Clustering algorithms

Konstantinos Koutroumbas

Unit 1

– General concepts
– Problem formulation

koutroum@noa.gr

Bibliography

Course grades

70%: Final exams (obligatory)

20%: Project (obligatory)

20%: Homeworks

Programming language

MATLAB

Bibliography

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.

Clustering criterion

Concepts that need to be clarified:

- Entity
- Measure of proximity
- Cluster
- Clustering criterion

Measure of proximity

Clustering definition

Input: A set *E* of entities.

<u>Clustering:</u>
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.

Clustering criterion

Concepts that need to be clarified:

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

Measure of proximity

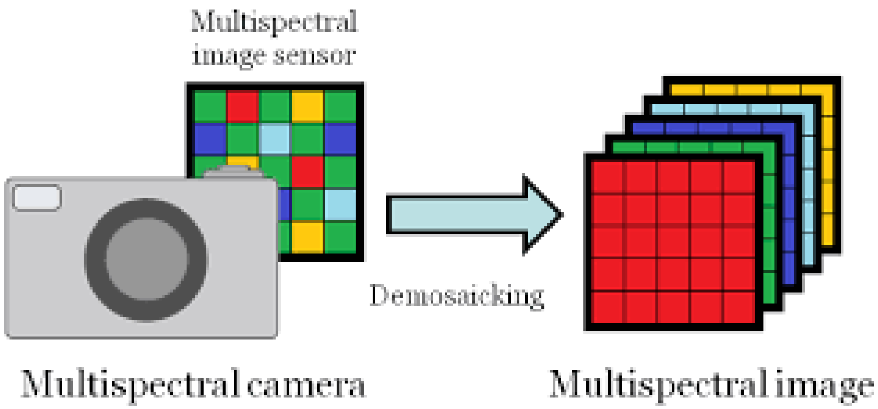
Entities

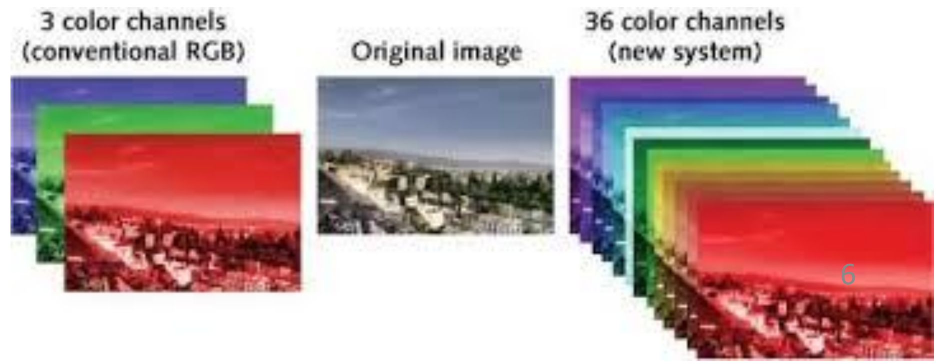
Entities

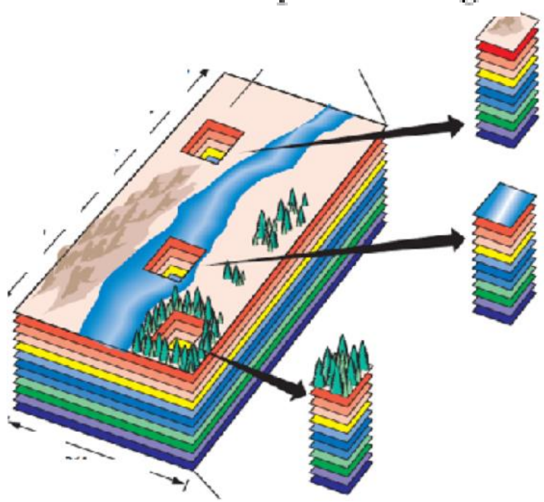
They are **application-dependent**.

They can be e.g.,

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

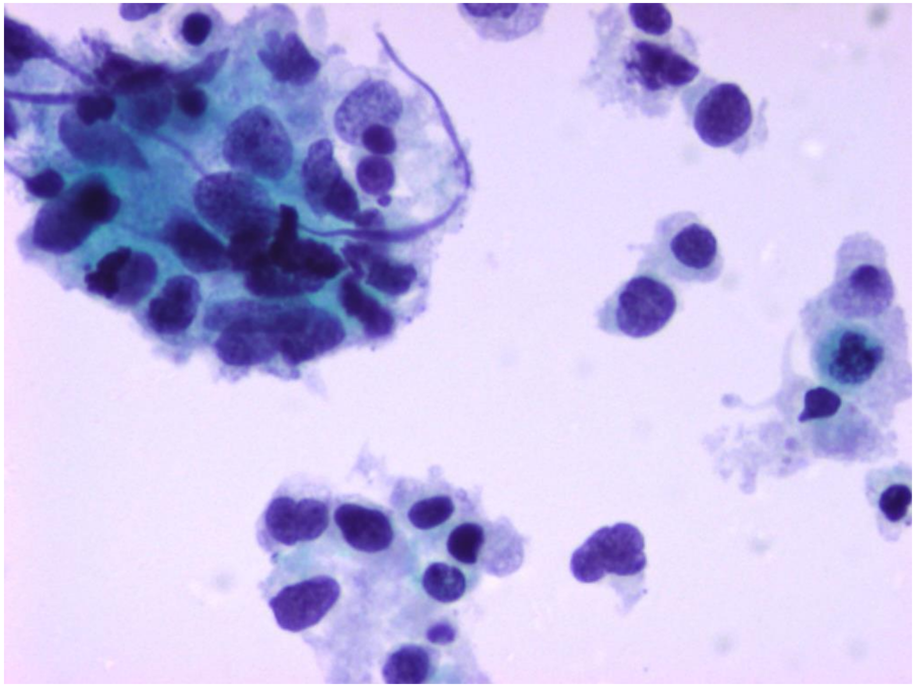


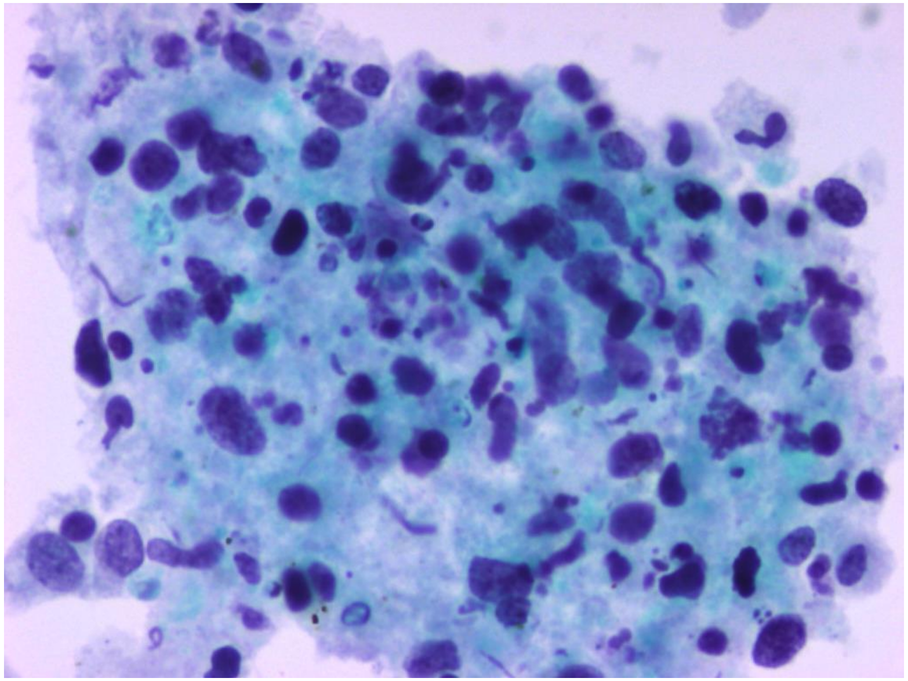




Entities

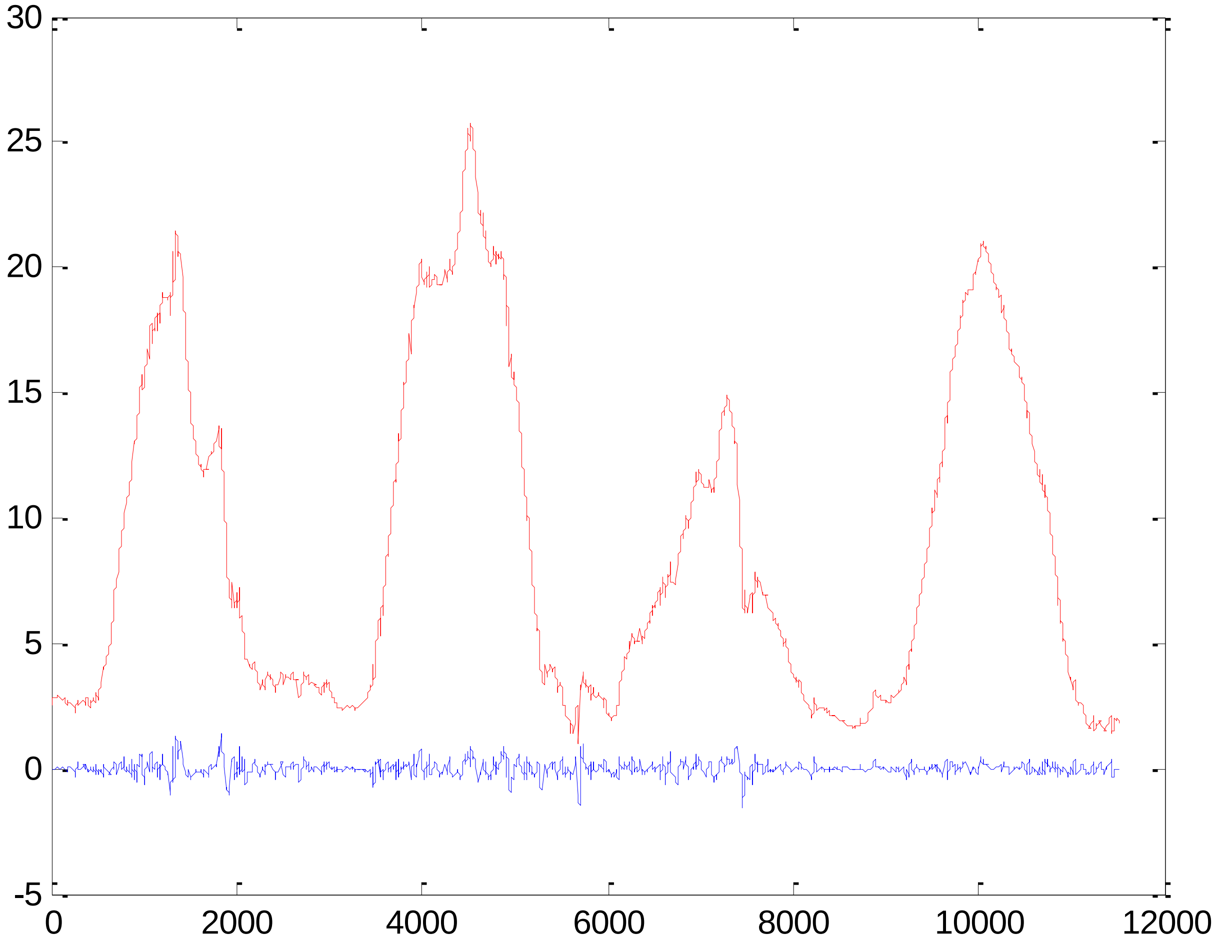
(B) Human cells or tissues.





Entities

(C) Time series.



Entities

(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 a

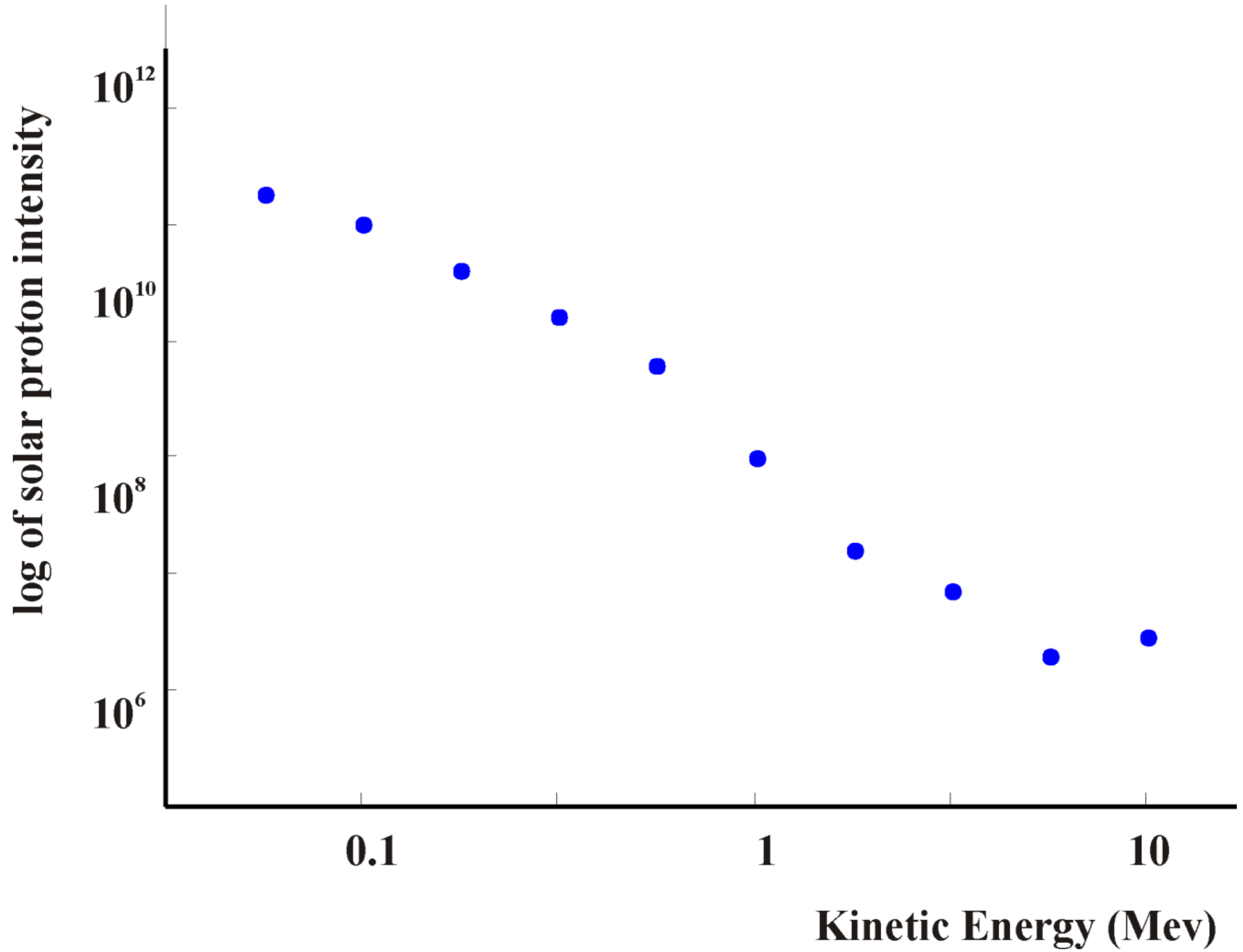
b c d e f g h i j k l m n o p

q r s t u v w x y z & 1 2 3 4

5 6 7 8 9 0 ($ £ . , ! ?)

Entities

(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).

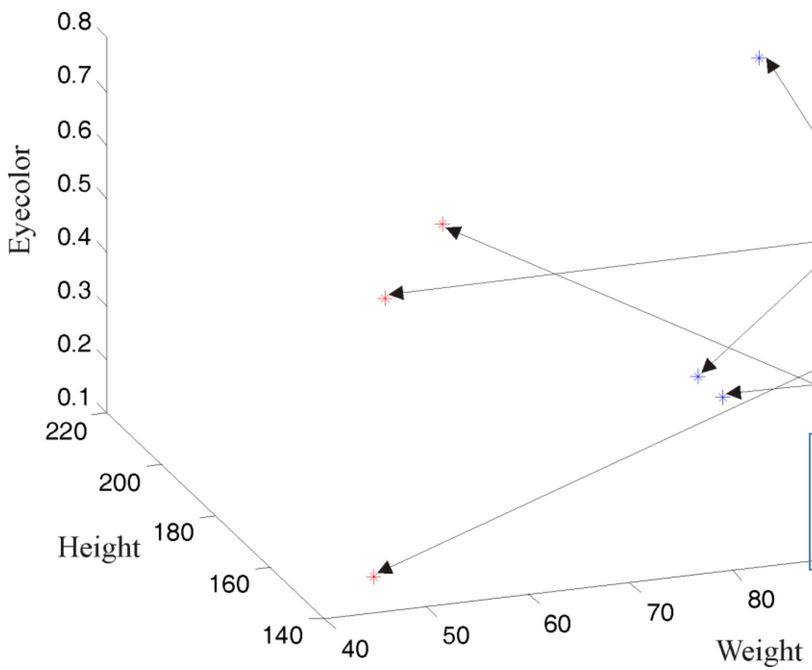
$$i\text{-}th\ entity \Leftrightarrow \boldsymbol{x}_i = [x_{i1}, x_{i2},, \cdots, x_{il},]^T$$

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

AN IMPORTANT ISSUE: ENTITY REPRESENTATION (cont.)

Examples:

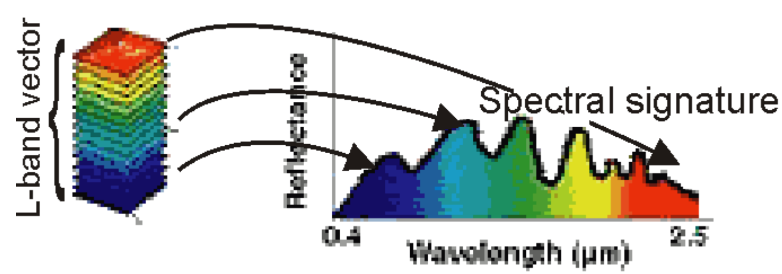
1. Set of athletes: Features weight, height, eyecolor



	Weight	Height	Eyecolor
Athlete1	90	190	.2
Athlete2	50	155	.6
Athlete3	100	205	.7
Athlete4	48	152	.1
Athlete5	95	200	.1
Athlete6	57	160	.7

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~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.

Clustering criterion

Concepts that need to be clarified:

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

Measure of proximity

Measure of proximity: similarity - dissimilarity

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

- $s(\mathbf{x},\mathbf{y})$: similarity measure

  (the closer the $\mathbf{x}$ and $\mathbf{y}$, the larger the value of $s(\mathbf{x},\mathbf{y})$ is)
- $d(\mathbf{x},\mathbf{y})$: dissimilarity measure

  (the farther the $\mathbf{x}$ and $\mathbf{y}$, the larger the value of $d(\mathbf{x},\mathbf{y})$ 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.

Clustering criterion

Concepts that need to be clarified:

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

Measure of proximity

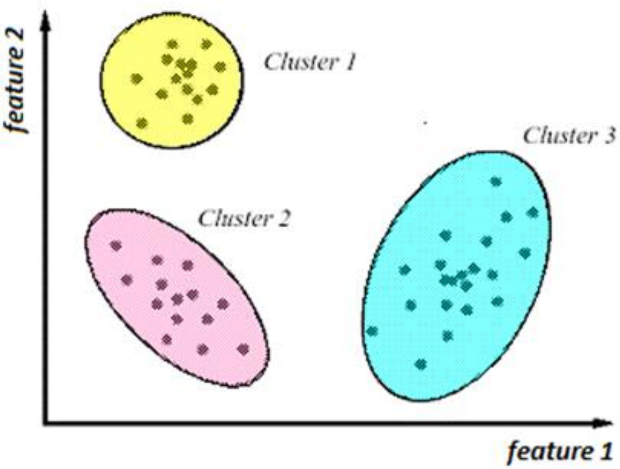
Cluster definition

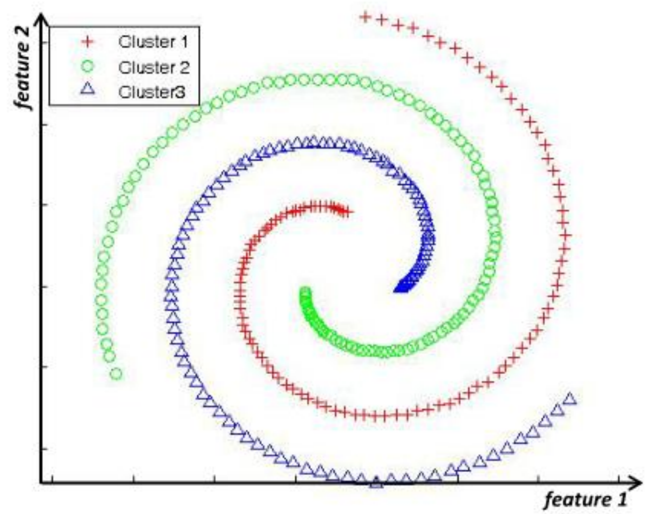
Cluster definition remarks

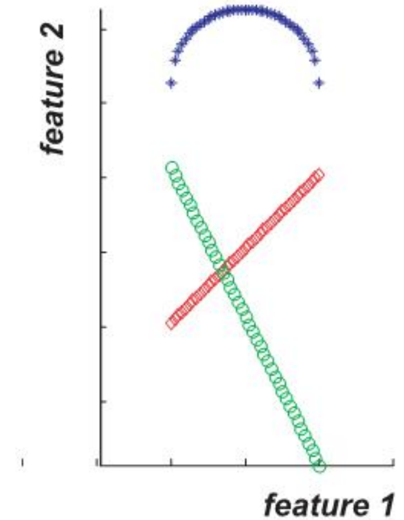
✓ There is no rigorous definition for the cluster.

✓ 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.







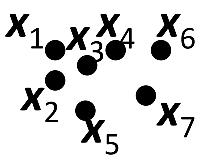
In this sense:

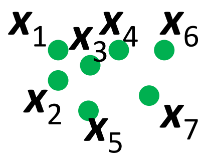
The process of clustering aims at the identification of aggregations of points in an l-dim. space.

Cluster definition - representation

Cluster representation

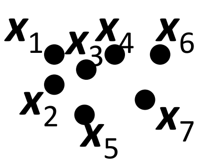
✓ via all its points (non-parametric representation)

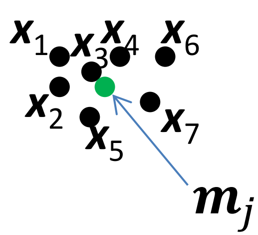


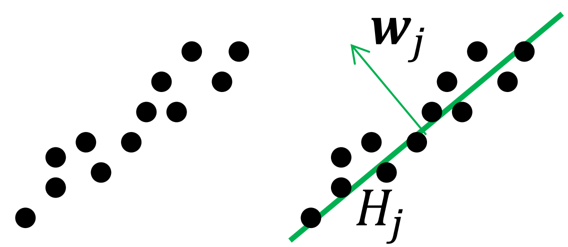


$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6, \boldsymbol{x}_7\}$

✓ via a set of parameters (parametric representation)







Parameters:

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

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

Parameters:

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

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

✓ $\boldsymbol{\theta}_j$: 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.

Clustering criterion

Concepts that need to be clarified:

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

Measure of similarity

Clustering criterion

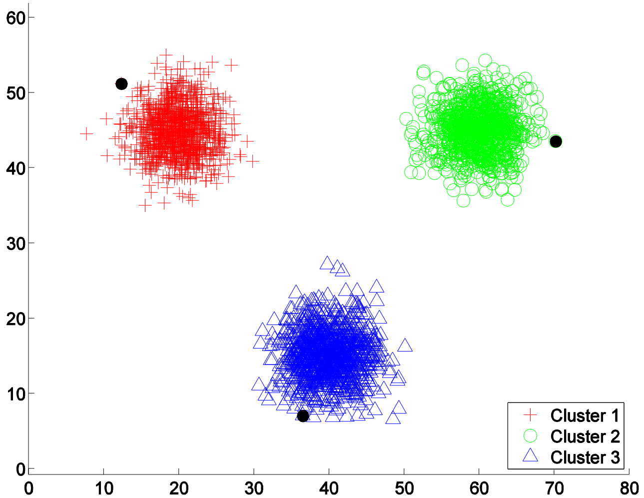
It is highly responsible for the kind of clustering that will result.

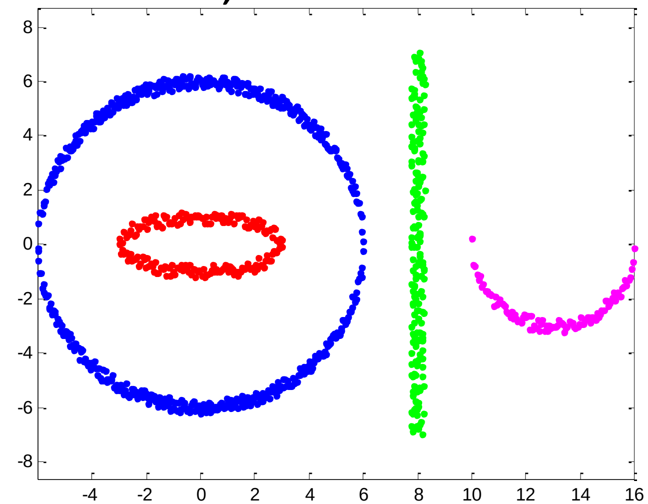
It gives meaning to the adjective "**sensible**".

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

The type of criterion that will be selected *<u>should take into account</u>* 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.

Clustering criterion

Measure of proximity

Concepts that need to be clarified:

- Entity
- Measure of proximity
- Cluster
- Clustering criterion

Clustering: An ill-posed problem

Consider the sequence of real numbers

1, 4, 9, 16, 25, …

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 $b_n=n^2+(n-1)*(n-2)*(n-3)*(n-4)*(n-5)$ also produces the sequence 1, 4, 9, 16, 25, …

BUT, according to this law, it is $b_6=6^2+(6-1)*(6-2)*(6-3)*(6-4)*(6-5)=156$.

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

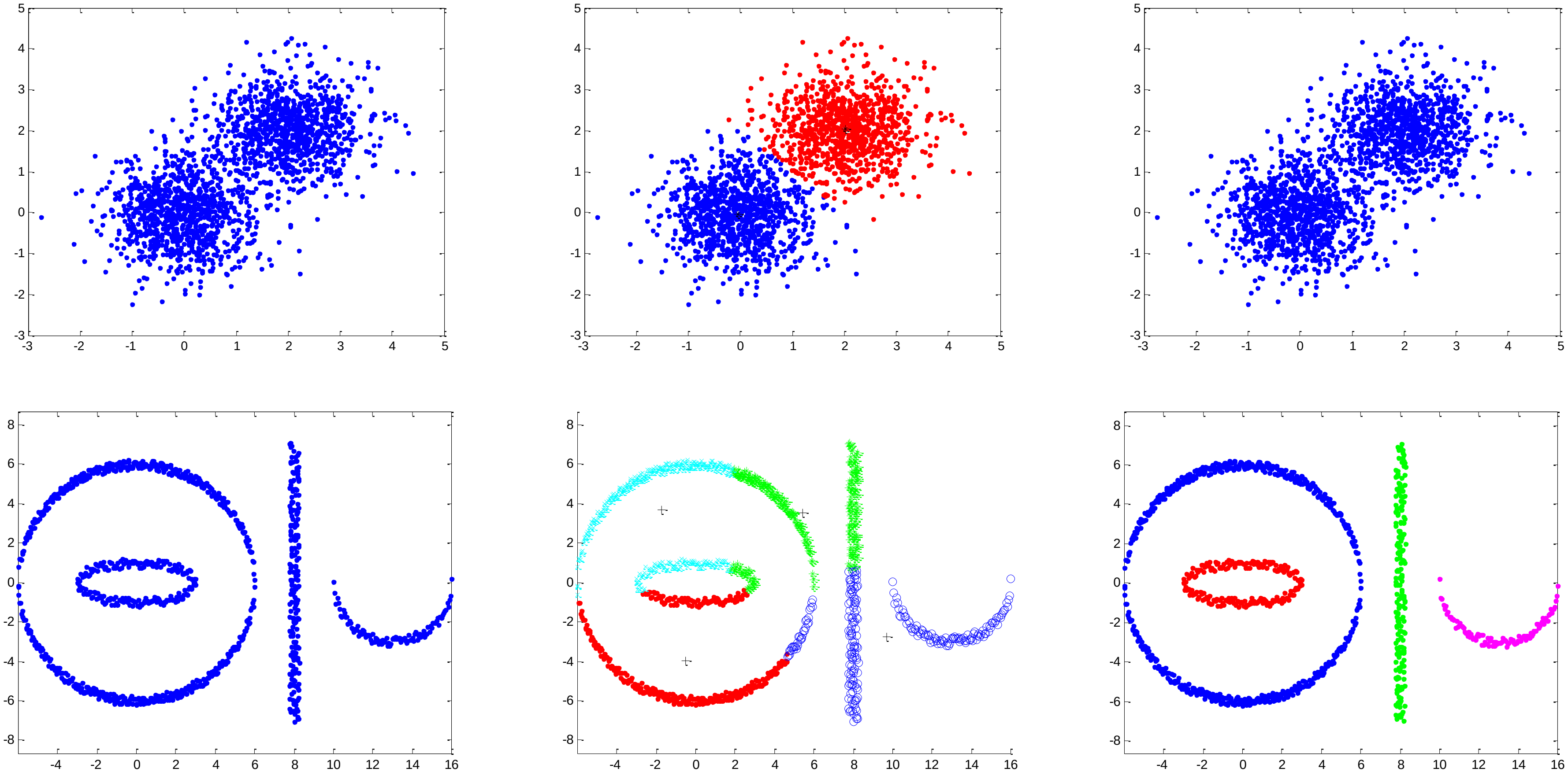
Actually, the selection of the law is subjective!!!!!

Clustering: An ill-posed problem

Switching to the clustering problem:

Clustering according to

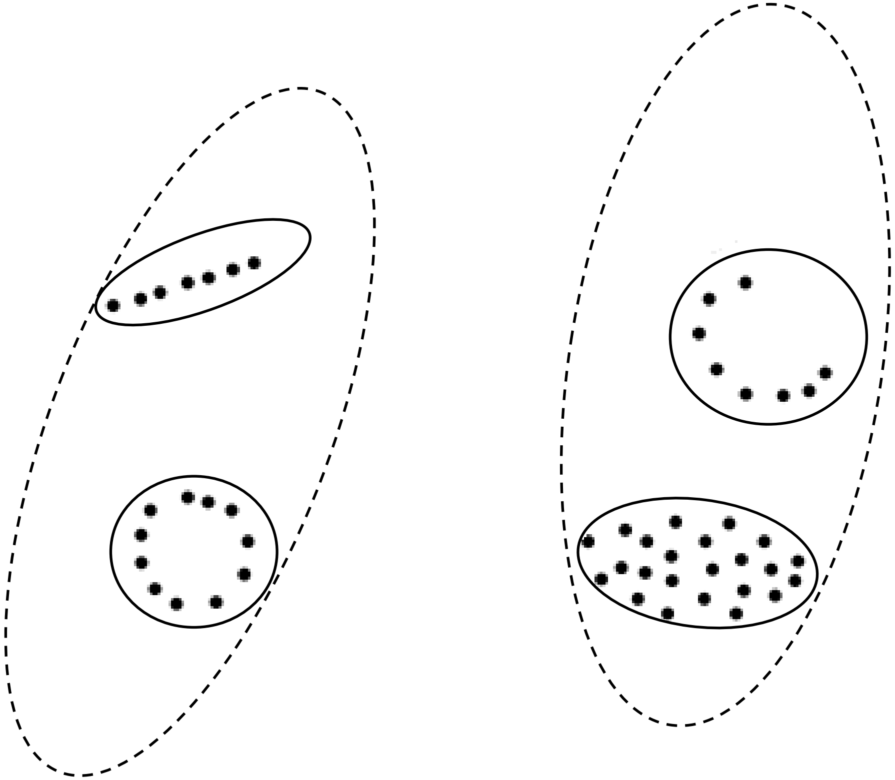
Data set	**(a)** a criterion that favors **compact** clusters.	**(b)** a criterion that favors **various-shaped** clusters.



Clustering: An ill-posed problem

Another kind of subjectivity:

How many clusters are below?



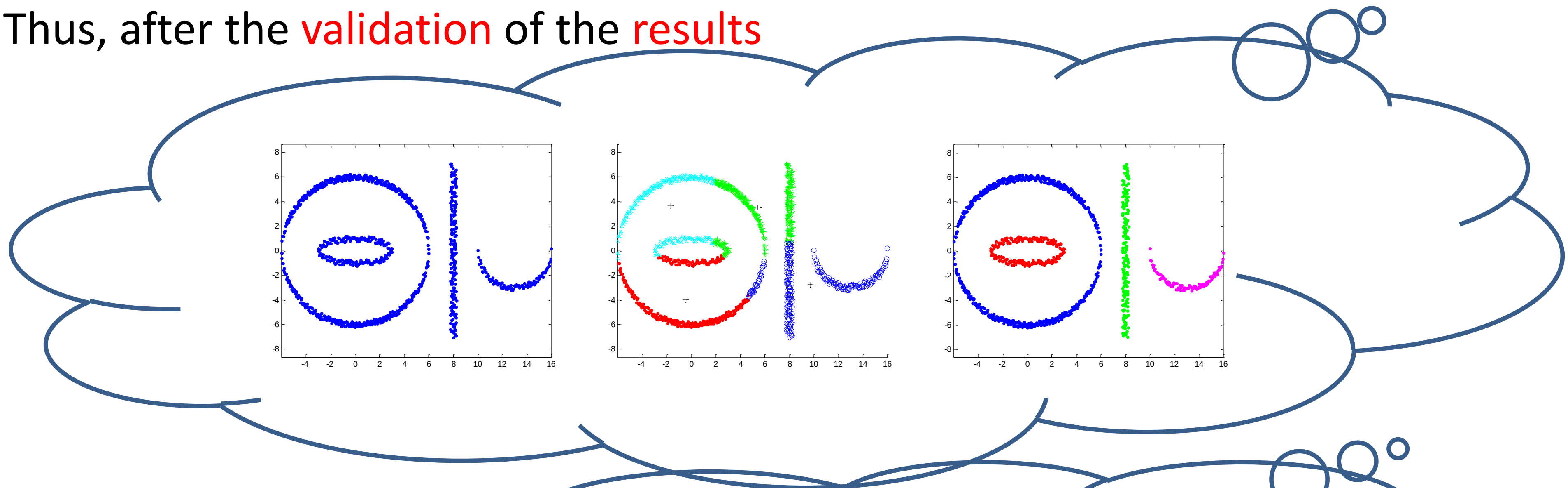
4 or **2** ???

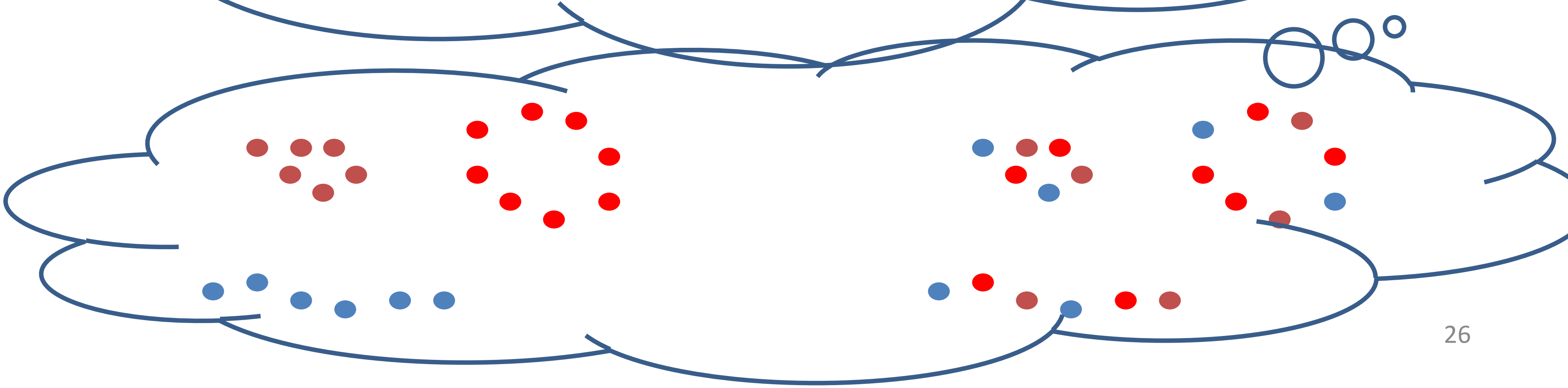
Clustering validation - interpretation

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 will be interpreted by an expert in the field of application.

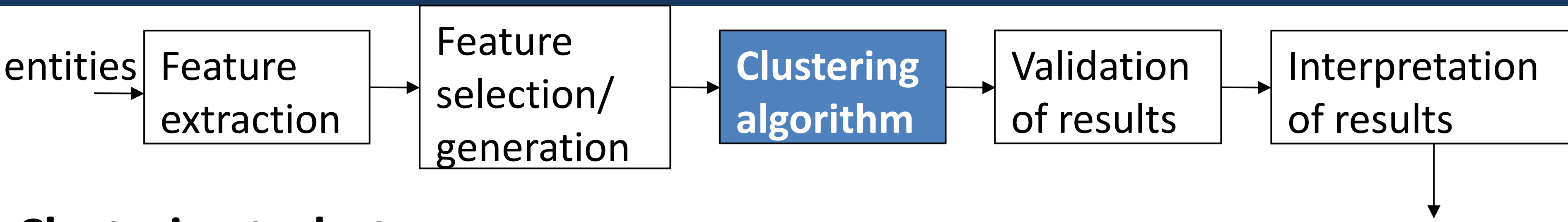
Example:

- ❑ 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.

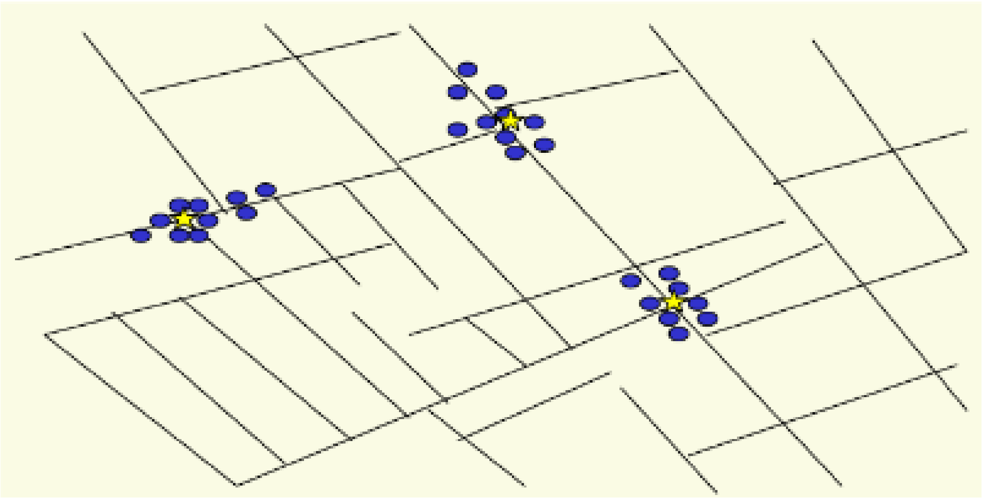
Validation of the results.

Interpretation of the results.

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?

(*) Nina Mishra HP Labs

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 – Relation between data vectors and clusters

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

Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$

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

$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$

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

$$u_j : \boldsymbol{x} \rightarrow \{0,1\}, \qquad j = 1, \dots, m$$

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

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

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

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

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(A) Hard Clustering (cont.):

Remarks:

- If $u_j(\boldsymbol{x}_i) = 1 \Rightarrow \boldsymbol{x}_i$ *belongs* to the *j*-th cluster.
- If $u_j(\boldsymbol{x}_i) = 0 \Rightarrow \boldsymbol{x}_i$ *does not belong* to the *j*-th cluster.
- Consider a specific $\boldsymbol{x}_i$. A value of $u_j(\boldsymbol{x}_i)$ 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) <u>Probabilistic Clustering:</u> 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.

$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$

Let $P_j(\boldsymbol{x}_i)$ be the **probability** that $\boldsymbol{x}_i$ **belongs** to cluster C_j.

The clustering is defined via the following relations:

$$\sum_{j=1}^{m} P_j(\boldsymbol{x}_i) = 1, \qquad i = 1, \dots, N$$

$$0 < \sum_{i=1}^{N} P_j(\boldsymbol{x}_i) < N, \qquad j = 1, 2, \dots, m$$

$P_j(\boldsymbol{x}_i)$ quantifies our **degree of certainty** that $\boldsymbol{x}_i$ **belongs** to the j-th cluster.

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(A1) Probabilistic Clustering (cont.):

Remarks:

- Values of $P_j(\boldsymbol{x}_i)$ close to 1 $\Rightarrow$ High *degree of confidence* that $\boldsymbol{x}_i$ belongs to the *j*-th cluster.

- Values of $P_j(\boldsymbol{x}_i)$ close to 0 $\Rightarrow$ High *degree of confidence* that $\boldsymbol{x}_i$ **does not** belong to the *j*-th cluster.

- Values of $P_j(\boldsymbol{x}_i)$ close to $\frac{1}{m}$ for all clusters, $j = 1, \dots, m \Rightarrow$ Low *degree of confidence* that $\boldsymbol{x}_i$ belongs to a specific cluster.

- Consider a certain $\boldsymbol{x}_i$. A value of $P_j(\boldsymbol{x}_i)$ close to 1, for the *j*-th cluster implies values close to 0, for the remaining clusters.

Clustering Definitions – Relation between data vectors and clusters

(B) <u>Fuzzy Clustering:</u> Each point **belongs** to all clusters up to some degree.

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$$

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

$$u_j : \boldsymbol{x} \rightarrow [0,1], \qquad j = 1, \ldots, m$$

$$\sum_{j=1}^{m} u_j(\boldsymbol{x}_i) = 1, \qquad i = 1, \ldots, N$$

$$0 < \sum_{i=1}^{N} u_j(\boldsymbol{x}_i) < N, \qquad j = 1, 2, \ldots, m$$

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

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

Clustering definitions

Clustering Definitions – Relation between data vectors and clusters

(B) Fuzzy Clustering (cont.):

Remarks:

- Values of $u_j(\boldsymbol{x}_i)$ close to 1 $\Rightarrow$ High *degree of membership* of $\boldsymbol{x}_i$ to the *j*-th cluster.

- Values of $u_j(\boldsymbol{x}_i)$ close to 0 $\Rightarrow$ Low *degree of membership* of $\boldsymbol{x}_i$ to the *j*-th cluster.

- Consider a certain $\boldsymbol{x}_i$. A value of $u_j(\boldsymbol{x}_i)$ 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) <u>Possibilistic Clustering:</u> Each point is compatible with all clusters up to some "degree of compatibility".

$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$

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

$$u_j : \boldsymbol{x} \rightarrow (0,1], \qquad j = 1, \ldots, m$$

$$0 < \sum_{i=1}^{N} u_j(\boldsymbol{x}_i) < N, \qquad j = 1,2, \ldots, m$$

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

Types of features

With respect to their **domain**

Continuous (the domain is a continuous subset of $\Re$).
Discrete (the domain is a finite discrete set).
Binary or *dichotomous* (the domain consists of two possible values).

With respect to the **relative significance of the values they take**

Nominal (the values code states, e.g., the gender of a person).
Ordinal (the values are meaningfully ordered, e.g., the rating of the services of a hotel (poor, good, very good, excellent)).
Interval-scaled (the difference of two values is meaningful but their ratio is meaningless, e.g., temperature in ^{o}C).
Ratio-scaled (the ratio of two values is meaningful, e.g., weight).

Proximity measures: Definitions

(A) Between vectors

(1) Dissimilarity measure (between vectors of X) is a function

$$d: X \times X \rightarrow \Re$$

with the following properties

1. $\exists d_0 \in \Re: 0 \leq d_0 \leq d(\boldsymbol{x}, \boldsymbol{y}) < +\infty, \forall \boldsymbol{x}, \boldsymbol{y} \in X$
2. $d(\boldsymbol{x}, \boldsymbol{x}) = d_0, \forall \boldsymbol{x} \in X$
3. $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}), \forall \boldsymbol{x}, \boldsymbol{y} \in X$

Examples: Euclidean distance, Manhattan distance etc.

If in addition:

4. $d(\boldsymbol{x}, \boldsymbol{y}) = d_0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
5. $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}), \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$ (triangular inequality)

d is called **metric** dissimilarity measure.

Proximity measures : Definitions

(A) Between vectors

(2) Similarity measure (between vectors of X) is a function

$$s: X \times X \rightarrow \Re$$

with the following properties

1. $\exists s_0 \in \Re: 0 \leq s(\boldsymbol{x}, \boldsymbol{y}) \leq s_0 < +\infty, \forall \boldsymbol{x}, \boldsymbol{y} \in X$
2. $s(\boldsymbol{x}, \boldsymbol{x}) = s_0, \forall \boldsymbol{x} \in X$
3. $s(\boldsymbol{x}, \boldsymbol{y}) = s(\boldsymbol{y}, \boldsymbol{x}), \forall \boldsymbol{x}, \boldsymbol{y} \in X$

If in addition:

4. $s(\boldsymbol{x}, \boldsymbol{y}) = s_0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
5. $\frac{1}{s(\boldsymbol{x},\boldsymbol{z})} \leq \frac{1}{s(\boldsymbol{x},\boldsymbol{y})} + \frac{1}{s(\boldsymbol{y},\boldsymbol{z})}, \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$

s is called **metric** similarity measure.

Examples: inner product, Tanimoto distance etc.

NOTE:
Similarity measures and dissimilarity measures are also referred as **proximity measures**.

NOTATION:
- Similarity measure: *s*
  dissimilarity measure: *d*
- **proximity measures:** $\wp$

Proximity measures : Definitions

Exercise:

Consider the case where the elements of X are **scalars**.

Which of the following is

(a) a dissimilarity measure,

(b) a **metric** dissimilarity measure?

1. $d_1(x, y) = |x - y|$
2. $d_2(x, y) = |x^2 - y^2|$
3. $d_3(x, y) = \cos(x - y)$
4. $d_4(x, y) = \sin(|x - y|)$

Proximity measures: Definitions

(B) Between sets

Let $D_i \subset X, \quad i = 1, \dots, k$, and $U = \{D_1, \dots, D_k\}$.

A **proximity measure** (similarity or dissimilarity) $\wp$ on U is a function

$$\wp: U \times U \to \Re$$

For dissimilarity measure the following properties should hold

1. $\exists d_0 \in \Re: 0 \leq d_0 \leq d(D_i, D_j) < +\infty, \forall D_i, D_j \in X$
2. $d(D_i, D_i) = d_0, \forall D_i \in X$
3. $d(D_i, D_j) = d(D_j, D_i), \forall D_i, D_j \in X$

Question: What is the definition when $\wp$ stands for a similarity measure?

If in addition:

4. $d(D_i, D_j) = d_0 \Leftrightarrow D_i = D_j$
5. $d(D_i, D_k) \leq d(D_i, D_j) + d(D_j, D_k), \forall D_i, D_j, D_k \in X$

d is called **metric** dissimilarity measure.

Proximity functions between a point and a set

Remark: Having in mind that a cluster is actually a set C, a proximity function between a point $\boldsymbol{x}$ and a set C actually **quantifies** the resemblance/relation of $\boldsymbol{x}$ with the cluster C.

Let $X = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ and $\boldsymbol{x} \in X, C \subset X$

Definitions of $\wp(\boldsymbol{x}, C)$:

(a) All points of C **contribute** to the definition of $\wp(\boldsymbol{x}, C)$.

$$d^{ps}{}_{max}(\boldsymbol{x}, C) = max_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y})$$
$$s^{ps}{}_{max}(\boldsymbol{x}, C) = max_{\boldsymbol{y} \in C} s(\boldsymbol{x}, \boldsymbol{y})$$

- Max proximity function

$$\wp^{ps}{}_{max}(\boldsymbol{x}, C) = max_{\boldsymbol{y} \in C} \wp(\boldsymbol{x}, \boldsymbol{y})$$

$$d^{ps}{}_{min}(\boldsymbol{x}, C) = min_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y})$$
$$s^{ps}{}_{min}(\boldsymbol{x}, C) = min_{\boldsymbol{y} \in C} s(\boldsymbol{x}, \boldsymbol{y})$$

- Min proximity function

$$\wp^{ps}{}_{min}(\boldsymbol{x}, C) = min_{\boldsymbol{y} \in C} \wp(\boldsymbol{x}, \boldsymbol{y})$$

$$d^{ps}{}_{avg}(\boldsymbol{x}, C) = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} d(\boldsymbol{x}, \boldsymbol{y})$$
$$s^{ps}{}_{avg}(\boldsymbol{x}, C) = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} s(\boldsymbol{x}, \boldsymbol{y})$$

- Average proximity function

$$\wp^{ps}{}_{avg}(\boldsymbol{x}, C) = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} \wp(\boldsymbol{x}, \boldsymbol{y})$$

n_C is the cardinality of C.

Proximity functions between a point and a set

Definitions of $\wp(\boldsymbol{x}, C)$ (cont.):

(b) A representative of C, r_C, **contributes** to the definition of $\wp(\boldsymbol{x}, C)$.

In this case $\wp(\boldsymbol{x}, C) = \wp(\boldsymbol{x}, r_C)$

Typical **point** representatives are:

- The mean vector

$$\boldsymbol{m}_p = \frac{1}{n_C} \sum_{\boldsymbol{y} \in C} \boldsymbol{y}$$

n_C is the cardinality of C.

- The mean center

$$\boldsymbol{m}_C \in C: \sum_{\boldsymbol{y} \in C} d(\boldsymbol{m}_C, \boldsymbol{y}) \leq \sum_{\boldsymbol{y} \in C} d(\boldsymbol{z}, \boldsymbol{y}), \forall \boldsymbol{z} \in C$$

d: dissimilarity measure.

- The median center

$$\boldsymbol{m}_{med} \in C: med(d(\boldsymbol{m}_{med}, \boldsymbol{y}) | \boldsymbol{y} \in C) \leq med(d(\boldsymbol{z}, \boldsymbol{y}) | \boldsymbol{y} \in C), \forall \boldsymbol{z} \in C$$

NOTE: Other representatives (e.g., hyperplanes, hyperspheres) are useful in certain applications (e.g., object identification using clustering techniques).