Cluster Analysis: An Exciting and Mysterious Topic

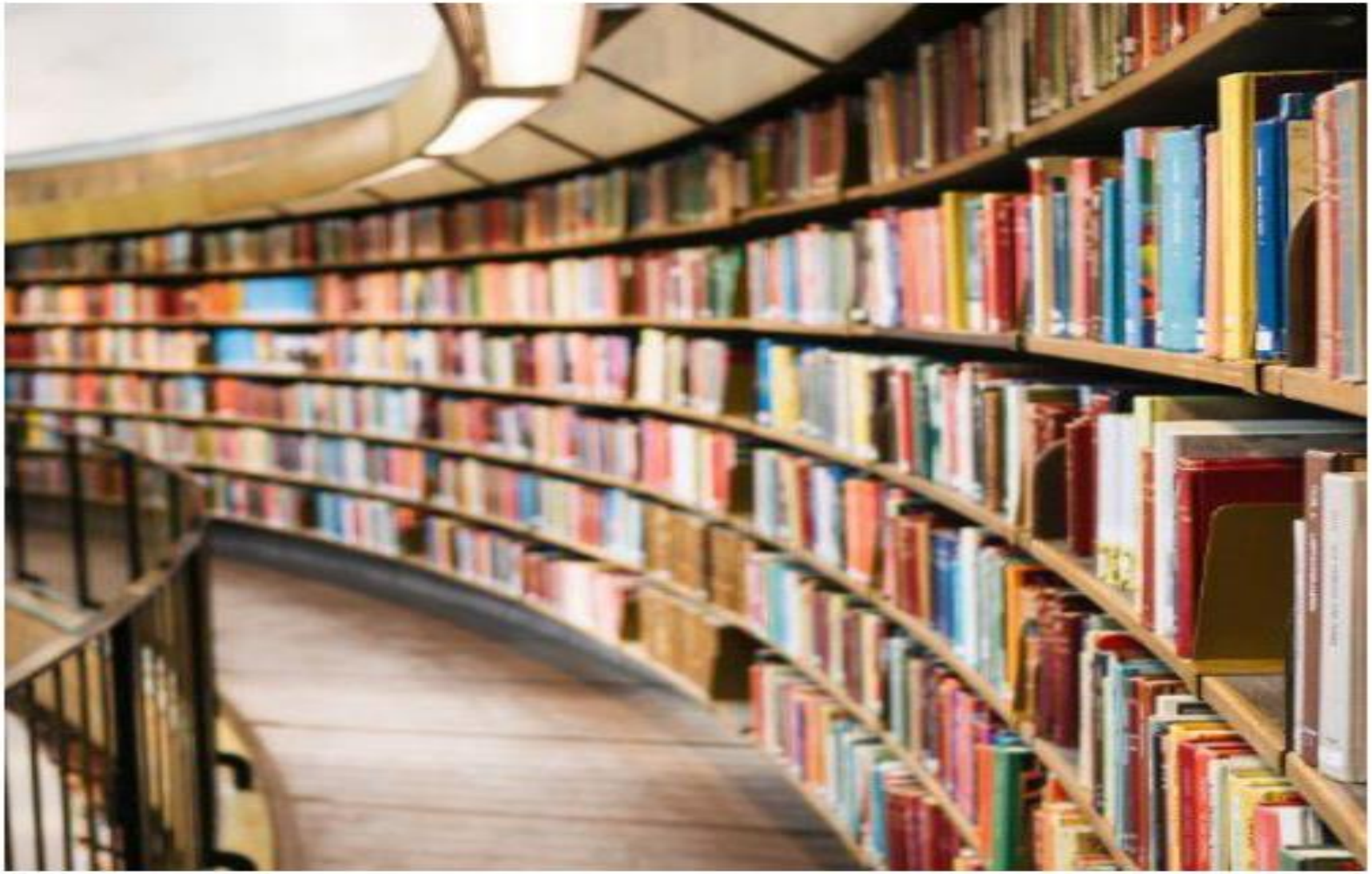
Mousa Golalizadeh

Department of Statistics
School of Mathematical Sciences
Tarbiat Modares University

Talk Given at SOAL
Department of Mathematical Sciences
Sharif University of Technology
12 May 2022

Motivations

Why Such Topic?

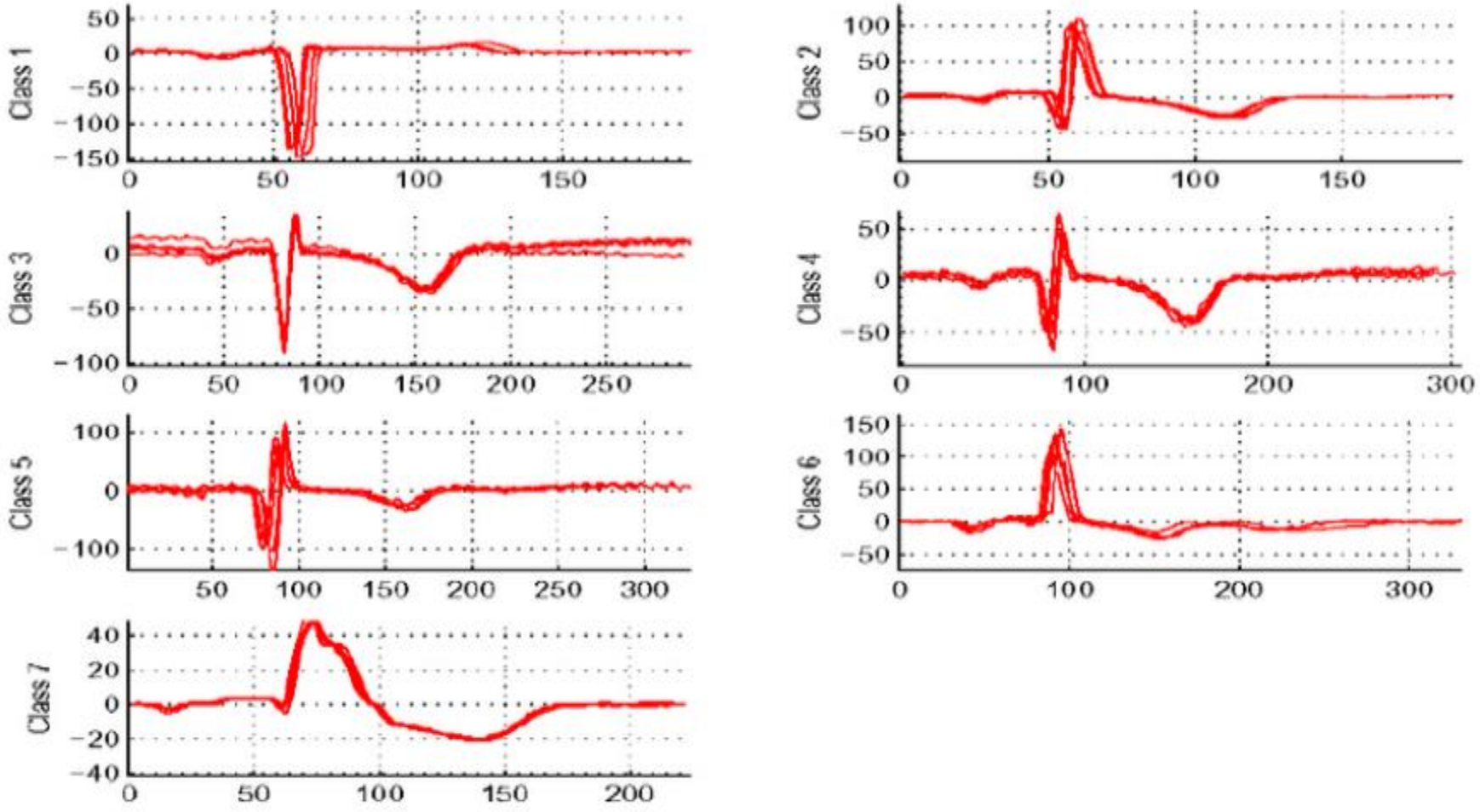


Why Such Topic?

Pixel Values

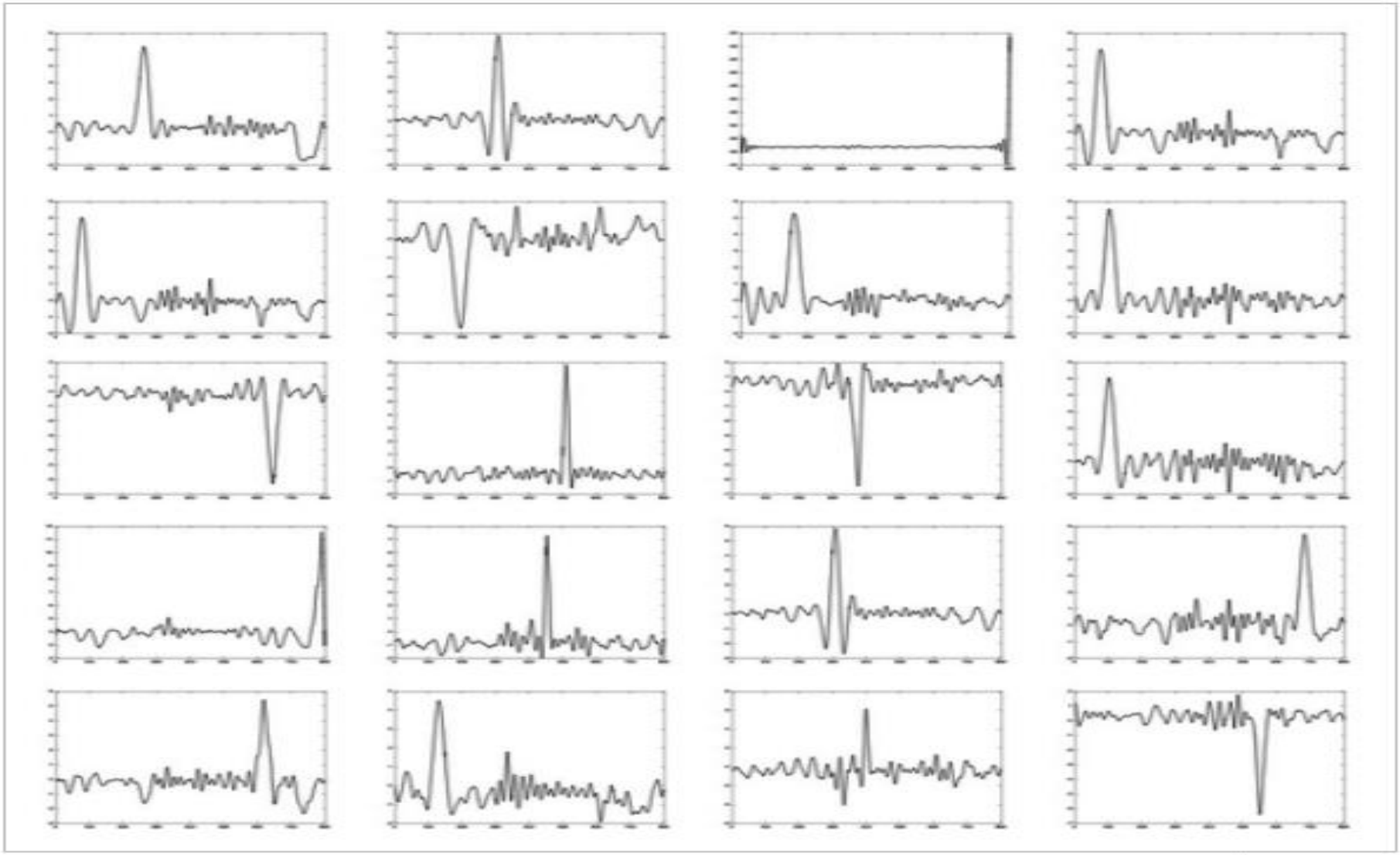
متغیر / نمونه	۱	۲	۳	۴	۵	۶	۷	۸
۱	۶	−۱	−۱	−۱	−۱	−۱	−۱	−۱
۲	۵	−۱	−۱	−۱	−۰/۸۱	−۰/۶۷	−۰/۸۱	−۰/۸۹
۳	۴	−۱	−۱	−۱	−۱	−۱	−۱	−۱
۴	۷	−۱	−۱	−۱	−۱	−۱	−۰/۲۷	۰/۶۸
۵	۳	−۱	−۱	−۱	−۱	−۱	−۰/۹۳	−۰/۲۰
۶	۶	−۱	−۱	−۱	−۱	−۱	−۰/۴۰	۰/۹۸
۷	۳	−۱	−۱	−۱	−۰/۸۳	۰/۴۴	۱	۱
۸	۱	−۱	−۱	−۱	−۱	−۱	−۱	−۱
۹	۰	−۱	−۱	−۱	−۱	−۱	−۰/۴۵	۰/۸۸
۱۰	۱	−۱	−۱	−۱	−۱	−۱	−۱	−۱
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
۷۲۸۲	۳	−۰/۷۰	۰/۱۲	۰/۳۶	۰/۳۶	۰/۵۵	/۷۸۰	۰/۷۷
۷۲۸۳	۳	−۱	−۰/۸۸	−۰/۳۳	۰/۲۷	۰/۳۳	۰/۷۵	۱
۷۲۸۴	۳	−۰/۹۸	−۰/۰۵	۰/۲۳	۰/۲۳	۰/۲۳	−۰/۳۵	−۰/۸۱
۷۲۸۵	۱	−۱	−۱	−۱	۱	−۱	−۱	−۱
۷۲۸۶	۱	−۱	−۱	−۱	−۱	−۱	−۱	−۰/۷۴
۷۲۸۷	۳	−۱	−۱	−۱	−۰/۹۹	−۰/۵۳	−۰/۲۱	۰/۶۲
۷۲۸۸	۳	−۱	−۱	−۱	−۰/۹۹	۰/۷۱	۰/۵۶	۰/۳۵
۷۲۸۹	۳	−۱	−۱	−۱	−۰/۷۸	−۰/۹۸	−۰/۸۳	۰/۰۷
۷۲۹۰	۰	−۱	−۱	−۱	−۱	−۱	−۰/۵۵	۰/۴۶
۷۲۹۱	۱	−۱	−۱	−۱	−۱	−۱	−۰/۱۱	۱

Why Such Topic?



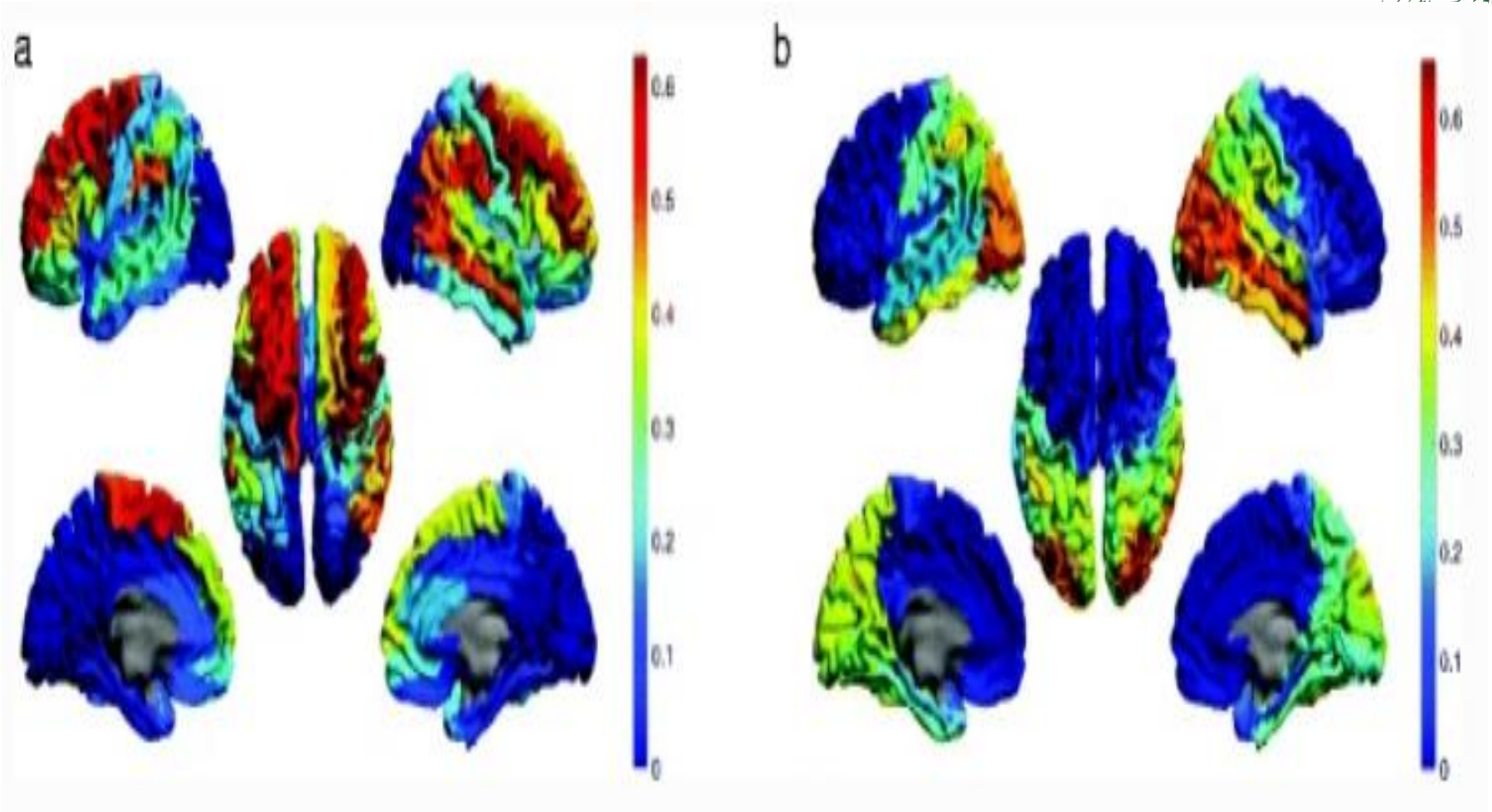
Example of ECG beats used for the hierarchical clustering

Why Such Topic?



Source: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/iet-sen.2016.0261

Why Such Topic?



Source: Märtens, M., Meier, J., Hillebrand, A., Tewarie, P., and Mieghem, P. V. (2017) Brain network clustering with information flow motifs. *Appl Netw Sci* **2,** 25. https://doi.org/10.1007/s41109-017-0046-z

Preliminaries

Facts (1)

Each noun is a label used to describe a class of things and so animals are called cats, dogs, horses, etc., and each name collects individuals into groups.

Facts (2)

- Naming and classifying are essentially synonymous as are *homogeneous groups* or *clusters*.

- Groups are *mutually exclusive* (an item belongs to a single group) rather than *overlapping*.

Question

What is difference between Clustering and Classification?

Classification problems

- Product categorization based on product attributes and/or text descriptions of product is an example of *classification*: deciding how to assign (known) labels to an object.

- Classification is an example of what is called *supervised learning*: to learn how to classify objects, you need a dataset of objects that have already been classified (called *training*).

Terminologies

Synonyms Terms:

grouping

assembling

gathering

categorizing

packing

classifying

…

Main Purpose

intelligent test

=

Set objects in distinct categories

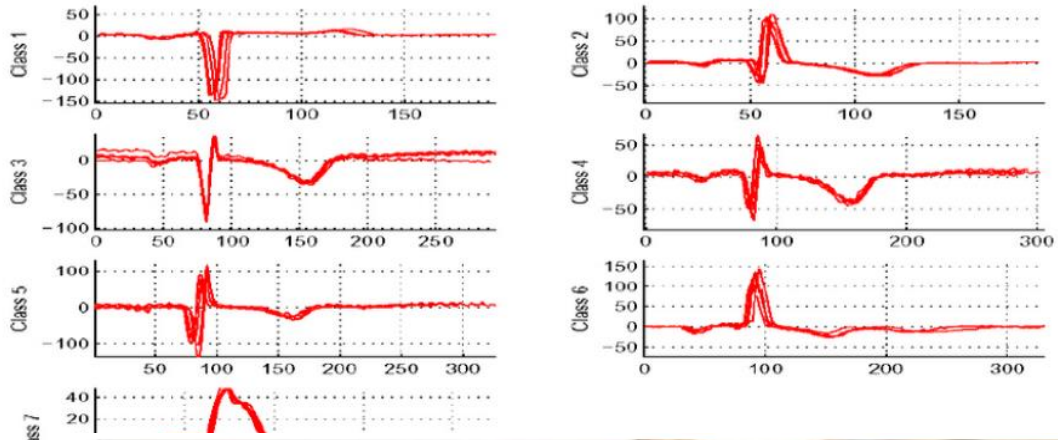
What is objects?

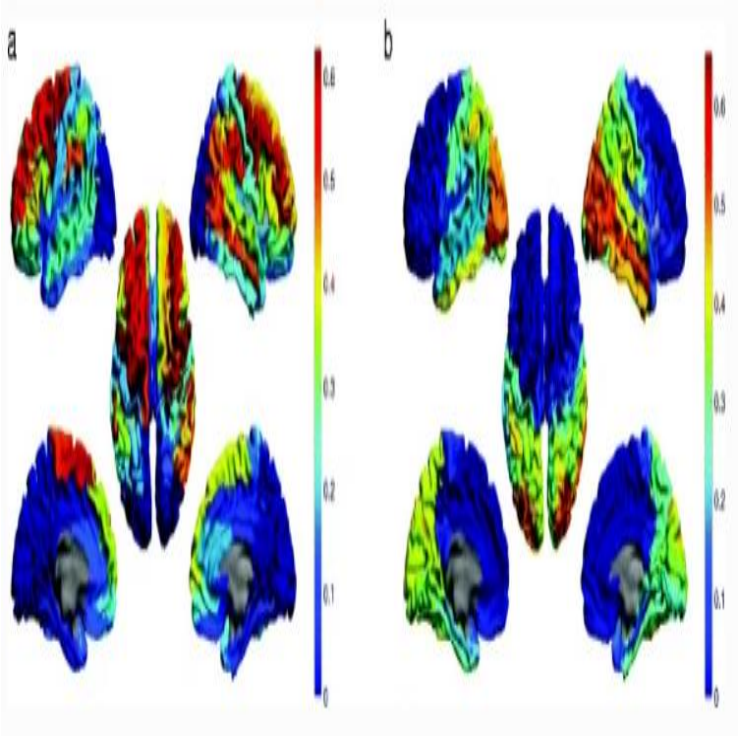
- A data object is a collection of one or more data points that create meaning as a whole.

- The most common example of a data object is a *data table*, but others include arrays, pointers, records, files, sets, and scalar types.

- Values in a data object may have their own unique IDs, data types, & attributes.

Recall





Examp

Key Question

Is cluster a

Quantitative

or

Qualitative?

Key Ingredients

Categorical

variable

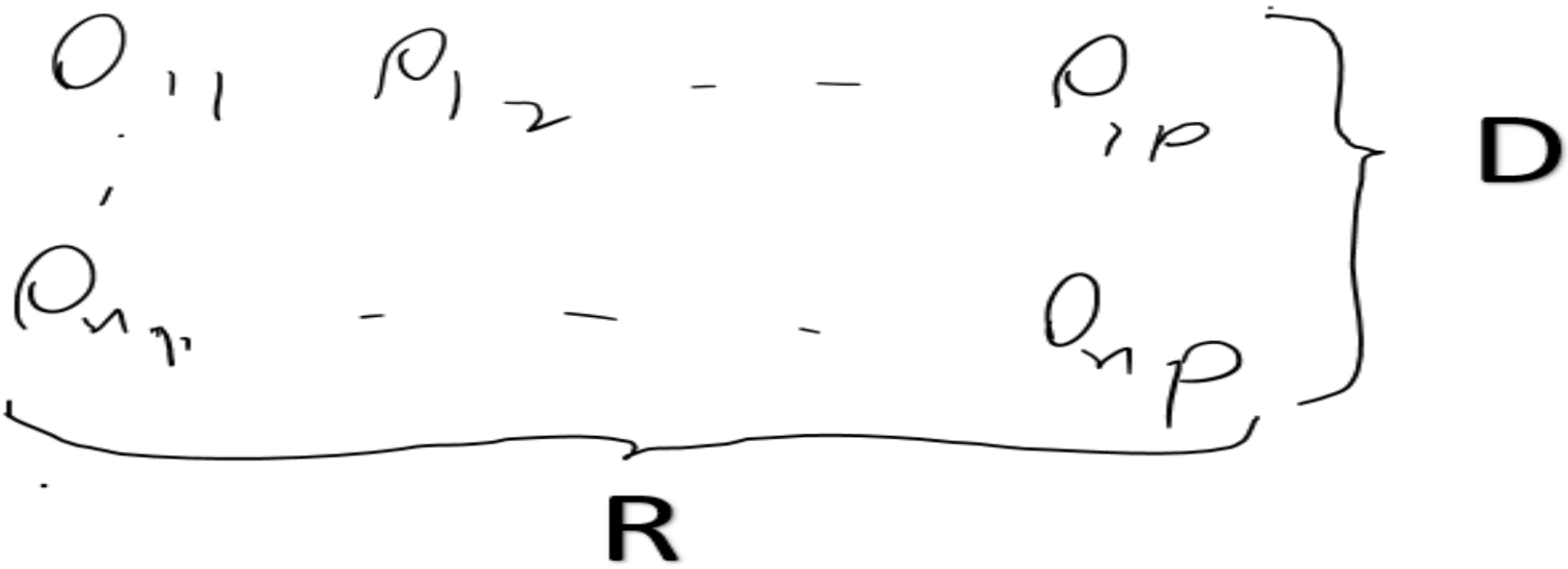
Quantitative

Synonyms Terms:

variable, feature, attribute, ….

Remark

Note that individual objects shared certain properties (Correlation) and sort of proximity (Distance)



Some Technical Issues

Structure

Data objects (observation) usually consist of:

Values: The data itself.

Unique IDs. Identifiers for each objects.

Attributes. Additional data within one Unique ID.

Data types. Classifications of data such as text, numeric, and boolean.

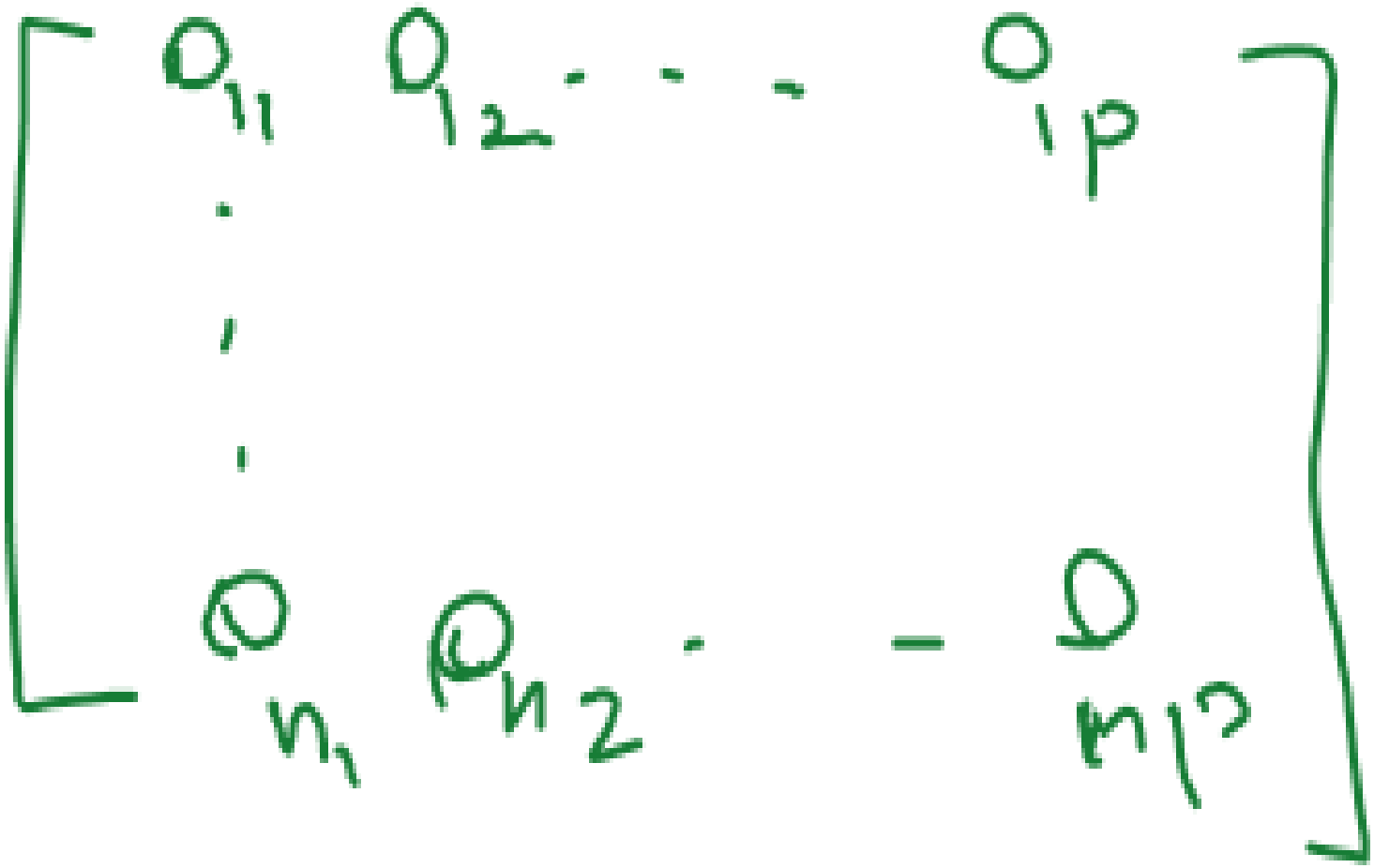
Data Type

Types of data used in cluster analysis are:

- **Interval-Scaled variables**
- **Binary variables**
- **Nominal, Ordinal, & Ratio variables**
- **Variables of mixed types**

Warning: Univariate or Multivariate Data?

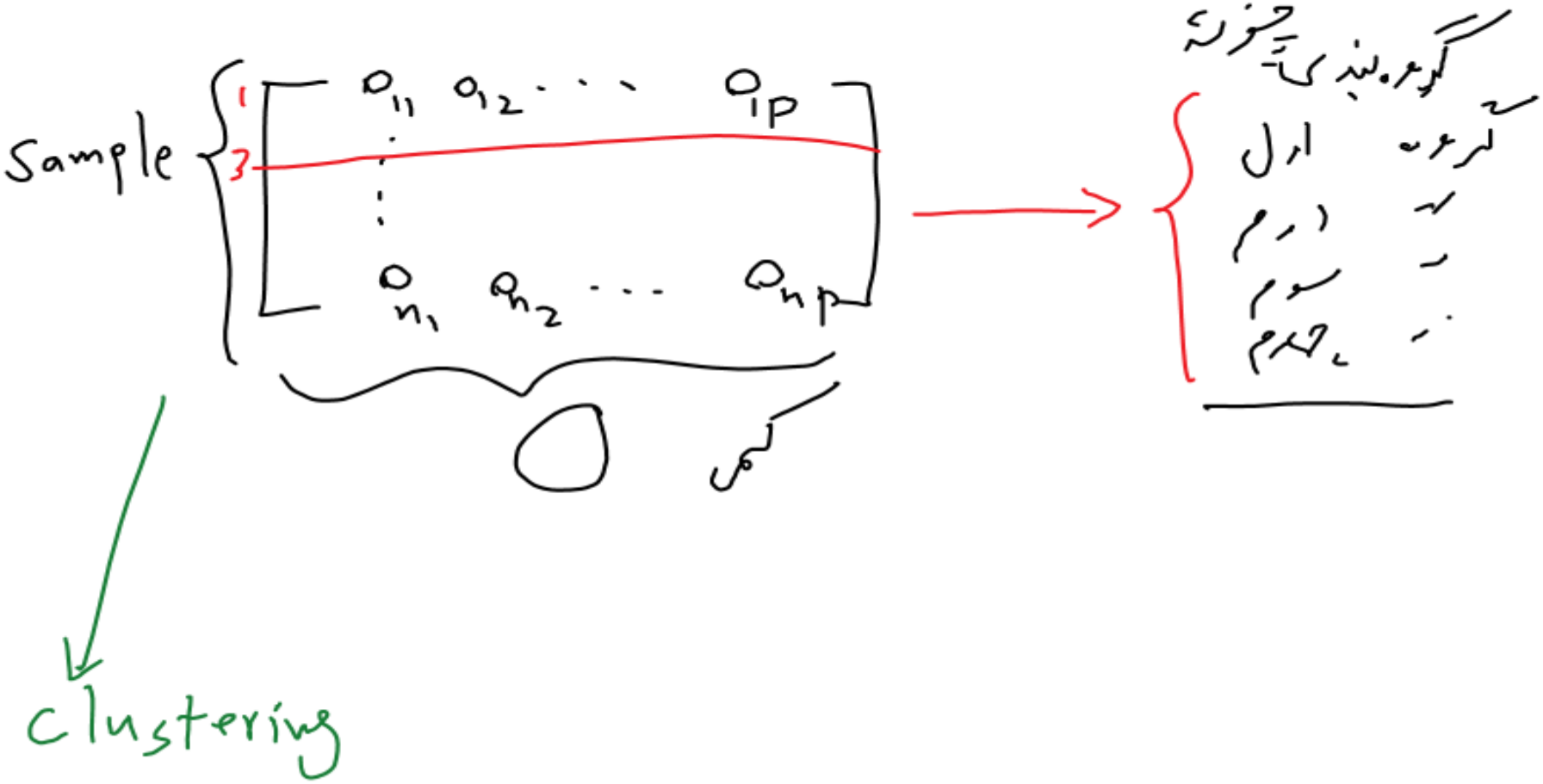
Data Matrix



An Example

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score
Joyner-Kersee (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51	7291
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12	6897
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20	6858
Sablovskaite (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24	6540
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90	6540
Schulz (GDR)	13.75	1.83	13.50	24.65	6.33	42.82	125.79	6411
Fleming (AUS)	13.38	1.80	12.88	23.59	6.37	40.28	132.54	6351
Greiner (USA)	13.55	1.80	14.13	24.48	6.47	38.00	133.65	6297
Lajbnerova (CZE)	13.63	1.83	14.28	24.86	6.11	42.20	136.05	6252
Bouraga (URS)	13.25	1.77	12.62	23.59	6.28	39.06	134.74	6252
Wijnsma (HOL)	13.75	1.86	13.01	25.03	6.34	37.86	131.49	6205
Dimitrova (BUL)	13.24	1.80	12.88	23.59	6.37	40.28	132.54	6171
Scheider (SWI)	13.85	1.86	11.58	24.87	6.05	47.50	134.93	6137
Braun (FRG)	13.71	1.83	13.16	24.78	6.12	44.58	142.82	6109
Ruotsalainen (FIN)	13.79	1.80	12.32	24.61	6.08	45.44	137.06	6101
Yuping (CHN)	13.93	1.86	14.21	25.00	6.40	38.60	146.67	6087
Hagger (GB)	13.47	1.80	12.75	25.47	6.34	35.76	138.48	5975
Brown (USA)	14.07	1.83	12.69	24.83	6.13	44.34	146.43	5972
Mulliner (GB)	14.39	1.71	12.68	24.92	6.10	37.76	138.02	5746
Hautenauve (BEL)	14.04	1.77	11.81	25.61	5.99	35.68	133.90	5734
Kytola (FIN)	14.31	1.77	11.66	25.69	5.75	39.48	133.35	5686
Geremias (BRA)	14.23	1.71	12.95	25.50	5.50	39.64	144.02	5508
Hui-Ing (TAI)	14.85	1.68	10.00	25.23	5.47	39.14	137.30	5290
Jeong-Mi (KOR)	14.53	1.71	10.83	26.61	5.50	39.26	139.17	5289
Launa (PNG)	16.42	1.50	11.78	26.16	4.88	46.38	163.43	4566

Initial Thought of Clustering

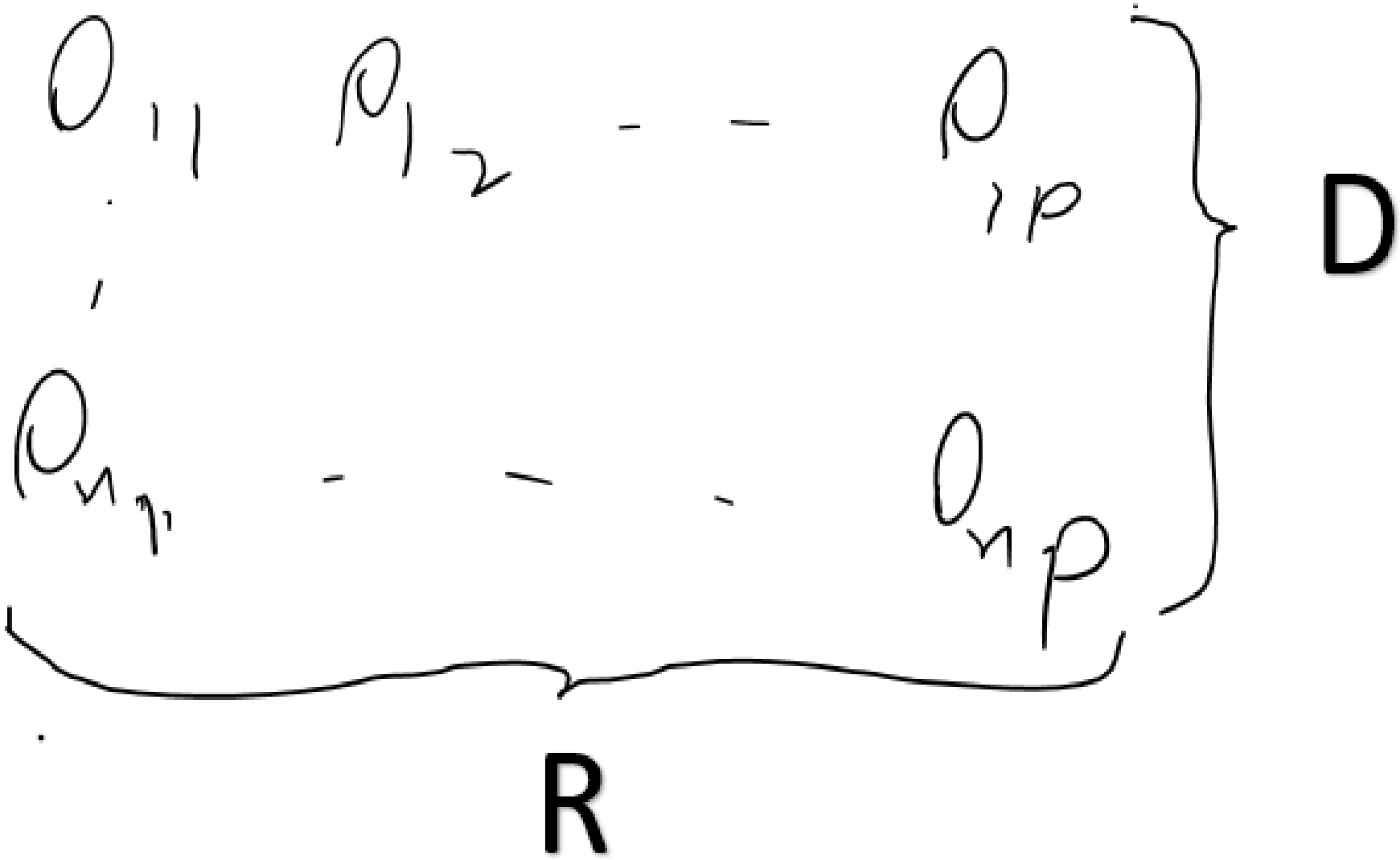


Note

Clustering, similar to other modeling techniques is a part of exploratory analysis a so unsupervised learning method.

Clustering is task of working without known targets

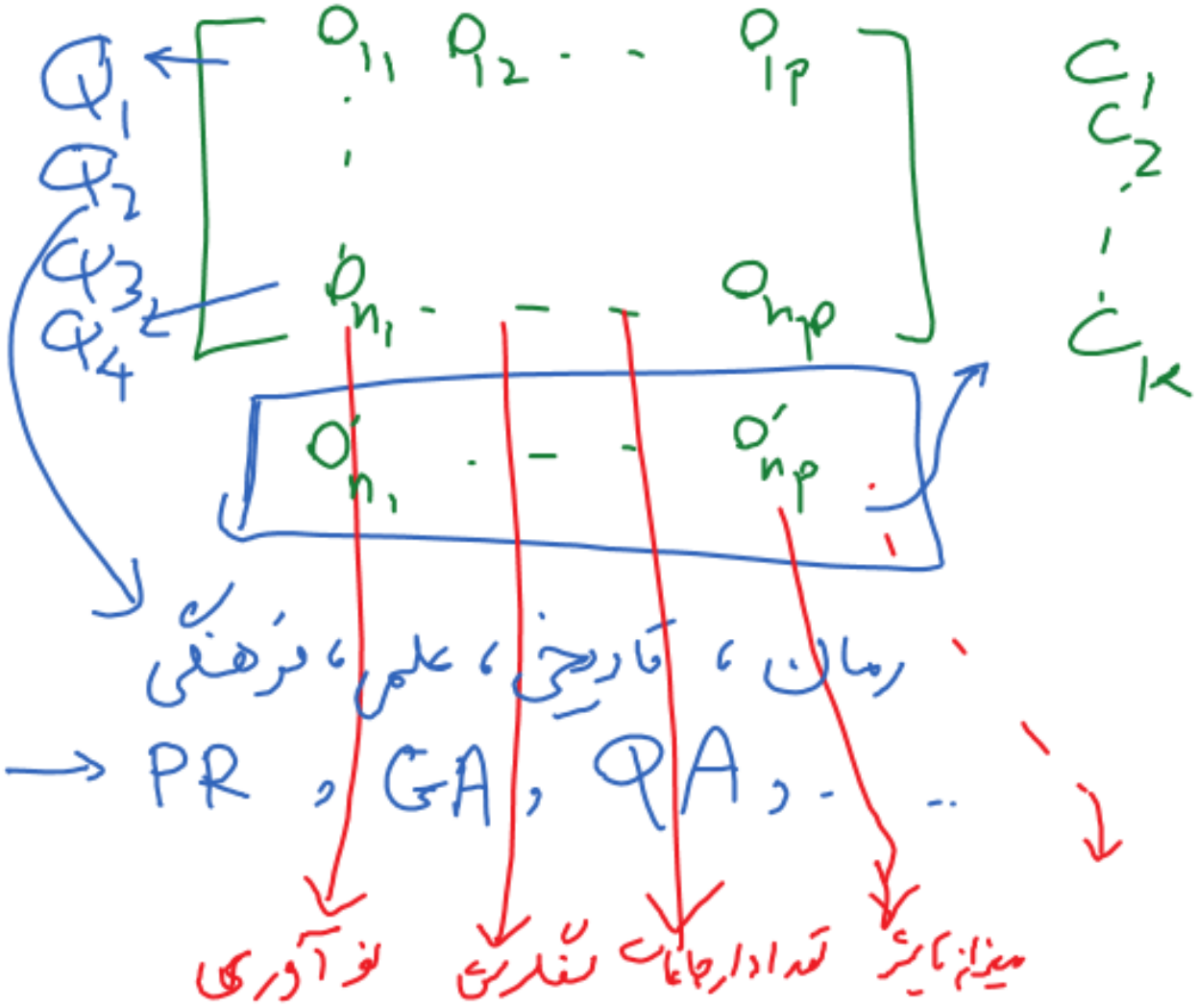
Useful Quantities



Another Example

chest	waist	hips	gender	chest	waist	hips	gender
34	30	32	male	36	24	35	female
37	32	37	male	36	25	37	female
38	30	36	male	34	24	37	female
36	33	39	male	33	22	34	female
38	29	33	male	36	26	38	female
43	32	38	male	37	26	37	female
40	33	42	male	34	25	38	female
38	30	40	male	36	26	37	female
40	30	37	male	38	28	40	female
41	32	39	male	35	23	35	female

Case Study


رمان ، تاریخی ، علمی ، فرهنگی
PR ، GA ، QA ، ...
نوآوری
تعداد ارجاعات
میزان تاثیر
WOS
ISI

Clustering and Distances

Distances (1)

- Cluster analysis is based on relative distances, homogeneity and degree of their separations.

- One needs the notions of *similarity* and *dissimilarity* to initiate clustering.

- Based on dissimilarity, as a *distance*, the points in a cluster are closer to each other than they are to points in others.

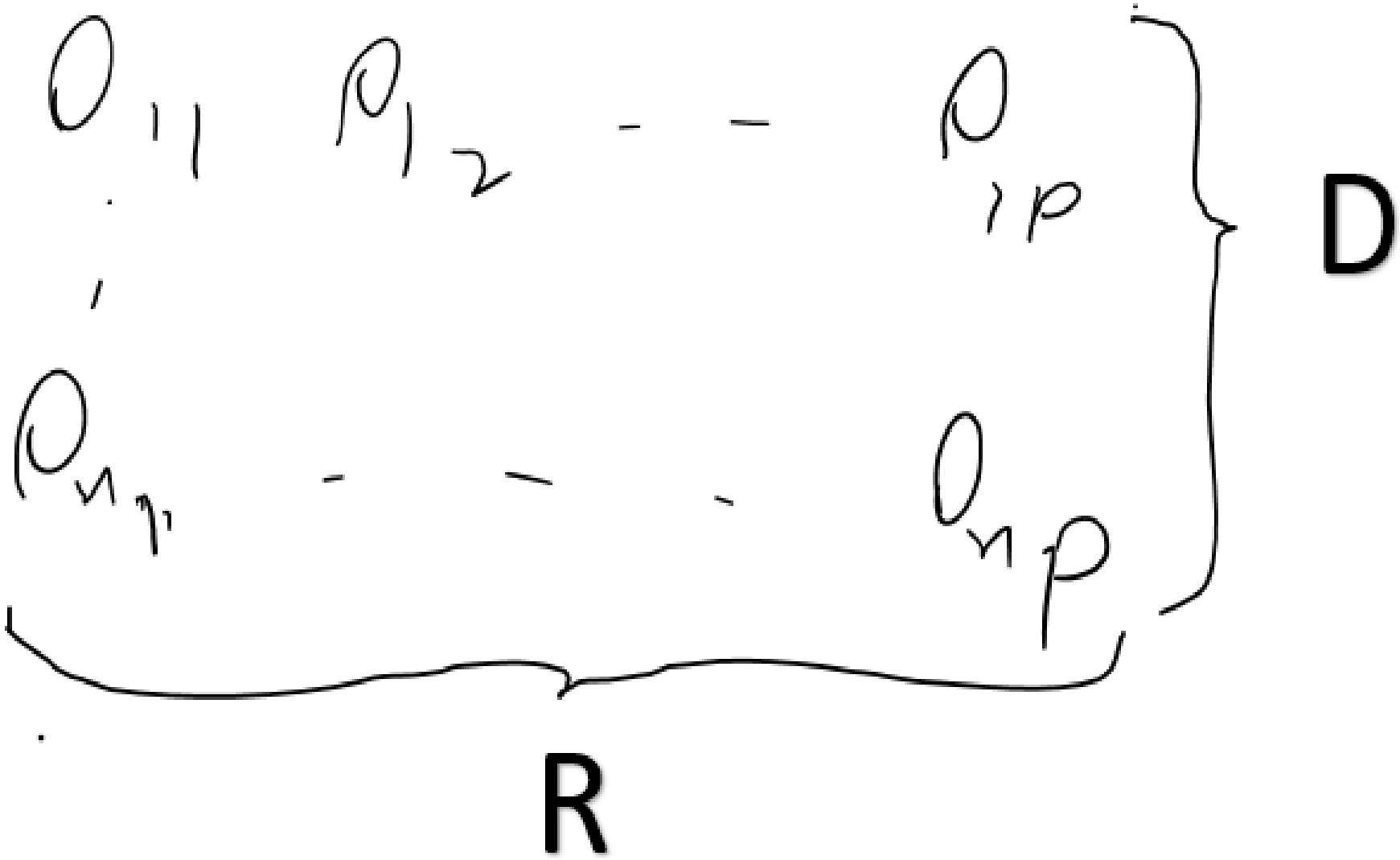
Distances (2)

Different application areas will have different notions of distance and dissimilarity.

A few of the most common distances are:

- Euclidean distance
- Hamming distance
- Manhattan (city block) distance
- Cosine similarity

Recall



Scaling

One way to try to make clustering more coordinate-free is to transform all columns to have a mean value of 0 and a standard deviation of 1.

But, be careful on optimization and reaching sensible results issues!

Visualizing Clustering Results

Objective

- Visualization is an effective way to get an overall view of the data or clusters.

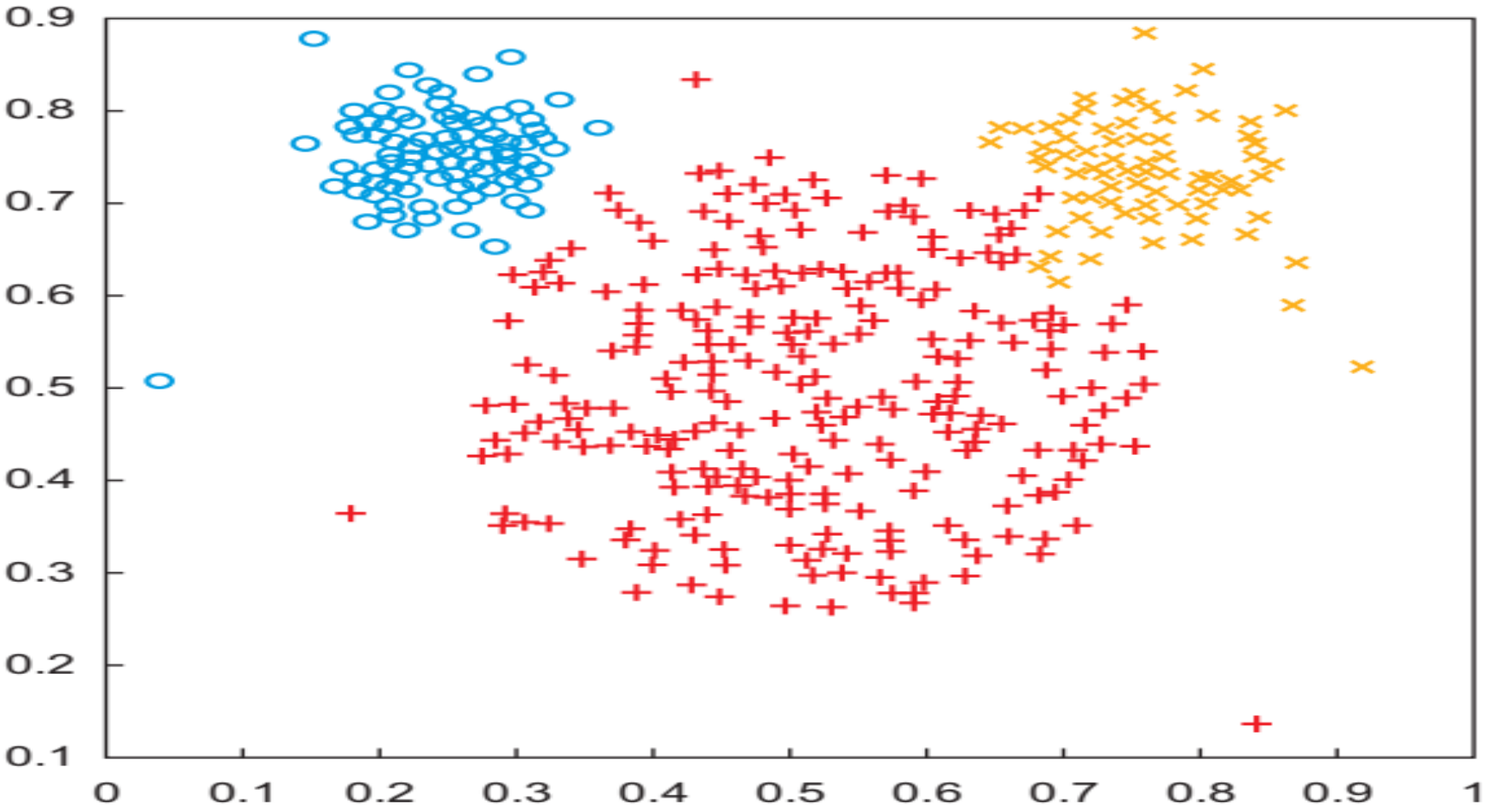
- We can try to visualize the clustering by projecting the data onto a space where one can see clear separation.

- We can also use hierarchical trees for illustrating samples being in the clusters.

Reminder

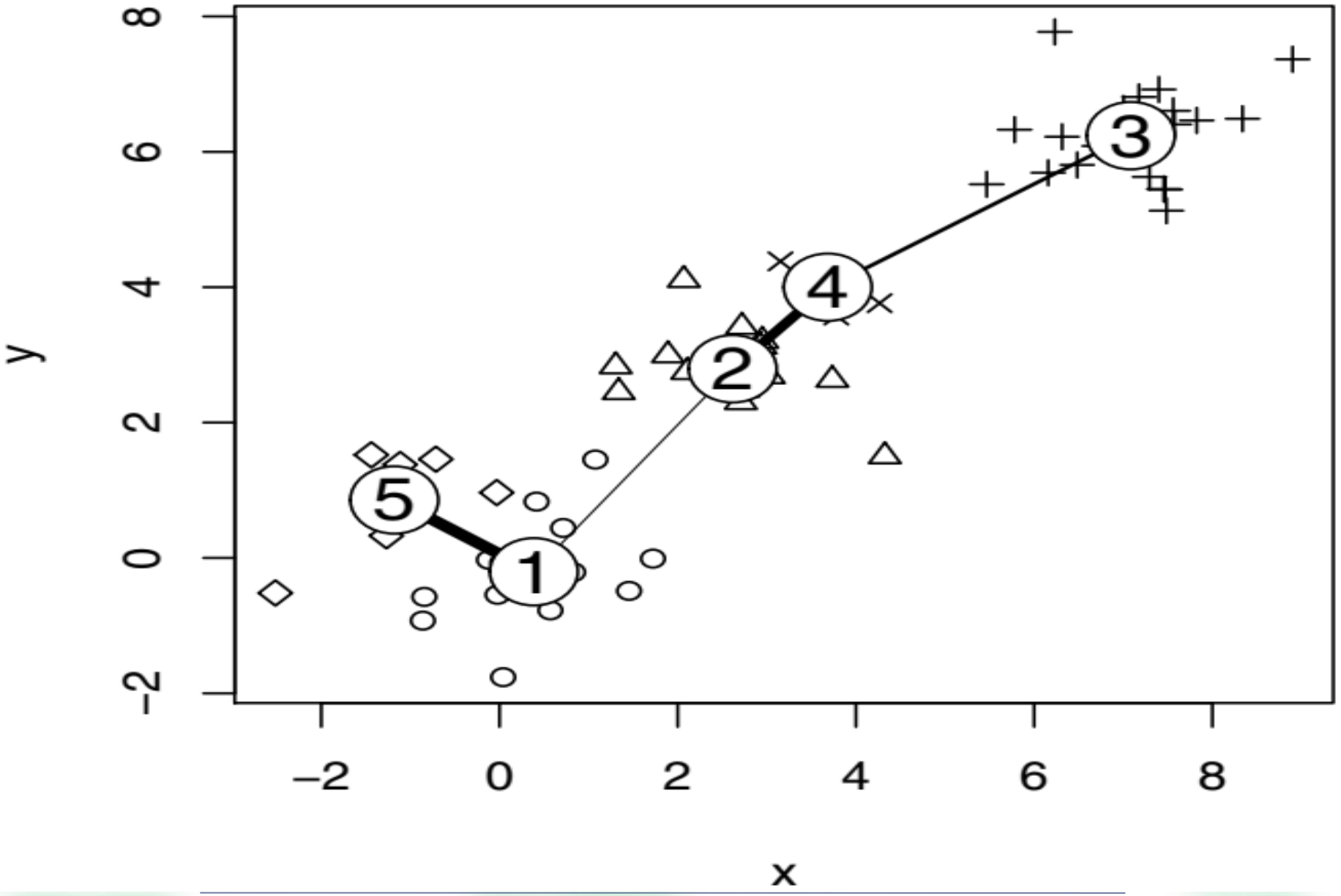
In clustering, we attempt to divide our data set into "*sufficiently distinct*" subsets, or *clusters*, wherein observations should be *similar* to those in the same cluster but *differ* greatly from the observations in other clusters.

No Name for Axis

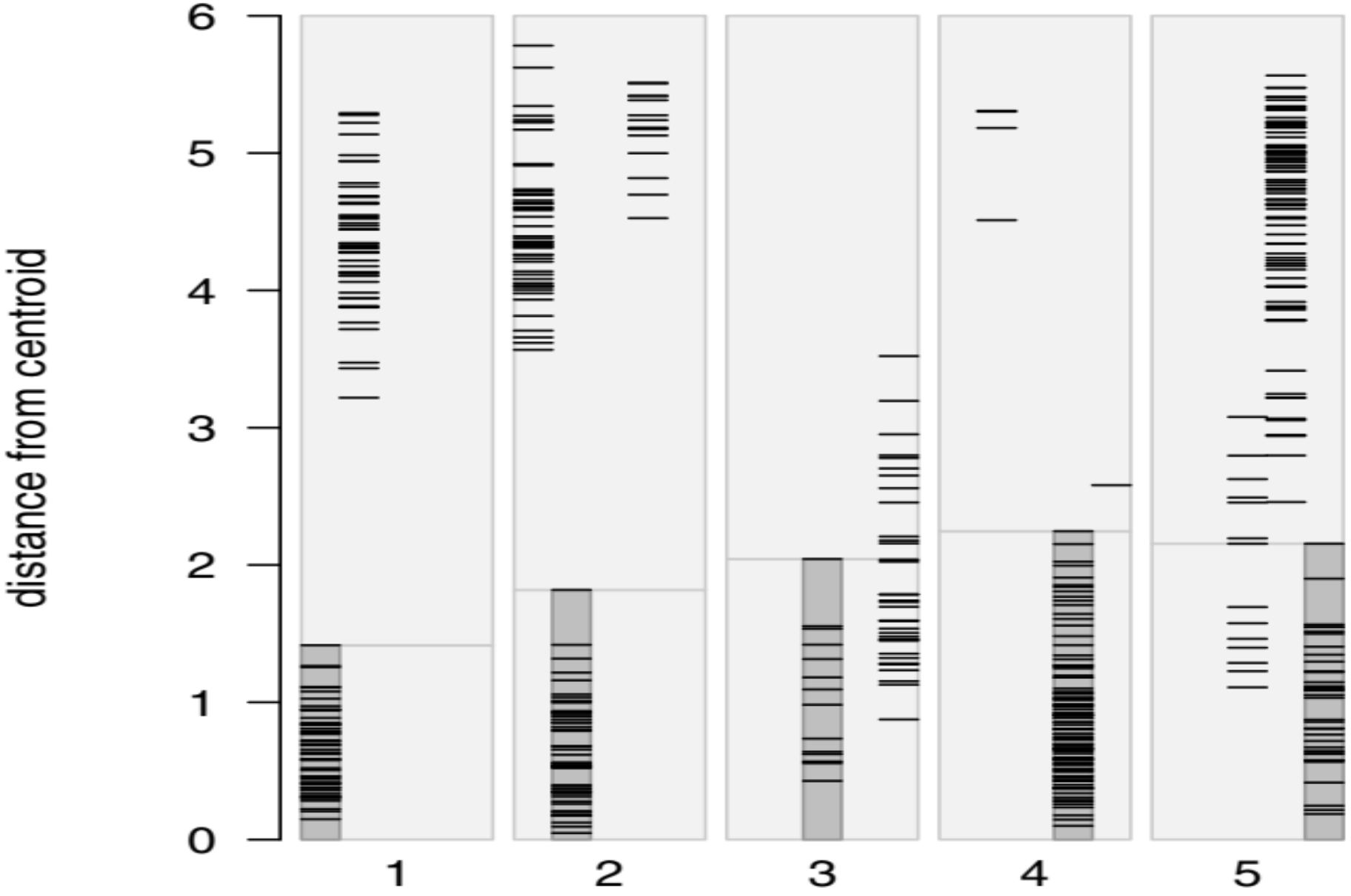


Dimension?

Neighboorhood graph



Strip Plot



Visualizing Clusters

- We can easily plot 1 and 2-D data to visualize possible clustering pattern.

How about clustering multivariate data?

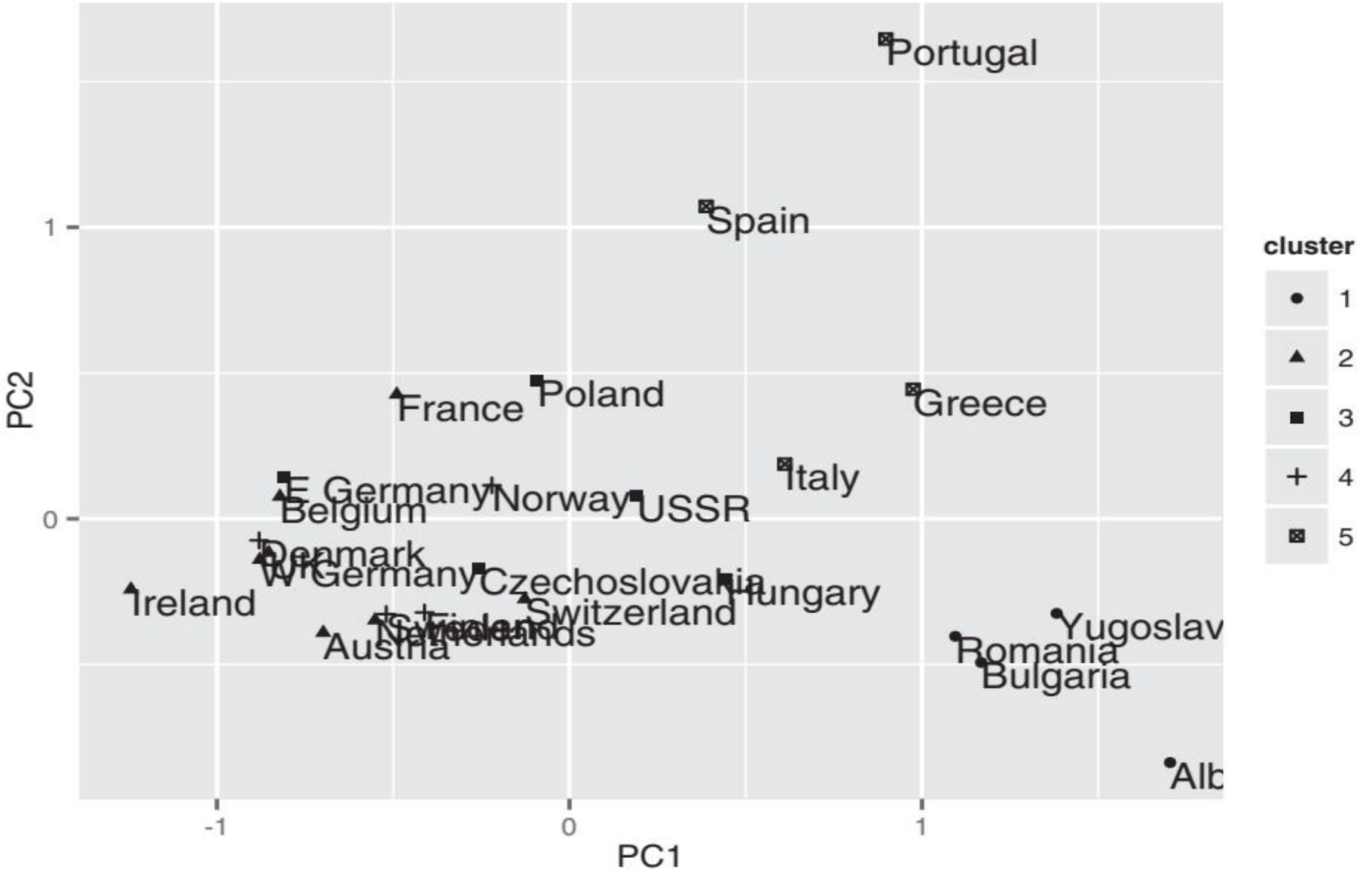
- We can project data onto any two PCs, but the first two are the most likely to show useful information.

About PC

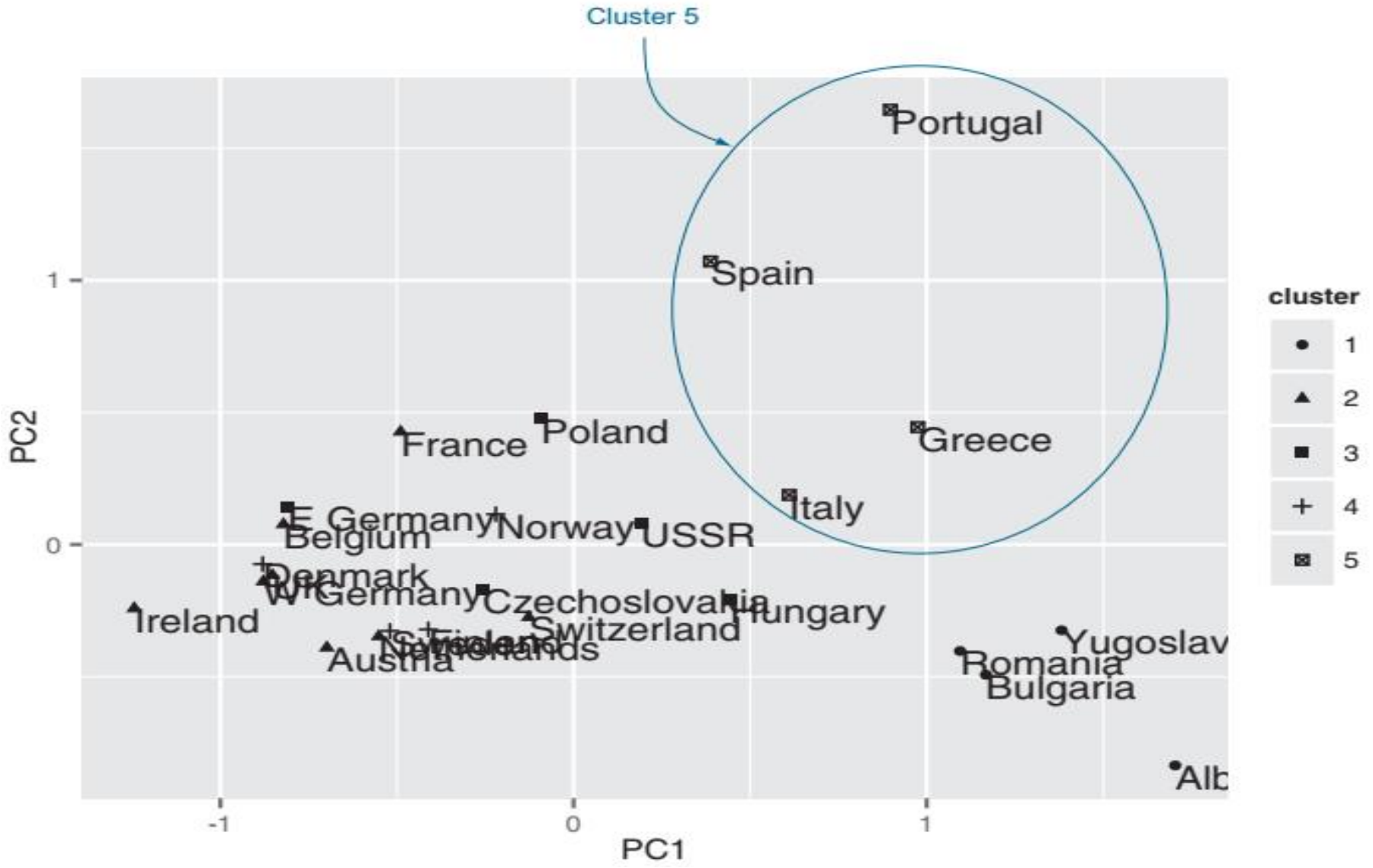
- If p is the number of variables that describe the data, then the principal components describe the hyperellipsoid in p-space that bounds the data.

- If you order the principal components by the length of the hyperellipsoid's corresponding axes (longest first), then the first two principal components describe a plane in p-space that captures as much of variation of data as can be captured in 2-D.

Mapping Using PCs



Clustering Using PC



Clustering Approaches

Methods

Two well-known clustering approaches are:

- Model-based
- No Model (Exploratory)

Popular types of clustering methods are:

Agglomerative hierarchical methods (AHM)
K-means type methods (KTM)
Maximum Likelihood methods (MLM)

A Quick View

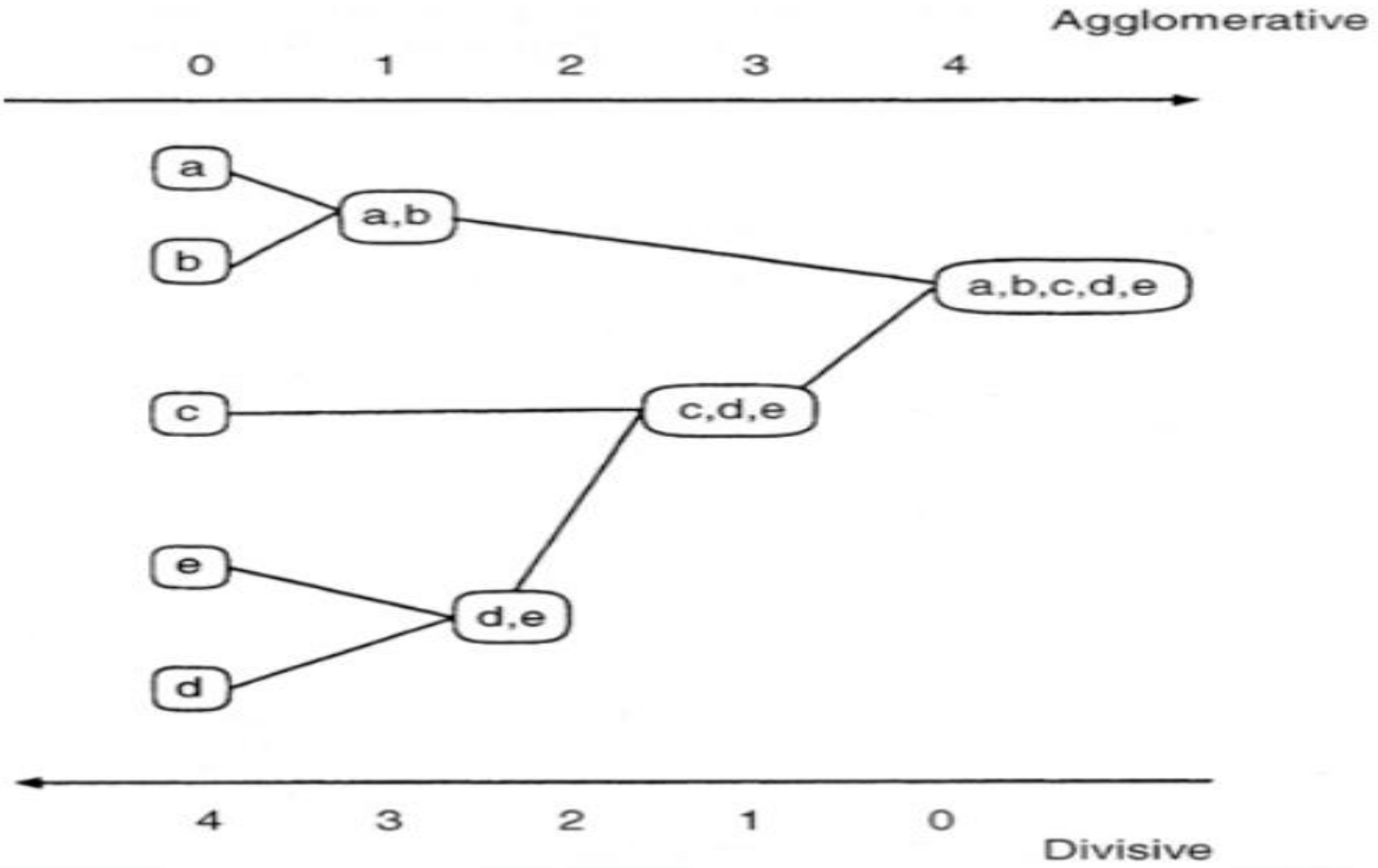
- *Hierarchical clustering* finds nested groups of clusters. An example of hierarchical clustering might be the standard plant taxonomy, which classifies plants by family, then genus, then species, etc.

- *k-means*, which is a quick and popular way of finding clusters in quantitative data.

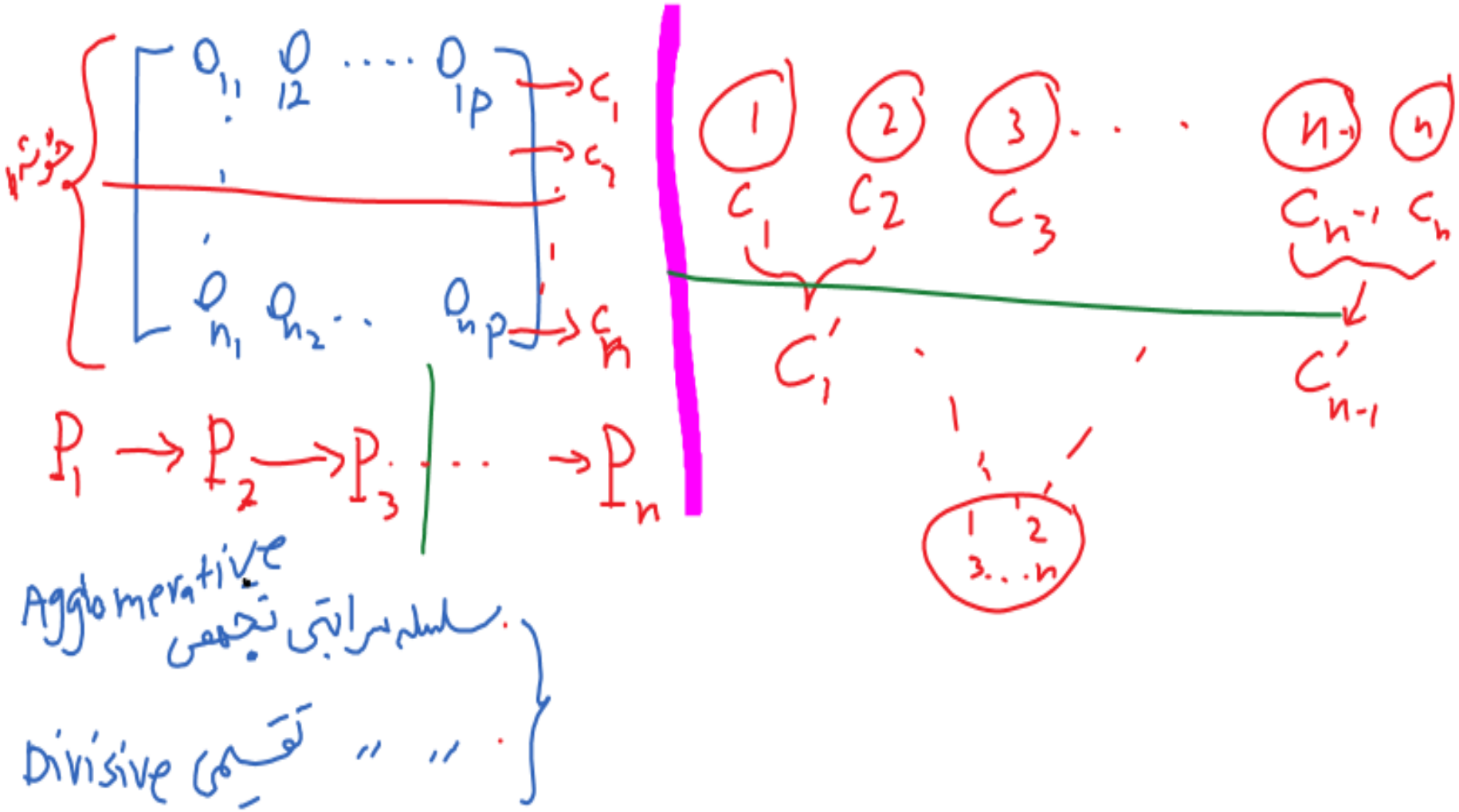
- *Model-based*, which take an statistical model into account, estimate parameters as well as probability of membership for each sample.

Hierarchical Clustering

Schematic



Hierarchical Clustering



Fact about AHM

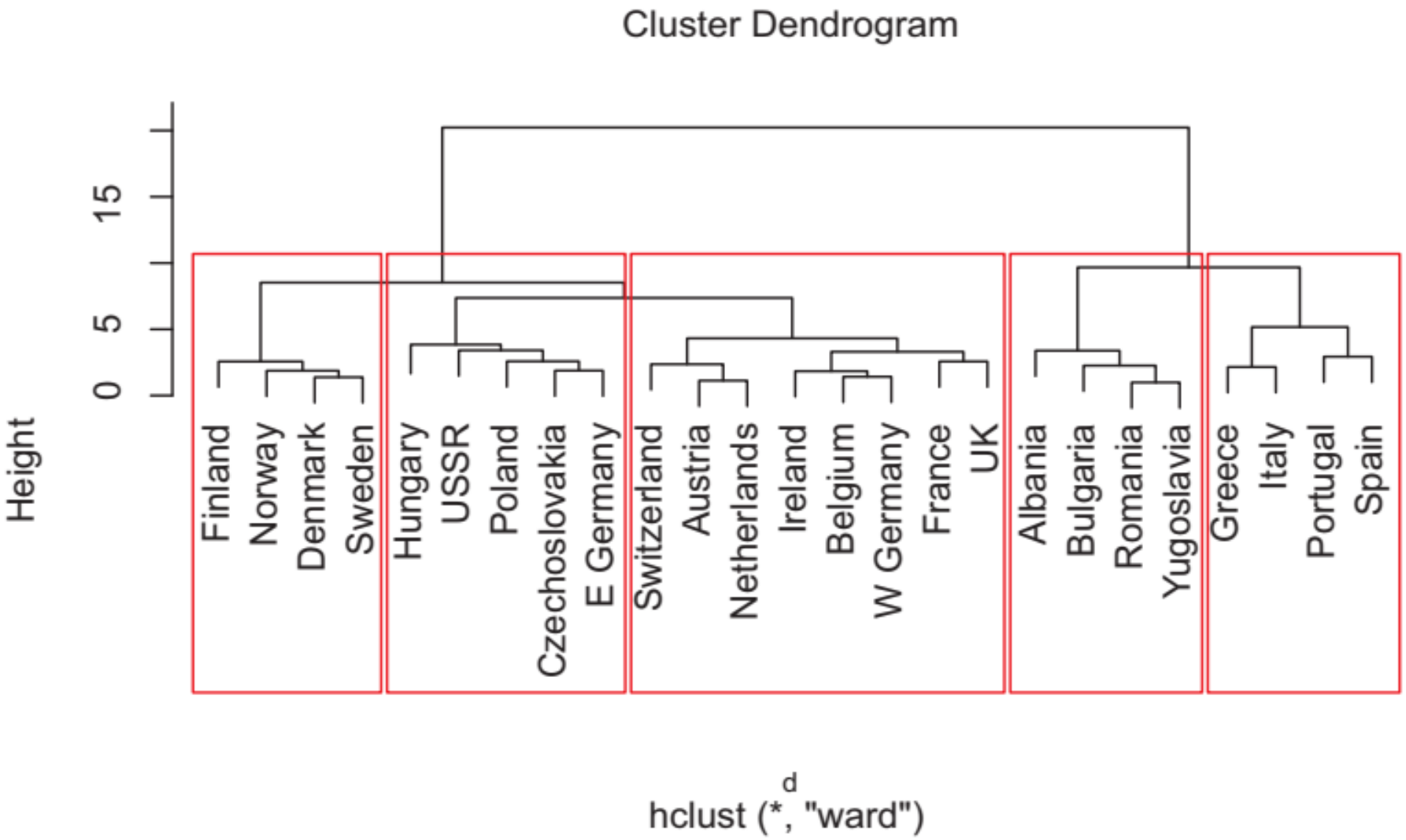
- Data are not divide into a particular clusters at a single step.
- It produces partitions by a series of successive fusions of individuals.
- It is irreversible; setting in a group cannot back.

Dendrogram

- Fusions at each stage of hierarchical analysis are monitored by a graph called dendrogram.

- The graph resembles an evolutionary tree, the subject more popular in biological application.

Hierarchical



Stop Where?

Two odds cases of hierarchical clustering are

- all in one cluster

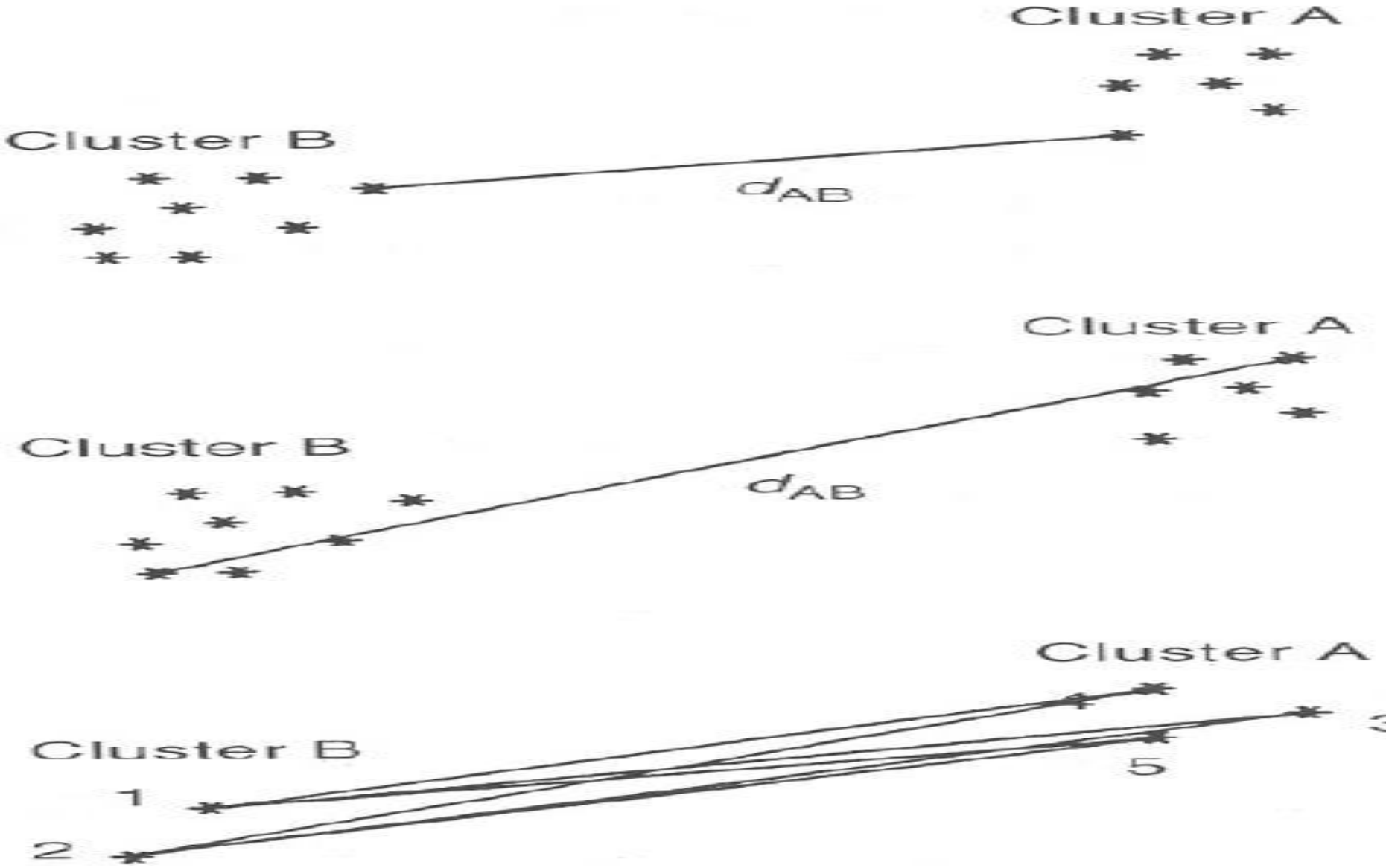
and

- each in one cluster.

Hence, one requires to decide which division is best.

Inter-cluster Dissimilarity (linkage type) in Hierarchical Clustering

Measuring Inter-cluster Dissimilarity (1)



Measuring Inter-cluster Dissimilarity (2)

- Intergroup dissimilarity measures are:

  single, complete and average linkage

- First two are invariant under monotone transformations of original inter-individual dissimilarities or distances.

Algorithm of AHM

P1,…,Pn are partitions representing single and n groups. Basis of all procedures are similar:

Start: each clusters C1,…,Cn has single member.

(I) Find nearest pairs of distinct clusters, merge them and omit one of them

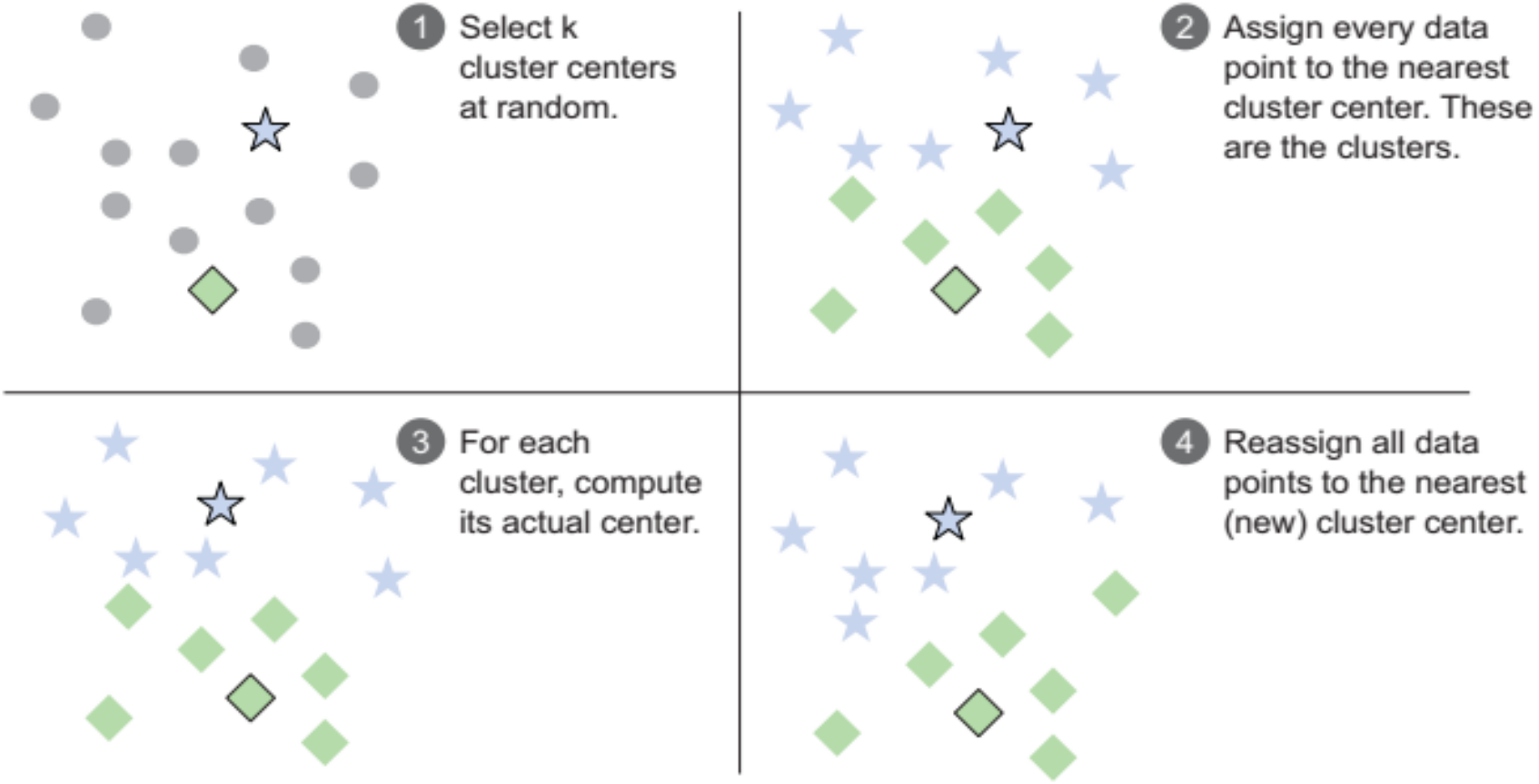
(II) If #clusters reaches one, stop. O.W. go to (I).

Note: Methods start with inter-individual distance matrix.

K-Means Clustering

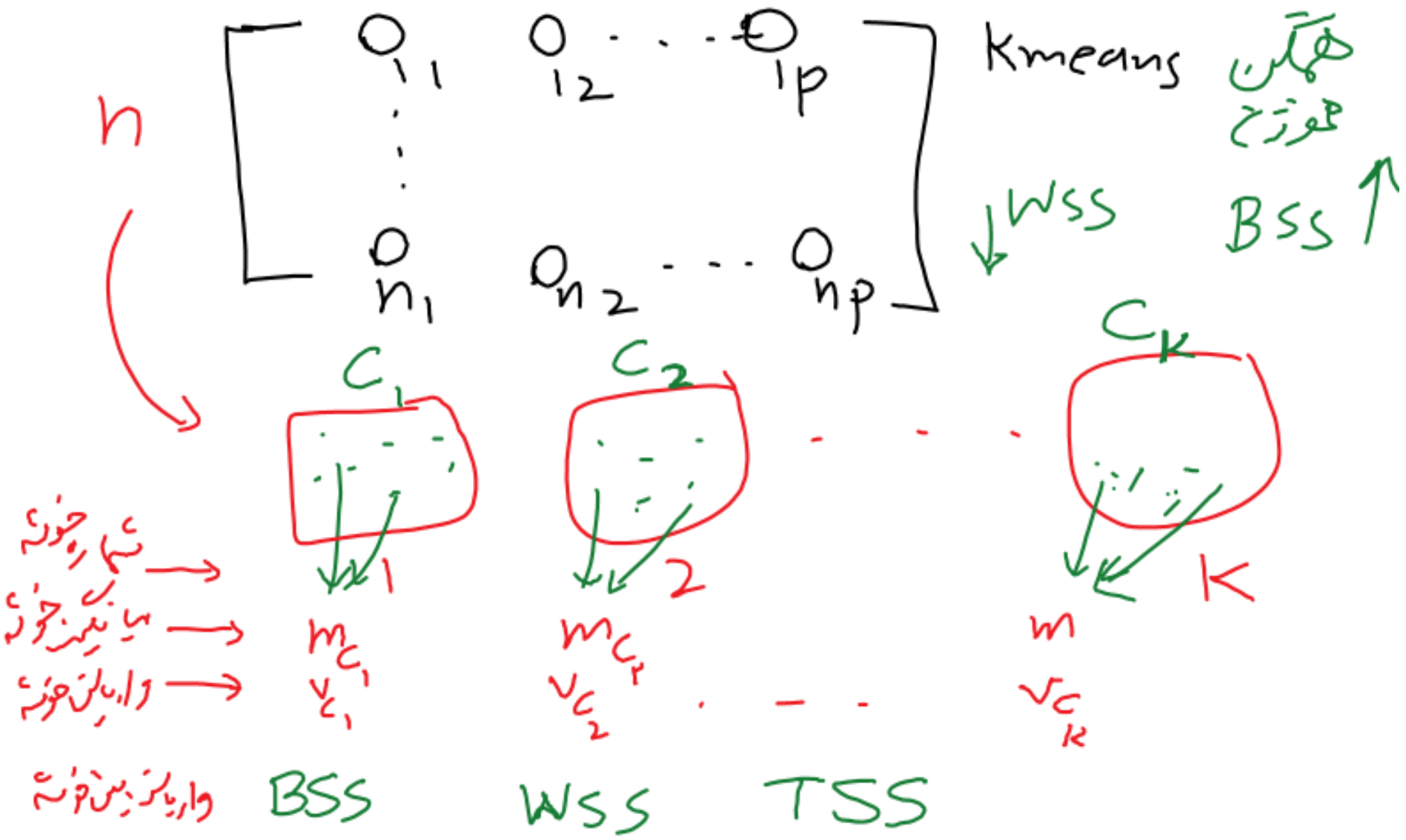
K-Means Algorithm

Cluster Dendrogram



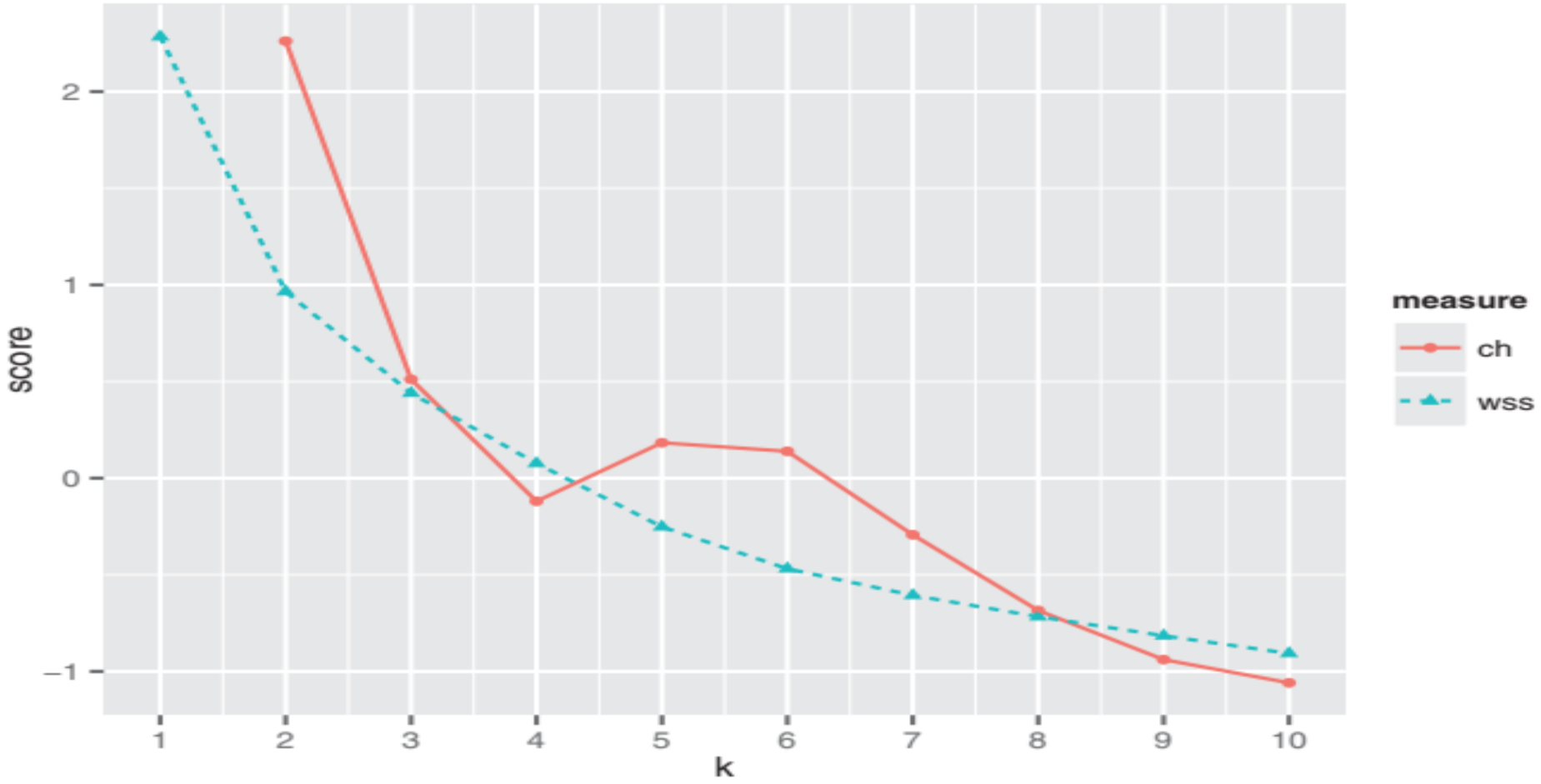
5 Repeat steps 3 and 4 until the points stop moving, or you have reached a maximum number of iterations.

K-Means Clustering



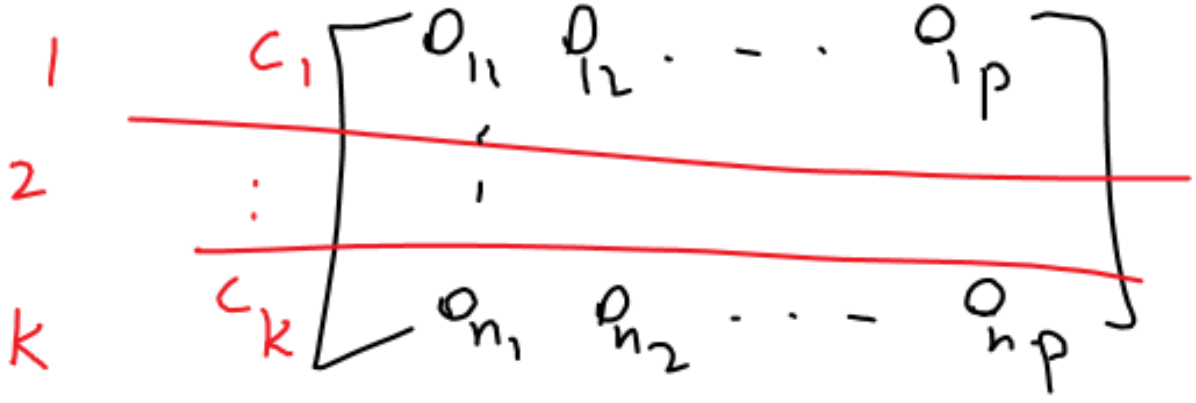
Numbers of Clusters

Total within sum of squares
Calinski-Harabasz index



Model-Based Clustering

Model-Based Clustering



مدل مبنا

توزیع آماری

MVN

k جامعه

$O \sim N(\mu_p, \Sigma_p)$ p-متغیر

یک جامعه $\rightarrow$ n تعداد نمونه

$$f(O) = \alpha_1 f_{C_1} + \alpha_2 f_{C_2} + \cdots + \alpha_k f_{C_k} = \sum_{i=1}^{k} \alpha_i f_{C_i}$$

$$\alpha_1 + \alpha_2 + \cdots + \alpha_k = 1$$

Technical Notations

The likelihood function is : $$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i; \boldsymbol{\theta}_{\gamma_i})$$

Using formal multivariate normal density, likelihood is

$$L(\boldsymbol{\theta}; \boldsymbol{\gamma}) = \text{const} \prod_{k=1}^{c} \prod_{i \in E_k} |\Sigma_k|^{1/2} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)' \sum_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\}$$

Clearly, this leads $\bar{\mathbf{X}}_j = n_i^{-1} \sum_{i \in E_i} \mathbf{X}_i$ replacing this gives logL

$$l(\theta, \gamma) = \text{const} - \frac{1}{2} \sum_{i=1}^{c} \text{trace}(\mathbf{W}_j \boldsymbol{\Sigma}_j^{-1} + n \log |\boldsymbol{\Sigma}_j|)$$

Where Wj is p.p matrix of SSP of vars for subpop j.

Fraley and Raftery (2002) used EM algorithm for MLE and also Bayesian method using BIC.

Technical Notations (2)

- size, shape and rotation are highly affecting the consequence of model-based clustering.

- Different models are proposed under the title mclust.

A Web

One can find a comparison of all the clustering algorithms in Scikit-learn at

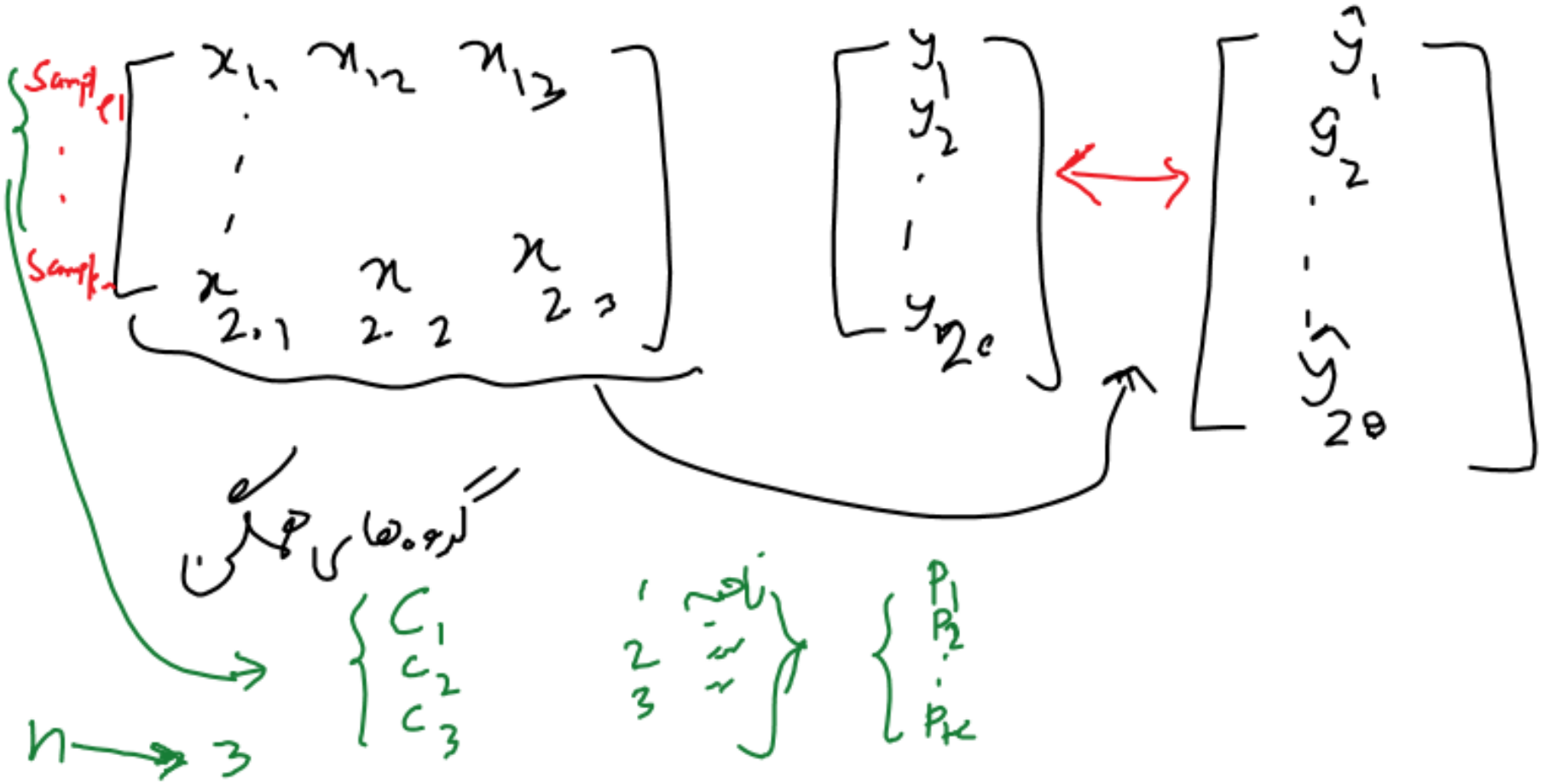
http://scikit-learn.org/stable/modules/clustering.html

Clustering's Results Evaluation

Critical Question

- whether a given cluster is “real”?
- does the cluster represent actual structure in the data?
- is it an artifact of clustering algorithm?

Cluster Evaluation (Simple)



Evaluating clustering (1)

- Clustering models are hard to evaluate because they're unsupervised.
- Clusters that items are assigned to are generated by modeling procedure, not supplied in a series of annotated examples.
- Evaluation is largely checking observable summaries about clustering.

Evaluating clustering (2)

INTRA-CLUSTER DISTANCES VERSUS CROSS-CLUSTER DISTANCES

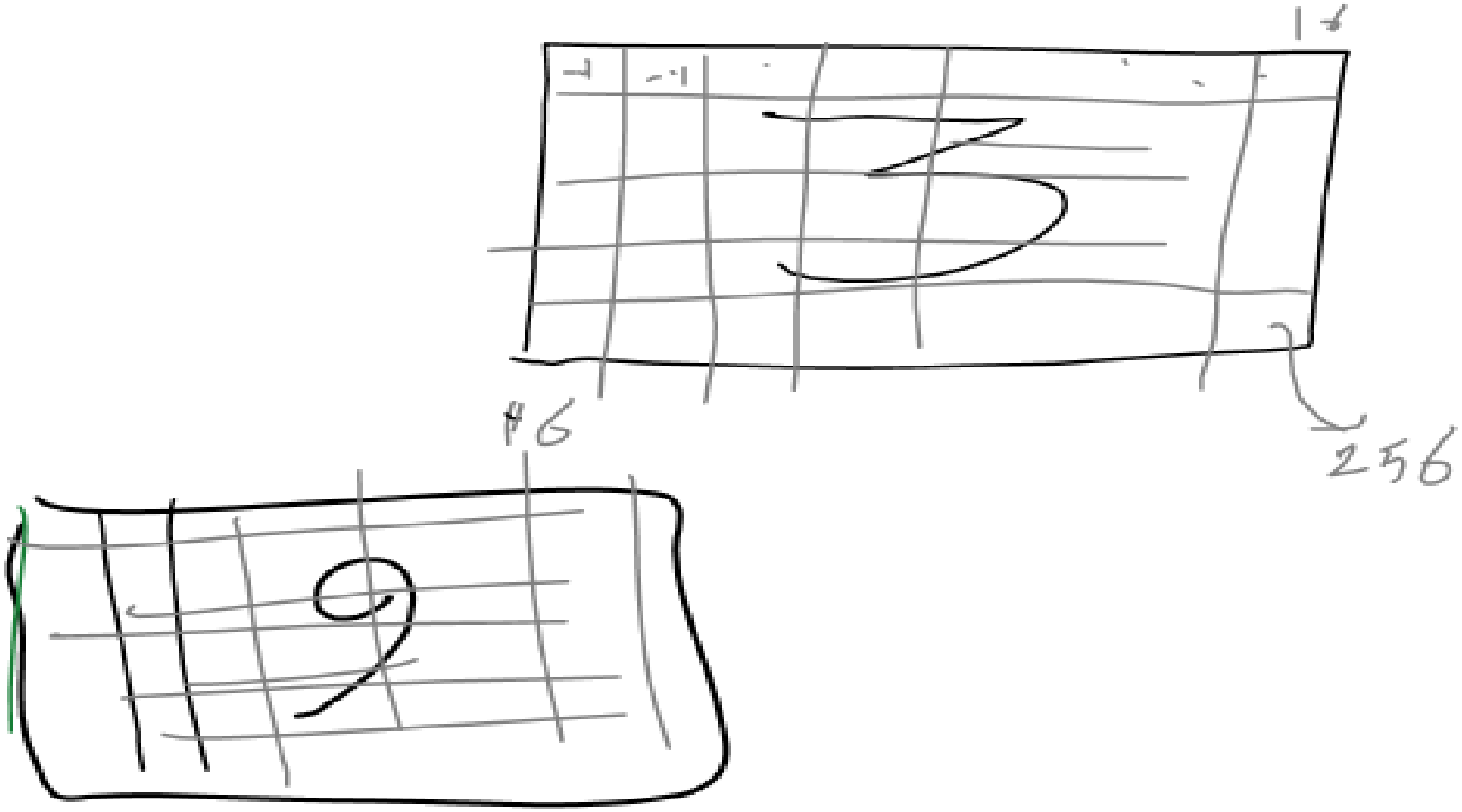
The traditional measure of this is comparing typical distance between two items in same cluster to typical distance between two items from different clusters.

TREATING CLUSTERS AS CLASSIFICATIONS OR SCORES

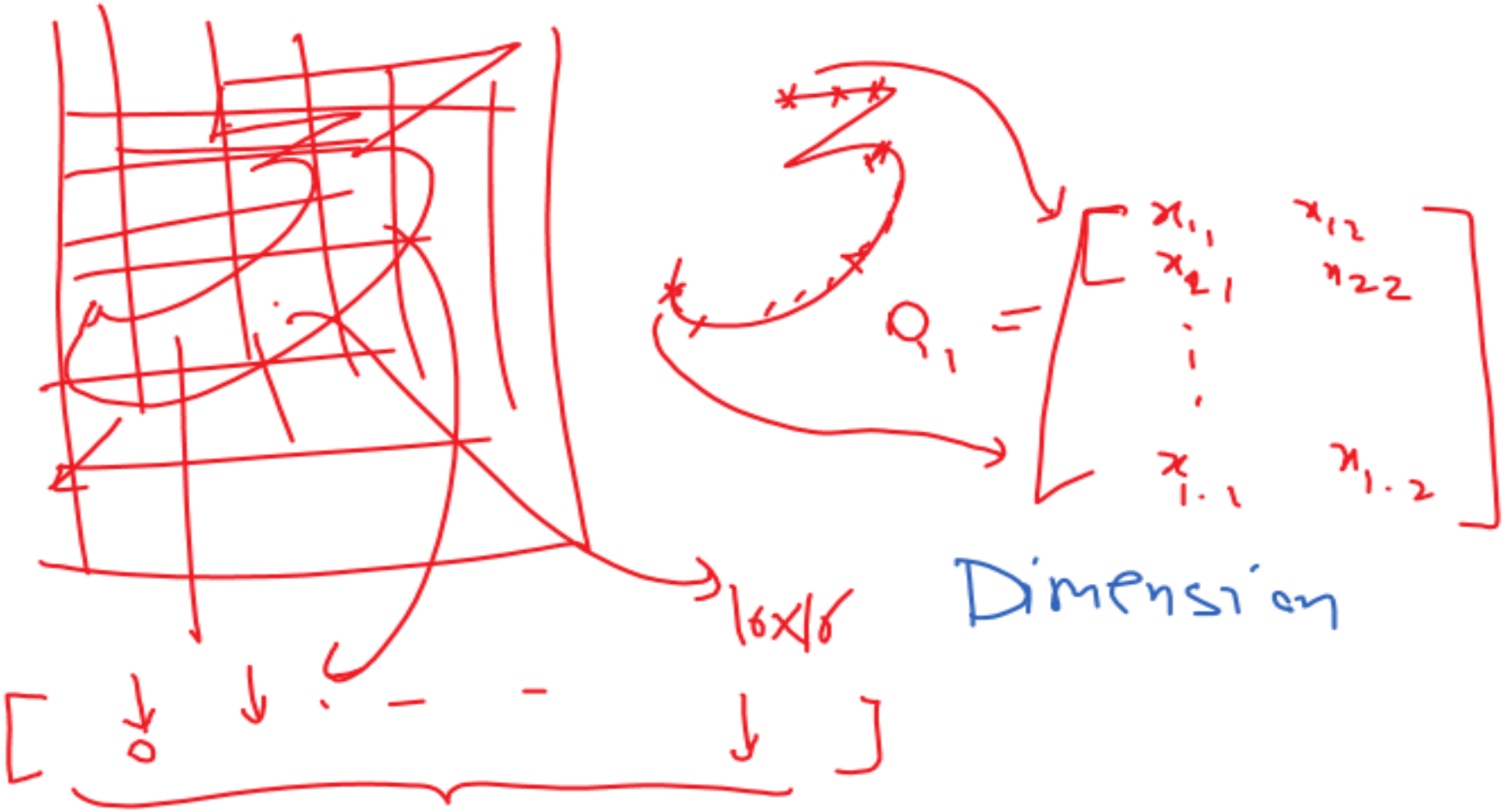
It is advised treating the cluster assignment as if it were a classification.

Clustering in New Era

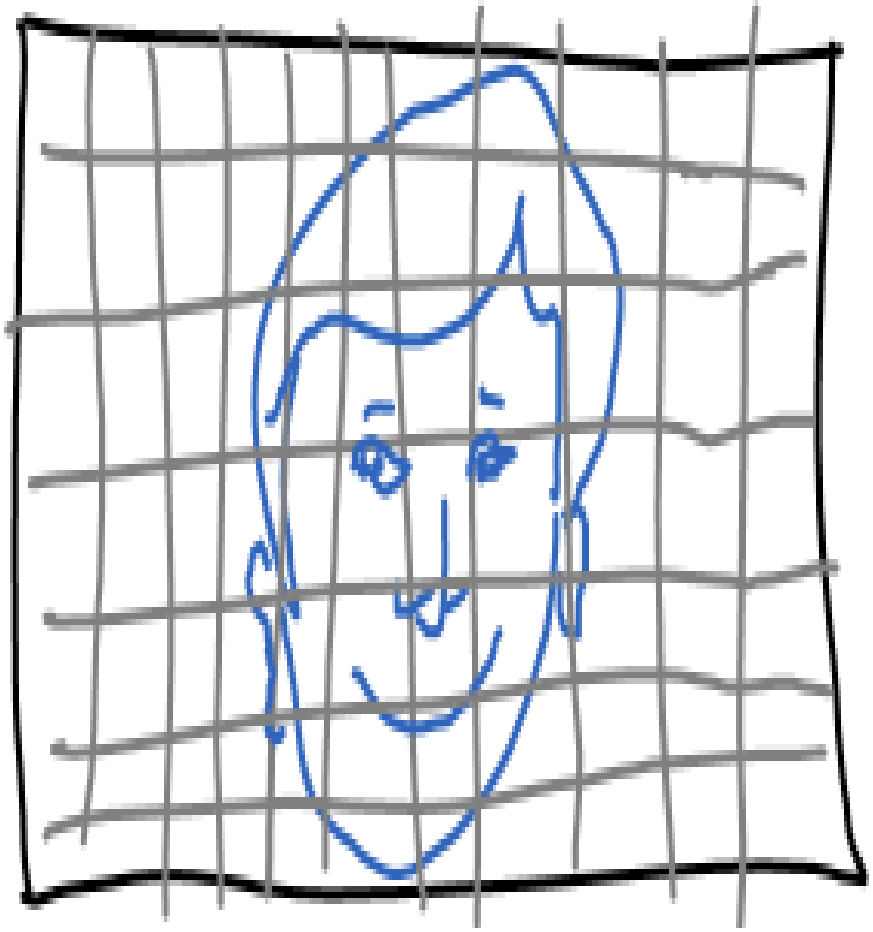
Clustering Images

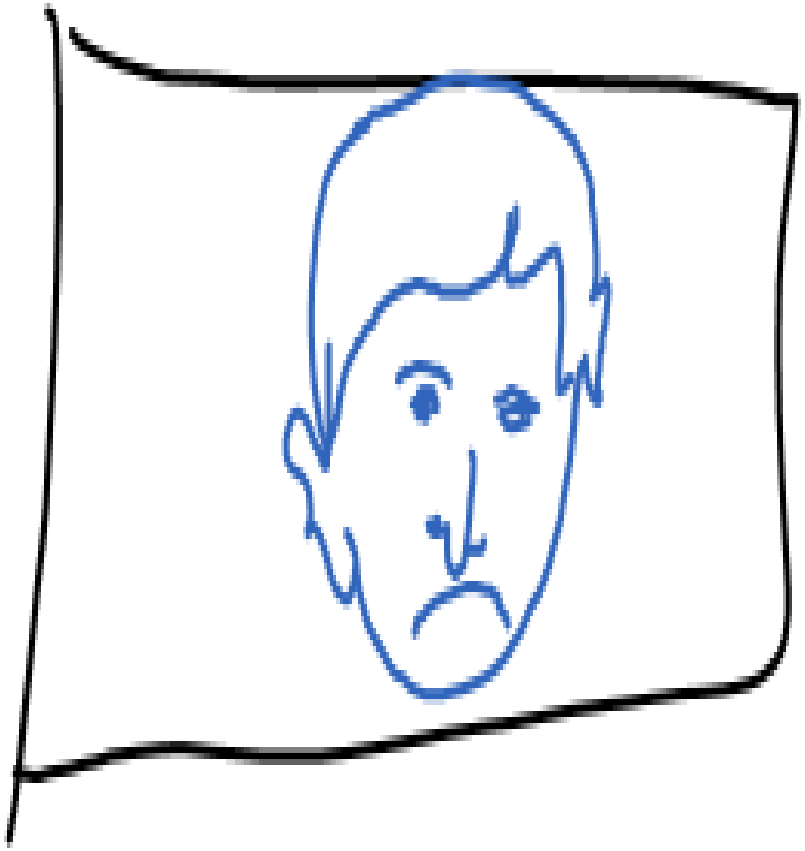


Tricks to Cluster Images

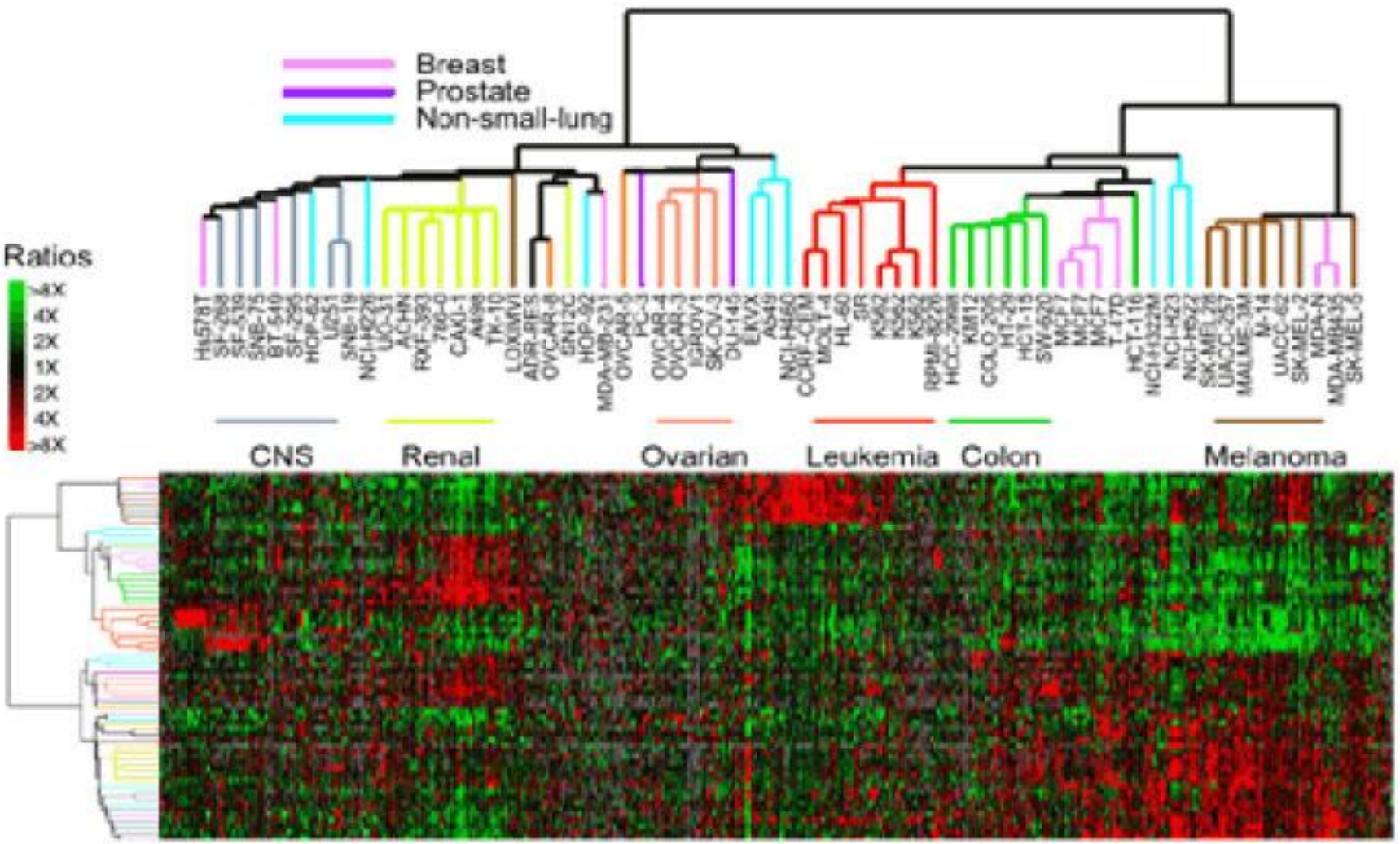


Clustering Shapes (Faces)





Microarray Clustering



KNN

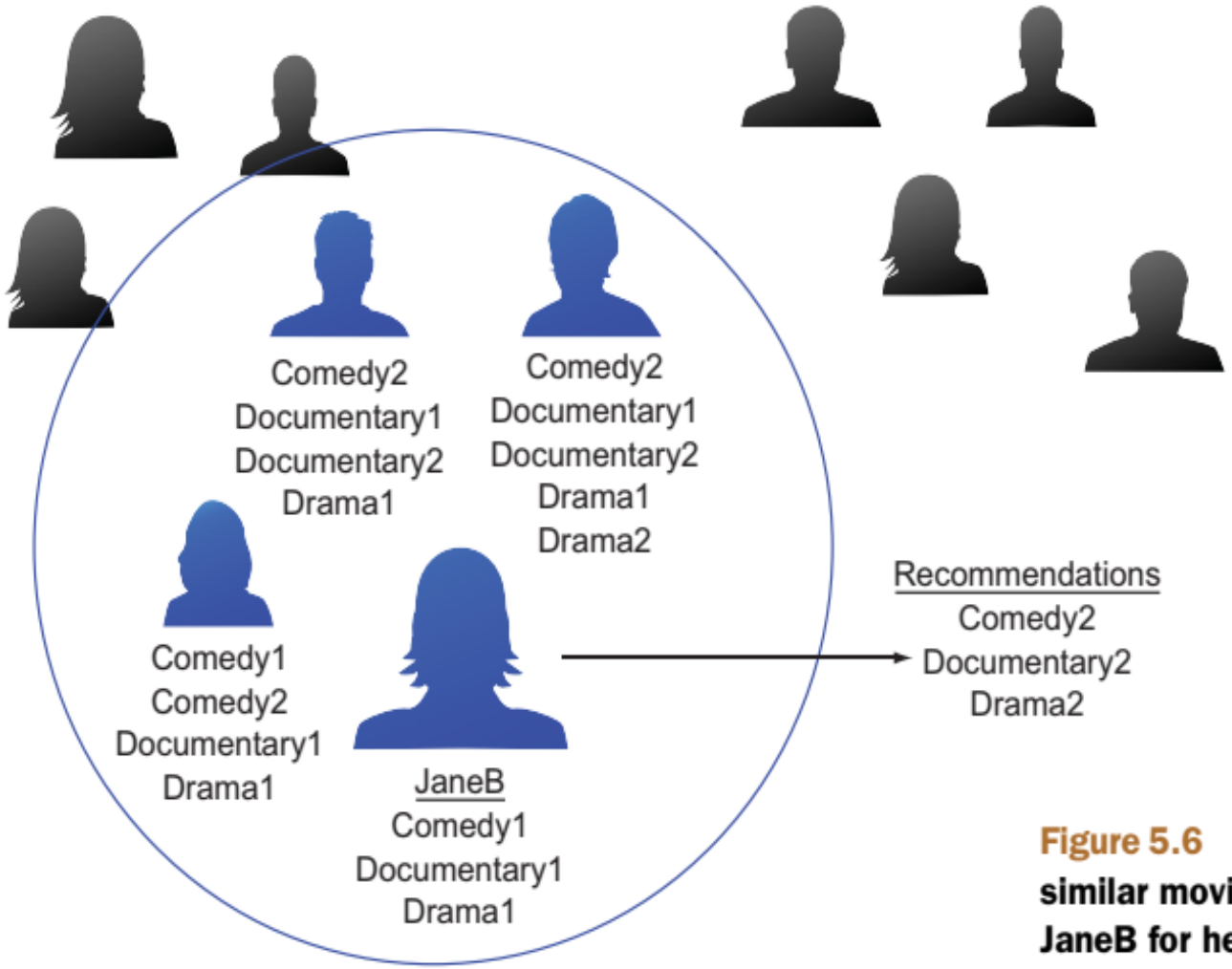
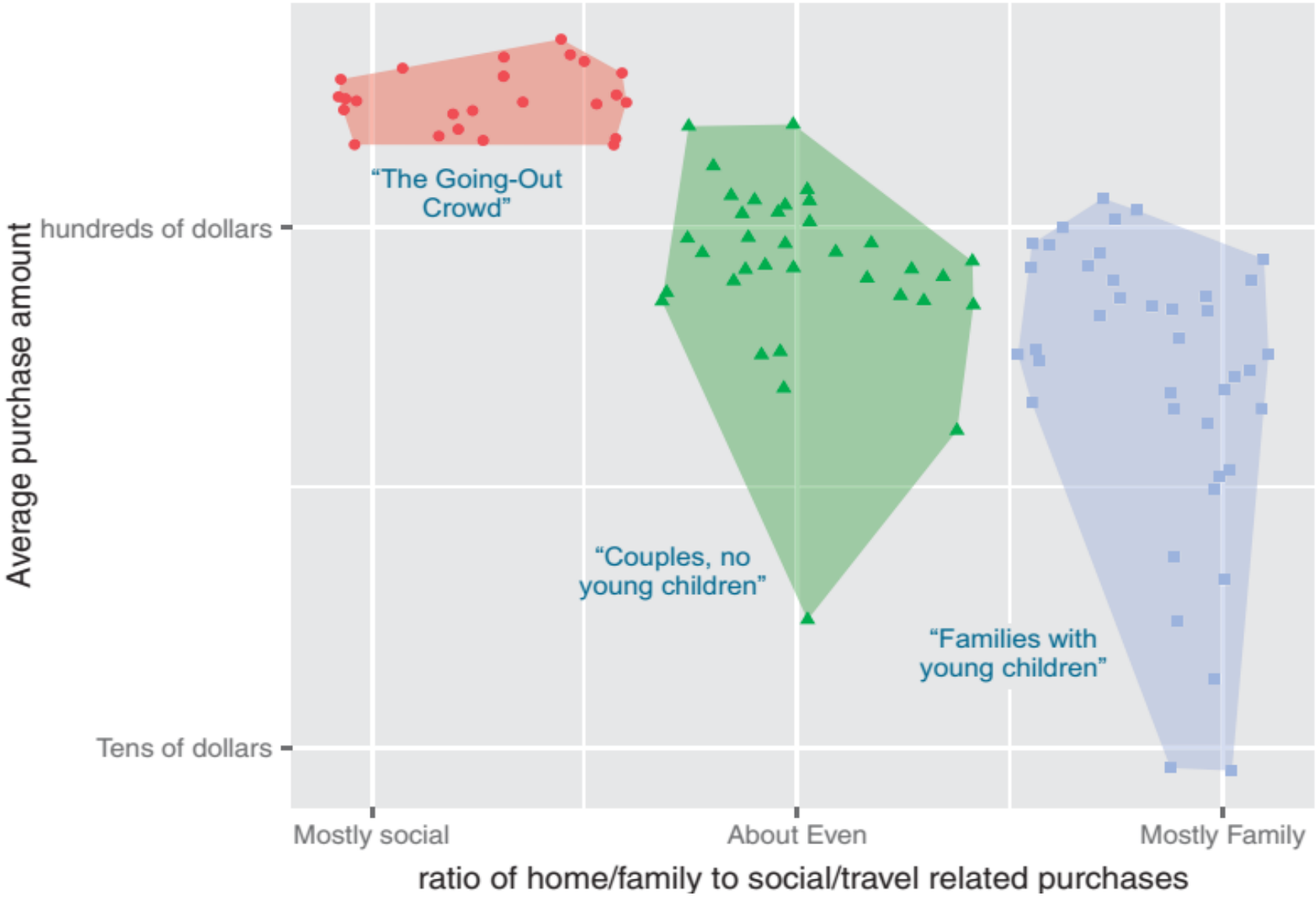


Figure 5.6 **Look to the customers with similar movie-watching patterns as JaneB for her movie recommendations.**

Association Rule



Other Clusterings

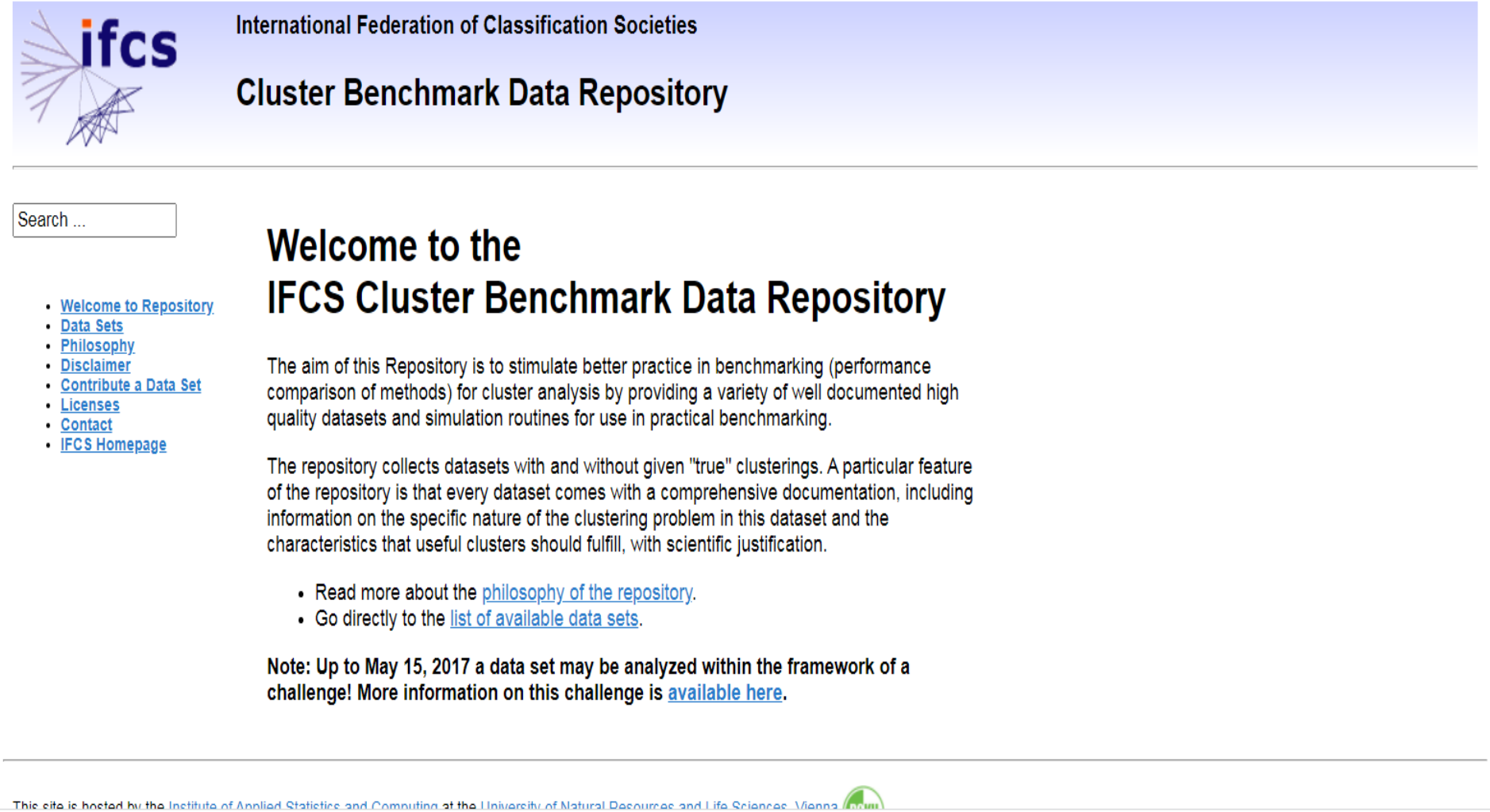
Text clustering

Fuzzy clustering

Manifold clustering

....

Database



https://ifcs.boku.ac.at/repository/

Time To Stop?

Summary (1)

- Cluster analysis is a topic worthy of a day discussion or a book in itself.
- Due to time limitation, we discussed only some selected approaches.
- Summaries here are only based on this talk.

Summary (2)

- Clustering is to discover or draw out similarities among subsets of your data.
- Points in the same cluster should be more similar (nearer) to each other than they are to points in other clusters.
- Different units cause different distances and potentially different clustering results.

Summary (3)

- Clustering is often used for data exploration or as a precursor to supervised learning methods.

- Clustering, like visualization, is more iterative and interactive, and less automated than supervised methods.

Summary (4)

- Different clustering algorithms will give different results. So, consider different approaches, with different #clusters.
- There are many heuristics methods for estimating the best number of clusters.
- Clustering is a task to do not to learn!

Acknowledges

You all

Dr. Daneshgar & Dr. Amini

My students in TMU

Authors

Time is yours for raising questions!