GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

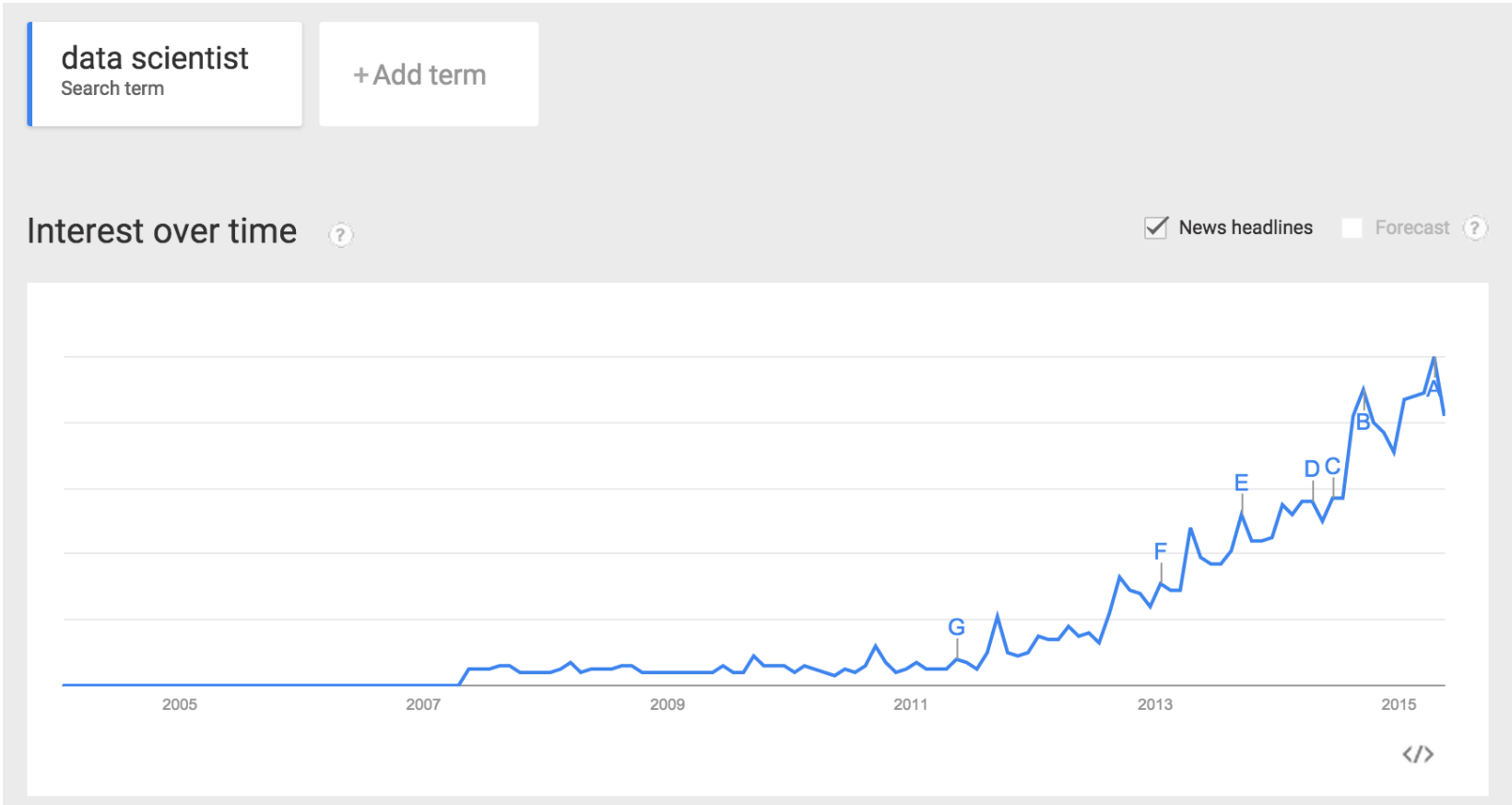
INTRO TO DATA SCIENCE

- What is Data Science?
- The Data Science Workflow
- Machine Learning
- Recommendation Systems & Case Studies
- Wrap Up

WHAT IS DATA SCIENCE?

… and what is it not?


Job Trends from Indeed.com
Data-scientist
Percentage of Matching Job Postings
0.04
0.03
0.02
0.01
0
Jan '06
Jan '07
Jan '08
Jan '09
Jan '10
Jan '11
Jan '12
Jan '13
Jan '14
Jan '15
indeed
one search. all jobs.


data scientist
Search term
+ Add term
Interest over time
News headlines
Forecast
A
B
C
D
E
F
G
2005
2007
2009
2011
2013
2015

DATA SCIENCE IS THE EXTRACTION OF KNOWLEDGE FROM BIG DATA.

DATA SCIENCE IS THE EXTRACTION OF KNOWLEDGE FROM BIG DATA.

SOMEWHAT TRUE

DATA SCIENCE IS THE EXTRACTION OF KNOWLEDGE FROM ~~BIG~~ DATA.

THAT'S BETTER!

- Scaling to big data is often desired, but it is not required.
- You can think of data science as being a new set of tools and techniques.

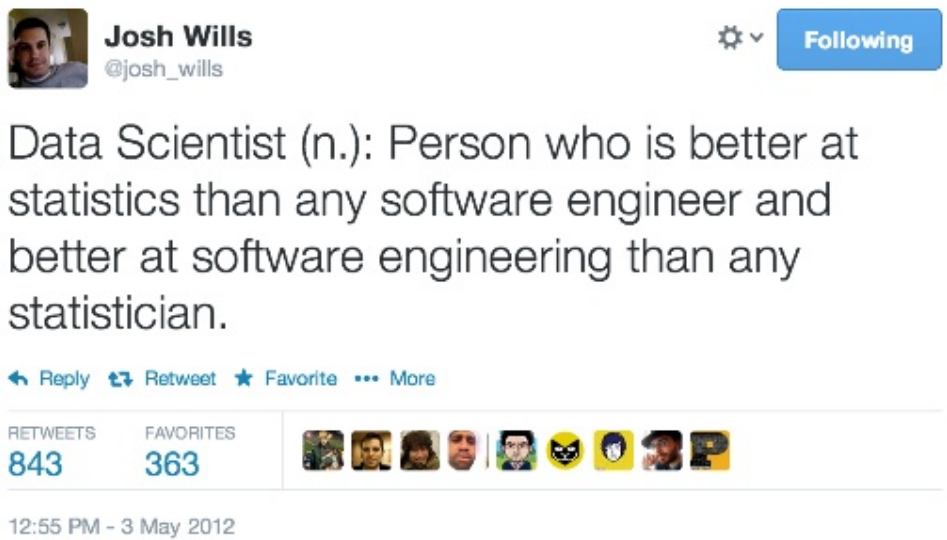
DATA SCIENCE IS JUST A BUZZWORD FOR STATISTICS.

DATA SCIENCE IS JUST A BUZZWORD FOR STATISTICS.

FALSE*!

- Data Science encompasses more than what most statisticians study
- Data Scientists generally care more about accuracy of a model than its interpretability

* Not all agree, e.g. http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf


Josh Wills
@josh_wills
Following
Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.
Reply
Retweet
Favorite
More
RETWEETS
843
FAVORITES
363
12:55 PM - 3 May 2012

DATA SCIENCE VS STATISTICS

Data modeling culture. Assumes the data is generated by an underlying stochastic model. **Interpretability** is essential. (98% of statistitians*)

Algorithmic modeling. The data cannot be characterized by a simple model. **Accuracy of predictions** is emphasized over understanding the underlying system. (2% of statisticians*, but most data scientists)

* Breiman, Leo. "Statistical Modeling: The Two Cultures". Statistical Science. 2001, Vol. 16, No. 3, 199-231. [https://projecteuclid.org/download/pdf_1/euclid.ss/1009213726]

DATA SCIENCE VS MACHINE LEARNING

Data Science includes the application of machine learning methods as part of a broad **process of data collection, data munging, prediction, and presentation** of results.

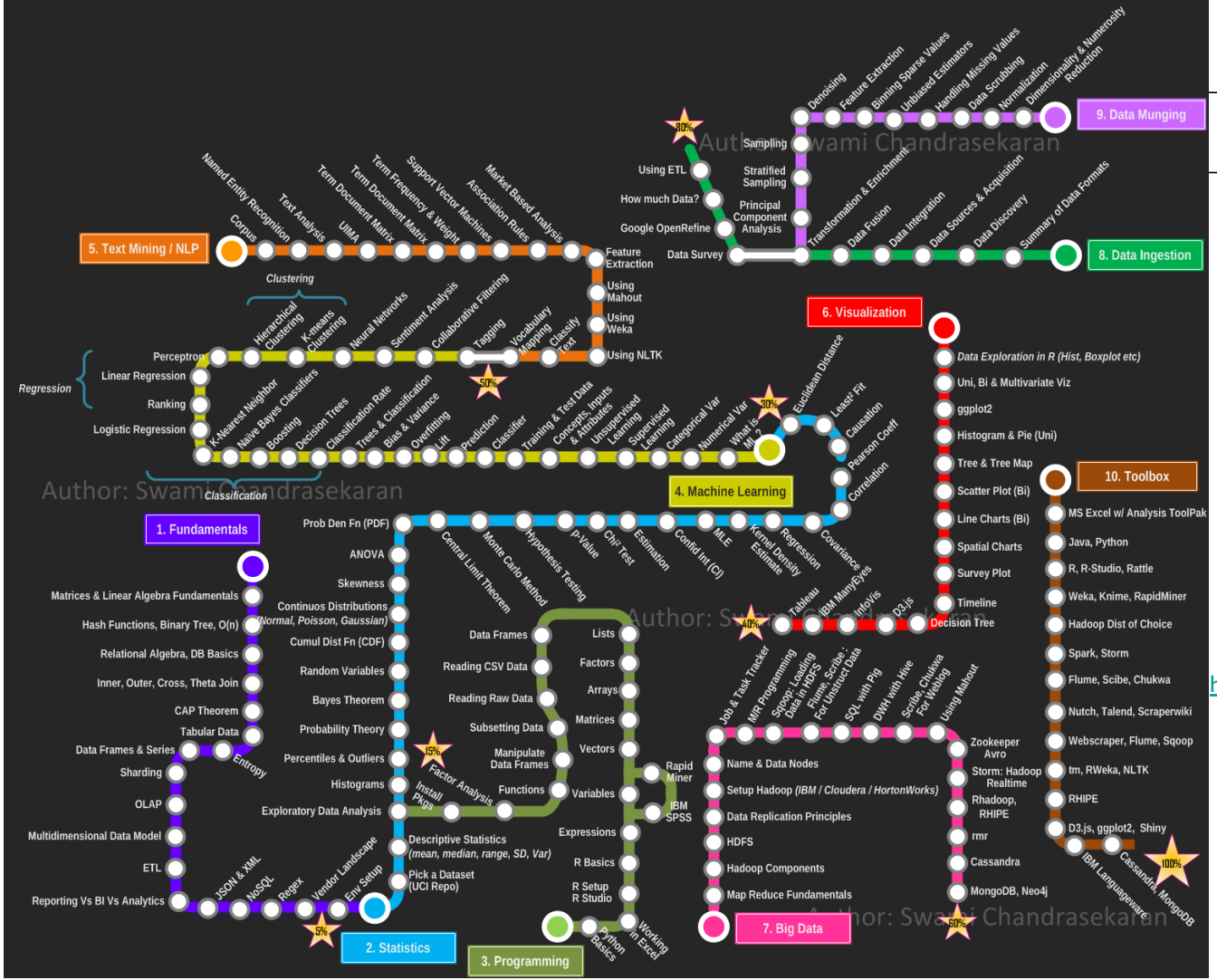
Machine Learning is "**the construction and study of algorithms** that can learn from and make predictions on data." (Wikipedia)

DATA SCIENCE VS SOFTWARE ENGINEERING

Data science uses programming to acquire and explore data. Oftentimes, the software is written for one-time use and is not maintained or intended for general usage. The goal of data science is to predict and present data, rather than to create software.

Software engineering is the study of the design, development, and maintenance of software.

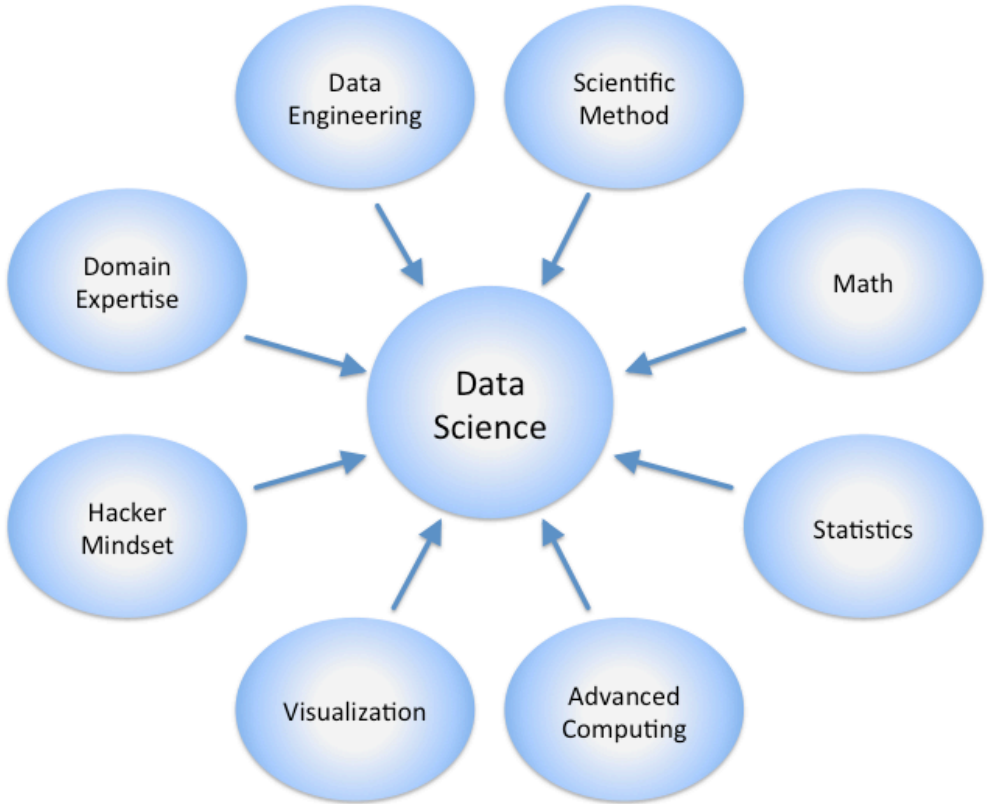
- Mathematics
- Statistics
- Information Theory
- Information Technology
- Signal Processing
- Probability Models
- Machine Learning
- Statistical Learning
- Computer Programming
- Data Engineering
- Pattern Recognition
- Visualization
- Predictive Analytics
- Uncertainty Modeling
- Data Warehousing
- Data Compression
- High Performance Computing
- Computer Vision
- Information Retrieval
- Natural Language Processing



Source:
Swami Chandrasekaran

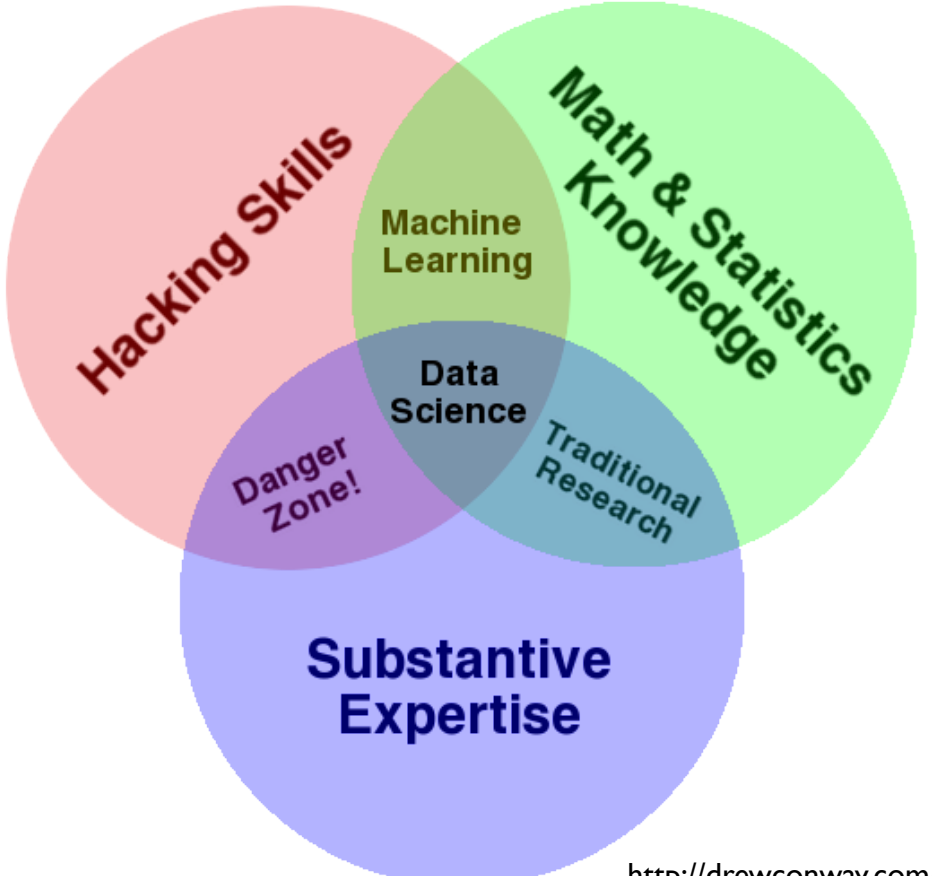
http://nirvacana.com/thoughts/becoming-a-data-scientist/

THE DATA SCIENCE SKILLSET

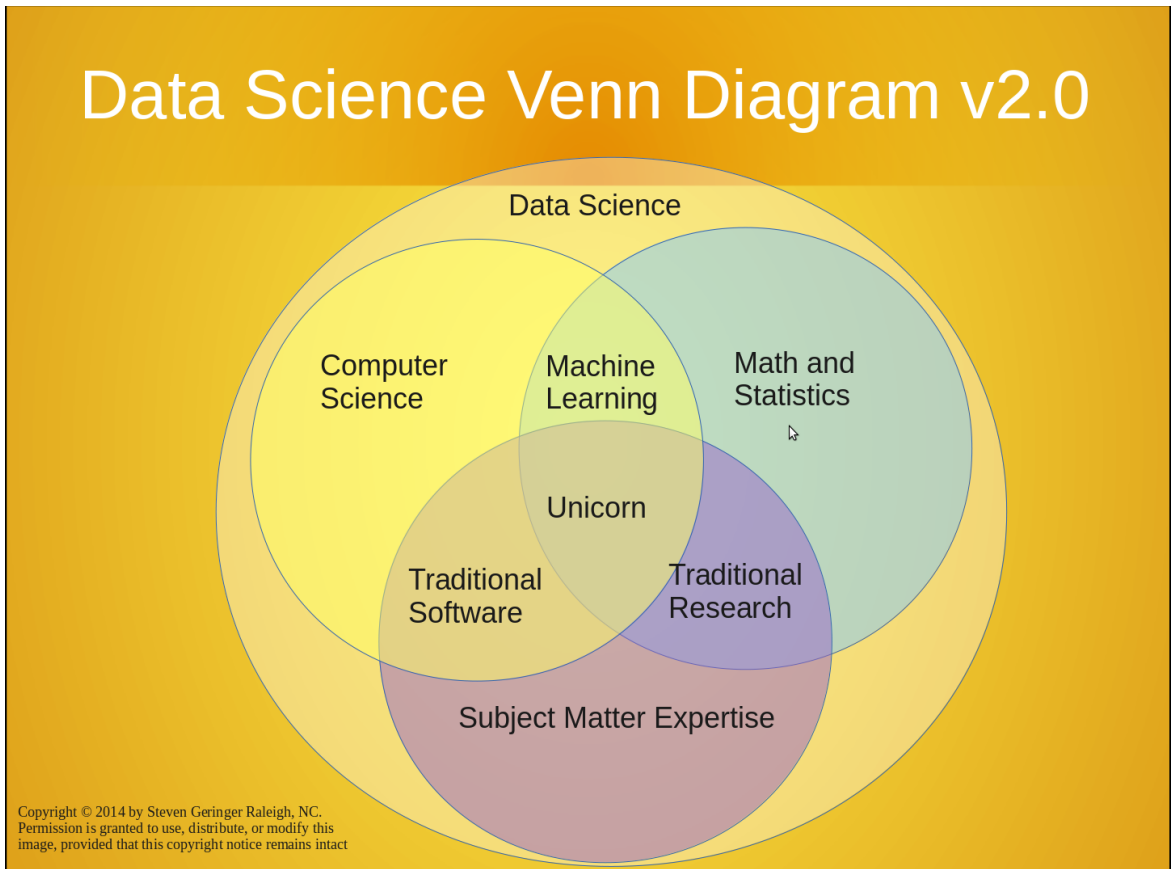


http://upload.wikimedia.org/wikipedia/commons/4/44/DataScienceDisciplines.png

THE DATA SCIENCE SKILLSET



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


Data Science Venn Diagram v2.0
Data Science
Computer Science
Machine Learning
Math and Statistics
Unicorn
Traditional Software
Traditional Research
Subject Matter Expertise
Copyright © 2014 by Steven Geringer Raleigh, NC. Permission is granted to use, distribute, or modify this image, provided that this copyright notice remains intact

Data Businessperson

Business person, Leader, Entrepreneur

Data Creative

Artist, Jack-of-All-Trades, Hacker

Data Researcher

Scientist, Researcher, Statistician

Data Engineer

Engineer, Developer

Source: 2012 survey of Data Community DC members

THE DATA SCIENCE WORKFLOW

MOST OF A DATA SCIENTIST'S TIME IS TYPICALLY SPENT ANALYZING DATA AND MAKING MODELS.

MOST OF A DATA SCIENTIST'S TIME IS TYPICALLY SPENT ANALYZING DATA AND MAKING MODELS.

FALSE

- Most time is typically spent collecting and cleaning hard-to-get, imperfect data!

DATA SCIENCE IS MORE OF AN ART THAN A SCIENCE

DATA SCIENCE IS MORE OF AN ART THAN A SCIENCE

TRUE

- Each problem requires different methods for gathering data, cleaning data, deciding on models, and assessing models.

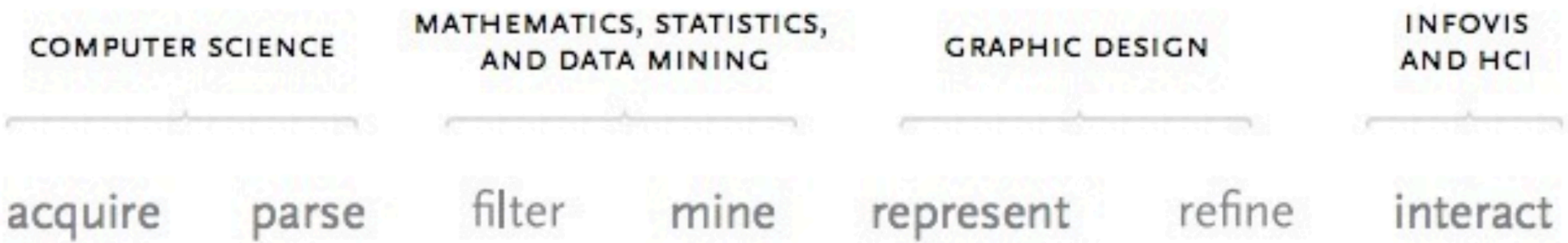
- Creativity, curiosity, and perseverance are essential to finding a good model.

PROBLEM: HOW CAN WE ESTIMATE THE CHANCE THAT A USER WILL MAKE A PURCHASE ON OUR WEBSITE?

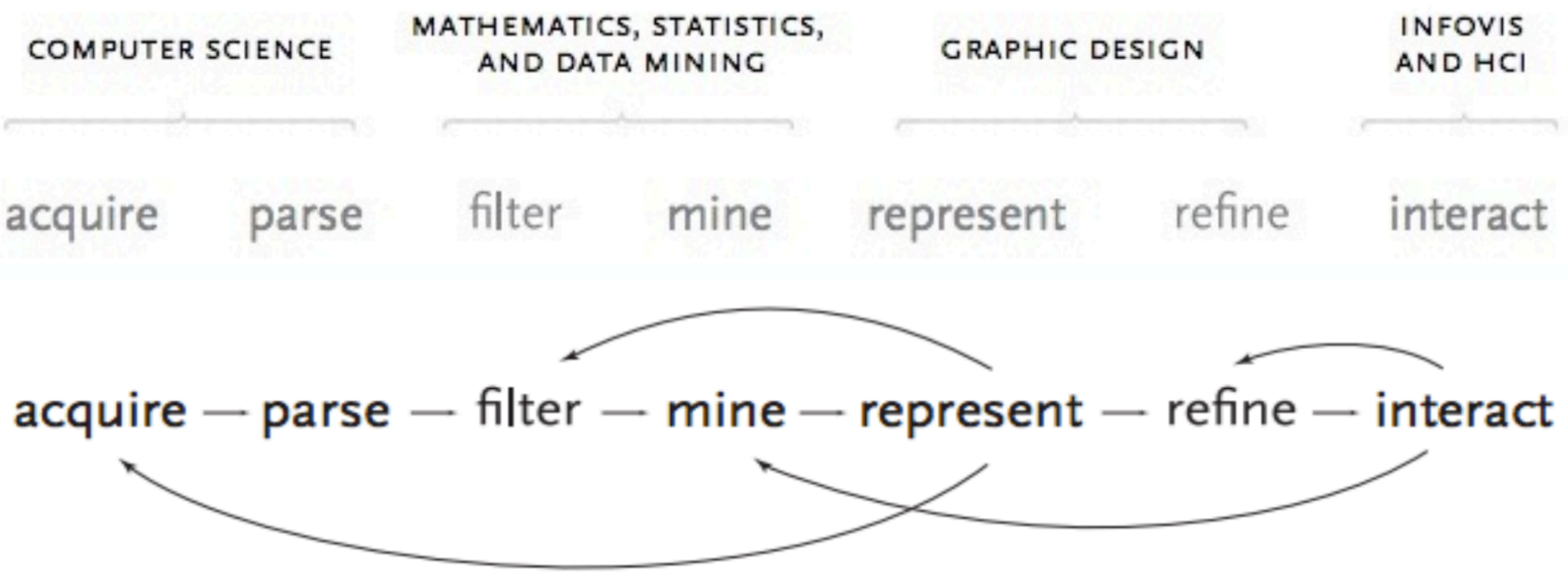
- First, you get the data in a form that you can work with …
- Second, you plot the data to get a feel for what's going on …
- Third, you iterate between graphics and models to build a succinct quantitative summary of the data …
- Finally, you look back at what you have done, and contemplate what tools you need to do better in the future.

http://had.co.nz/thesis/practical-tools-hadley-wickham.pdf

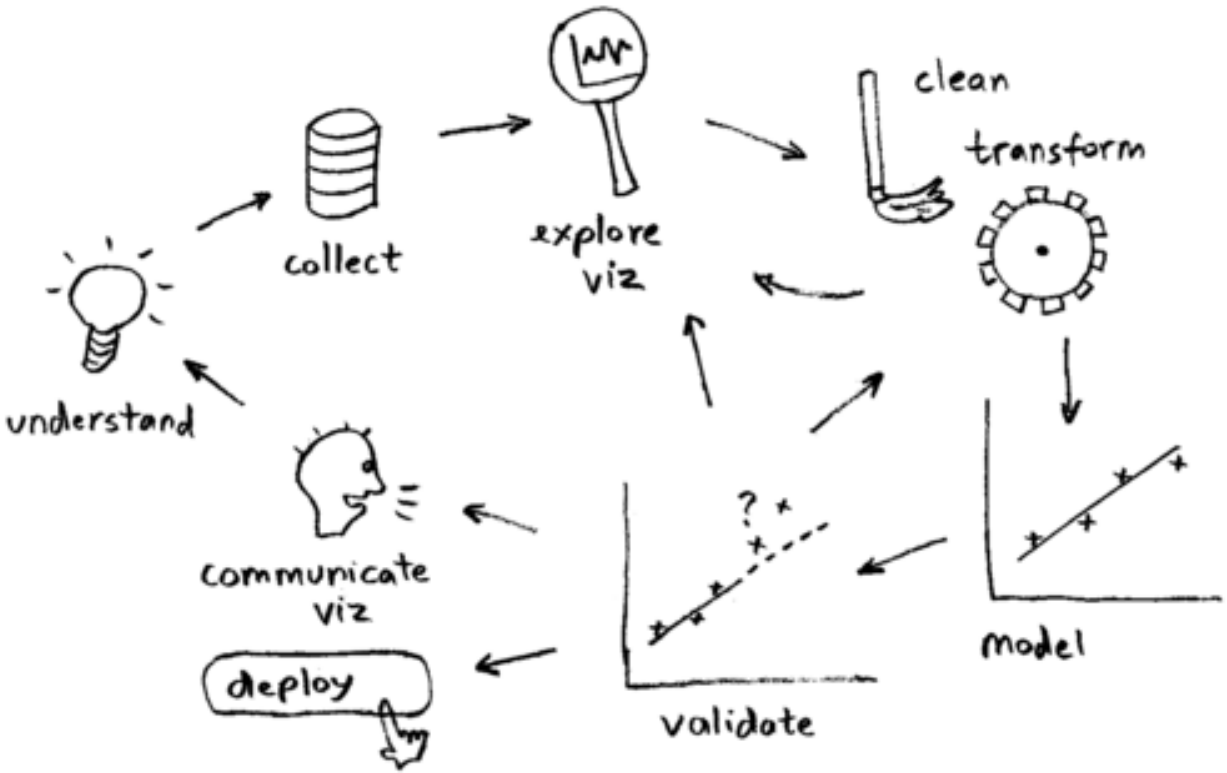
THE DATA SCIENCE WORKFLOW



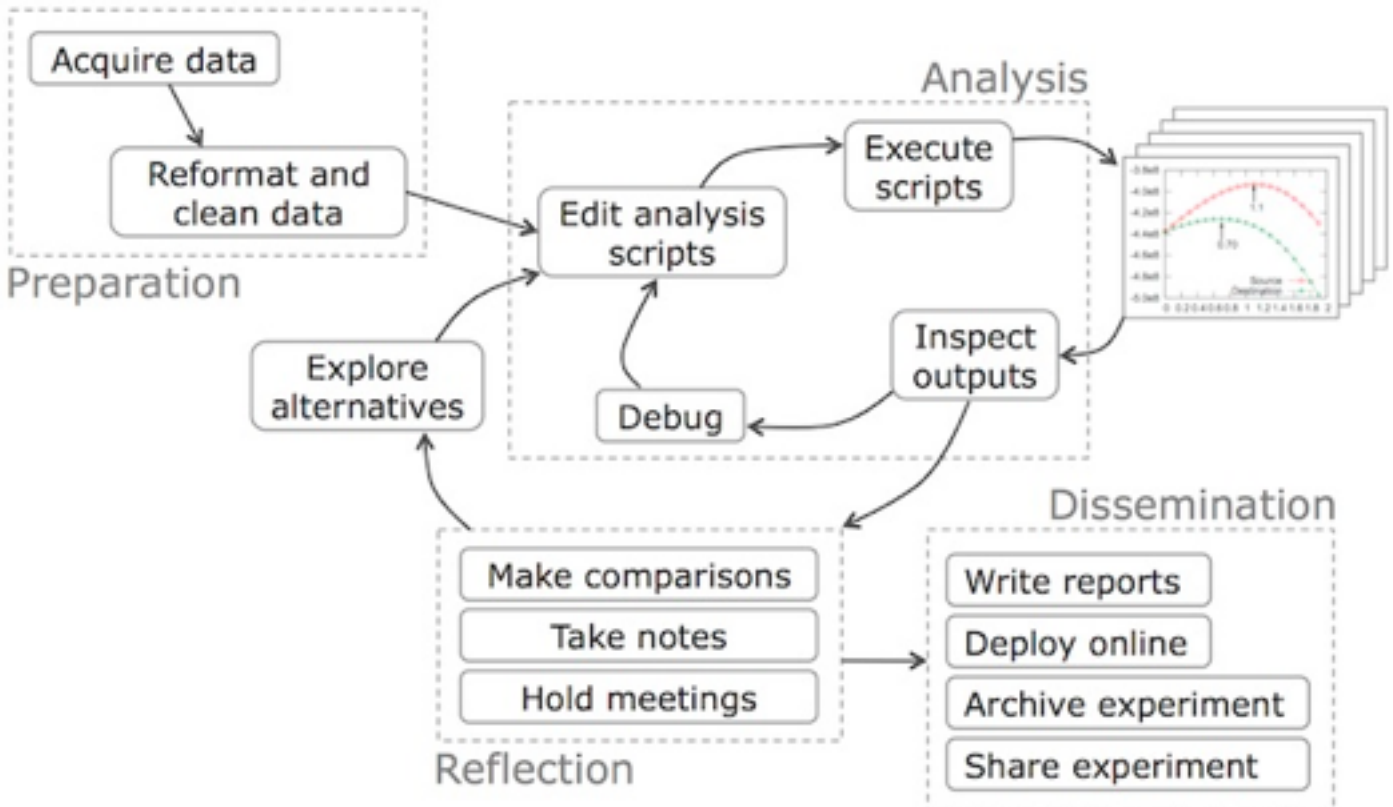
http://benfry.com/phd/dissertation-110323c.pdf



http://benfry.com/phd/dissertation-110323c.pdf

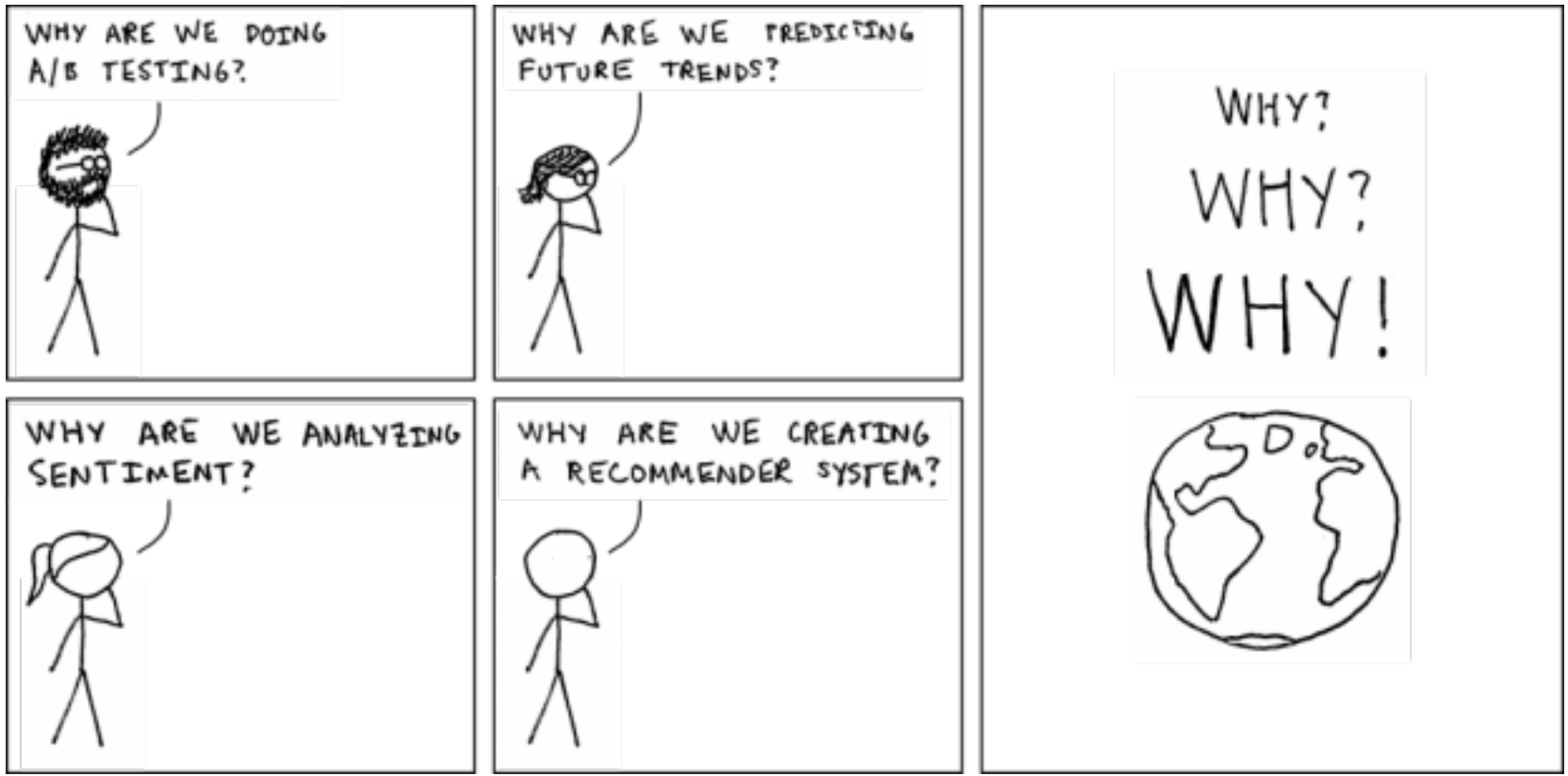


http://datascience.la/data-science-toolbox-survey-results-surprise-r-and-python-win/



source: Phillip Guo - BLOG@CACM Data Science Workflow: Overview and Challenge

THE FUN PART IS MODELING



http://xkcd.com

EXCEL IS A DATA SCIENTIST'S WEAPON OF CHOICE.

EXCEL IS A DATA SCIENTIST'S WEAPON OF CHOICE.

FALSE

- According to an informal survey, only ~10% of Data Scientists claim to use primarily use Excel – almost all use R and Python as their main analysis/visualization tool.

http://datascience.la/data-science-toolbox-survey-results-surprise-r-and-python-win/

Big Data Manipulation

- Hadoop
- Pig
- Hive
- Python

Statistical Analysis

- SAS
- SPSS
- R
- Tableau

Query & Reporting

- SQL
- Business Objects
- Cognos

Data Warehousing & Loading

- Teradata
- Informatica

Small-Scale Reporting & Financial Analysis

- Excel

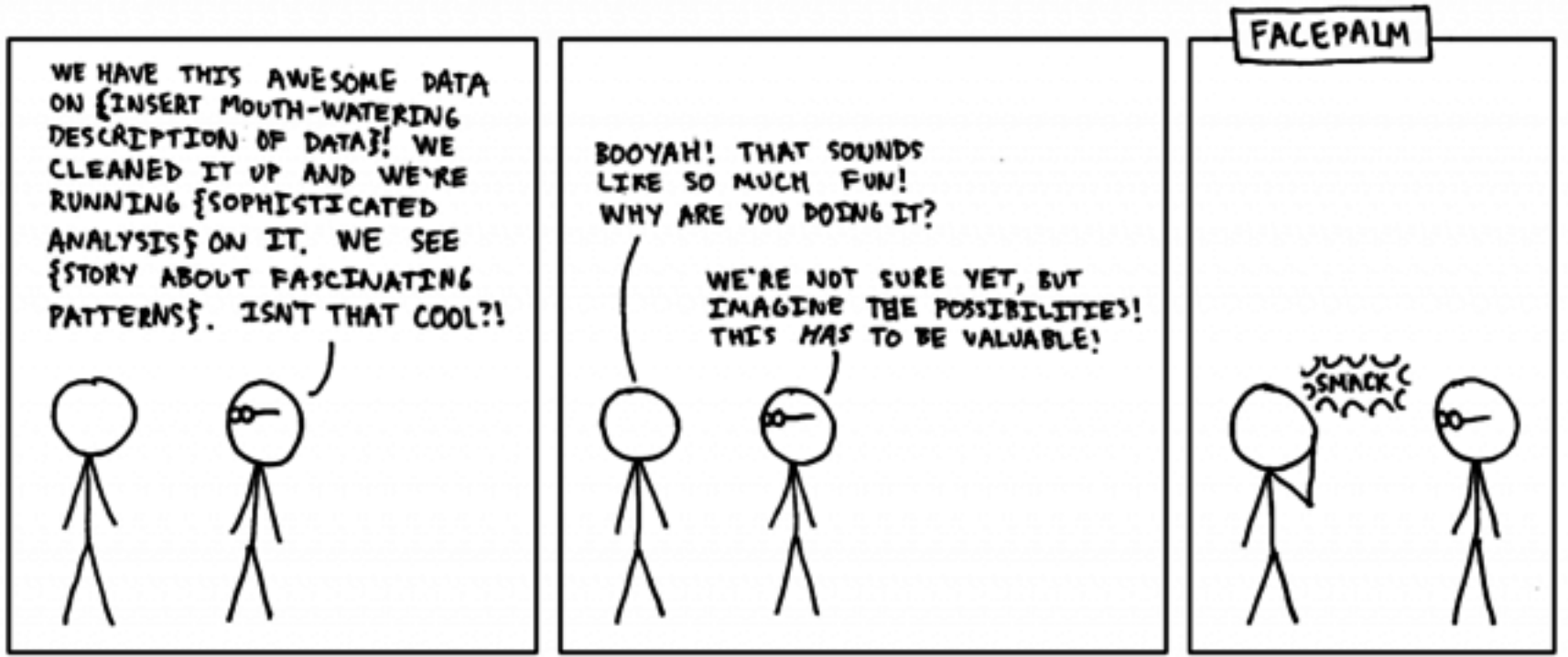
http://blogs.wsj.com/cio/2014/04/30/its-already-time-to-kill-the-data-scientist-title/

INTRO TO DATA SCIENCE

MACHINE LEARNING

How can machines learn?

THE FUN PART IS MODELING

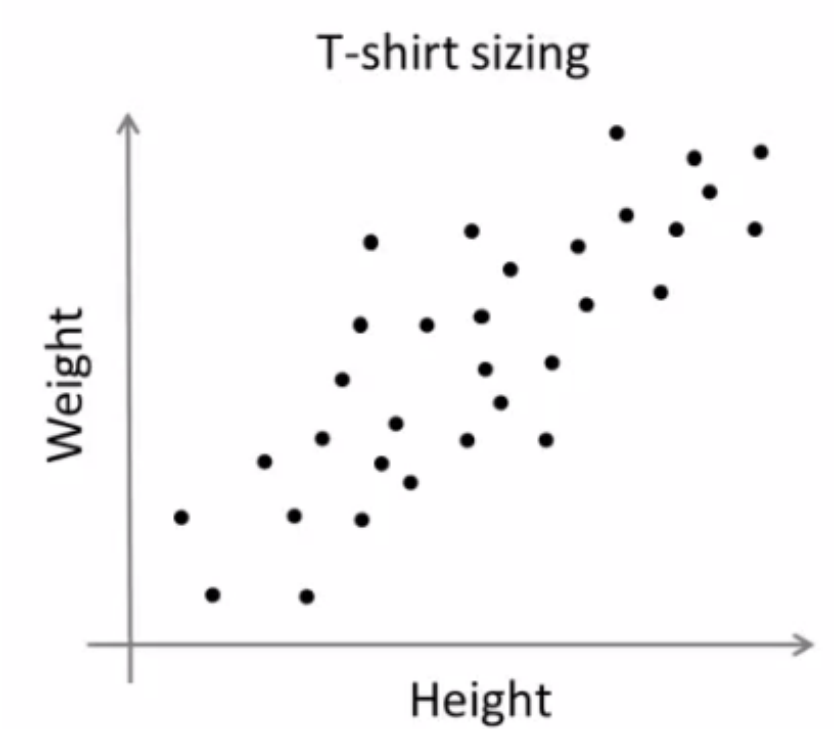


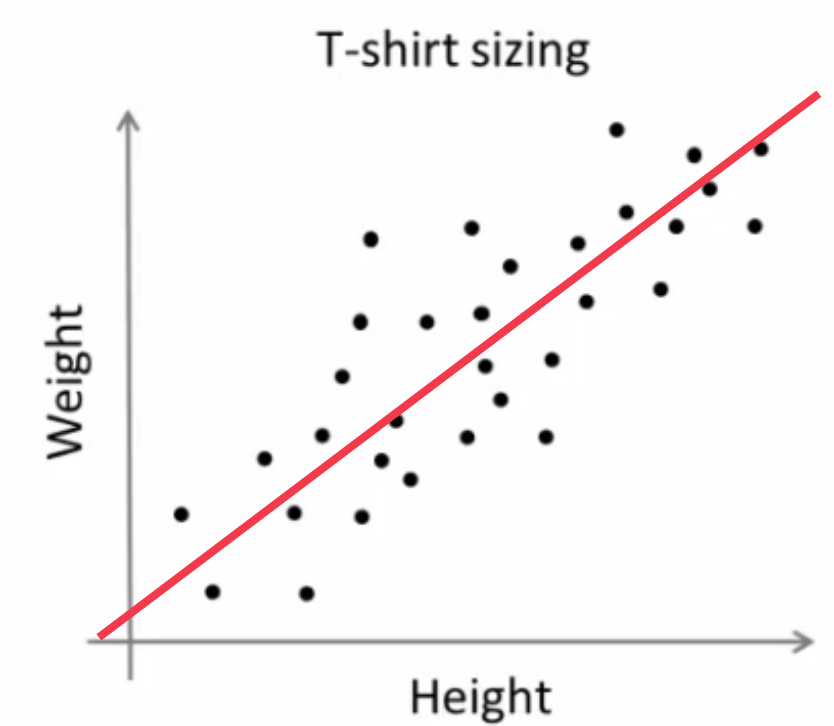
http://xkcd.com

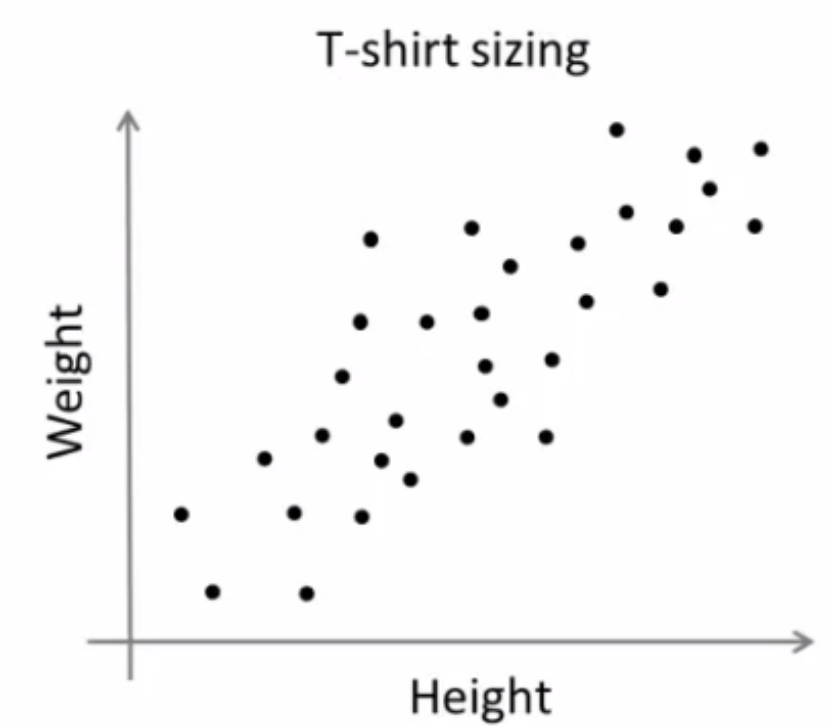
GIVEN DATA, HOW CAN WE POSSIBLY ACCURATELY EXTRAPOLATE BEYOND IT?

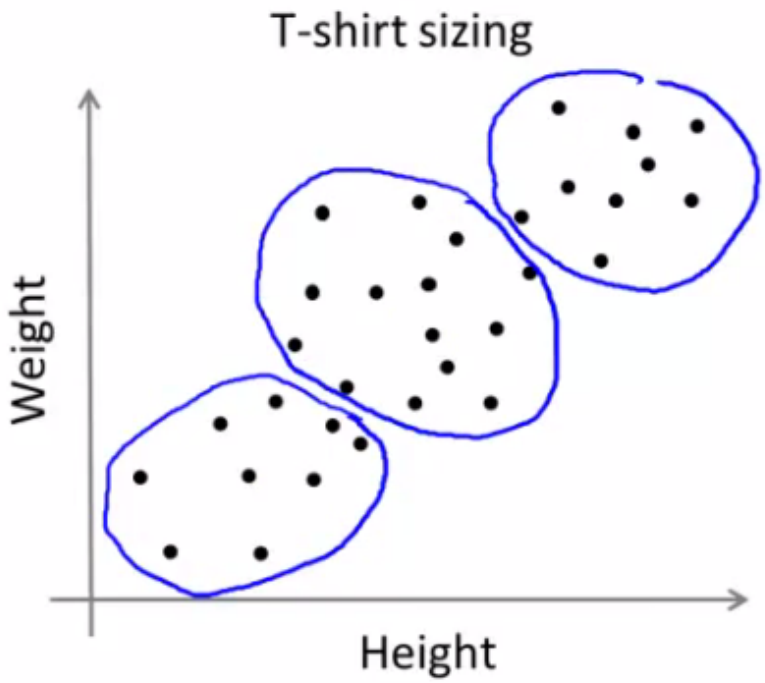
GIVEN DATA, HOW CAN WE POSSIBLY ACCURATELY EXTRAPOLATE BEYOND IT?

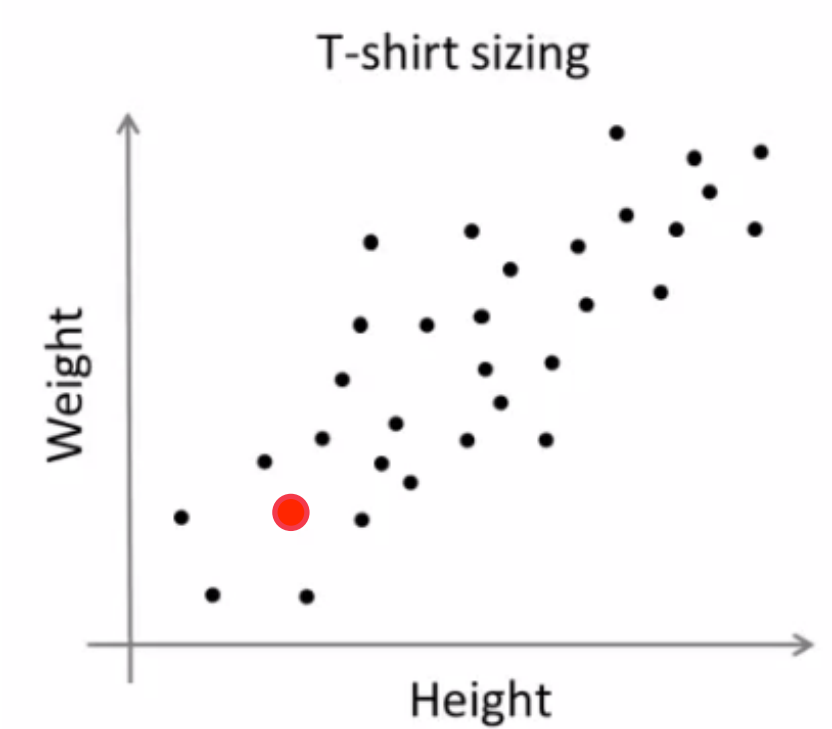
- Real-world data is not random.
- Different models make different assumptions about the distributions of the underlying data.


T-shirt sizing
Weight
Height


T-shirt sizing
Weight
Height


T-shirt sizing
Weight
Height


T-shirt sizing
Weight
Height


T-shirt sizing
Weight
Height

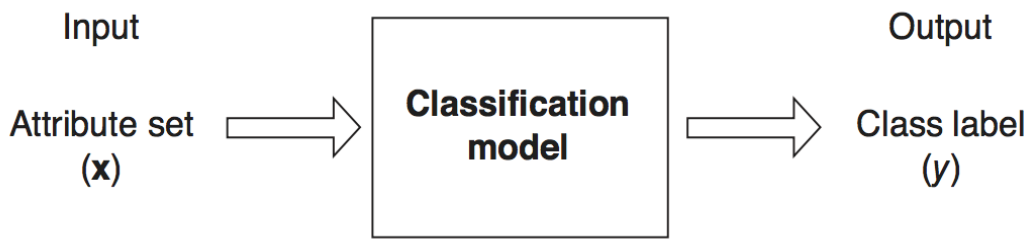


Figure 4.2. Classification as the task of mapping an input attribute set $\mathbf{x}$ into its class label y.

http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

	continuous	**categorical**
supervised	regression	classification
unsupervised	dimension reduction	clustering

- K-Nearest Neighbors
- Naïve Bayes
- Regression & Regularization
- Logistic Regression
- K-Means Clustering
- Ensemble Techniques
- Decision Trees & Random Forests
- Dimensionality Reduction
- Recommendation Systems
- Neural Networks

Analyzing the Analyzers
http://cdn.oreillystatic.com/oreilly/radarreport/0636920029014/Analyzing_the_Analyzers.pdf

Why Soft Skills Matter in the Data Sciences

http://data-informed.com/soft-skills-matter-data-science/

Data Scientist: The Sexiest Job of the 21st Century

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

Data Science For Business (less technical, more intuitive)
http://www.amazon.com/Data-Science-Business-data-analytic-thinking/dp/1449361323

Data Science From Scratch (easier technical introduction without Pandas)
http://www.amazon.com/Python-Data-Analysis-Wrangling-IPython/dp/1449319793/

Python for Data Analysis (using iPython/Pandas/NumPy)
http://www.amazon.com/Python-Data-Analysis-Wrangling-IPython/dp/1449319793/

An Introduction to Statistical Learning (machine learning with R)
http://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1461471370/

QUESTIONS?