An Introduction to Cluster Analysis

Cosmin Lazar

ASC Engineering Department Bosch Cluj

Introduction

"Cluster analysis foundations rely on
one of the most fundamental, simple
and very often unnoticed ways (or methods)
of understanding and learning,
which is grouping similar objects into groups."

Overview

- What is cluster analysis?
- Some definitions and notations
- How it works?
- Cluster Analysis Diagram
  - Objectives of cluster analysis
  - Research design issues
  - Assumptions in cluster analysis
  - Clustering methods
  - Interpreting the clusters
  - Validation

What is cluster analysis?

What is a cluster?

No general accepted definition!!!

What is cluster analysis?

What is a cluster?

No general accepted definition!!!

A cluster is …

What is cluster analysis?

What is a cluster?

No general accepted definition!!!

A cluster is …

D1: … comprised of a number of *similar* objects collected and grouped together

What is cluster analysis?

What is a cluster?

No general accepted definition!!!

A cluster is …

D1: … comprised of a number of *similar* objects collected and grouped together

D2: … a set of entities which are *alike*, and entities from different clusters are not *alike*

What is cluster analysis?

What is a cluster?

No general accepted definition!!!

A cluster is …

D1: … comprised of a number of *similar* objects collected and grouped together

D2: … a set of entities which are *alike*, and entities from different clusters are not *alike*

D3: … an aggregation of points in the test space such that the *distance* between any two points in the cluster is less that the distance between any point in the cluster and any point not in it.

What is cluster analysis?

What is a cluster?

No general accepted definition!!!

A cluster is …

D1: … comprised of a number of *similar* objects collected and grouped together

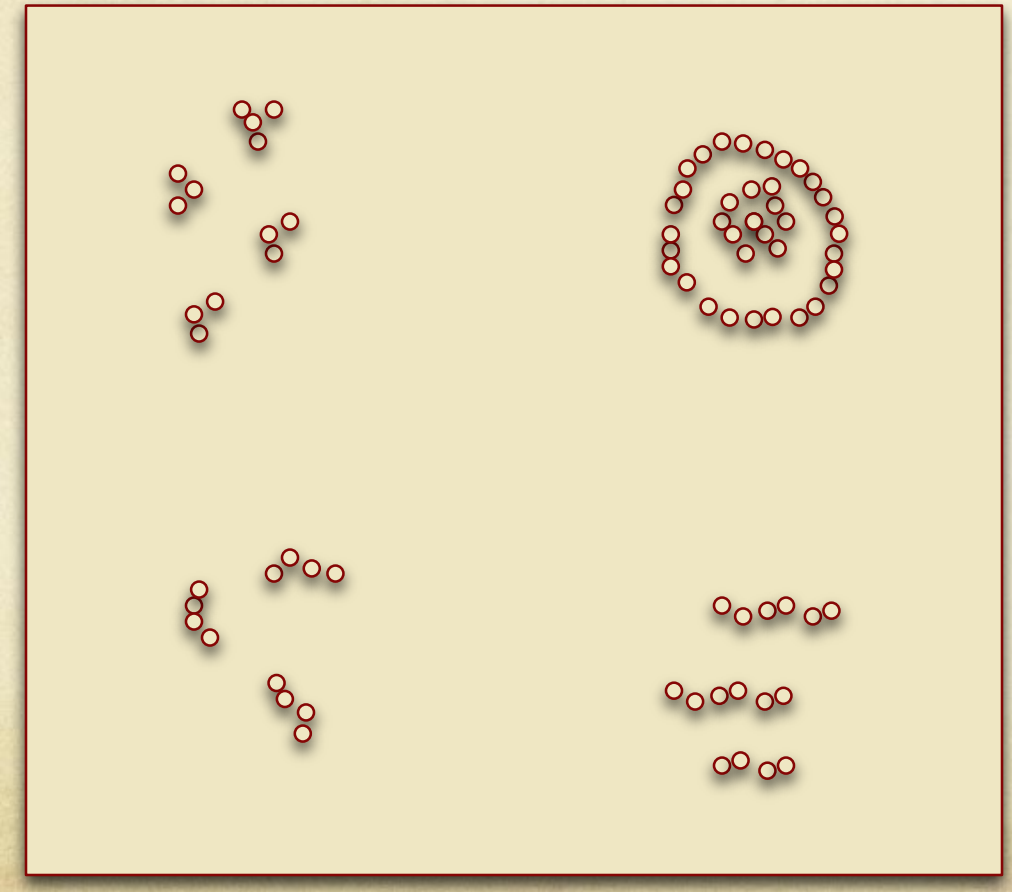
D2: … a set of entities which are *alike*, and entities from different clusters are not *alike*

D3: … an aggregation of points in the test space such that the *distance* between any two points in the cluster is less that the distance between any point in the cluster and any point not in it.

D4: … a connected region of a multidimensional space containing a relative *high density* of points, separated from other such regions by regions containing a relatively low density of points.

What is cluster analysis?

It is hard to give a general accepted definition of a cluster because objects can be grouped with different purposes in mind.



Humans are excellent cluster seekers

...only in two or three dimensions.

What is cluster analysis?

Cluster analysis is a multivariate data mining technique whose goal is to groups objects based on a set of user selected characteristics (or features)

Clusters should exhibit high internal homogeneity and high external heterogeneity

What does this mean?

When plotted geometrically, objects within clusters should be very close together and clusters will be far apart.

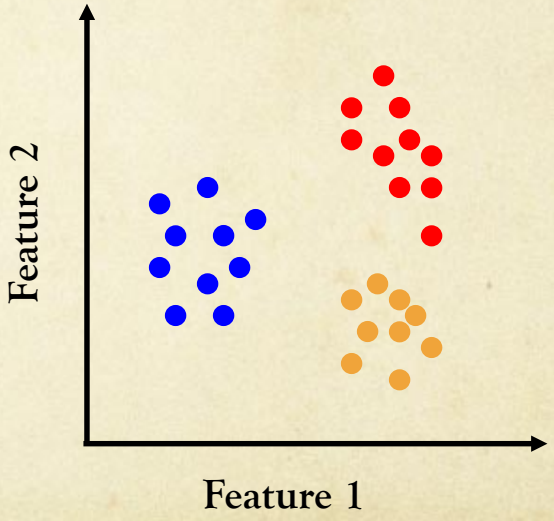
What is cluster analysis?

Cluster analysis is a multivariate data mining technique whose goal is to groups objects based on a set of user selected characteristics (or features)

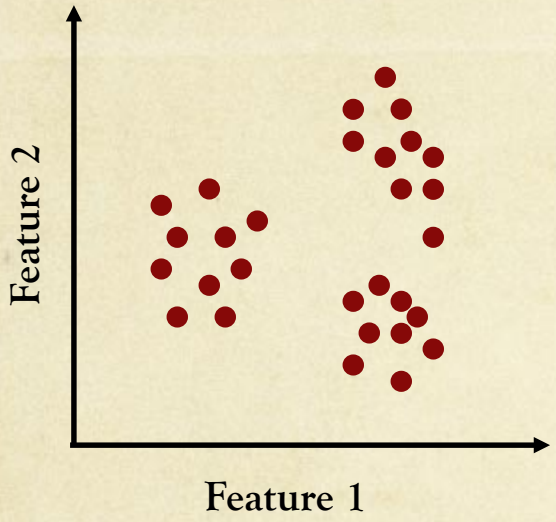
Clusters should exhibit high internal homogeneity and high external heterogeneity

What does this mean?

When plotted geometrically, objects within clusters should be very close together and clusters will be far apart.



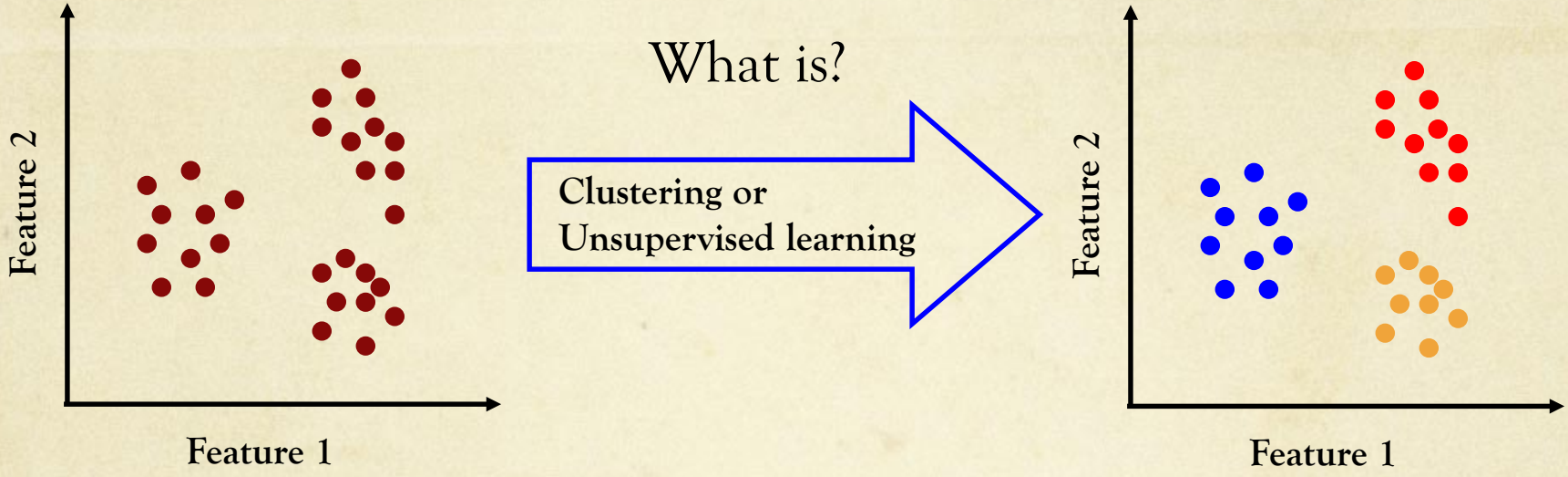
What is cluster analysis?



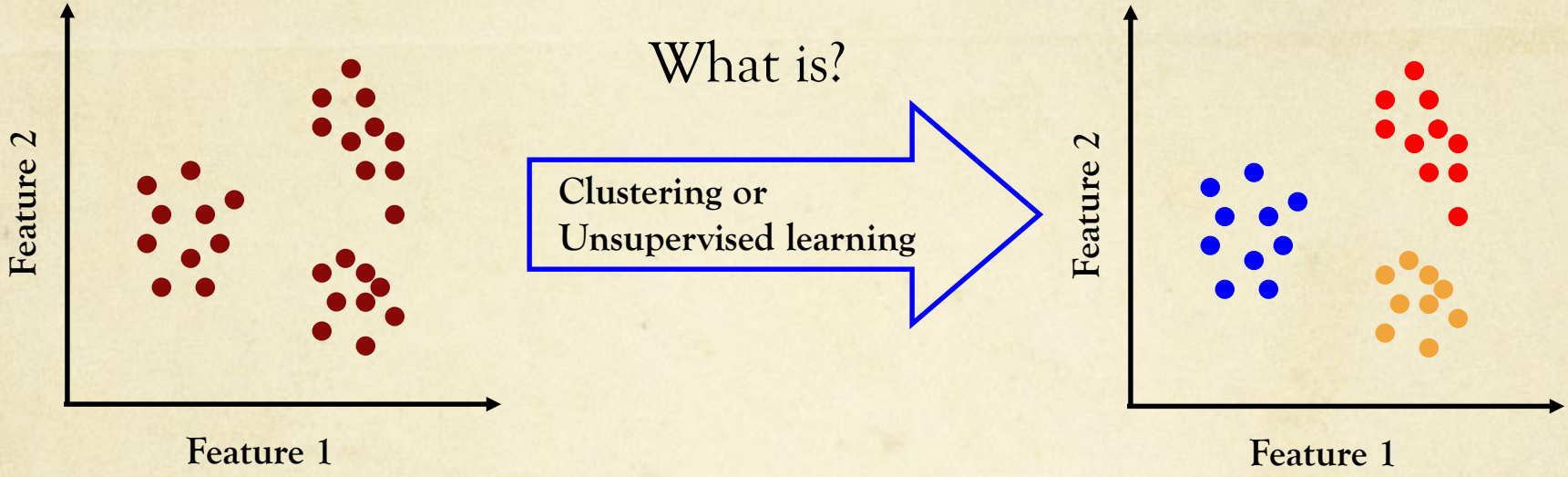
What is?

Clustering or
Unsupervised learning

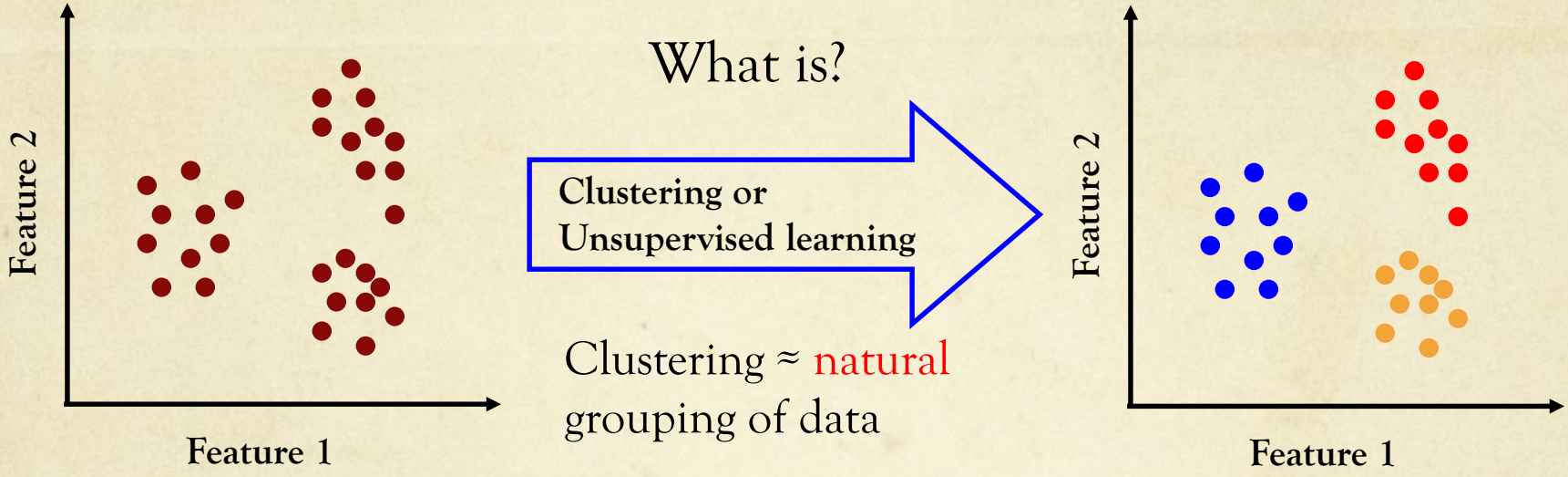
What is cluster analysis?



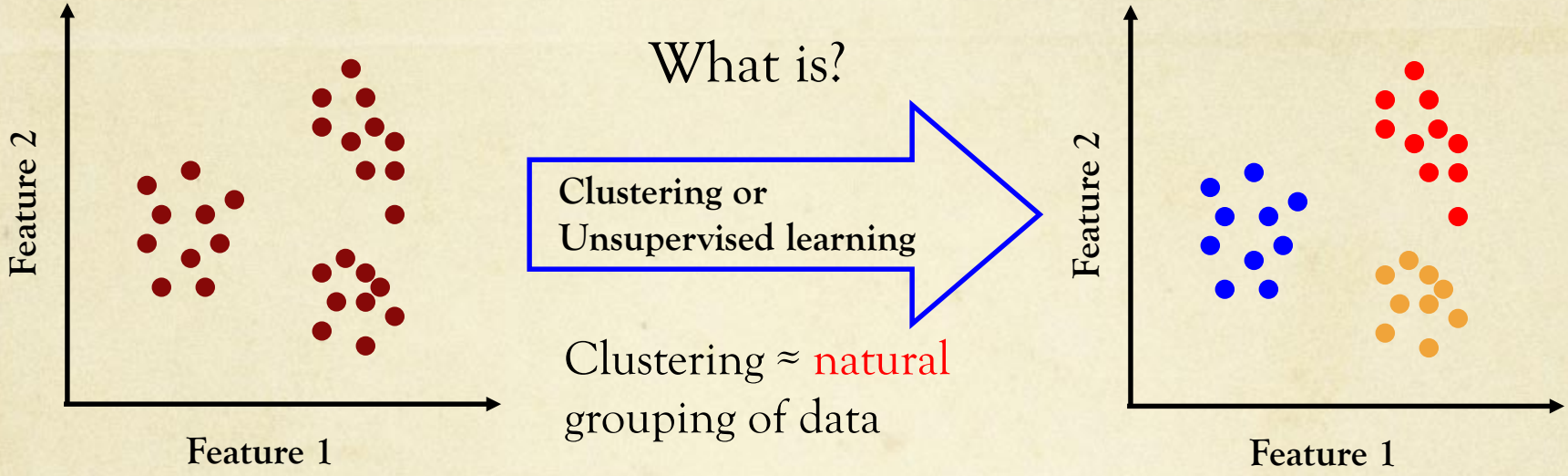
What is cluster analysis?



What is cluster analysis?

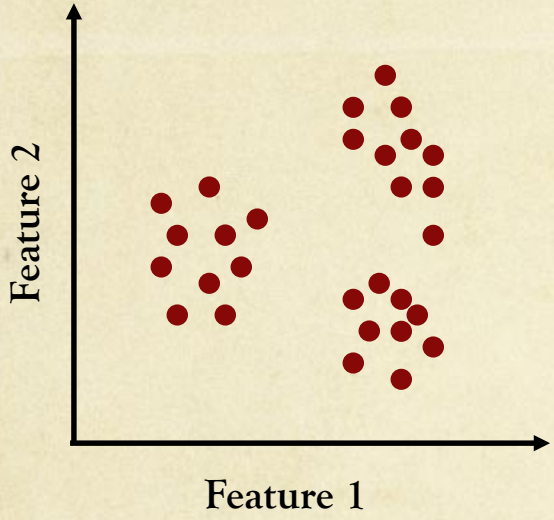


What is cluster analysis?



What is not?

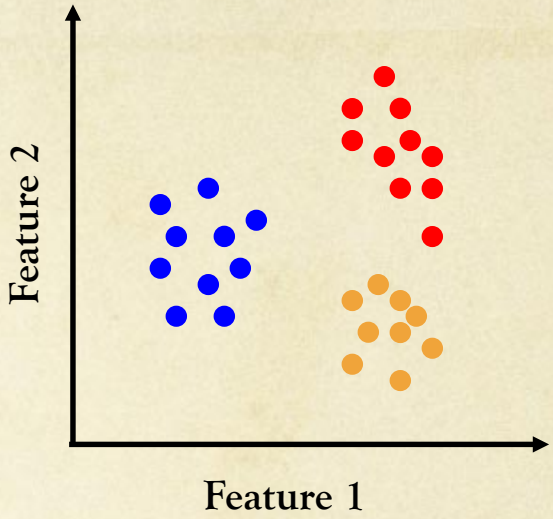
What is cluster analysis?



What is?

Clustering or
Unsupervised learning

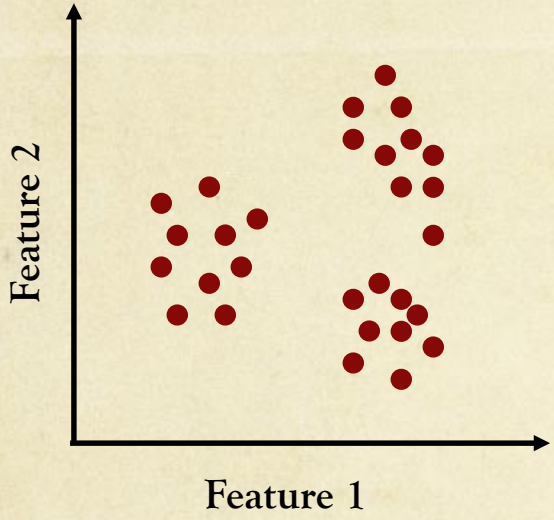
Clustering ≈ natural grouping of data



What is not?

Supervised learning

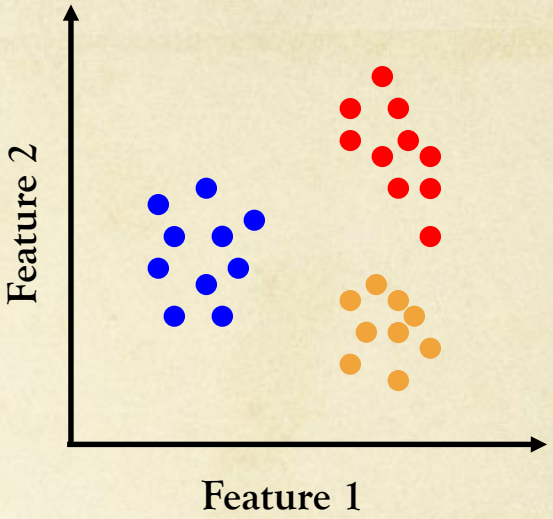
What is cluster analysis?



What is?

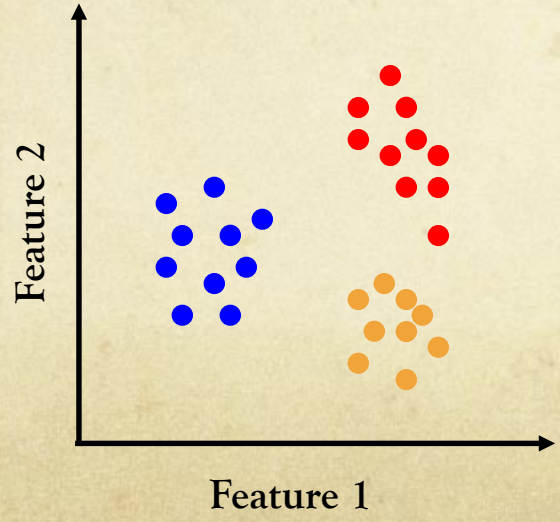
Clustering or
Unsupervised learning

Clustering ≈ natural grouping of data

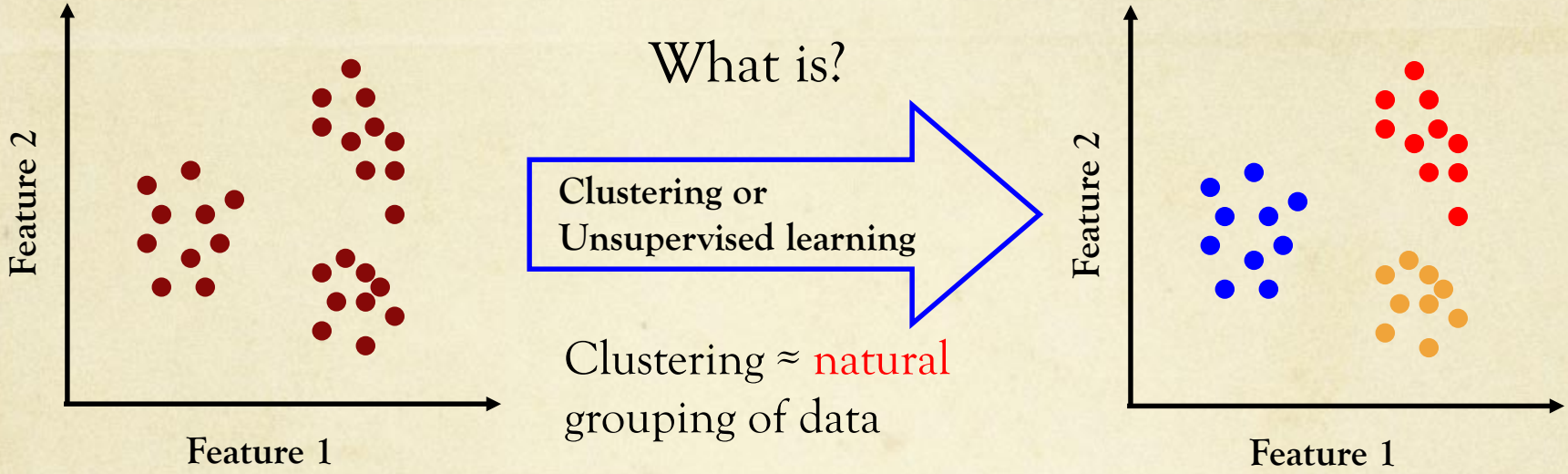


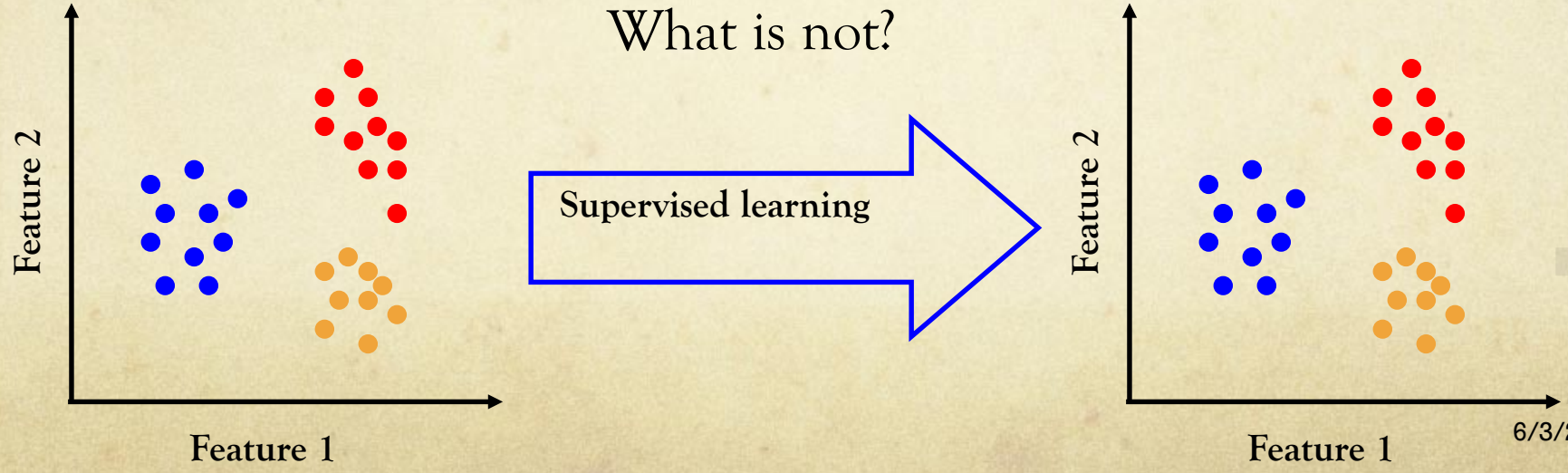
What is not?

Supervised learning

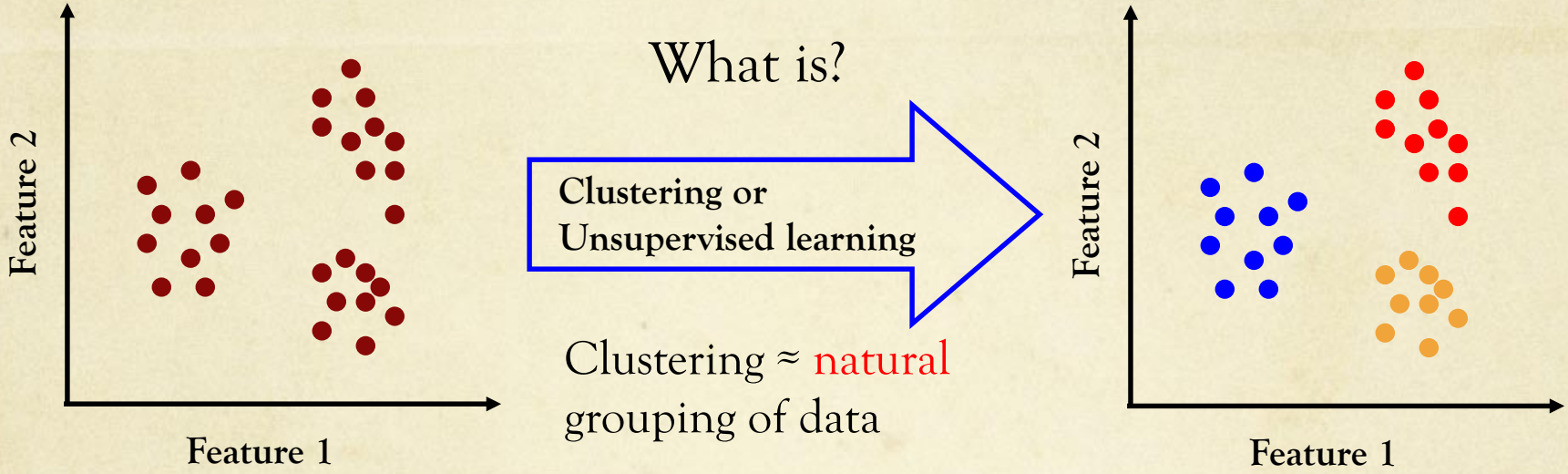


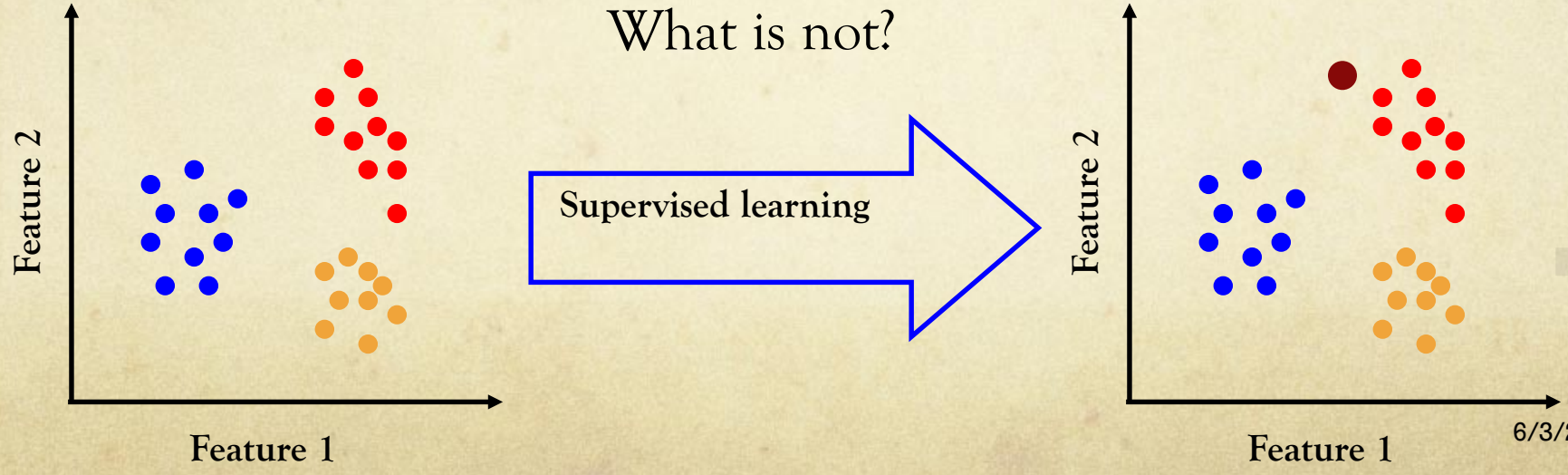
What is cluster analysis?



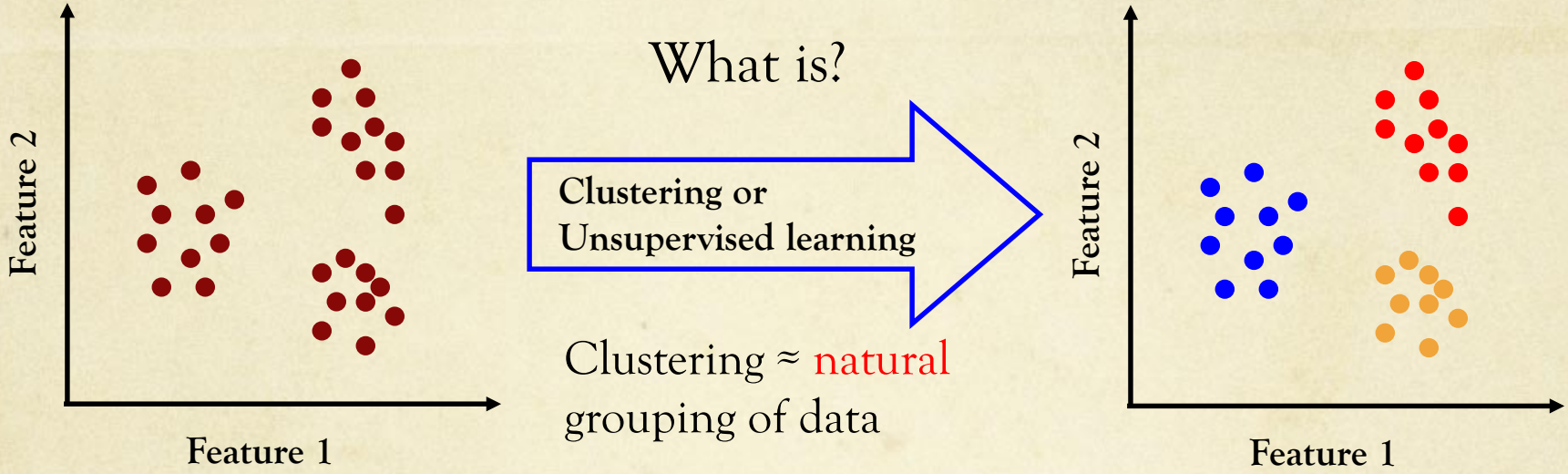


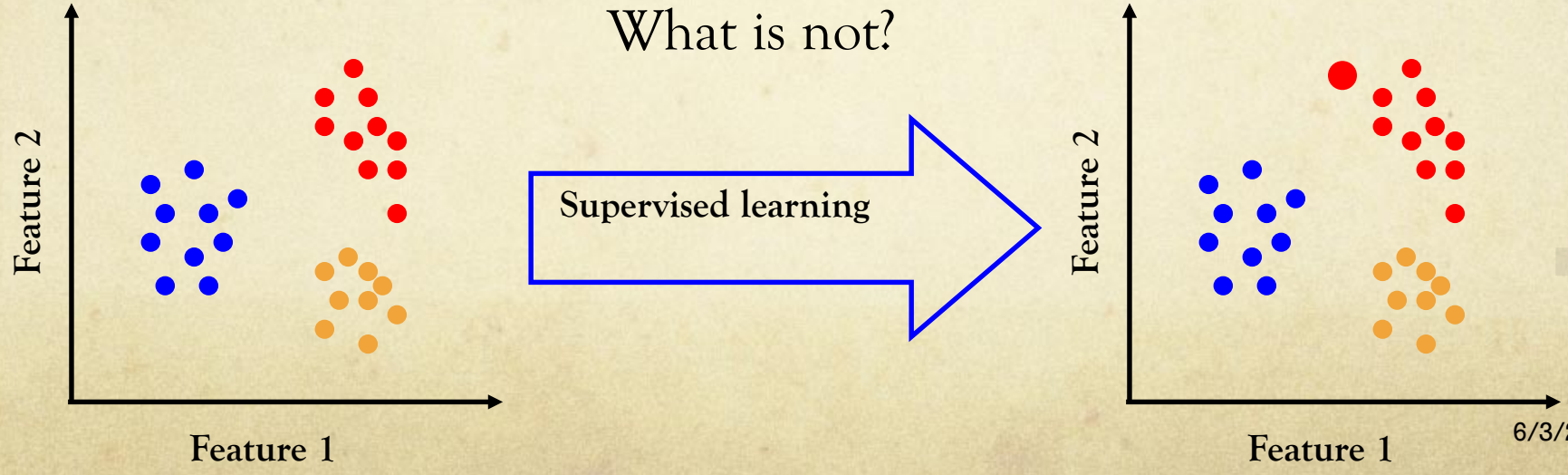
What is cluster analysis?





What is cluster analysis?



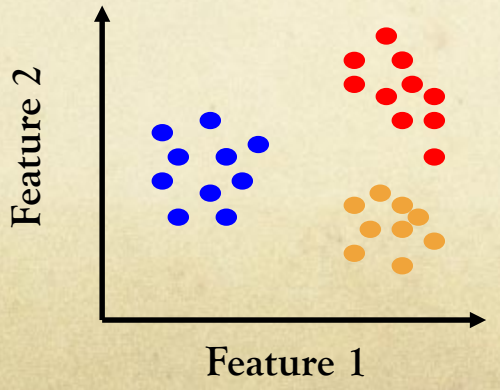


Overview

- What is cluster analysis? ☑
- Some definitions and notations
- How it works?
- Cluster Analysis Diagram
  - Objectives of cluster analysis
  - Research design issues
  - Assumptions in cluster analysis
  - Clustering methods
  - Interpreting the clusters
  - Validation

Definitions & notations

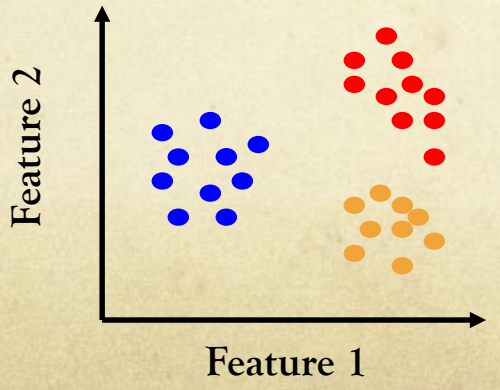
	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	



Definitions & notations

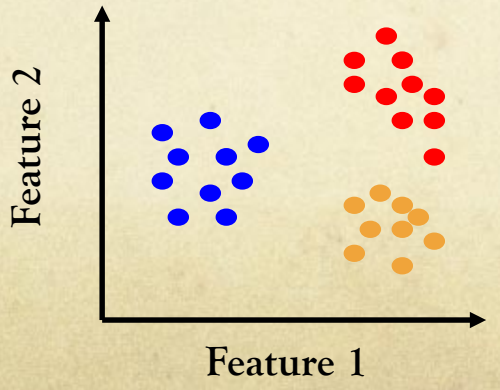
	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

- Objects or elementary data



Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

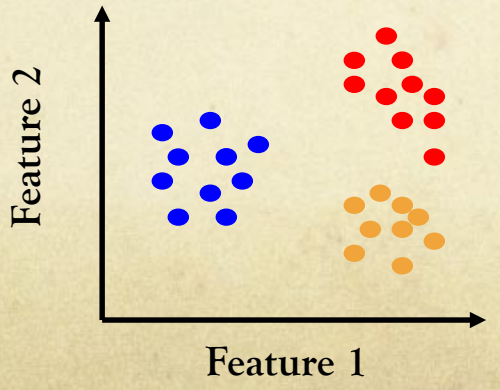


- Objects or elementary data

Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

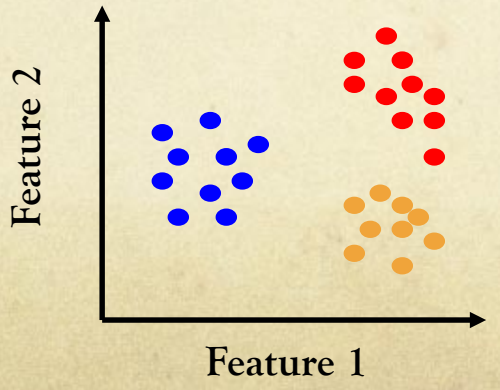
- Objects or elementary data



Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

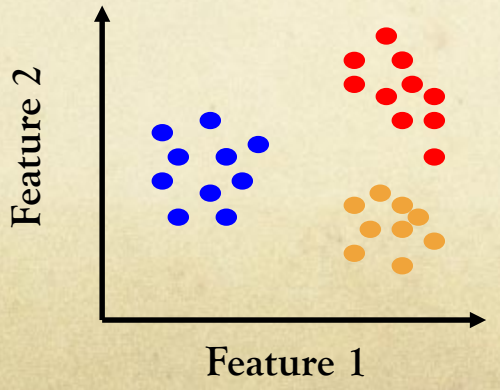
- Objects or elementary data
- Features or cluster variate



Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

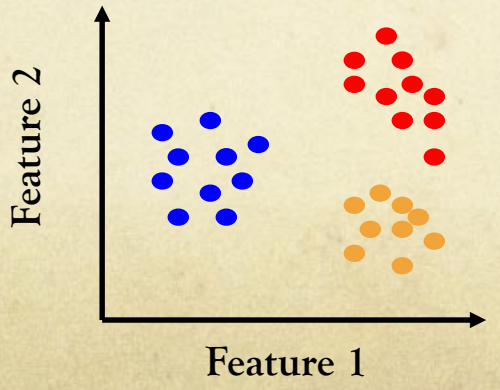
- Objects or elementary data
- Features or cluster variate



Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

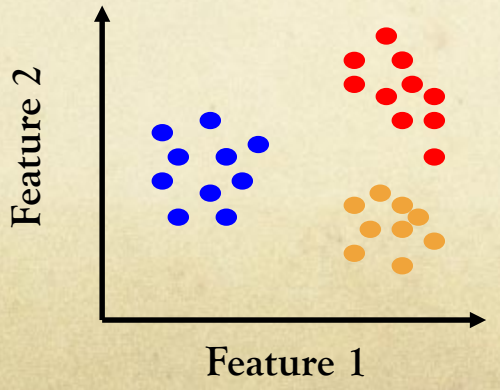
- Objects or elementary data
- Features or cluster variate



Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

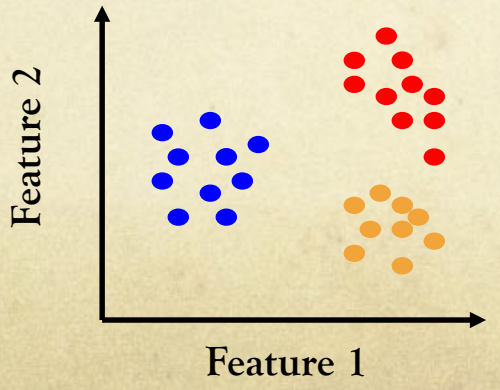
- Objects or elementary data
- Features or cluster variate
- Data dimension



Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

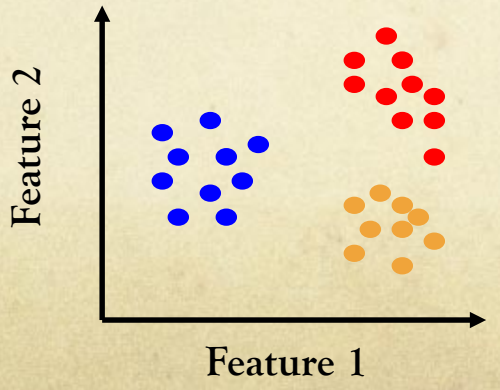
- Objects or elementary data
- Features or cluster variate
- Data dimension



Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

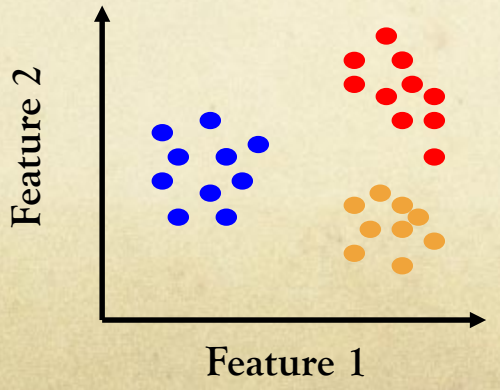
- Objects or elementary data
- Features or cluster variate
- Data dimension



Definitions & notations

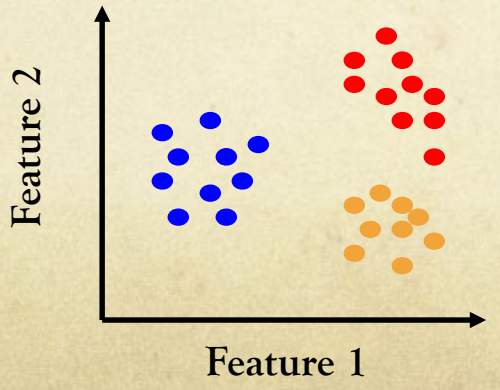
	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

- Objects or elementary data
- Features or cluster variate
- Data dimension
- Similarity measure



Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

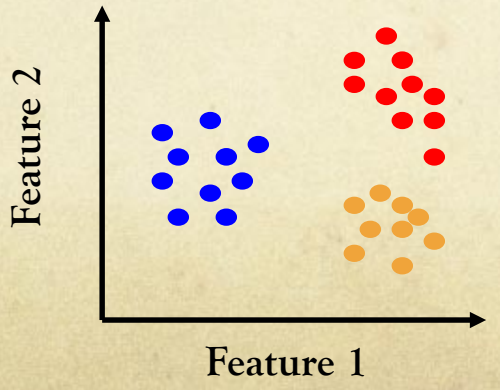


- Objects or elementary data
- Features or cluster variate
- Data dimension
- Similarity measure

Definitions & notations

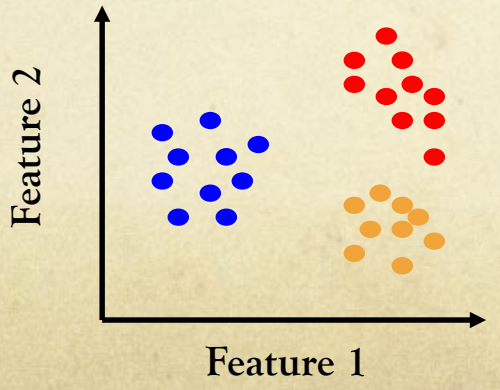
	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

- Objects or elementary data
- Features or cluster variate
- Data dimension
- Similarity measure



Definitions & notations

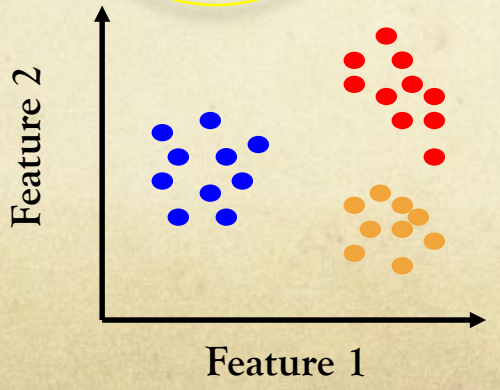
	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	



- Objects or elementary data
- Features or cluster variate
- Data dimension
- Similarity measure
- Cluster

Definitions & notations

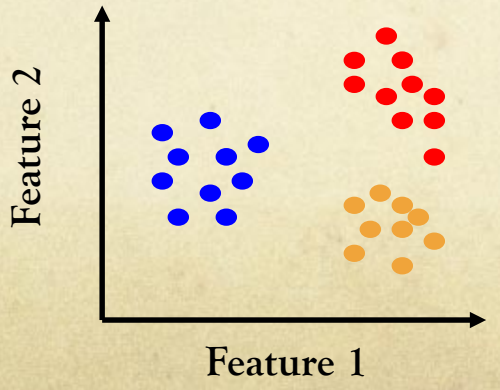
	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	



- Objects or elementary data
- Features or cluster variate
- Data dimension
- Similarity measure
- Cluster

Definitions & notations

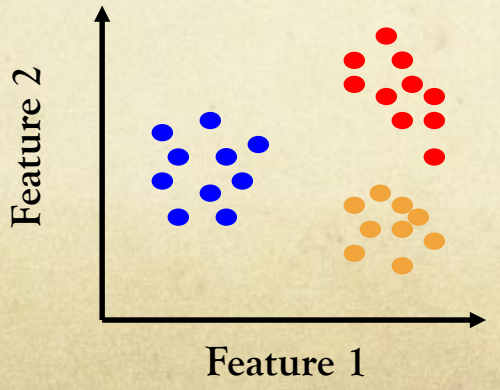
	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	



- Objects or elementary data
- Features or cluster variate
- Data dimension
- Similarity measure
- Cluster

Definitions & notations

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	



- Objects or elementary data
- Features or cluster variate
- Data dimension
- Similarity measure
- Cluster
- Cluster seed
- Cluster centroid
- Cluster solution
- Outlier

Definitions & notations

Dimensions

	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

Number of variables per sample
1 - Univariate data
2 - Bivariate data
3 - Trivariate data
>3 Multi&HyperVariate data

Remark: Quantitative variables (can do math on them)

	$Pixel_1$	$Pixel_2$	$Pixel_3$	...
Red	$Value_{11}$	$Value_{21}$	$Value_3$	...
Green	$Value_{12}$	$Value_{22}$	$Value_{32}$	...
Blue	$Value_{13}$	$Value_{23}$	$Value_{33}$	...

Definitions & notations

Dimensions	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

Number of variables per sample
1 - Univariate data
2 - Bivariate data
3 - Trivariate data
>3 Multi&HyperVariate data

Remark: Quantitative variables (can do math on them)

An example

	$Pixel_1$	$Pixel_2$	$Pixel_3$	...
Red	$Value_{11}$	$Value_{21}$	$Value_3$	...
Green	$Value_{12}$	$Value_{22}$	$Value_{32}$	...
Blue	$Value_{13}$	$Value_{23}$	$Value_{33}$	...

Definitions & notations

Dimensions	$Sample_1$	$Sample_2$	$Sample_3$	...
$Variable_1$	$Value_{11}$	$Value_{21}$	$Value_{31}$	
$Variable_2$	$Value_{12}$	$Value_{22}$	$Value_{32}$	
$Variable_3$	$Value_{13}$	$Value_{23}$	$Value_{33}$	
...	.	.	.	
	.	.	.	
	.	.	.	

Number of variables per sample
1 - Univariate data
2 - Bivariate data
3 - Trivariate data
>3 Multi&HyperVariate data

Remark: Quantitative variables (can do math on them)

An example

RGB images are trivariate data

	$Pixel_1$	$Pixel_2$	$Pixel_3$	...
Red	$Value_{11}$	$Value_{21}$	$Value_3$	...
Green	$Value_{12}$	$Value_{22}$	$Value_{32}$	...
Blue	$Value_{13}$	$Value_{23}$	$Value_{33}$	...

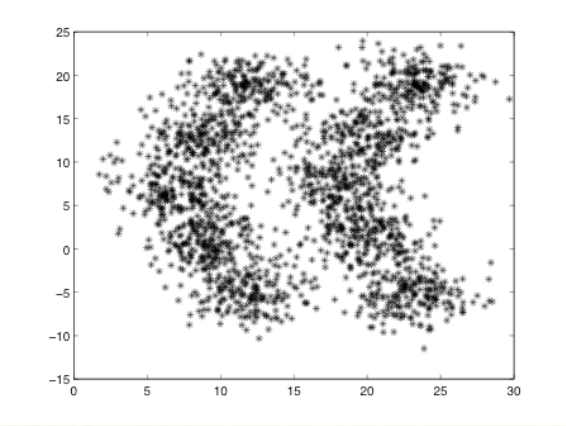
Overview

- What is cluster analysis? ☑
- Some definitions and notations ☑
- How it works?
- Cluster Analysis Diagram
  - Objectives of cluster analysis
  - Research design issues
  - Assumptions in cluster analysis
  - Clustering methods
  - Interpreting the clusters
  - Validation

How does it work?

What does ***natural grouping*** mean?

Example

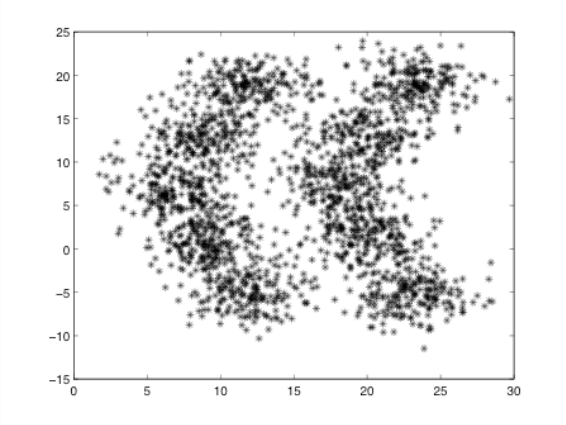


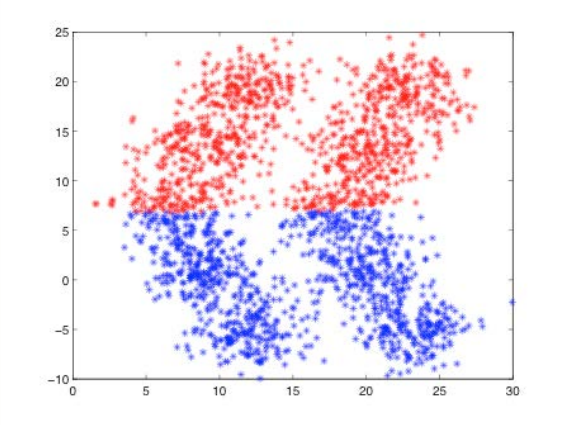
How does it work?

What does ***natural grouping*** mean?

Example

For some clustering algorithms, natural grouping means this…



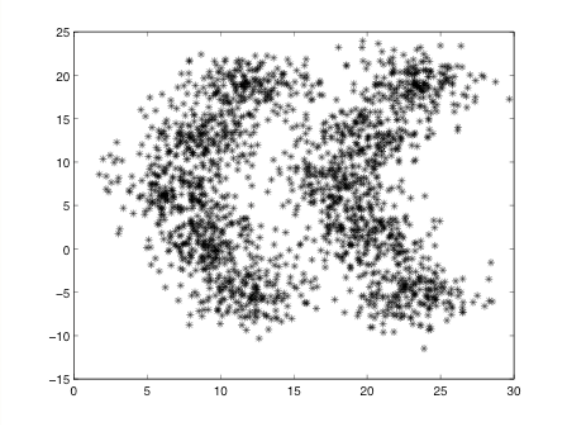


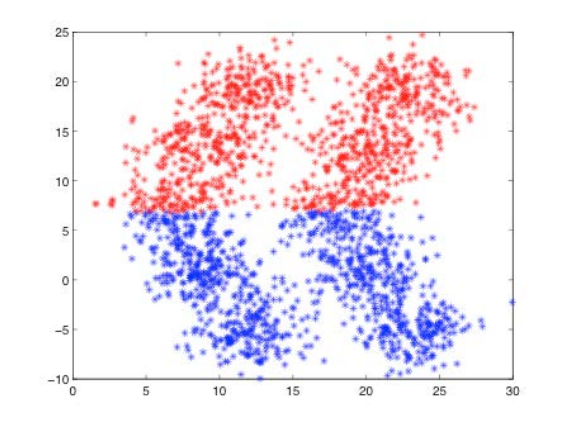
How does it work?

What does ***natural grouping*** mean?

Example

For some clustering algorithms, natural grouping means this…

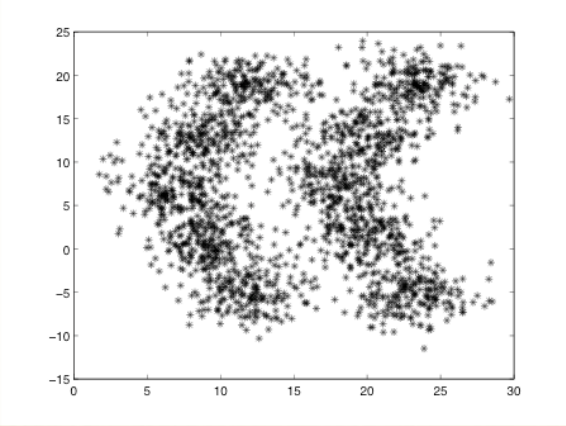




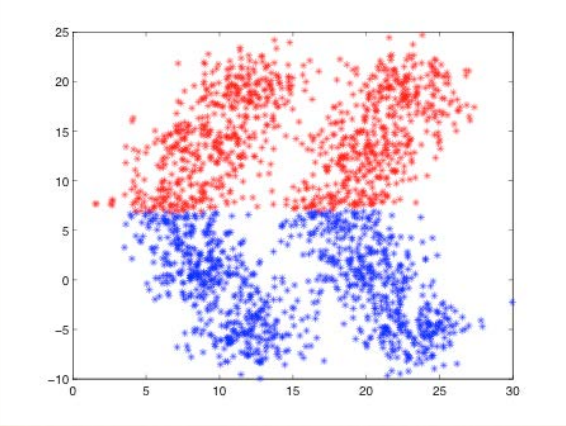
How does it work?

What does ***natural grouping*** mean?

Example



For some clustering algorithms, natural grouping means this…

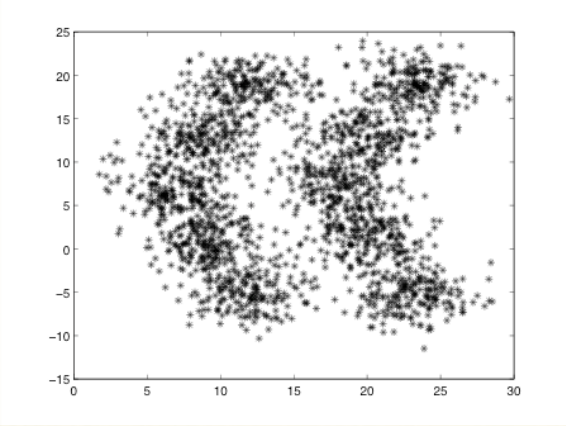


Actually, natural grouping means this…

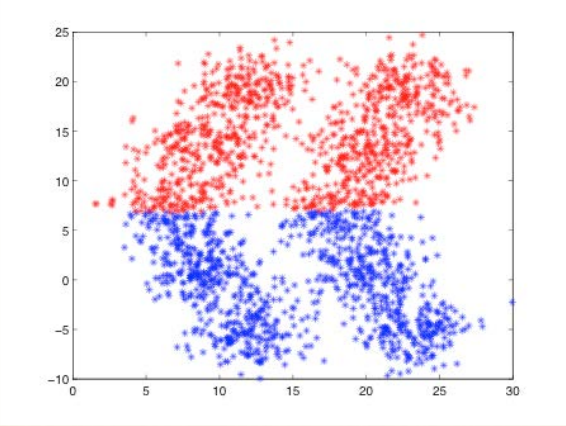
How does it work?

What does ***natural grouping*** mean?

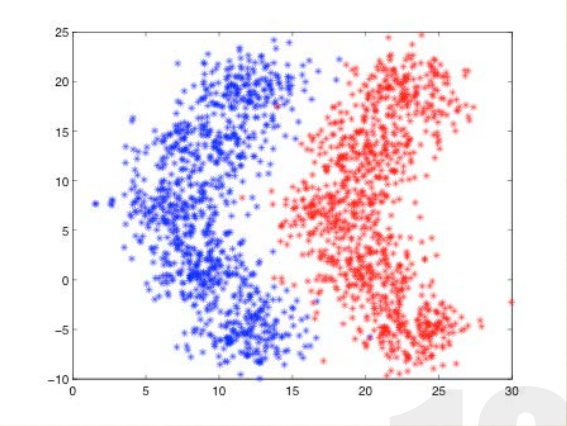
Example



For some clustering algorithms, natural grouping means this…



Actually, natural grouping means this…



How does it work?

A simple example:

Suppose that a biologist wants to determine the subspecies in a population of birds belonging the same specie

A small sample of 8 birds is selected as a pilot test

For each of the 8 birds, two characteristics of their beak are measured: V1 - length and V2 - width.

Clustering variables	Objects							
	S1	S2	S3	S4	S5	S6	S7	S8
V1	3.1	3.3	3.2	3.8	3.65	3.7	3.75	3.78
V2	1.1	1.2	1.05	1.1	1.2	1.05	1.6	1.62

How does it works?

A simple example:

Clustering variables	Objects							
	S1	S2	S3	S4	S5	S6	S7	S8
V1	3.1	3.3	3.2	3.8	3.65	3.7	3.75	3.78
V2	1.1	1.2	1.05	1.1	1.2	1.05	1.6	1.62

Objective

Identify structures (classes) in the data by grouping the most similar objects into groups

Three questions to be answered:

Q1: how does he measure the similarity between individuals?

Q2: how clusters should be formed?

Q3: how many clusters?

How does it work?

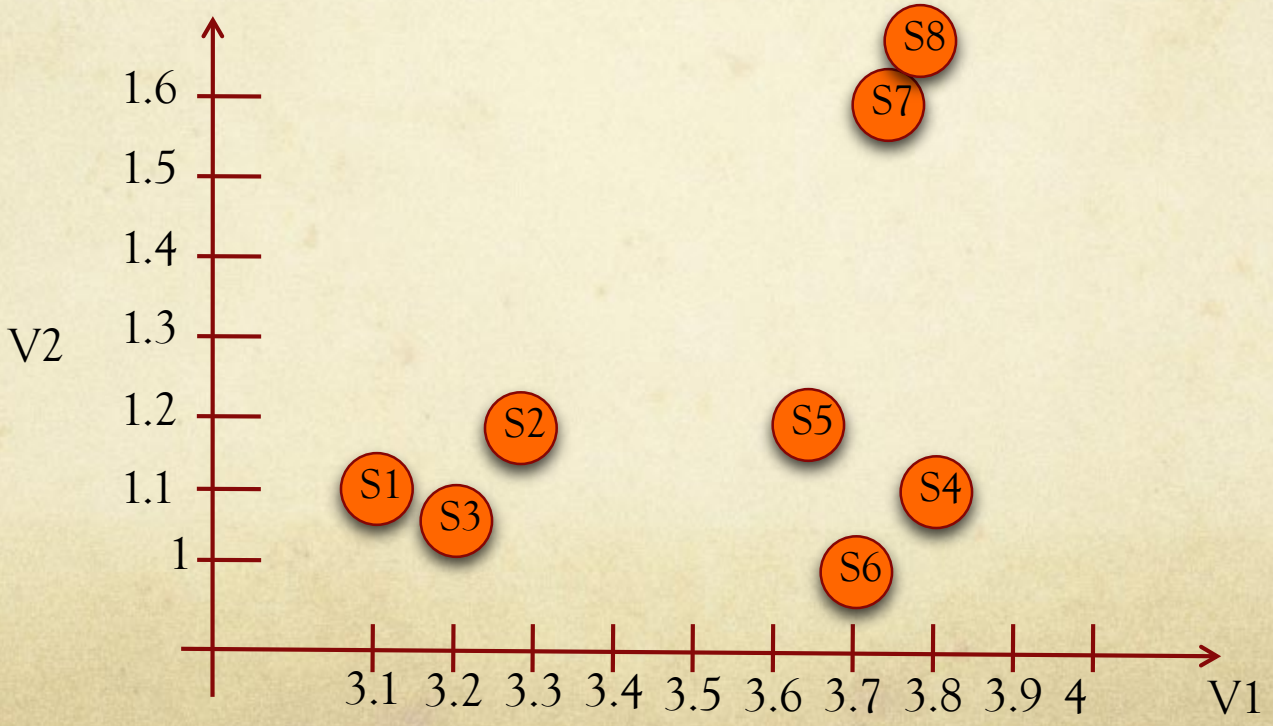
Q1: how does he measure the similarity between objects?

Clustering variables	Subjects							
	S1	S2	S3	S4	S5	S6	S7	S8
V1	3.1	3.3	3.2	3.8	3.65	3.7	3.75	3.78
V2	1.1	1.2	1.05	1.1	1.2	1.05	1.6	1.62

How does it work?

Q1: how does he measure the similarity between objects?

Clustering variables	Subjects							
	S1	S2	S3	S4	S5	S6	S7	S8
V1	3.1	3.3	3.2	3.8	3.65	3.7	3.75	3.78
V2	1.1	1.2	1.05	1.1	1.2	1.05	1.6	1.62



How does it work?

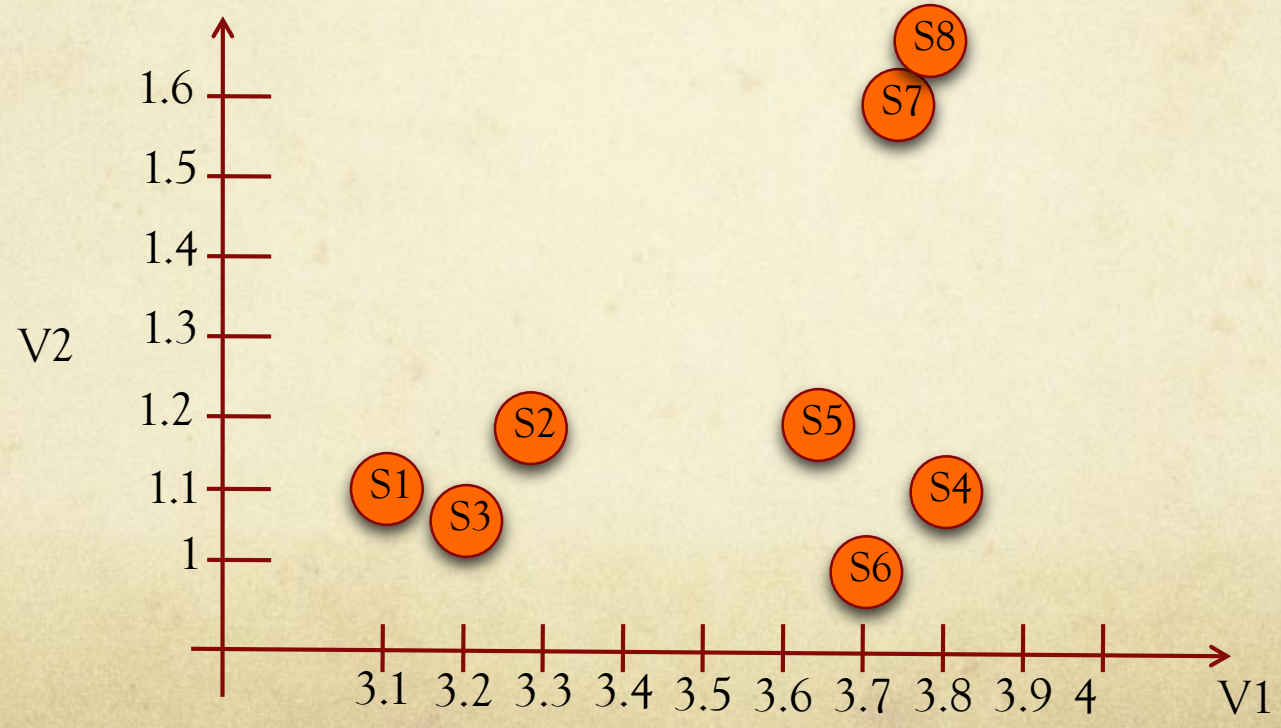
Q1: how does he measure the similarity between objects?

A1: build similarity matrix between all pairs of observations

Observations	Observations							
	S1	S2	S3	S4	S5	S6	S7	S8
S1	.....	.....	.....	.....	.....	.....	.....	.....
S2	0.22	.....	.....	.....	.....	.....	.....	.....
S3			.....	.....	.....	.....	.....	.....
S4				.....	.....	.....	.....	.....
S5					.....	.....	.....	.....
S6						.....	.....	.....
S7							.....	.....
S8								.....

How does it work?

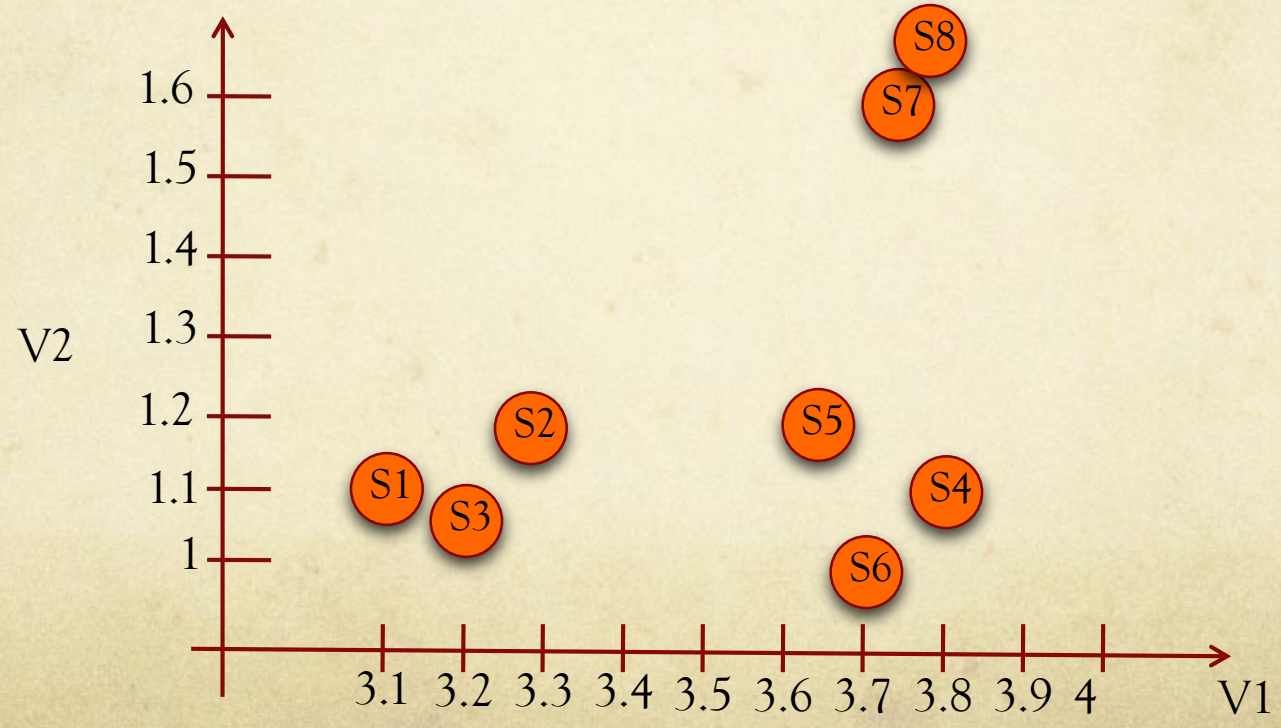
Q2: how does he form the clusters?



How does it work?

Q2: how does he form the clusters?

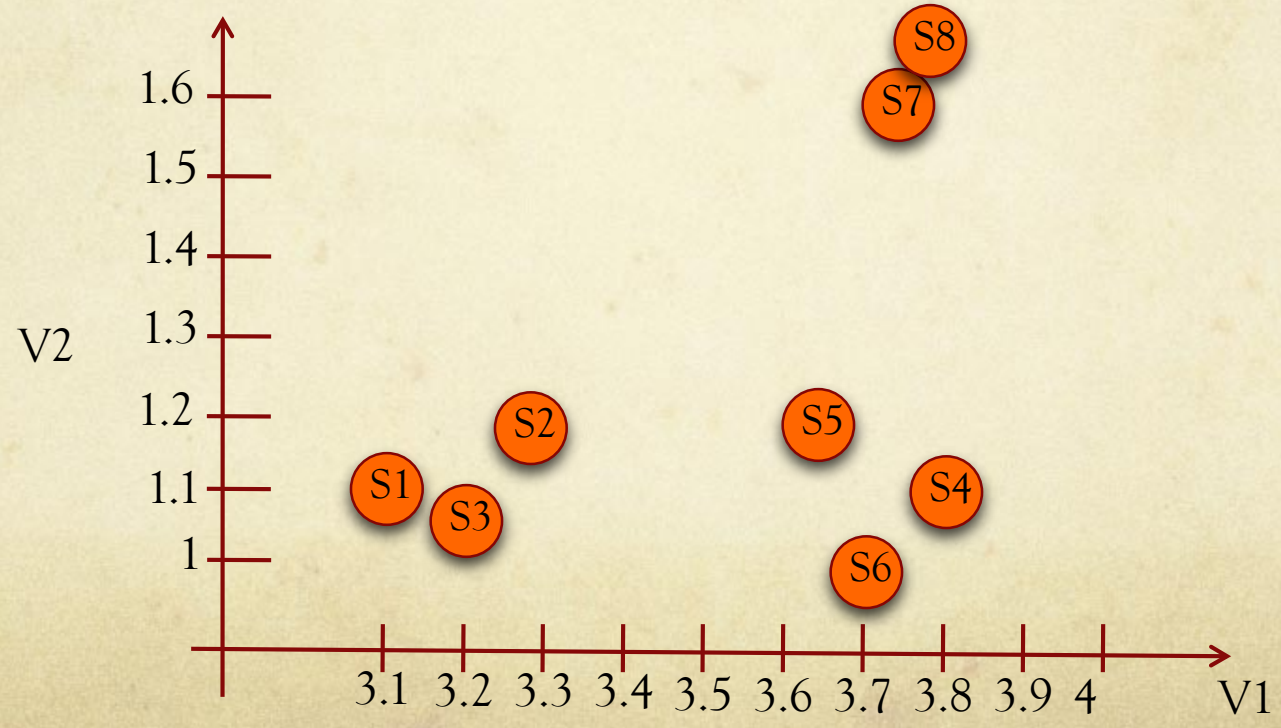
A21: group observations which are most similar into clusters



How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

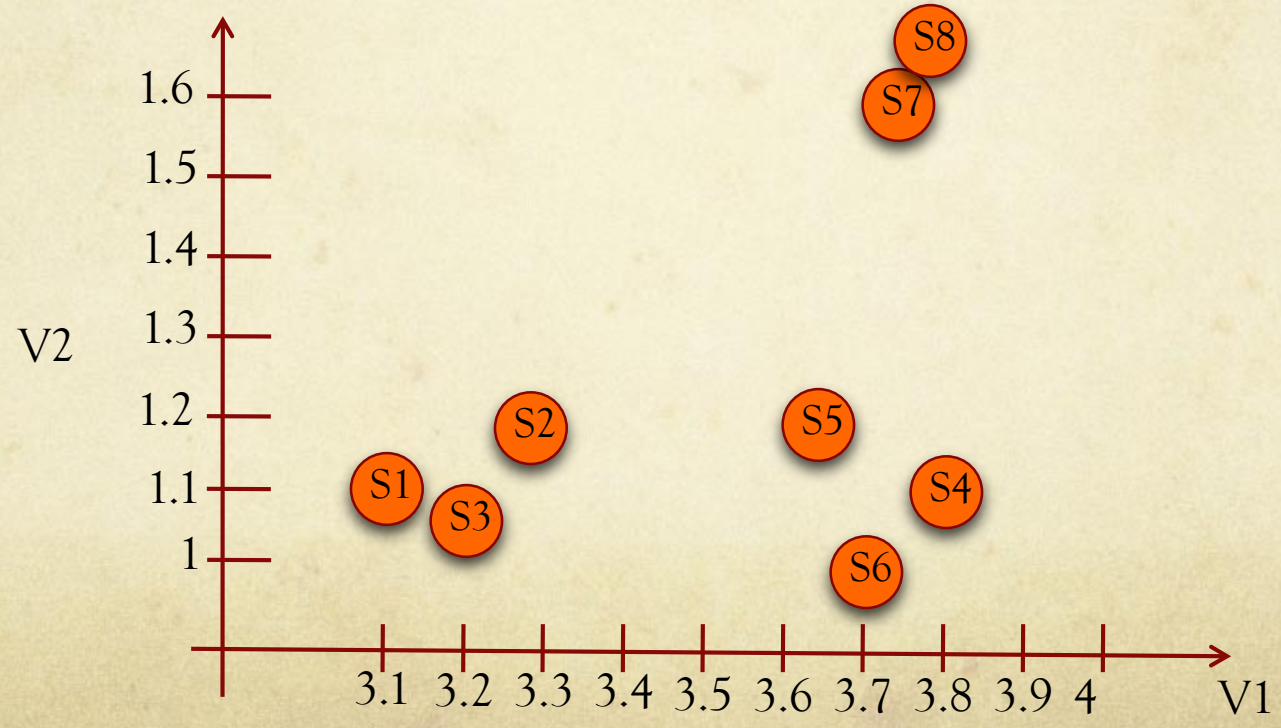


How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another

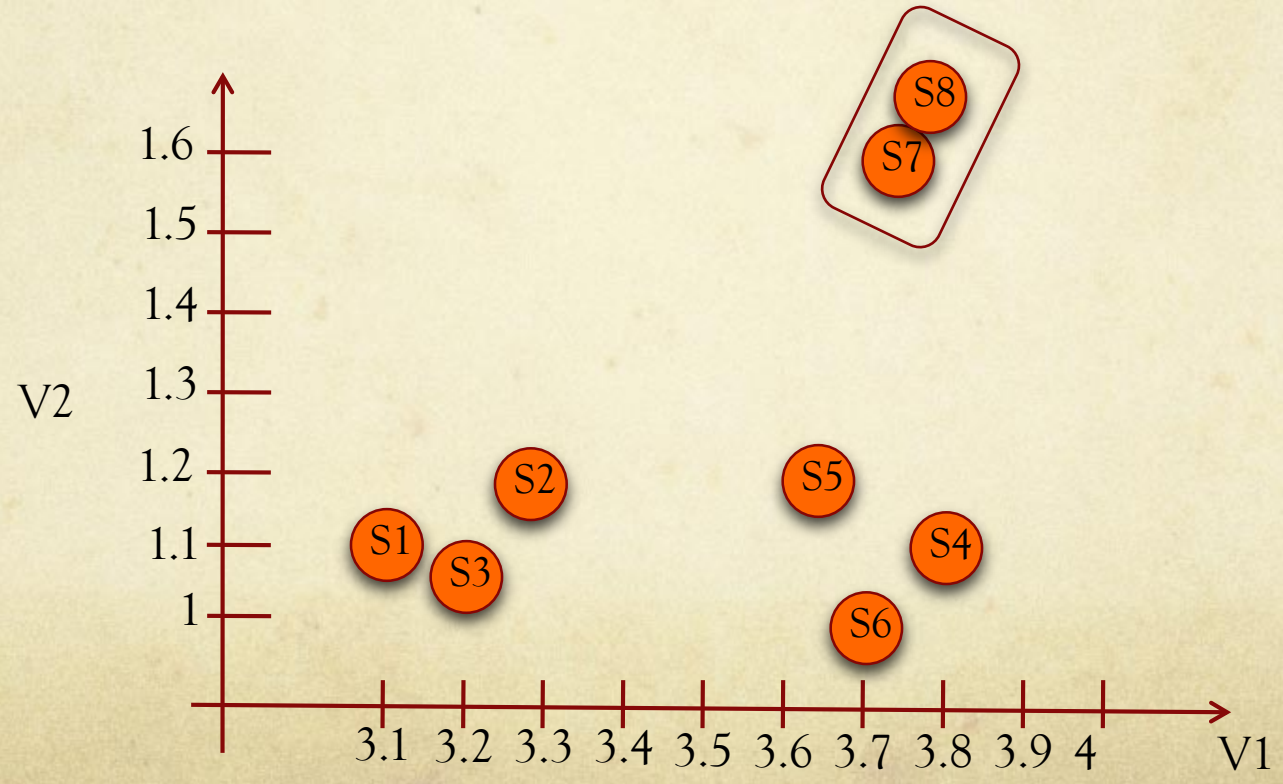


How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another

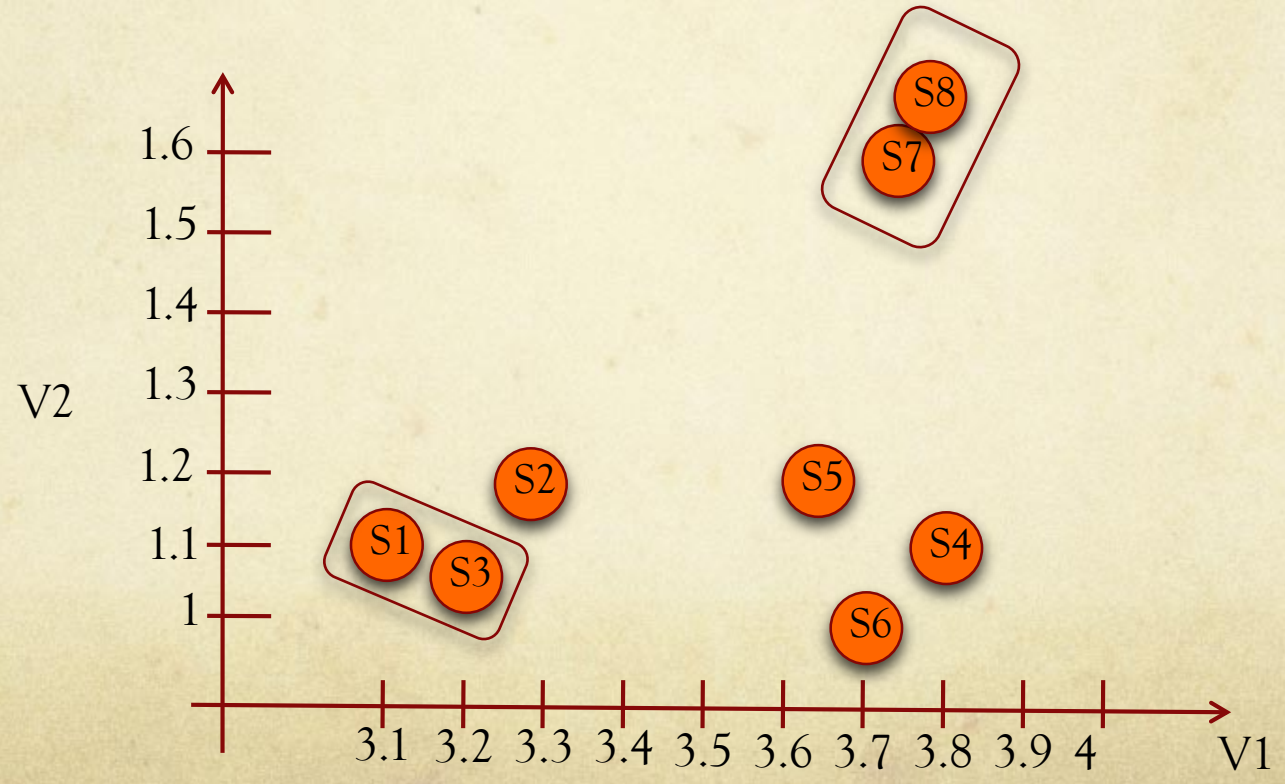


How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another

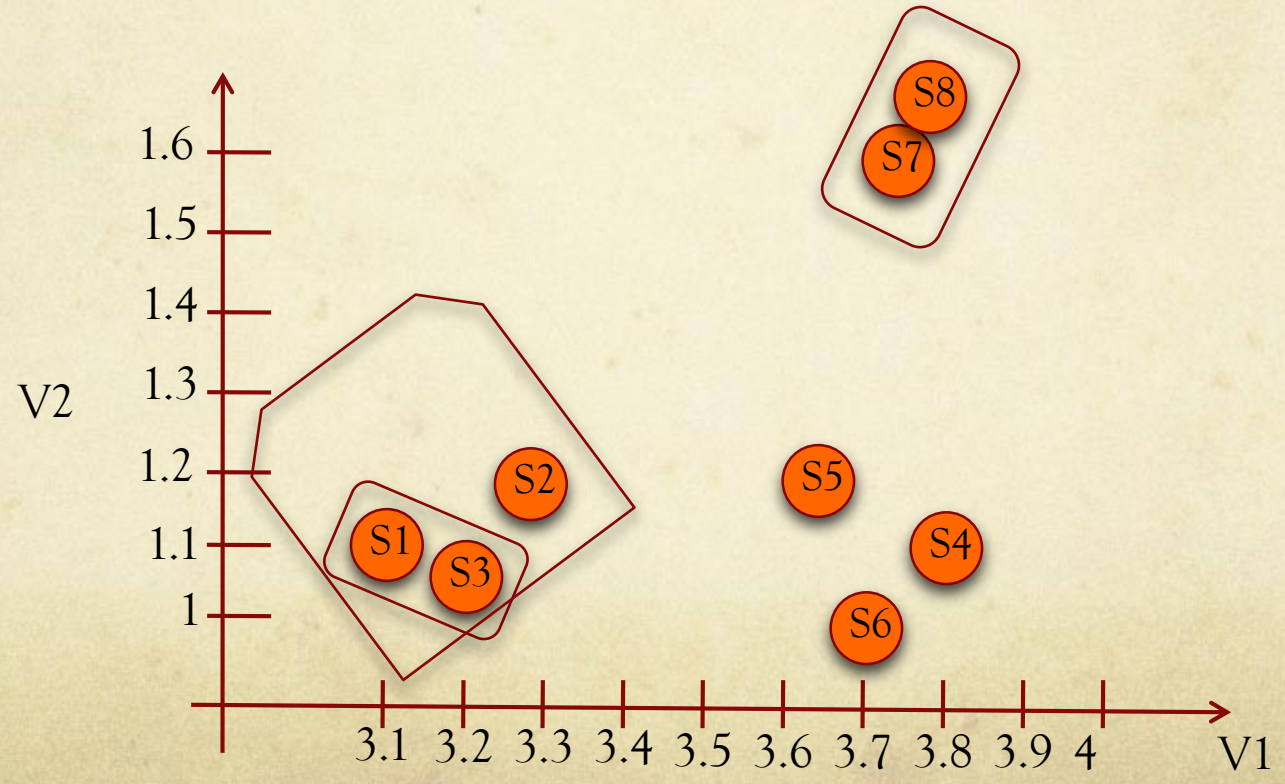


How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another

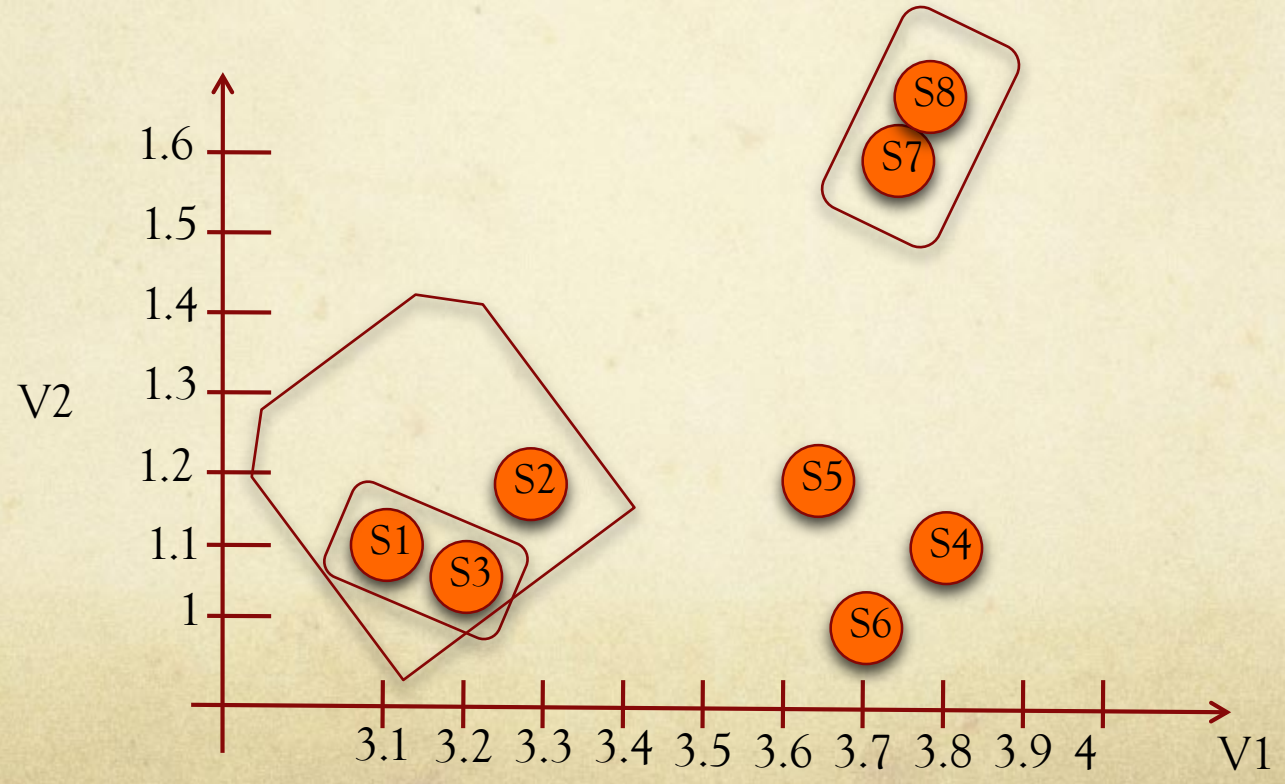


How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another

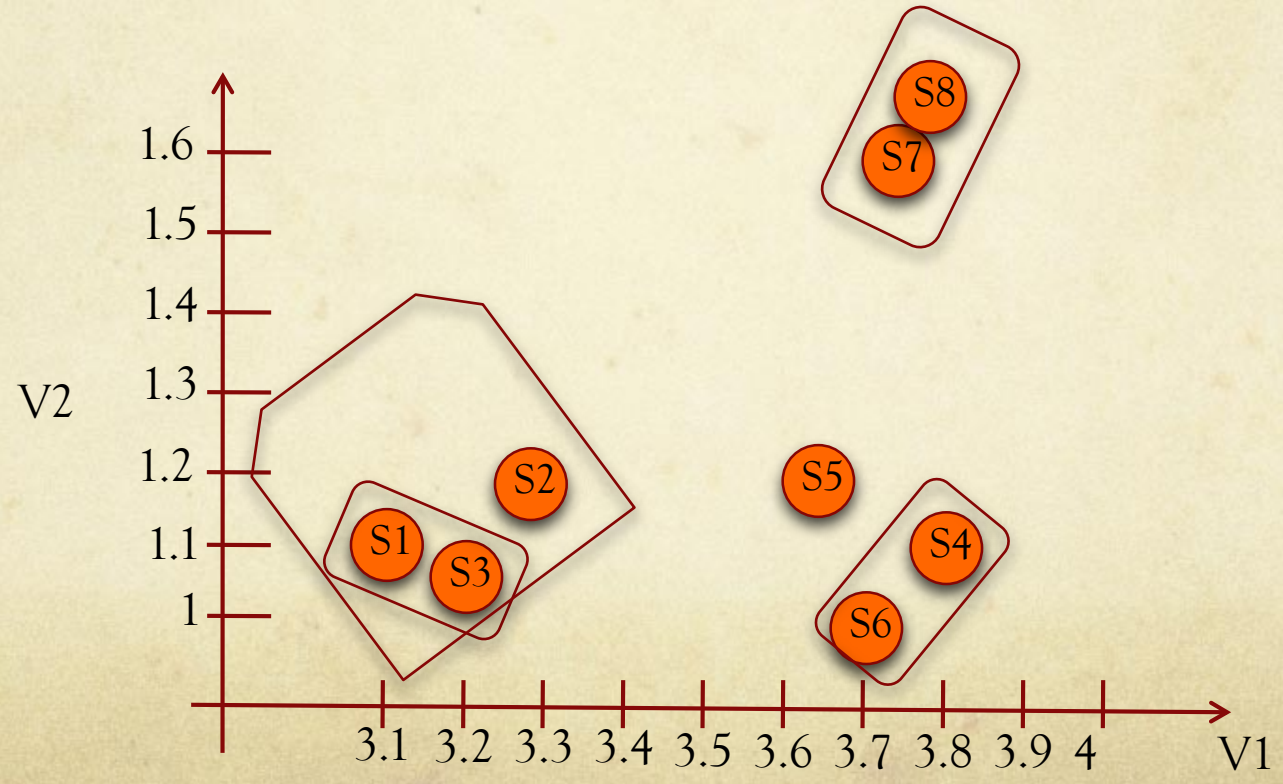


How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another

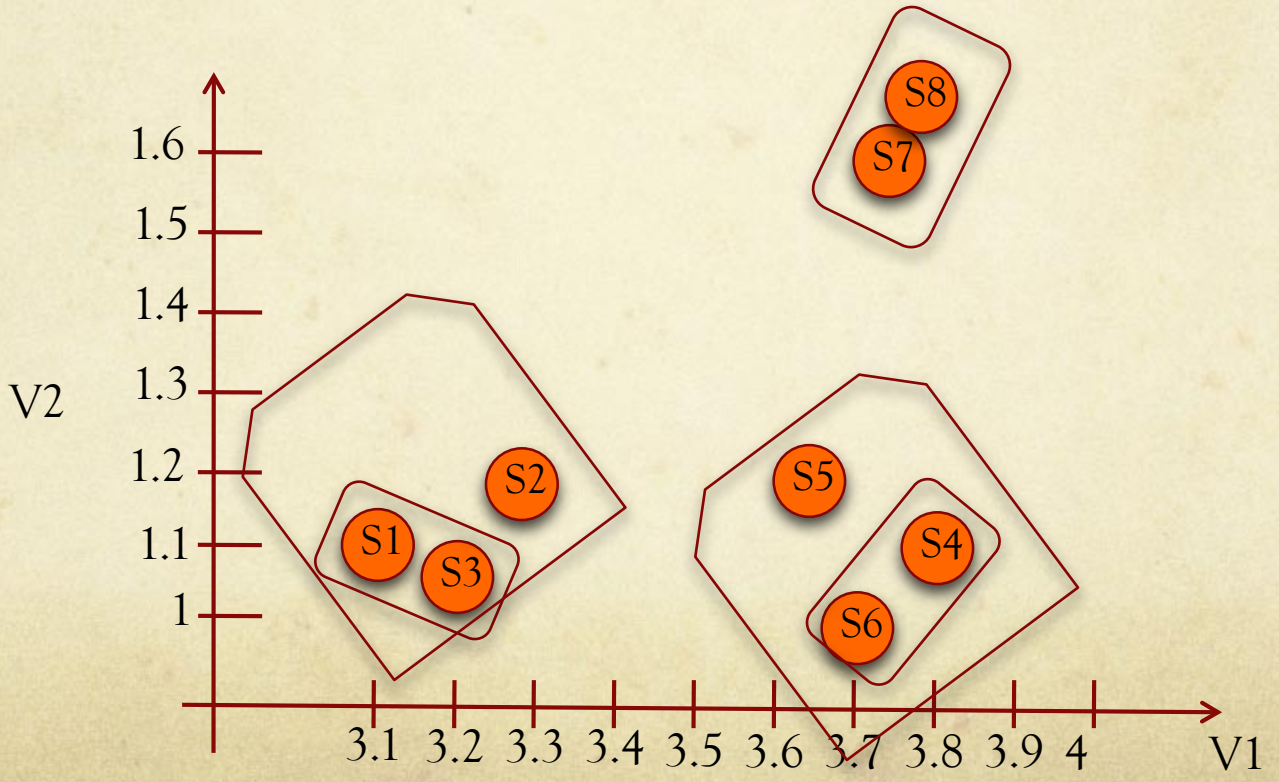


How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another

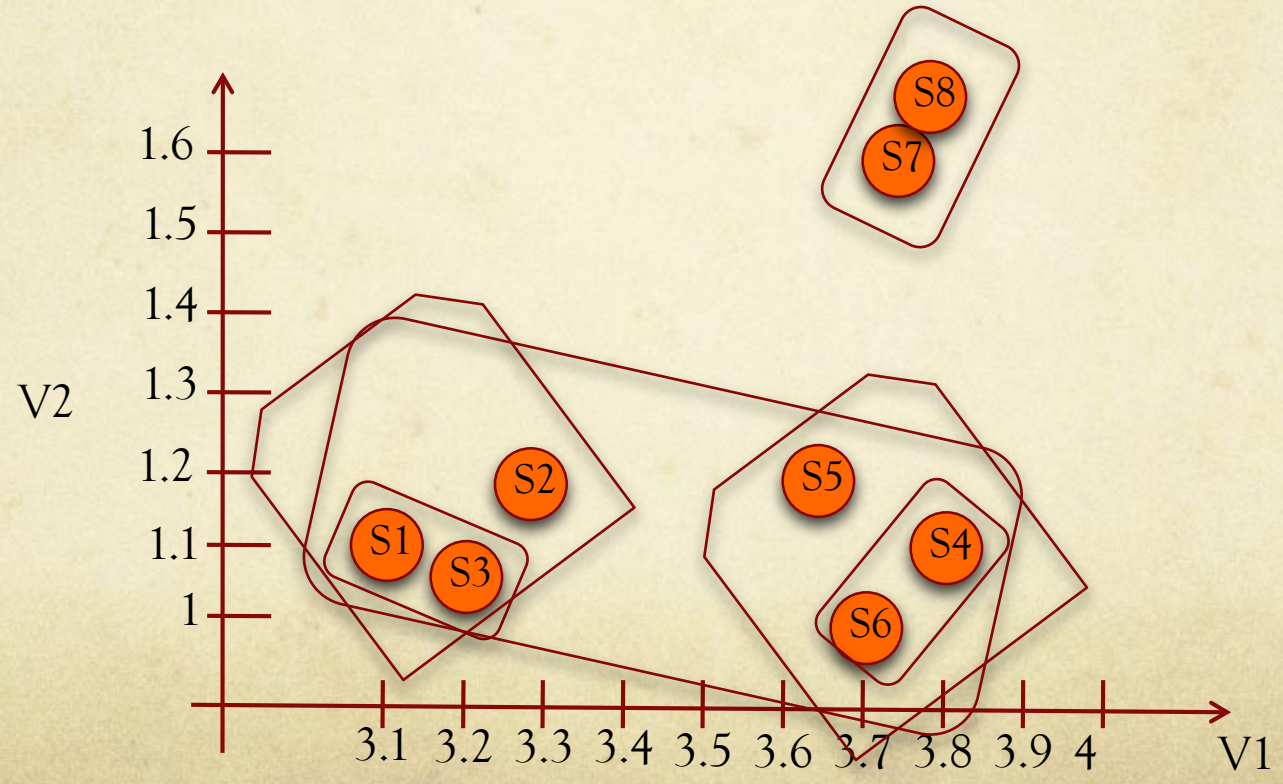


How does it work?

Q2: how does he form the clusters?

A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another

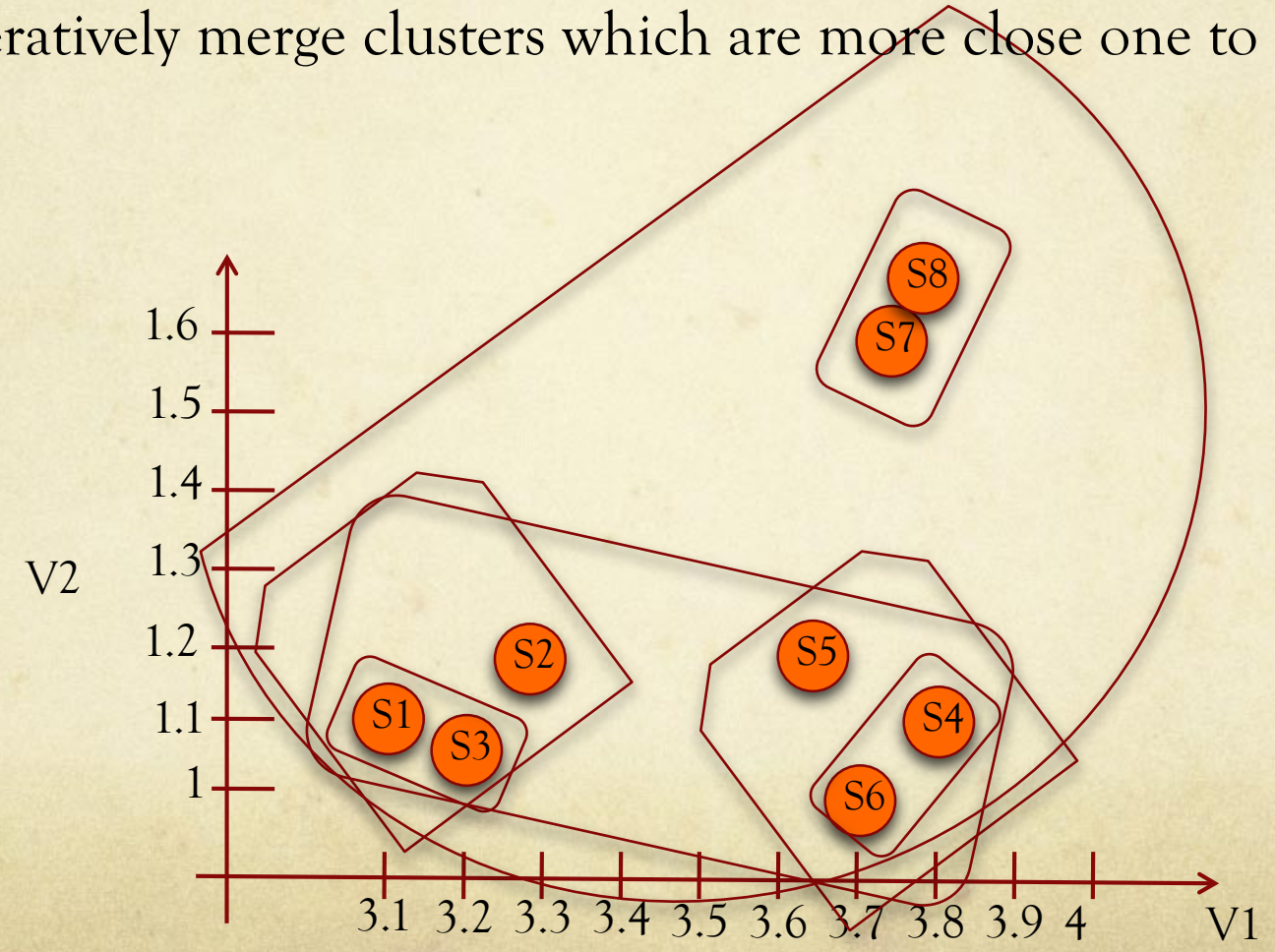


How does it work?

Q2: how does he form the clusters?

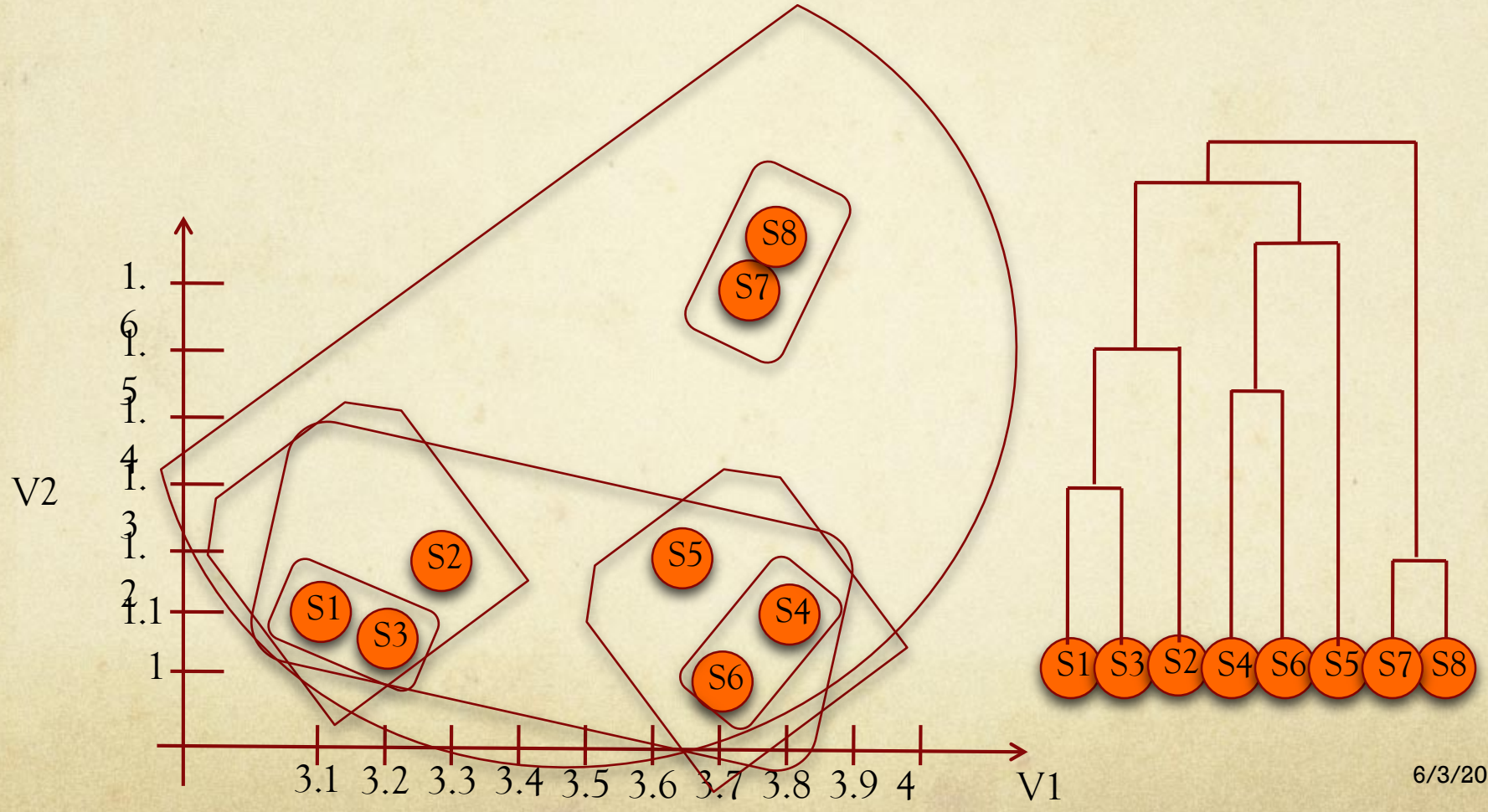
A21: group observations which are most similar into clusters

A22: iteratively merge clusters which are more close one to another



How does it work?

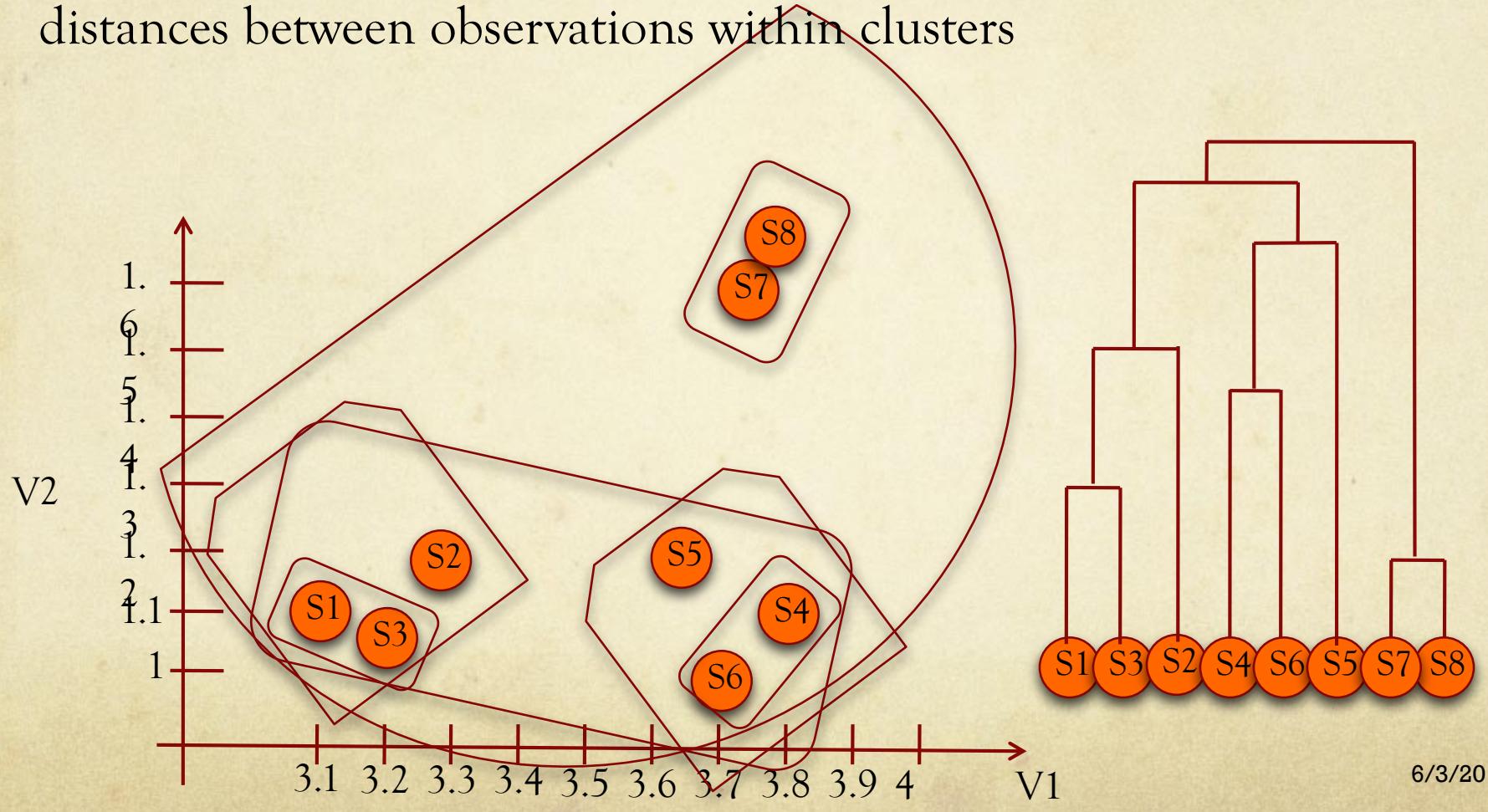
Q3: how to determine the number of clusters in the final solution?



How does it work?

Q3: how to determine the number of clusters in the final solution?

A3: measuring homogeneity of a cluster solution by averaging all distances between observations within clusters



Overview

- What is cluster analysis? ☑
- Some definitions and notations ☑
- How it works? ☑
- Cluster Analysis Diagram
  - Objectives of cluster analysis
  - Design issues
  - Assumptions in cluster analysis
  - Clustering methods
  - Interpreting the clusters
  - Validation

Cluster Analysis Diagram

Stage 1: Objectives of Cluster Analysis

Stage 2: Design Issues and Pre-Processing

Stage 3: Assumptions in Cluster Analysis

Stage 4: Deriving Clusters and Assessing Overall Fit

Stage 5: Interpreting the Clusters

Stage 6: Validating and Profiling the Clusters

Cluster Analysis – Objectives

Stage 1: Objectives of Cluster Analysis

Select objectives

Taxonomy description	– for exploratory purposes and the formation of a taxonomy (an empirically based classification of objects)
Data simplification	– a researcher could face a large number of observations that are meaningless unless classified into manageable groups
Hypothesis generation or testing	– a researcher wishes to develop hypothesis concerning the nature of the data or to examine previously stated hypothesis
Relationship identification	– a researcher wishes to reveal relationships among observations that are not possible with individual observations

Overview

- What is cluster analysis? ☑
- Some definitions and notations ☑
- How it works? ☑
- Cluster Analysis Diagram
  - Objectives of cluster analysis ☑
  - Design issues
  - Assumptions in cluster analysis
  - Clustering methods
  - Interpreting the clusters
  - Validation

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Five questions to be asked before starting:

1. What variables are relevant?
2. Is the sample size adequate?
3. Can outliers be detected and if so should they be removed?
4. How should object similarity be measured?
5. Should data be standardized?

Cluster Analysis – Design issues

Stage 2: Design Issues and Pre-Processing

Q1: What variables are relevant?

Select clustering variables

Theoretical, conceptual and practical considerations must be observed when selecting variables for clustering analysis

Cluster Analysis – Design issues

Stage 2: Design Issues and Pre-Processing

Q1: What variables are relevant?

Select clustering variables

Theoretical, conceptual and practical considerations must be observed when selecting variables for clustering analysis

Feature selection methods enable users to select the most relevant variables to be used in cluster analysis

Cluster Analysis – Design issues

Stage 2: Design Issues and Pre-Processing

Q1: What variables are relevant?

Select clustering variables

Theoretical, conceptual and practical considerations must be observed when selecting variables for clustering analysis

Feature selection methods enable users to select the most relevant variables to be used in cluster analysis

Feature extraction methods enable users to derive new features from the existing features which could be more relevant then the existing features for cluster analysis

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q2: Is the sample size adequate?

A2: the sample size must be large enough to provide sufficient representation of small groups within the population and represent the underlying structure

Remark - the issue of sample size do not relates to any statistical inference issues

Optimal sample size - the researcher should

- ensure the sample size is sufficiently large to adequately represent all relevant groups
- specify the group sizes necessary for relevance for the questions being asked

Remark:

1. Interest is focus on the identification of small groups – large sample size
2. Interest is focus on the identification of large groups – small sample size

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q3: Can outliers be detected and if so should they be removed?

What outliers can be?

1. Truly aberrant observation not representative for the population	- distort the actual structure and result in unrepresentative clusters – should be removed
2. Representative observations of small or insignificant groups	- should be removed so that the resulting clusters represent more accurately relevant groups
3. An undersampling of the actual group in the population that causes poor representation of the group	- they represent valid and relevant groups - should be included in the clustering solution

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Three ways to measure inter-objects similarities

- correlation measures
- distance measures

} require metric data

- association measures

} require non-metric data

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Correlation measures

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Correlation measures

$$CC(X_i, X_j) = \frac{\sum_{k=1}^{d}(X_i - \mu_{X_i})(X_j - \mu_{X_j})}{\frac{1}{d-1}\sum_{k-1}^{d}(X_i - \mu_{X_i})\sum_{k=1}^{d}(X_j - \mu_{X_j})}$$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Correlation measures

$$CC(X_i, X_j) = \frac{\sum_{k=1}^{d}(X_i - \mu_{X_i})(X_j - \mu_{X_j})}{\frac{1}{d-1}\sum_{k-1}^{d}(X_i - \mu_{X_i})\sum_{k=1}^{d}(X_j - \mu_{X_j})}$$

$$SA(X_i, X_j) = a\cos(CC(X_i, X_j))$$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Correlation measures

Pearson's correlation coefficient

$$CC(X_i, X_j) = \frac{\sum_{k=1}^{d}(X_i - \mu_{X_i})(X_j - \mu_{X_j})}{\frac{1}{d-1}\sum_{k-1}^{d}(X_i - \mu_{X_i})\sum_{k=1}^{d}(X_j - \mu_{X_j})}$$

$$SA(X_i, X_j) = a\cos(CC(X_i, X_j))$$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Correlation measures

Pearson's correlation coefficient

$$CC(X_i, X_j) = \frac{\sum_{k=1}^{d}(X_i - \mu_{X_i})(X_j - \mu_{X_j})}{\frac{1}{d-1}\sum_{k-1}^{d}(X_i - \mu_{X_i})\sum_{k=1}^{d}(X_j - \mu_{X_j})}$$

Spectral angle

$$SA(X_i, X_j) = a\cos(CC(X_i, X_j))$$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

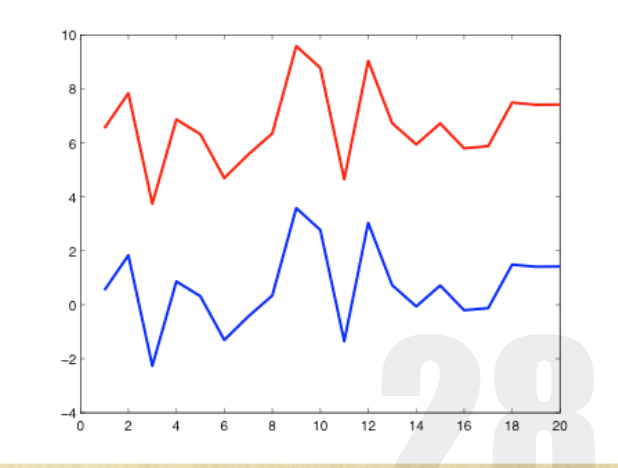
Correlation measures

Pearson's correlation coefficient

$$CC(X_i, X_j) = \frac{\sum_{k=1}^{d}(X_i - \mu_{X_i})(X_j - \mu_{X_j})}{\frac{1}{d-1}\sum_{k-1}^{d}(X_i - \mu_{X_i})\sum_{k=1}^{d}(X_j - \mu_{X_j})}$$

Spectral angle

$$SA(X_i, X_j) = a\cos(CC(X_i, X_j))$$



Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

- r - metrics

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \left\{ X_k^n, X_k \in \Re^d \right\}$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \left\{ X_k^n, X_k \in \Re^d \right\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \left\{ X_k^n, X_k \in \Re^d \right\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \left\{ X_k^n, X_k \in \Re^d \right\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \{X_k^n, X_k \in \Re^d\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Minkowski metrics

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \{X_k^n, X_k \in \Re^d\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Minkowski metrics $r \geq 1$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \{X_k^n, X_k \in \Re^d\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Minkowski metrics $r \geq 1$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \{X_k^n, X_k \in \Re^d\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Minkowski metrics $r \geq 1$

Fractionary metrics

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \{X_k^n, X_k \in \Re^d\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Minkowski metrics $r \geq 1$

Fractionary metrics $r < 1$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \{X_k^n, X_k \in \Re^d\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Minkowski metrics $r \geq 1$

Fractionary metrics $r < 1$

$r = 1$ Manhattan distance

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \{X_k^n, X_k \in \Re^d\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Minkowski metrics $r \geq 1$

Fractionary metrics $r < 1$

$r = 1$ Manhattan distance

$r = 2$ Euclidian distance

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

r - metrics

Let $X = \{X_k^n, X_k \in \Re^d\}$

then $L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$

Metric exponent

Minkowski metrics $r \geq 1$

Fractionary metrics $r < 1$

$r = 1$ Manhattan distance

$r = 2$ Euclidian distance

$r \geq 3$ High order metrics

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

Mahalanobis distance

$$MD(X_i, X_j) = \frac{\sum_{k=1}^{d}(X_i - X_j)}{\frac{1}{d-1}\sum_{k-1}^{d}(X_i - \mu_{X_i})\sum_{k=1}^{d}(X_j - \mu_{X_j})}$$

$$CC(X_i, X_j) = \frac{\sum_{k=1}^{d}(X_i - \mu_{X_i})(X_j - \mu_{X_j})}{\frac{1}{d-1}\sum_{k-1}^{d}(X_i - \mu_{X_i})\sum_{k=1}^{d}(X_j - \mu_{X_j})}$$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

$$L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$$

Mahalanobis distance

$$MD(X_i, X_j) = \frac{\sum_{k=1}^{d} (X_i - X_j)}{\frac{1}{d-1} \sum_{k-1}^{d} (X_i - \mu_{X_i}) \sum_{k=1}^{d} (X_j - \mu_{X_j})}$$

Pearson's correlation coefficient

$$CC(X_i, X_j) = \frac{\sum_{k=1}^{d} (X_i - \mu_{X_i})(X_j - \mu_{X_j})}{\frac{1}{d-1} \sum_{k-1}^{d} (X_i - \mu_{X_i}) \sum_{k=1}^{d} (X_j - \mu_{X_j})}$$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q4: How should object similarity be measured?

Distance measures

L_1 - metrics

$$L_r(X_i, X_j) = \left(\sum_{k=1}^{d} (x_{ik} - x_{jk})^r \right)^{1/r}$$

Mahalanobis distance

$$MD(X_i, X_j) = \frac{\sum_{k=1}^{d} (X_i - X_j)}{\frac{1}{d-1} \sum_{k-1}^{d} (X_i - \mu_{X_i}) \sum_{k=1}^{d} (X_j - \mu_{X_j})}$$

Pearson's correlation coefficient

$$CC(X_i, X_j) = \frac{\sum_{k=1}^{d} (X_i - \mu_{X_i})(X_j - \mu_{X_j})}{\frac{1}{d-1} \sum_{k-1}^{d} (X_i - \mu_{X_i}) \sum_{k=1}^{d} (X_j - \mu_{X_j})}$$

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing

Q5: Should data be standardized?

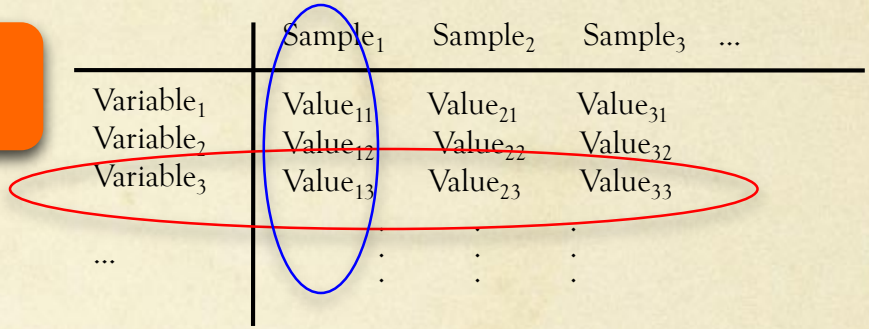
Remark1: Distance measures used to estimate inter-object similarities are sensitive to different scales or magnitudes among the variables.

Remark2: In general, variable with a larger dispersion (standard deviation) will have a bigger impact on the clustering results.

A5: Clustering variables that are not all of the same scale should be standardized.

Cluster Analysis – Design Issues

Stage 2: Design Issues and Pre-Processing



Q5: Should data be standardized?

Standardization techniques:

- Z - score

$$V_i = \frac{V_i - \mu_{V_i}}{\sigma_{V_i}}$$

- Range scaling

$$V_i = \frac{V_i - \min(V_i)}{\max(V_i) - \min(V_i)}$$

- Variable standardization
- Sample standardization

Overview

- What is cluster analysis? ☑
- Some definitions and notations ☑
- How it works? ☑
- Cluster Analysis Diagram
  - Objectives of cluster analysis ☑
  - Design issues ☑
  - Assumptions in cluster analysis
  - Clustering methods
  - Interpreting the clusters
  - Validation

Cluster Analysis – Assumptions

Stage 3: Assumptions in Cluster Analysis

1. It is always assumed that the sample is representative for the population

2. It is assumed that variables are not correlated; if variables are correlated, remove correlated variables or use distance measures that compensates for the correlation such as Mahanalobis distance

Overview

- What is cluster analysis? ☑
- Some definitions and notations ☑
- How it works? ☑
- Cluster Analysis Diagram
  - Objectives of cluster analysis ☑
  - Design issues ☑
  - Assumptions in cluster analysis ☑
  - Clustering methods
  - Interpreting the clusters
  - Validation

Cluster Analysis – Methods

Stage 4: Deriving Clusters and Assessing Overall Fit

Methods:

Cluster Analysis – Methods

Stage 4: Deriving Clusters and Assessing Overall Fit

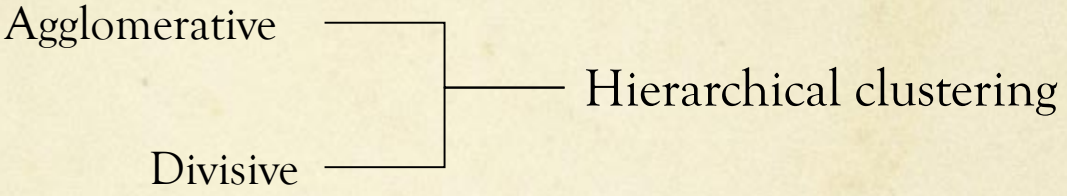
Methods:

Hierarchical clustering

Cluster Analysis – Methods

Stage 4: Deriving Clusters and Assessing Overall Fit

Methods:



Cluster Analysis – Methods

Stage 4: Deriving Clusters and Assessing Overall Fit

Methods:

- Agglomerative
- Divisive

→ Hierarchical clustering

Partitional clustering

Cluster Analysis – Methods

Stage 4: Deriving Clusters and Assessing Overall Fit

Methods:

- Agglomerative
- Divisive

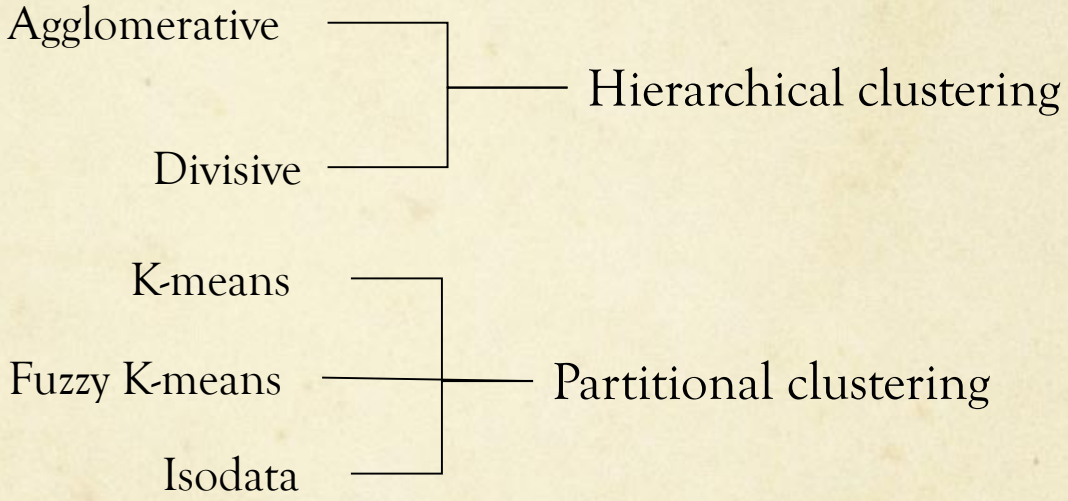
→ Hierarchical clustering

Partitional clustering

Cluster Analysis – Methods

Stage 4: Deriving Clusters and Assessing Overall Fit

Methods:



Cluster Analysis – Methods

Stage 4: Deriving Clusters and Assessing Overall Fit

Methods:

- Agglomerative
- Divisive

→ Hierarchical clustering

- K-means
- Fuzzy K-means
- Isodata

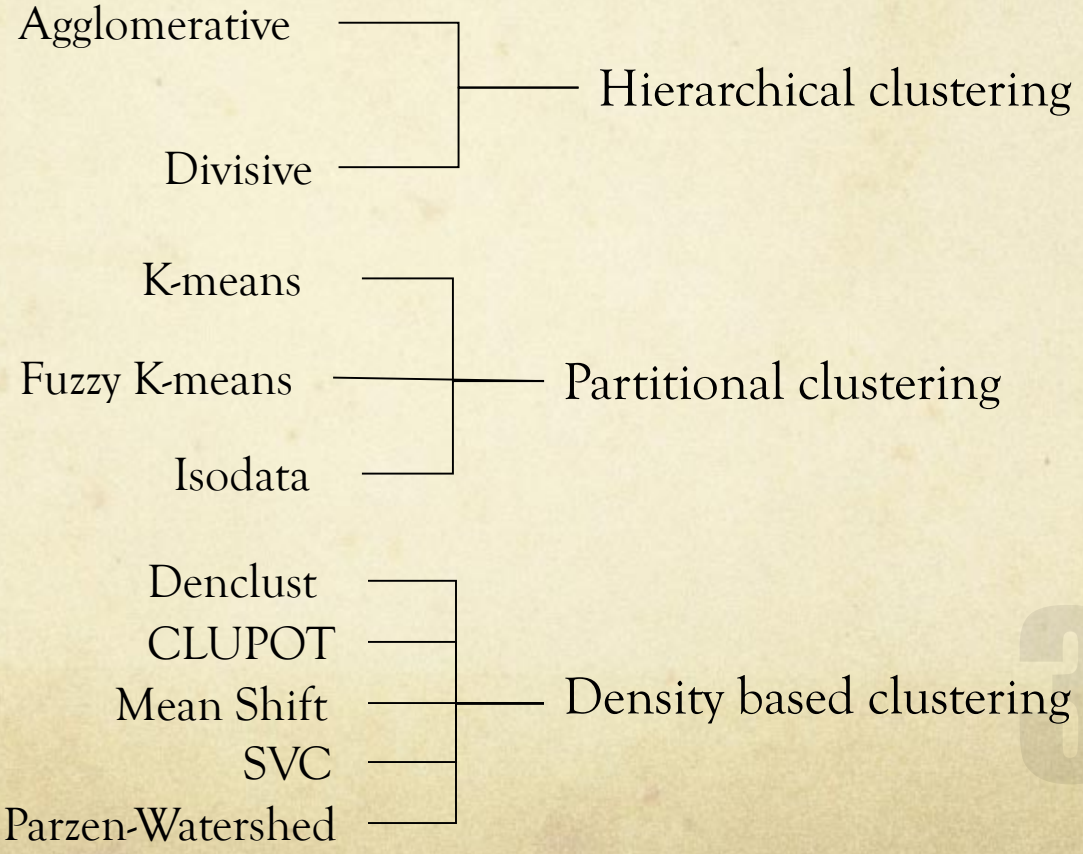
→ Partitional clustering

Density based clustering

Cluster Analysis – Methods

Stage 4: Deriving Clusters and Assessing Overall Fit

Methods:



Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

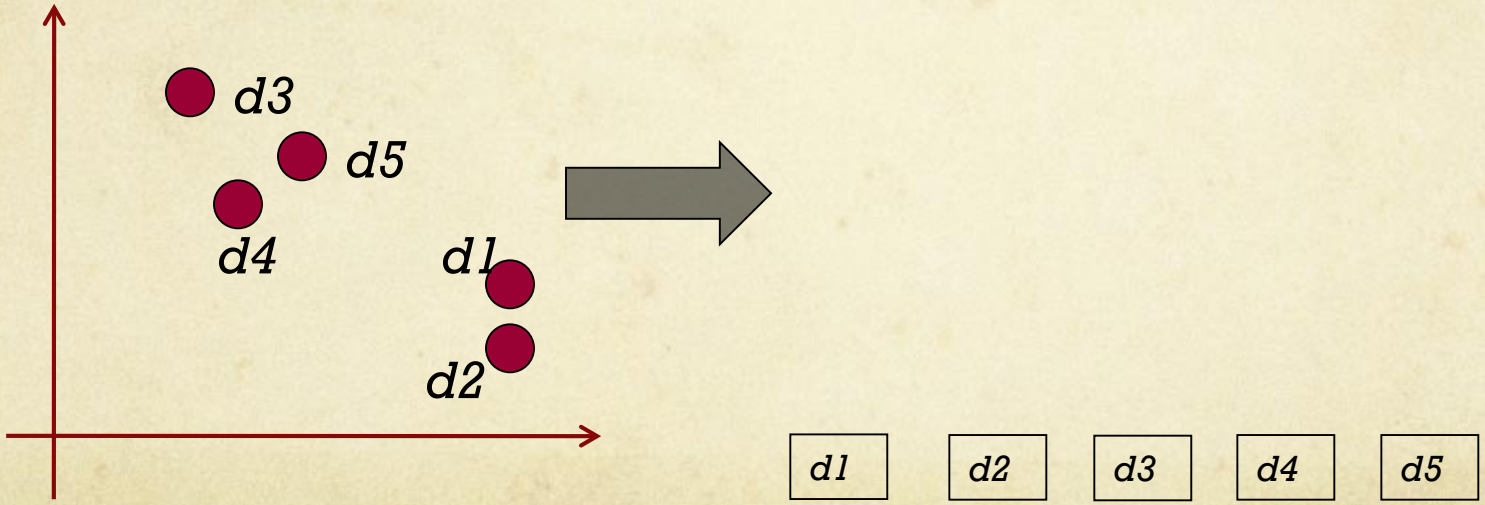
Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

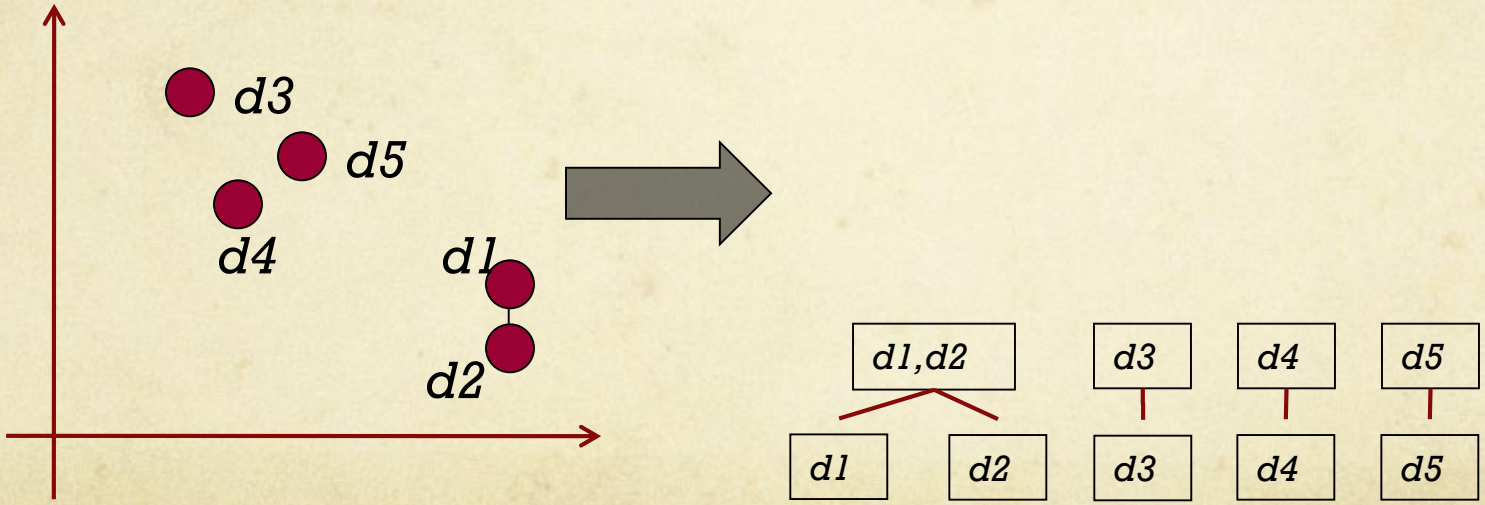


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

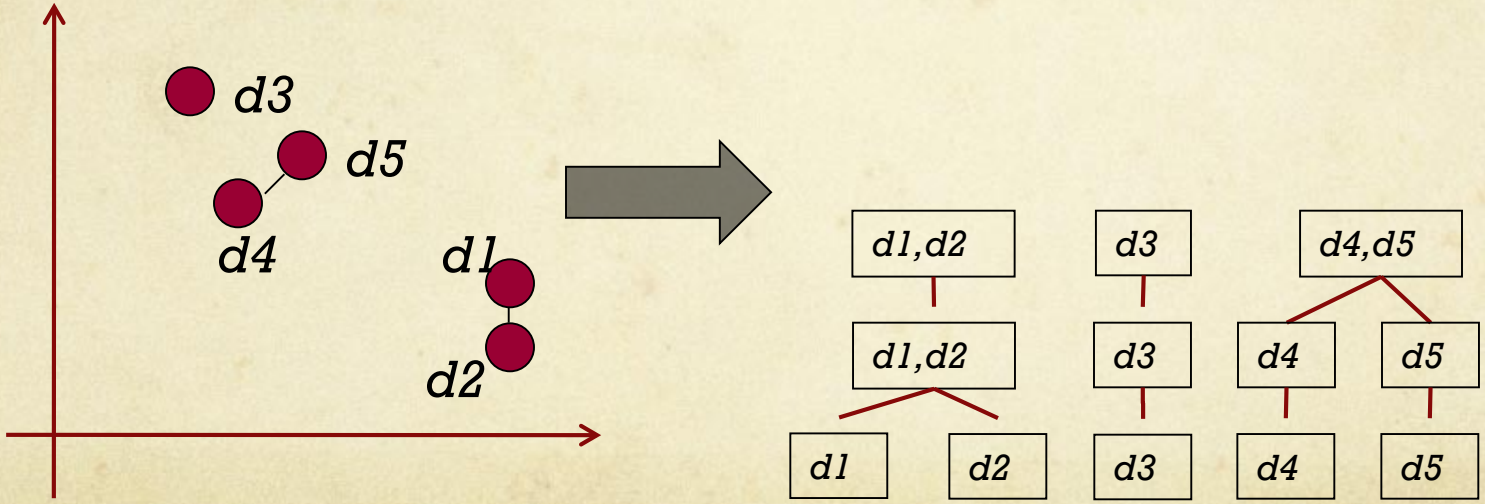


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

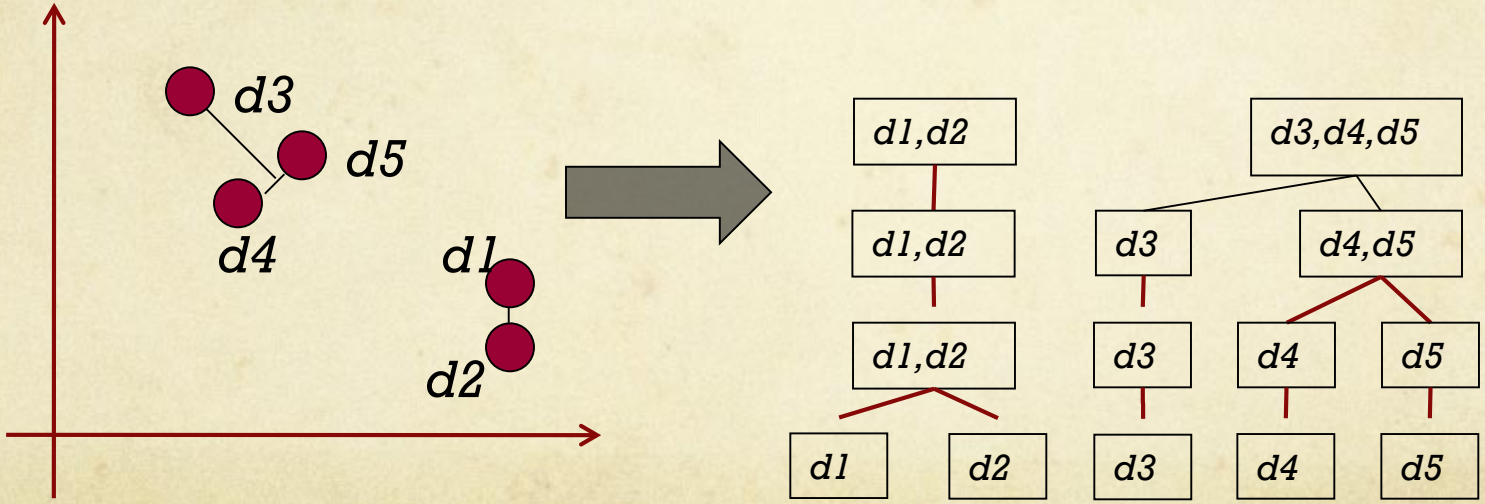


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

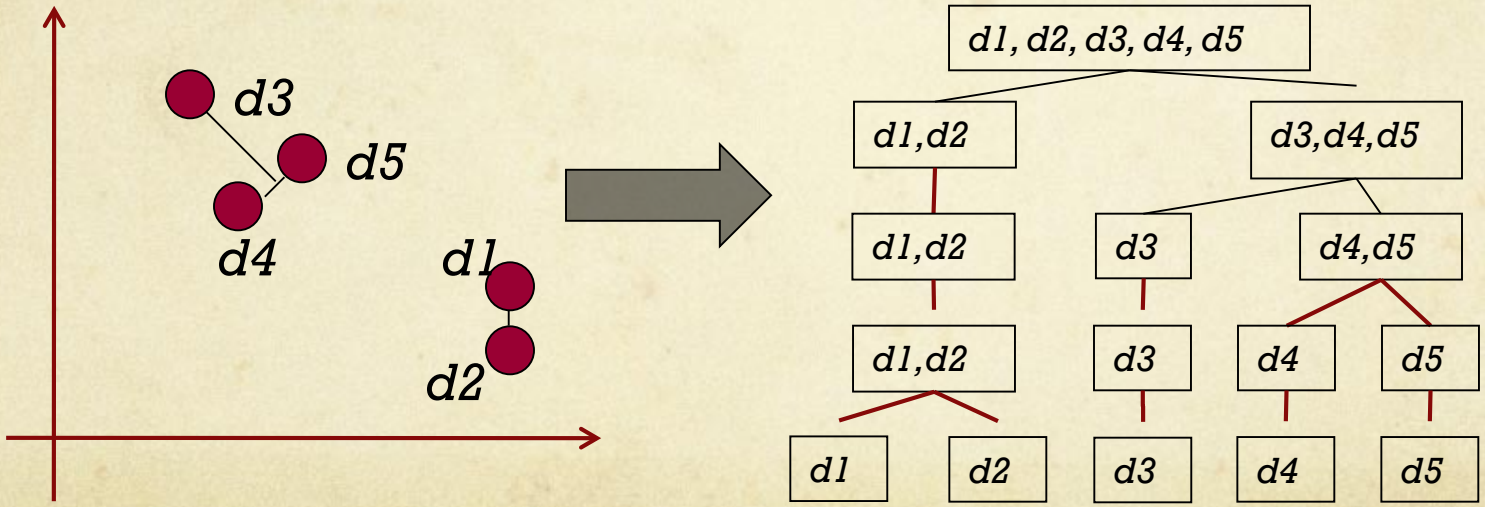


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

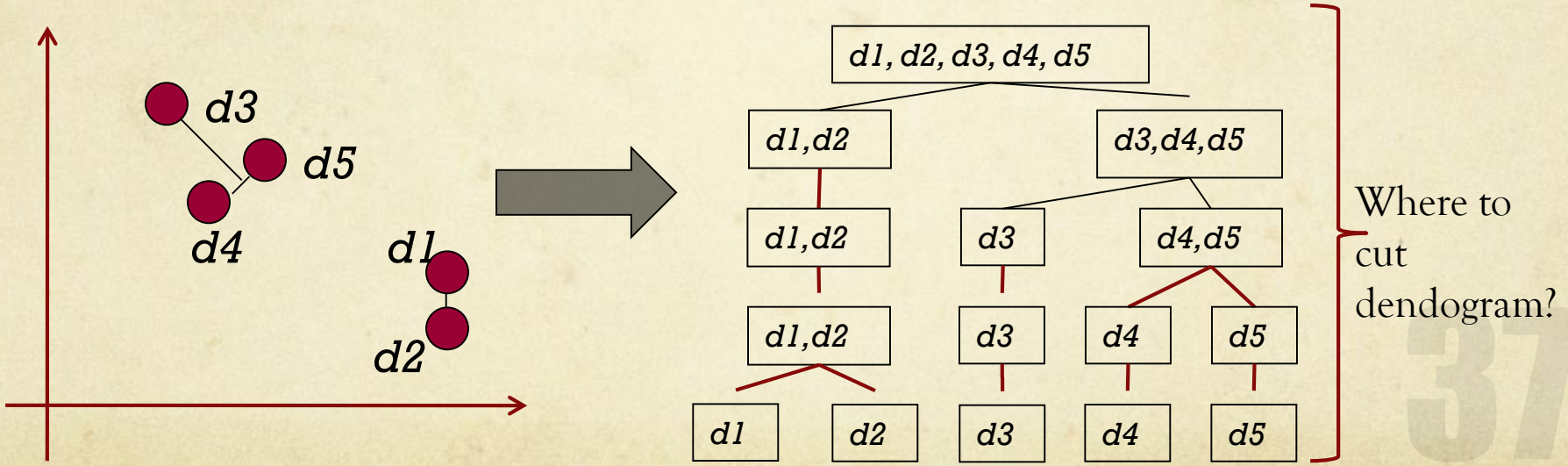


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

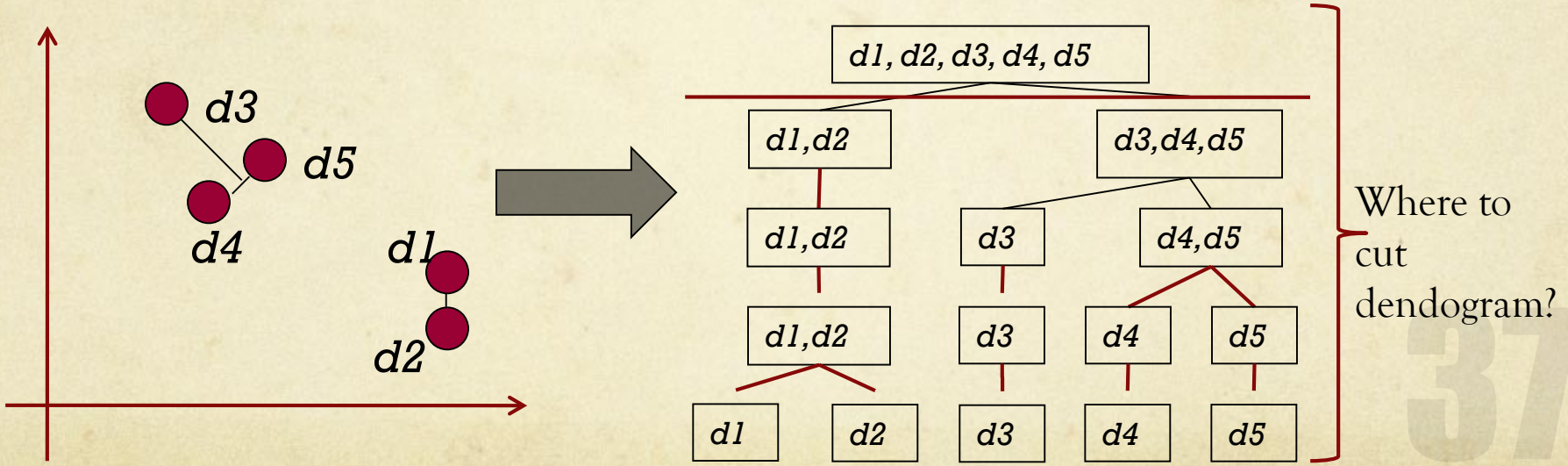


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

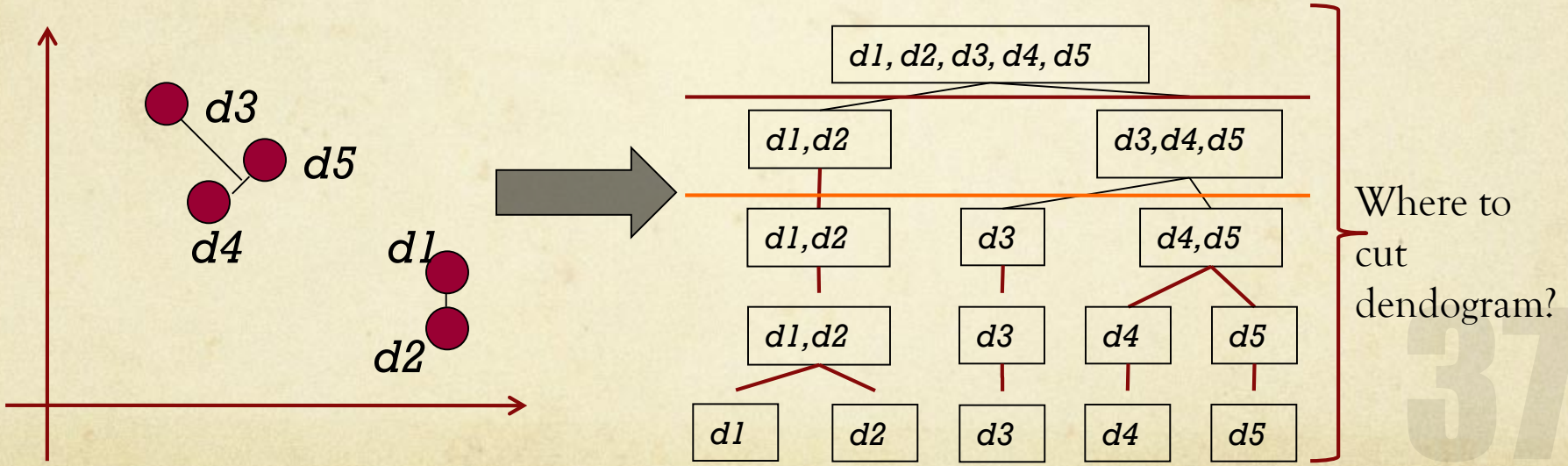


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

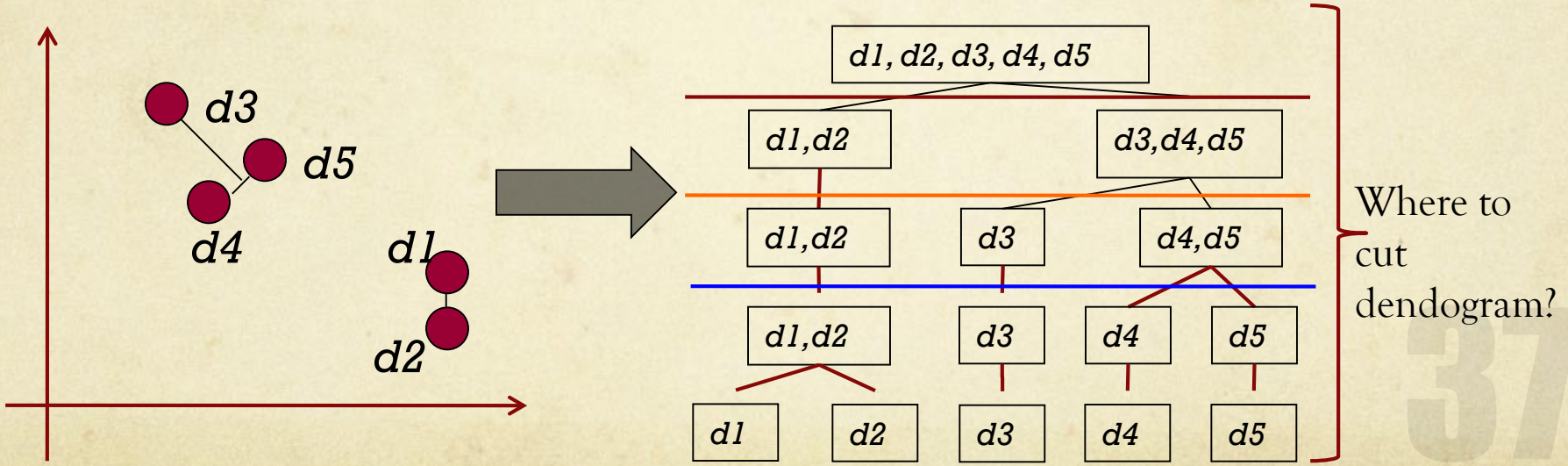


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.

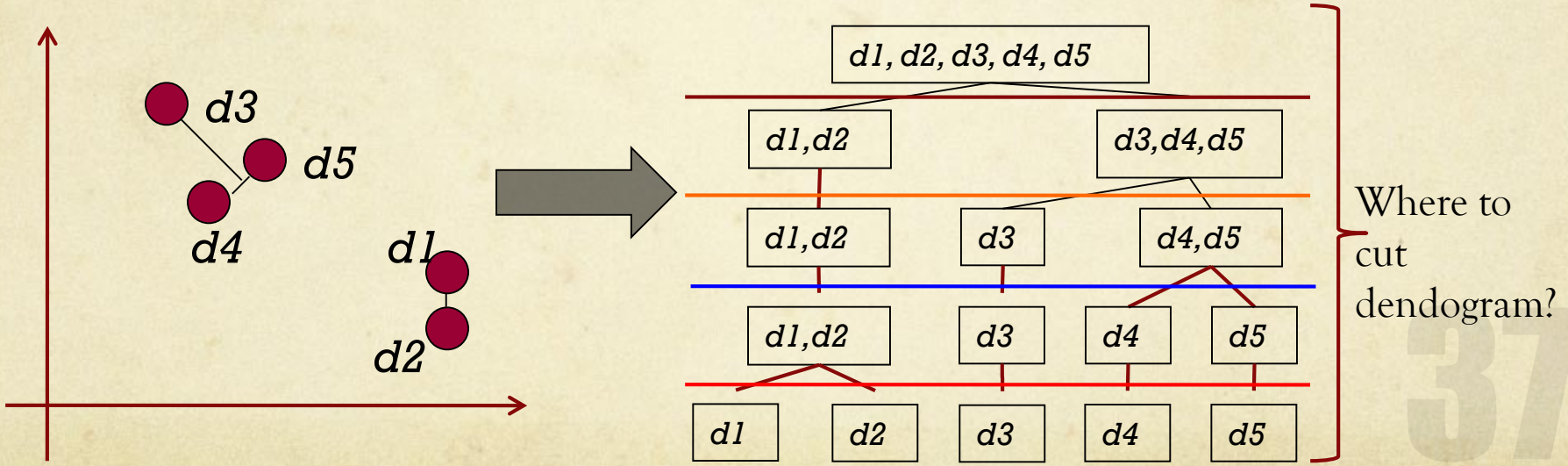


Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Principle: compute the Distance-Matrix between all objects (initially one object = one cluster). Find the two clusters with the closest distance and put those two clusters into one. Compute the new Distance-Matrix.



Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Single-link (nearest neighbor method)

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Single-link (nearest neighbor method)

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

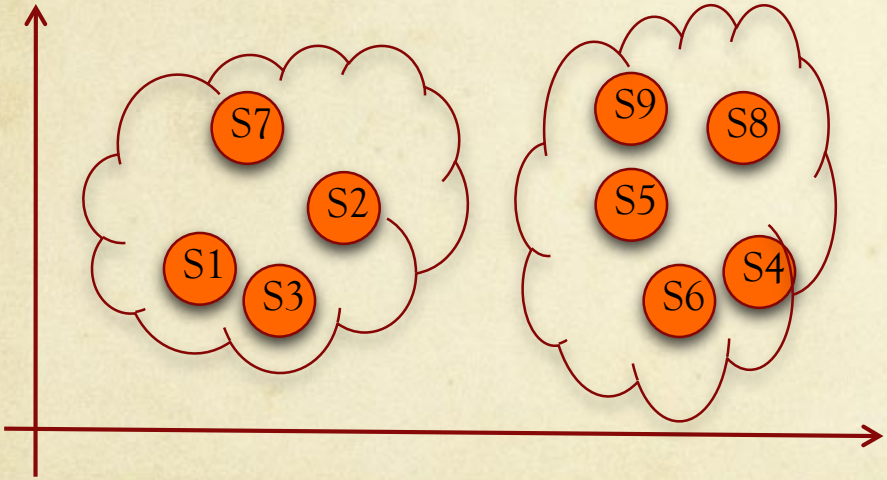
Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Single-link (nearest neighbor method)

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$



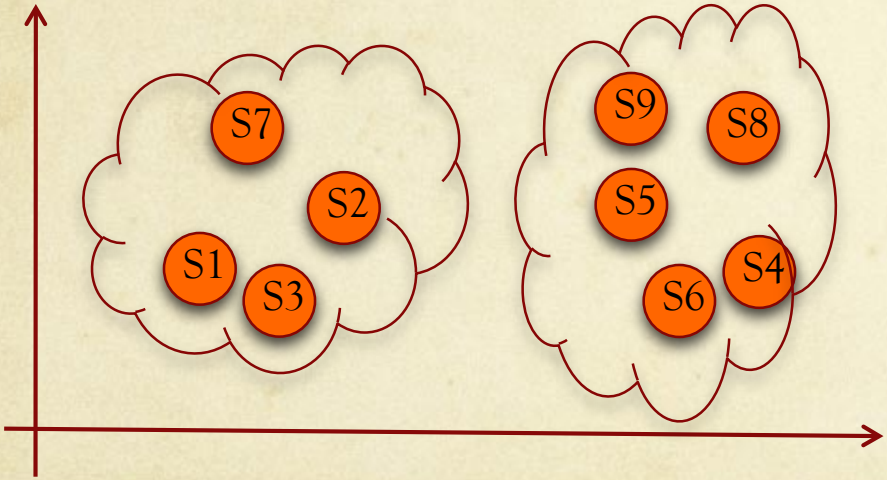
Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Single-link (nearest neighbor method)

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$



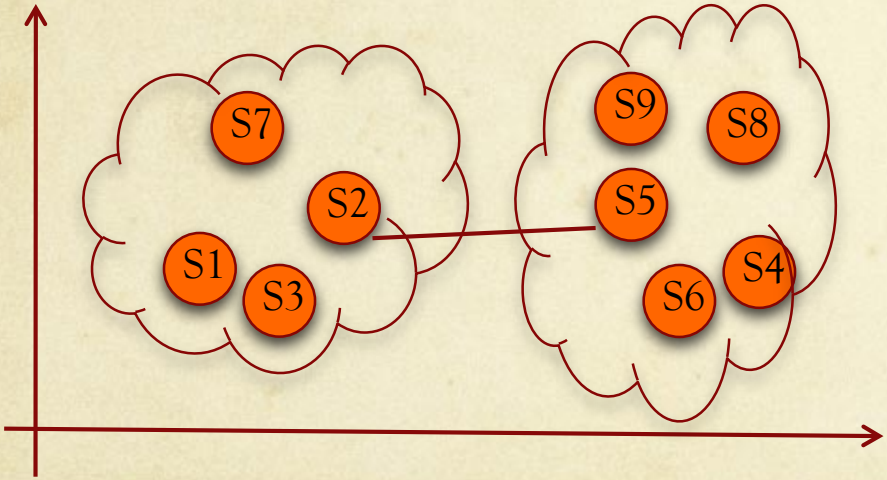
Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Single-link (nearest neighbor method)

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$



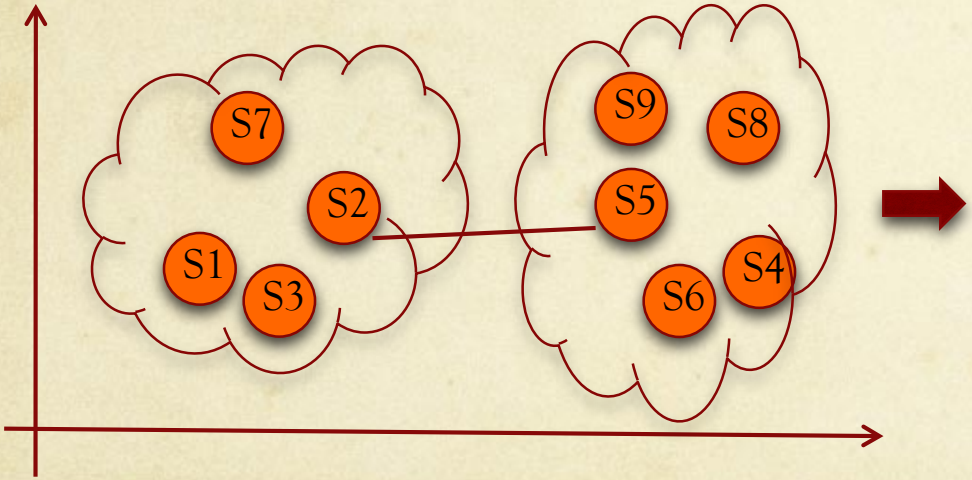
Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Single-link (nearest neighbor method)

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$



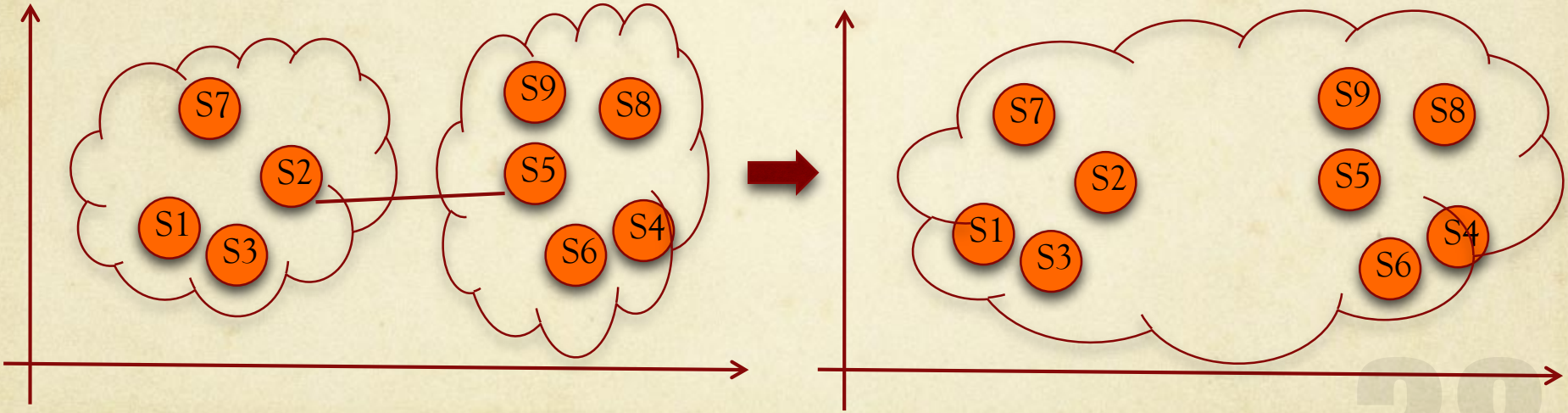
Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Single-link (nearest neighbor method)

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$



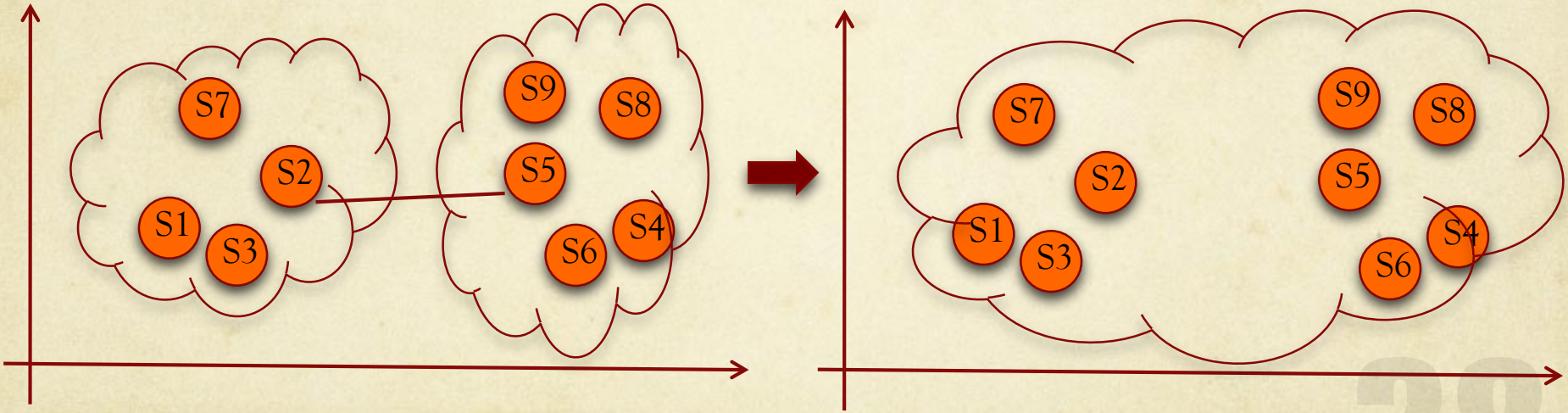
Cluster Analysis – Methods

Hierarchical clustering

Agglomerative (bottom - up)

Single-link (nearest neighbor method)

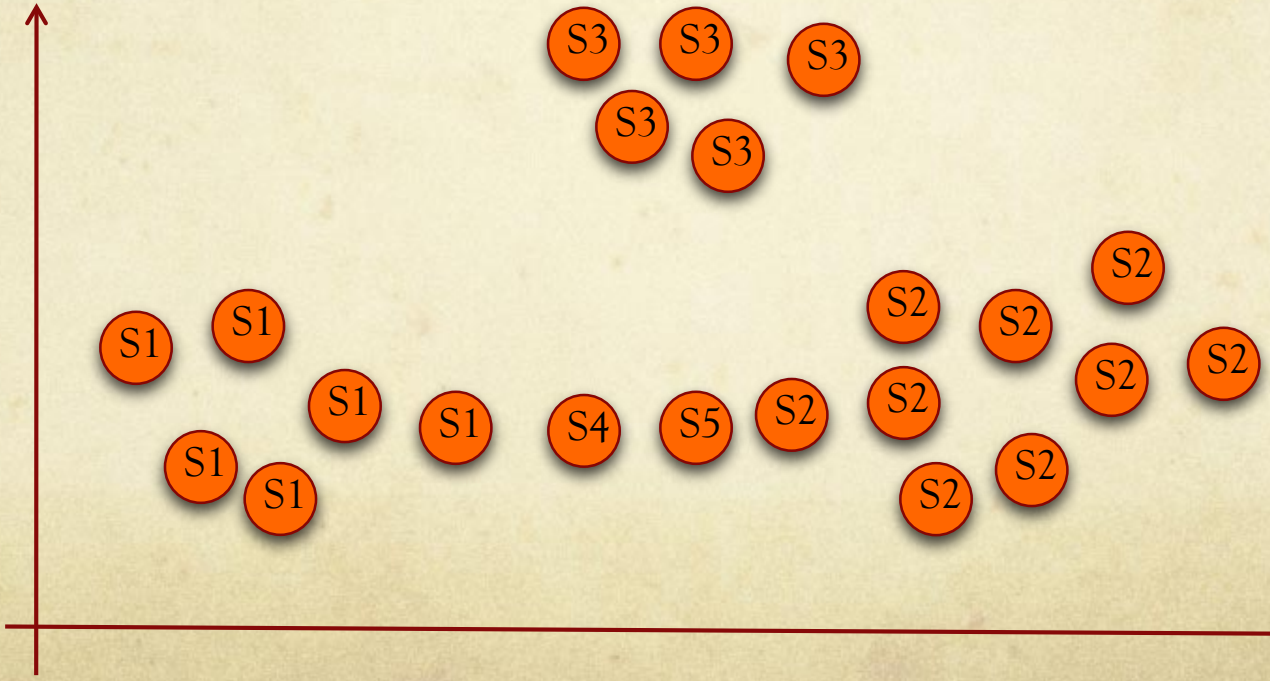
$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$



Drawback: can result in long and thin clusters due to chaining effect

Cluster Analysis – Methods

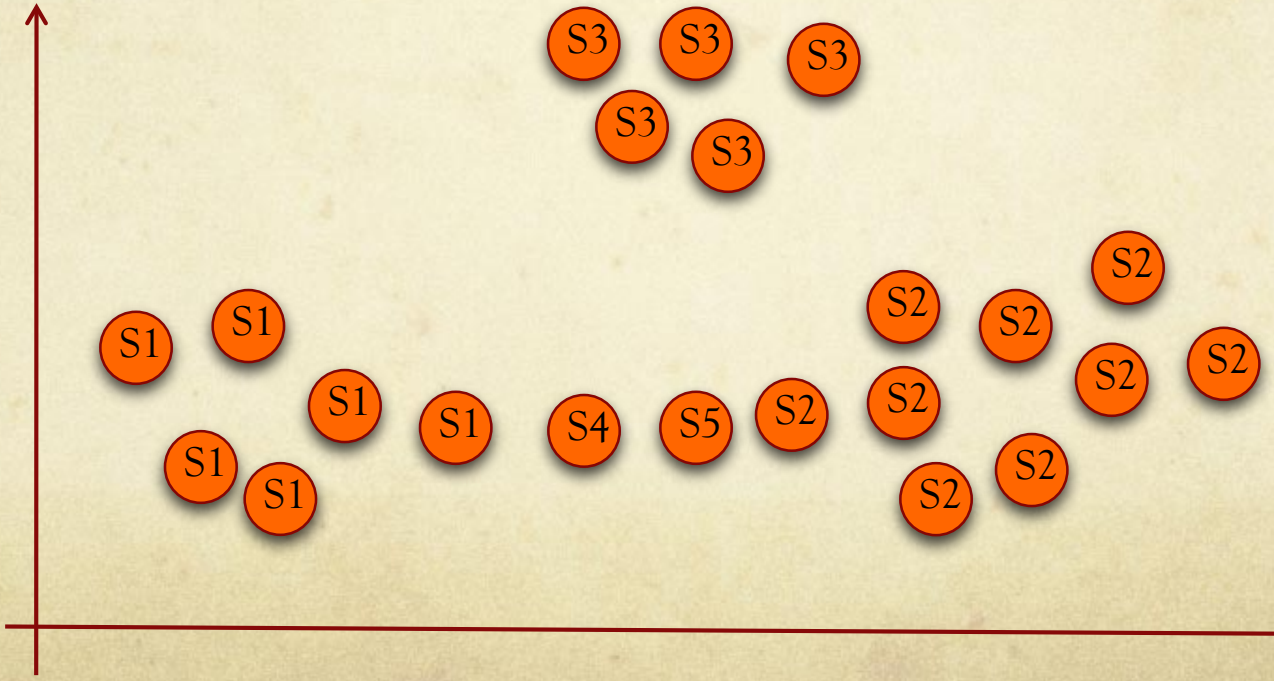
Single-link (nearest neighbor method)



Cluster Analysis – Methods

Single-link (nearest neighbor method)

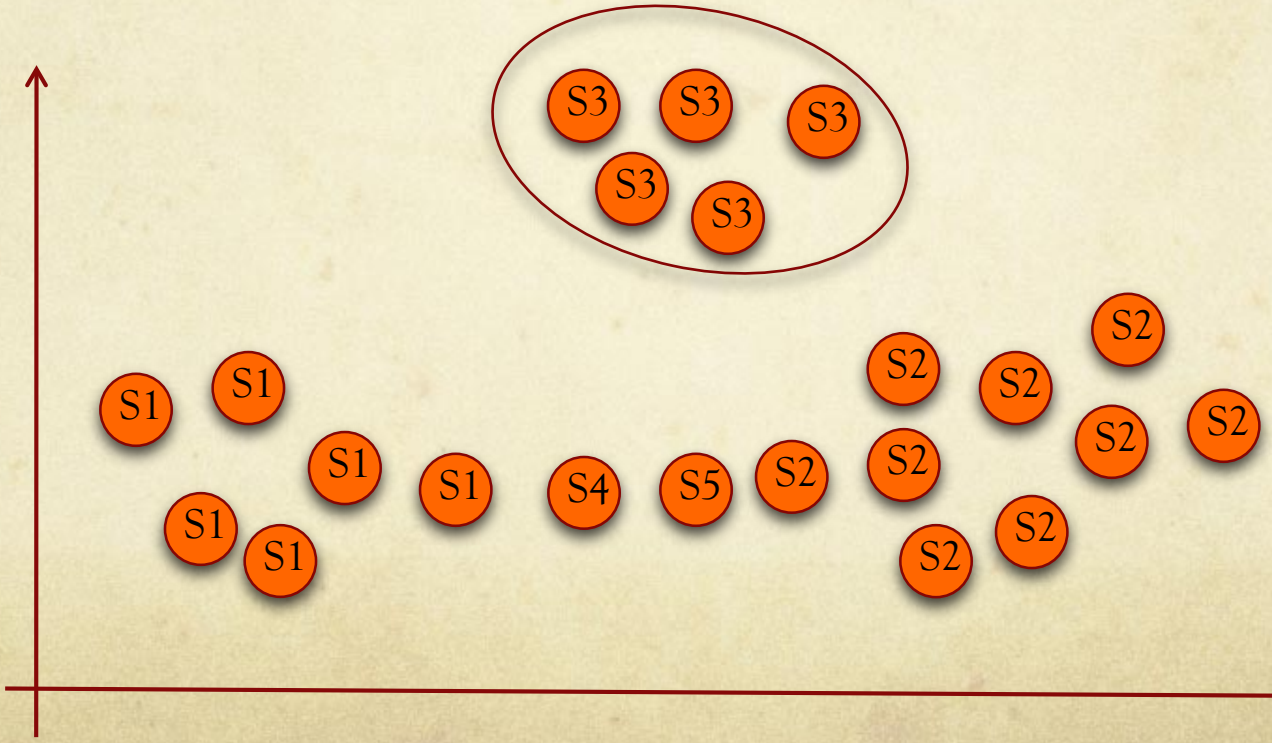
Drawback: can result in long and thin clusters due to chaining effect



Cluster Analysis – Methods

Single-link (nearest neighbor method)

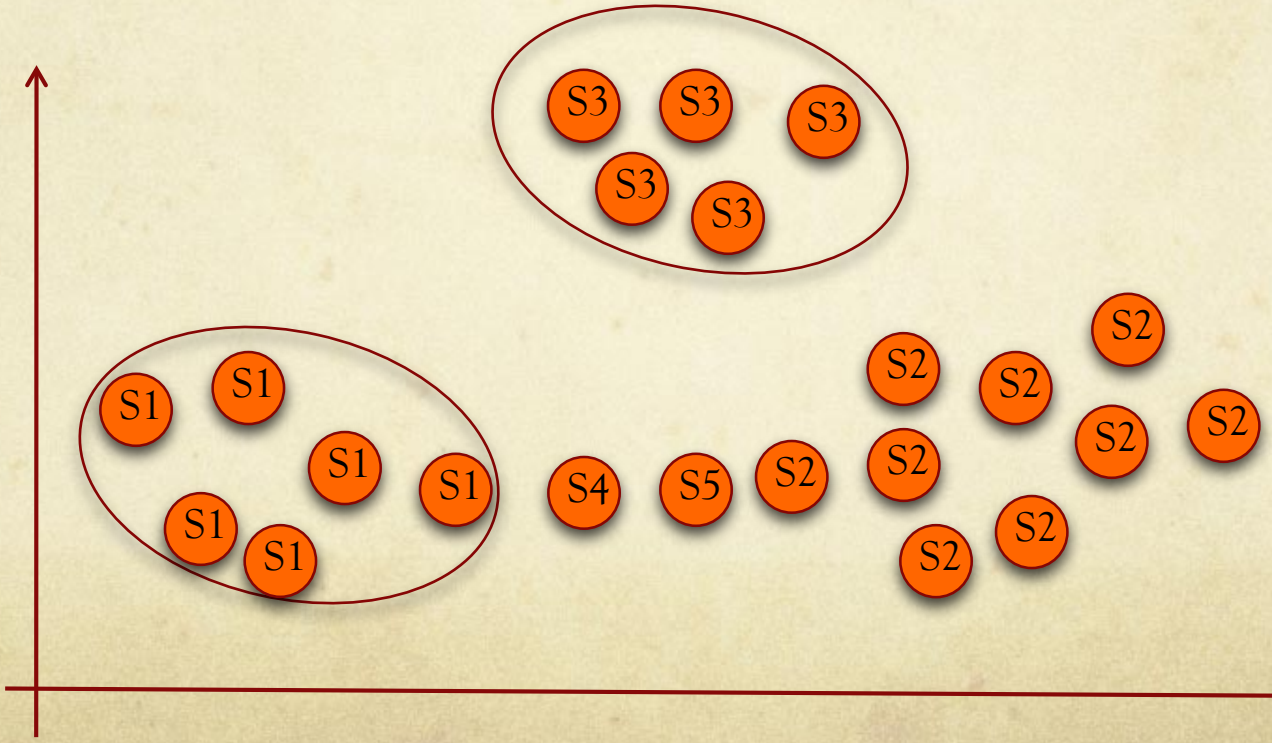
Drawback: can result in long and thin clusters due to chaining effect



Cluster Analysis – Methods

Single-link (nearest neighbor method)

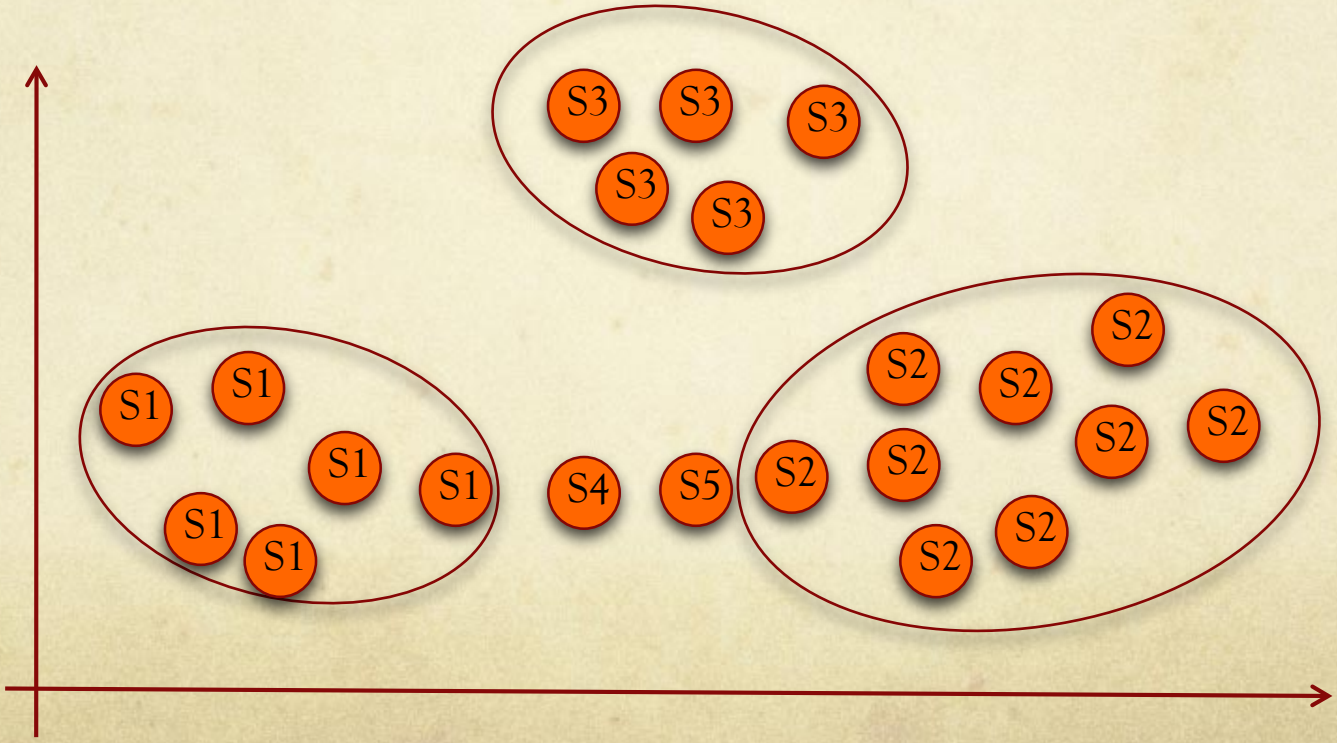
Drawback: can result in long and thin clusters due to chaining effect



Cluster Analysis – Methods

Single-link (nearest neighbor method)

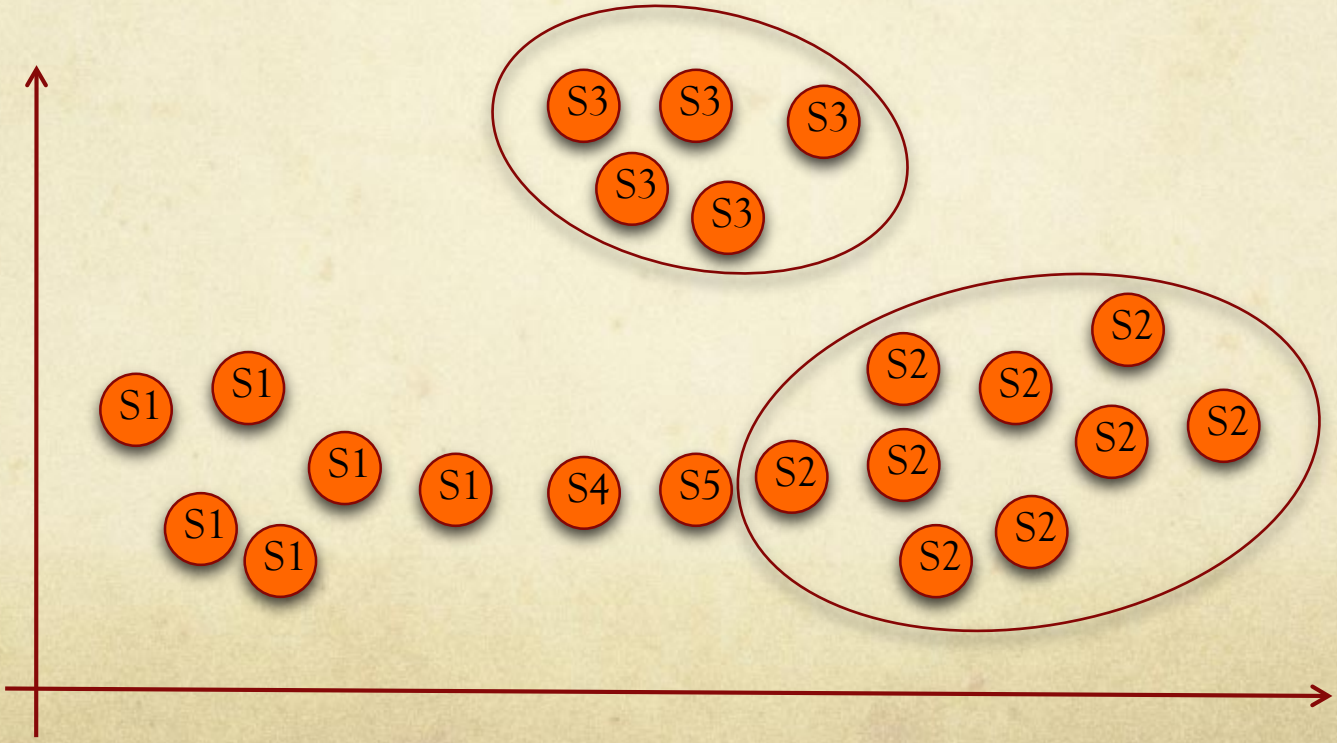
Drawback: can result in long and thin clusters due to chaining effect



Cluster Analysis – Methods

Single-link (nearest neighbor method)

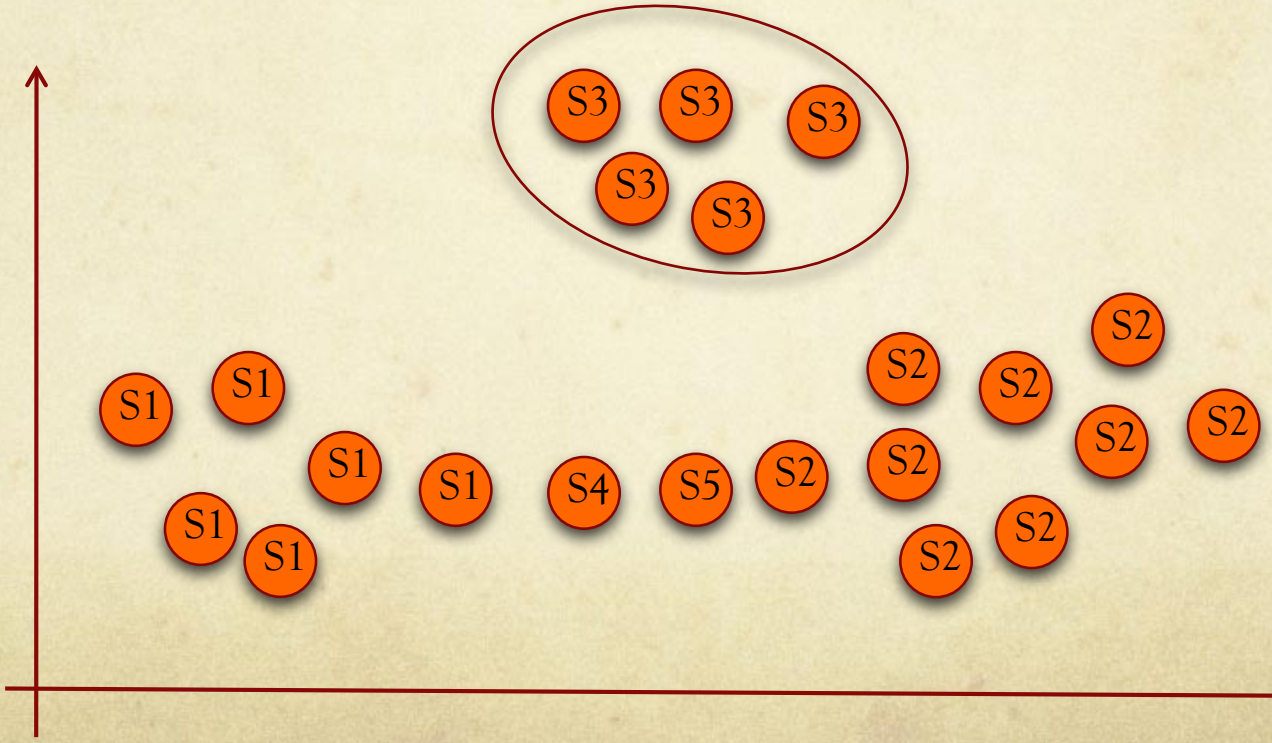
Drawback: can result in long and thin clusters due to chaining effect



Cluster Analysis – Methods

Single-link (nearest neighbor method)

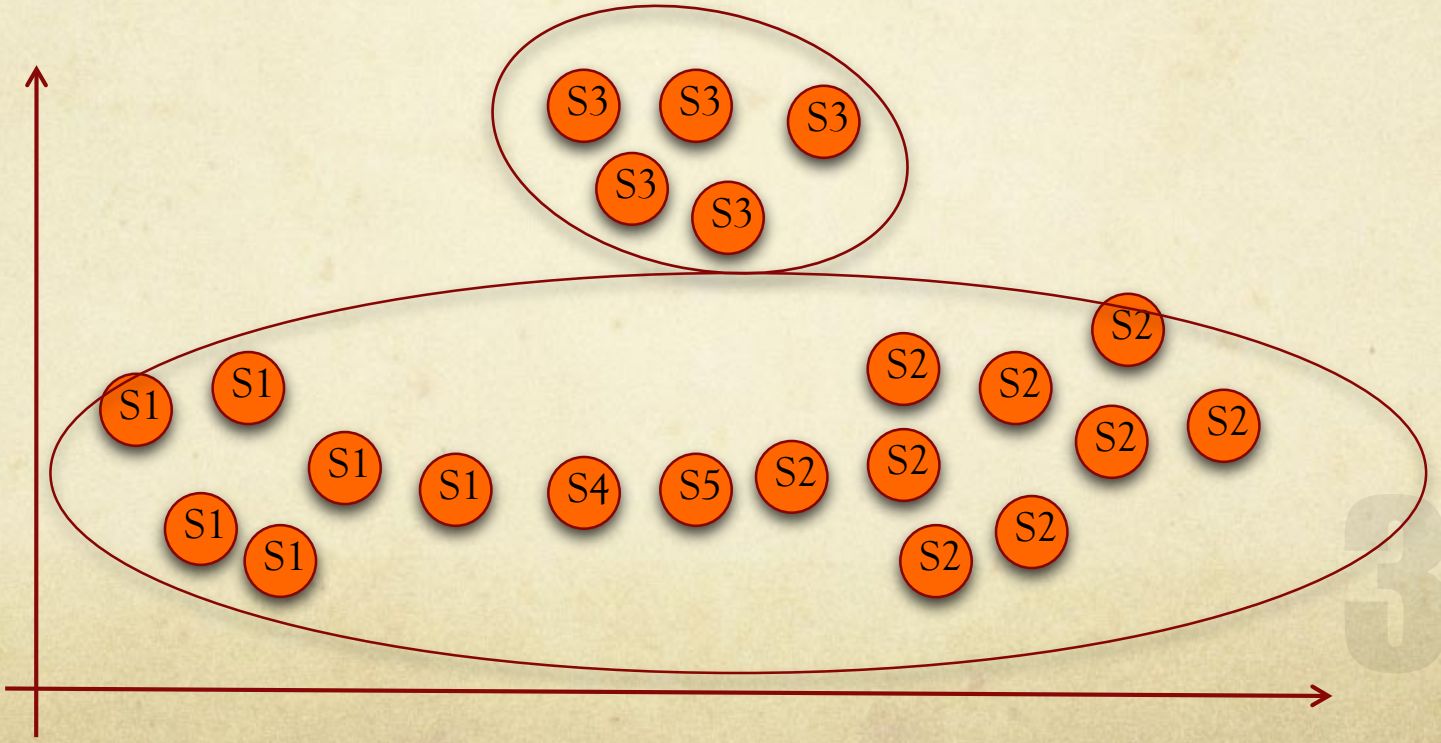
Drawback: can result in long and thin clusters due to chaining effect



Cluster Analysis – Methods

Single-link (nearest neighbor method)

Drawback: can result in long and thin clusters due to chaining effect



Cluster Analysis – Methods

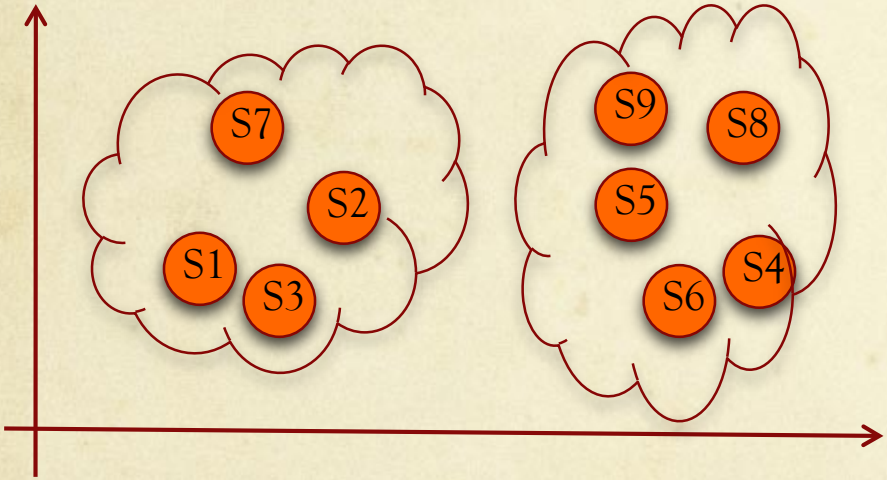
Complete-linkage (furthest-neighbor or diameter method)

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

Cluster Analysis – Methods

Complete-linkage (furthest-neighbor or diameter method)

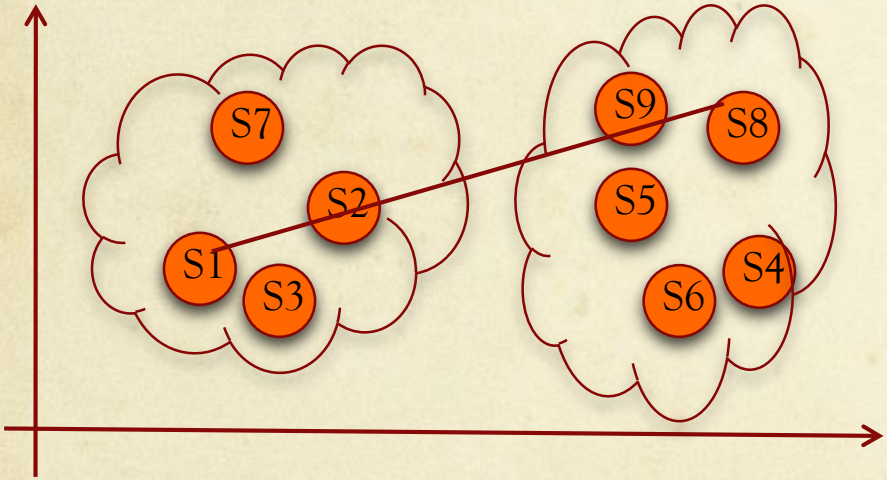
$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$



Cluster Analysis – Methods

Complete-linkage (furthest-neighbor or diameter method)

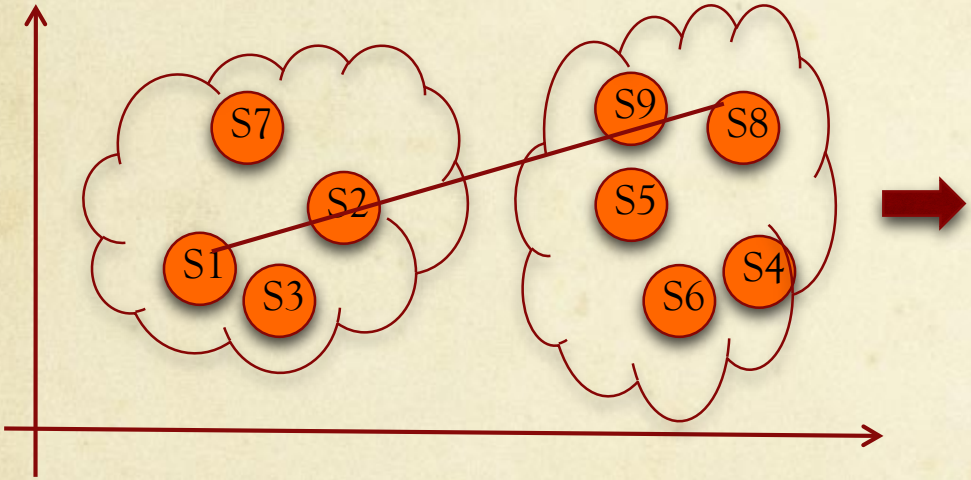
$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$



Cluster Analysis – Methods

Complete-linkage (furthest-neighbor or diameter method)

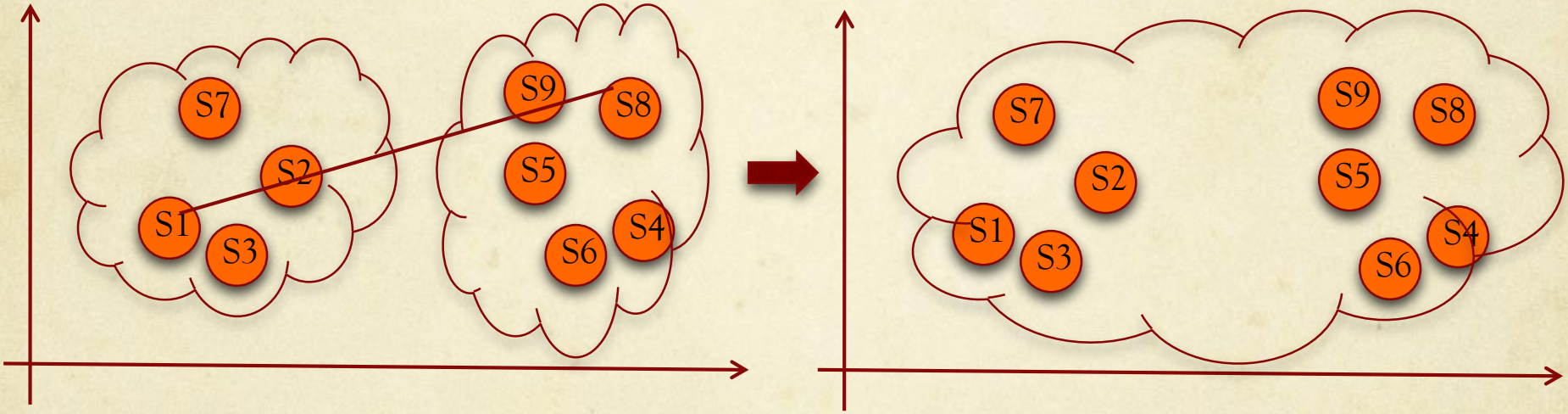
$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$



Cluster Analysis – Methods

Complete-linkage (furthest-neighbor or diameter method)

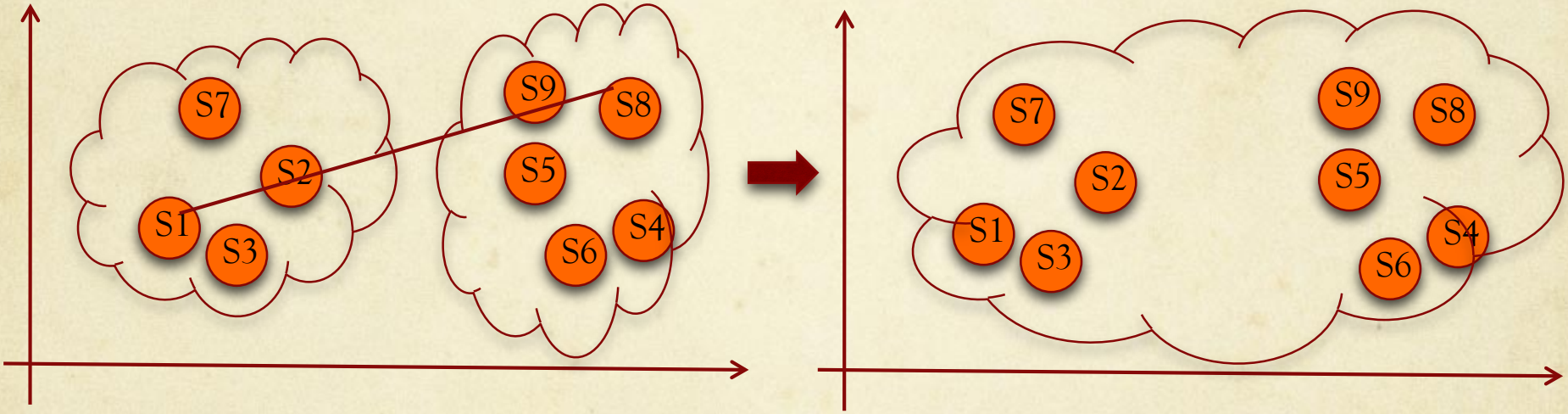
$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$



Cluster Analysis – Methods

Complete-linkage (furthest-neighbor or diameter method)

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$



Drawback: makes spherical clusters

Cluster Analysis – Methods

Average-linkage (Centroid method)

Similarity between clusters is the average distance between all objects in one cluster and all objects in other cluster

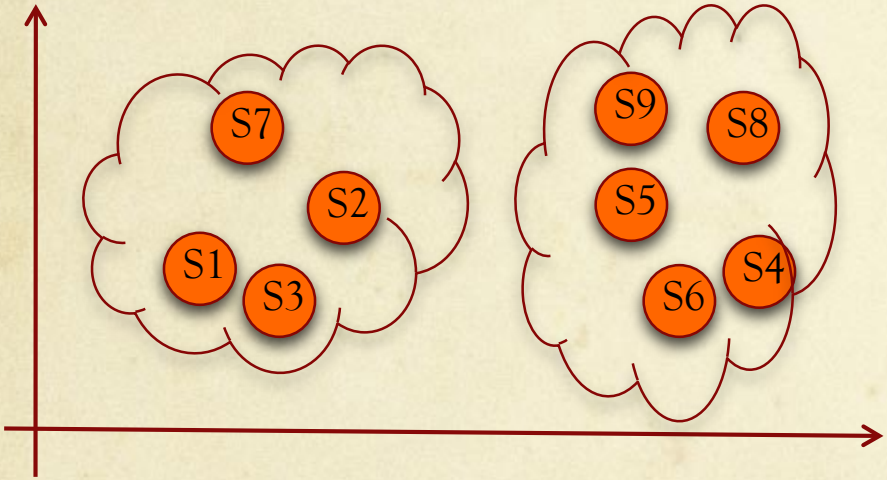
Advantage: less affected by outliers

Drawback: generates clusters with approximately equal within cluster variation

Cluster Analysis – Methods

Average-linkage (Centroid method)

Similarity between clusters is the average distance between all objects in one cluster and all objects in other cluster



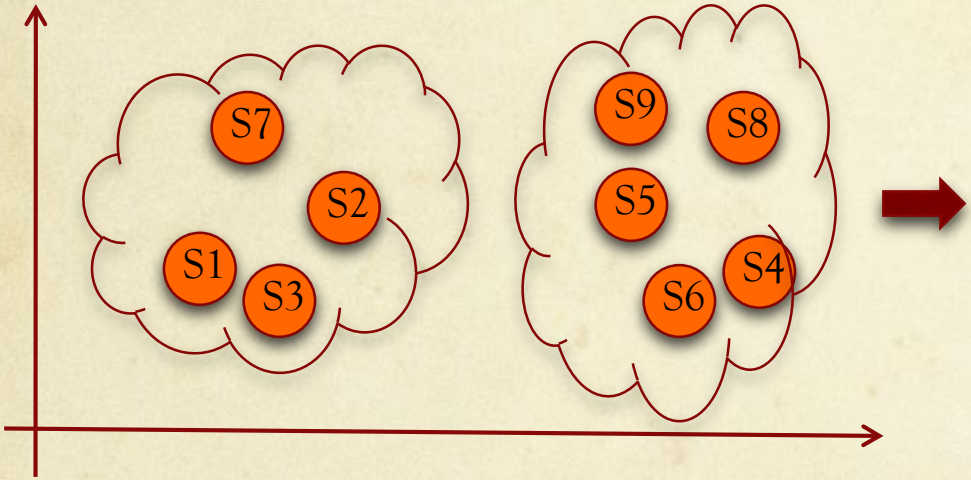
Advantage: less affected by outliers

Drawback: generates clusters with approximately equal within cluster variation

Cluster Analysis – Methods

Average-linkage (Centroid method)

Similarity between clusters is the average distance between all objects in one cluster and all objects in other cluster



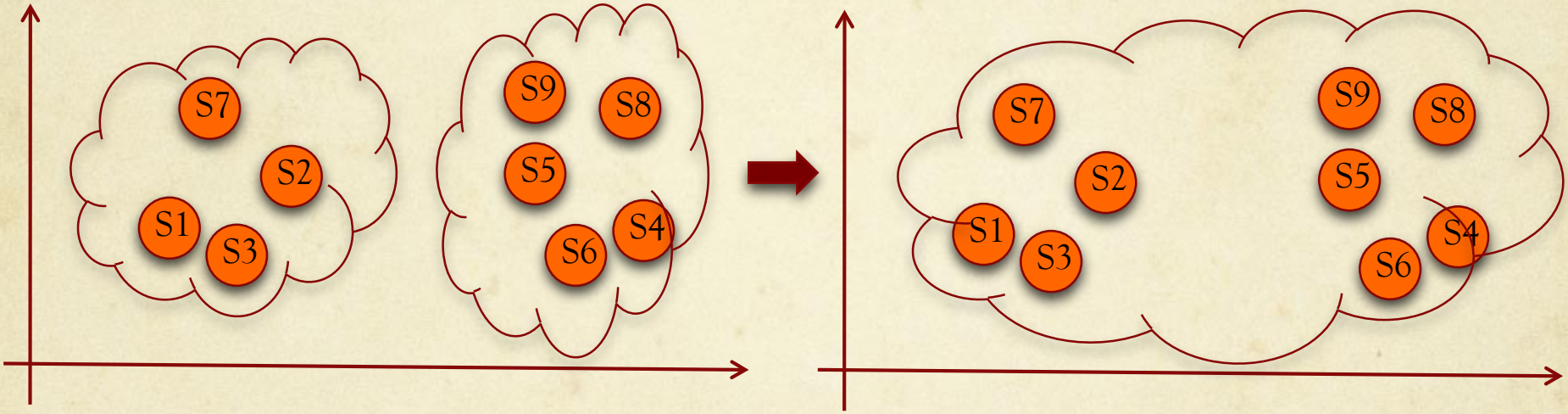
Advantage: less affected by outliers

Drawback: generates clusters with approximately equal within cluster variation

Cluster Analysis – Methods

Average-linkage (Centroid method)

Similarity between clusters is the average distance between all objects in one cluster and all objects in other cluster



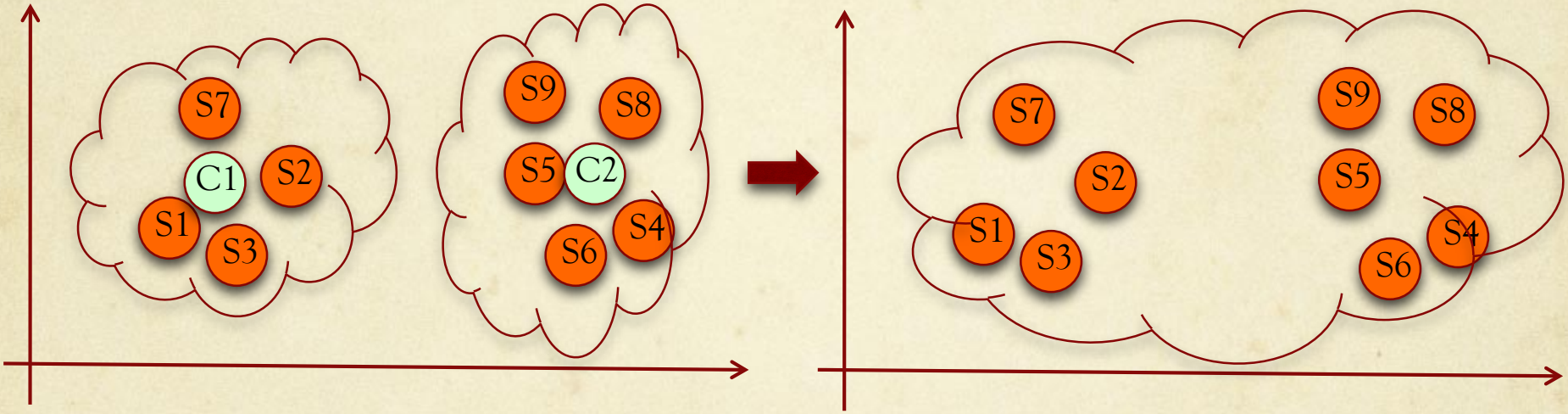
Advantage: less affected by outliers

Drawback: generates clusters with approximately equal within cluster variation

Cluster Analysis – Methods

Average-linkage (Centroid method)

Similarity between clusters is the average distance between all objects in one cluster and all objects in other cluster



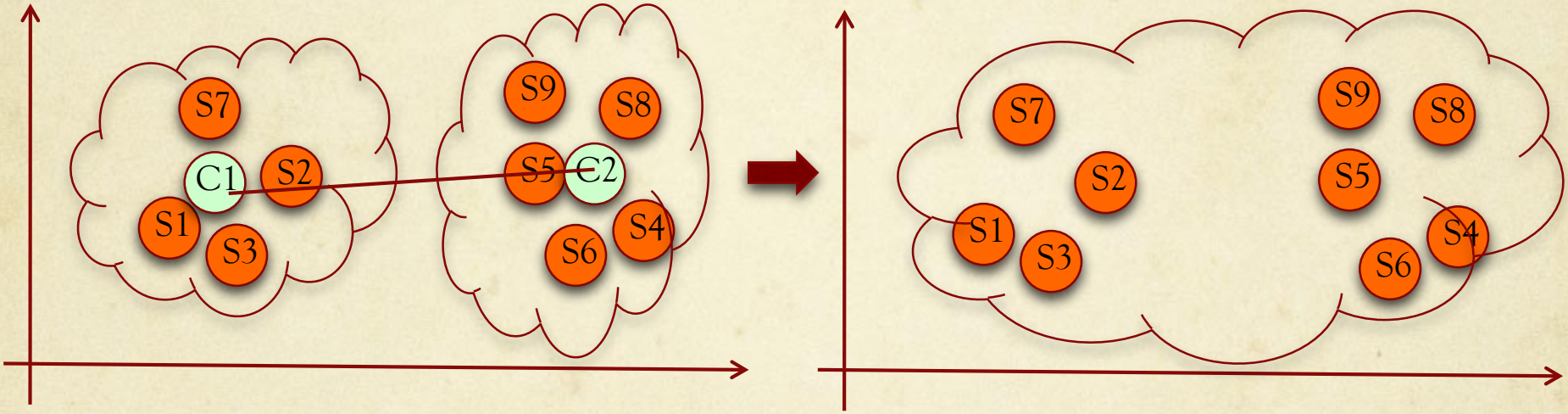
Advantage: less affected by outliers

Drawback: generates clusters with approximately equal within cluster variation

Cluster Analysis – Methods

Average-linkage (Centroid method)

Similarity between clusters is the average distance between all objects in one cluster and all objects in other cluster



Advantage: less affected by outliers

Drawback: generates clusters with approximately equal within cluster variation

Cluster Analysis – Methods

Divisive (top-down)	- divisive algorithms need much more computing power so in practical only agglomerative methods are used
Computational complexity	- $O(n^2)$ - optimal
Drawbacks	- computation of similarity matrix between all pairs of points; for large datasets this is computational expensive

Cluster Analysis – Methods

Partitional clustering

- A typical clustering analysis approach via partitioning data set iteratively
- Statement of the problem: given a *K*, find a partition of *K clusters* to optimize the chosen partitioning criterion
- In principle, partitions achieved via minimizing the sum of squared distances in each cluster

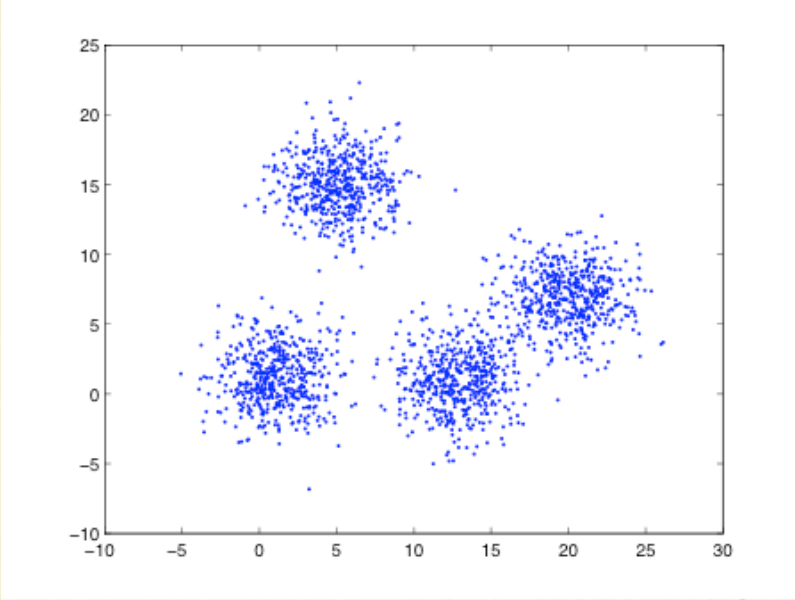
$$E = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} \| \mathbf{x} - \mathbf{m}_i \|^2$$

K-means - (MacQueen'67): each cluster is represented by the centre of the cluster and the algorithm converges to stable centers of clusters

Cluster Analysis – Methods

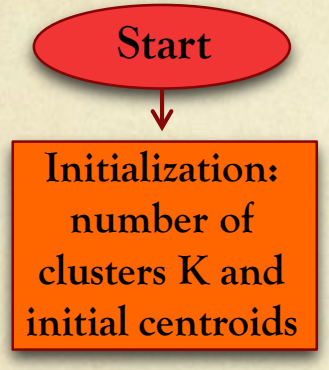
K-means algorithm

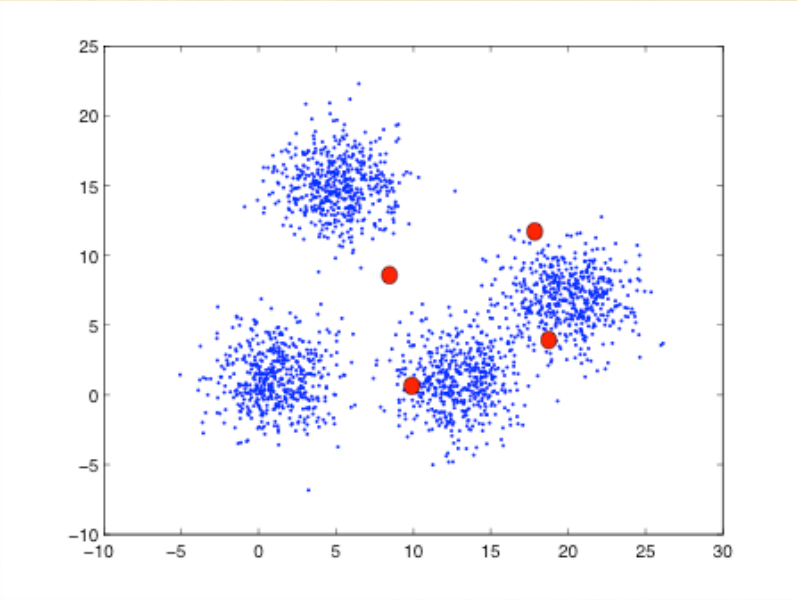
Start



Cluster Analysis – Methods

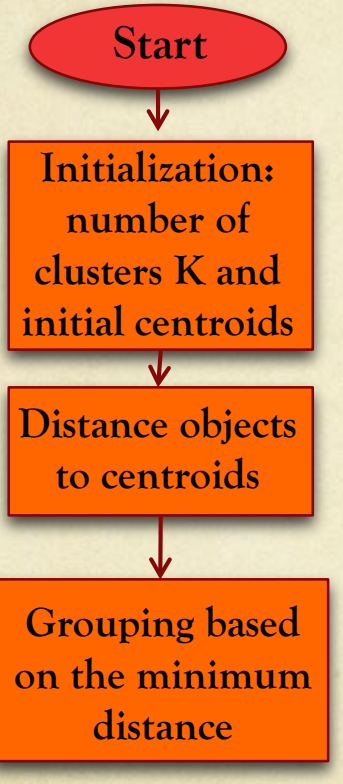
K-means algorithm

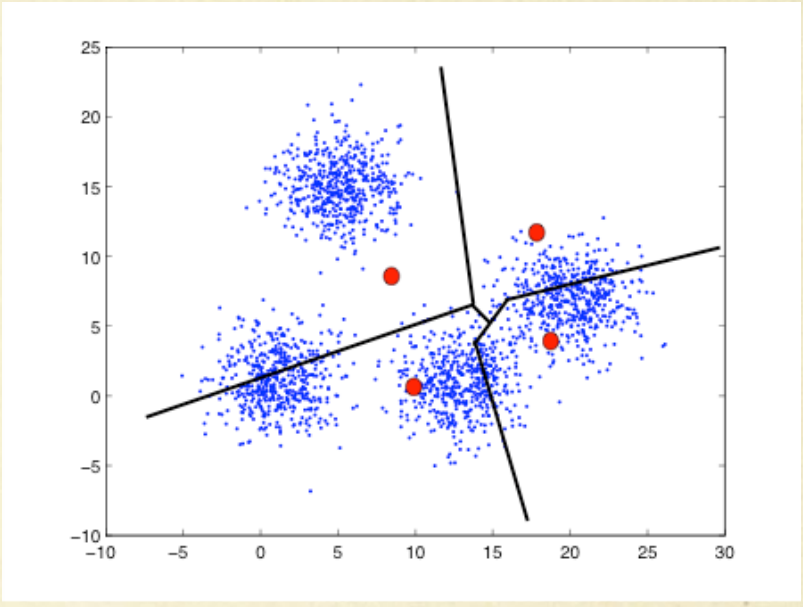




Cluster Analysis – Methods

K-means algorithm

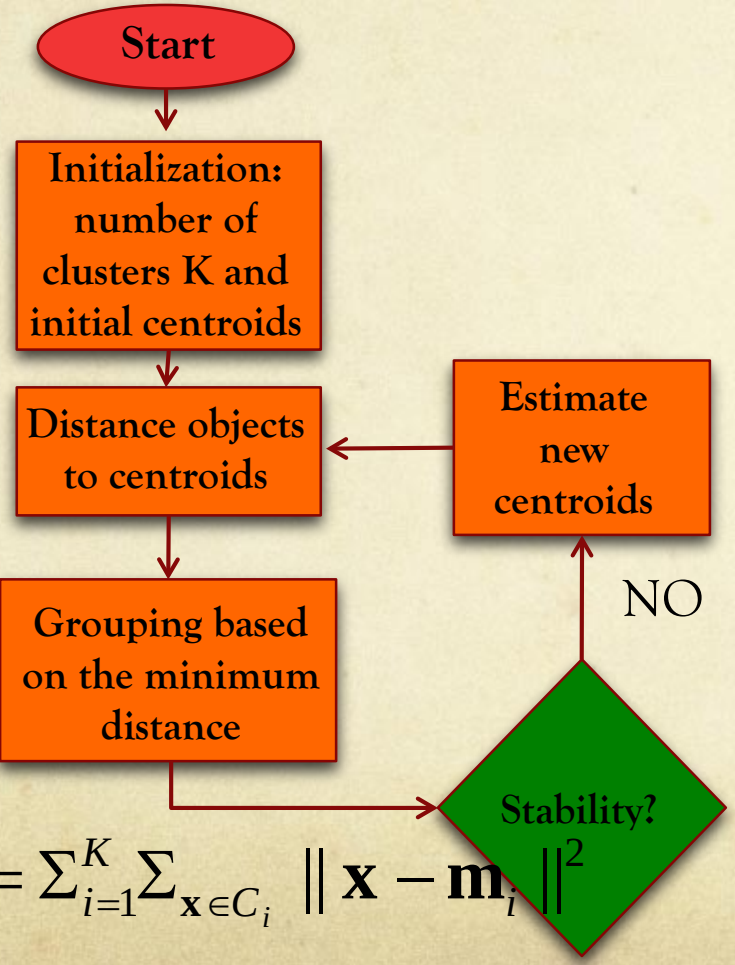




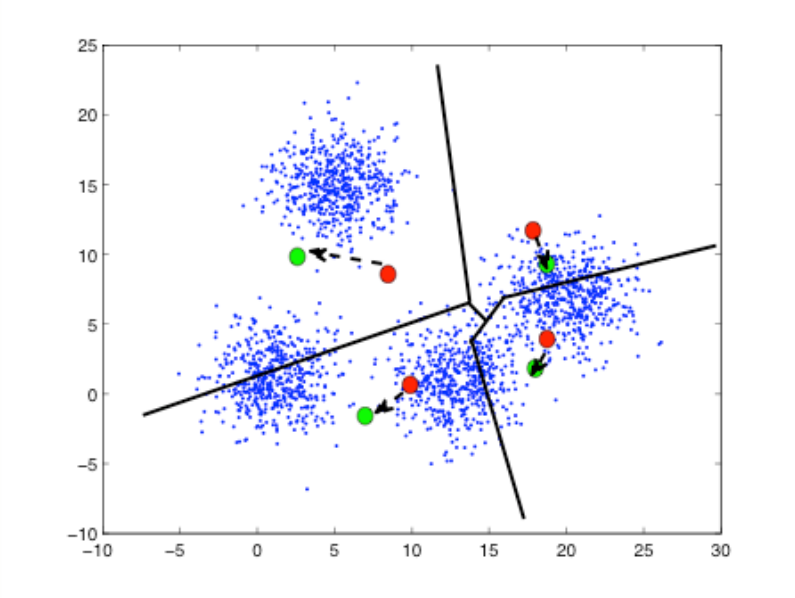
$$E = \Sigma_{i=1}^{K} \Sigma_{\mathbf{x} \in C_i} \| \mathbf{x} - \mathbf{m}_i \|^2$$

Cluster Analysis – Methods

K-means algorithm

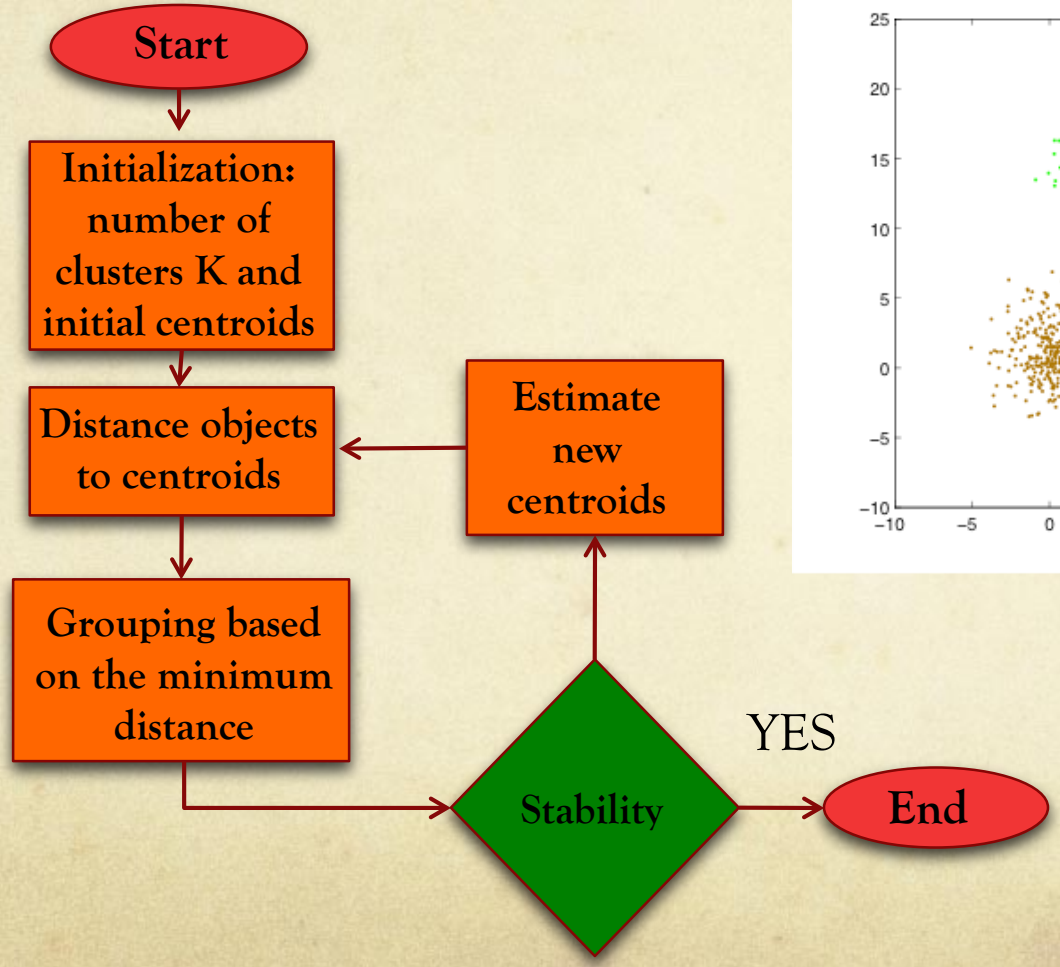


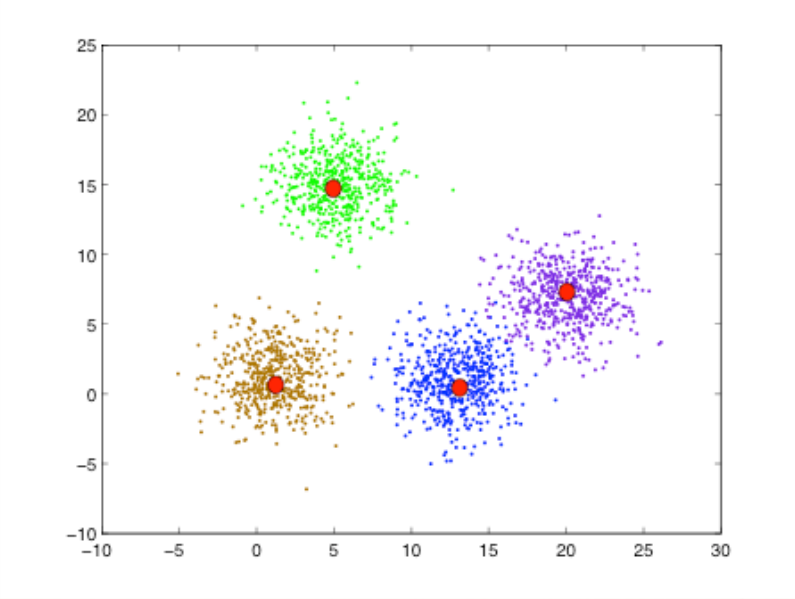
$$E = \Sigma_{i=1}^{K} \Sigma_{\mathbf{x} \in C_i} \| \mathbf{x} - \mathbf{m}_i \|^2$$



Cluster Analysis – Methods

K-means algorithm





Cluster Analysis – Methods

Drawbacks

Sensitive to initial seed points

Converge to a local optimum that may be unwanted solution

Need to specify *K*, the *number* of clusters, in advance

Unable to handle noisy data and outliers

Not suitable for discovering clusters with non-convex shapes

Applicable only when mean is defined, then what about categorical data?

Advantages

Efficient in computation

$O(tKn)$, where n is number of objects, K is number of clusters, and t is number of iterations. Normally, $K, t << n$

Cluster Analysis – Methods

Drawbacks

Sensitive to initial seed points

Converge to a local optimum that may be unwanted solution

Need to specify *K*, the *number* of clusters, in advance

Unable to handle noisy data and outliers

Not suitable for discovering clusters with non-convex shapes

Applicable only when mean is defined, then what about categorical data?

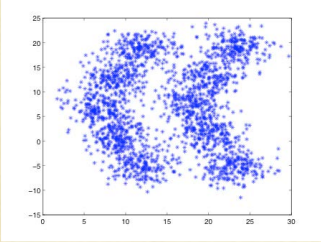
Advantages

Efficient in computation

$O(tKn)$, where n is number of objects, K is number of clusters, and t is number of iterations. Normally, $K, t << n$

Cluster Analysis – Methods

Drawbacks



Sensitive to initial seed points

Converge to a local optimum that may be unwanted solution

Need to specify *K*, the *number* of clusters, in advance

Unable to handle noisy data and outliers

Not suitable for discovering clusters with non-convex shapes

Applicable only when mean is defined, then what about categorical data?

Advantages

Efficient in computation

O(*tKn*), where *n* is number of objects, *K* is number of clusters, and *t* is number of iterations. Normally, *K*, *t* << *n*

Overview

- What is cluster analysis? ☑
- Some definitions and notations ☑
- How it works? ☑
- Cluster Analysis Diagram
  - Objectives of cluster analysis ☑
  - Design issues ☑
  - Assumptions in cluster analysis ☑
  - Clustering methods ☑
  - Interpreting the clusters
  - Validation

Cluster Analysis – Methods

Density based clustering

- Clustering based on density (local cluster criterion), such as density-connected points or based on an explicitly constructed density function

- Major features
  - Discover clusters of arbitrary shape
  - Handle noise (outliers)

DBSCAN - Ester, et al. 1996 - http://www2.cs.uh.edu/~ceick/7363/Papers/dbscan.pdf
DENCLUE - Hinneburg & D. Keim 1998 - http://www2.cs.uh.edu/~ceick/7363/Papers/dbscan.pdf

Parzen Watershed - http://www.ecmjournal.org/journal/smi/pdf/smi97-01.pdf

MeanShift - http://courses.csail.mit.edu/6.869/handouts/PAMIMeanshift.pdf

Support Vector Clustering - http://jmlr.csail.mit.edu/papers/volume2/horn01a/rev1/horn01ar1.pdf

Cluster Analysis – Methods

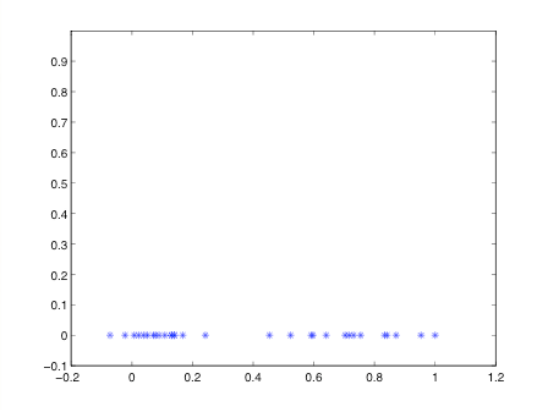
Density is the number of points within a specified space range

Cluster Analysis – Methods

Density is the number of points within a specified space range

Density estimation

An example on univariate data



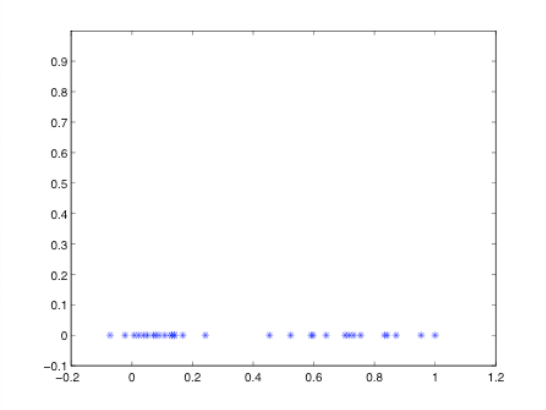
Cluster Analysis – Methods

Density is the number of points within a specified space range

Density estimation

From histograms…

An example on univariate data



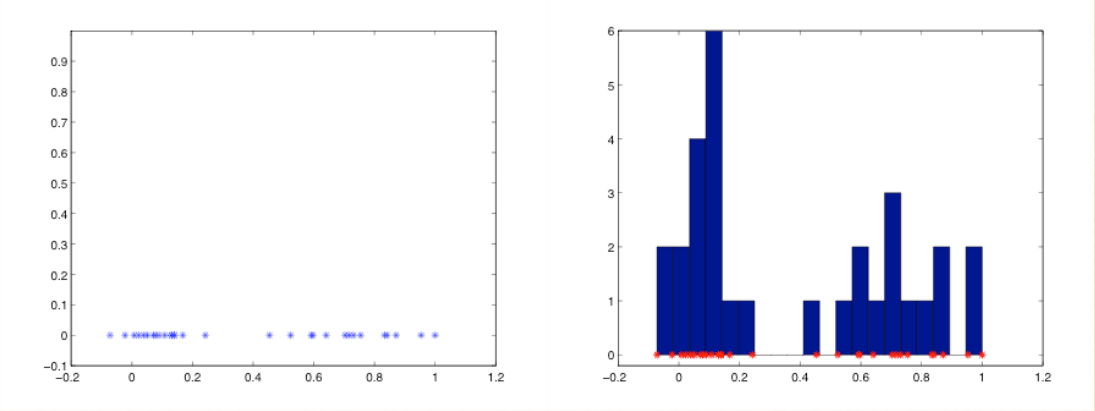
Cluster Analysis – Methods

Density is the number of points within a specified space range

Density estimation

From histograms…

An example on univariate data

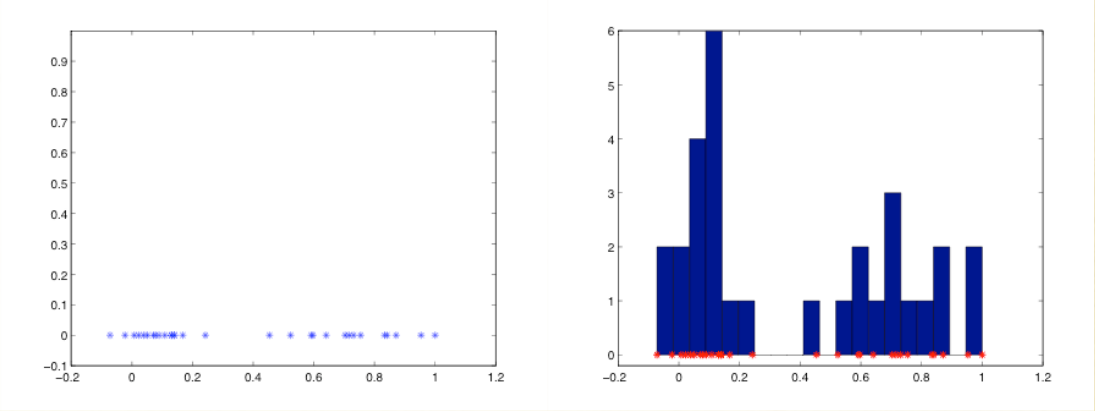


Cluster Analysis – Methods

Density is the number of points within a specified space range

Density estimation

An example on univariate data



From histograms...

...to kernel density estimation (Parzen window technique)

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$

$k(r)$ - kernel function or parzen window

Cluster Analysis – Methods

Density is the number of points within a specified space range

Density estimation

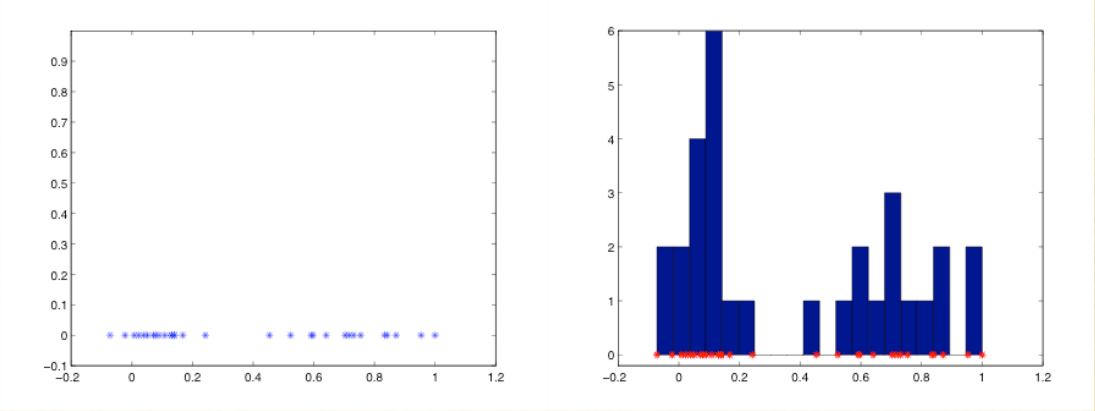
From histograms...

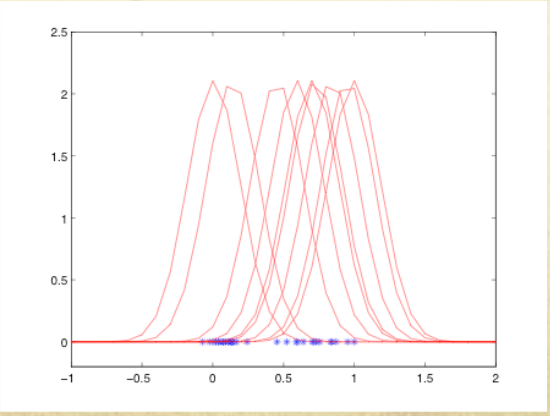
...to kernel density estimation (Parzen window technique)

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$

$k(r)$ - kernel function or parzen window

An example on univariate data





Cluster Analysis – Methods

Density is the number of points within a specified space range

Density estimation

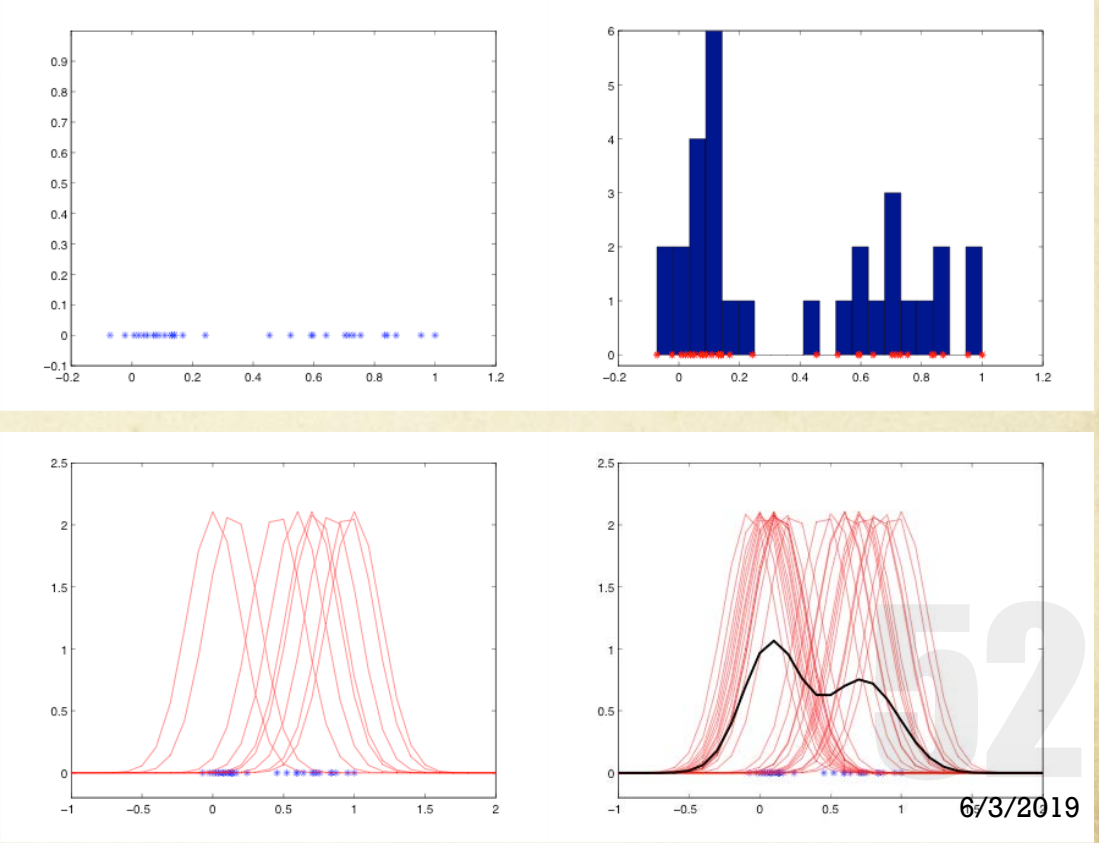
From histograms...

...to kernel density estimation (Parzen window technique)

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$

$k(r)$ - kernel function or parzen window

An example on univariate data



Cluster Analysis – Methods

Why is density estimation computational expensive in high dimensional spaces?

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$

$k(r)$ - kernel function or parzen window

Cluster Analysis – Methods

Why is density estimation computational expensive in high dimensional spaces?

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$

$k(r)$ - kernel function or parzen window

S – discrete space 1D with n = 10 discrete values

Cluster Analysis – Methods

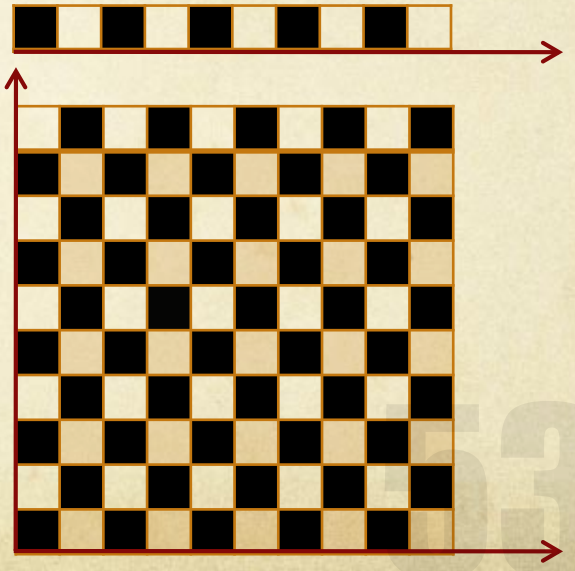
Why is density estimation computational expensive in high dimensional spaces?

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$

$k(r)$ - kernel function or parzen window

S – discrete space 1D with n = 10 discrete values

S2 – discrete space 2D with n^2 discrete values



Cluster Analysis – Methods

Why is density estimation computational expensive in high dimensional spaces?

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$

$k(r)$ - kernel function or parzen window

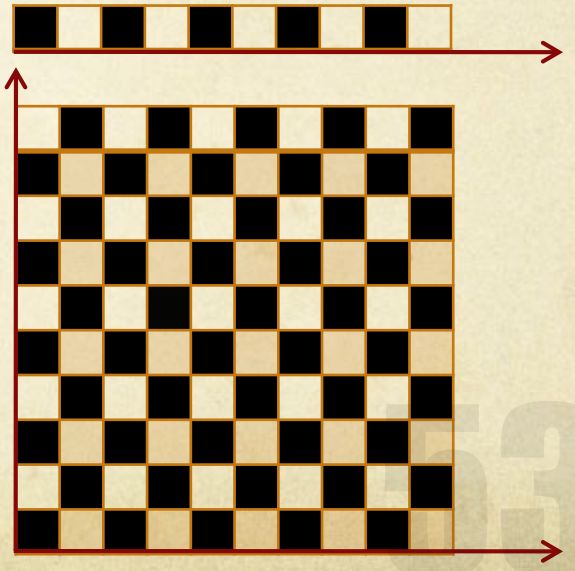
S – discrete space 1D with n = 10 discrete values

S2 – discrete space 2D with n^2 discrete values

.

.

.



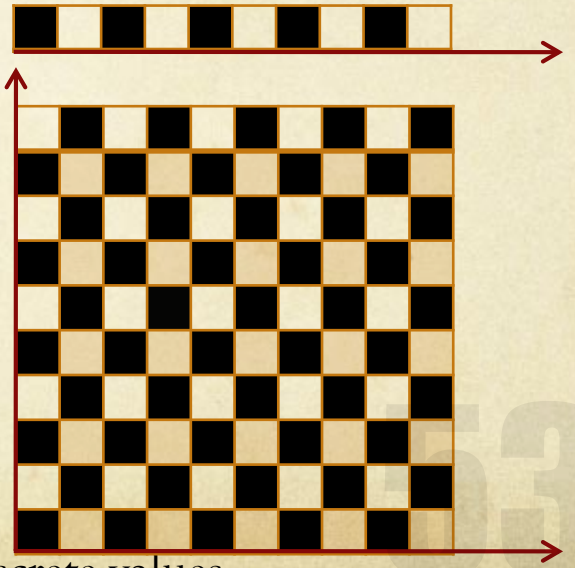
Cluster Analysis – Methods

Why is density estimation computational expensive in high dimensional spaces?

$$f(x) = \sum_i K(x - x_i) = \sum_i k\left(\frac{\|x - x_i\|^2}{h^2}\right)$$

$k(r)$ - kernel function or parzen window

S – discrete space 1D with n = 10 discrete values



S2 – discrete space 2D with n^2 discrete values

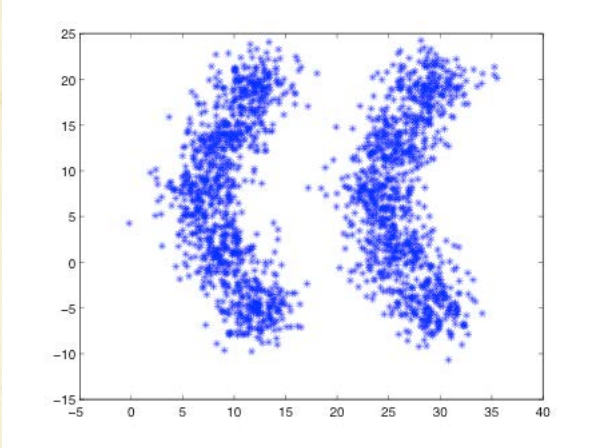
.
.
.

S10 – discrete space 2D with n^10 = 10.000.000.000 discrete values

Parzen Watershed algorithm

In based on the density estimation of the *pdf* in the feature space

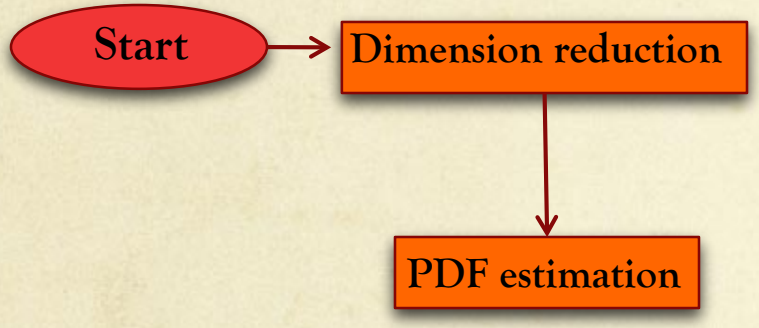
Algorithm

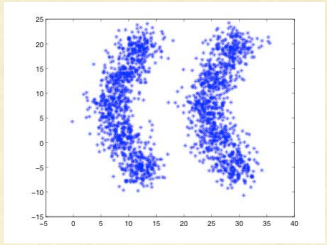


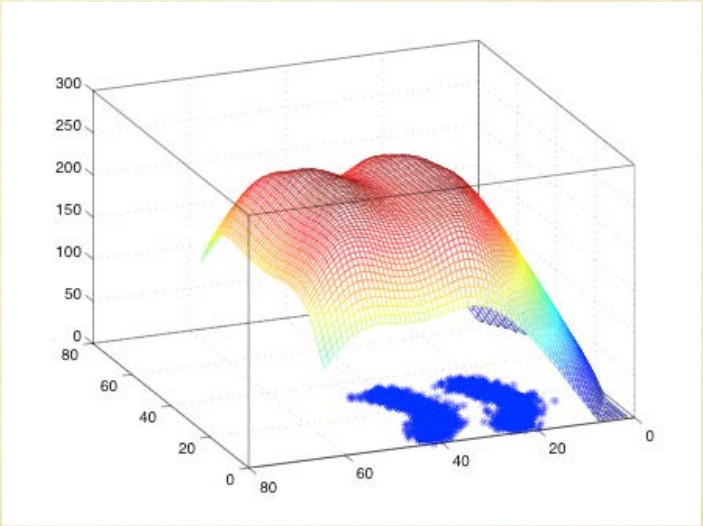
Parzen Watershed algorithm

In based on the density estimation of the *pdf* in the feature space

Algorithm



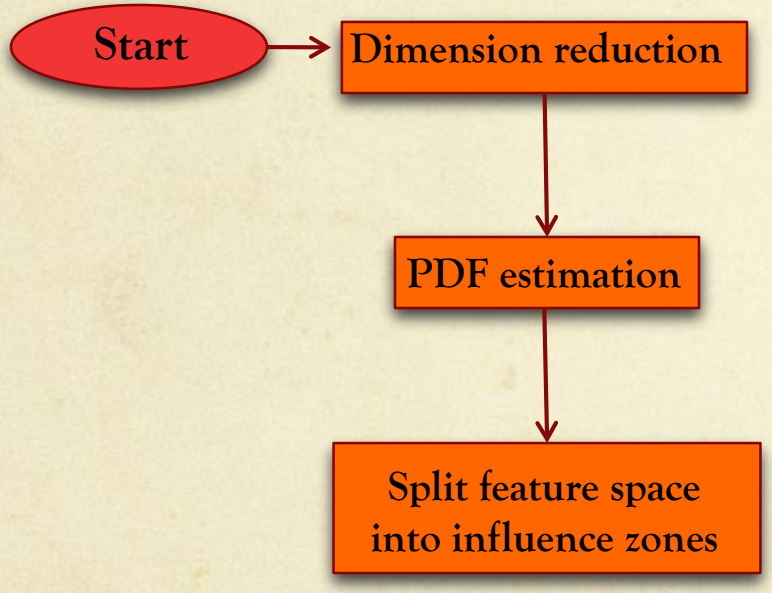


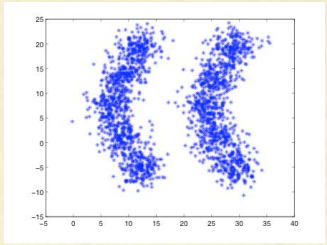


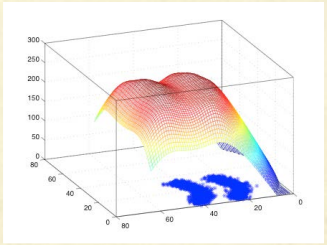
Parzen Watershed algorithm

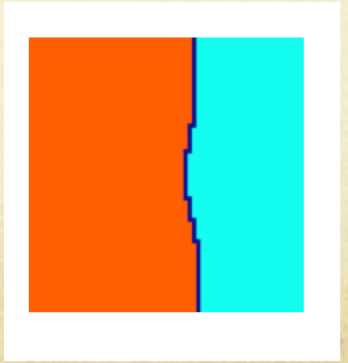
In based on the density estimation of the *pdf* in the feature space

Algorithm





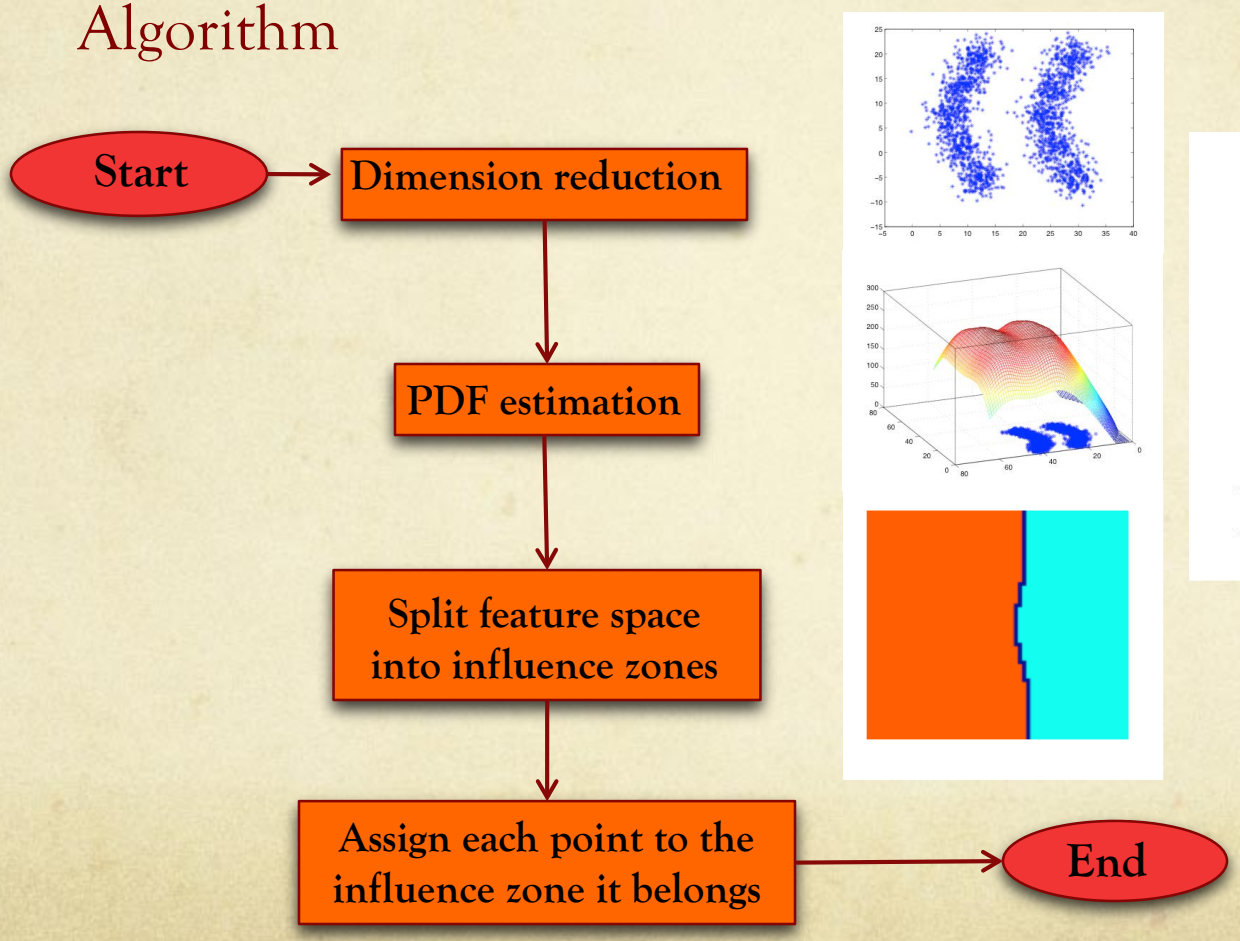


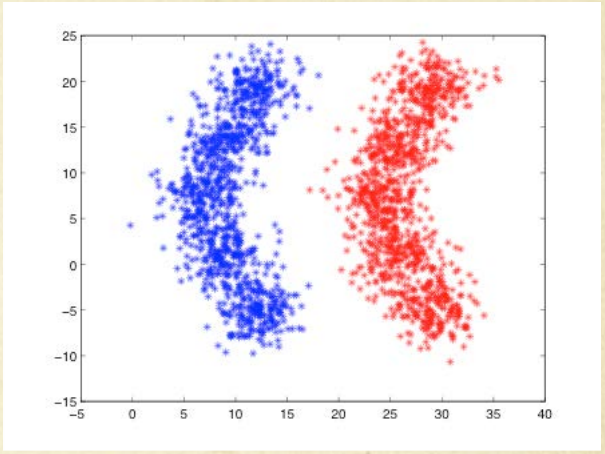


Parzen Watershed algorithm

In based on the density estimation of the *pdf* in the feature space

Algorithm





Parzen Watershed algorithm

Strengths :

- Application independent tool
- Suitable for real data analysis
- Does not assume any prior shape (e.g. elliptical) on data clusters
- Can handle arbitrary feature spaces
- Only ONE parameter to choose
- *H (window size) has a physical meaning, unlike K-Means*

Weaknesses :

- The window size (bandwidth selection) is not trivial
- Inappropriate window size can cause modes to be merged, or generate additional "shallow" modes -> Use adaptive window size
- Low dimension feature space
- Computational complexity high

Cluster Analysis – Interpreting the clusters

Stage 5: Interpreting the clusters

The cluster centroid (a mean profile of the cluster on each cluster variable) is particularly useful in the interpretation stage

Interpretation involves:

- Examining and distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters
- Cluster solution failing to reveal significant differences indicate that other solutions should be examined
- The cluster centroid should also be assessed for correspondence to researcher's prior expectation based on theory or practical experience

Overview

- What is cluster analysis? ☑
- Some definitions and notations ☑
- How it works? ☑
- Cluster Analysis Diagram
  - Objectives of cluster analysis ☑
  - Design issues ☑
  - Assumptions in cluster analysis ☑
  - Clustering methods ☑
  - Interpreting the clusters ☑
  - Validation

Cluster Analysis – Validation

Stage 6: Validating and Profiling the Clusters

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.
Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without reference to external information.- Use only the data*
4. *Comparing the results of two different sets of cluster analyses to determine the stability of the solution.*
5. *Determining the 'correct' number of clusters.*

Cluster analysis – Validation

Indices for cluster validation

- Cross validation
- External index – used to measure the extent to which cluster labels match externally supplied class labels
  - Labels provided by experts or ground truth
- Internal index – based on the intrinsic content of the data. Used to measure the goodness of a clustering structure *without respect to external information*
  - Davies Bound – index , Dunn – index, C – index, Silhouette coefficient etc.
- Relative index – used to compare the results of different clustering algorithms
  - Internal or external indices

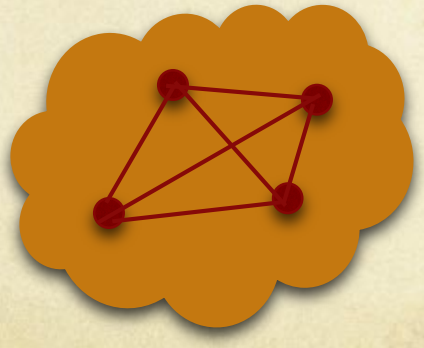
Cluster analysis – Validation

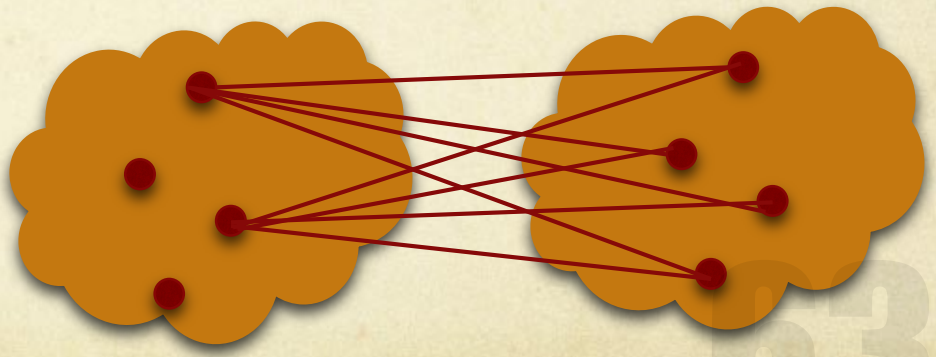
Internal indices – example: silhouette coefficient

$$sc = 1 - \frac{c}{s}$$

Cluster cohesion is the mean value of the distances of all pairs of points within a cluster

Cluster separation is the mean value of the distances between the points in the cluster and points outside the cluster





c – the smallest the better

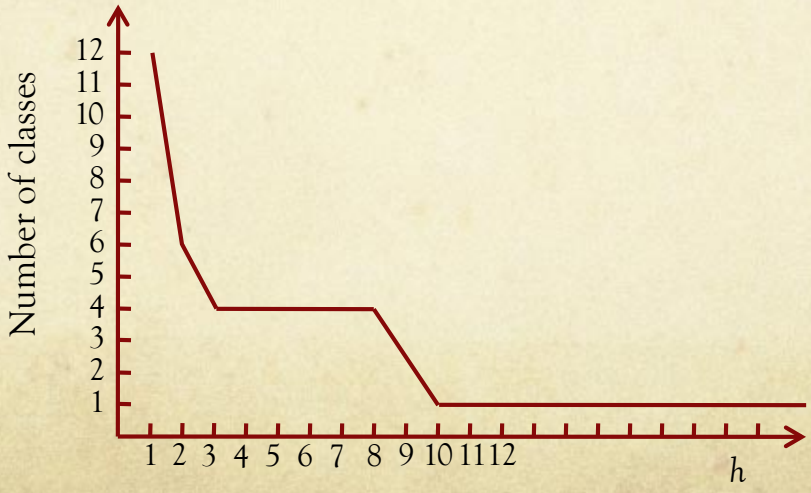
s – biggest the better

Cluster validation

- K – means, hierarchical
  - Davies Bound – index , Dunn – index, C – index, Silhouette coefficient etc.
- Density based clustering
  - Stability of the number of classes

$No\_Of\_Classes = f(h)$

h - window size



Overview

- What is cluster analysis? ☑
- Some definitions and notations ☑
- How it works? ☑
- Cluster Analysis Diagram
  - Objectives of cluster analysis ☑
  - Design issues ☑
  - Assumptions in cluster analysis ☑
  - Clustering methods ☑
  - Interpreting the clusters ☑
  - Validation ☑

Area of applications