Festival della Scienza

Genova, 21 ottobre _ 5 novembre 2021

Mappe

La Matematica nell'analisi del genoma

Claudia Angelini

Istituto per le Applicazioni del Calcolo *M. Picone*

Consiglio Nazionale delle Ricerche

Consiglio Nazionale delle Ricerche

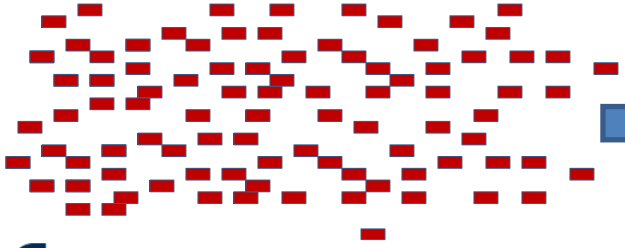
CNR Campania REte Outreach

I sequenziatori di nuova generazione

2007-2012

2019-

Sequenziamento



GGTCTGGATGC
CGGTCTGGATGC
GCGGTCTGGATG
GCGGTCTGGAT
GGCGGTCTGGAT
GGCGGTCTGGA
TCTATGCGGGCCCCT
TCTATGCGGGCCCC
ATCTATGCGGGCC
TATCTATGCGGGC
TTATCTATGCGGG
CTTATCTATGCGGG

Come mettere insieme tutta questa informazione?

Milioni o centinaia di milioni di sequenze

Consiglio Nazionale delle Ricerche

Festival della Scienza

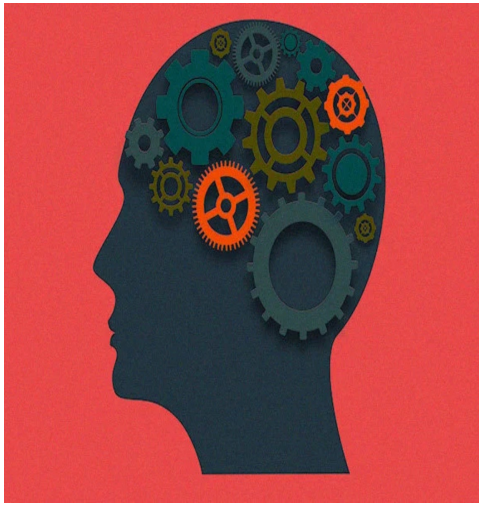
Risolvere un puzzles

Riferimento

Dati

Assemblamento

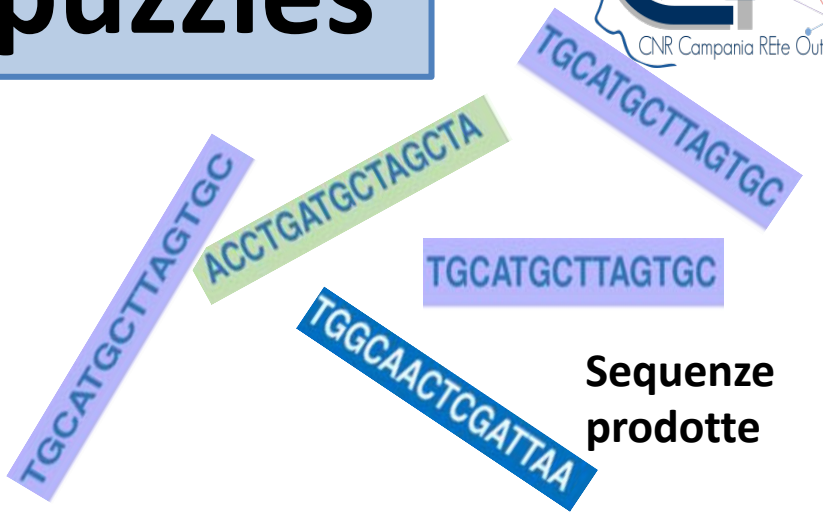
Allineamento



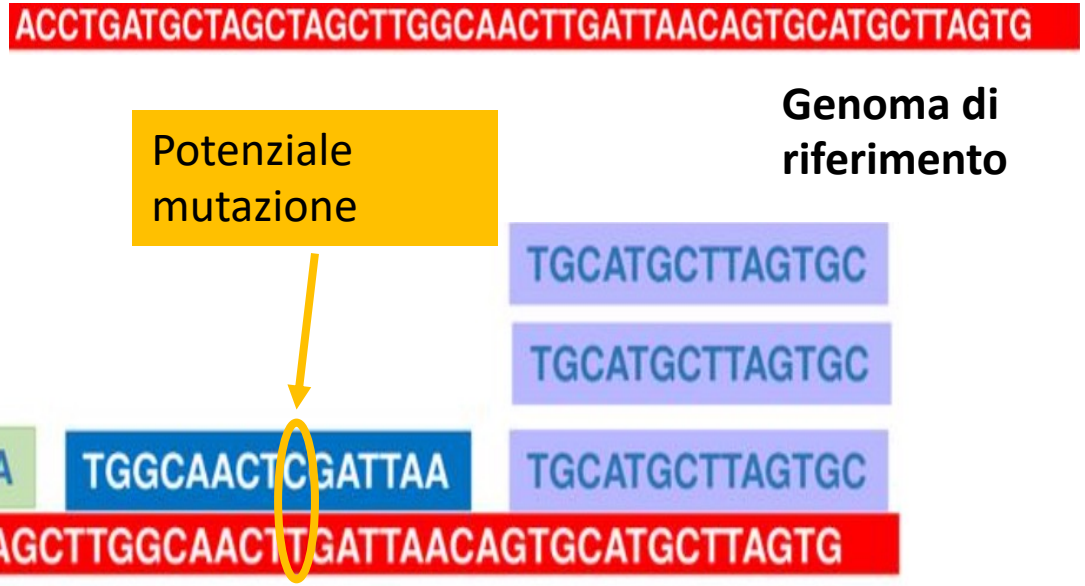
Algoritmi

Risolvere il DNA puzzles

Negli studi sul **genoma umano** si devono allineare **decine di milioni o centinaia di milioni di sequenze** su un riferimento che ha circa **3 miliardi di basi**

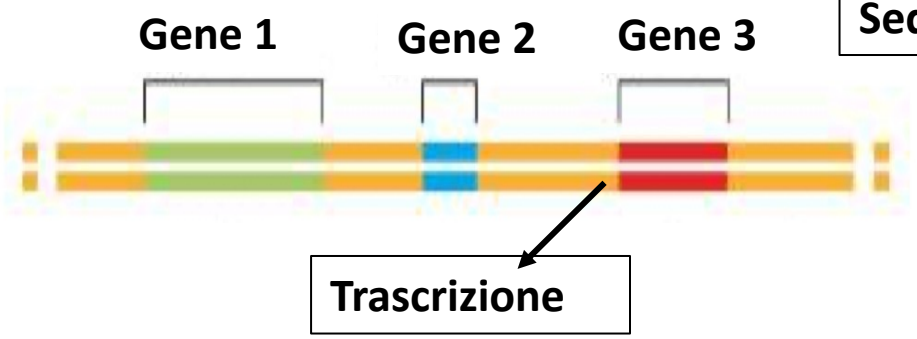


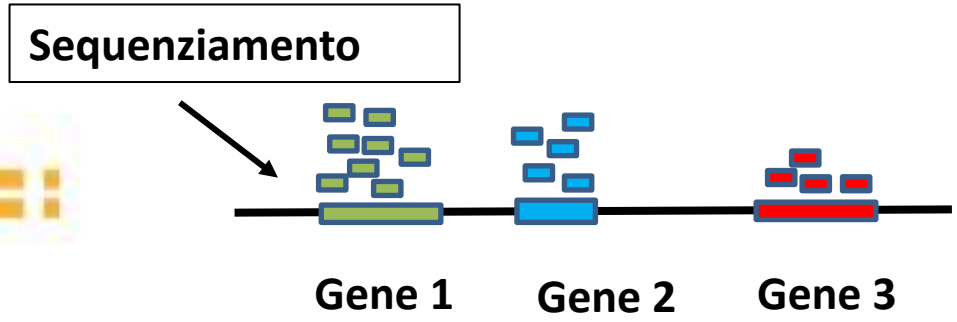
Sequenze prodotte

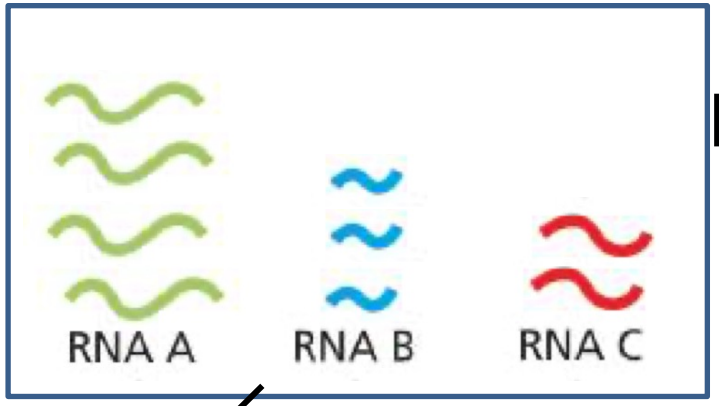


Allineamento al genoma di riferimento

La misurazione dell'espressione genica







Gene 1	Gene 2	Gene 3	-----
4	3	2	….

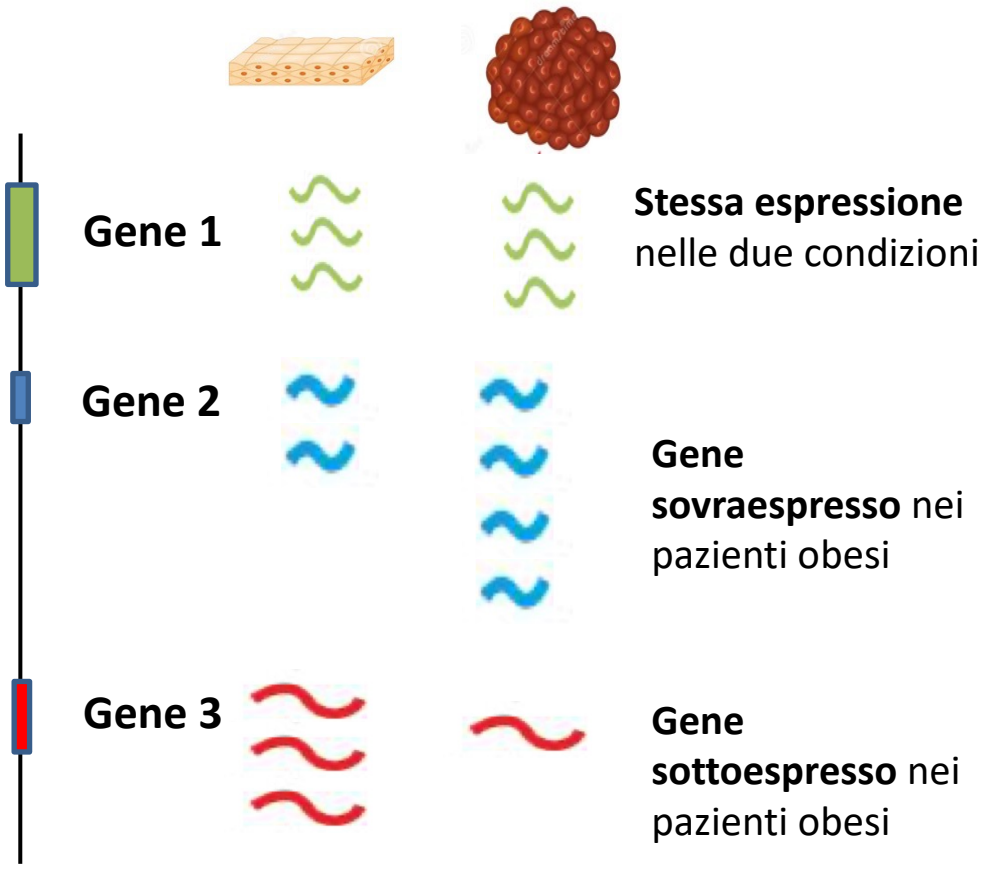
I sequenziatori sono in grado di misurare contemporaneamente l'**espressione** di tutti i geni di un dato **individuo** in una data **condizione**

Traduzione: gli RNA codificanti vengono trasformati in **proteine**

Espressione genica e regolazione cellulare

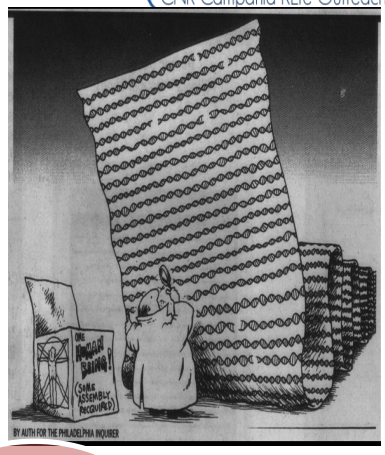
E’ possibile misurare i **livelli di espressione** genica per ciascun individuo, tessuto, condizione sperimentale (**profili di espressione**)

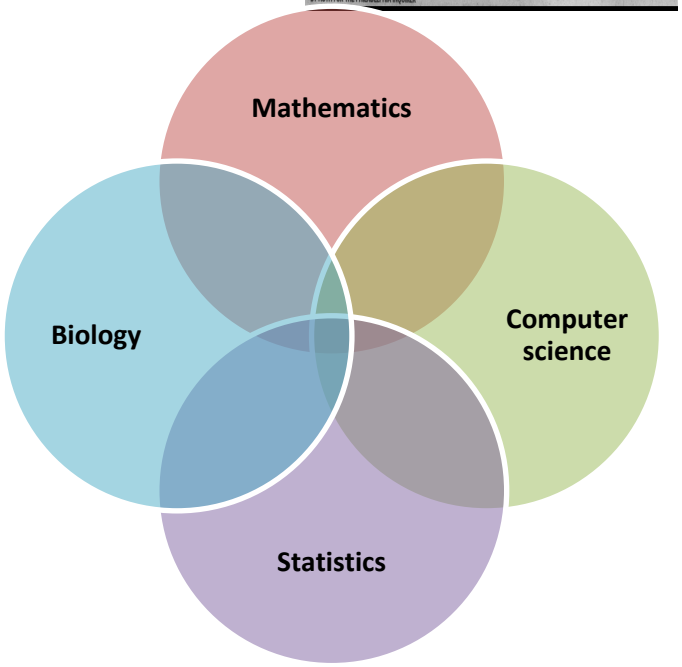
Le **variazioni dell’espressione** di uno o più geni possono portare a **de-regolazioni** dei meccanismi cellulari e possono essere **associate a fenotipi**, quali ad esempio l’obesità



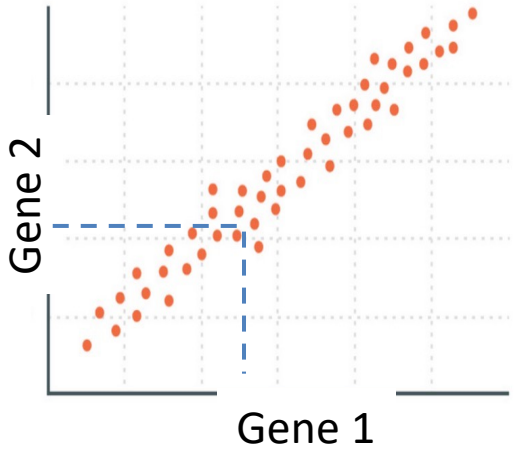
Cosa può fare la matematica?

- Sviluppare **metodi computazionali** ed **algoritmi** per assemblare dati con grande affidabilità ed efficienza
- **Decriptare l'informazione** contenuta nei genomi, trascrittomi, etc., ed identificare differenze e strutture all'interno di tali dati
- **Estrarre informazioni e conoscenza** dai dati attraverso la loro **analisi statistica**
- **Integrare** diversi tipi di dati ed informazioni in **sistemi intelligenti**
- Sviluppare **modelli dinamici** e sistemi computazionali in grado di **simulare** l'effetto di specifiche variazioni, l'azione di farmaci e terapie, la progressione di patologie, etc





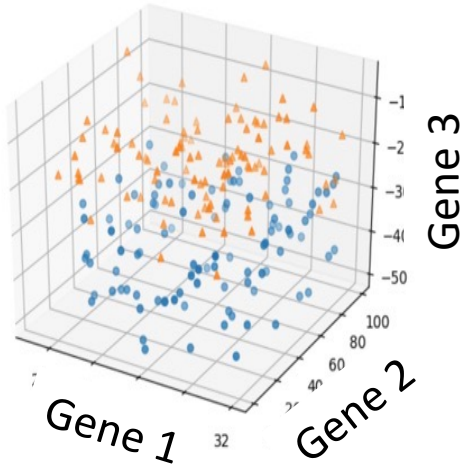
Visualizzazione dei dati



Campione	Gene 1	Gene 2
Individuo 1	17	14
Individuo 2	48	56
Individuo 3	20	25
……	…	…
Individuo **N**	33	47

Rappresentazione di una tabella di dati di **2** dimensioni

Ogni punto rappresenta un individuo.
La posizione del punto dipende dalle variabili misurate



Campione	Gene 1	Gene 2	Gene 3
Individuo 1	17	14	20
Individuo 2	48	56	34
Individuo 3	20	25	15
……	…	…	…
Individuo **N**	33	47	27

Rappresentazione di una tabella di dati di **3** dimensioni

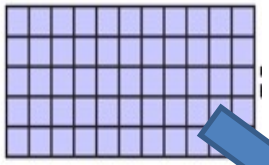
Dati in Alta dimensione

- Come rappresentare i dati quando il numero delle dimensioni (variabili) è maggiore di 3?
- Come rappresentare i dati quando il numero delle dimensioni **P** è estremamente grande?

→ **La maledizione della Dimensionalità**

Matrice **X** dei dati

P: Numero di dimensioni o variabili



Campione	Gene 1	Gene 2	Gene 3	….	Gene P
Individuo 1	17	14	20	…	…
Individuo 2	48	56	34	…	
Individuo 3	20	25	15	…	
……	…	…	…	…	…
Individuo **N**	33	47	27	…	…

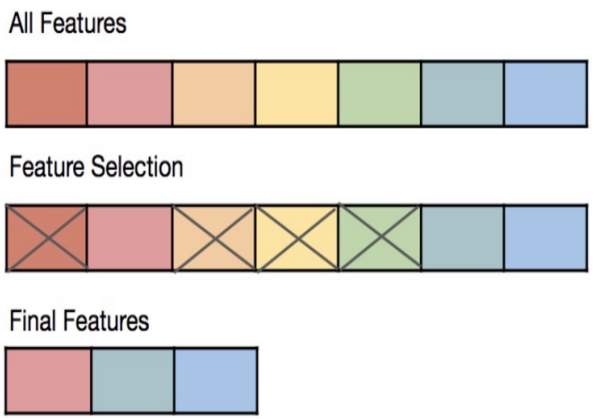
N: Numero di campioni

Quando si misura l'espressione dei geni **P** è circa **30.000**

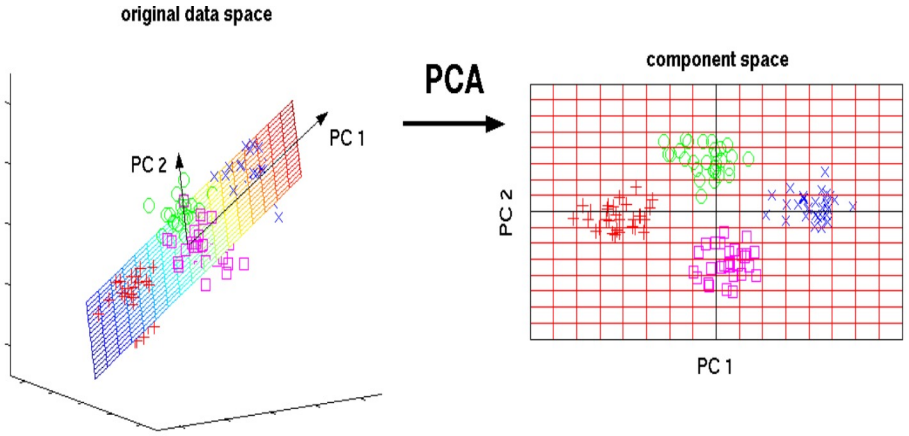
Un individuo/campione può essere visto come un punto in uno spazio di **P** dimensioni

Riduzione della Dimensionalità

- Molte variabili sono **poco rilevanti** oppure contengono informazione ridondante → **solo alcune sono importanti**, oppure si possono opportunamente combinare
- Si cerca di **approssimare** i dati in **spazi di dimensione inferiore** (solitamente anche molto piccola) cercando di catturare la maggior parte delle informazioni



Selezioni delle variabili o features



Proiezione in sotto-spazi di dimensione inferiore

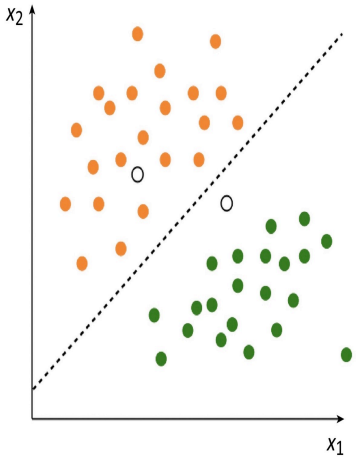
Festival della Scienza

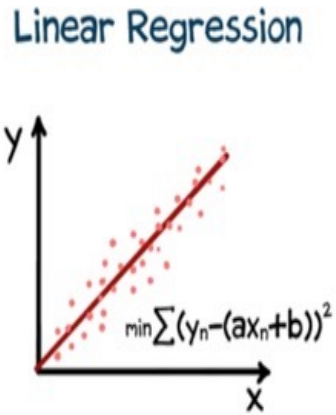
L'apprendimento statistico

CNR Campania REte Outreach

Apprendimento supervisionato

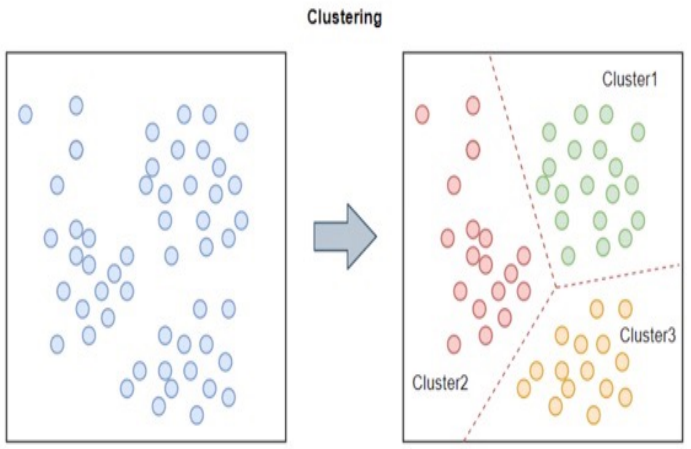
Classificazione





Consente di fare **predizione** su nuovi dati **dopo aver imparato** una regola a partire da **esempi noti**, ad esempio la **classificazione** consente di riconoscere elementi simili ad esempi precedentemente appresi.

Apprendimento non supervisionato



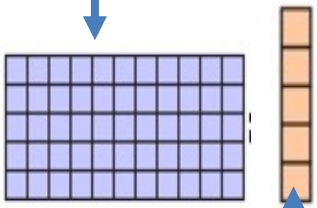
Consente di individuare **strutture** nascoste all'interno di insiemi di dati in cui **non sono note informazioni aggiuntive**, ad esempio il **clustering** consente di raggruppare elementi simili tra loro sulla base di caratteristiche misurate

Consiglio Nazionale delle Ricerche

Classificazione vs Clustering

Classificazione e **Clustering** sono due tra i più importanti esempi di apprendimento **supervisionato** e **non supervisionato**

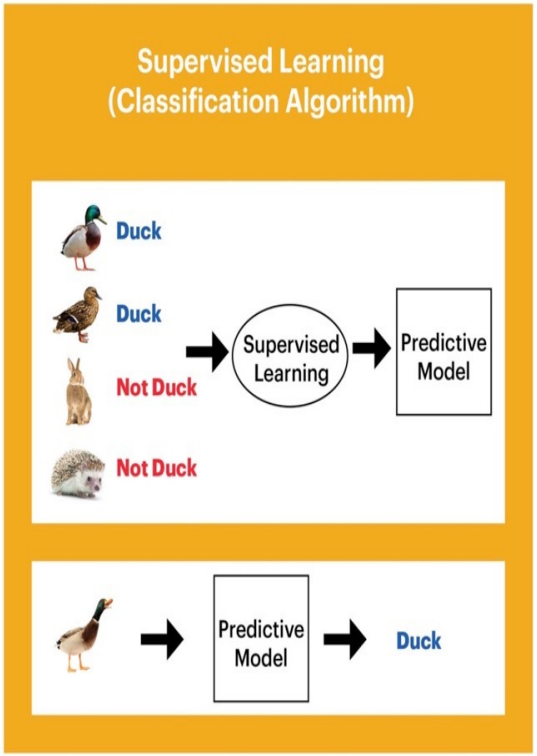
Matrice **X** dei dati

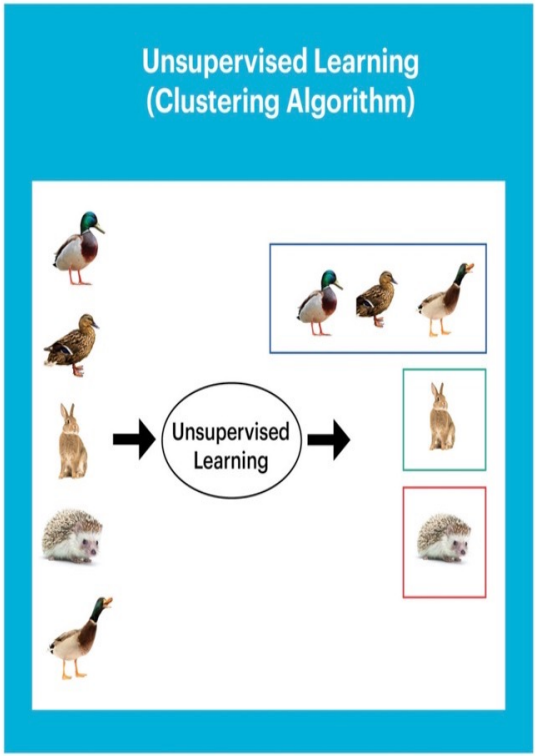


Vettore responso o delle class-labels

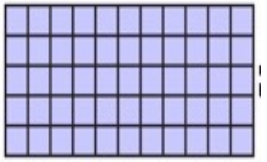
Richiedono una fase di **training**

Sono solitamente seguiti da una fase di **predizione**





Matrice **X** dei dati

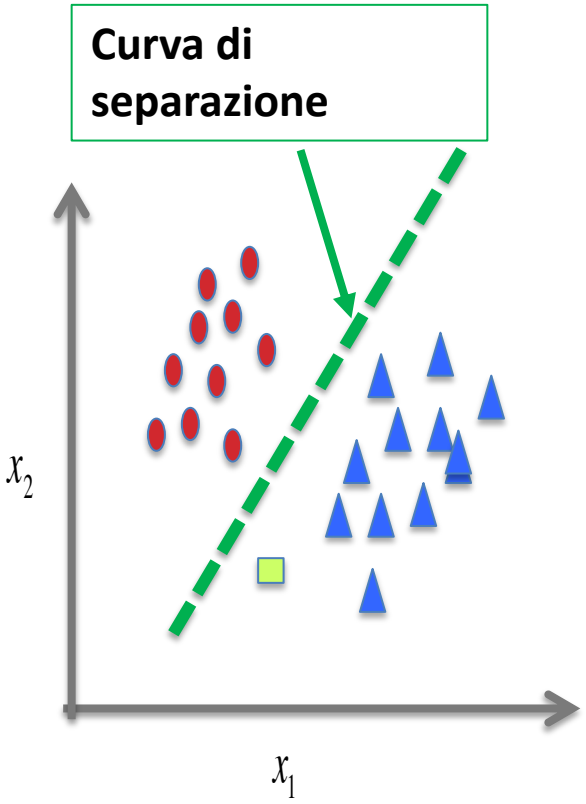


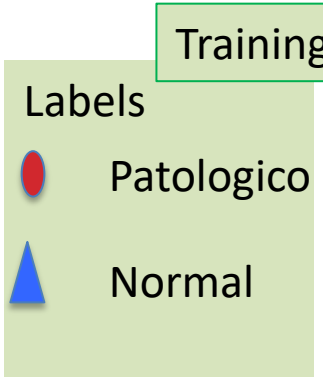
La classificazione applicata allo studio delle patologie

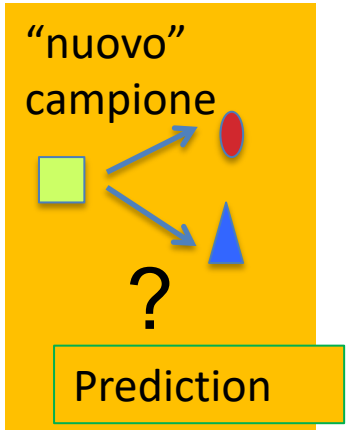
1) A partire dal profilo genetico si individua un sottoinsieme di geni in grado di discriminare due o più tipi di patologie (**biomarcatori**)
2) Si **addestra il classificatore** a riconoscere i particolare tipo di patologia,
3) Si **valida** il classificatore con nuovi dati misurandone l'accuratezza

Una volta validato il classificatore darà in gradi di fare previsioni ….

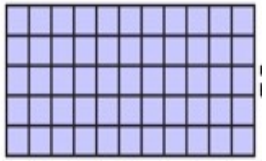
→ le previsioni dovranno essere estremamente accurate per poter essere utilizzate in campo clinico







Il Clustering



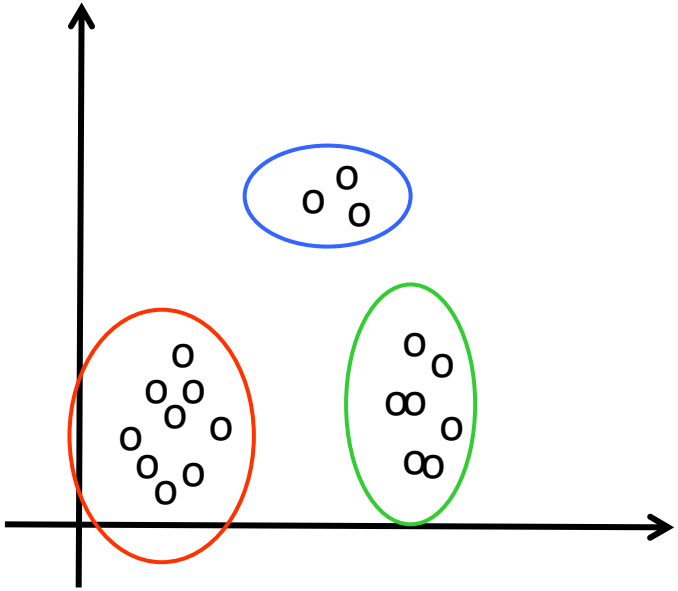
X: matrice **N** campioni con **p** caratteristiche

Una clusterizzazione è una suddivisione delle **n** campioni in **K gruppi**, tale che

✓I campioni appartenenti ad uno **stesso gruppo** devono risultare **simili** tra loro

✓I campioni appartenenti a **gruppi diversi** devono risultare **differenti**

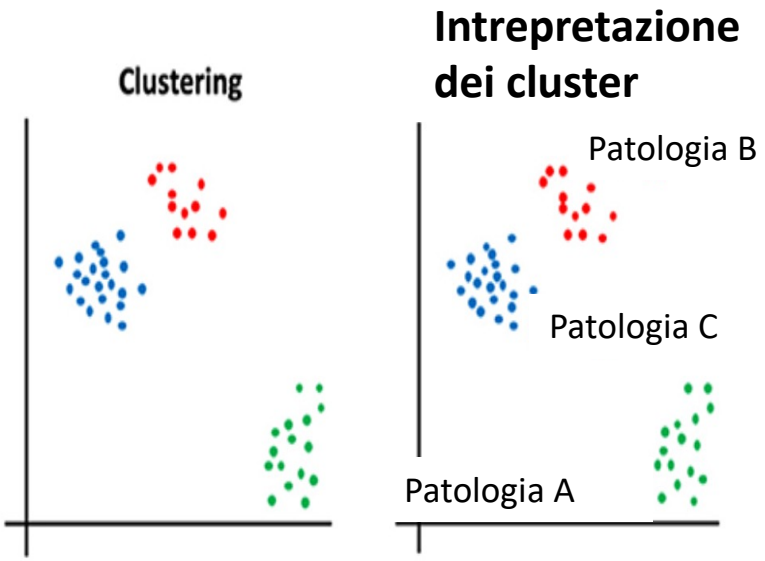
❑ In genere si assume anche che ogni campione appartenga ad uno (ed uno solo) gruppo (mutua esclusione).

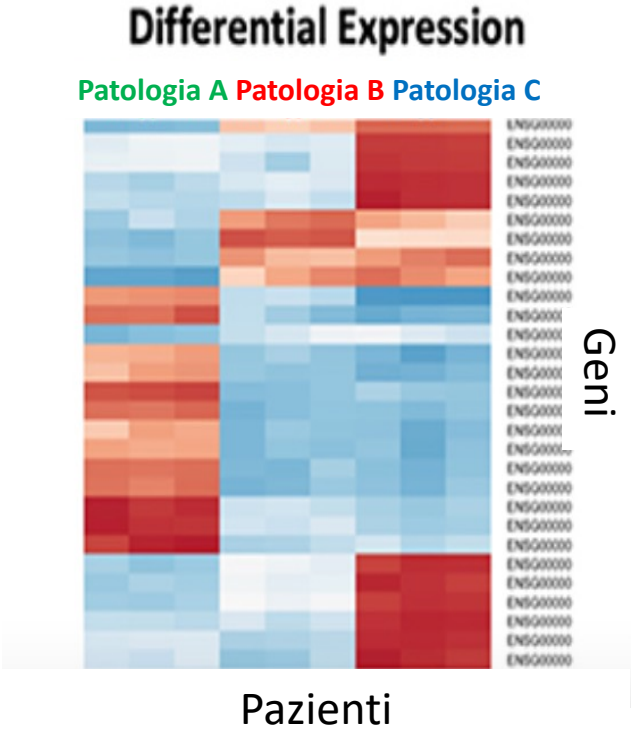


- A differenza della classificazione non sono noti esempi con labels su cui imparare,
- Si cerca un raggruppamento "*naturale*" a partire dalla matrice **X** dei dati
- Inoltre, in generale non è noto neanche il numero K di gruppi che sono presenti

Il clustering applicato allo studio delle patologie

1) A partire dal **profilo genetico** si individua un sottoinsieme di geni in grado di meglio **discriminare** i campioni
2) Si utilizza un algoritmo per individuare il **numero di clusters** più idoneo e successivamente **per assegnare i campioni ai diversi cluster**
3) Si cerca di **interpretare** i clusters ottenuti da un punto di vista biologico





E' possibile individuare sotto-tipi di una data patologia, separare pazienti che rispondono ad una terapia da quelli resistenti, etc