GENERAL ASSEMBLY

INTRO TO DATA SCIENCE REVIEW

AGENDA

10 Databases in 10 minutes

Course Recap

Course Review

Next Steps

DATABASES

DATABASES

- What does it mean to be familiar with a database?
- What do data scientists need to know about databases?
- How to evaluate a database?

DATABASES

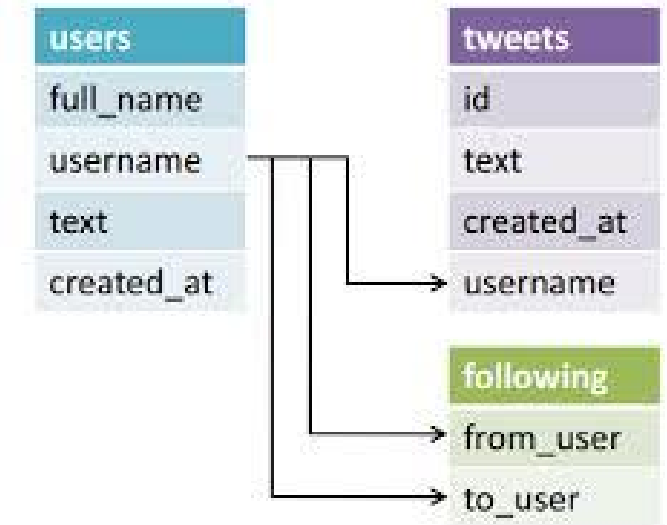
10 Databases in 10 minutes[1]

MySQL **PostGreSQL** **SQLite**	**MongoDB** **Cassandra** **Redis** **HBase** **Riak** **Couchbase** **Neo4j**

1. with a nod to *Seven Databases in Seven Weeks*

DATABASES: 10 DATABASES IN 10 MINUTES

Relational (RDBMS)



DATABASES: 10 DATABASES IN 10 MINUTES

Relational (RDBMS)

		pros	cons
SQLite	powerful, **embedded** RDBMS	lightweight mostly fully-functional DB good for prototyping, testing	lightweight- size restricted to 2GB can be slow
MySQL	most popular and commonly used open-source RDBMS	good for read-heavy applications easy to work with good security scalable	not 100% SQL compliant not suited for high-concurrent (high read/write) apps
PostGreSQL	most advanced, SQL-compliant and open-source objective RDBMS	big community extensible can handle complex procedures	not suited for read-heavy operations not as easy as MySQL to administer not easy to set up replication

Also see *Amazon RDS*

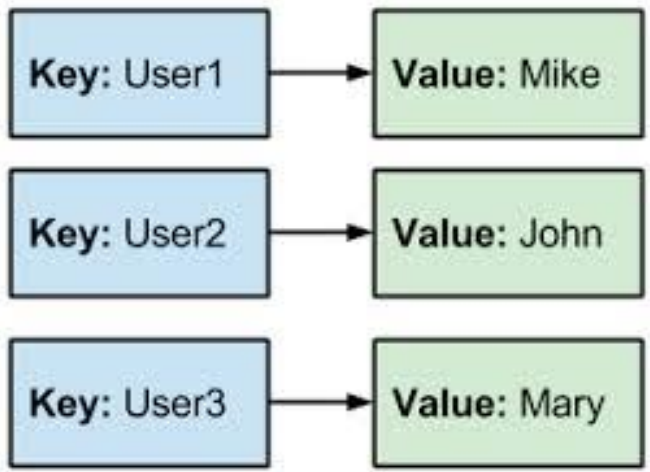
DATABASES: 10 DATABASES IN 10 MINUTES

NOSQL: NON-RELATIONAL DATA STORES

Key-value	**Redis** **Riak** MemcacheDB
Document	**MongoDB** **Couchbase** RavenDB
Column-oriented	**Cassandra** **HBase** AWS DynamoDB
Graph	**Neo4j** OrientDB Titan

DATABASES: 10 DATABASES IN 10 MINUTES

Key-value Stores



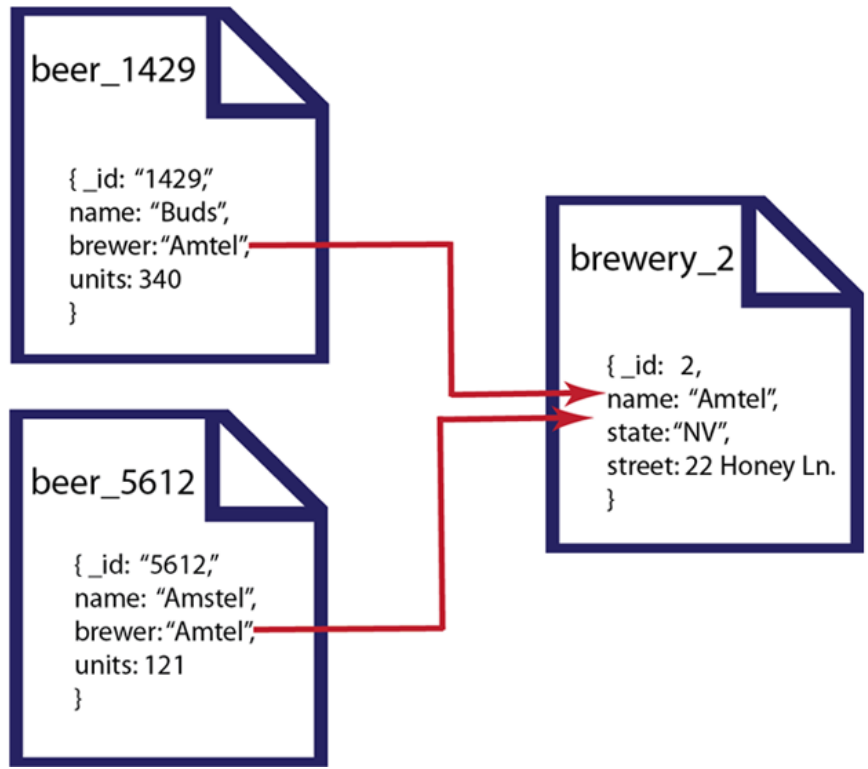
DATABASES: 10 DATABASES IN 10 MINUTES

Key-value Stores

		pros	cons
Redis	Fast, in-memory Good for caching Partition Tolerance Master-Slave Replication	***fast***!! lots of great built-in data structures easy to administer good community support	sharding, failover not yet fully realized
Riak	Primarily used for extreme high availability	fault tolerant good at replication (DR) good support from Basho open-source edition tunable trade-offs for distribution and replication	be ready to pay for cross data center replication (enterprise)

DATABASES: 10 DATABASES IN 10 MINUTES

Document oriented



DATABASES: 10 DATABASES IN 10 MINUTES

Document-oriented Stores

		pros	cons
MongoDB	Arguably most popular of NoSQL DBs Schema-free document store	master/slave replication with failover large community of users good support from MongoDB (10gen) auto-sharding	scaling with write-heavy applications can get tricky don't use built-in map-reduce!
Couchbase	High performance key-value/document store with flexible, but slow, indexes	master-master replication replication supports filtering or selective replication write operations do not block reads	not as popular as Mongo, smaller community of support

DATABASES: 10 DATABASES IN 10 MINUTES

Column oriented databases

Row Store v. Column Store

Record #	Name	Address	City	State
0003623	ABC	125 N Way	Cityville	PA
0003626	Newburg	1300 Forest Dr.	Troy	VT
0003647	Flotsam	5 Industrial Pkwy	Springfield	MT
0003705	Jolly	529 S 5th St.	Anywhere	NY

Record #	Name	Address	City	State
0003623	ABC	125 N Way	Cityville	PA
0003626	Newburg	1300 Forest Dr	Troy	VT
0003647	Flotsam	Industrial Pkwy	Springfield	MT
0003705	Jolly	529 S 5th St.	Anywhere	NY

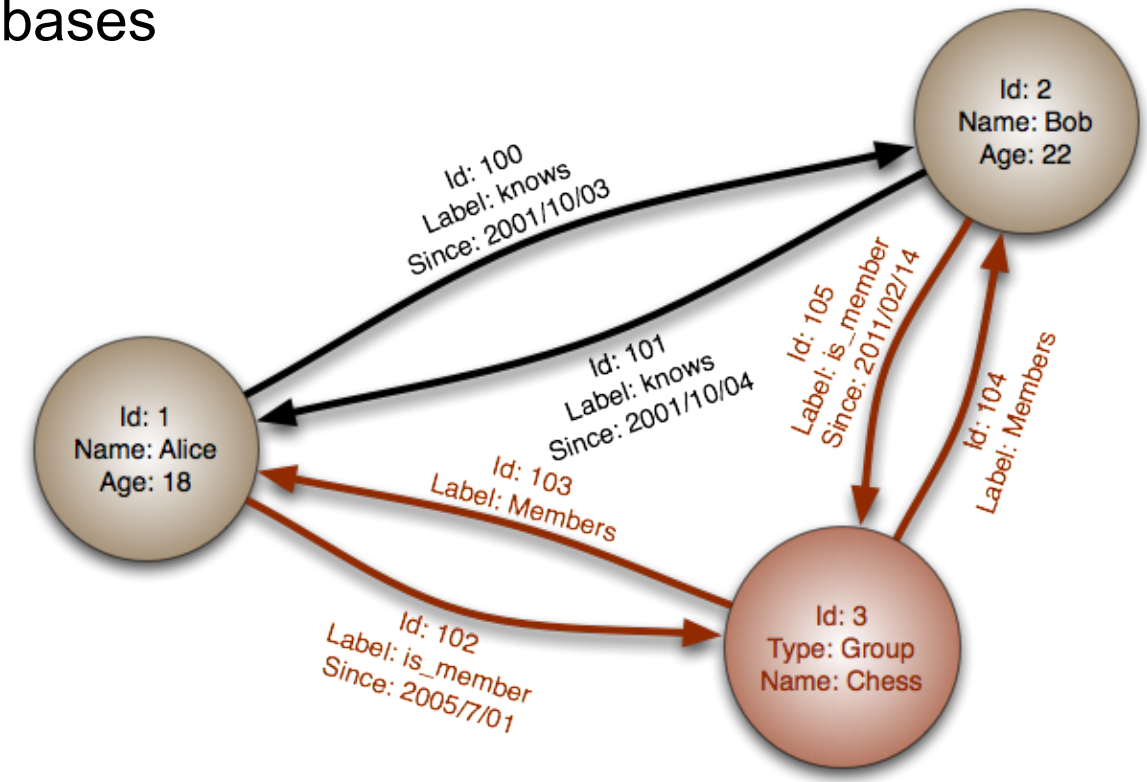
DATABASES: 10 DATABASES IN 10 MINUTES

Column oriented databases

		pros	cons
Cassandra	Store huge datasets in "almost" SQL Based on DynamoDB	tunable trade-offs for distribution and replication excellent at cross datacenter replication good for write-heavy applications	not suited for read-heavy applications learning curve for efficient usage best to have enterprise support
HBase	Based on Google's BigTable Billions of rows by millions of columns	uses HDFS as storage map/reduce with Hadoop best if you use the Hadoop/HDFS stack already	lots of “moving parts” (e.g. zookeeper) dependent on HDFS, Hadoop complex to administer

DATABASES: 10 DATABASES IN 10 MINUTES

Graph databases



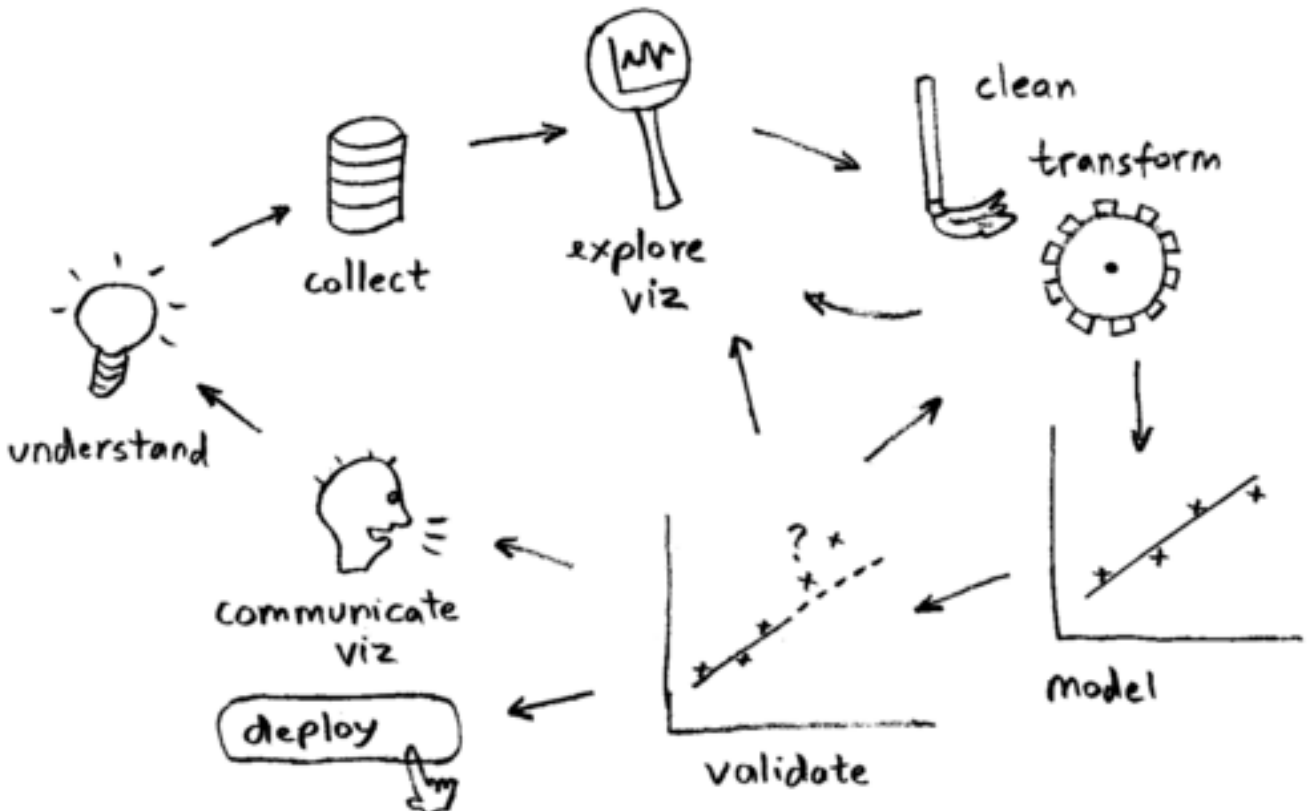
DATABASES: 10 DATABASES IN 10 MINUTES

Graph databases

		pros	cons
Neo4j	“World’s leading graph database” Native graph processing	good at graph-style interconnected data path finding optimized for reads clustering, replication, caching, online backup	graph DBs still relatively immature query-syntax has a learning curve relatively small user community

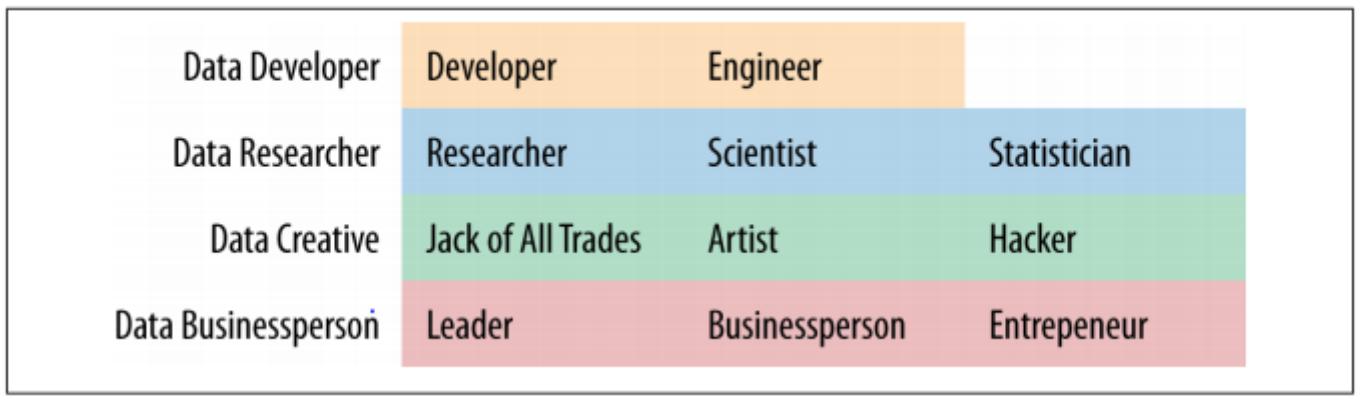
COURSE RECAP

DATA SCIENCE WORKFLOW

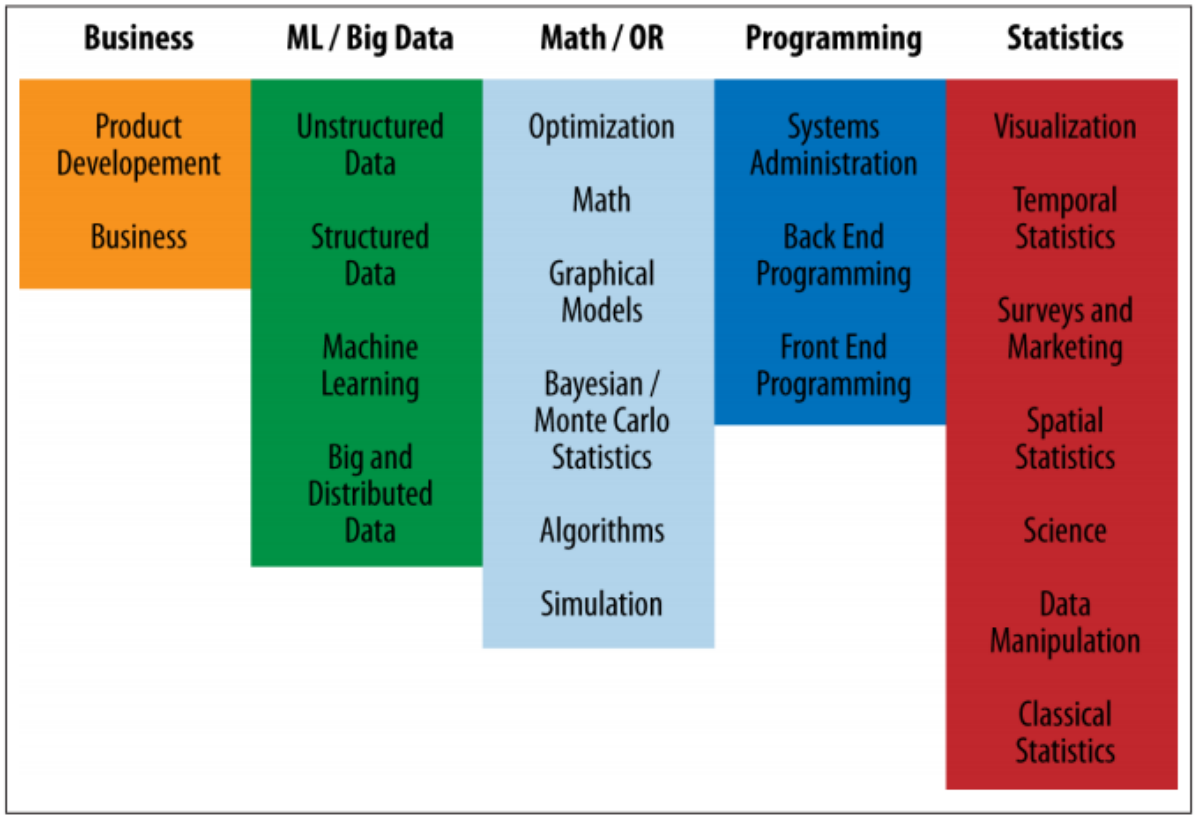


source: DATA SCIENCE TOOLBOX SURVEY RESULTS… SURPRISE! R AND PYTHON WIN

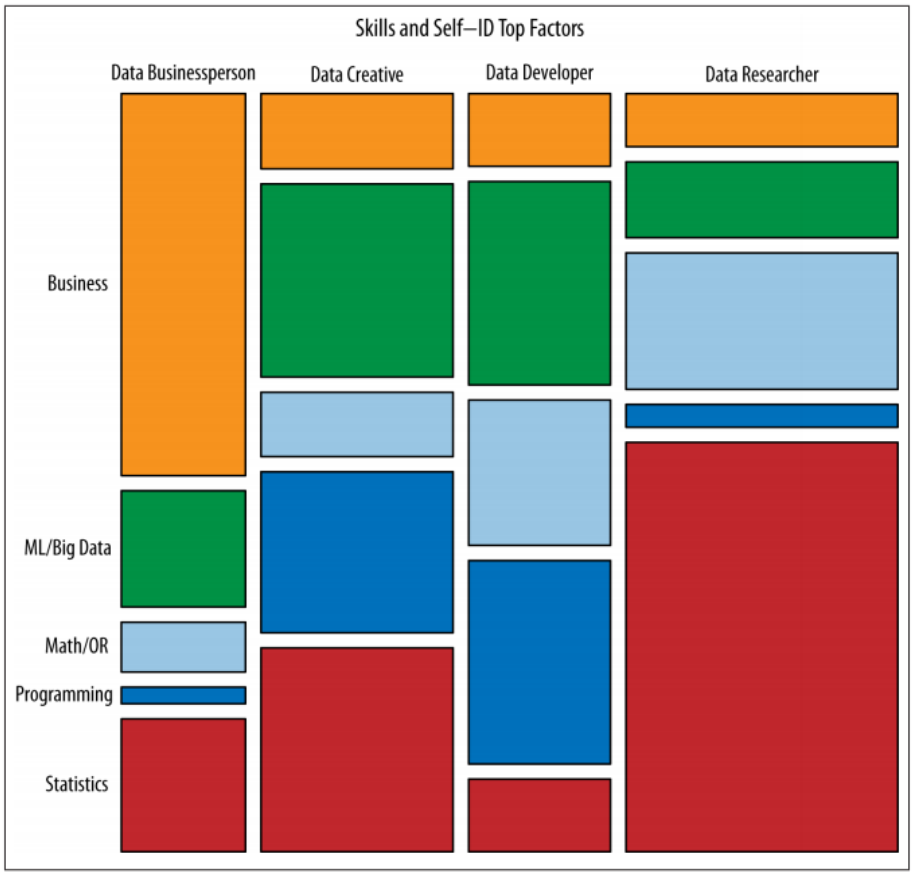
Data Science Types



Data Science Skills



Data Science Skills by Type



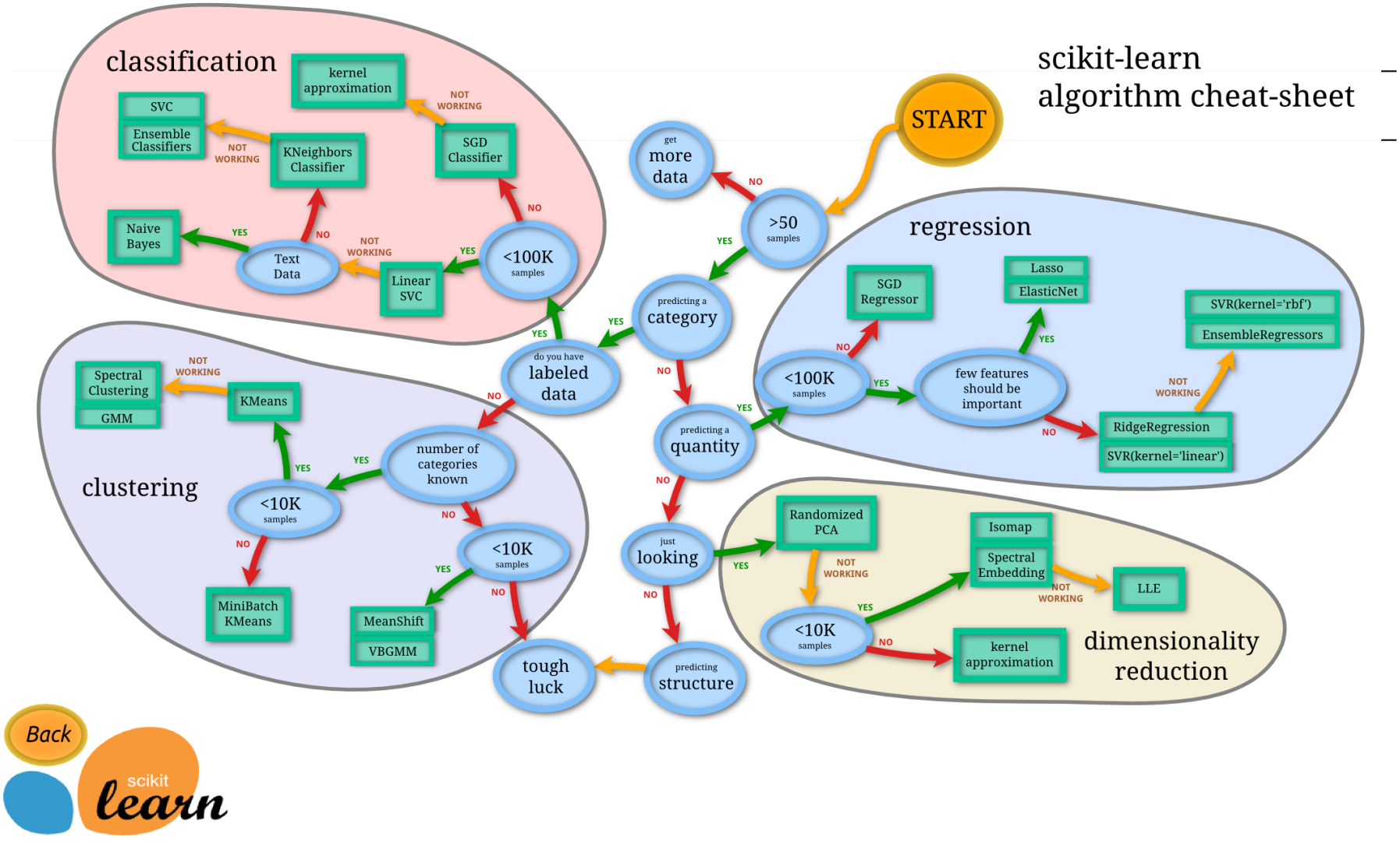
COURSE REVIEW

WHAT WE'VE COVERED: MODELS

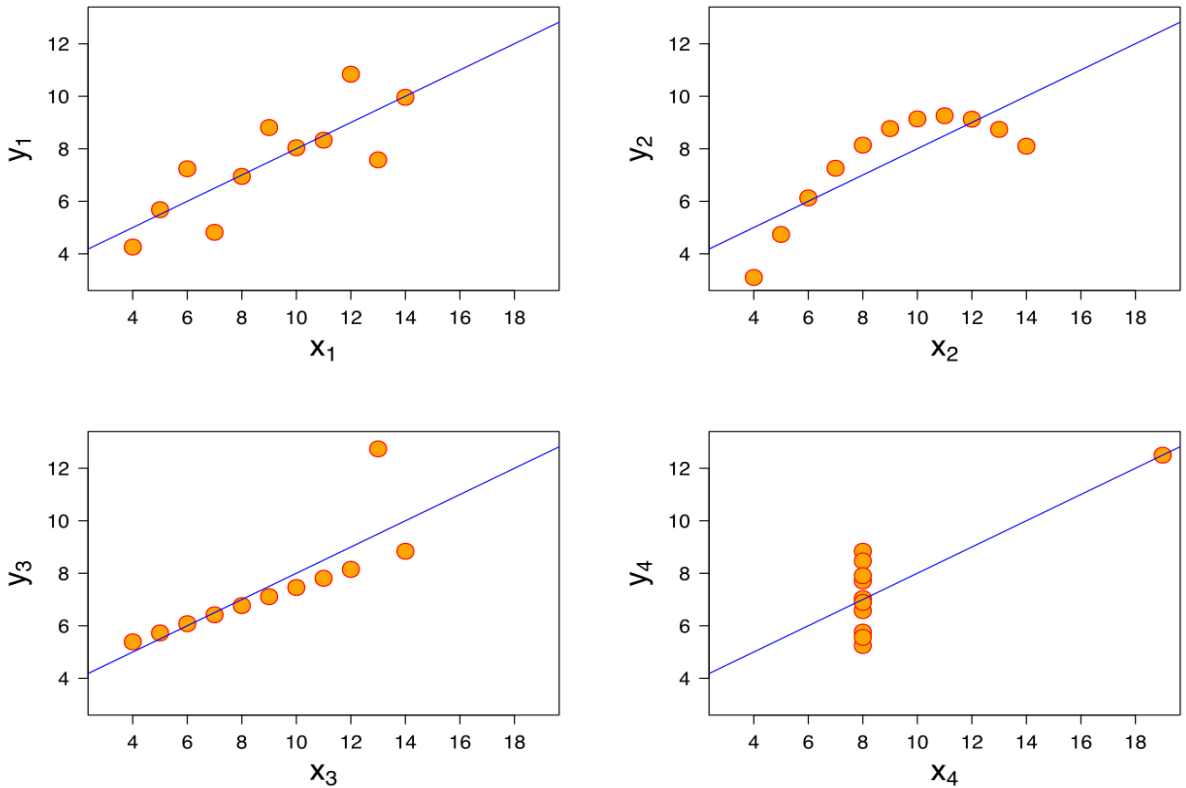
Supervised Learning	Linear Models	Regression	Simple Linear Regression
			Multiple Linear Regression
			Polynomial, Ridge, Lasso
		Classification	Logistic Regression
	Non-linear Models	Classification	K-Nearest Neighbors Naive Bayes
		Regression/ Classification	Decision Trees, Random Forests
Unsupervised Learning		Clustering	K-Means LDA (Topic Modeling)
		Dimensionality Reduction	PCA

WHAT WE'VE ALSO COVERED

- Exploratory Data Analysis
- Data Visualization
- Model Selection
- Cross Validation
- Bayesian Analysis (A/B Testing)
- Natural Language Processing
- Time Series Analysis
- GeoSpatial Problems
- Recommendation Engines, Collaborative Filtering
- MapReduce


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
Naive Bayes
YES
Text Data
NO
KNeighbors Classifier
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
NOT WORKING
Linear SVC
YES
classification
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
Lasso
ElasticNet
NO
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
NO
RidgeRegression
SVR(kernel='linear')
regression
NO
number of categories known
YES
<10K samples
NO
KMeans
YES
Spectral Clustering
GMM
NOT WORKING
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
clustering
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

ANSCOMBE'S QUARTET



LINEAR REGRESSION MODEL

Simple Linear Regression

$$y = \alpha + \beta x$$

Multiple Linear Regression

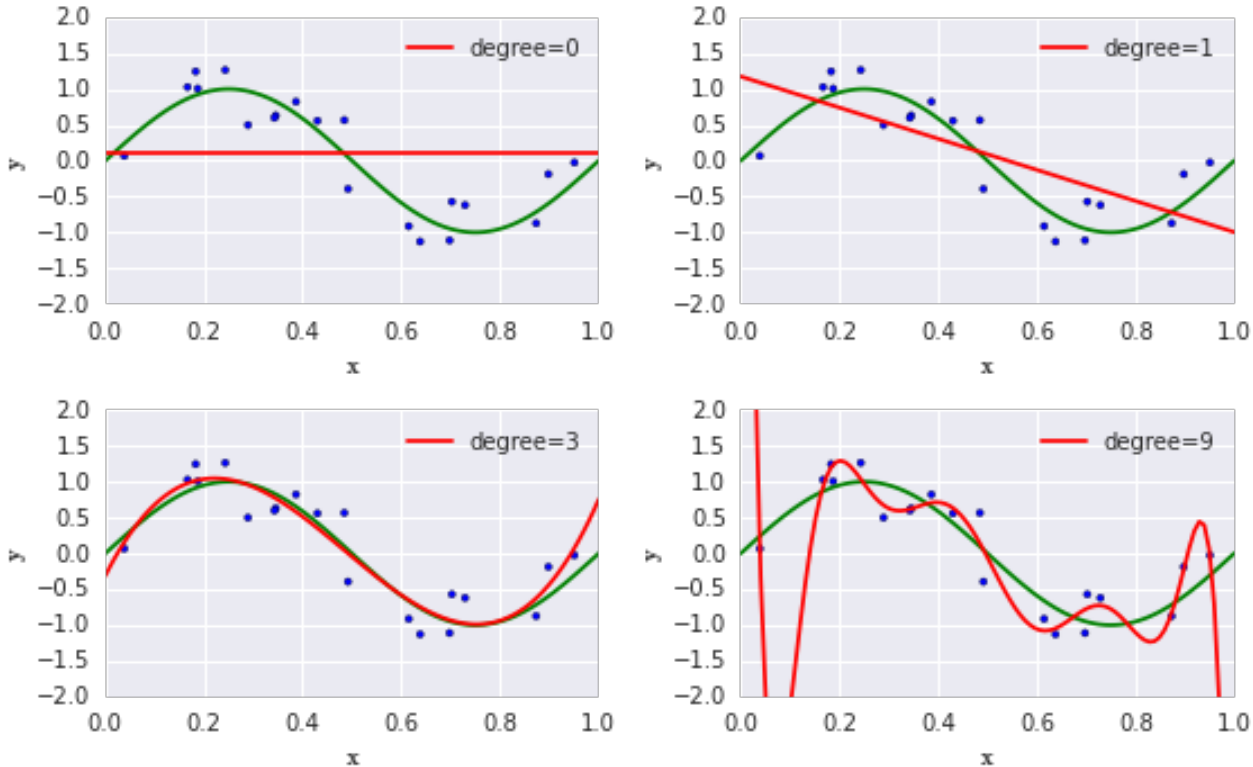
$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

LINEAR REGRESSION ...

- How does sales volume change with changes in price? How is this affected by changes in weather?
- Is there a relationship between the amount of a drug absorbed and body weight of a patient?
- Can we explain the effect of education on income?
- How does the energy released by an earthquake vary with the depth of its epicenter?

Assessing Model Accuracy

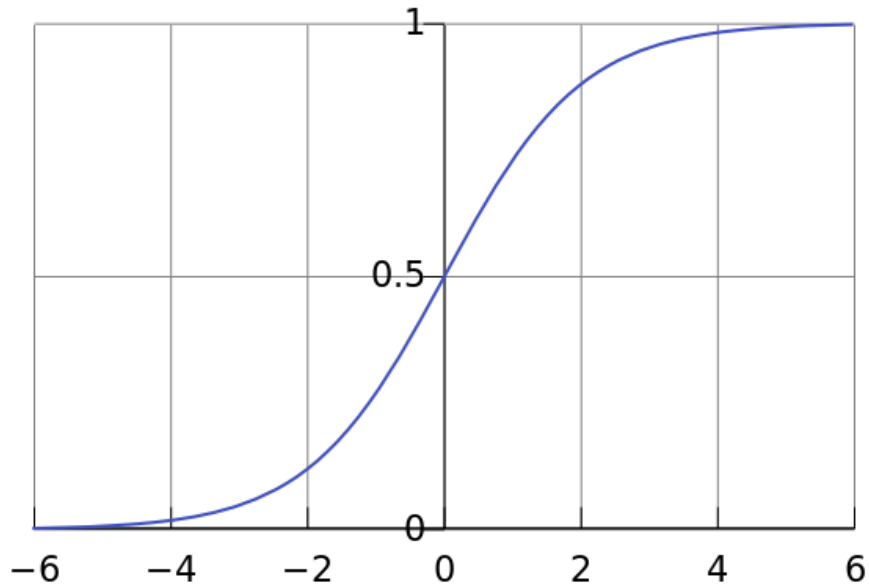
How to decide among multiple models?



THE LOGISTIC FUNCTION

- The logistic function always returns a value between zero and one.

$$F(t) = \frac{1}{1 + e^{-t}}$$

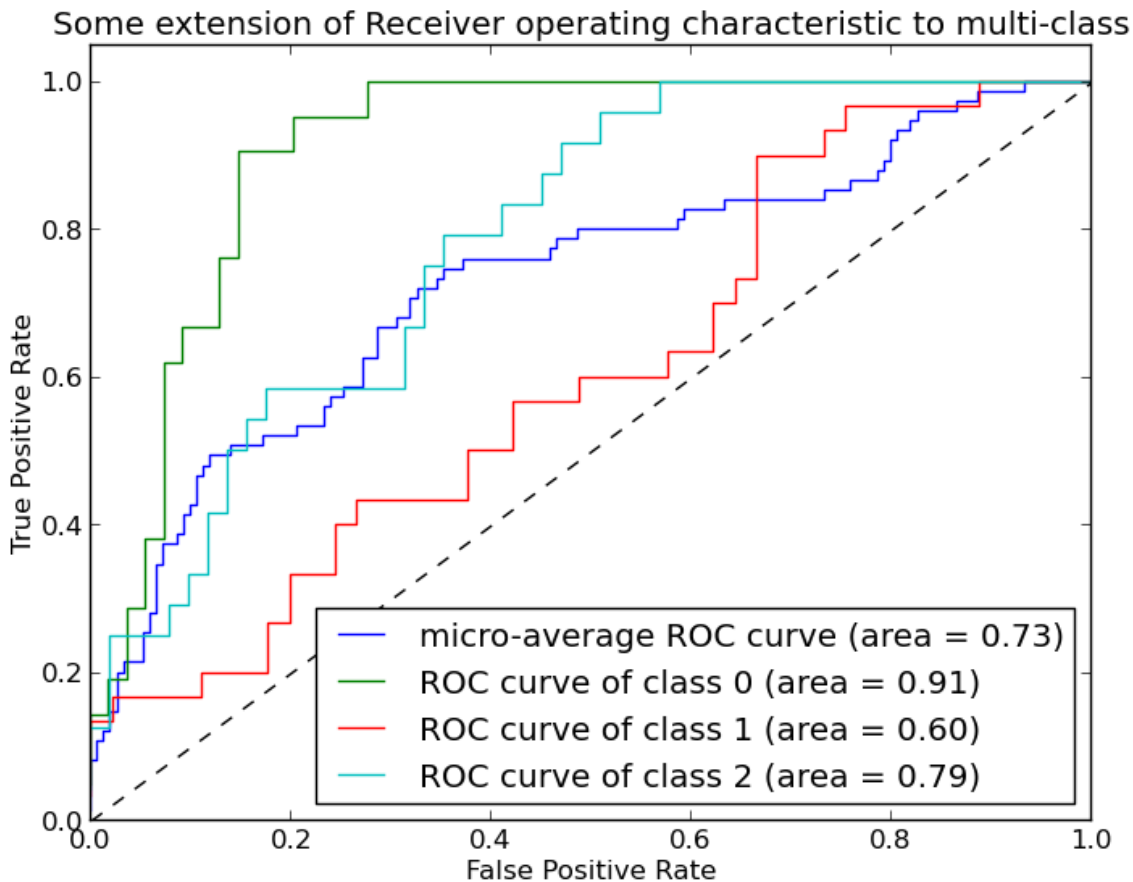


THE LOGISTIC FUNCTION

- The **logit function** is the inverse of the logistic function. It links back to a linear combination of the explanatory variables so that the parameter values can be solved.

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta x)}}$$

$$g(x) = ln \frac{F(x)}{1 - F(x)} = \beta_0 + \beta x$$


Some extension of Receiver operating characteristic to multi-class
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
micro-average ROC curve (area = 0.73)
ROC curve of class 0 (area = 0.91)
ROC curve of class 1 (area = 0.60)
ROC curve of class 2 (area = 0.79)

BAYESIAN INFERENCE

What can we say about classification using Bayes' theorem?

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

Bayes' theorem can help us to determine the probability of a record belonging to a class, *given* the data we observe.

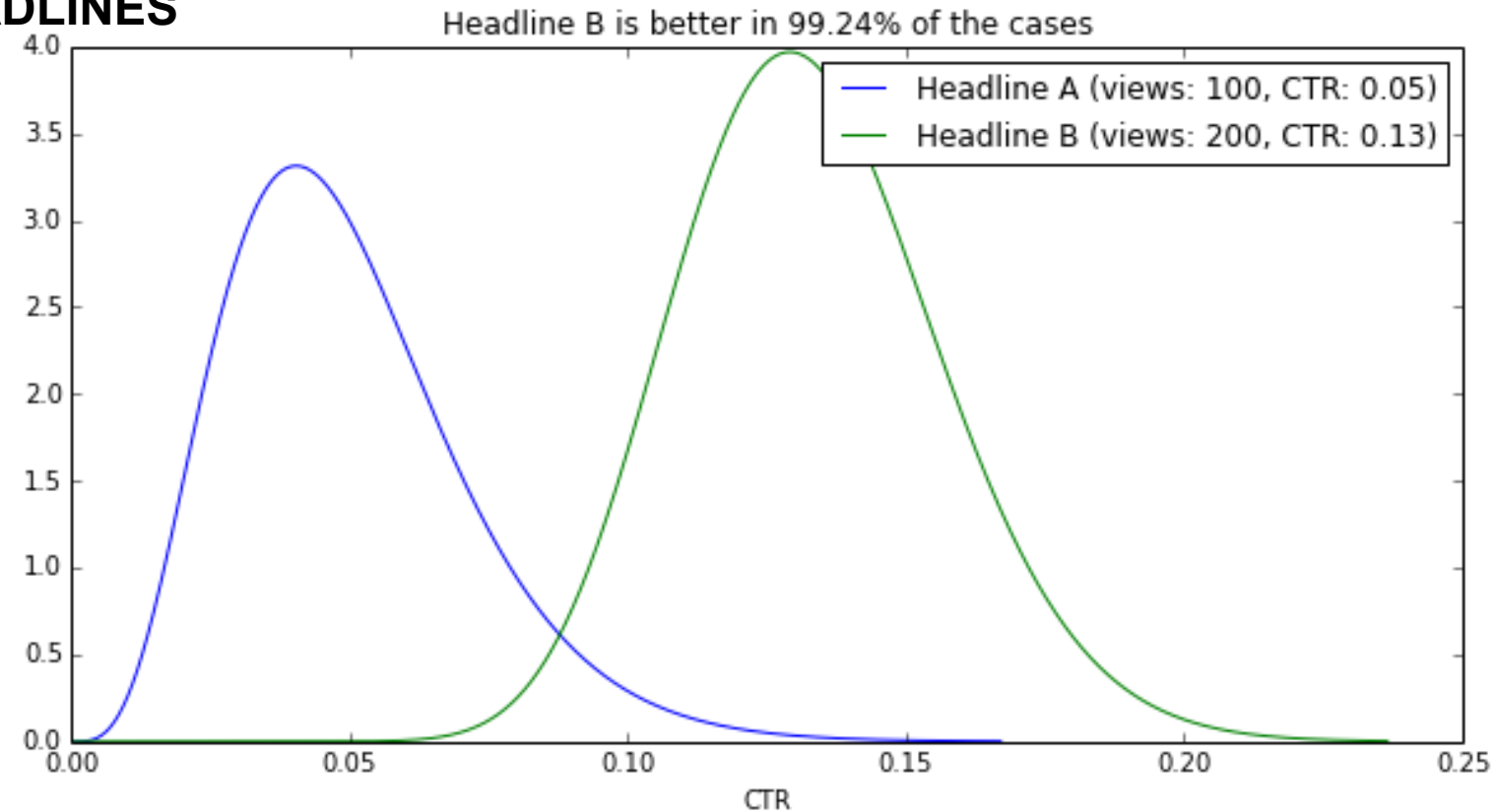
source: Data Analysis with Open Source Tools, by Philipp K. Janert. O'Reilly Media, 2011.

THE POSTERIOR

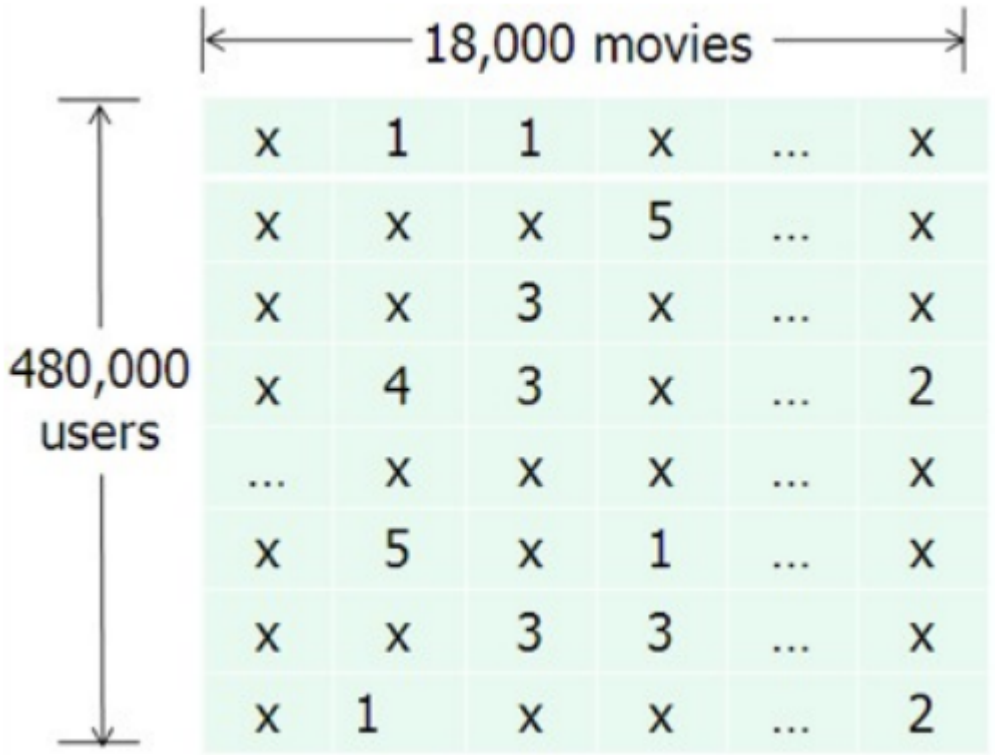
$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The goal of any Bayesian computation is to learn the ***posterior distribution*** of a particular variable.

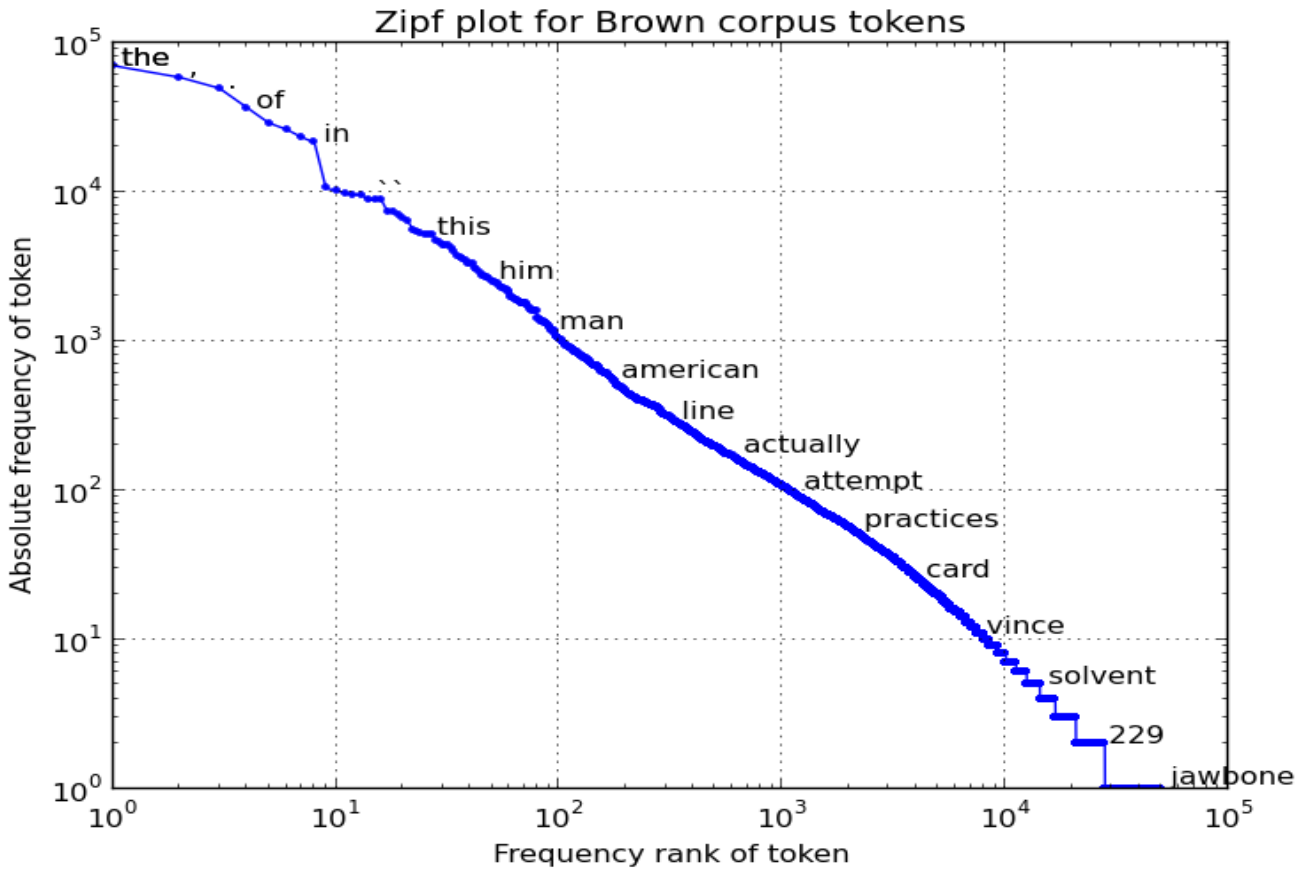
Beta distribution: Comparing Two Headlines



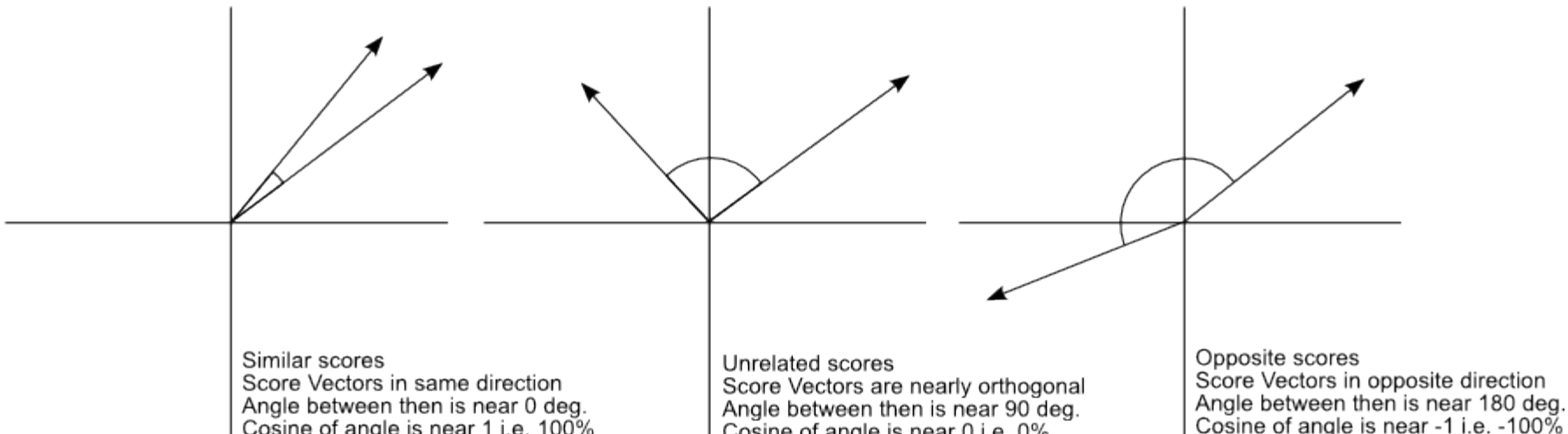
COLLABORATIVE FILTERING



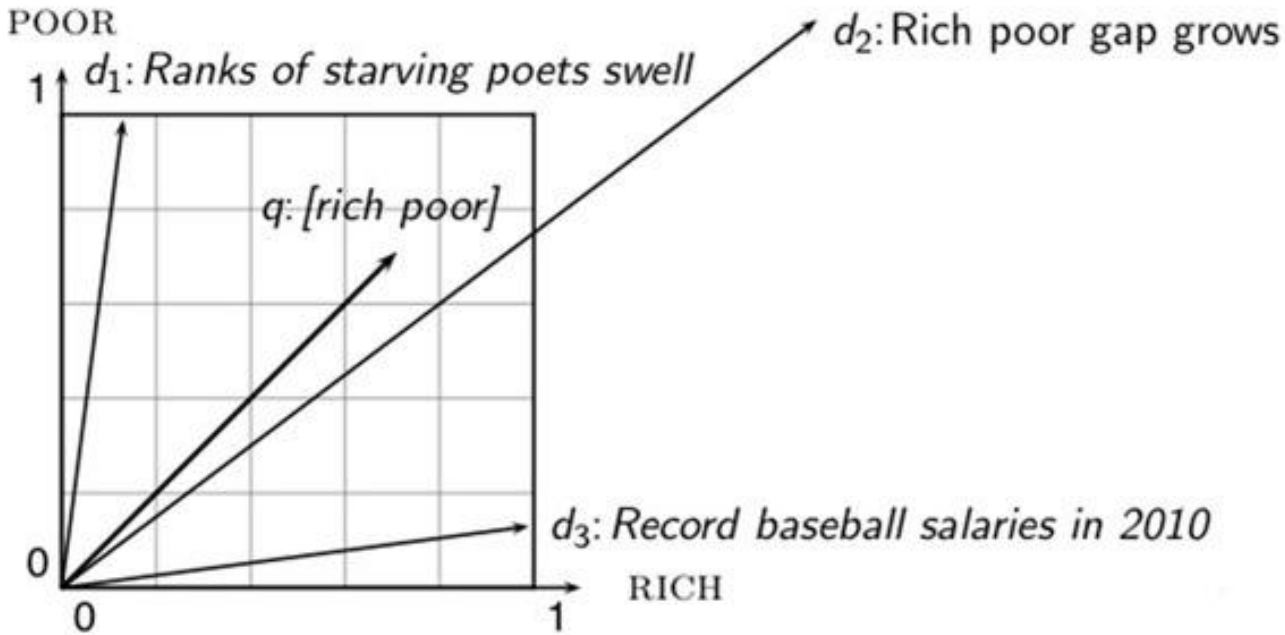
NLP: ZIPF'S LAW



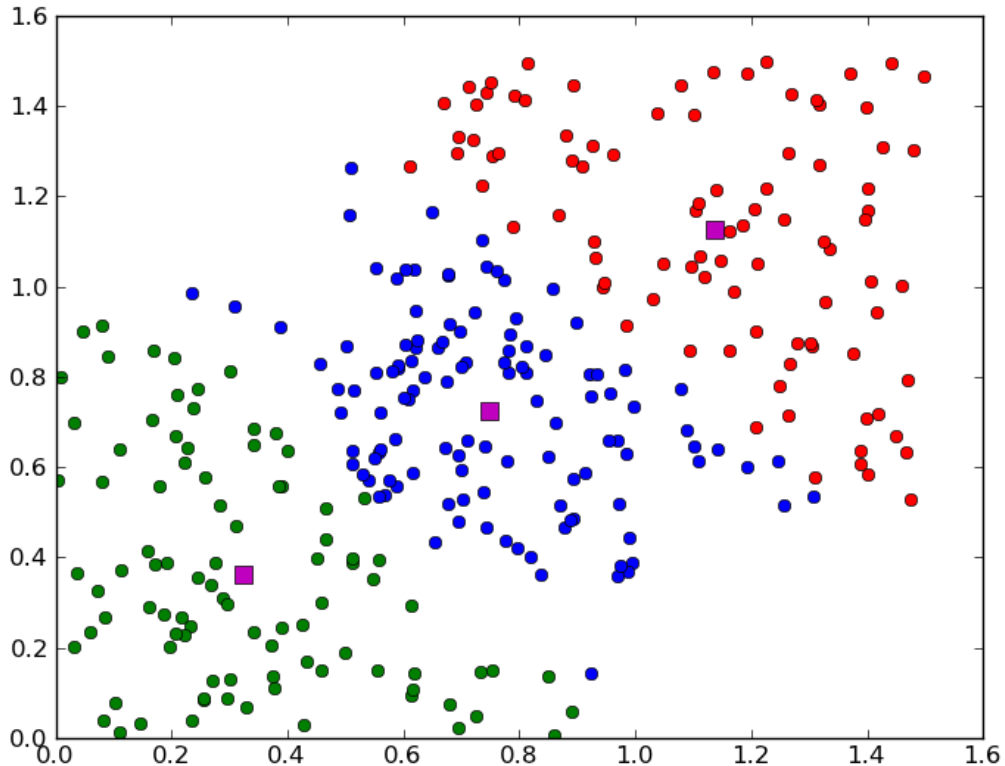
COSINE SIMILARITY



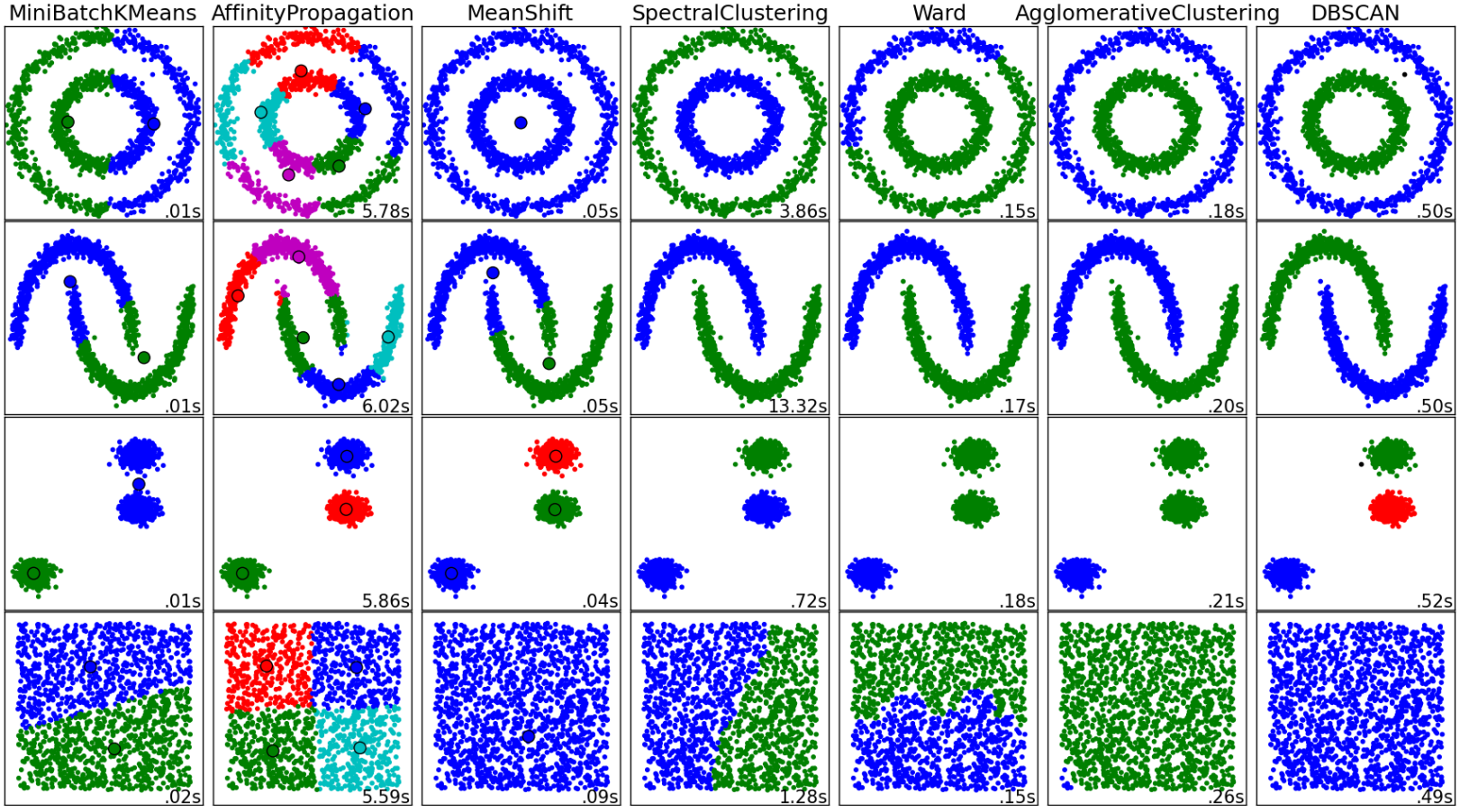
COSINE SIMILARITY



K-Means Clustering



Clustering Algorithms



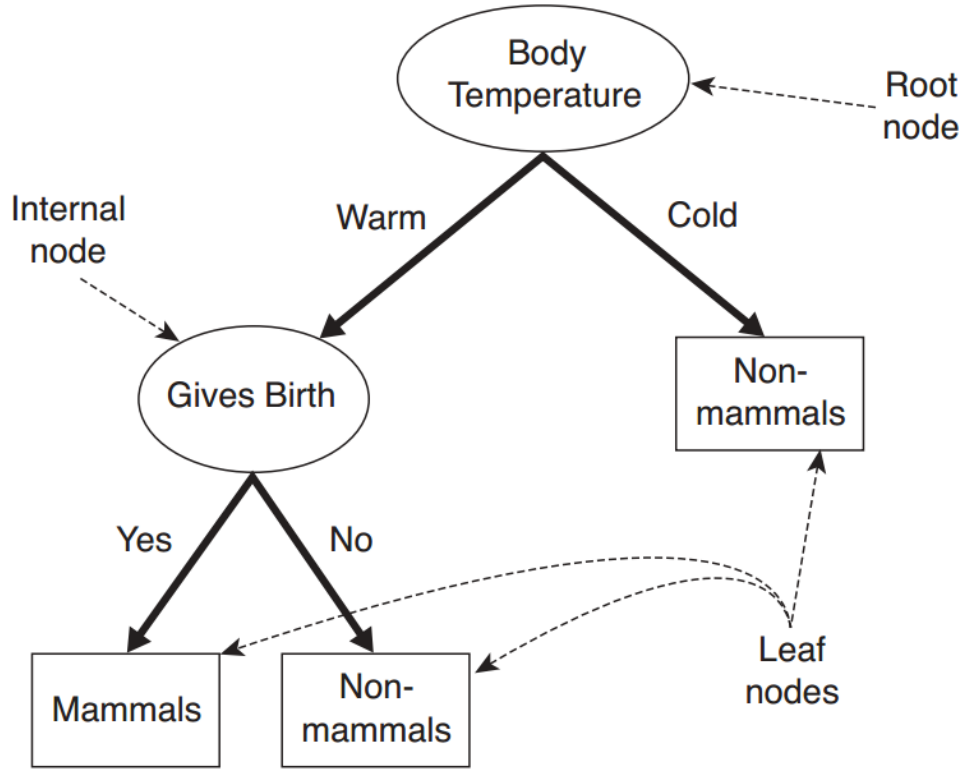
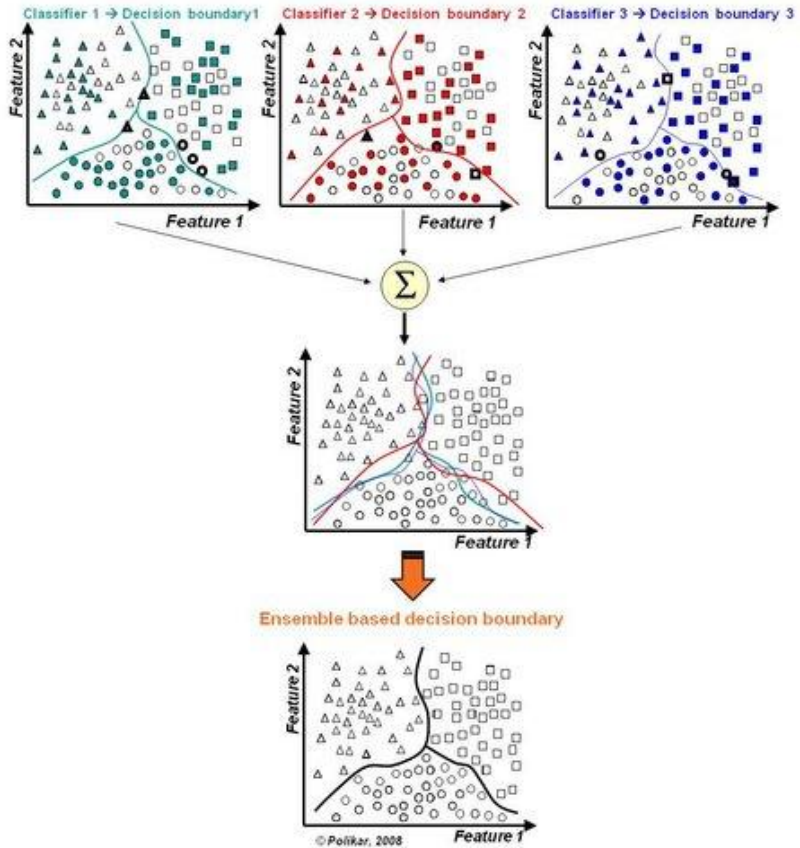
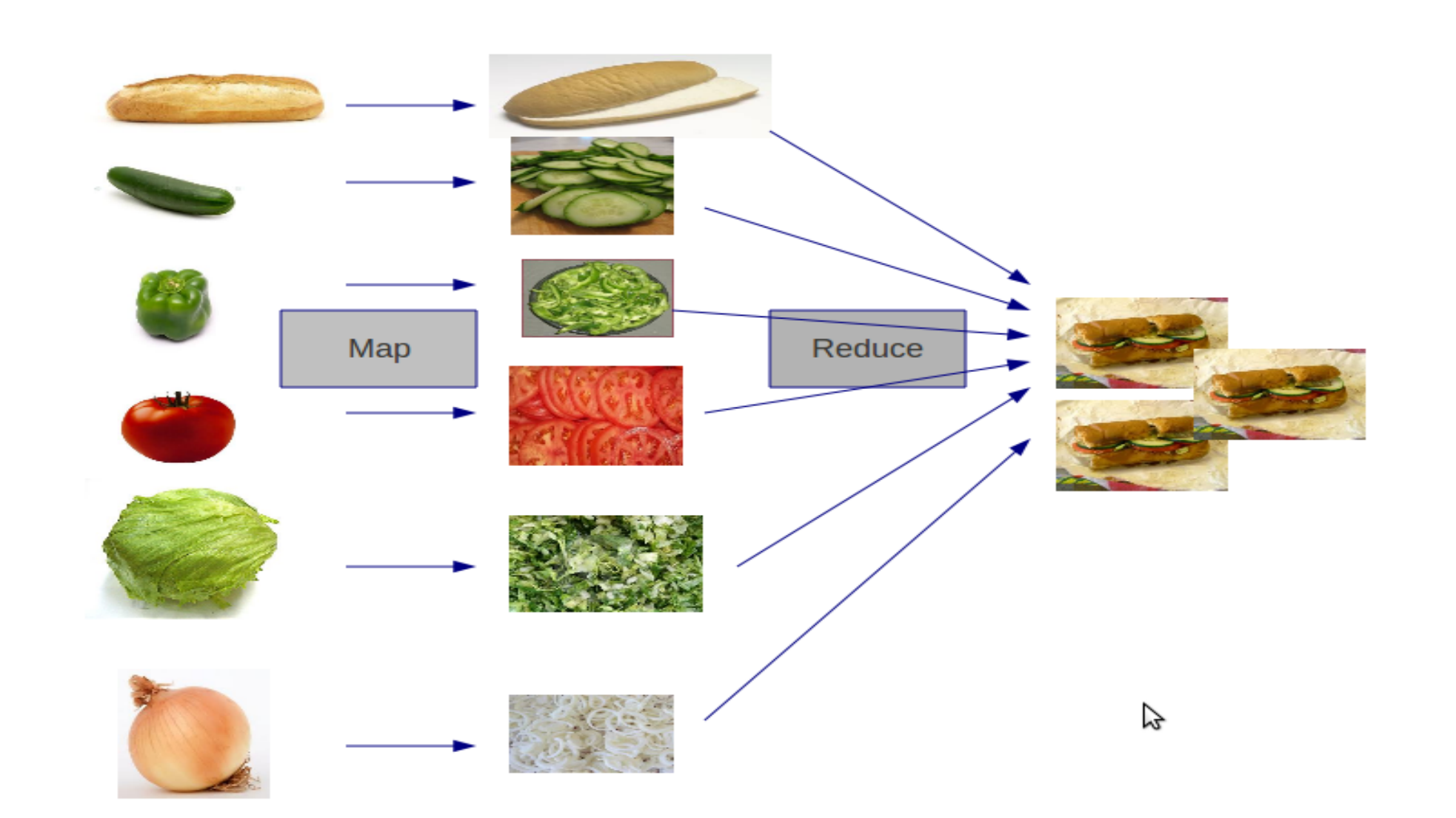


Figure 4.4. A decision tree for the mammal classification problem.

RANDOM FORESTS

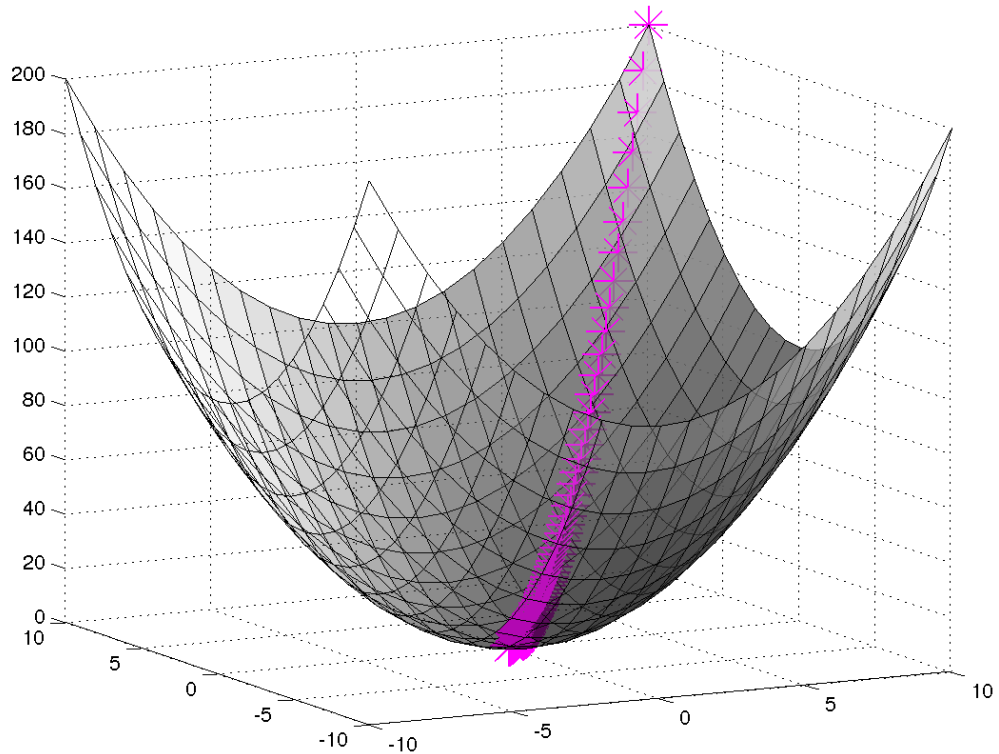



Map
Reduce

WHAT WE DIDN'T COVER

Bayes Networks
Probabilistic Graphical Models
Streaming/Sketch algorithms
Neural Networks
Sampling Methods
Optimization Methods
Gradient Descent
Support Vector Machines
Mixture Models, EM
Linear Algebra
Experiment Design

GRADIENT DESCENT

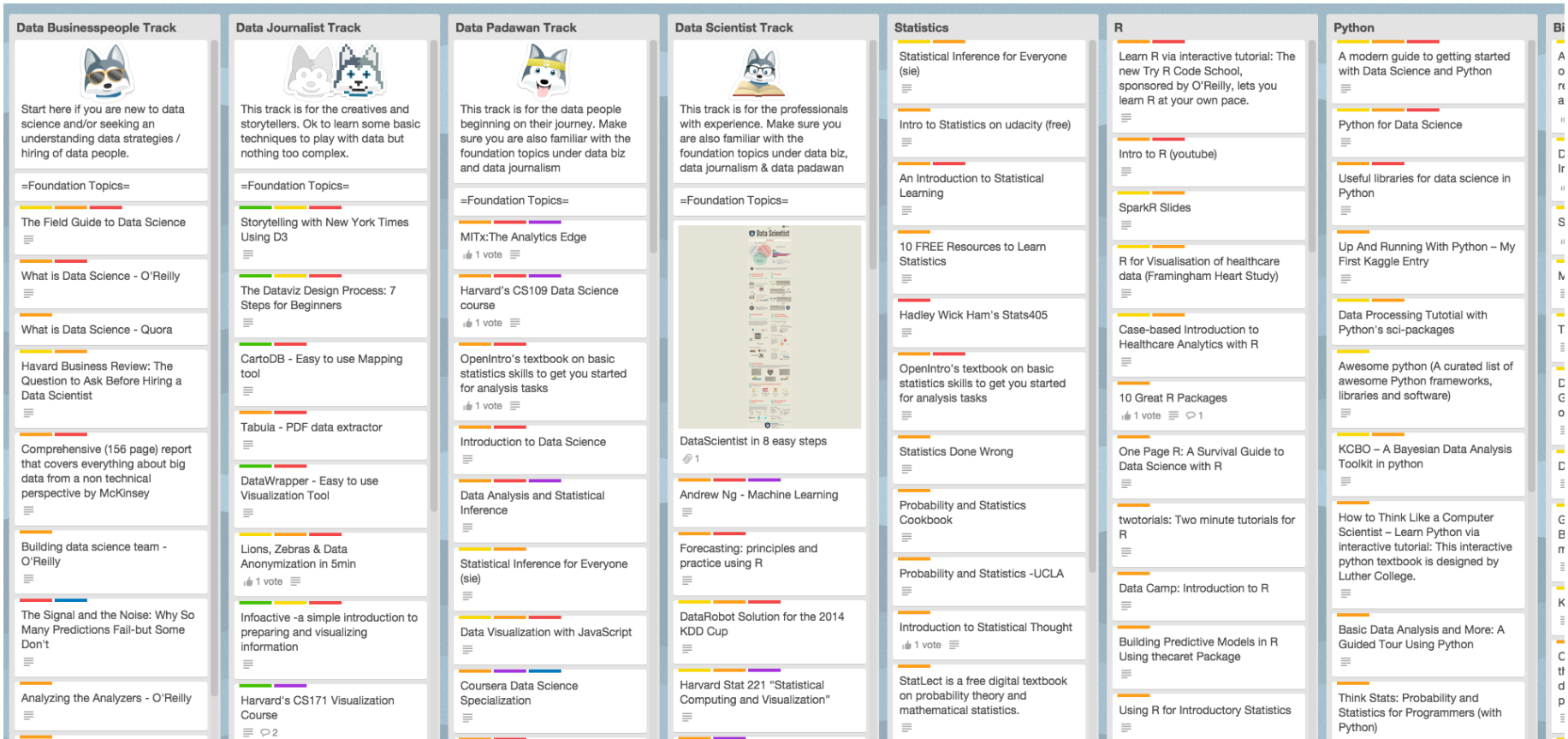


NEXT STEPS

WHAT'S NEXT?

- Networking
- Meetups (speaking, attending)
- Online Courses
- Focus on one or two topics and dig deep

TRELLO DATA SCIENCE KANBAN



Data Businesspeople Track

Start here if you are new to data science and/or seeking an understanding data strategies / hiring of data people.

=Foundation Topics=

The Field Guide to Data Science

What is Data Science - O'Reilly

What is Data Science - Quora

Havard Business Review: The Question to Ask Before Hiring a Data Scientist

Comprehensive (156 page) report that covers everything about big data from a non technical perspective by McKinsey

Building data science team - O'Reilly

The Signal and the Noise: Why So Many Predictions Fail-but Some Don't

Analyzing the Analyzers - O'Reilly

Data Journalist Track

This track is for the creatives and storytellers. Ok to learn some basic techniques to play with data but nothing too complex.

=Foundation Topics=

Storytelling with New York Times Using D3

The Dataviz Design Process: 7 Steps for Beginners

CartoDB - Easy to use Mapping tool

Tabula - PDF data extractor

DataWrapper - Easy to use Visualization Tool

Lions, Zebras & Data Anonymization in 5min

1 vote

Infoactive -a simple introduction to preparing and visualizing information

Harvard's CS171 Visualization Course

2

Data Padawan Track

This track is for the data people beginning on their journey. Make sure you are also familiar with the foundation topics under data biz and data journalism

=Foundation Topics=

MITx:The Analytics Edge

1 vote

Harvard's CS109 Data Science course

1 vote

OpenIntro's textbook on basic statistics skills to get you started for analysis tasks

1 vote

Introduction to Data Science

Data Analysis and Statistical Inference

Statistical Inference for Everyone (sie)

Data Visualization with JavaScript

Coursera Data Science Specialization

Data Scientist Track

This track is for the professionals with experience. Make sure you are also familiar with the foundation topics under data biz, data journalism & data padawan

=Foundation Topics=

DataScientist in 8 easy steps

1

Andrew Ng - Machine Learning

Forecasting: principles and practice using R

DataRobot Solution for the 2014 KDD Cup

Harvard Stat 221 "Statistical Computing and Visualization"

Statistics

Statistical Inference for Everyone (sie)

Intro to Statistics on udacity (free)

An Introduction to Statistical Learning

10 FREE Resources to Learn Statistics

Hadley Wick Ham's Stats405

OpenIntro's textbook on basic statistics skills to get you started for analysis tasks

Statistics Done Wrong

Probability and Statistics Cookbook

Probability and Statistics -UCLA

Introduction to Statistical Thought

1 vote

StatLect is a free digital textbook on probability theory and mathematical statistics.

R

Learn R via interactive tutorial: The new Try R Code School, sponsored by O'Reilly, lets you learn R at your own pace.

Intro to R (youtube)

SparkR Slides

R for Visualisation of healthcare data (Framingham Heart Study)

Case-based Introduction to Healthcare Analytics with R

10 Great R Packages

1 vote

1

One Page R: A Survival Guide to Data Science with R

twotorials: Two minute tutorials for R

Data Camp: Introduction to R

Building Predictive Models in R Using thecaret Package

Using R for Introductory Statistics

Python

A modern guide to getting started with Data Science and Python

Python for Data Science

Useful libraries for data science in Python

Up And Running With Python – My First Kaggle Entry

Data Processing Tutotial with Python's sci-packages

Awesome python (A curated list of awesome Python frameworks, libraries and software)

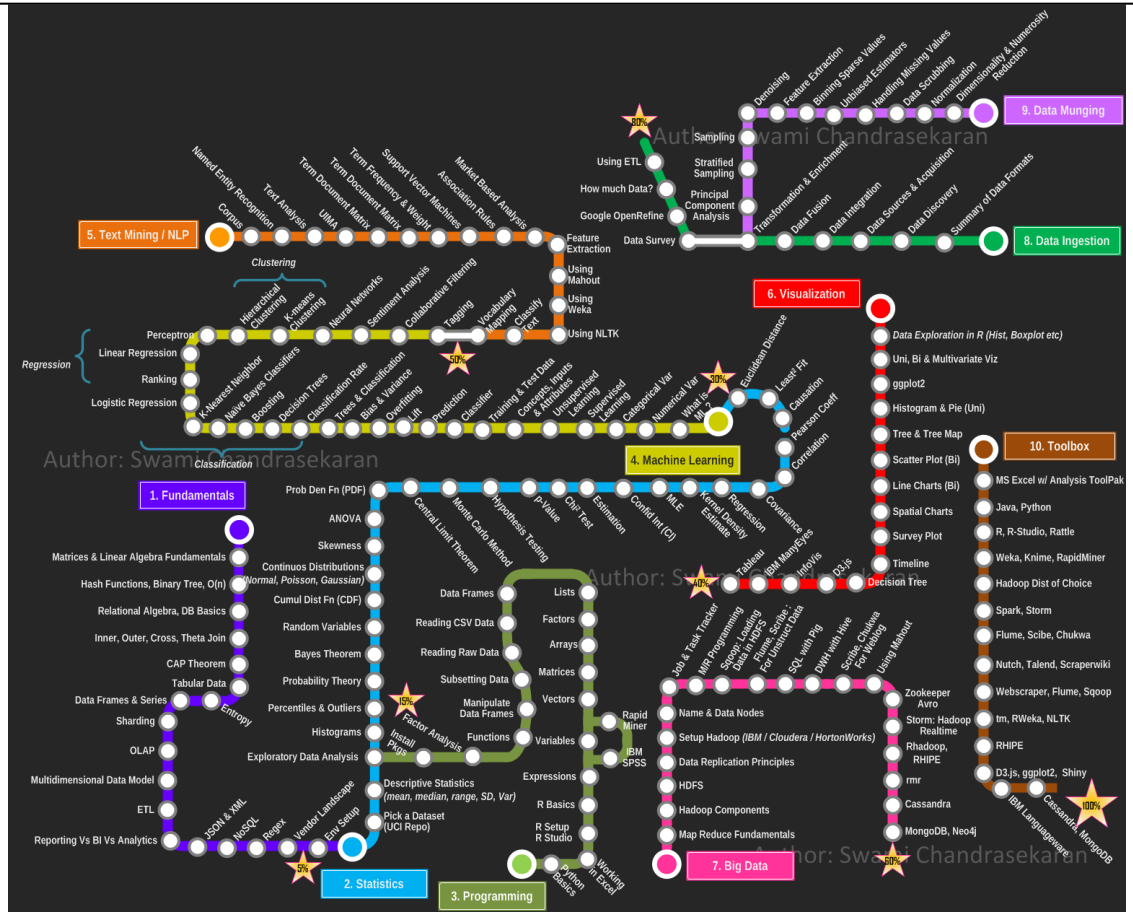
KCBO – A Bayesian Data Analysis Toolkit in python

How to Think Like a Computer Scientist – Learn Python via interactive tutorial: This interactive python textbook is designed by Luther College.

Basic Data Analysis and More: A Guided Tour Using Python

Think Stats: Probability and Statistics for Programmers (with Python)

BECOMING A DATA SCIENTIST



WHAT'S NEXT?

Some discussion questions:

- In what contexts should click data be handled in real time?
- Which is better: good data or good models? Is there a universal good model? Are there any models that are definitely not so good?
- How would you improve a spam detection algorithm that uses Naive Bayes?
- CTRs for ads have been surprisingly low this week. What steps would you take to find out why?
- What would be some good metrics for a monthly subscription service or product?

PROJECT PRESENTATIONS

PROJECT PRESENTATION

Suggested template:

Problem statement:
Source of data
Approach
Conclusions
Future Work

PROJECT PRESENTATION: RUBRIC

	Project		
		Score (1-5)	**Instructor /TA comments**
Presentation & communication: How well did the presenter tell the story?	How well did the presentation demonstrate a clear description of problem statement/ questions to answer? How focused were the project goals?		
	How well did the technical overview describe the algorithms and methods used?		
	How well did the presentor handle audience questions?		
Data Visualization: **Were visuals used at all?** **If so, how much did they contribute to telling the story?** **If not, should they have been?**	How effective were the data visualizations in furthering presenter's points? Were they easy to interpret?		
	Were chart or plot axes and components clearly labled?		
Quality of Data:	Did the project use a sufficient amount of data? Was it sufficiently robust?		
	If applicable, were missing or null values addressed?		
Quality of Model:	How well is the statistical model / implementation described? Is it based on reliable assumptions?		
	Have the model outputs been tested independently of the model (i.e cross-validation)?		
	Is there good reasoning for explaining the statistical methods that were tried and rejected?		
Overall	Rate the overall quality of the project		