HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

IT4142E
Introduction to Data Science

Chapter 6: Introduction to Machine Learning

Lecturer:

Muriel VISANI: murielv@soict.hust.edu.vn

Acknowledgements:

Khoat Than
Viet-Trung Tran

Department of Information Systems
School of Information and Communication Technology - HUST

Contents of the course

- Chapter 1: Overview
- Chapter 2: Data scraping
- Chapter 3: Data cleaning, pre-processing and integration
- Chapter 4: Introduction to Exploratory Data Analysis
- Chapter 5: Introduction to Data visualization
- Chapter 6: Introduction to Machine Learning
  - Performance evaluation
- Chapter 7: Introduction to Big Data Analysis
- Chapter 8: Applications to Image and Video Analysis

Goals of this chapter

Goal	Description of the goal
M1	**Understand and be able to design and manage the systems which are based on Data Science (DS)**
M1.2	Identify, compare, and categorize the data type and systems in practice
M1.3	Be able to design systems based on DS in their future organizations

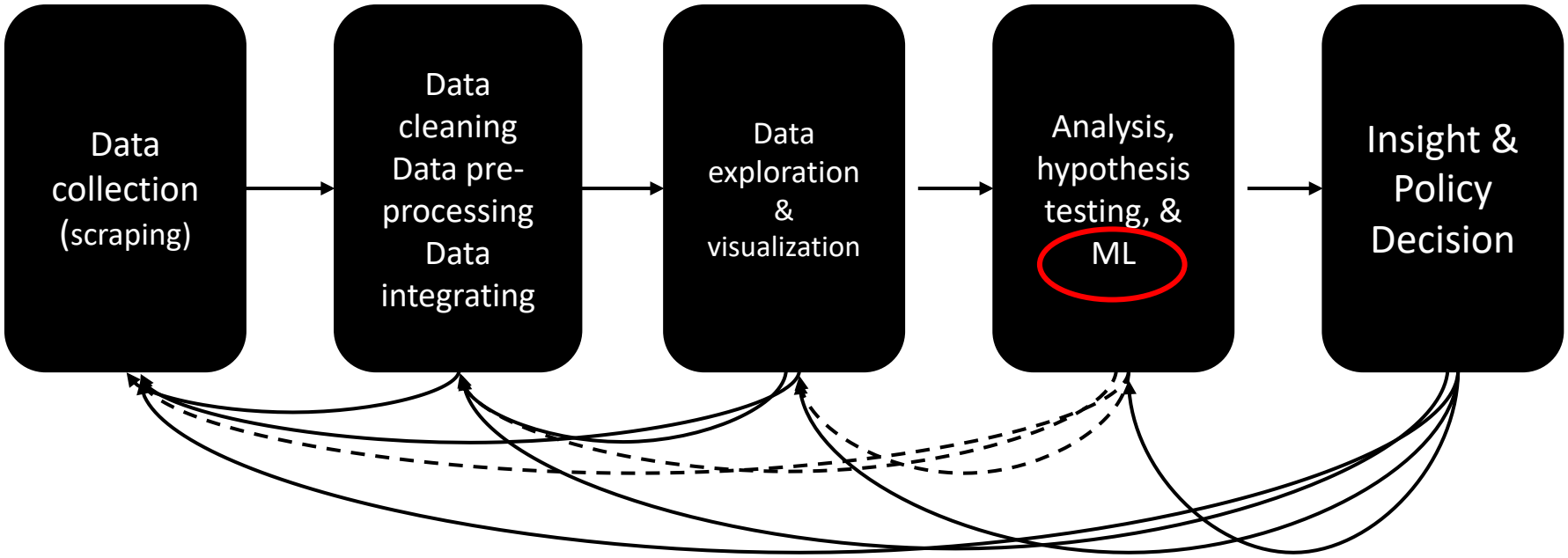
Contents of this chapter

- **Chapter 6: Introduction to Machine Learning**
  - Introduction and definitions
    - What is Machine Learning?
    - Supervised *vs.* unsupervised learning
    - Focus of this chapter
  - Part 1: Unsupervised **clustering**
    - Objective
    - Main issues and useful definitions
    - Methods
      - Partitioning methods
      - Hierarchical methods
      - Grid-based methods
      - Density-based methods
    - Performance evaluation
  - Part 2: Supervised **classification**
    - Objective
    - Methods
    - Performance evaluation
  - Summary
  - Homework

Introduction and definitions

What is Machine Learning?

Recall: DS methodology



What is Machine Learning?

- Machine learning is a sub-domain of Artificial Intelligence
- Sometimes related to cognitive sciences (*e.g.* neural networks)
- Based on **inference**
- Solves a wide variety of problems
  - Used in almost any application domains
    - Smart cars, diagnostic assistance for doctors, spam filters, targeted advertising, etc.

Different types of inference

- **Deductive** (Logical) inference:
  - From A and A $\rightarrow$ B, infer B
  - Deducing the consequences from the causes (premises)
  - Logical rules (see course on AI)

- **Abductive** inference :
  - From B andA $\rightarrow$ B, infer A
  - Making hypothesis about the causes, from the consequences
  - Application to diagnostic systems

Different types of inference

- **Inductive** inference:
  - This is the type of inference mostly used in Machine Learning
  - Knowledge is extracted on the basis of specific examples

Valid only in probabilistic terms

Watch out for hasty generalizations!

Learning for inductive inference

- Learning serves for **inductive** inference
- The aim is to extract a model from data samples (observations)
  - These observations are contained in a learning dataset

- A **model** is used to extract knowledge
  - The model is learned from the training dataset
  - The model can be improved with experience
    - *i.e.* with more observation in the learning dataset

Machine learning (ML) process

ML is a **two-step process**:

Goal: extracting a model to generate knowledge

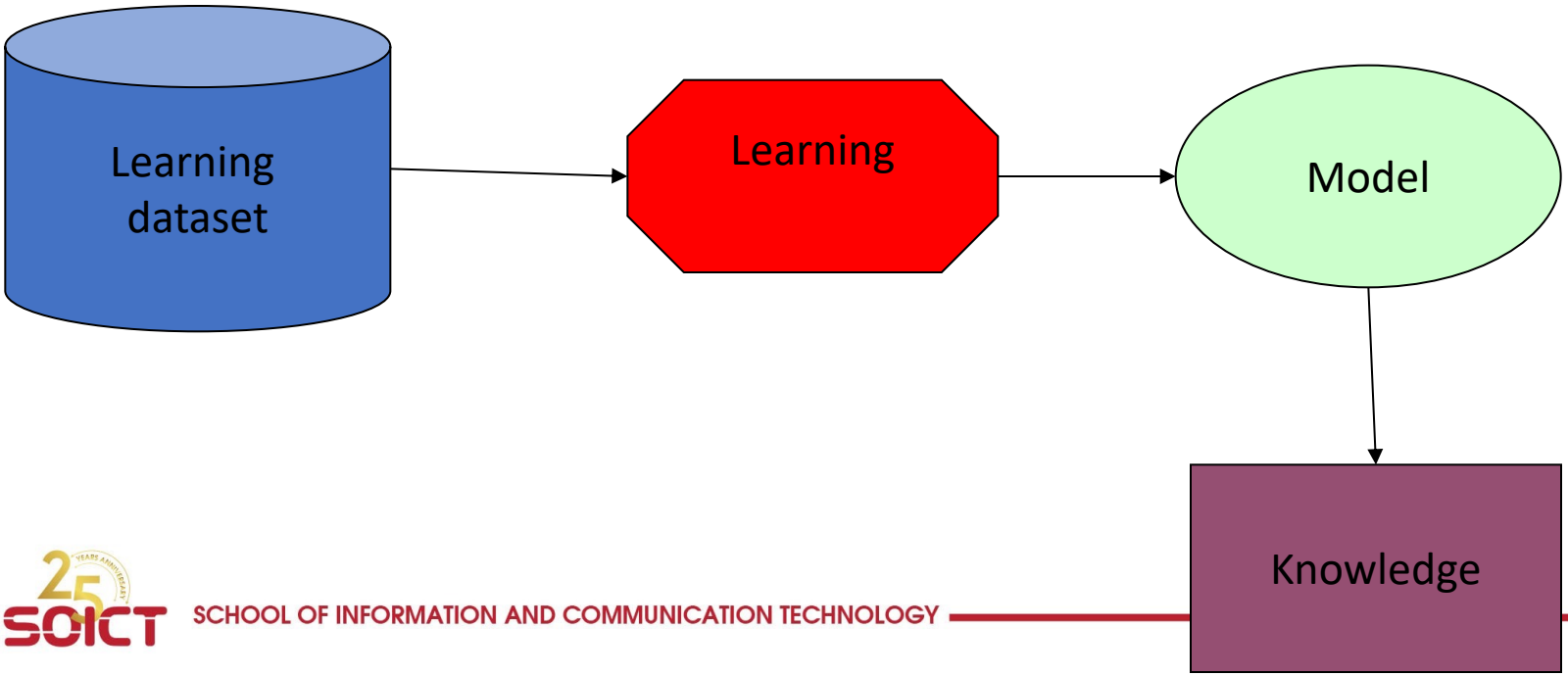
Machine learning (ML) process

Step 1: Learning the model

ML is a **two-step process**:

Goal: extracting a model to generate knowledge

- on the learning dataset (model fitting)



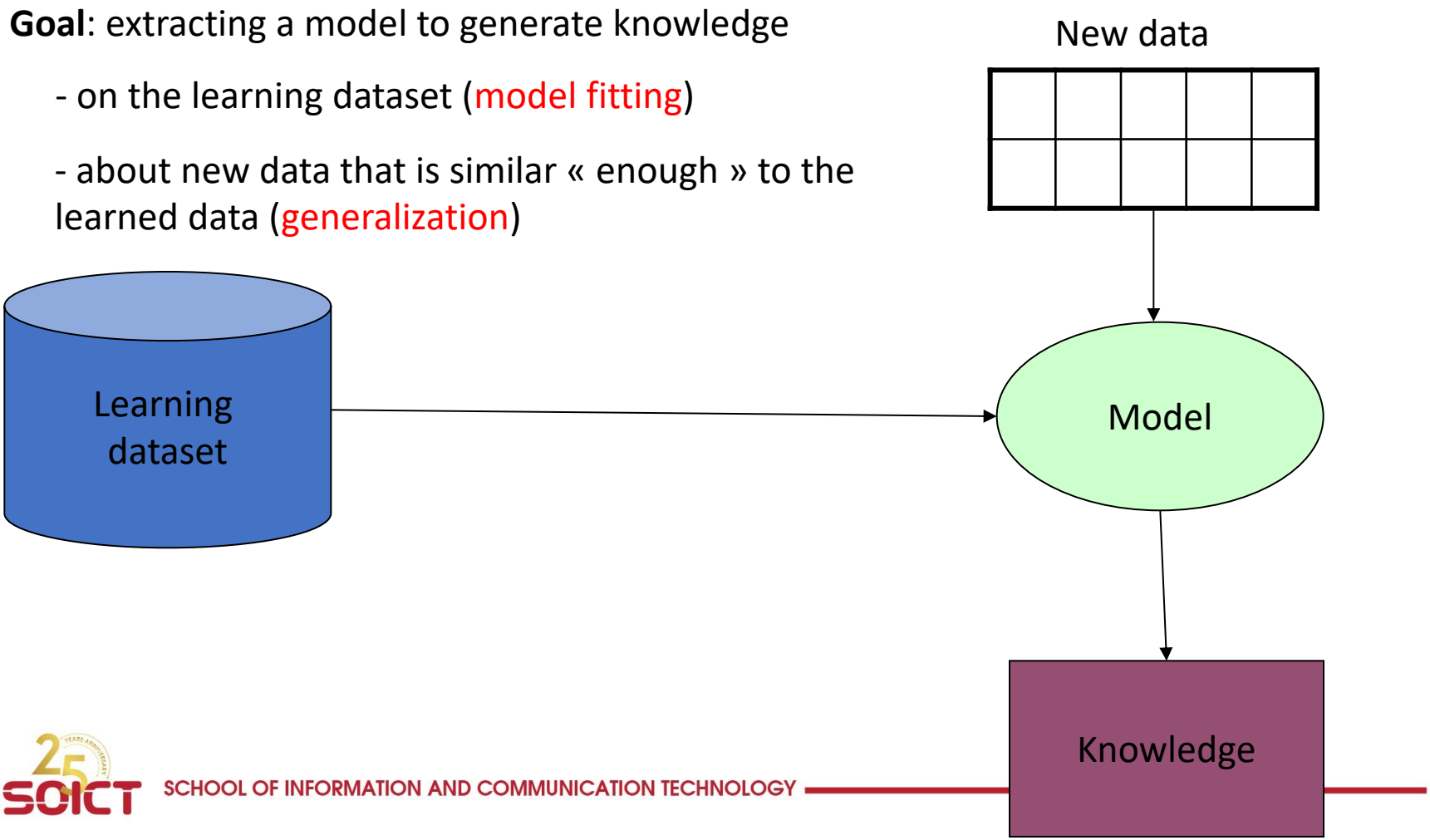
Machine learning (ML) process

ML is a **two-step process**:

Goal: extracting a model to generate knowledge

- on the learning dataset (model fitting)

- about new data that is similar « enough » to the learned data (generalization)



Goals of ML

- As stated in Chapter 1, the main objectives of Machine Learning are:
  - Description
  - Segmentation
  - Association
  - Prediction
- Most of the time, the goal is to…
  - Infer the value of a new variable (response variable)…
  - …based on the observed values of explanatory variables…
  - …for the learning dataset and/or new data

- **NB**: sometimes, illustrative variables can be used for the analysis of the results: variables used for understanding the results but not for learning
  - Example: I made clusters of customers based **only** on what they buy
    - Later on, I got information about their salaries
    - I try to study the link between the clusters (response variable) and salary (illsutrative variable)

Introduction and definitions

Supervised *vs.* unsupervised learning

Supervised *vs.* unsupervised: different goals

- Depending on the final objective, we might have to use a method based on supervised, unsupervised (or semi-supervised learning)
  - Supervised learning serves for **prediction/estimation**
  - Unsupervised learning might be used for **description, segmentation or association**
  - Semi-supervised learning can be used for **segmentation** or **prediction/estimation**
    - Out of the scope of this course

Supervised learning

- **Supervised** learning methods
  - The learning dataset contains the values of the response variable(s)

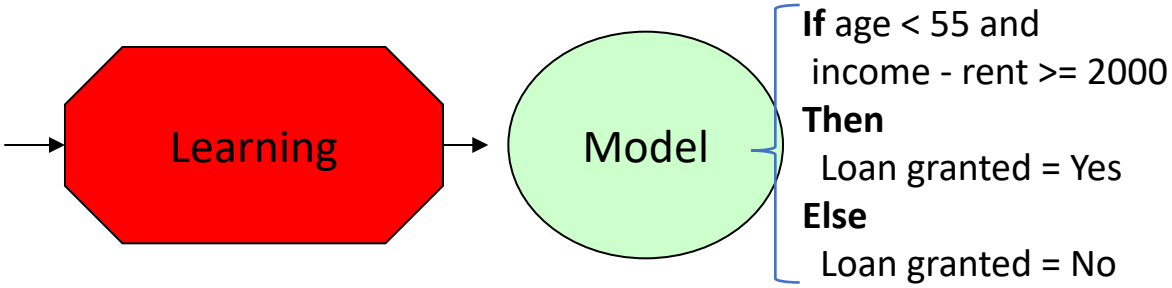
Supervised learning

- **Supervised** learning methods
  - The learning dataset contains the values of the response variable(s)
  - Example

Step 1: Learning the model

Learning dataset

Age	Rent	Income	Loan granted?
36	0	1299	Yes
55	240	2500	No
40	768	3000	Yes
39	0	2000	Yes
44	334	512	No
26	631	722	No



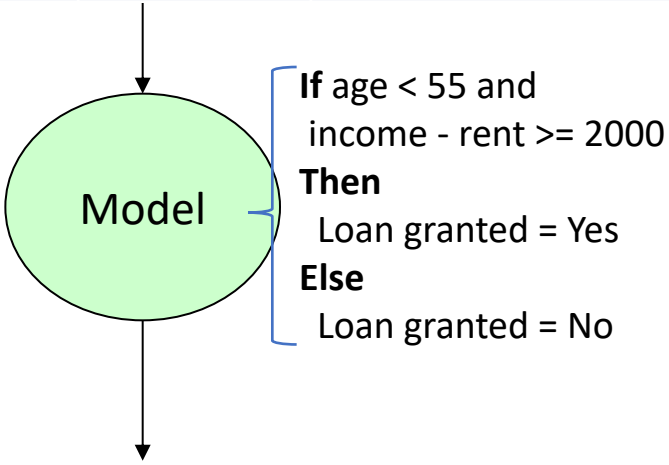
Supervised learning

- **Supervised** learning methods
  - The learning dataset contains the values of the response variable(s)
  - Example

New example

Age	Rent	Income	Loan granted?
30	260	1900	**???**

Step 2: Applying the model



The model cannot always be expressed using rules
- it depends on the model
- *e.g.* decision trees give rules but neural networks don't

New example

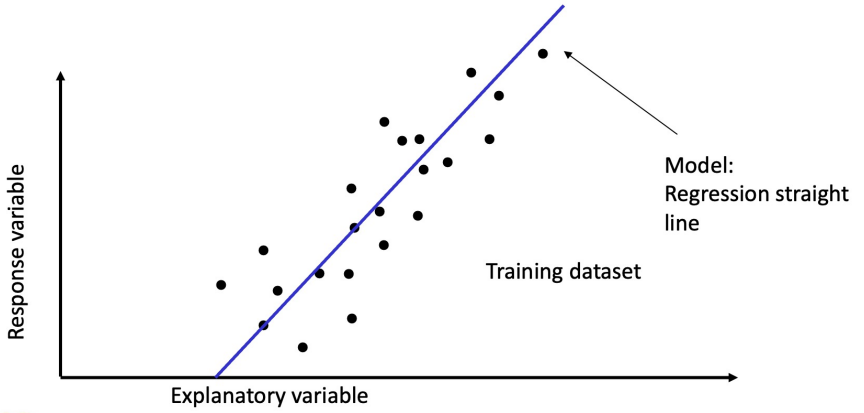
Age	Rent	Income	Loan granted?
30	260	1900	**No**

Supervised learning

- **Supervised learning** often serves for prediction
  - Main **tasks**: regression and classification
- **Regression** (linear regression, regression trees, neural networks...)
  - The response variable is **numeric**
  - The explanatory variables are the attributes observed
- **Classification** (Bayesian classifiers, classif trees, neural networks...)
  - Consists in assigning new data into pre-defined classes
  - The response variable is **categorical:** the class
  - The explanatory variables are the observed attributes

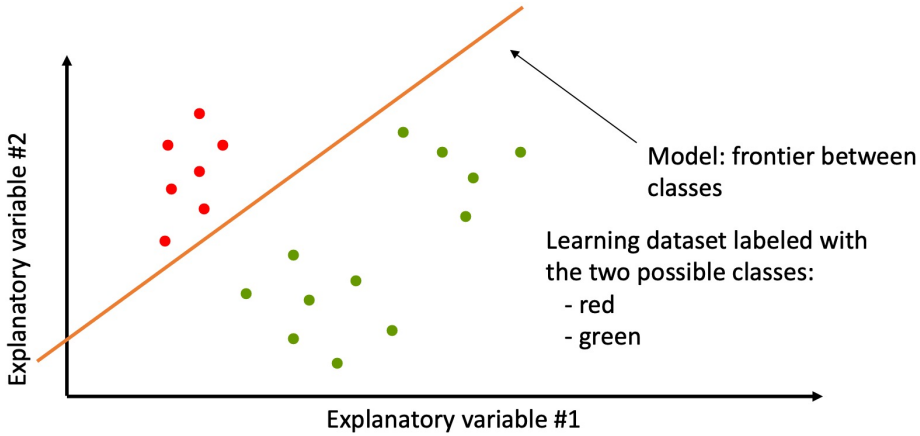
Questions

- From the data in your capstone project, give an example of possible:
  - Regression task:
    - Training dataset
    - Response variable
    - Explanatory variable(s)



Example: linear regression

  - Classification task:
    - Training dataset
    - Response variable
    - Explanatory variable(s)



Example: classification using a frontier-based model

Unsupervised learning

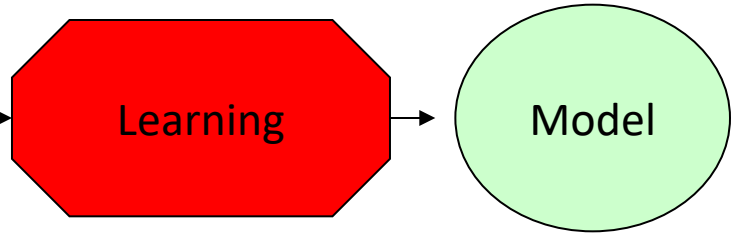
- **Unsupervised** learning methods
  - The learning dataset **does not contain** the values of the response variable(s)

Unsupervised learning

- **Unsupervised** learning methods
  - The learning dataset **does not contain** the values of the response variable(s)
  - Example

Step 1: Learning the model

Learning dataset		
Age	**Rent**	**Income**
36	0	1299
55	240	2500
40	768	3000
39	0	2000
44	334	512
26	631	722



	Customer Segments		
	Age	**Rent**	**Income**
Segment #1	[25 ; 45]	[250 ; 700]	<1000
Segment #2	[33 ; 60]	<400	>1200
Segment #3	[25 ; 55]	[500 ; 1000]	>2500

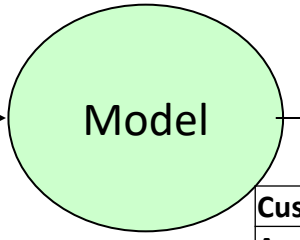
Unsupervised learning

- **Unsupervised** learning methods
  - The learning dataset **does not contain** the values of the response variable(s)
  - Example

Step 2: Applying the model

- to the training dataset…

Learning dataset		
Age	Rent	Income
36	0	1299
55	240	2500
40	768	3000
39	0	2000
44	334	512
26	631	722



	Customer Segments		
	Age	Rent	Income
Segment #1	[25 ; 45]	[250 ; 700]	<1000
Segment #2	[33 ; 60]	<400	>1200
Segment #3	[25 ; 55]	[500;1000]	>2500

Learning dataset			
Age	Rent	Income	Segment
36	0	1299	2
55	240	2500	2
40	768	3000	3
39	0	2000	2
44	334	512	1
26	631	722	1

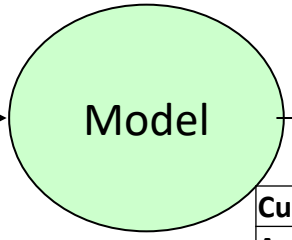
Unsupervised learning

- **Unsupervised** learning methods
  - The learning dataset **does not contain** the values of the response variable(s)
  - Example

Step 2: Applying the model

- to the training dataset… **and / or** to new data !

New sample			
Age	Rent	Income	Segment
30	260	900	???



New sample			
Age	Rent	Income	Segment
30	260	900	1

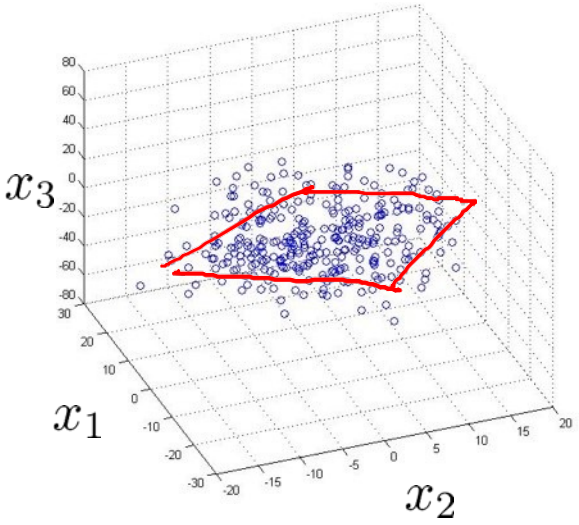
	Customer Segments		
	Age	Rent	Income
Segment #1	[25 ; 45]	[250 ; 700]	<1000
Segment #2	[33 ; 60]	<400	>1200
Segment #3	[25 ; 55]	[500;1000]	>2500

The model cannot always be expressed using data « slices »

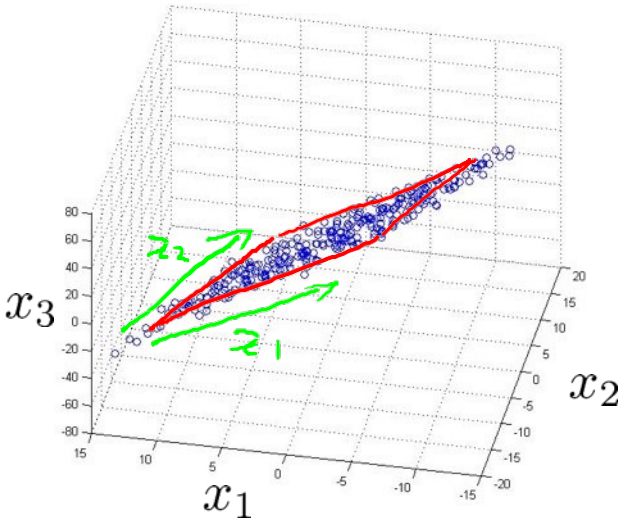
- it depends on the model

Unsupervised learning: goals

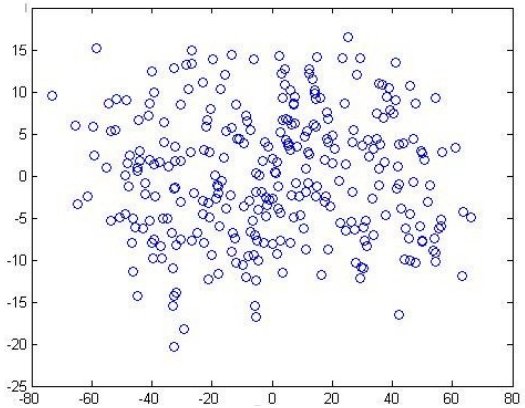
- Unsupervised learning might be used for
  - **Description**
    - Example: Principal Component Analysis for 3D -> 2D



Initial space *E* in 3D



PCA projection hyperplane E_q in 2D



New 2D representation space E_q

- Some very clear explanations about PCA:
  - https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Unsupervised learning: goals

- Unsupervised learning might be used for
  - Description
  - **Association**

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

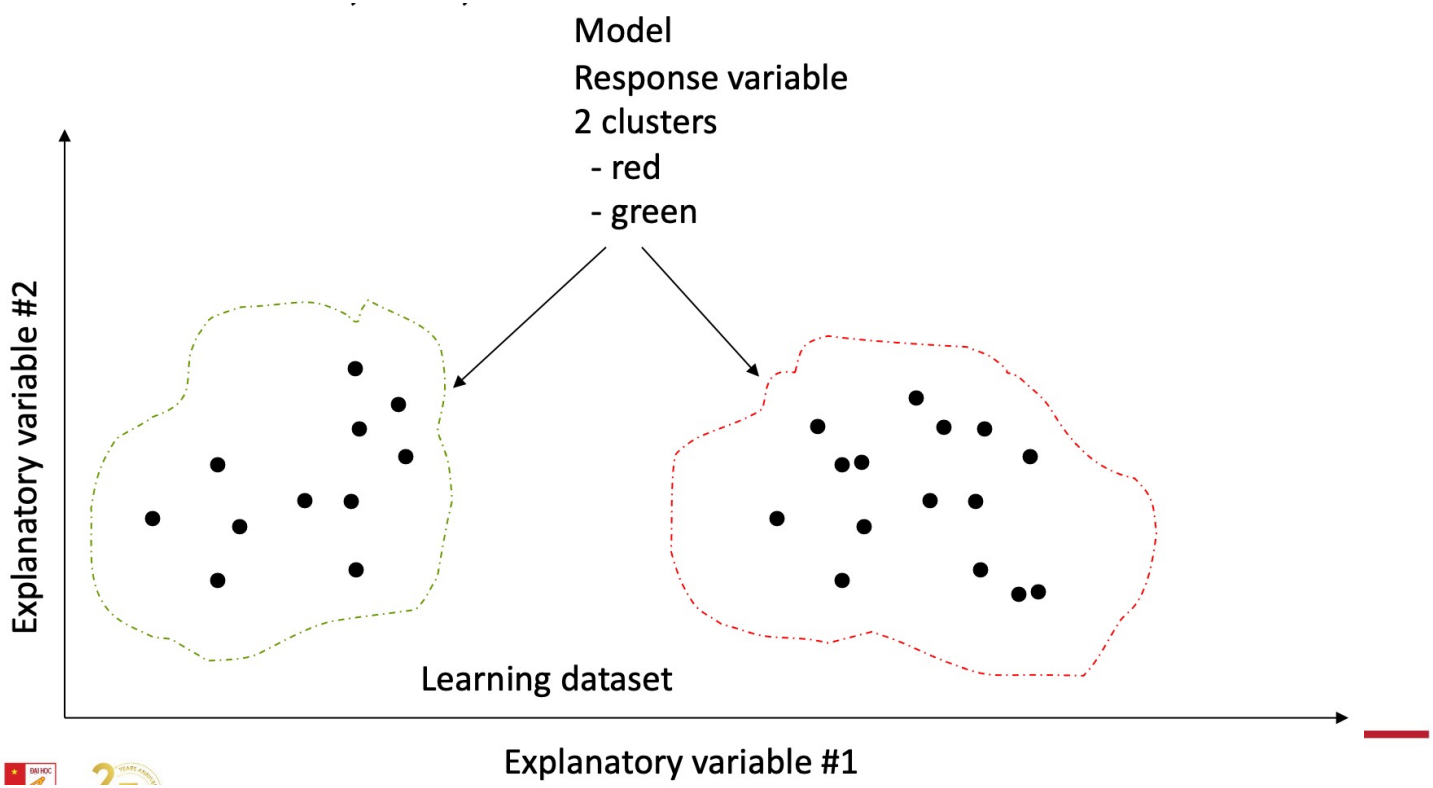
$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

For further reading on association rules:
http://www.saedsayad.com/association_rules.htm

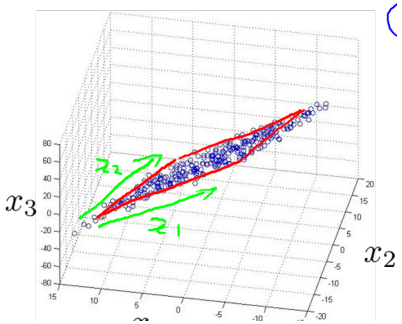
Unsupervised learning: goals

- Unsupervised learning might be used for
  - Description
  - Association
  - **Segmentation** (clustering)



Questions

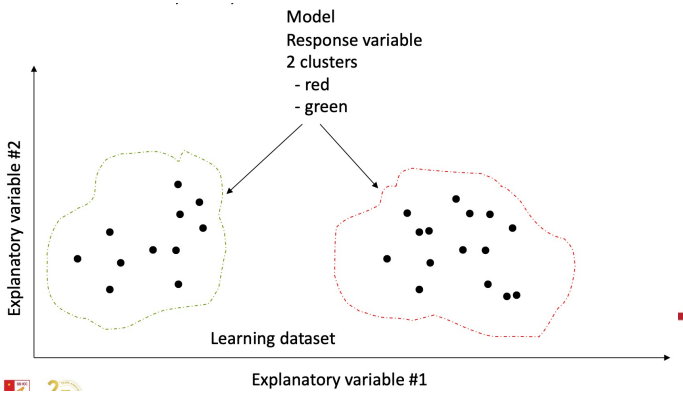
- From the data in your capstone project, give an example of possible:
  - Dimensionality reduction task:
    - Training dataset
    - Variables to be « reduced »



  - Association rules:
    - Training dataset
    - Examples of association rules from that dataset

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

  - Clustering task:
    - Training dataset
    - Explanatory variables



Introduction and definitions

Focus of this chapter

Focus of this chapter

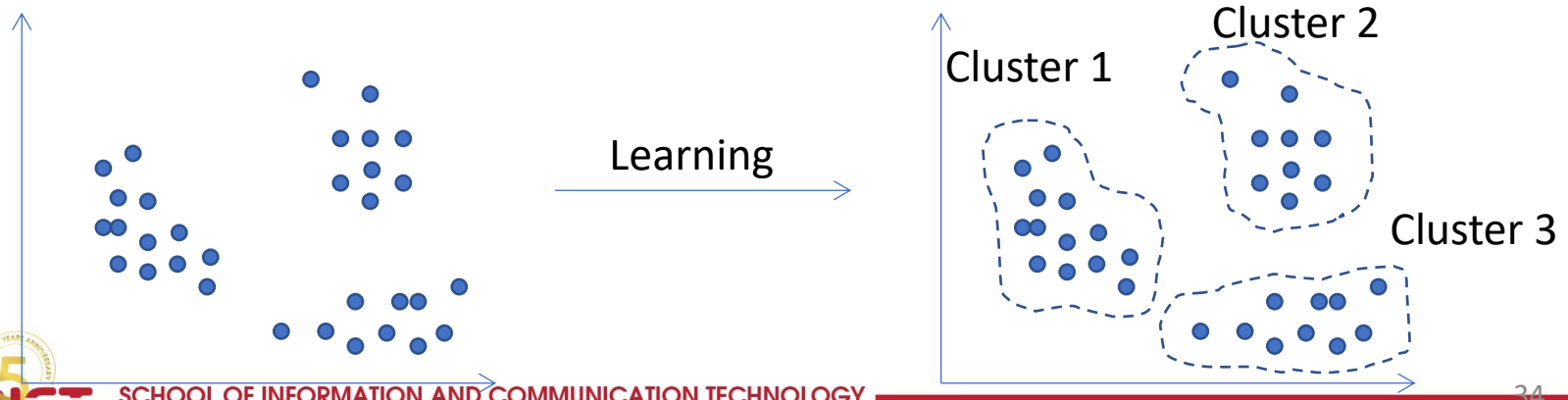
- In the rest of this chapter, we will focus **on clustering and classification** methods. Why?
  1. Because we don't have enough time to cover everything (this is only an intro course and there are thousands of methods)
  2. Because clustering and classification are what we need most for the following chapters (especially for image analysis)
  3. Because regression is easily self-taught
     - See Chapter 3 of the reference book: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)*
  4. Because association rules can be self-taught
     - See Chapter 14.2 of the reference book: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)*
     - See Chapters 6 and 7 of the reference book: *Data mining: concepts and techniques.* (more advanced)
  5. Because dimensionality reduction can also be self-taught
     - See Chapters 14.5 to 14.9 of the reference book: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*

Part 1: unsupervised clustering

Objective

Motivation

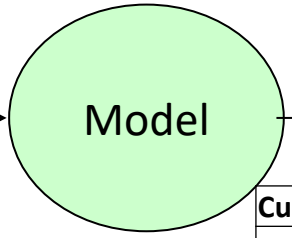
- Decomposing a data set into groups (clusters) helps make sense of the entire collection of observations
  - Groups are made out of **similar** observations
  - These groups are not pre-defined: the machine has to **learn** them
- Clustrering **segments** the dataset into homogeneous groups
  - Clustering is often used for **description** purposes
    - Some authors consider it as EDA (see Chapter 4)
      - *E.g.,* clustering can help identify outliers



Objective of unsupervised clustering

- Main objectives
  - First, learning homogeneous groups of data (**clusters**)
  - Second, assigning the data to the clusters
  - The response variable is **categorical:** the cluster
    - It is initially unknown (unsupervised learning)
  - The explanatory variables are the observed attributes
  - Reminder: example of clustering

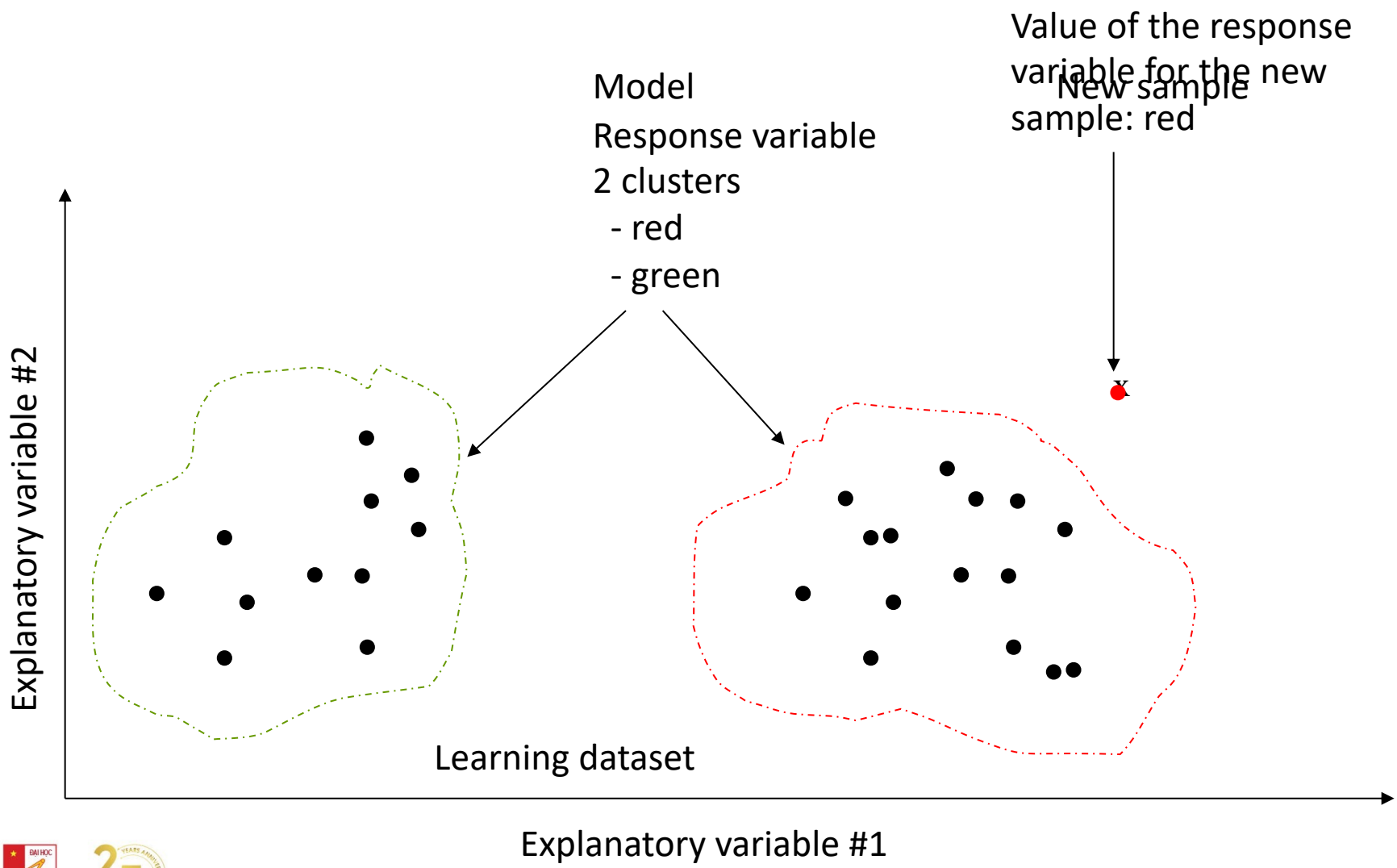
New sample			
Age	Rent	Income	Segment
30	260	900	???



New sample			
Age	Rent	Income	Segment
30	260	900	1

	Customer Segments		
	Age	Rent	Income
Segment #1	[25 ; 45]	[250 ; 700]	<1000
Segment #2	[33 ; 60]	<400	>1200
Segment #3	[25 ; 55]	[500;1000]	>2500

Unsupervised clustering: illsutration



Classification *vs.* clustering

- In both cases, the goal is to label the data, but:
  - With **classification**, the labels (**classes**) are **pre-defined**...
    - ... and known through observation (training dataset)
    - Class examples: (sick/not sick), (roses / orchid / iris),...
    - The objective is to **predict** the class of new samples
  - ... Whereas with **clustering**, the labels (**clusters**) are initially **unknown**
    - They are not given with the training dataset
    - The objective is to learn the clusters that make most sense
    - Cluster examples: customer segments, images with similar contents...
    - The objective is to **describe** and « summarize » the data (all data: training dataset and new samples)

Part 1: unsupervised clustering

Main issues and useful definitions

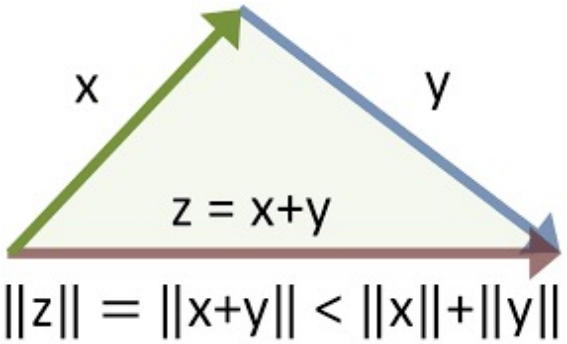
Clustering: main issues to address

- What is « **similar** »?
- How do we group observations?
- How do we evaluate the quality of the grouping?

Notion of similarity - dissimilarity

- Any clustering method requires a similarity or dissimilarity measure
  - The **similarity** between x_i and y_i is high when x_i and y_i are close (in the representation space)
  - The **dissimilarity** between x_i and y_i is high when x_i and y_i are far (in the representation space)
  - A **distance** is a special case of a dissimilarity measure
    - Distances verify the triangle inequality, but dissimilarity measures do not necessarily verify it

Triangle inequality



Some useful distances / similarity measures

Minkowski distances:

$$D(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^r \right)^{1/r}$$

- Euclidean distance

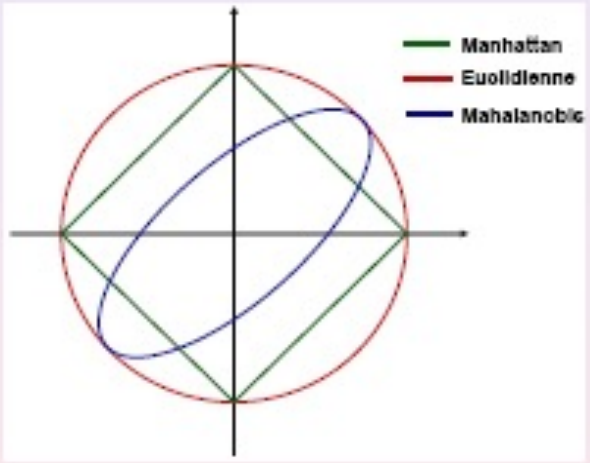
$$D(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

- Manhattan / city-block distance

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} |x_i - y_i|$$

Mahalanobis distance

$$D(\boldsymbol{x}, \boldsymbol{y}) = [\det V]^{1/m} (\boldsymbol{x} - \boldsymbol{y})^{\mathrm{T}} V^{-1} (\boldsymbol{x} - \boldsymbol{y})$$ with *V*: covariance matrix



Unit circles:

circles with a radius of 1

Cosine similarity measure

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\sum x_i y_i}{\sqrt{\sum (x_i)^2} \times \sqrt{\sum (y_i)^2}}$$

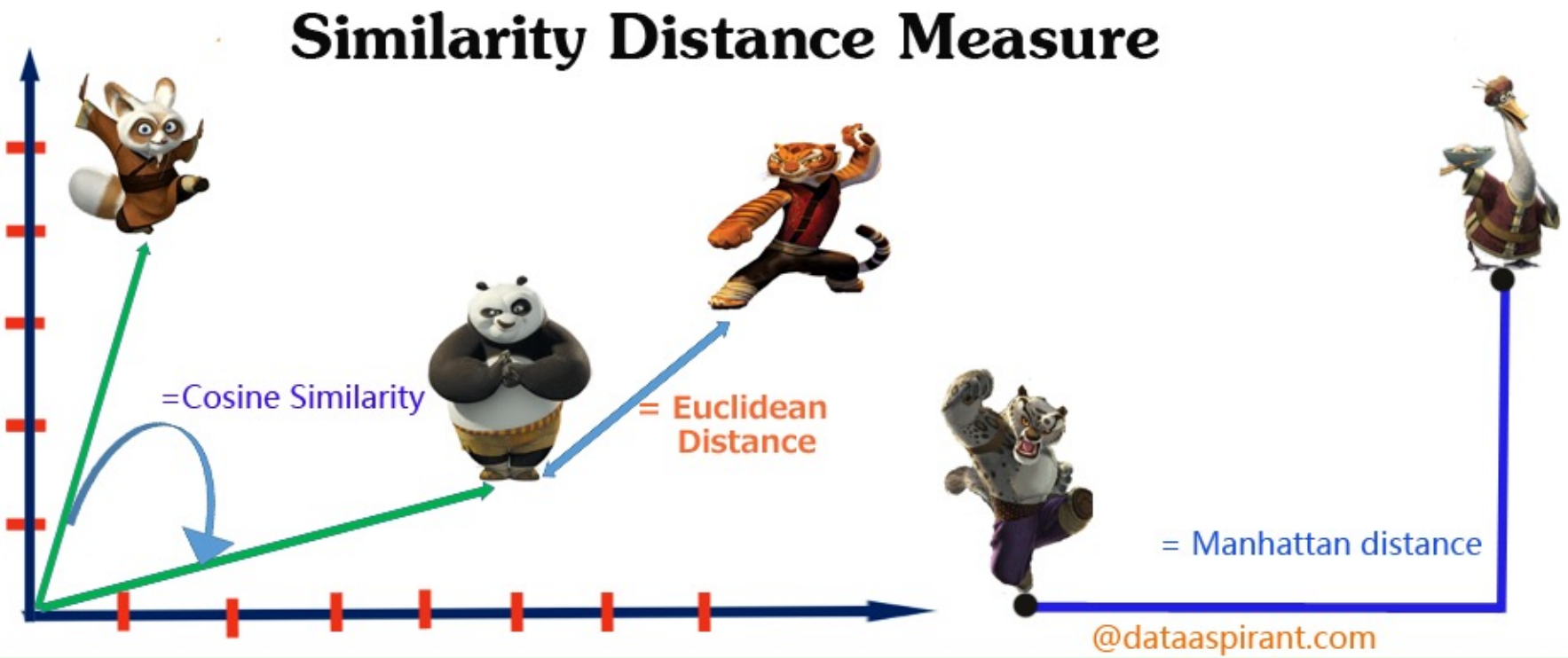
The computed similarity resides on the interval `[-1,1]`, where vectors with the same orientation have a similarity equal to `1`, orthogonal orientation a similarity equal to `0`, and opposite orientation a similarity equal to `-1`. The cosine distance seeks to express vector dissimilarity in positive space and does so by subtracting

Cosine distance

$$d(\mathbf{x}, \mathbf{y}) = 1 - s(\mathbf{x}, \mathbf{y})$$

Some useful distances / similarity measures

- Summary (partial)
- Now, the main difficulty is to pick the distance / similarity that is best for our problem
    - Quite a tough question, but for later…



Part 1: Unsupervised clustering

Methods

Clustering: main issues to address

- What is « **similar** »?
- How do we group observations?
- How do we evaluate the quality of the grouping?

Different types of clustering methods

- Partitioning methods
  - « cuts » the data into groups of similar observations
- Hierarchical methods
  - Builds a hierarchical structure of clusters
- Grid-based methods
  - Divides the representation space into cells and merges neighbouring cells to create clusters
- Density-based methods
  - Defines clusters as high-density regions

Part 1: unsupervised clustering

Partitioning methods

Partitioning methods

- Partitioning methods
  - « cuts » the data into groups of similar observations
  - « Exact » method: evaluates all possible partitions
    - Too much time-consuming!
  - Heuristic methods:
    - k-means, k-medoids, CLARA, CLARANS, ISODATA, …

Partitioning methods: k-means

- **Requires**
  - A distance / dissimilarity
  - The desired number of clusters: *k*
  - An initialization of the *k* cluster centroids
- **Produces**
  - Final estimate of cluster centroids
  - An assignment of each point to one of the *k* clusters
    - (hard clustering)

Partitioning methods: k-means

- There are many variants to the k-means algorithm

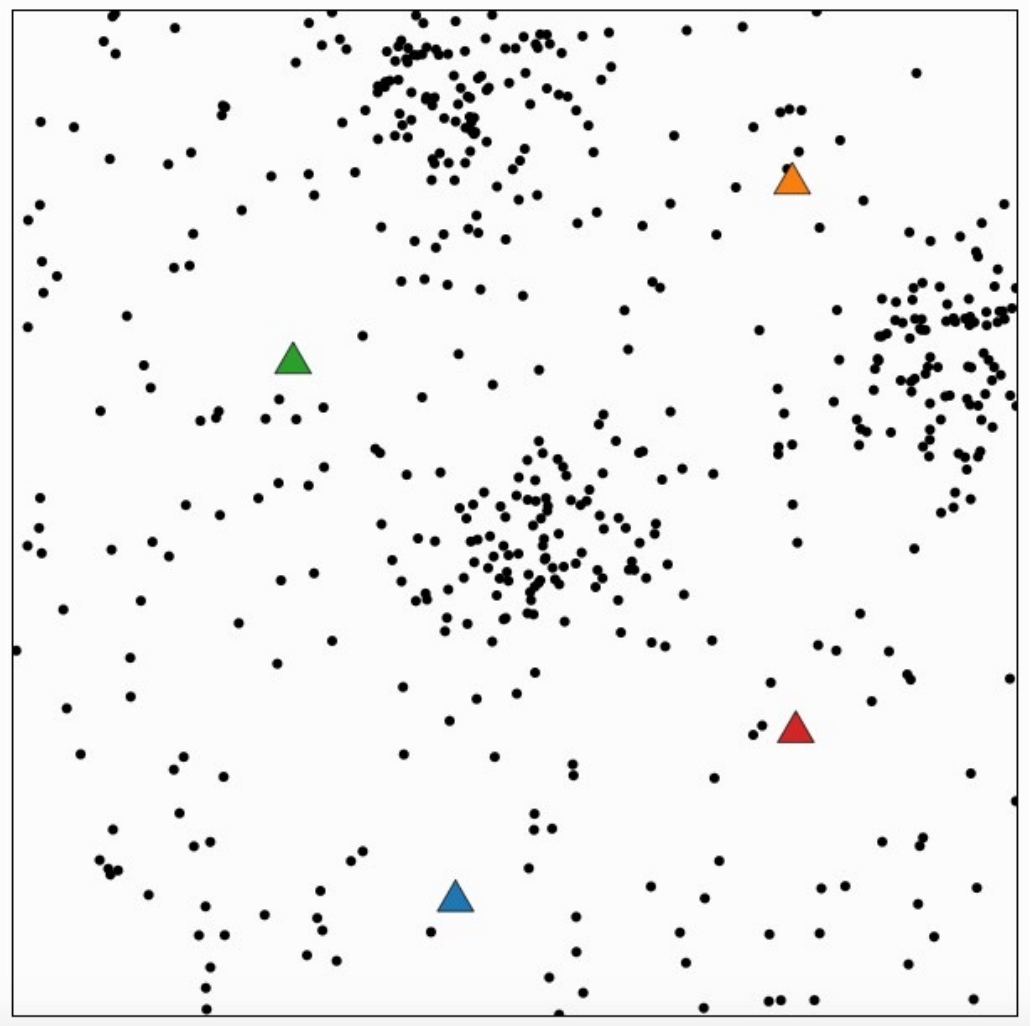
- The most « basic » algorithm is:
  1. Initialize randomly the k centers of the k clusters
  2. Assign each object to the cluster with the nearest center
     - Minimizing:

$$I = \sum_{j=1}^{k} \sum_{x_i \in K_j} (x_i - \mu_j)^2$$ where μ_j is the centroid of cluster K_j

  3. Recalculate the center of gravity (average) μ_j of each cluster K_j
  4. Iterate steps 2 and 3 until observations no longer change clusters

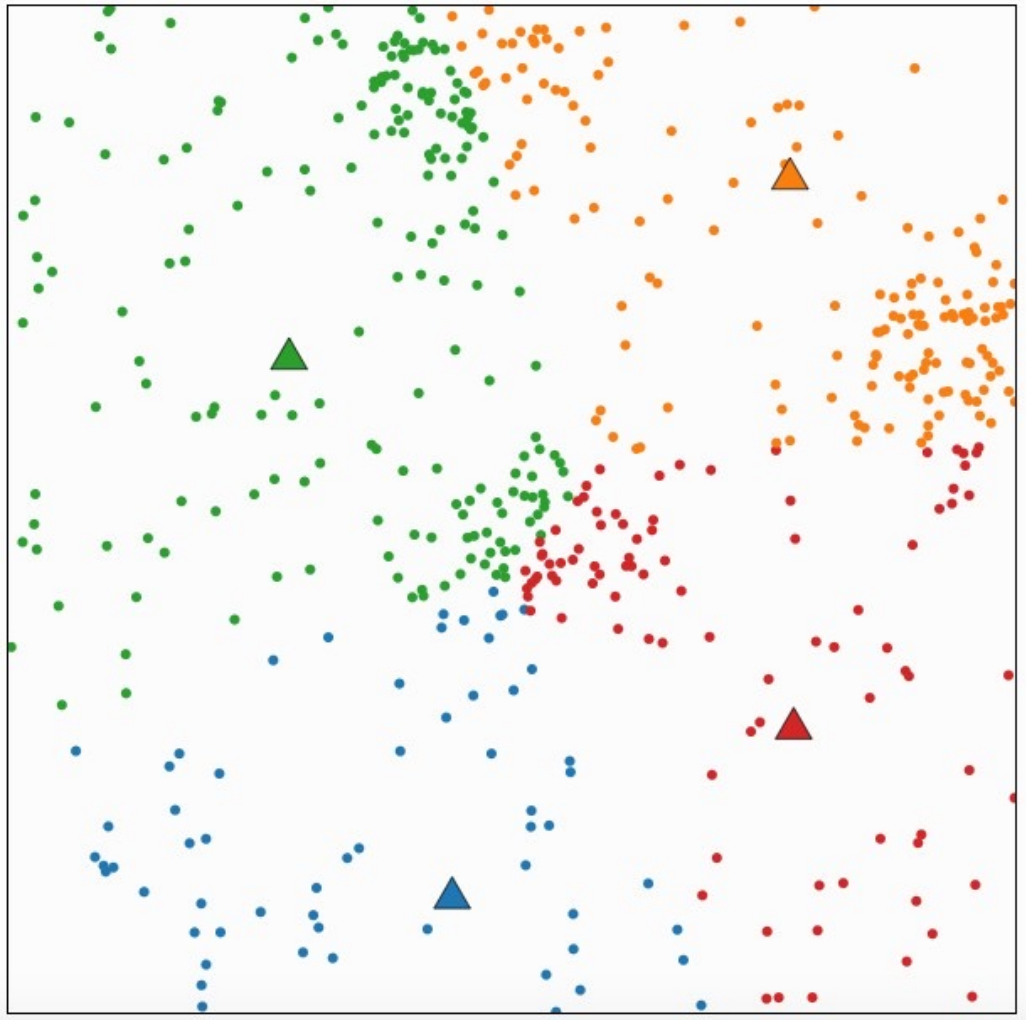
Partitioning methods: k-means

- Example with
  - k=4
  - Euclidean distance



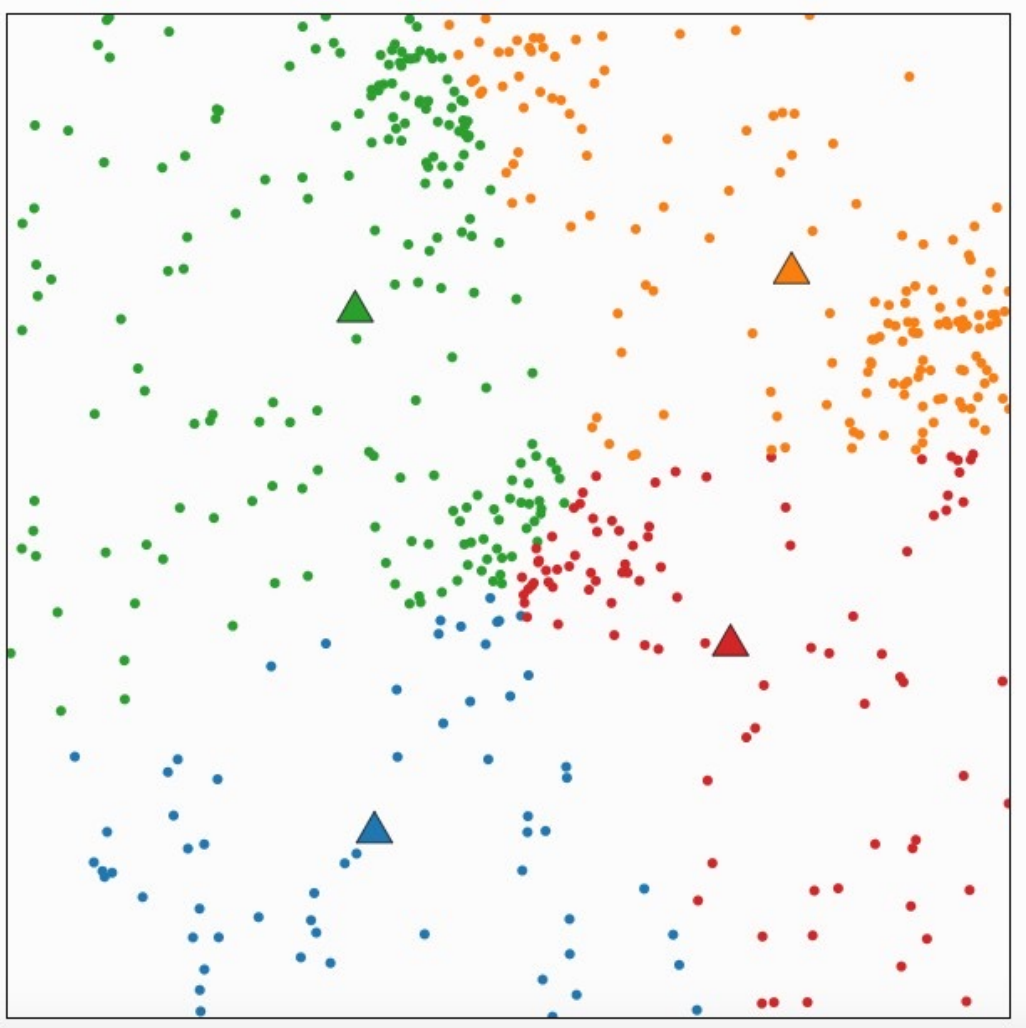
Partitioning methods: k-means

- Example with
  - k=4
  - Euclidean distance



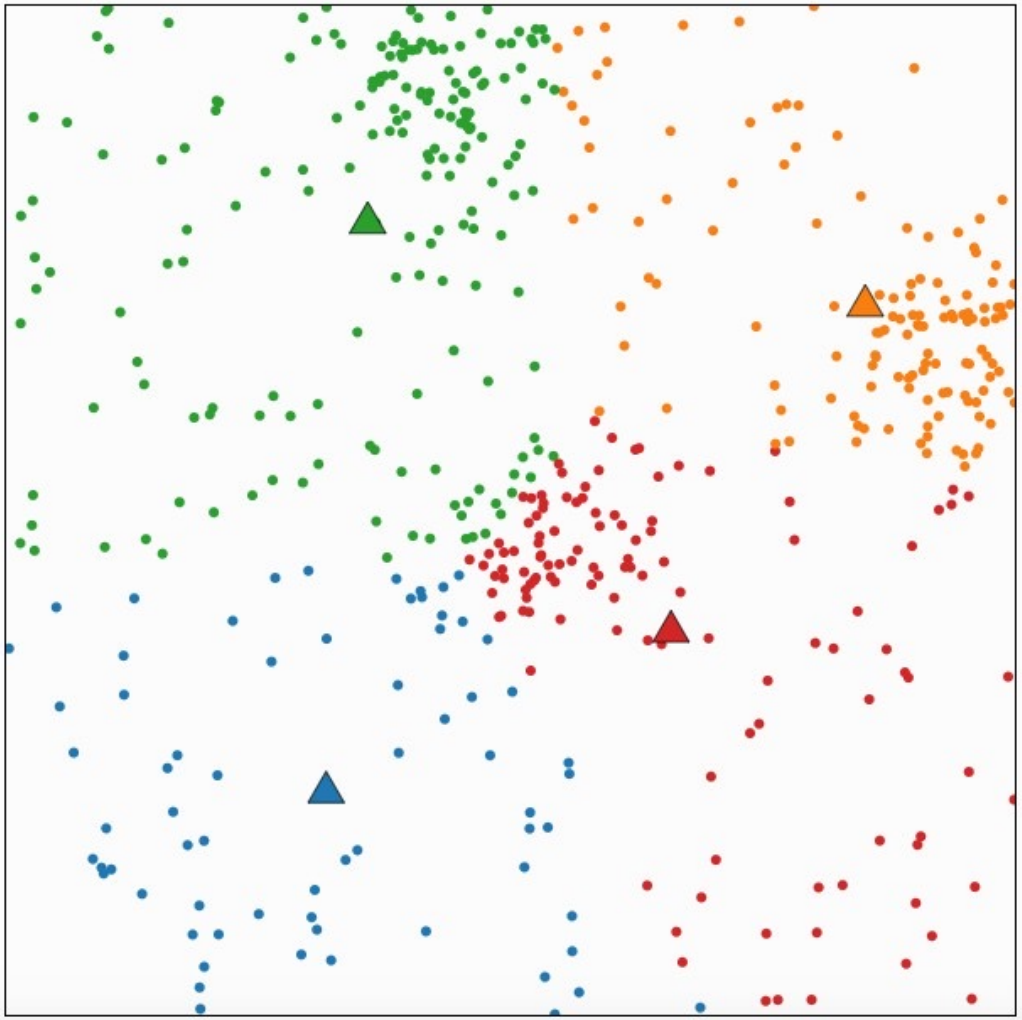
Partitioning methods: k-means

- Example with
  - k=4
  - Euclidean distance



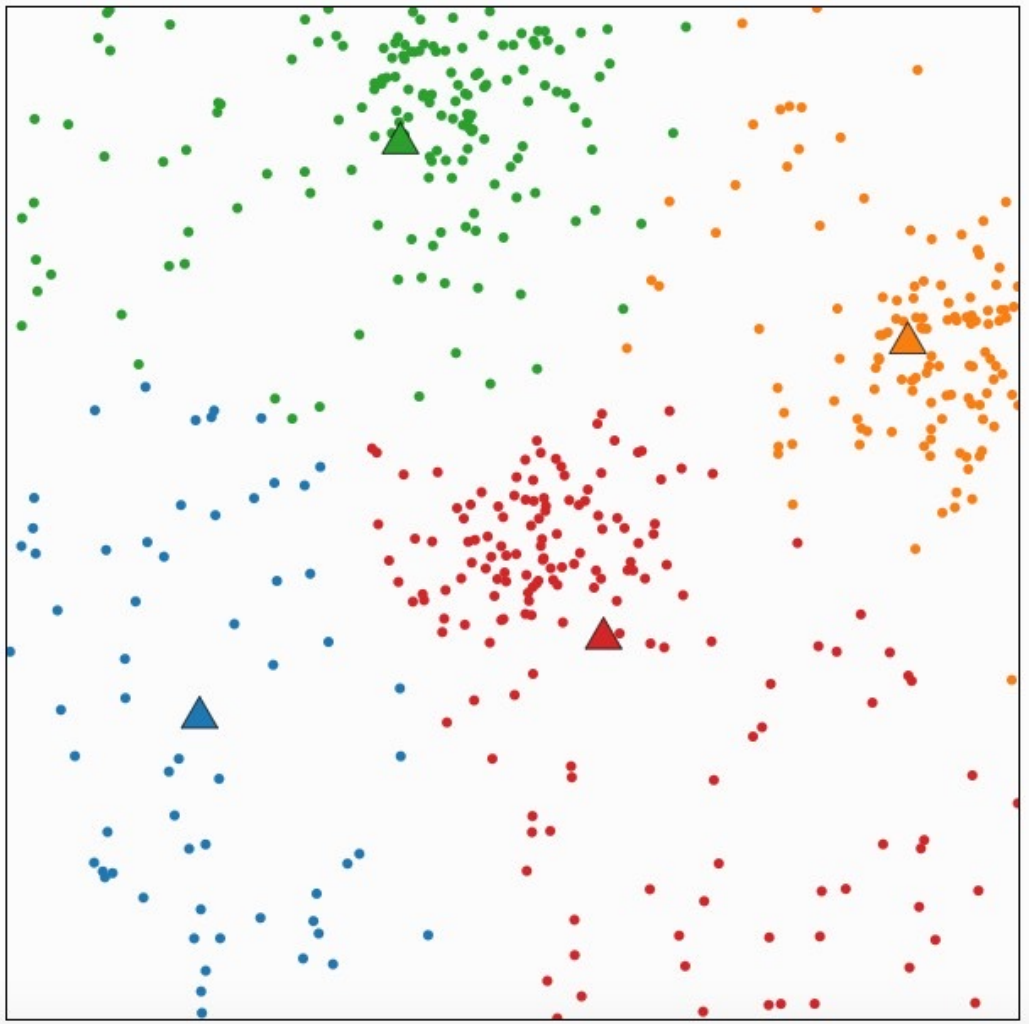
Partitioning methods: k-means

- Example with
  - k=4
  - Euclidean distance



Partitioning methods: k-means

- Example with
  - k=4
  - Euclidean distance



Partitioning methods: k-means

☺ Simple to implement

☺ Relatively low complexity O(*Nkl*) where:

- *N* is the number of objects and *l* is the number of iterations
- *l* and *k* are generally much lower compared to N

☹ Sensitive to outliers

☹ Sensitive to the initialization of the centroids

- Running the algorithm twice will not produce the same output

☹ Highly depends on the distance used

☹ Converges to a local optimum

☹ Need to pre-define the number *k* of clusters

Other partitioning methods

- K-medoids: similar to k-means, but each cluster is represented by its medoid
  - The medoid of cluster K_j is the observation from K_j with the minimum average distance with the other observations in K_j
  - Less sensitive to outliers, but much more computationally expensive
  - Example of algorithm: PAM (Partitioning Around Medoids)
- Variants of PAM for reducing its computational complexity
  - CLARA, CLARANS, etc.

Part 1: unsupervised clustering

Hierarchical methods

Hierarchical methods

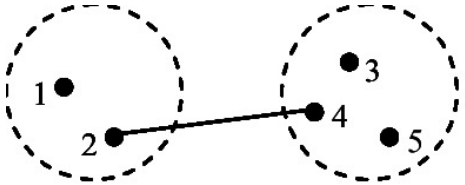
- Hierarchical methods create a hierarchical decomposition of clusters into sub-clusters
- Two types of approaches:
  - **Bottom-up** approaches (agglomeration): successively merge similar clusters
  - **Top-down** approaches (division): successively divide clusters into sub-cluster.
- Many methods: AHC, AGNES, DIANA, BIRCH, ROCK, CURE, R-tree, SS-tree, SR-tree, ...

Hierarchical methods: bottom-up

- An example of bottom-up method: **AHC** (Agglomerative Hierarchical Clustering)
  1. Assign each observation to a cluster
  2. Merge two closest clusters
  3. Repeat step 2 until there is only one cluster left

- Step 2 requires a distance between 2 **clusters** K_i, K_j: Different possibilities:
  - Single-linkage:

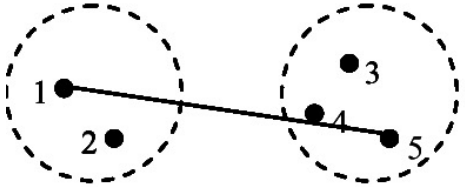
$$D(K_i, K_j) = \min_{x_l \in K_i, x_m \in K_j} D(x_l, x_m)$$



d_{24}

  - Complete-linkage:

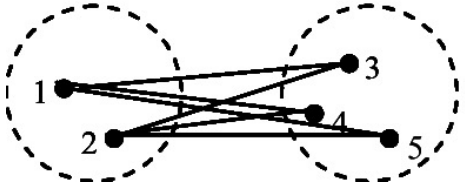
$$D(K_i, K_j) = \max_{x_l \in K_i, x_m \in K_j} D(x_l, x_m)$$



d_{15}

  - Average-linkage:

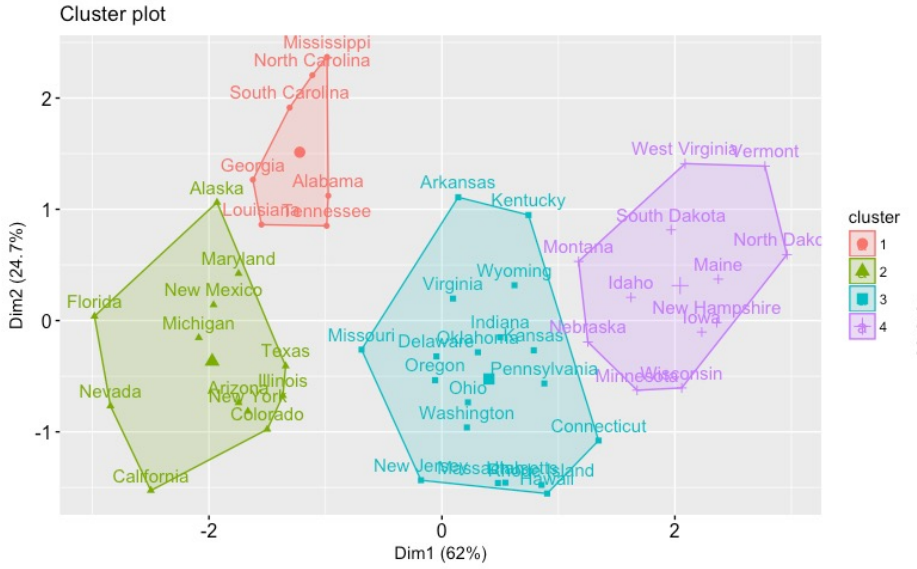
$$D(K_i, K_j) = \text{mean}_{x_l \in K_i, x_m \in K_j} D(x_l, x_m)$$

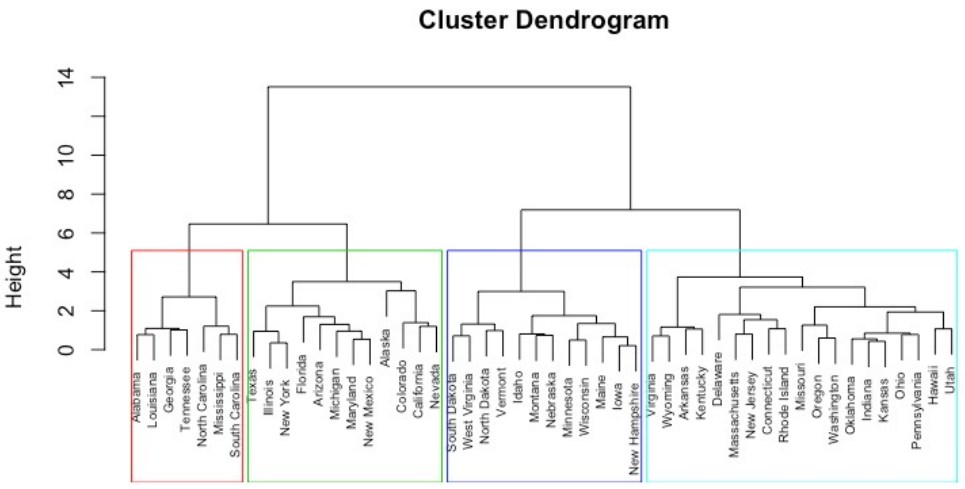


$$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$$

Hierarchical methods: **bottom-up**

- An example of application of **AHC**
  - « Criminal profile » of the different US states
    - https://uc-r.github.io/hc_clustering
  - The dendogram helps us:
    - Choose the number of clusters
    - Understanding the clusters distribution
    - When necessary, zoom in/out for each cluster





Hierarchical methods: **bottom-up**

- **AHC**

☺ Simple to implement

☺ No need to pre-define the number *k* of clusters

☹ Sensitive to outliers

☹ Highly depends on the distance used

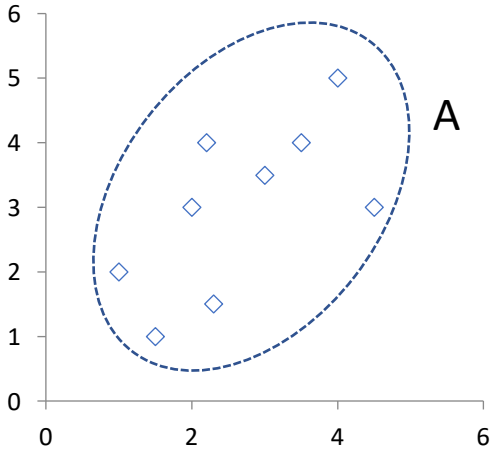
☹ Very high complexity $O(N^2)$

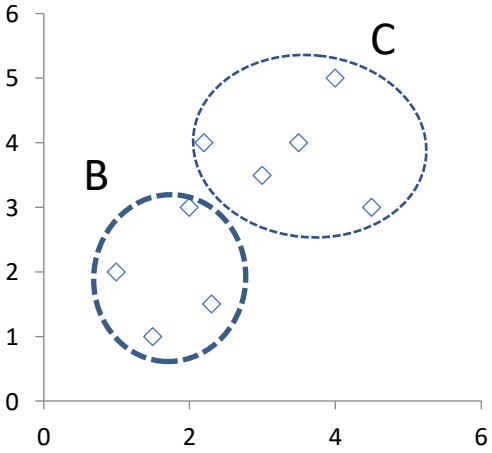
Hierarchical methods: top-down

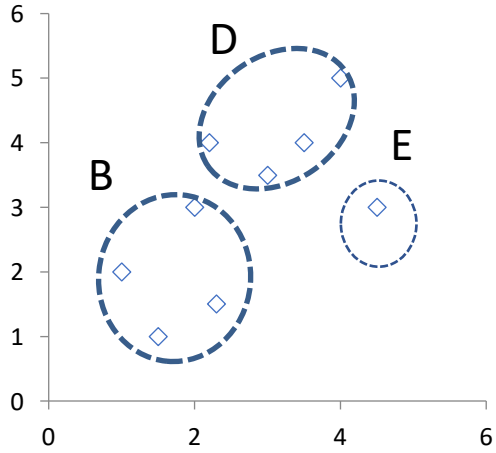
- An example of top-down method: **DIANA**
  - Start with a cluster that groups all the observations from the training dataset
  - Iteratively, **divide** the cluster with maximum diameter into two sub-clusters
    - The diameter of a cluster being the greatest distance between any two objects in the cluster
  - Stop when all clusters contain only one observation, or the maximum desired number of clusters is reached (if this parameter is pre-defined)

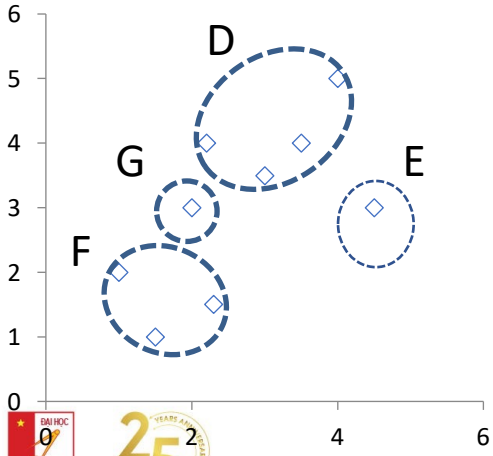
Hierarchical methods: **top-down**

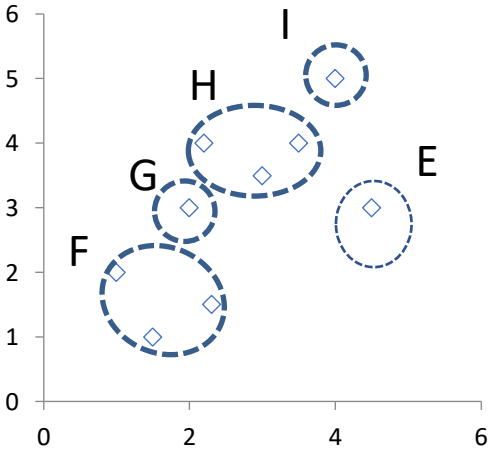
- **DIANA** example with K=5

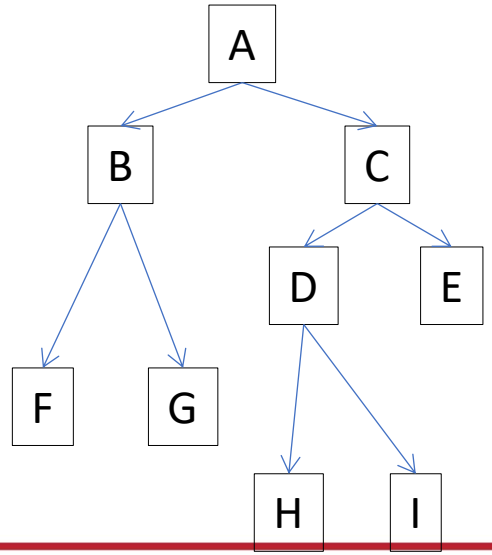












Hierarchical methods: **top-down**

- **DIANA** algorithm for cluster division (with K the cluster with largest diameter at the current iteration)

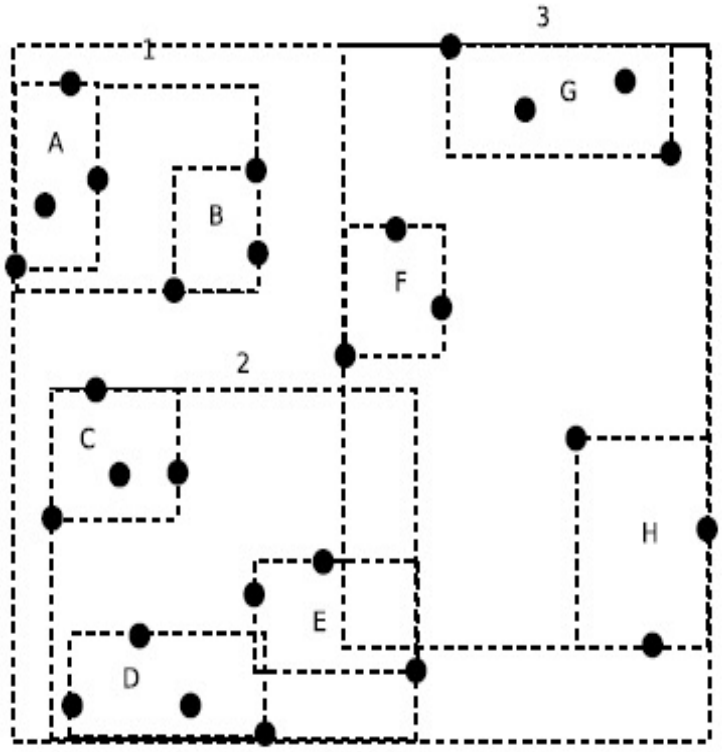
1. Identify $x' \in K$ having the largest average dissimilarity with other K objects ; x' starts a new cluster K'
2. For each $x_i \notin K'$, calculate

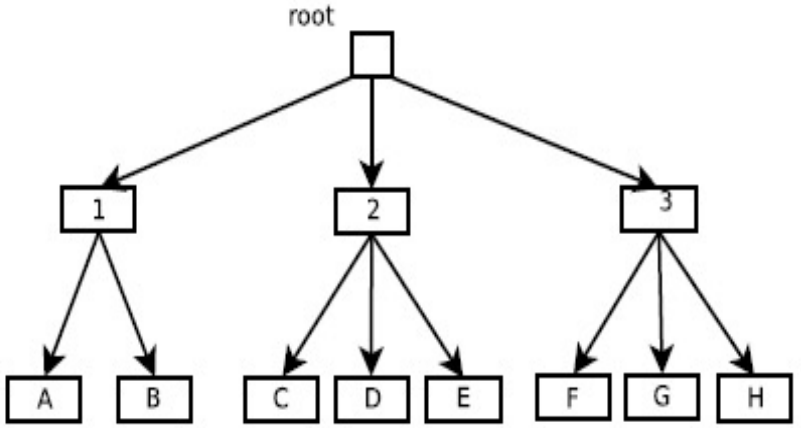
   $$d_i = \text{average}[D(x_i,x_j) \mid x_j \in K \setminus K'] - \text{average}[D(x_i,x_j) \mid x_j \in K']$$
3. Choose x_i^* for which d_i is the largest($d_i^* = \text{Argmax}_i(d_i)$).

   If $d_i^* > 0$, then add x_i^* in K'
4. Repeat steps 2 and 3 until $d_i^* < 0$

Other top-down hierarchical methods

- Many other top-down hierarchical methods: BIRCH, R-tree, SS-tree, SR-tree…
  - There is no « best » method: everything depends on your problem!
  - Example: R-tree



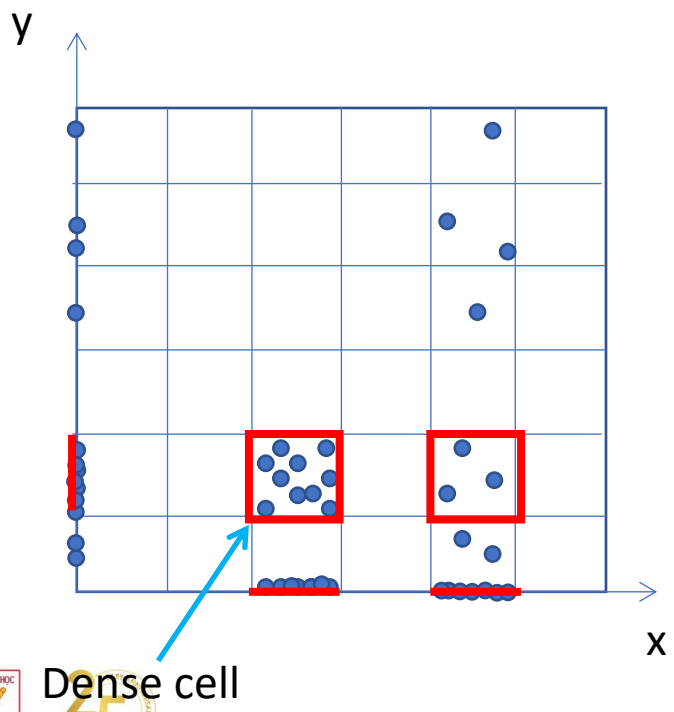


Part 1: unsupervised clustering

Grid-based methods

Grid-based methods

- Grid-based methods
  - Divides the representation space into cells and merges neighbouring cells to create clusters
  - Several methods: STING, CLIQUE…



- Example: CLIQUE method: iteratively:
  - Projecting data on x and y axes
  - Determining dense 1-dimensional cells
  - Intersection of 1-dimensional dense cells to form 2-dimensional candidates
  - Determine among the candidates the cells that are really dense
  - Merge neighbouring dense cells into the same cluster

Part 1: unsupervised clustering

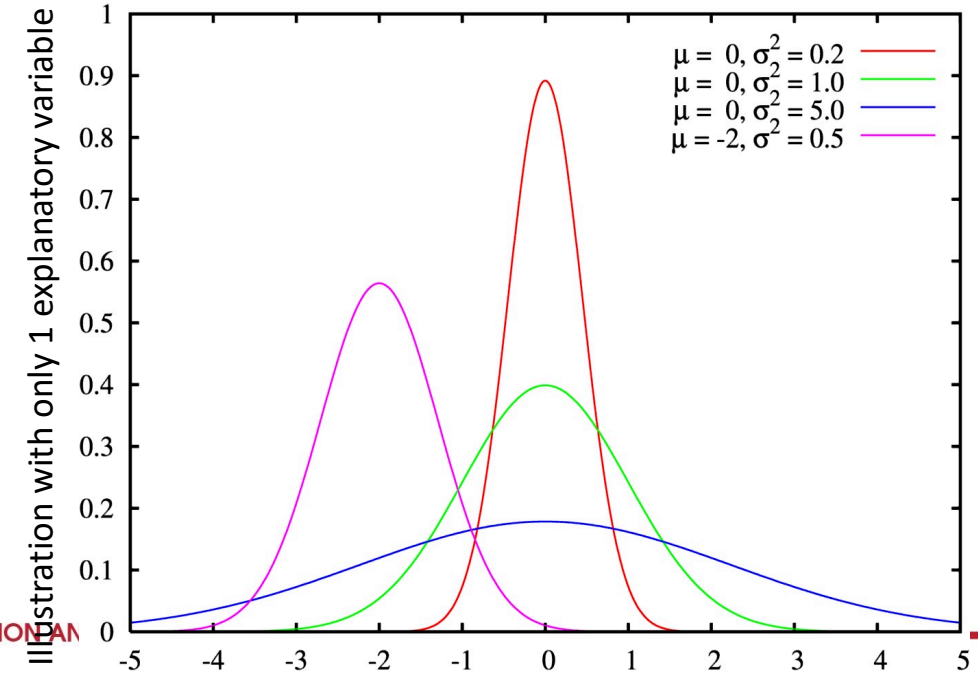
Density-based methods

Density-based methods

- Density-based methods
  - Defines clusters as high-density regions
    - Same idea as grid-based methods, but more refined
- There are two types of methods
  - **Non-parametric** approaches:
    - DBSCAN, DENCLUE, OPTICS, ...
  - **Parametric** Approaches: assume that data is distributed according to a known model
    - EM (Expectation-Maximization)

Density-based methods

- Example of density-based, **parametric** method: **EM**
  - Observations are assumed to be distributed according to a mixture model of *K* Gaussian distributions (each Gaussian = 1 cluster)
  - EM algorithm estimates the optimal parameters of the Gaussian mixture (averages and covariance matrices) **iteratively**, until convergence



Density-based methods

- Example of density-based, **parametric** method: **EM**

  - Algorithm
    - **Step 1 (Parameter Initialization)**
      - Initialize all μ_i and σ_i
    - **Step 2 (Expectation)**
      - For each point x,
        - For each cluster i,
          - Calculate the probability that x belongs cluster i
    - **Step 3 (Maximization)**
      - For each cluster i,
        - Calculate the mean μ_i according to the probabilities that all points belong to cluster i
    - Repeat Step 2 and Step 3 until the parameters converge

One possible implementation:

$$p(x \in C_i) = \frac{p(x|<\mu_i, \sigma_i>)}{\sum_j p(x|<\mu_j, \sigma_j>)}$$

One possible implementation:

$$\mu_i = \sum_x x \cdot p(x|<\mu_i, \sigma_i>)$$

Image from Raymond Wong

More explanations about the EM algorithm:
https://www.youtube.com/watch?v=REypj2sy_5U

Part 1: unsupervised clustering

Performance evaluation

Clustering: main issues to address

- What is « **similar** »?
- How do we group observations?
- How do we evaluate the quality of the grouping?

Clustering performance measures

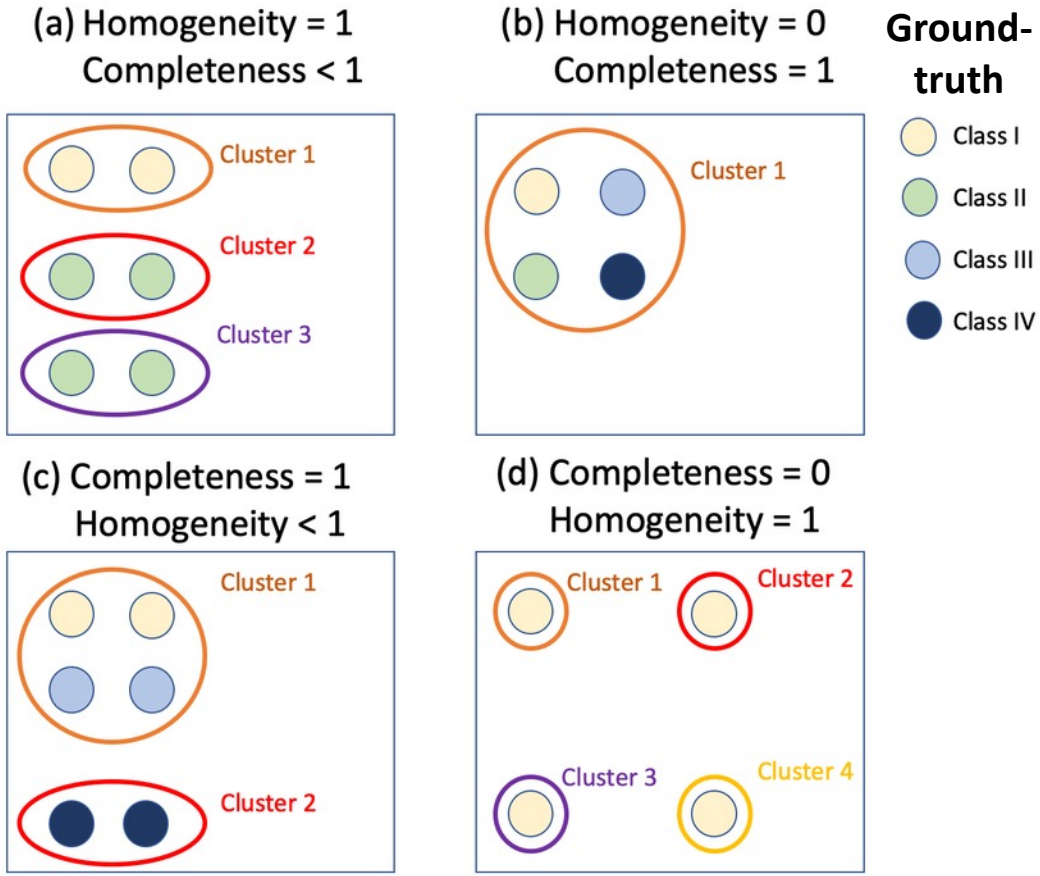
- There are two types of performance measures for clustering
  - Extrinsic Measures: require **ground truth** labels
    - Labels can be obtained through manual annotation
    - Labels can be seen as a special kind of **illustrative variable**
    - Examples of extrinsic measures:
      - Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure
  - Intrinsic Measures*:* do **not require** ground truth labels
    - Examples of intrinsic measures
      - Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

Questions

- From the data in your capstone project, give an example of possible ground-truth label:
  - What information is used to define the ground-truth?
  - Who can give you that information?

Clustering performance measures

- Examples of Extrinsic measures
  - **Homogeneity**
    - All clusters contain only observations that are members of a single (ground-truth) class
    - Mathematically defined using conditional entropies
    - Permutations of the class or cluster label values won't change the score
  - **Completeness**
    - All of the (ground-truth) classes contain only data points which were put into a single cluster by the algorithm



Clustering performance measures

- Examples of Extrinsic measures
  - **V-measure**
    - « Summarizes » homogeneity and completeness into a single (scalar) value
    - Harmonic mean of homogeneity and completeness

$$V = \frac{2}{\frac{1}{Homogeneity} + \frac{1}{Completeness}} = \frac{2 \cdot Homogeneity \cdot Completeness}{Homogeneity + Completeness}$$

Clustering performance measures

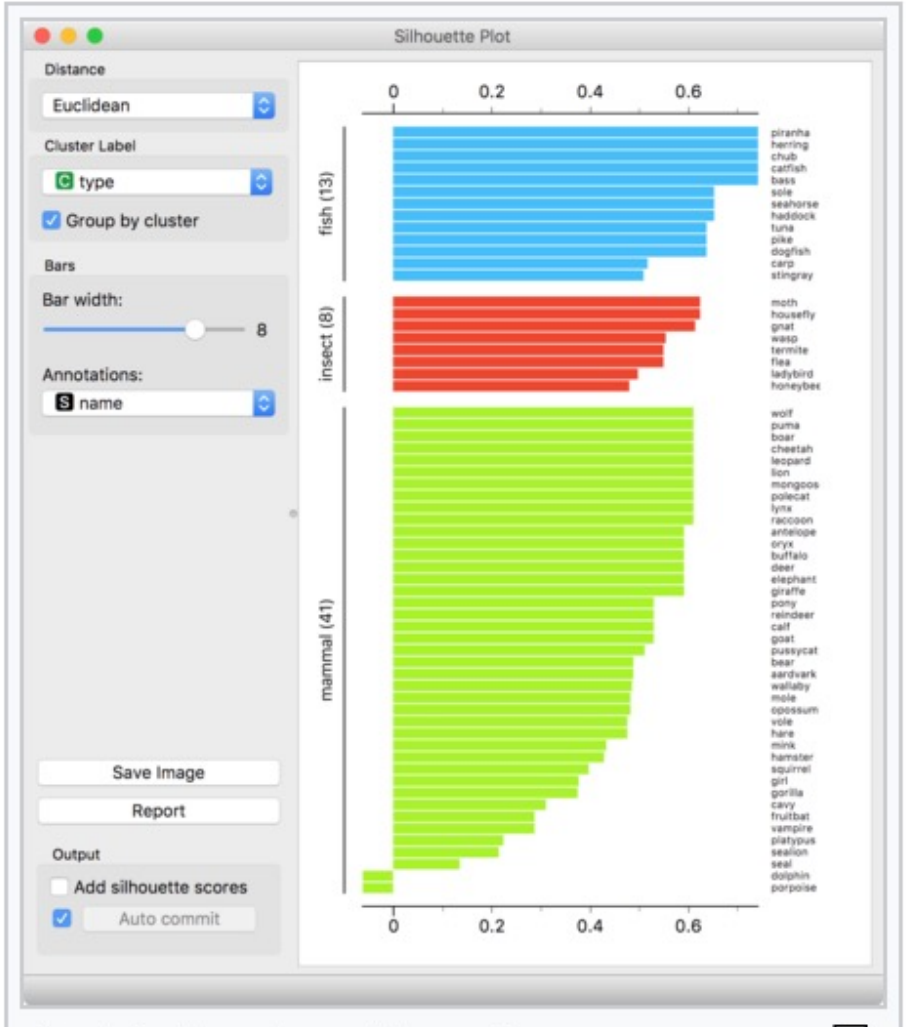
- Example of Intrinsic measure: **silhouette** value
  - The problem with extrinsic measures is, we don't often have access to the real (ground-truth) classes!
  - The silhouette value is a measure of how similar an observation is to its own cluster (cohesion) compared to other clusters (separation)

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases} \quad \text{, thus} \quad -1 \leq s(i) \leq 1$$

  - Where *a(i)* is a measure of how well the observation x_i belongs to its own cluster K_j
  - *b(i)* is the distance between x_i and its closest cluster $K_i \neq K_j$
  - For the details about a(i) and b(i), just look on Wikipedia
    - https://en.wikipedia.org/wiki/Silhouette_(clustering)

Clustering performance measures

- Example of Intrinsic measure: **silhouette** value



Wikipedia

A plot showing silhouette scores from three types of animals from a Zoo. At the bottom of the plot, silhouette identifies two outliers in the “green” cluster.

Summary

Summary

- In this lecture, you've learned about:
  - Basic notions of machine Learning:
    - Explanatory variables, response variables, illustrative variables
    - Supervised *vs.* unsupervised learning
  - Unsupervised clustering
    - Some examples of clustering problems
    - How to address the main issues of clustering
      - How to define "similar"?
        -> overview of different similarity measures / distances
      - How to group the data?
        -> overview of different clustering methods
        - Partitioning methods
        - Hierarchical methods
        - Grid-based methods
        - Density-based methods
      - How to evaluate the quality of the groups?
        -> performance evaluation
        - Extrinsic evaluation measures
        - Intrinsic evaluation measures

Homework

Homework: tutorials using Python

- Partitioning methods: example with *k-means*
  - https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py
- Hierarchical method: example with AHC
  - https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/
- Density-based method: example with EM algorithm
  - https://www.python-course.eu/expectation_maximization_and_gaussian_mixture_models.php

Questions

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!!!

soict.hust.edu.vn/ fb.com/groups/soict