Technology Fundamentals for Analytics

Jason Kuruzovich

Overview

- Introductions
- Let's get excited
- Course Syllabus
- Appetite wetting
- Lab 1. Installation of key software

Me

- Director of the Severino Center for Technological Entrepreneurship
- Associate Professor of Information Systems, Marketing, and Entrepreneurship
- Research on marketing, multichannel retailing, most recently entrepreneurship

Startup Tech Valley | Mission | Format | Events | Entrepreneurs | Sponsors | Get Invo

Presenting at the April 2nd Event:

Dumbstruck – Application to capture reactions to messages you send from your mobile device.
Peter Allegretti

Schodack Central School District 'Accelerator' – Schodack schools are working with start-ups to provide them space to grow their business while providing our students with real-world opportunities.
Bob Horan

Smart Kids NY – A classroom visiting series introducing elementary students to new technologies through hands-on demos, experiments and field trips led by young entrepreneurs.
Peg Zokowski

Vistex Composites – A disruptive technology for the manufacture of advanced thermoset and thermoplastic composite materials.
Casey Hoffman

Make It Private – A secure online communication tool used for protecting one's privacy.
Laben Coblentz

Rollio – A mobile utility tool and an analytics based web-app for client meetings and business trip optimization
Jake Soffer

<Advertisement>
Sept 3
5:30
Brown's
Revolution Hall

FREE BEER* + GREAT NETWORKING + INSPIRING TALKS

Agenda

- Information and society
- What do we mean by data scientist/analytics?
- Let's get excited
- Analytics Overview
  - Visualization, Statistics, and Machine Learning
- Big Data
- Where to go from here…becoming a data science rock star….

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

Internet 0.1 beta (18th Century)

Internet 0.1 Beta (18th Century)

Chain of towers or optical telegraph capable of transmitting 1-3 symbols per min

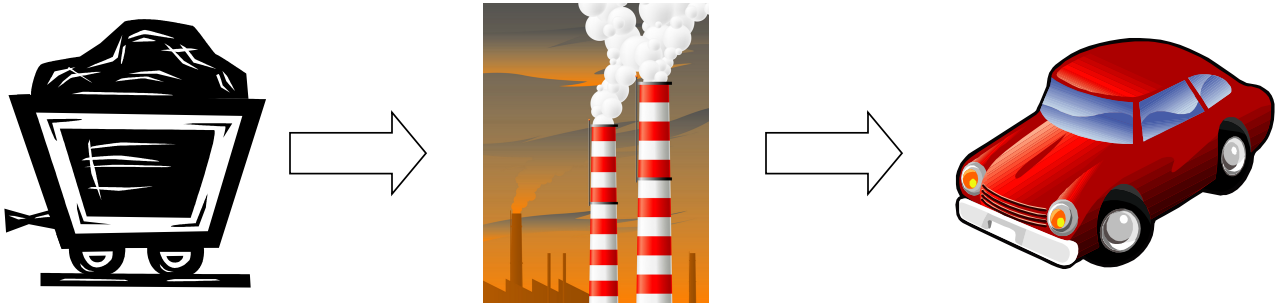
Towers 5-20 KM apart

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

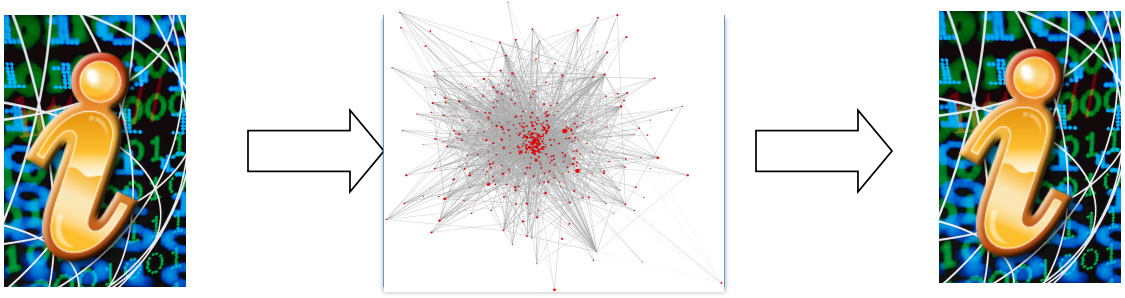
-Eric Schmidt, Former CEO of Google

Information Economy

TRADITIONAL PRODUCTION PROCESS



INFORMATION BASED BUSINESS PROCESS



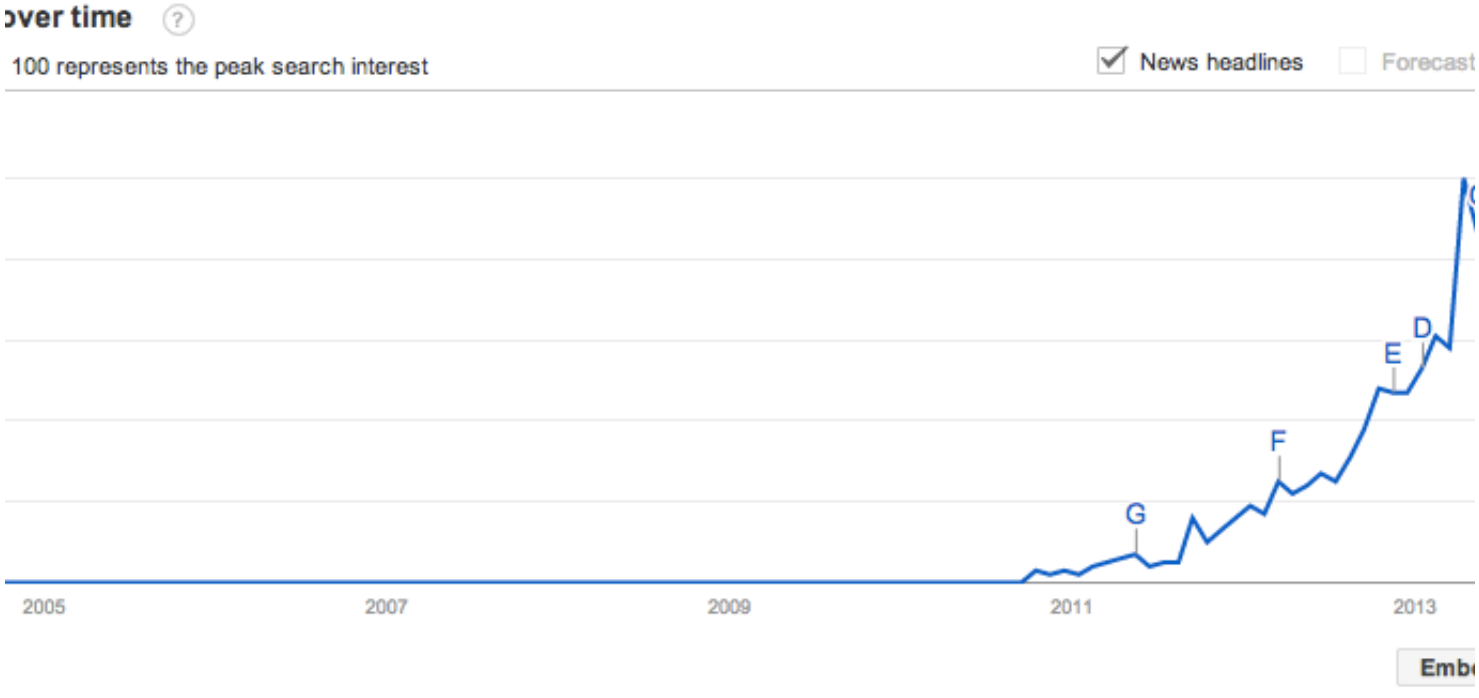
INFORMATION TECHNOLOGY

Internet of Things

- Equipping all objects in the world with minuscule identifying devices could be transformative of daily life. For instance, business may no longer run out of stock or generate waste products, as involved parties would know which products are required and consumed.
- http://www.thetileapp.com

The Facebook Social Network

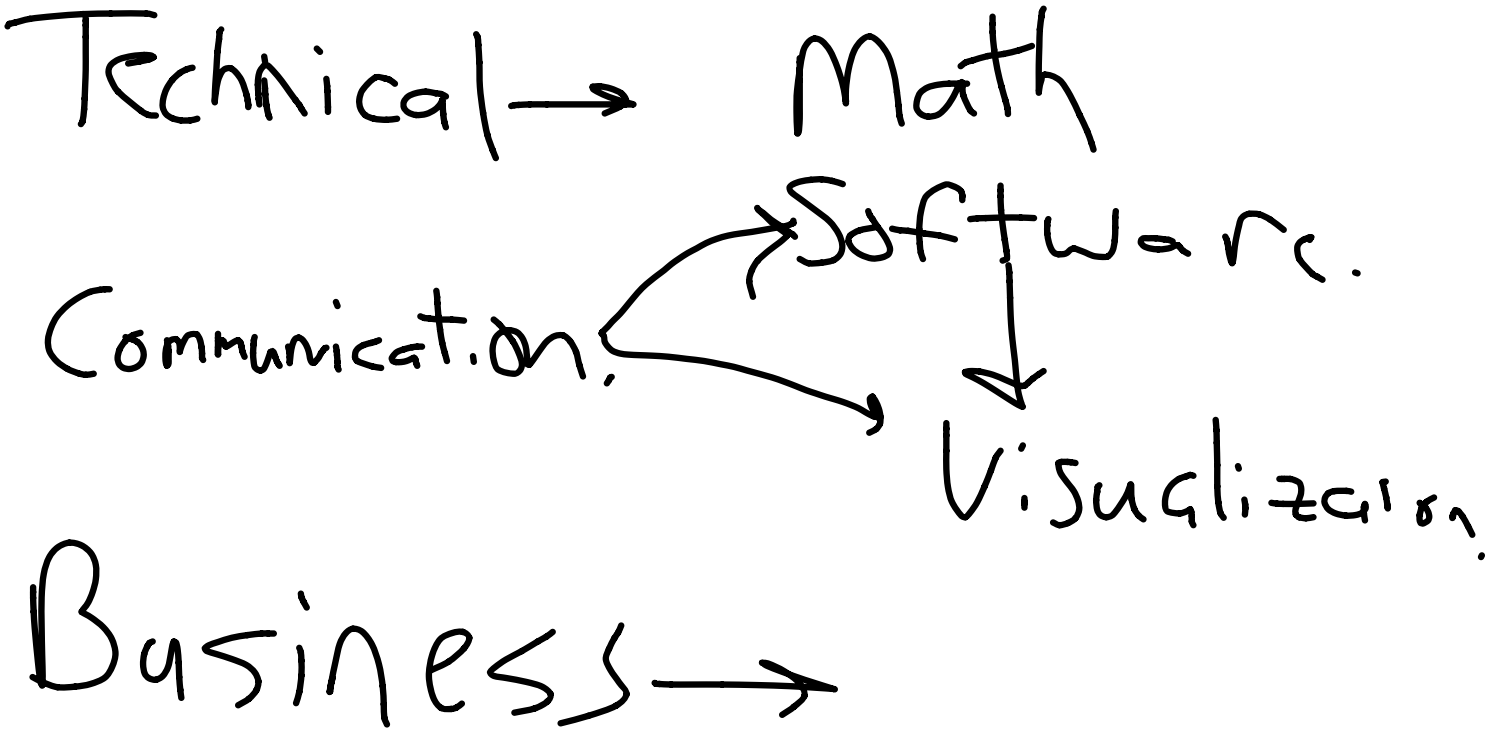
What is a “Data Scientist”?



What is a "Data Scientist"?

Work with / interpret / Pattern Recog.

What skills are needed as a data scientist?



Data Scientist

- "A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product."

  - Hilary Mason, chief scientist at bit.ly

"data wrangling" "data jujitsu" "data munging"

Data Scientist

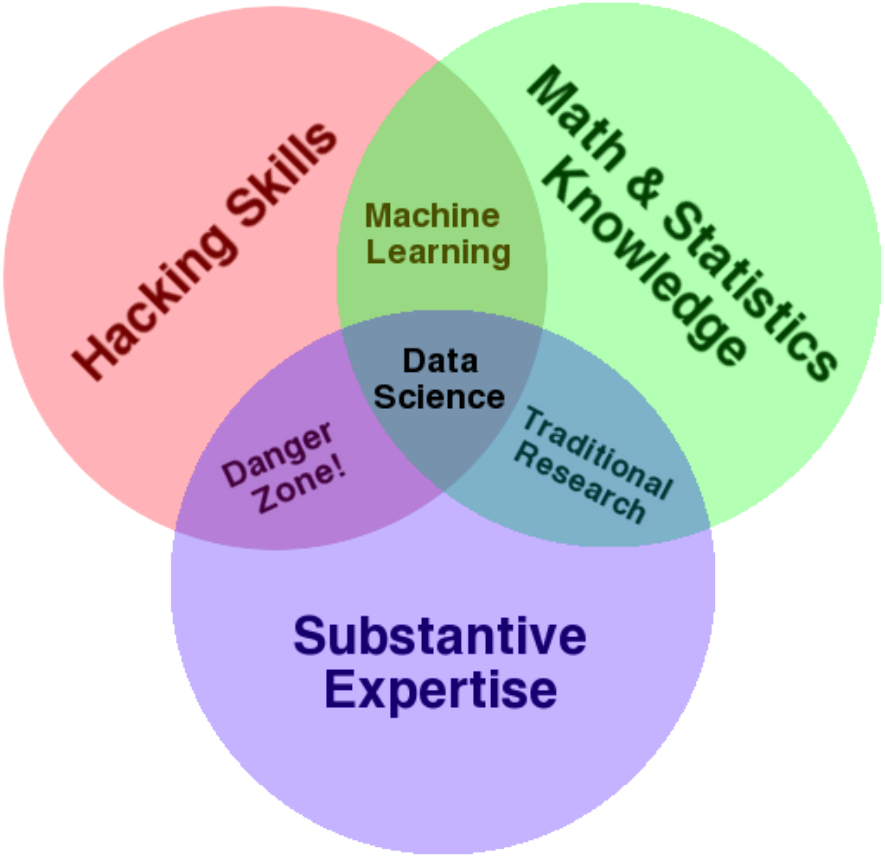
Data science requires skills ranging from traditional computer science to mathematics to art. Describing the data science group he put together at Facebook (possibly the first data science group at a consumer-oriented web property), Jeff Hammerbacher said:

"... on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization."

Key Tools of the Data Scientist

- Data Munging - parsing, scraping, and formatting data
- Statistics - traditional analysis you're used to thinking about
- Visualization - graphs, tools, etc.

Data Science Venn Diagram



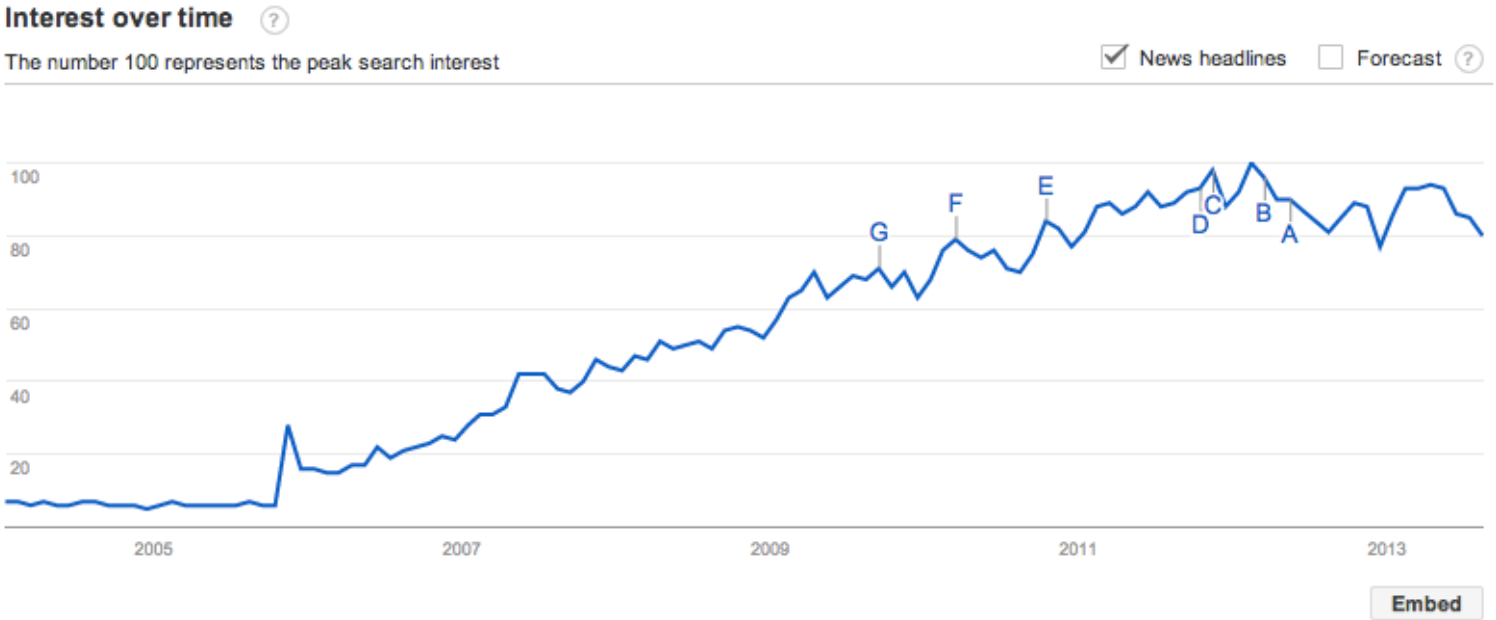
Source:
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Abstractions vs Tools

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

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

What do we mean by “Analytics”?
(Term Frequency in News)



Source:
http://www.google.com/trends/explore?q=analytics#q=analytics&cmpt=q

What do we mean by "Analytics"?

Process for interpreting Data.

Active → Prediction.

Analytics as Data Apps

- The web is full of "data-driven apps."

"The thread that ties most of these applications together is that data collected from users provides added value. Whether that data is search terms, voice samples, or product reviews, the users are in a feedback loop in which they contribute to the products they use. That's the beginning of data science."

Prediction is a Great Business Model!

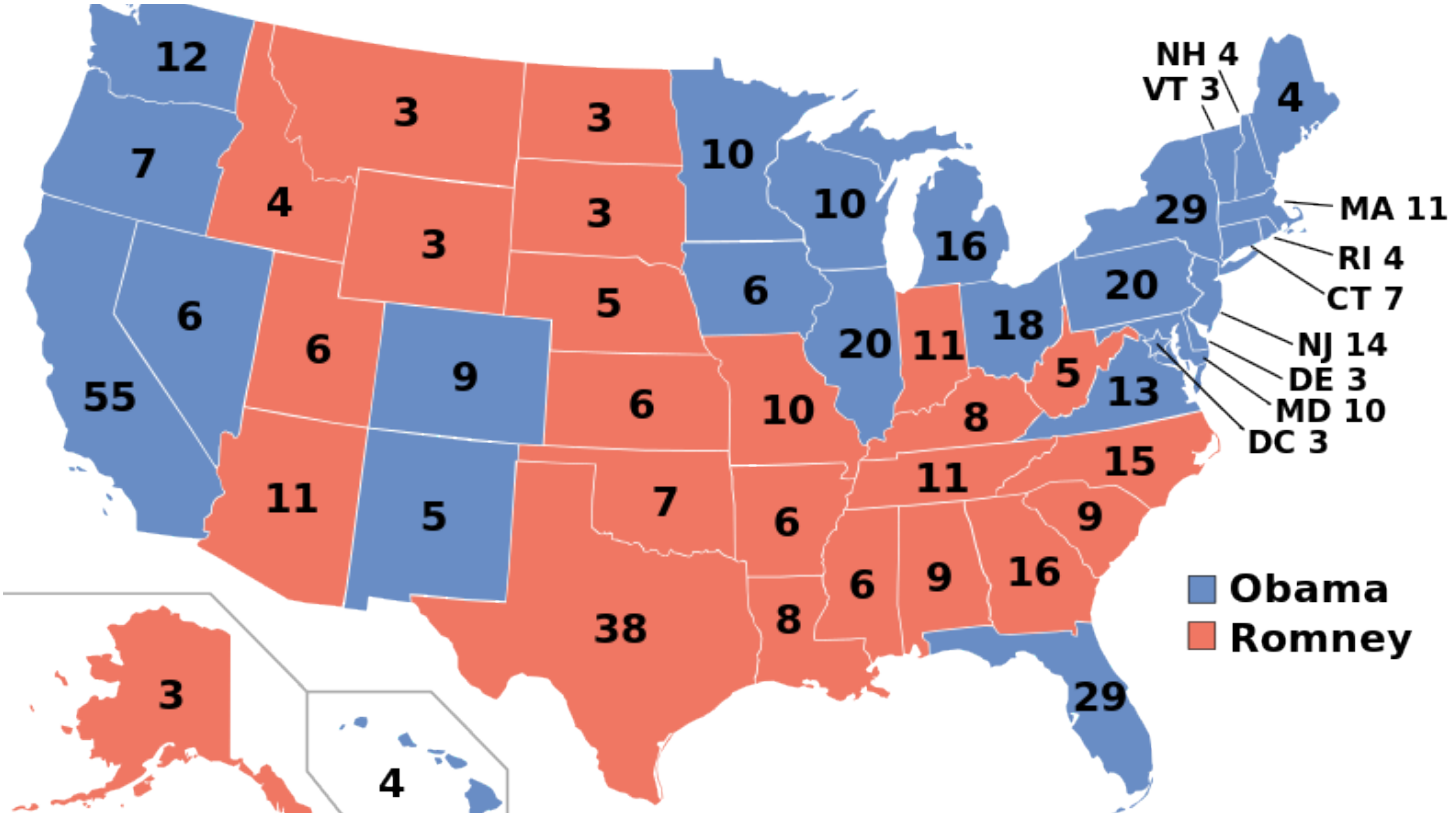
June 18, 2012, 4:19 PM

People Will Pay for 'Transparently Useless Advice' About Chance Events

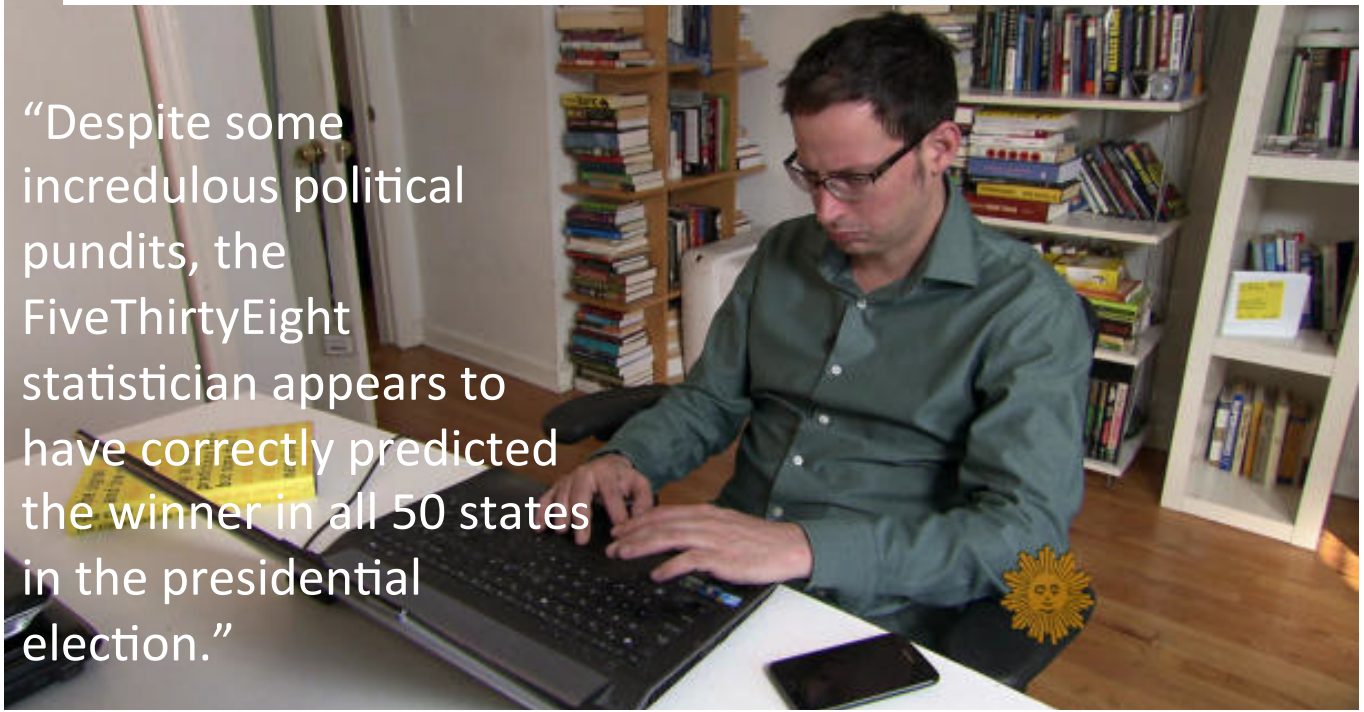
Before each flip, each participant got the chance to pay to read a written prediction of what the result of the next toss would be. If they didn't purchase the "tip," they still got to read the prediction, afterward, free of charge.

Obviously, the chances of the "expert" advice being right, one time, is —wait for it— 50%. Still, when the advice was correct after the first coin toss, roughly 12% of players paid for advice in the second round (compared to roughly 4% if the prediction was wrong). Two correct predictions in a row inspired about 20% of players to buy a prediction for round three. The proportion of people buying bets continued to climb until, after four correct predictions, more than 40% of players were paying for advice for round five.

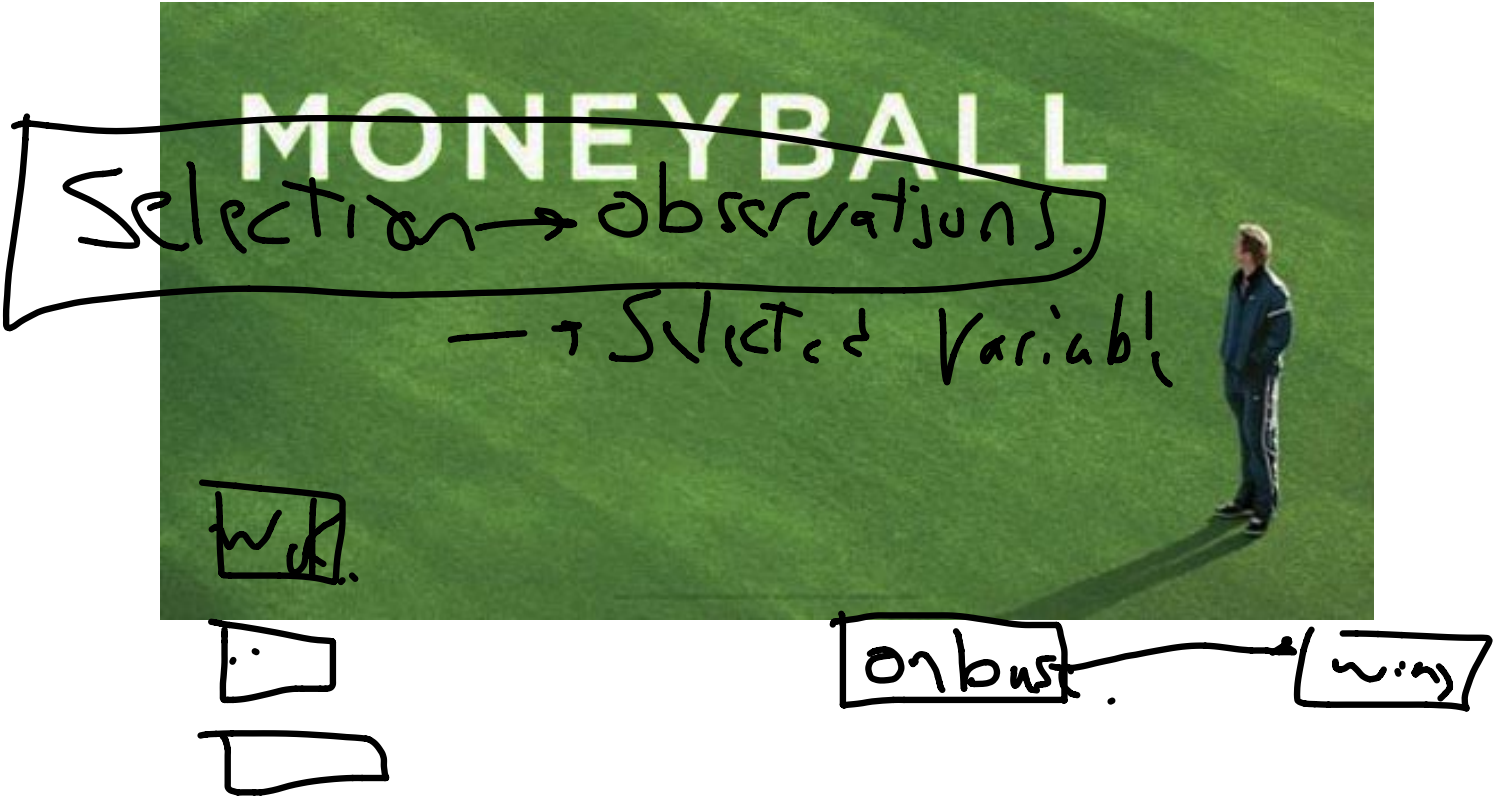
Electoral College



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/



Introduction to Moneyball

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

Obama's Tech

- "The 2012 campaign took advantage of advanced set-top-box monitoring technology to figure out what shows the voters they wanted to reach were watching and when, resulting in a smarter and cheaper — if potentially more invasive — way to beam commercials into their homes. The system gave Obama a significant advantage over Mitt Romney, according to Democrats and many Republicans (at least those who were not on Romney's media team)."

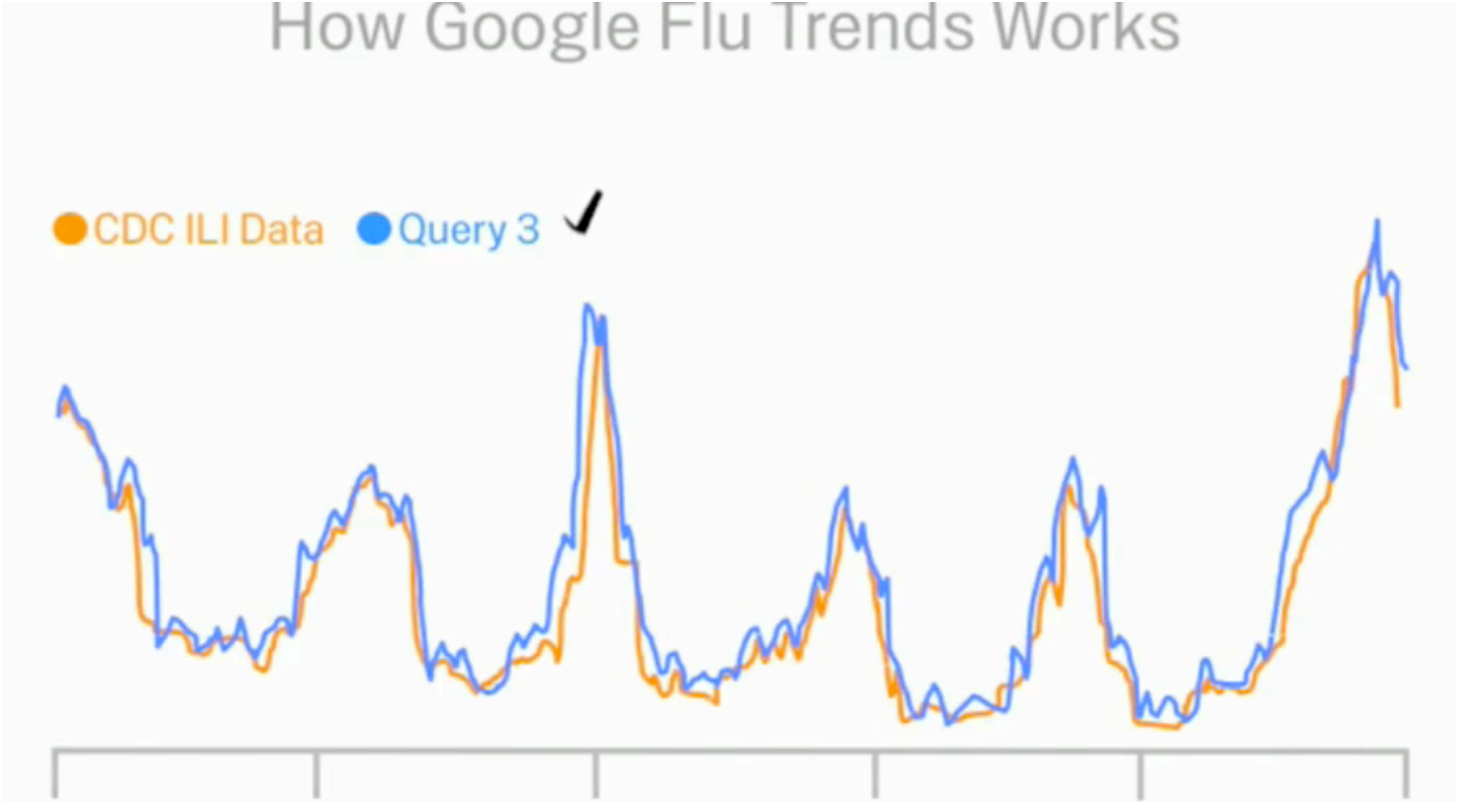
Source: http://www.nytimes.com/2013/06/23/magazine/the-obama-campaigns-digital-masterminds-cash-in.html?pagewanted=all&_r=0

Obama's Tech

- The campaign couldn't call the more than 150 million registered voters, obviously. But they could call enough of them in swing states (up to 11,000 a night) to figure out how they — and other people who lived near them, looked like them and earned like them — were likely to vote with an increasing degree of accuracy. In 2008, Wagner and his small team combined information from those calls with any other data they could find — census data, state voter lists and the like — and fed it into algorithms that produced support scores.

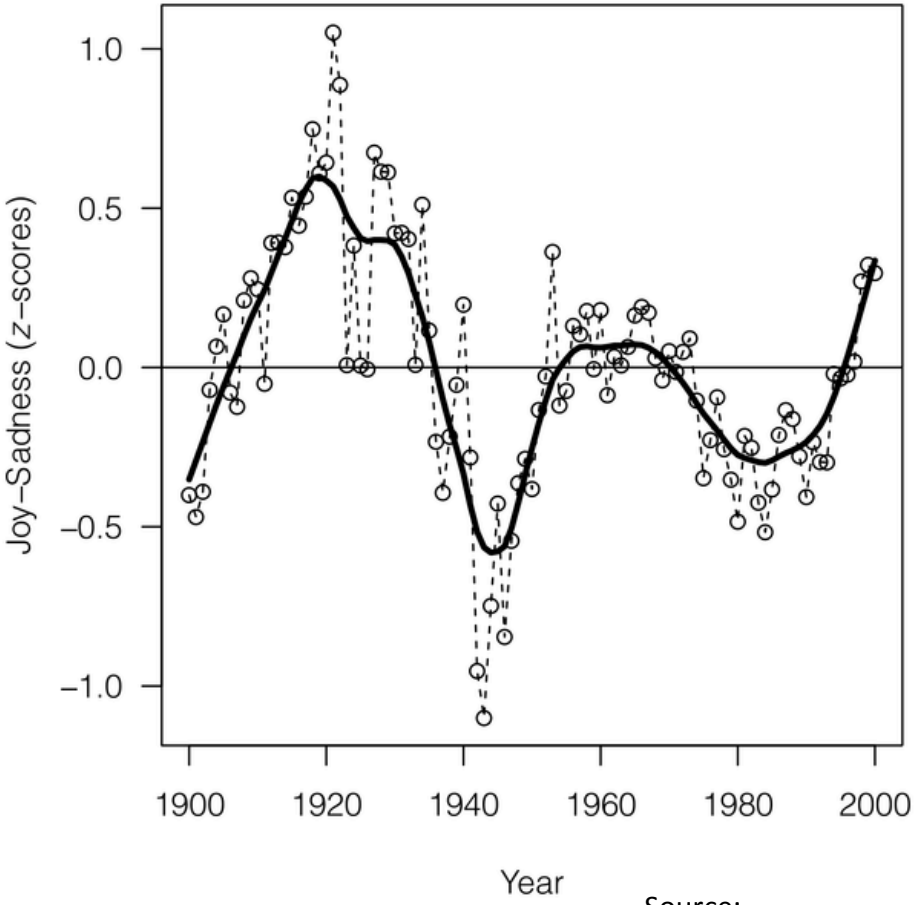
Source: http://www.nytimes.com/2013/06/23/magazine/the-obama-campaigns-digital-masterminds-cash-in.html?pagewanted=all&_r=0

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

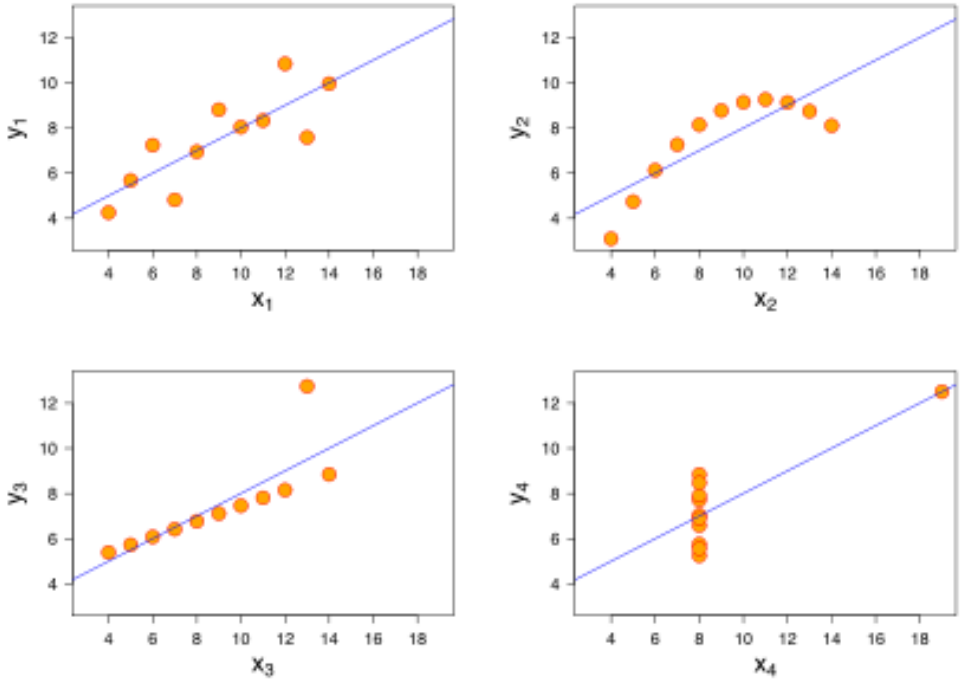
Analytics is the discovery and communication of meaningful patterns in data.

-Wikipedia

We can discover meaning through visualization, statistics, data mining

Why Visualize?

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



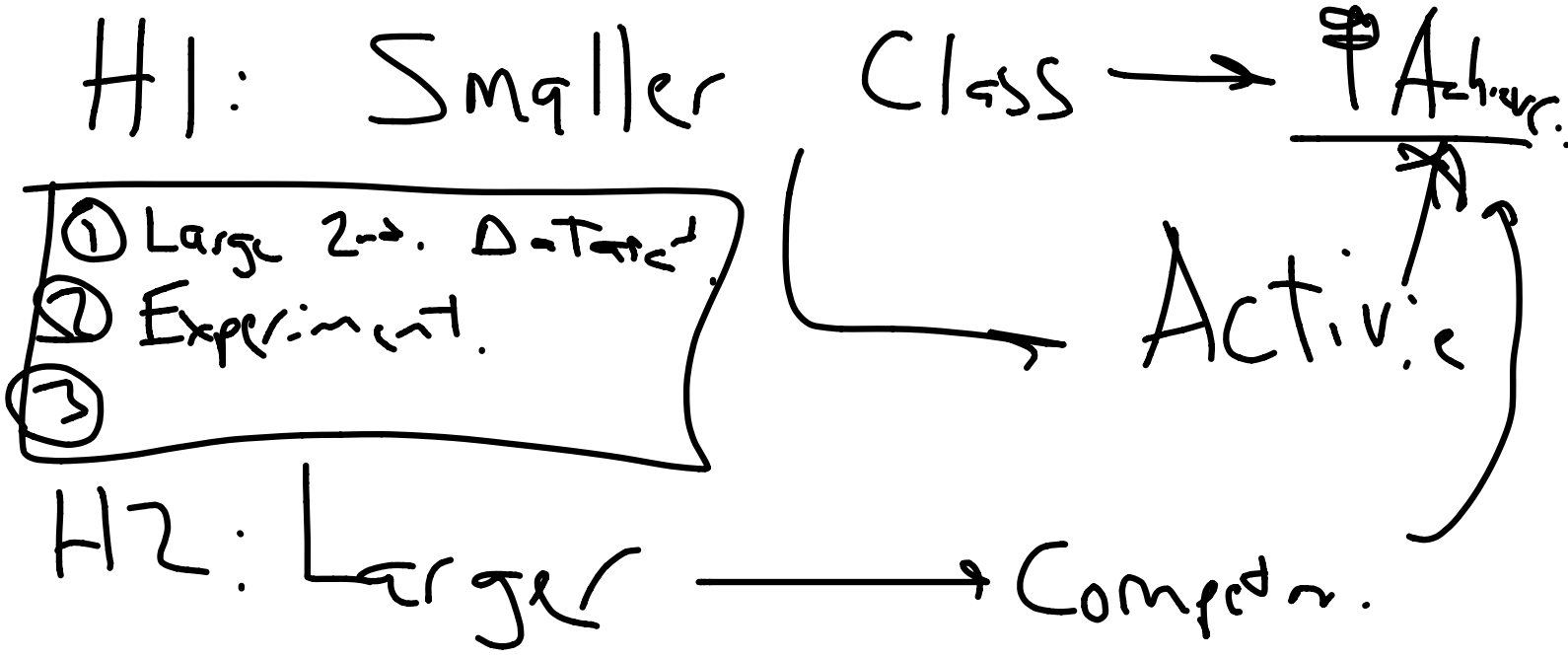
	Goals
Statistics	EXPLAIN the role of specific constructs
Machine Learning	CALCULATE an ACCURATE PREDICTION

	Features	Model
Statistics	MEASURE VALIDATED CONSTRUCTS of interest used by OTHER RESEARCHERS	DATA REDUCTION and EASY UNDERSTAND RELATIONSHIP ANALYSIS (SEM or REGRESSION)
Predictive Analytics	CONSIDER ALL AVAILABLE DATA (there might be some relevant nuggets)	Complex BLACK BOX methods like NEURAL NETWORKS and SUPPORT VECTOR MACHINES

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?

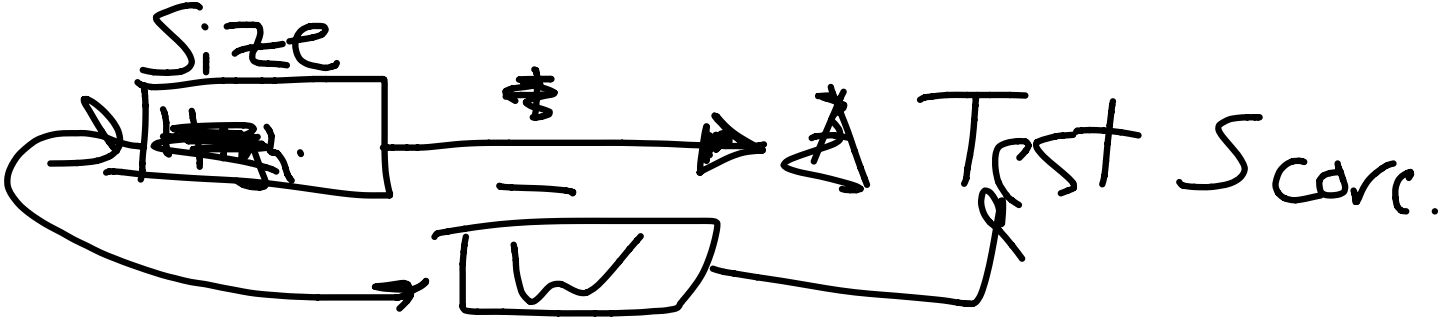
How could we find out the role of class size in student achievement?



What would your hypothesis be?

Observational Study

- In an observational study, the researchers merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations. There is no human intervention.



What would your hypothesis be?

Observational Study

- Examine different class sizes across different populations of children
  - Limitations
    - Could be other socio-economic factors involved that are not captured

Designed Experiment Study

- From within the same population of students, split children and give some children a big class and some a small class
  - Limitations:
    - There could be great resistance from administrators, parents, students
    - No way to make “double blind”

Natural Experiment

- A natural experiment is an empirical study in which the experimental conditions (i.e., which units receive which treatment) are determined by nature or by other factors out of the control of the experimenters and yet the treatment assignment process is arguably exogenous or "as-if random."
  - Example: what if one year a classroom has 40 students and the next year it has 20, just through random variation in population

LIFE & CULTURE | October 8, 2013, 8:32 p.m. ET

Secrets of the Happiest Commuters

Take Control of the Details; Women Travel Differently

- A person who commutes an hour each way has to make 40% more money to be as satisfied with life as a person who lives near the office, according to research co-authored by Alois Stutzer, an economics professor at the University of Basel in Switzerland.

- What type of study?

Smoking Ban

- In Helena, Montana during the six-month period from June 2002 to December 2002 a smoking ban was in effect in all public spaces in Helena including bars and restaurants. Helena is geographically isolated and served by only one hospital. It was observed that the rate of heart attacks dropped by 60% while the smoking ban was in effect.

What type of study?

Correlation does not imply causation.

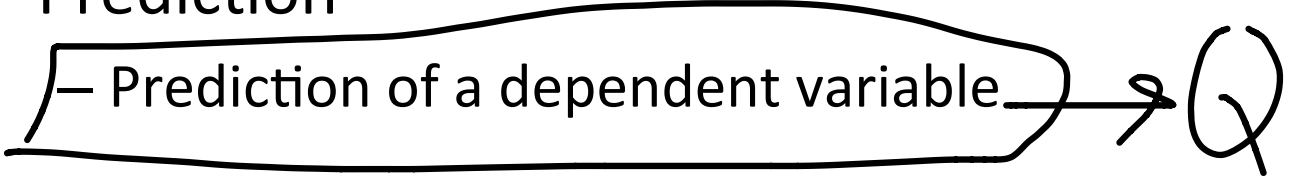
A good final note for statistics…next on to machine learning…

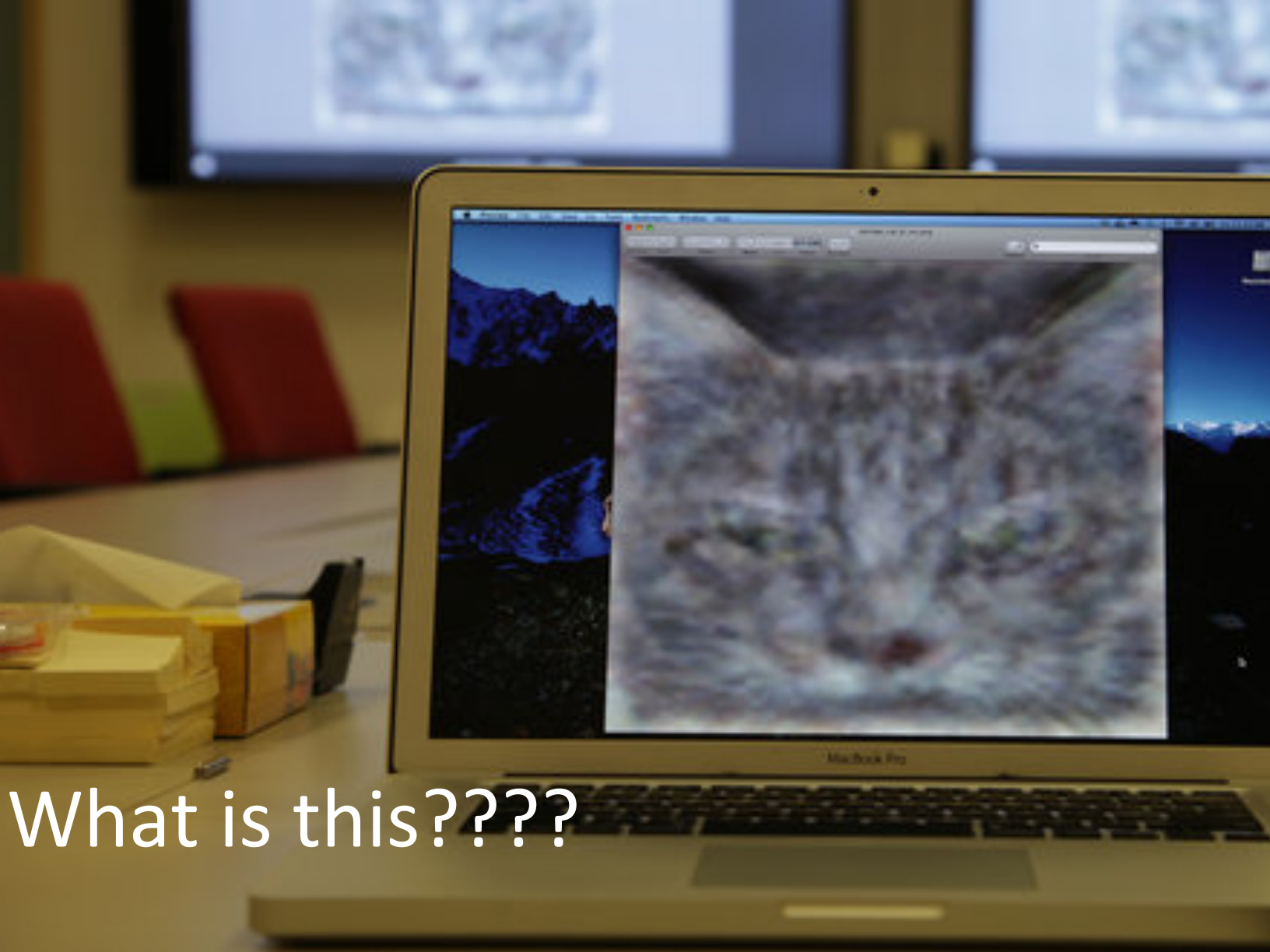
"Field of study that gives computers the ability to learn without being explicitly programmed”

-Arthur Samuel (on machine learning)

Main Categories of Machine Learning Problems

- Classification
  - Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.
- Prediction
  - Prediction of a dependent variable




What is this????

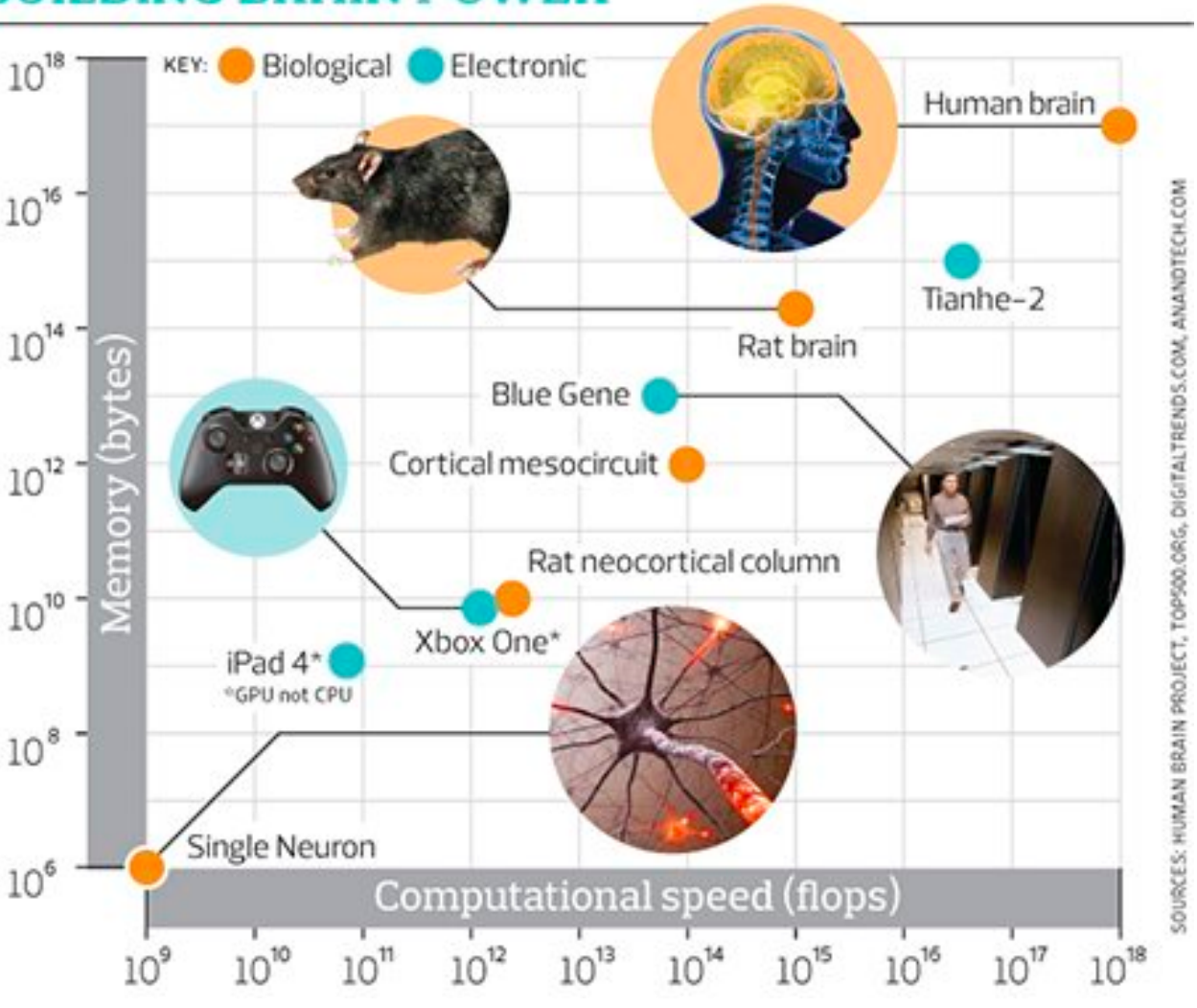
When computer scientists at Google's mysterious X lab built a neural network of 16,000 computer processors with one billion connections and let it browse YouTube, it did what many web users might do — it began to look for cats.

Building high-level features using large scale unsupervised learning

- "Contrary to what appears to be a widely-held intuition, our experimental results reveal that it is possible to train a face detector without having to label images as containing a face or not. Control experiments show that this feature detector is robust not only to translation but also to scaling and out-of-plane rotation."

Concept	Random guess	Same architecture with random weights	Best linear filter	Best first layer neuron	Best neuron	Best neuron without contrast normalization
Faces	64.8%	67.0%	74.0%	71.0%	**81.7%**	78.5%
Human bodies	64.8%	66.5%	68.1%	67.2%	**76.8%**	71.8%
Cats	64.8%	66.0%	67.8%	67.1%	**74.6%**	69.3%

BUILDING BRAIN POWER



What do we mean by
“**unsupervised learning?**”

Unsupervised Learning

Unsupervised Learning

- In machine learning, the problem of unsupervised learning is that of trying to find hidden structure in unlabeled data.
  - Example: Could you create categories of customers based on their purchase behavior and demographics

Supervised vs. Unsupervised Learning of Cats

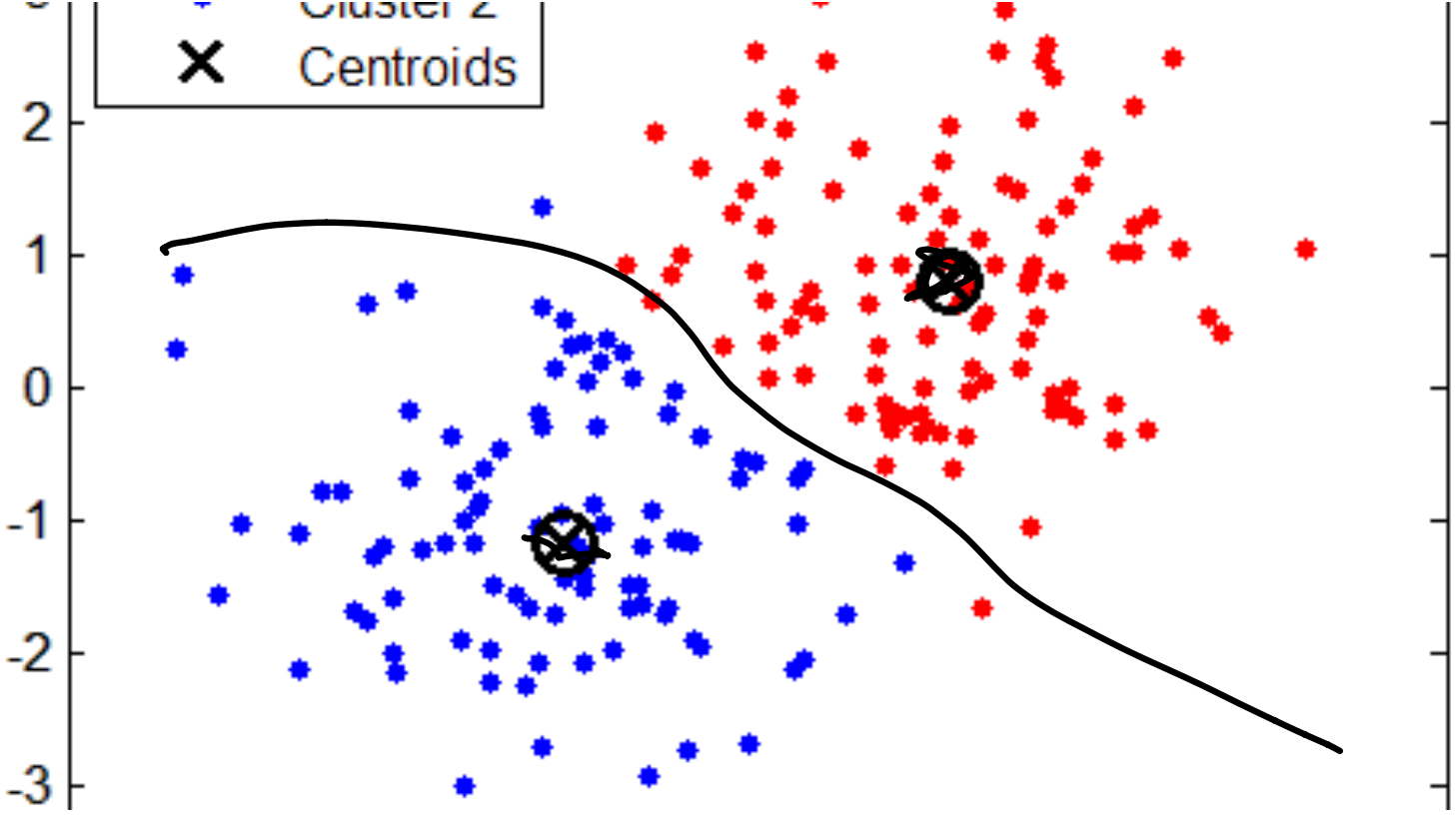
- Unsupervised Learning
  - Here are a bunch of images…..classify them into different object classes
- Supervised Learning
  - Here are a bunch of images and we have already classified them into which ones are cats
  - We will train our algorithm using a subset of data where we will give you the answers
  - Then we will test you on a group were we don't give you the answer

Classification

- Classification – Unsupervised Learning
  - K-means Clustering
  - Hierarchical Clustering

- Classification - Supervised Learning
  - Logistic Regression (2 categories DV)
  - Naïve Bayes
  - Support vector machines

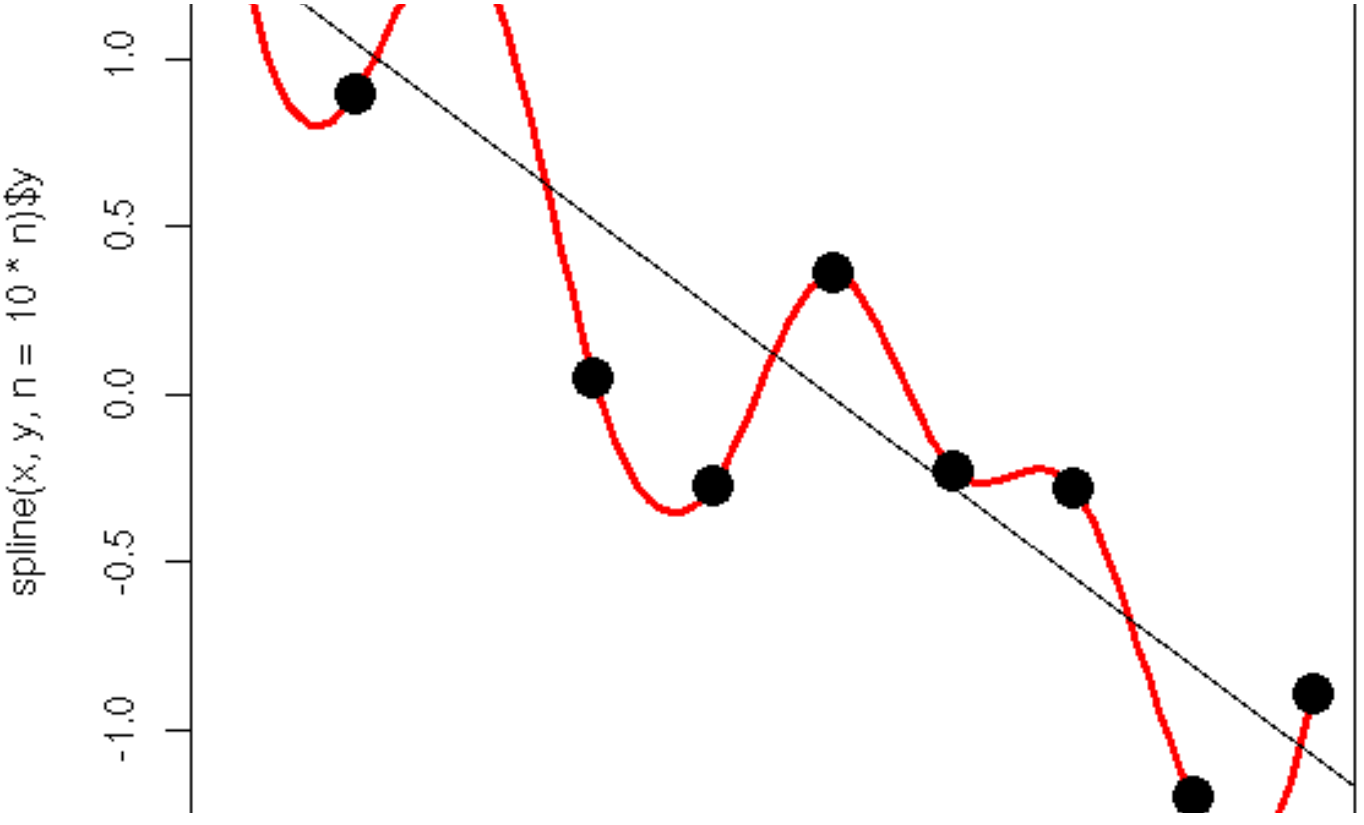
K-Means Clustering



Supervised Learning

- Machine learning focuses on prediction (not understanding relationship
- Cross Validation is necessary to prevent overfitting
    - Training Data – Used to tune the associated algorithm to make accurate predictions
    - Test Data –Used to assess the capabilities of the model

Overfitting



Supervised vs. Unsupervised Learning

Supervised Learning

- In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

Supervised Learning Example

- Categories/Group membership
  - How can we classify individual samples into different class?

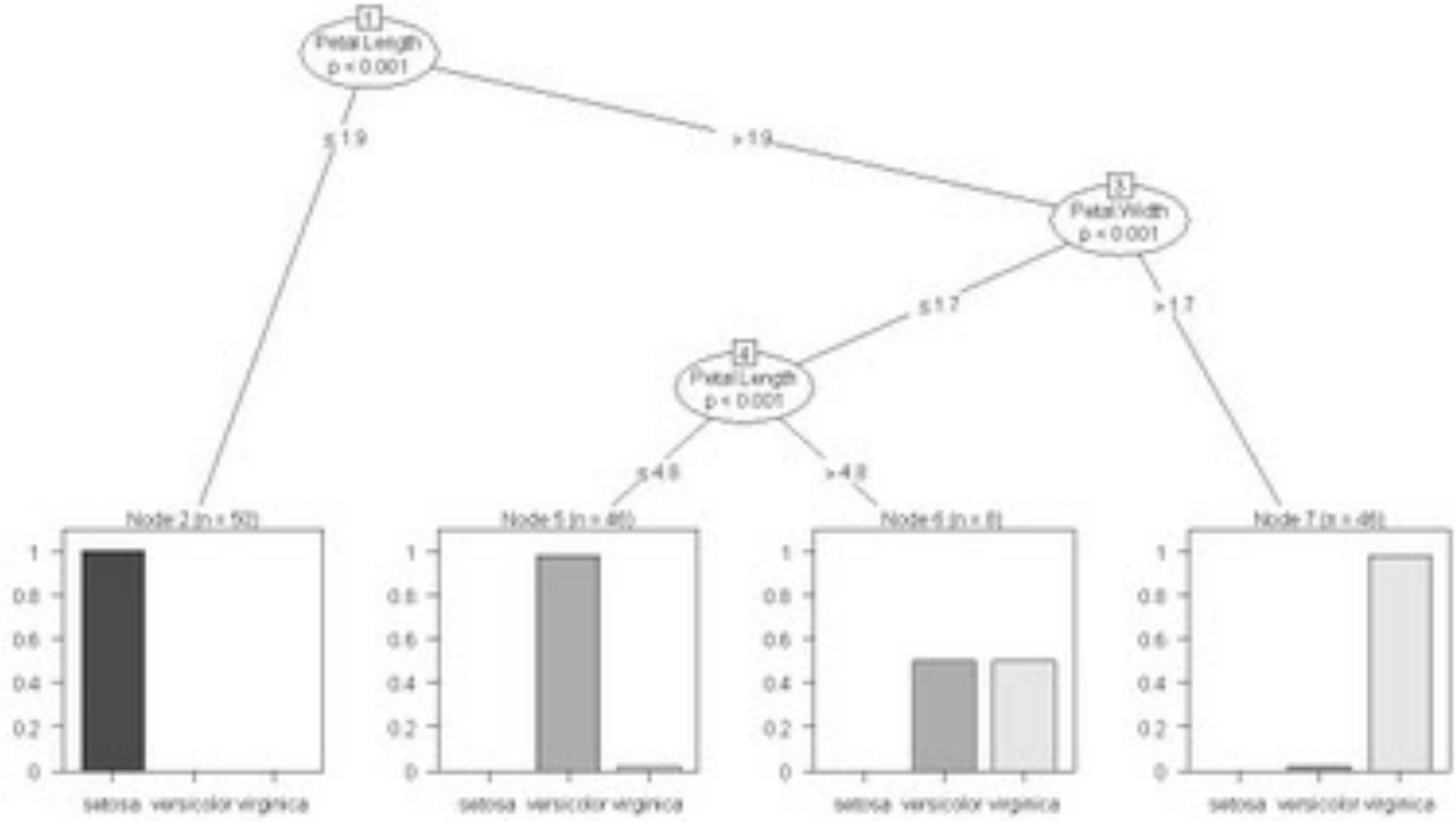
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

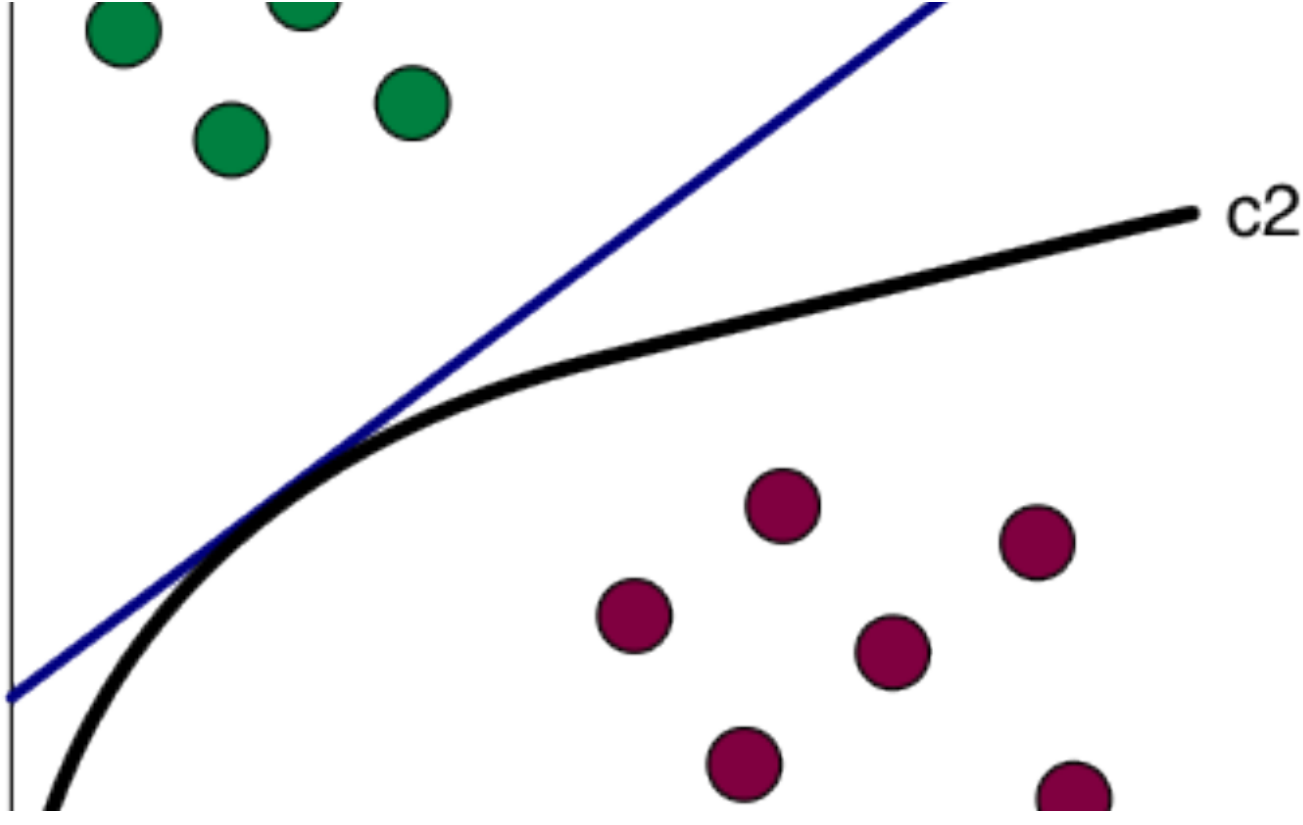
Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	481455

http://archive.ics.uci.edu/ml/datasets/Iris

Decision Tree



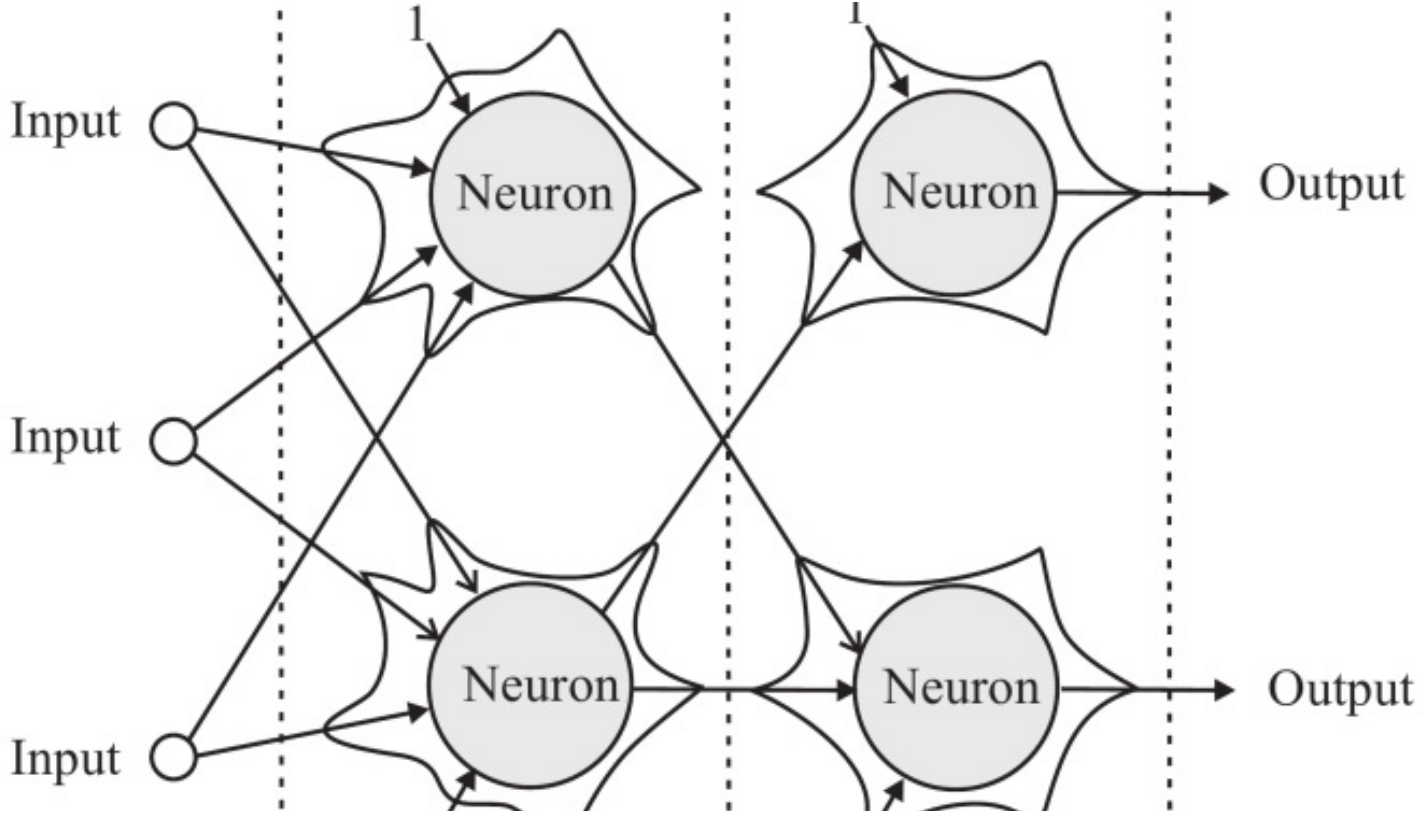
Linear (c1) and Nonlinear (c2) Classification



Neural Networks - Prediction

- artificial neural networks are computational models inspired by animal central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network.

Neural Networks

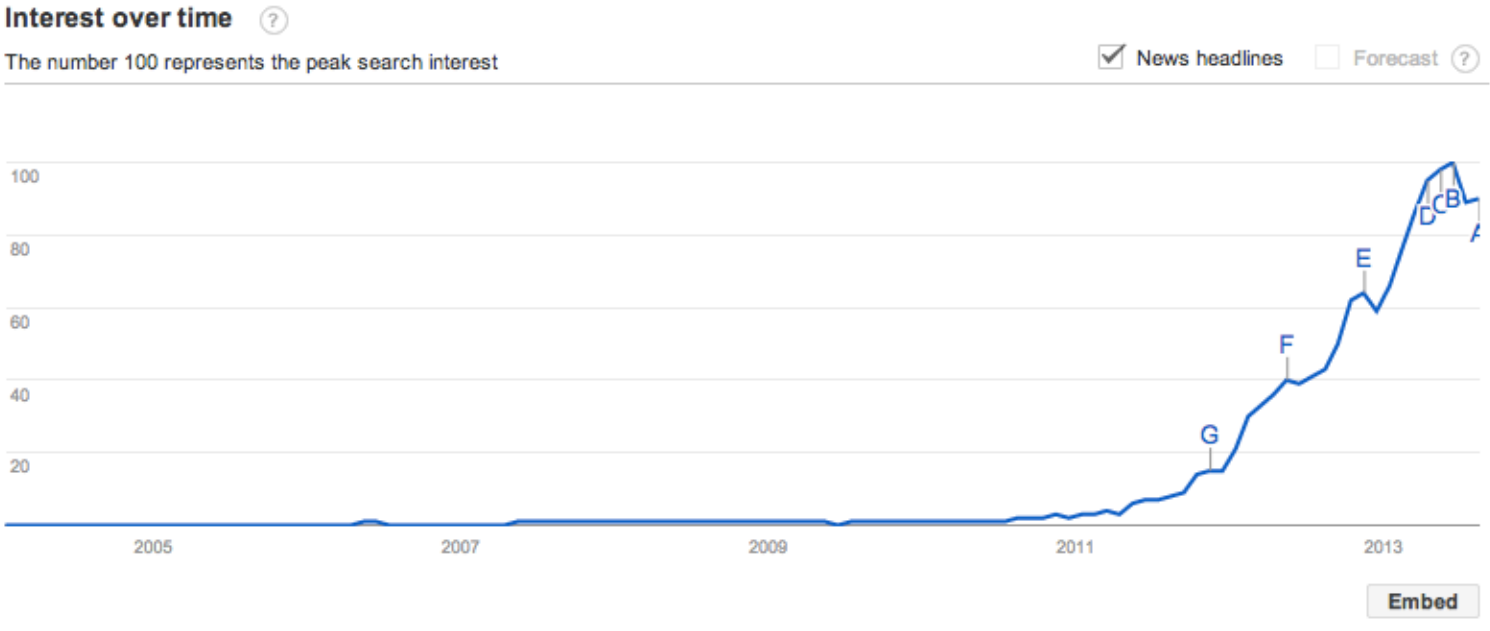


Neural Network vs. Regression

- Neural Networks are likely to provide better predictions than regression based analyses
- However, unlike regression we can't directly understand the "hidden" layer resulting from the analysis

Big Data

What is “big data”?



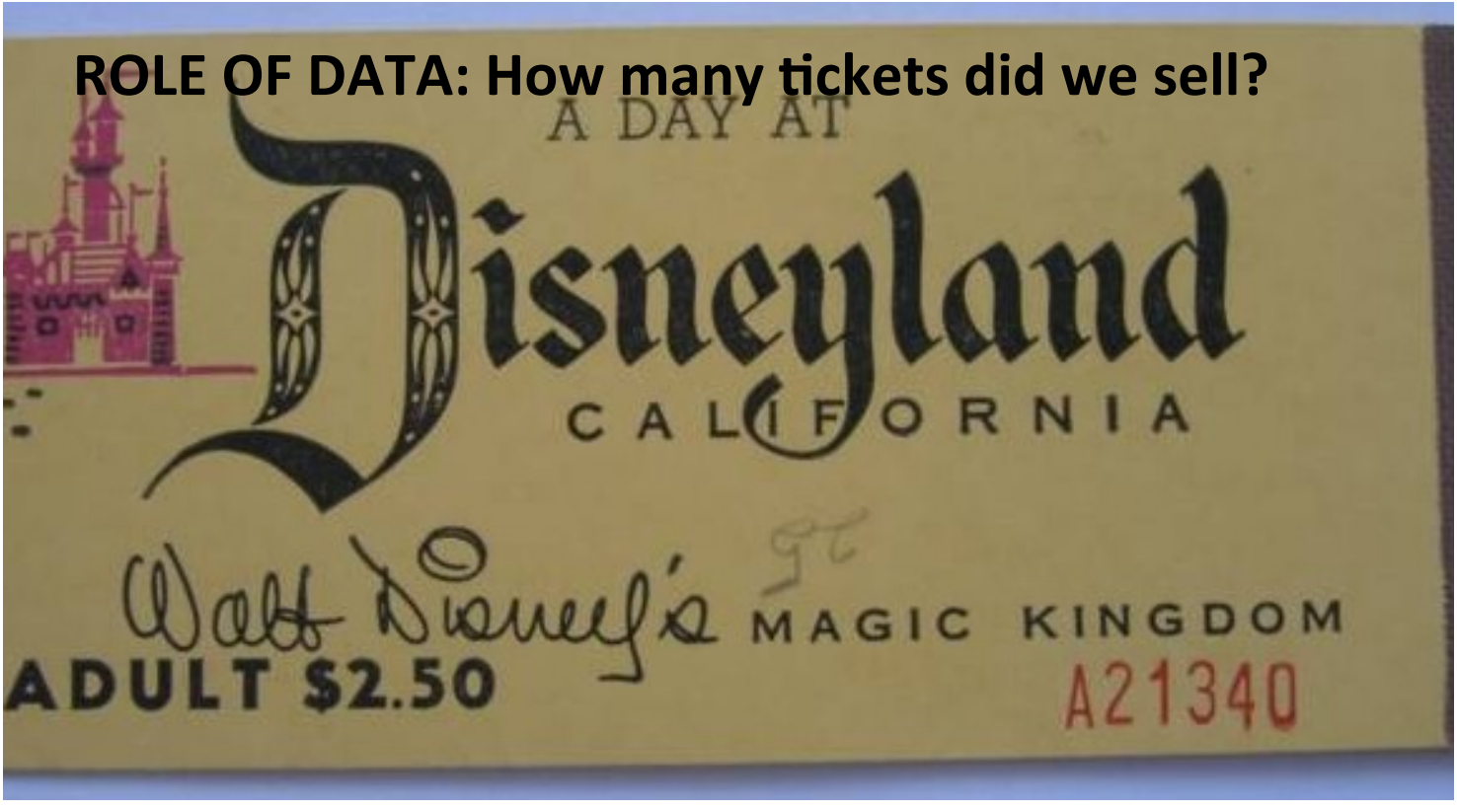
Source:
http://www.google.com/trends/explore?q=analytics#q=%22big%20data%22&cmpt=q

What is “big data”?

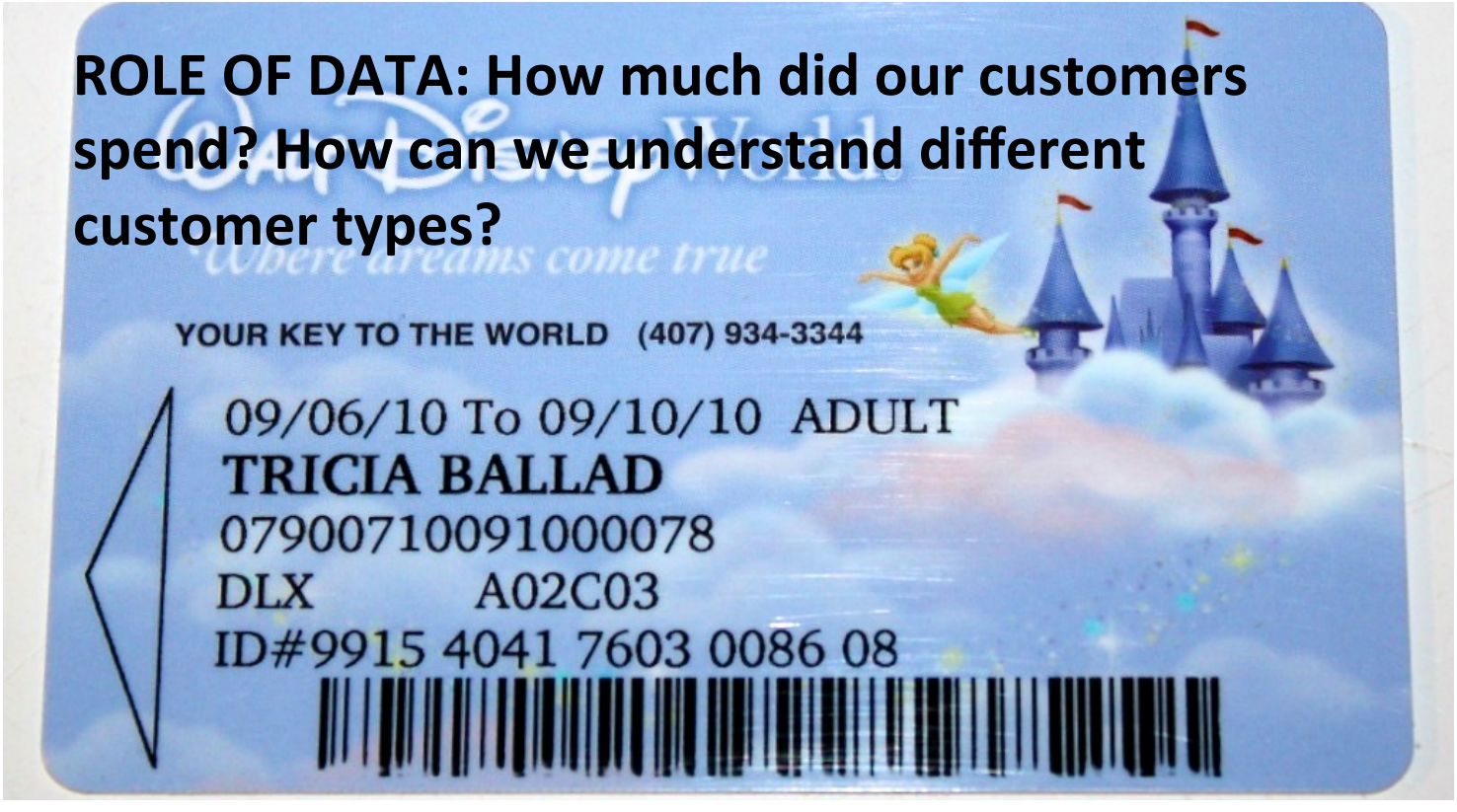
“Big Data”

- The term “Big Data” has been used to capture the data management challenges along three dimensions: volumes, velocity and variety (and sometimes veracity)
- Numerous specific new technology products have been developed to handle the challenges of big data
  - Hadoop
  - Cassandra
  - MongoDB

Disney


ROLE OF DATA: How many tickets did we sell?
A DAY AT
Disneyland
CALIFORNIA
Walt Disney's MAGIC KINGDOM
ADULT $2.50
A21340

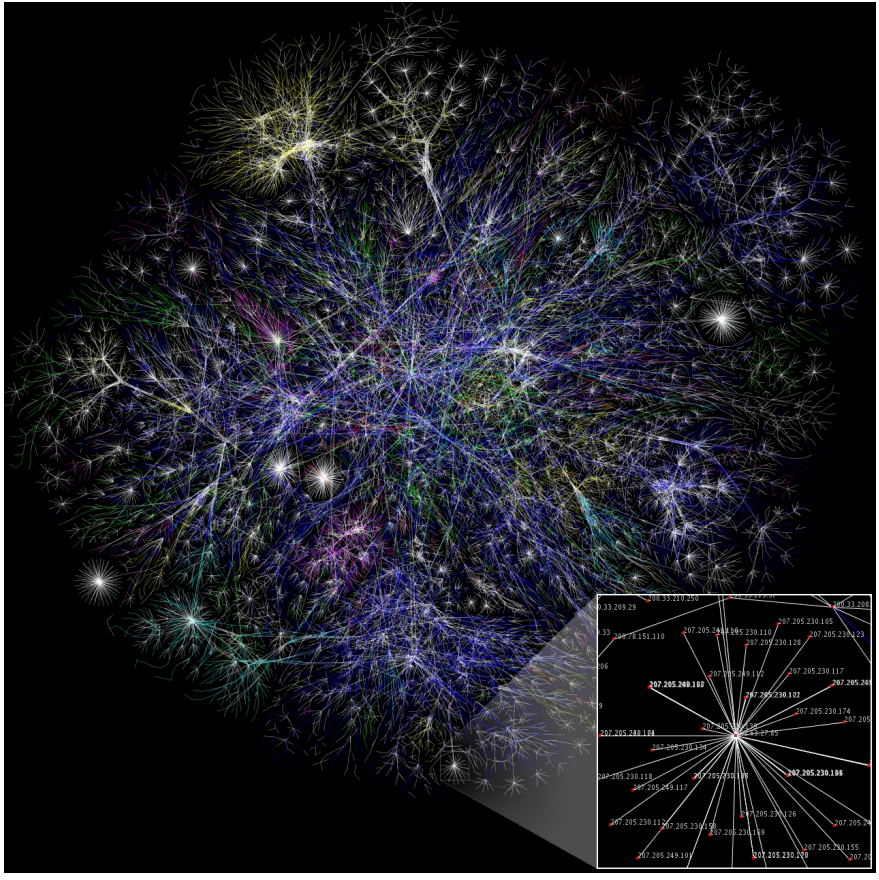
Disney – Data Warehouse Stage



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?**

Web as Information Source



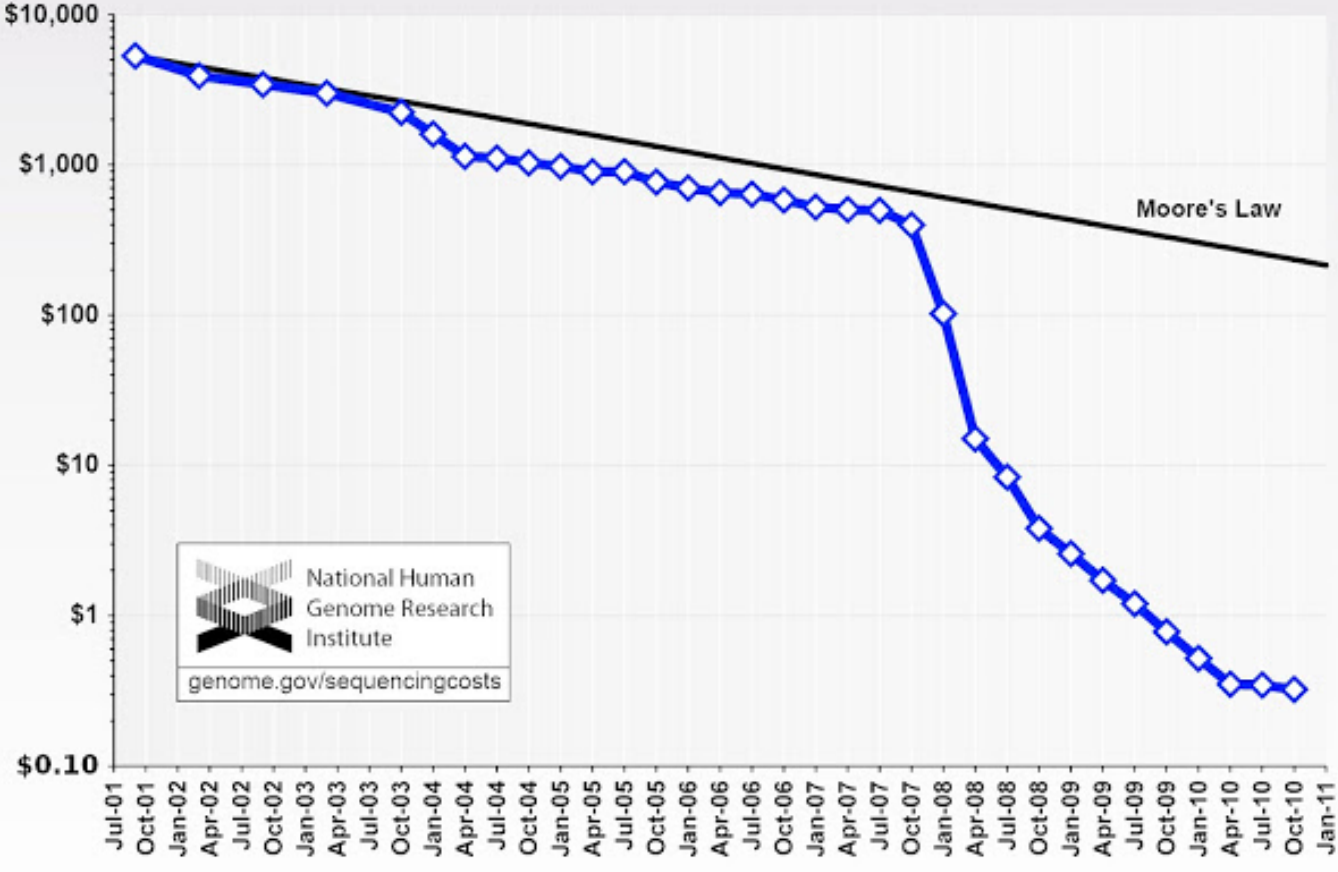
Google processes about 24 petabytes (1PB=1024 TB) of information a day

Approximate size of web 1024 petabytes (1 exobyte)

To think of it easier; if you filled a room that was 8' X 10' X 8' (ceiling) 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".

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

Cost per Megabase of DNA Sequence



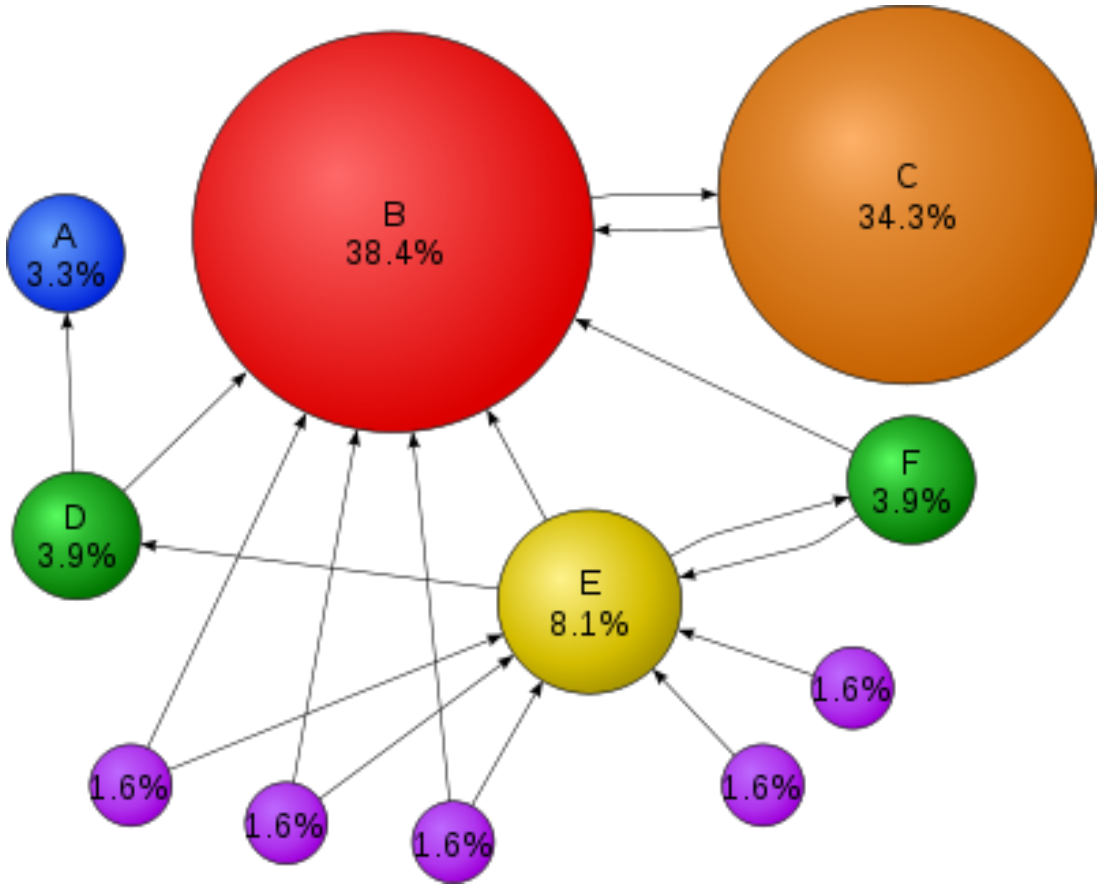
Big Data and Astronomy

The Murchison Widefield Array is the first Square Kilometre Array precursor to enter full operations, generating a vast torrent of information that needs to be stored for later retrieval by researchers.

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

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

Summarization: What is this?

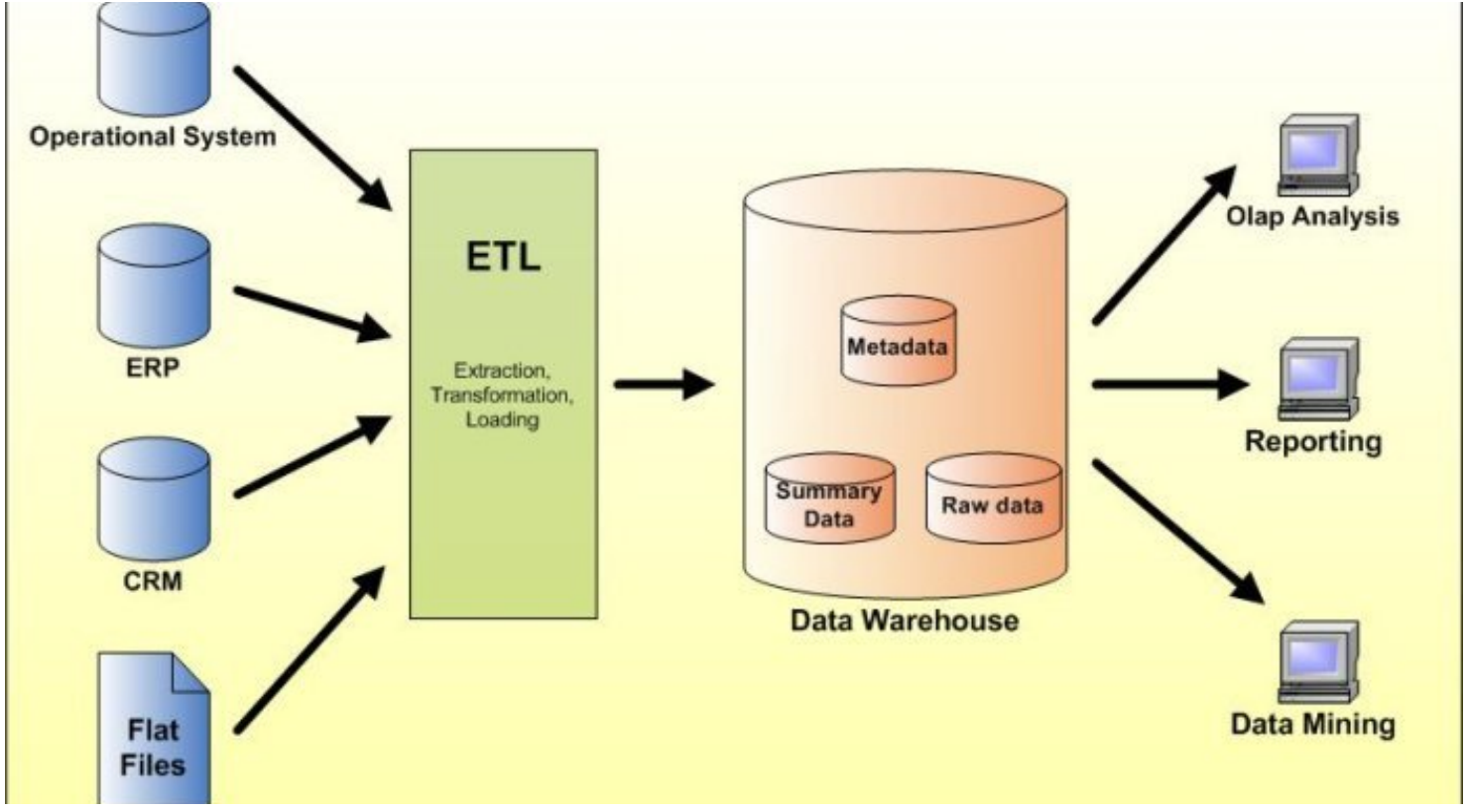


PageRank

- PageRank is an algorithm used by the Google web search engine to rank websites in their search engine results. PageRank was named after Larry Page,[1] one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google:
  - 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.

Big Data and Entrepreneurs

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

Big Data 1.0

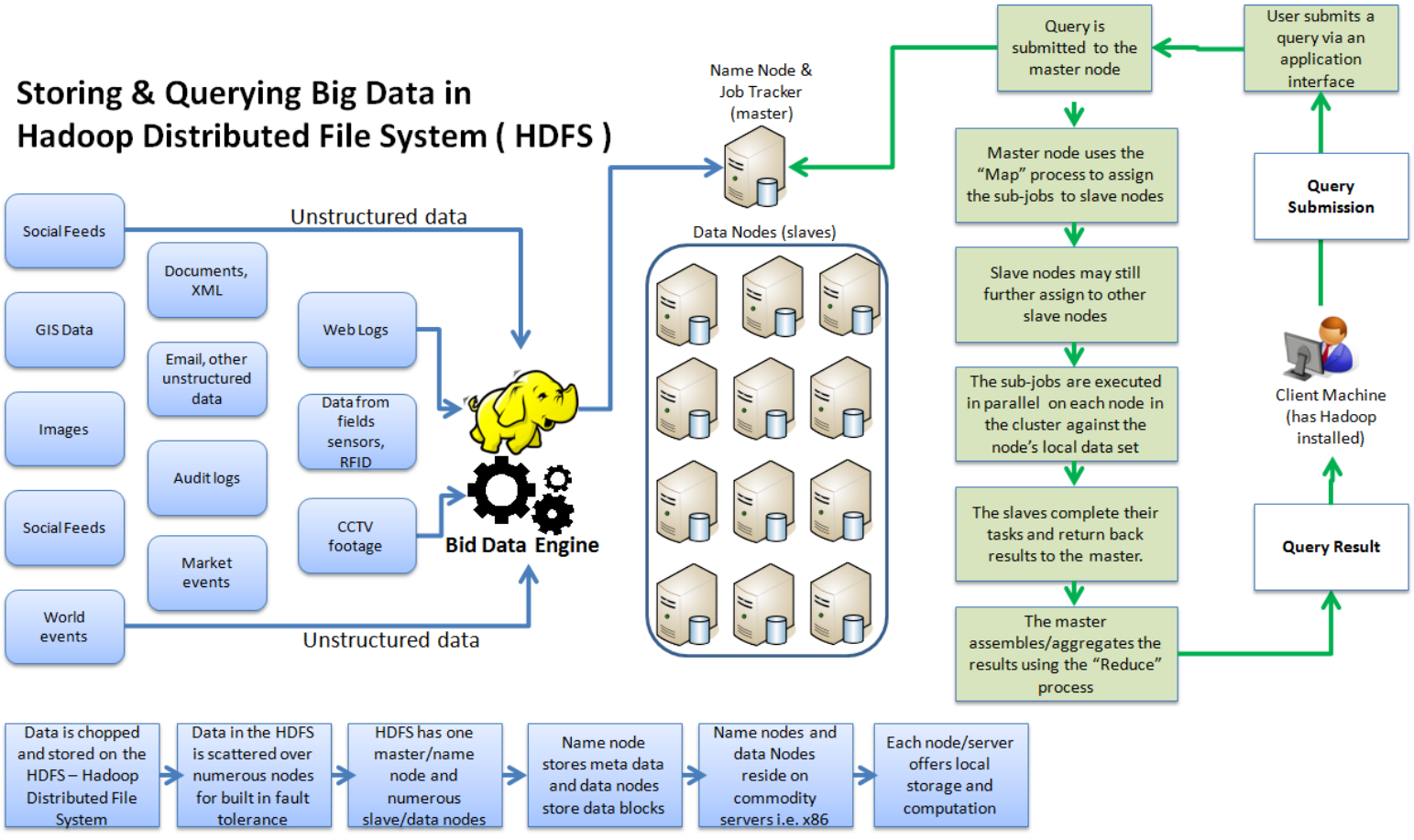
- Online Transaction Processing Systems (OLTP)
  - CRM, ERP, etc
- Extract, Transform, Load (ETL)
  - Extract data from all OLTP systems and put into a Data Warehouse
- Data Warehouse
  - System for creating reports, exploring existing data

Data Warehouse Limitations

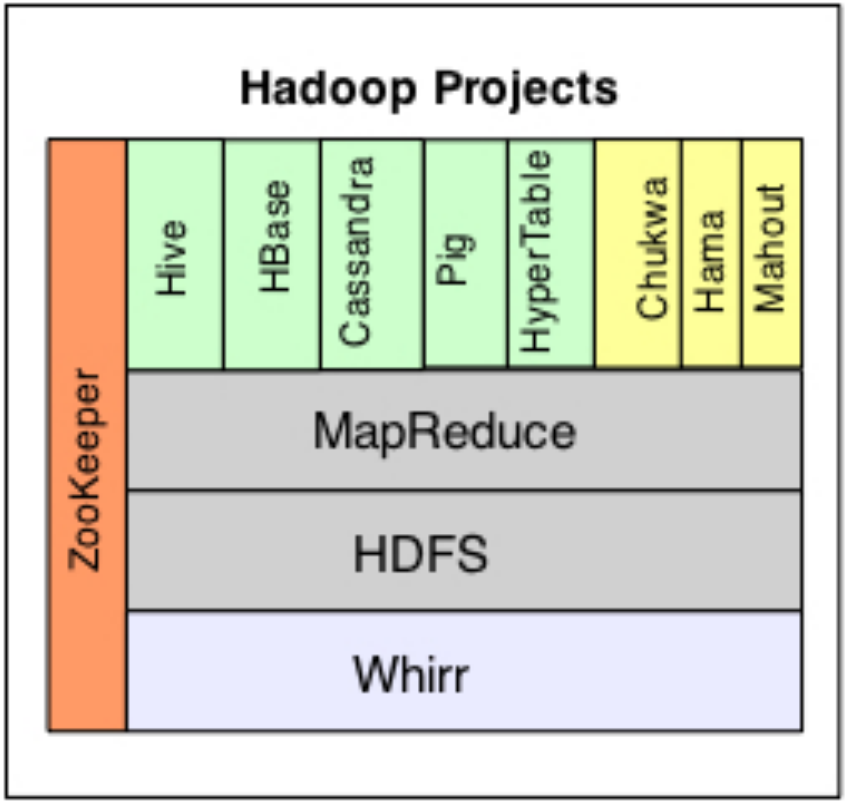
- Not setup for extremely large datasets
  - Weblogs capturing all data
  - Data can be too large to hold in memory for a single location
  - Data from non-transactional sources (Twitter, logs, email) doesn't fit as well with

Big Data 2.0

HADOOP and HDFS



Hadoop Is Really an Ecosystem!



<- Core Analytics

<-Core Data

Hadoop Ecosystem

Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. Pig's language layer currently consists of a textual language called Pig Latin.

The **Apache Hive™** data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL.

The Apache Mahout™ machine learning library's goal is to build scalable machine learning libraries.

Where to go from here…becoming a data science rockstar….

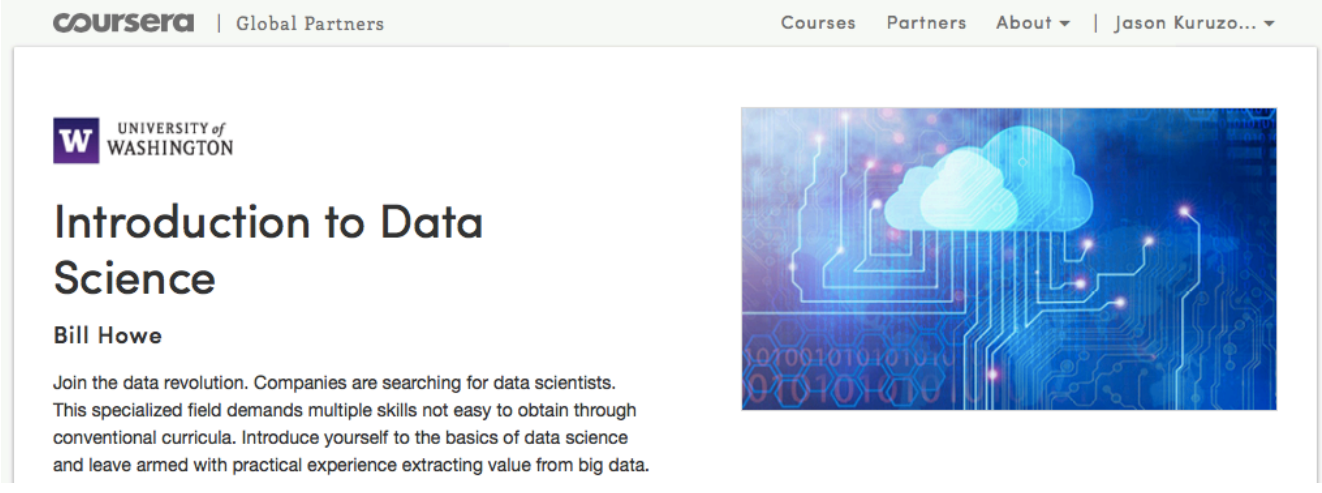
Tremendous Access to Data

- GeoData http://www.factual.com/
- InfoChimps - http://www.infochimps.com
- Government Sources http://data.gov
- APIs
  - Twitter https://dev.twitter.com/
  - Facebook https://developers.facebook.com/docs/reference/apis/
- Structured Data
  - http://www.w3.org/wiki/SparqlEndpoints
- Contests http://www.kaggle.com

Tremendous Access to Training

http://www.ats.ucla.edu/stat/r/

Data Science Overview



https://www.coursera.org/course/datasci

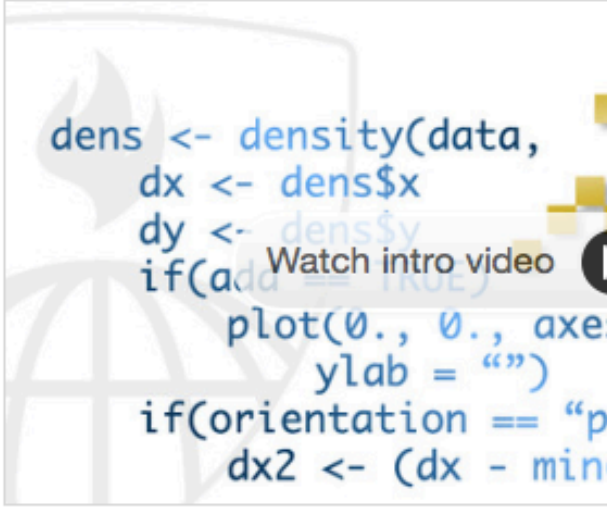
Data Analysis with R

Global Partners Courses Partners About |

INS
OL
H

ing for Data

learning the fundamental computing skills necessary
lysis. You will learn to program in R and to use R for
functions, making informative graphs, and applying
ethods.



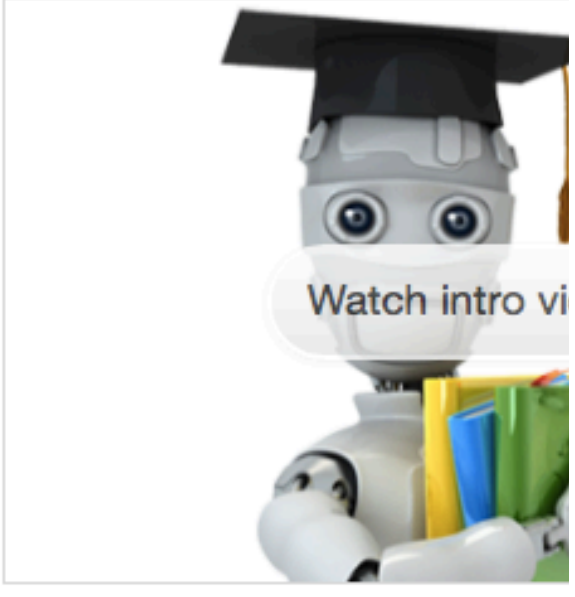
Machine Learning

bal Partners Courses Partners Abo

d

earning

tive machine learning techniques, and gain
n and getting them to work for yourself.



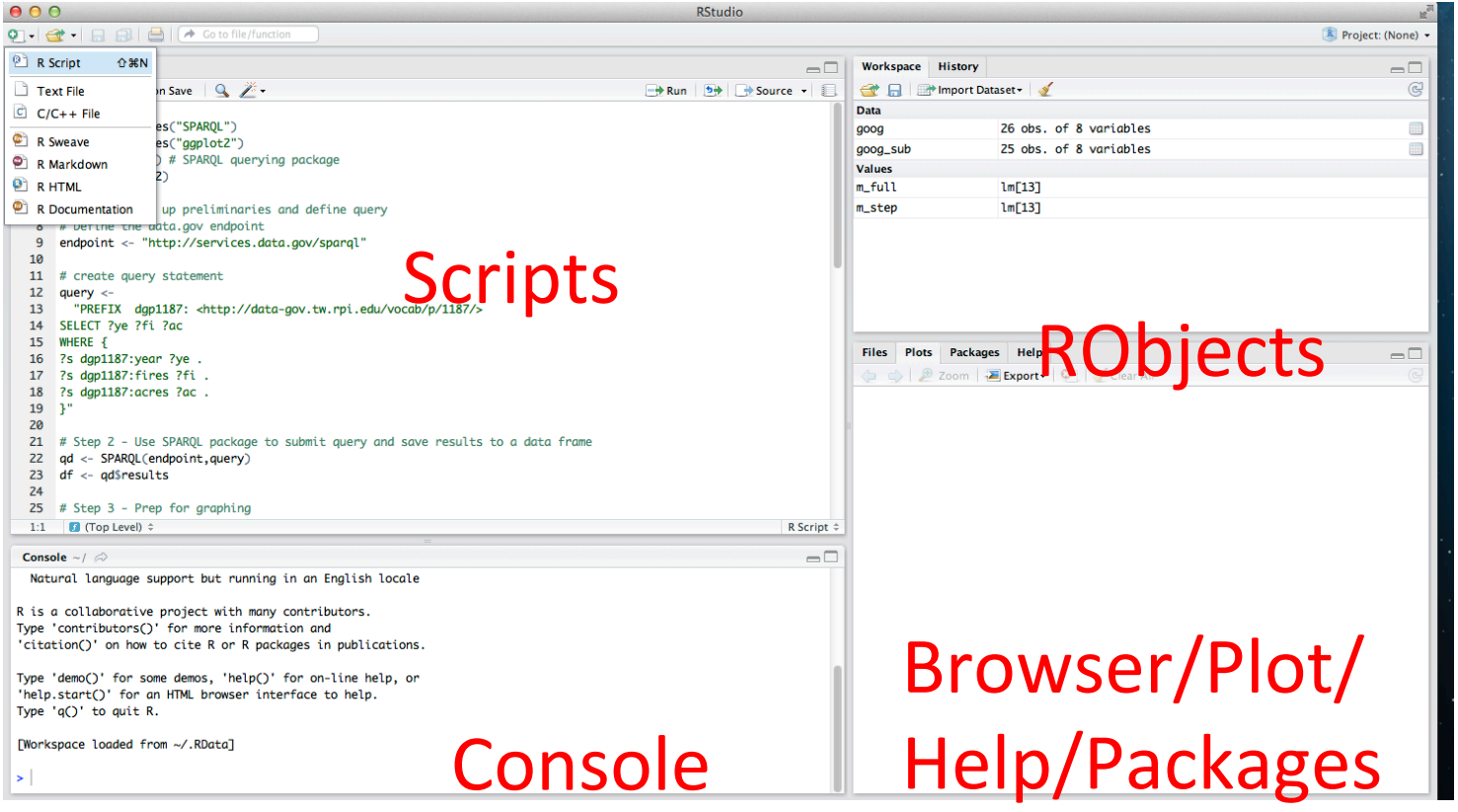
Tips for Learning New Skills

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

Some Software Packages that are Free

- Download R
- Download R Studio
- Find a Kaggle Competition
  - Titanic
- Find Existing Solutions on GitHub
- Diagnose Solutions

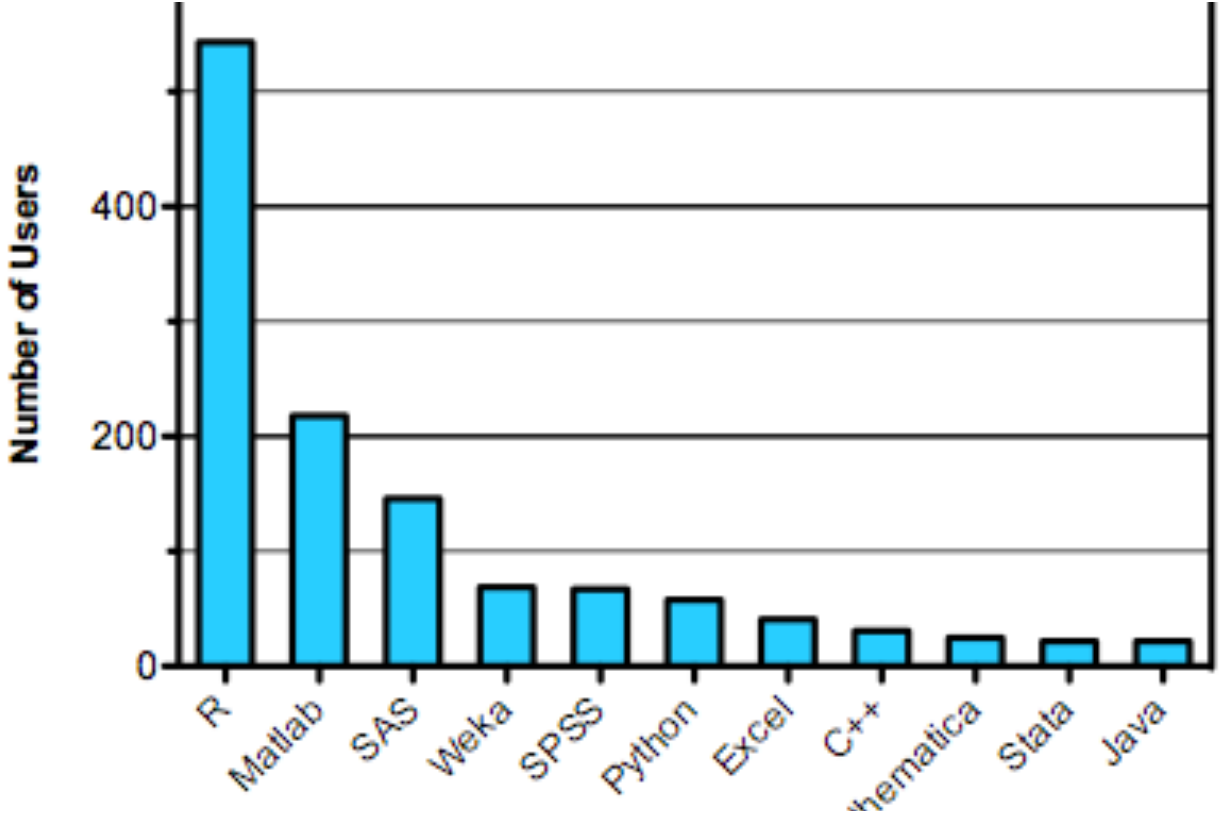
Working with RSTUDIO



Tips

- Always work in a script
- Highlight just a few lines and run those when debugging, writing code
- Double click on objects
- When not sure, try gui menu (stwd) or help(command)

kaggle


Number of Users
400
200
0
R
Matlab
SAS
Weka
SPSS
Python
Excel
C++
Stata
Java

kaggle

Titanic: Machine Learning from Disaster

https://www.kaggle.com/wiki/Tutorials

11 months to go

Friday, September 28, 2012

Knowledge • 7,181 teams

Competition Details » Get the Data » Make a submis

Predict survival on the Titanic (with tutorials in
Python and an introduction to Random Forests

Titanic – The Data

VARIABLE DESCRIPTIONS:

- survival Survival (0 = No; 1 = Yes)
- pclass Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- name Name
- sex Sex
- age Age
- sibsp Number of Siblings/Spouses Aboard
- parch Number of Parents/Children Aboard
- ticket Ticket Number
- fare Passenger Fare
- cabin Cabin
- embarked Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Titanic – The Process

1. Identify cross validation dataset (separate into test and train)
2. Create features which can be used to predict the outcomes
3. Select different features and models to see how effectively outcomes can be predicted

Titanic

Python

- http://nbviewer.ipython.org/urls/raw.github.com/agconti/kaggle-titanic/master/Titanic.ipynb

R

https://www.kaggle.com/c/titanic-gettingStarted/forums/t/3702/basic-r-code

RPI – Data Science Rockstars

[ProtoML (An RCOS Team)]

- [http://rcos.rpi.edu/projects/protoml/]

Leaderboard

1. ProtoML
2. jarfo
3. HiDLoN
4. FirfiD
5. mouse
6. Domcastro & Sayani

267 teams

4578 entries

Given samples from a pair of variables A, B, find whether A is a cause of B.

Come to our NIPS workshop (dec 9 or 10 in Tahoe).

$10,000 • 267 teams

http://www.kaggle.com/c/cause-effect-pairs

RPI

The Rensselaer IDEA: Harnessing the Power of Data to Change the World

The Rensselaer Institute for Data Exploration and Applications (IDEA) Anchors a New Era of Research and Discovery at the Nation's Oldest Technological Research University

An abundance of data in organizations is creating a high demand for professionals with the skills to harness data to gain insights and drive competitive actions in businesses. The M.S. in Business Analytics provides students with the knowledge and essential skills needed to respond to the new challenges characteristic of the data intensive, decision-making environments in the world today.

Where is “Data Science” happening in the Wild?

- Find something!

Software

- Install
  - RSTUDIO
  - MAMP/WAMP

Next Time

- Relational Algebra (Wikipedia Entry) [link]
- SQL (Wikipedia Entry) [link]
- Relational Algebra I Video [link]
- Relational Algebra II Video [link]
- Introduction to SQL Video [link]
- MySQL Tutorial [link]
- Chapters 1&2 in Provost and Fawcett

[LAB 2 SQL and Movies]