

Clustering algorithms

Konstantinos Koutroumbas

Unit 1

- General concepts
- Problem formulation

Course grades

70%: Final exams (obligatory)

20%: Project (obligatory)

20%: Homeworks

Programming language

MATLAB

Suggested bibliography

1. S. Theodoridis, K. Koutroumbas, “Pattern Recognition”, 4th ed., Academic Press, 2008.
2. C. C. Aggarwal, C. K. Reddy, editors “Data Clustering: Algorithms and Applications”, CRC Press, 2014.
3. G. Gan, C. Ma, J. Wu, “Data Clustering: Theory, Algorithms and Applications”, ASA-SIAM, 2007.

Clustering definition

Input: A set E of **entities**.

Clustering:

Grouping of the **entities** into “**sensible**” **clusters** (groups), so that:

- “**more similar**” entities to belong to the **same cluster** and
- “**less similar**” entities to belong to **different clusters**.

Why clustering?

Difficulty:

- We are **surrounding** by a **wealth** of pieces of **information** (of any kind).
- The human brain is unable to study each one of them separately in an attempt to extract “knowledge” from each one of them.

Idea to address the difficulty:

- **Cluster similar entities** to a group and study the properties of the group as a whole.
- All the **knowledge extracted** for the group characterizes each one of its constituent entities.

Clustering definition

Input: A set E of **entities**.

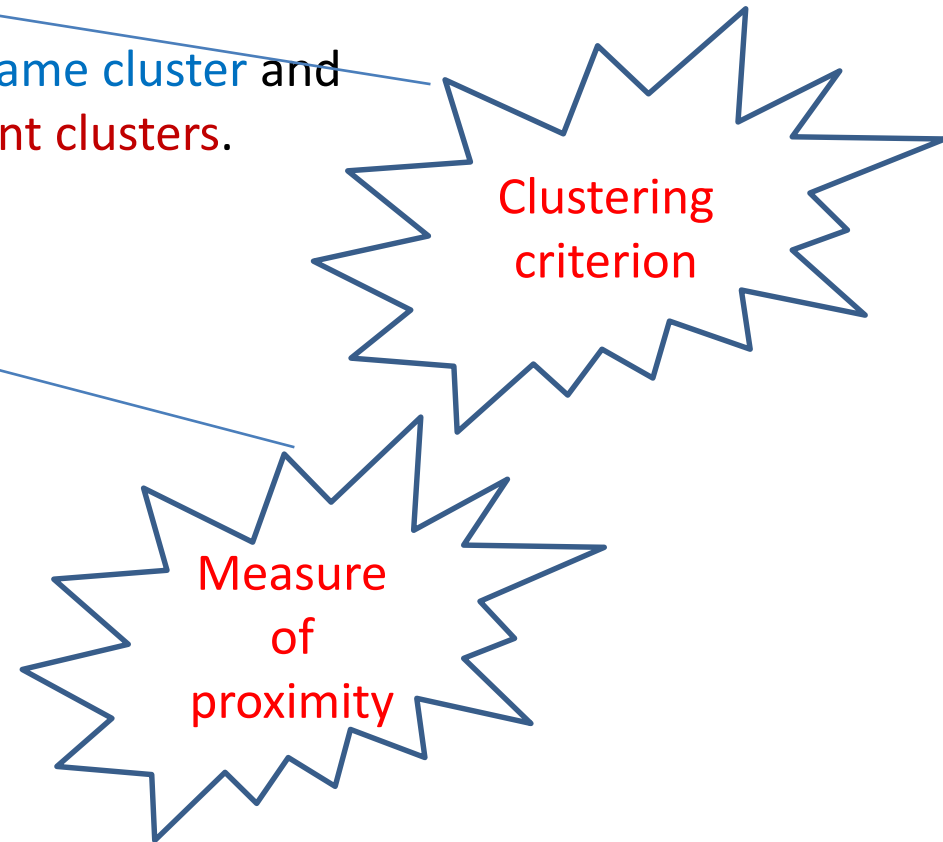
Clustering:

Grouping of the **entities** into “**sensible**” **clusters** (groups), so that:

- “**more similar**” entities to belong to the **same cluster** and
- “**less similar**” entities to belong to **different clusters**.

Concepts that need to be clarified:

- Entity
- Measure of proximity
- Cluster
- Clustering criterion



Clustering definition

Input: A set E of entities.

Clustering:

Grouping of the **entities** into “sensible” clusters (groups), so that:

- more similar entities to belong to the same cluster and
- less similar entities to belong to different clusters.

Concepts that need to be clarified:

- **Entity**
- Measure of proximity
- Cluster
- Clustering criterion



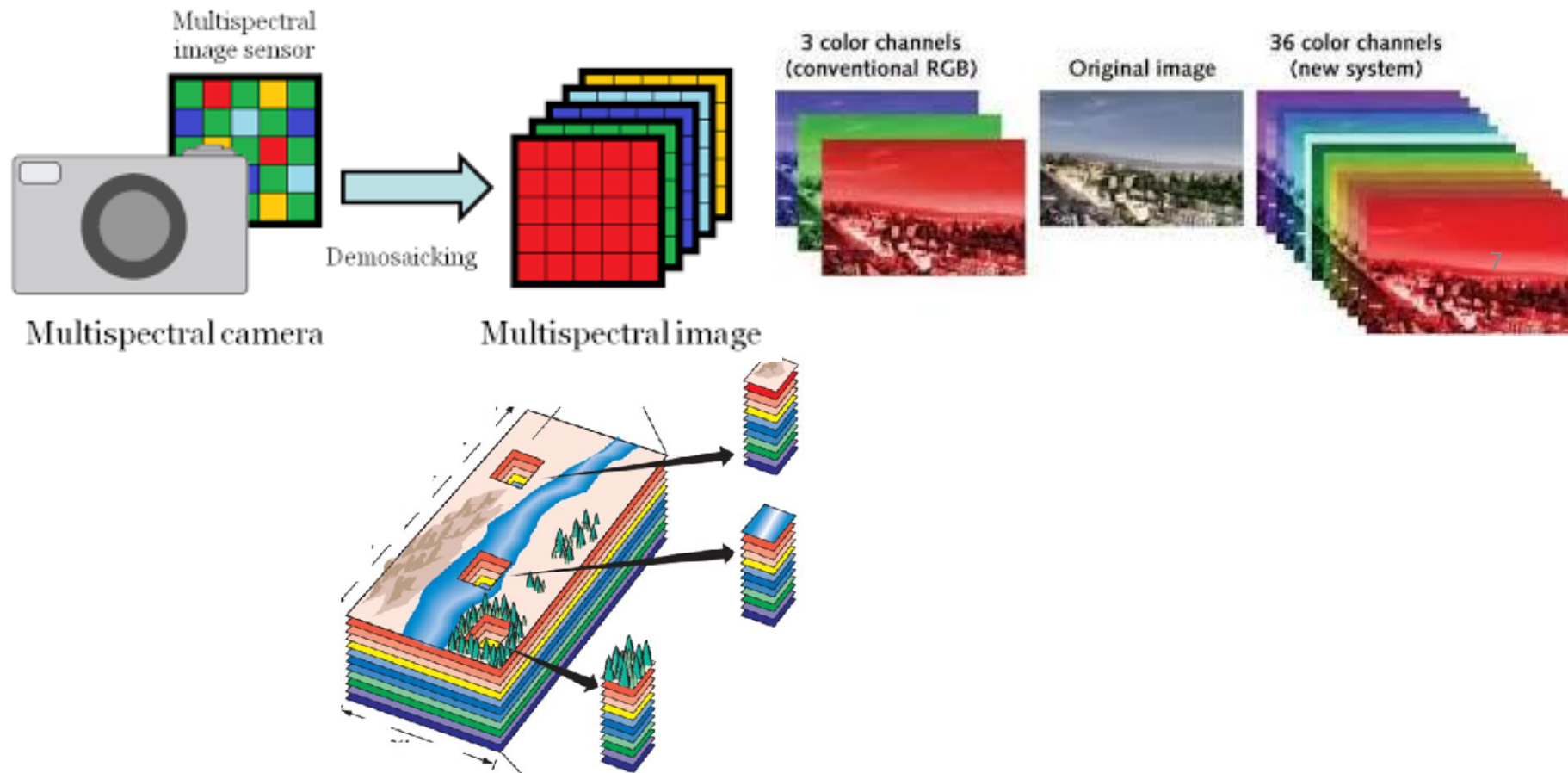
Entities

Entities

They are **application-dependent**.

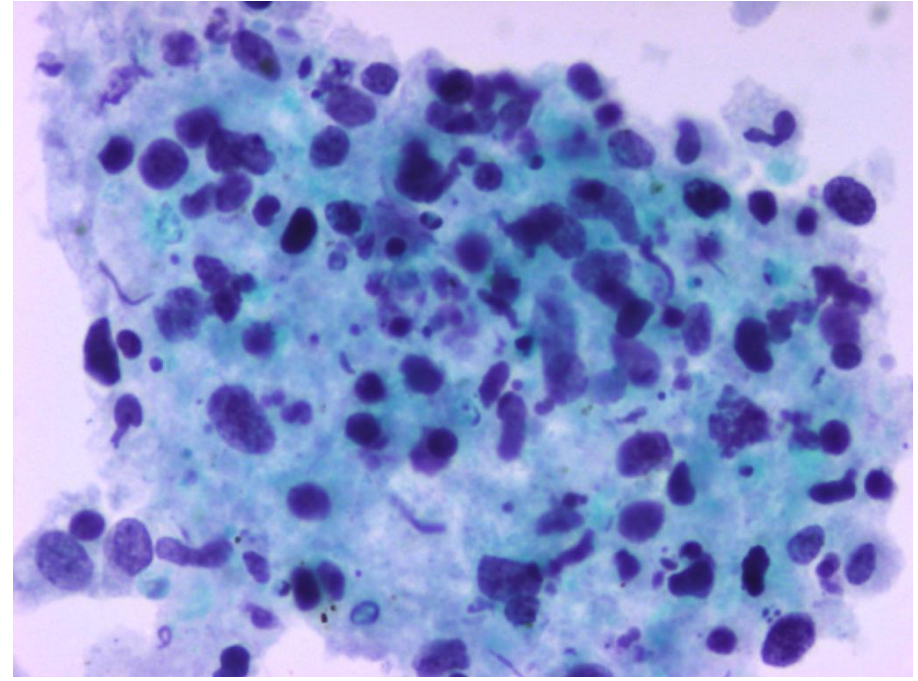
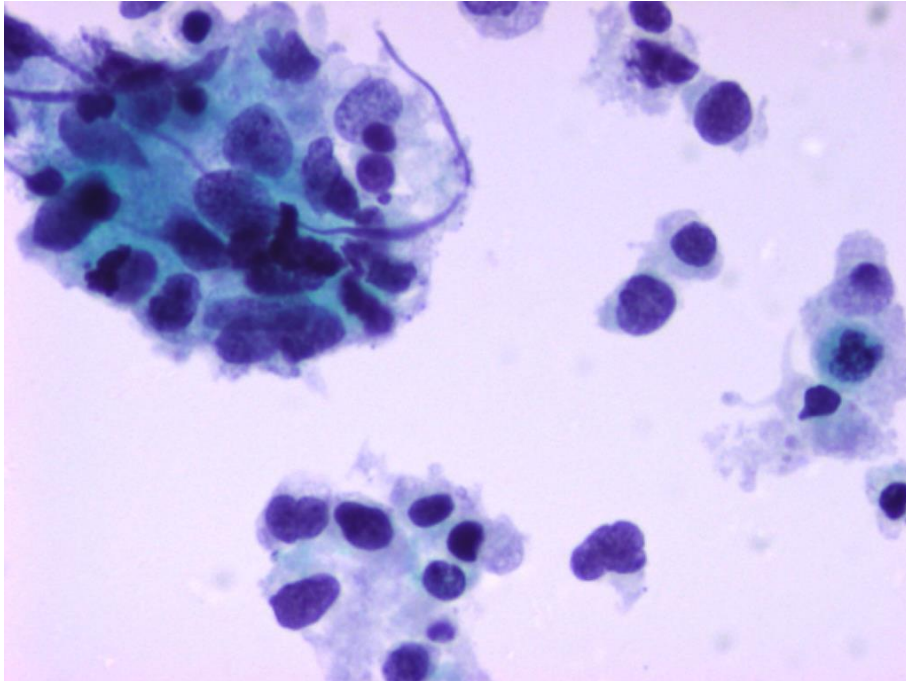
They can be e.g.,

(A) Images (grayscale, multispectral, hyperspectral ...): **whole** images, **parts** of images, image **pixels** ...

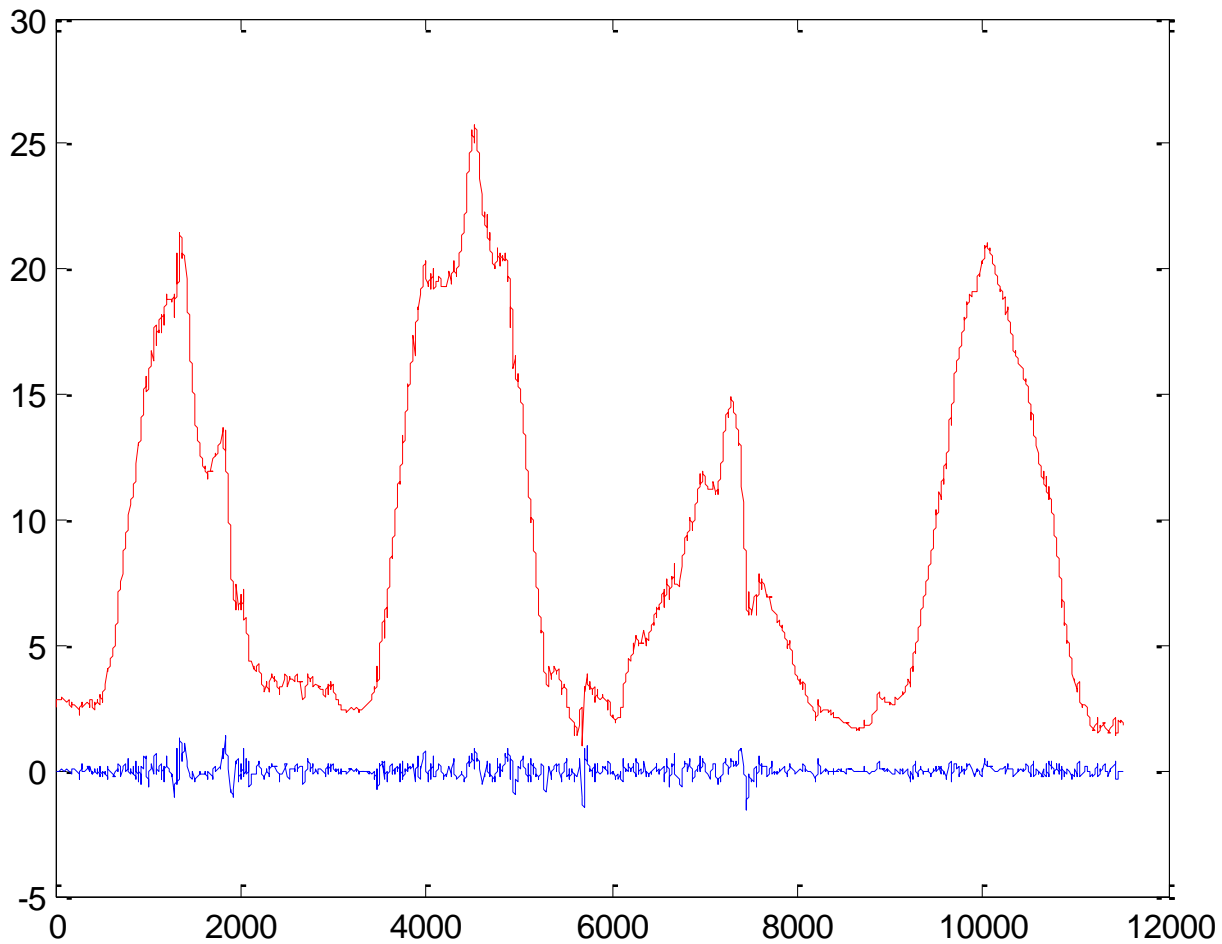


Entities

(B) Human **cells** or **tissues**.



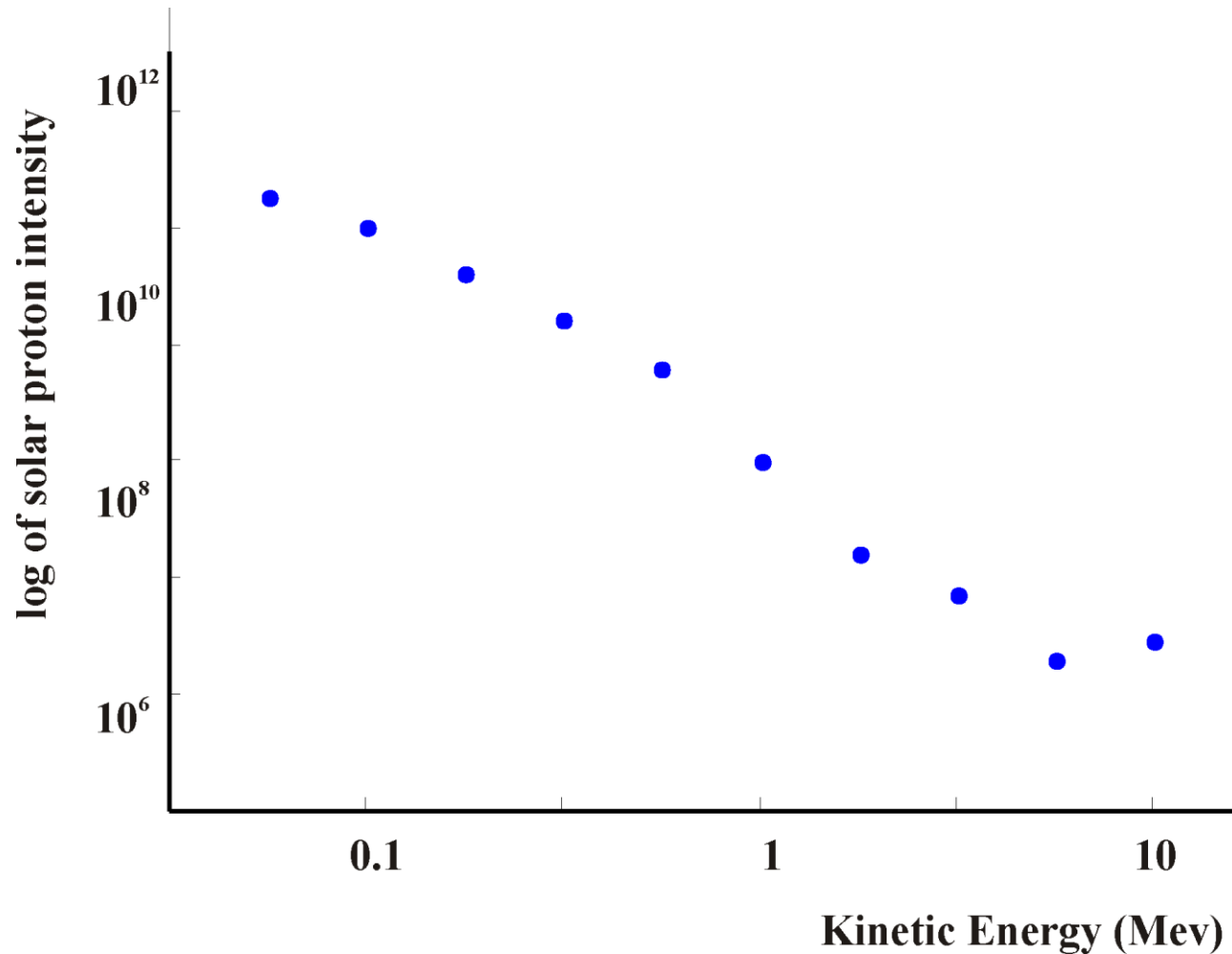
(C) Time series.



(D) Text

A B C D E F G H I J K L M N O
P Q R S T U V W X Y Z Å Ø Ü å
b c d e f g h i j k l m n o p
q r s t u v w x y z & 1 2 3 4
5 6 7 8 9 0 (€ £ • ¬ ! ?)

(E) Pairs of measures of related quantities (each point correspond to a pair)



Entities

(F) Consumers **personal preferences**

(G) . . .

Entity representation

AN IMPORTANT ISSUE: ENTITY REPRESENTATION

How the **entities** involved in a **clustering problem** are **represented**?

Since in most of the cases we deal with *statistical processing methods*, we need “**numerical**” **representations**.

In general, we work as follows:

For each entity, we **select** a set of l quantities (measurements - **features**), **the same for all entities**.

Then, each **entity** is **represented** as a **point** in an l -dimensional space (**feature space**).

$$i\text{-th entity} \Leftrightarrow \mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{il}]^T$$

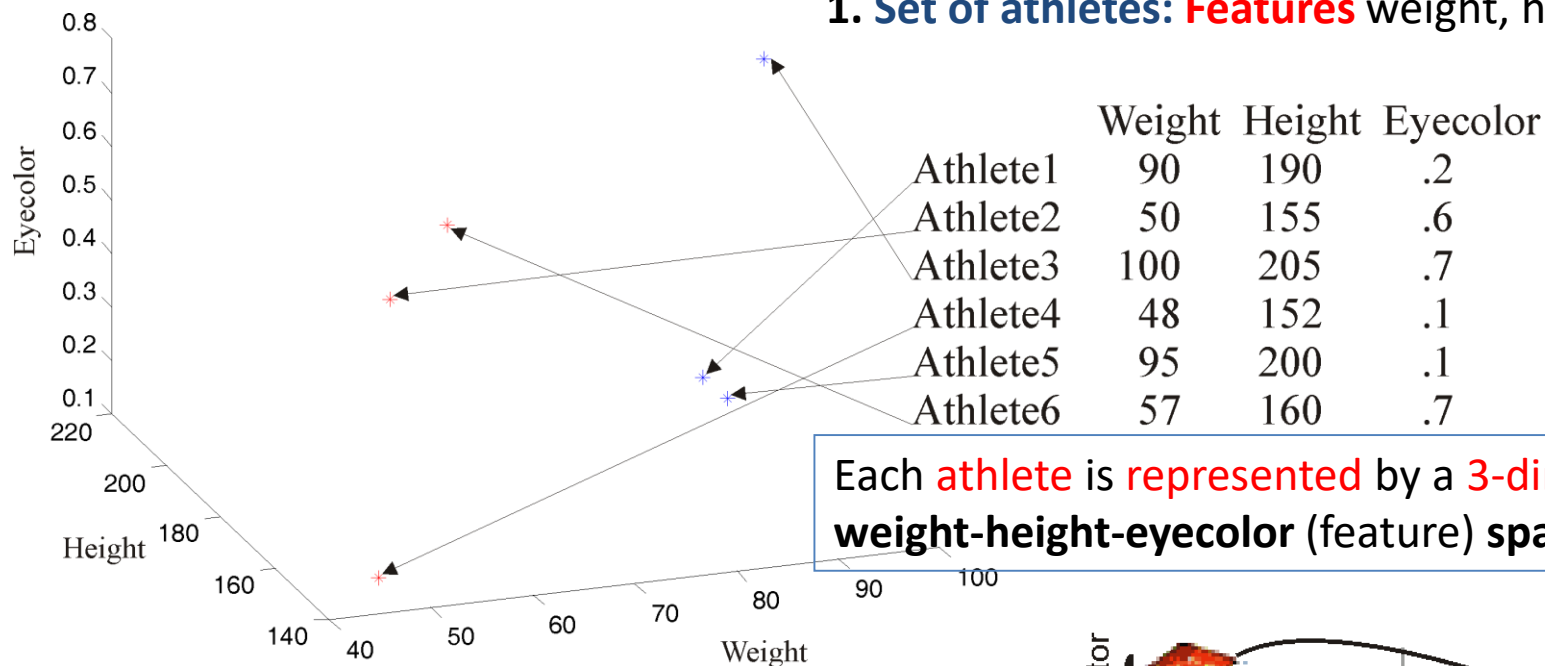
IMPORTANT NOTE: The **adoption** of suitable **features** is a **highly application dependent** stage.

Entity representation

AN IMPORTANT ISSUE: ENTITY REPRESENTATION (cont.)

Examples:

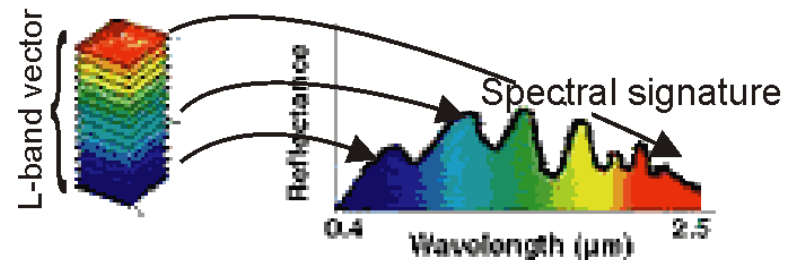
1. Set of athletes: **Features** weight, height, eyecolor



Each **athlete** is represented by a **3-dim. vector** in the **weight-height-eyecolor (feature) space**.

2. Set of pixels (HSI): **Features** spectral measurements

Each **pixel** is represented by an **L -dim. vector** (typically $L \sim 200$) in the **spectral (feature) space**.



Clustering definition

Data: A set E of entities.

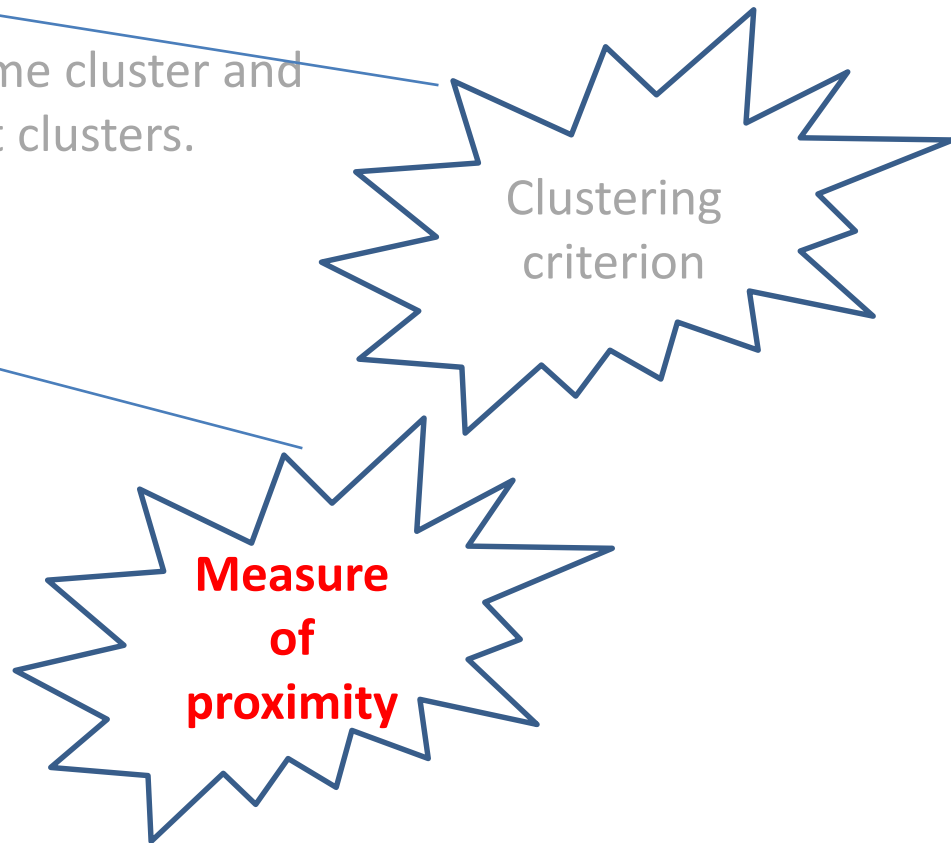
Clustering:

Grouping of the entities into “**sensible**” clusters (groups), so that:

- more similar entities to belong to the same cluster and
- less similar entities to belong to different clusters.

Concepts that need to be clarified:

- Entity
- **Measure of proximity**
- Cluster
- Clustering criterion



Measure of proximity: similarity - dissimilarity

It **quantifies** the **proximity** between **two vectors** (two entities)

- $s(x, y)$: **similarity measure**
(the **closer** the x and y , the **larger** the value of $s(x, y)$ is)
- $d(x, y)$: **dissimilarity measure**
(the **farther** the x and y , the **larger** the value of $d(x, y)$ is)

Entities that **belong** to the **same cluster** exhibit **high similarity** values and **low dissimilarity** values.

Entities that **belong** to **different clusters** exhibit **low similarity** values and **high dissimilarity** values.

Examples:

Dissimilarity measures: (a) Euclidean distance, (b) Manhattan distance etc.

Similarity measures: inner product etc.

NOTE: Measures of **similarity** (**dissimilarity**) are also **defined** between **(a) a vector and a set** and **(b) two sets**.

Clustering definition

Data: A set E of entities.

Clustering:

Grouping of the entities into “sensible” **clusters** (groups), so that:

- more similar entities to belong to the same cluster and
- less similar entities to belong to different clusters.

Concepts that need to be clarified:

- Entity
- Measure of proximity
- **Cluster**
- Clustering criterion



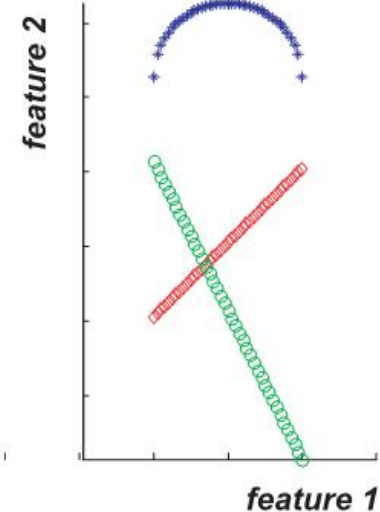
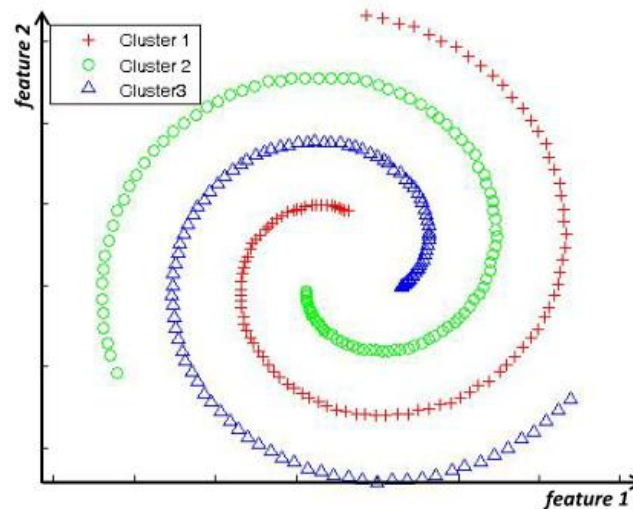
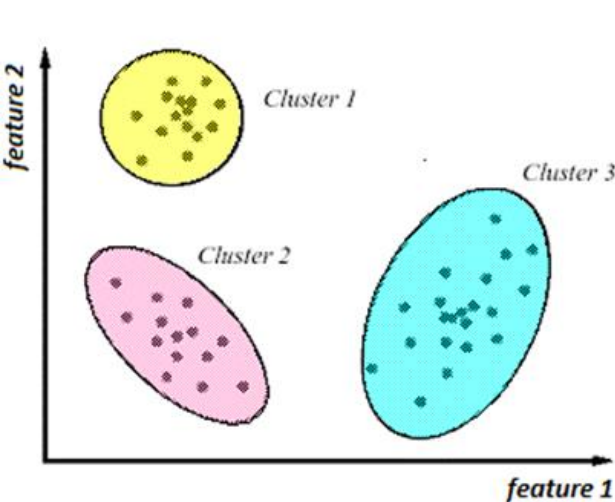
Clustering
criterion

Measure
of
proximity

Cluster definition

Cluster definition remarks

- ✓ There is **no rigorous** definition for the **cluster**.
- ✓ However, we usually have in mind an **aggregation of points** around :
 - a **specific point** in the feature space (usually modeled by a **normal** distribution).
 - a **manifold** (e.g. a hyperplane, a hypersphere etc) in the feature space.



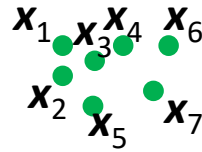
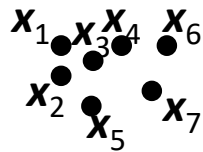
In this sense:

The process of **clustering** aims at the **identification** of **aggregations of points** in an l -dim. space.

Cluster definition - representation

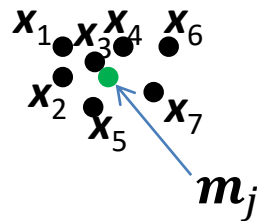
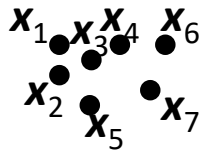
Cluster representation

- ✓ via all its points (**non-parametric** representation)



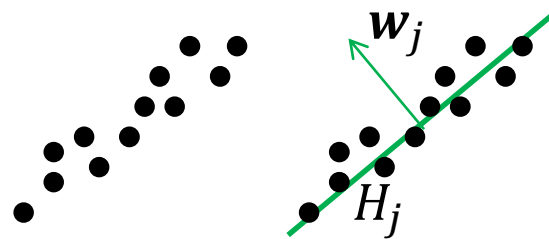
$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$

- ✓ via a set of parameters (**parametric** representation)



Parameters:

$$\mathbf{m}_j = [m_{j1}, m_{j2}, \dots, m_{jl}]^T$$



$$H_j: \mathbf{w}_j^T \mathbf{x}_i + w_{j0} = 0$$

Parameters:

$$\mathbf{w}_j = [w_{j1}, w_{j2}, \dots, w_{jl}]^T, w_{j0}$$

Cluster definition - representation

Cluster representation (cont.)

✓ **ONLY** For the **parametric** representation:

In general, a cluster C_j may be represented by:

k -dim. linear manifold ($k < l$) ⁽¹⁾	Compact set in k -dim. lin. manifold
point (0-dim.) l parameters	point
line (1-dim.)	line segment $2l$ parameters
plane (2-dim.)	polygon
hyperplane (k -dim, $k < l$) $(l + 1)$ parameters (for $k = l - 1$)	Polyhedron

✓ θ_j : The **vector** containing the **parameters** describing cluster C_j .

Clustering definition

Data: A set E of entities.

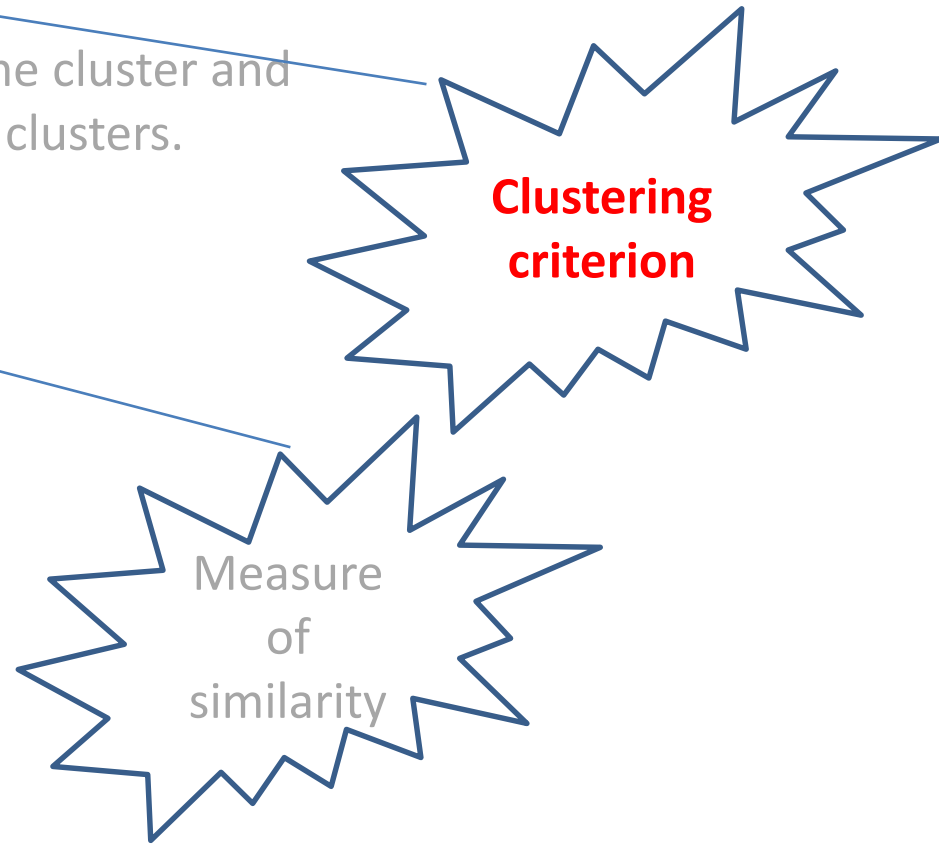
Clustering:

Grouping of the entities into “**sensible**” clusters (groups), so that:

- more similar entities to belong to the same cluster and
- less similar entities to belong to different clusters.

Concepts that need to be clarified:

- Entity
- Measure of similarity
- Cluster
- **Clustering criterion**



Clustering criterion

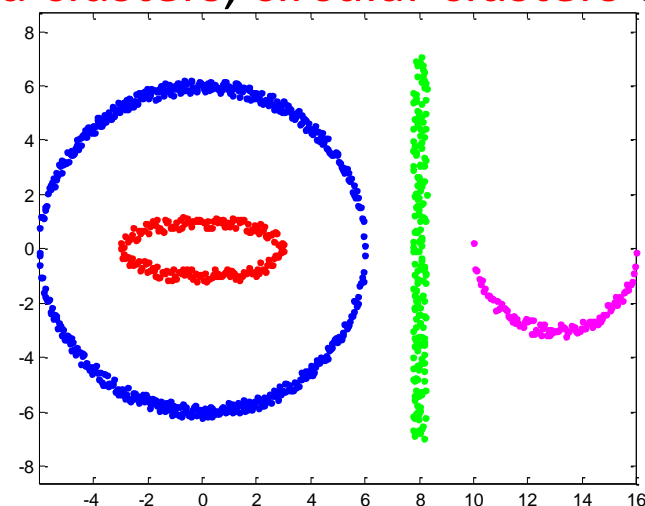
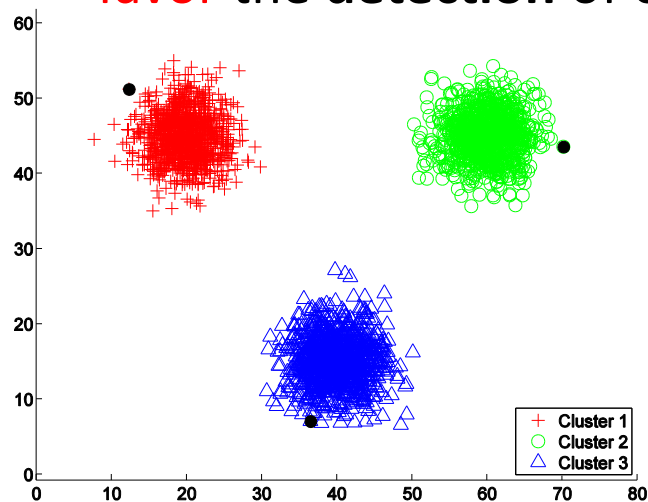
It is **highly responsible** for the **kind/quality of clustering** that will result.

It gives meaning to the adjective “**sensible**”.

It is **expressed** via a **cost function** or a **set of rules**.

The **type of criterion** that will be selected should take into account the **expected type of clusters** formed in the data set.

For example, **other criteria** favor the **detection** of **compact clusters** while **other** favor the **detection** of e.g. **elongated clusters, circular clusters** etc.



Clustering definition

Input: A set E of entities.

Clustering:

Grouping of the entities into “sensible” clusters (groups), so that:

- more similar entities to belong to the same cluster and
- less similar entities to belong to different clusters.

Concepts that need to be clarified:

- Entity
- Measure of proximity
- Cluster
- Clustering criterion



Clustering
criterion

Measure
of
proximity

Clustering: An ill-posed problem

Consider the sequence of real numbers

$$1, 4, 9, 16, 25, \dots$$

What is the next number in the sequence?

36?

Why?

Because implicitly has been assumed the law $a_n = n^2$.

Thus, $a_6 = 6^2 = 36$.

However, e.g., the law $b_n = n^2 + (n-1) \cdot \dots \cdot (n-5)$ also produces the sequence 1, 4, 9, 16, 25, ...

BUT, according to this law, it is

$$b_6 = 6^2 + (6-1) \cdot (6-2) \cdot (6-3) \cdot (6-4) \cdot (6-5) = 156 !!$$

Which of the two is the **correct** answer?

Actually, the **selection of the law** is **subjective**!!!!

Clustering: An ill-posed problem

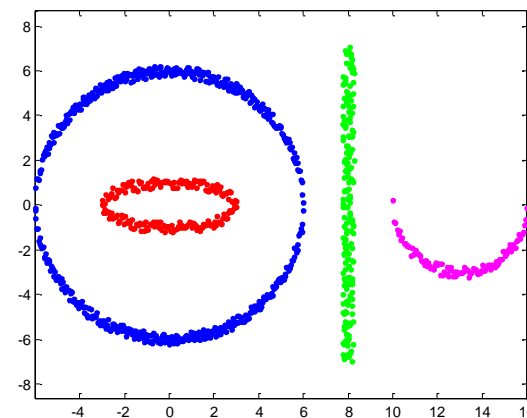
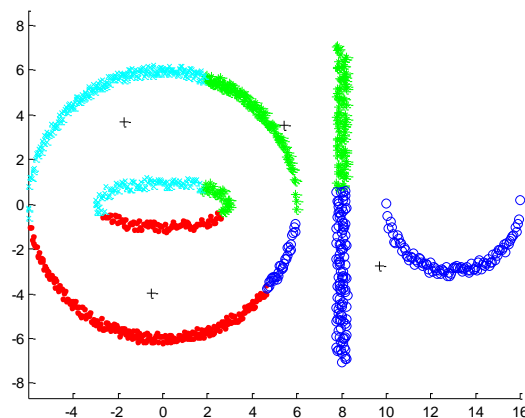
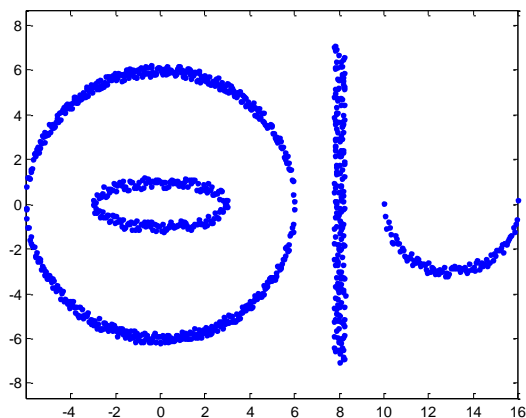
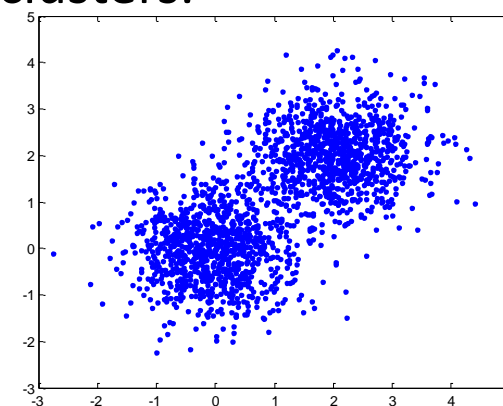
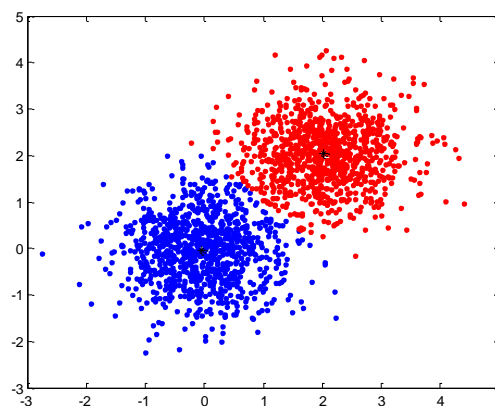
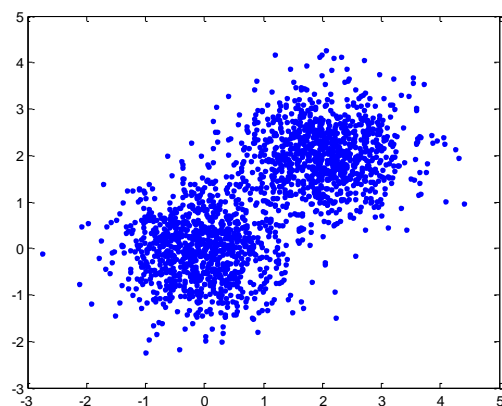
Switching to the clustering problem:

Clustering according to

Data set

(a) a criterion that favors compact clusters.

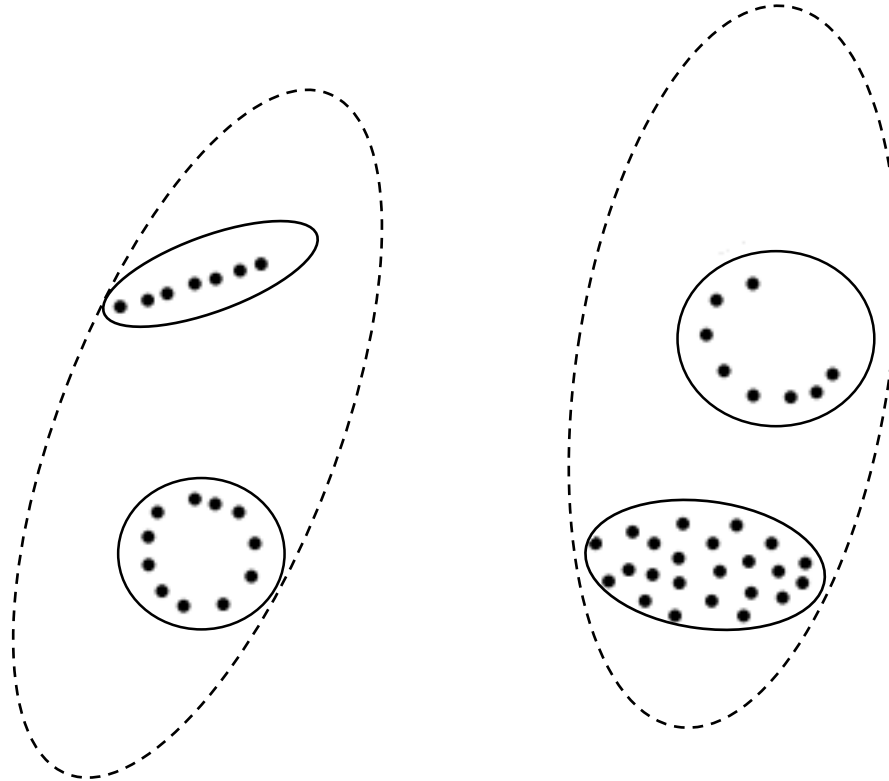
(b) a criterion that favors various-shaped clusters.



Clustering: An ill-posed problem

Another kind of **subjectivity**:

How many clusters are below?



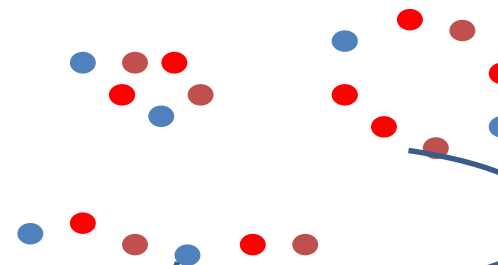
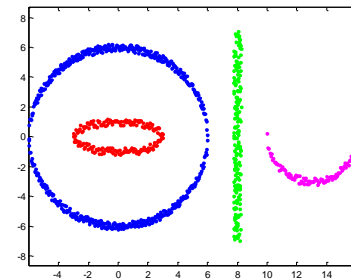
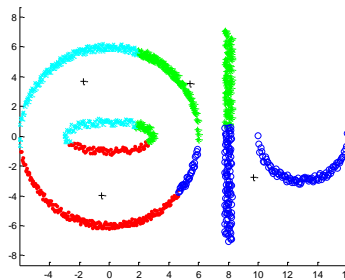
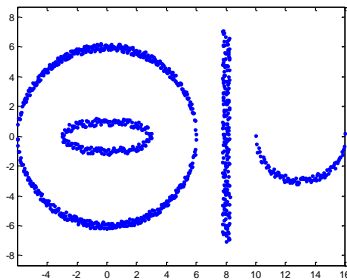
4 or **2** ???

Clustering validation - interpretation

Question: If there exists so much degree of **subjectivity** what is the usefulness of clustering after all?

Answer: The use of the “**clustering tool**” should be **done with care**.

Thus, after the **validation** of the **results**



Clustering validation - interpretation

The resulting clustering need to be **interpreted** by an **expert in the field of application**.

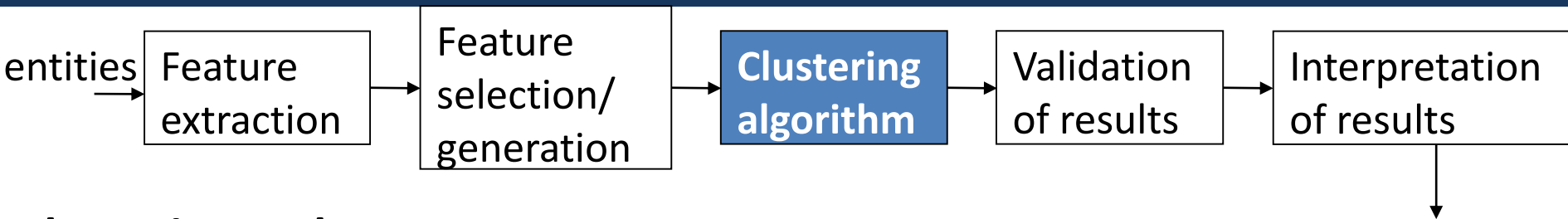
Example:

- ❑ Consider a **set of patients** that have been infected by the **same disease** and they follow the **same treatment**.
- ❑ The **patients** are **clustered** to **groups** according to the reactions to the treatment.
- ❑ Only **specialized** doctors can **identify** correctly the resulting **clusters**.

Thus,

- a cluster may contain patients with e.g., low blood pressure, low fat that exhibit reaction A,
- another cluster may consists e.g., of aged patients with high levels of insulin which exhibit reaction B, etc.

Clustering stages



Clustering task stages

Feature extraction – Selection/generation: Information rich features-**Parsimony**

Proximity Measure adoption: This quantifies the term **similar** or **dissimilar**.

Clustering Criterion adoption: This consists of a cost function or some type of rules.

Clustering Algorithm: This consists of the set of **steps** followed to reveal the clustering structure of the data set under study, based on the adopted **similarity measure** and the adopted **clustering criterion**.

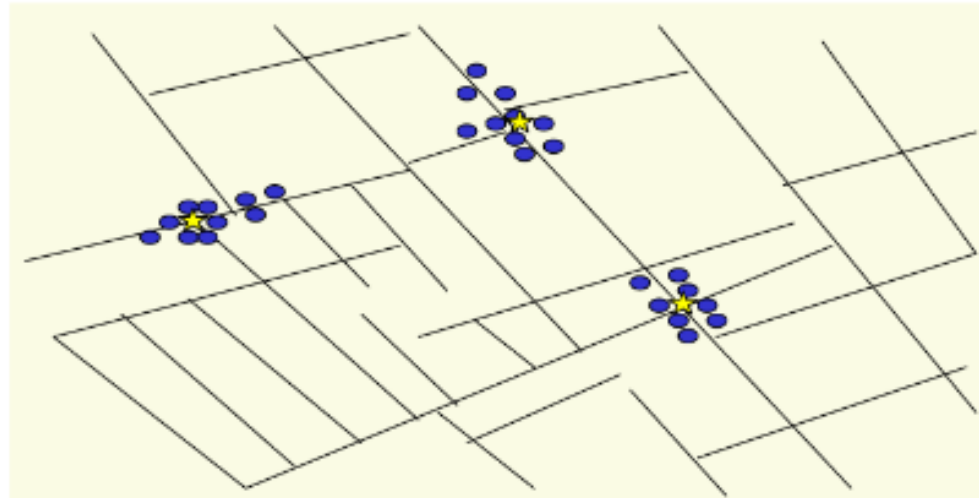
Validation of the results.

Interpretation of the results from an expert.

Clustering: A historic example(*)

Dr John Snow plotted the **location** of **cholera deaths** on a **map** during an outbreak at London in the 1850s

The **locations** were **clustered** around certain intersections where there were **polluted wells!!!**



Questions:

- Which are the **entities** and how they are **represented**?
- Which **dissimilarity measure** could be used?
- What is the **form** of the **resulted clusters**?
- What kind of clusters should be able to reveal the adopted **clustering criterion**?

Clustering: Application areas

Application areas:

- Engineering
- Bioinformatics
- Social Sciences
- Medicine
- Data and Web Mining
- Zoology
- Archaeology
- . . .

Clustering: Application areas

What can we do with clustering:

- Data reduction:

- Represent all **entities** in a certain cluster C by a **set of properties** that are **shared** by the majority of the **entities in C** .

- Hypothesis generation:

- Consider a **set of companies**, each one represented by its **size**, its degree of **activities abroad**, its ability to **complete successfully research projects**.
- After performing clustering, a **cluster** results, whose majority of entities are **large companies** with a high degree of **activities abroad**.
- This suggests the **hypothesis** that “**large companies have activities abroad**”.

- Hypothesis testing:

- Verify the **hypothesis** that “**large companies have activities abroad**”.
- Consider a **set of companies**, each one represented by its **size**, its degree of **activities abroad**, its ability to **complete successfully research projects**.
- After performing clustering, **if** a **cluster** results, whose majority of entities are **large companies** with a high degree of **activities abroad**, the hypothesis is verified.

Clustering: Application areas

What can we do with clustering (cont.):

- Prediction based on groups:

Example (*movie recommendation system*):

- Consider a set of movie watchers each one represented by (a) certain “general features” (e.g., age, gender, nationality etc) (b) its degree of preference to each one out of a set of movies.
- Cluster the movie watchers (taking into account both the “general features” and the preferences) and identify the resulting clusters (e.g., male teenagers prefer action movies, females below 12 years old prefer movies with princesses etc).
- If a new watcher asks for a movie, the system may ask its “general features” and propose movies based on the previously identified clusters.

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(A) Hard Clustering: Each point belongs exclusively to a single cluster.

Let $X = \{x_1, x_2, \dots, x_N\}$

An m -clustering R of X , is defined as the partition of X into m sets (clusters), C_1, C_2, \dots, C_m , (that is, $R = \{C_1, C_2, \dots, C_m\}$) so that

$$C_i \neq \emptyset, i = 1, \dots, m$$

$$\bigcup_{i=1}^m C_i = X$$

$$C_i \cap C_j = \emptyset, i \neq j, i, j = 1, 2, \dots, m$$

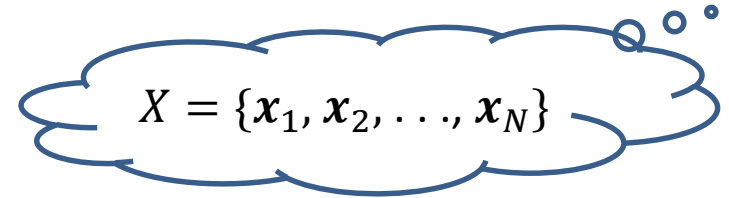
In addition, data in C_i are more similar to each other and less similar to the data in the rest of the clusters.

Quantifying the terms similar-dissimilar depends on the types of clusters that are expected to underlie the structure of X .

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(A) Hard Clustering: Each point belongs exclusively to a single cluster.



An **alternative formulation**:

A hard clustering of X into m clusters is characterized by m functions, each one corresponding to a cluster.

$$u_j: \mathbf{x} \rightarrow \{0,1\}, \quad j = 1, \dots, m$$

$$\sum_{j=1}^m u_j(\mathbf{x}_i) = 1, \quad i = 1, \dots, N$$

$$0 < \sum_{i=1}^N u_j(\mathbf{x}_i) < N, \quad j = 1, 2, \dots, m$$

$u_j(\mathbf{x}), j = 1, \dots, m$: **Membership functions.**

Thus, each \mathbf{x}_i belongs exclusively to the j -th cluster if $u_j(\mathbf{x}_i) = 1$.

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(A) Hard Clustering (cont.):

Remarks:

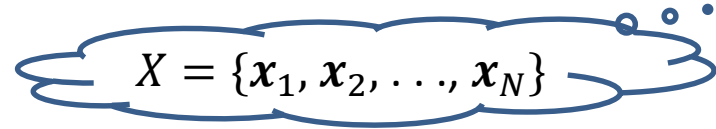
- If $u_j(\mathbf{x}_i) = 1 \Rightarrow \mathbf{x}_i$ *belongs* to the j -th cluster.
- If $u_j(\mathbf{x}_i) = 0 \Rightarrow \mathbf{x}_i$ *does not belong* to the j -th cluster.
- Consider a specific \mathbf{x}_i . A value of $u_j(\mathbf{x}_i)$ equal to 1, for the j -th cluster implies values equal to 0, for the remaining clusters.

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(A1) Probabilistic Clustering: Each point belongs exclusively to a single cluster.

- However, we are **not certain** to **which cluster** a **data point belongs**.
- This uncertainty/ignorance of ours is modeled via a **probabilistic** framework.


$$X = \{x_1, x_2, \dots, x_N\}$$

Let $P_j(x_i)$ be the **probability** that x_i belongs to cluster C_j .

The clustering is defined via the following relations:

$$\sum_{j=1}^m P_j(x_i) = 1, \quad i = 1, \dots, N$$

$$0 < \sum_{i=1}^N P_j(x_i) < N, \quad j = 1, 2, \dots, m$$

$P_j(x_i)$ quantifies our **degree of certainty** that x_i belongs to the j -th cluster.

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(A1) Probabilistic Clustering (cont.):

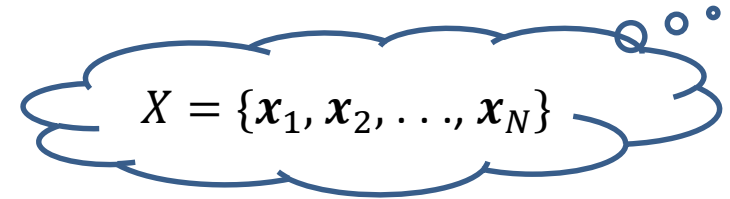
Remarks:

- Values of $P_j(\mathbf{x}_i)$ close to **1** \Rightarrow **High degree of confidence** that \mathbf{x}_i belongs to the **j -th cluster**.
- Values of $P_j(\mathbf{x}_i)$ close to **0** \Rightarrow **High degree of confidence** that \mathbf{x}_i **does not** belong to the **j -th cluster**.
- Values of $P_j(\mathbf{x}_i)$ close to $\frac{1}{m}$ for all clusters, $j = 1, \dots, m \Rightarrow$ **Low degree of confidence** that \mathbf{x}_i belongs to a specific **cluster**.
- Consider a certain \mathbf{x}_i . A value of $P_j(\mathbf{x}_i)$ close to **1**, for the **j -th cluster** implies values close to **0**, for the **remaining clusters**.

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(B) Fuzzy Clustering: Each point belongs to all clusters up to some degree.



A fuzzy clustering of X into m clusters is characterized by m functions, each one corresponding to a cluster.

$$u_j: \mathbf{x} \rightarrow [0,1], \quad j = 1, \dots, m$$

$$\sum_{j=1}^m u_j(\mathbf{x}_i) = 1, \quad i = 1, \dots, N$$

$$0 < \sum_{i=1}^N u_j(\mathbf{x}_i) < N, \quad j = 1, 2, \dots, m$$

$u_j(\mathbf{x}), j = 1, \dots, m$: Membership functions.

Thus, each \mathbf{x}_i belongs to the j -th cluster up to some degree indicated by the value of $u_j(\mathbf{x}_i)$.

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(B) Fuzzy Clustering (cont.):

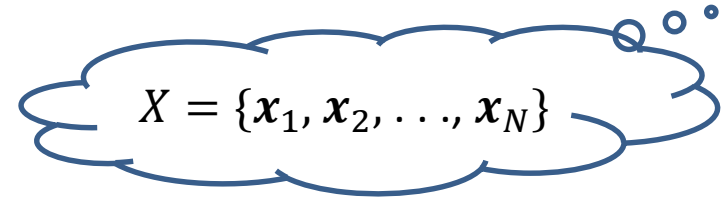
Remarks:

- Values of $u_j(\mathbf{x}_i)$ close to 1 \Rightarrow High *degree of membership* of \mathbf{x}_i to the j -th cluster.
- Values of $u_j(\mathbf{x}_i)$ close to 0 \Rightarrow Low *degree of membership* of \mathbf{x}_i to the j -th cluster.
- Consider a certain \mathbf{x}_i . A value of $u_j(\mathbf{x}_i)$ close to 1, for the j -th cluster implies values close to 0, for the remaining clusters.

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(C) Possibilistic Clustering: Each point is compatible with all clusters up to some “degree of compatibility”.



A possibilistic clustering of X into m clusters is characterized by m functions, each one corresponding to a cluster.

$$u_j: \mathbf{x} \rightarrow (0,1], \quad j = 1, \dots, m$$

$$0 < \sum_{i=1}^N u_j(\mathbf{x}_i) < N, \quad j = 1, 2, \dots, m$$

Remark: The degree of compatibility of a \mathbf{x}_i with the j -th cluster is independent from the degrees of compatibility of \mathbf{x}_i with all the remaining clusters.