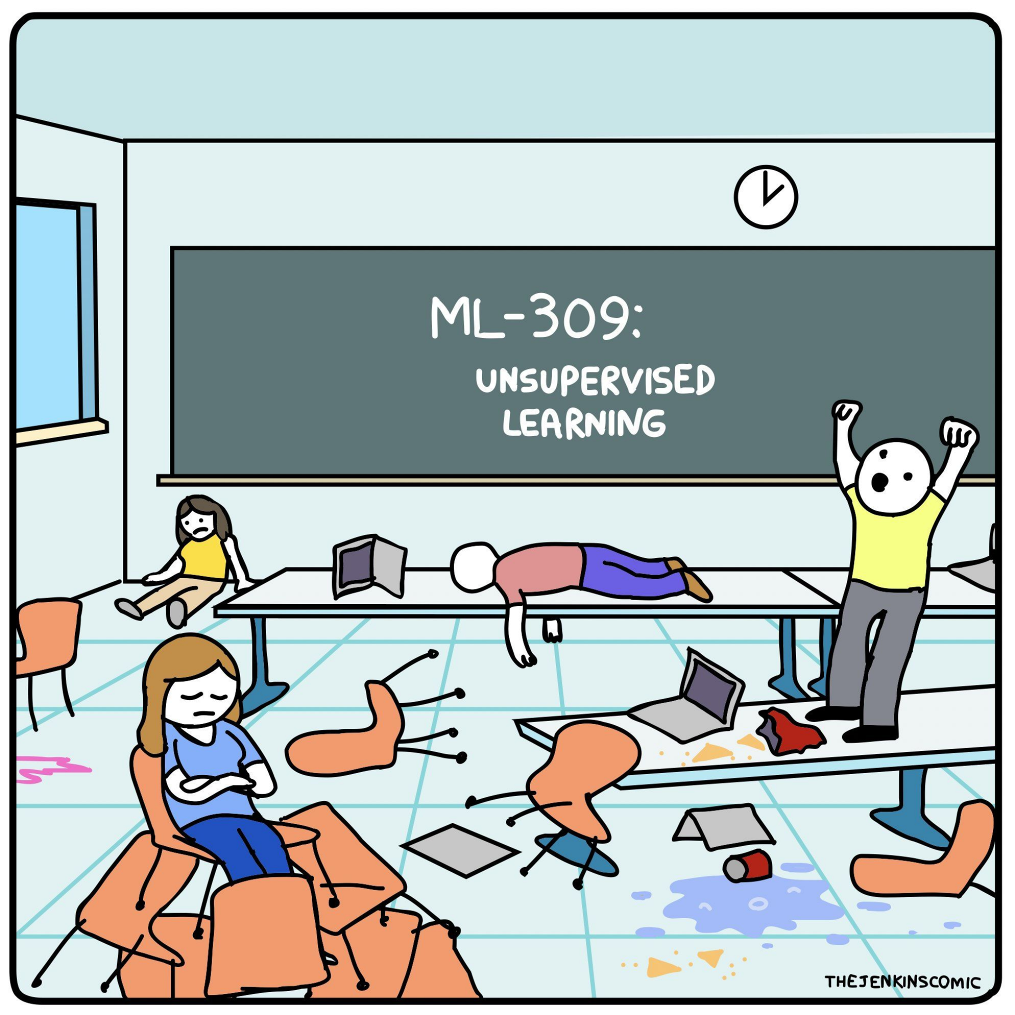
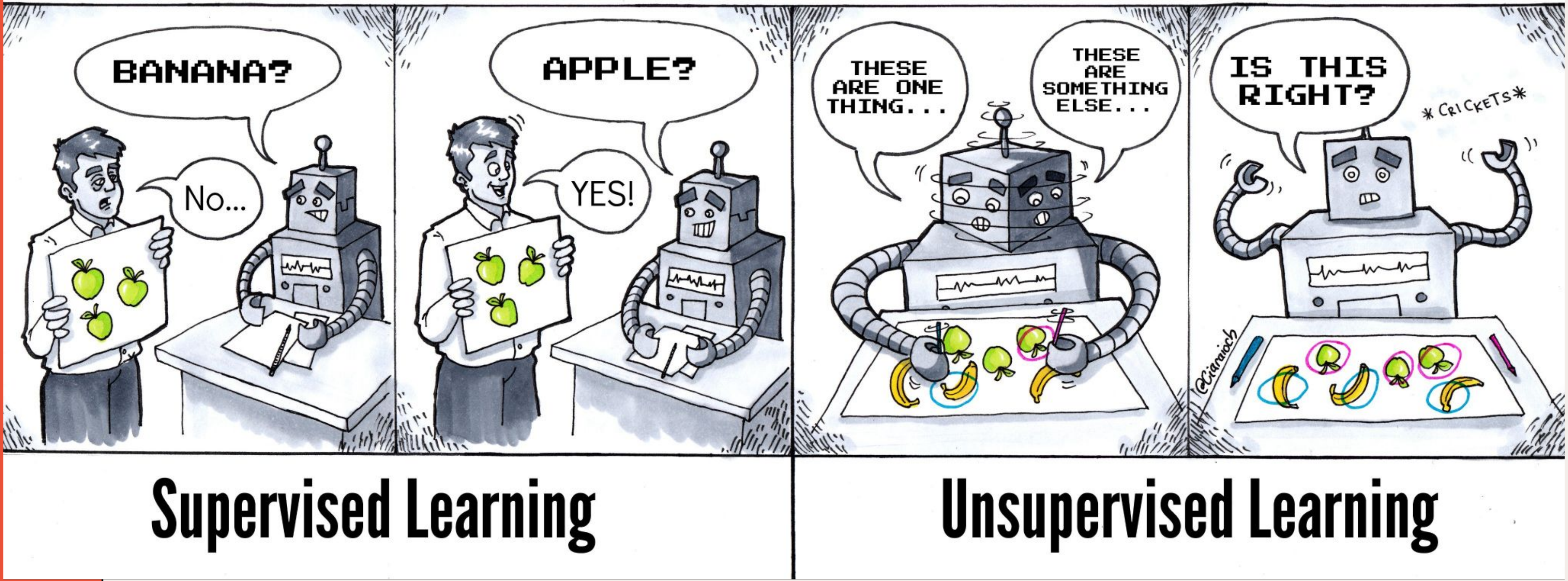
<welcome to>

SIG AI

Meeting 2: 03/07/24


ML-309:
UNSUPERVISED LEARNING
THEJENKINSCOMIC


BANANA?
No...
APPLE?
YES!
THESE ARE ONE THING...
THESE ARE SOMETHING ELSE...
IS THIS RIGHT?
CRICKETS
@ciaraioch
Supervised Learning
Unsupervised Learning

Three main types of Machine Learning Algorithms

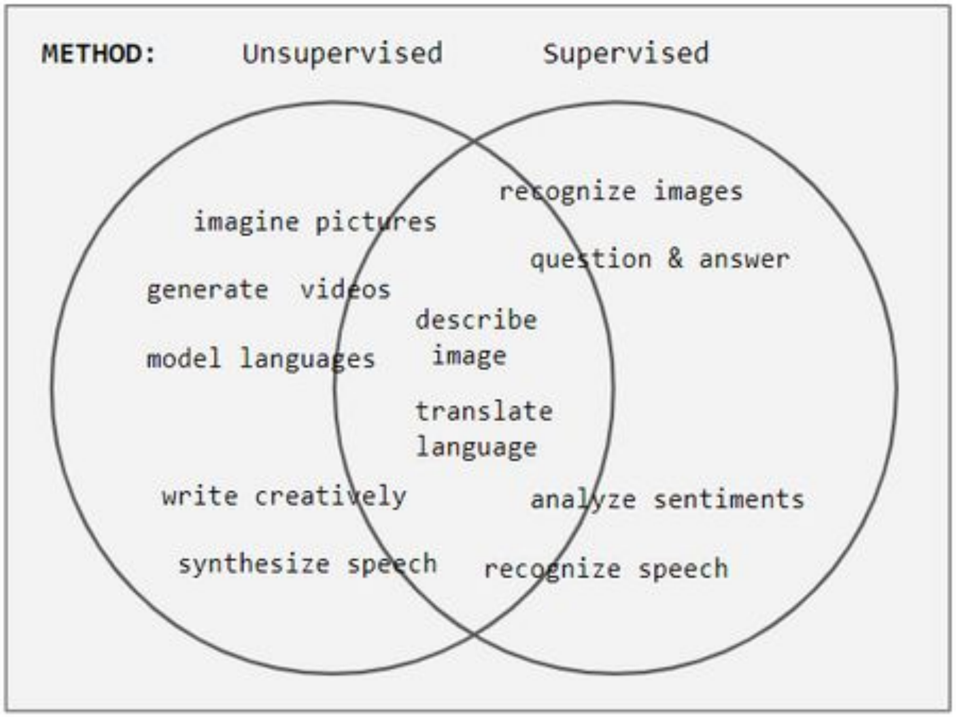
UNSUPERVISED: ME TOO. BUT AT LEAST THEY TOLD YOU THE ANSWERS
SUPERVISED: THEY GAVE ME SO MUCH TO READ, AND TESTS!
REINFORCEMENT: AT LEAST Y'ALL DON'T MAKE YOUR OWN BOOK
@OURSKY.HK

What is Unsupervised Learning?

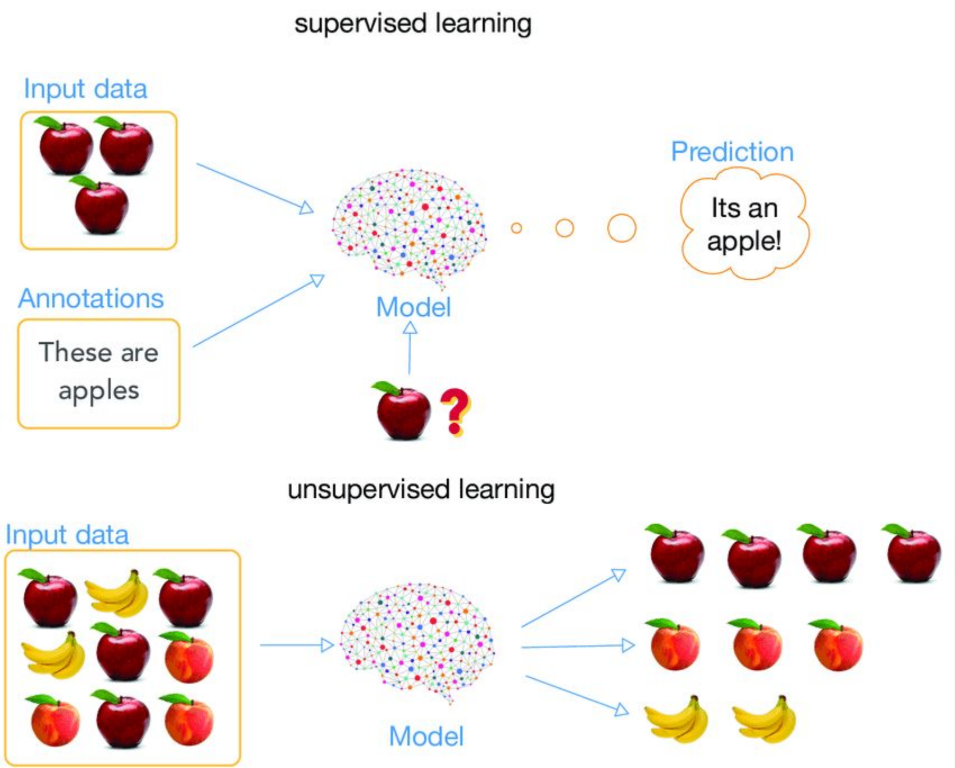
- Unsupervised learning in artificial intelligence is a type of machine learning that learns from data without human supervision. Unlike supervised learning, unsupervised machine learning models are given unlabeled data and allowed to discover patterns and insights without any explicit guidance or instruction.
- Unsupervised learning algorithms are better suited for more complex processing tasks, such as organizing large datasets into clusters. They are useful for identifying previously undetected patterns in data and can help identify features useful for categorizing data.
- Now, in what world is unlabelled raw data a good starting point, given we're all about data engineering and pre-processing?

What use-case can you think of, where using unlabelled data might be beneficial?

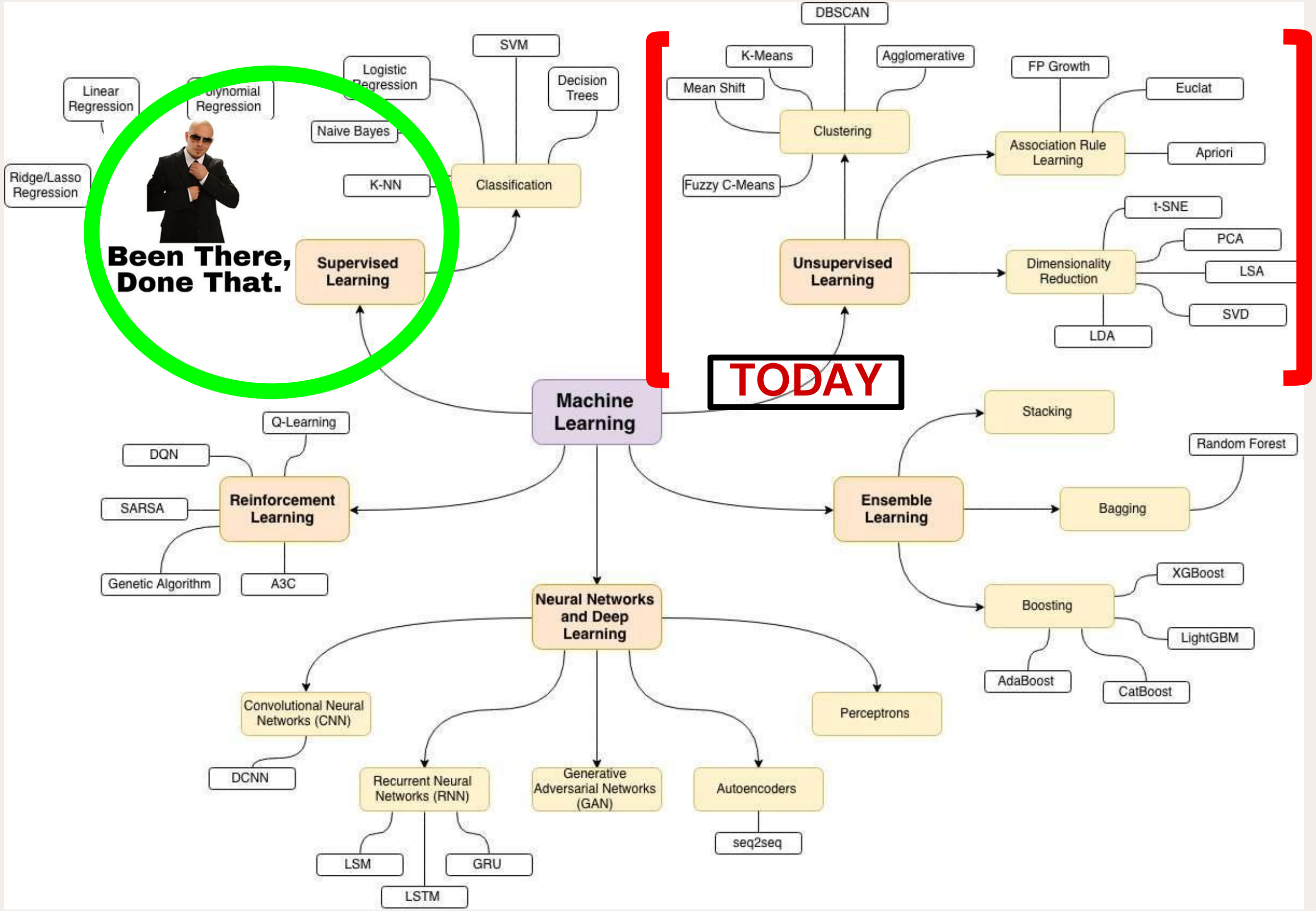
Supervised versus Unsupervised



Supervised versus Unsupervised



*



Been There, Done That.

TODAY

Machine Learning

Supervised Learning

Classification

Logistic Regression

SVM

Decision Trees

Naive Bayes

K-NN

Linear Regression

olynomial Regression

Ridge/Lasso Regression

Unsupervised Learning

Clustering

DBSCAN

K-Means

Agglomerative

Mean Shift

Fuzzy C-Means

Association Rule Learning

FP Growth

Euclat

Apriori

Dimensionality Reduction

t-SNE

PCA

LSA

SVD

LDA

Reinforcement Learning

Q-Learning

DQN

SARSA

Genetic Algorithm

A3C

Ensemble Learning

Stacking

Bagging

Random Forest

Boosting

XGBoost

LightGBM

AdaBoost

CatBoost

Neural Networks and Deep Learning

Convolutional Neural Networks (CNN)

DCNN

Recurrent Neural Networks (RNN)

LSM

GRU

LSTM

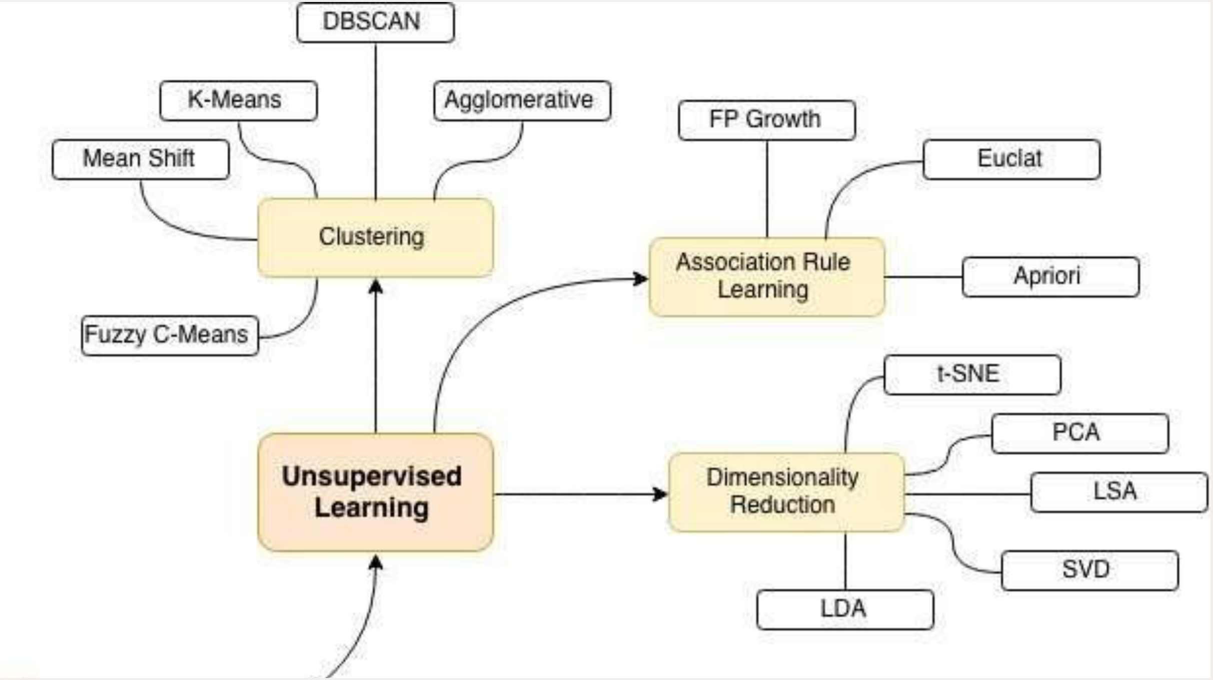
Generative Adversarial Networks (GAN)

Autoencoders

seq2seq

Perceptrons

What are we tackling today?



Ok, what are the different types?

Clustering

It is a technique used to **group data points or items into several groups (clusters) based on the similarity of features**. The goal is to partition the data into sets such that the data points in the same set are more similar to each other than to those in other sets. One popular method is K-Means clustering.

Real-world applications of clustering include customer segmentation, document clustering, and image segmentation.

Association Rule Learning

It is a method for discovering interesting **relations between variables in large databases**. It's a popular tool used for mining widespread relationships hidden in large data sets.

Real-world applications of associative rule learning include market basket analysis, web usage mining, and intrusion detection.

Dimensionality Reduction

It is the process of **reducing the number of random variables under consideration by obtaining a set of principal variables**. It can be divided into feature selection and feature extraction. Principal Component Analysis (PCA) is a popular method for feature extraction.

Real-world applications of dimensionality reduction include image compression, simplifying machine learning tasks, and exploratory data analysis.

Ok, what are the different types? *

Clustering

K-means
K-Means is a clustering algorithm that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

Association Rule Learning

Apriori
The Apriori algorithm is used for mining frequent itemsets and devising association rules in a transaction database. The algorithm uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function that exploits the downward closure property of support.

Dimensionality Reduction

PCA
PCA is a dimensionality reduction technique that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Clustering

Clustering is a technique for exploring raw, unlabeled data and breaking it down into groups (or clusters) based on similarities or differences. It is used in a variety of applications, including customer segmentation, fraud detection, and image analysis. Clustering algorithms split data into natural groups by finding similar structures or patterns in uncategorized data.

- Exclusive clustering: Data is grouped in a way where a single data point can only exist in one cluster. This is also referred to as “hard” clustering.
- Overlapping clustering: Data is grouped in a way where a single data point can exist in two or more clusters with different degrees of membership. This is also referred to as “soft” clustering.
- Hierarchical clustering: Data is divided into distinct clusters based on similarities, which are then repeatedly merged and organized based on their hierarchical relationships.
- Probabilistic clustering: Data is grouped into clusters based on the probability of each data point belonging to each cluster.

Clustering

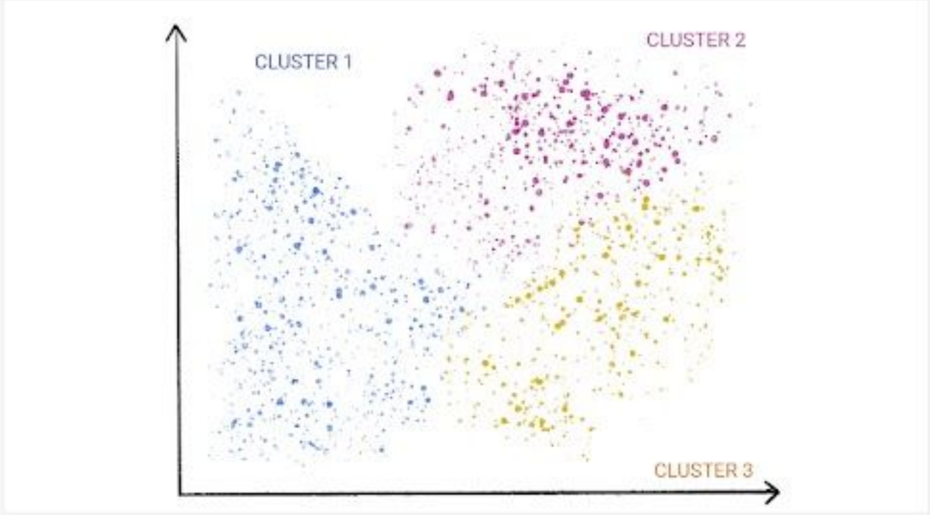


Figure 1. An ML model clustering similar data points.

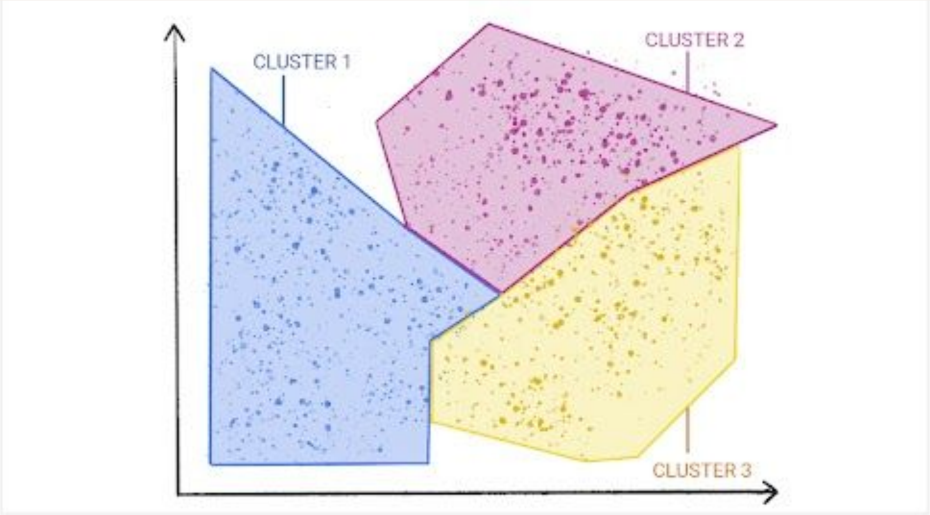
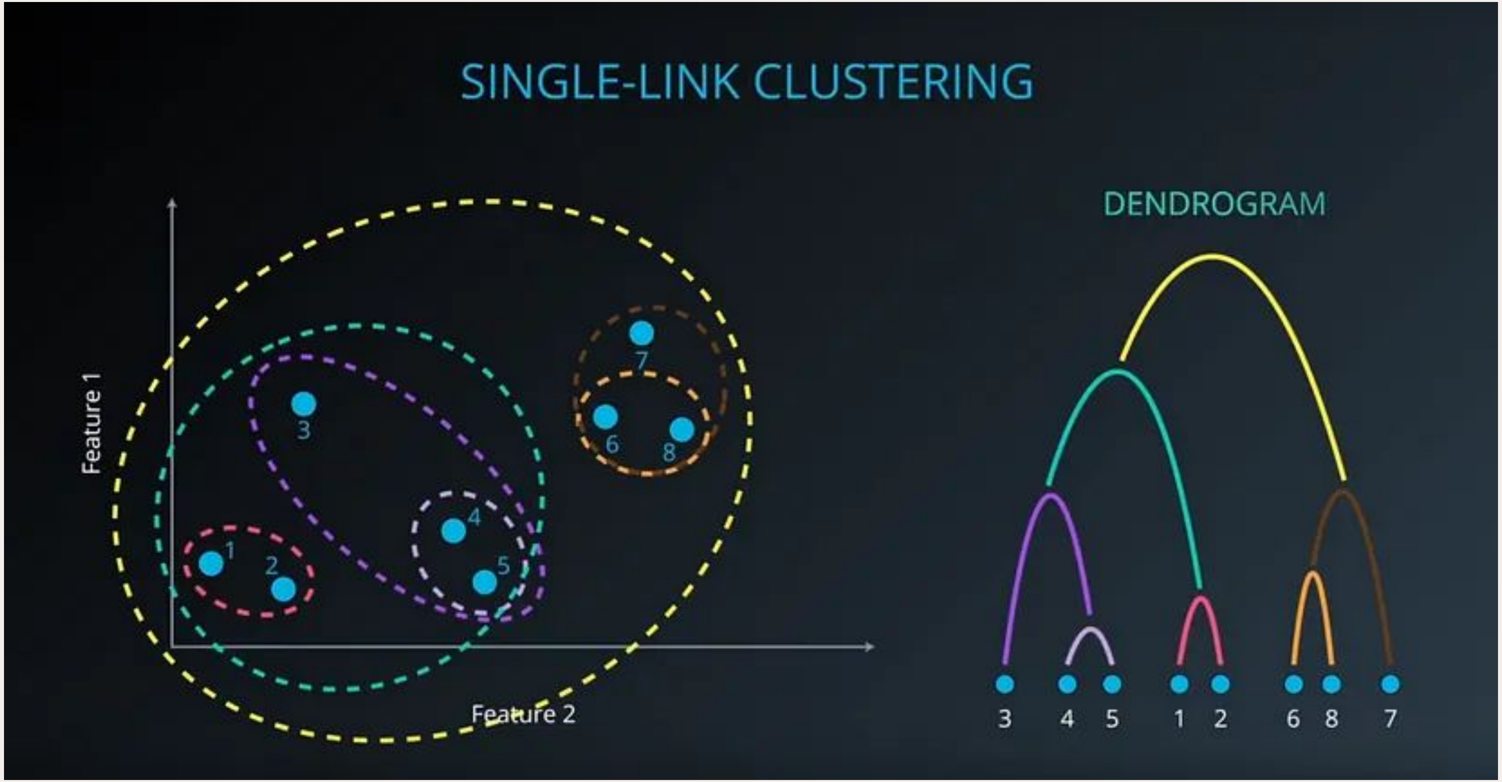
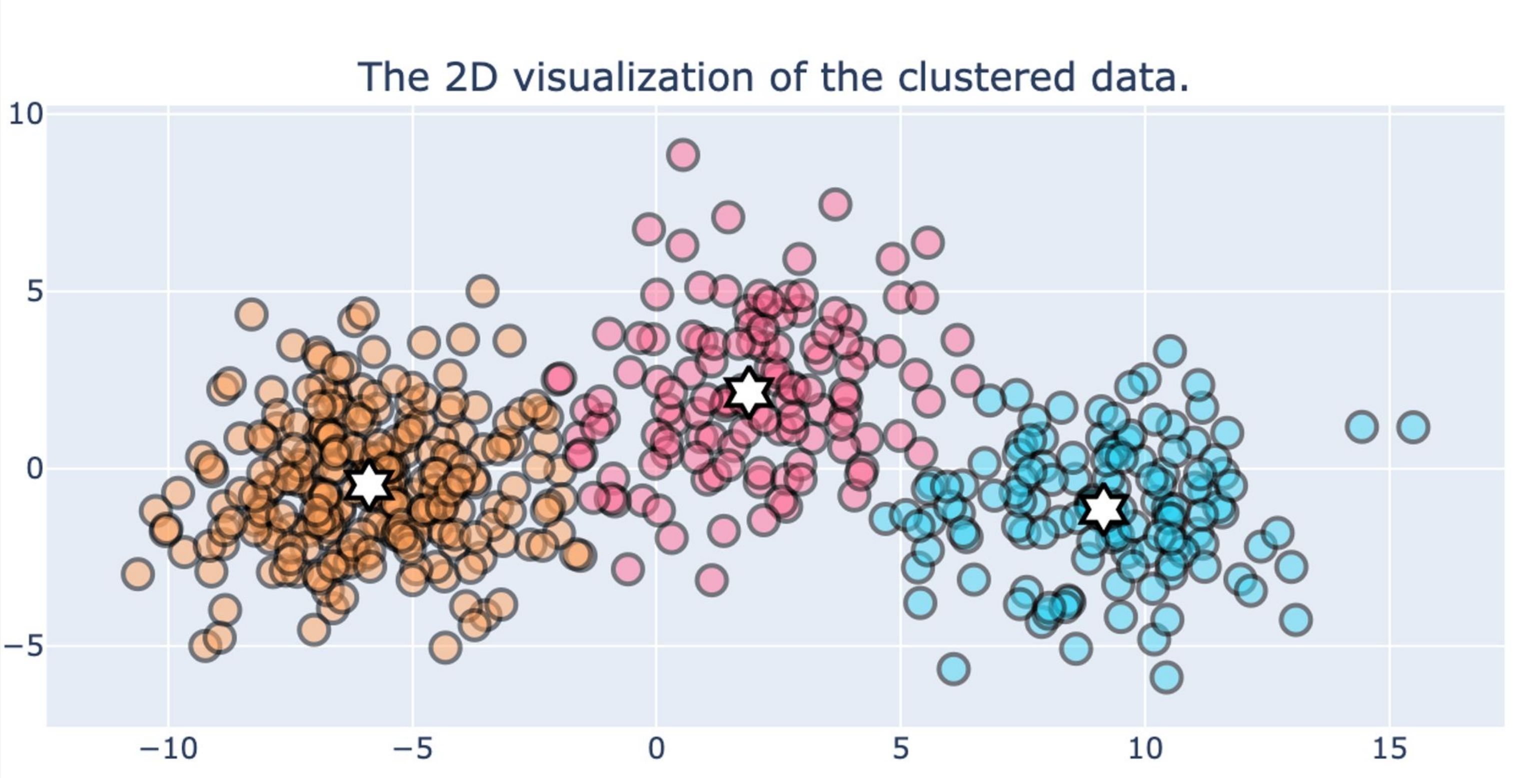


Figure 2. Groups of clusters with natural demarcations.

Clustering

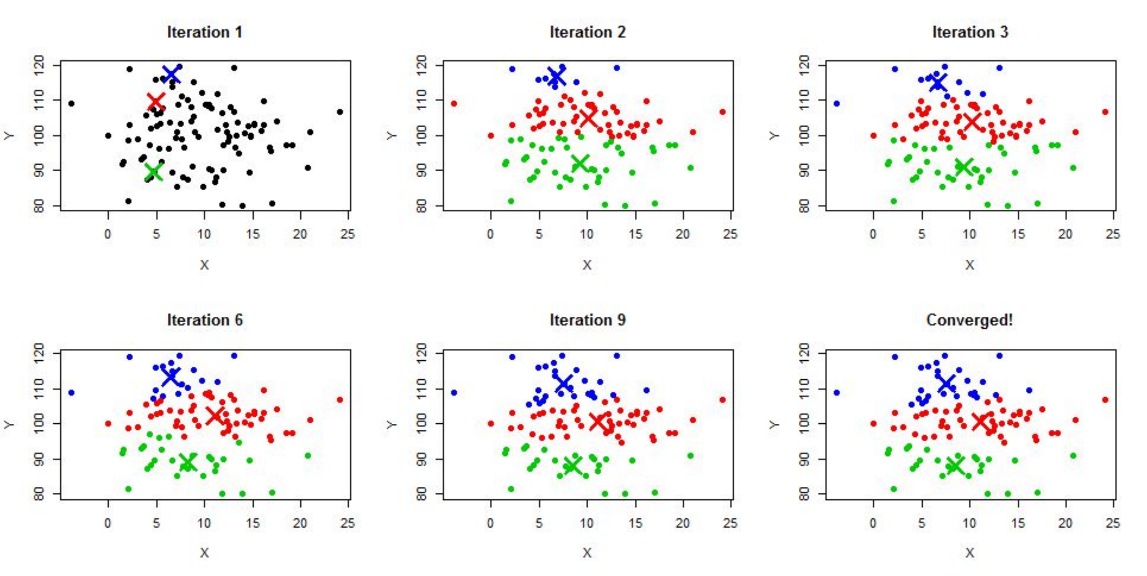

SINGLE-LINK CLUSTERING
Feature 1
Feature 2
1
2
3
4
5
6
7
8
DENDROGRAM
3
4
5
1
2
6
8
7

Clustering


The 2D visualization of the clustered data.
10
5
0
−5
−10
−5
0
5
10
15

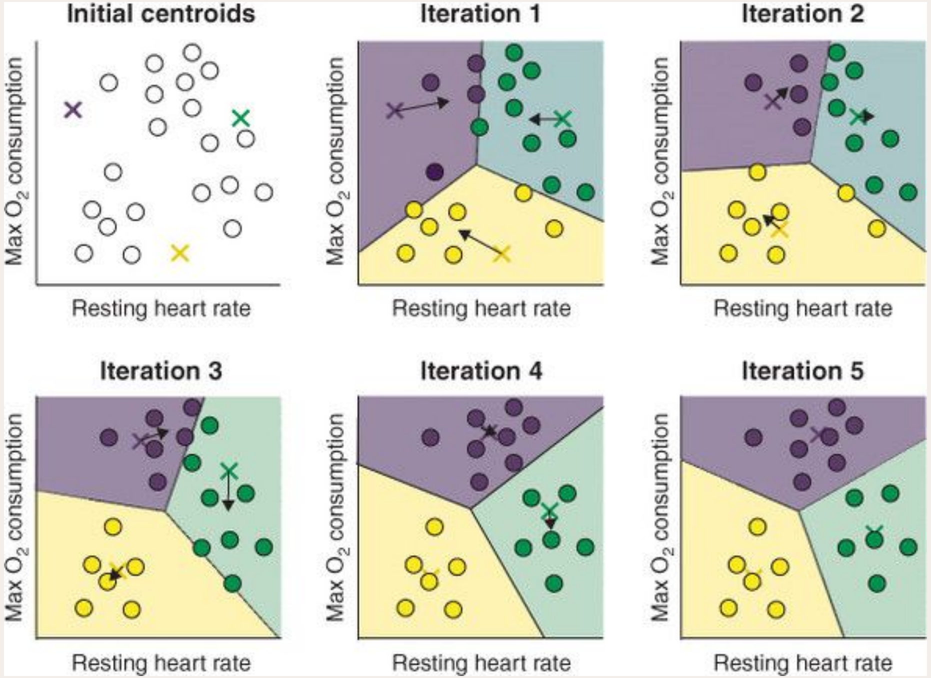
K-means

- K-Means clustering algorithm is easily the most popular and widely used, primarily because of the intuition and the ease of implementation. It is a **centroid-based algorithm** where the user must define the required number of clusters it wants to create.
- K-Means clustering is an **iterative algorithm** that creates non-overlapping clusters meaning each instance in your dataset can only belong to one cluster exclusively.



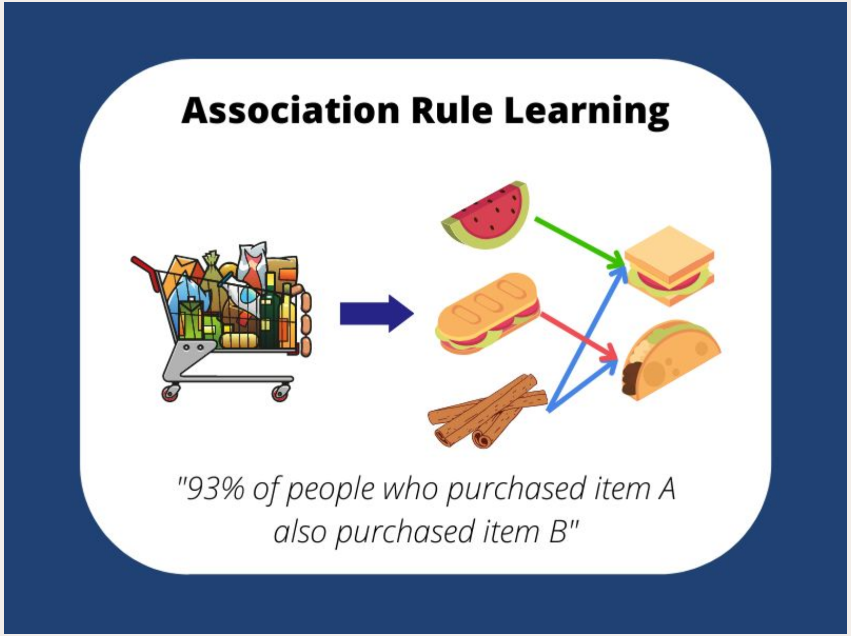
What do we mean by centroid based?

- Imagine you have a bunch of points on a 2D plane. You want to group these points into ‘K’ number of clusters. You start by randomly picking ‘K’ points on this plane. These points are your initial ‘centroids’. Then, for each point on the plane, you calculate its distance to all the centroids and assign it to the cluster of the closest centroid. Now, you have your initial clusters.
- But, these clusters might not be optimal. So, you calculate the mean (average) position of all points in each cluster and move the centroid to that mean position. This might change the cluster assignment of some points. So, you repeat the process of assigning points to the closest centroid and recalculating the centroids until the centroids don’t change anymore (or change very little). That’s when you say the algorithm has ‘converged’ and your clusters are ready.



Association Rule Learning

- This type of unsupervised machine learning takes a rule-based approach to discovering interesting relationships between features in a given dataset. It works by using a measure of interest to identify strong rules found within a dataset.
- We typically see association rule mining used for market basket analysis: this is a data mining technique retailers use to gain a better understanding of customer purchasing patterns based on the relationships between various products.



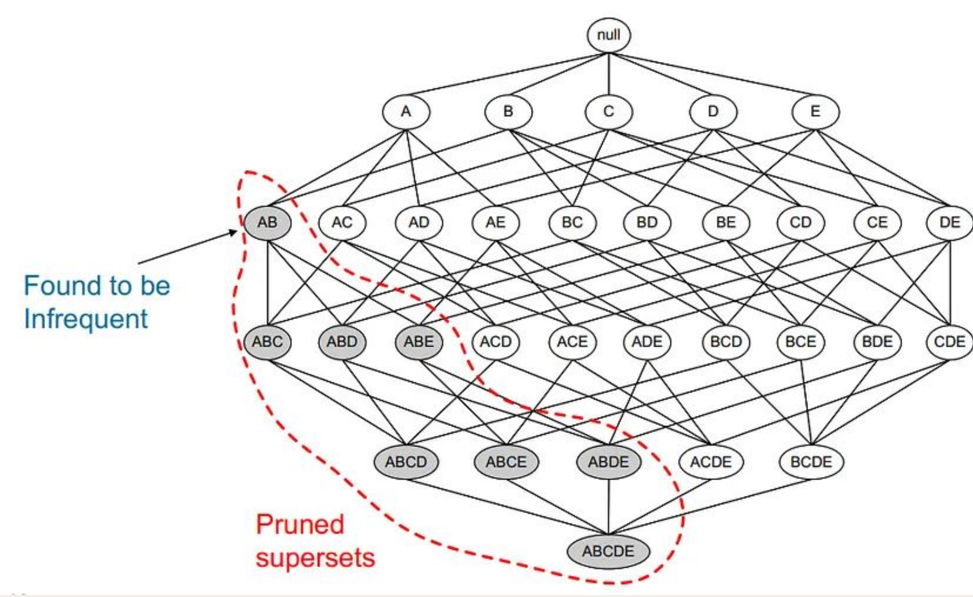
"93% of people who purchased item A also purchased item B"

Association Rule Learning

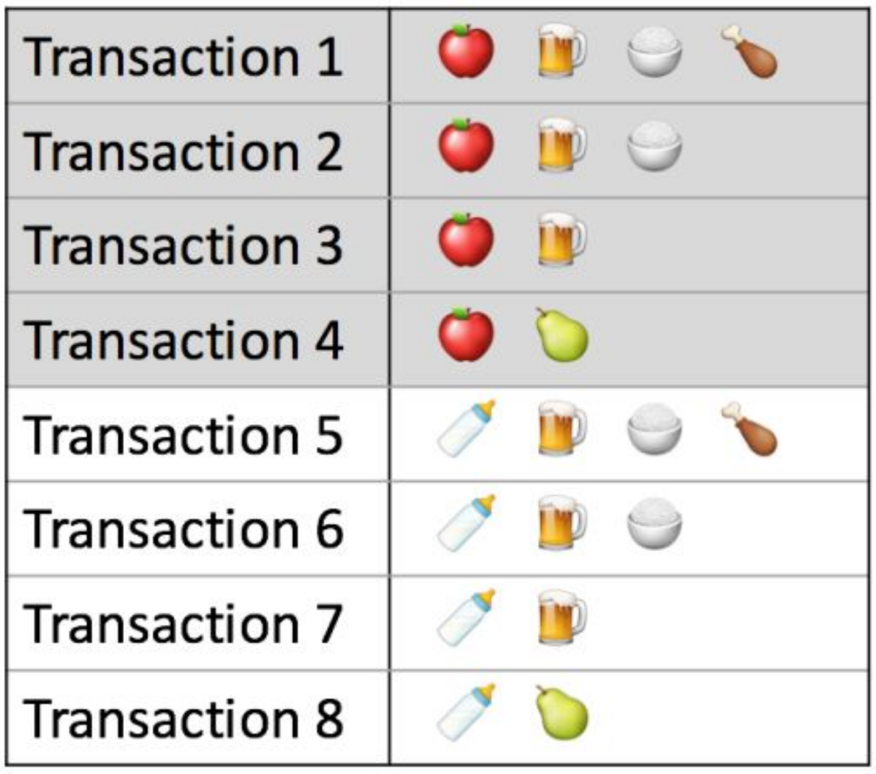
CUSTOMERS WHO BOUGHT THIS ITEM:
ALSO BOUGHT:
piccolo

Apriori

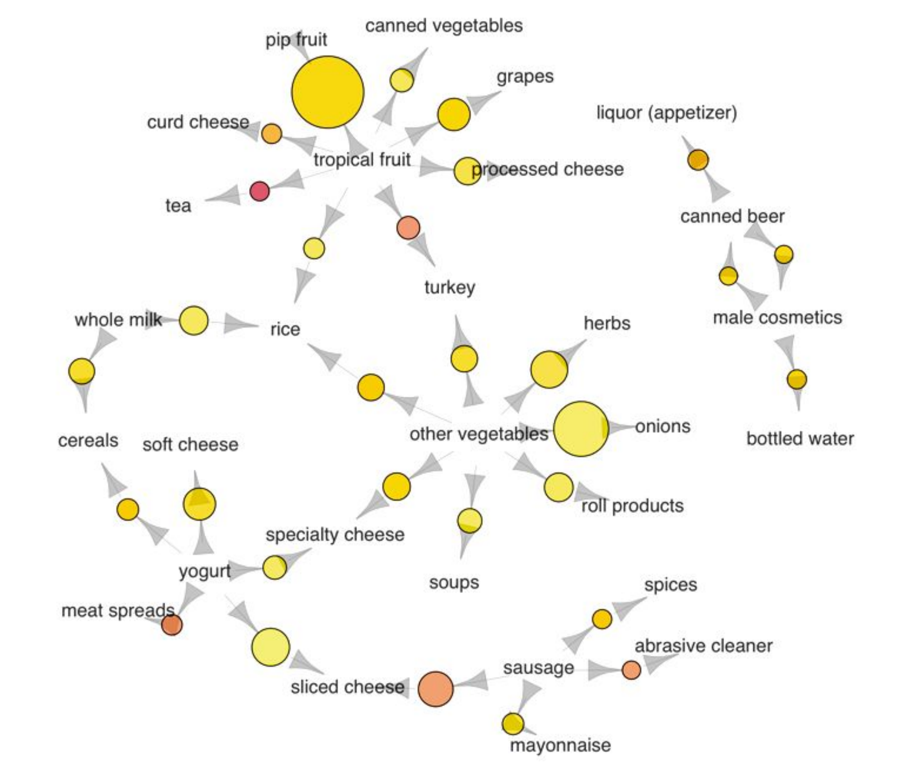
- The main idea of Apriori is
- All non-empty subsets of a frequent itemset must also be frequent.
- It's a bottom-up approach. We started from every single item in the itemset list. Then, the candidates are generated by self-joining. We extend the length of the itemsets one item at a time. The subset test is performed at each stage and the itemsets that contain infrequent subsets are pruned. We repeat the process until no more successful itemsets can be derived from the data.



Apriori

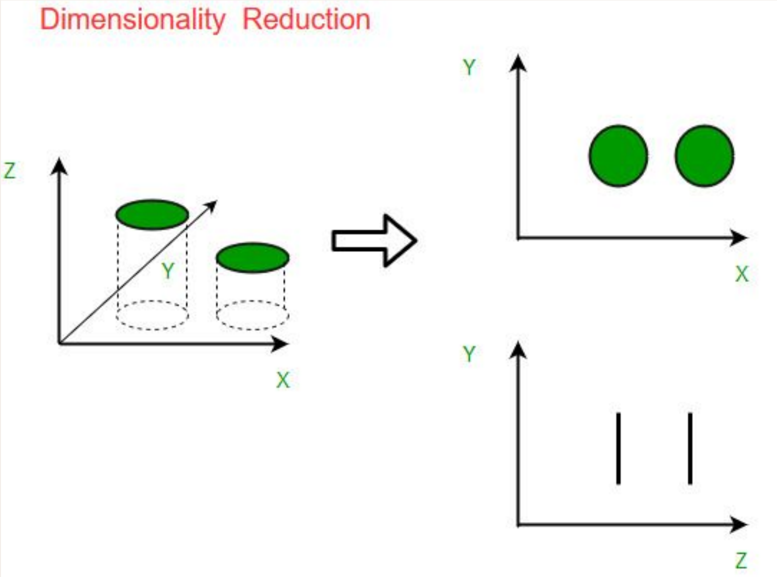


Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

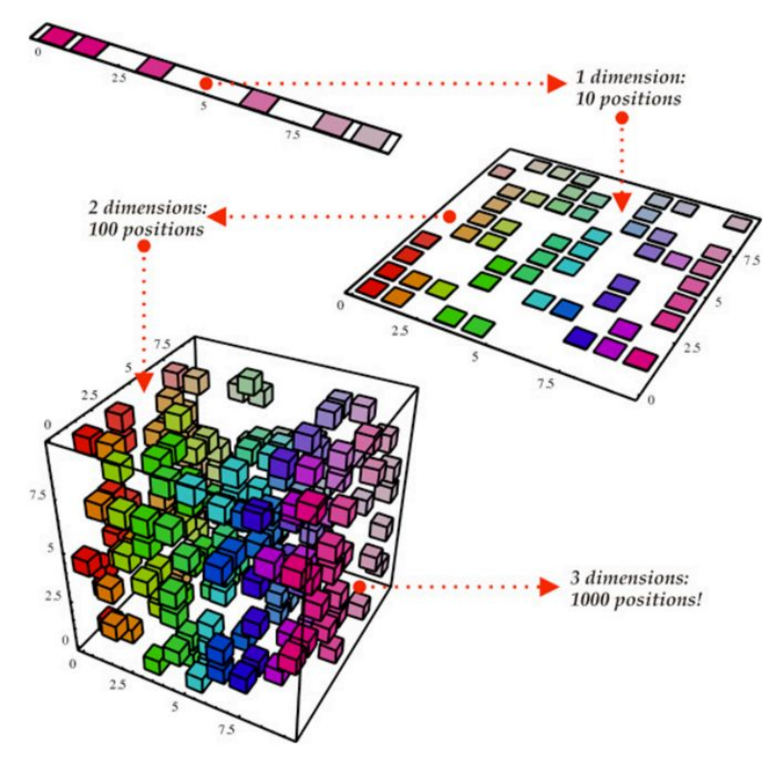


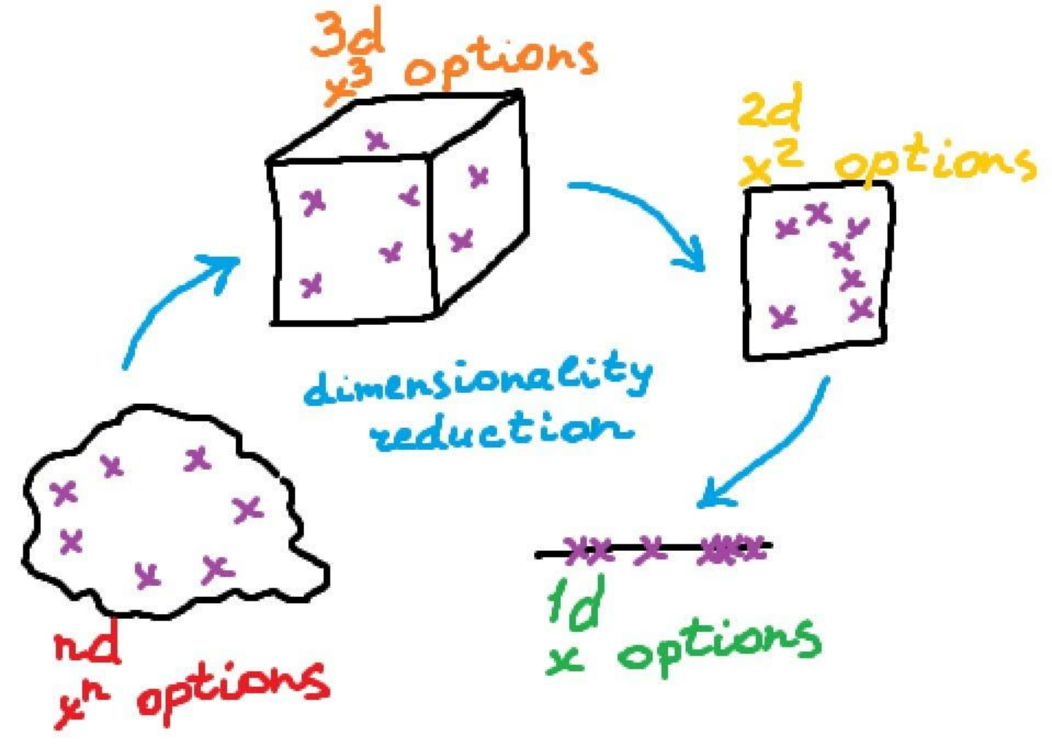
Dimensionality Reduction

- Dimensionality reduction is simply, the process of reducing the dimension of your feature set. Your feature set could be a dataset with a hundred columns (i.e features) or it could be an array of points that make up a large sphere in the three-dimensional space. reduces the number of features, or dimensions, in a dataset. More data is generally better for machine learning, but it can also make it more challenging to visualize the data.
- Dimensionality reduction extracts important features from the dataset, reducing the number of irrelevant or random features present.



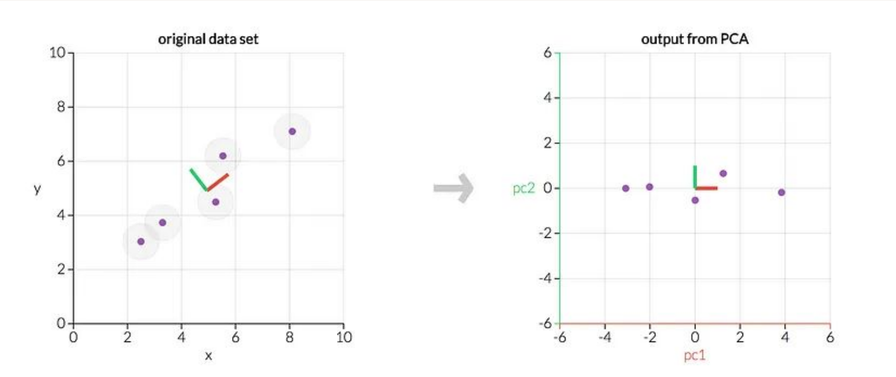
Dimensionality Reduction



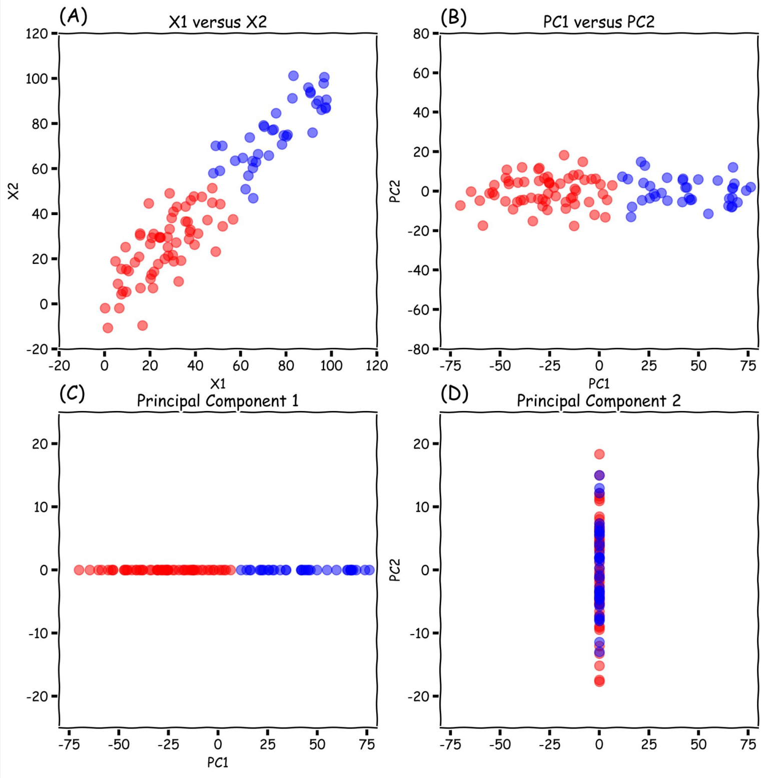


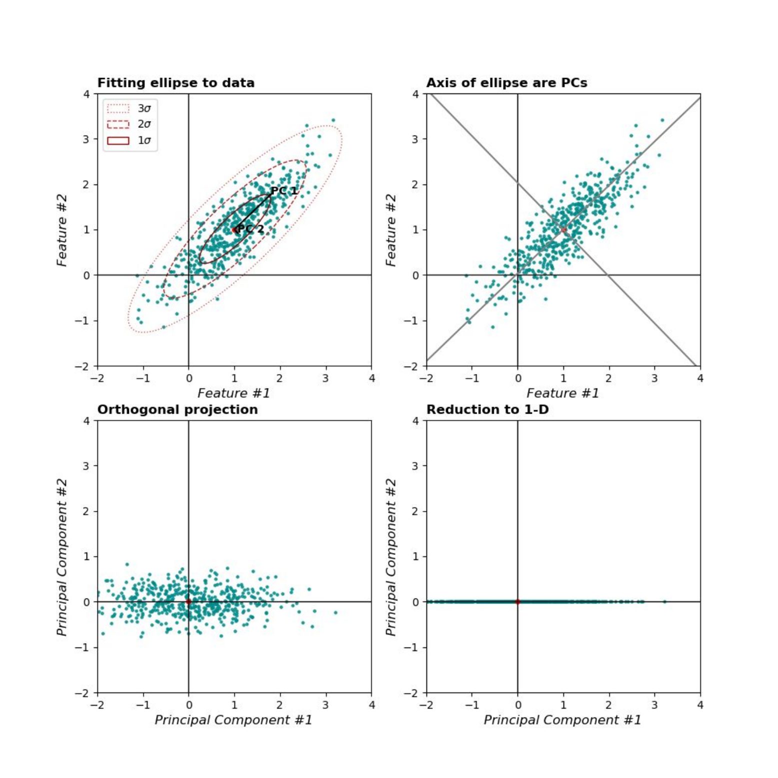
PCA

- Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- Principal component analysis is a technique for feature extraction — so it combines our input variables in a specific way, then we can drop the “least important” variables while still retaining the most valuable parts of all of the variables!
- As an added benefit, each of the “new” variables after PCA are all independent of one another. This is a benefit because the assumptions of a linear model require our independent variables to be independent of one another. If we decide to fit a linear regression model with these “new” variables (see “principal component regression” below), this assumption will necessarily be satisfied.



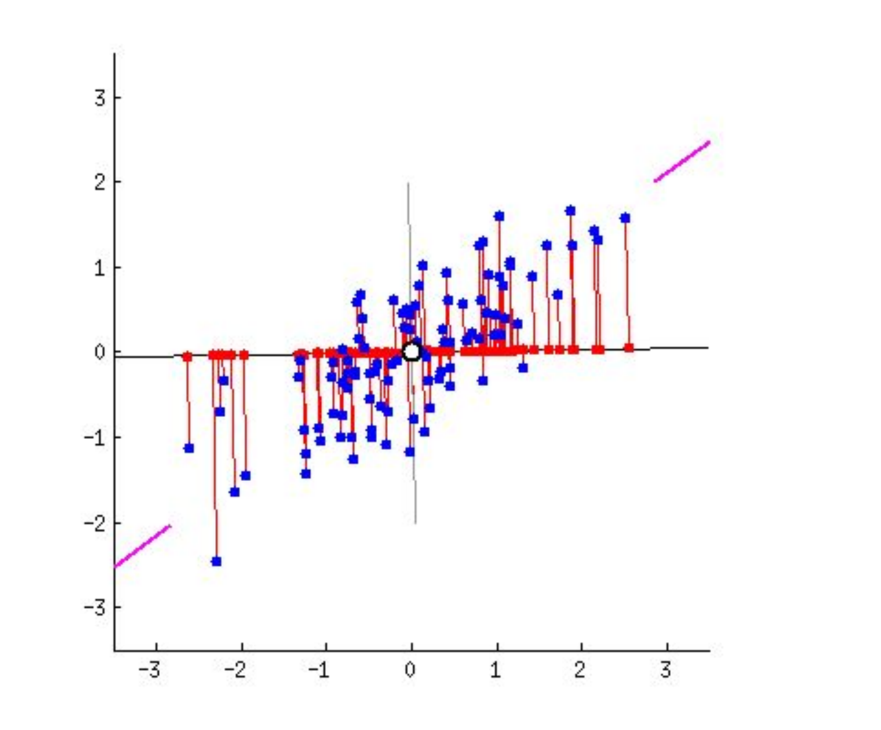
PCA


(A)
X1 versus X2
X2
X1
120
100
80
60
40
20
0
-20
-20
0
20
40
60
80
100
120
(B)
PC1 versus PC2
PC2
PC1
80
60
40
20
0
-20
-40
-60
-80
-75
-50
-25
0
25
50
75
(C)
Principal Component 1
PC1
20
10
0
-10
-20
-75
-50
-25
0
25
50
75
(D)
Principal Component 2
PC2
20
10
0
-10
-20
-75
-50
-25
0
25
50
75


Fitting ellipse to data
3σ
2σ
1σ
PC 1
PC 2
Feature #2
Feature #1
Axis of ellipse are PCs
Feature #2
Feature #1
Orthogonal projection
Principal Component #2
Principal Component #1
Reduction to 1-D
Principal Component #2
Principal Component #1
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4

PCA

As there are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set. For example, let's assume that the scatter plot of our data set is as shown below, can we guess the first principal component ? Yes, it's approximately the line that matches the purple marks because it goes through the origin and it's the line in which the projection of the points (red dots) is the most spread out. Or mathematically speaking, it's the line that maximizes the variance (the average of the squared distances from the projected points (red dots) to the origin).


3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3