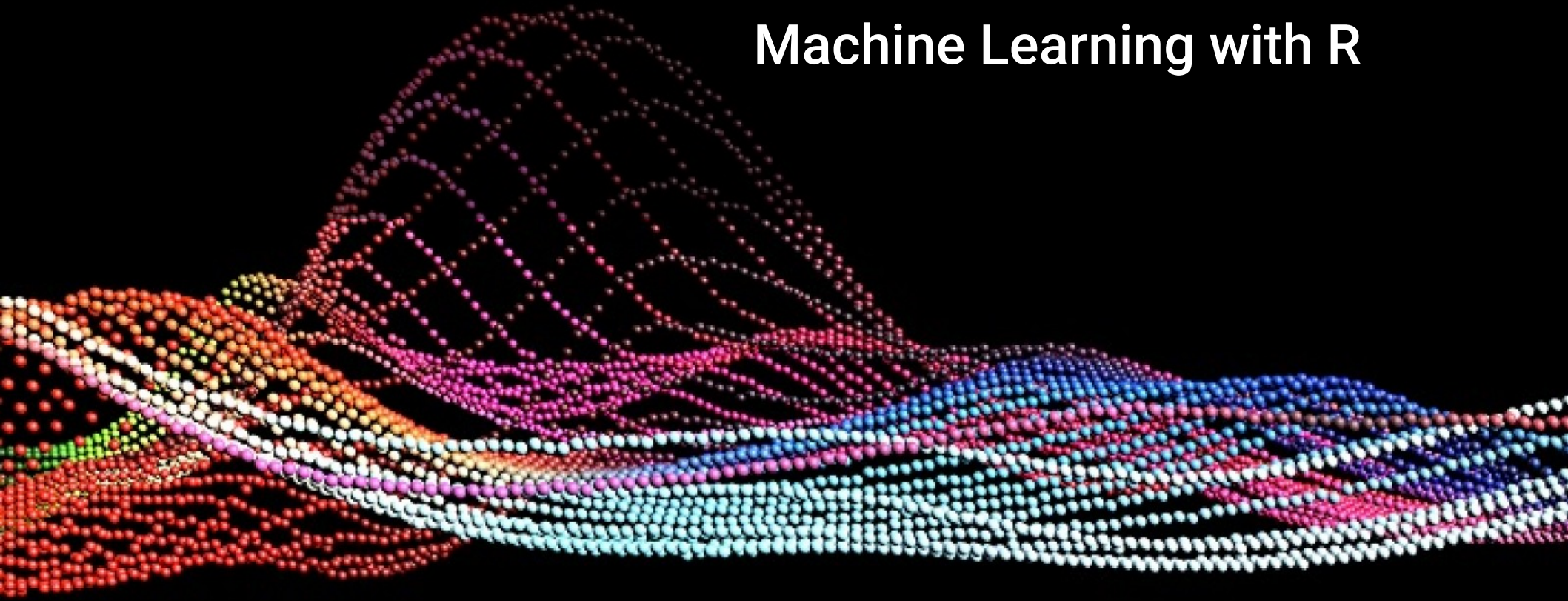
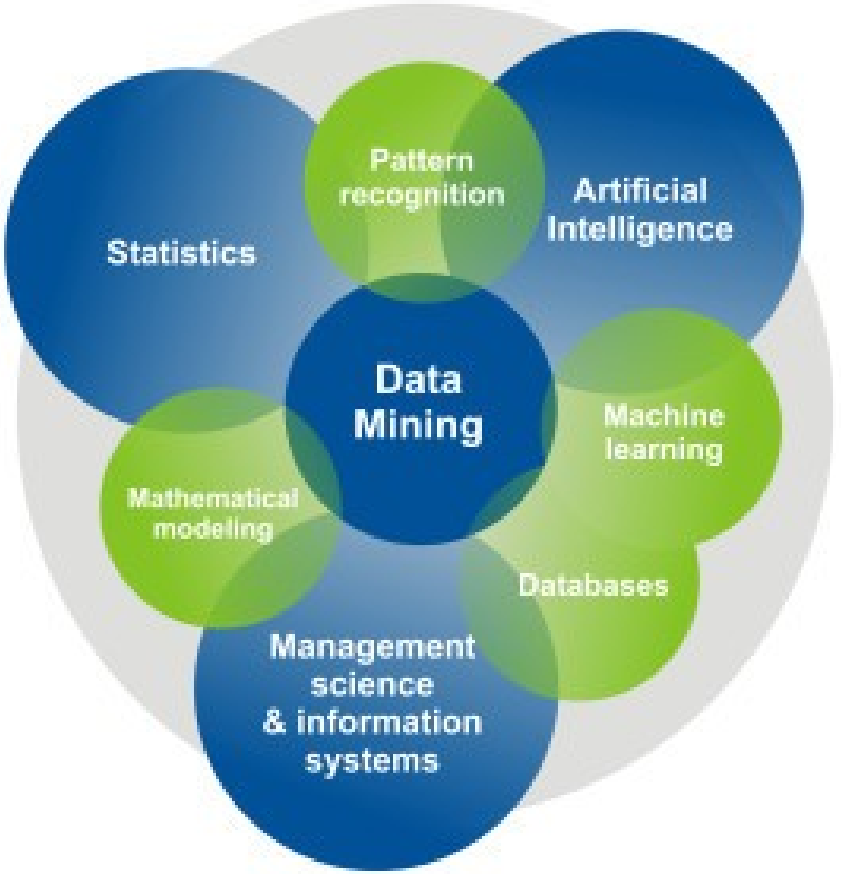
Machine Learning with R



Machine Learning and Data Mining



- No clear distinction between DM and ML
- The same tools, processes and models are used in both
- It’s not a subject matter distinction but rather a contextual one

Machine Learning and Data Mining

- Data Mining:
  - Search for non-obvious correlations in large data sets
- Machine Learning
  - Developing models to make data based decisions
- Many of the same algorithms are used in both
  - Many DM algorithms are called unsupervised learning when used in ML

Machine Learning and Data Mining

- Characterized as “brute force” methods
  - Process a lot of data using a lot of computing power
- Ideas were developed in the 1960-80s
  - Impractical because they were computationally intensive
  - The hardware of the 20th century could not deliver
- From this point on, we will include DM under the category of unsupervised ML

Cognition

*Systems that **think** like humans*

- emulate human thought processes
- can pass a generalized Turing test
- solve problems, creativity, generate ideas
- usually depicted in movies and fiction
- usually referred to as "cognitive science"

*Systems that **reason** rationally*

- machine learning
- expert systems
- data analysis and modeling
- pattern recognition
- speech recognition and computer vision

Human

- interact with people as if it were a person
- autonomous activity
- interact with environment and learn
- adapt behaviour from experience
- handle new and unknown situations

Rational

- human-machine interfaces
- industrial and commercial robots
- task automation (eg. GPS and ABS in cars)
- monitoring and control systems
- guidance to humans in task performance

*Systems that **act** like humans*

Activity

*Systems that **act** rationally*

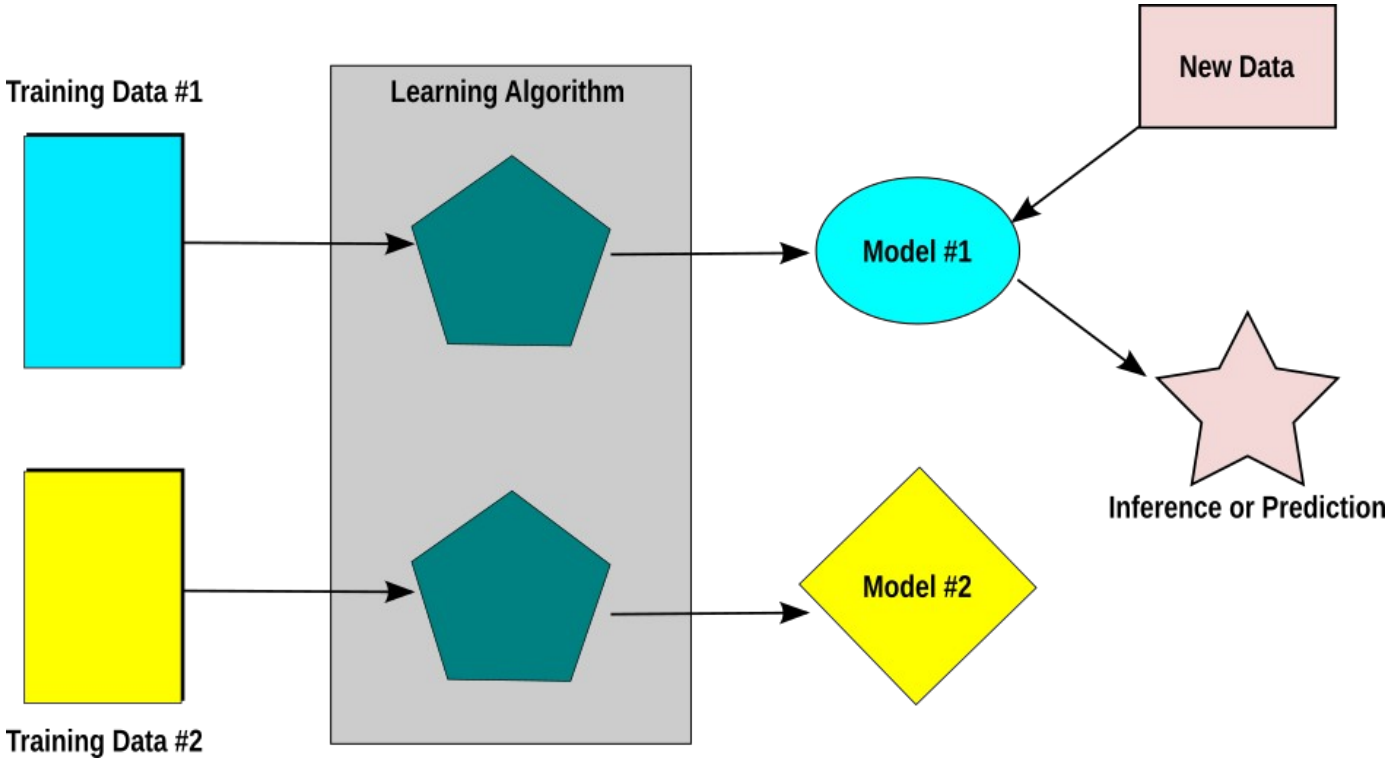
Terminology

- ML Model
  - An decision making algorithm developed by a ML *algorithm* that is *trained* on a data set
- An ML algorithm is a way of developing a model
  - Stochastic regression, decision trees, clustering algorithms etc
- Training - using a data set to develop a model by applying a ML algorithm
  - Industry standard procedures for training and testing
  - Each algorithm has “parameters” that can be adjusted to influence the shape of the model (and “hyperparameters that are guesses used to help choose parameters)

Terminology

- Training
  - Feeding data to an algorithm to create a model – different data produces different models
  - Computationally expensive from hours to weeks
- Prediction or Inference
  - Created model used to answer questions
  - Classification "do these medical results indicate cancer?"
  - "What are the recommended movies for this user"

Terminology



Terminology

Sample or input—One data point that goes into your model.

- *Prediction or output*—What comes out of your model.
- *Target*—The truth. What your model should ideally have predicted, according to an external source of data.
- *Prediction error or loss value*—A measure of the distance between your model's prediction and the target.
- *Classes*—A set of possible labels to choose from in a classification problem.
- *Ground-truth or annotations*—All targets for a dataset, typically collected by humans.

Terminology

- *Binary classification*—Each input sample should be categorized into two exclusive categories.
- *Multiclass classification*—Each input sample should be categorized into one of more than two categories
- *Multilabel classification*—Each input sample can be assigned multiple labels.
- *Scalar regression*—A task where the target is a continuous scalar value.
- *Vector regression*—A task where the target is a set of continuous values
- *Mini-batch or batch*—A small set of samples (typically between 8 and 128) that are processed simultaneously by the model.

Types of ML

- Reinforcement Learning
  - Used when developing “intelligent agents”
- Supervised Learning
  - Training takes place with datasets that have labelled targets
  - Model ranked on how close the computed results are to the targets
  - Deep Learning - Methods based on layered neural net approach
- Unsupervised
  - No labels: essentially data mining
  - Models evaluated on how much they converge to an answer

THE 4 TYPES OF
MARKET SEGMENTATION

GEOGRAPHIC

- Zip code/post code
- City
- Country
- Population density
- Distance from a certain location (like your office or store)
- Climate
- Time zone
- Dominate language

DEMOGRAPHIC

- Age
- Gender
- Income
- Occupation
- Family size
- Race
- Religion
- Marital Status
- Education
- Ethnicity

PSYCHOGRAPHIC

- Values
- Goals
- Needs
- Pain points
- Hobbies
- Personality traits
- Interests
- Political party affiliation
- Sexual orientation

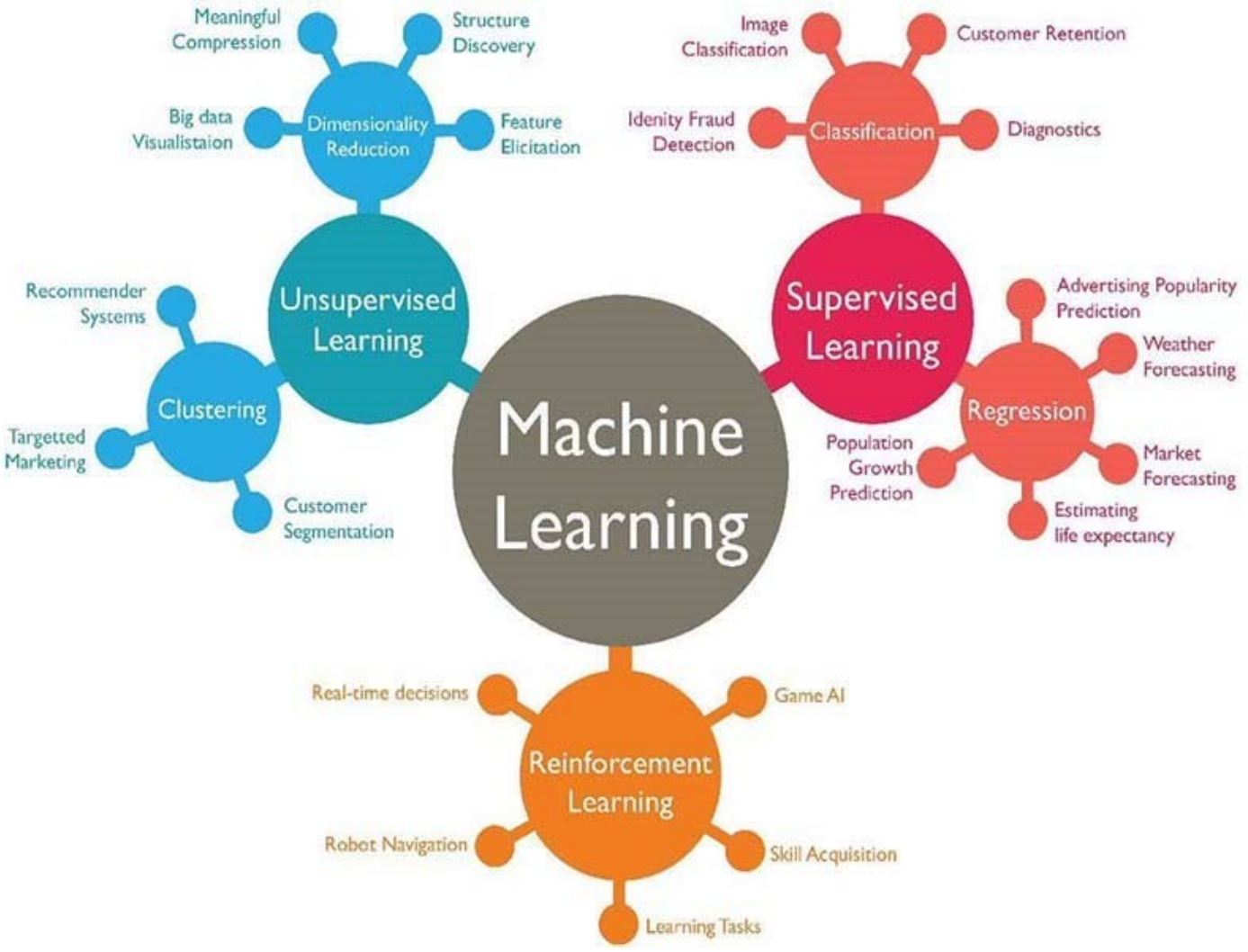
BEHAVIORAL

- Purchasing habits
- Brand interactions
- Spending habits
- Customer loyalty
- Actions taken on a website

ReferralCandy Blog

Psychographics table

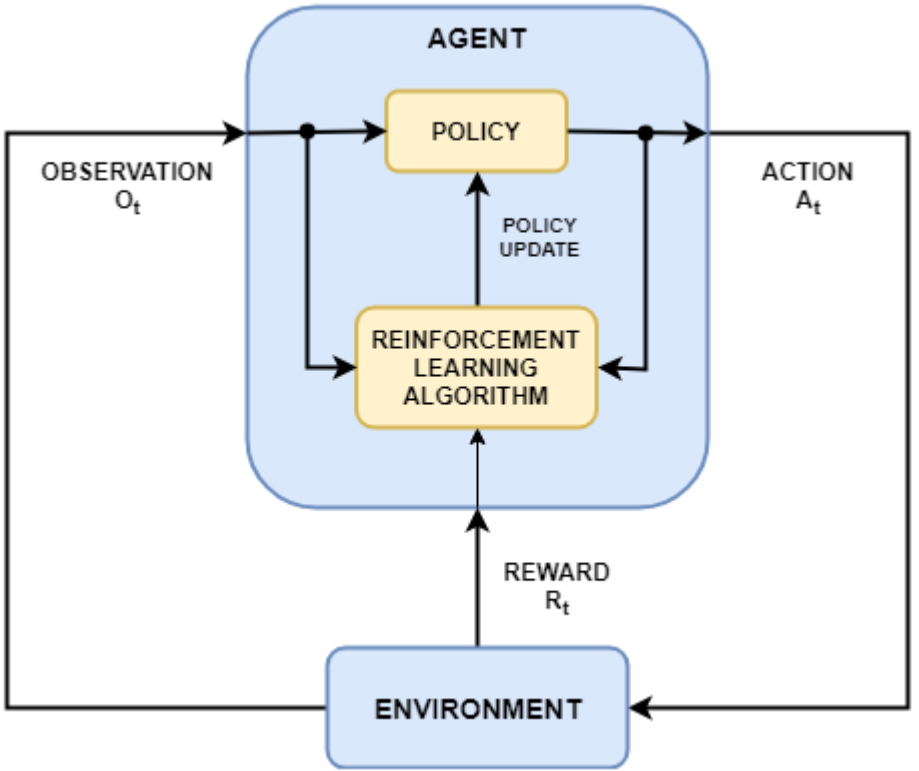
Resigned	Rigid, strict, authoritarian and chauvinist values, oriented to the past and to Resigned roles. Brand choice stresses safety, familiarity and economy. (Older)
Struggler	Alienated, Struggler, disorganised - with few resources apart from physical/mechanical skills (eg car repair). Heavy consumers of alcohol, junk food and lotteries, also trainers. Brand choice involves impact and sensation.
Mainstreamer	Domestic, conformist, conventional, sentimental, passive, habitual. Part of the mass, favouring big and well-known value for money 'family' brands. Almost invariably the largest 4Cs group.
Aspirer	Materialistic, acquisitive, affiliative, oriented to extrinsics ... image, appearance, charisma, persona and fashion. Attractive packaging more important than quality of contents. (Younger, clerical/sales type occupation)
Succeeder	Strong goal orientation, confidence, work ethic, organisation ... support status quo, stability. Brand choice based on reward, prestige - the very best . Also attracted to 'caring' and protective brands ... stress relief. (Top management)
Explorer	Energy - autonomy, experience, challenge, new frontiers. Brand choice highlights difference, sensation, adventure, indulgence and instant effect - the first to try new brands. (Younger - student)
Reformer	Freedom from restriction, personal growth, social awareness, value for time, independent judgement, tolerance of complexity, anti-materialistic but intolerant of bad taste. Curious and enquiring, support growth of new product categories. Select brands for intrinsic quality, favouring natural simplicity, small is beautiful.(Highier education)


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
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Reinforcement Learning

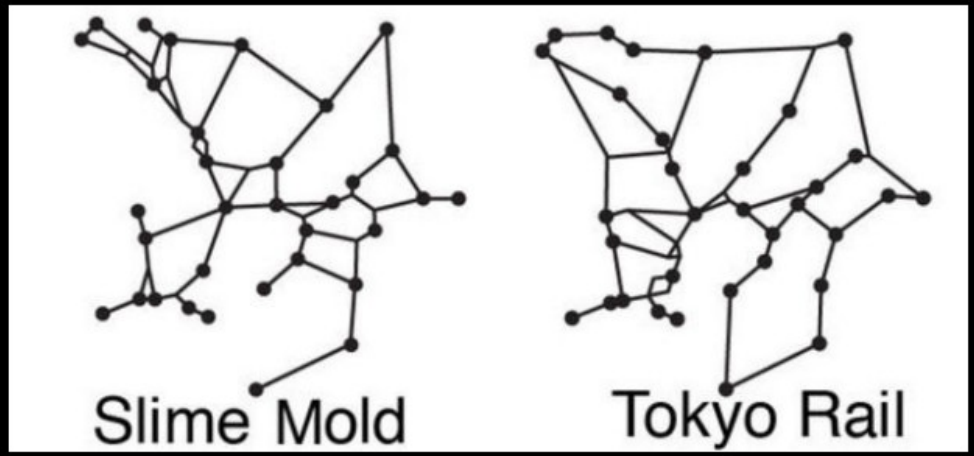
- An agent receives information about its environment and learns to choose actions that will maximize some reward
  - Eg. Chess and game playing programs
  - Robots moving in a warehouse
  - Solving mazes
- Agents develop a *policy* concerning how to change its *state* by making decisions
  - States are evaluated using a reward or cost function
  - Also apply a heuristic or *hint* to improve decisions

Reinforcement Learning



Reinforcement Learning

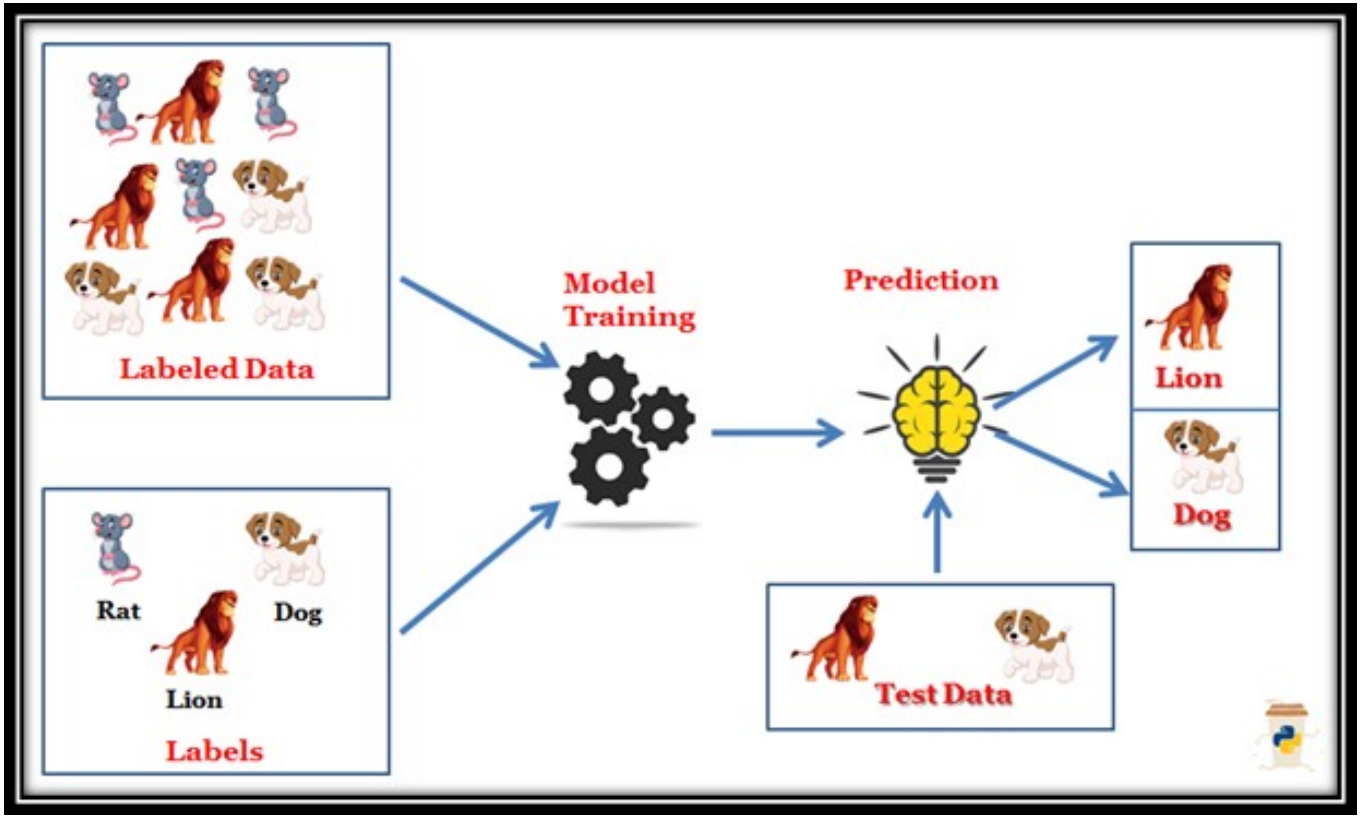
The slime mould creates networks between food sources. Scientists set up a simulacrum of Tokyo with food sources at the points representing subway stations.



The network designed by the slime mould was more efficient than the actual subway system.

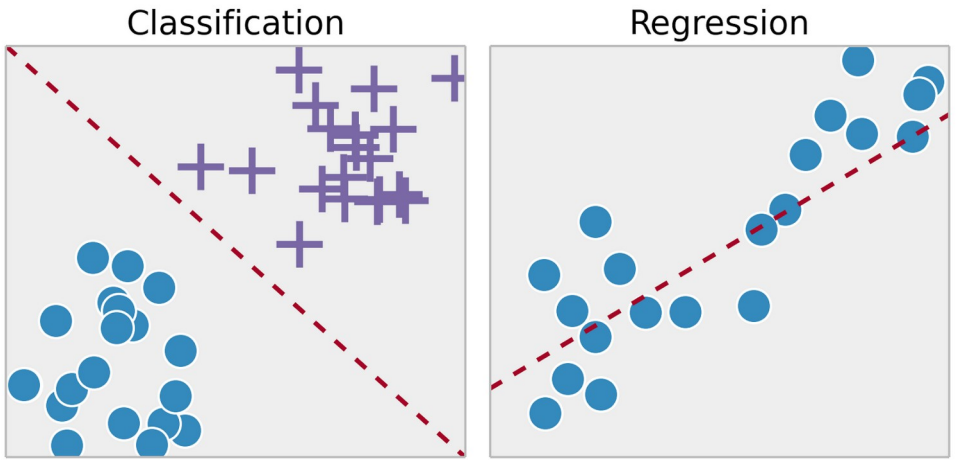
Studying how the slime mould creates networks has produced much more efficient algorithms for network design.

Supervised Learning



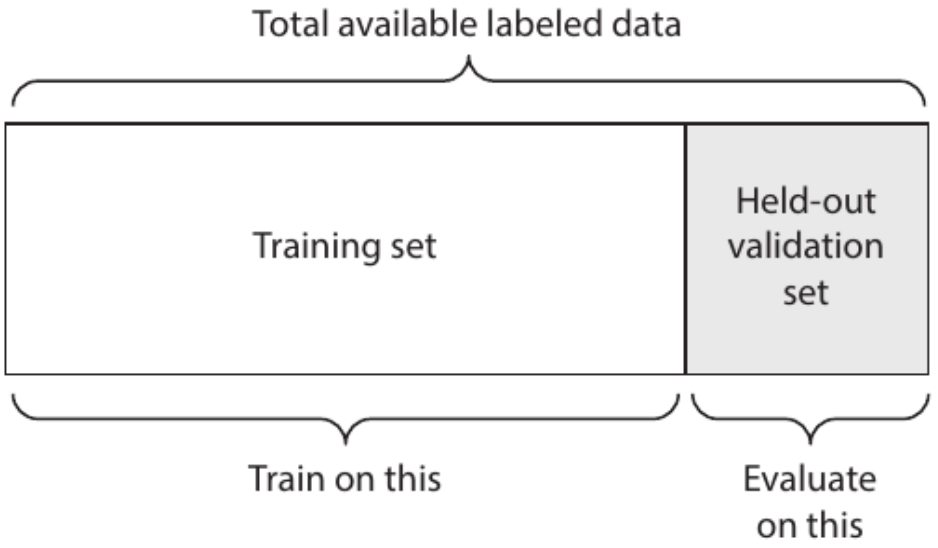
Supervised Learning

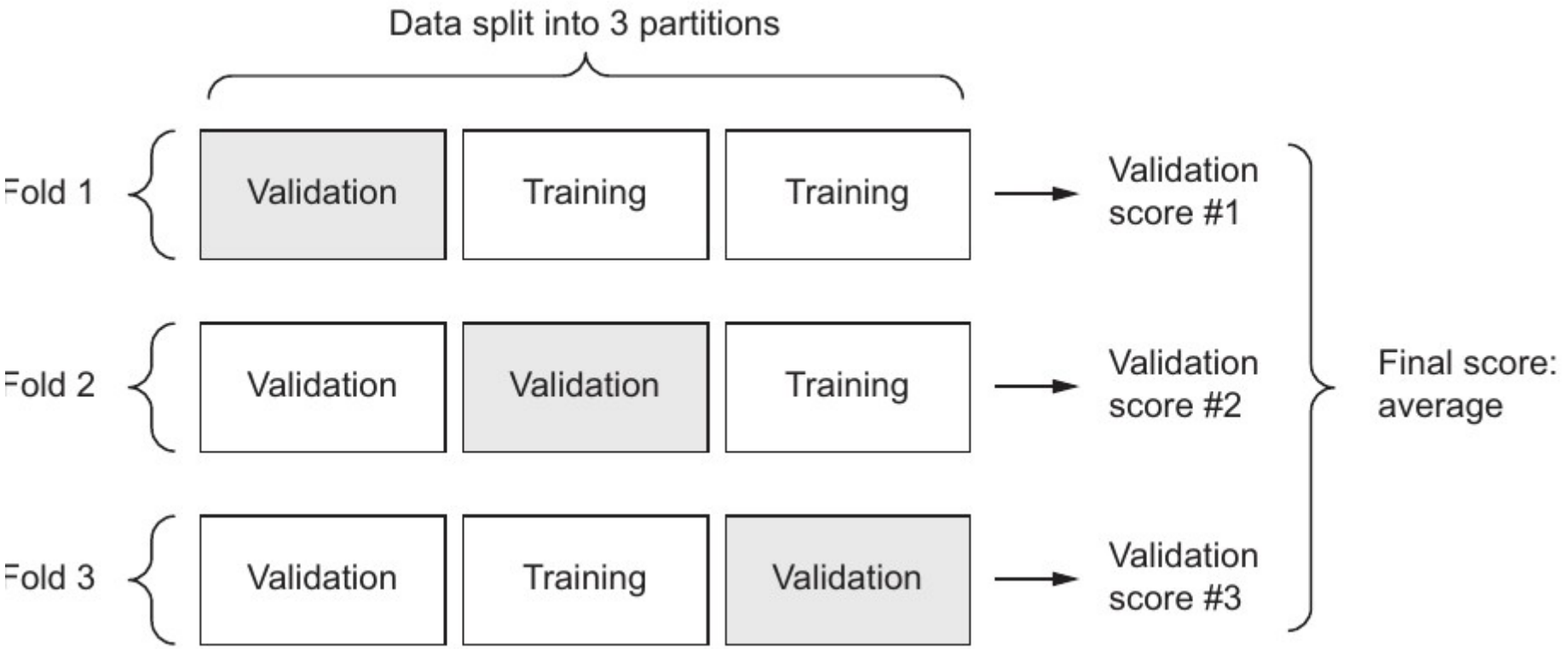
- There are two basic types of SML
  - Regression – compute a result from an n-dimensional input
  - Classification – identify what type an n-dimensional input is



Supervised Learning Process

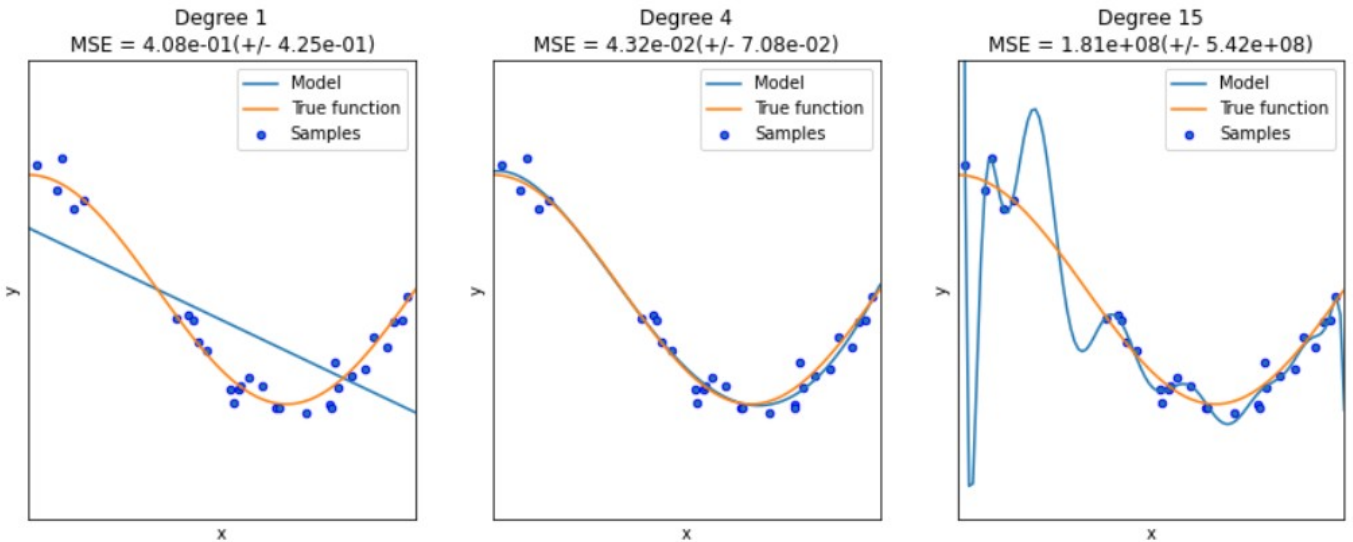
- Data set divided in n partition
  - N-1 parts used to train the model
  - The other part to test the model – n=2 shown below




Data split into 3 partitions
Fold 1
Validation
Training
Training
Validation score #1
Fold 2
Validation
Validation
Training
Validation score #2
Fold 3
Validation
Training
Validation
Validation score #3
Final score: average

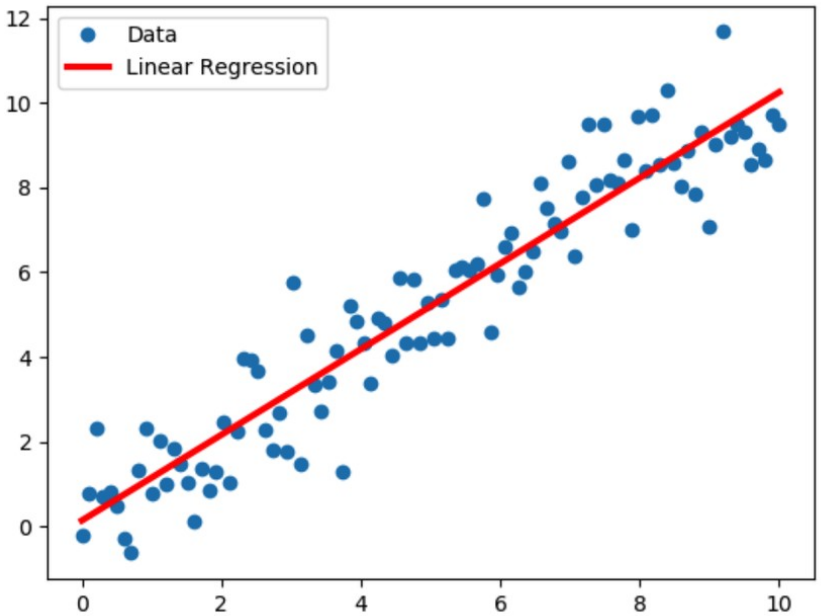
Overfitting and Underfitting

- The training dataset is a sample of a population



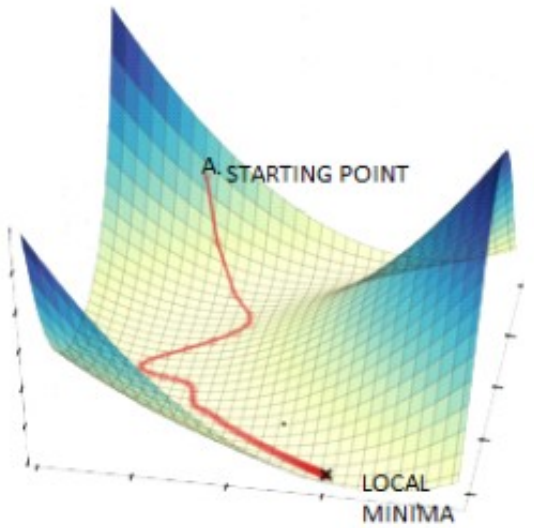
Linear Regression

- Exactly the same as in statistics
  - Loss function is usually the mean square error



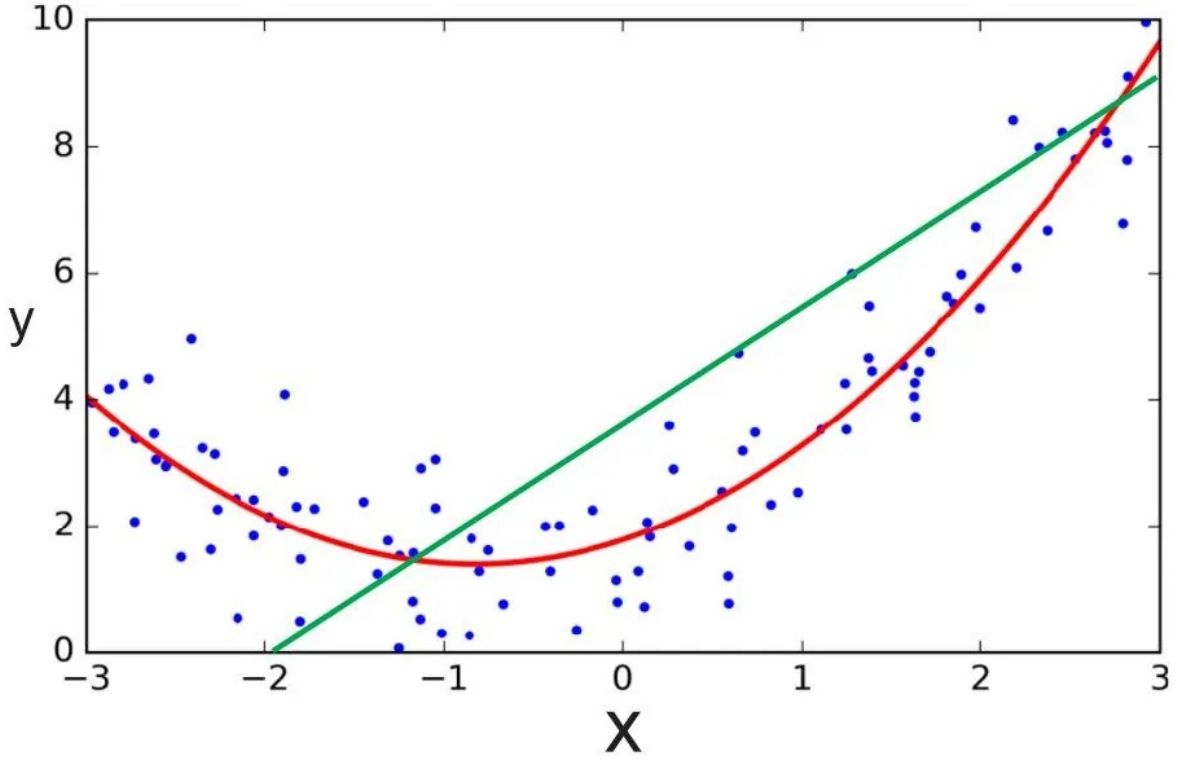
Multi-Linear Regression

- Best fit across multiple dimensions
  - Loss function is usually the Gradient Descent Algorithm
  - There are a number of tune-able parameters
  - eg. “alpha” – the learning rate



Polynomial Regression

- Regression to polynomial curves



Regression Types

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

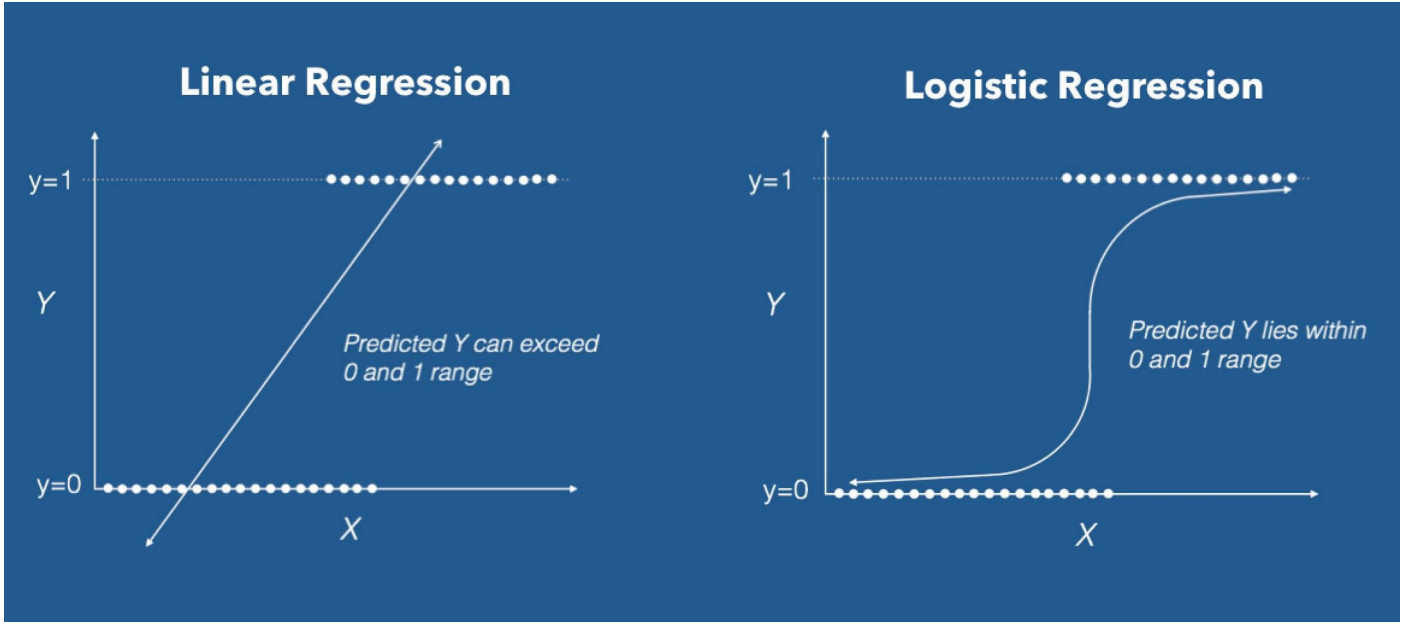
$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

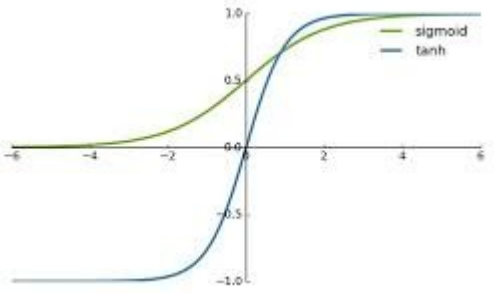
Logistic Regression

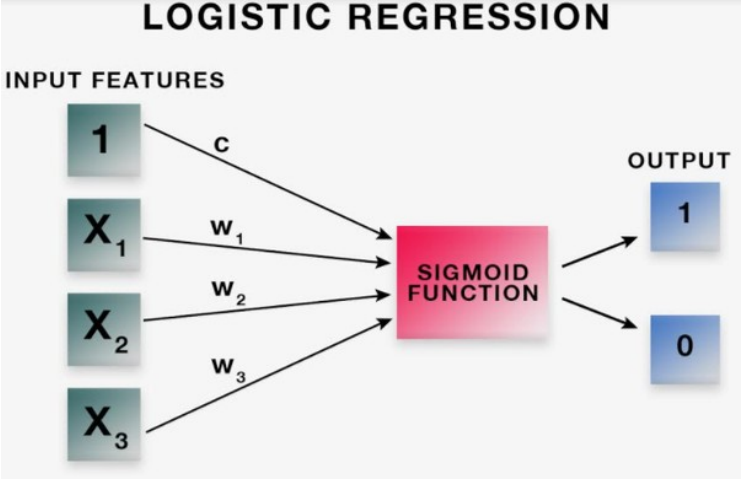
- Classification of data into two exclusive classes
  - Uses regression but maps computed values to each label



Logistic Regression Function

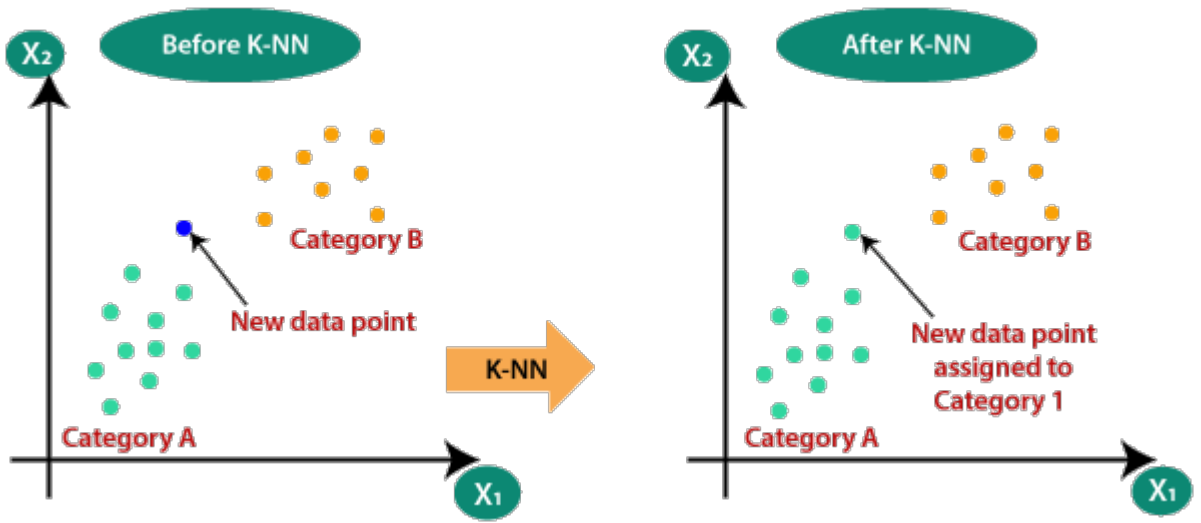
- Uses an activation function
  - Either the sigmoid or tanh





K Nearest Neighbour

- Simplest SML algorithm
- Lazy learning – uses minimal training
  - Uses a set of labelled observations
  - New observation is added and labelled based on its nearest n neighbours
  - Requires a distance metric to compute what “nearest” means
- Each added observation is used in computing values for following observations


Before K-NN
X2
Category B
New data point
Category A
X1
K-NN
After K-NN
X2
Category B
New data point
assigned to
Category 1
Category A
X1

Advantages of KNN

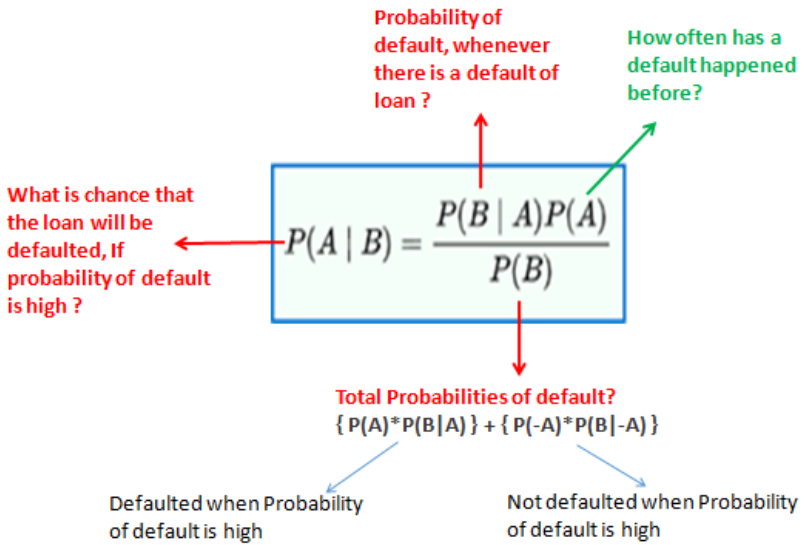
- K-NN is intuitive and simple
- K-NN has no assumptions or training step
- It constantly evolves
- Has only two hyperparameters – the distance metric and number of neigbours
- Very easy to implement for multi-class problem

Dis-advantages of KNN

- As dataset grows, efficiency or speed of algorithm declines very fast
- K-NN works best with few dimensions, does not scale well to larger dimensions
- Unbalanced data biases the learning
  - Over representation of a class biases choices to that class
- Very sensitive to outliers
- No capability to deal with missing data

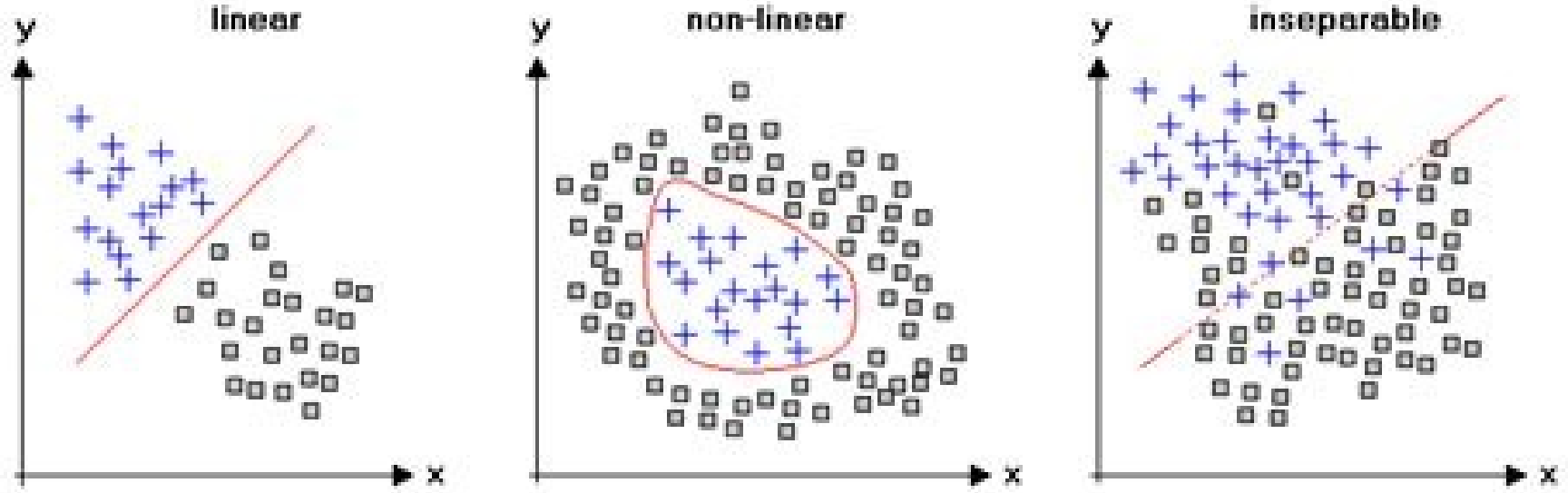
Naive Bayes Classification

- Simple model based on Bayes theorem
  - Assumes independence of dimensions
  - Scaleable and simple to implement
  - Works well in many complex situations
  - Email spam detection for example based on content

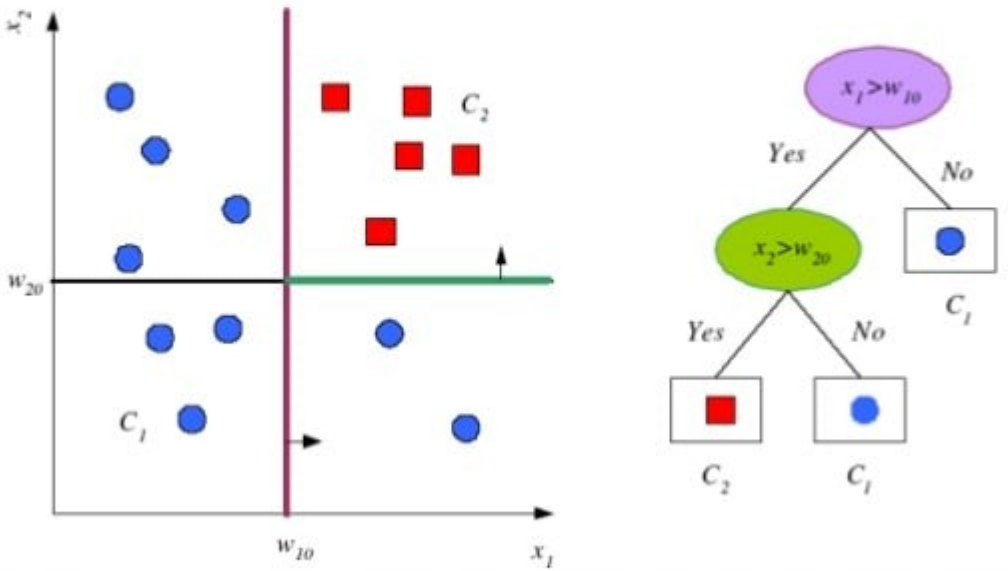


Decision Trees

- Can be used or regression or classification
  - Primarily used in classification “playing 20 questions with the data”
  - eg. Healthcare diagnosis based on symptoms
  - This is a non-linear classifier
- Usually used to implement some chain of decisions
  - The ML problem is: Given a set of labelled observations, what choices on each input dimension gives the best classification?
    - *ie. What choices give the fewest number of misclassifications*

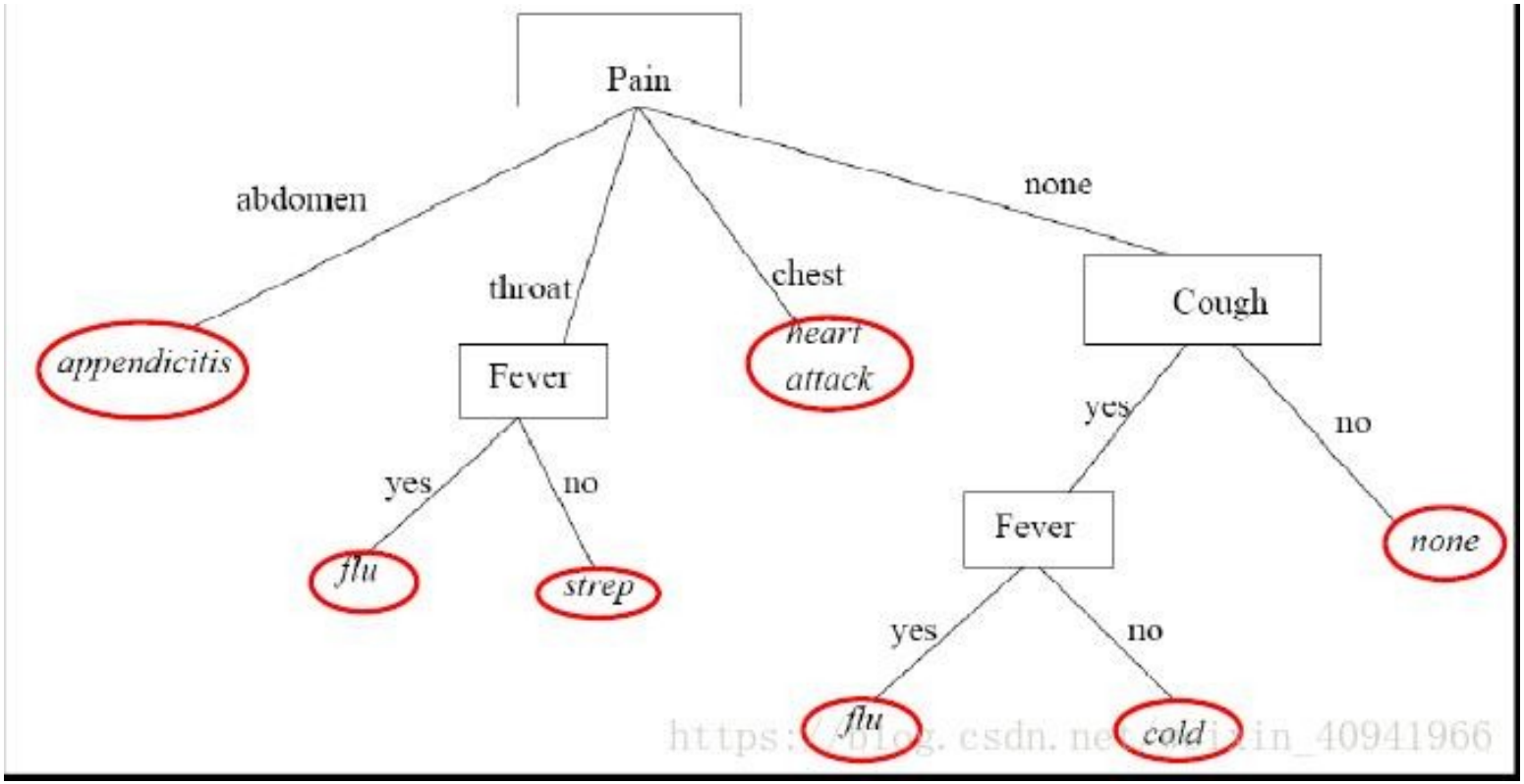

linear
y
x
non-linear
y
x
inseparable
y
x

Decision Tree

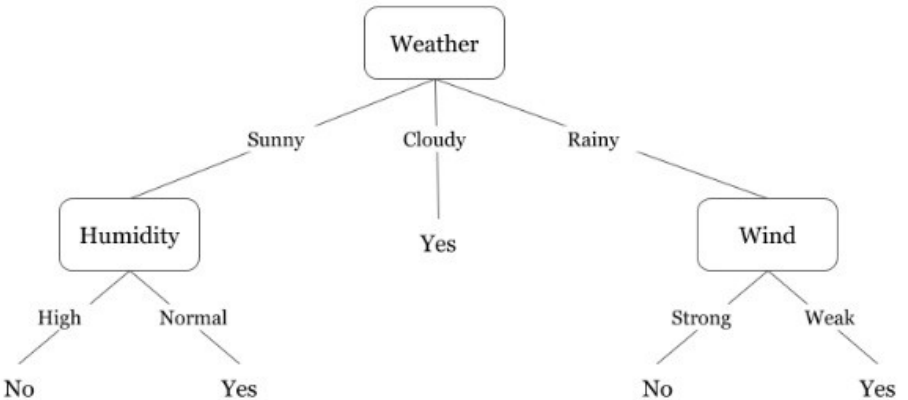


w_{10} x_1

10 Lec 2: Decision Trees - Nearest Neighbors

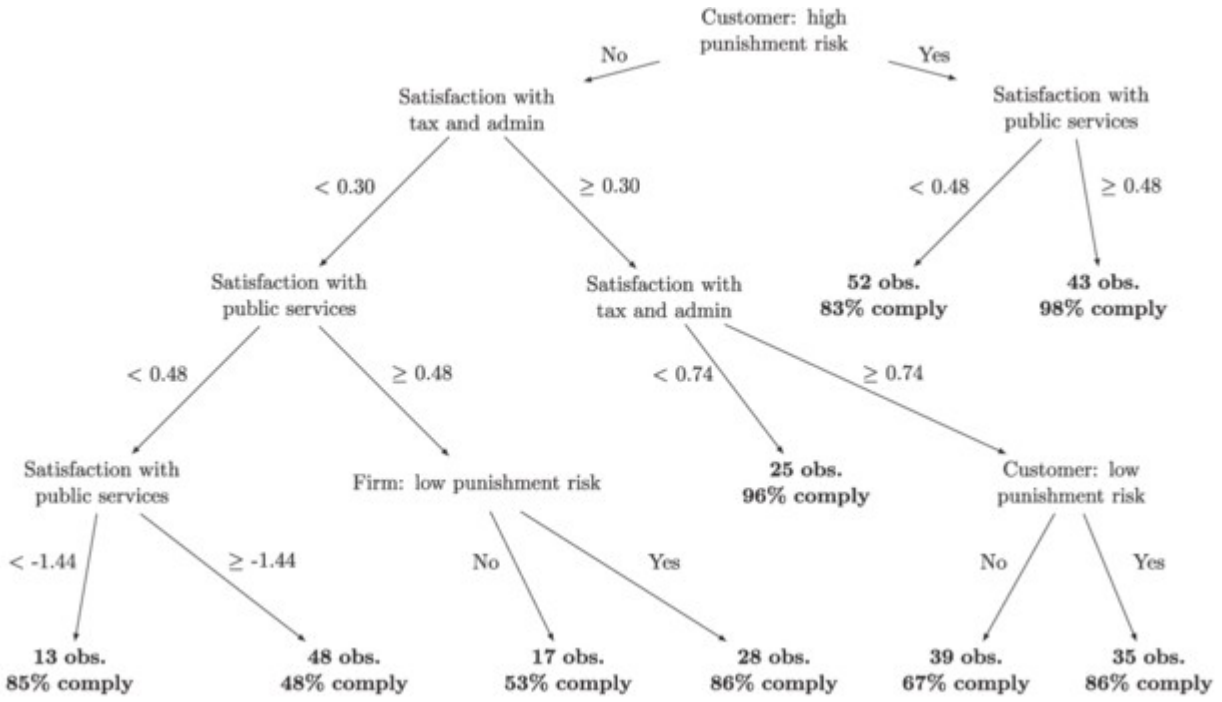


Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No



Regression Trees

- The same as for decision tree except that we are producing values
  - This might be thought of as a fuzzy classification
  - The final results are the means of the different categories
  - Instead of yes/no we use an expected value (loosely speaking)


Customer: high punishment risk
No
Yes
Satisfaction with tax and admin
Satisfaction with public services
< 0.30
≥ 0.30
< 0.48
≥ 0.48
Satisfaction with public services
Satisfaction with tax and admin
52 obs. 83% comply
43 obs. 98% comply
< 0.48
≥ 0.48
< 0.74
≥ 0.74
Satisfaction with public services
Firm: low punishment risk
25 obs. 96% comply
Customer: low punishment risk
< -1.44
≥ -1.44
No
Yes
No
Yes
13 obs. 85% comply
48 obs. 48% comply
17 obs. 53% comply
28 obs. 86% comply
39 obs. 67% comply
35 obs. 86% comply

Decision Trees

- We successively partition the observations
  - For a dimension, we look for the split that produces the lowest *entropy* – measure of mis-classification
  - Once we have split the observations in two, we then repeat the process on each partition using a different dimension
  - There are different entropy measures used
  - Data can be numeric, ordinal or categorical (unlike logistic regression)
  - This works well for clustered data, especially non-linear
  - Performance is not good with linear data in some cases

Decision Trees Advantages

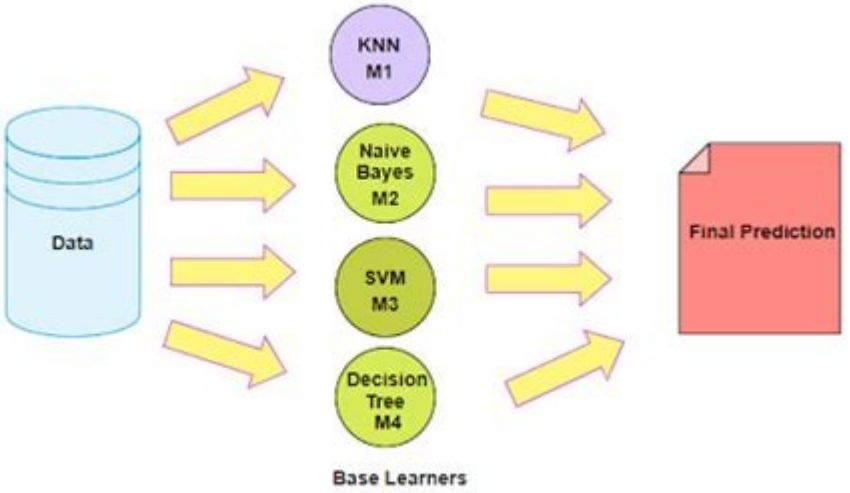
- Requires less effort for data preparation during pre-processing compared to other algorithms
- Does not require normalization or scaling of data.
- Can handle both continuous and categorical variables.
- Can automatically handle missing values.
- Very intuitive and easy to explain to technical teams as well as stakeholders.
- Usually robust to outliers and can handle them automatically.

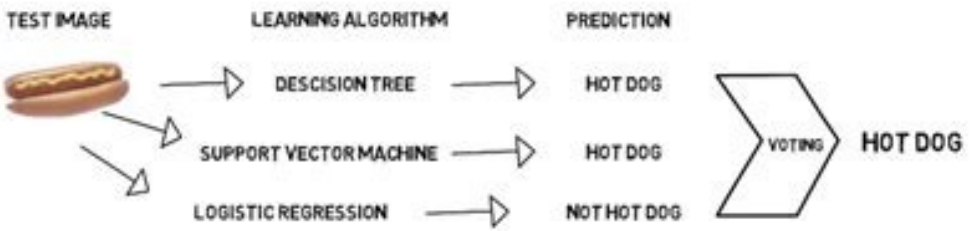
Decision Trees Disadvantages

- Overfitting is the main problem which ultimately leads to wrong predictions.
- Unstable - Adding a new data point can lead to re-generation of the overall tree and all nodes need to be recalculated and recreated.
- High variance in the output which leads to application errors
- Affected by noise: Little bit of noise can make it unstable which leads to wrong predictions.
- Not suitable for large datasets: If data size is large, then one single tree may grow complex and lead to overfitting.

Ensemble Learning

- Creating multiple models for a given problem to mediate the disadvantages of decision trees
  - Goal is to reduce the bias of any one model
  - Models can use different algorithms (Bayes, Decision Tree etc)
    - *Downside is that this requires a LOT more work*
  - Models can be generated by the same algorithm but use different training data sets
    - *Downside is that the data sets may not be available*
  - Bagging – (bootstrap aggregation) simulating having different data sets
  - Boosting – weighting certain classifiers that perform well


KNN
M1
Naive
Bayes
M2
SVM
M3
Decision
Tree
M4
Data
Final Prediction
Base Learners

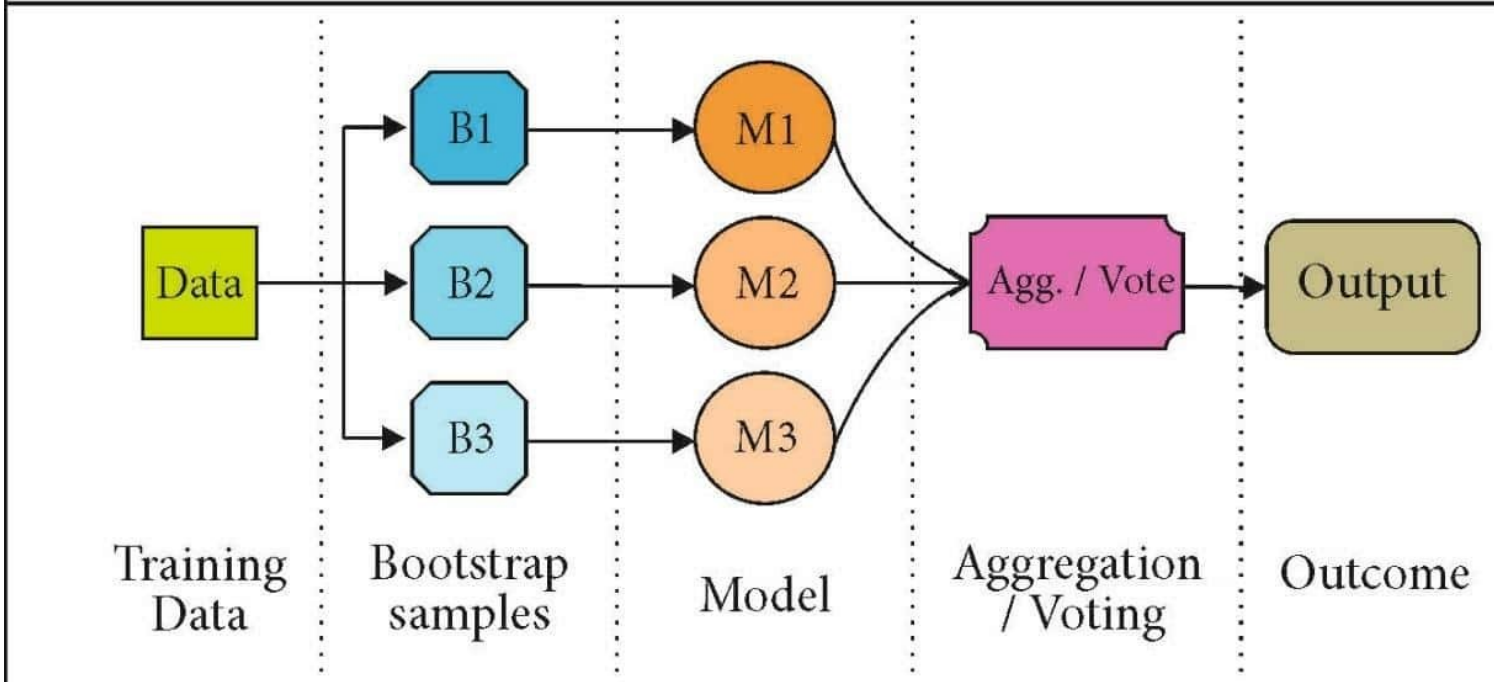

TEST IMAGE
LEARNING ALGORITHM
PREDICTION
DESCISION TREE
HOT DOG
SUPPORT VECTOR MACHINE
HOT DOG
LOGISTIC REGRESSION
NOT HOT DOG
VOTING
HOT DOG

Bootstrap Bagging

- To “bag” a training set
  - Assume the training set *is* the population and not a sample
  - Each bootstrap sample is done with replacement from the original set
  - The results are computed and the results combined with some sort of *voting* process
    - For example, taking the average of the results or the most commonly occurring result.

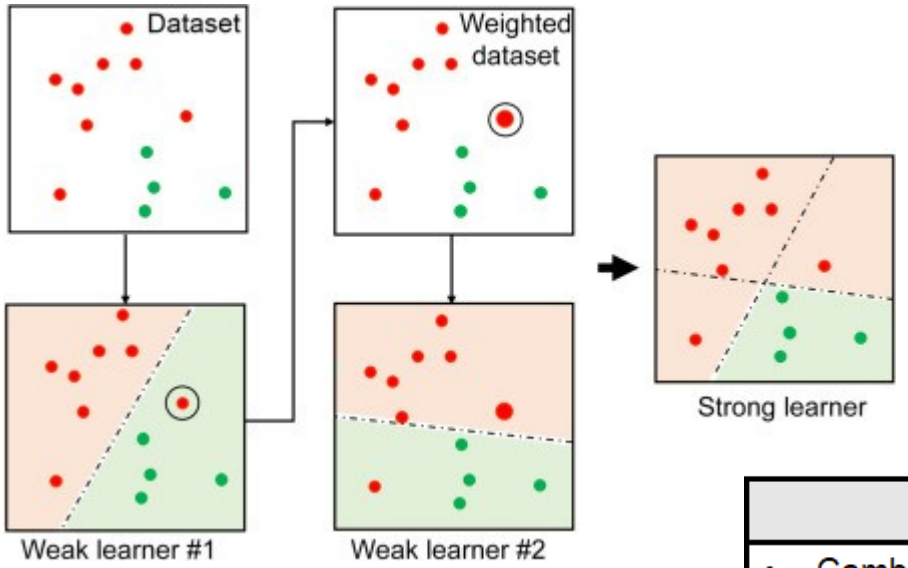
BAGGING Algorithm

Bootstrap Aggrigating



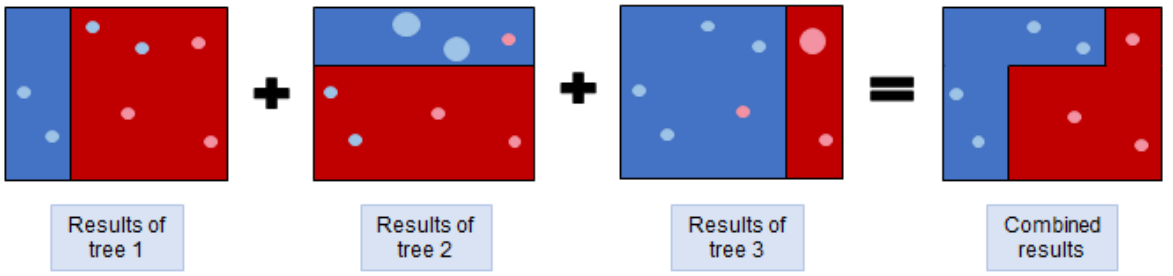
Boosting

- Each classifier is used in isolation
  - These are call weak classifiers
  - Each is treated as a decision *stump* – a tree of depth 1
  - Each weak classifier is a *say* in the final decision based on successes (ie. low entropy)
  - Each sample is weighted as to how wrong the classifier was
- The classifiers are then combined
  - Each classifier uses the weight of a sample from the previous classifier to focus on *correcting* wrong classifications



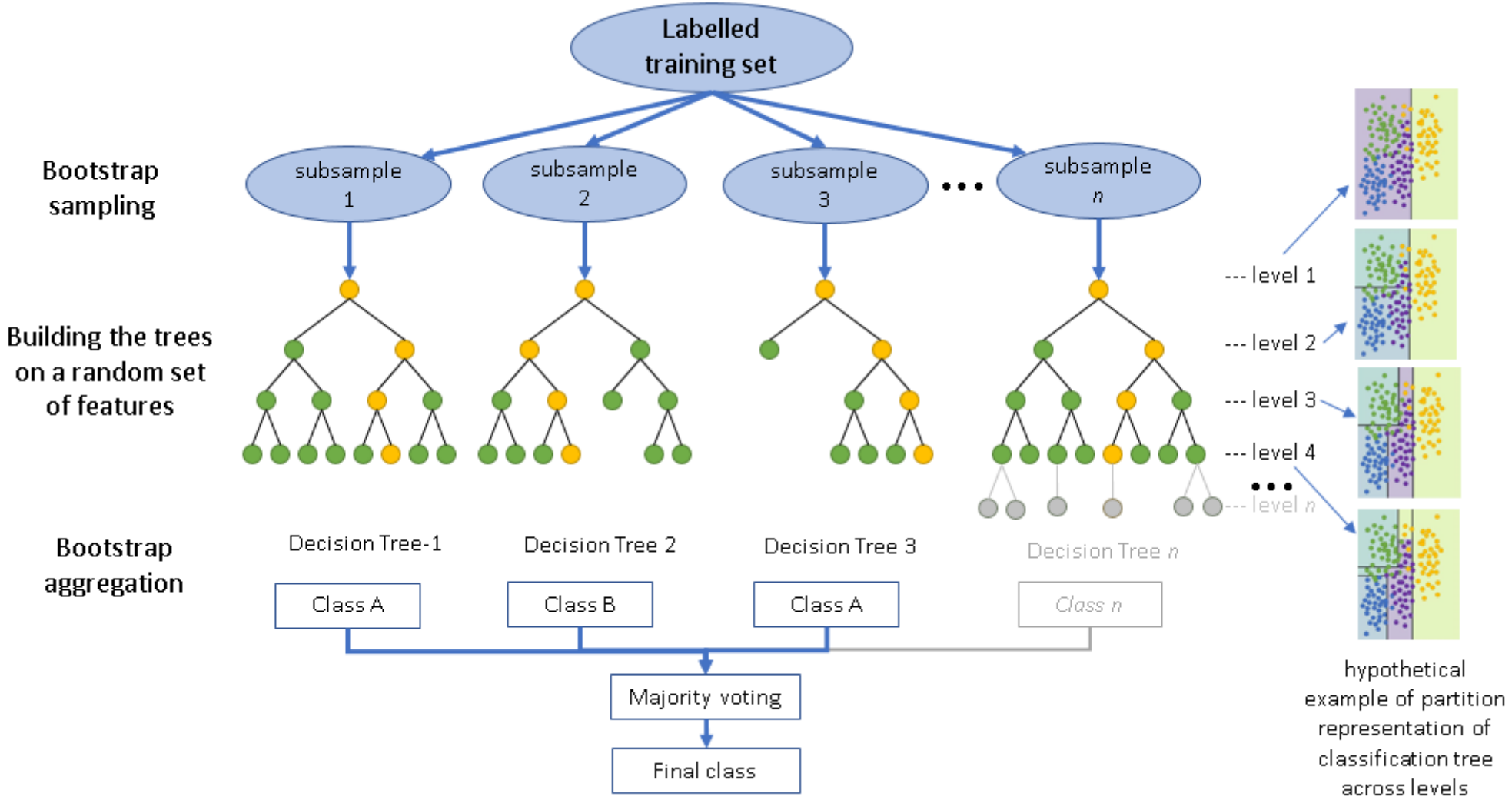
AdaBoost:

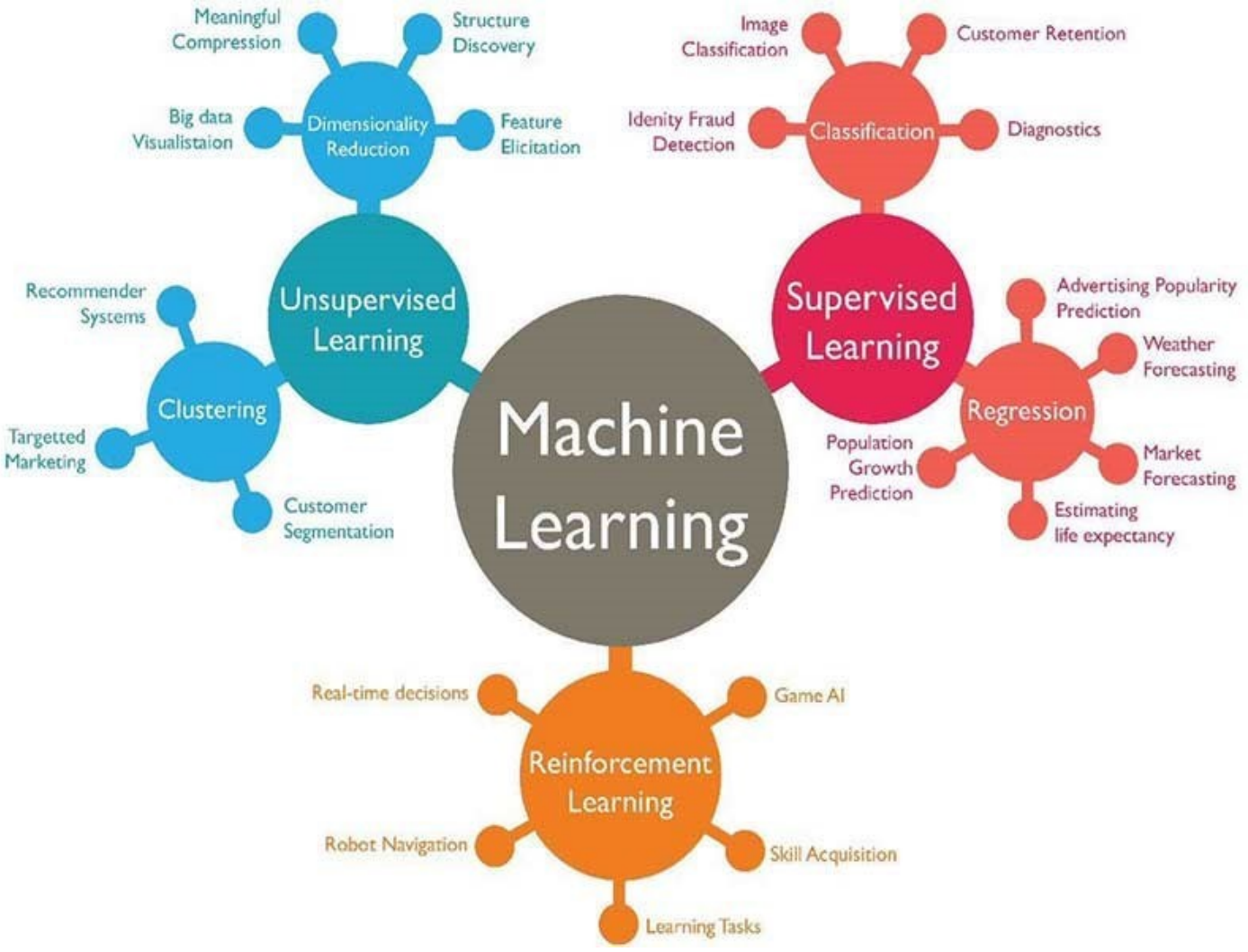
- Combining **weak learners** (decision trees)
- Assigning **weights to incorrect values**
- **Sequential tree growing** considering past mistakes



Random Forest

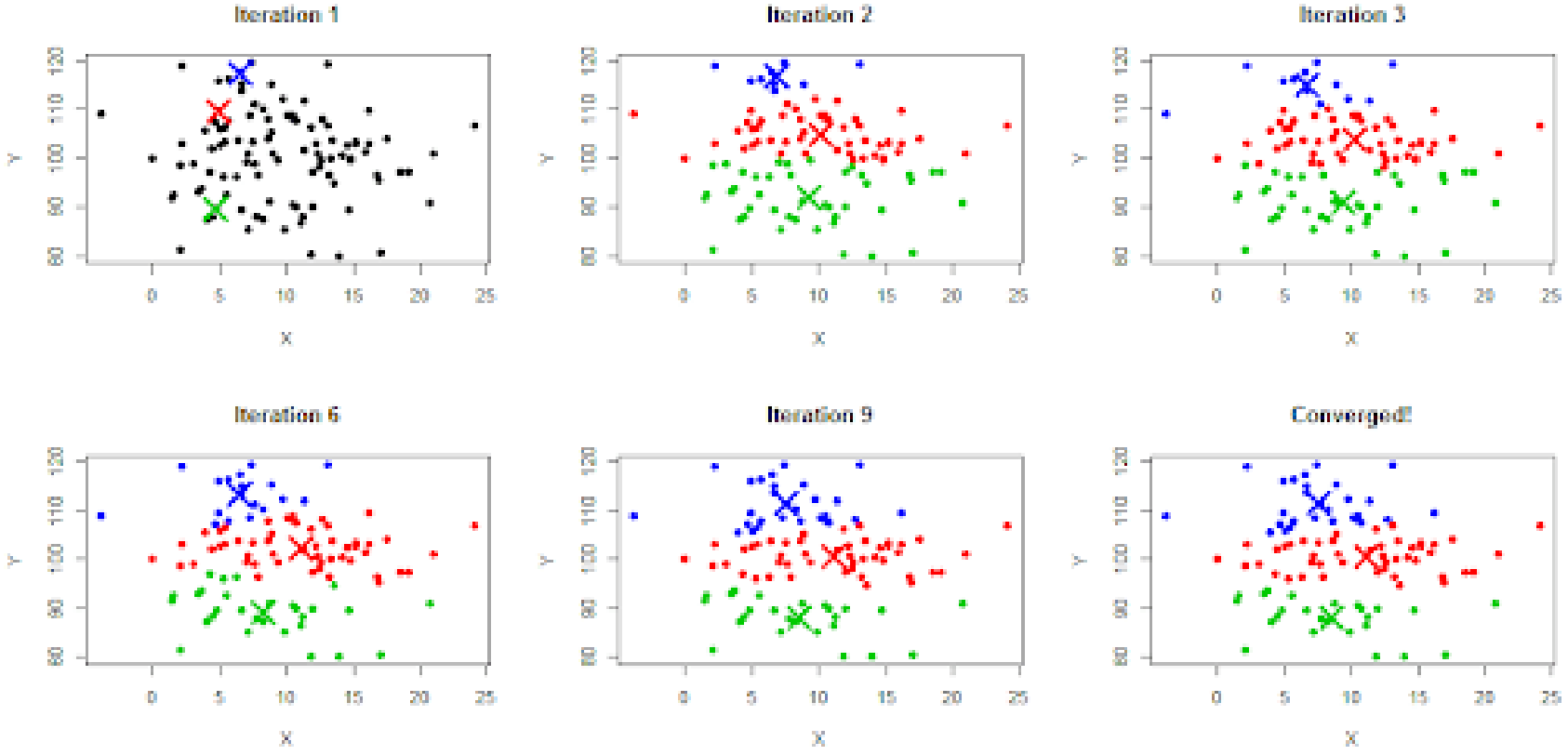
- Decision trees tend to overfit the training data
  - A random forest is a ensemble of decision trees
  - Often the data sets are bagged
  - But we may still be biased by a strong classifier
  - We need to reduce the variance
- A random forest modifies the ensemble
  - Each tree uses a sample of the dimensions for classification
  - We can either boost the result or use some aggregation method to get a final model


Labelled training set
Bootstrap sampling
subsample 1
subsample 2
subsample 3
subsample n
Building the trees on a random set of features
--- level 1
--- level 2
--- level 3
--- level 4
--- level n
Bootstrap aggregation
Decision Tree-1
Decision Tree 2
Decision Tree 3
Decision Tree n
Class A
Class B
Class A
Class n
Majority voting
Final class
hypothetical example of partition representation of classification tree across levels


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
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Unsupervised K-means Clustering

- Data set is by choosing k arbitrary cluster points
  - These will be the initial centroids of the clusters
- Computation process
  - 1. For each point, the distance from that point to each centroid is computed
    - *The point is then assigned to the cluster where the distance to that centroid is a minimum (some points may be reassigned)*
  - 2. Then new centroids are computed with the all of the points in that cluster
  - We then go back to step 1 and repeat until no points are reassigned

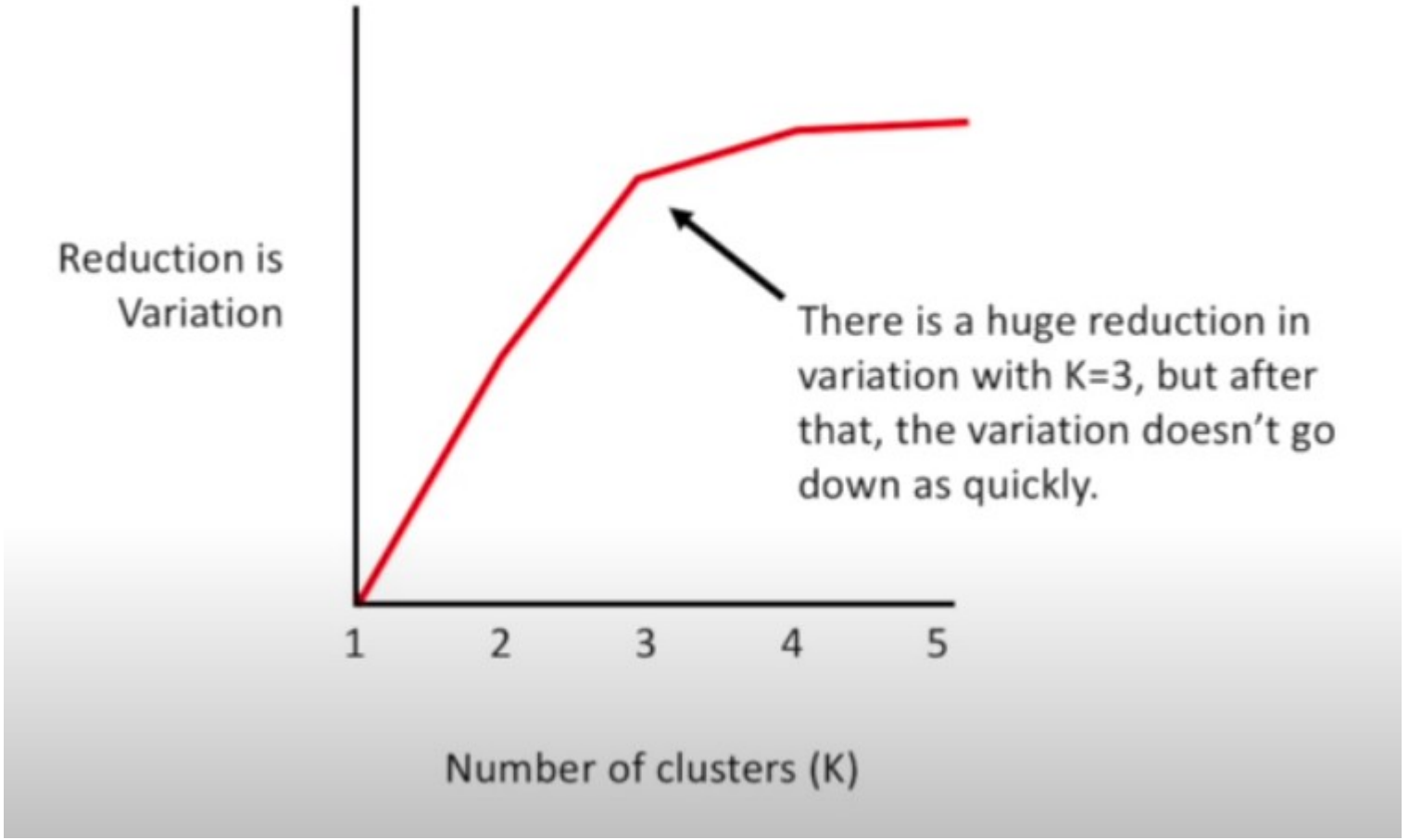

Iteration 1
Iteration 2
Iteration 3
Iteration 6
Iteration 9
Converged!
X
Y

Unsupervised K-means Clustering

- In other words, when subsequent applications of the algorithm do not change the clustering, we are done
- To evaluate how good the clustering is we evaluate the amount of variance in each cluster and between clusters
  - We generally want to select the initial partitioning that produces the clustering with the lowest intra-custer variance
  - We may need to evaluate different sets of starting centroids to get convergence and minimize variance

Unsupervised K-means Clustering

- If k=1 then we have one cluster and maximal variance
- If k = number of data points, we get 0 variance
- We choose k based on where we stop getting significant reductions in variance


Reduction is Variation
There is a huge reduction in variation with K=3, but after that, the variation doesn't go down as quickly.
1
2
3
4
5
Number of clusters (K)

Advantages of K-means Clustering

- Relatively simple to implement.
- Scales to large data sets.
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Disadvantages of K-means Clustering

- Choosing k manually.
- Being dependent on initial values.
- Clustering data of varying sizes and density.
  - Has trouble where data clusters are of different sizes and densities
- Outliers can skew centroids or create spurious clusters
- Scaling with number of dimensions – at higher dimensions it becomes difficult to distinguish between examples

Hierarchical Clustering

- Different approach to clustering
  - Produces a hierarchical cluster model
  - Much like a file system
- Process
  - 1. Make each data point a single-point cluster → forms N clusters
  - 2. Take the two closest data points and make them one cluster → forms N-1 clusters
  - 3. Take the two closest clusters and make them one cluster → Forms N-2 clusters.
  - 4. Repeat step-3 until you are left with only one cluster.

Hierarchical Clustering - Linkage

- Common linkage methods are:
  - Complete-linkage: the distance between two clusters is defined as the longest distance between two points in each cluster.
  - Single-linkage: the distance between two clusters is defined as the shortest distance between two points in each cluster.
  - Average-linkage: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.
  - Centroid-linkage: finds the centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.

Hierarchical Clustering - Linkage

- Common linkage methods are:
  - Complete-linkage: the distance between two clusters is defined as the longest distance between two points in each cluster.
  - Single-linkage: the distance between two clusters is defined as the shortest distance between two points in each cluster.
  - Average-linkage: the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.
  - Centroid-linkage: finds the centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.

Dendrogram

- A Dendrogram is a tree diagram of hierarchical relationships between data sets of data
    - Contains the memory of hierarchical clustering algorithm
    - *Distance between data points represents dissimilarities.*
    - *Height of the blocks represents the distance between clusters*
    - *Clades are the branch and are arranged according to how similar (or dissimilar) they are. Clades that are close to the same height are similar to each other; clades with different heights are dissimilar — the greater the difference in height, the more dissimilarity.*