

Peschiamo nel mare della complessità...



Maria Francesca Carfora

Istituto per le Applicazioni del Calcolo *M. Picone*

Consiglio Nazionale delle Ricerche



Ciliegia di Vignola



La Ciliegia di Vignola IGP si riferisce ai frutti freschi delle seguenti cultivar di ciliegio: Early Bigi e Lory, ..., Mora di Vignola (precoci), ... Durone Nero, Black Star, ... Summer Charm (Staccato) (tardive). La zona di produzione della Ciliegia di Vignola IGP interessa alcuni comuni delle province di Modena e Bologna posti nella fascia altimetrica che va dai 30 fino ai 950 metri s.l.m., nella regione Emilia-Romagna.

La Ciliegia di Vignola IGP è un frutto dal sapore dolce e fruttato. La sua polpa è consistente e croccante, ... La buccia ha un colore che varia dal rosso brillante al rosso scuro, tranne che per la varietà Durone della Marca, dove invece risulta sempre lucente ma di colore giallo e rosso brillante. I calibri minimi sono di 20-23 mm a seconda della varietà, mentre i massimi possono superare 28 mm.





Ciliegia di Vignola



**fenotipo/
genotipo**

Info geografiche

Aspetto, sapore...

La Ciliegia di Vignola IGP è un frutto dal **sapore dolce e fruttato**. La **sua polpa è consistente e croccante**, ... La **buccia ha un colore che varia** dal rosso brillante al rosso scuro, tranne che per la varietà Durone della Marca, dove invece risulta sempre lucente ma di colore giallo e rosso brillante. I **calibri** minimi sono di 20-23 mm a seconda della varietà, mentre i massimi possono superare 28 mm.

Morfometria



Un esempio di gestione ed analisi di dati complessi

OLIVEHEALTH

[HOMEPAGE](#)

[ATTIVITA'](#)

[PARTENARIATO](#)

[NEWS](#)

[ARCHIVIO NEWS](#)

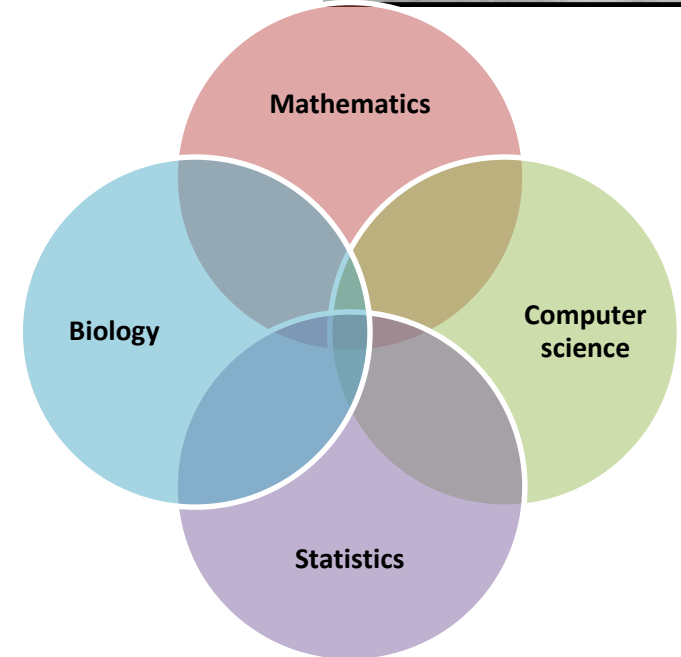
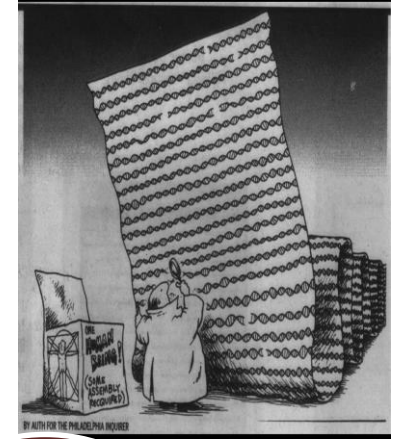
[CONTATTI](#)



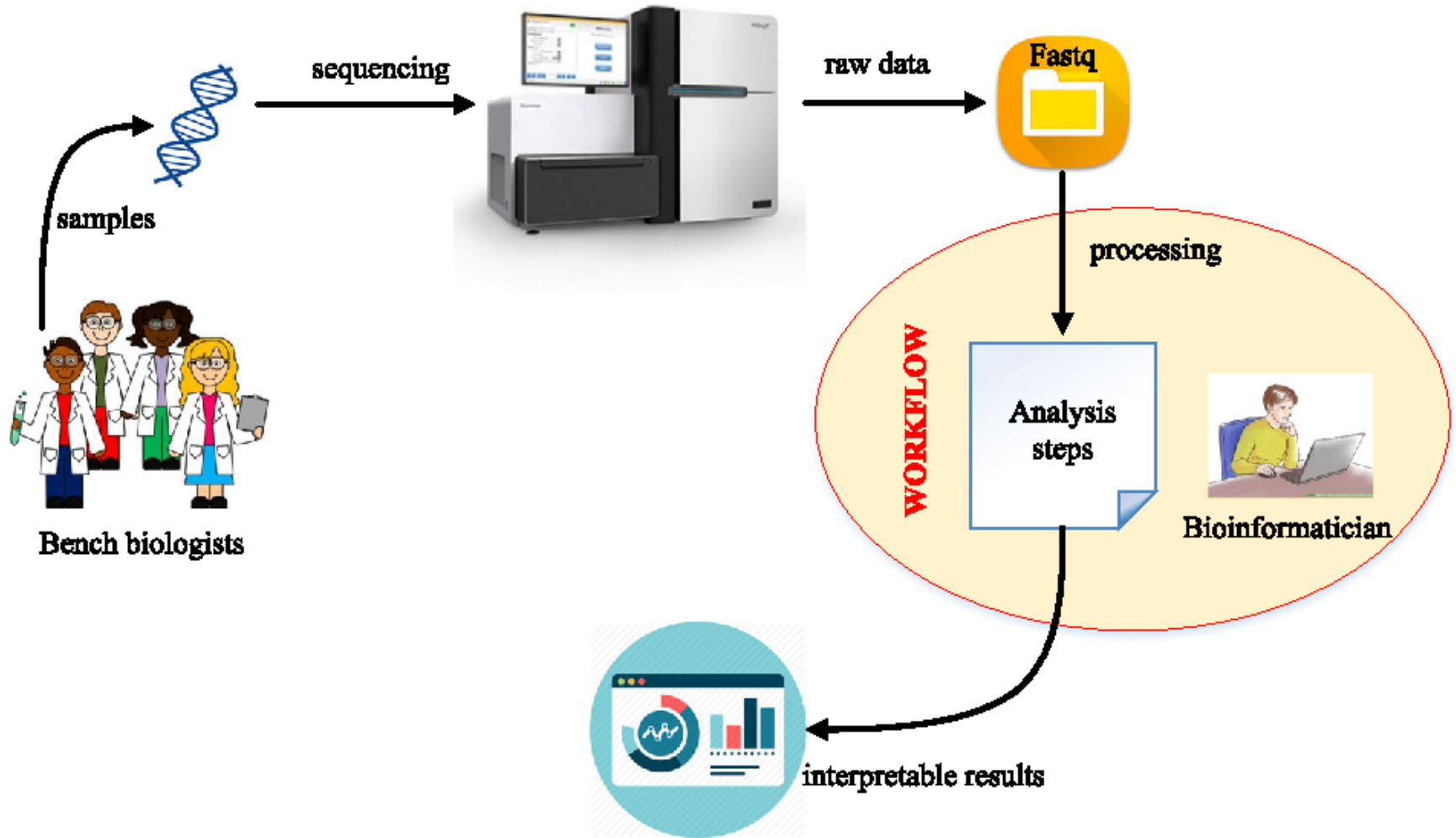
Identificazione delle componenti salutistiche della
filiera olivicola campana

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



Analisi di dati genici



Sequenziatori



2007-2012

2019-

**Come mettere
insieme tutta questa
informazione?**



Sequenziamento

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

**Milioni o
centinaia di
milioni di
sequenze**



Banche dati e software

Hit Table						
Query Centric View						
Distances						
Info						
☒	E Value ▲	% Pairwise Id...	Name	Organism	Bit-Score	Description
☒	2.86e-15	100.0%	AF483338	Raphus cucullatus	89.651	Raphus cucullatus cytochrome b (c
☒	1.22e-13	98.0%	KX902236	Raphus cucullatus	84.241	Raphus cucullatus mitochondrion,
☒	6.30e-11	93.9%	AY090673	Aegothales bennettii	75.224	Aegothales bennettii clone cb.232
☒	6.30e-11	93.9%	AY090675	Aegothales bennettii	75.224	Aegothales bennettii clone cb.632
☒	6.30e-11	93.9%	AY090676	Aegothales bennettii	75.224	Aegothales bennettii clone cb.632
☒	6.30e-11	93.9%	AY090680	Aegothales bennettii	75.224	Aegothales bennettii clone cb.339
☒	6.30e-11	93.9%	X95774	Aegothales bennettii	75.224	Aegothales bennettii mitochondria
☒	7.67e-10	91.8%	AB276048	Phoebastria immutabilis	71.618	Phoebastria immutabilis mitochond
☒	7.67e-10	91.8%	AB276049	Phoebastria immutabilis	71.618	Phoebastria immutabilis mitochond
☒	7.67e-10	91.8%	AB276050	Phoebastria immutabilis	71.618	Phoebastria immutabilis mitochond
☒	7.67e-10	91.8%	AB276051	Phoebastria nigripes	71.618	Phoebastria nigripes mitochondria
☒	7.67e-10	91.8%	AB426114	Phoebastria nigripes	71.618	Phoebastria nigripes mitochondria
☒	7.67e-10	91.8%	AB426115	Phoebastria nigripes	71.618	Phoebastria nigripes mitochondria
☒	7.67e-10	91.8%	AB426116	Phoebastria nigripes	71.618	Phoebastria nigripes mitochondria
☒	7.67e-10	91.8%	AF082066	Mycteria americana	71.618	Mycteria americana isolate B5670
☒	7.67e-10	91.8%	AY274042	Treron sieboldii	71.618	Treron sieboldii NADH dehydroge

Selected sequences are only summaries [Download Full Sequence\(s\)](#)

[Alignment View](#) [Dotplot](#) [Dotplot \(Self\)](#) [DNA Fold](#) [Distances](#) [Text View](#) [Download](#) [Lineage](#) [Info](#)

Displays the text view of documents of

Extract R.C. Translate Add/Edit Annotation Allow Editing Annotate & Predict Save

Consensus Identity

1 10 20 30 40 49

1. unknown CATCCGTTGCCACACATGTCGTGATGTACAGTACGGCTGATTAATCCG

2. AF483338 CATCCGTTGCCACACATGTCGTGATGTACAGTACGGCTGATTAATCCG

General

Colors: ACGT

☒ Graphs Options

☒ Annotations Options

☒ Consensus Options

☒ Highlighting Options

☐ Complement Options

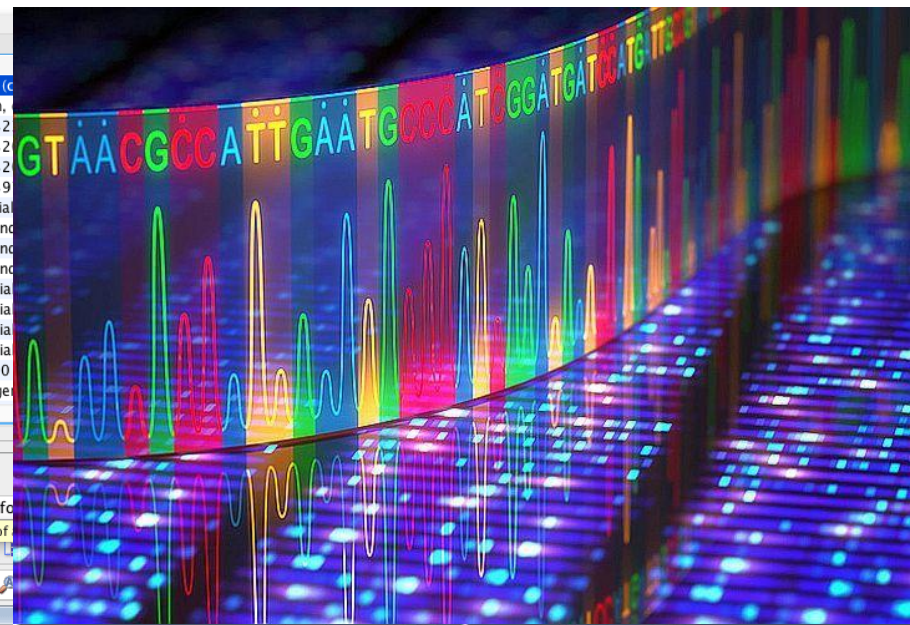
☐ Translation

☐ Wrap

☒ Show Names

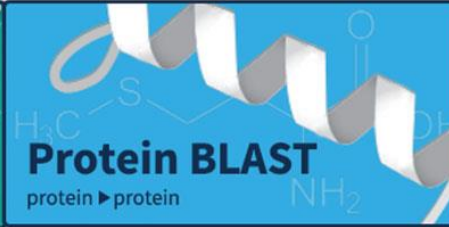
☐ Show Description

Alt click on a sequence position or annotation, or select a region to zoom in. Alt-shift click to zoom out.





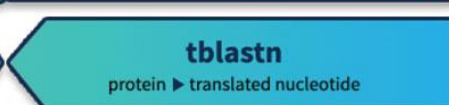
Nucleotide BLAST
nucleotide ► nucleotide



Protein BLAST
protein ► protein



blastx
translated nucleotide ► protein



tblastn
protein ► translated nucleotide

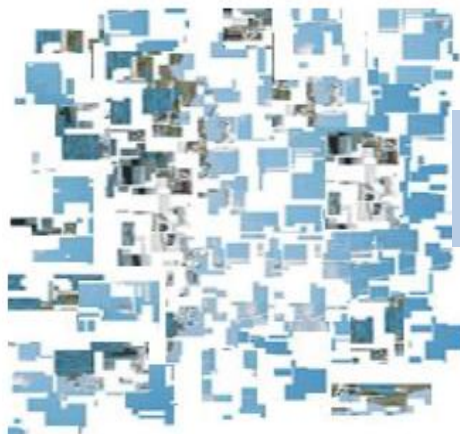


Risolvere un puzzle

Riferimento



Dati



Assemblamento



Allineamento



Algoritmi

Risolvere il DNA puzzle

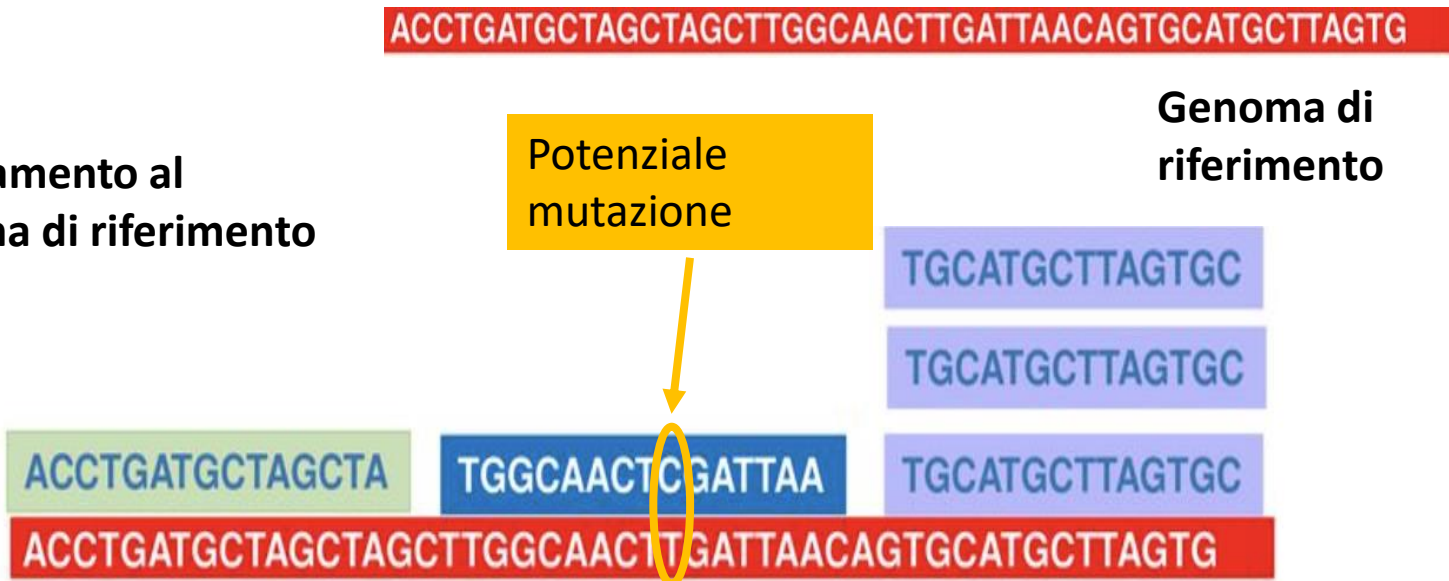
Negli studi sul **genoma umano** si devono allineare **decine di milioni o centinaia di milioni di sequenze** su un riferimento che ha circa **3 miliardi di basi**



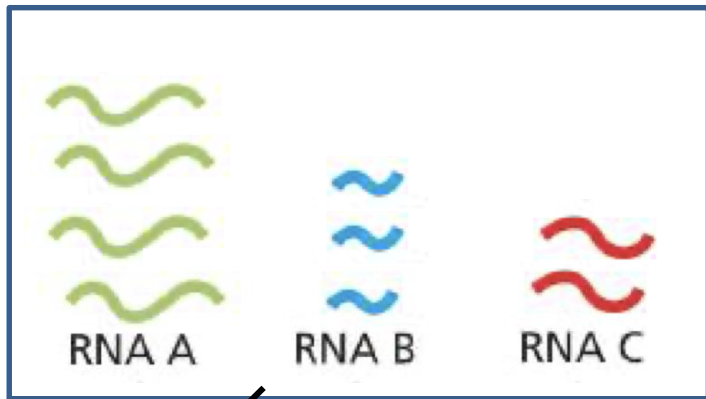
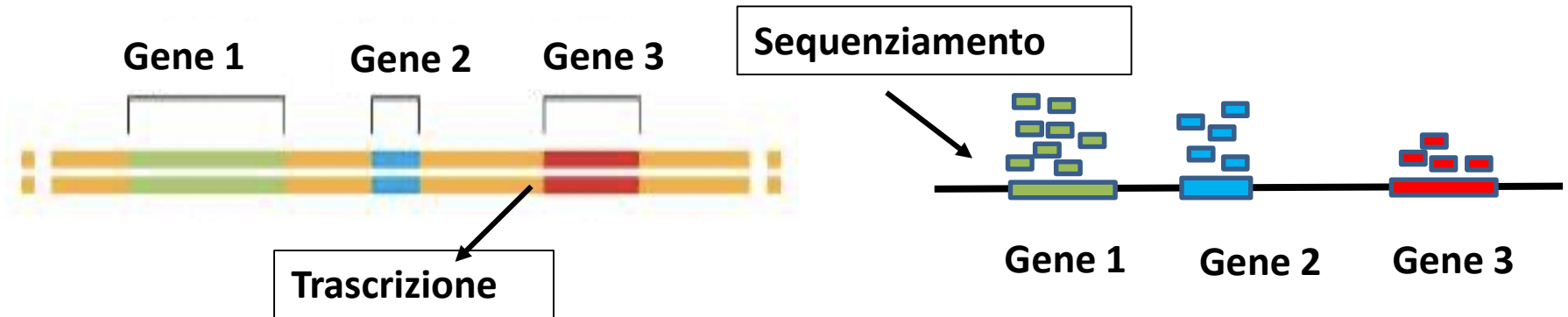
Allineamento al
genoma di riferimento

Potenziale
mutazione

Genoma di
riferimento



La misurazione dell'espressione genica



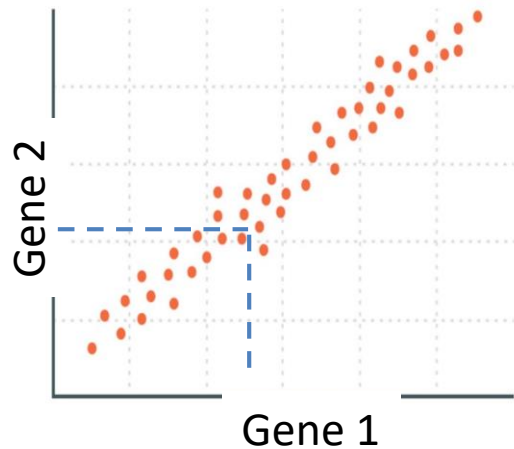
Gene 1	Gene 2	Gene 3	-----
4	3	2	

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**



Visualizzazione dei dati

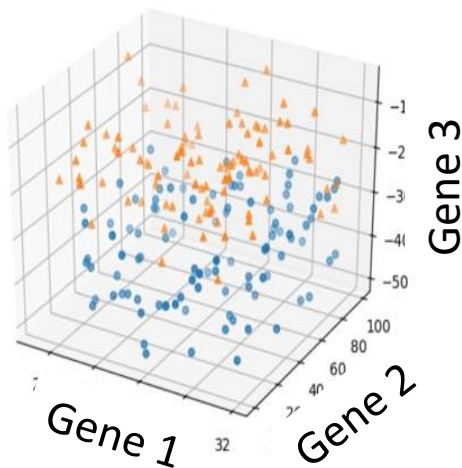


Campione	Gene 1	Gene 2
Individuo 1	17	14
Individuo 2	48	56
Individuo 3	20	25
.....	...	...
Individuo N	33	47

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



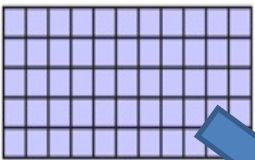
- Come rappresentare i dati quando il numero delle dimensioni (variabili) è maggiore di 3?
- Come rappresentare i dati quando il numero delle dimensioni **P** è estremamente grande?

→ 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

All Features



Feature Selection

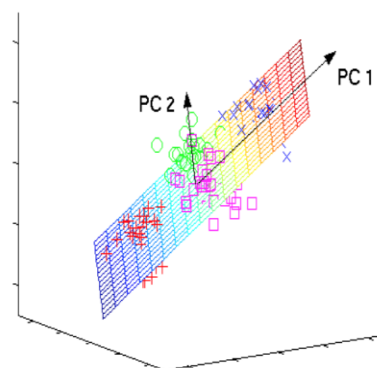


Final Features



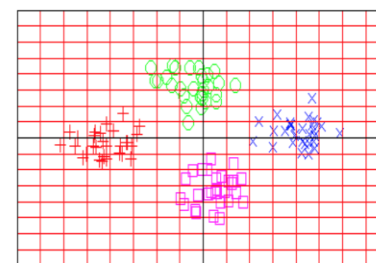
Selezioni delle
variabili o features

original data space



PCA

component space



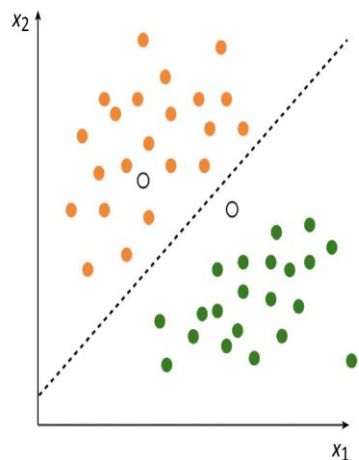
PC 1

Proiezione in sotto-spazi
di dimensione inferiore

L'apprendimento statistico

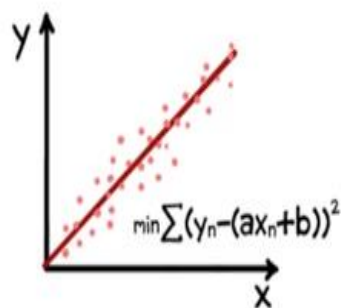
Apprendimento supervisionato

Classificazione



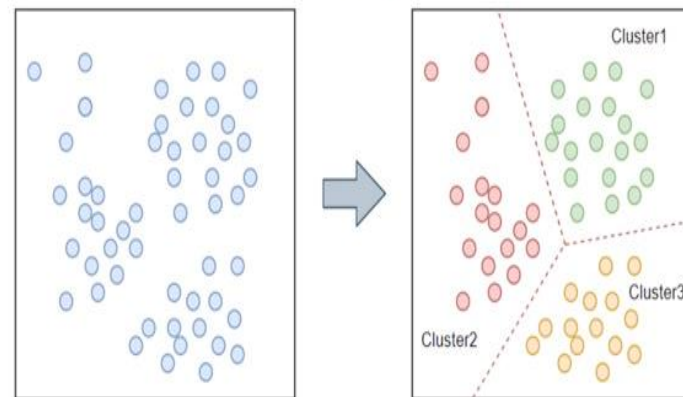
Consente di fare **predizione** su nuovi dati **dopo aver imparato** una regola a partire da **esempi noti**, ad esempio la **classificazione** consente di riconoscere elementi simili ad esempi precedentemente appresi.

Linear Regression



Apprendimento non supervisionato

Clustering



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