Data Science in 7 Steps

Mário Antunes

June 23, 2017

Instituto de telecomunicações
Universidade de Aveiro

About me

Name: Mário Antunes

Occupation: PhD Student, researcher, jack of all trades

Areas of interest: Artificial intelligence, Machine learning, text mining, stream mining, IoT, M2M, ...

DS Portugal

- Community of data science enthusiasts.
- Mission: share knowledge between peers informally.
- Several successful meetups in the past.
- Everyone is welcomed to join and share knowledge.

Terminology

Dataset organized set of examples, typically composed of features and labels

Feature single property of an example (input variable)

Label classification category of an example (output variable)

Example single instance of a dataset

Outline

1. Step 0 - Demystifying machine learning
2. Step 1 - Basic Introduction
3. Step 2 - Acquiring data
4. Step 3 - Preprocessing data
5. Step 4 - Learn a machine learning model
6. Step 5 - Evaluate a machine learning model
7. Step 6 - Profit
8. Step 7 - Advance topics and discussing

Step 0 - Demystifying machine learning

Demystifying machine learning

- I do not like equations!

Demystifying machine learning

- I do not like equations!
- Never had formal training!

Demystifying machine learning

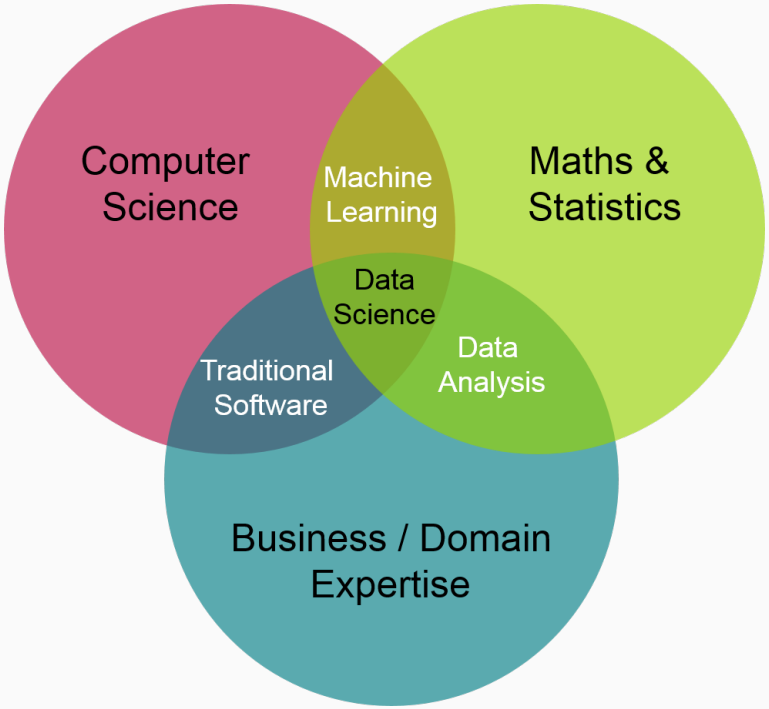
- I do not like equations!
- Never had formal training!
- It is such a complex topic!

Demystifying machine learning

- I do not like equations!
- Never had formal training!
- It is such a complex topic!
- Can I be a data scientist?

Demystifying machine learning

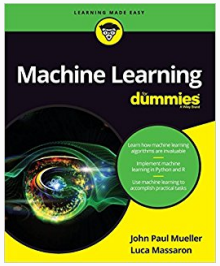
- I do not like equations!
- Never had formal training!
- It is such a complex topic!
- Can I be a data scientist?
- Yes!

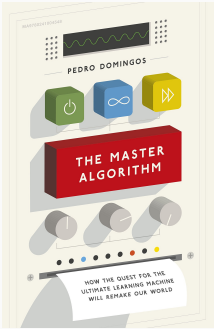


How to start?

- Several tools, frameworks and libraries ready to use
- Books, on-line courses, stack-overflow
- Meeting groups (remember DSPT)
- More than just a trend it is a necessity

Materials





coursera

Step 1 - Basic Introduction

Five main learning techniques

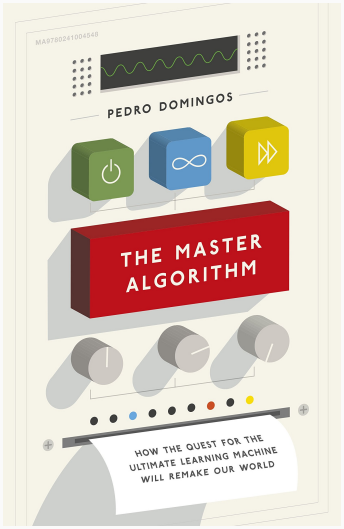
Induction symbolic reasoning

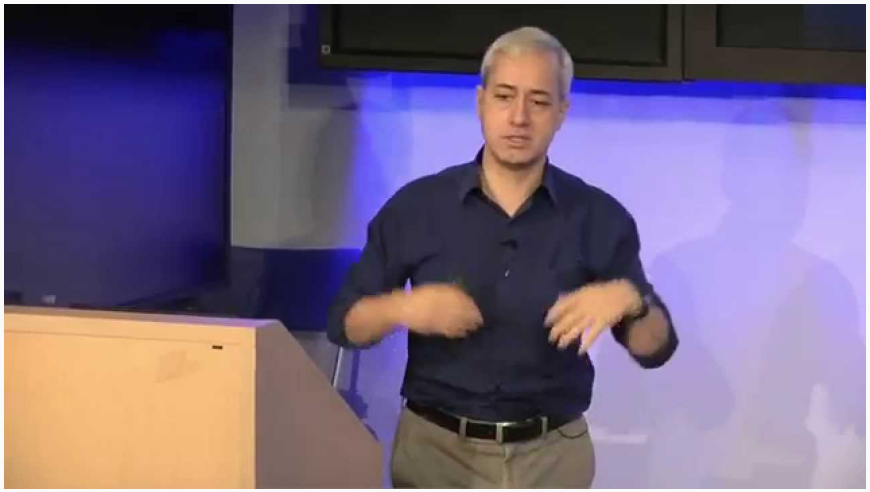
Neural Networks connections modelled on brain's neurons

Evolutionary algorithms learn from random generations (genetic algorithm)

Bayesian inference probabilistic models based on bayes' theorem

Analogy learns by finding *similar* examples



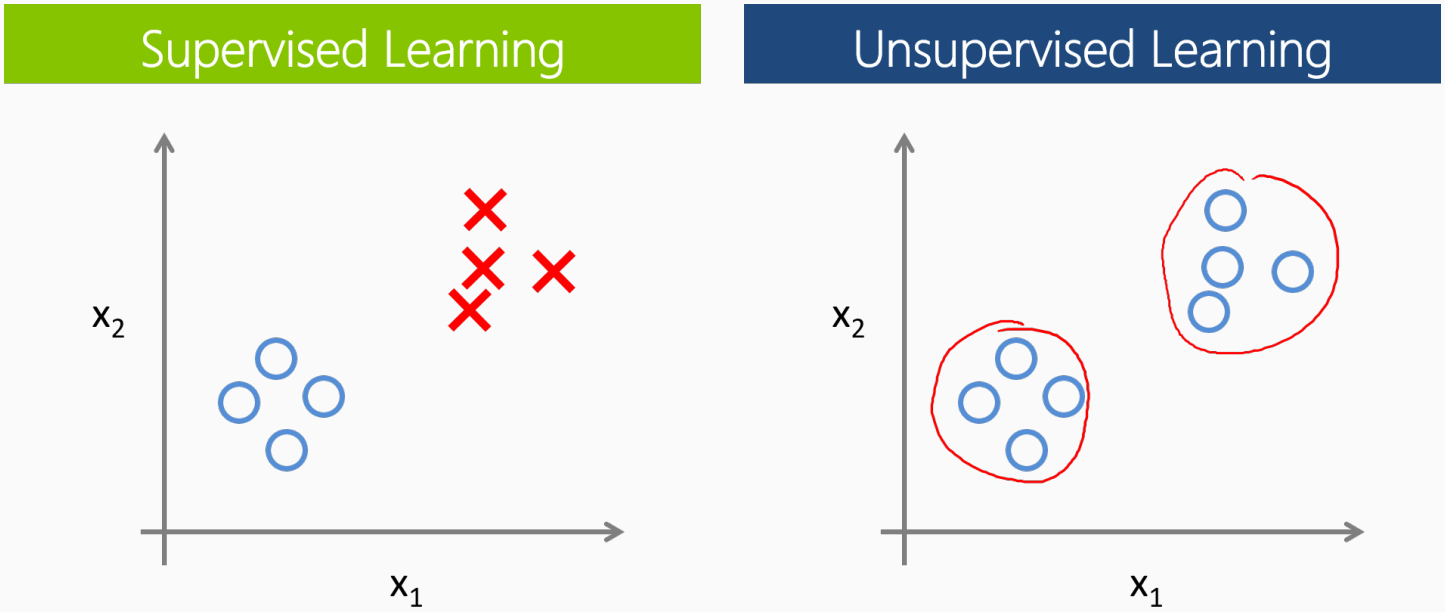


Taxonomy

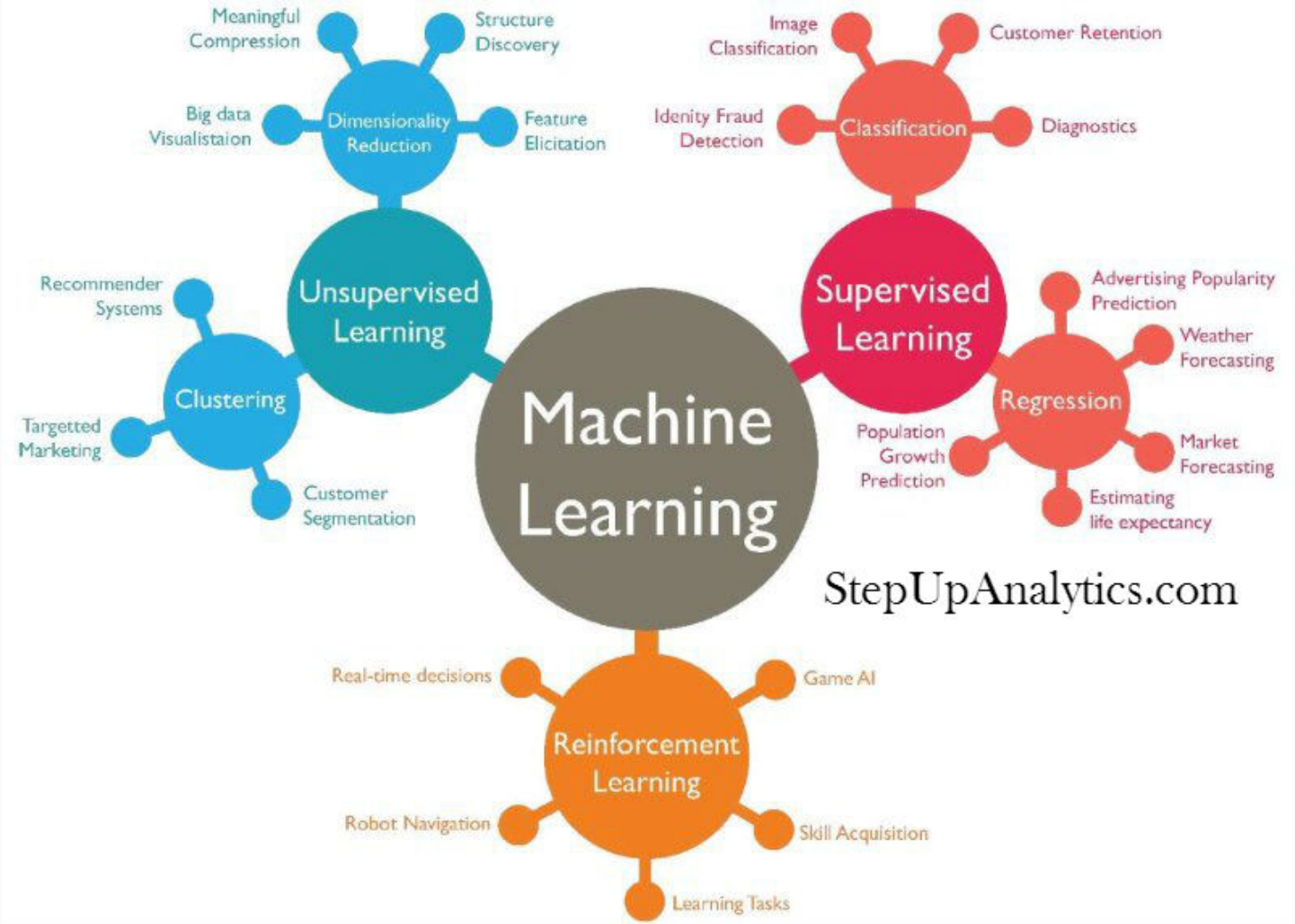
Supervised learning algorithm learns from input and output data

Unsupervised learning algorithm learns with input data only

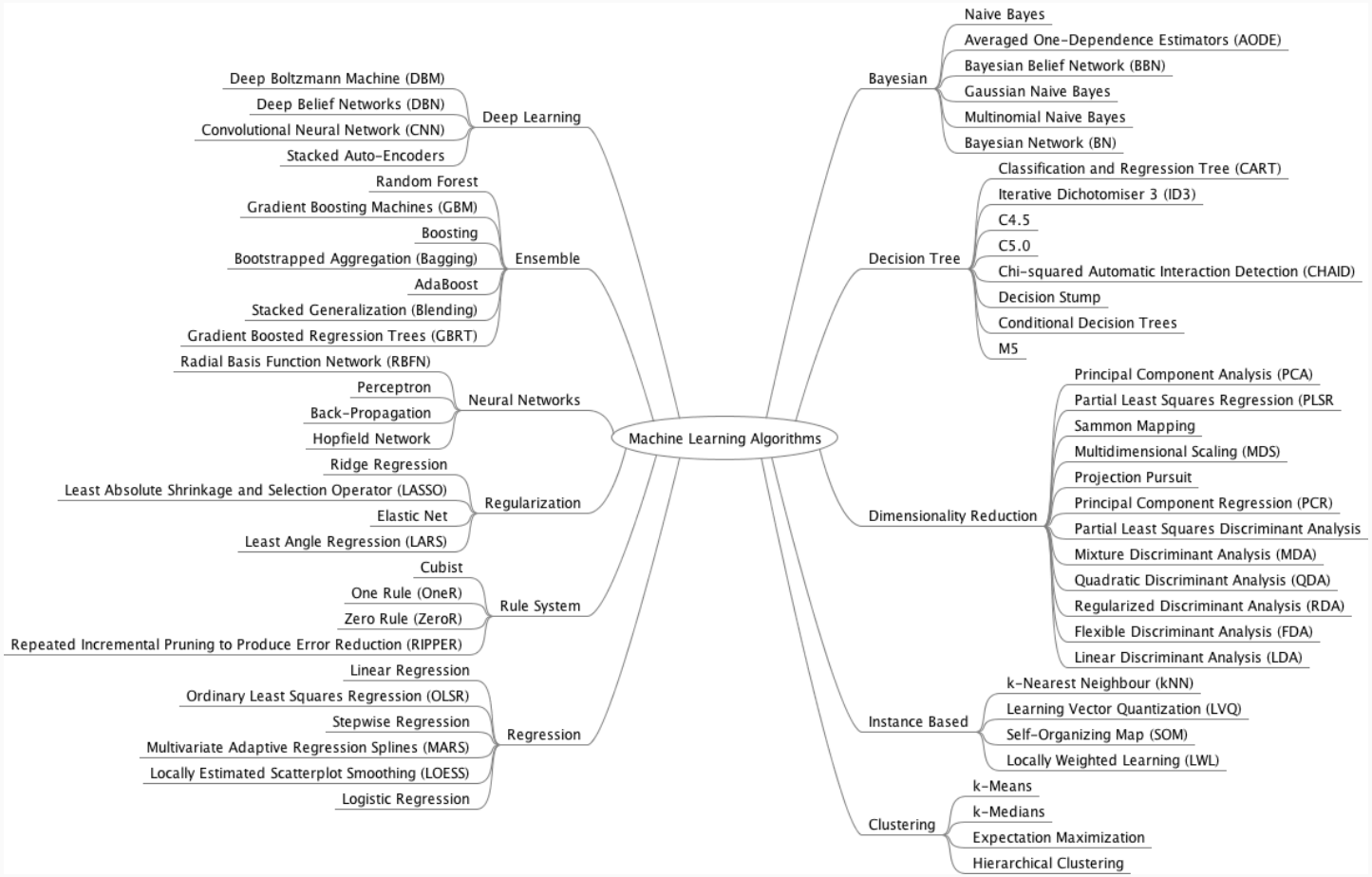
Reinforcement learning algorithms learns based on a reward system



Taxonomy

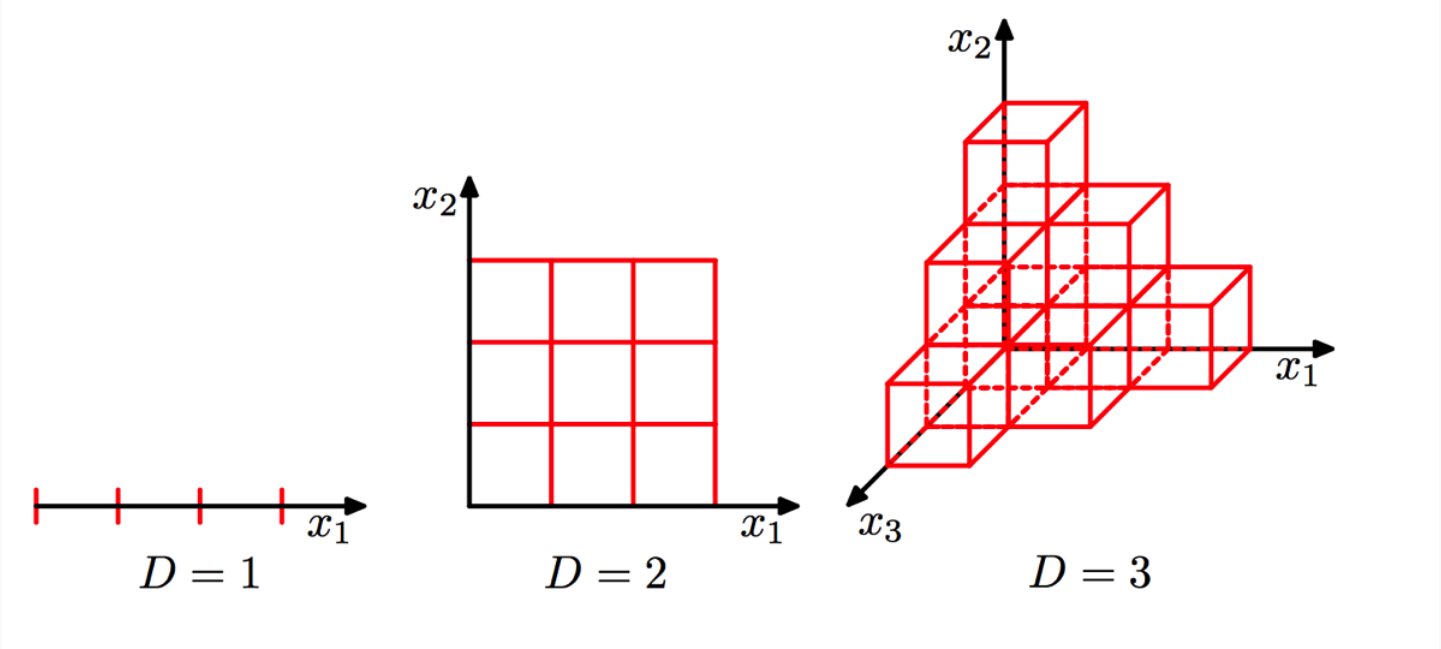

Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
StepUpAnalytics.com
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Taxonomy


Machine Learning Algorithms
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering

Curse of Dimensionality

- As the number of features (or dimensions) grows, the amount of data we need to learn accurately grows exponentially.



No free lunch

"No Free Lunch" :(

D. H. Wolpert. The supervised learning no-free-lunch theorems. In Soft Computing and Industry, pages 25–42. Springer, 2002.

Our model is a simplification of reality

⇩

Simplification is based on assumptions (model bias)

⇩

Assumptions fail in certain situations

Roughly speaking:

"No one model works best for all possible situations."

Step 2 - Acquiring data

Acquiring data

- The most important step in data science
- Google's Research Director Peter Norvig said:

  We don't have better algorithms. We just have more data.

- Unfortunately there is no recipe for data acquisition

Tips

- Public available dataset

Tips

- Public available dataset

- Gather data from web (scrapper, crawler, APIs)

Tips

- Public available dataset
- Gather data from web (scrapper, crawler, APIs)
- Gather data from sensors

Tips

- Public available dataset
- Gather data from web (scrapper, crawler, APIs)
- Gather data from sensors
- Better yet, known what you want to learn and gather the right data

Step 3 - Preprocessing data

Preprocessing data

- The world is a messy and noisy place

Preprocessing data

- The world is a messy and noisy place

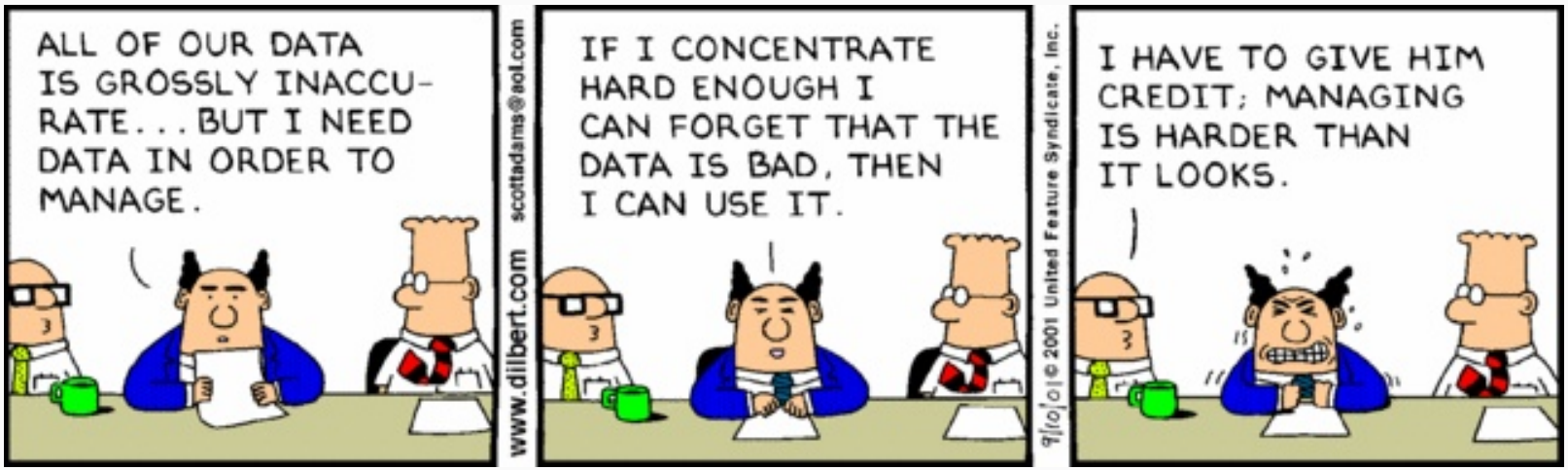
- As such the data that you acquire needs to be clean

Preprocessing data

1. Obtain meaningful data (called ground truth)
2. Acquire enough data for the task at hand (requires experience)
3. Organize data in the correct format
4. Deal with bad data
5. Create new features (advanced)

Deal with bad data

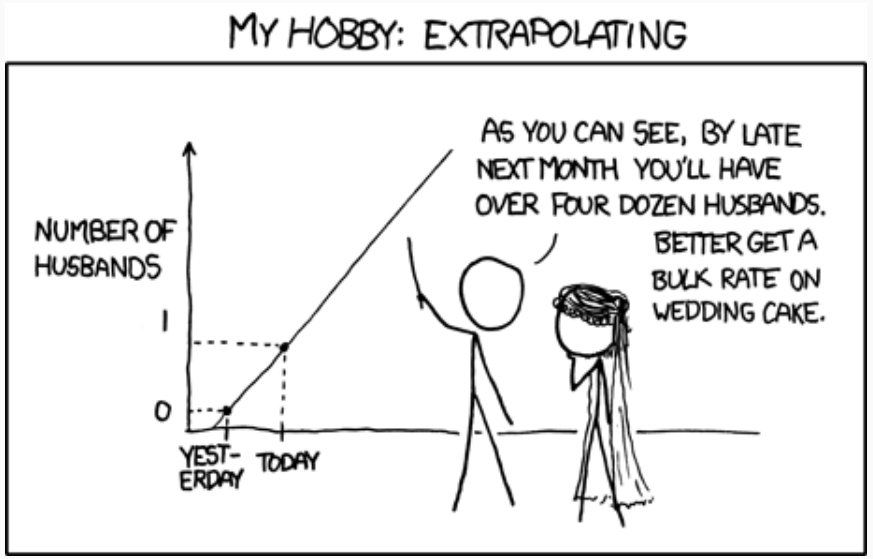
- Bad data is not criminal, just refers to:
  1. Mislabel or missing data
  2. Redundancy of information
  3. Outliers



Missing data

- When a feature has more than 90% examples missing, just drop it.
- Replacement strategies:
  - Replace with a computed value:
    - Mean, median, most common
    - Value outside of range
    - $\emptyset$
  - Interpolate missing value

"Well, this certainly explains much of the company's missing data. Who else thought the 'DEL' key on their computer was for delegating work?"



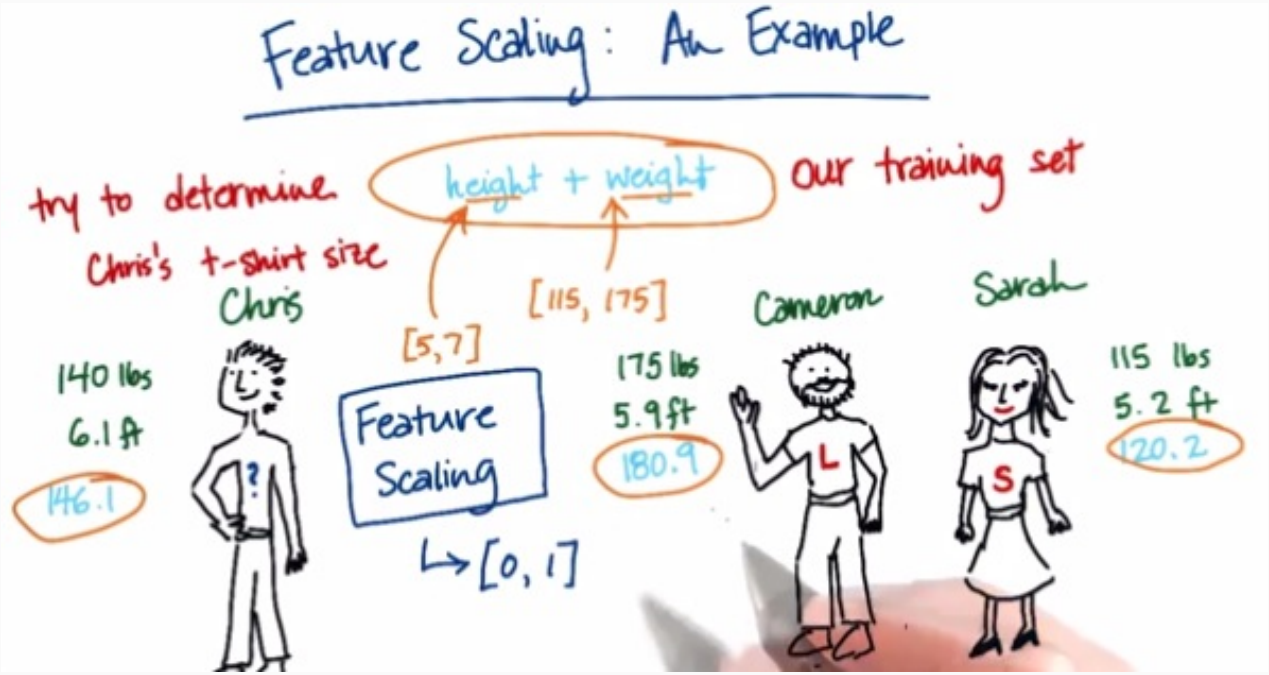
Transform Distributions

- Normalization:

$$x_i = \frac{x_i - min(X)}{max(X) - min(X)}$$

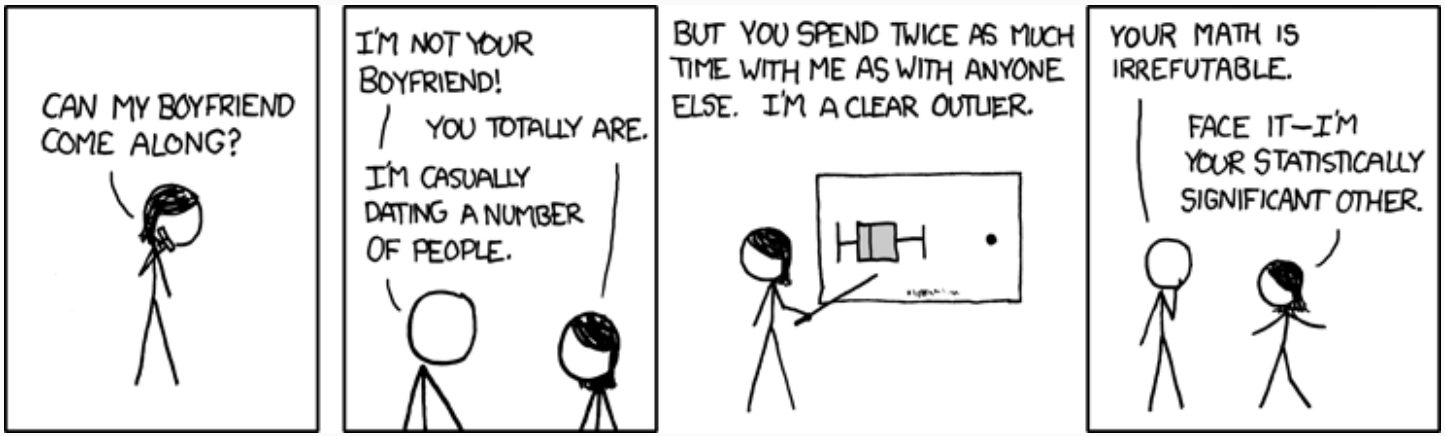
- Z-score normalization:

$$x_i = \frac{x_i - \mu}{\sigma}$$



Removing outliers

- Plot examples and manual remove outliers (small datasets)
- Remove [1%, 5%, 10%] of the extremes
- Use unsupervised learning methods to cluster data

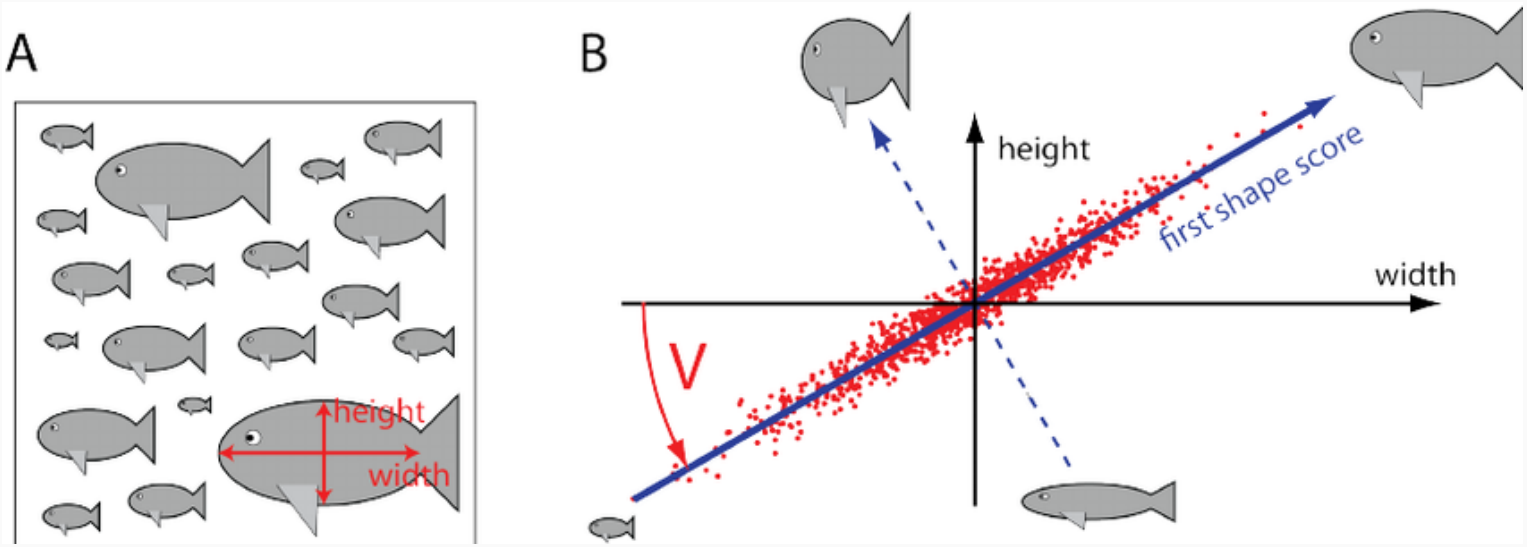


Create features automatically (Advanced)

- Polynomial expansion:

$$(x + y)^2 = x^2 + 2xy + y^2$$

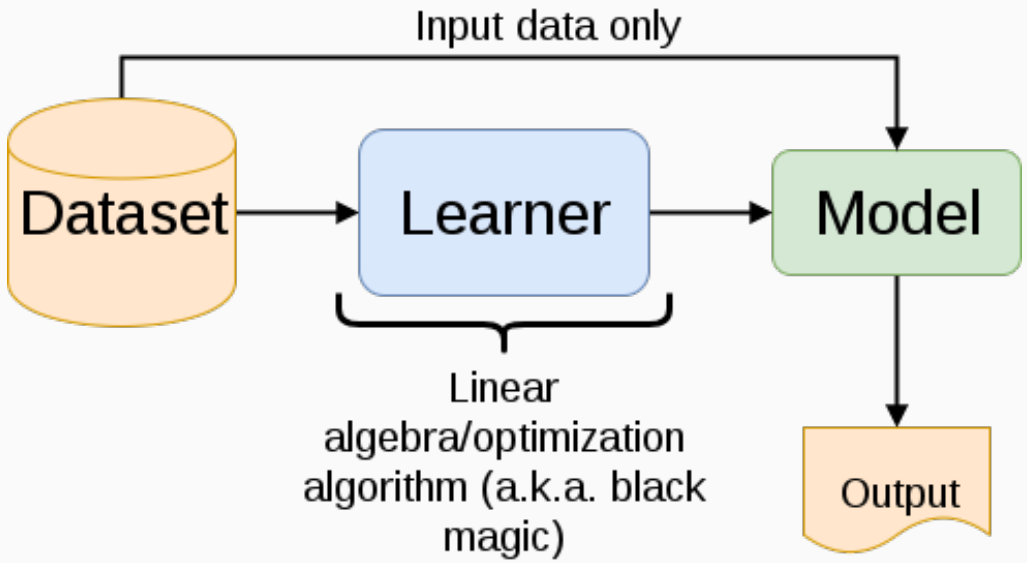
- Principal Component Analysis (PCA):



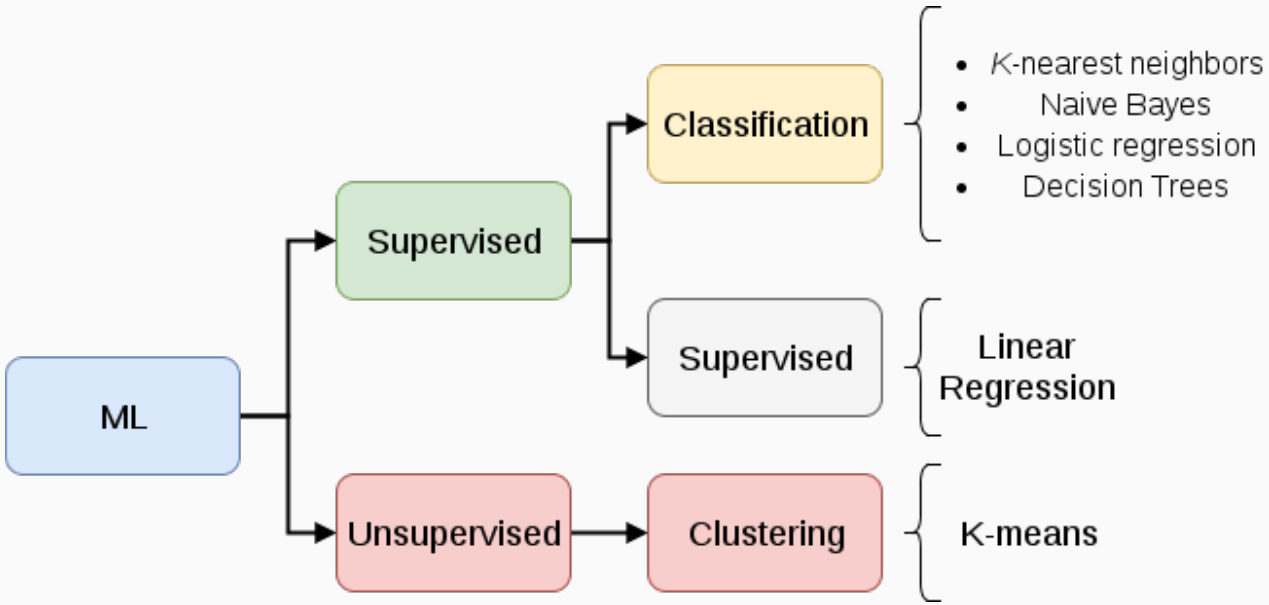
Step 4 - Learn a machine learning model

Learn a machine learning model

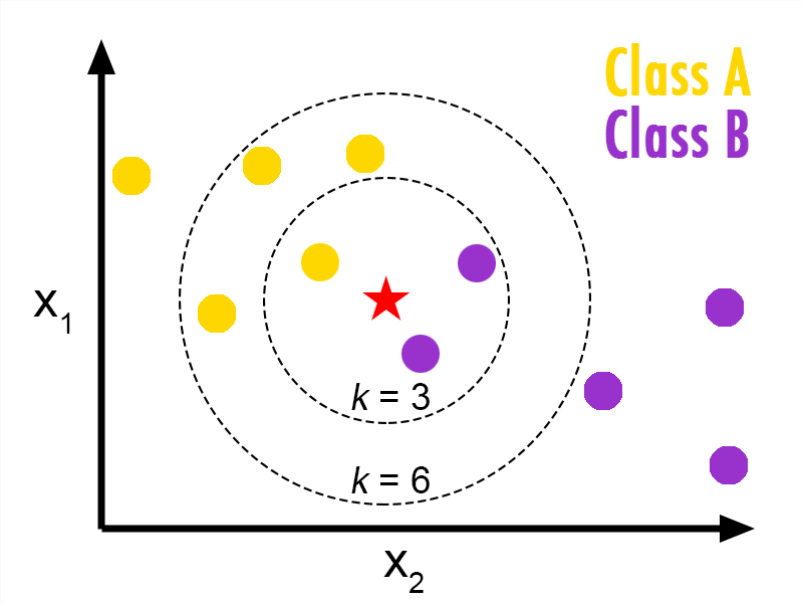
- There is a lot going on inside the learners code
- But you do not have to worry about it
- Just take into account:
  - Different types of learners (with different assumptions)
  - Different types of data (numerical, stream, images, stock...)



Taxonomy


ML
Supervised
Unsupervised
Classification
Supervised
Clustering
K-nearest neighbors
Naive Bayes
Logistic regression
Decision Trees
Linear Regression
K-means

K-nearest neighbour

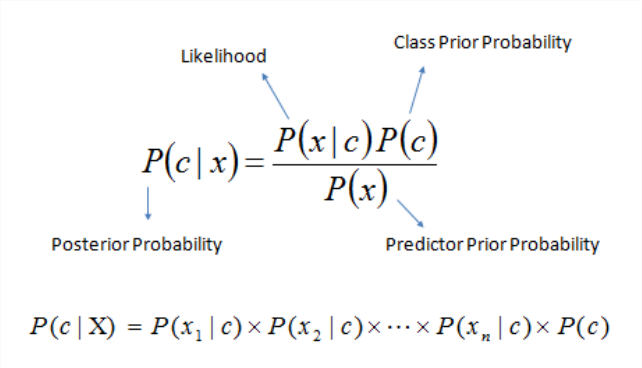


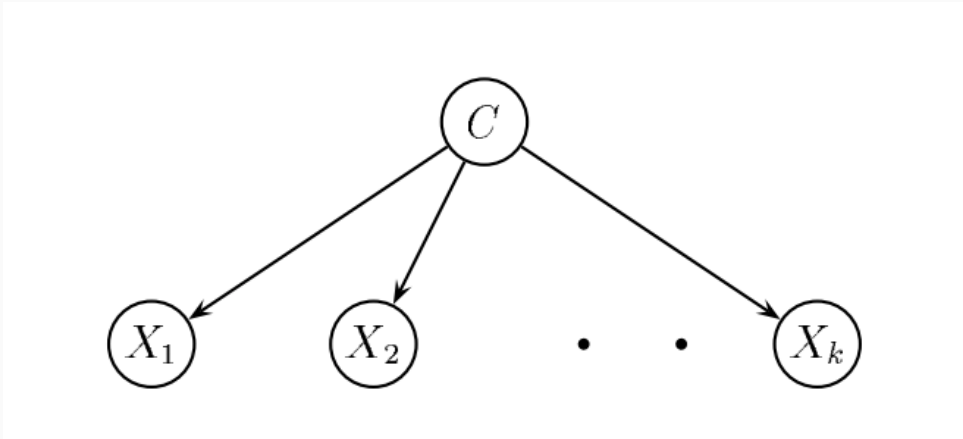
- Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_1)^2}$
- Manhattan $\sum_{i=1}^{k}|x_i - y_1|^2$
- Minkowski $\left(\sum_{i=1}^{k}|x_i - y_1|^q\right)^{\frac{1}{q}}$

Naive bayes

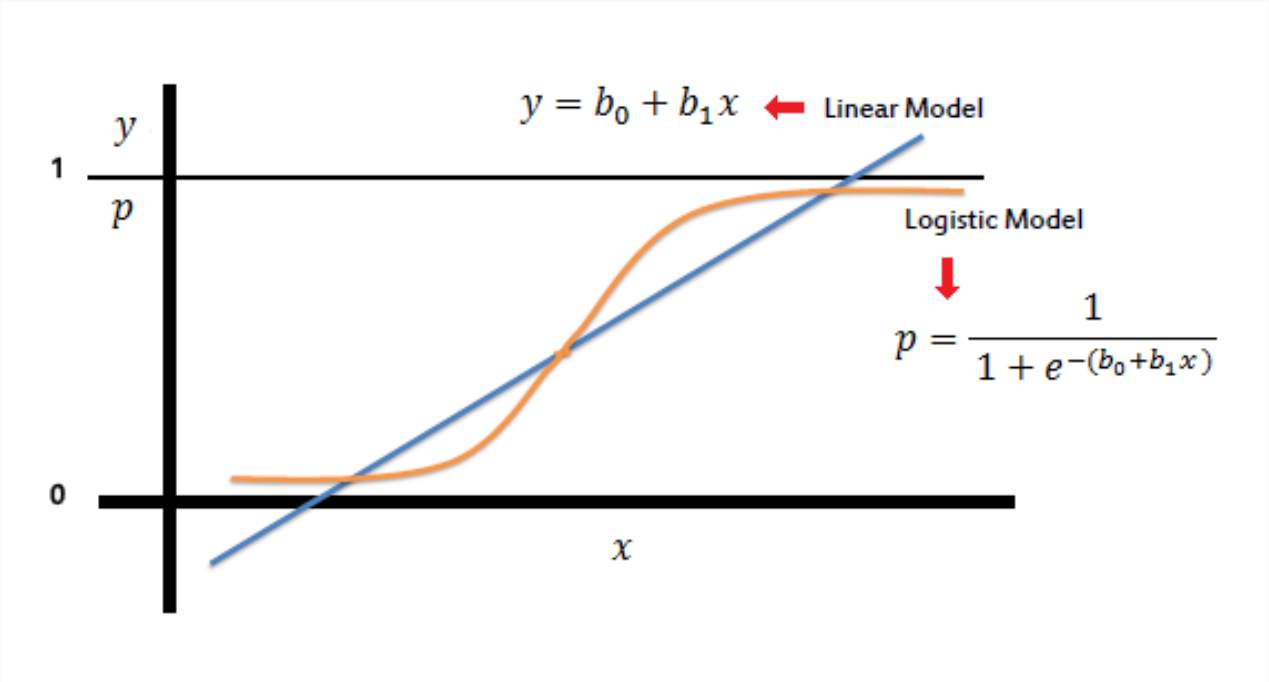
- George E. P. Box said:

  All models are wrong; some models are useful.

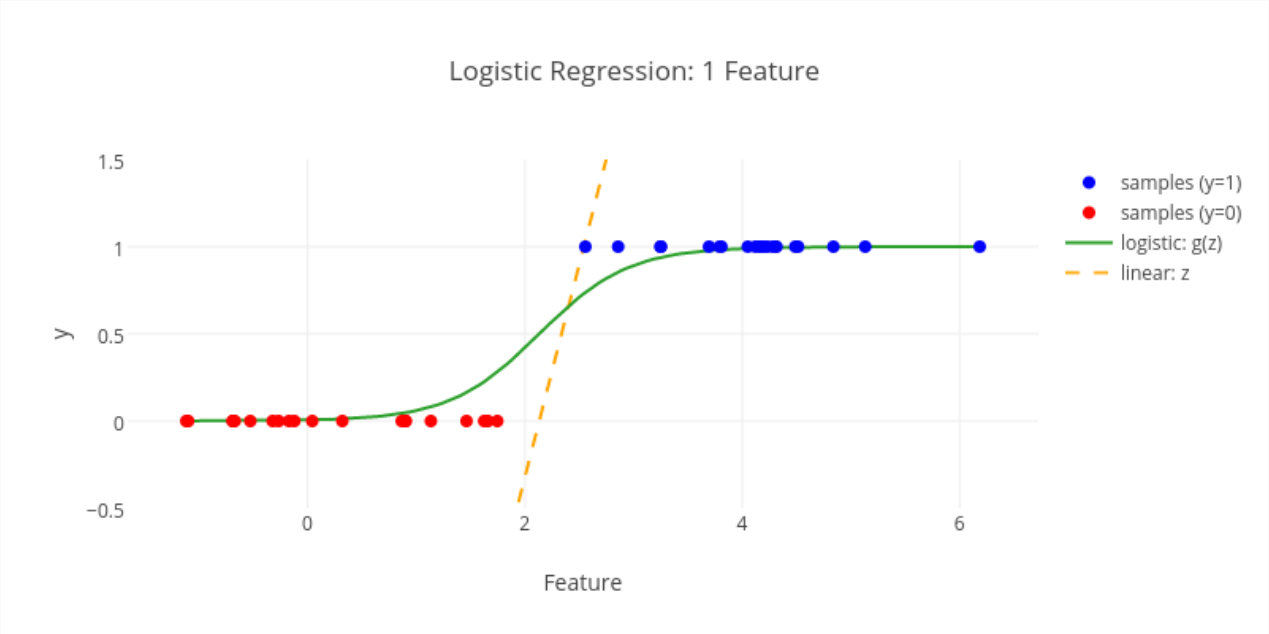




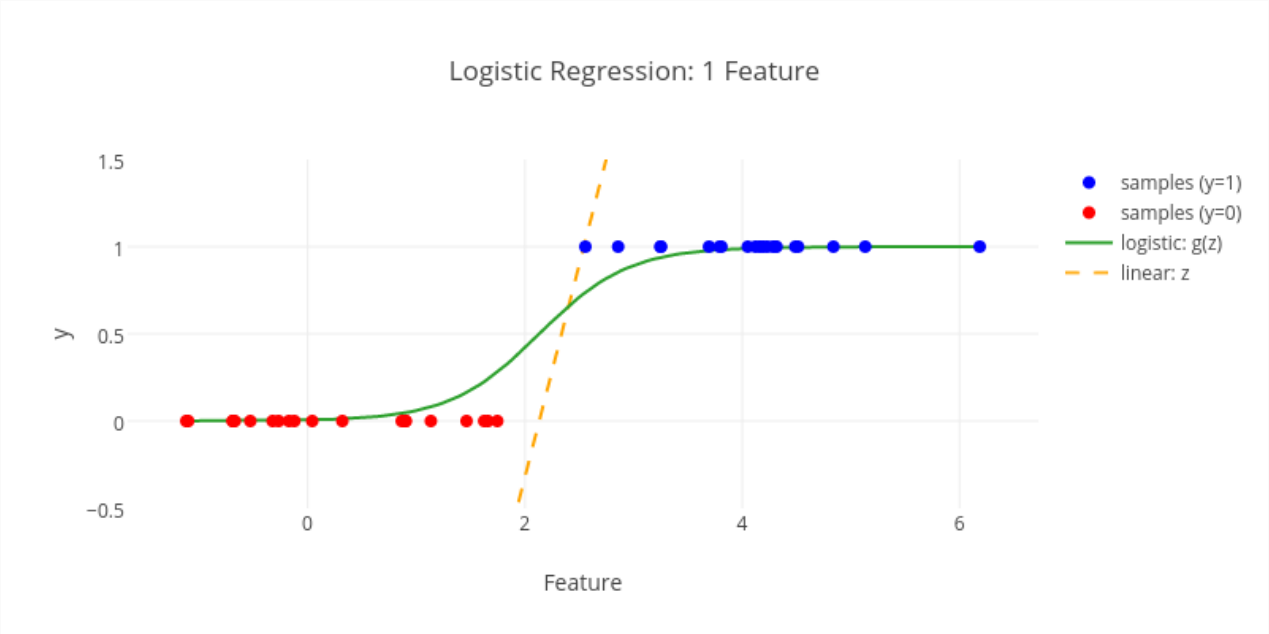
Logistic regression



Logistic regression

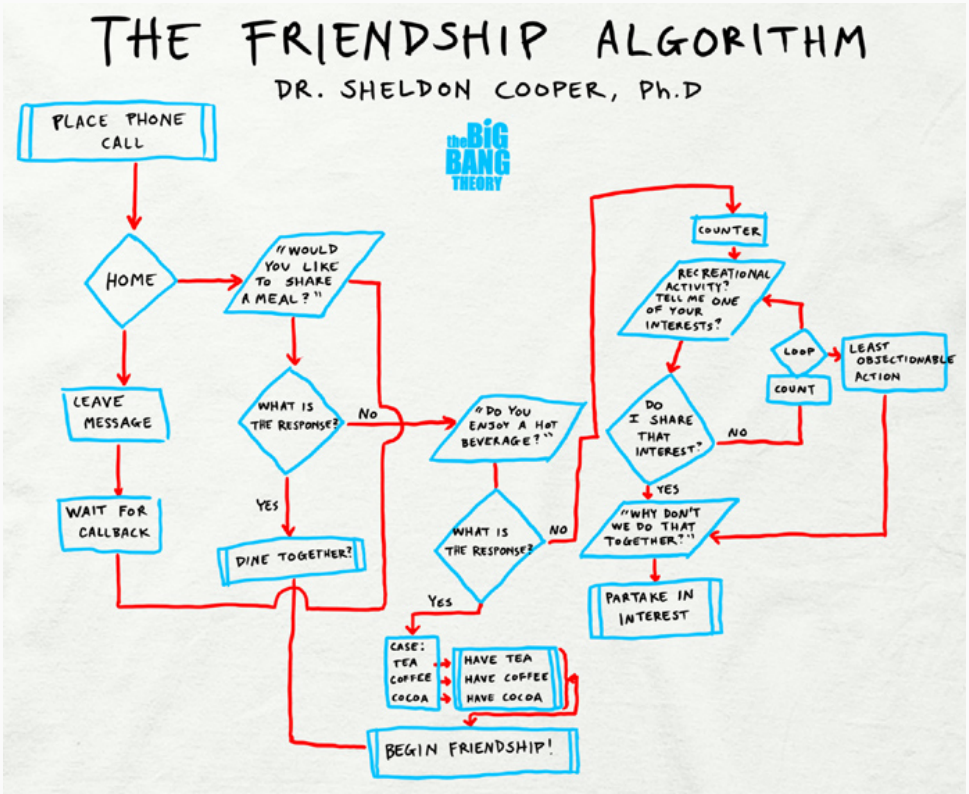


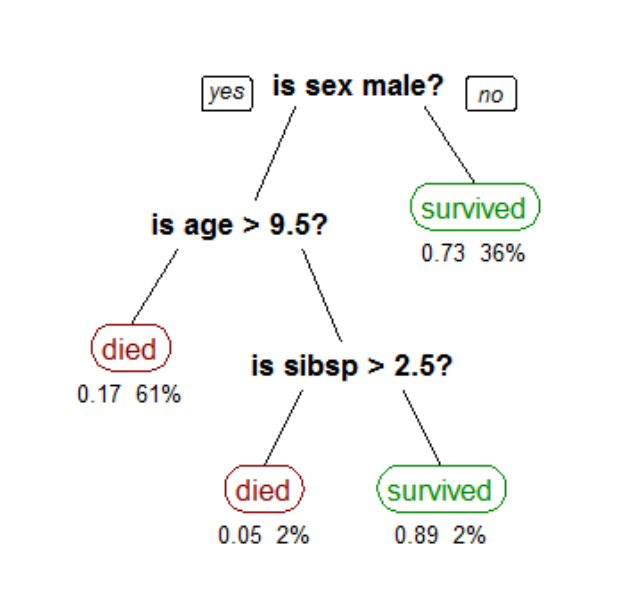
Logistic regression



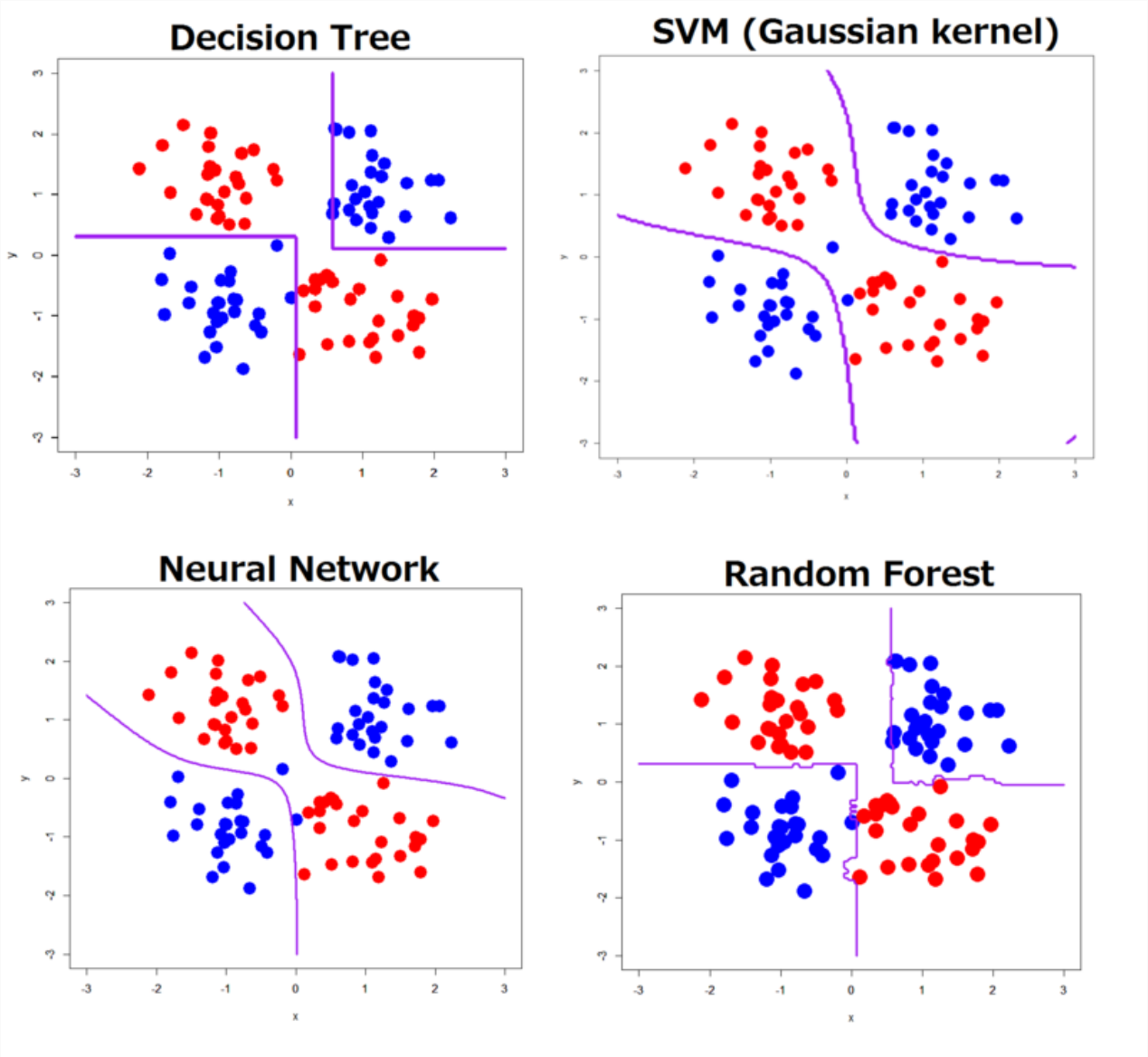
- Support Vector Machine (SVM) and Neural Networks are extensions of this...

Decision Trees

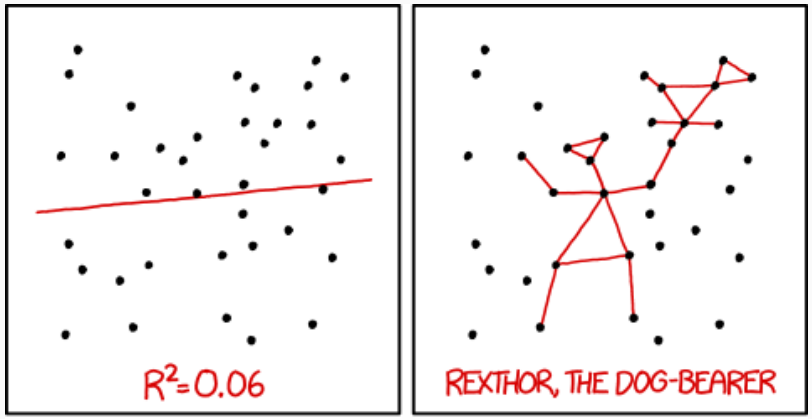




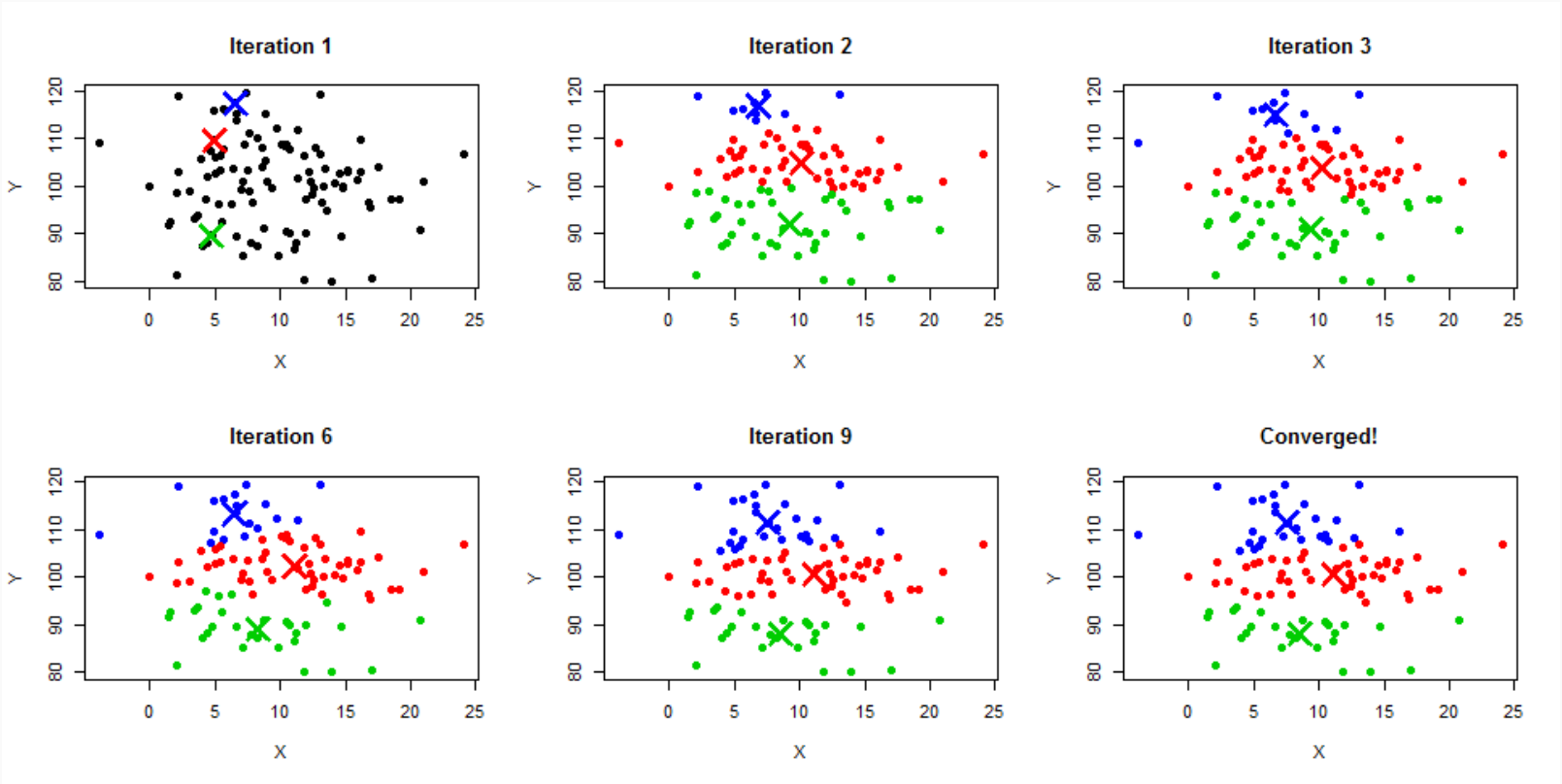
Model decision boundary



Linear regression



K-means

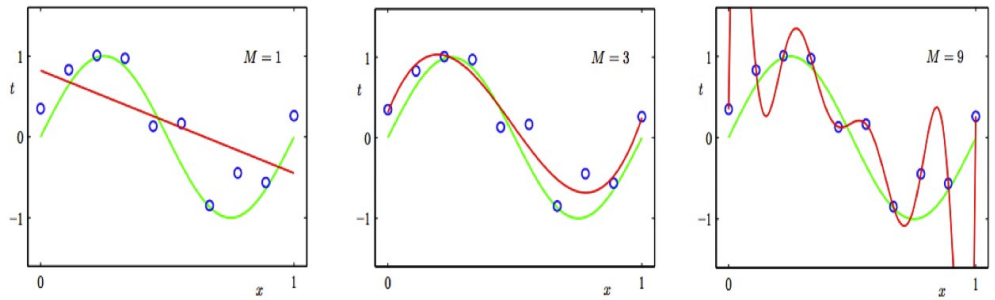

Iteration 1
Iteration 2
Iteration 3
Iteration 6
Iteration 9
Converged!
X
Y

Step 5 - Evaluate a machine learning model

Generalization, over-fitting, under-fitting

Under- and Over-fitting examples

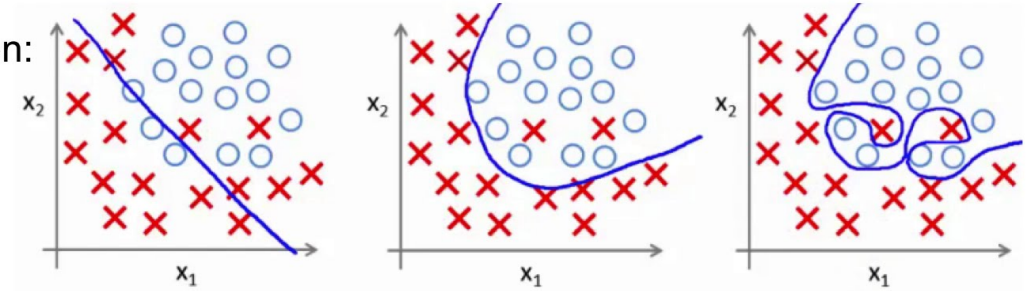
Regression:



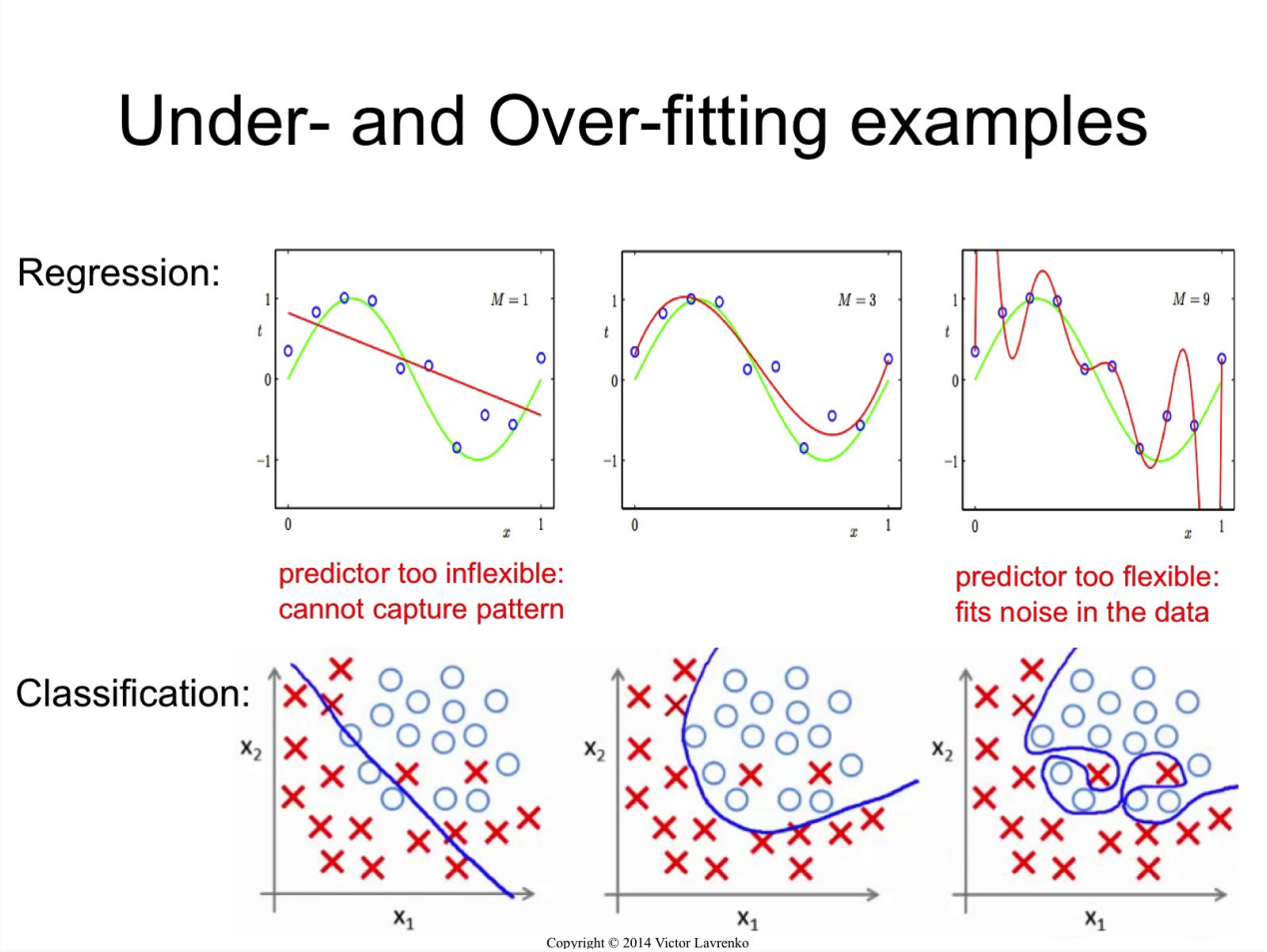
predictor too inflexible:
cannot capture pattern

predictor too flexible:
fits noise in the data

Classification:

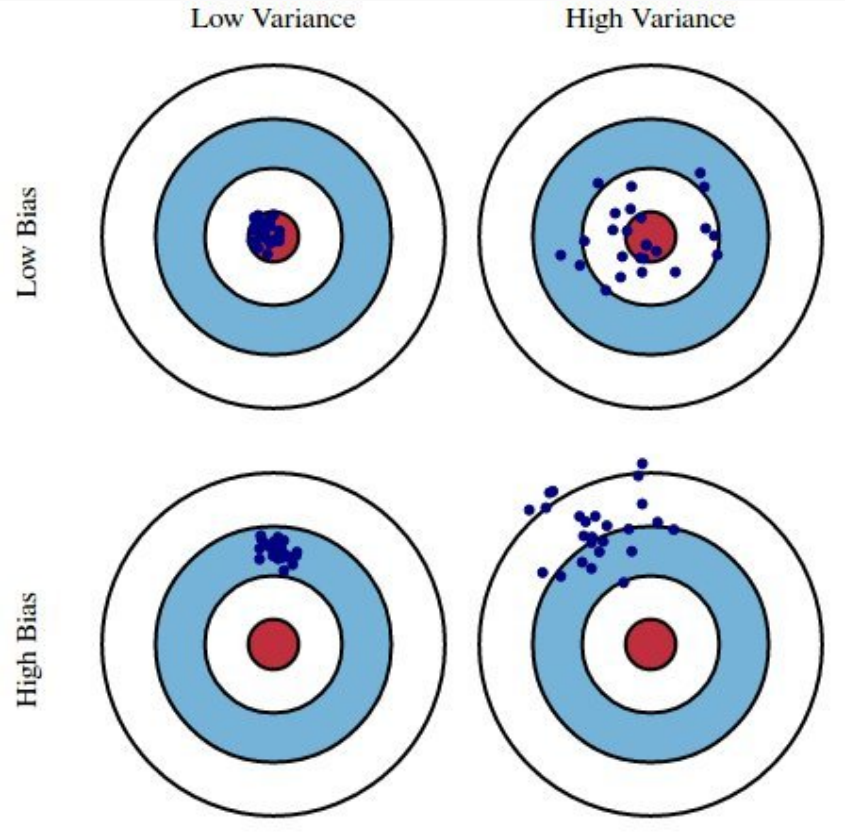


Generalization, over-fitting, under-fitting



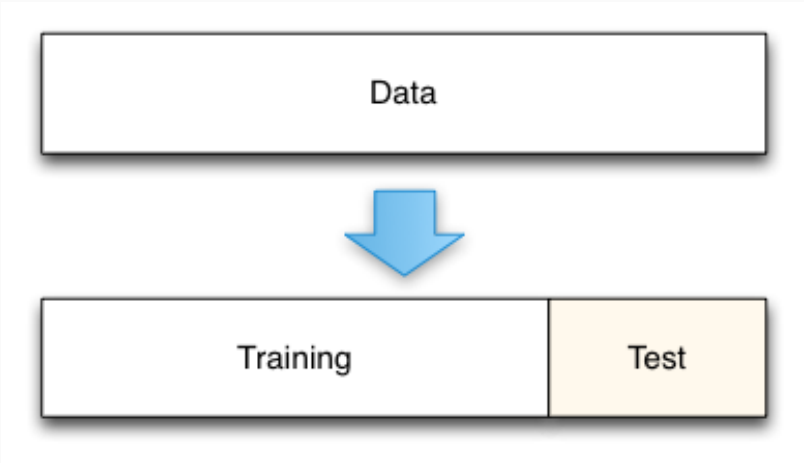
- *Generalization* = *min*(*overfitting* + *underfitting*)
- Over-fitting = high variance
- Under-fitting = high bias

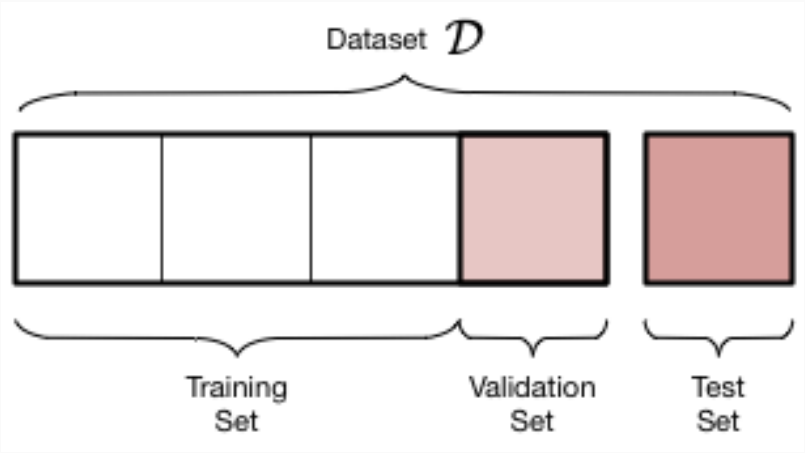
Bias and variance



Validation

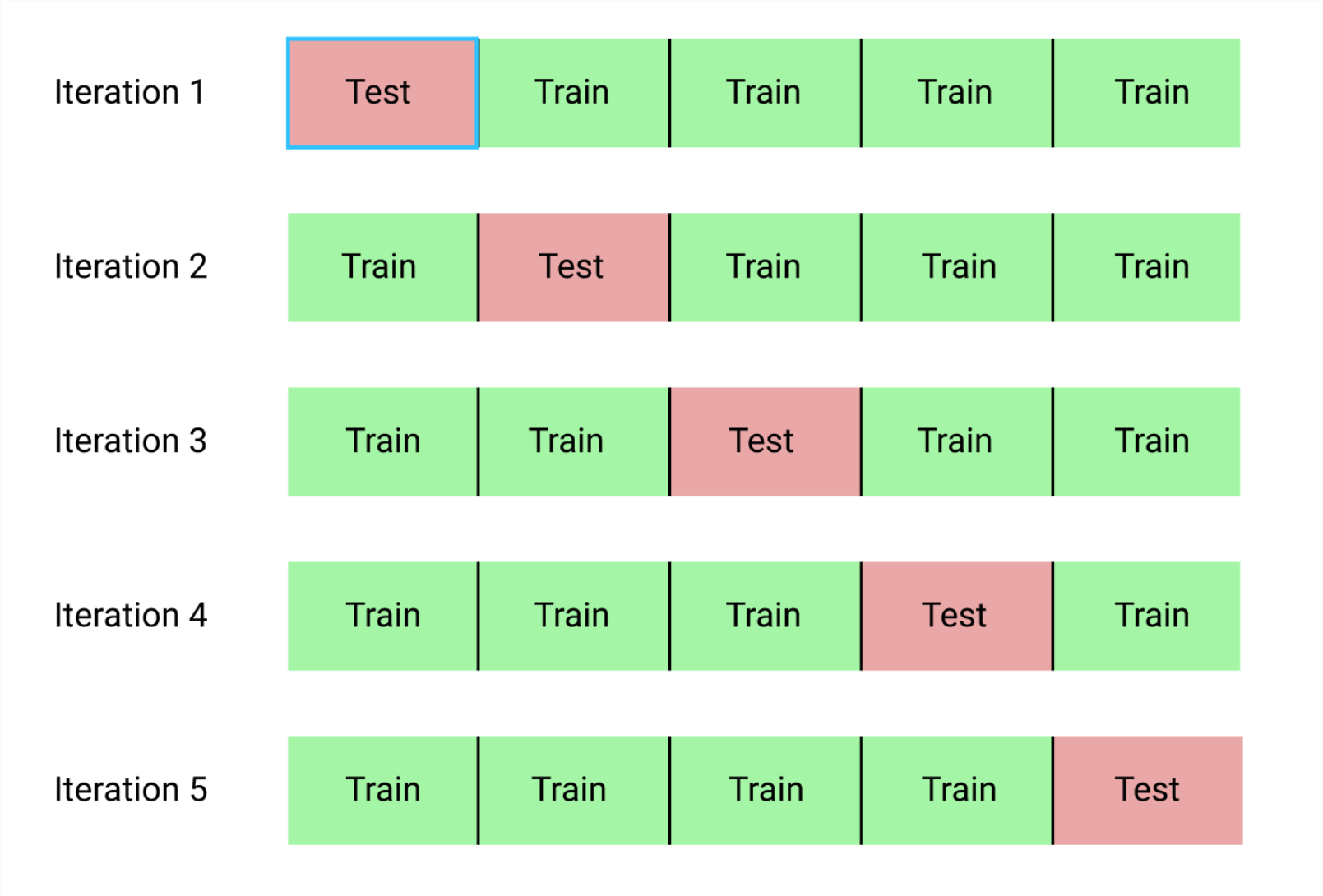
- Split-sample or holdout validation





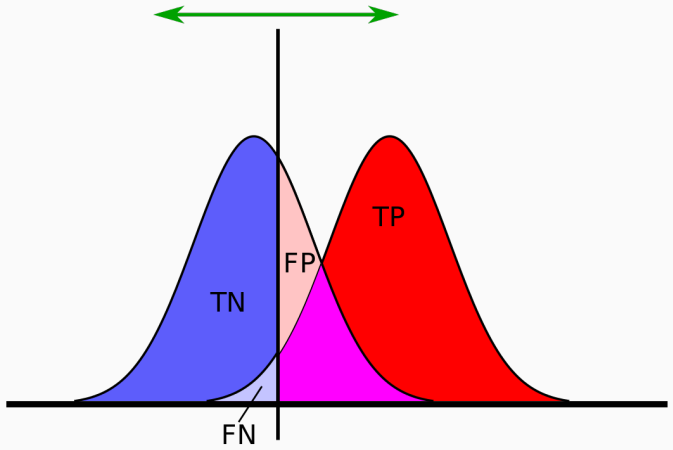
Validation

- K-fold cross validation

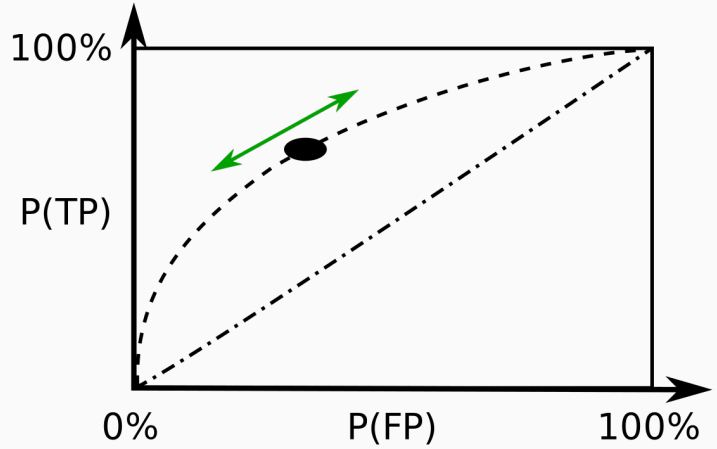


Metrics

- Classification Accuracy
- Area Under ROC Curve
- Confusion Matrix



TP	FP
FN	TN



Step 6 - Profit

Profit

What can you do with machine learning?

Users' Similarity

Group users from social media based on their similarity.

Step 7 - Advance topics and discussing

Users' Similarity

Are we doomed?