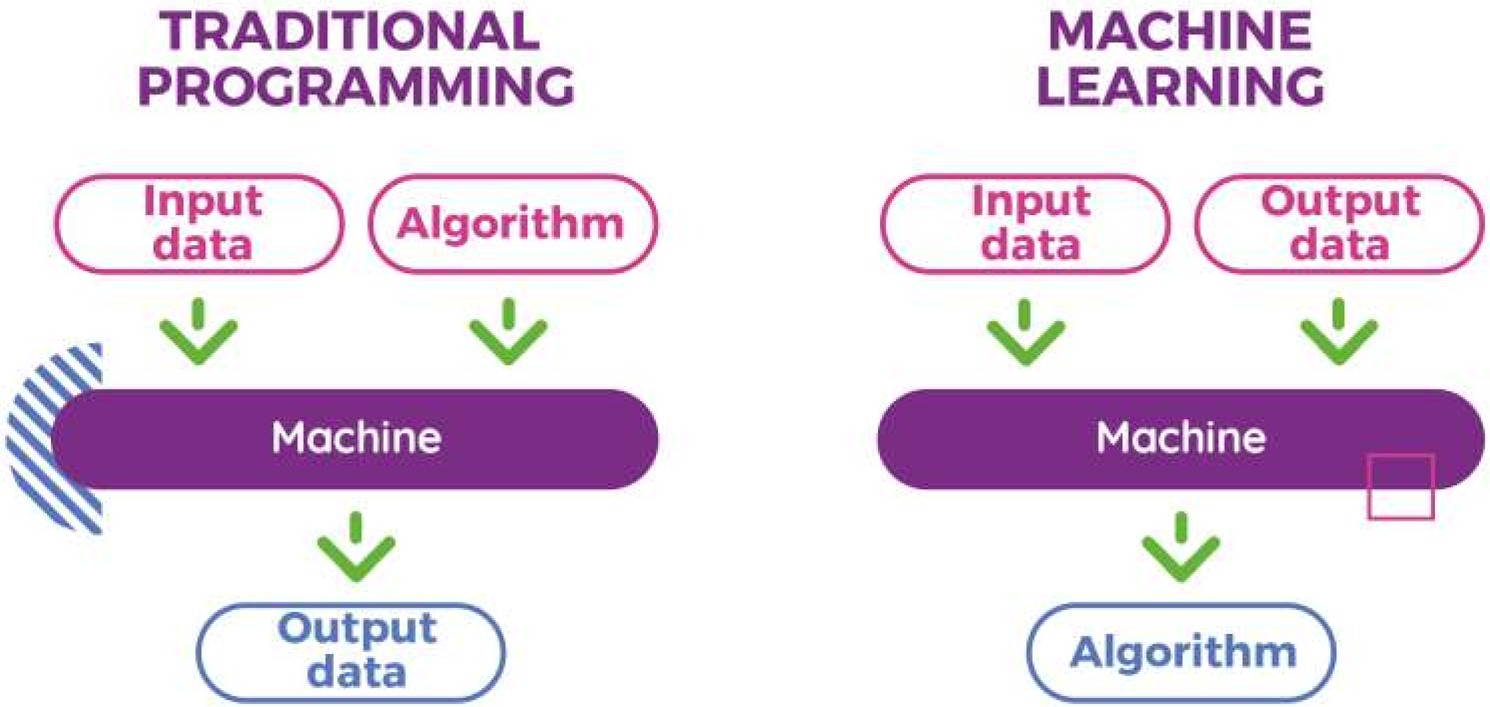
Machine Learning: Unsupervised

Andi Sulasikin S.T., M.Sc.
Data Science Lead

Jakarta Smart City

Dinas Komunikasi, Informatika dan Statistik, Provinsi DKI Jakarta

Machine Learning



Machine Learning

Machine learning is the subfield of computer science that gives "computers the ability to learn without being explicitly programmed." - term coined by Arthur Samuel 1959 while at IBM

The study of algorithms that can learn from data.

Supervised Learning vs. Unsupervised Learning

Supervised learning: discover patterns in the data that relate data attributes with a target (class) attribute.

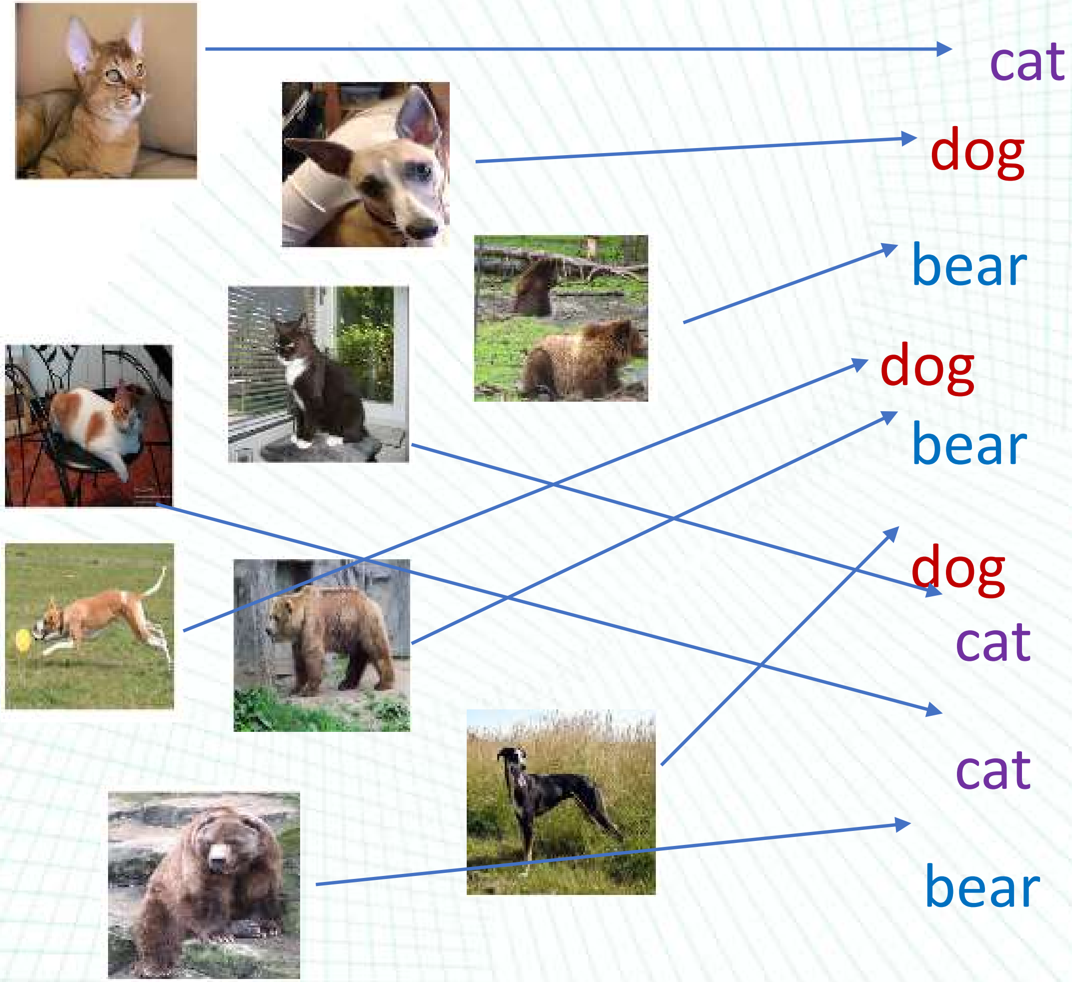
- These patterns are then utilized to predict the values of the target attribute in future data instances

Unsupervised learning: The data have no target attribute**.**

- We want to explore the data to find some intrinsic structures in them.

Supervised Learning vs. Unsupervised Learning

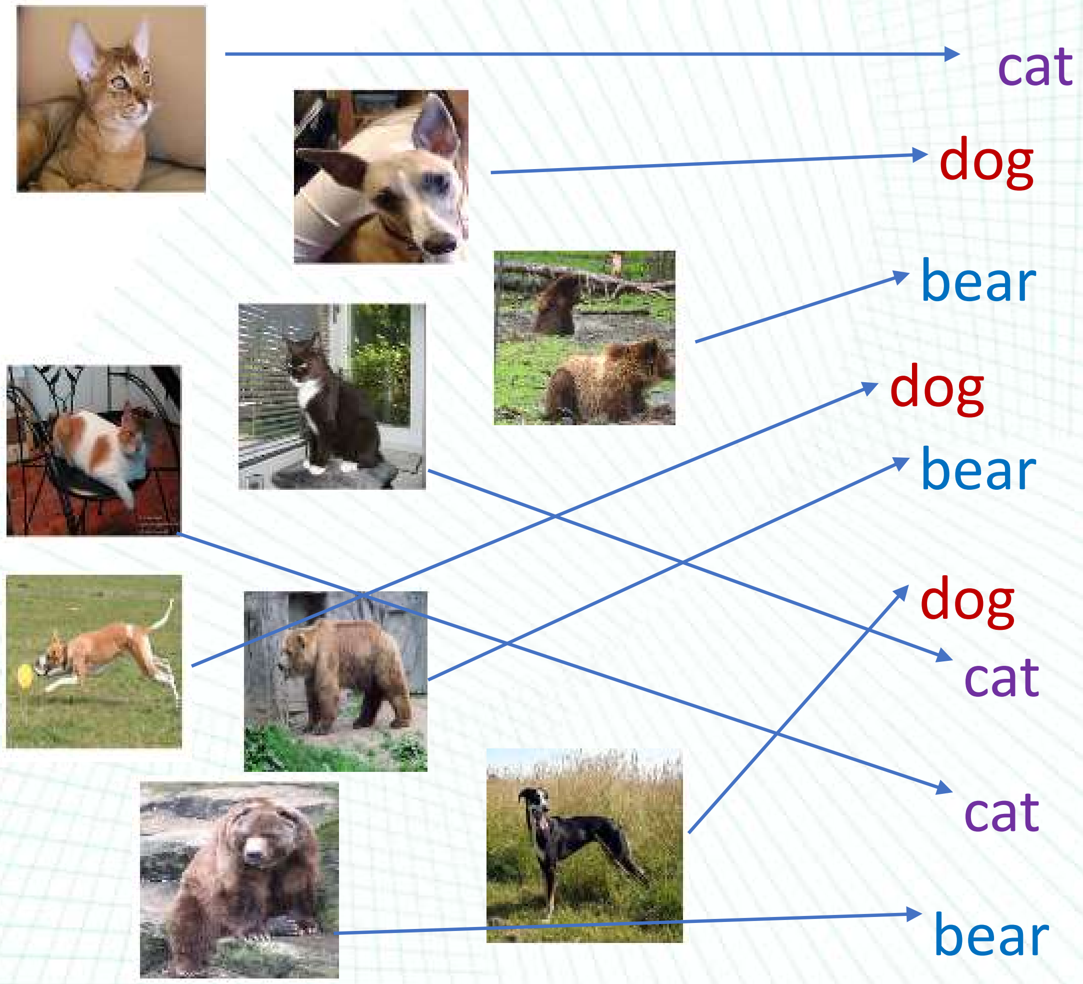
$x \to y$



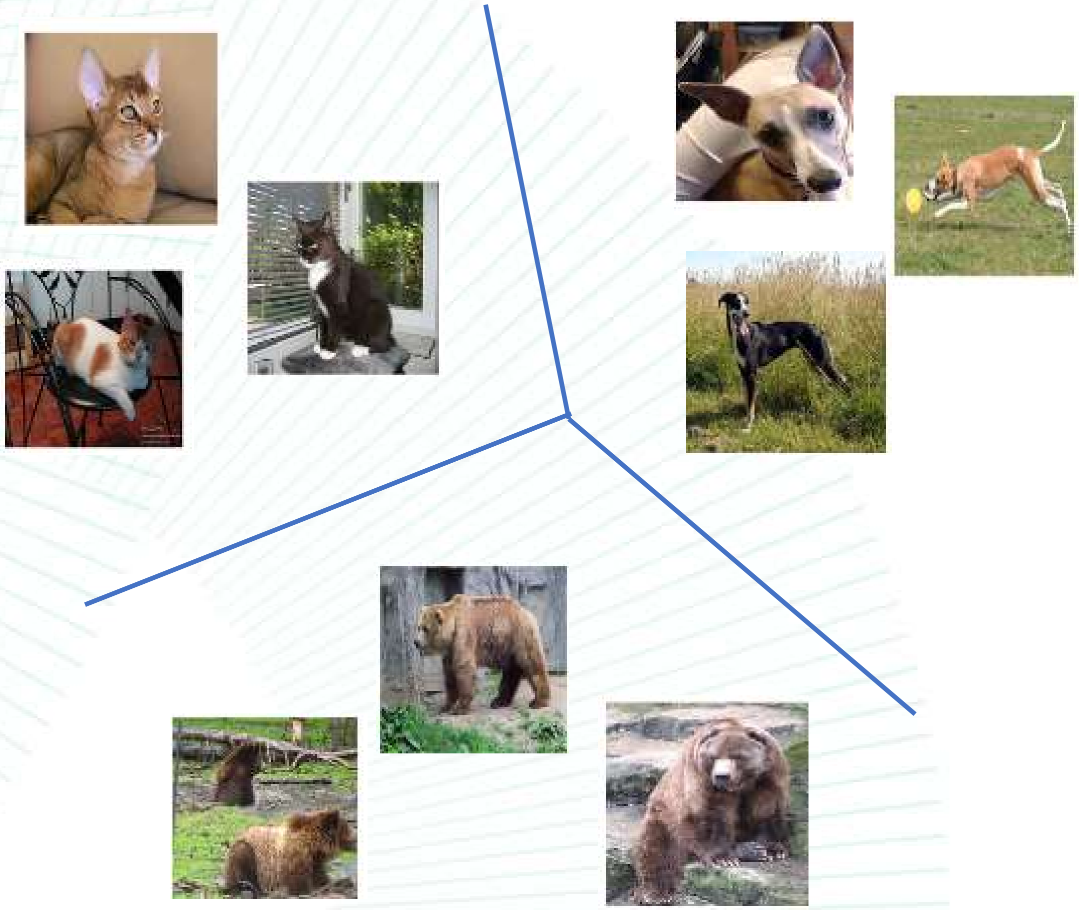
x

Supervised Learning vs. Unsupervised Learning

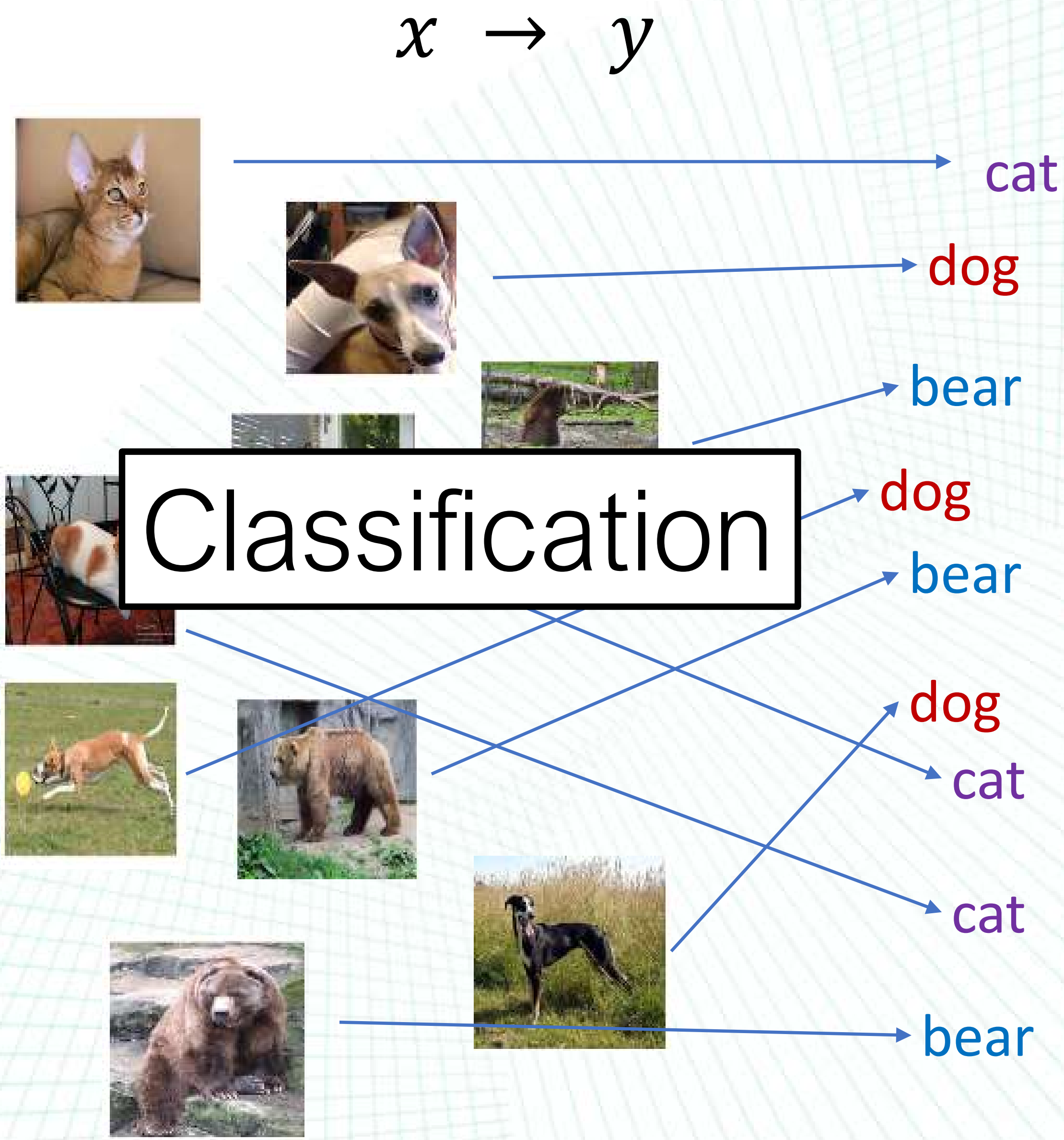
$x \rightarrow y$



x



Supervised Learning vs. Unsupervised Learning



$x \rightarrow y$

cat

dog

bear

dog

bear

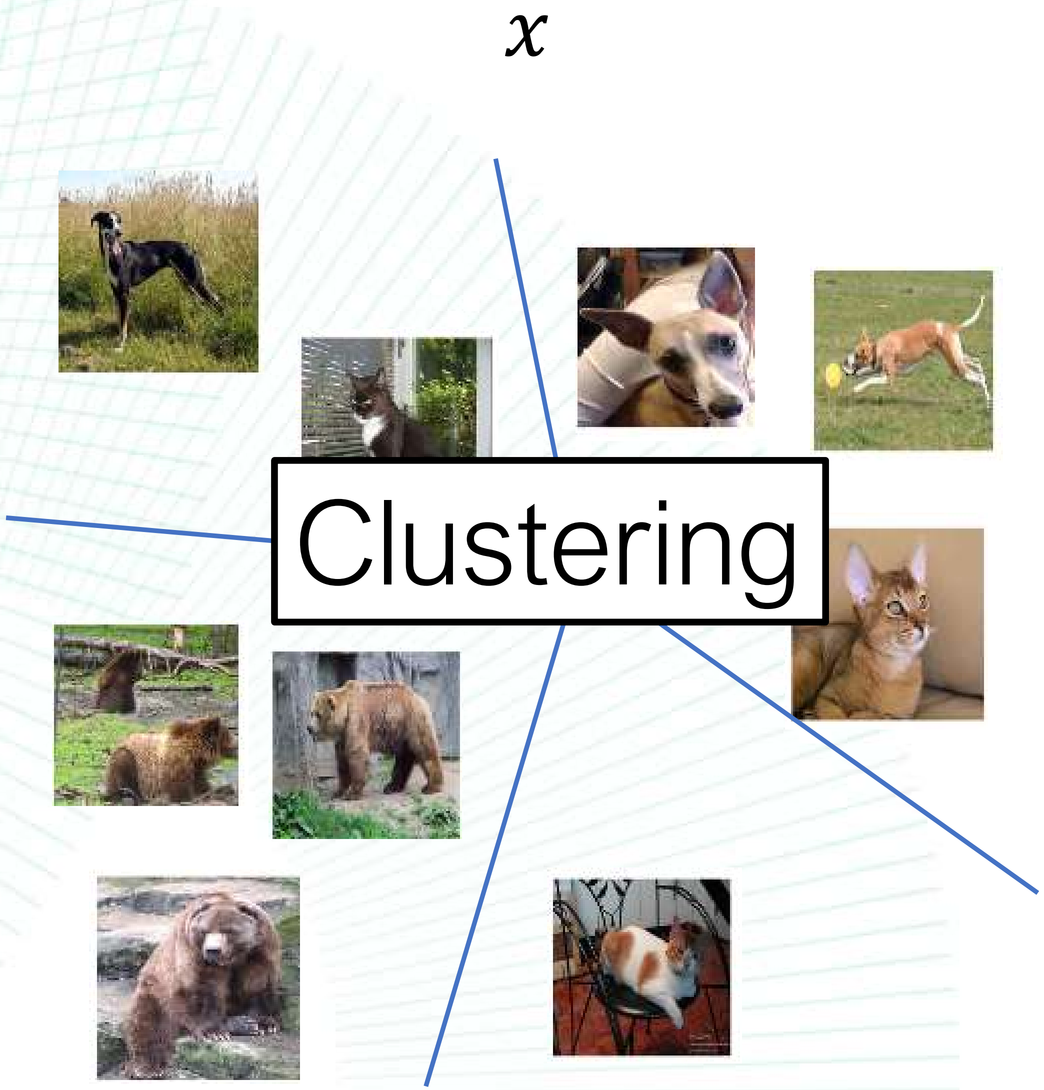
dog

cat

cat

bear

Classification



x

Clustering

Supervised Learning vs. Unsupervised Learning

Parameter	Supervised Learning	Unsupervised Learning
Dataset	Labelled Dataset	Unlabeled Dataset
Method of Learning	Guided Learning	Algorithm learns by itself using dataset
Complexity	Simpler Method	Computationally Complex
Accuracy	More Accurate	Less Accurate

When to Use Unsupervised Learning

When you do not need to supervise the model. Instead, you need to allow the model to work on its own to discover information. It mainly deals with the unlabeled data.

Why We Need to Use Unsupervised Learning

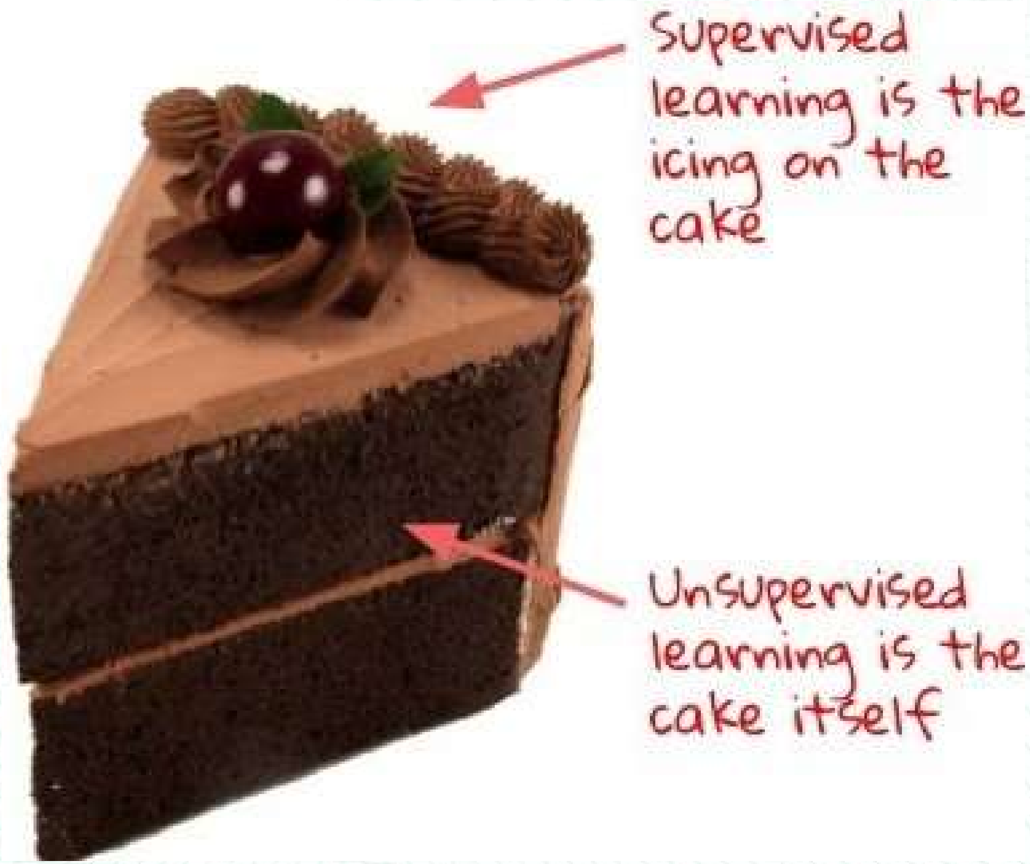
Here, are **prime reasons for using Unsupervised Learning:**

- finds all kind of unknown patterns in data.
- help you to find features which can be useful for categorization.
- It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.
- It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention.

Disadvantages of Unsupervised Learning

- You may never know the method of the data was sorted by algorithm
- It provides less accurate outputs
- Output obtained may not be what the user was expecting due to data interpretation mismatch
- Output obtained has to be understood by user and mapped with corresponding labels

The Future of ML is Unsupervised Learning


Supervised learning is the icing on the cake
Unsupervised learning is the cake itself

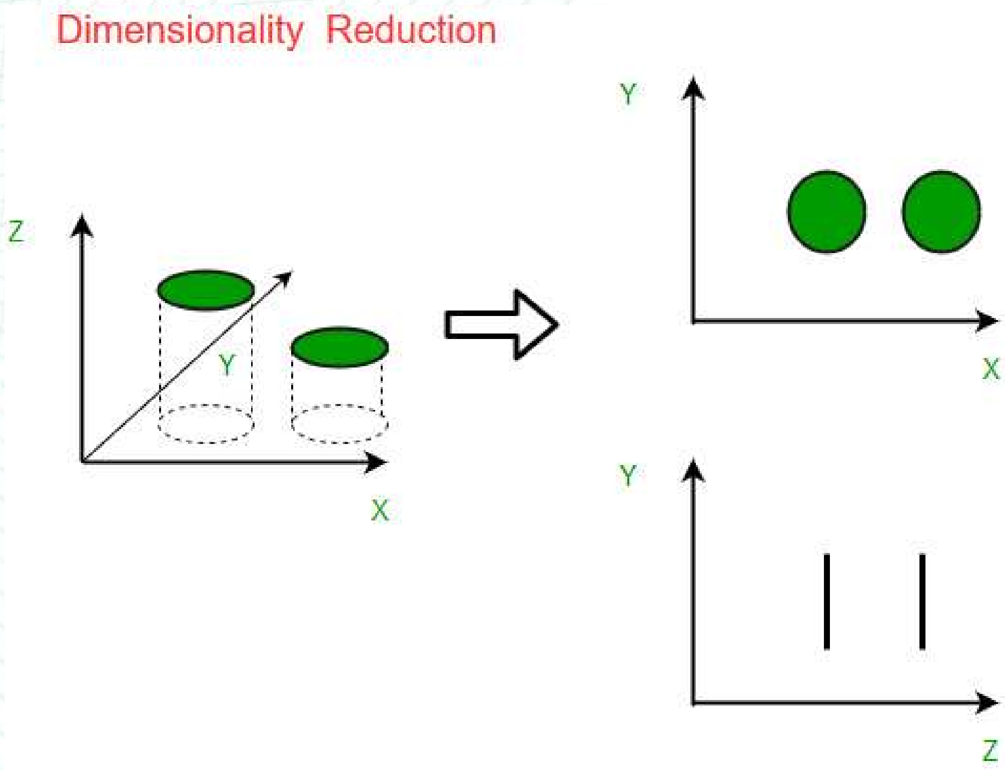
Humans learn mostly through unsupervised learning: we absorb vast amounts of data from our surroundings without needing a label.

To reach true machine intelligence (i.e., a machine that thinks and learns for itself), **ML needs to get better at unsupervised learning** - it should learn without us having to feed it labels or explicit instructions.

Dimensionality Reduction

Dimensionality reduction refers to techniques for reducing the number of input variables in training data.

When dealing with high dimensional data, it is often useful to reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the “essence” of the data. This is called dimensionality reduction. — Page 11, Machine Learning: A Probabilistic Perspective, 2012.

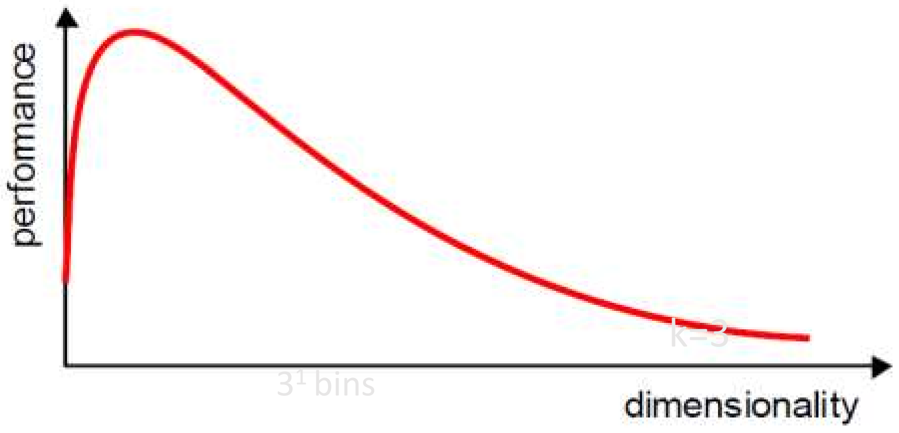


Curse of Dimensionality

- Increasing the number of features will not always improve classification accuracy.

- In practice, the inclusion of more features might actually lead to worse performance.

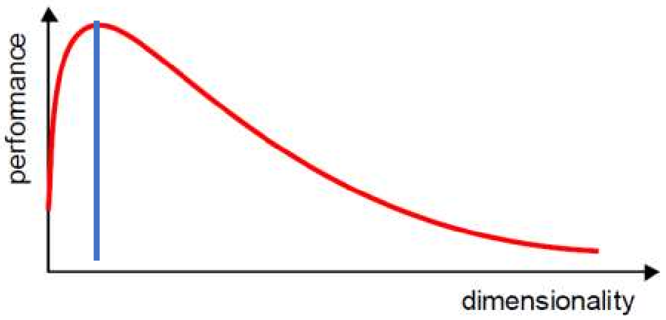
- The number of training examples required increases exponentially with dimensionality d (i.e., kd).



Dimensionality Reduction

What is the objective?

Choose an optimum set of features of lower dimensionality to improve classification accuracy.



Different methods can be used to reduce dimensionality:

- **Feature extraction**
- **Feature selection**

Dimensionality Reduction (Cont'd)

Feature extraction: finds a set of new features (i.e., through some mapping f()) from the existing features.

The mapping f() could be **linear or non-linear**

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ . \\ x_N \end{bmatrix} \xrightarrow{f(\mathbf{x})} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_K \end{bmatrix}$$

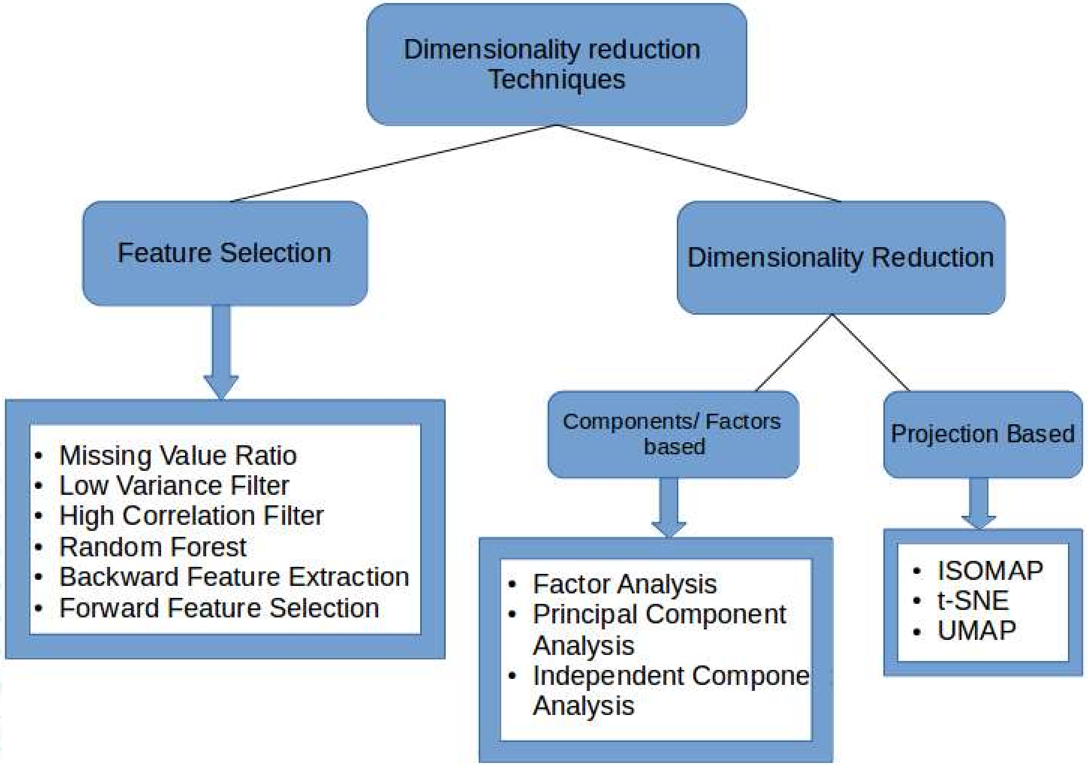
K<<N

Feature selection: chooses a subset of the original features.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ . \\ x_N \end{bmatrix} \to \mathbf{y} = \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ . \\ . \\ x_{i_K} \end{bmatrix}$$

K<<N

Dimensionality Reduction Techniques



Dimensionality reduction Techniques

Feature Selection

- Missing Value Ratio
- Low Variance Filter
- High Correlation Filter
- Random Forest
- Backward Feature Extraction
- Forward Feature Selection

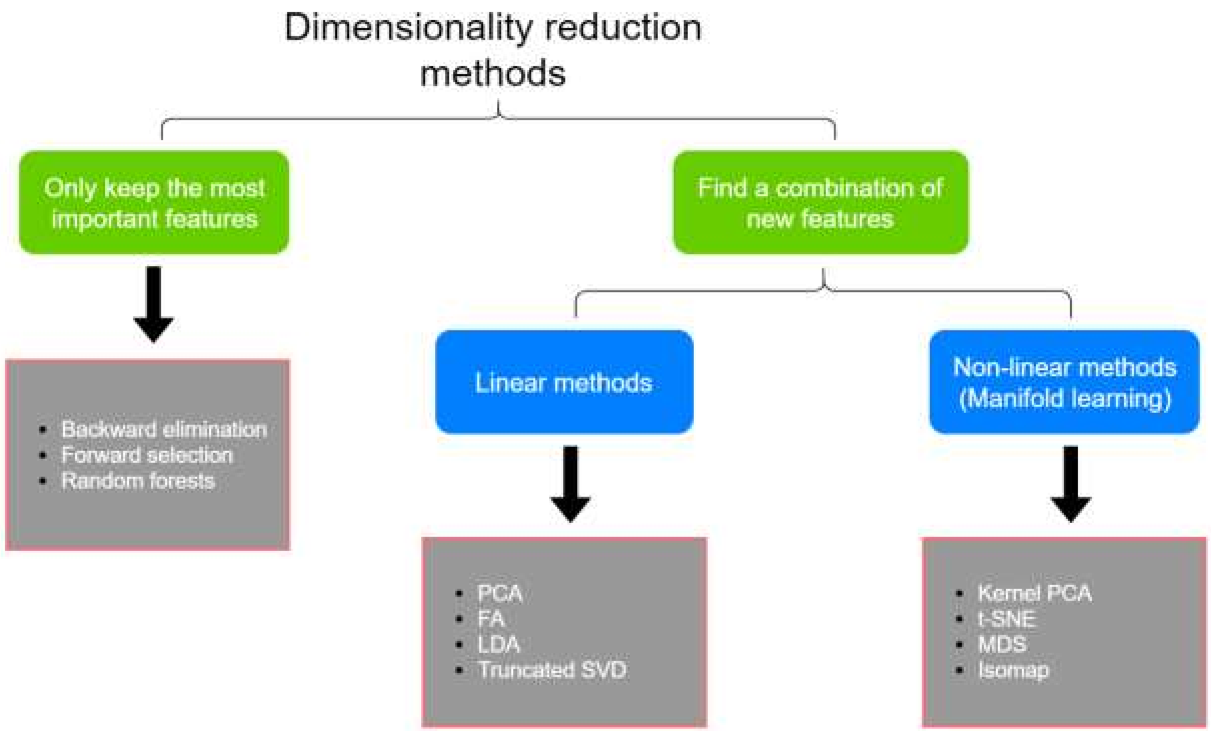
Dimensionality Reduction

Components/ Factors based

- Factor Analysis
- Principal Component Analysis
- Independent Compone Analysis

Projection Based

- ISOMAP
- t-SNE
- UMAP



Dimensionality reduction methods

Only keep the most important features

- Backward elimination
- Forward selection
- Random forests

Find a combination of new features

Linear methods

- PCA
- FA
- LDA
- Truncated SVD

Non-linear methods (Manifold learning)

- Kernel PCA
- t-SNE
- MDS
- Isomap

Image copyright: Rukshan Pramoditha

The Importance of Dimensionality Reduction

- A lower number of dimensions in data means less training time and less computational resources and increases the overall performance of machine learning algorithms
- avoids the problem of overfitting
- extremely useful for data visualization
- takes care of multicollinearity
- very useful for factor analysis
- removes noise in the data
- can be used for image compression
- can be used to transform non-linear data into a linearly-separable form

Principal Component Analysis (PCA)

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's **often used to make data easy to explore and visualize.**

Objective of PCA is to perform dimensionality reduction while preserving as much of the randomness in the high-dimensional space as possible

Principal Component Analysis (PCA)

It **takes your cloud of data points**, and rotates it such that the maximum variability is visible.

PCA is mainly concerned with **identifying correlations** in the data.

Measuring Correlation

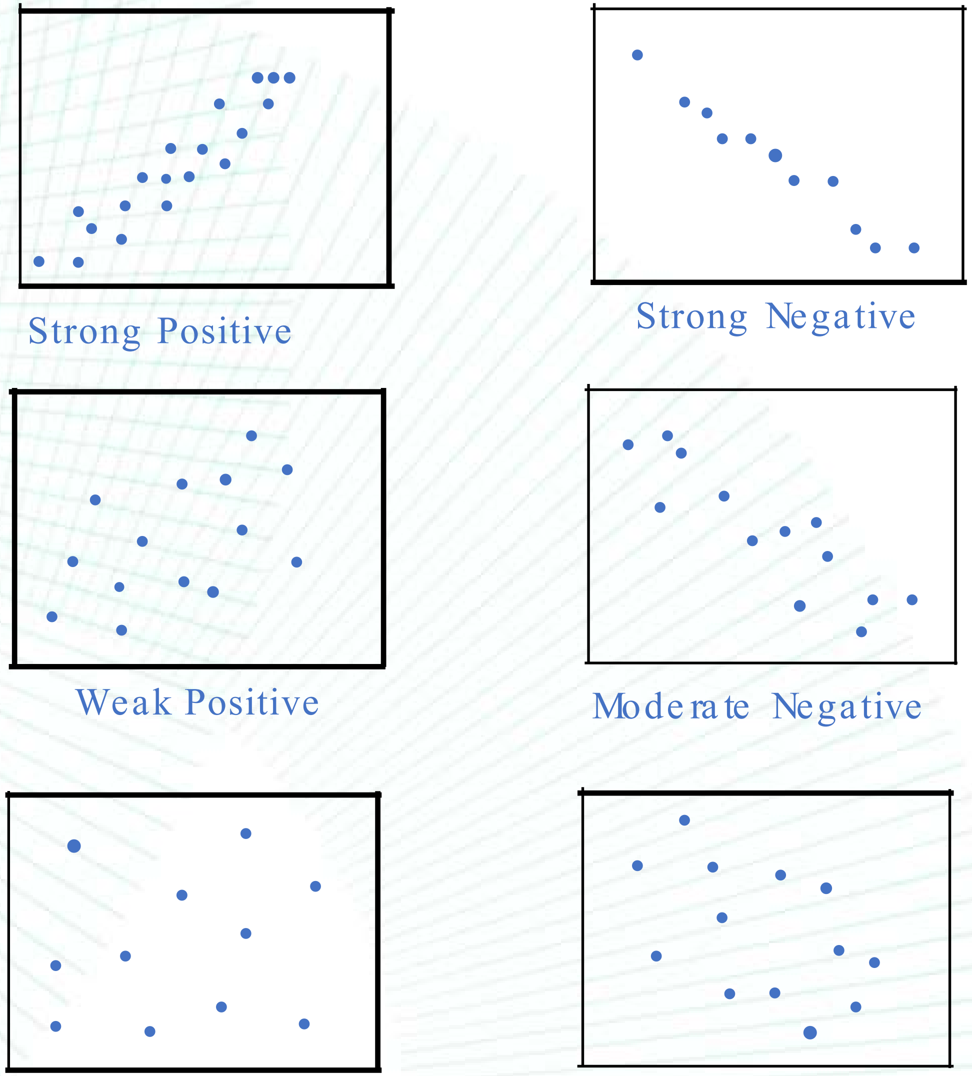
Degree and type of relationship between any two or more quantities (variables) in which they vary together over a period

Correlation can vary from +1 to -1.

Values close to +1 indicate a h igh degree of positive correlation, and values close to -1 indicate a high degree of negative correlat ion.

Values close to zero indicate poor correlation of either kind, and 0 indicates no correlation at all

Degree of Correlation



Strong Positive

Strong Negative

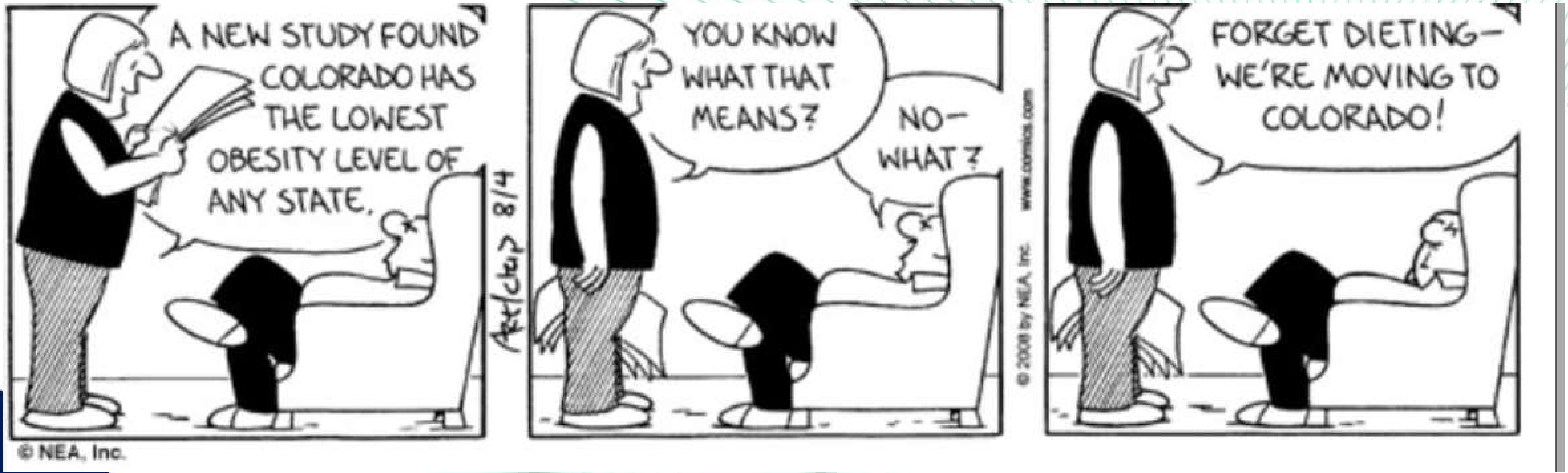
Weak Positive

Moderate Negative

None

Weak Negative

Beware: Correlation does not imply causation

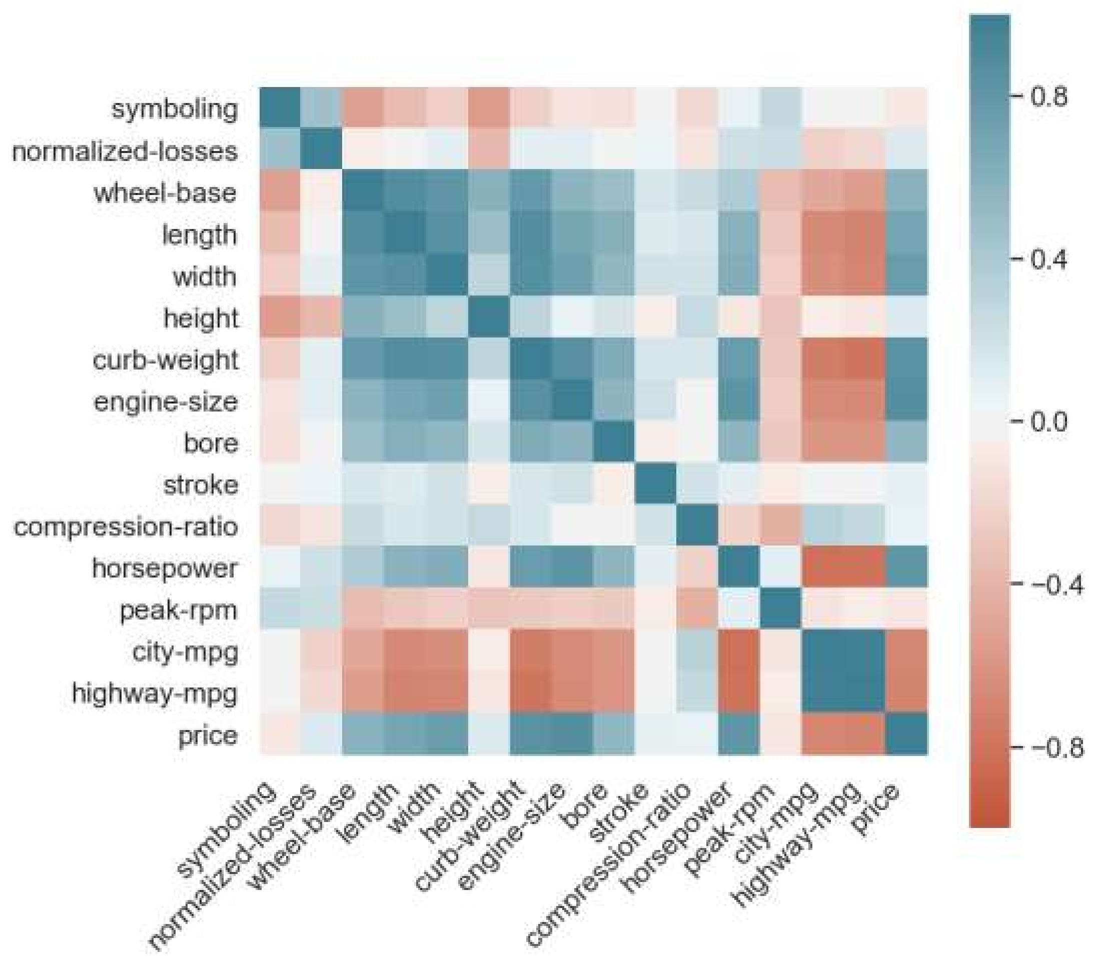

A NEW STUDY FOUND COLORADO HAS THE LOWEST OBESITY LEVEL OF ANY STATE.
8/4
YOU KNOW WHAT THAT MEANS?
NO— WHAT?
FORGET DIETING— WE'RE MOVING TO COLORADO!

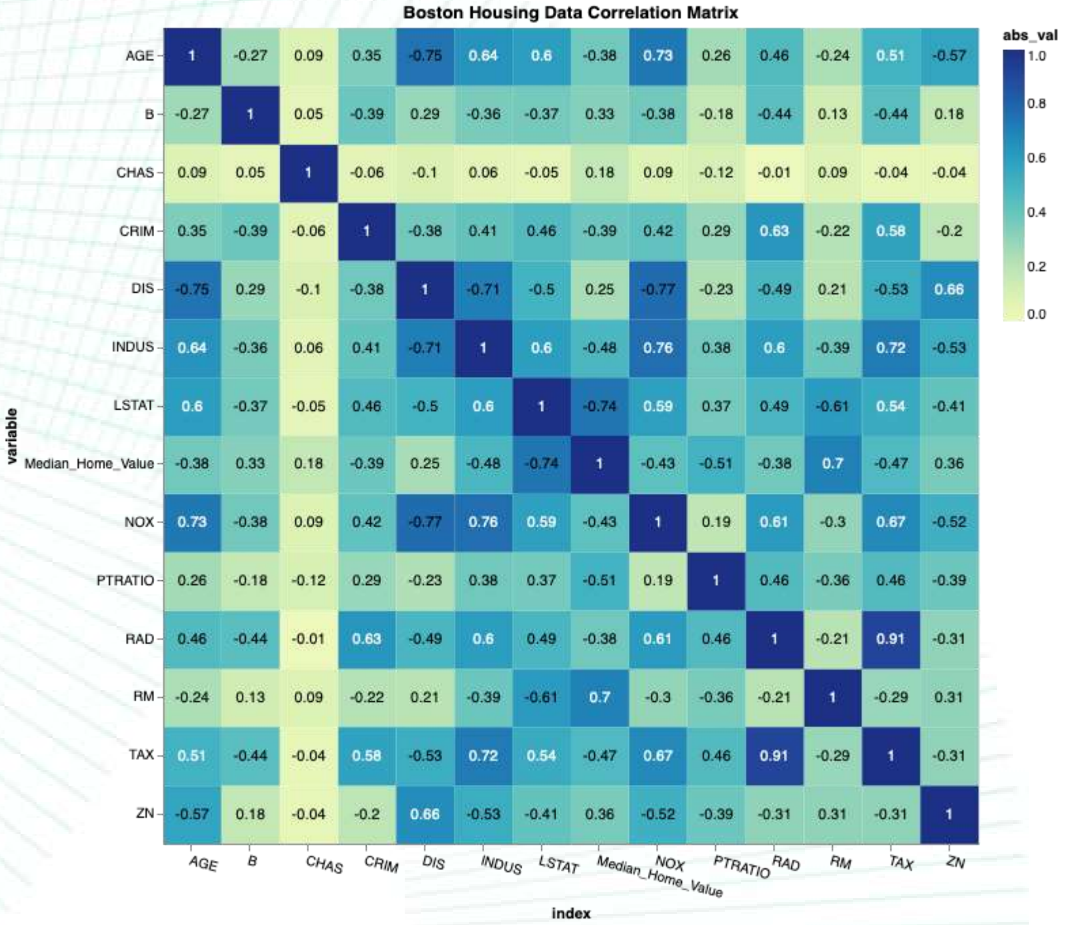
© 2008 by NEA, Inc. www.comics.com

© NEA, Inc.

Correlation Matrix

It shows at a glance how variables correlate with each other

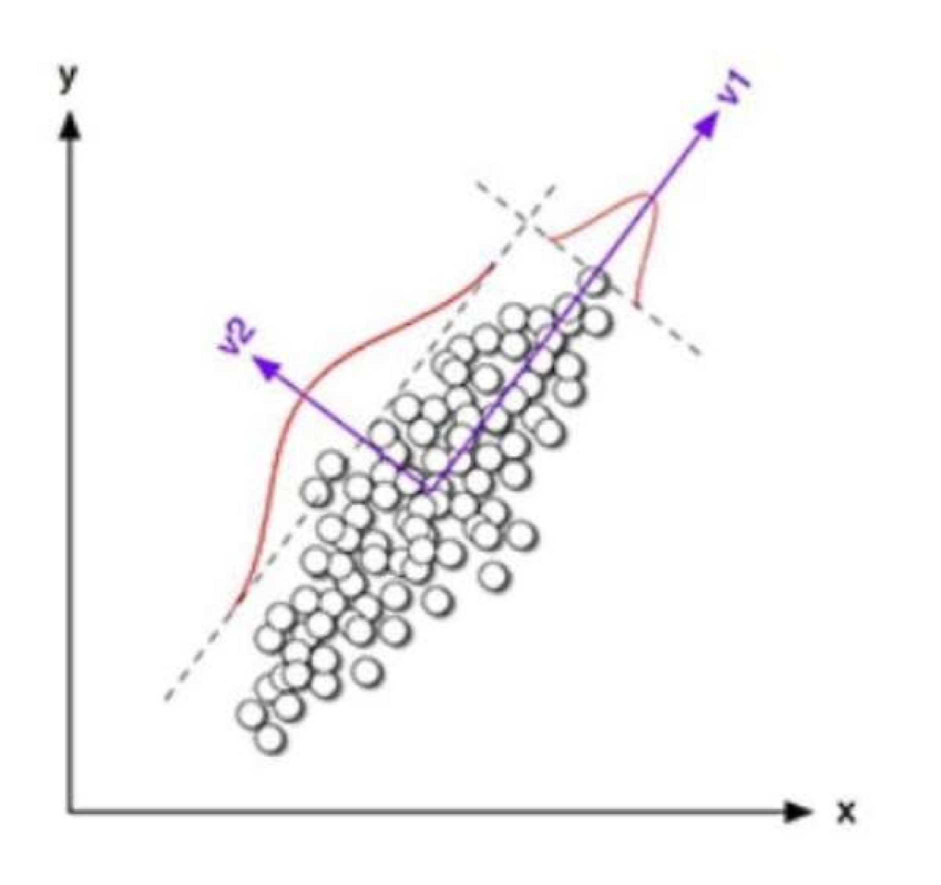




Boston Housing Data Correlation Matrix

	AGE	B	CHAS	CRIM	DIS	INDUS	LSTAT	Median_Home_Value	NOX	PTRATIO	RAD	RM	TAX	ZN
AGE	1	-0.27	0.09	0.35	-0.75	0.64	0.6	-0.38	0.73	0.26	0.46	-0.24	0.51	-0.57
B	-0.27	1	0.05	-0.39	0.29	-0.36	-0.37	0.33	-0.38	-0.18	-0.44	0.13	-0.44	0.18
CHAS	0.09	0.05	1	-0.06	-0.1	0.06	-0.05	0.18	0.09	-0.12	-0.01	0.09	-0.04	-0.04
CRIM	0.35	-0.39	-0.06	1	-0.38	0.41	0.46	-0.39	0.42	0.29	0.63	-0.22	0.58	-0.2
DIS	-0.75	0.29	-0.1	-0.38	1	-0.71	-0.5	0.25	-0.77	-0.23	-0.49	0.21	-0.53	0.66
INDUS	0.64	-0.36	0.06	0.41	-0.71	1	0.6	-0.48	0.76	0.38	0.6	-0.39	0.72	-0.53
LSTAT	0.6	-0.37	-0.05	0.46	-0.5	0.6	1	-0.74	0.59	0.37	0.49	-0.61	0.54	-0.41
Median_Home_Value	-0.38	0.33	0.18	-0.39	0.25	-0.48	-0.74	1	-0.43	-0.51	-0.38	0.7	-0.47	0.36
NOX	0.73	-0.38	0.09	0.42	-0.77	0.76	0.59	-0.43	1	0.19	0.61	-0.3	0.67	-0.52
PTRATIO	0.26	-0.18	-0.12	0.29	-0.23	0.38	0.37	-0.51	0.19	1	0.46	-0.36	0.46	-0.39
RAD	0.46	-0.44	-0.01	0.63	-0.49	0.6	0.49	-0.38	0.61	0.46	1	-0.21	0.91	-0.31
RM	-0.24	0.13	0.09	-0.22	0.21	-0.39	-0.61	0.7	-0.3	-0.36	-0.21	1	-0.29	0.31
TAX	0.51	-0.44	-0.04	0.58	-0.53	0.72	0.54	-0.47	0.67	0.46	0.91	-0.29	1	-0.31
ZN	-0.57	0.18	-0.04	-0.2	0.66	-0.53	-0.41	0.36	-0.52	-0.39	-0.31	0.31	-0.31	1

Steps for PCA



1. Standardize the data
2. Calculate the covariance matrix
3. Find the eigenvalues and eingenvectors of the covariance matrix
4. Plot the eigenvectors/ principal components over the scaled data

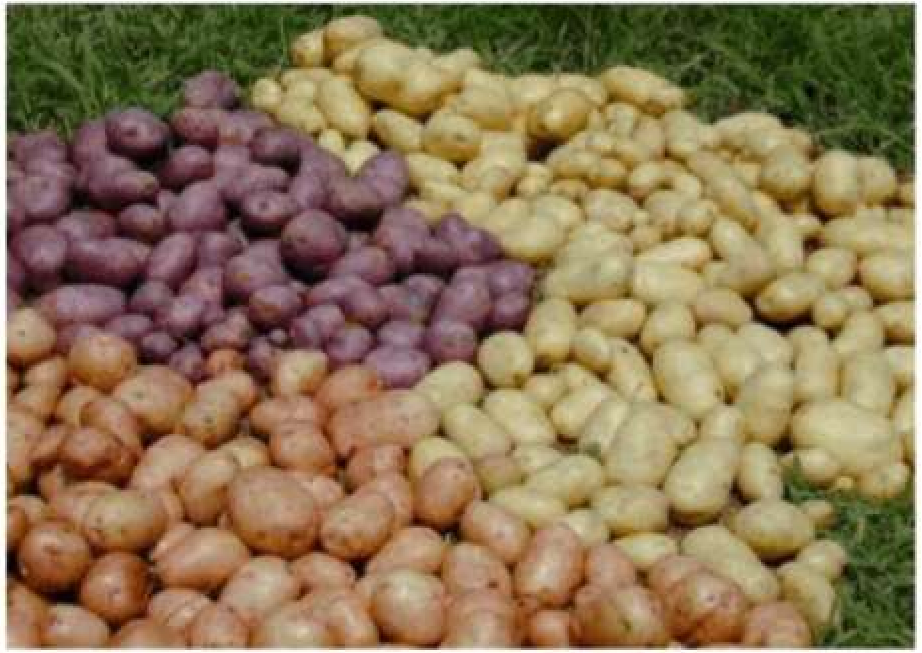
Clustering

Clustering

- **Clustering is a technique for finding similarity groups in data, called clusters. I.e.,**
  - it groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.
- **Clustering is often called an unsupervised learning task as no class values denoting an a priori grouping of the data instances are given, which is the case in supervised learning.**
- **Due to historical reasons, clustering is often considered synonymous with unsupervised learning.**
  - In fact, association rule mining is also unsupervised
- **This chapter focuses on clustering.**

An Illustration

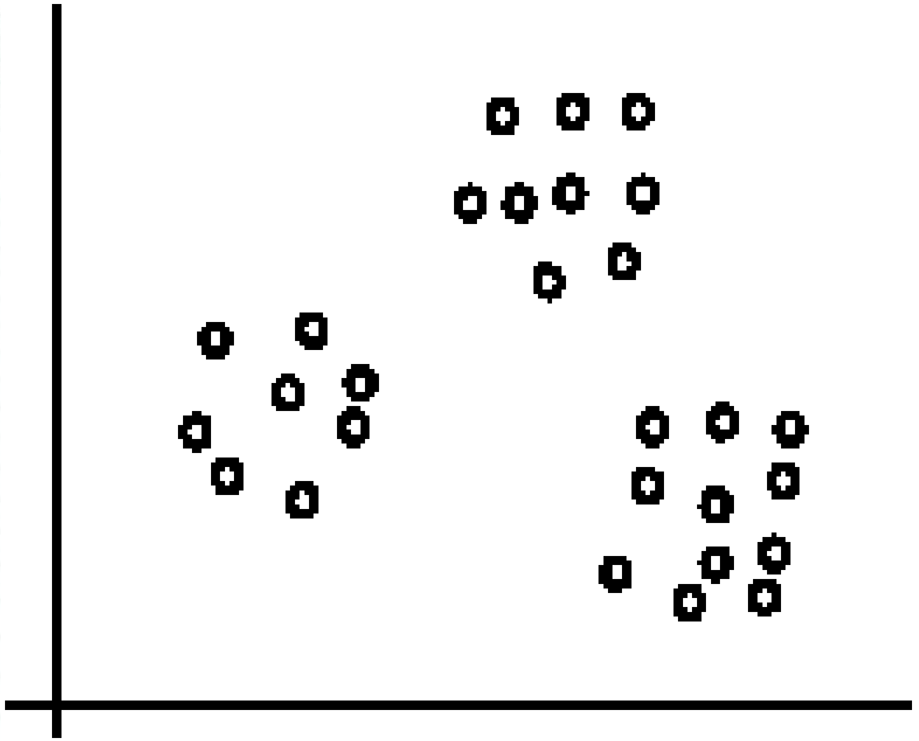
sample



Cluster/group

An Illustration

The data set has three natural groups of data points, i.e., 3 natural clusters.



Aspects of Clustering

A clustering algorithm

- Partitional clustering
- Hierarchical clustering
- …

A distance (similarity, or dissimilarity) function

Clustering quality

- Inter-clusters distance ⇒ maximized
- Intra-clusters distance ⇒ minimized

The quality of a clustering result depends on the algorithm, the distance function, and the application.

Types of Clustering

A clustering is a set of clusters

Important distinction between hierarchical and partitional sets of clusters

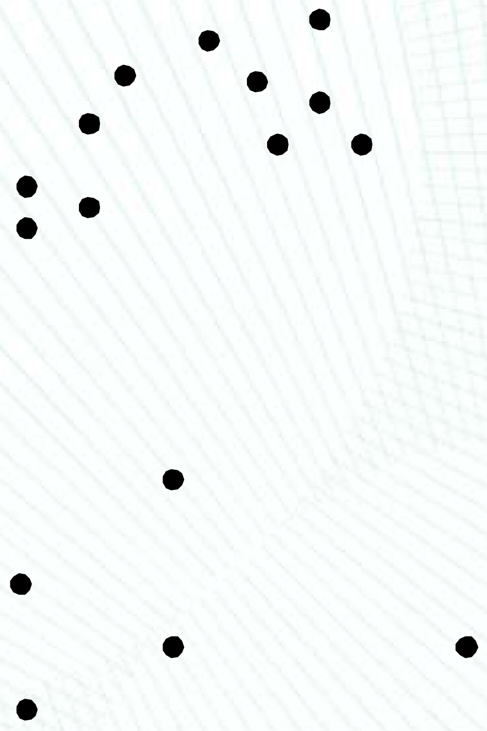
Partitional Clustering

- A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

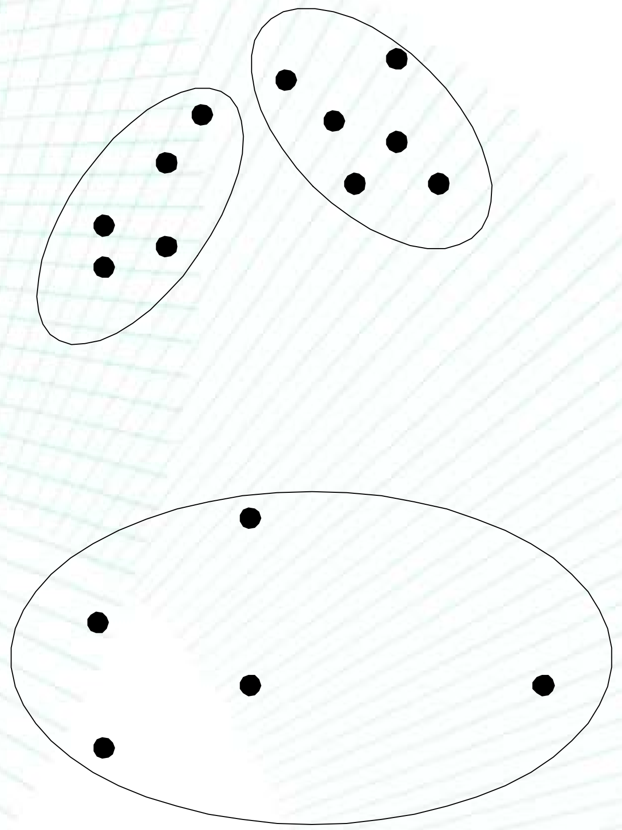
Hierarchical clustering

- A set of nested clusters organized as a hierarchical tree

Partitional Clustering



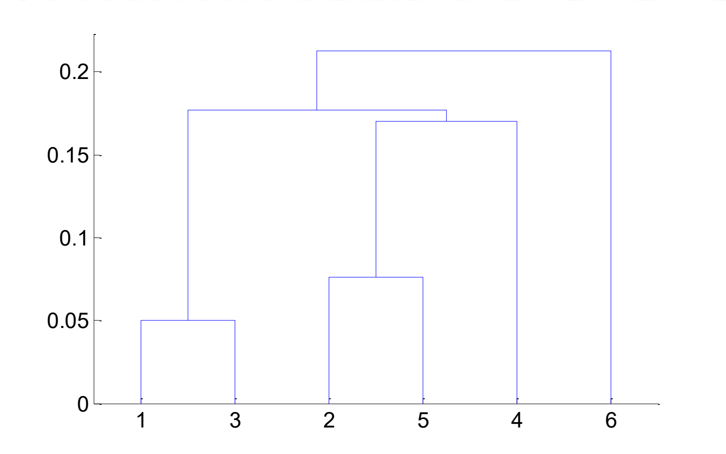
Original Points

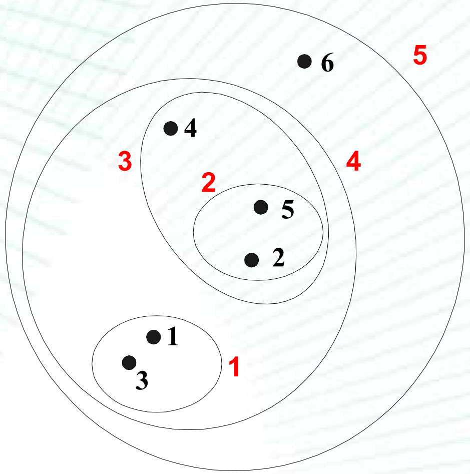


A Partitional Clustering

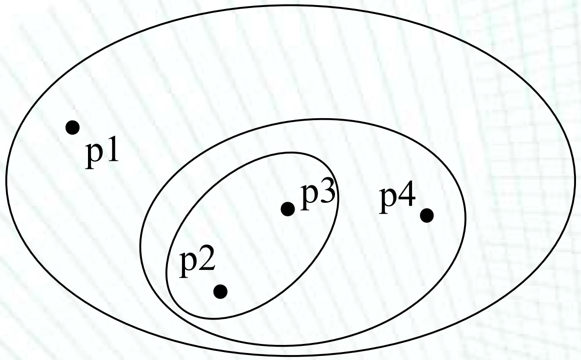
Hierarchical Clustering

- **Produces a set of nested clusters organized as a hierarchical tree**
- **Can be visualized as a dendrogram**
  - A tree like diagram that records the sequences of merges or splits

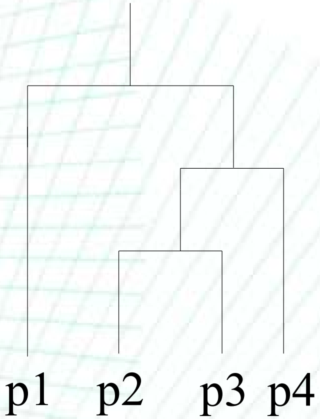




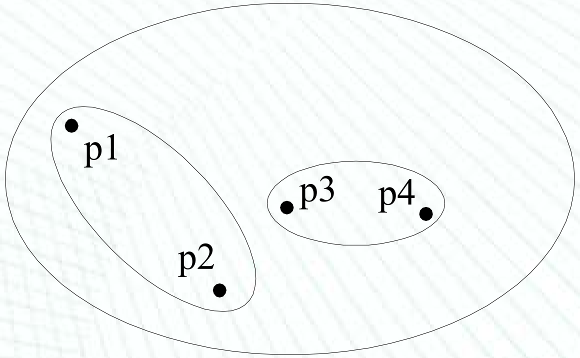
Hierarchical Clustering



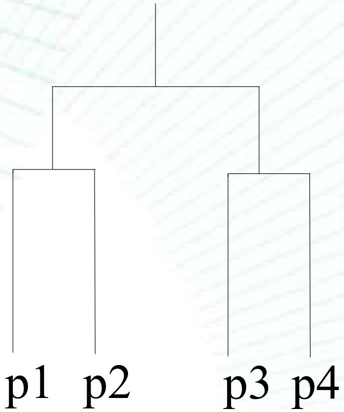
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Hierarchical Clustering

- **Two main types of hierarchical clustering**
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- **Traditional hierarchical algorithms use a similarity or distance matrix**
  - Merge or split one cluster at a time

Clustering Algorithms

- **K-means and its variants**
- **Hierarchical clustering**
- **Density-based clustering**

K-Means Clustering

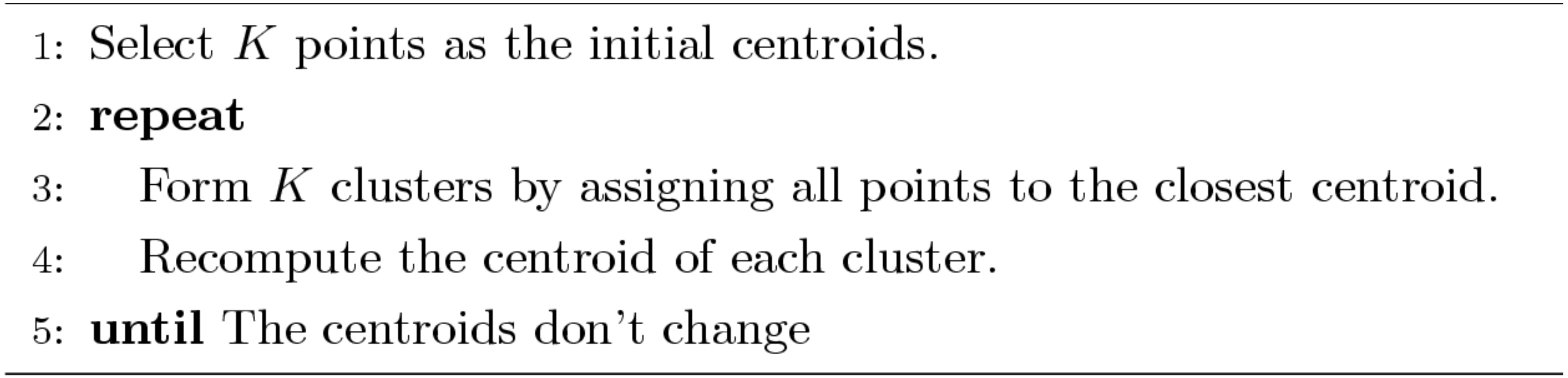
- **K-means is a partitional clustering algorithm**
- **Let the set of data points (or instances) D be**

  {x1, x2, …, xn},

  where xi = (xi1, xi2, …, xir) is a vector in a real-valued space X ⊆ Rr, and r is the number of attributes (dimensions) in the data.
- **The k-means algorithm partitions the given data into k clusters.**
  - Each cluster has a cluster **center**, called centroid.
  - k is specified by the user

K-Means Clustering

Basic algorithm



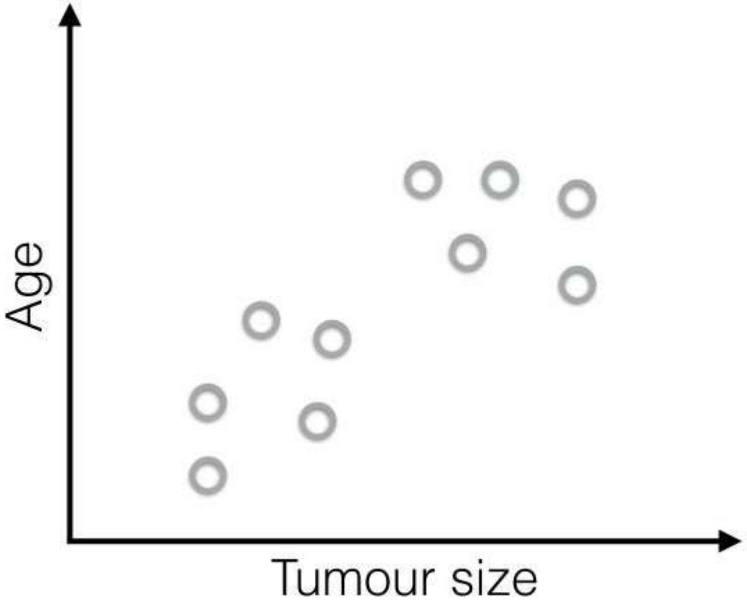
```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

K-Means Clustering

*although they do have something in common with k-nearest neighbour, but **they are not the same**

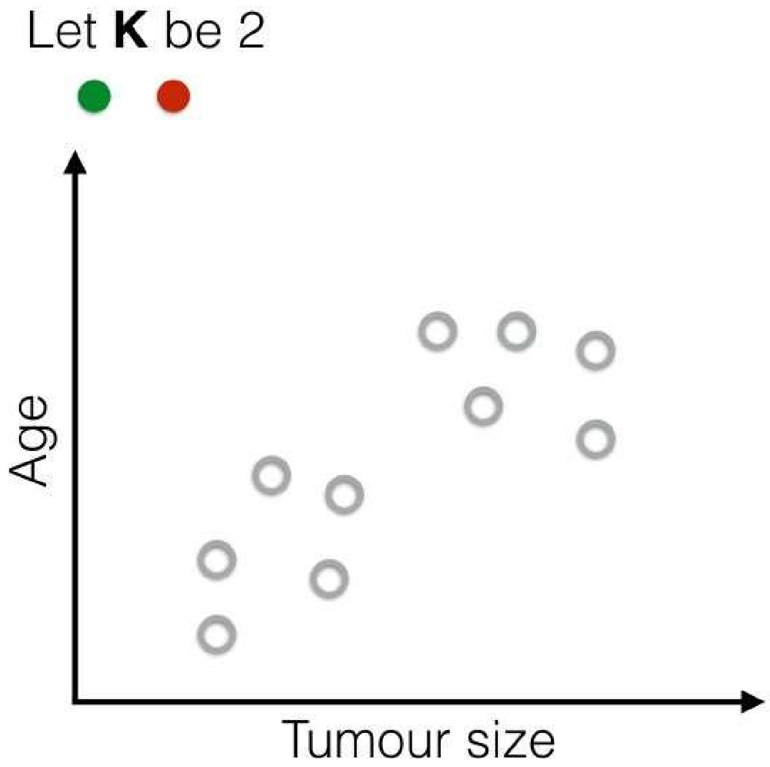
Steps of K-Means Clustering

1.Choose **K**, the number of potential clusters



Steps of K-Means Clustering

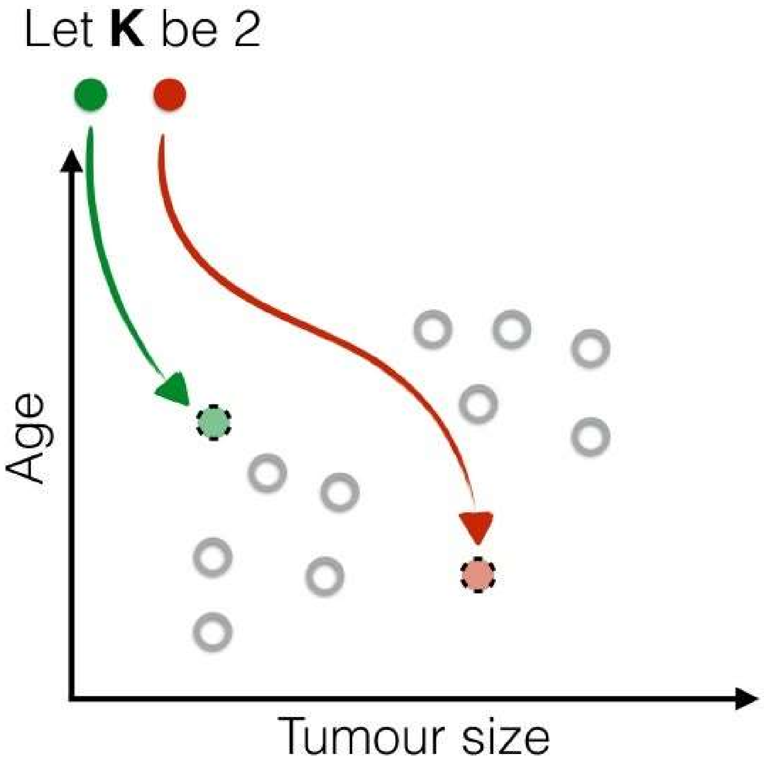
1.Choose **K**, the number of potential clusters



Steps of K-Means Clustering

1.Choose **K**, the number of potential clusters

2.Initialise cluster centers randomly within the data

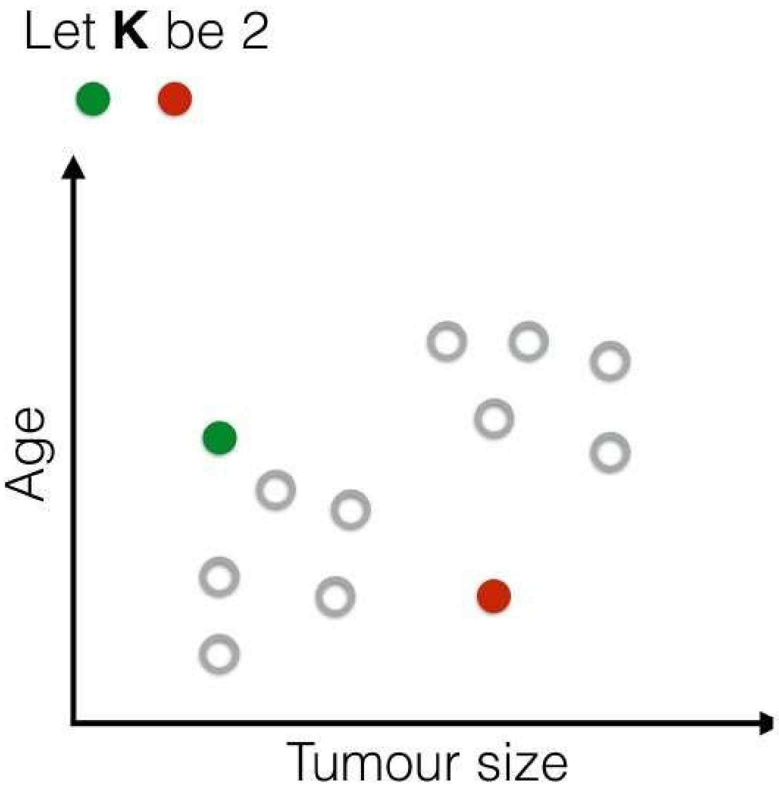


Steps of K-Means Clustering

1. Choose **K**, the number of potential clusters

2. Initialise cluster centers randomly within the data

3. Instances are clustered to the nearest cluster centre

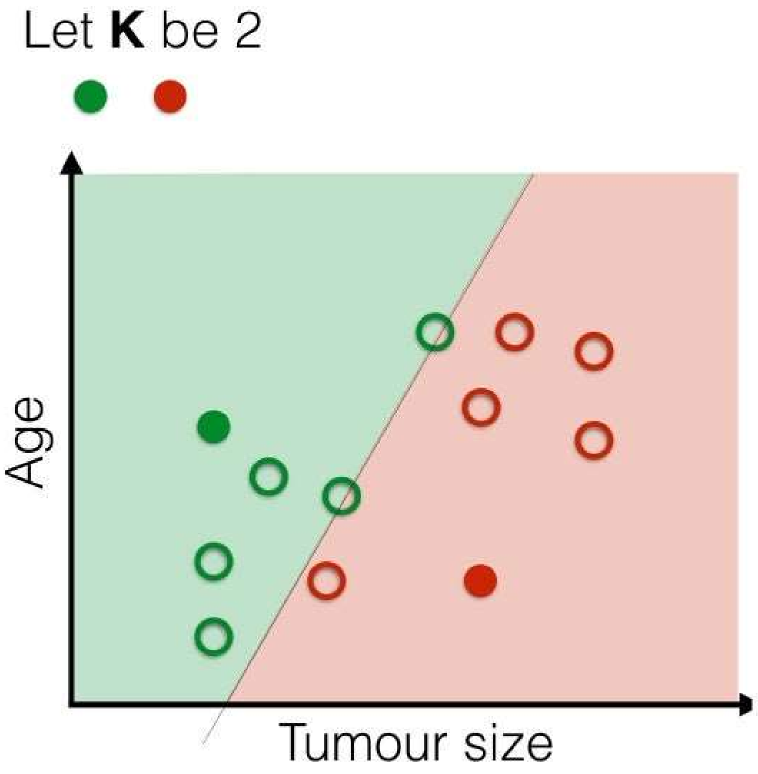


Steps of K-Means Clustering

1.Choose **K**, the number of potential clusters

2.Initialise cluster centers randomly within the data

3.Instances are clustered to the nearest cluster centre



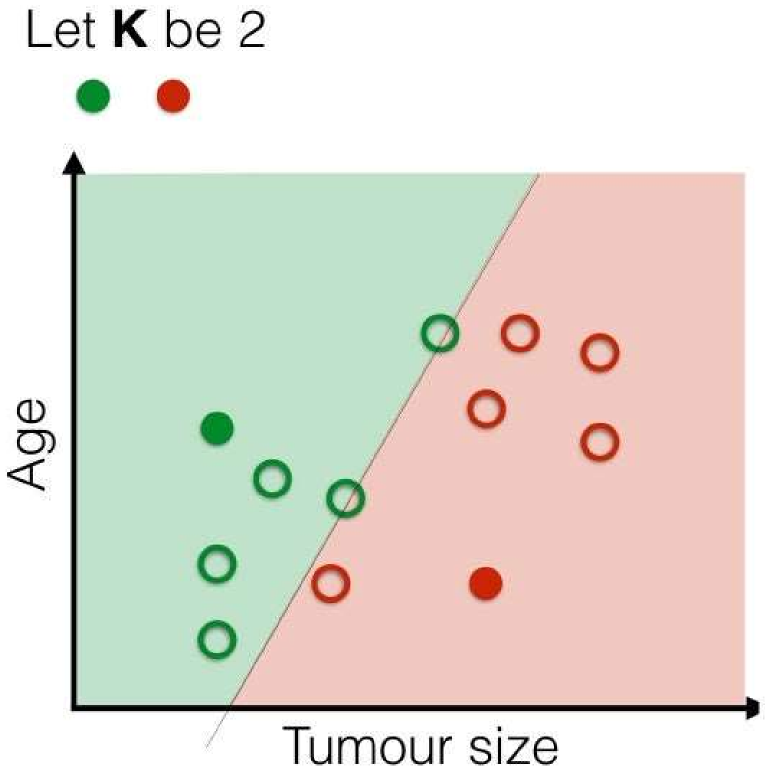
Steps of K-Means Clustering

1.Choose **K**, the number of potential clusters

2.Initialise cluster centers randomly within the data

3.Instances are clustered to the nearest cluster centre

4.Centroids of each of the **K** clusters become new cluster centers



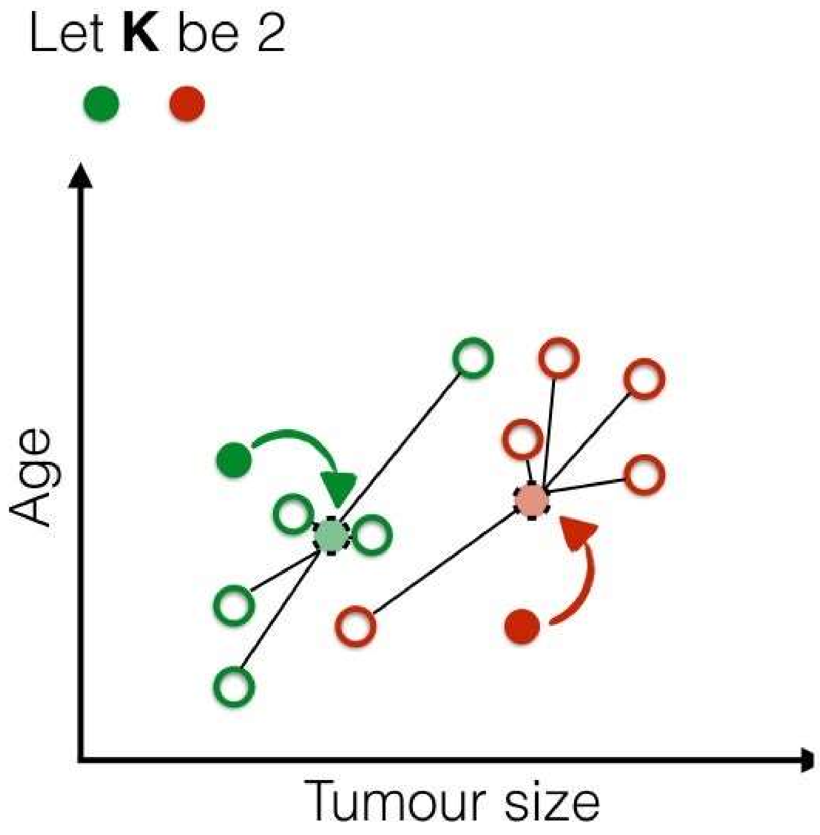
Steps of K-Means Clustering

1.Choose **K**, the number of potential clusters

2.Initialise cluster centers randomly within the data

3.Instances are clustered to the nearest cluster centre

4.Centroids of each of the **K** clusters become new cluster centers



Steps of K-Means Clustering

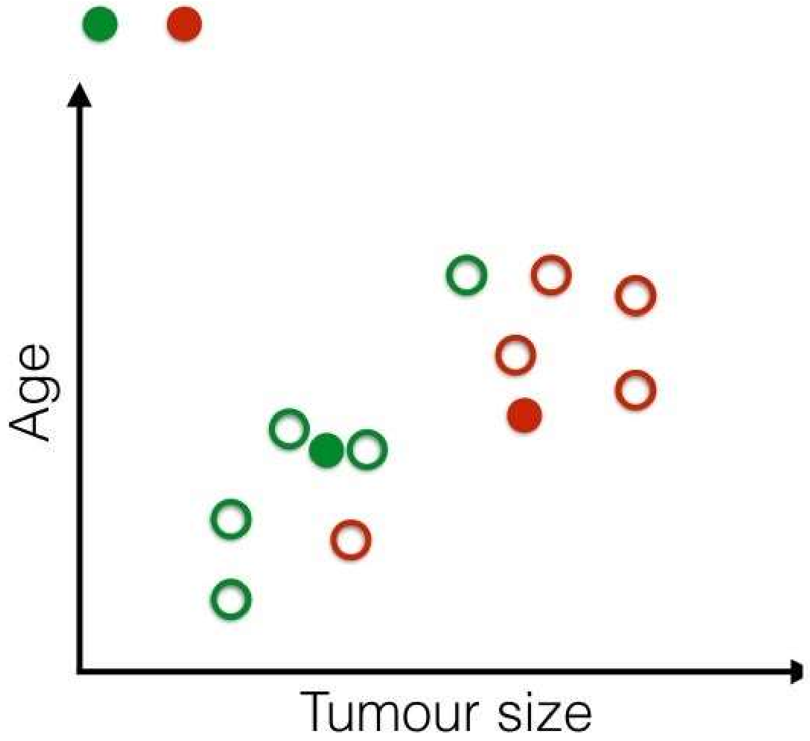
1.Choose **K**, the number of potential clusters

2.Initialise cluster centers randomly within the data

3.Instances are clustered to the nearest cluster centre

4.Centroids of each of the **K** clusters become new cluster centers

Let **K** be 2



Steps of K-Means Clustering

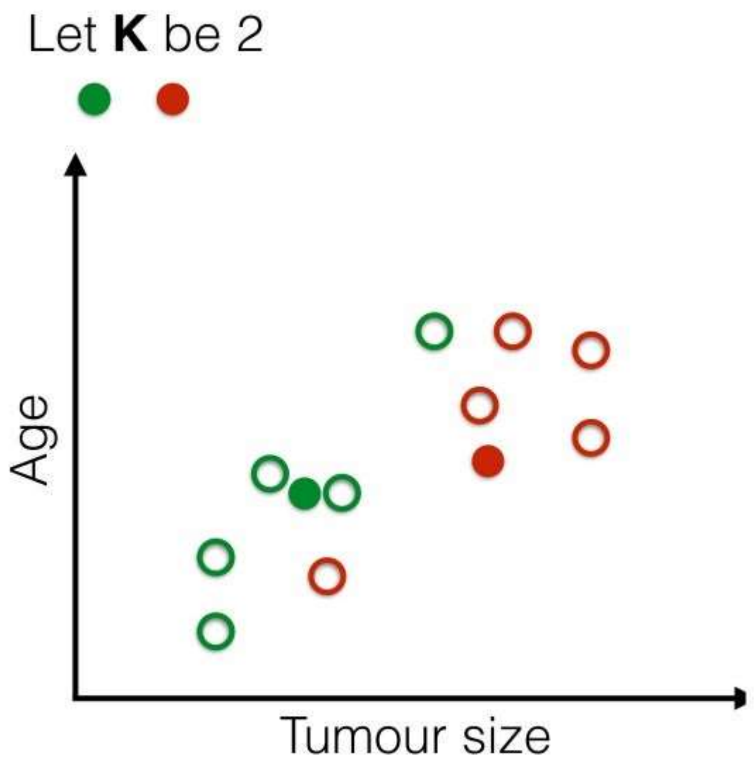
1.Choose **K**, the number of potential clusters

2.Initialise cluster centers randomly within the data

3.Instances are clustered to the nearest cluster centre

4.Centroids of each of the **K** clusters become new cluster centers

5.Steps **3/4** are repeated until convergence



Steps of K-Means Clustering

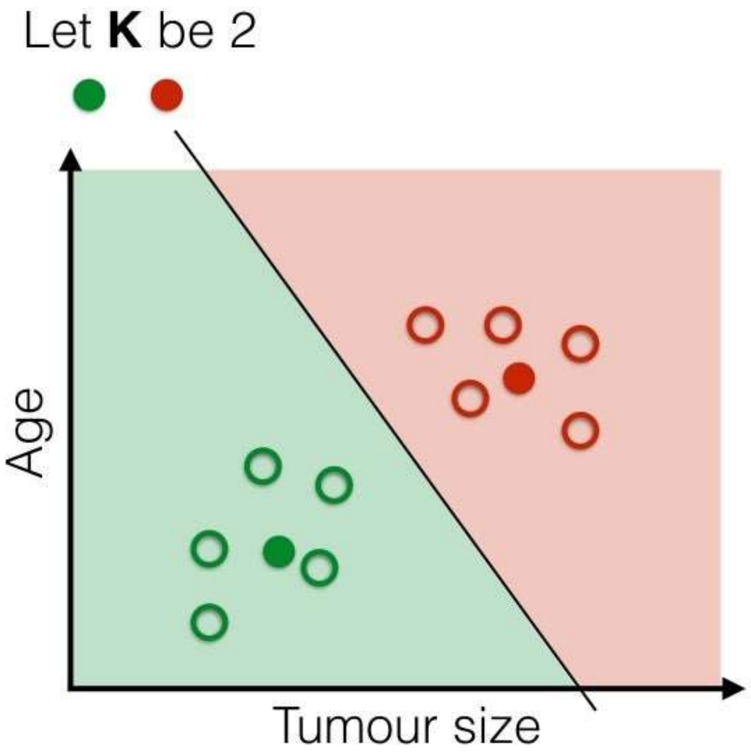
1.Choose **K**, the number of potential clusters

2.Initialise cluster centers randomly within the data

3.Instances are clustered to the nearest cluster centre

4.Centroids of each of the **K** clusters become new cluster centers

5.Steps **3/4** are repeated until convergence



Limitations of K-Means

- **K-means has problems when clusters are of differing**
  - Sizes
  - Densities
  - Non-globular shapes

- **K-means has problems when the data contains outliers.**

Application of Unsupervised algorithm in Public Sector

Mapping of Flood Prone Area in Jakarta using Fuzzy C-Means

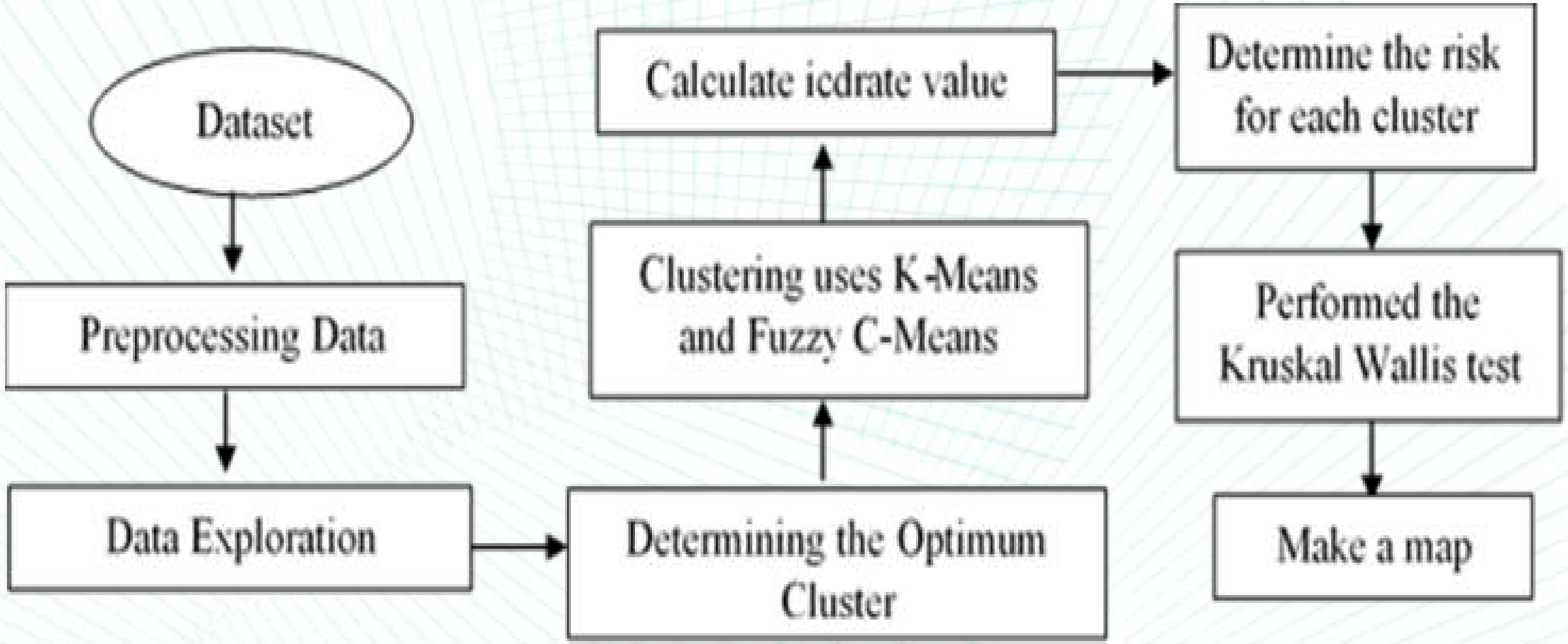
Background

The Jakarta Provincial Government needs **evidence-based policies to deal with potential floods to protect residents from the threat of flood disasters**. Mapping flood-prone areas in Jakarta can be a reference to minimize the significant loss and harm due to flooding.

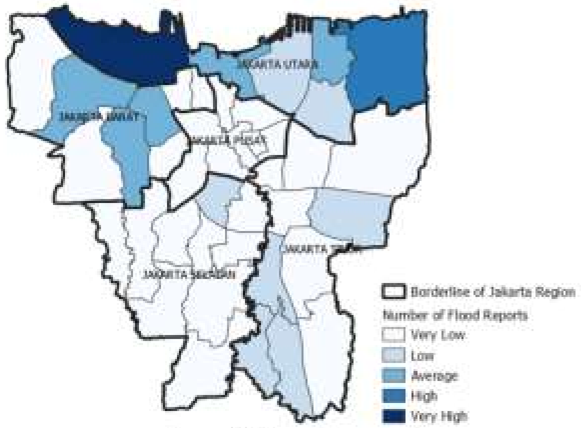
Dataset

Sub-District	Groundwater Use (mm^3)	Number of Flood Report	Land Subsidence (m)	Water Level (m)
Cakung	16228.0	3	1.7	15
Cempaka Putih	1444.6	0	1.9	0
.....	.....	.....	.....	.....
Tanah Abang	326.9	4	1	19.25
Tanjung Priok	402.3	5	2.2	18
Tebet	17798.4	2	1.2	30

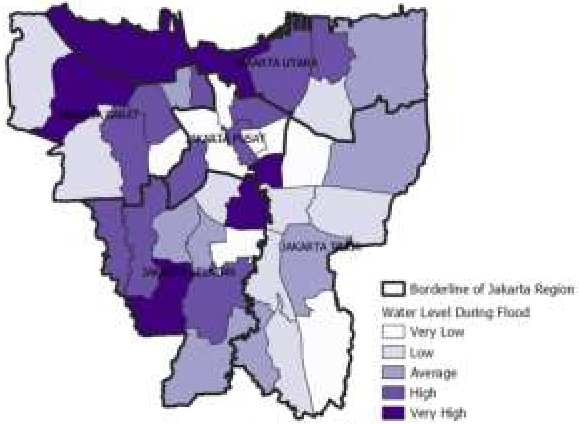
Methodology


Dataset
Preprocessing Data
Data Exploration
Determining the Optimum Cluster
Clustering uses K-Means and Fuzzy C-Means
Calculate icdrate value
Determine the risk for each cluster
Performed the Kruskal Wallis test
Make a map

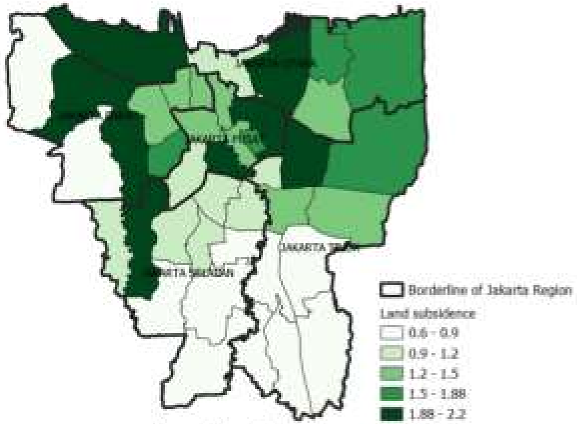
Data Exploration



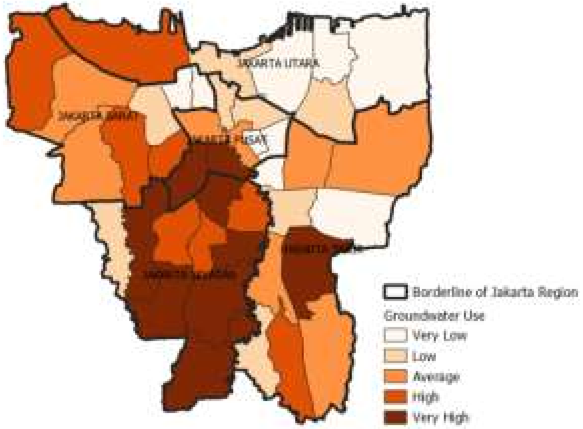
number of flood reports



water level



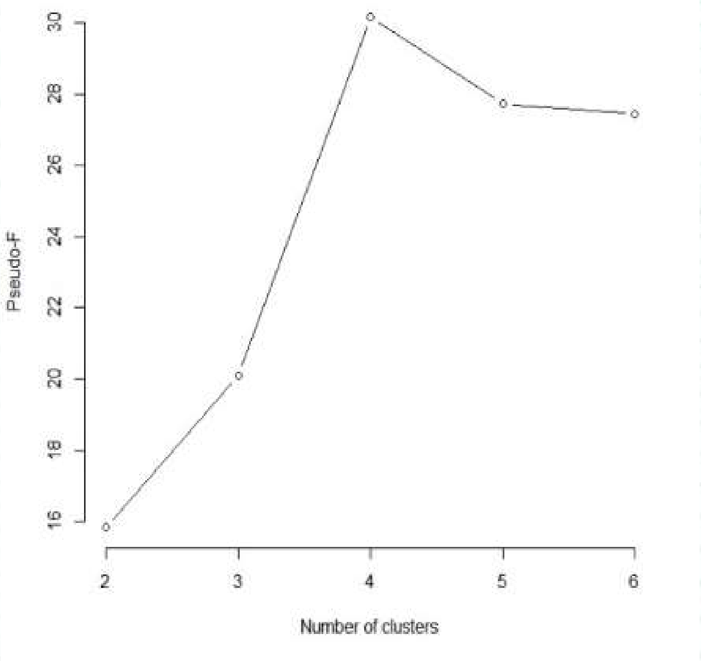
land subsidence



groundwater use

Clustering Result

- Optimum Cluster using Pseudo-F



- Selecting Clustering Model with Icdrate

The evaluation results using the Icdrate value show that the Fuzzy C-Means has a smaller Icdrate value than the K-Means, namely 0.204673. Therefore, the clustering method using Fuzzy C-Means gives better results than the K-Means. In this study, the Fuzzy C-Means is the best method to classify flood-prone areas in Jakarta

Clustering Result

- Results of grouping using the Fuzzy C-Means method

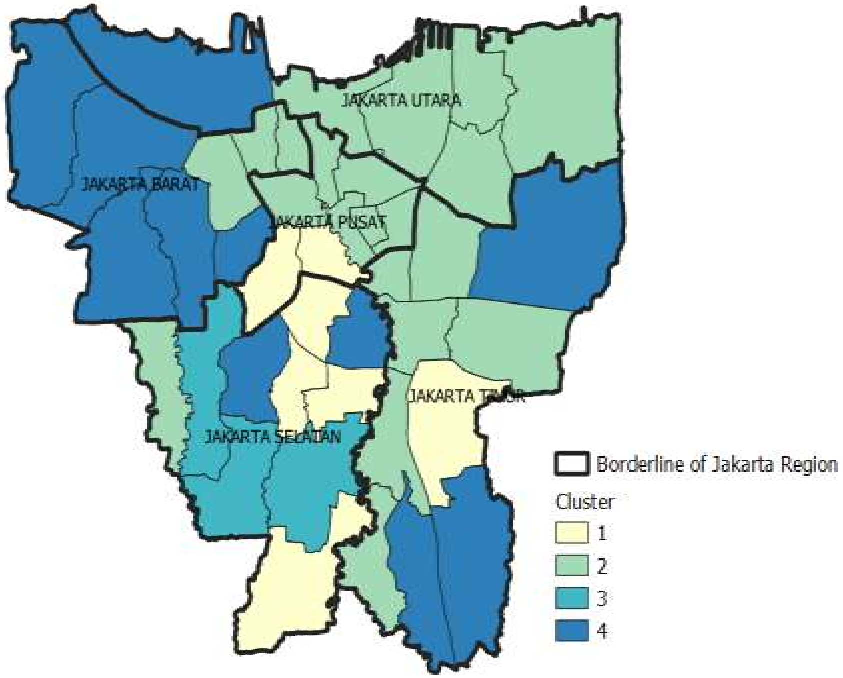
Cluster	Numbers of Cluster	Members of Cluster
1	7	Jagakarsa, Makasar, Mampang Prapatan, Menteng, Pancoran, Setiabudi, Tanah Abang
2	21	Cempaka Putih, Cilincing, Duren sawit, Gambir, Grogol Petamburan, Jatinegara, Johar Baru,
		Kelapa gading, Kemayoran, Koja, Kramat Jati, Matraman, Pademangan, Pasar Rebo,
		Pesanggrahan, Pulo Gadung, Sawah Besar, Senen, Taman Sari, Tambora, Tanjung Priok
3	3	Cilandak, Keb Lama, Pasar Minggu
4	11	Cakung, Cengkareng, Cipayung, Ciracas, Kalideres, Keb Baru, Kebon Jeruk, Kembangan, Palmerah, Penjaringan, Tebet

- The variable average value, total score, and risk level in each cluster

Cluster	Water Level (m)	Land Subsidence (m)	Ground-water Use (mm^3)	Number of Flood Report	Total Score	Risk
1	9.90	1.10	31755.69	2.86	7	Low
2	12.82	1.46	3876.39	4.38	10	Low Enough
3	32.25	1.20	79915.75	2.67	11	High Enough
4	14.59	1.33	17363.51	5.45	12	High

Clustering Result

- Cluster map based on risk level: yellow = low, dark blue = high, light blue = high enough, and green = low enough



Discussion

- Data exploration shows that the condition of the river in Jakarta is getting shallower. Therefore, a river normalization program needs to be carried out to optimize the capacity of the river.
- The government needs to pay attention to water pumps to function correctly so that standing water due to flooding can be immediately resolved
- The Pseudo-F value indicates four is the optimum number of clusters
- The Fuzzy C-Means method results in a lower Icdrate value than the K-Means, suggesting that the Fuzzy C-Means are better than the K-Means. The analysis results showed different results from previous studies [8], [9]. With the Fuzzy C-Means method, 42 sub-districts in Jakarta are grouped into four clusters.

Discussion

- Districts with a high and relatively high risk of flooding are primarily located in West Jakarta and South Jakarta. The government prevention efforts are by paying attention to the characteristics of each cluster
- To prevent sub-districts in clusters 1, 2, and 3 from becoming high flood risk areas, the government needs to limit groundwater use by inspecting illegal ground wells and looking for alternative water sources.
- The government can reduce flooding in cluster 4 by building higher embankments, replacing temporary embankments with permanent embankments, and maintaining embankments in North Jakarta

Conclusion

- The Fuzzy C-Means method is suitable for mapping flood-prone areas in Jakarta.
- It is possible if a sub-district becomes a flood-prone area if its location is close to a high flood risk sub-districts without any preventive and remedial efforts from the local government.
- Further works can be carried out using other clustering methods for comparison
- In addition, future works need to add more variables related to flooding, such as the number of affected hamlets, inundation time, and the number of residents.

Hands On

Let’s open Google Colab!

Thank you!