

Machine Learning and Data Mining

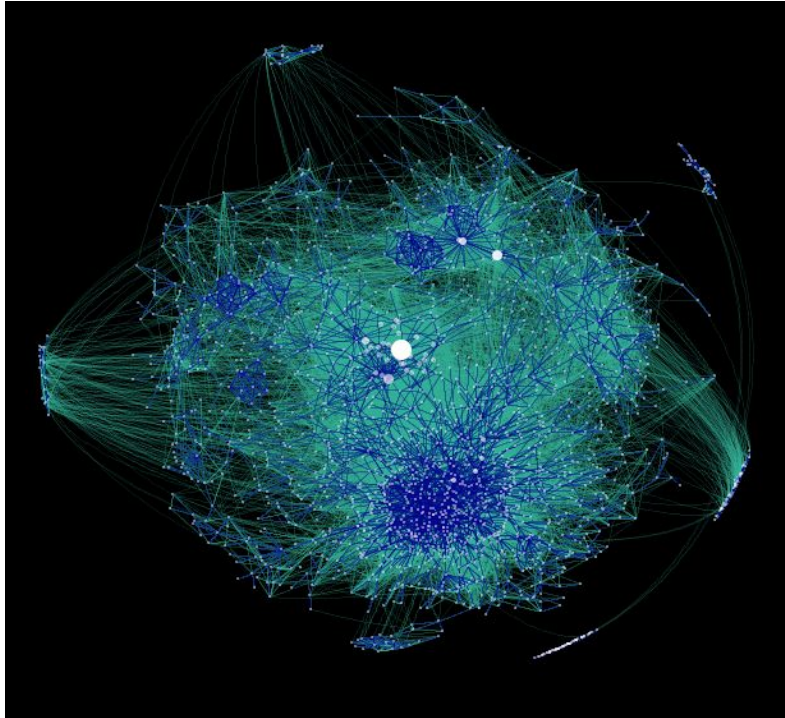
Massinissa Hamidi
Univ. Évry Paris-Saclay

Based on the slides of Farida Zehraoui

Outline

- Introduction to machine learning for computational biology
- Unsupervised learning approaches
 - k-means algorithm
 - Hierarchical clustering
 - Spectral Clustering
- Supervised learning approaches
 - Decision trees
 - Ensemble methods

Data of all kinds!



Blogosphere: representing the links between blogs on the internet

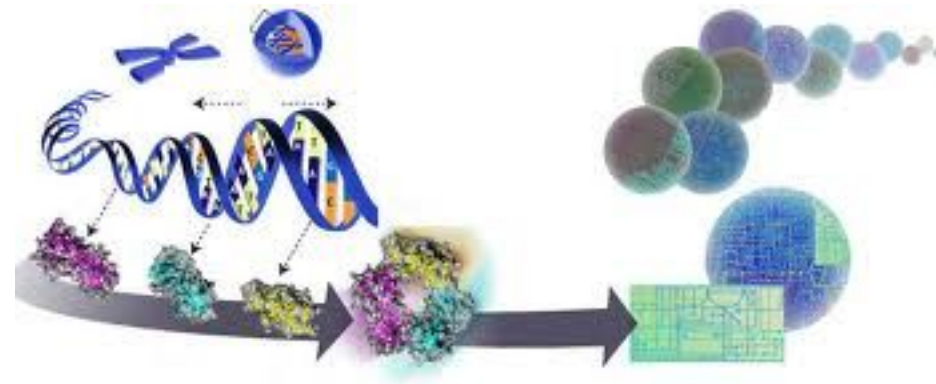
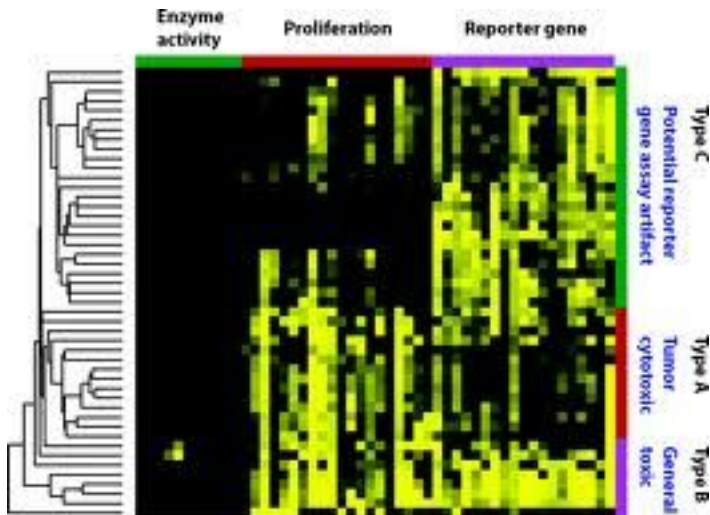
Internet data: social networks, etc...



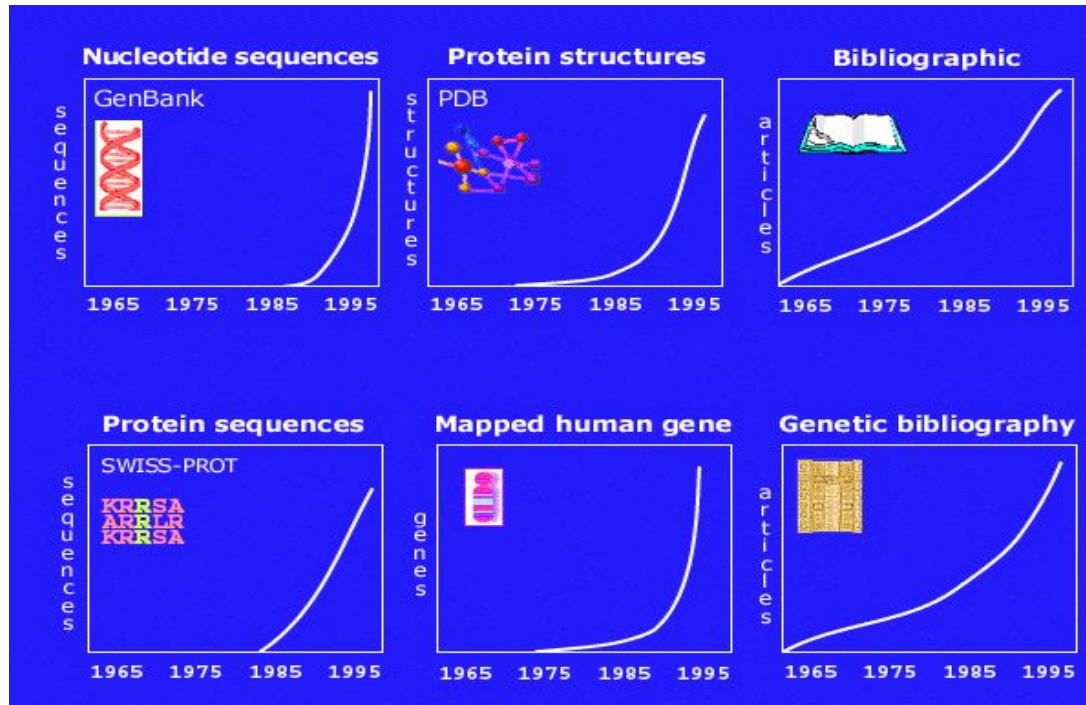
Epidemiologic data

More data:

text in natural language, biological data

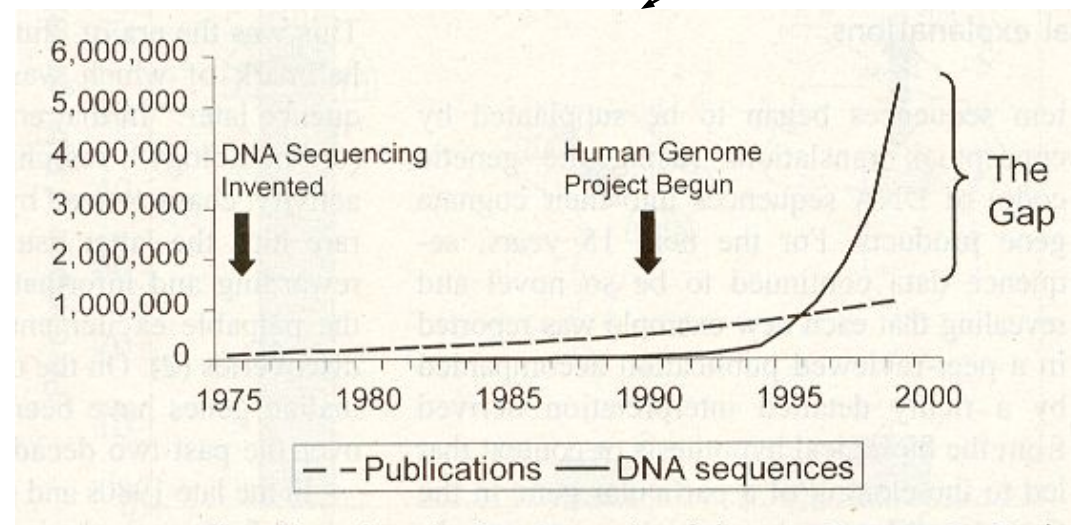


Exponential growth of biological information



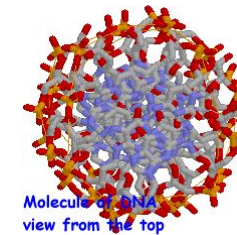
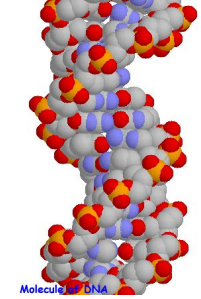
Cumulative increases of published articles in molecular biology and genetics and DNA sequence records in GENBANK

The exponential growth of biological information. These graphs show the unprecedented growth of sequences, structures, and literature over the last fifteen years. The development of efficient storage and management tools has been an important cornerstone in bioinformatics and computational biology.

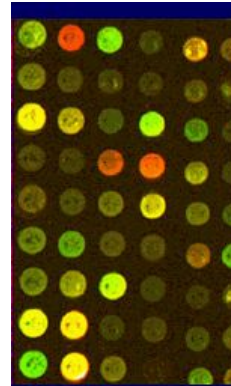
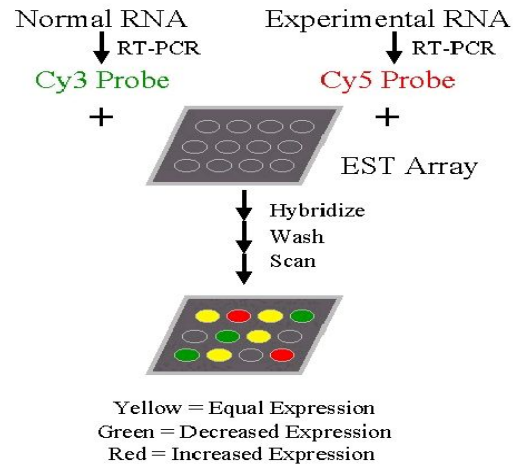


Computational biology

- Data production at different levels: molecules, cells, organs, organisms and populations
- Integration of several data sources: structure and function data, gene expression data, pathway data, clinical data, ...
- Need of prediction of Molecular Function and Structure, phenotypes, ...
- Computational biology: synthesis (simulation tools) and **analysis / prediction** (machine learning approaches)



Microarray processing

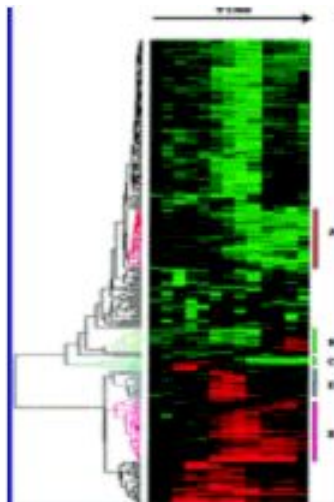


Microarray Scanner

	A	B	C	D	E	F	G
1	YORF	NAME	GWEIGHT	spo0	spo30	spo2	spo5
2	EWEIGHT			1	1	1	1
3	YAL003W	EFB1	1	0.23	-1.79	-1.29	-1.56
4	YAL004W	YAL004W	1	0.41	-0.38	-0.89	-1.06
5	YAL005C	SSA1	1	0.61	-0.07	-1.29	-1.29
6	YAL010C	MDM10	1	0.16	-0.15	-0.76	-1.25



Cluster

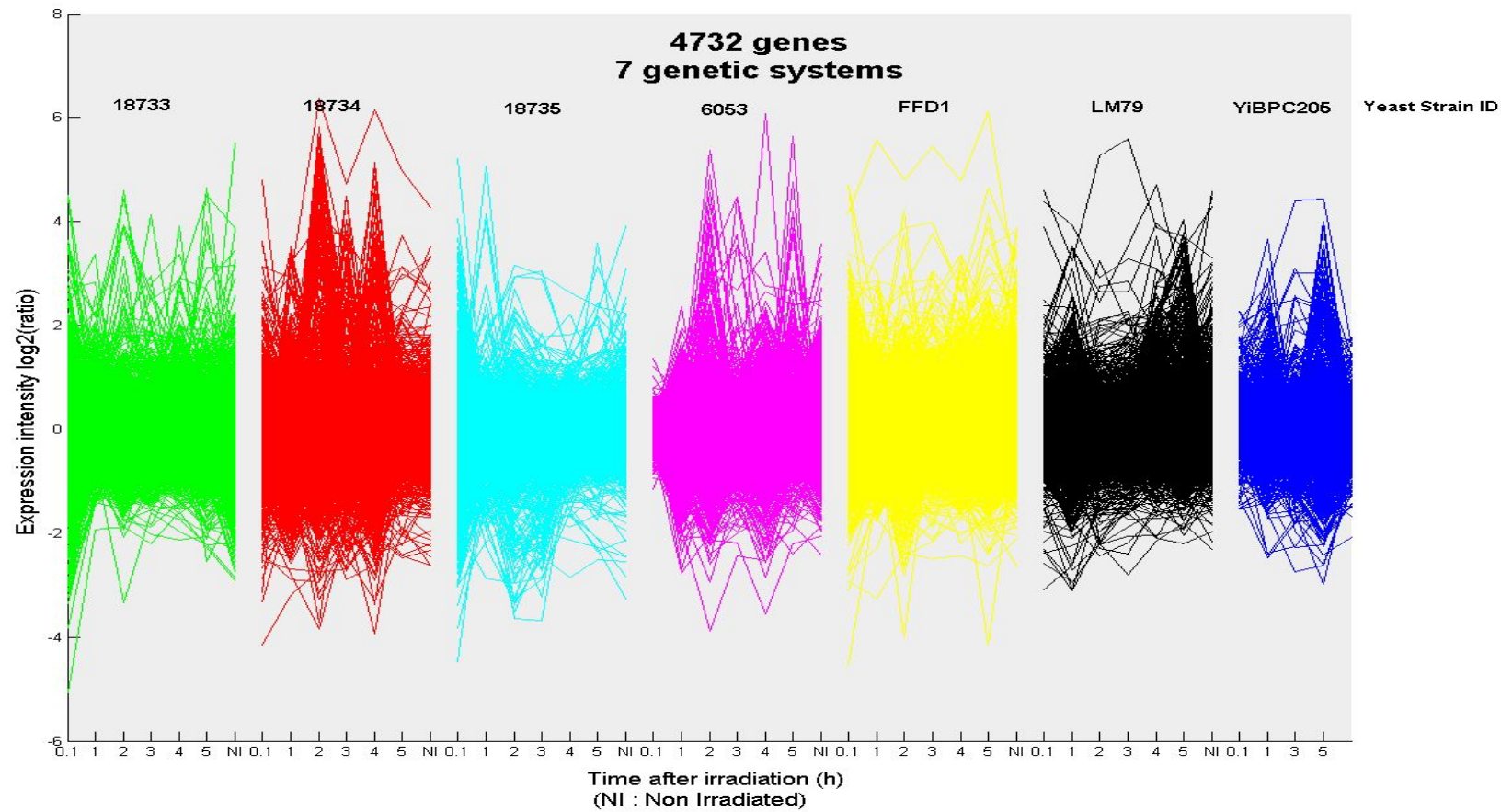


TreeView

From: Shin-Mu Tseng

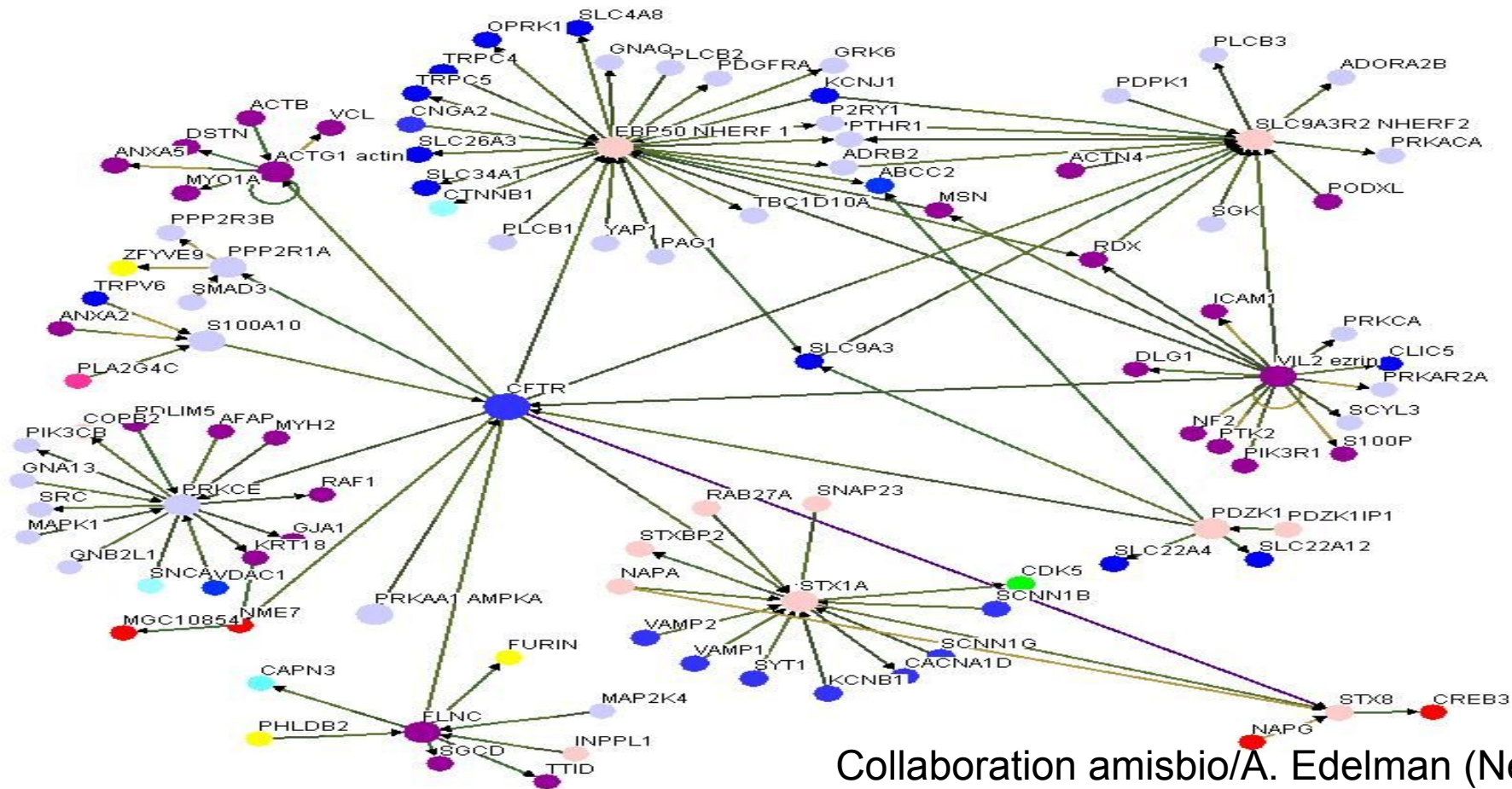
tsengsm@mail.ncku.edu.tw

Gene expression time series



Collaboration amisbio/M. Dutreix (Curie)

Protein-protein interaction network



10

Data everywhere!

Patients data

Poll data

Images

Documents

Astronomy

Environment data (remote sensing or satellites)

Biological data

Données de production, de fabrication

Marketing data

Données scolaires, universitaires

Voice

Econometric data

Logs d'utilisateurs

Internet...

D'autres ?

Données musique, jeux ...

Data analysis

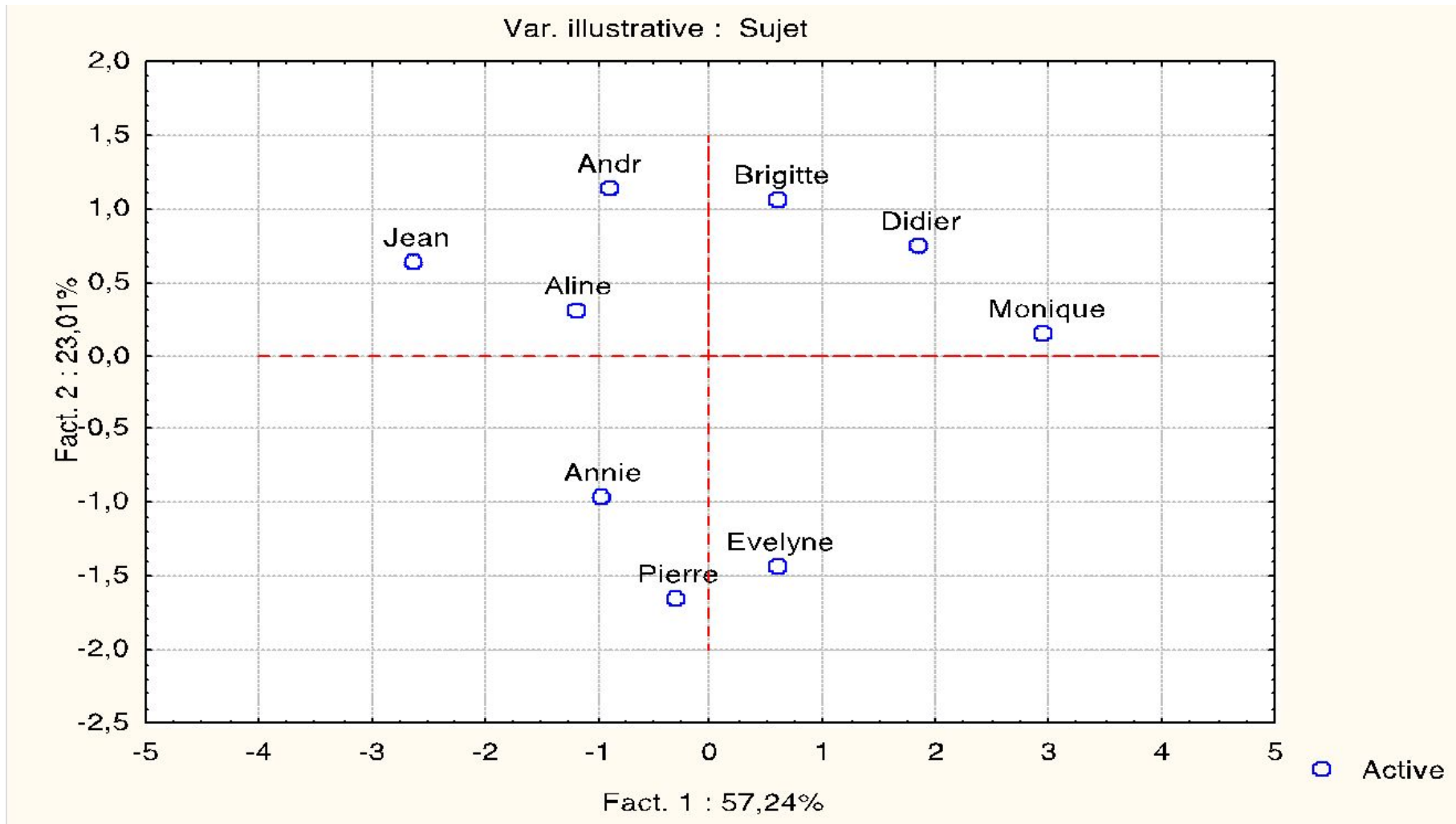
Data visualization: project data into a plan (dimension reduction)

Data classification: identify groups of observations

Analyze correlations

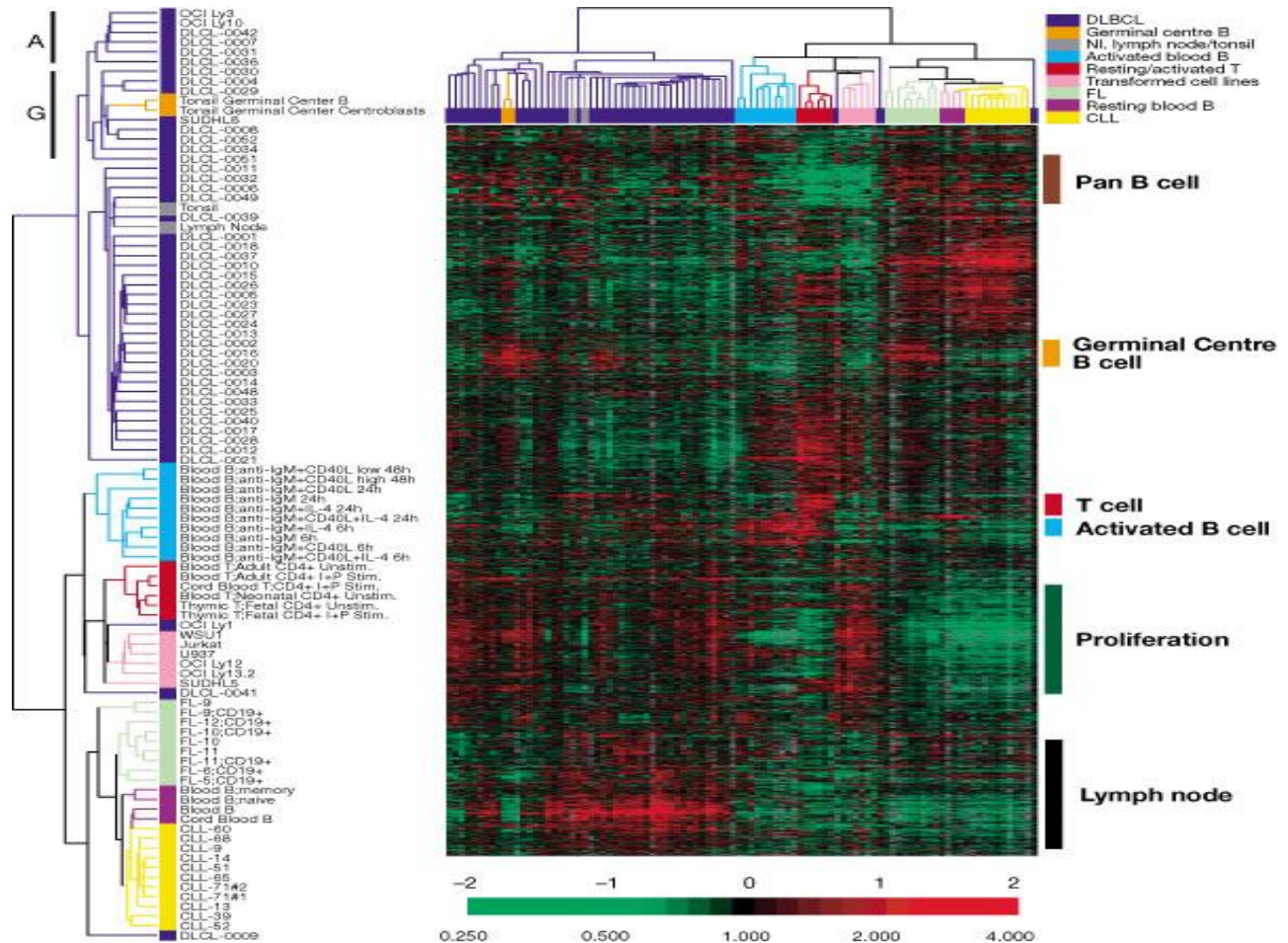
Explain variables

Visualisation of students in a plan (according to their grades)

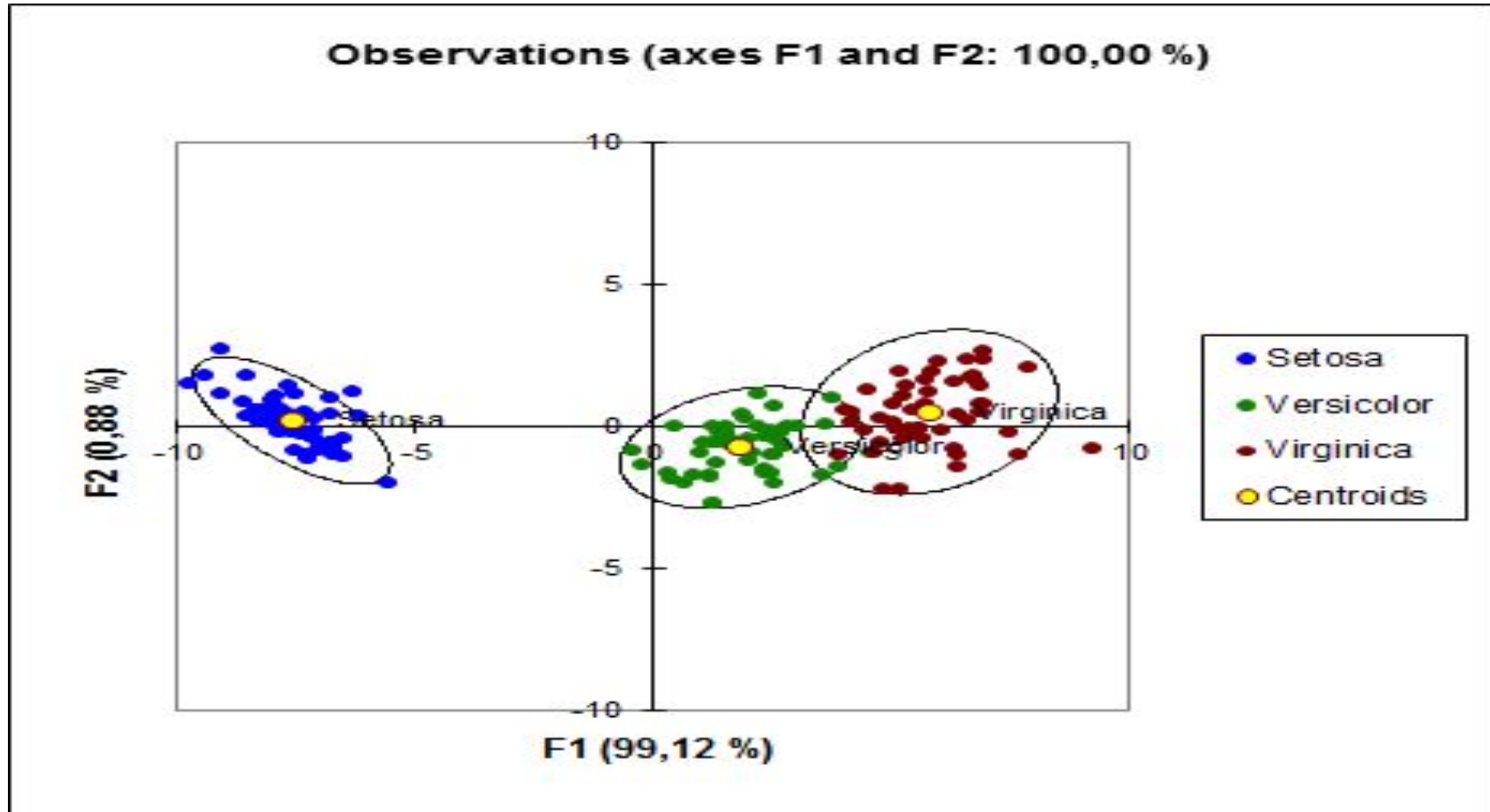


Grouping genes according to their expression

Nature Feb, 2000
Paper by
Allzadeh. A et al
*Distinct types of
diffuse large
B-cell lymphoma
identified by gene
expression
profiling*



Discriminative Analysis



Data

- Data : set of objects described by their attributes
- An attribute (variable) characterizes the objects
 - Examples: color, temperature, etc.

Attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

Attribute types

- Real-value attributes
 - gene expression, ...
- Binary attributes
 - Has Alzheimer's disease (True/False), ...
- Nominal attributes
 - Biological process, ...
- Ordinal attributes
 - expression level (low, high), ...
- Mixed types attributes


Why "Machine Learning"?

- Machine Learning is used when:
 - Observations / data are available
 - Human expertise does not exist (or not sufficient)
 - Solution changes in time
 - ...

"Machine Learning"

Learning models from data

- Data is cheap and abundant (data warehouses, data marts);
- knowledge is expensive and scarce.

 Build a model that is *a good approximation* to the information contained in data.

What is Machine Learning?

Machine Learning: Optimize a performance criterion using data or past experience.

Use two main domains: Statistics / Computer science

□ **Statistics:** Inference from a sample

□ **Computer science:** Efficient algorithms to

- Solve the optimization problem
- Represent and evaluate the model for inference

Machine learning

Learn = Optimize & Generalize



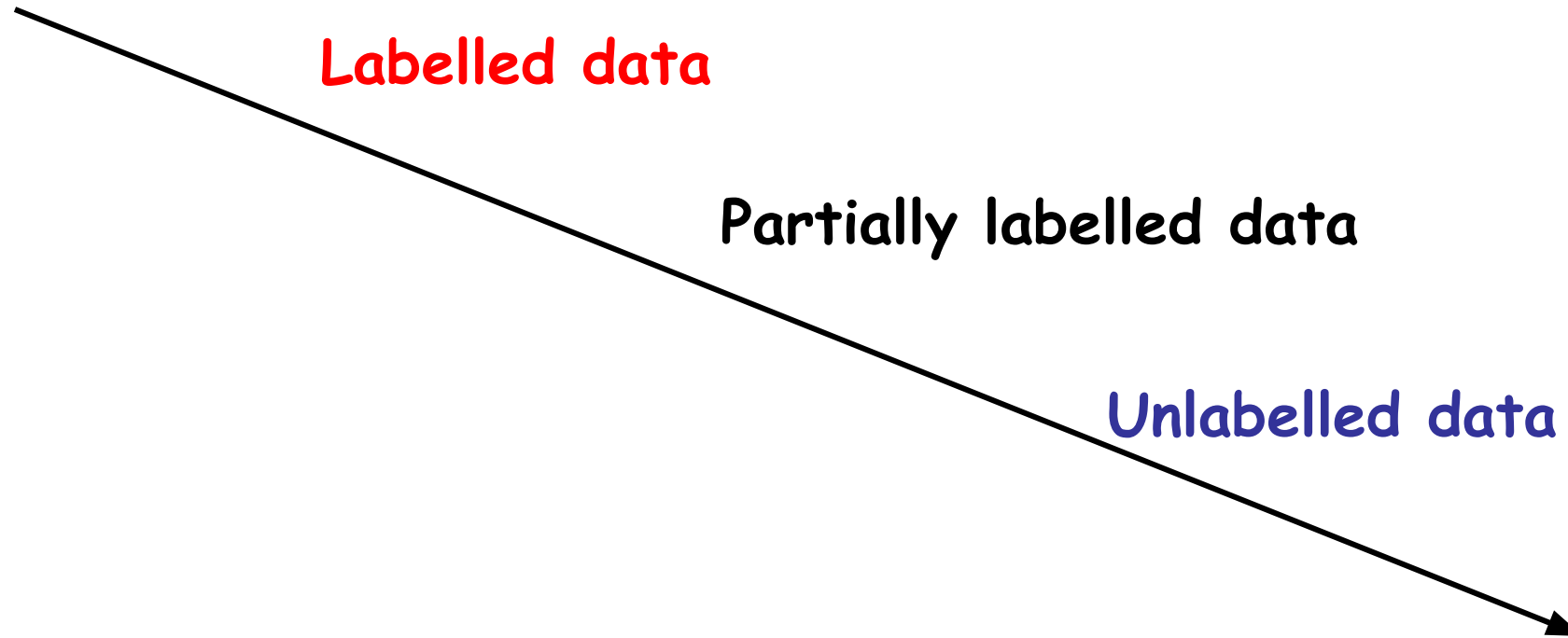
Optimization:

"Classical" methods
(mathematical programming
...) and less classical (genetic
algorithms ...)

Generalization Theory:

Minimization of estimated
generalization error

From supervised learning to unsupervised learning



Types of Learning

- **Supervised learning**
 - Training data includes desired outputs
- **Unsupervised learning**
 - Training data does not include desired outputs
- **Semi-supervised learning**
 - Training data includes a few desired outputs
- **Reinforcement learning**
 - Rewards from sequence of actions

Supervised vs. unsupervised Learning

- **Supervised learning:** classification and regression (from examples).
 - **Supervision:** The data are associated to pre-defined labels (classes / continuous values).
- **Unsupervised learning (clustering)**
 - **Class labels are unknown**
 - Given a set of data, the goal is to provide clusters from the data

Unsupervised Learning

Unsupervised learning

Representation problem

- How to represent the objects?
- How to define the similarity or the dissimilarity between objects?

Optimization problem

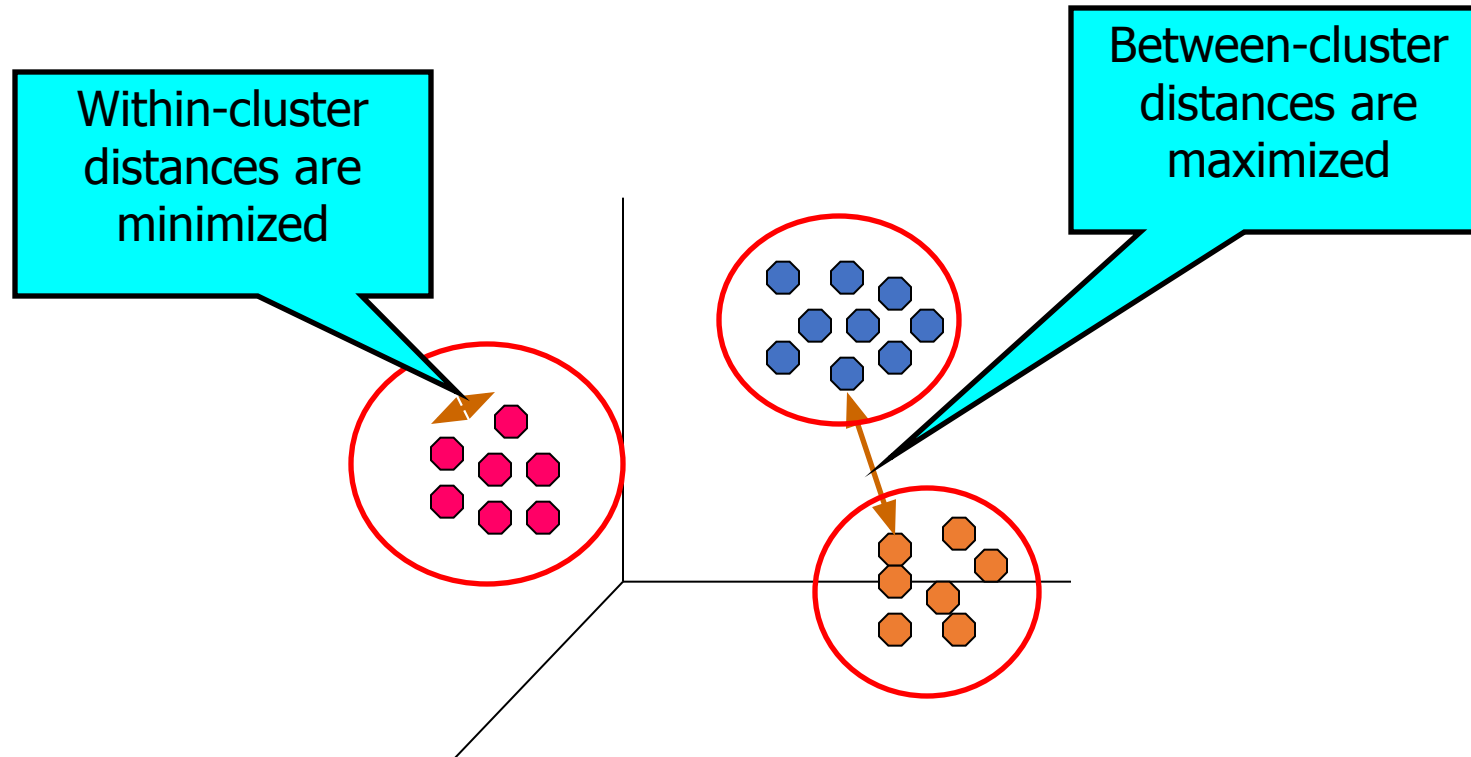
- What is a cluster?
- How to formulate the optimization problem for the partition?

Validation problem

- How to compare two clustering results?
- How to evaluate/ validate a clustering?

What is clustering?

- A **grouping** of data objects: the objects **within a group are similar** to one another and **different from the objects** in other groups



Why do we cluster?

- **Clustering:** given a collection of data objects, group them so that:
 - Similar to one another within the same cluster
 - Dissimilar to the objects in other clusters
- Clustering results are used:
 - As a **tool** to get insight into data distribution
 - Visualization of clusters may unveil important information
 - As a **preprocessing step** for other algorithms
 - Efficient indexing or dimension reduction tool

Distance functions

- The function $d(.,.)$ is a distance if:
 - $d(i, j) \geq 0$ (non-negativity)
 - $d(i, i) = 0$ (isolation)
 - also, $d(x, y) = 0 \Rightarrow x = y$ (identity-discerning)
 - $d(i, j) = d(j, i)$ (symmetry)
 - $d(i, j) \leq d(i, h) + d(h, j)$ (triangular inequality)
- The definitions of distance functions are different for real, boolean, categorical, and ordinal variables.
- Weights may be associated with some variables.

Distance/similarity

- **Dissimilarity** : distance without triangular inequality
- **Similarity** : function s from X^*X to R^+ such that:
 1. s symmetric : $(x,y) \in X^*X ; s(x,y) = s(y,x)$
 2. $(x,y) \in X^*X$ with $x \neq y ; s(x,x) = s(y,y) > s(x,y)$.

Data Structures

- *Data* matrix

attributes/variables

objects

$$\begin{bmatrix} x_{11} & \dots & x_{1\ell} & \dots & x_{1d} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{i\ell} & \dots & x_{id} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{n\ell} & \dots & x_{nd} \end{bmatrix}$$

- *Distance* matrix

objects

objects

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Distance functions for real-valued vectors

- *Minkowski* (L_p norms) distance:

$$L_p(x,y)=\left(|x_1-y_1|^p+|x_2-y_2|^p+\dots+|x_d-y_d|^p\right)^{1/p}=\left(\sum_{i=1}^d (x_i-y_i)^p\right)^{1/p}$$

where p is a positive integer

- If $p = 1$, L_1 is the *Manhattan* distance:

$$L_1(x,y)=|x_1-y_1|+|x_2-y_2|+\dots+|x_d-y_d|=\sum_{i=1}^d |x_i-y_i|$$

Distance functions for real-valued vectors

- If $p = 2$, L_2 is the **Euclidean distance**:

$$d(x,y)=\sqrt{(|x_1-y_1|^2+|x_2-y_2|^2+...+|x_d-y_d|^2)}$$

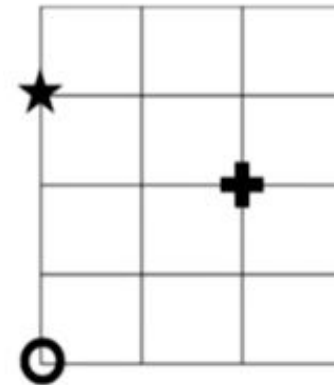
- **Weighted distances:**

- Euclidean distance:

$$d(x,y)=\sqrt{(w_1|x_1-y_1|^2+w_2|x_2-y_2|^2+...+w_d|x_d-y_d|^2)}$$

- Manhattan distance

$$d(x,y)=w_1|x_1-y_1|+w_2|x_2-y_2|+...+w_d|x_d-y_d|$$



Is circle closer to star or cross?

- Euclidean distance
 - Cross – 2.8
 - Star – 3
- Manhattan Distance
 - Cross – 4
 - Star - 3

The k-means problem

- Given a set X of n points in a d -dimensional space and an integer k
- Choose a set of k points $\{k_1, k_2, \dots, k_k\}$ in the d -dimensional space to form clusters $\{C_1, C_2, \dots, C_k\}$ such that

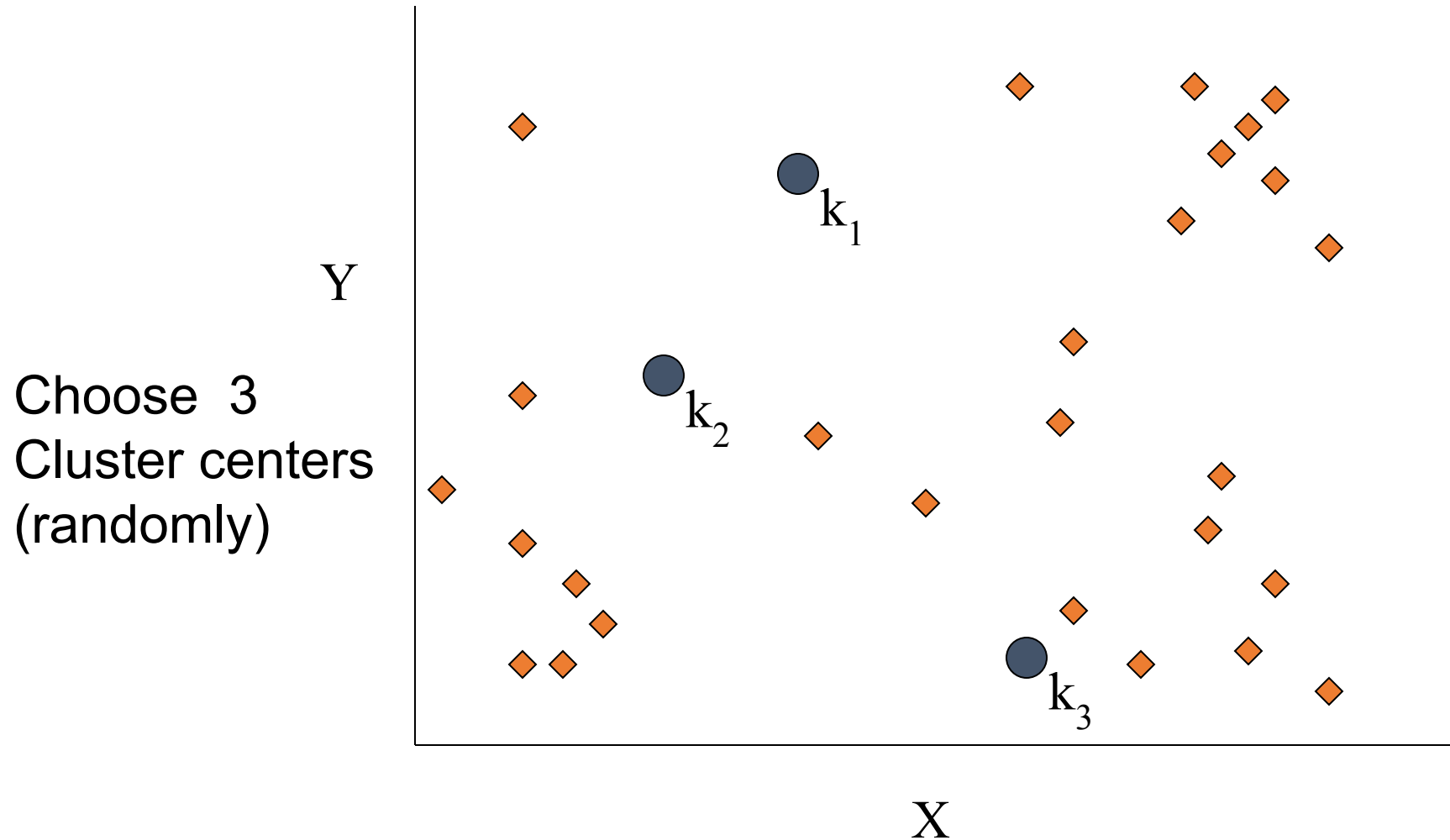
$$Cost(C) = \sum_{i=1}^k \sum_{x \in C_i} L_2^2(x - k_i)$$

is minimized

The k-means algorithm

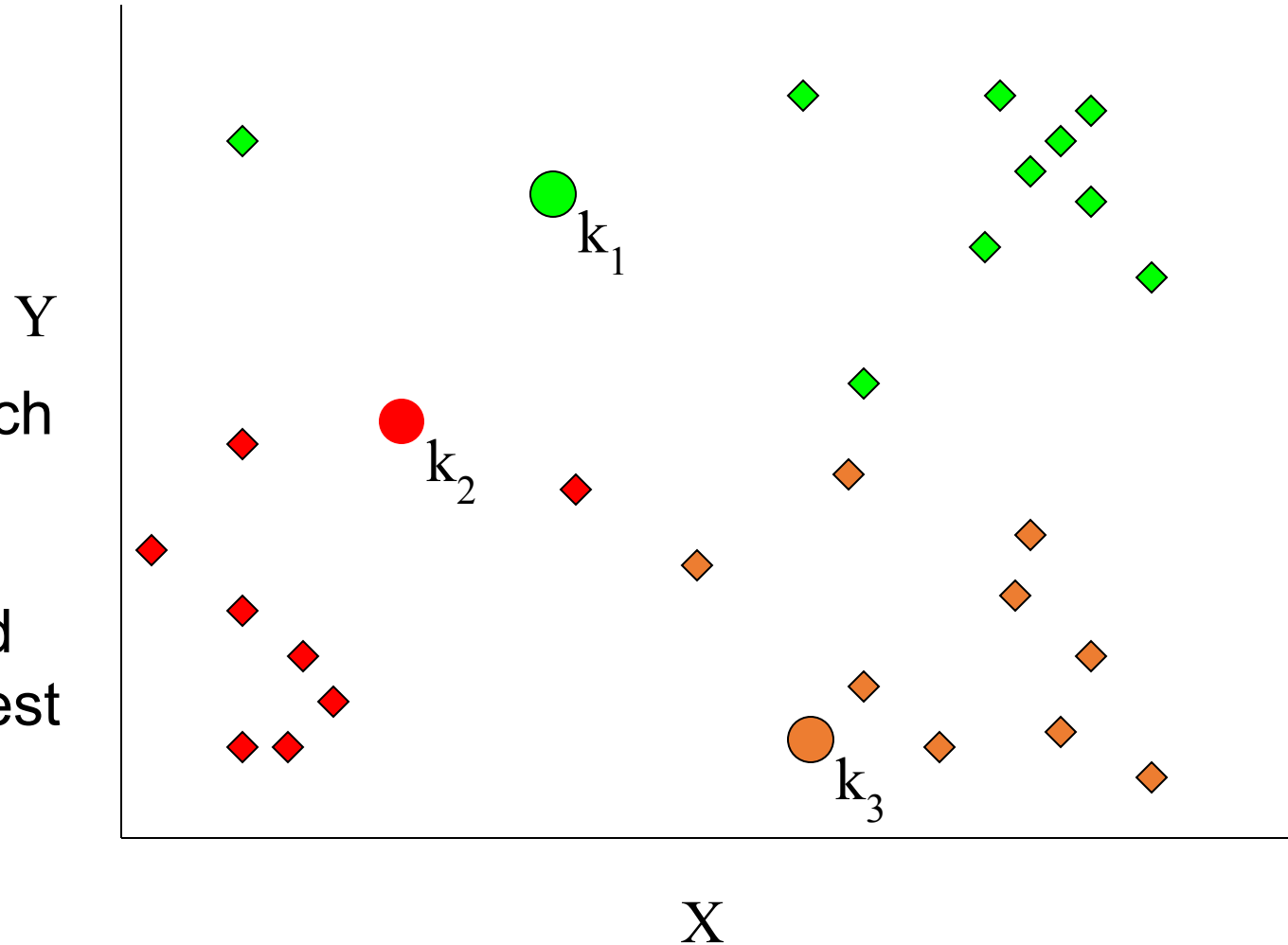
- Randomly initialize k cluster centers $\{c_1, \dots, c_k\}$
- For each i , the cluster C_i is the set of points in X that are closer to k_i than they are to k_j for all $i \neq j$
- For each i , k_i is the center of cluster C_i (mean of the vectors in C_i)
- Repeat until convergence

Example K-means, step 1

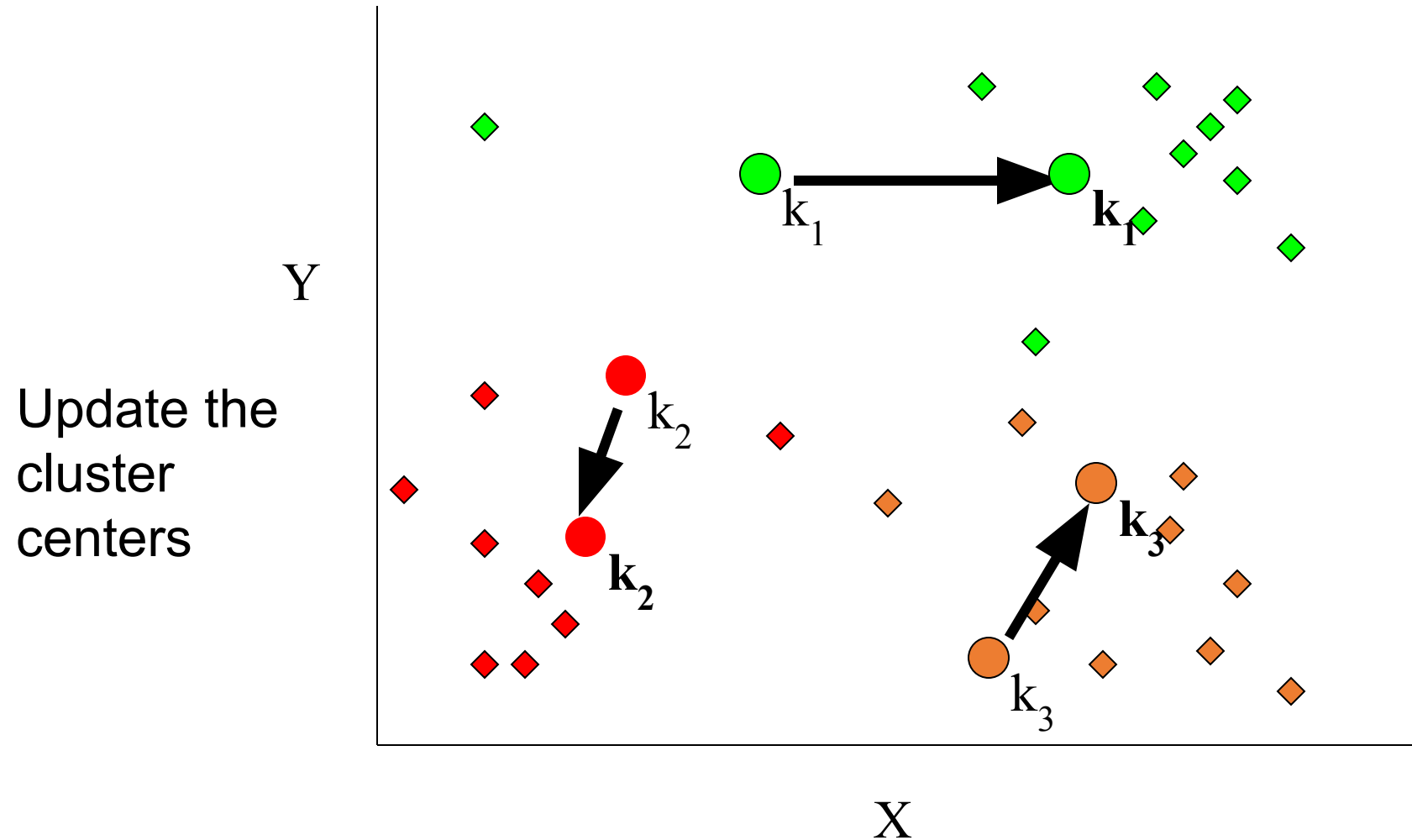


Example K-means, step 2

Allocate each point to the cluster represented by the closest center

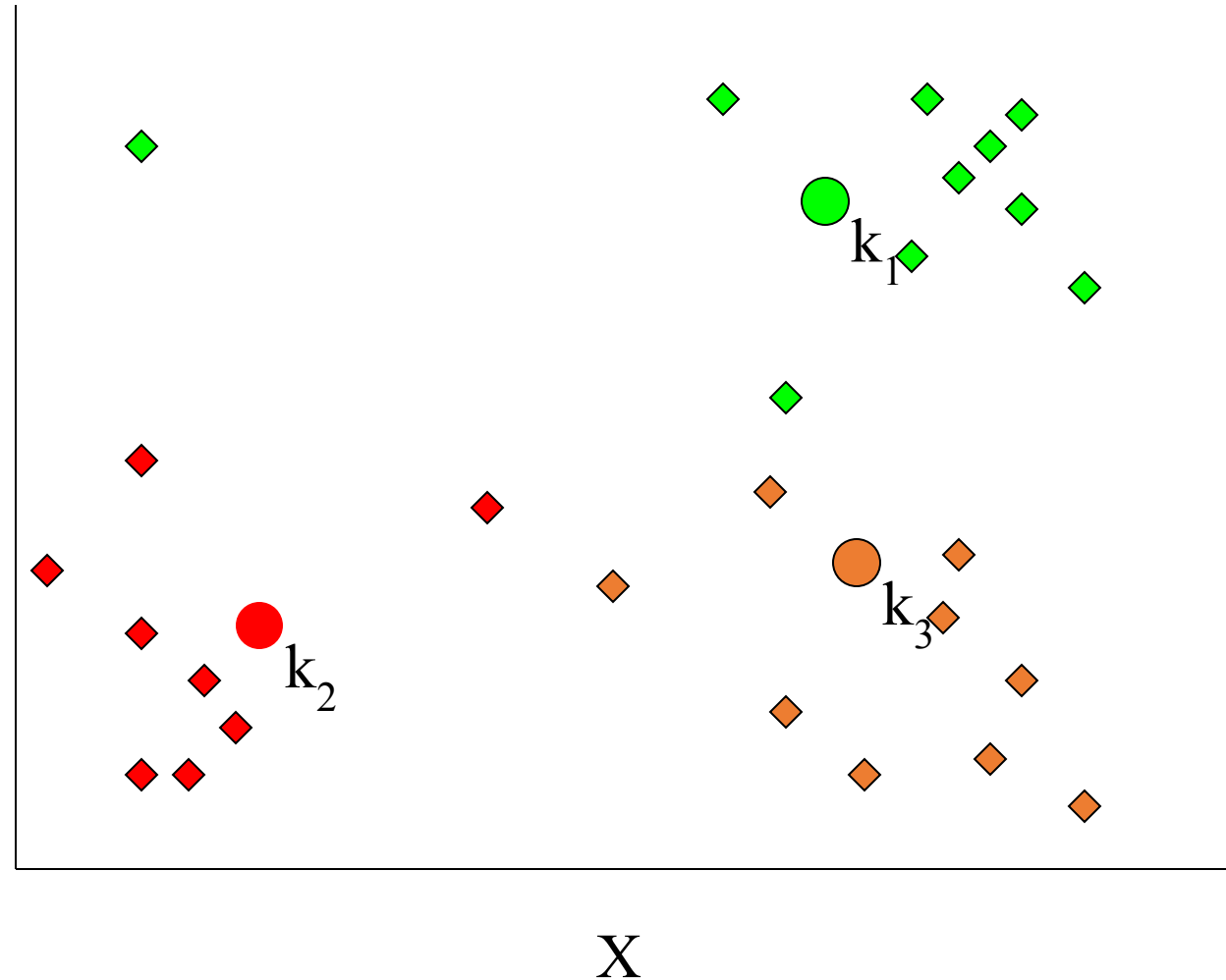


Example K-means, step 3

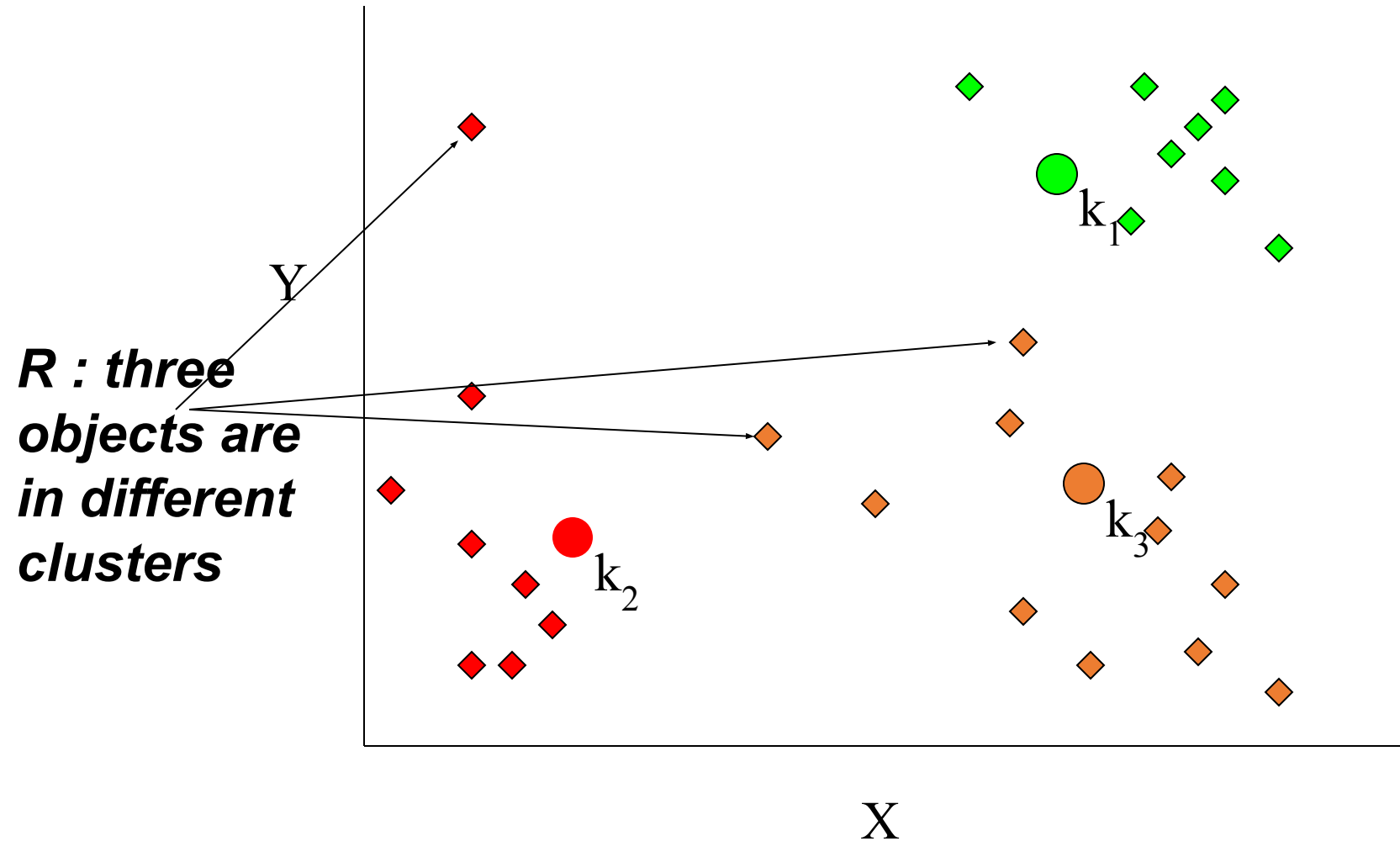


Example K-means, step 4

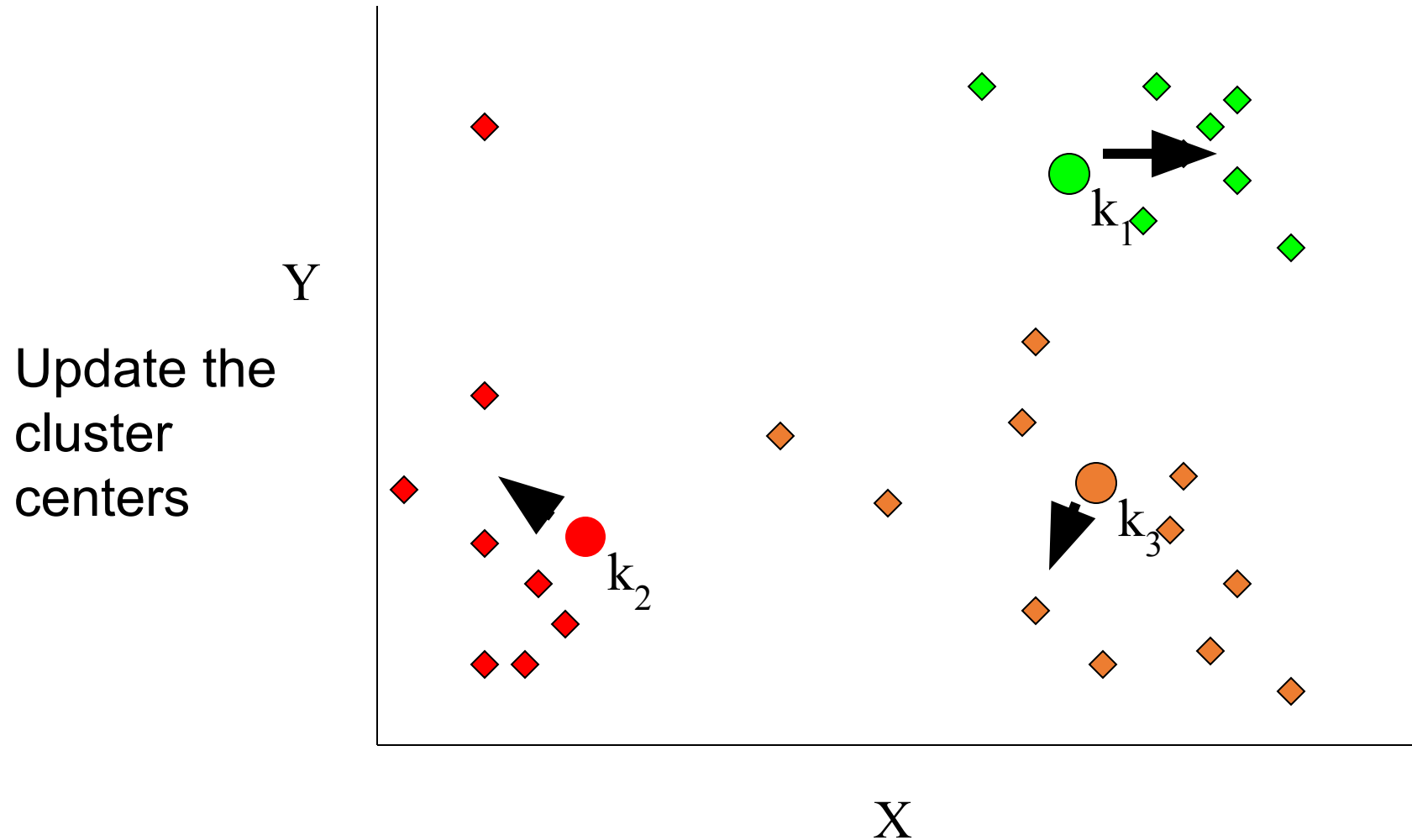
Reallocate the
objects to
clusters
represented by
the new centers



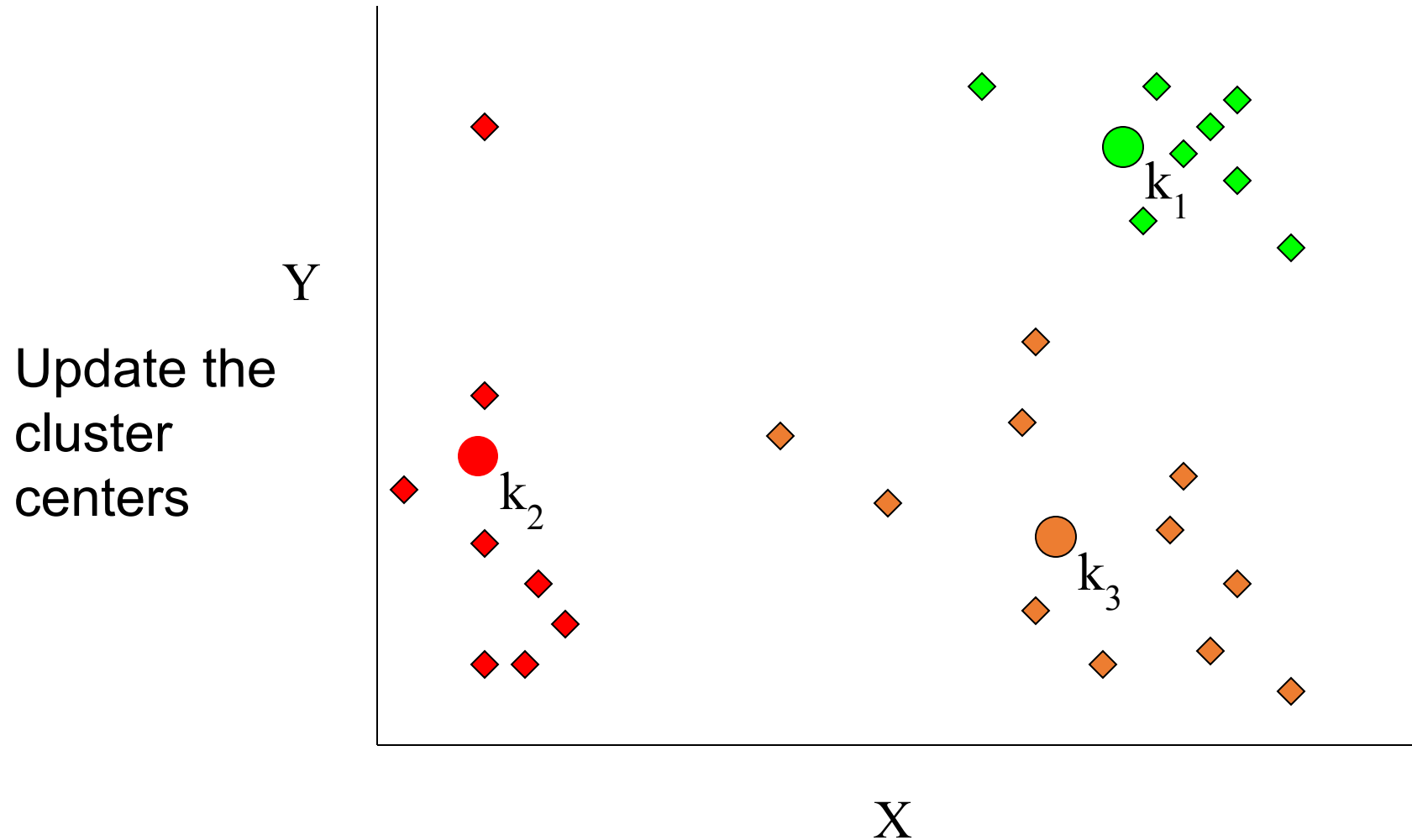
Example K-means, step 4



Example K-means, step 4b



Example K-means, step 5



Exemple

Voici un tableau de données : il s'agit des cigarettes représentées par leur contenu en nicotine et en goudron.

CIGARETTES	Nicotine (cg)	Goudron (mg)
Royal anis	4.5	4.9
Rothmans	11	14
Chesterfield Lights	6	8
Benson & Hedges	11	13
Peter Stuyvesant	10	12.7
Gitanes	10	12
Malboro	10	14
Lucky Strike	9	14
Light Delight	5	7

Soit $k = 2$. Utiliser l'algorithme des k-moyennes pour faire un clustering sur ces données.

- Choisir l'initialisation suivante : $M1 = \text{Gitanes}$ et $M2 = \text{Lucky Strike}$

Exemple

Voici un tableau de données : il s'agit des cigarettes représentées par leur contenu en nicotine et en goudron.

CIGARETTES	Nicotine (cg)	Goudron (mg)
Royal anis	4.5	4.9
Rothmans	11	14
Chesterfield Lights	6	8
Benson & Hedges	11	13
Peter Stuyvesant	10	12.7
Gitanes	10	12
Malboro	10	14
Lucky Strike	9	14
Light Delight	5	7

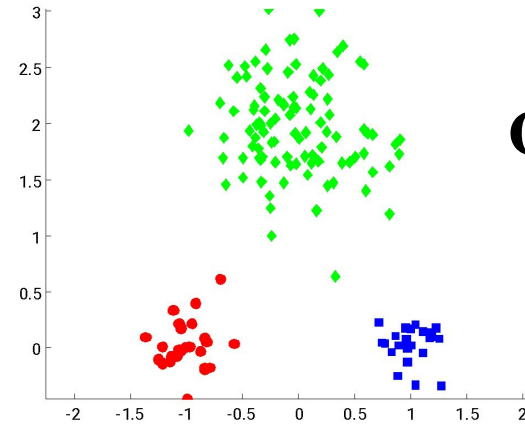
Soit $k = 2$. Utiliser l'algorithme des k-moyennes pour faire un clustering sur ces données.

- Choisir l'initialisation suivante : $M1 = \text{Gitanes}$ et $M2 = \text{Lucky Strike}$

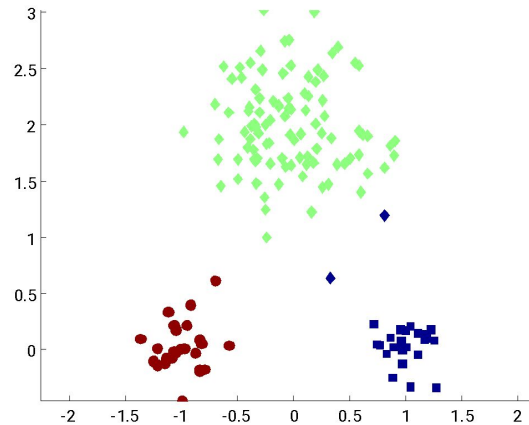
Properties of the k-means algorithm

- Finds a local optimum
- Converges often quickly
- The choice of initial points influences the clustering result

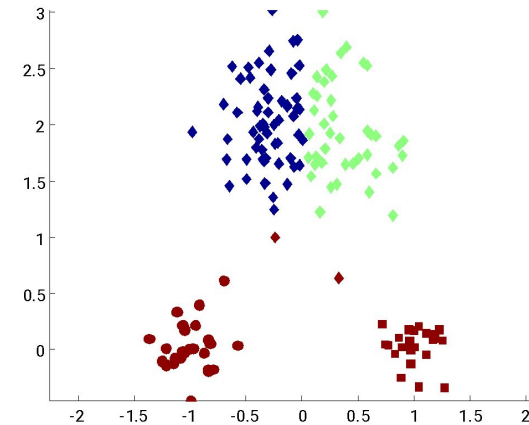
K-means Clustering results with different random initializations



Original Points



**Optimal
Clustering**



**Sub-optimal
Clustering**

Some alternatives to random initialization

- Multiple runs
- Select original set of points. E.g., choose the most distant points as cluster centers (kmeans++ algorithm)