

# Modulo 7: Clustering

*Unsupervised Machine Learning*

*Ing. Marco Pirrone Ph.D.*

*Ing. Massimo Romano Ph.D.*



DIPARTIMENTO DI METODI E MODELLI PER  
L'ECONOMIA IL TERRITORIO E LA FINANZA  
MEMOTEF



SAPIENZA  
UNIVERSITÀ DI ROMA



Learning  
Session

# *Clustering*

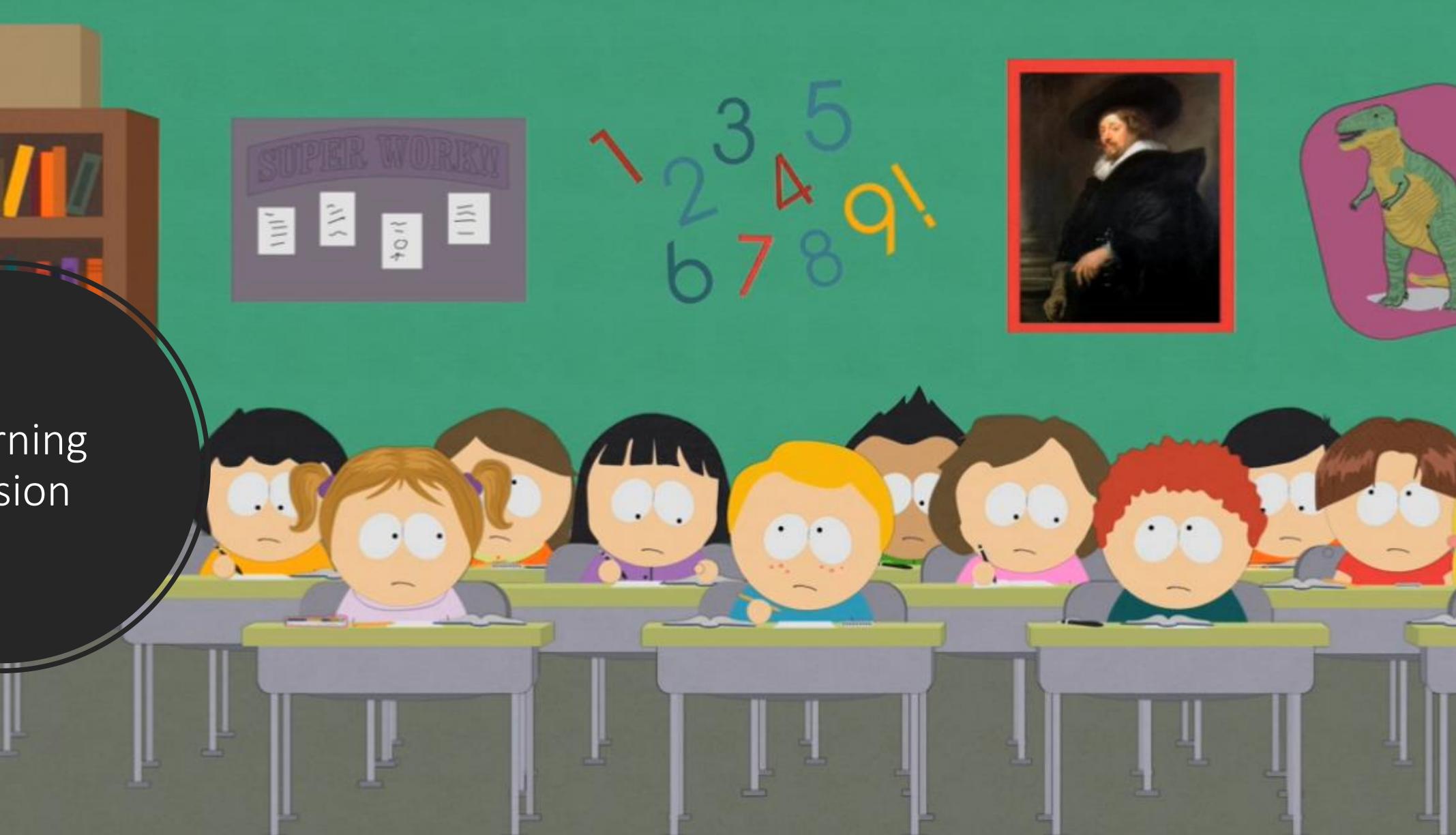
## *Unsupervised Machine Learning*



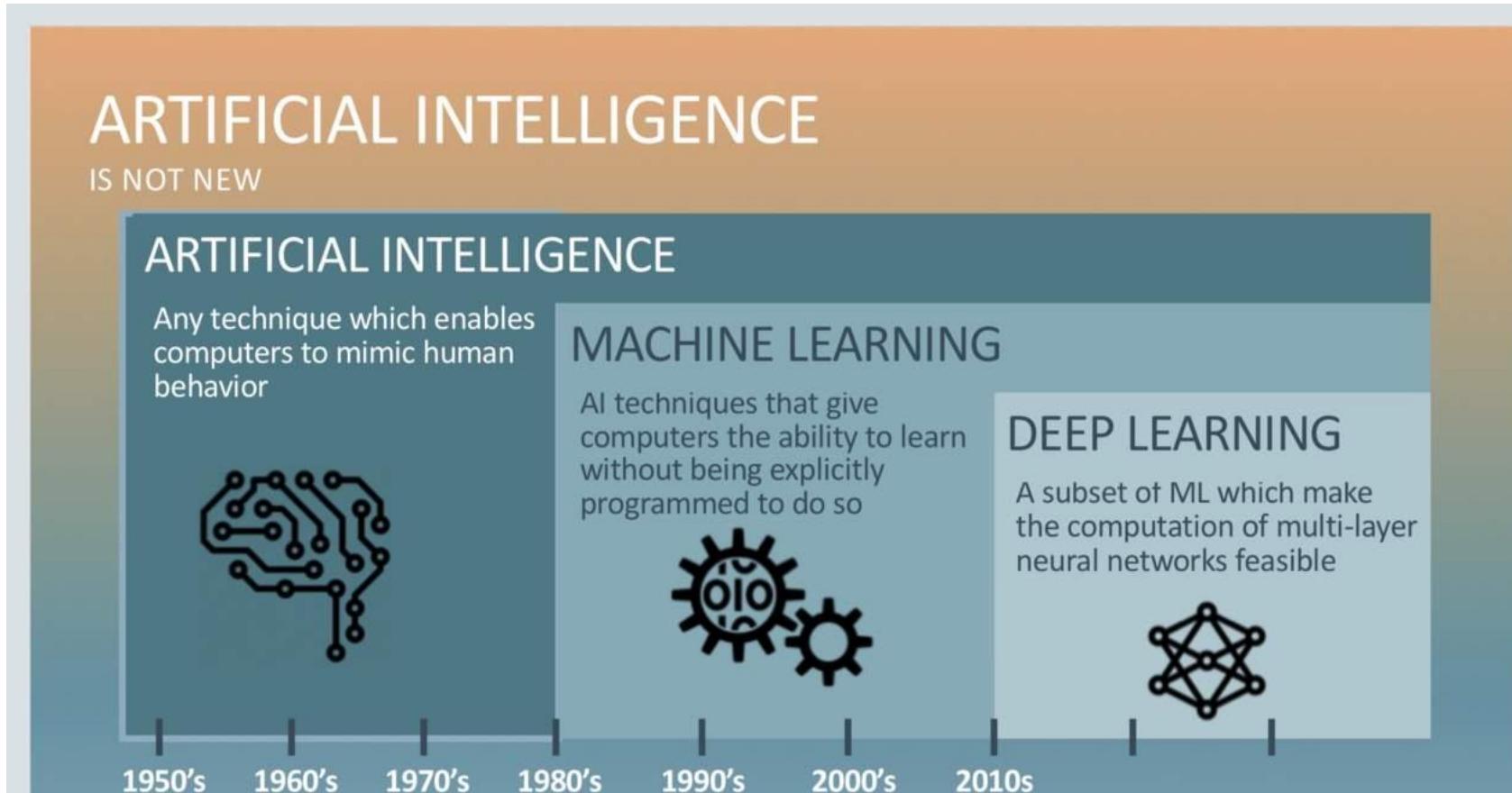
LAB Session



# Learning Session



# Machine Learning



# Machine Learning

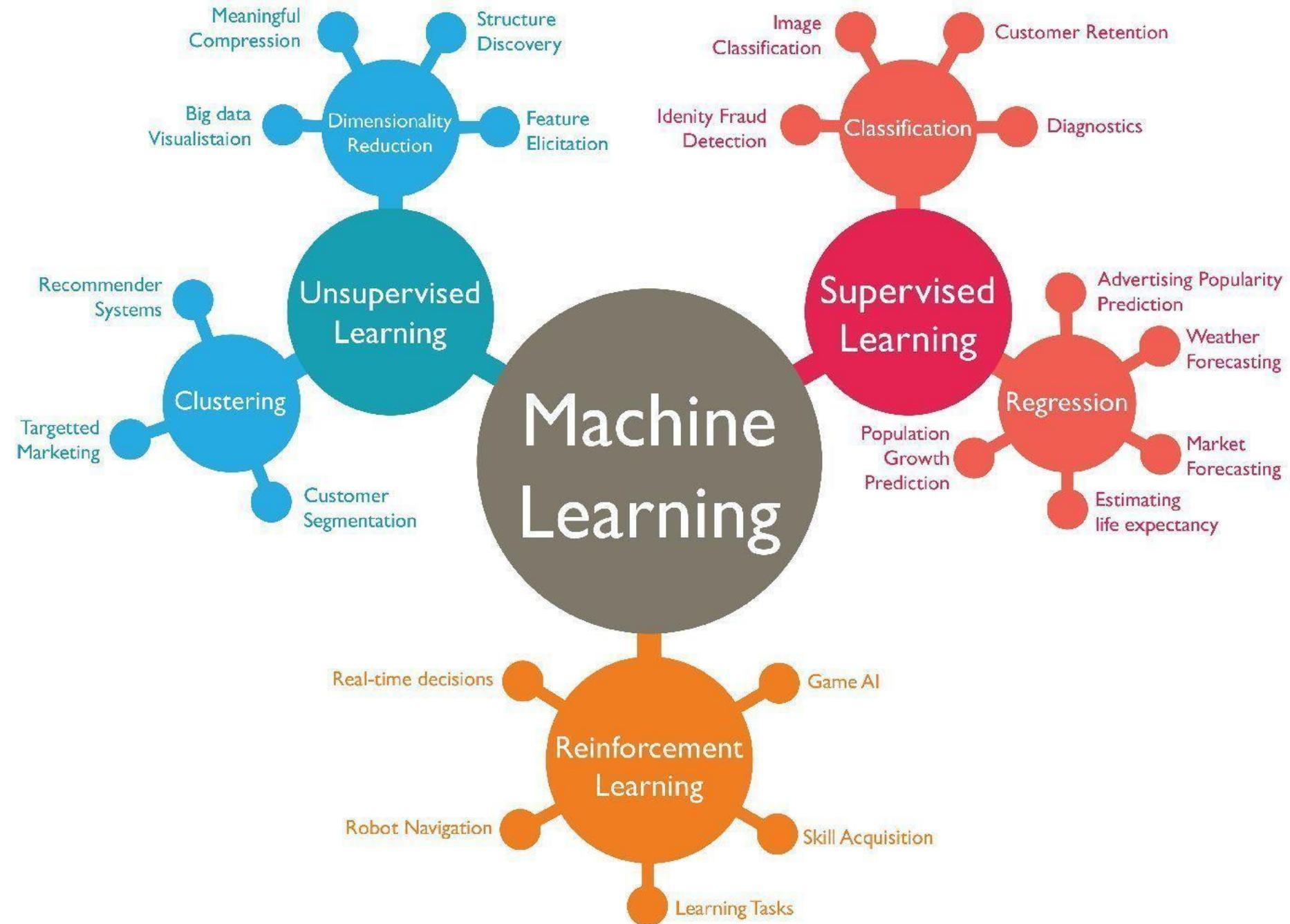
Si occupa della realizzazione di sistemi che si basano su **osservazioni, esempi ed esperienza** per la sintesi di **nuova conoscenza**

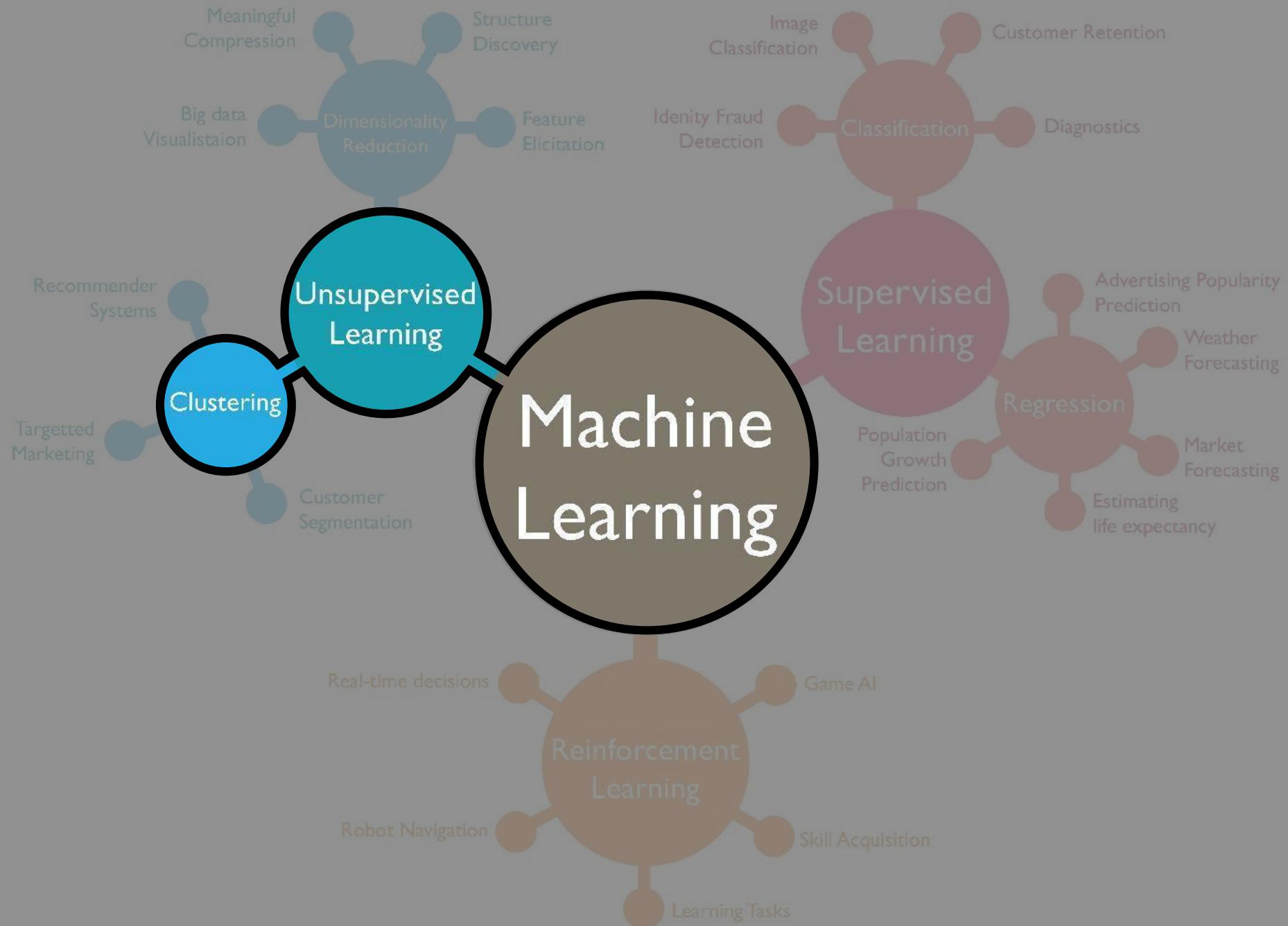
# Machine Learning – utile quando..

- Difficoltà di formalizzazione (es. riconoscere un amico in una foto)
- Elevato numero di variabili in gioco (es. meteo)
- Mancanza di teoria (es. andamento mercati finanziari)
- Personalizzazione (documenti interessanti o meno)

# Supervised learning vs. unsupervised learning

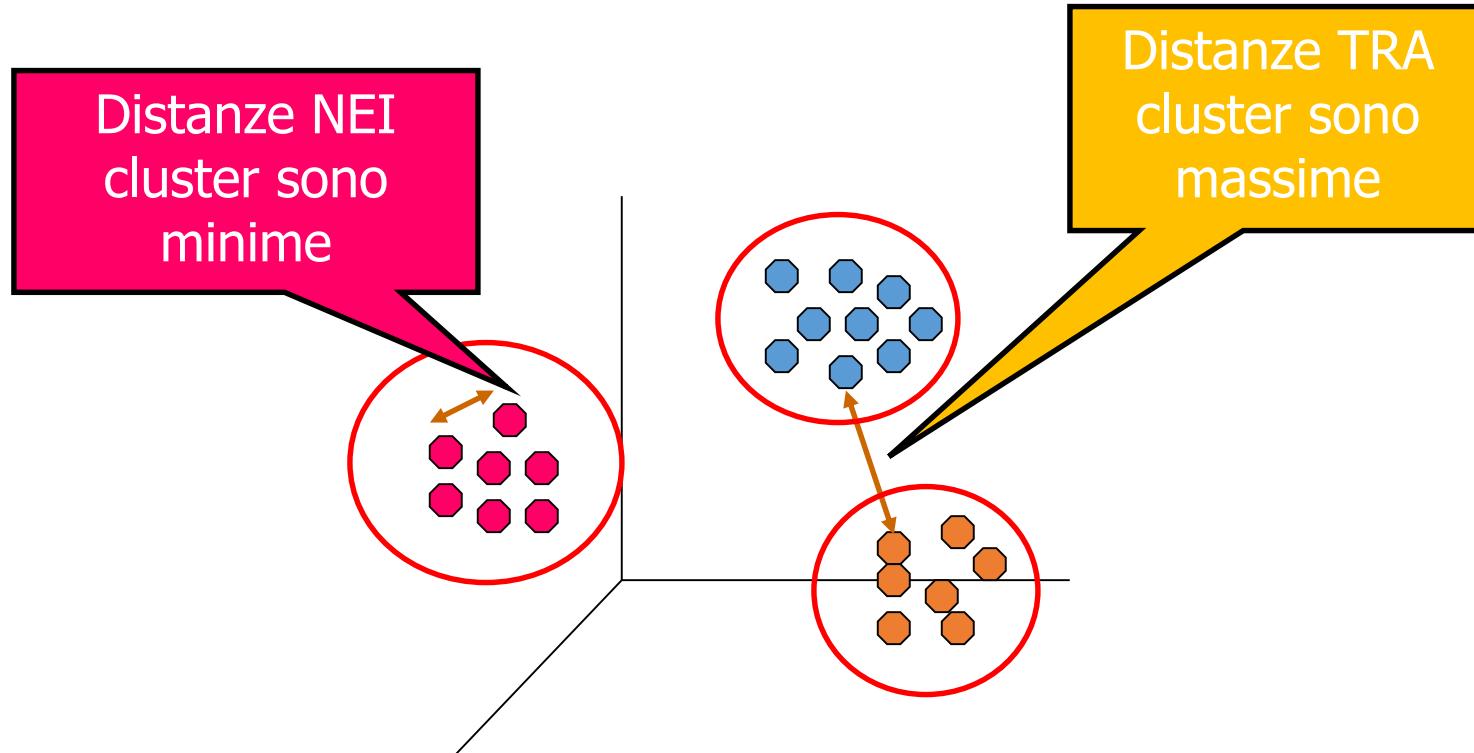
- **Supervised learning:** L'apprendimento supervisionato è una tecnica di apprendimento automatico che mira a istruire un sistema informatico in modo da consentirgli di elaborare automaticamente previsioni sui valori di uscita di un sistema rispetto ad un input sulla base di una serie di esempi ideali, **costituiti da coppie di input e di output**, che gli vengono inizialmente forniti
- **Unsupervised learning:** Le tecniche di apprendimento non supervisionato mirano ad estrarre, in modo automatico, della conoscenza a partire da basi di dati. Questo avviene **senza una specifica conoscenza dei contenuti da analizzare**
- **Reinforcement learning:** Questa tecnica di programmazione si basa sul presupposto di potere **ricevere degli stimoli dall'esterno a seconda delle scelte dell'algoritmo**. Quindi una scelta corretta comporterà un premio mentre una scelta scorretta porterà ad una penalizzazione del sistema





# Cos'è Cluster Analysis?

- Trovare gruppi di unità statistiche tali che le unità di un gruppo siano simili (o correlate) tra loro e diverse da (o estranee a) le unità in altri gruppi



# Applicazioni della Cluster Analysis

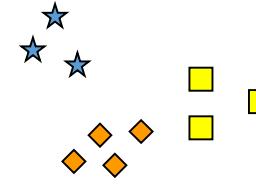
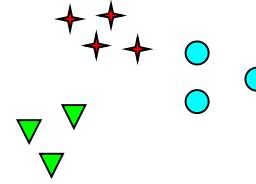
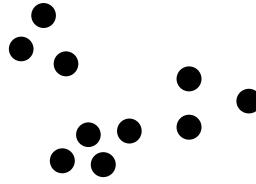
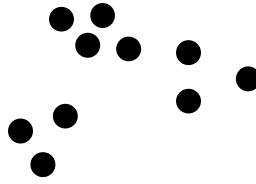
- **Inferenza**

- Cluster di documenti da ricerche web
- Gruppi di geni e proteine che hanno funzioni simili
- Gruppi di azioni con fluttuazioni di prezzo simili

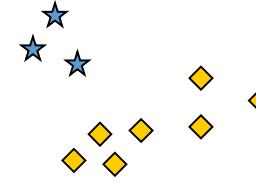
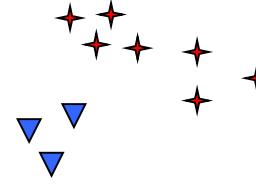
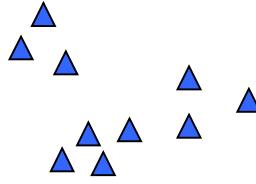
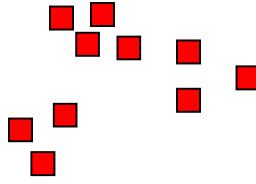
- **Sintesi**

- Ridurre la dimensione di dataset eccessivamente grandi

# Quanti cluster?



Sei Cluster



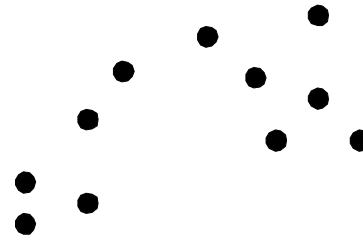
Due Cluster

Quattro Cluster

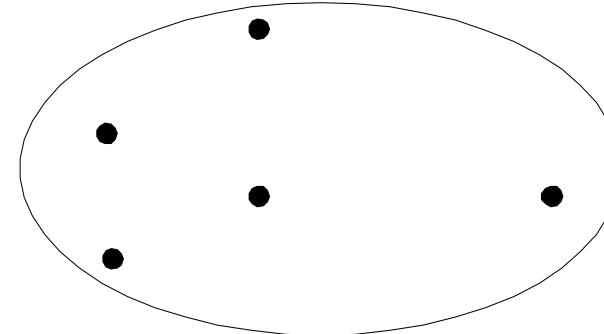
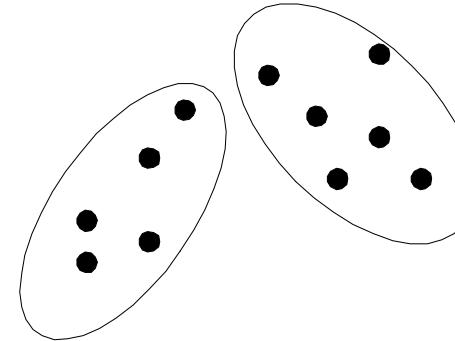
# Clustering

- Si definisce **clustering** un insieme di cluster
- Clustering partizionale
  - Una divisione in cluster che non si sovrappongono
- Clustering gerarchico
  - Un insieme di cluster innestati organizzati come un albero gerarchico
- Cluster basato sulla densità

# Clustering Partizionale

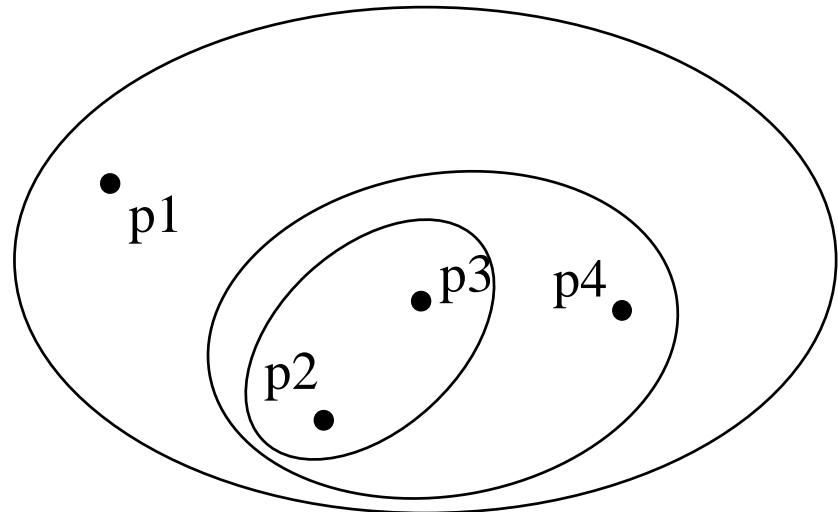


Punti originari

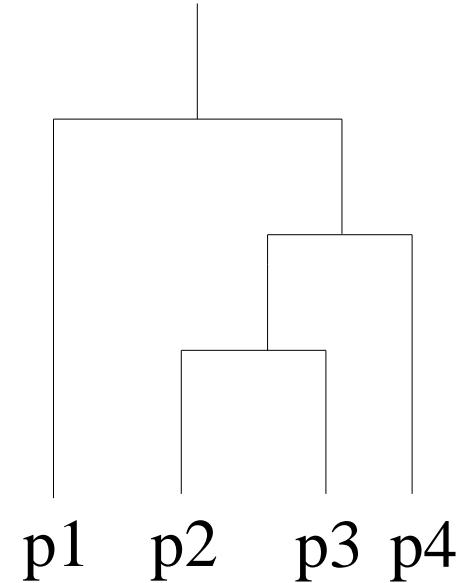


Clustering partizionale

# Clustering Gerarchico



Clustering gerachico



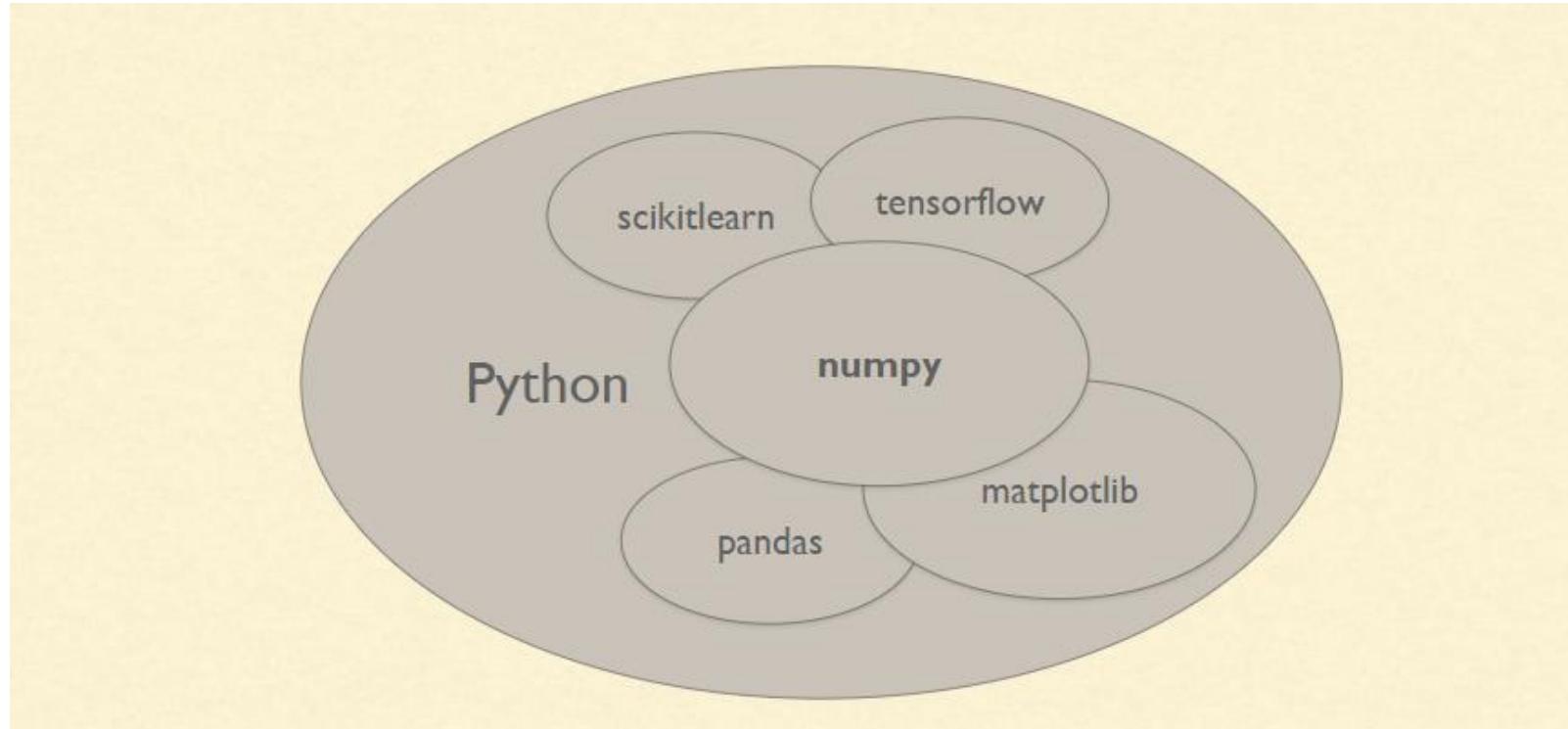
Dendrogramma

# LAB Session



# LAB Session – Framework

**Prerequisiti:** Python, Jupyter.



Jupyter Notebook 5.7.4  
Pandas 0.23.4  
NumPy 1.14.2  
Matplotlib 2.2.3  
Seaborn 0.9.0  
SciPy 1.1.0  
Scikit-Learn 0.20.2

# LAB Session – NumPy & Pandas

- **NumPy** (Numeric Python o Numeric Python) è una estensione open source per Python che fornisce supporto per il trattamento di array e matrici multidimensionali di grandi dimensioni. *Array e matrici sono una parte essenziale dell'ecosistema Machine Learning.*
- **Pandas** (PANel + DAta) è una libreria basata su **NumPy** per la manipolazione dei dati che possono assumere il formato di tabelle numeriche o serie temporali: **DataFrame** e **Series**.

# LAB Session – Matplotlib & Seaborn

- **Matplotlib** è una libreria Python per la manipolazione e stampa di grafici 2D. Un modulo chiamato **pyplot** che semplifica la definizione di stili, le proprietà dei caratteri, gli assi di formattazione, ecc.
- **Seaborn** è una libreria per la visualizzazione dei dati basata su **Matplotlib**, fornendo un'interfaccia di più alto livello. E' progettato per funzionare molto bene con gli oggetti **Pandas** (DataFrame).

# LAB Session – Codice Sorgente

Disponibile su Git Hub:

<https://github.com/gotamo/master-and-skills>

 gotamo	new version	Latest commit ca9dc6a 18 hours ago
..		
 exercise	first	2 months ago
 lecture	new version	18 hours ago
 samples	first	2 months ago

# Learning Session



# Algoritmi di Clustering

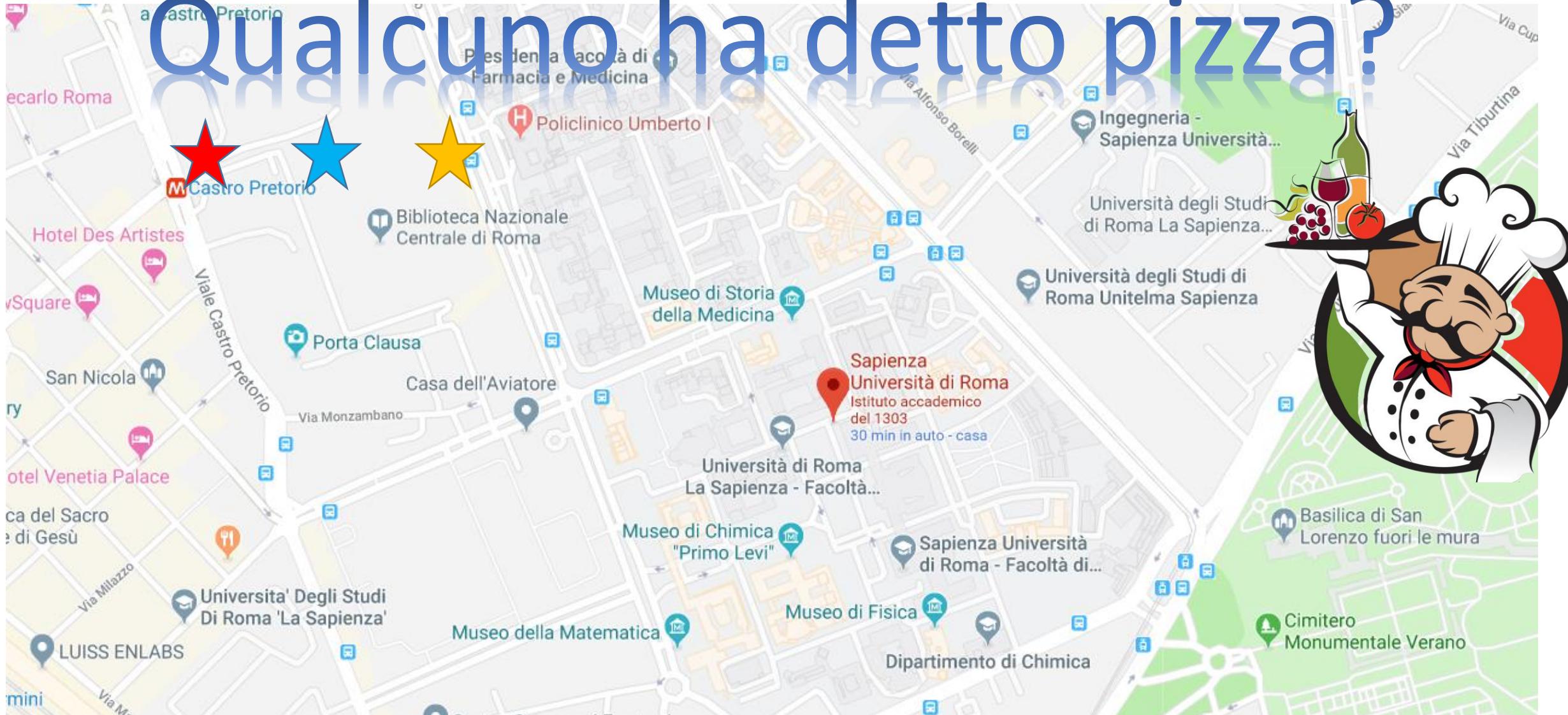
- Cluster Partizionale: K-means e le sue varianti
- Clustering gerarchico
- Clustering basato sulla densità
- Misure di validità dei Cluster

# Algoritmi di Clustering

- K-means e le sue varianti (partizionale)
- Clustering gerarchico
- Clustering basato sulla densità
- Misure di validità dei Cluster

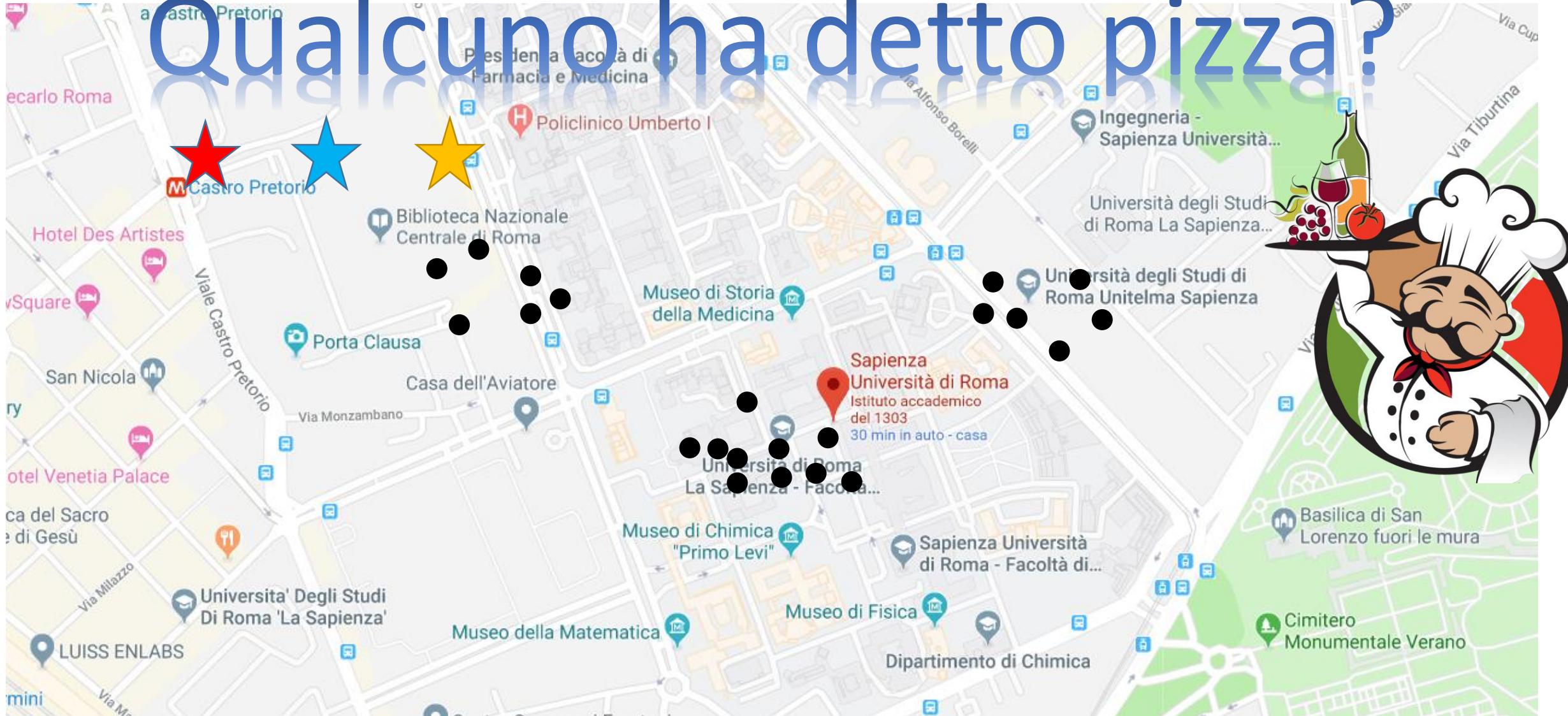
K-means nella vita reale

# Qualcuno ha detto pizza?



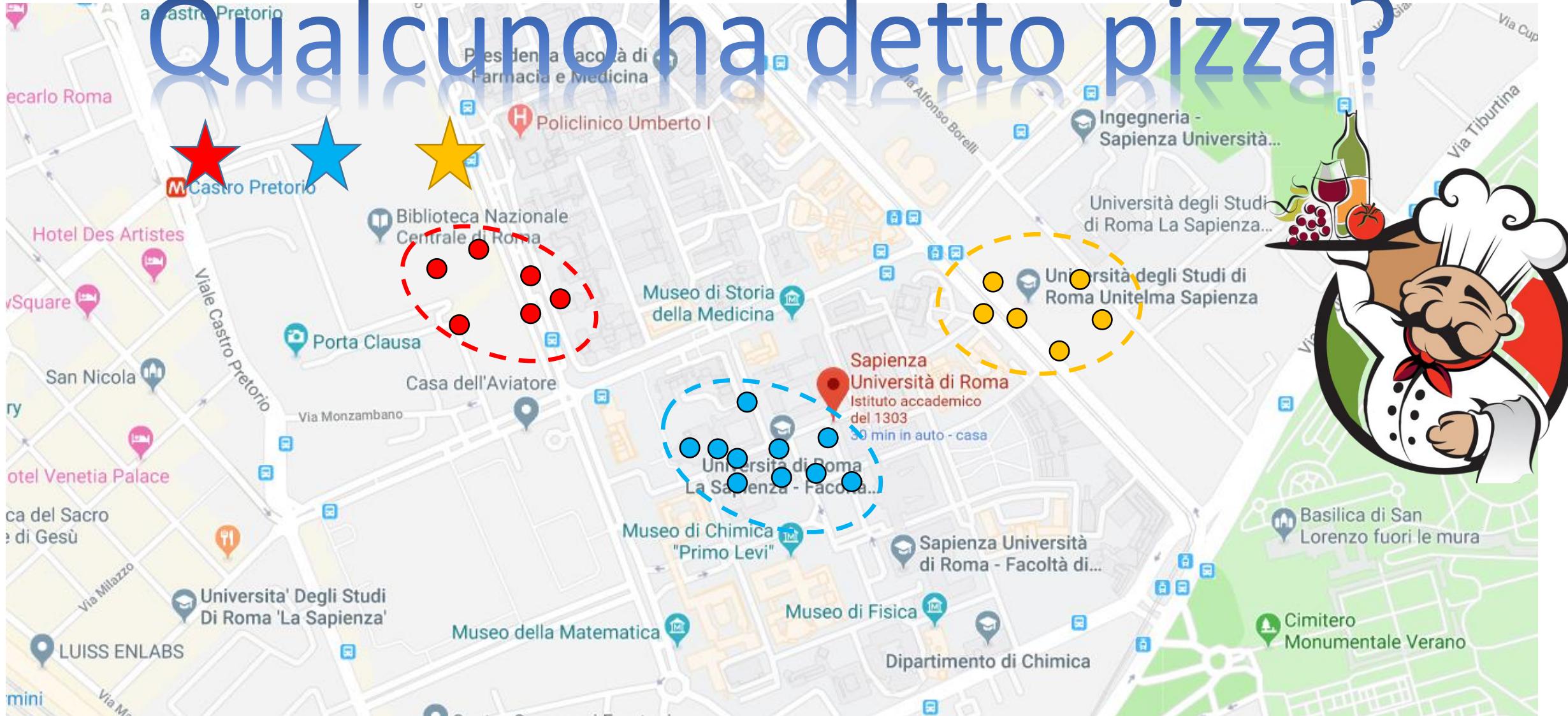
K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



K-means nella vita reale

# Qualcuno ha detto pizza?



# K-means Clustering (Partizionale)

- Ogni cluster è associato ad un centroide (baricentro)
- Il numero di cluster, K, è specificato inizialmente

## Algoritmo

Seleziona in modo casuale **K punti** che rappresentano i centroidi iniziali

Ripeti

- Assegna ciascun oggetto al centroide più vicino
- Ricalcola i centroidi dei cluster trovati

fino a quando gli assegnamenti non cambiano

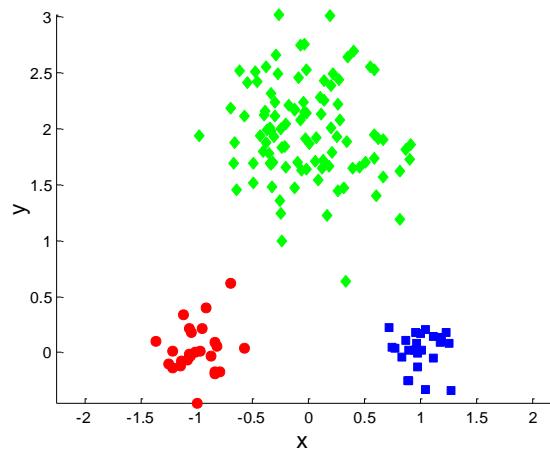
# K-means Clustering – Dettagli

- I **centroidi iniziali** sono spesso scelti **in maniera casuale**.
- Il centroide è (tipicamente) la **media dei punti nel gruppo**.
- La ‘**vicinanza**’ è misurata in molti modi (es.usando la distanza **Euclidea, correlazione**, etc.)
- Spesso la condizione di arresto viene modificata in '**Fino a quando relativamente pochi punti cambiano cluster**'

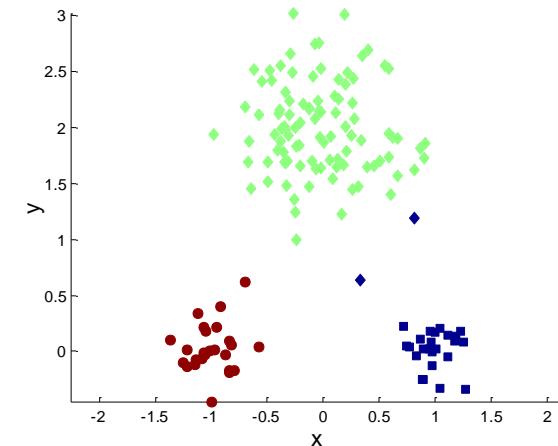
# K-means Clustering – Pro e Cons

- **Forze**
  - Relativamente efficiente
    - La convergenza avviene di solito dopo poche operazioni
    - La complessità è  $O( n * K * I * d )$   
 $n$  = numero di unità,  $K$  = numero di cluster,  
 $I$  = numero di iterazioni,  $d$  = numero di attributi
  - Applicabile e convergente per le più comuni misure di prossimità
- **Debolezze**
  - Può essere applicato solo quando il tipo di dato permette di definire la **media** (o baricentro)
  - Bisogna specificare in **anticipo**  $K$
  - Molto sensibile alla **scelta iniziale dei centroidi**
  - Molto sensibile alle **caratteristiche del cluster reali**
  - Molto sensibile alla presenza di **outlier (anomalie)**

# Importanza della scelta dei centroidi iniziali

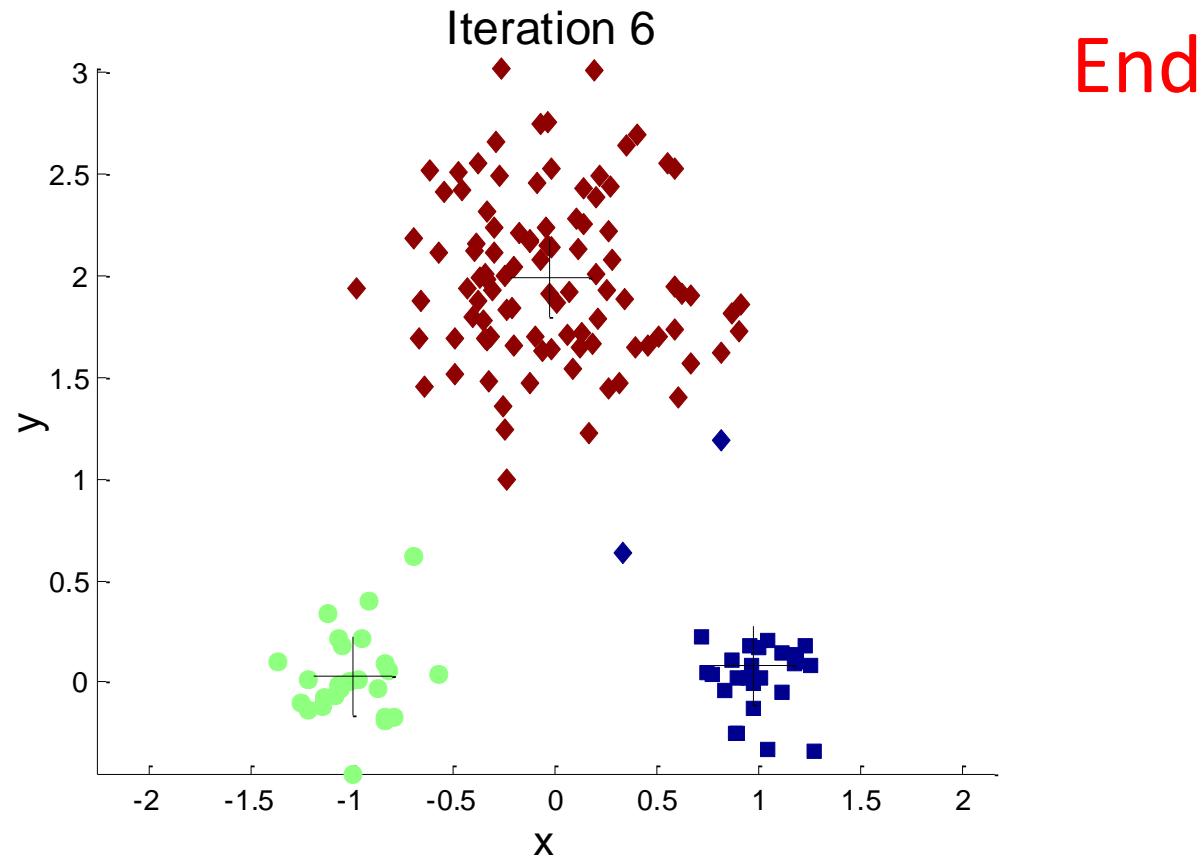


Gruppi reali

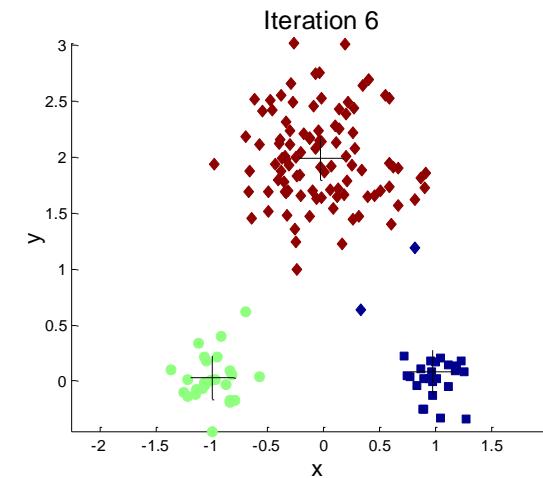
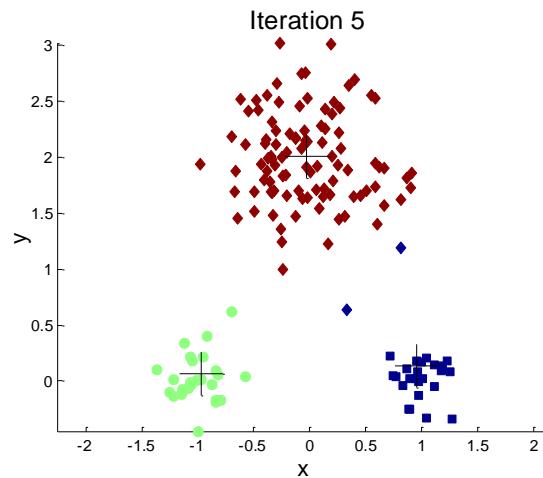
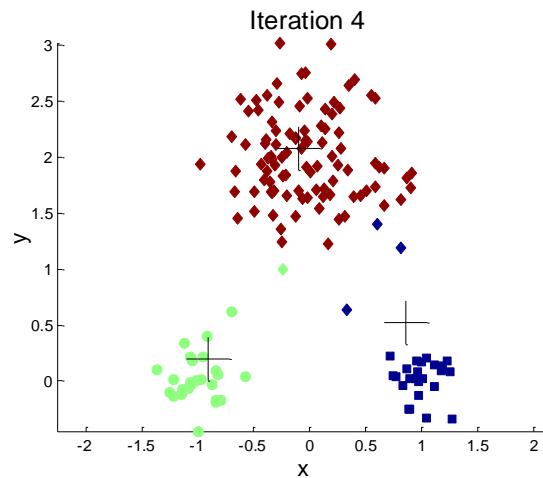
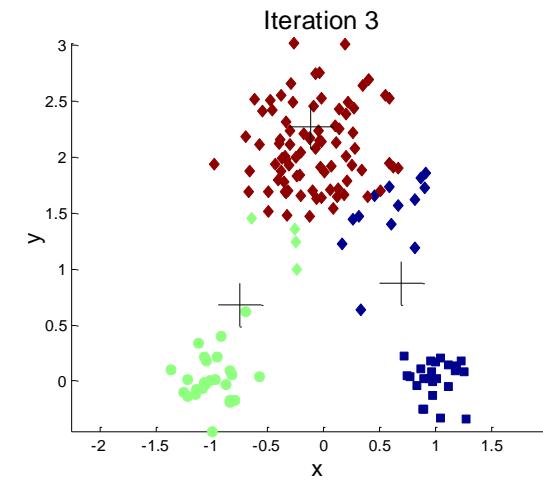
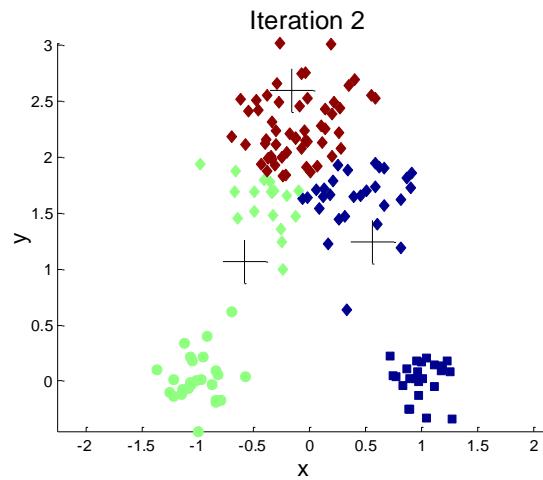
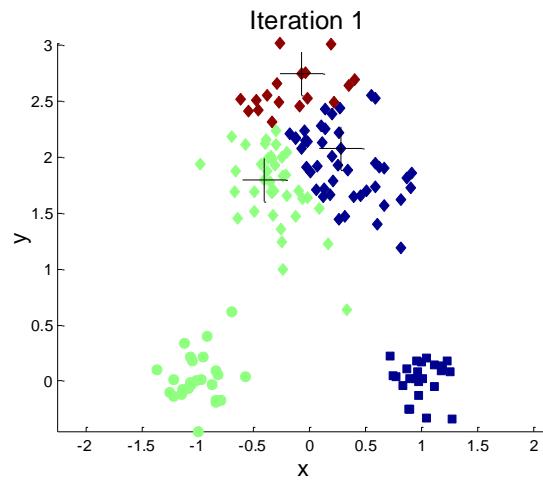


Clustering ottimale

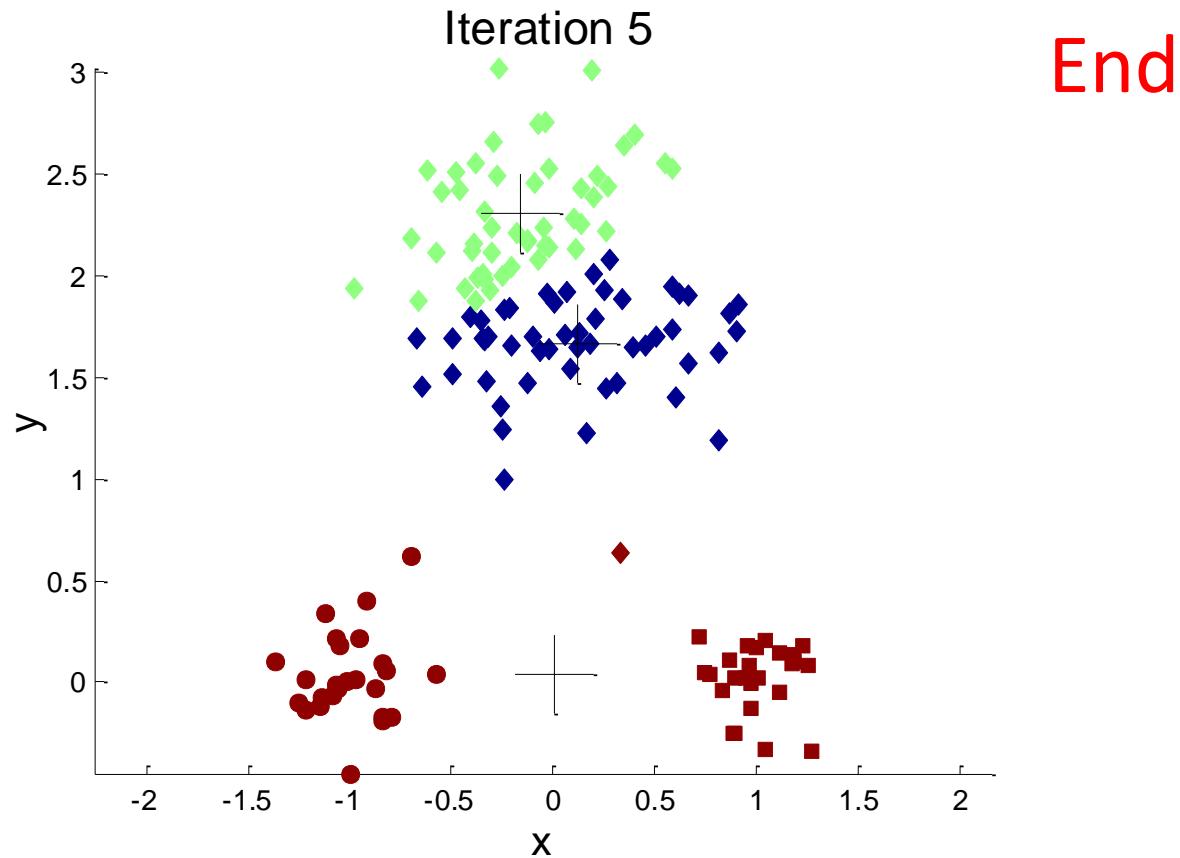
# Importanza della scelta dei centroidi iniziali



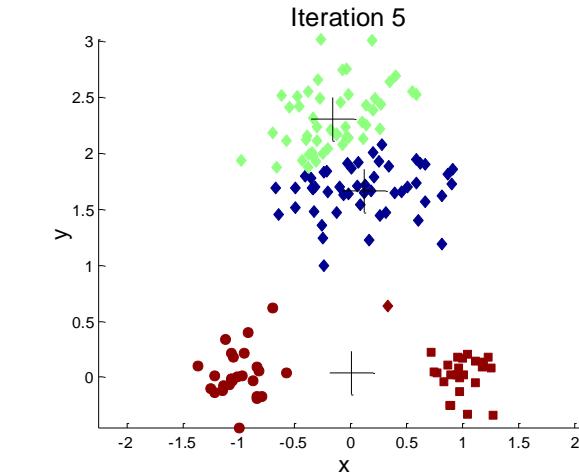
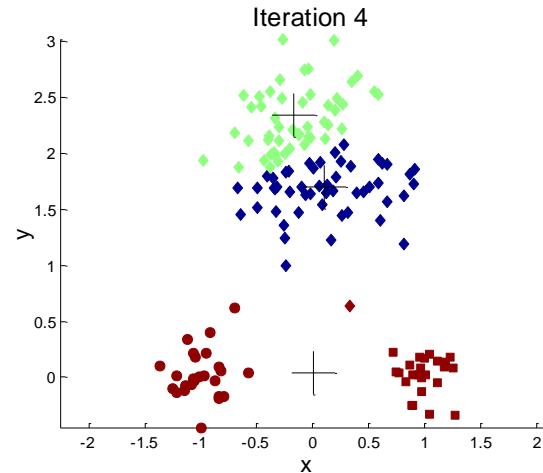
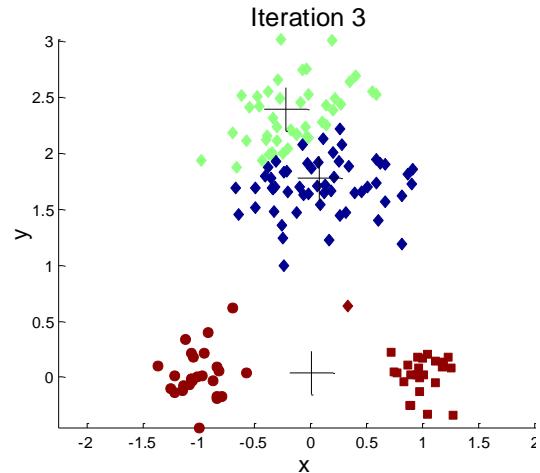
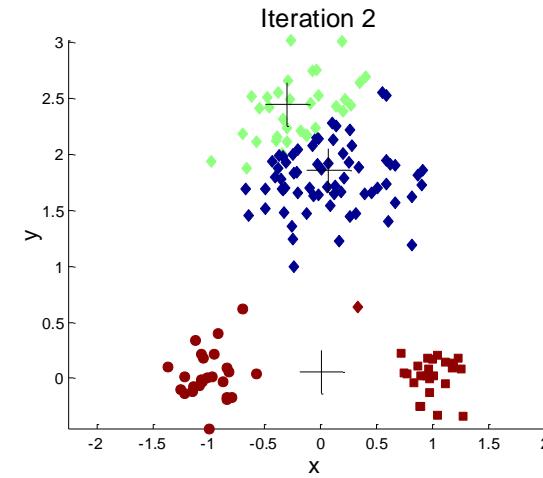
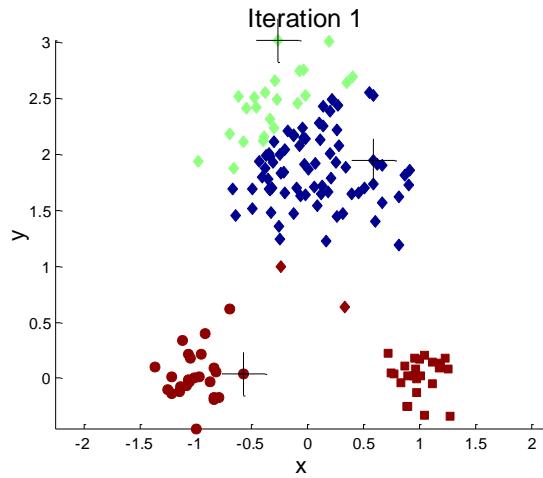
# Importanza della scelta dei centroidi iniziali



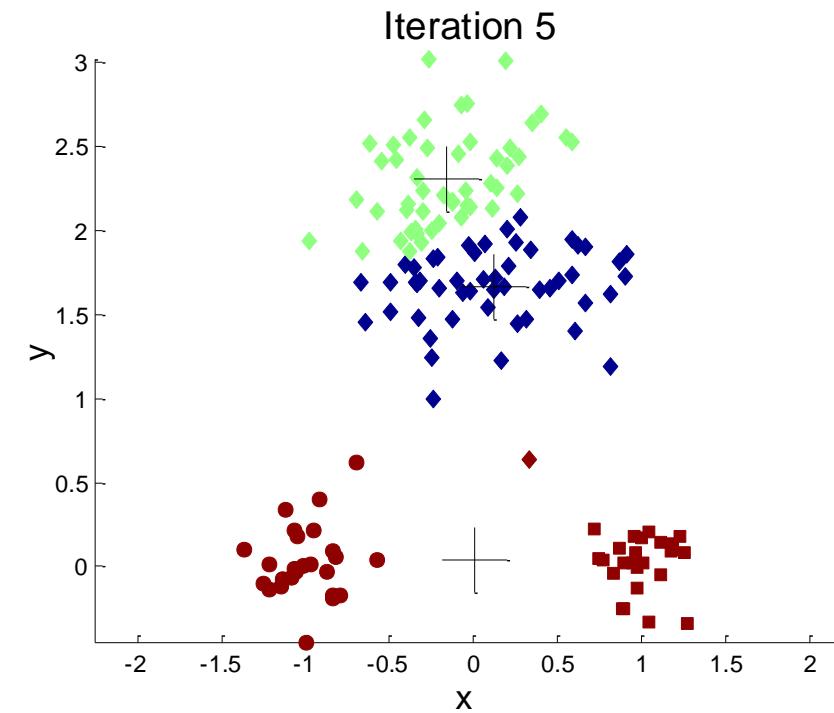
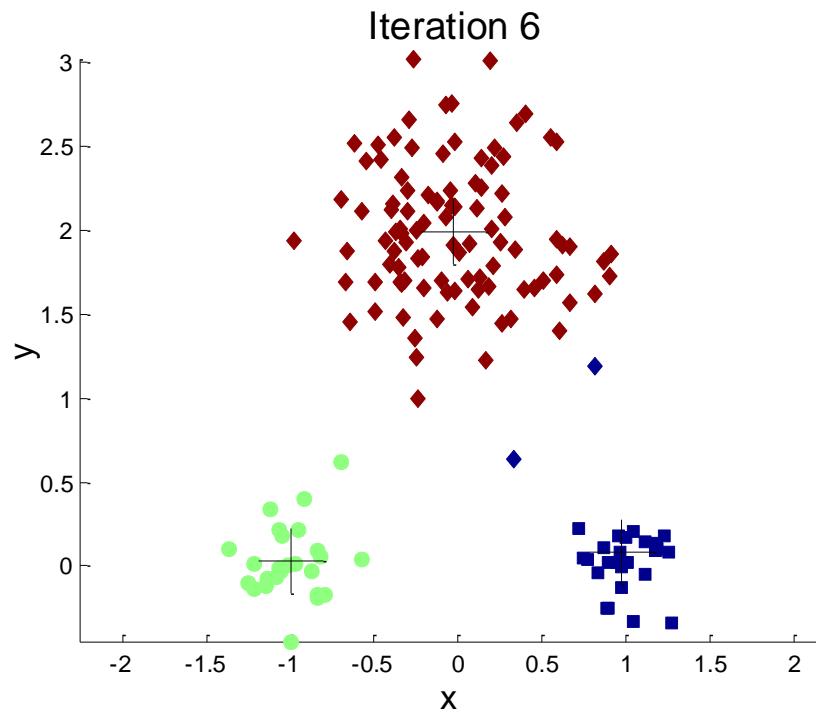
# Importanza della scelta dei centroidi iniziali



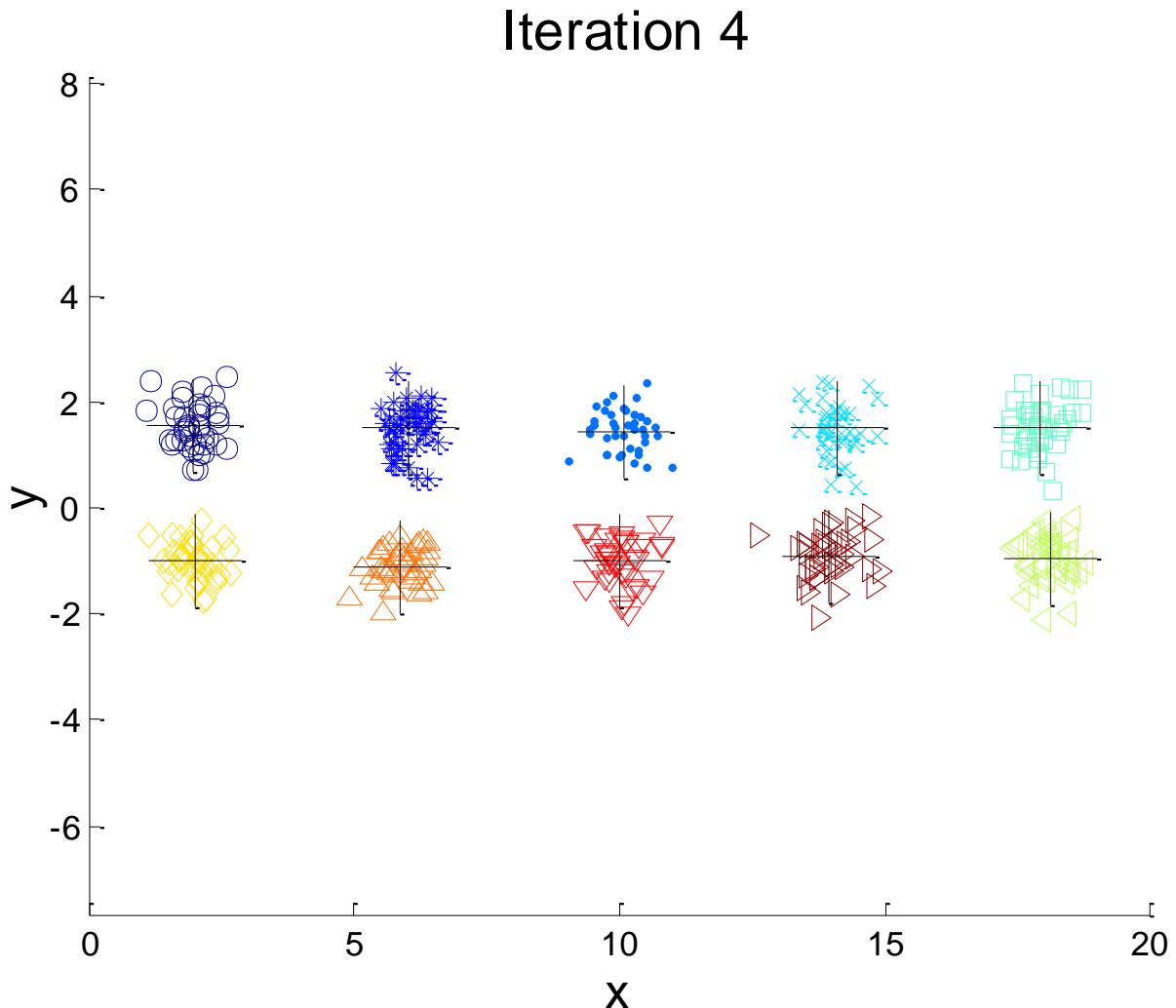
# Importanza della scelta dei centroidi iniziali



# Importanza della scelta dei centroidi iniziali

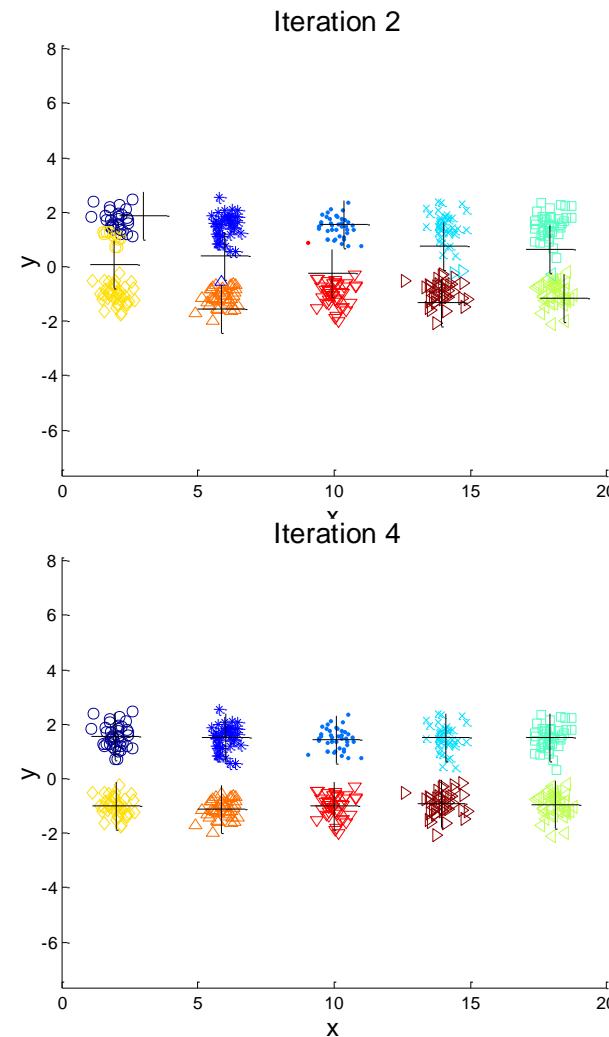
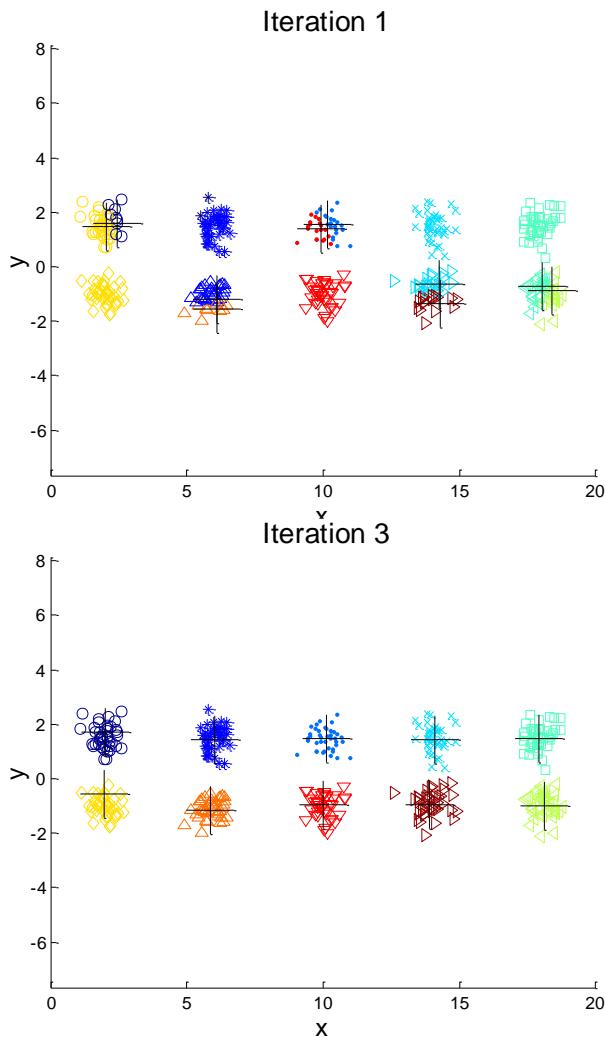


# Importanza della scelta dei centroidi iniziali



Due centroidi iniziali in  
un cluster di ogni  
coppia di cluster

# Importanza della scelta dei centroidi iniziali



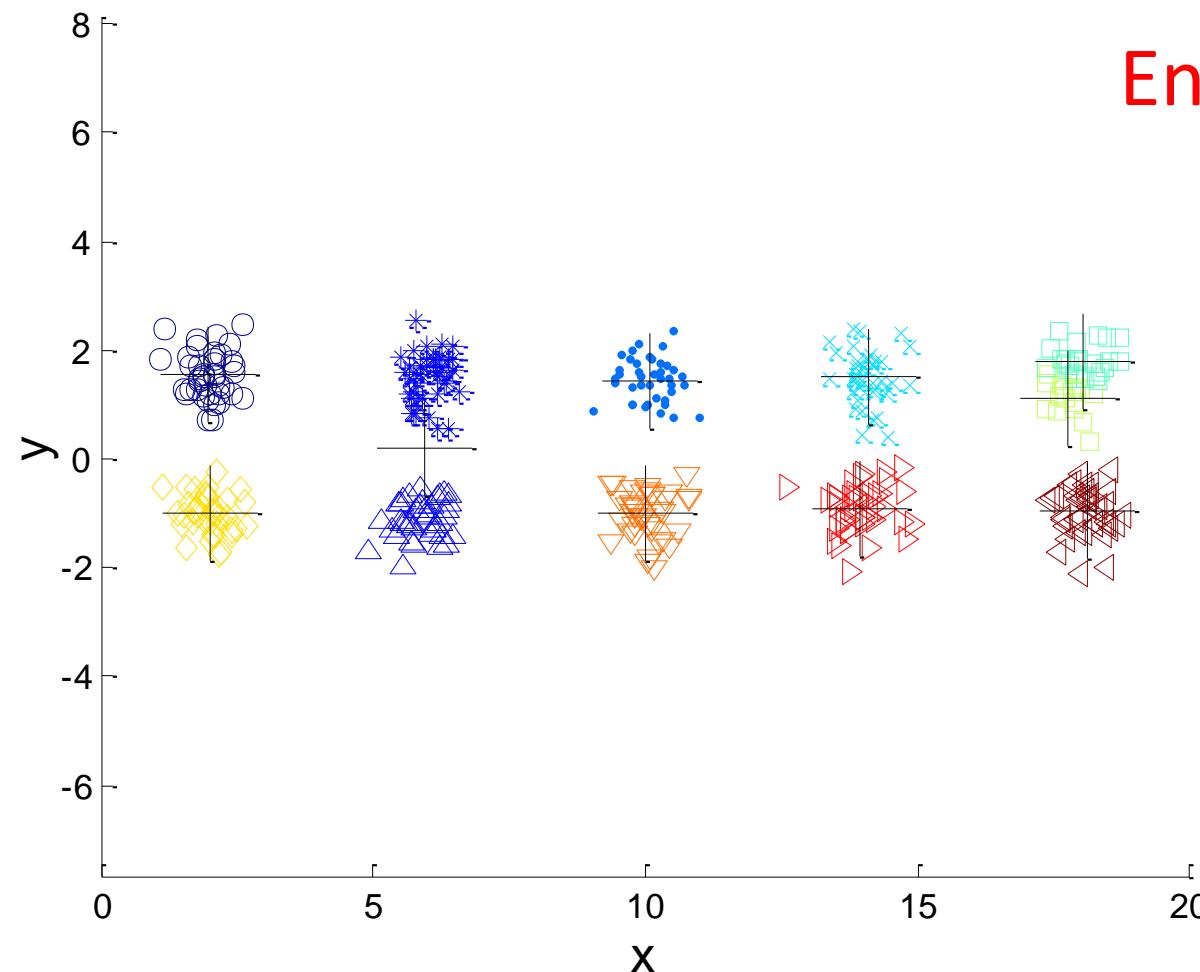
Due centroidi iniziali in  
un cluster di ogni  
coppia di cluster

# Importanza della scelta dei centroidi iniziali

Alcune coppie di cluster con tre centroidi iniziali, e altre con uno

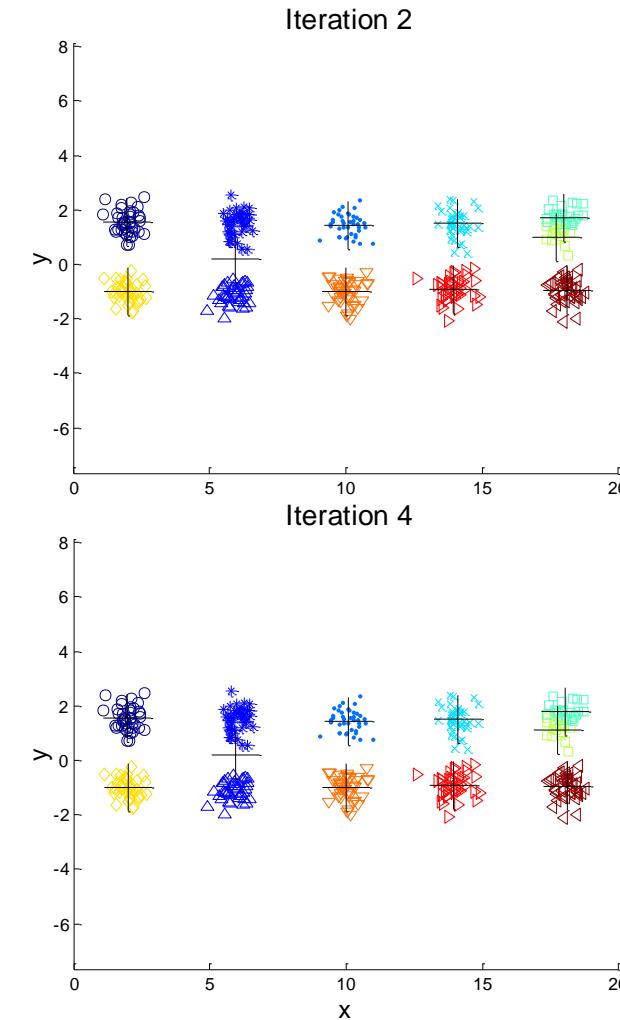
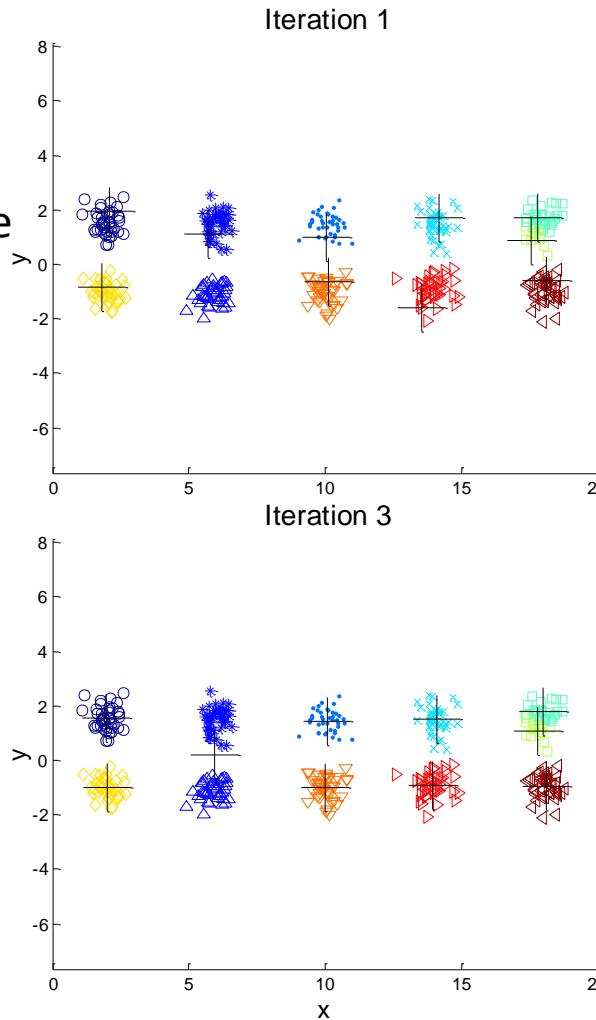
Iteration 4

End

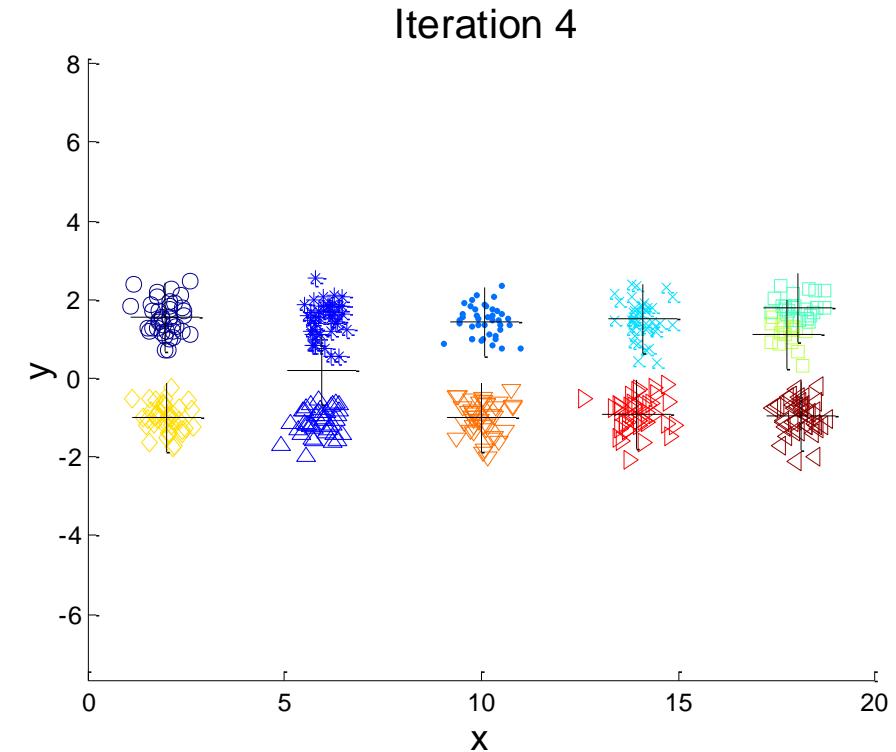
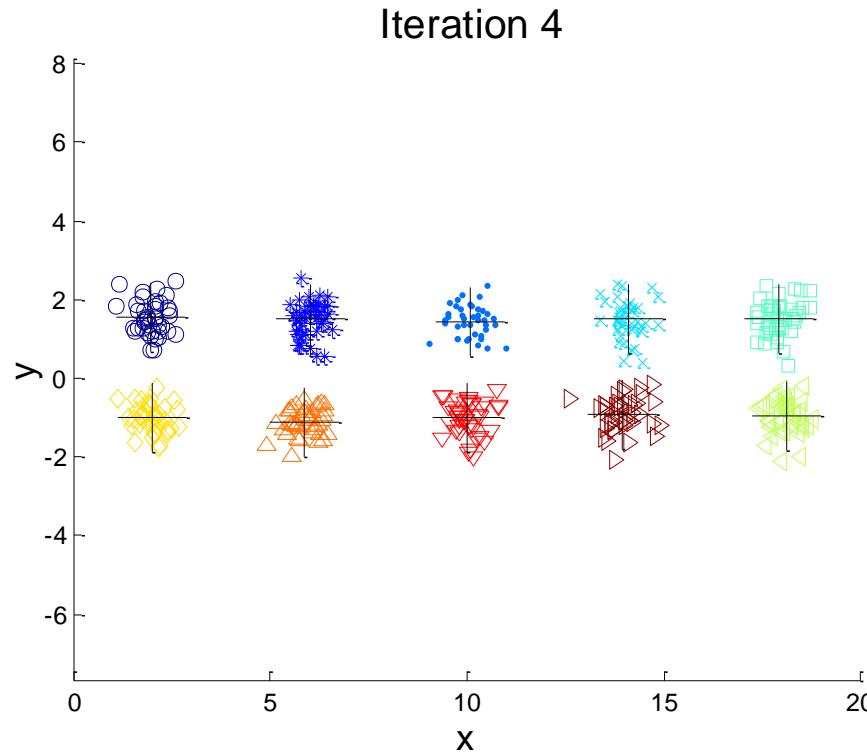


# Importanza della scelta dei centroidi iniziali

Alcune coppie di cluster  
con tre centroidi iniziali, e  
altre con uno



# Importanza della scelta dei centroidi iniziali



# K –means e la scelta iniziale dei centroidi

- Run Multiple (fare più prove con lo stesso numero di centroidi ma sceglindoli diversamente)
- Si usi il **clustering gerarchico** per determinare i centroidi iniziali (lo vedremo più avanti)
- Valutare la curva SSE al variare di K
- Pre e Post-processing
- Bisecting K-means (non così sensibile alla scelta dei centroidi iniziale)
- ...

# Valutazione dei cluster K-means

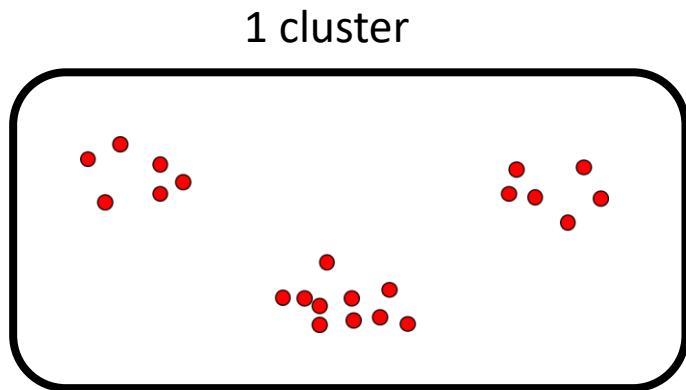
- E' la misura più comune: **Somma degli errori quadratici (SSE)**

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist^2(m_i, x)$$

- $x$  è un punto del cluster  $C_i$  e  $m_i$  è centroide di  $C_i$
- **Dati due clustering, possiamo scegliere quello con SSE minore**
- **Di solito** SSE si riduce incrementando K,

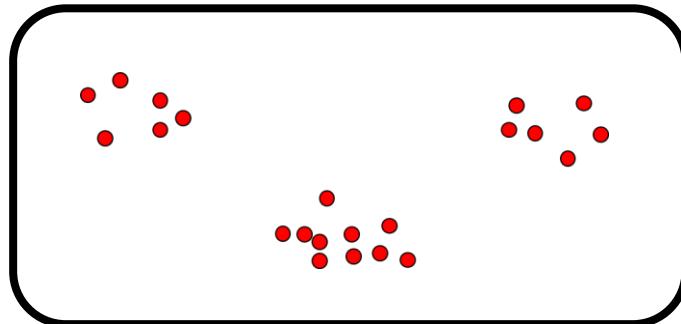
Si noti però che: un buon raggruppamento con un certo K può avere un SSE inferiore a un clustering scorretto con K più elevato

# Valutare la curva SSE al variare di K (K ottimale)

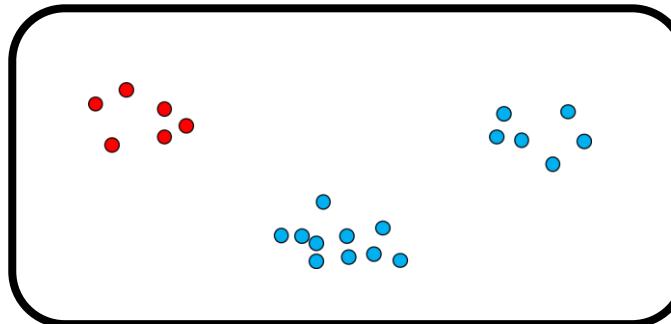


# Valutare la curva SSE al variare di K (K ottimale)

1 cluster

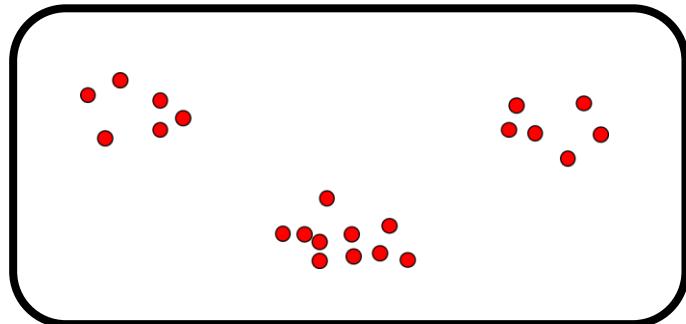


2 cluster

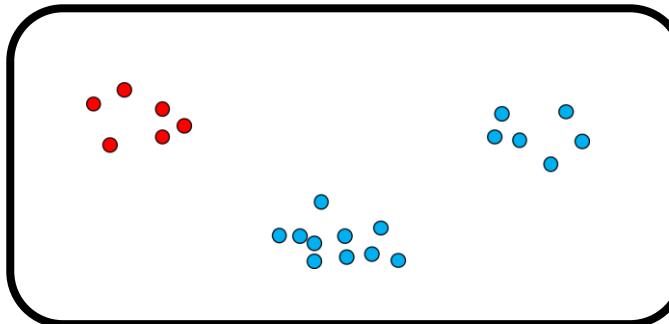


# Valutare la curva SSE al variare di K (K ottimale)

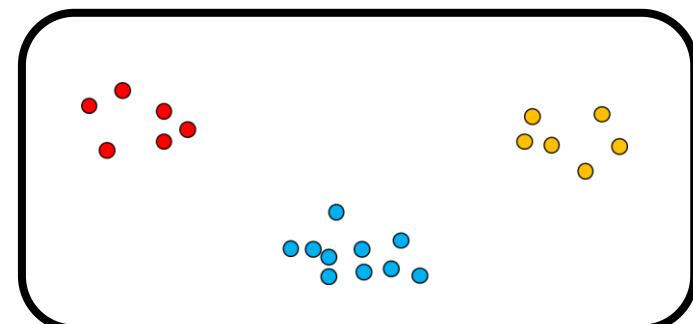
1 cluster



2 cluster

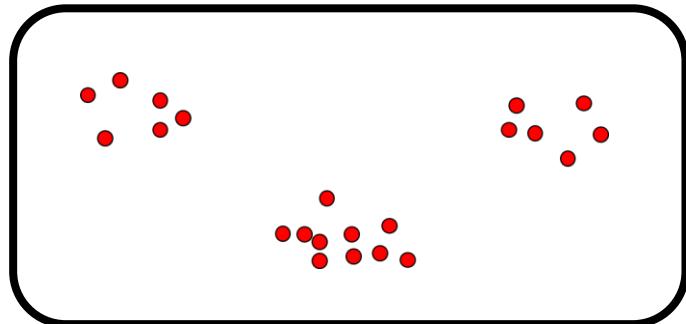


3 cluster

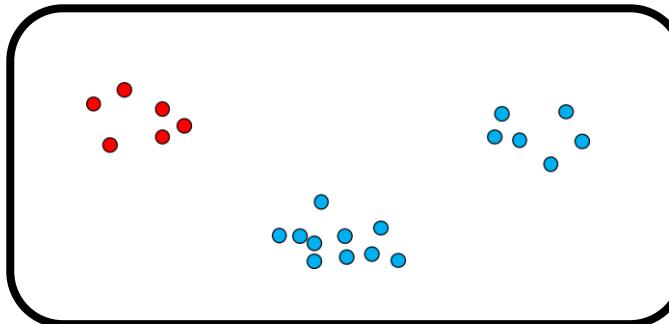


# Valutare la curva SSE al variare di K (K ottimale)

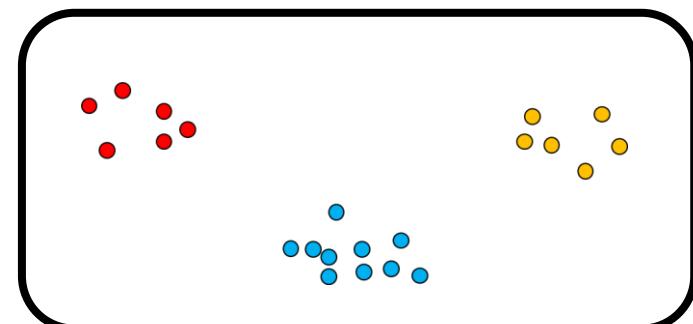
1 cluster



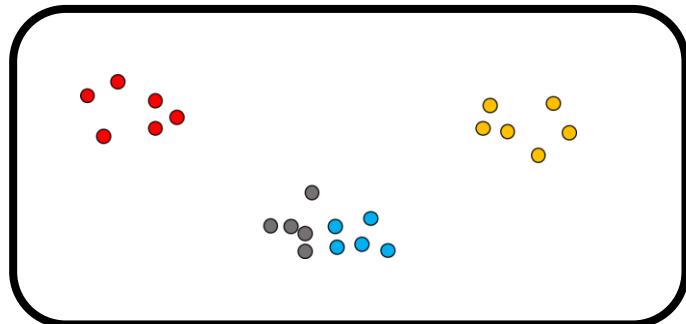
2 cluster



3 cluster

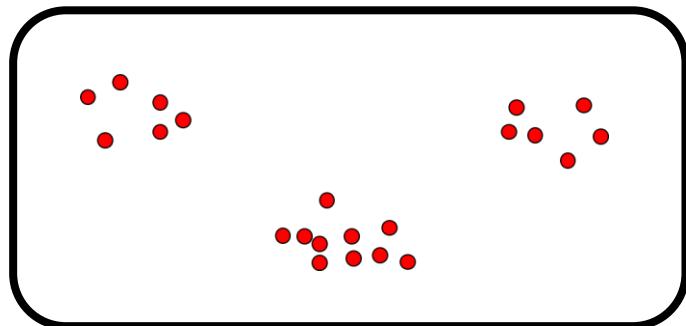


4 cluster

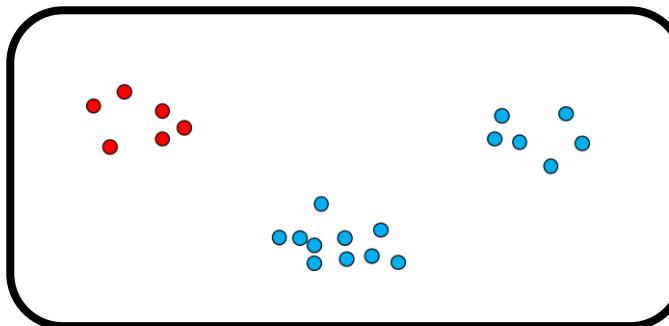


# Valutare la curva SSE al variare di K (K ottimale)

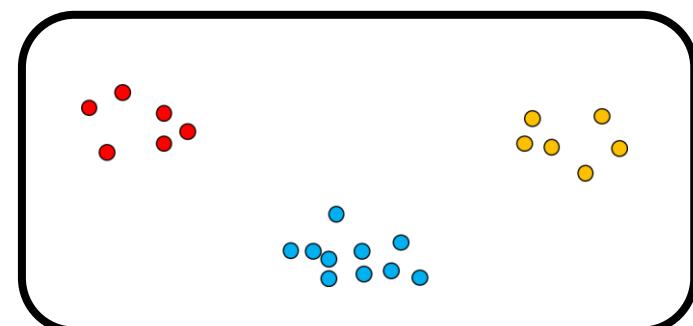
1 cluster



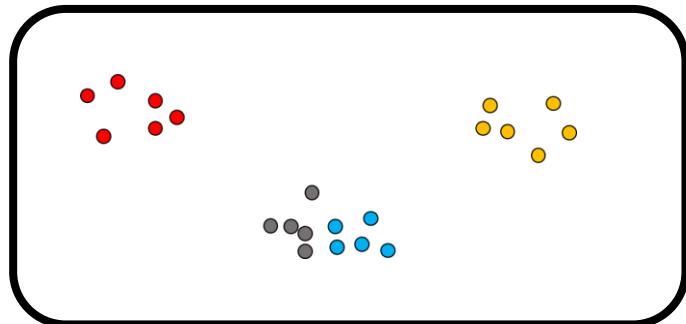
2 cluster



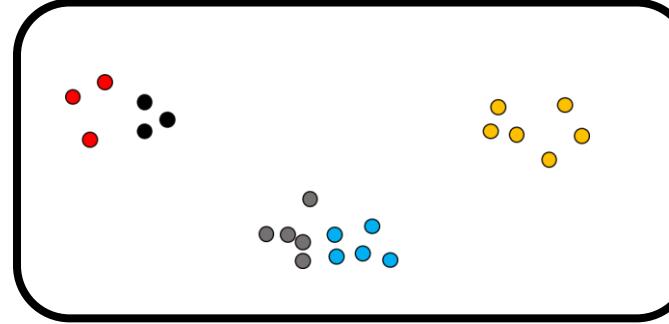
3 cluster



4 cluster

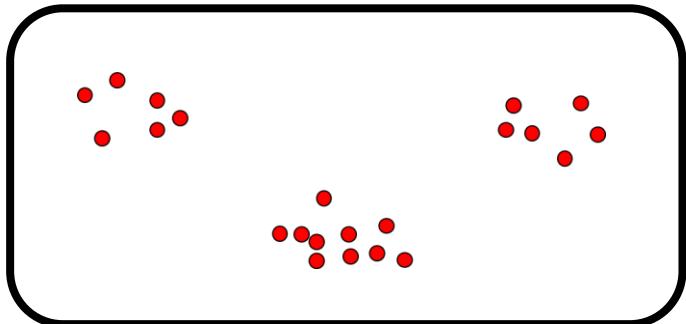


5 cluster

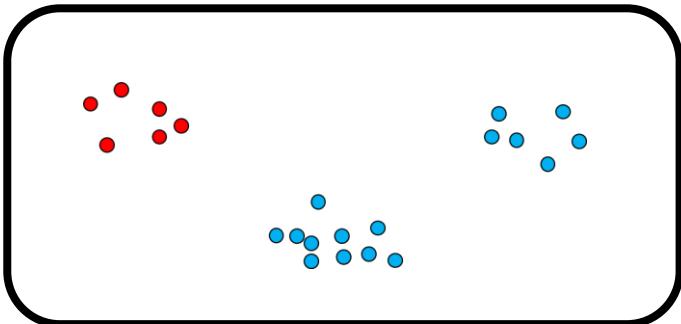


# Valutare la curva SSE al variare di K (K ottimale)

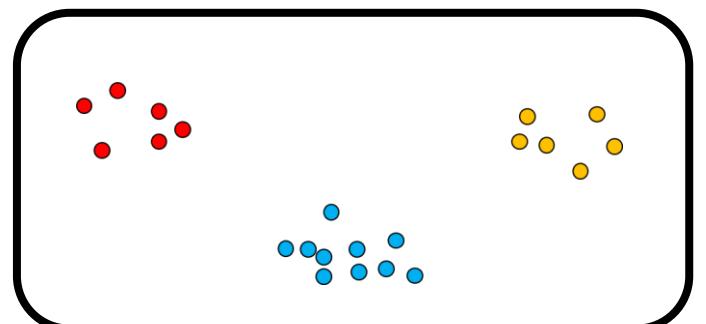
1 cluster



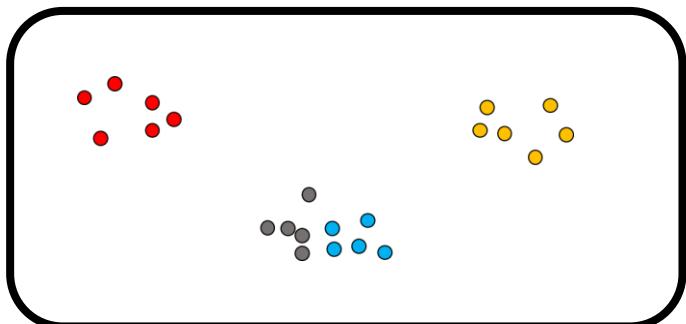
2 cluster



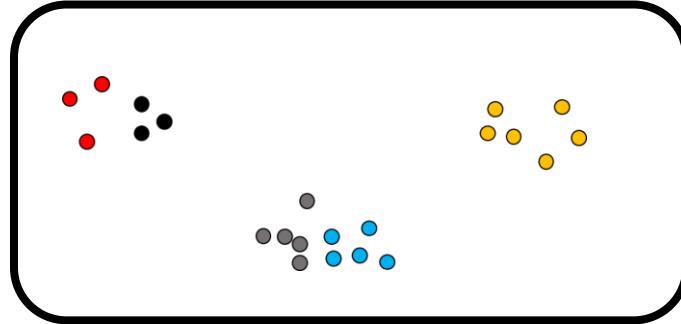
3 cluster



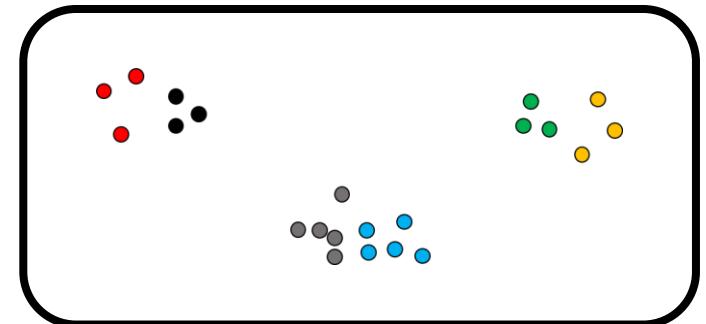
4 cluster



5 cluster

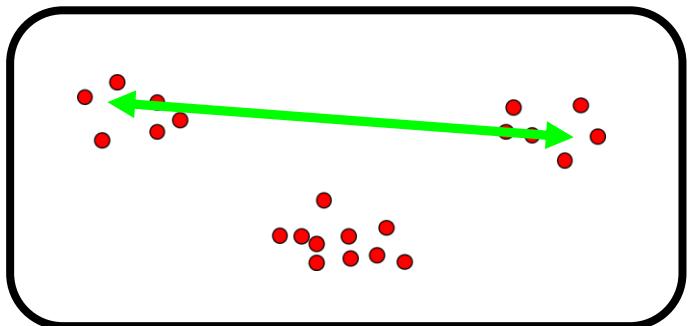


6 cluster

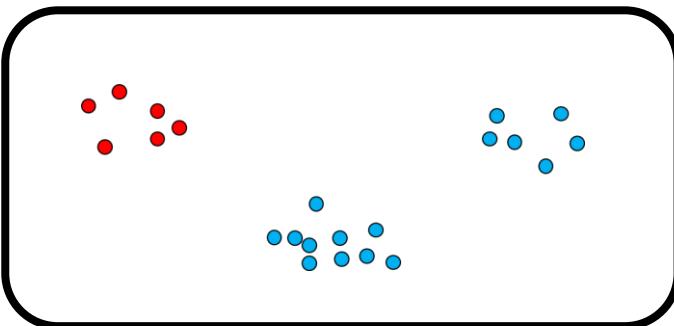


# Valutare la curva SSE al variare di K (K ottimale)

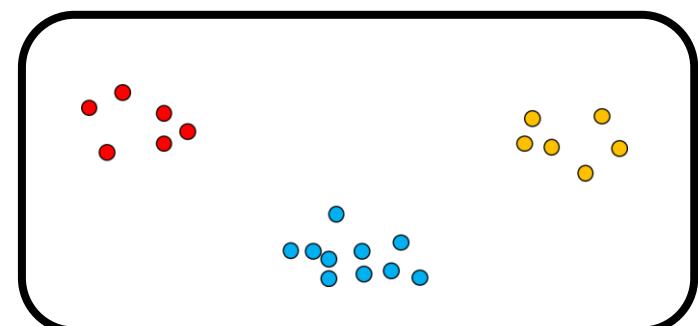
1 cluster



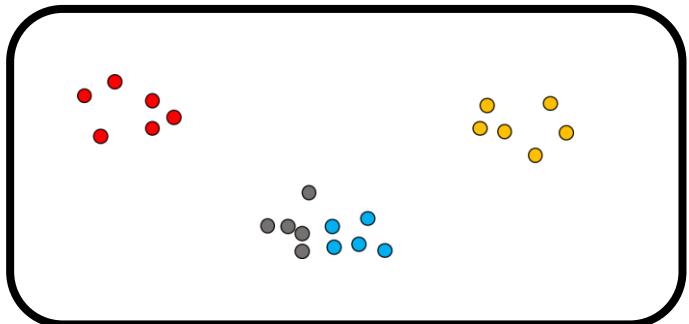
2 cluster



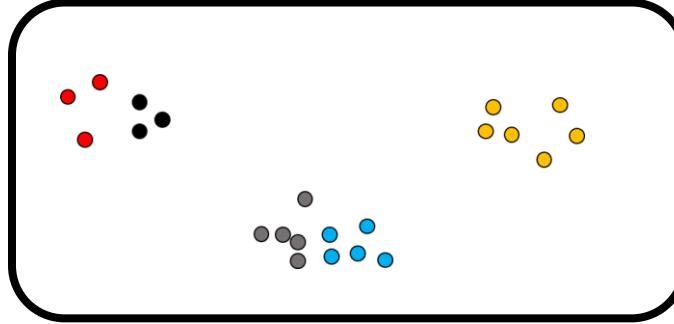
3 cluster



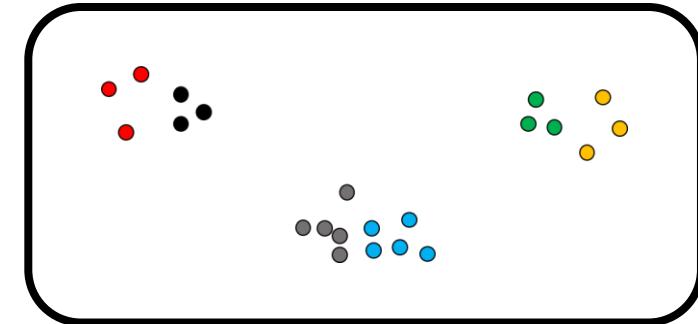
4 cluster



5 cluster

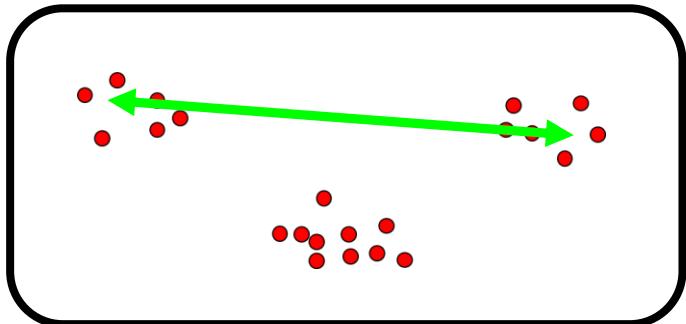


6 cluster

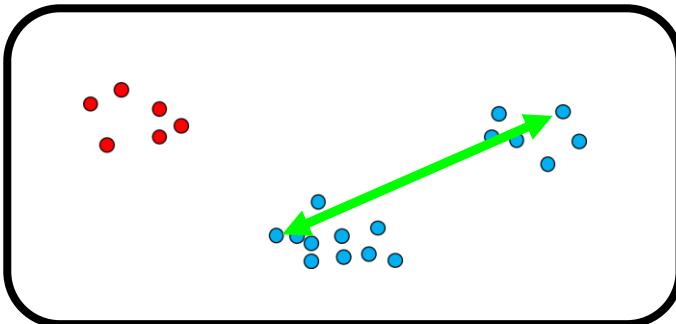


# Valutare la curva SSE al variare di K (K ottimale)

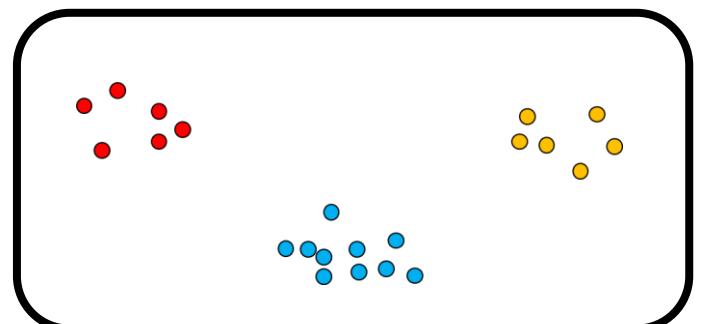
1 cluster



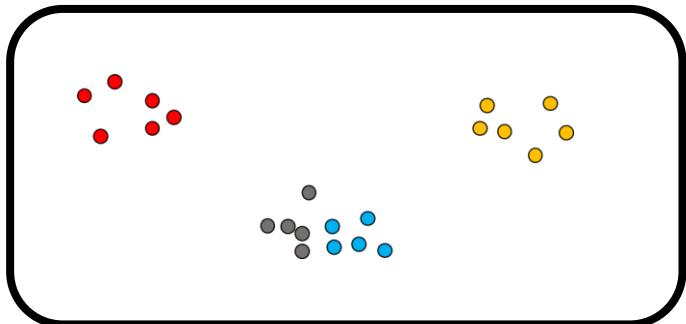
2 cluster



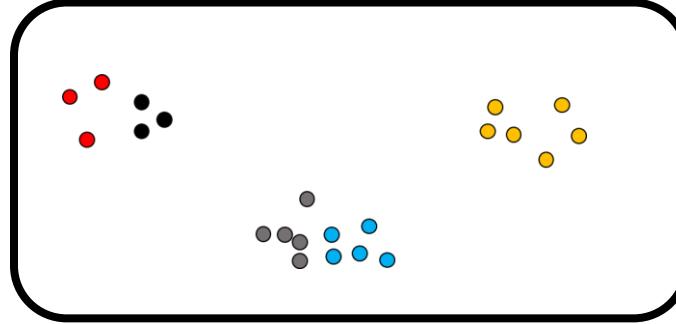
3 cluster



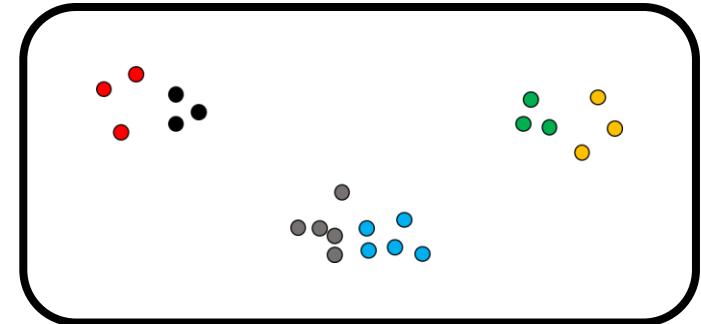
4 cluster



5 cluster

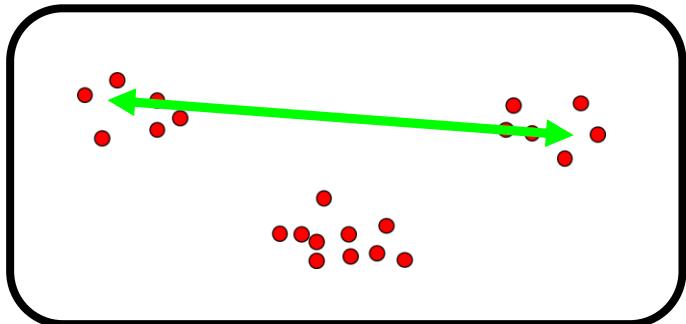


6 cluster

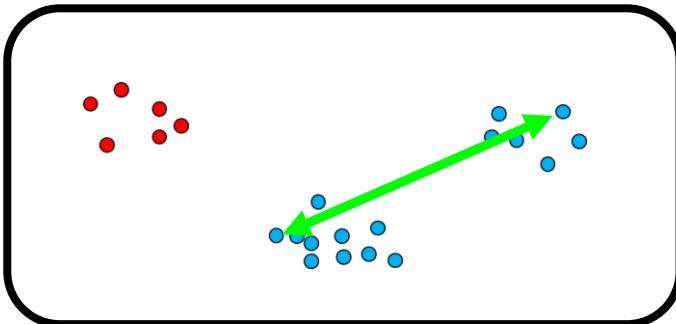


# Valutare la curva SSE al variare di K (K ottimale)

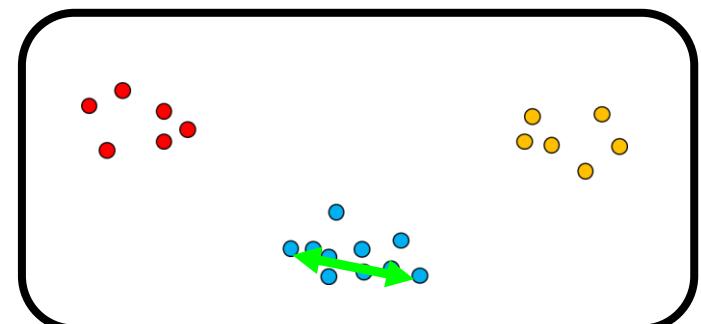
1 cluster



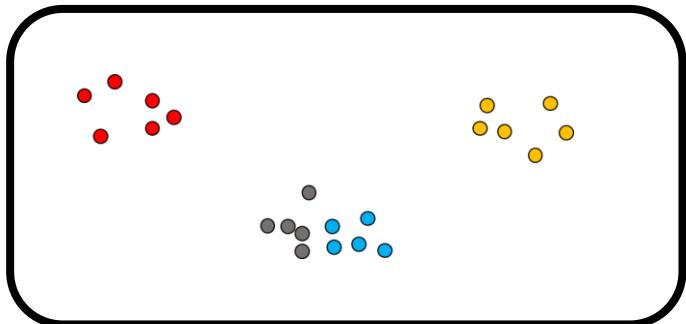
2 cluster



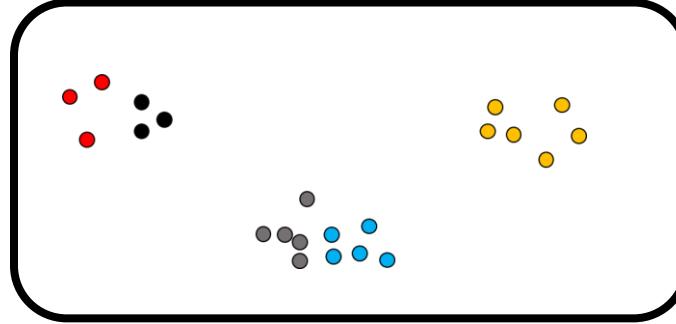
3 cluster



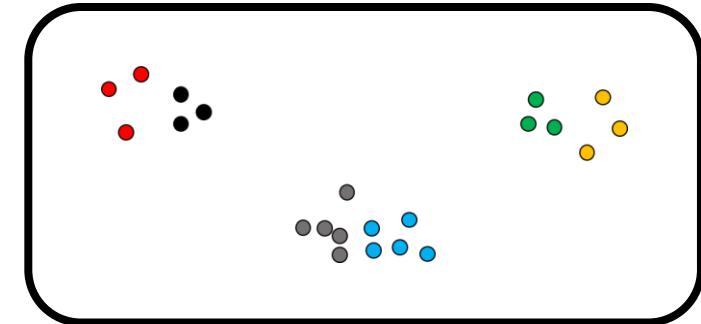
4 cluster



5 cluster

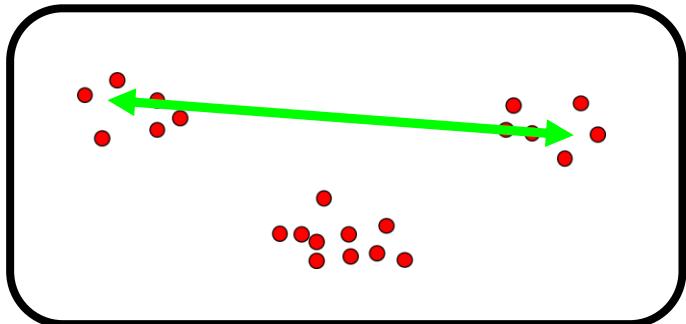


6 cluster

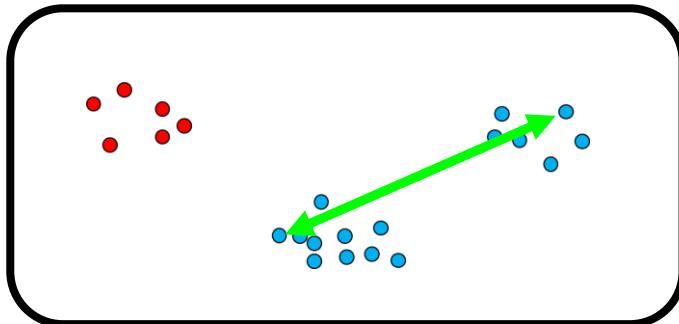


# Valutare la curva SSE al variare di K (K ottimale)

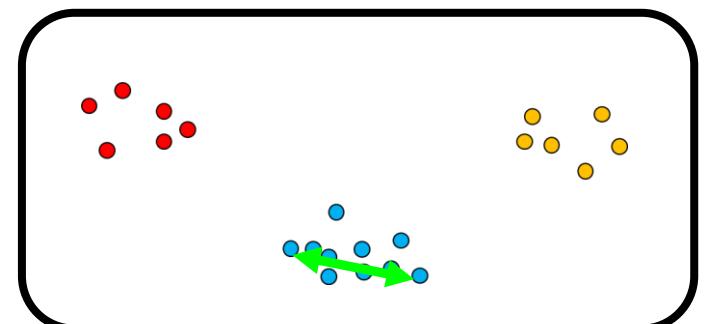
1 cluster



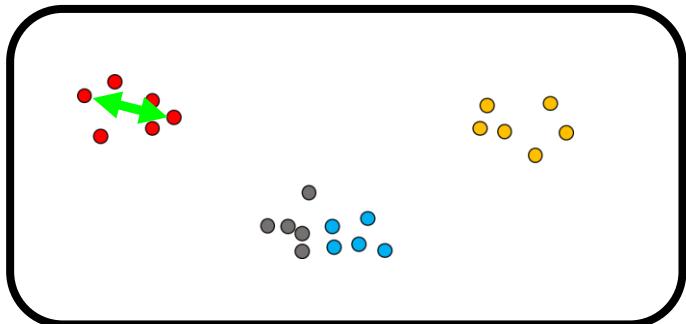
2 cluster



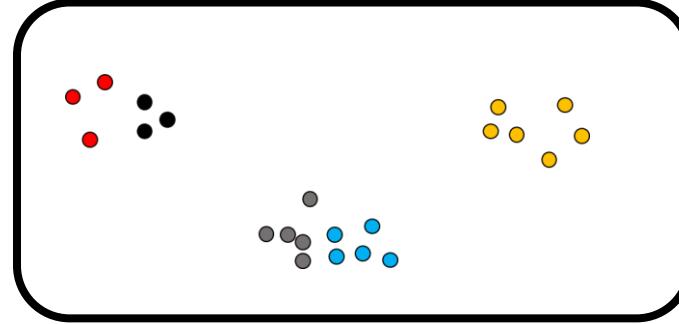
3 cluster



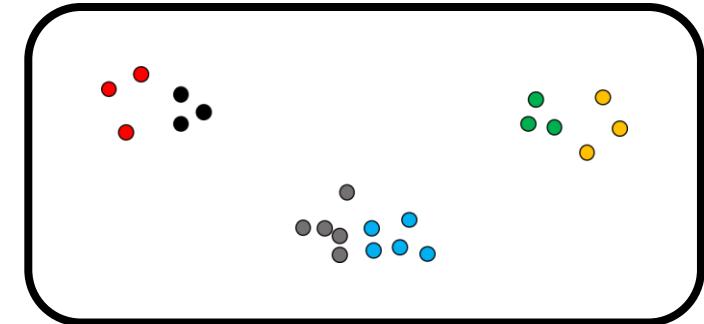
4 cluster



5 cluster

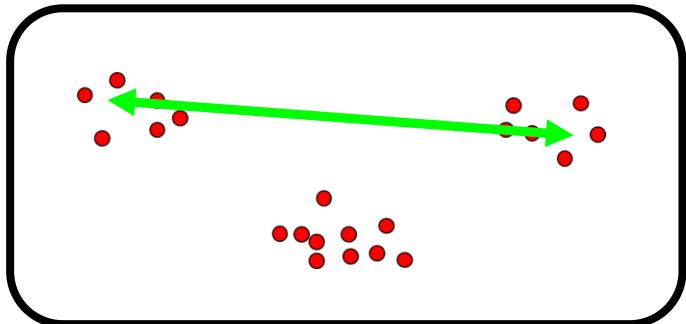


6 cluster

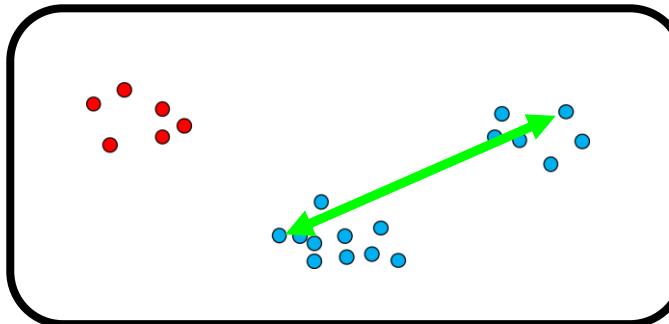


# Valutare la curva SSE al variare di K (K ottimale)

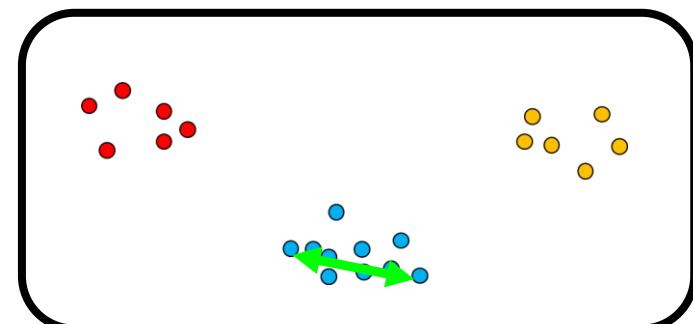
1 cluster



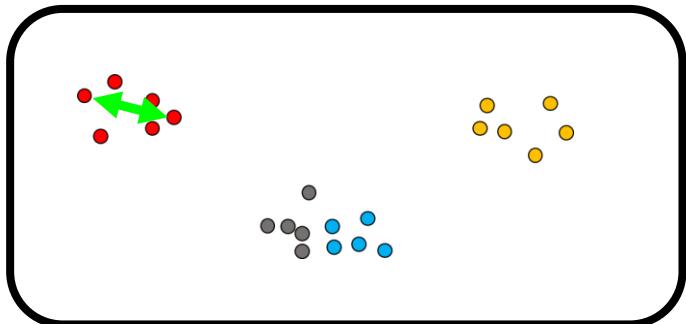
2 cluster



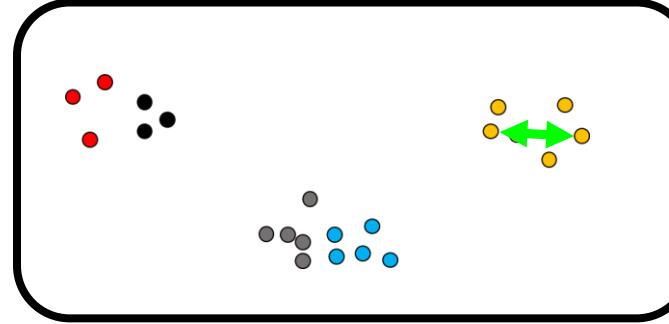
3 cluster



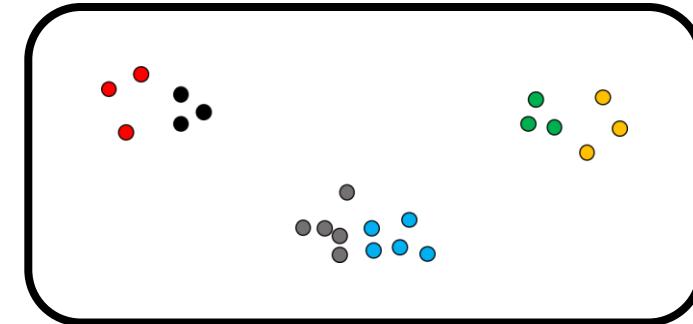
4 cluster



5 cluster

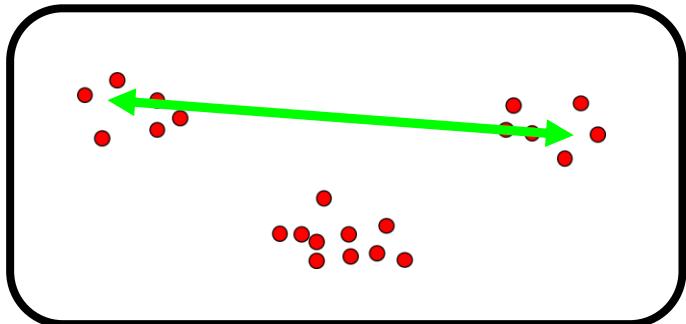


6 cluster

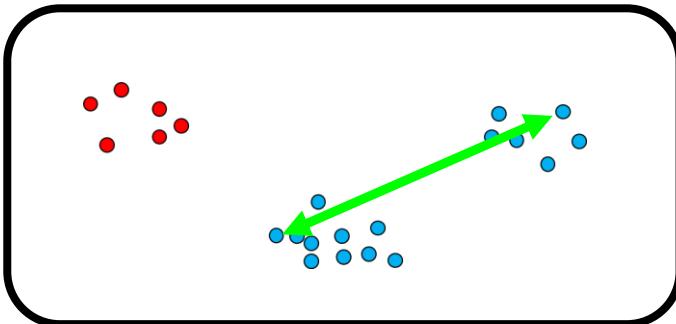


# Valutare la curva SSE al variare di K (K ottimale)

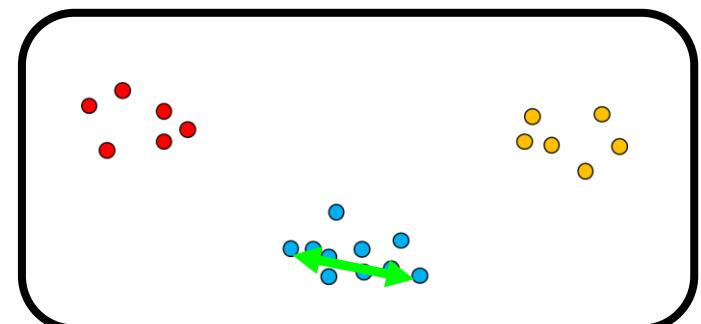
1 cluster



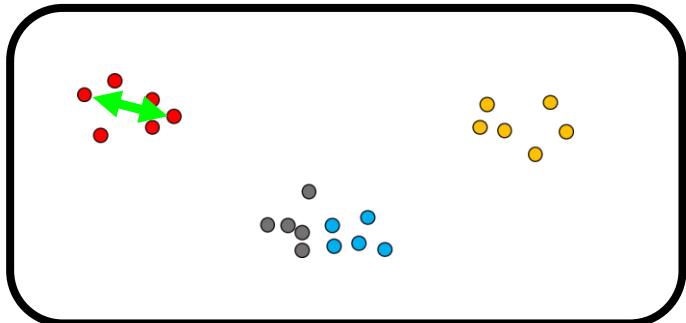
2 cluster



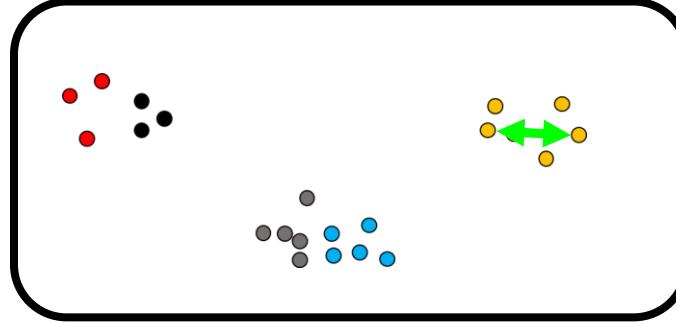
3 cluster



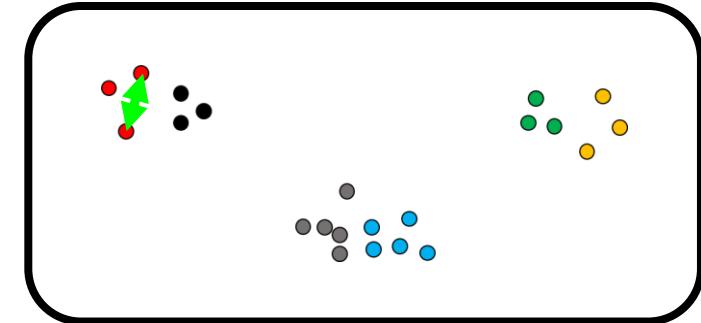
4 cluster



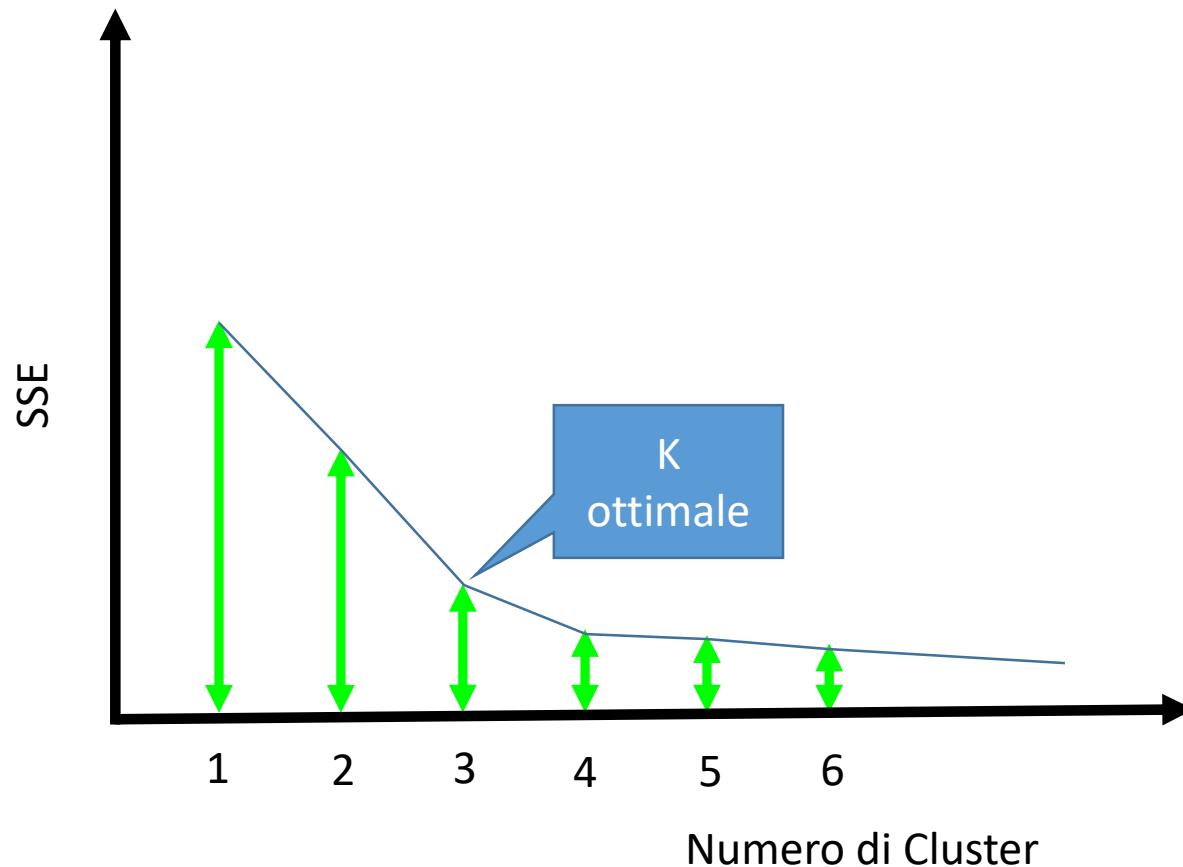
5 cluster



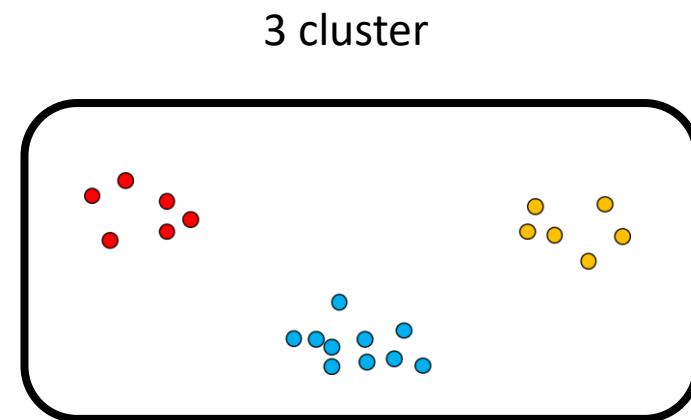
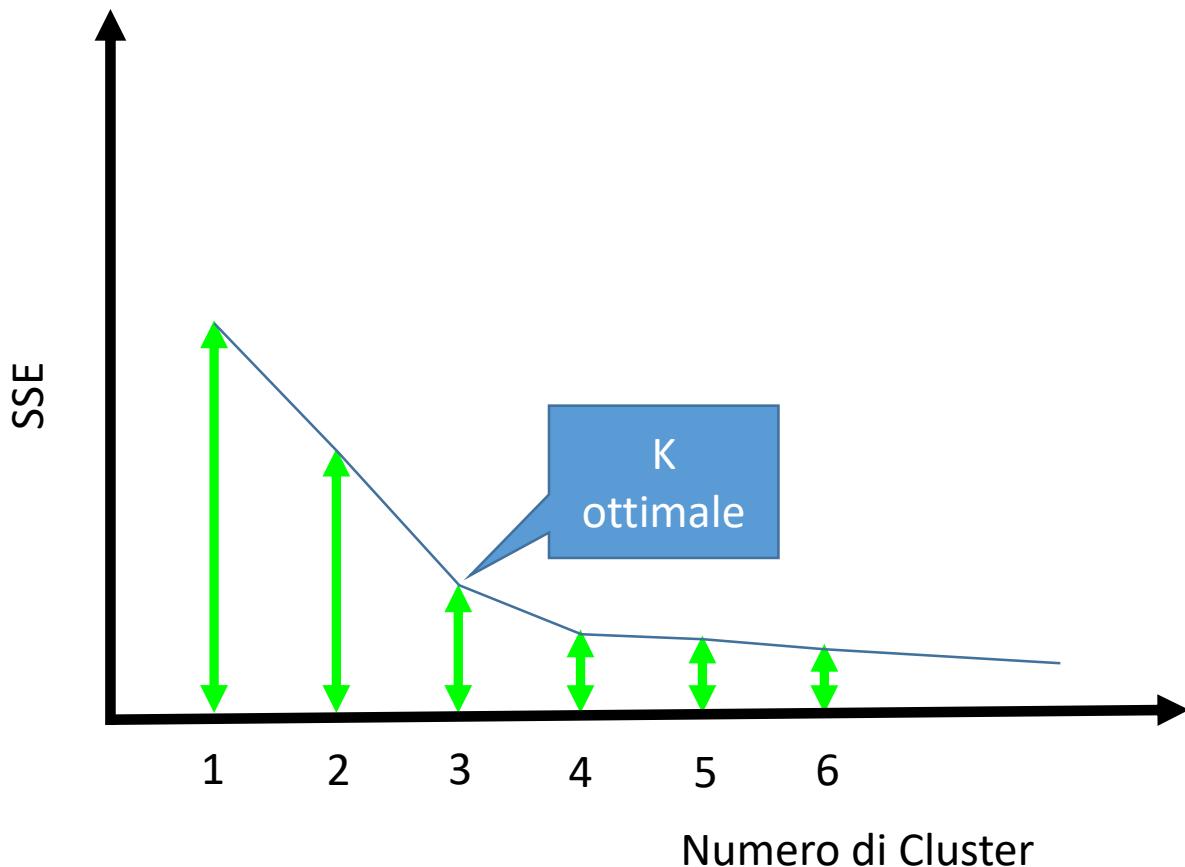
6 cluster



# Valutare la curva SSE al variare di K (K ottimale)

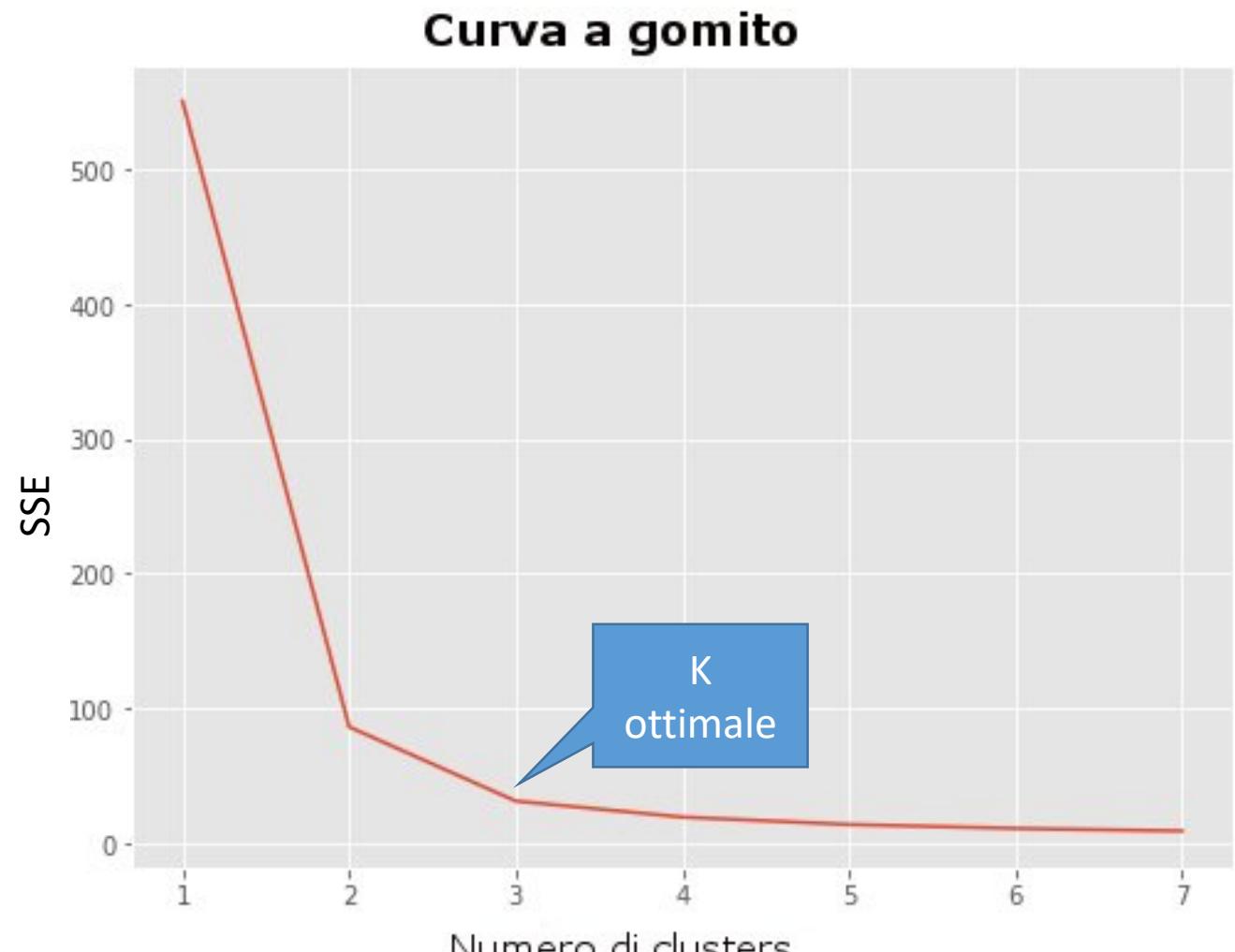


# Valutare la curva SSE al variare di K (K ottimale)



# Valutare la curva SSE al variare di K (K ottimale)

- L'obiettivo di questo processo è trovare quindi il punto in cui l'aumento di k causerà una diminuzione molto piccola di SSE, mentre la diminuzione di k aumenterà bruscamente la somma (il K per cui l'angolazione della curva si stabilizza)
- Questo punto dolce è chiamato il “**punto di gomito**” (nella figura K=3)



## Ancora sul problema dei centroidi iniziali

Se ci sono K cluster 'reali' la probabilità di selezionare un centroide per ogni cluster è piccola.

Supponendo che i cluster abbiano la stessa cardinalità n

$$P = \frac{\# \text{ modi di scegliere un centroide per cluster}}{\# \text{ modi di scegliere un centroide}} = \frac{K! n^K}{(Kn)^K} = \frac{K!}{K^K}$$

Es: K = 10, la probabilità è  $10!/10^{10} = 0.0003688$

# Pre e Post-Processing

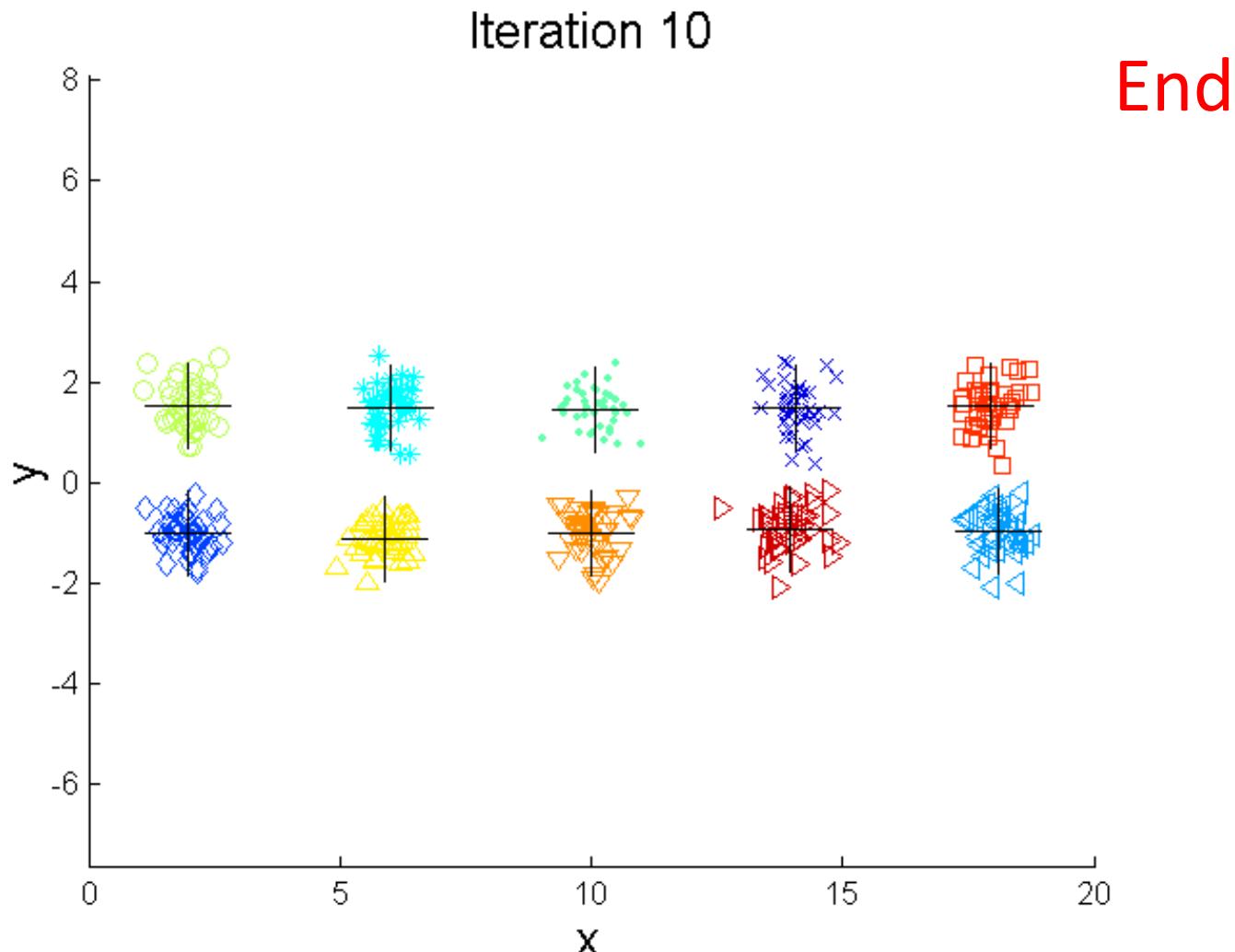
- Pre-processing
  - Normalizzare i dati
  - Rimuovere outlier (anomalie)
- Post-processing
  - Eliminare cluster piccoli che possono essere outlier
  - Spezzare cluster deboli (es. SSE>>)
  - Unire cluster vicini (es. SSE<<)

# Bisecting K-means

## Algoritmo:

- Si inizi con un cluster che contiene tutte le unità
- Bisezionare il cluster usando k-means ( ovvero K=2)
- Dei due cluster formati, bisezionare quello con SSE più elevato
- Procedere fino ad ottenere K cluster

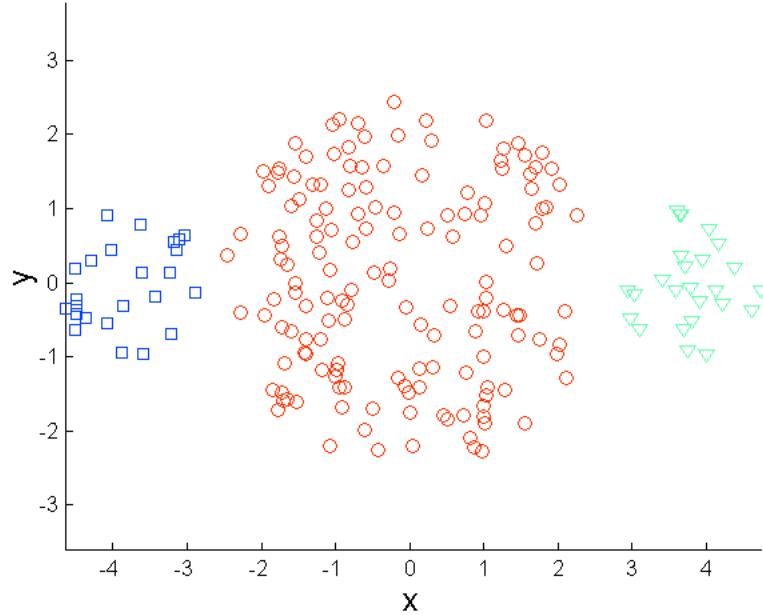
## Esempio: Bisecting K-means



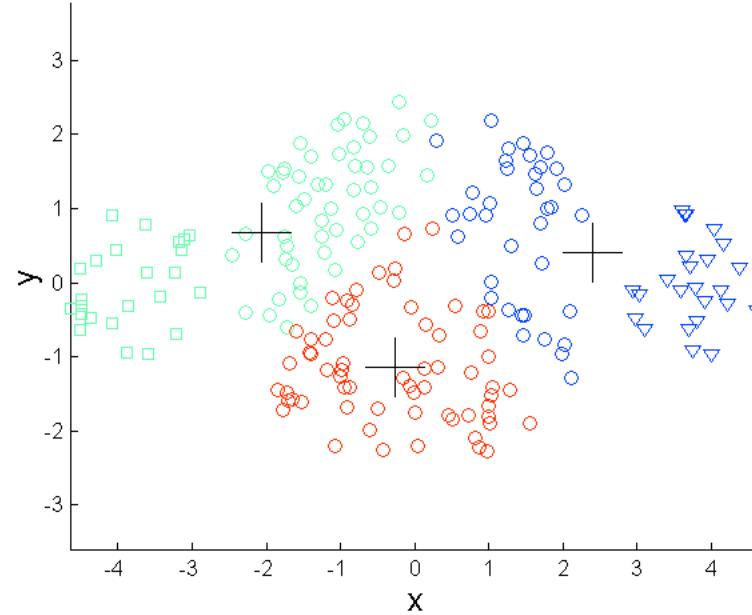
# Limitazioni di K-means

- K-Means ha problemi quando i cluster sono di differenti:
  - dimensioni
  - densità
  - forme non globulari
- K-means ha problemi quando i dati contengono outlier.

# Limitazioni di K-means: diverse dimensioni

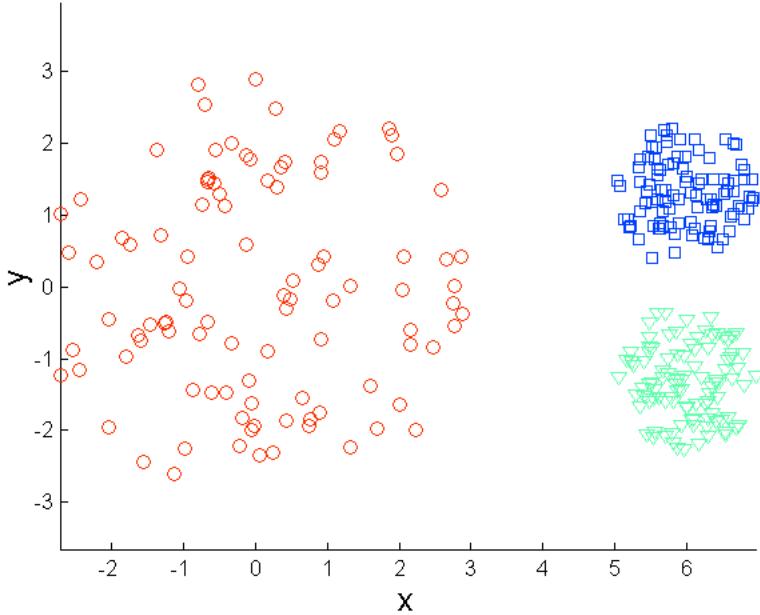


Gruppi reali

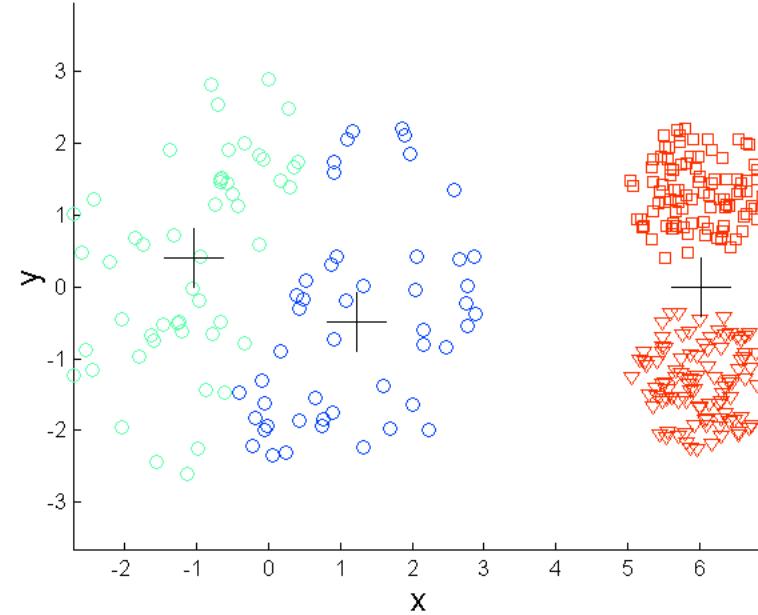


K-means (3 Cluster)

# Limitazioni di K-means : densità differenti

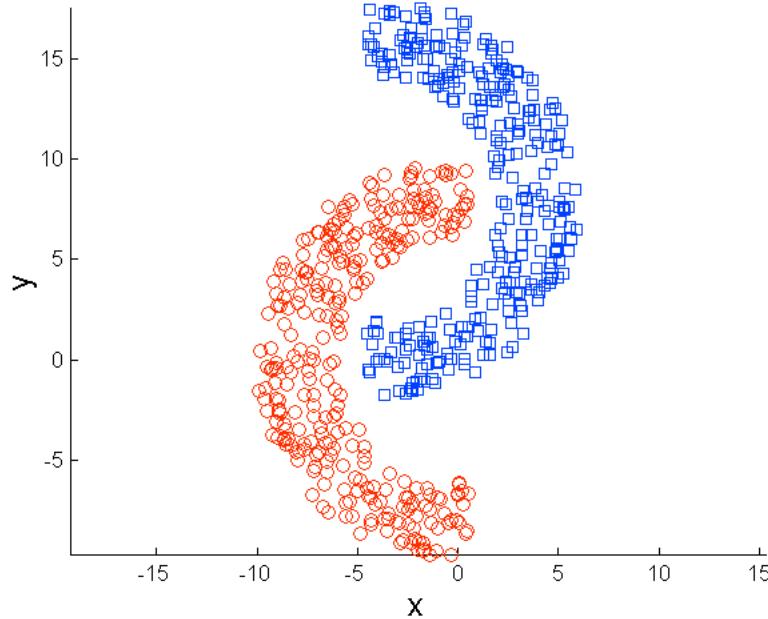


Gruppi reali

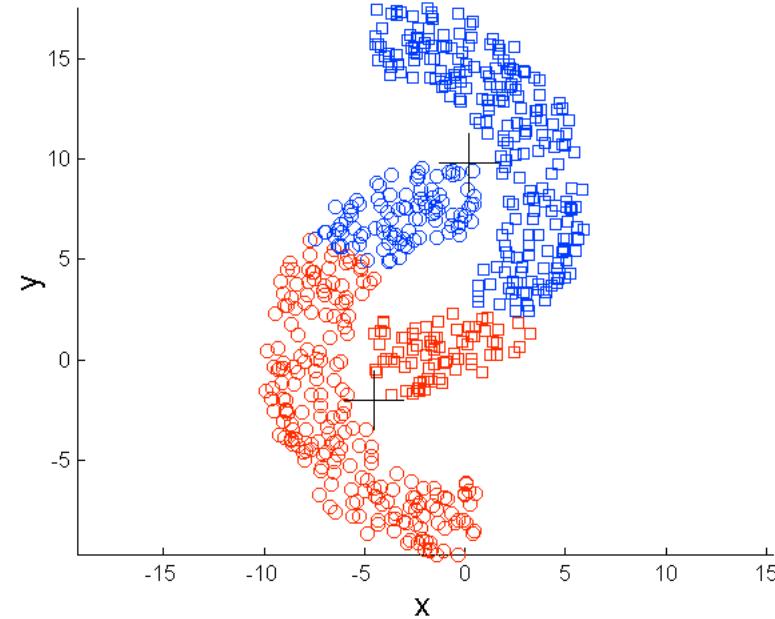


K-means (3 Cluster)

# Limitazioni di K-means : forme Non-globulari



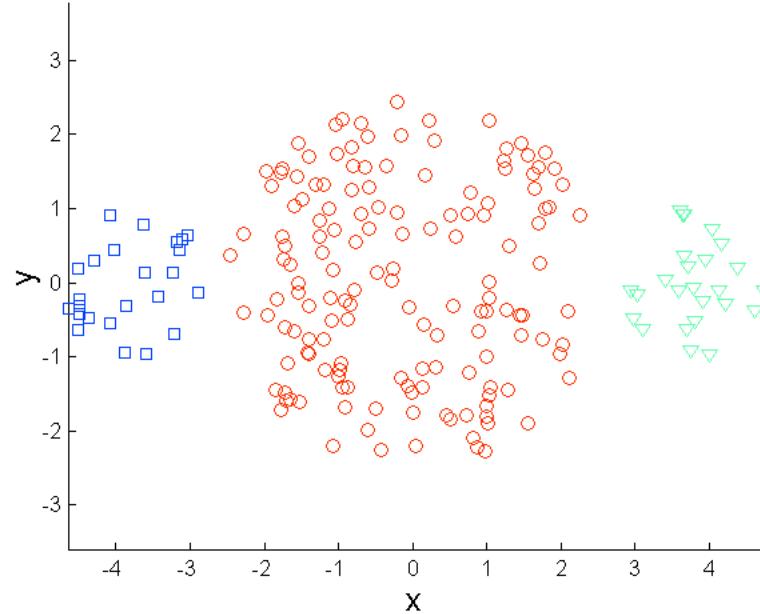
Gruppi reali



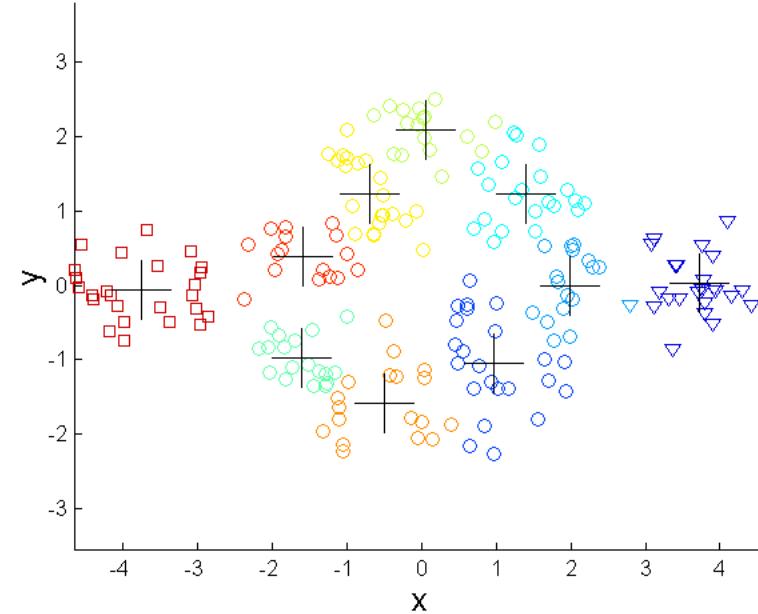
K-means (2 Cluster)

# Superare le limitazioni di K-means

Una soluzione è quella di utilizzare molti cluster. E' necessario però ricostruire i cluster successivamente.



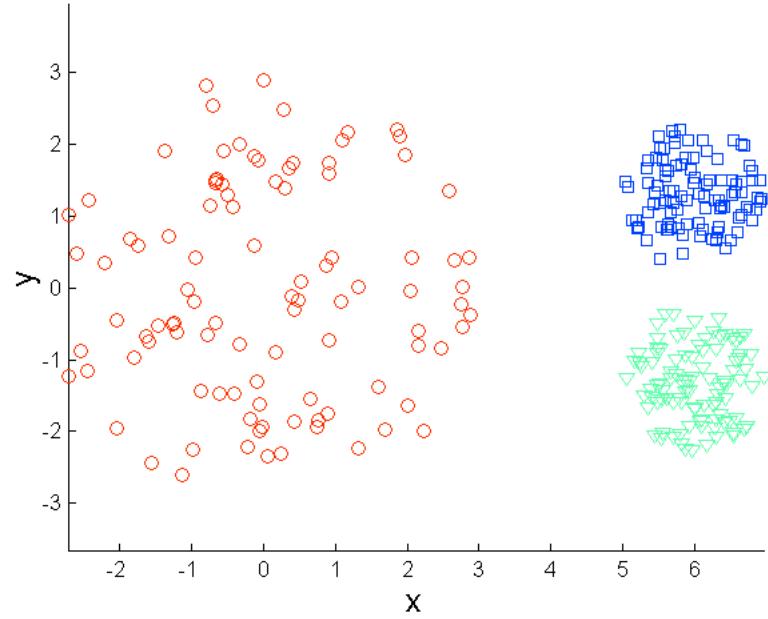
Gruppi reali



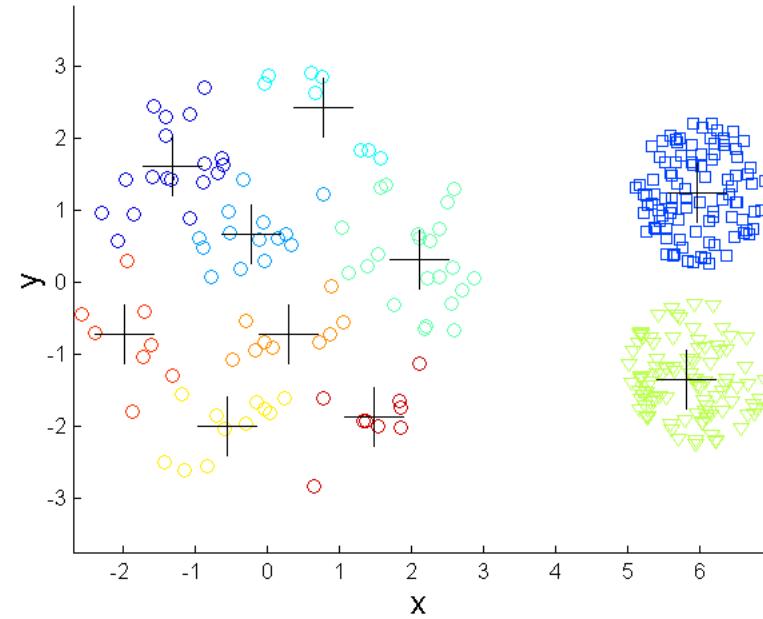
K-means Clusters

# Superare le limitazioni di K-means

Una soluzione è quella di utilizzare molti cluster. E' necessario però ricostruire i cluster successivamente.



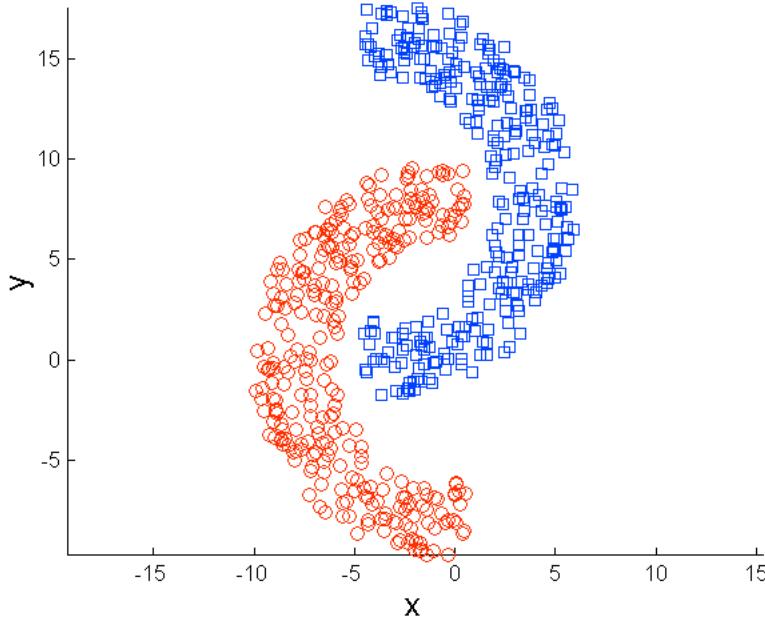
Gruppi reali



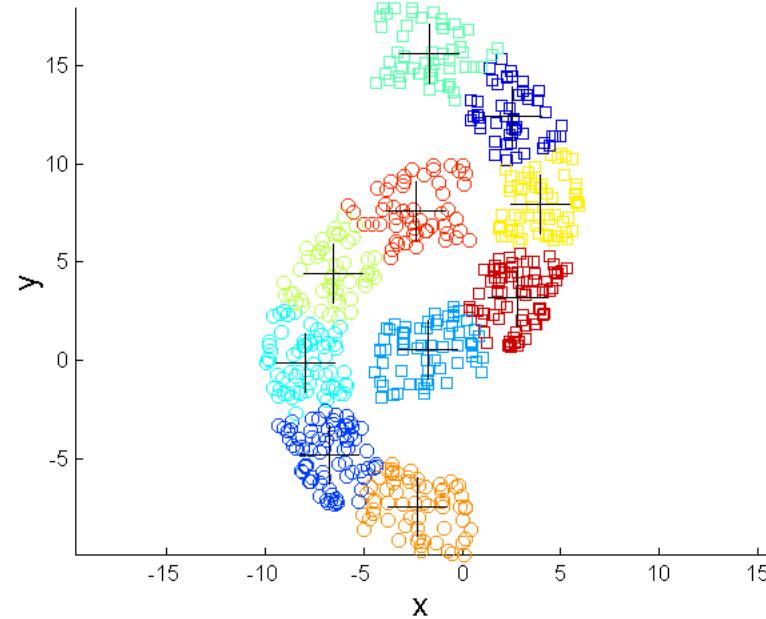
K-means Clusters

# Superare le limitazioni di K-means

Una soluzione è quella di utilizzare molti cluster. E' necessario però ricostruire i cluster successivamente.



Gruppi reali



K-means Clusters

# LAB Session



# LAB Session – Scikit-Learn



- **Scikit-Learn** è un progetto open source per Python che fornisce supporto ai principali algoritmi di Machine Learning e contiene tutti i moduli necessari per realizzare un progetto di apprendimento automatico (di base).
- È dotato di vari algoritmi di *classificazione*, *regressione* e *clustering* tra cui: support vector machines, random forests, gradient boosting, k-means and DBSCAN, etc.
- Il pacchetto è scritto prevalentemente in Python ma incorpora librerie C ++ come LibSVM e LibLinear per le Support Vector Machines e l'implementazione di modelli lineari generalizzati.
- Il pacchetto dipende da **Pandas** (principalmente per l'elaborazione di DataFrame), **NumPy** (per il costrutto narray) e **SciPy** (per matrici sparse).

# LAB Session – Scikit-Learn



## Classification

Identifying to which category an object belongs to.

**Applications:** Spam detection, Image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, ...

— Examples

## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, ridge regression, Lasso, ...

— Examples

## Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, ...

— Examples

## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

**Algorithms:** PCA, feature selection, non-negative matrix factorization.

— Examples

## Model selection

Comparing, validating and choosing parameters and models.

**Goal:** Improved accuracy via parameter tuning

**Modules:** grid search, cross validation, metrics.

— Examples

## Preprocessing

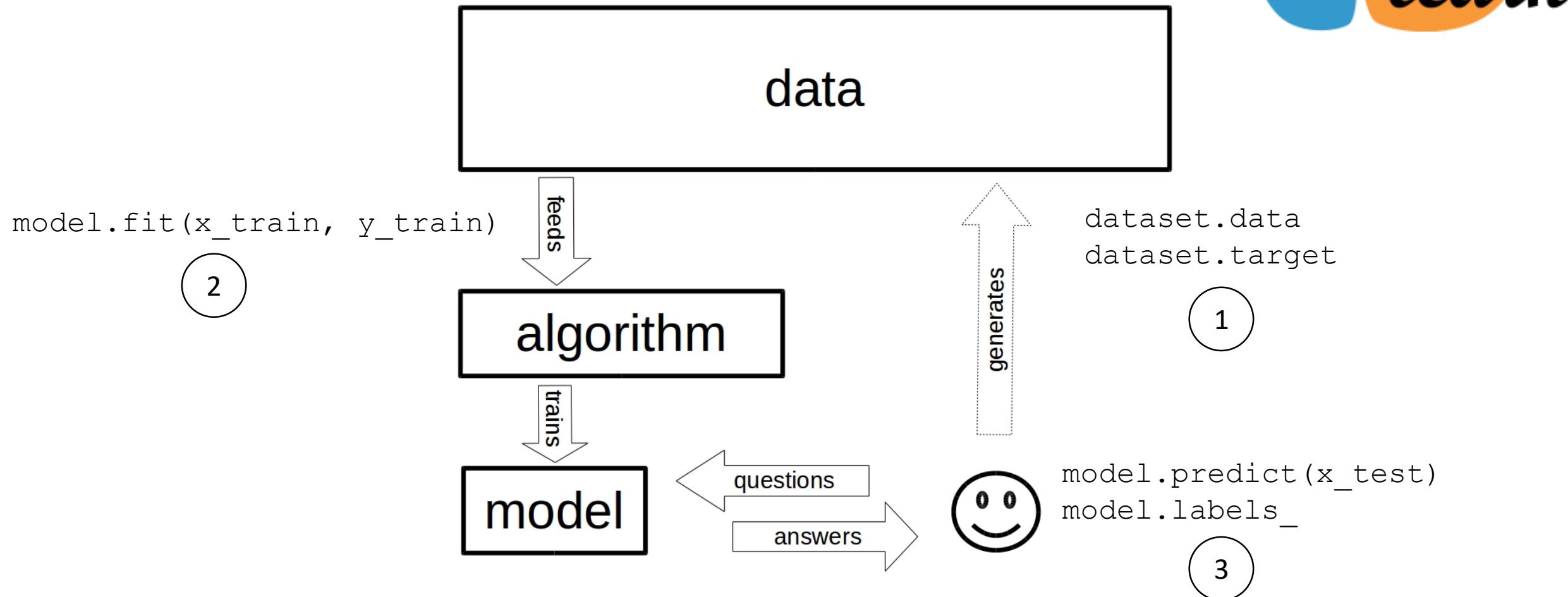
Feature extraction and normalization.

**Application:** Transforming input data such as text for use with machine learning algorithms.

**Modules:** preprocessing, feature extraction.

— Examples

# LAB Session – K-Mean



# Learning Session

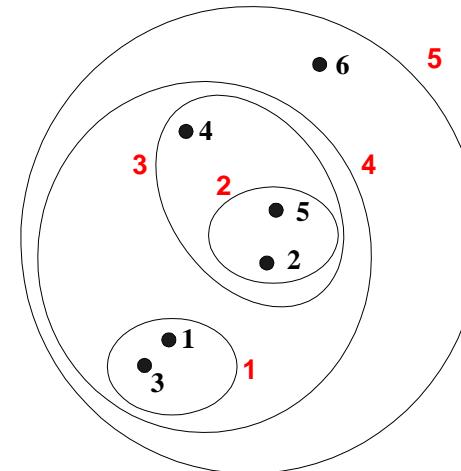
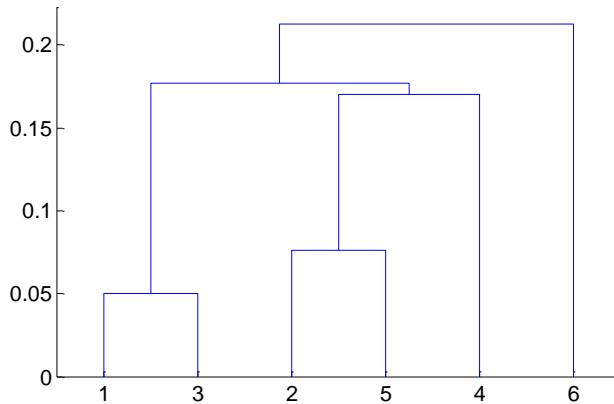


# Algoritmi di Clustering

- K-means e le sue varianti (partizionale)
- Clustering gerarchico
- Clustering basato sulla densità
- Misure di validità dei Cluster

# Clustering gerarchico

- Produce un insieme di cluster **innestati, organizzati come un albero gerarchico**
- Può essere visualizzato con un dendrogramma



# Clustering gerarchico

- Due tipi principali di clustering gerarchico
  - Agglomerativo:
    - Inizia con i punti considerati come singoli cluster
    - Ad ogni passo, unisce la coppia più vicina di cluster fino a quando rimane un solo cluster (o k cluster)
    - E' necessaria una nozione di prossimità tra cluster
  - Divisivo:
    - Inizia con un cluster, (contiene tutti i punti)
    - Ad ogni passo, divide un cluster fino a quando ogni cluster contiene un solo punto (o ci sono k cluster)
    - E' necessario scegliere quale cluster spezzare ad ogni passo
- Gli algoritmi gerarchici utilizzano una matrice di **similarità** o di **dissimilarità (distanza)**
- Si unisce o si divide un gruppo alla volta

# Algoritmo Agglomerativo

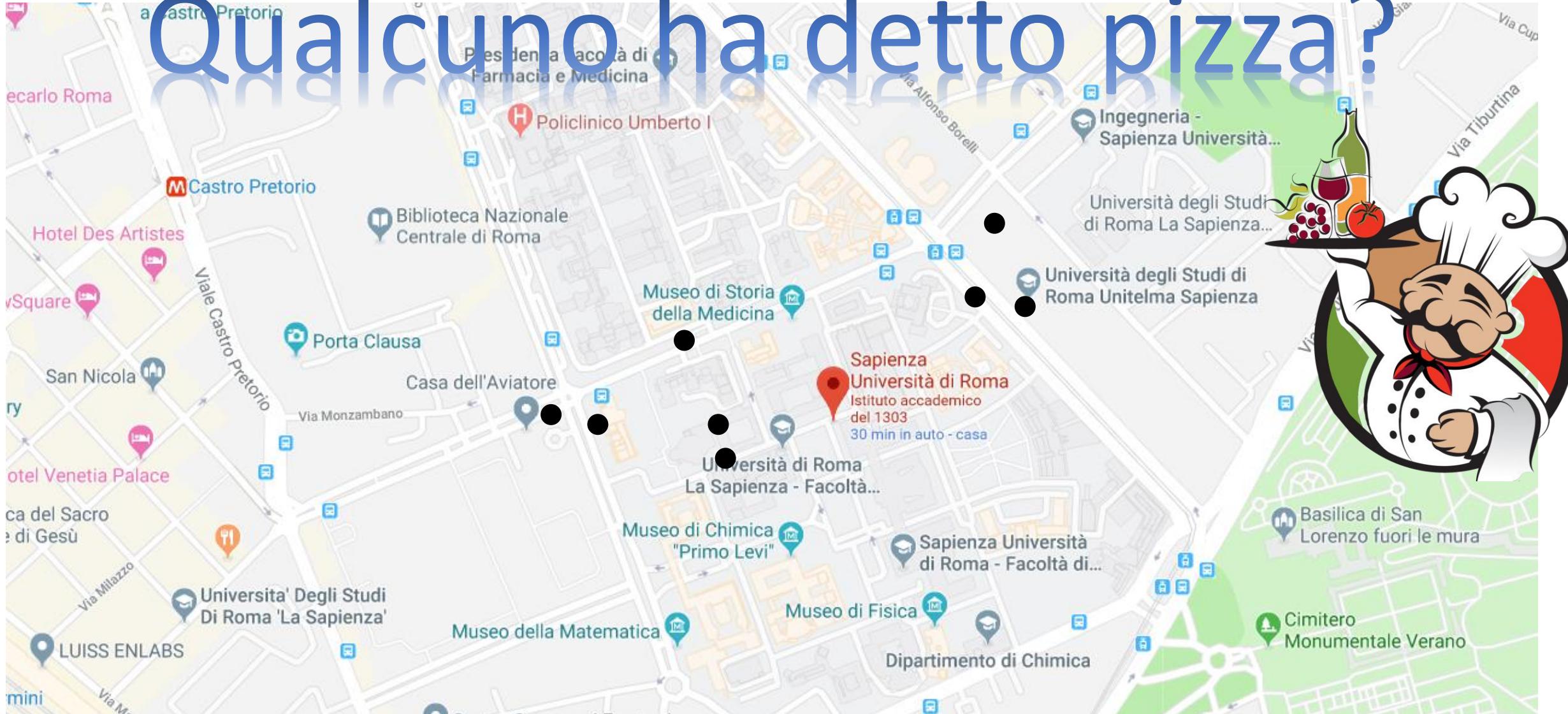
- La tecnica di clustering gerarchico più diffusa
- Algoritmo di base
  1. Calcola una matrice di prossimità (similarità o dissimilarità)
  2. Considera inizialmente ogni punto come un cluster
  3. **Ripeti**
  4. Unisci i due cluster più vicini
  5. Aggiorna la matrice di prossimità
  6. **Fino a** quando rimane un solo cluster.

Punto chiave  
la matrice di prossimità.

Approcci differenti nella definizione di prossimità portano a soluzioni differenti

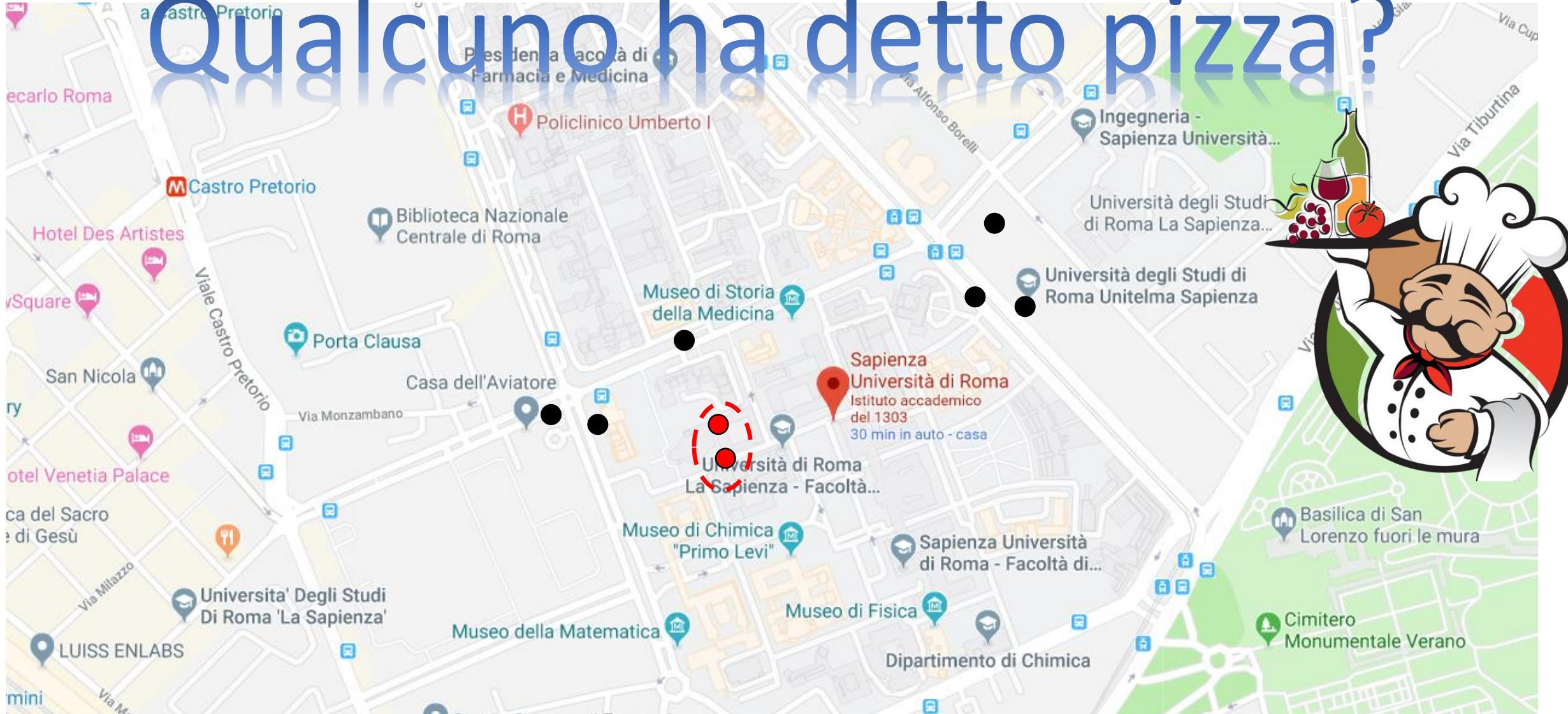
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



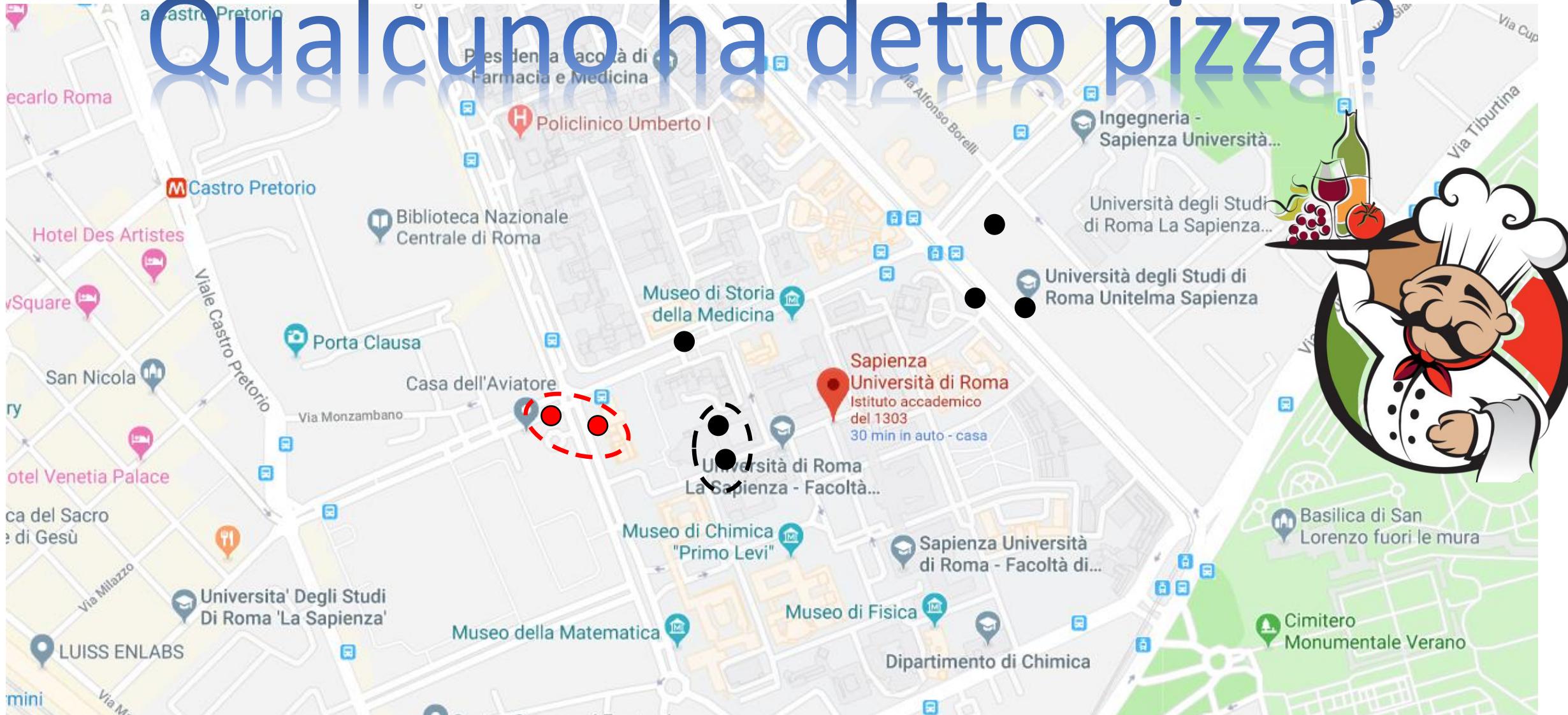
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



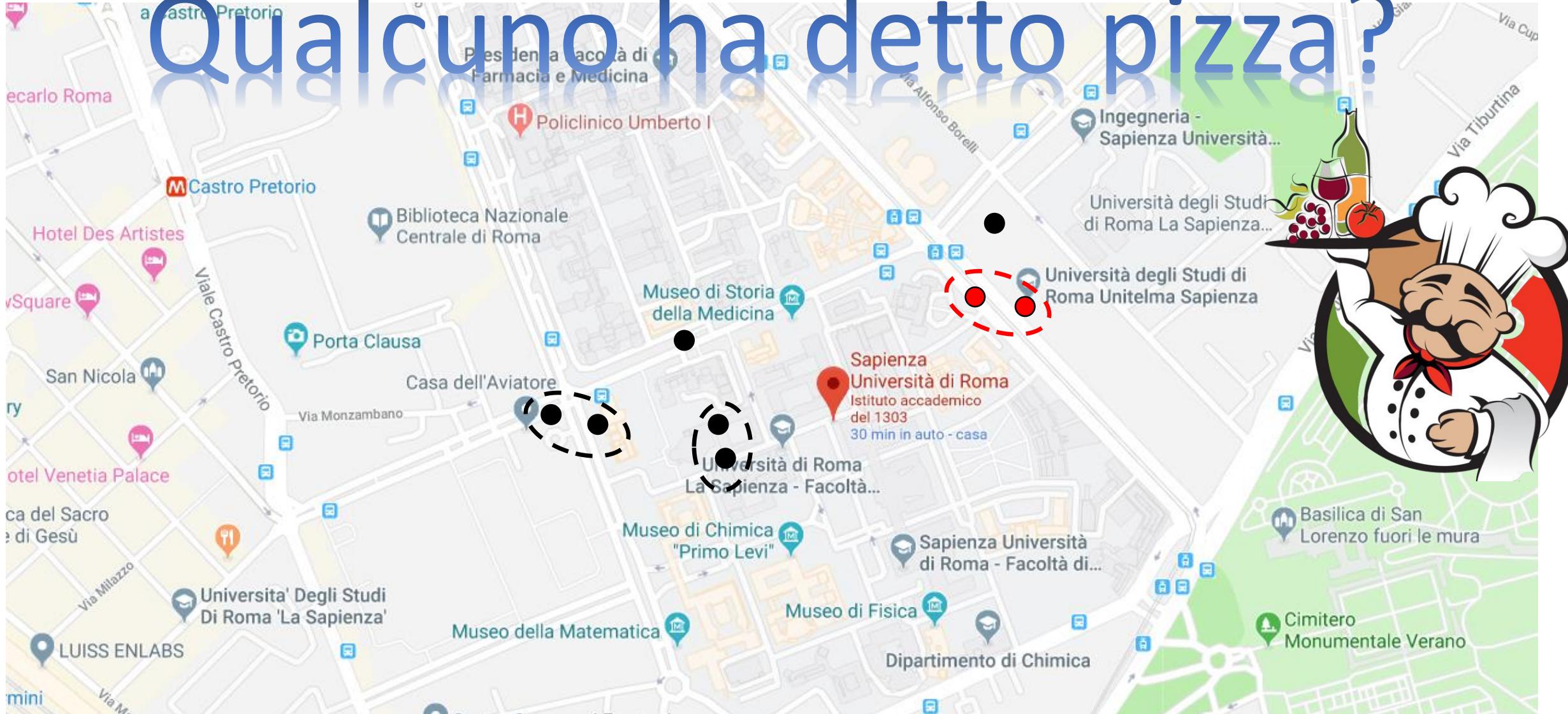
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



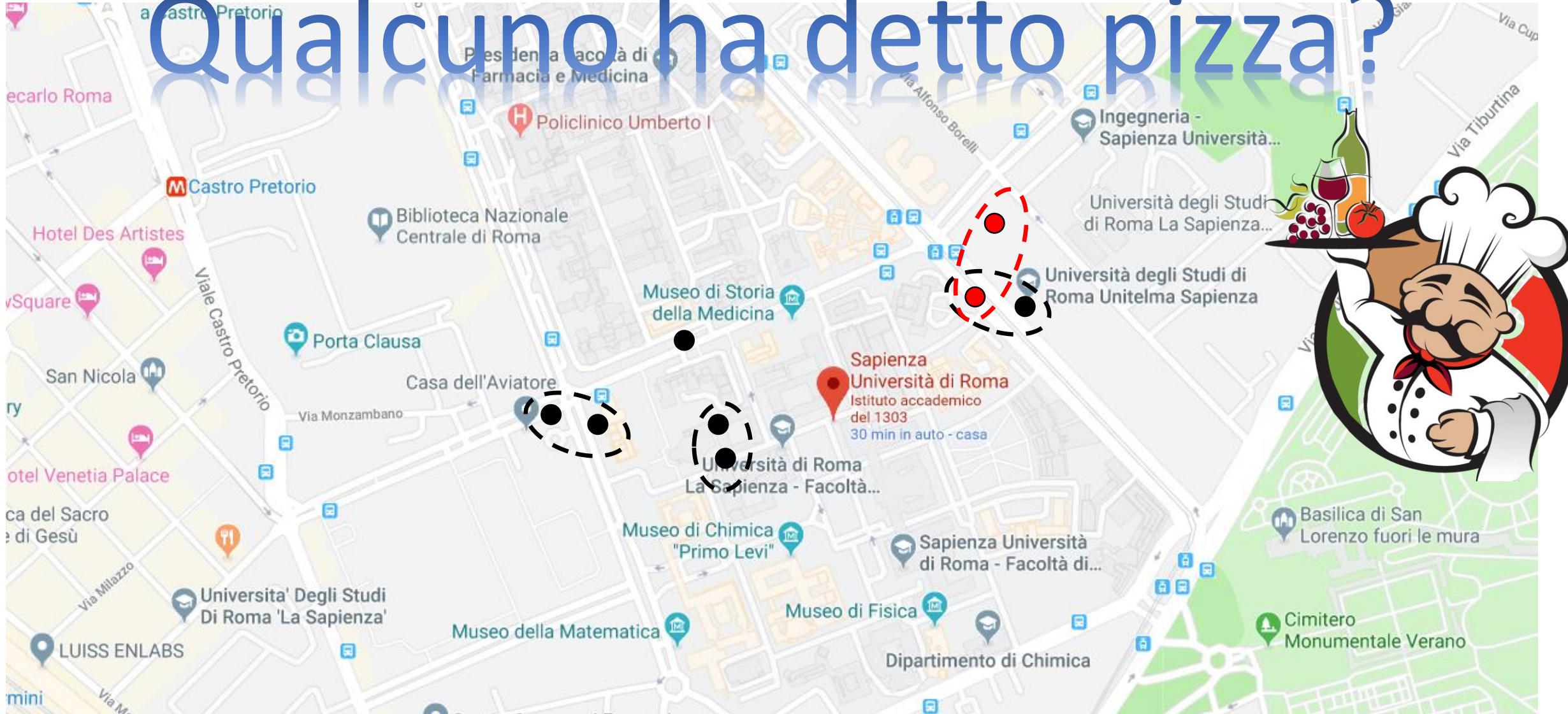
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



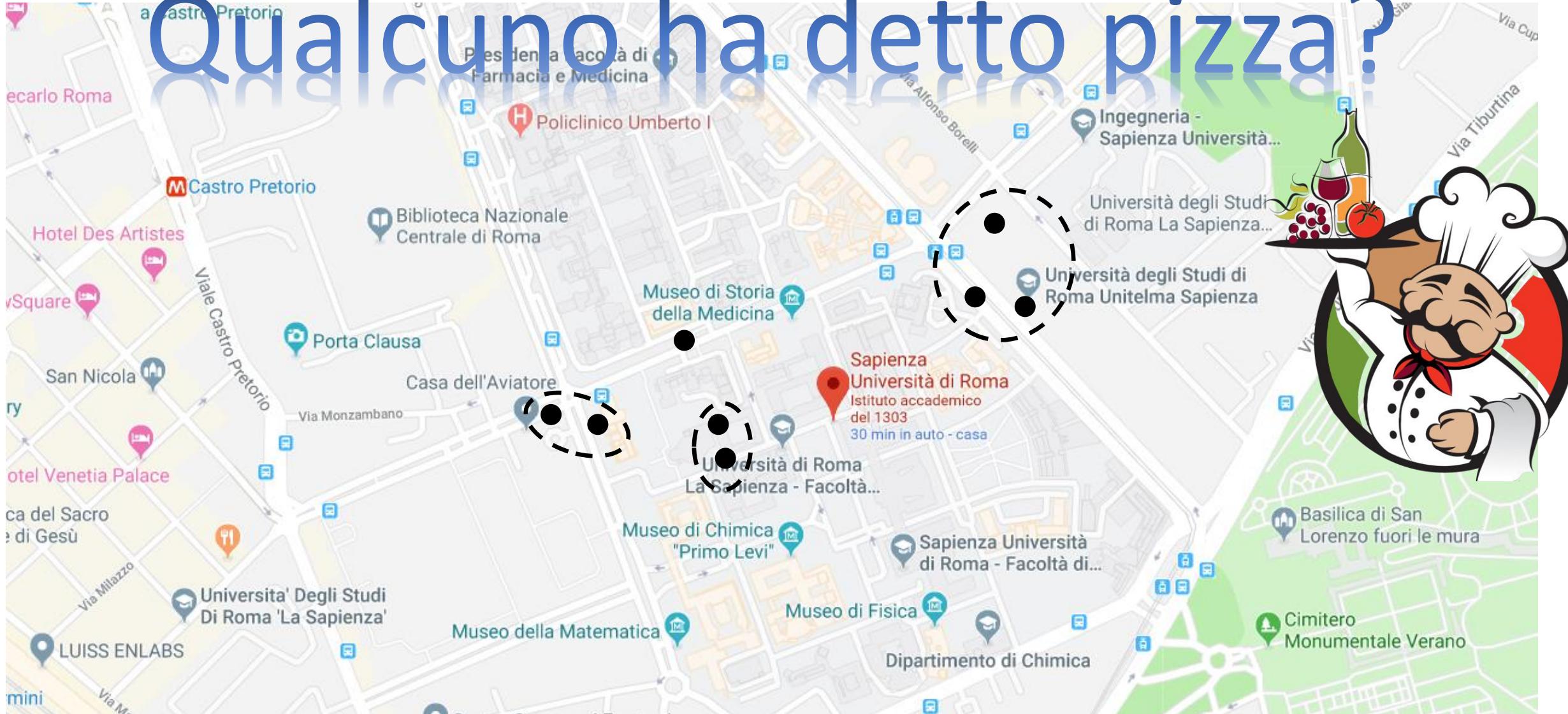
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



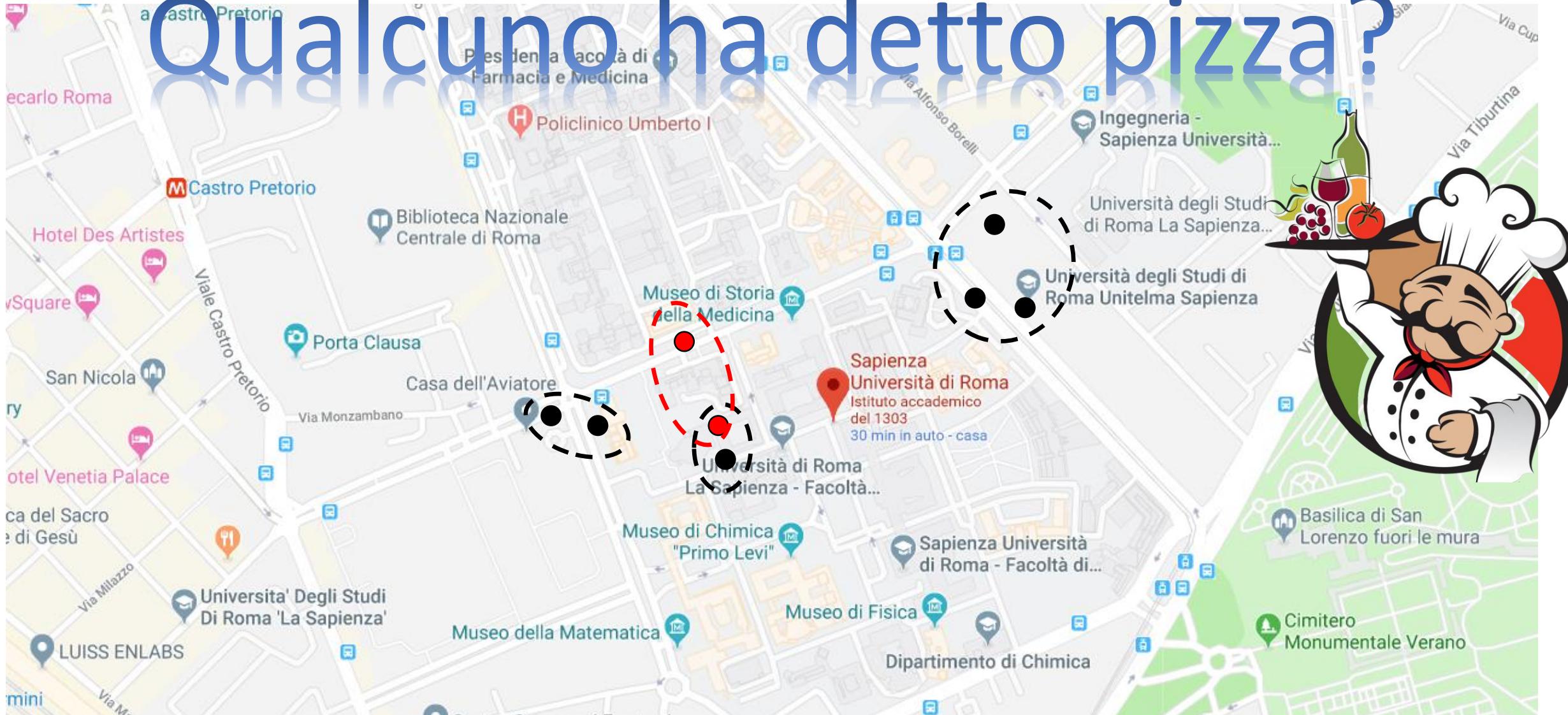
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



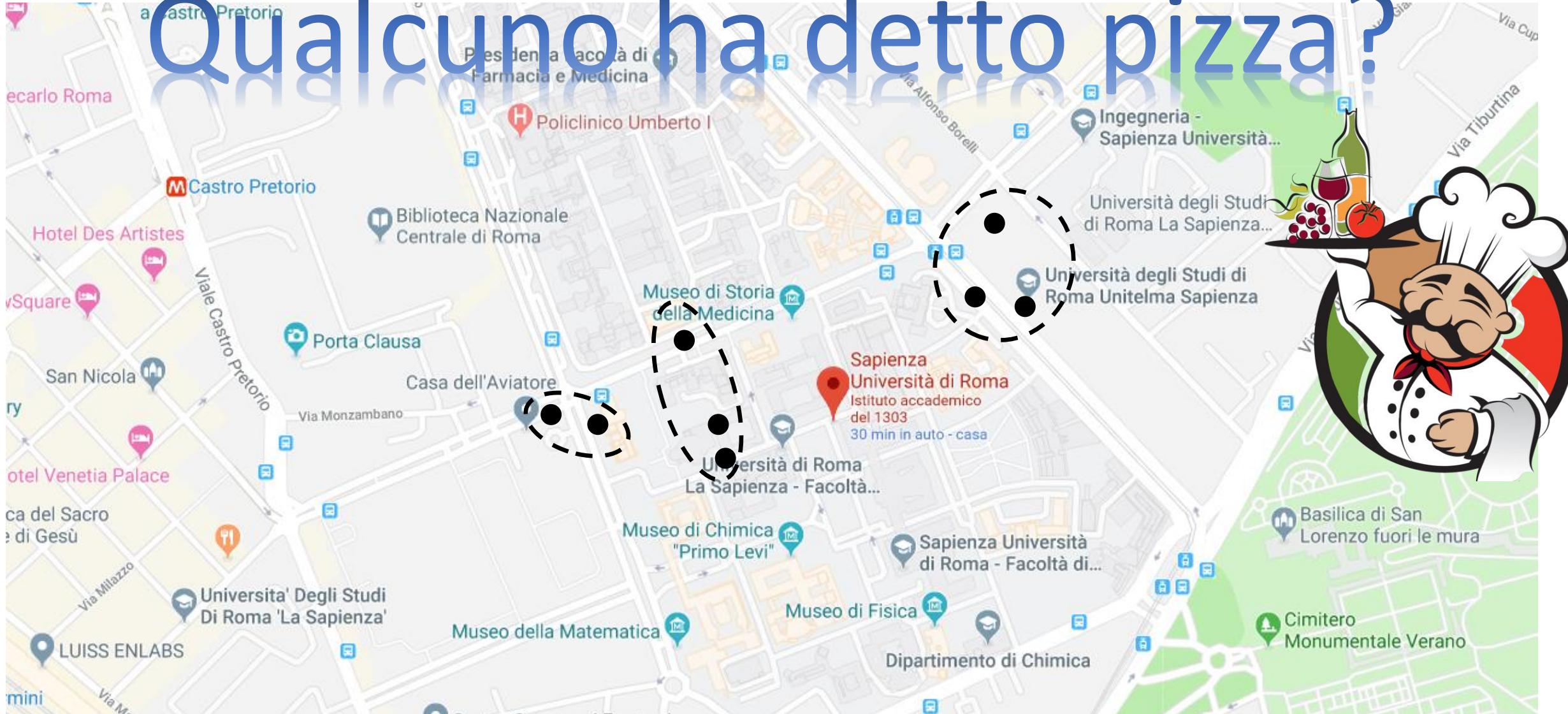
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



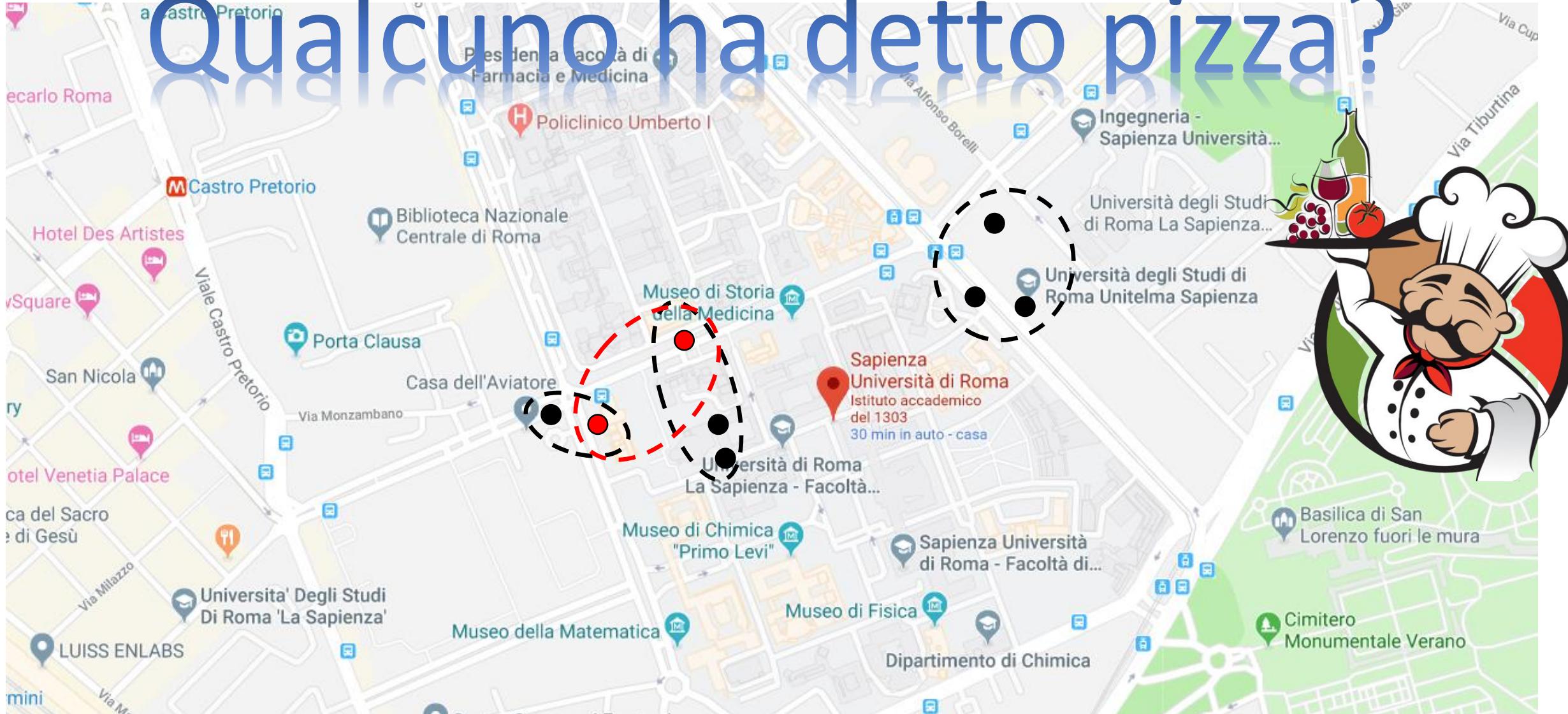
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



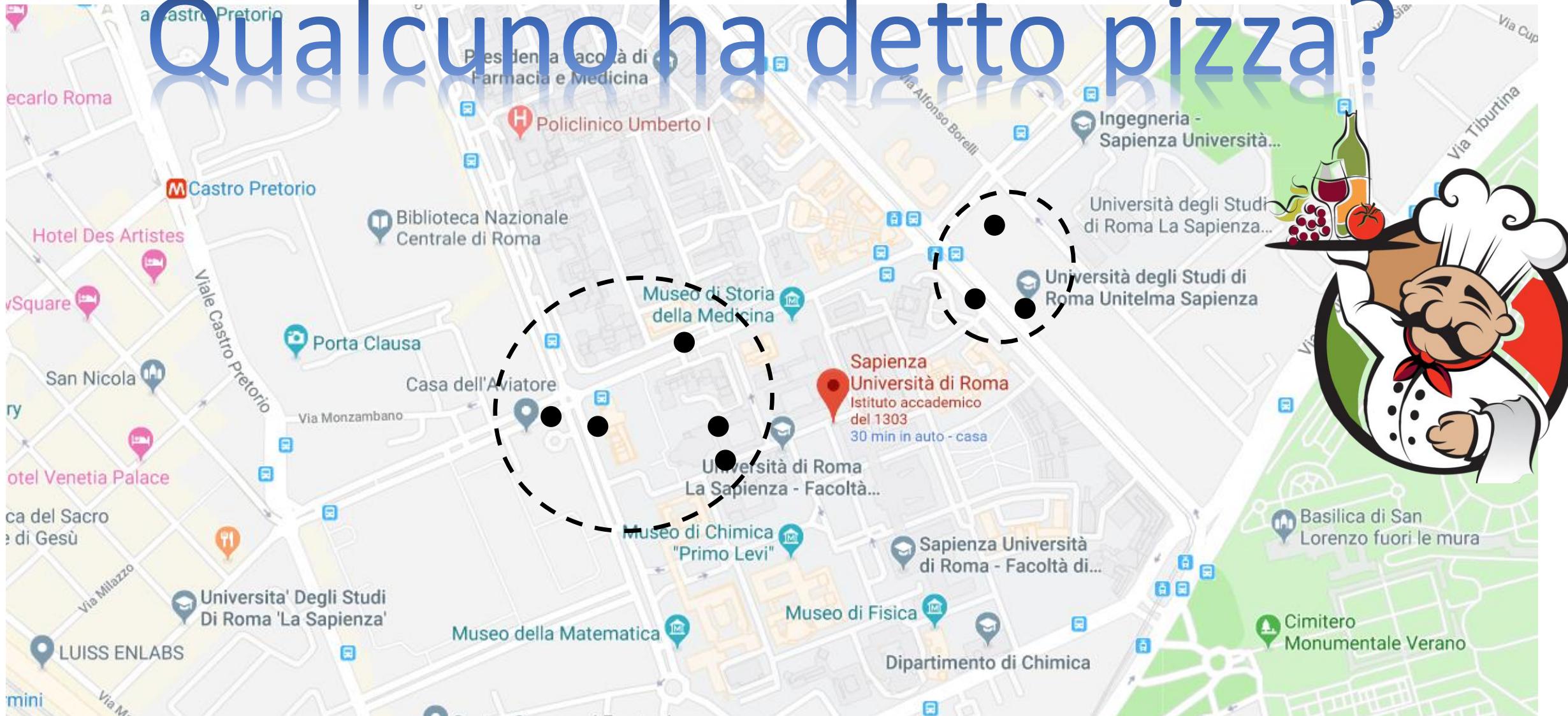
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



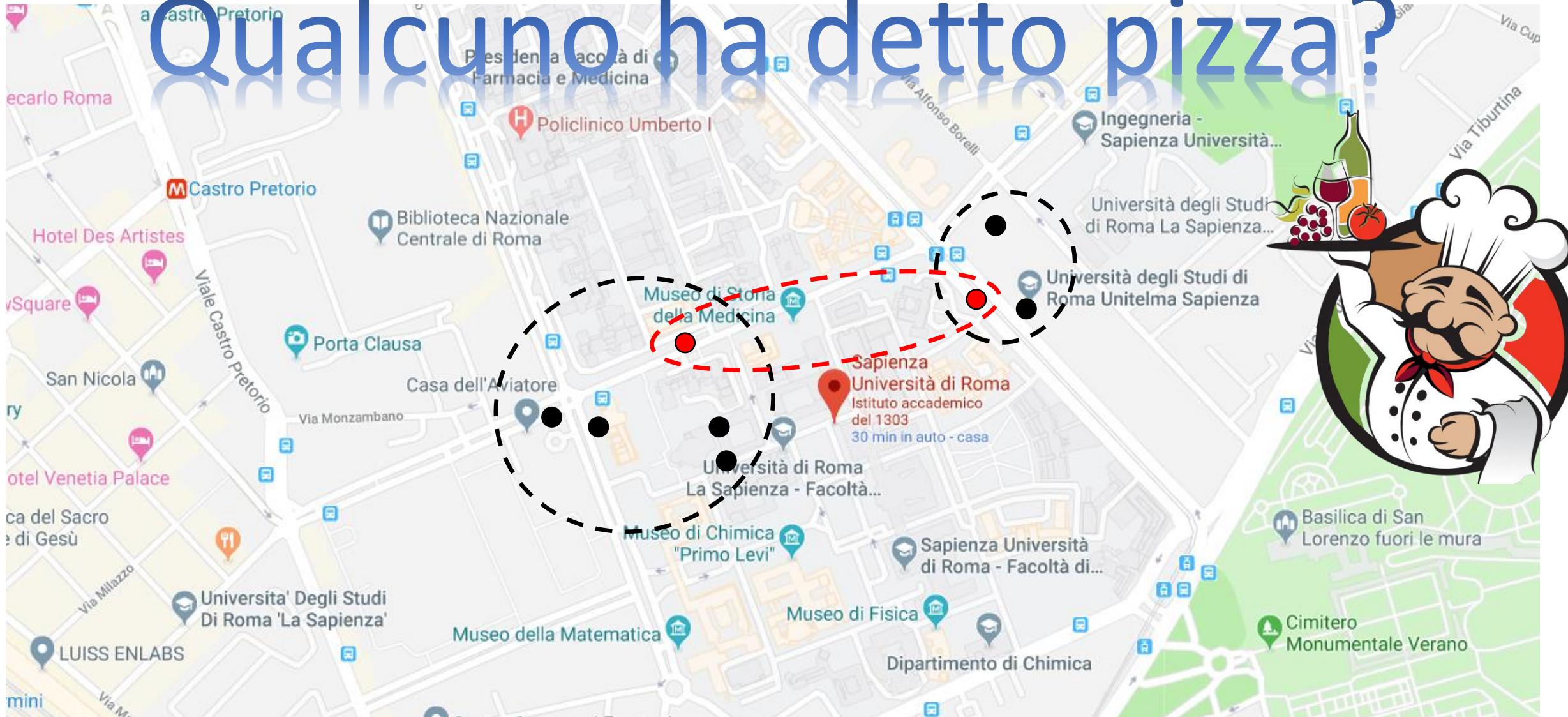
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



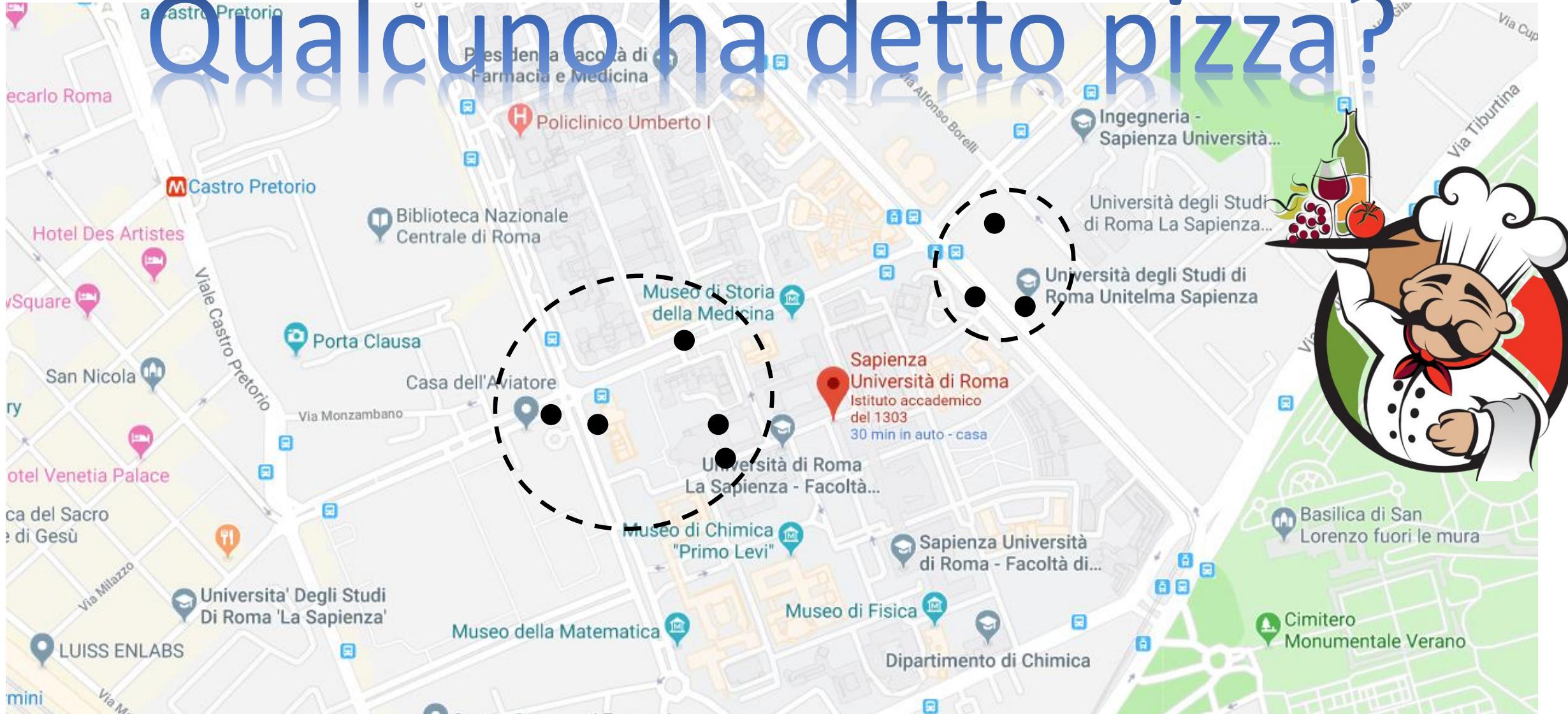
Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?



Clustering Gerarchico nella vita reale

# Qualcuno ha detto pizza?

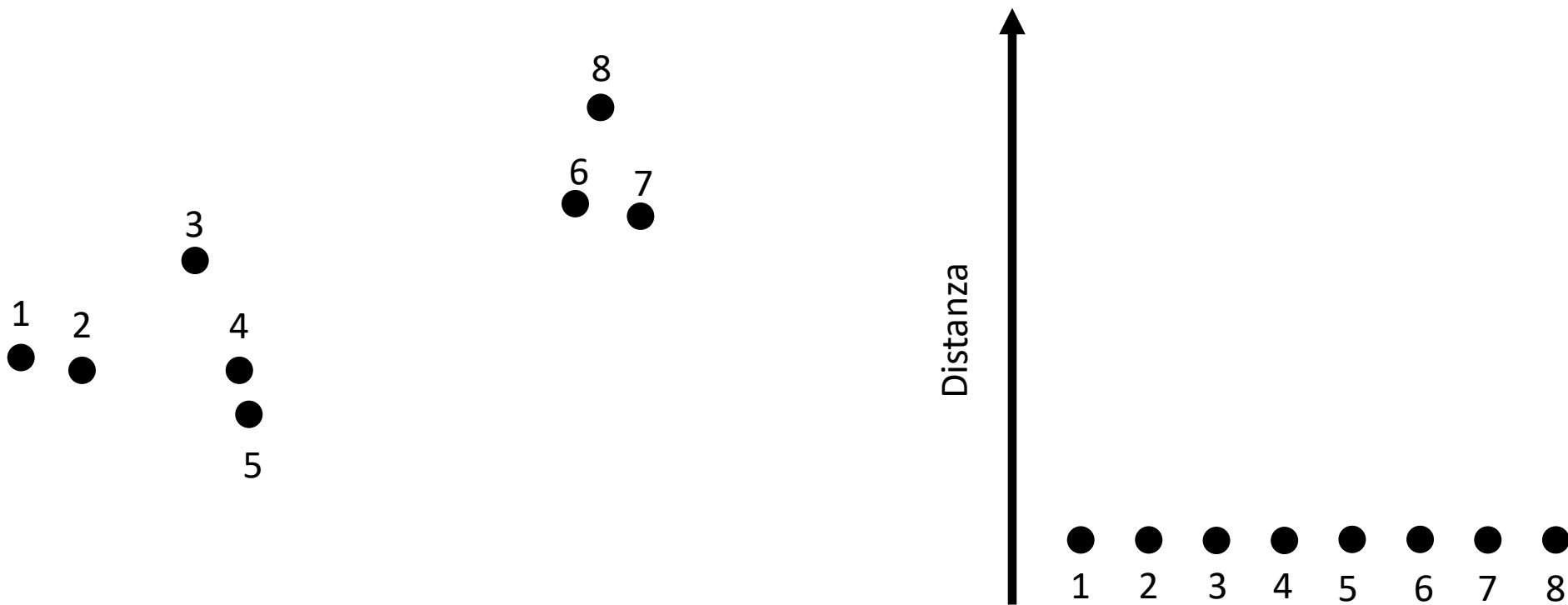


# Vantaggi del Clustering gerarchico

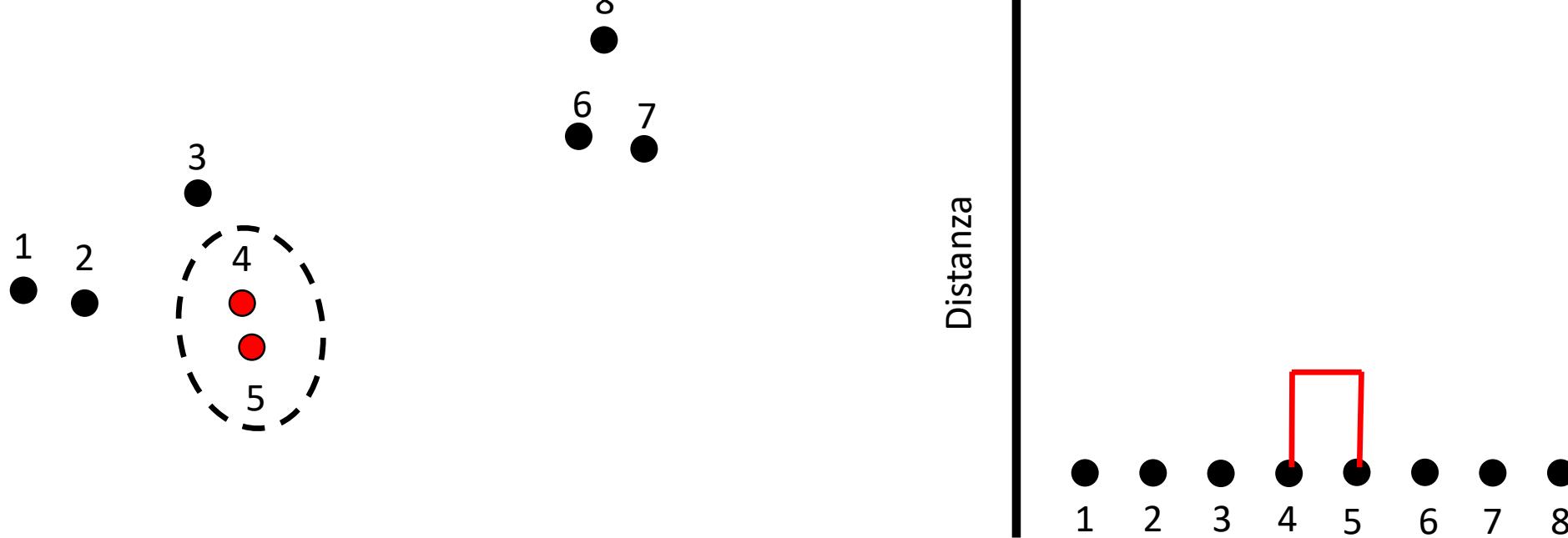
Non è necessario assumere un numero fissato di cluster ( $K$ ):

Qualsiasi numero desiderato di cluster può essere ottenuto con un 'taglio' del dendrogramma al livello appropriato

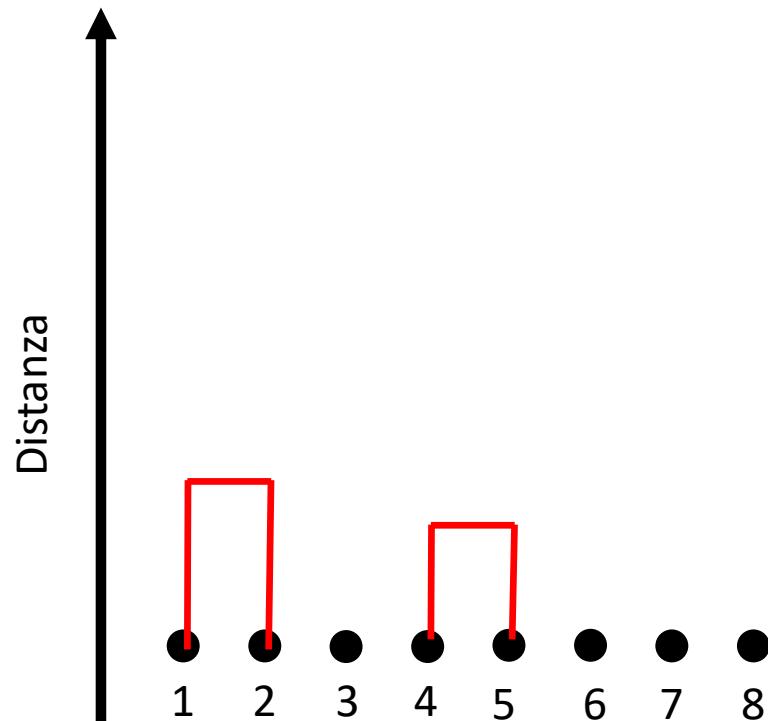
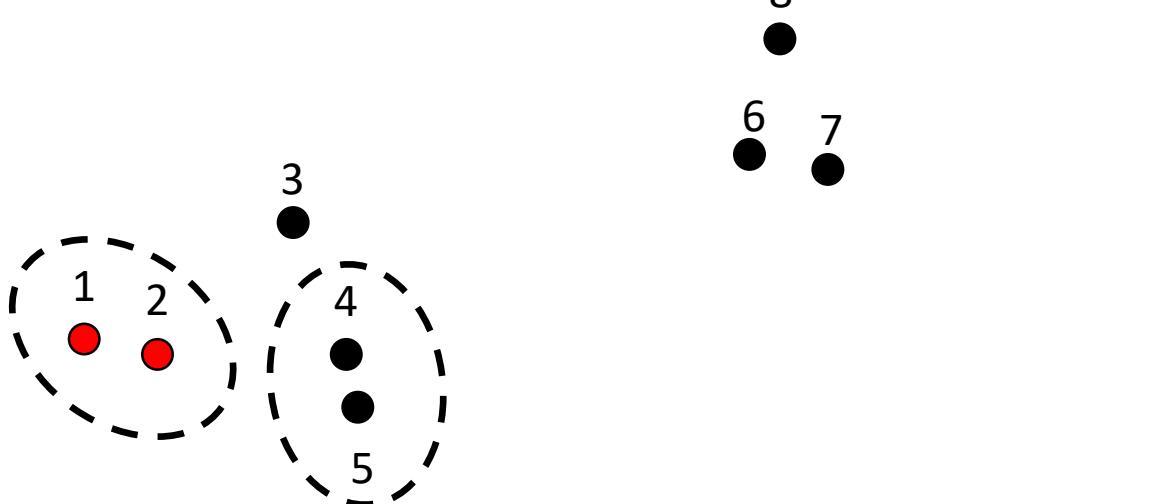
# Clustering gerarchico – dendrogramma



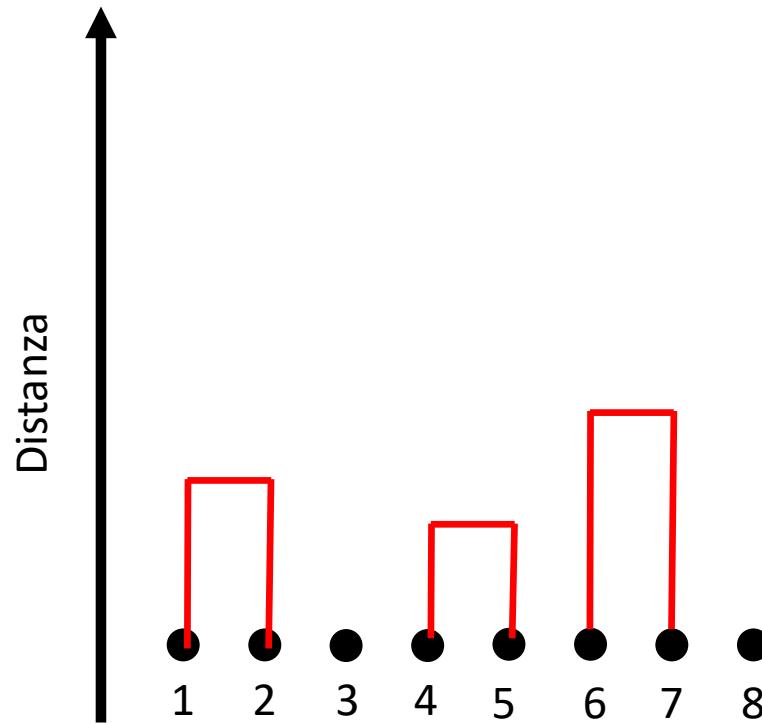
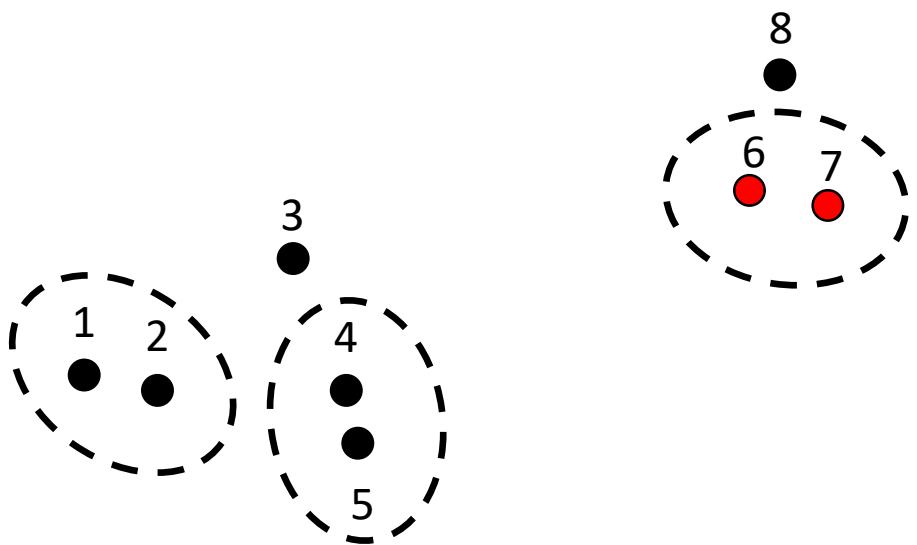
# Clustering gerarchico – dendrogramma



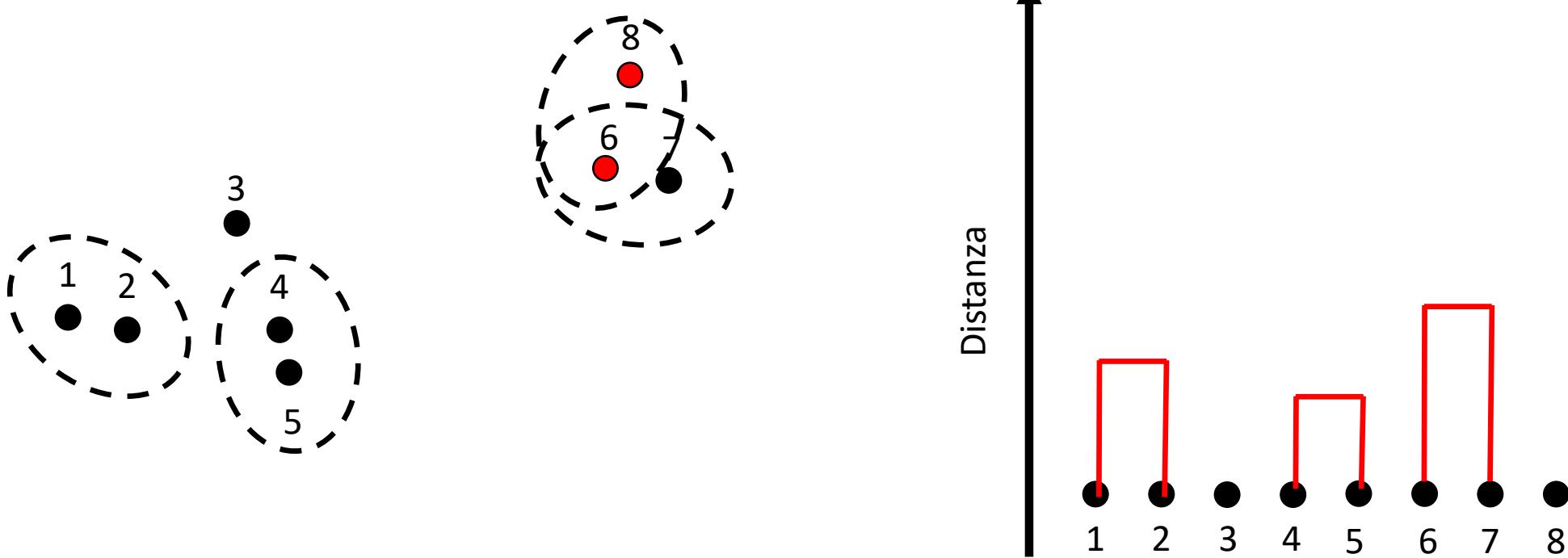
# Clustering gerarchico – dendrogramma



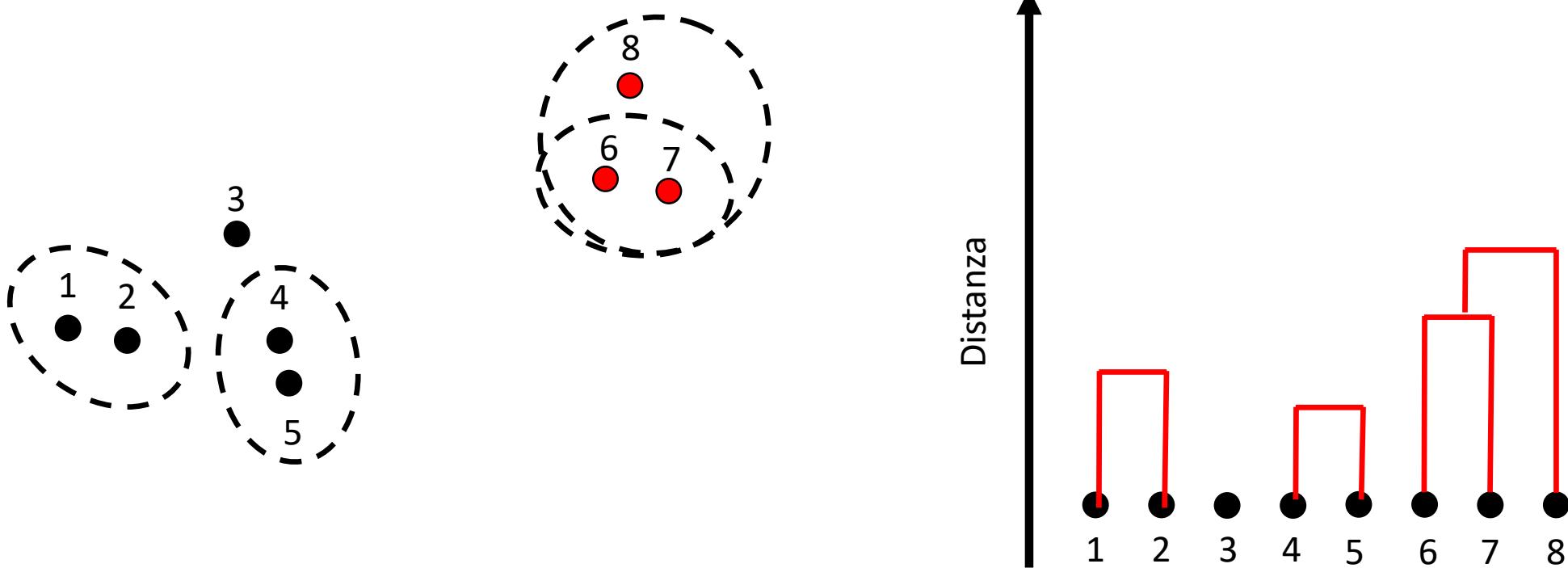
# Clustering gerarchico – dendrogramma



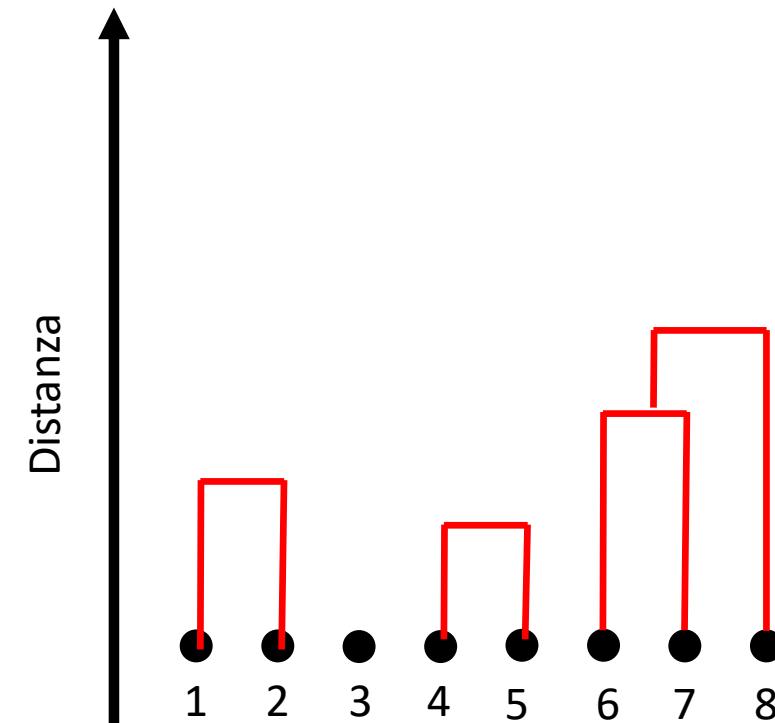
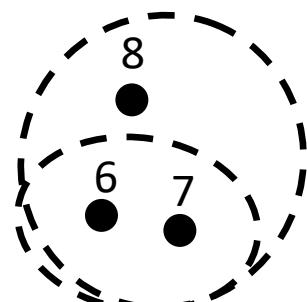
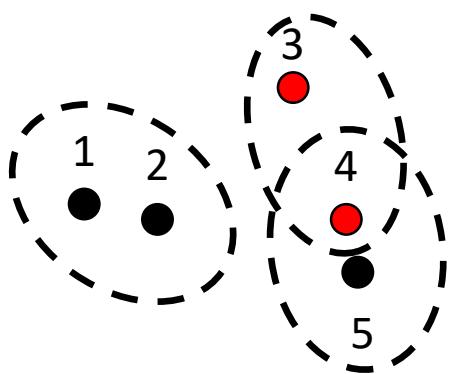
# Clustering gerarchico – dendrogramma



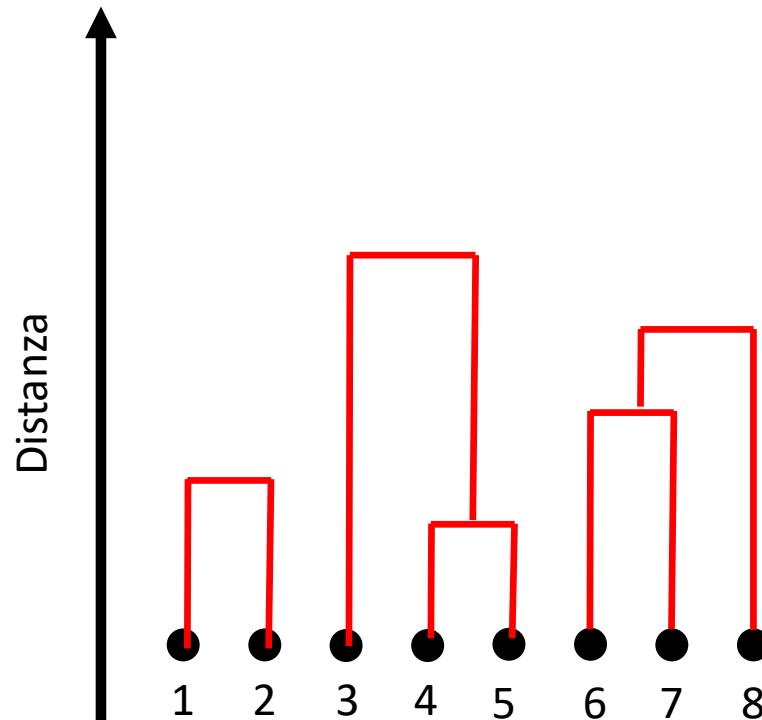
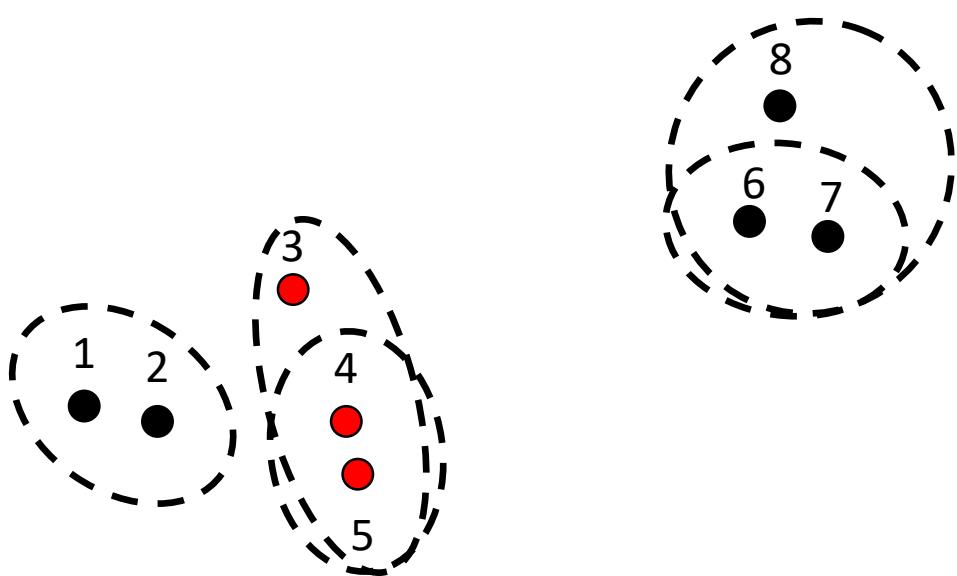
# Clustering gerarchico – dendrogramma



# Clustering gerarchico – dendrogramma



# Clustering gerarchico – dendrogramma

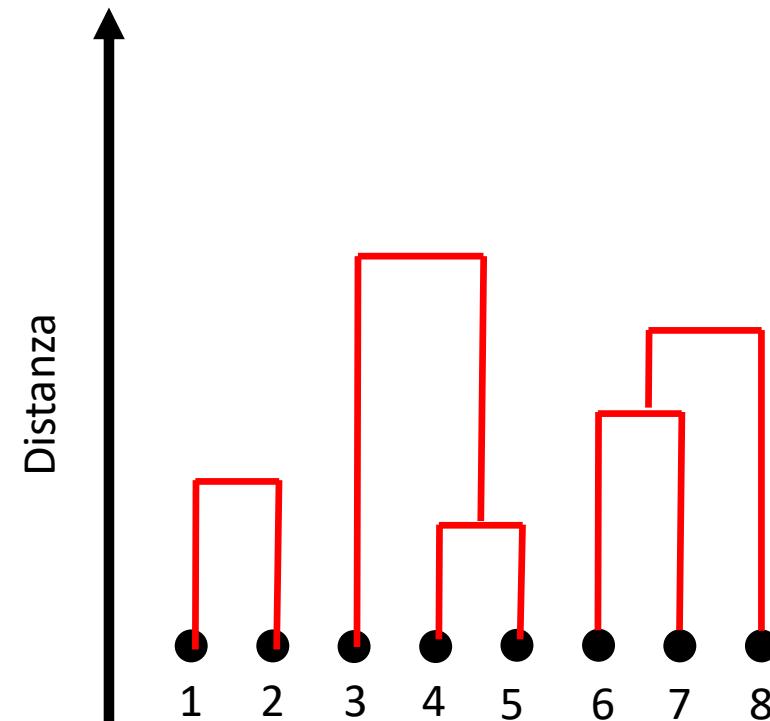
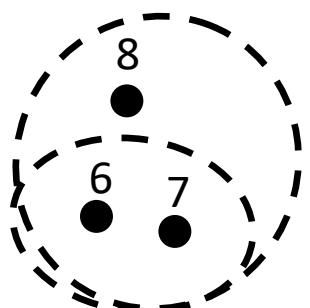
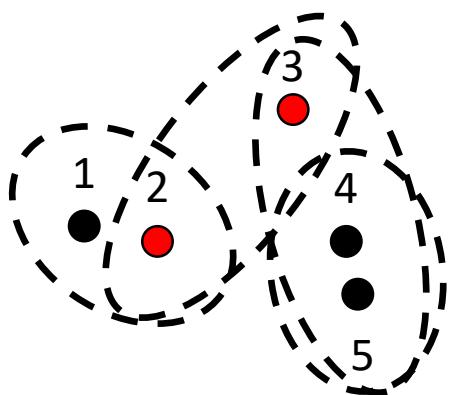


DIPARTIMENTO DI METODI E MODELLI  
L'ECONOMIA IL TERRITORIO E LA FINANZA  
MEMOTEF

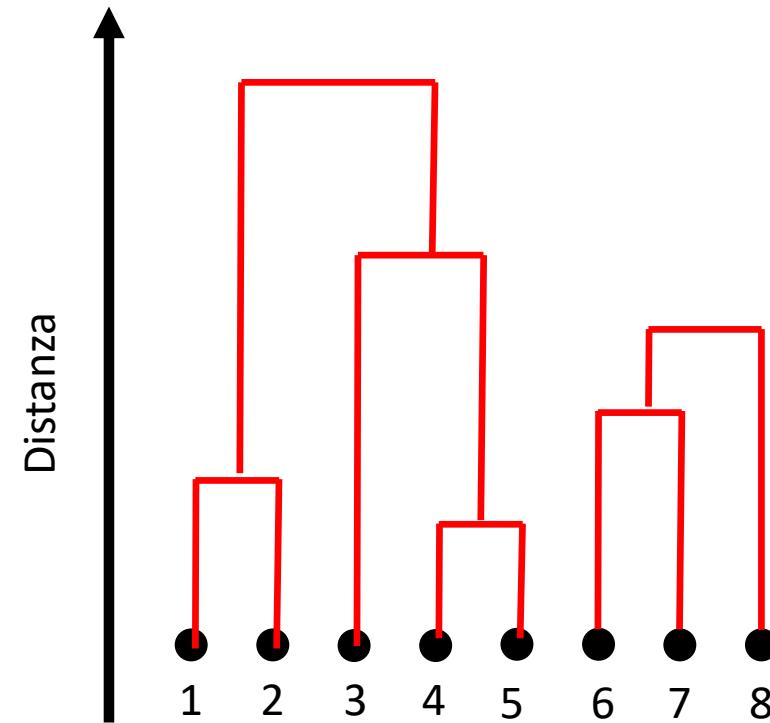
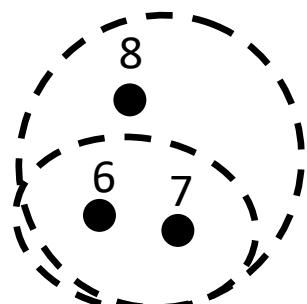
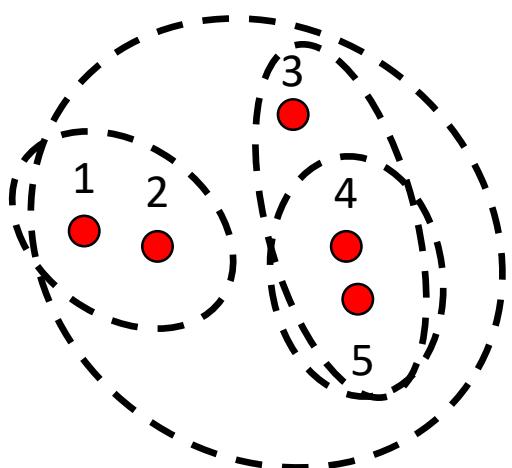


SAPIENZA  
UNIVERSITÀ DI ROMA

# Clustering gerarchico – dendrogramma



# Clustering gerarchico – dendrogramma

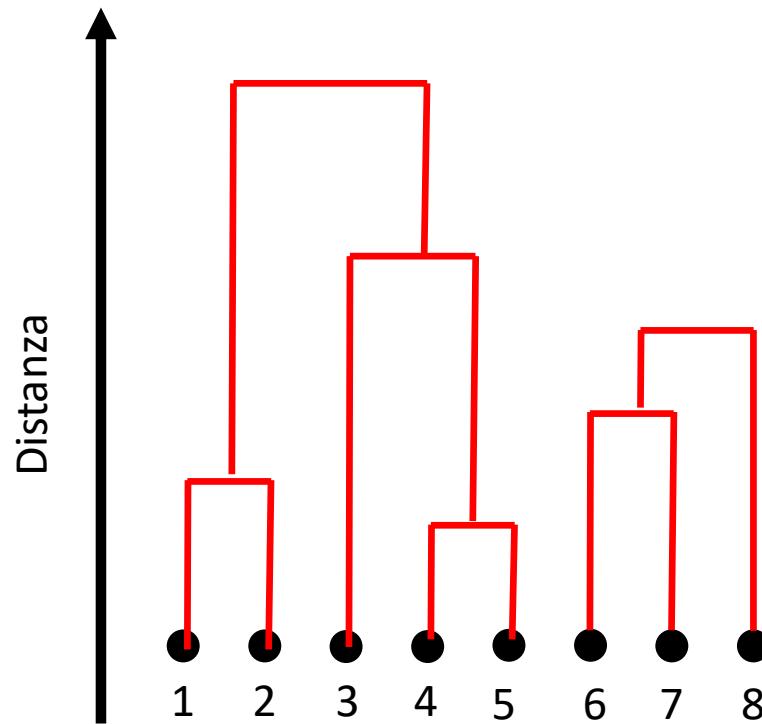
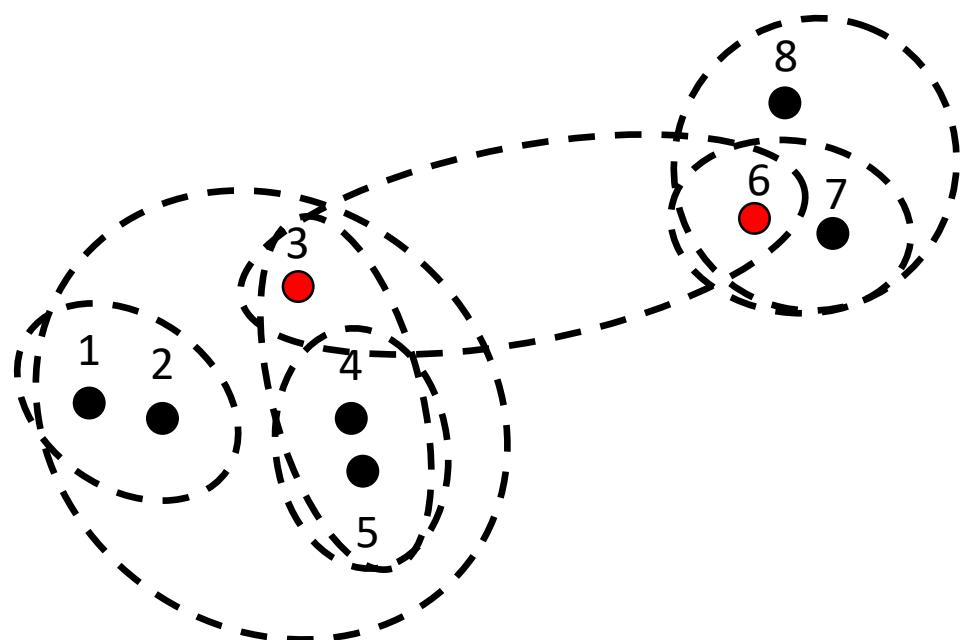


DIPARTIMENTO DI METODI E MODELLI  
L'ECONOMIA IL TERRITORIO E LA FINANZA  
MEMOTEF



SAPIENZA  
UNIVERSITÀ DI ROMA

# Clustering gerarchico – dendrogramma

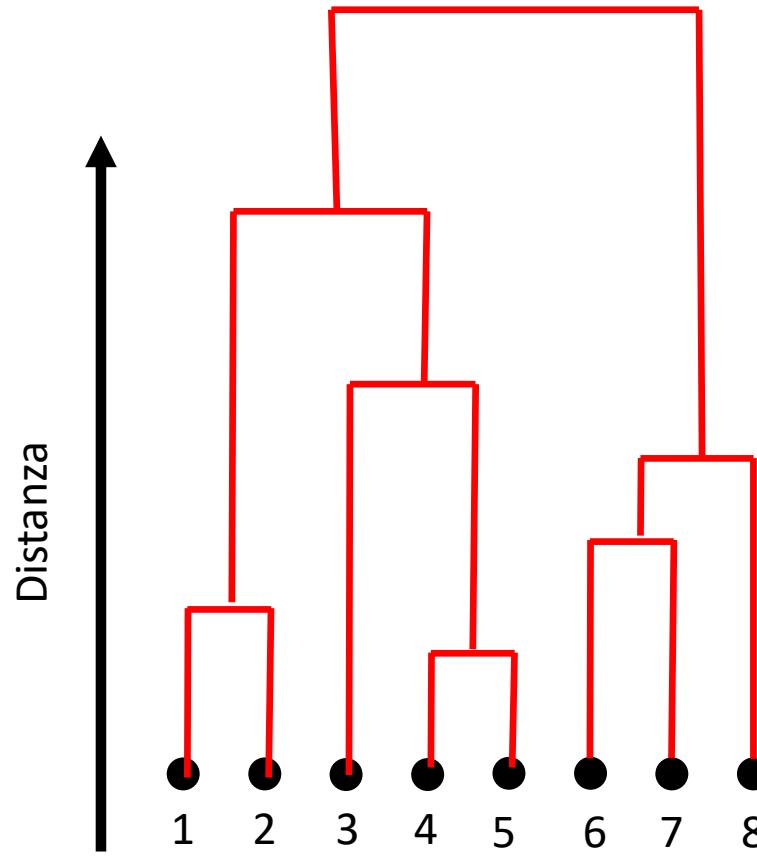
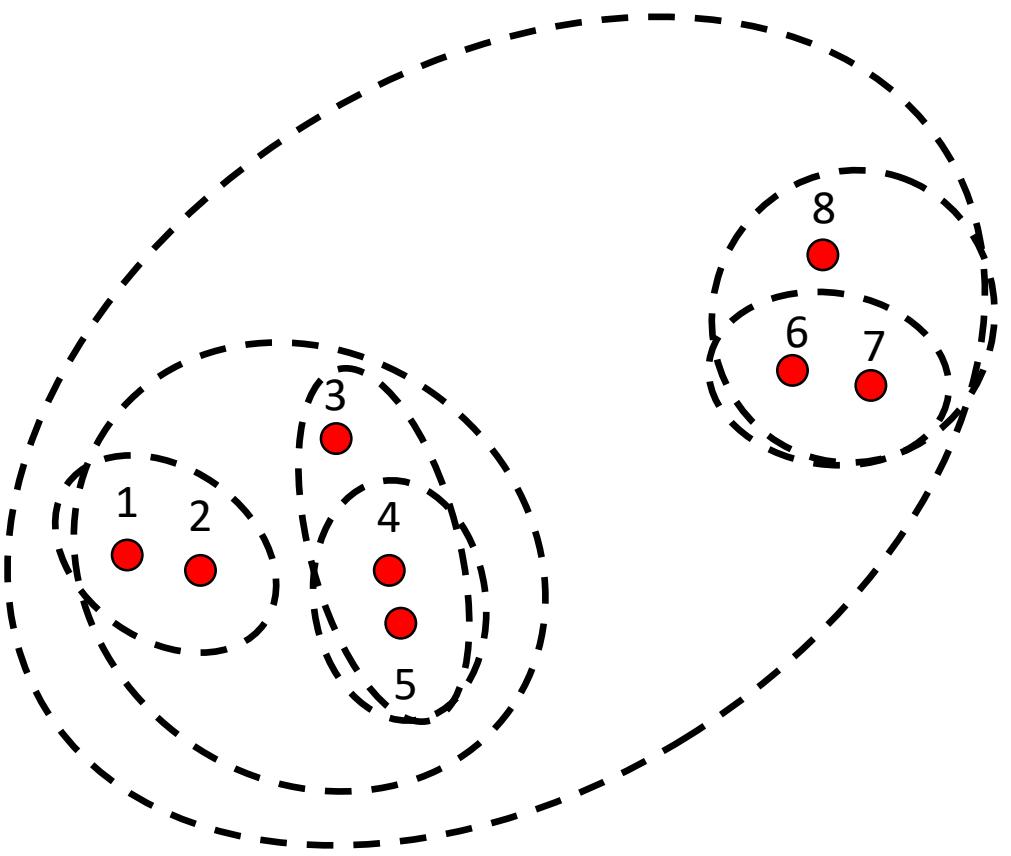


DIPARTIMENTO DI METODI E MODELLI  
L'ECONOMIA IL TERRITORIO E LA FINANZA  
MEMOTEF

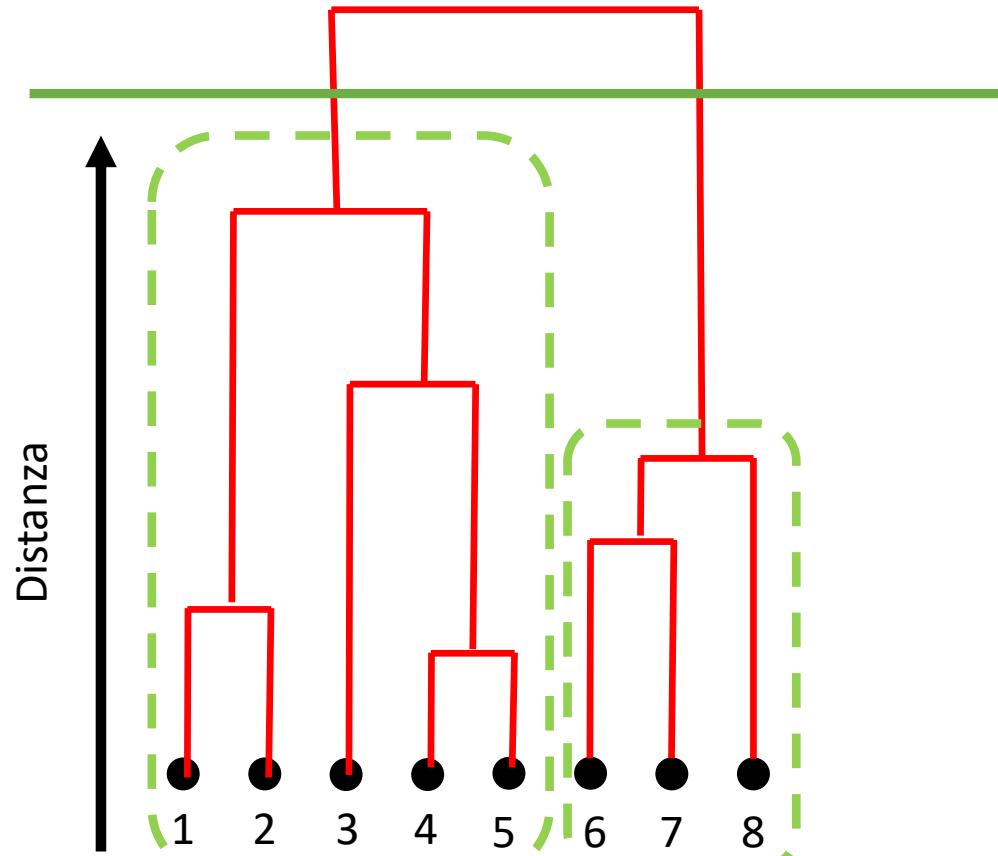
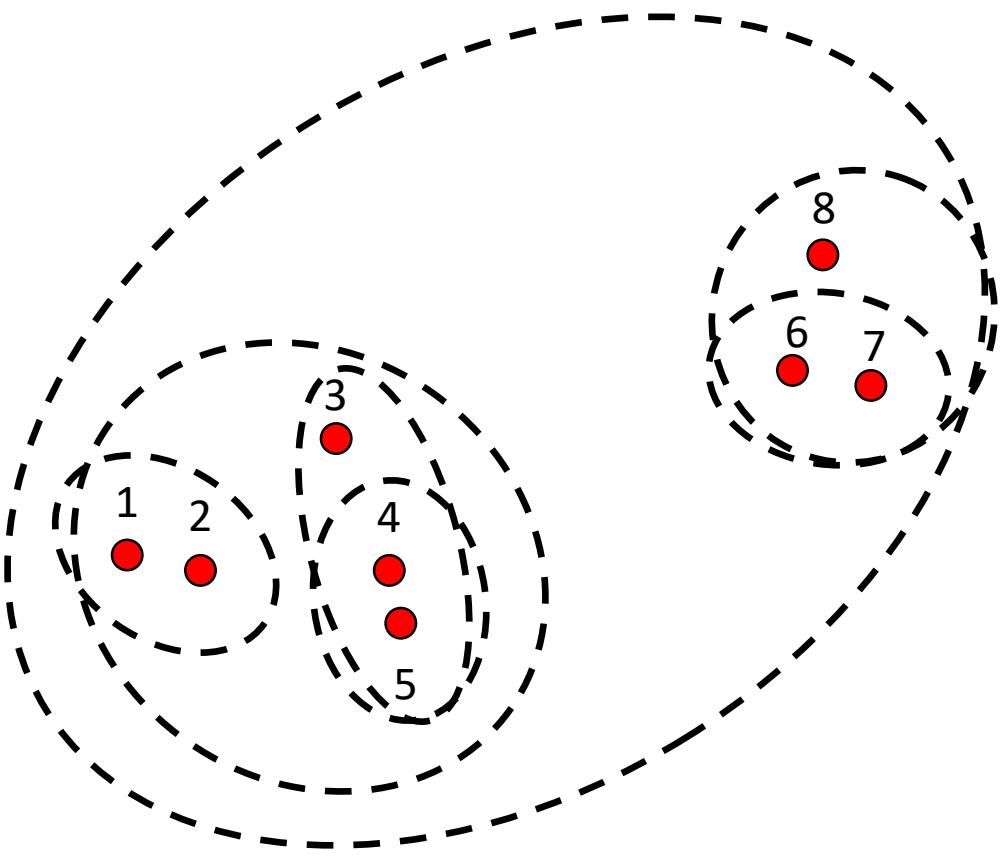


SAPIENZA  
UNIVERSITÀ DI ROMA

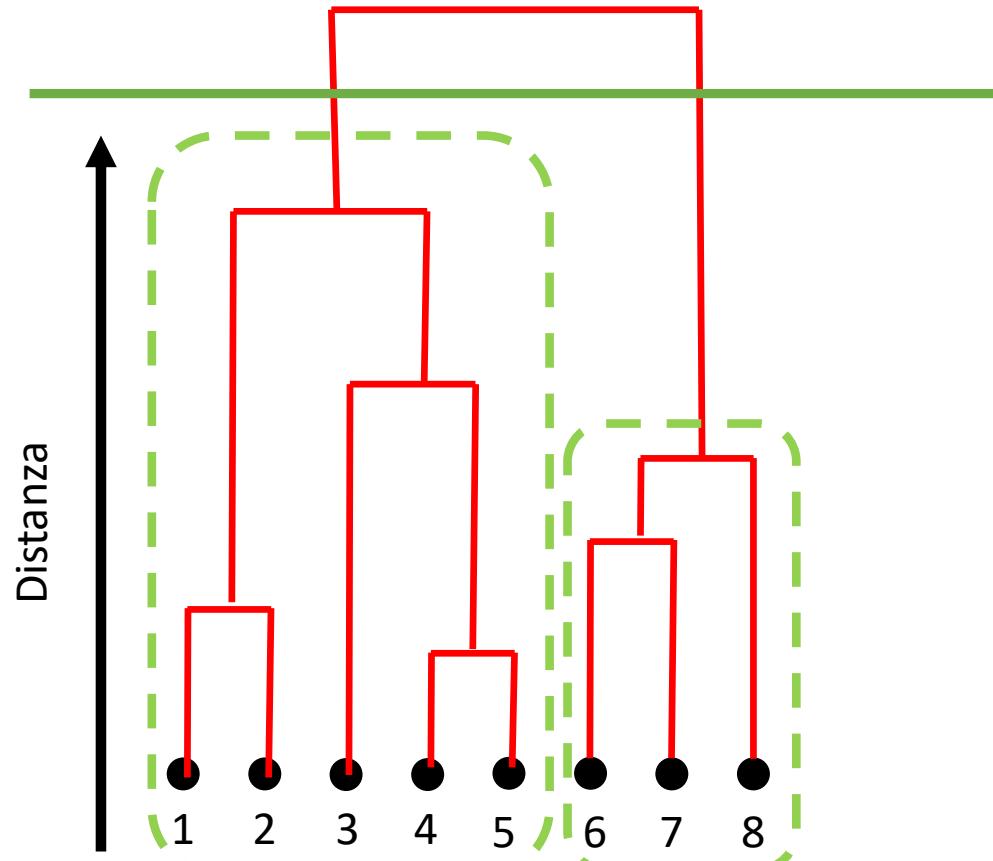
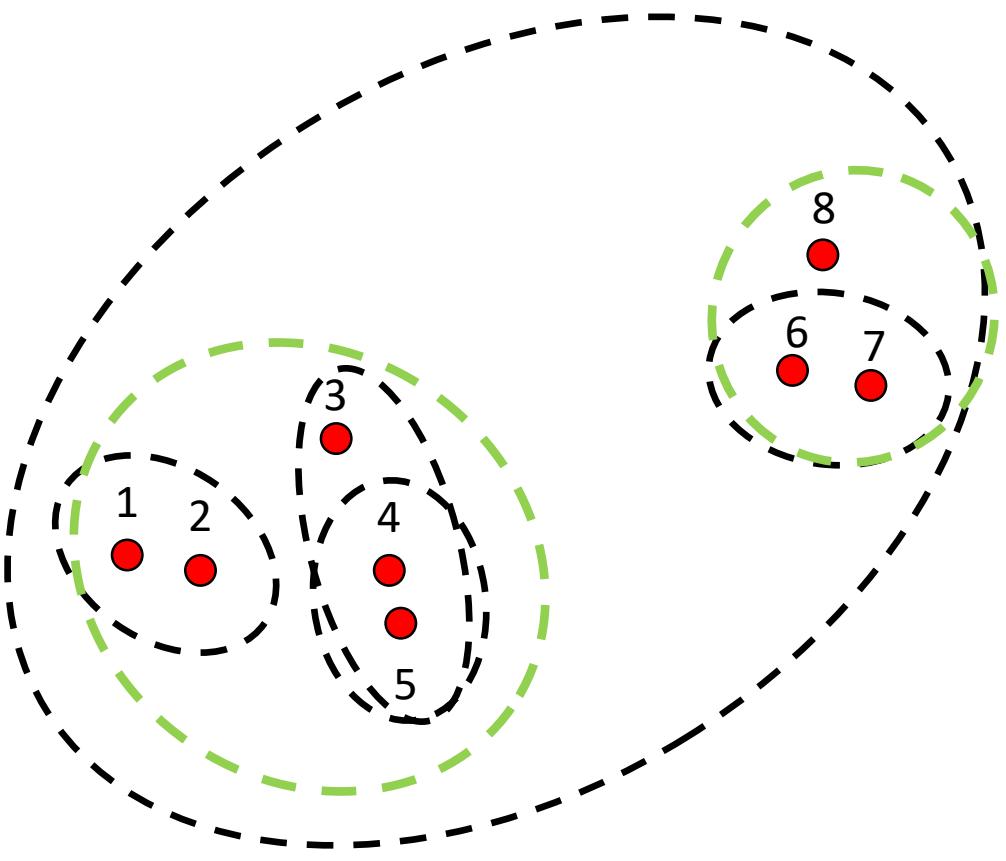
# Clustering gerarchico – dendrogramma



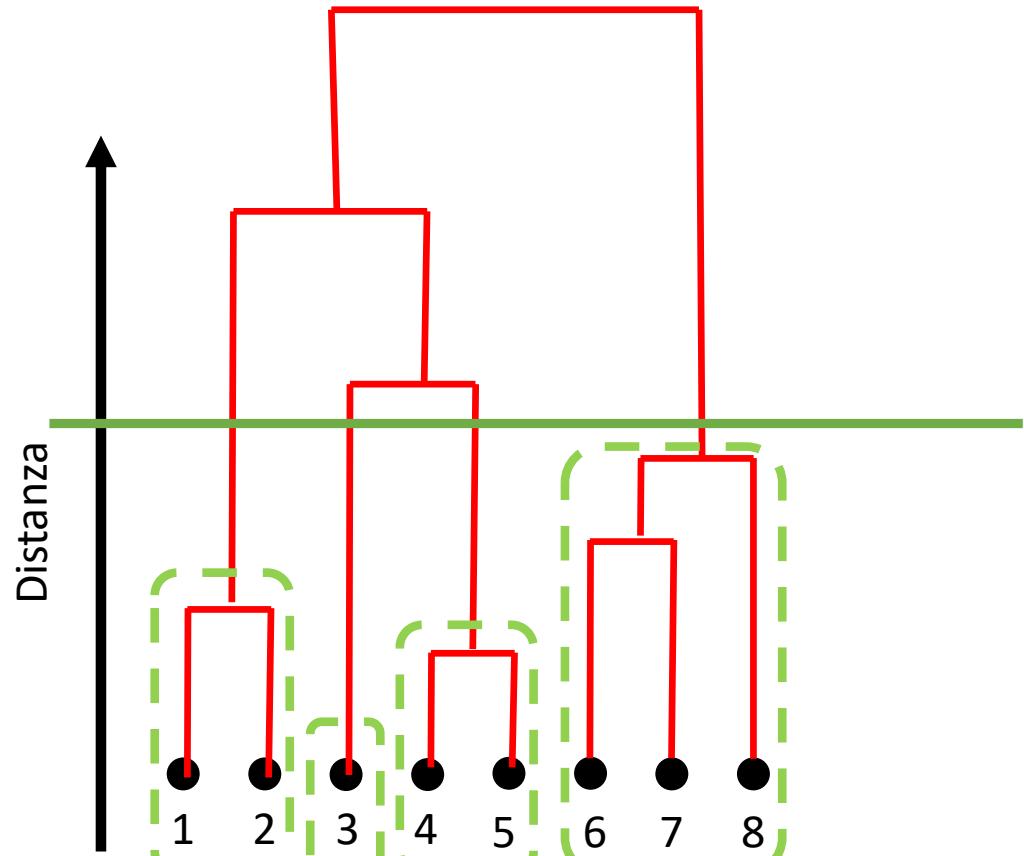
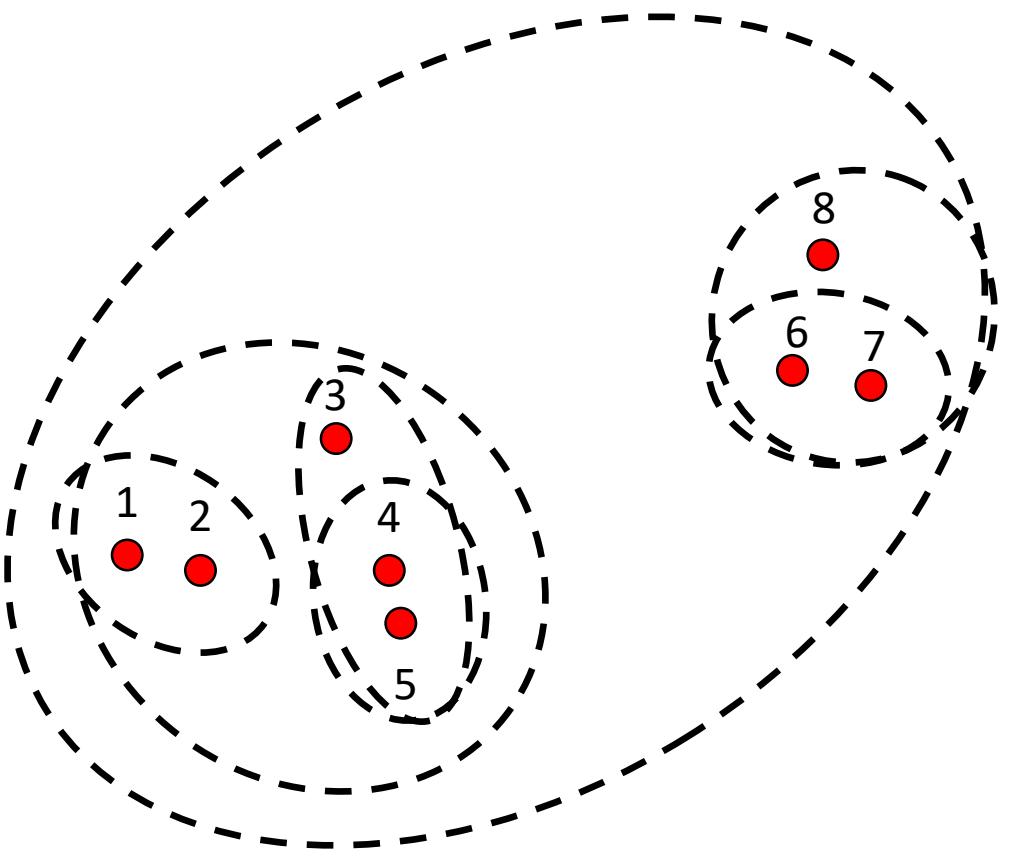
# Clustering gerarchico – dendrogramma



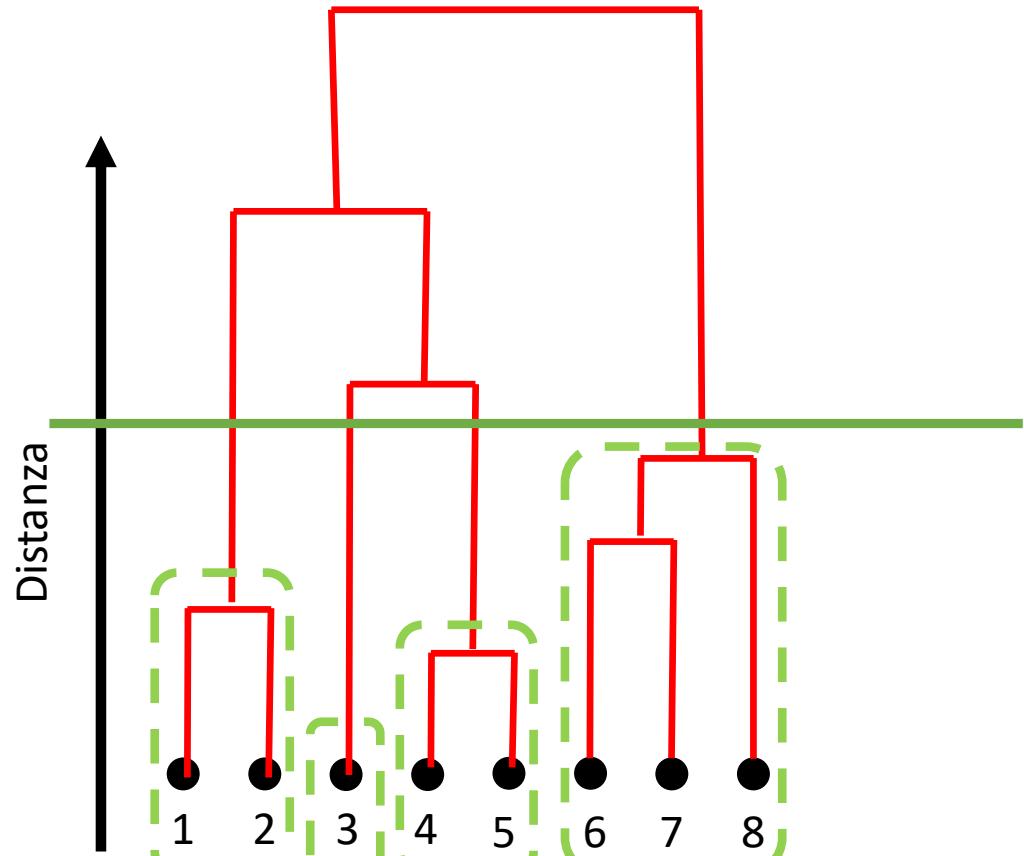
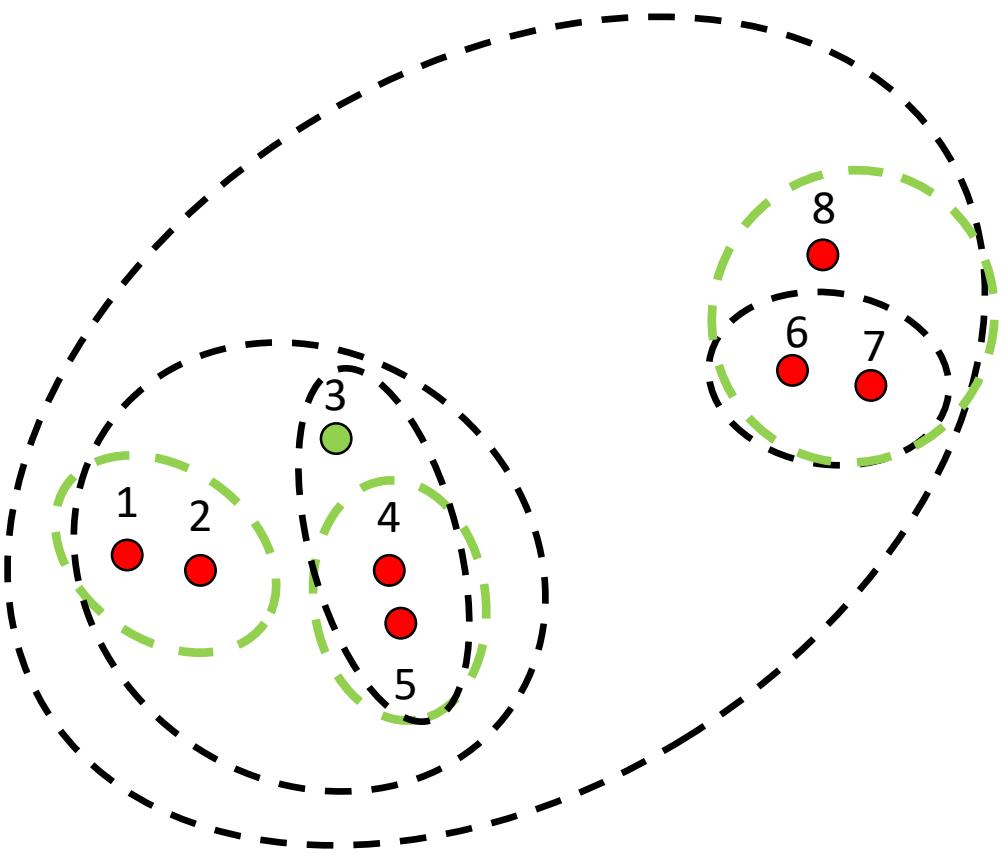
# Clustering gerarchico – dendrogramma



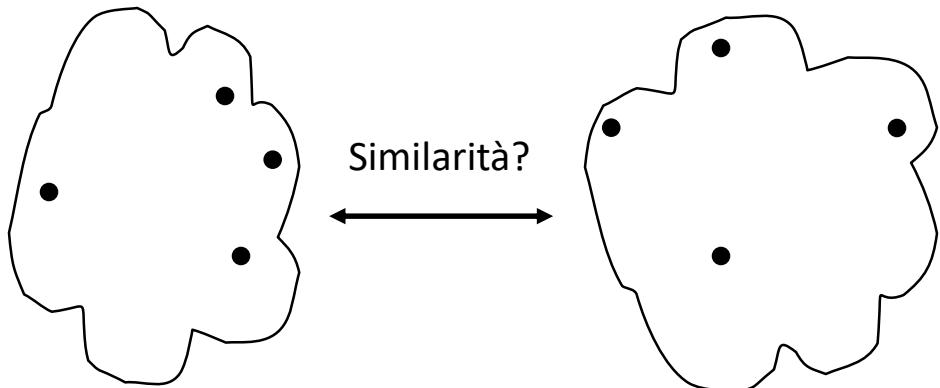
# Clustering gerarchico – dendrogramma



# Clustering gerarchico – dendrogramma



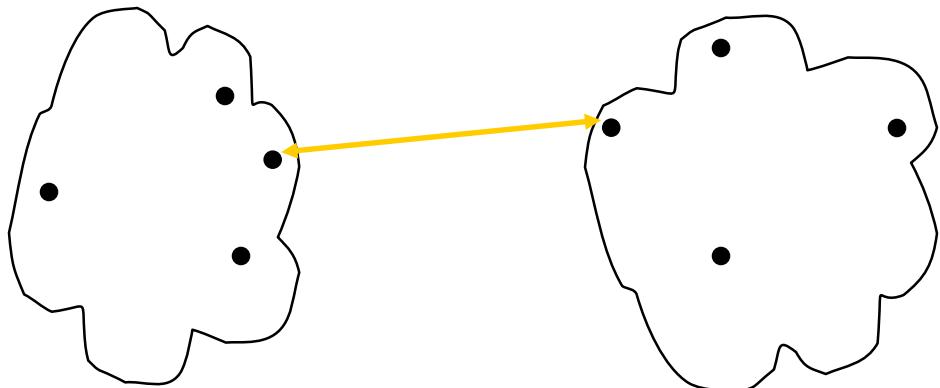
# Come definire la similarità tra gruppi



- MIN
  - MAX
  - Media del gruppo
  - Distanza tra i centroidi
  - Altri metodi basati su una funzione obiettivo
    - Ward's Method → SSE
- Matrice di prossimità

	C1	C2	C3	C4	C5	...
C1						
C2						
C3						
C4						
C5						
.						

# Come definire la similarità tra gruppi

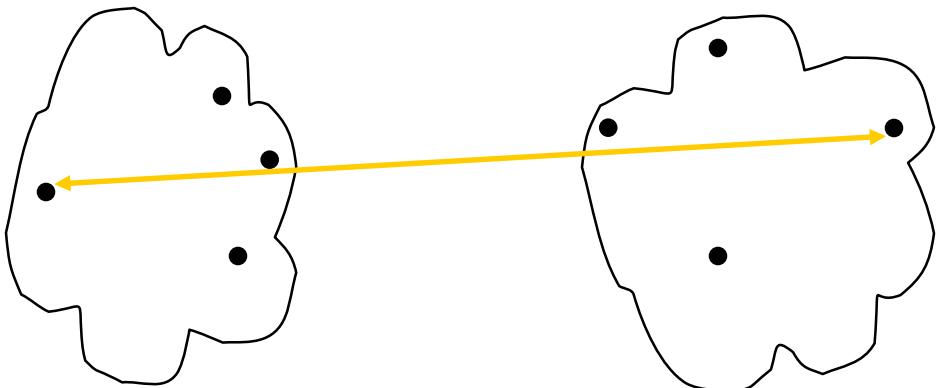


- MIN
- MAX
- Media del gruppo
- Distanza tra i centroidi
- Altri metodi basati su una funzione obiettivo
  - Ward's Method → SSE

	C1	C2	C3	C4	C5	...
C1						
C2						
C3						
C4						
C5						
.	.	.	.	.	.	.

Matrice di prossimità

# Come definire la similarità tra gruppi

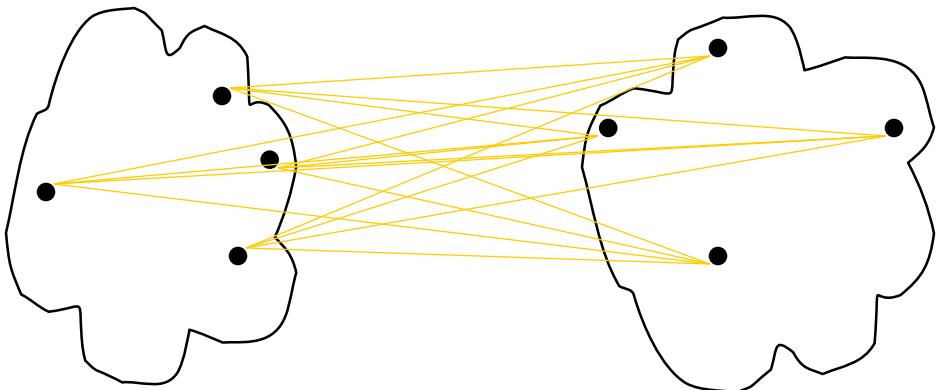


- MIN
- MAX
- Media del gruppo
- Distanza tra i centroidi
- Altri metodi basati su una funzione obiettivo
  - Ward's Method → SSE

	C1	C2	C3	C4	C5	...
C1						
C2						
C3						
C4						
C5						
.						

Matrice di prossimità

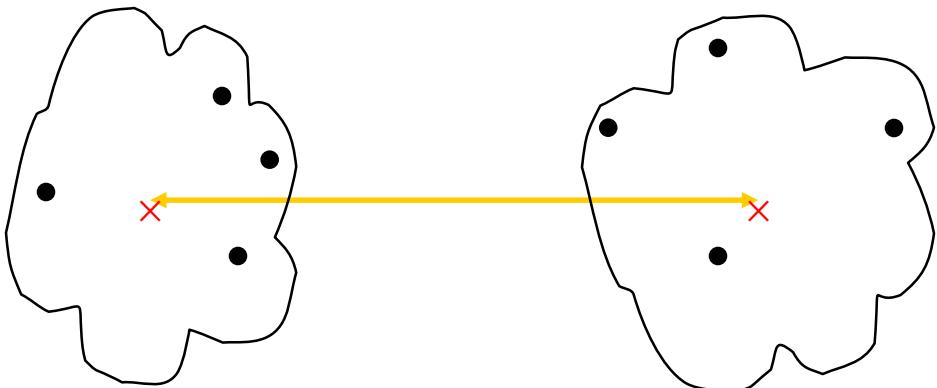
# Come definire la similarità tra gruppi



- MIN
  - MAX
  - **Media del gruppo**
  - Distanza tra i centroidi
  - Altri metodi basati su una funzione obiettivo
    - Ward's Method → SSE
- Matrice di prossimità

	C1	C2	C3	C4	C5	...
C1						
C2						
C3						
C4						
C5						
.						

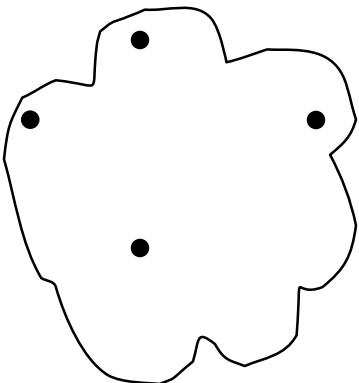
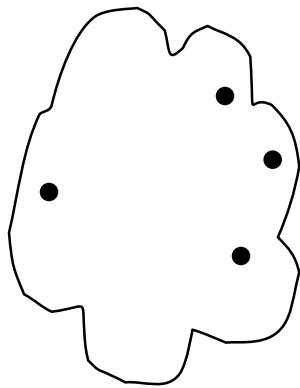
# Come definire la similarità tra gruppi



- MIN
- MAX
- Media del gruppo
- Distanza tra i centroidi
- Altri metodi basati su una funzione obiettivo
  - Ward's Method → la distanza tra due cluster si basa su SSE

	C1	C2	C3	C4	C5	...
C1						
C2						
C3						
C4						
C5						
.						
.						
Matrice di prossimità						

# Come definire la similarità tra gruppi



	C1	C2	C3	C4	C5	...
C1						
C2						
C3						
C4						
C5						
.	.	.	.	.	.	.

- MIN
- MAX
- Media del gruppo
- Distanza tra i centroidi
- Altri metodi basati su una funzione obiettivo
  - Ward's Method → la distanza tra due cluster si basa su SSE

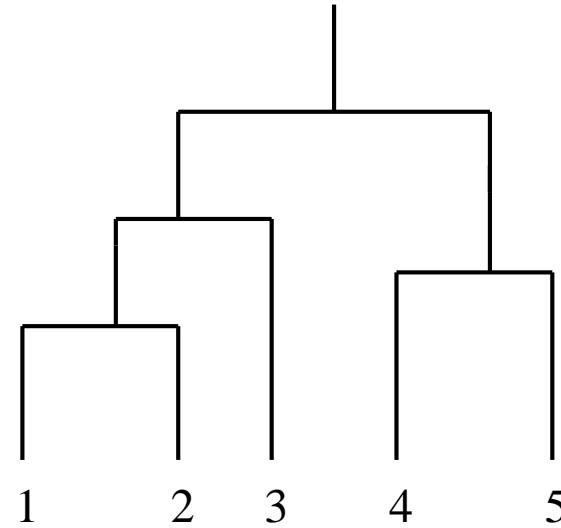
Matrice di prossimità

# Clustering gerarchico: MIN o legame singolo

- La similarità (distanza) di due cluster è definita dai due punti (uno per gruppo) più simili (o meno distanti) nei diversi cluster
- Determinata da un solo paio di punti, cioè un legame singolo

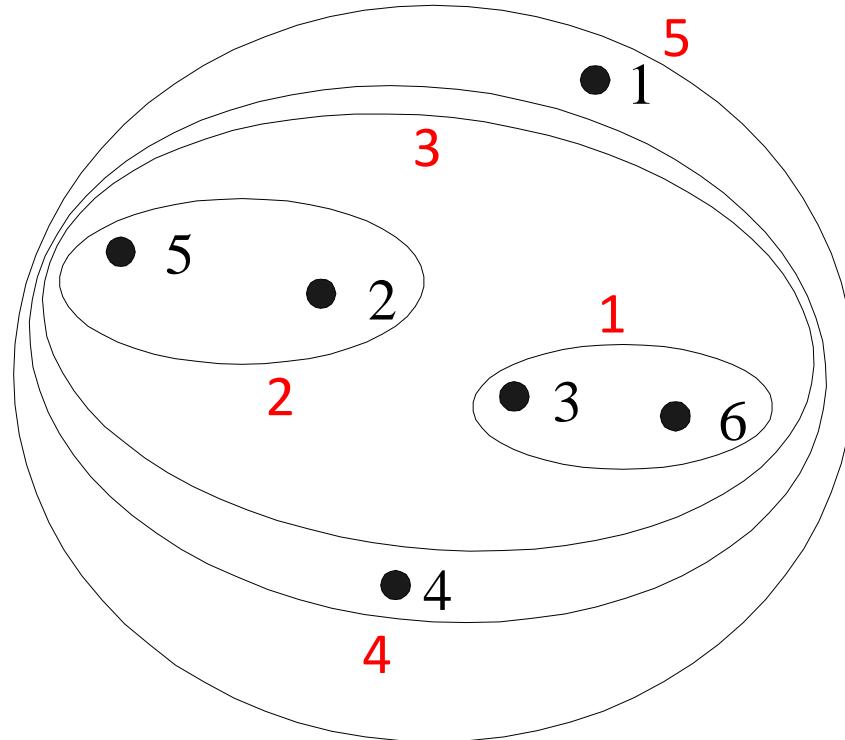
Matrice di similarità

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



# Clustering gerarchico: MIN o legame singolo

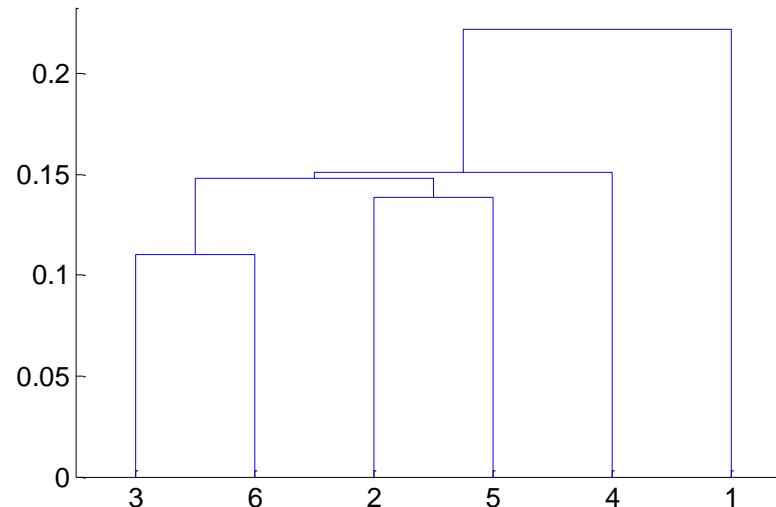
Distanze



Cluster innestati

Es. passo 3

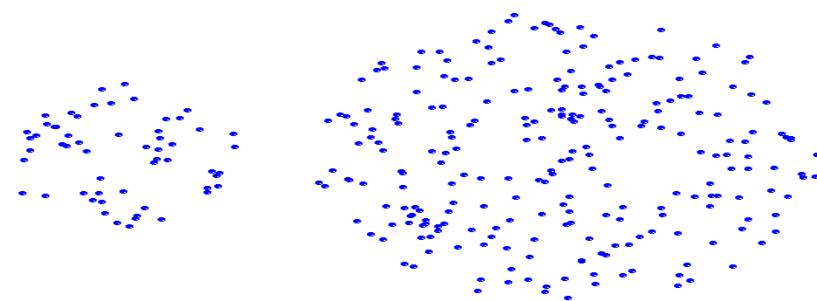
$$\text{Dist}(\{3,6\}, \{2,5\}) = \min(\text{dist}(\{2,3\}), \text{dist}(\{2,6\}), \text{dist}(\{5,3\}), \text{dist}(\{5,6\}))$$



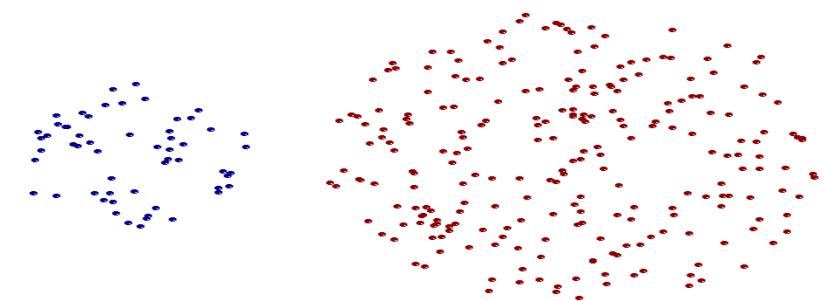
Dendrogramma

# Vantaggi di MIN

- Gestisce forme non-ellittiche



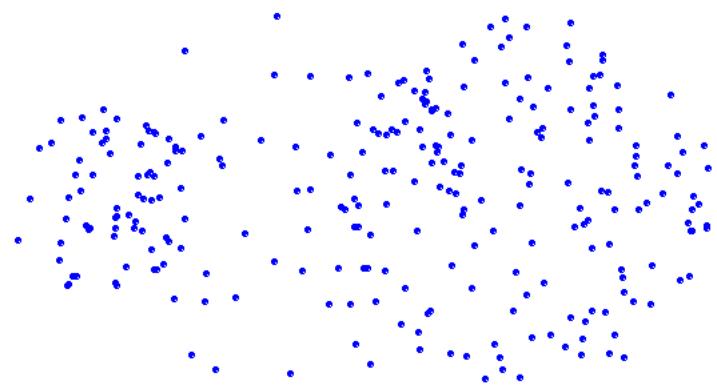
Gruppi reali



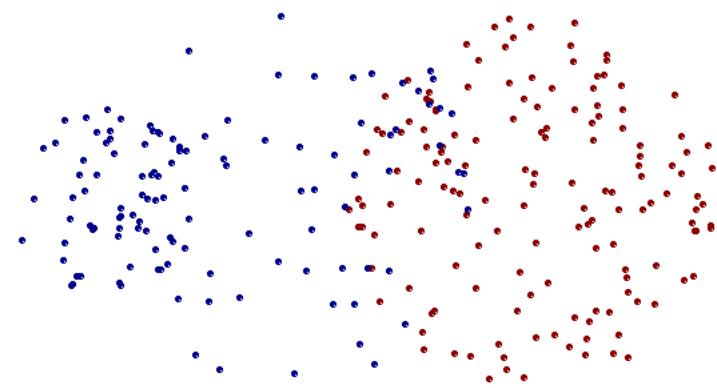
Due Cluster

# Limitazioni di MIN

- Sensibile a rumore e outlier



Gruppi reali



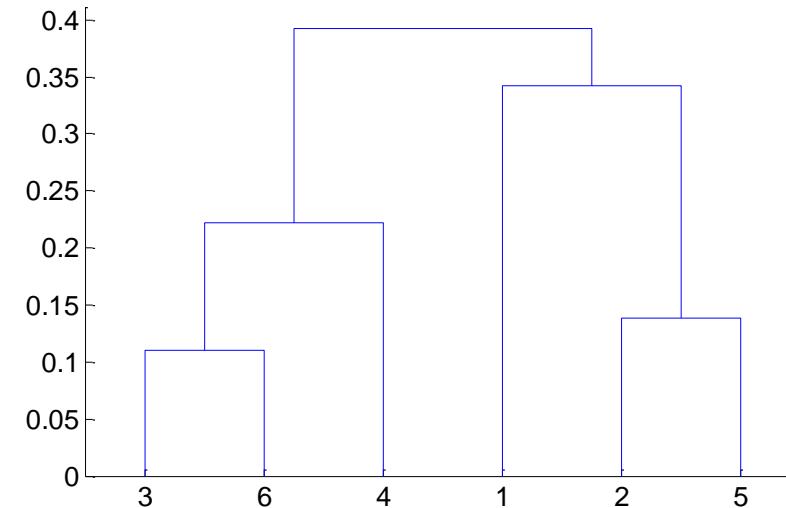
Due Cluster

# Clustering gerarchico: MAX o legame completo

- La similarità (distanza) di due gruppi si basa sui due punti meno simili (più distanti) nei due cluster

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Matrice di dissimilarità



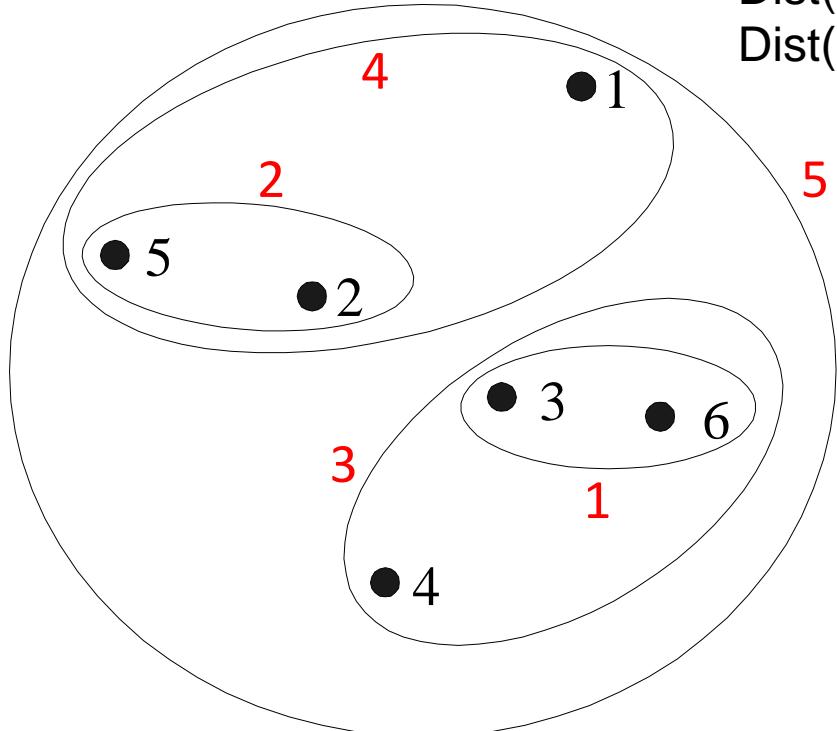
# Clustering gerarchico: MAX o legame completo

Passo 3

$$\text{Dist}(\{3,6\}, \{4\}) = \max(\text{dist}(\{3,4\}), \text{dist}(\{6,4\}))$$

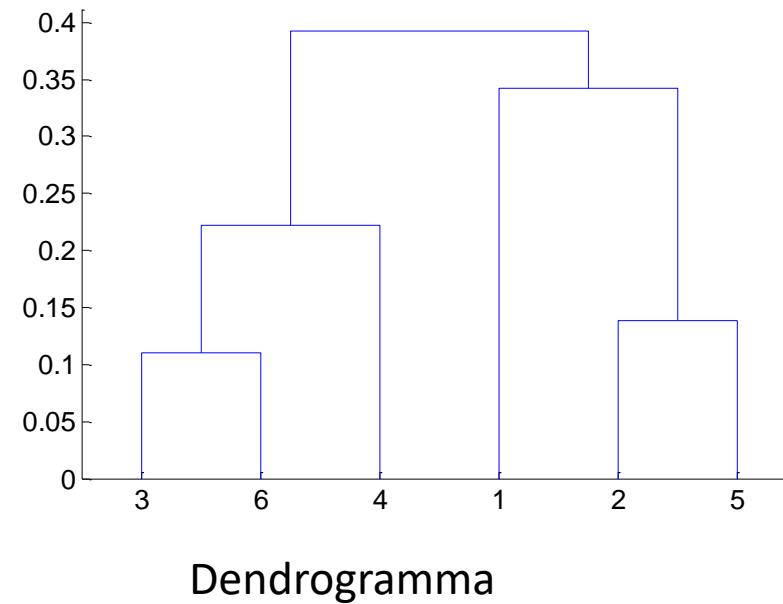
$$\text{Dist}(\{3,6\}, \{2,5\}) = \max(\text{dist}(\{3,2\}), \text{dist}(\{3,5\}), \text{dist}(\{6,2\}), \text{dist}(\{6,5\}))$$

$$\text{Dist}(\{3,6\}, \{1\}) = \max(\text{dist}(\{3,1\}), \text{dist}(\{6,1\}))$$



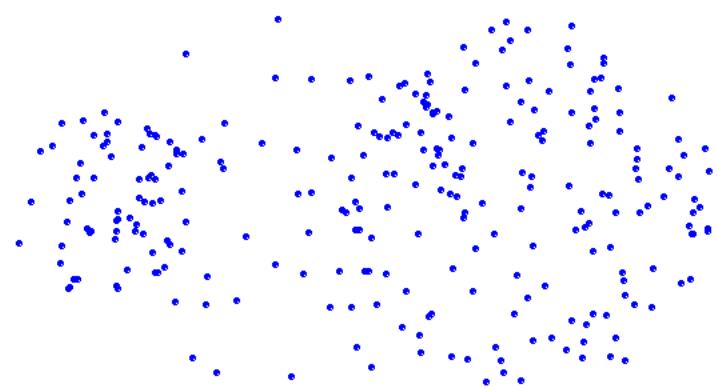
End

Cluster innestati

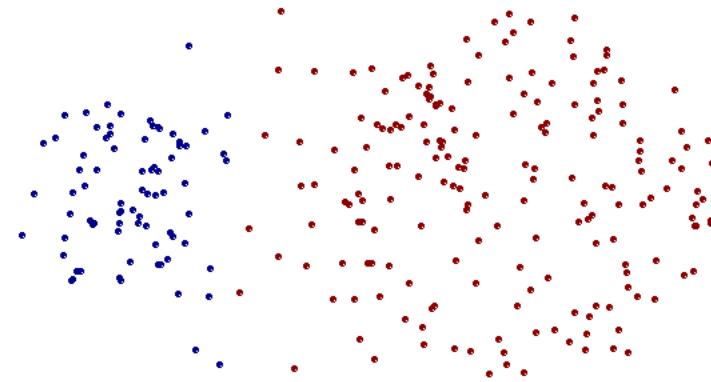


# Vantaggi di MAX

- Meno sensibile a rumore statistico e outlier



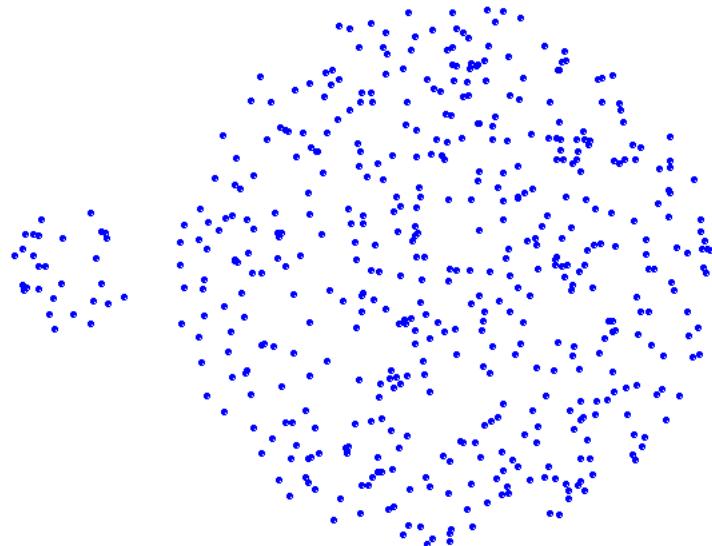
Punti originari



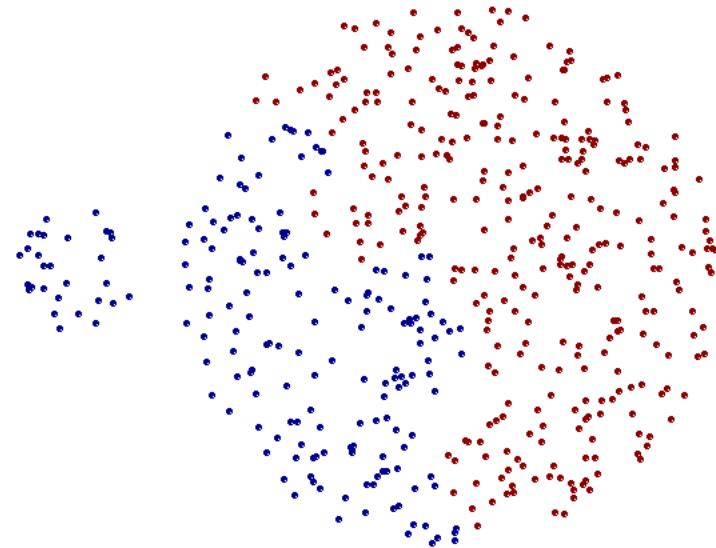
Due Cluster

# Limitazioni di MAX

- Tende a spezzare grandi cluster
- Sbilanciato verso cluster globulari (con queste forme si ottengono risultati migliori)



Punti originari



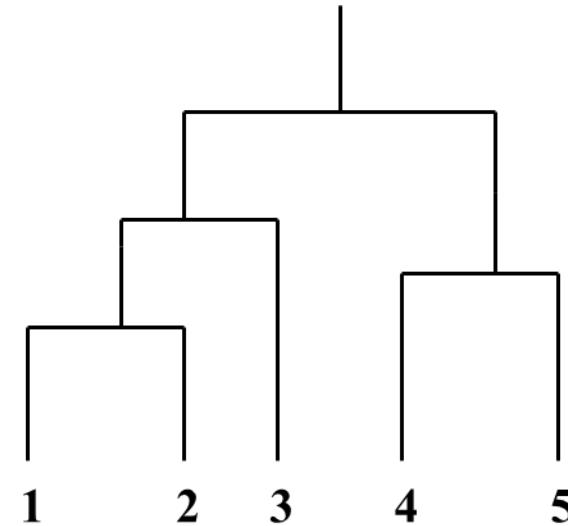
Due Cluster

# Clustering gerarchico: Media

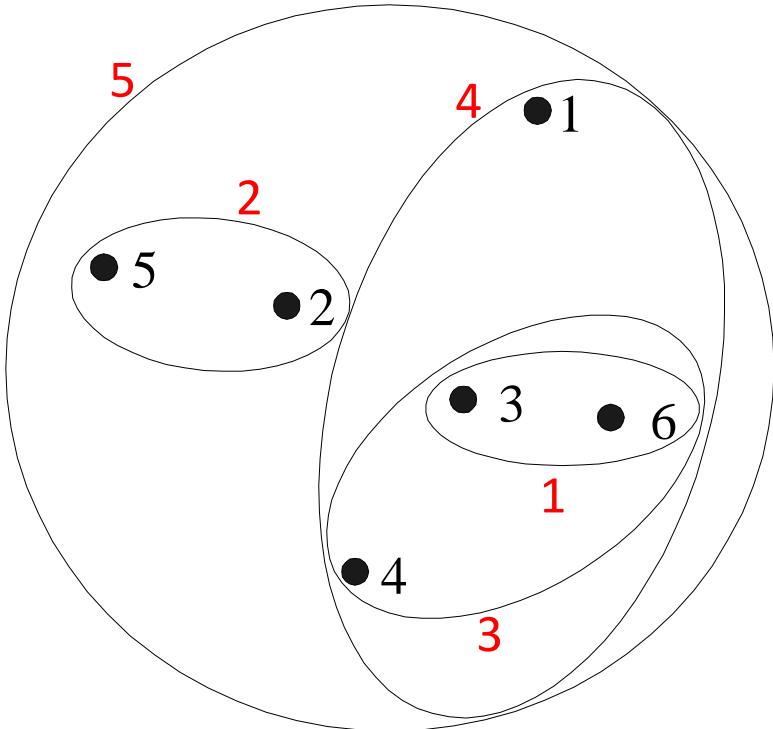
- La similarità di due cluster è la media delle similarità a coppie tra i punti nei due gruppi.

$$\text{prosimità(Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{prossimità}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

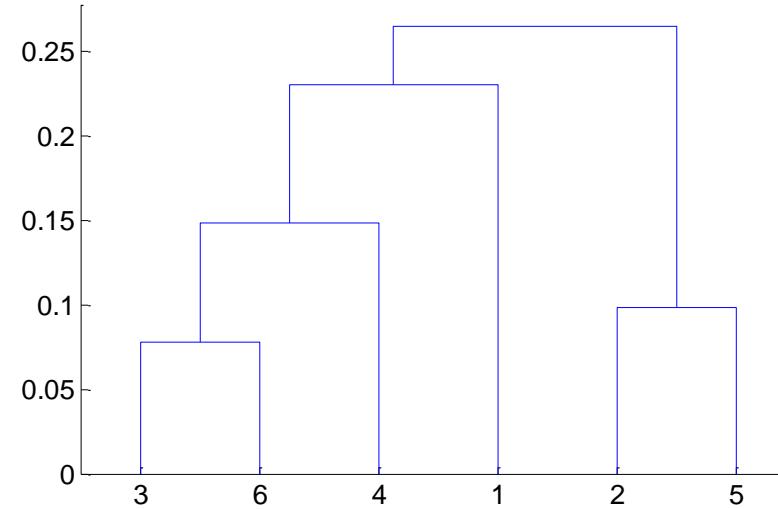


# Clustering gerarchico: Media



End

Cluster innestati



Dendrogramma

# Clustering gerarchico : Media

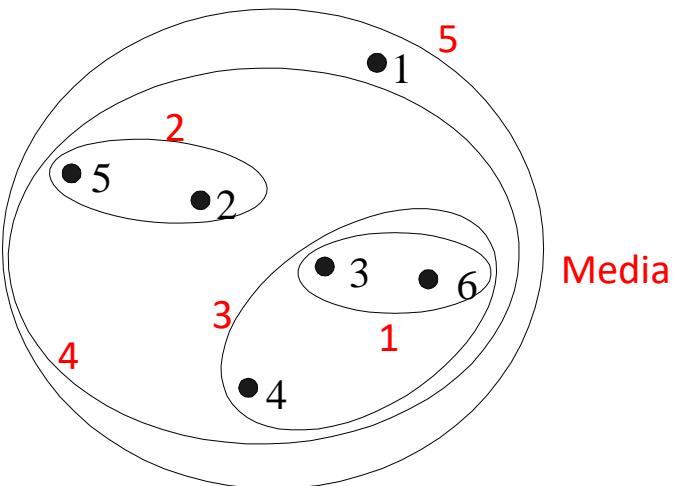
- Compromesso tra legame singolo e legame completo
- Punto di forza
  - Meno sensibile a rumore e outliers
- Limitazioni
  - Sbilanciato verso cluster globulari

# Clustering gerarchico: Metodo di Ward

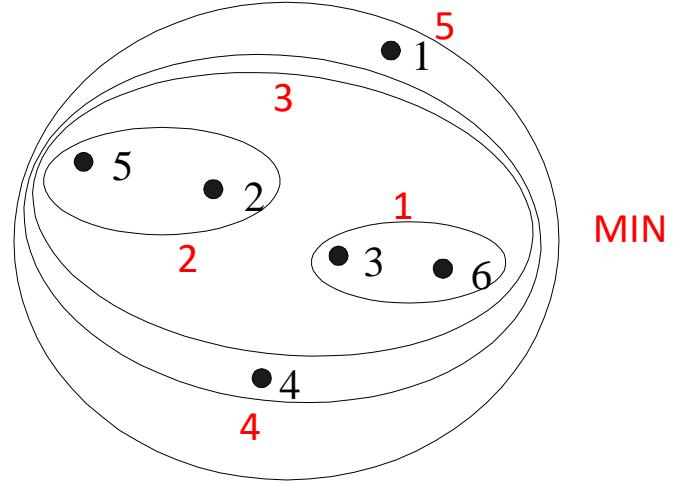
- La similarità di due gruppi si basa sull'incremento di SSE (somma degli errori quadratici) quando i due gruppi vengono uniti: **tanto minore l'incremento tanto più elevata la similarità**
- Simile al criterio della media se la distanza tra punti è la distanza Euclidea al quadrato.
- Meno sensibile a rumore e outliers e **sbilanciato verso cluster globulari**
- **Analogo gerarchico di K-means (può essere usato per scegliere i centroidi in K-means)**

## Clustering gerarchico: Comparazioni

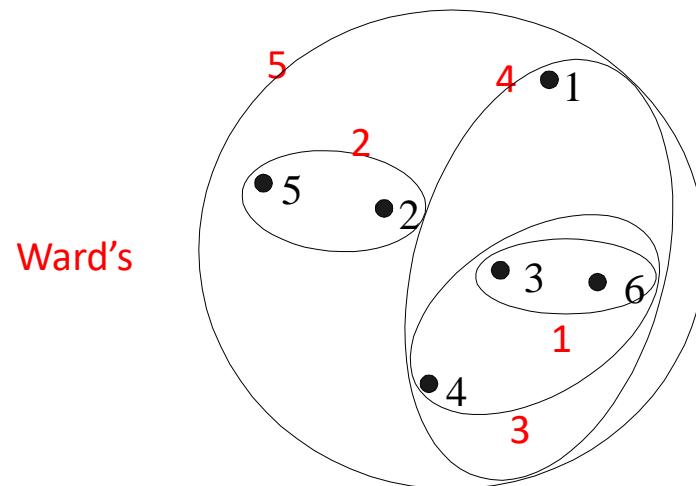
End



MIN



MAX



Ward's

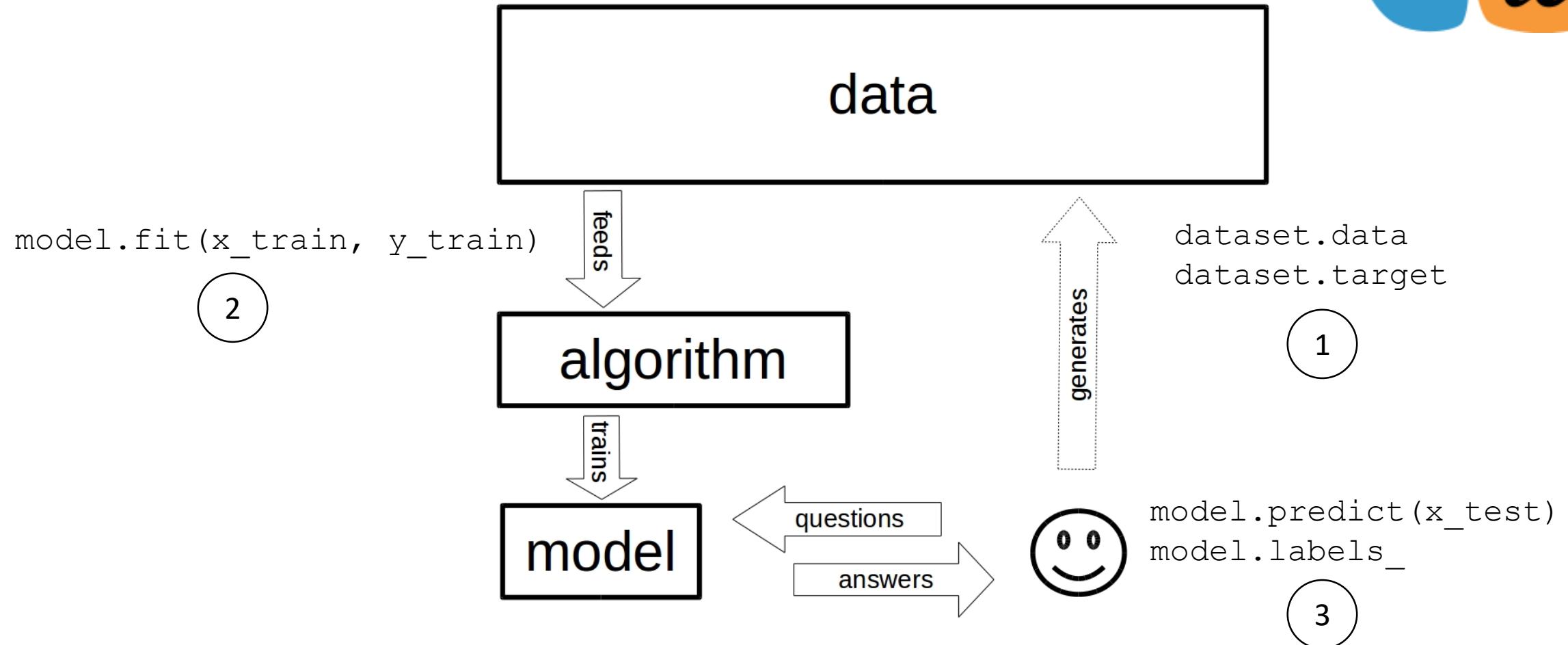
## Clustering gerarchico: Problemi e limitazioni

- Una volta presa la decisione di combinare due cluster, **non può essere annullata**
- Schemi differenti hanno problemi di vario tipo:
  - Sensibili a rumore e outlier
  - Difficoltà nel gestire cluster di dimensioni differenti e forme convesse
  - Spezzano grandi cluster

# LAB Session



# LAB Session – Clustering Gerarchico





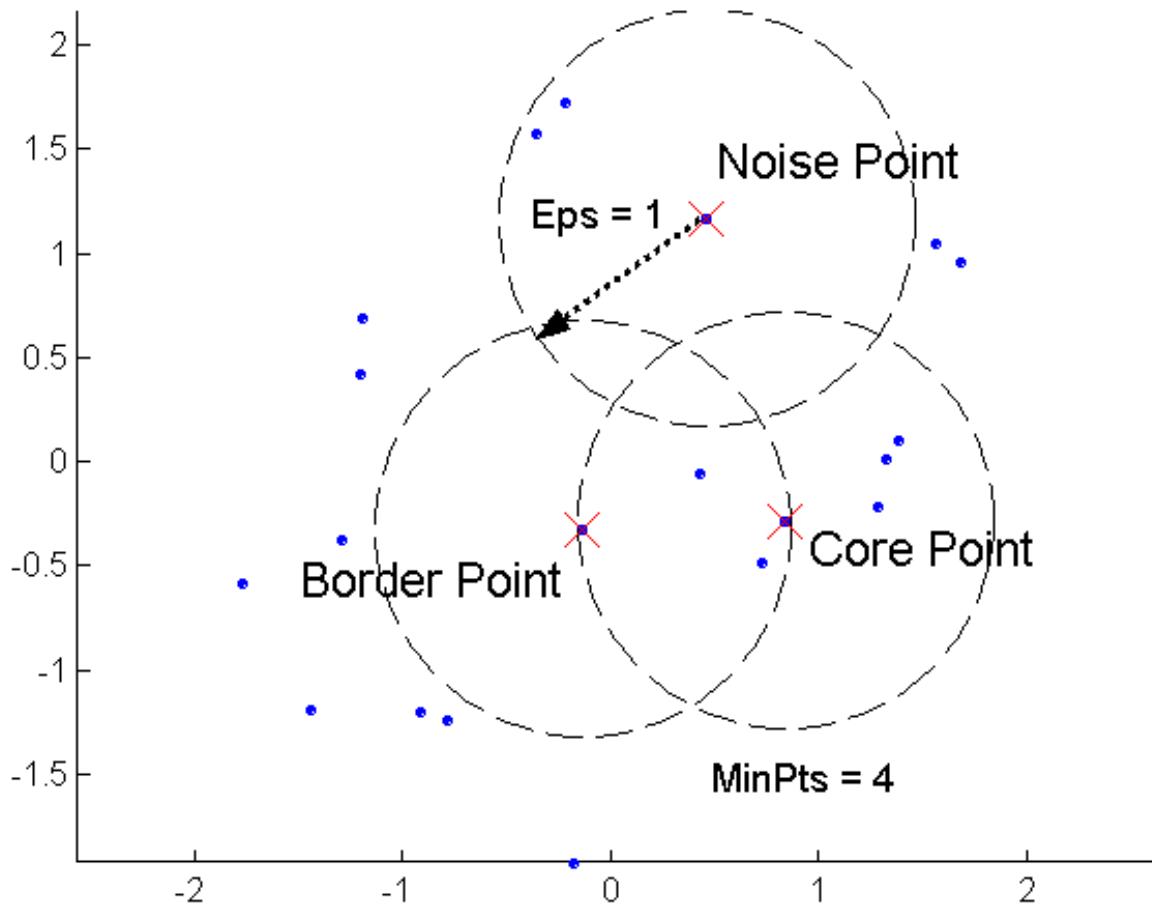
Learning  
Session



# Algoritmi di Clustering

- K-means e le sue varianti (partizionale)
- Clustering gerarchico
- Clustering basato sulla densità
- Misure di validità dei Cluster

# DBSCAN: Core, Border, e Noise Points

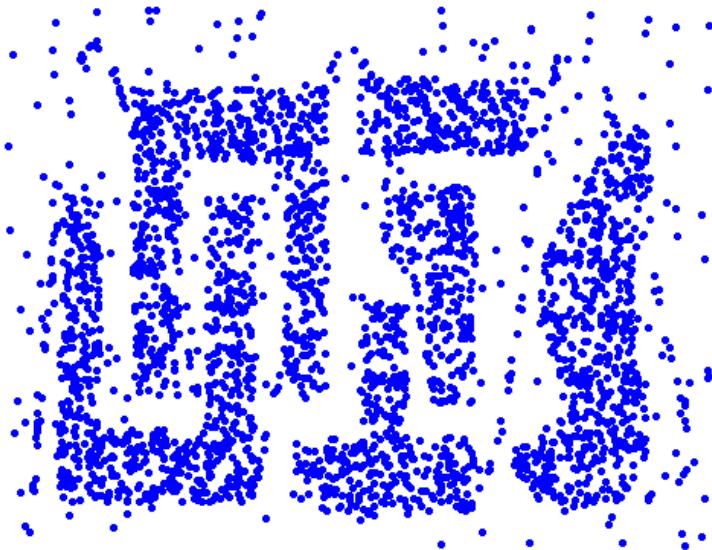


- Si basa sulla densità = numero di punti entro un raggio specificato ( $Eps$ )
- Un punto è definito **core point** se ha più di un numero specificato di punti ( $MinPts$ ) entro  $Eps$
- Un **border point** ha meno punti di  $MinPts$  entro  $Eps$ , ma nelle sue vicinanze (all'interno del cerchio di raggio  $Eps$ ) si trovano dei core points.
- Un **noise point** è ogni altro punto che non è né core point o border point.

# Algoritmo DBSCAN - Algoritmo

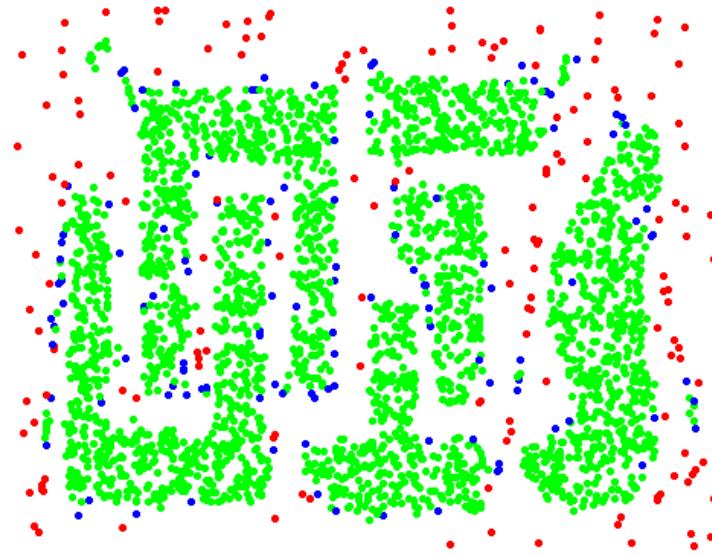
- identificare ciascun punto e definire se sia core, border o noise point
- eliminare i noise point
- formare un **cluster** con **tutti i core point che si trovano ciascuno nel cerchio di raggio Eps dell'altro**
- assegnare ciascun **border point** al **cluster** cui appartiene il **core point più vicino**

# DBSCAN: Core, Border e Noise Points



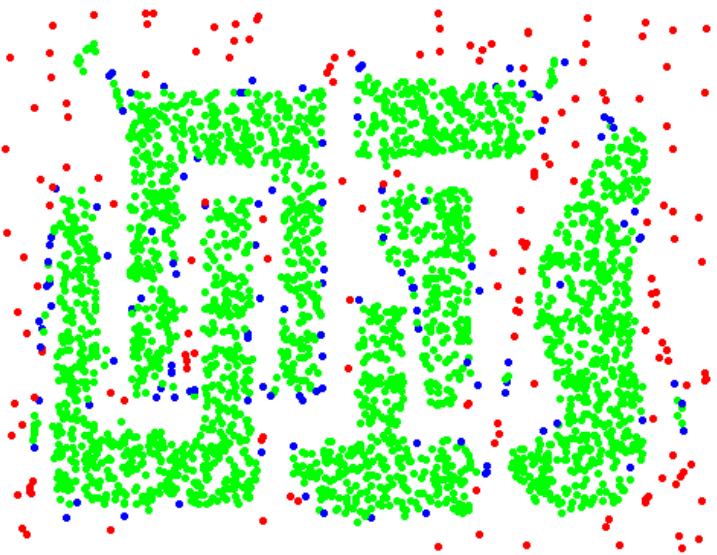
Punti originari

Eps = 10, MinPts = 4

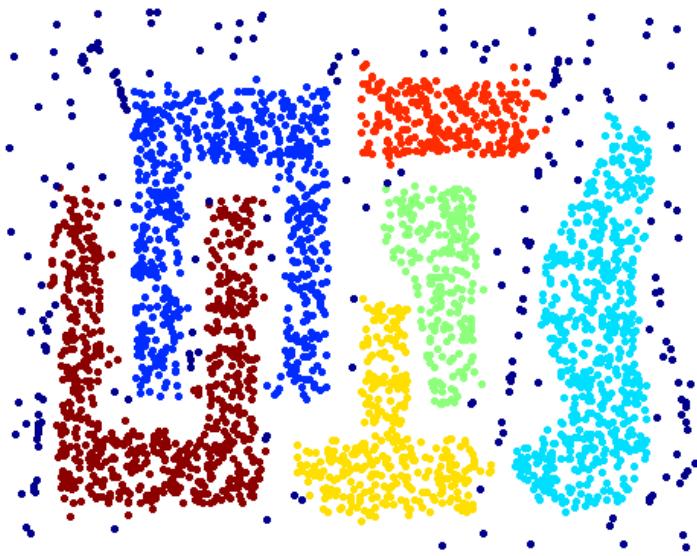


Tipi: **core**, **border** e **noise**

# DBSCAN: Core, Border e Noise Points



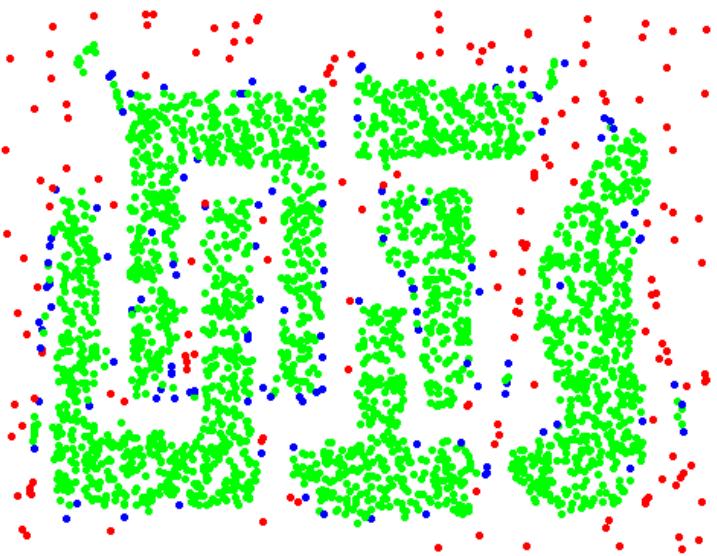
Eps = 10, MinPts = 4



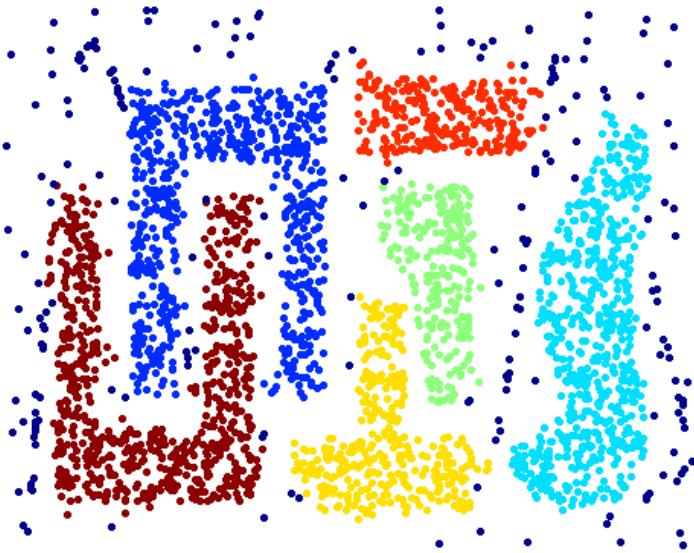
Cluster

# DBSCAN: Core, Border e Noise Points

- Resistente al rumore
- Può gestire gruppi di forma e dimensione differenti



Eps = 10, MinPts = 4

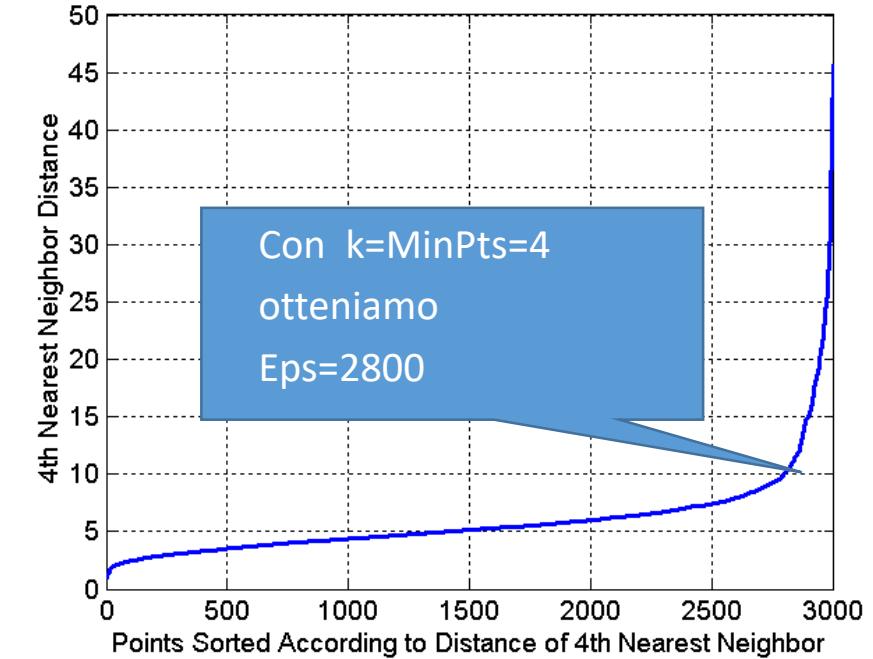


Cluster

DBSCAN **NON** è in grado di classificare insiemi di dati con grandi differenze nelle densità

## DBSCAN: Determinare EPS e MinPts (sorted k-dist graph)

- Si indichi con  $d$  la distanza di un punto  $p$  dal suo  $k$ -esimo nearest neighbor: entro una distanza  $d$  da  $p$  sono quindi contenuti  $k+1$  punti (**a meno di casi particolari in cui più punti siano alla stessa distanza  $d$  da  $p$** ).
- Fissato  $K$ , si disegnano su un grafico le distanze (ordinate) di ogni punto dal suo  $k^{\text{th}}$  nearest neighbor
- Si scegli Eps nel **punto di gomito**



# LAB Session





Learning  
Session



# Algoritmi di Clustering

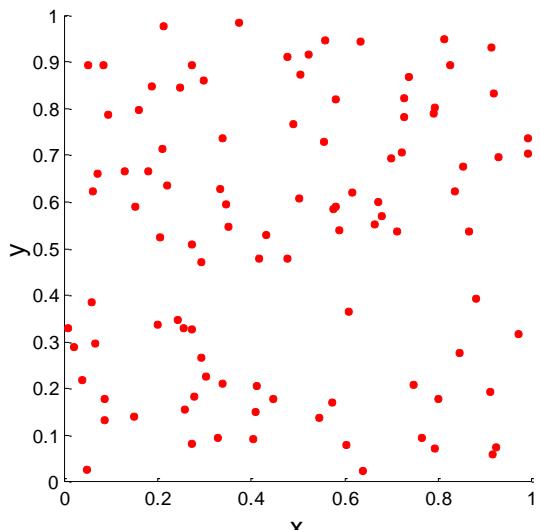
- K-means e le sue varianti (partizionale)
- Clustering gerarchico
- Clustering basato sulla densità
- Misure di validità dei Cluster

# Validazione dei cluster: a cosa serve?

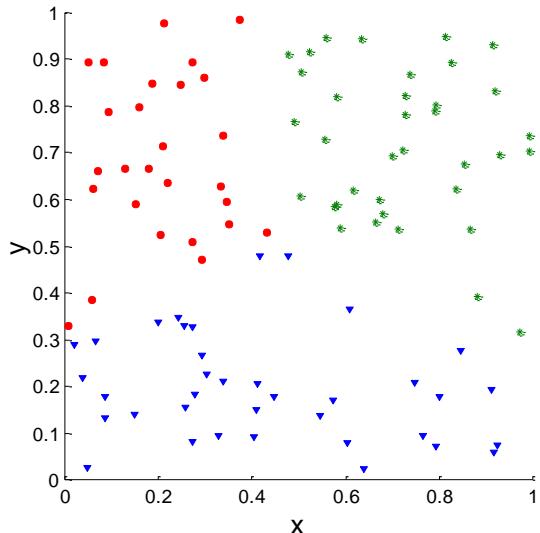
- Evitare di trovare pattern in dati casuali o relative a rumore
- Comparare algoritmi di clustering
- Comparare i cluster

# Cluster in Dati Casuali

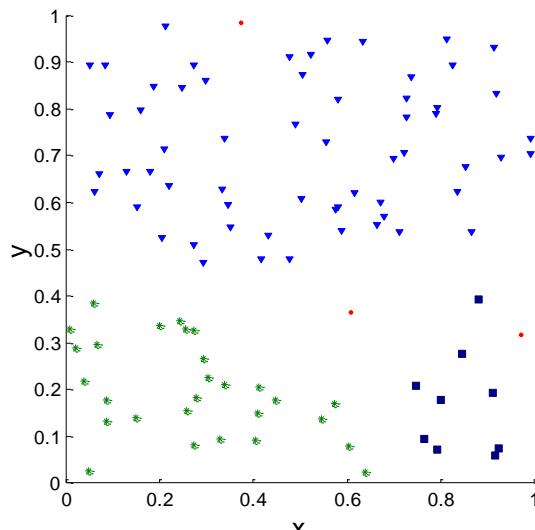
Punti casuali



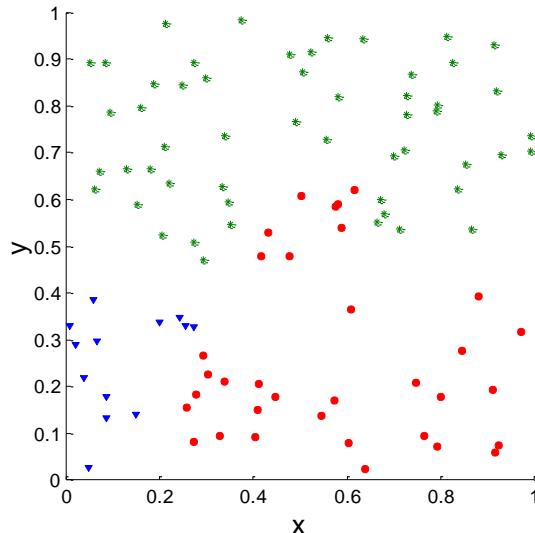
K-means



DBSCAN



Gerarchico  
Legame completo



# Misure di validità dei Cluster

Le misure numeriche usate per valutare la bontà di un clustering, sono classificate in:

- **Indici esterni:** Usati per misurare fino a che punto le etichette/raggruppamenti del cluster corrispondono a quelle fornite da una fonte esterna (conoscenza pregressa).
- **Indici interni:** Usati per misurare la bontà di un clustering senza riferimento a informazione esterna.
  - Correlazione
  - Similarità
  - Sum of Squared Error (SSE)
- **Indici relativi:** utilizzati per comparere due diversi clustering o cluster
  - Possono basarsi sia su indici interni che esterni

A volte si parla di **criteri** piuttosto che di **indici**. In generale possiamo dire che il criterio è la strategia e l'indice la misura che la implementa

## Misure di validità dei Cluster

- **Indici interni:** Usati per misurare la bontà di un clustering senza riferimento a informazione esterna.
  - Correlazione
  - Similarità
  - Sum of Squared Error (SSE)

## Misure di validità dei Cluster

- **Indici interni:** Usati per misurare la bontà di un clustering senza riferimento a informazione esterna.
  - Correlazione
  - Similarità
  - Sum of Squared Error (SSE)

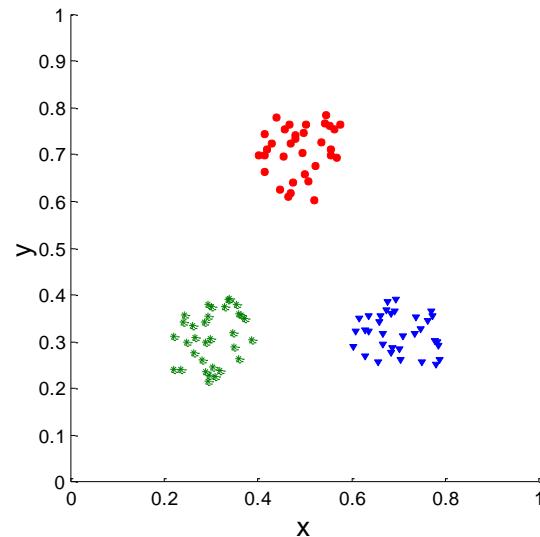
# Indici Interni: la correlazione

- Due matrici
  - **Matrice di prossimità** (o matrice di similarità)
    - Similarità tra ogni coppia di oggetti.  $\approx 1$  per coppie che appartengono allo stesso cluster
  - **Matrice di incidenza**
    - La cella della matrice è 1 se i punti associati sono nello stesso cluster. E' nulla altrimenti
- Calcola la correlazione tra le celle corrispondenti delle due matrici
  - Poichè le matrici sono simmetriche, solo  $n(n-1) / 2$  celle sono considerate.
- Alta correlazione indica che i punti che appartengono allo stesso cluster sono vicini tra loro.

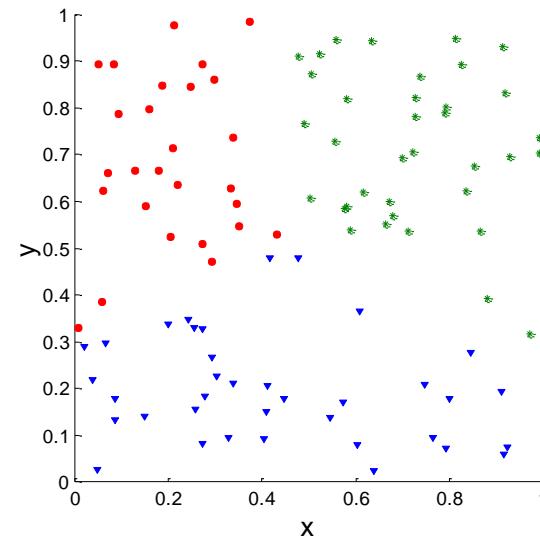
Non è una misura che funziona bene nel caso di cluster a densità differente.

# Indici Interni: la correlazione

## Correlazioni per K-means clusterings in due data set.



Alta Correlazione



Bassa correlazione

## Misure di validità dei Cluster

- **Indici interni:** Usati per misurare la bontà di un clustering senza riferimento a informazione esterna.



Correlazione

- Similarità
- Sum of Squared Error (SSE)

## Misure di validità dei Cluster

- **Indici interni:** Usati per misurare la bontà di un clustering senza riferimento a informazione esterna.

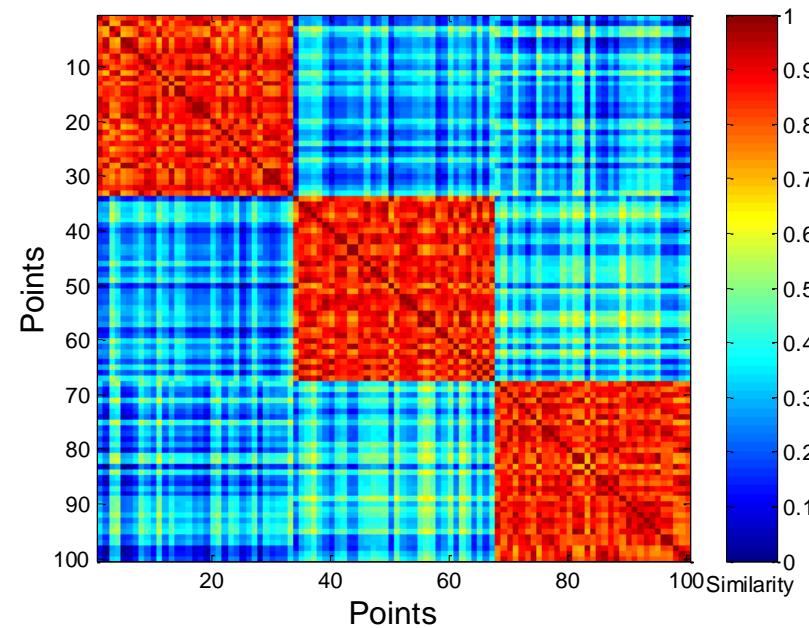
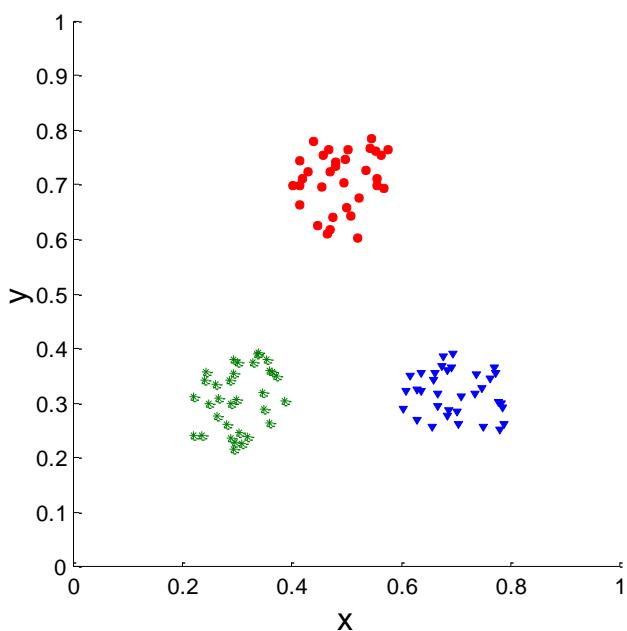


Correlazione

- Similarità
- Sum of Squared Error (SSE)

## Indici Interni: la similarità

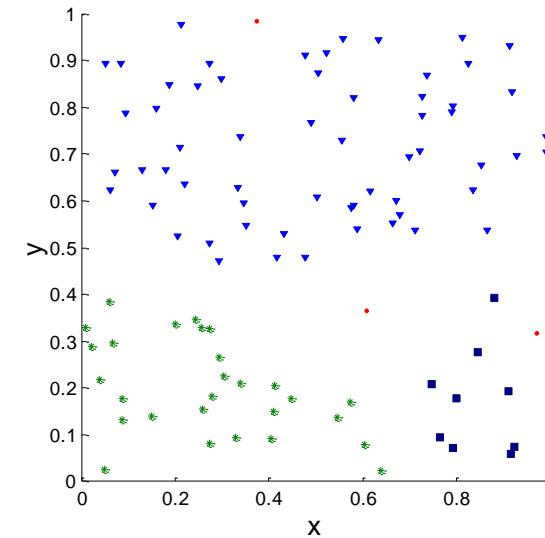
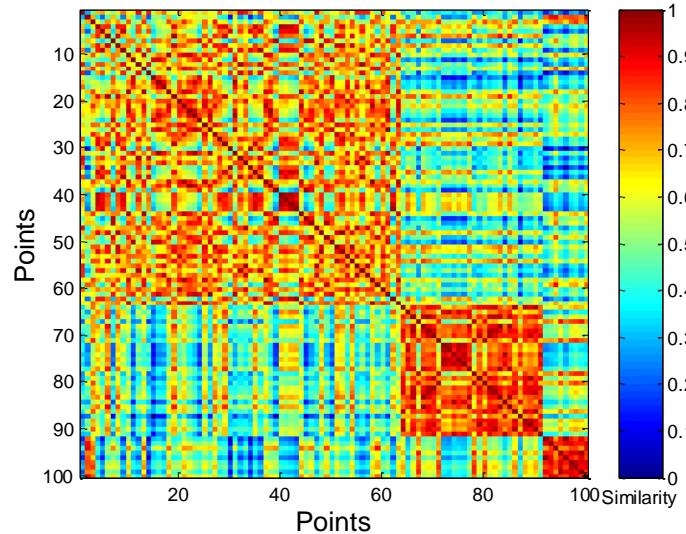
- Ordinare la matrice di similarità in base alle etichette di gruppo ed ispezionarla visivamente (ovvero si ordina la matrice di similarità in base ai raggruppamenti dettati dai cluster).



Se i dati sono distribuiti uniformemente la matrice è più “sfumata”

# Indici Interni: la similarità

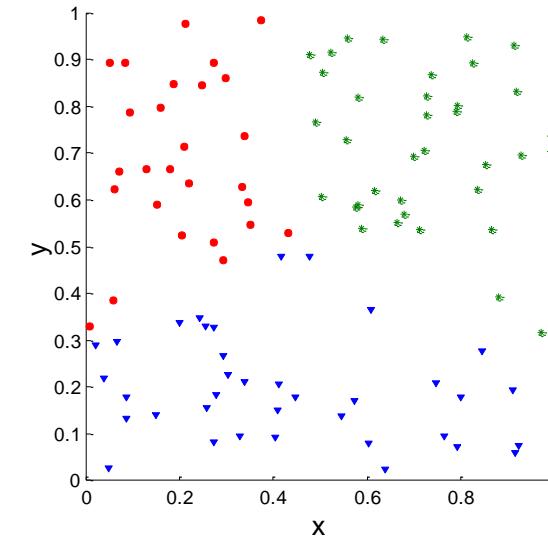
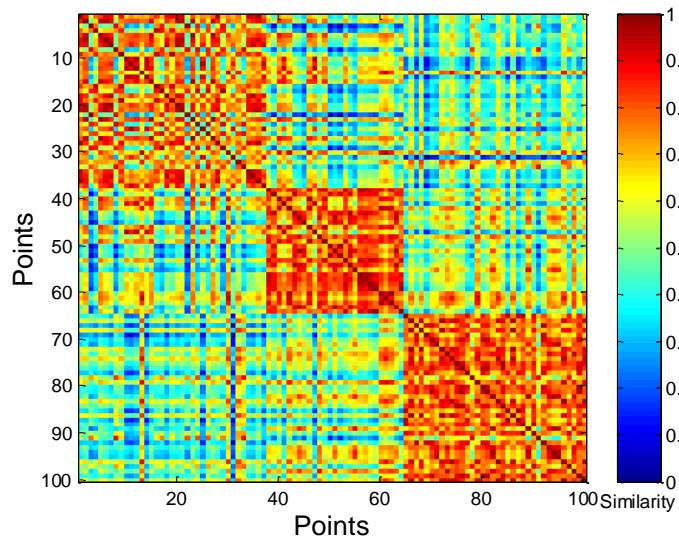
- Cluster nei dati casuali non sono così ben definiti



DBSCAN

# Indici Interni: la similarità

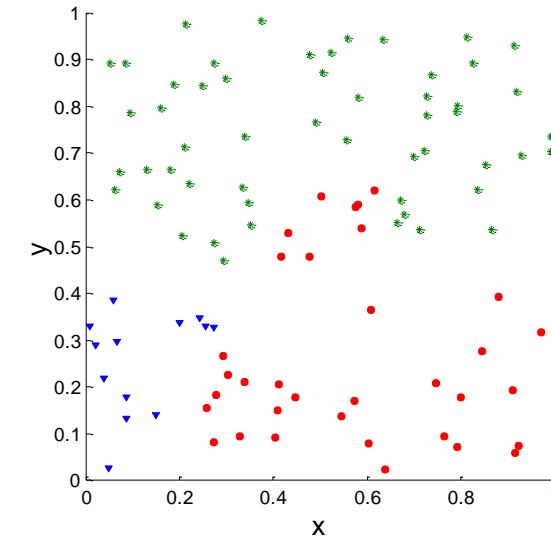
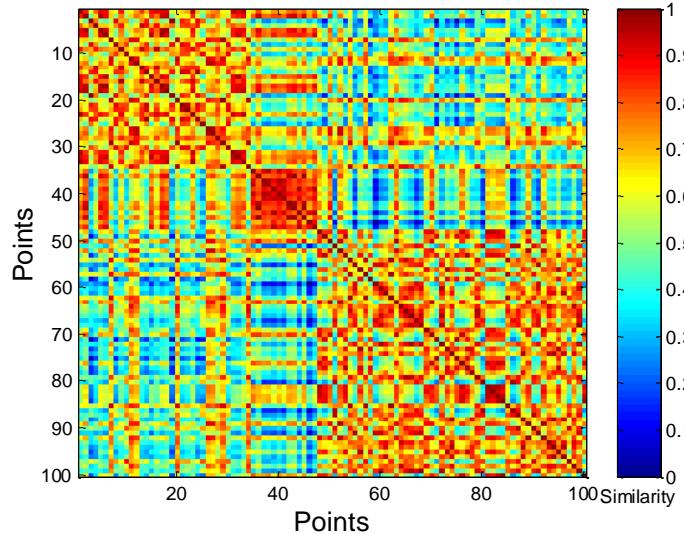
- Cluster nei dati casuali non sono così ben definiti



## K-means

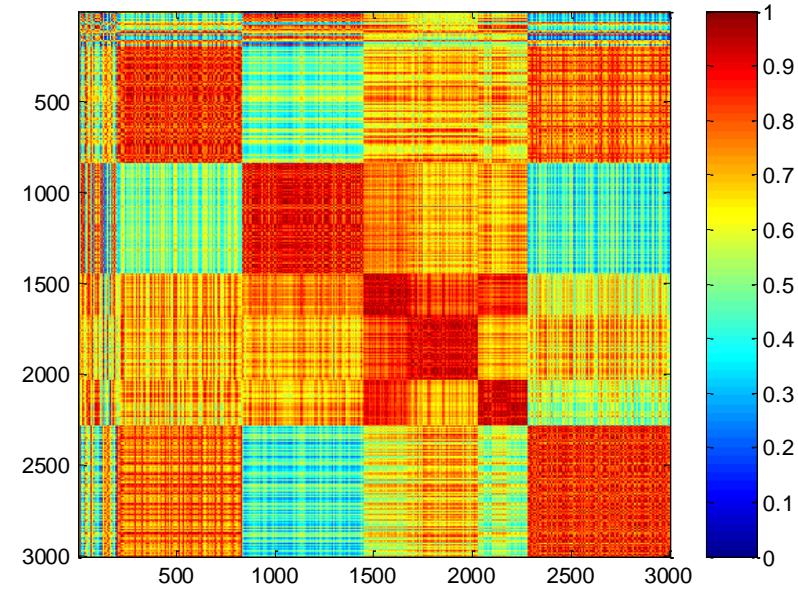
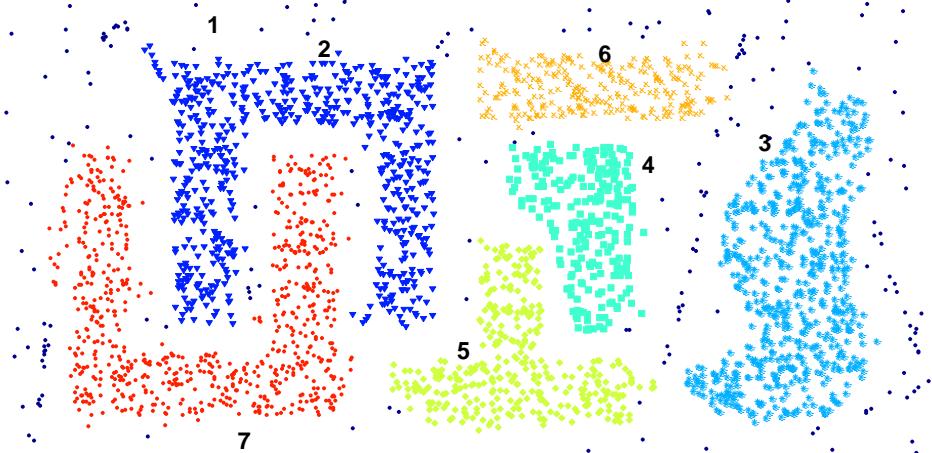
# Indici Interni: la similarità

- Cluster nei dati casuali non sono così ben definiti



Legame Completo

# Indici Interni: la similarità (non adatta al DBSCAN)



DBSCAN

## Misure di validità dei Cluster

- **Indici interni:** Usati per misurare la bontà di un clustering senza riferimento a informazione esterna.



Correlazione



Similarità

- Sum of Squared Error (SSE)

## Misure di validità dei Cluster

- **Indici interni:** Usati per misurare la bontà di un clustering senza riferimento a informazione esterna.



Correlazione

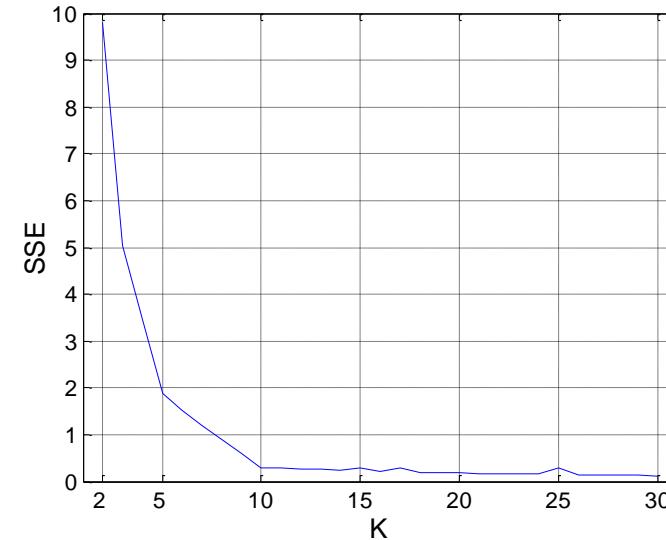
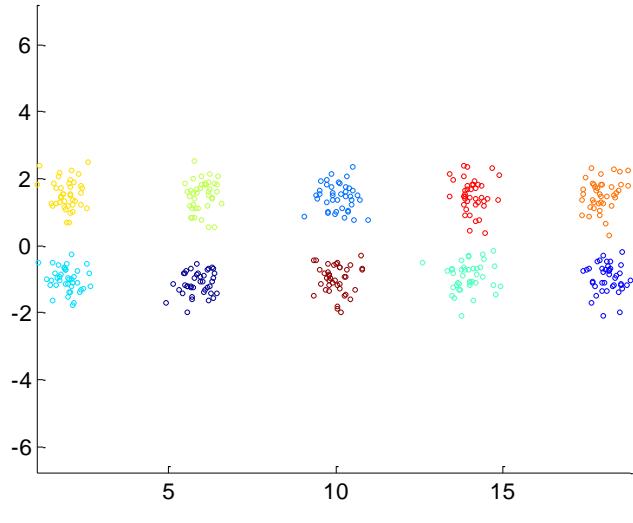


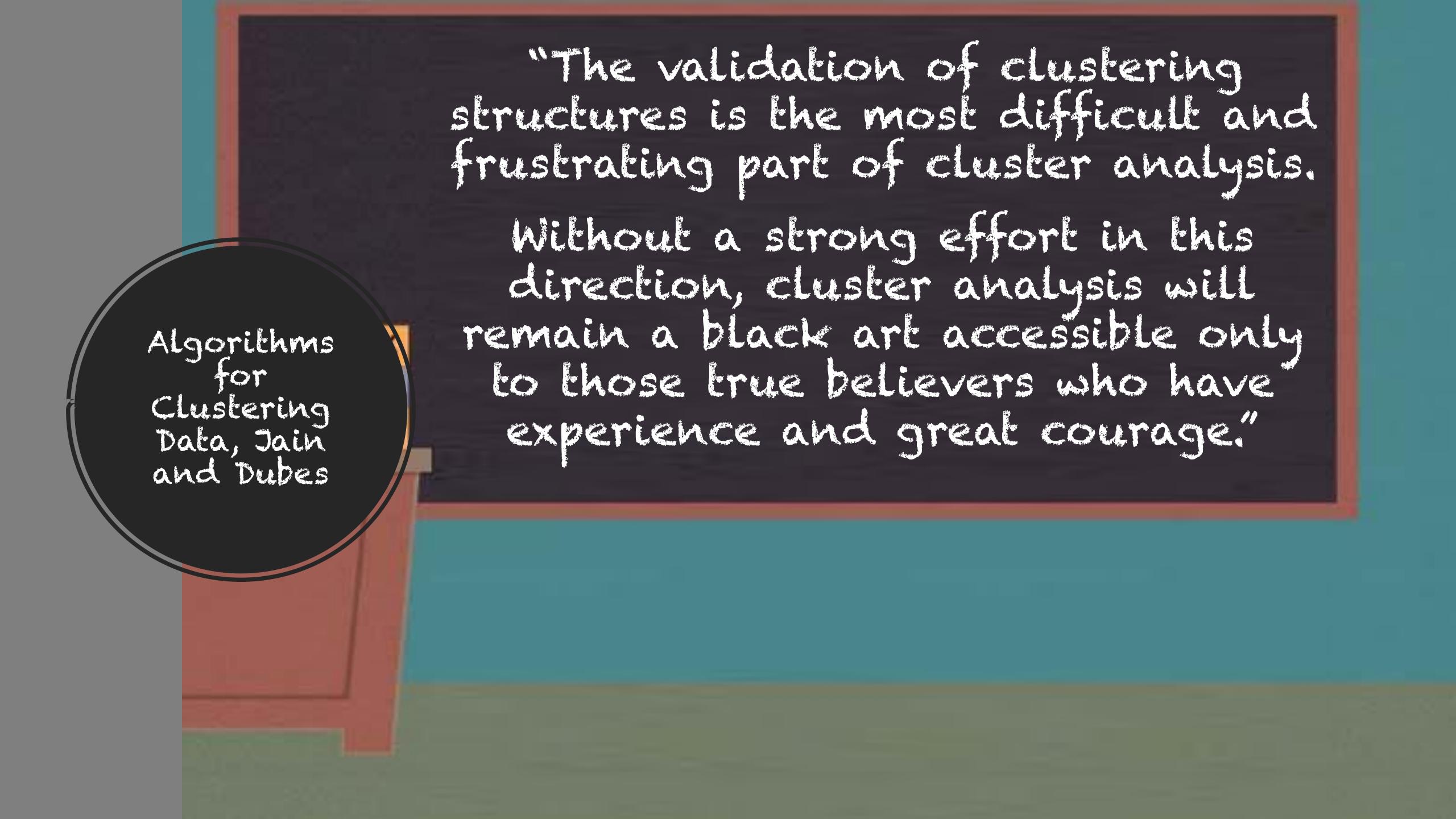
Similarità

- Sum of Squared Error (SSE)

## Indici Interni: SSE

- Utilizzato per comparare clustering differenti
- Può anche essere utilizzato per stimare il numero di cluster ottimale nel K-means (visto in precedenza)





Algorithms  
for  
Clustering  
Data, Jain  
and Dubes

"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

"La convalida delle strutture di clustering è la parte più difficile e frustrante dell'analisi dei cluster.

Senza un forte sforzo in questa direzione, la cluster analysis rimarrà un'arte oscura accessibile solo a quei veri credenti che hanno esperienza e grande coraggio."



That's  
all  
folks!

Marco  
Pirrone

Massimo  
Romano





# GRAZIE PER L'ATTENZIONE!

Ing. Marco Pirrone Ph.D.

(<https://www.linkedin.com/in/marcopirrone/>)

Ing. Massimo Romano Ph.D.

(<https://www.linkedin.com/in/massimoromano/>)

Seguici su

[www.masterandskills.com](http://www.masterandskills.com)



DIPARTIMENTO DI METODI E MODELLI PER  
L'ECONOMIA IL TERRITORIO E LA FINANZA  
MEMOTEF



SAPIENZA  
UNIVERSITÀ DI ROMA