporary way for you to try out a recent development version of the IPython/Jupyter notebook.

WARNING

Don't rely on this server for anything you want to last - your server will be *deleted after 10 minutes of inactivity*.

Your server is hosted thanks to Rackspace, on their on-demand bare metal servers, OnMetal.

Run some Python code!

To run the code below:

1. Click on the cell to select it.
2. Press `SHIFT+ENTER` on your keyboard or press the play button (▶) in the toolbar above.

A full tutorial for using the notebook interface is available here.

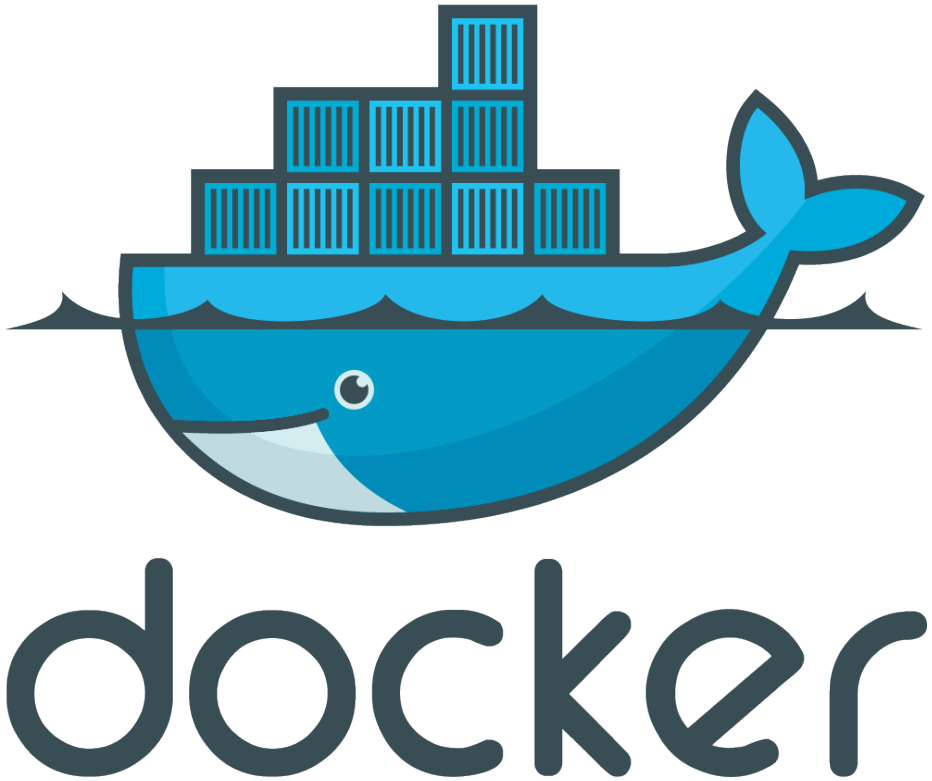
In [ ]:

```
%matplotlib notebook

import pandas as pd
import numpy as np
import matplotlib

from matplotlib import pyplot as plt
```

www.jupyter.com


docker

Azure
Machine Learning

Goal is content and technology for learning analytics that can be used online and in classrooms

Assignment (see course website)

Assignment

- https://gist.github.com/jkuruzovich/672945f488de09dee47010a6f3f343ba

Reading

- https://gist.github.com/jkuruzovich/6cc69cda7778fa58c66ac3da0bd2ce39