Reti neurali e medicina di precisione

Biennale Tecnologia,
Torino 19/04/2024

Michele Caselle – University of Torino and INFN
caselle@to.infn.it

Patologie complesse e medicina di precisione

Come in altri campi anche in medicina è iniziata da alcuni anni l'era della "medicina di precisione" che richiede grandi moli di dati e sofisticati programmi per estrarre informazioni da questi dati.

Questo approccio è particolarmente importante per le cosiddette "patologie complesse" dovute a molti fattori e che hanno caratteristiche diverse da individuo a individuo.

I metodi che si usano per studiare queste patologie combinano idee e strumenti della Fisica dei Sistemi Complessi e del Machine Learning e si basano sui più recenti risultati di Biologia Molecolare.

Medicina di precisione per il cancro

L’esempio principale di patologia complessa è il cancro. In questo ambito la Medicina di precisione sta ottenendo successi insperati.

L’obiettivo è, partendo dalla biopsia del tumore, identificare il particolare sottotipo tumorale del paziente, predire il grado di pericolosità del tumore, ottimizzare il trattamento terapeutico e prevedere il rischio di resistenza ai farmaci o di recidiva del tumore.

E, soprattutto, fare tutto questo il più rapidamente possibile

Medicina di precisione per il cancro

L’obiettivo di questo intervento è raccontarvi come funziona tutto questo e lo farò attraverso tre passi:

- un breve ripasso di biologia
- un poco di Data Mining e di Fisica dei Sistemi Complessi
- un paio di esempi di utilizzo delle reti neurali per risolvere problemi complessi

Geni e Proteine

Ogni cellula del nostro corpo ha esattamente lo stesso DNA (lo stesso contenuto di geni) di tutte le altre cellule. Ma cellule di organi diversi possono essere anche molto diverse tra loro.

Questo è dovuto al fatto che le funzioni della cellula sono svolte non direttamente dai geni ma dalle proteine che vengono prodotte usando i geni come stampo.

Questo processo è detto espressione genica ed è finemente regolato. Dei circa 20.000 geni a disposizone, ogni cellula ne usa solo una parte, esattamente quelli di cui ha bisogno.

Nei tumori questa regolazione salta, la cellula “impazzisce” ed usa geni che dovrebbero stare spenti.

Geni e Proteine

.

Ogni tumore è diverso dagli altri, ci sono infiniti modi diversi in cui una cellula può impazzire.

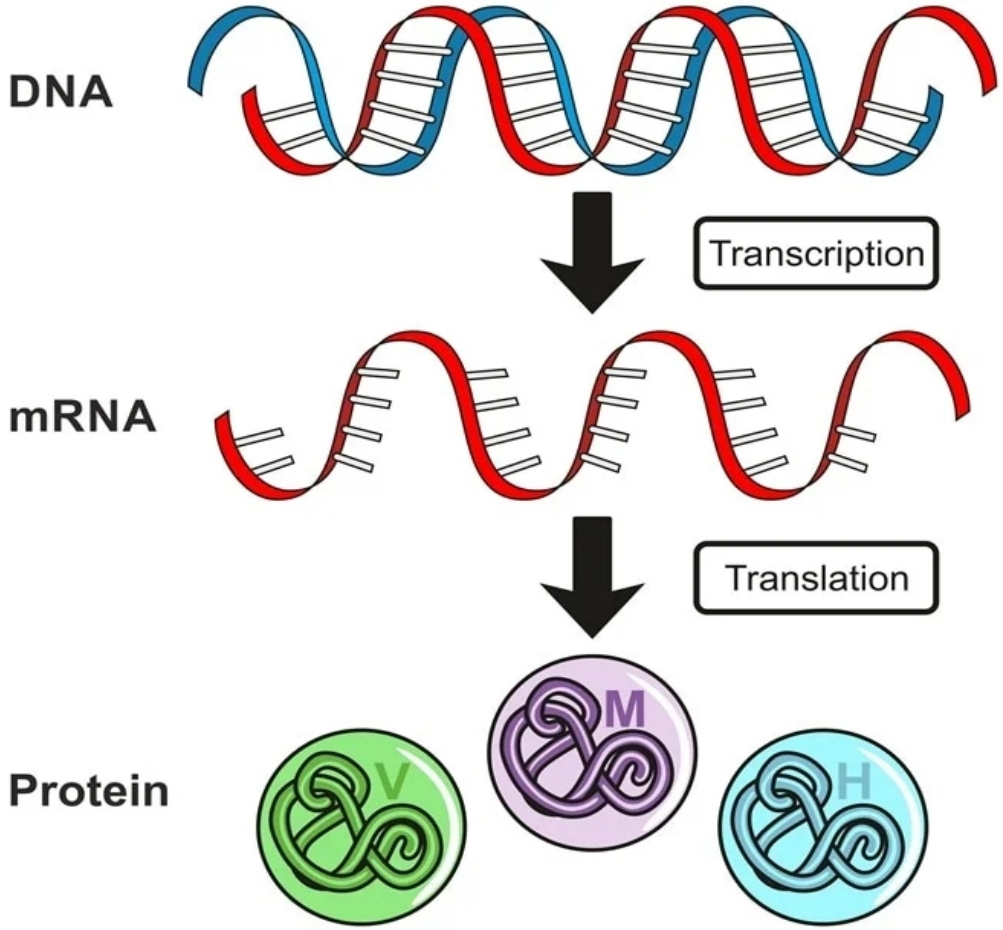
Il nostro obiettivo è capire il più rapidamente possibile quale meccanismo di controllo (di solito più di uno) è saltato nel particolare paziente che stiamo curando.

Espressione genica

La “espressione dei geni” avviene attarverso due passaggi principali:

Trascrizione (dal DNA al mRNA)

Traduzione (dal mRNA alle Proteine)



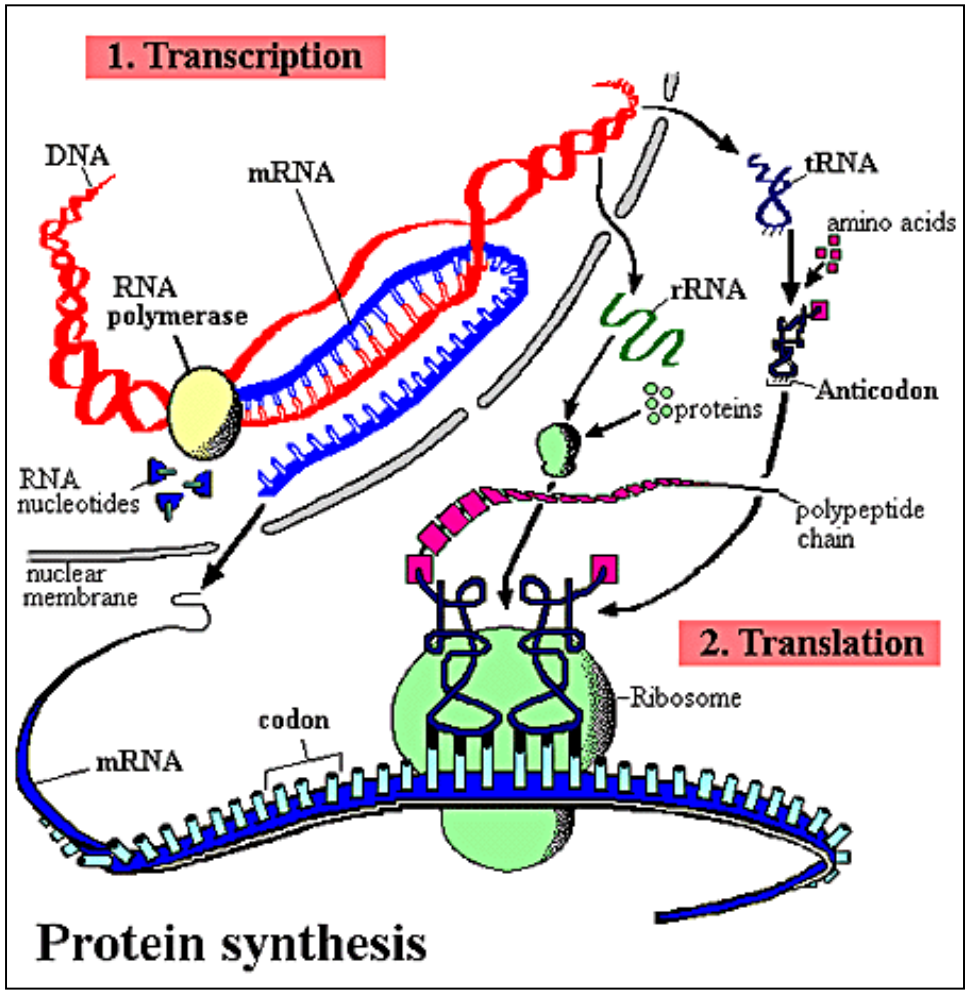
Gene Expression. Image Credit: udaix/Shutterstock.com

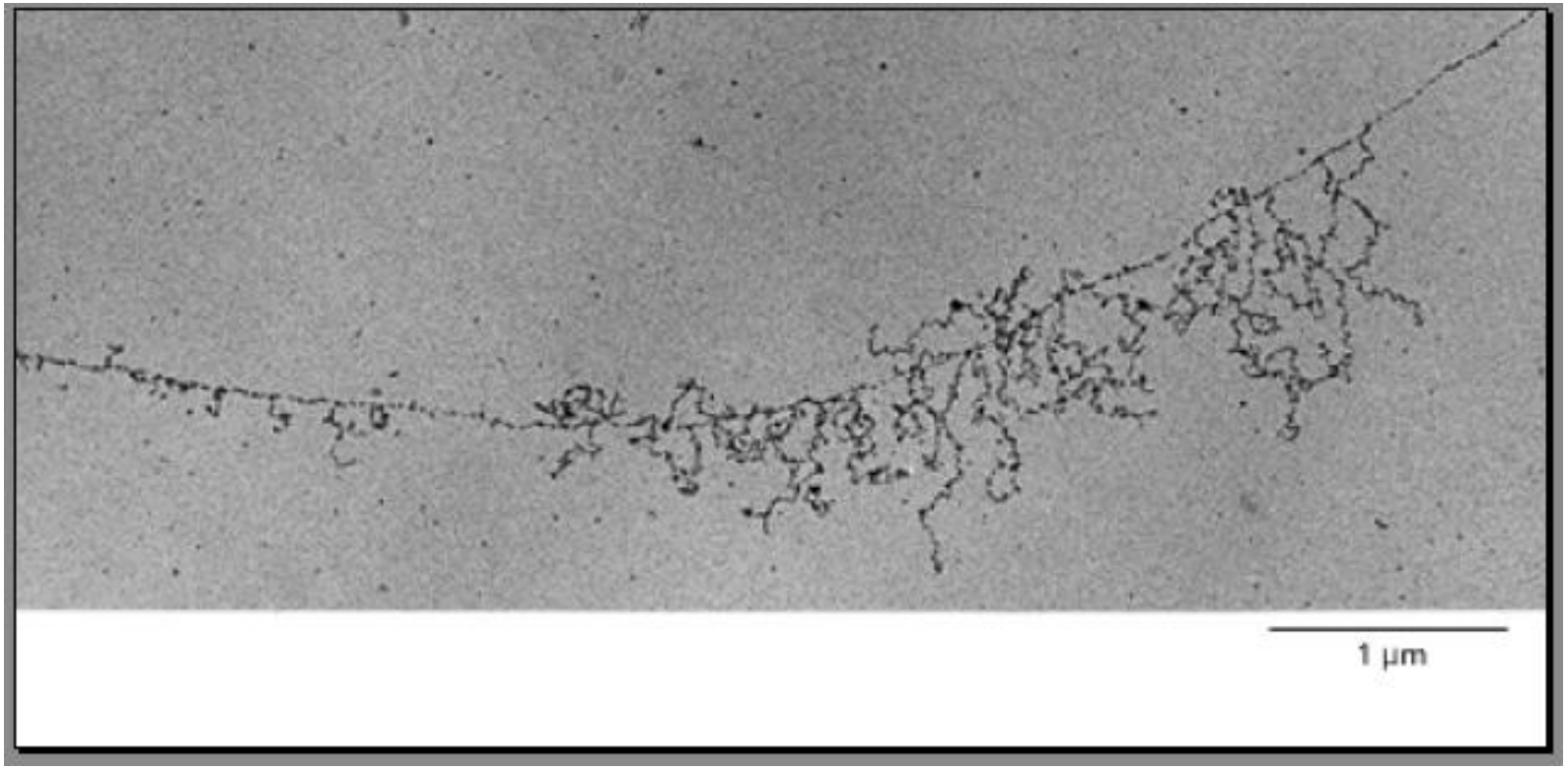
Espressione genica

La “espressione dei geni” avviene attarverso due passaggi principali:

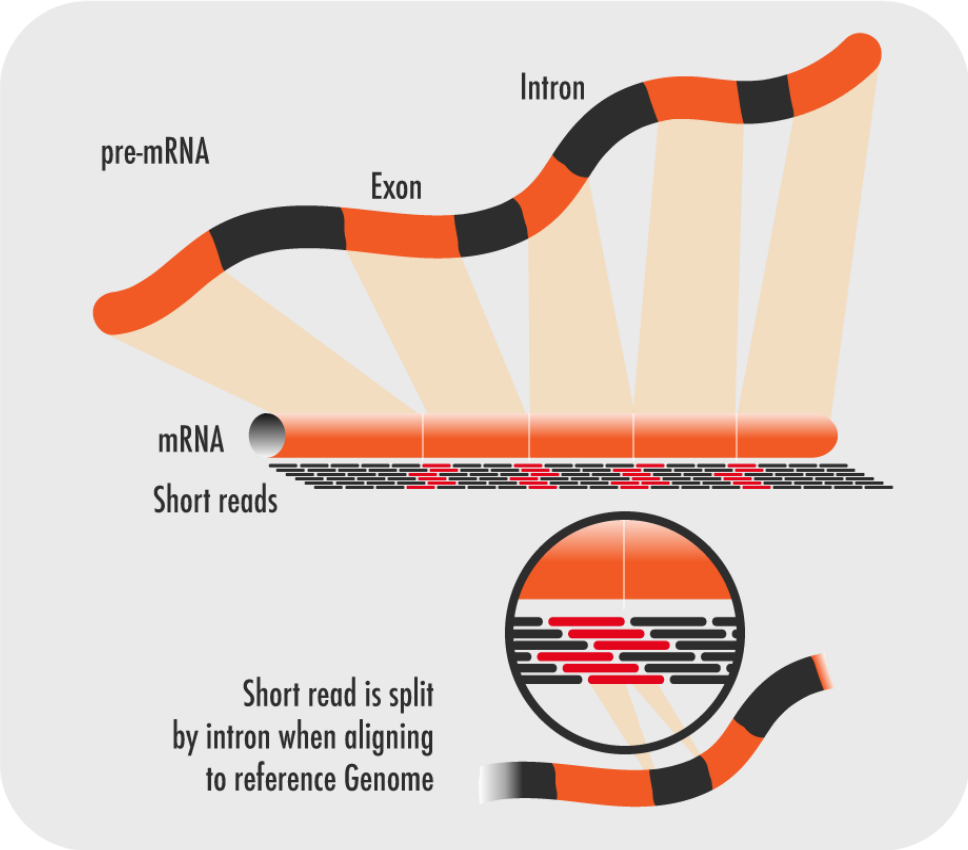
Trascrizione (dal DNA al mRNA)

Traduzione (dal mRNA alle Proteine)




1 µm

RNA-seq



Grazie ad una tecnica chiamata "RNA-seq" è possibile ottenere una "foto" del contenuto di mRNA di un tessuto o addirittura di una singola cellula.

Questa è l'informazione da cui partiamo per il nostro studio

RNA-seq

Pazienti

geni

	P1	P2	P3	P4	P5
	100	10	0	36	21
	45	0	6	124	0
	163	6	0	72	0
	43	32	12	0	2
	0	0	0	0	80

RNA-seq permette di contare quanti mRNA di un particolare gene sono presenti in un particolare momento nel tessuto che stiamo studiando.

Il risultato si presenta come una matrice di numeri che quantificano il livello di espressione di ognuno dei geni in ognuno dei pazienti

RNA-seq

Questi valori numerici vengono di solito rappresentati mediante colori:

rosso= alto livello di espressione,

blu= basso livello.

L'obiettivo della ricerca è riuscire a mettere vicino

pazienti e geni con lo stesso schema di colori, come nella figura a fianco.

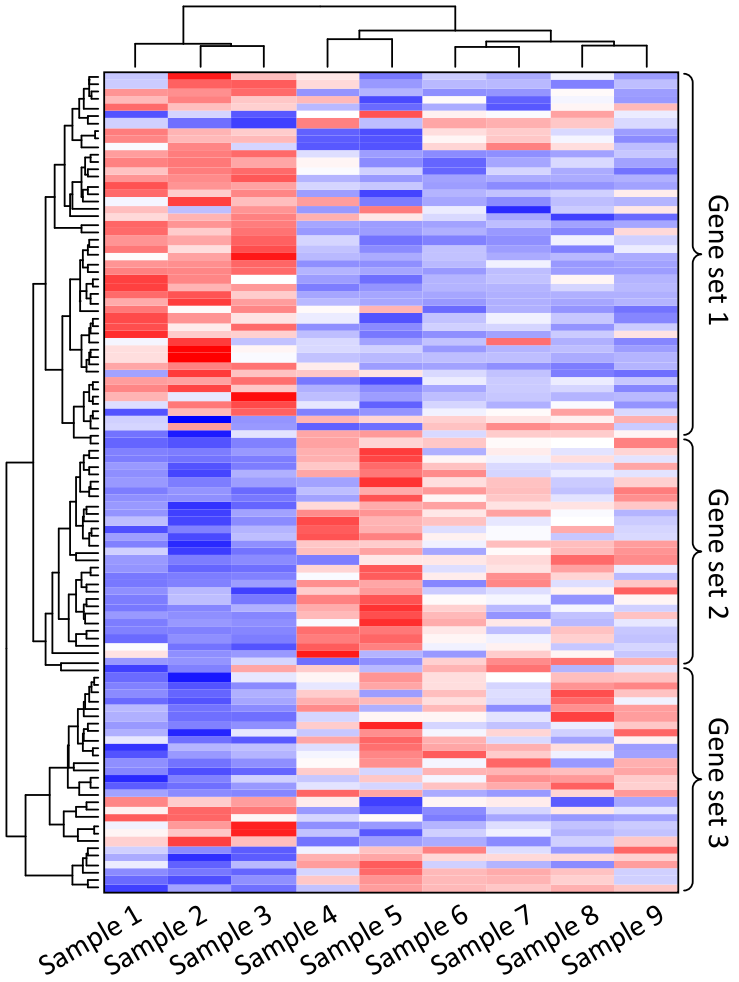
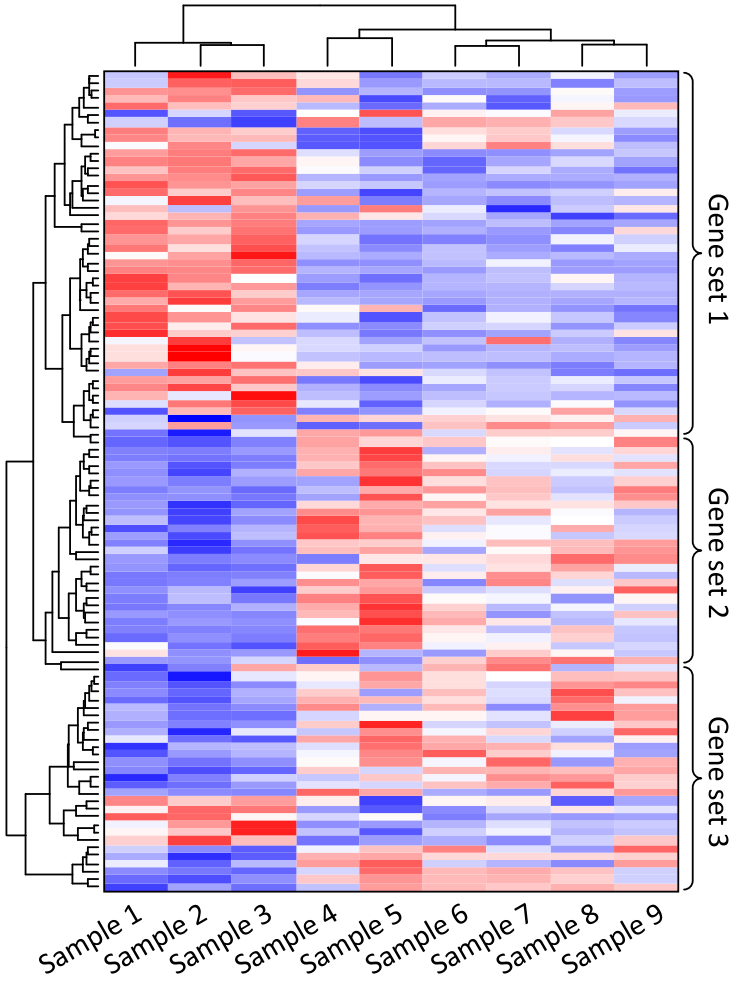


Image by Thomas Shafee - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=56957936

Clustering

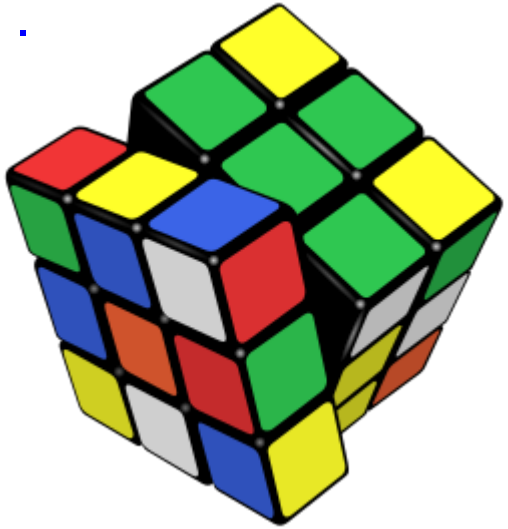
I programmi che svolgono questo compito si chiamano in gergo “*algoritmi di clustering*”

Sono programmi sofisticati che usano metodi e idee tipici del machine learning, in particolare del cosiddetto ”*apprendimento non supervisionato*”



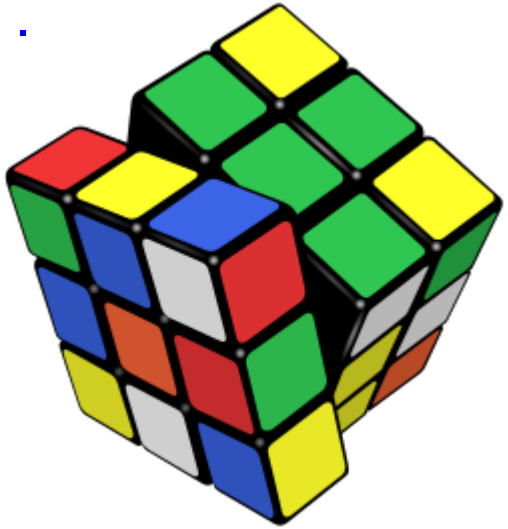
Apprendimento non supervisionato

Raggruppare i dati mettendo vicino i pazienti simili è un po’ come risovere il cubo di Rubik. A ogni mossa cerchiamo di avvicinare quadretti dello stesso colore .



Apprendimento non supervisionato

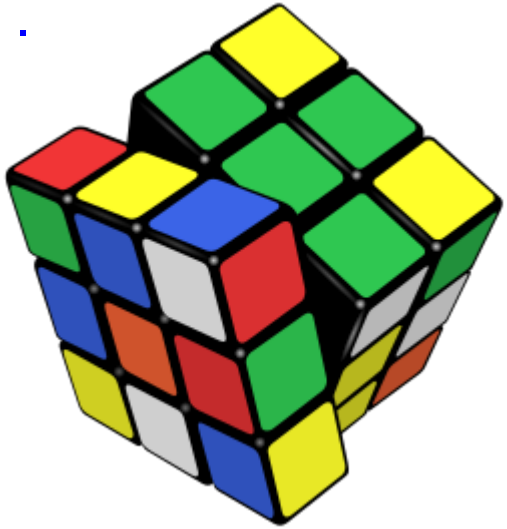
Raggruppare i dati mettendo vicino i pazienti simili è un po’ come risovere il cubo di Rubik. A ogni mossa cerchiamo di avvicinare quadretti dello stesso colore .



---------->

Apprendimento non supervisionato

Raggruppare i dati mettendo vicino i pazienti simili è un po' come risovere il cubo di Rubik. A ogni mossa cerchiamo di avvicinare quadretti dello stesso colore .



---------->

Ma se abbiamo a che fare con **20.000 geni e qualche migliaio di pazienti** non possiamo provare tutte le combinazioni! **Ci vuole un'idea**

Topic Modelling

Un problema simile lo devono affrontare i programmi che riconoscono automaticamente l'argomento di cui parla un testo sulla base solo del loro contenuto di parole.

Il cosiddetto "***Topic Modelling***".

L'analogia è notevole:

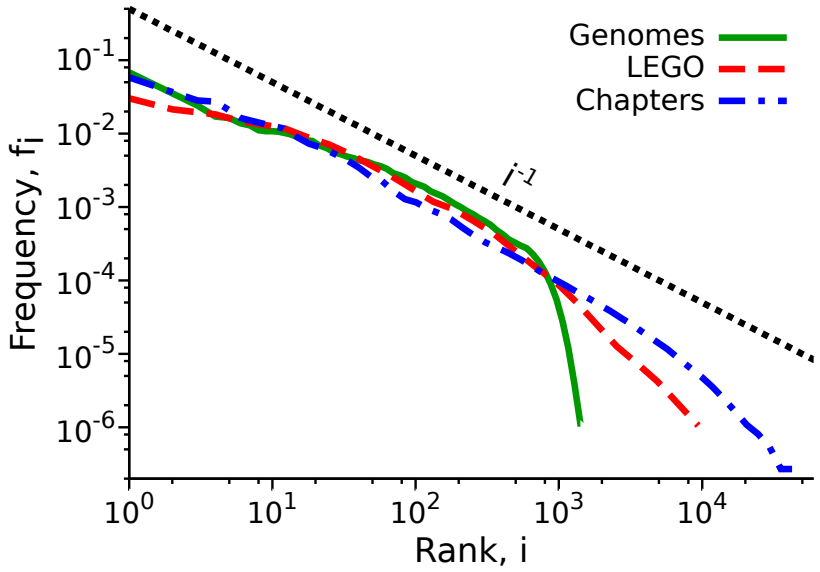
Geni	---------->	Parole
Pazienti	---------->	Testi
Livello di espressione di un gene	---------->	Frequenza con cui compare la parola nel testo

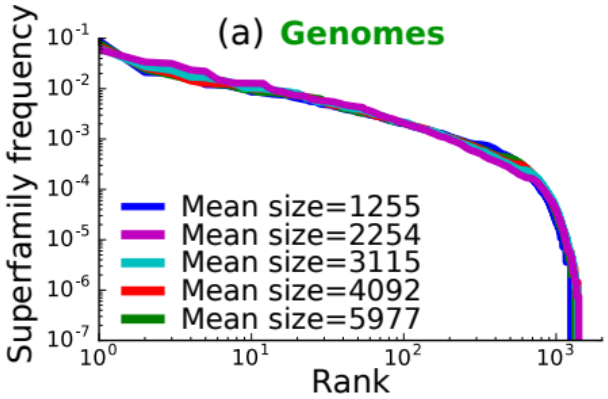
Topic Modelling

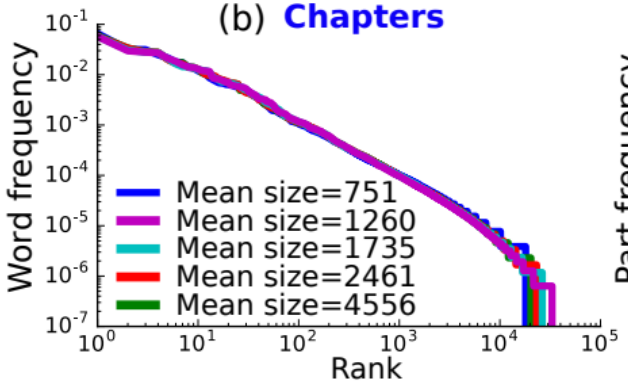
Il programma è in grado di riconoscere documenti simili tra loro (clusters) e fornisce un elenco di argomenti (topics) di cui parlano, fornendo anche la percentuale di "appartenenza" del documento ad un certo argomento.

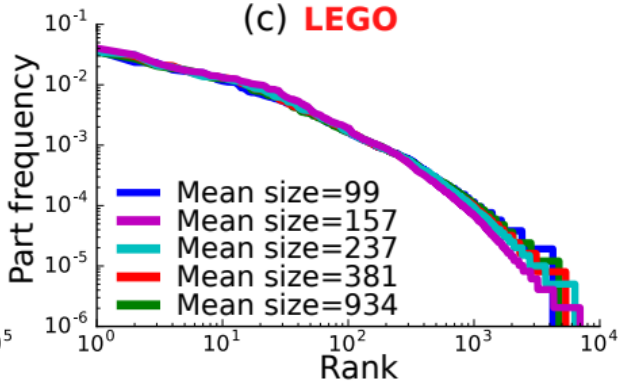
In modo simile se applicato ai deti genomici il programma riconosce pazienti simili tra loro (clusters), distingue una serie di gruppi di geni (topic) che servono a riconoscere queste somiglianze e dice, per ogni gruppo di pazienti qual è il peso relativo dei vari gruppi di geni nel determinare la suddivisione dei pazienti.

Universalità dei Sistemi Complessi

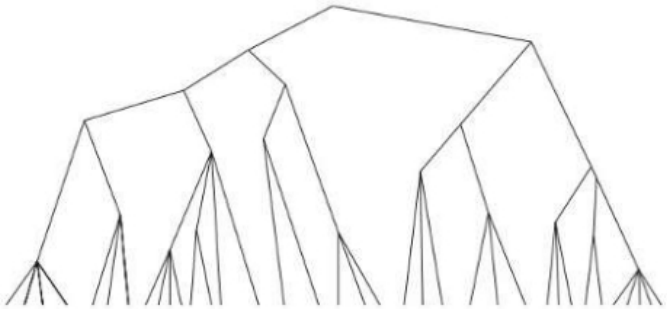


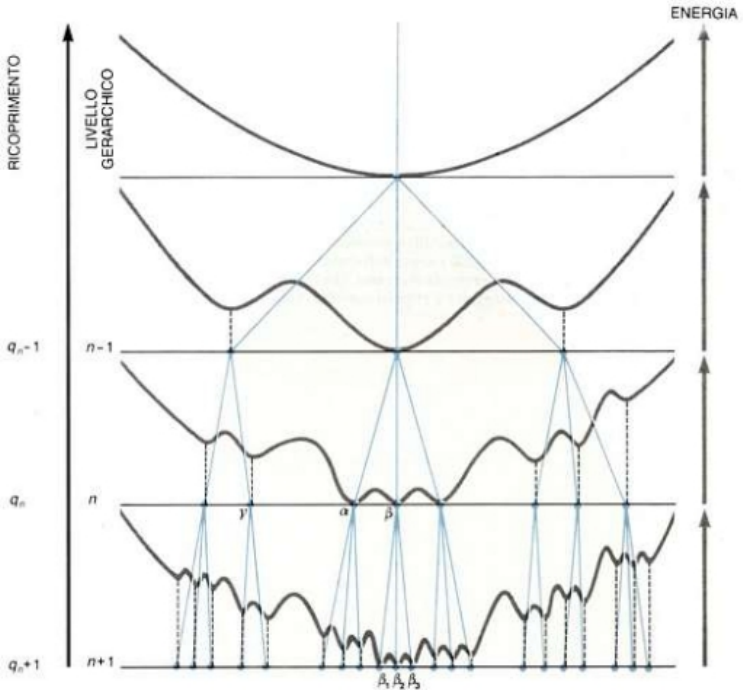


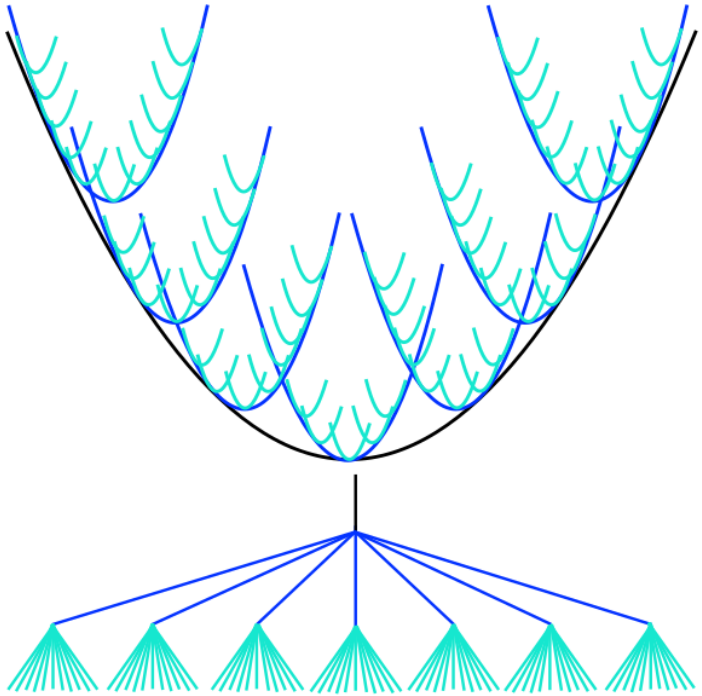




La soluzione di Parisi: trovare un ordine (gerarchico) nel disordine!

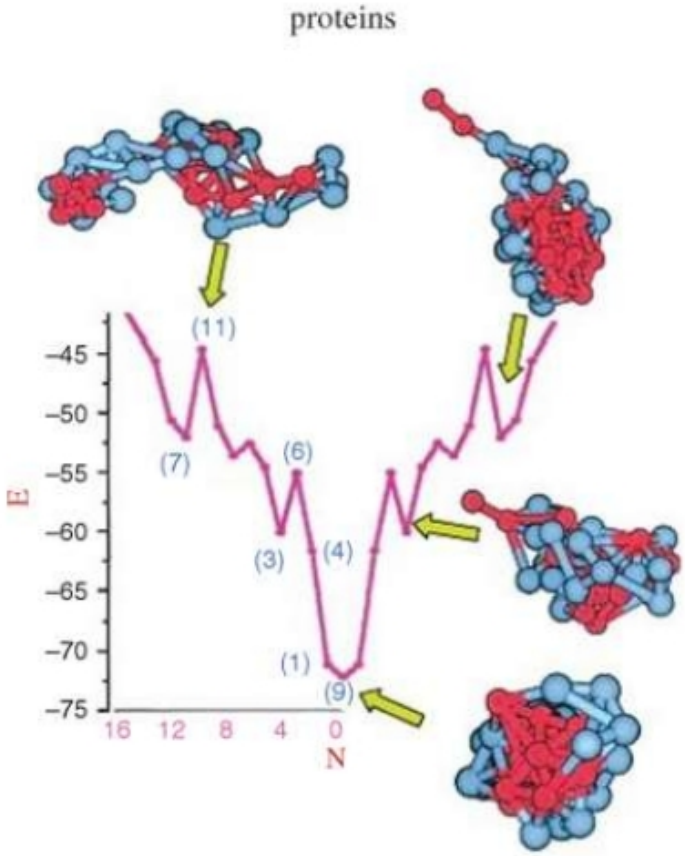


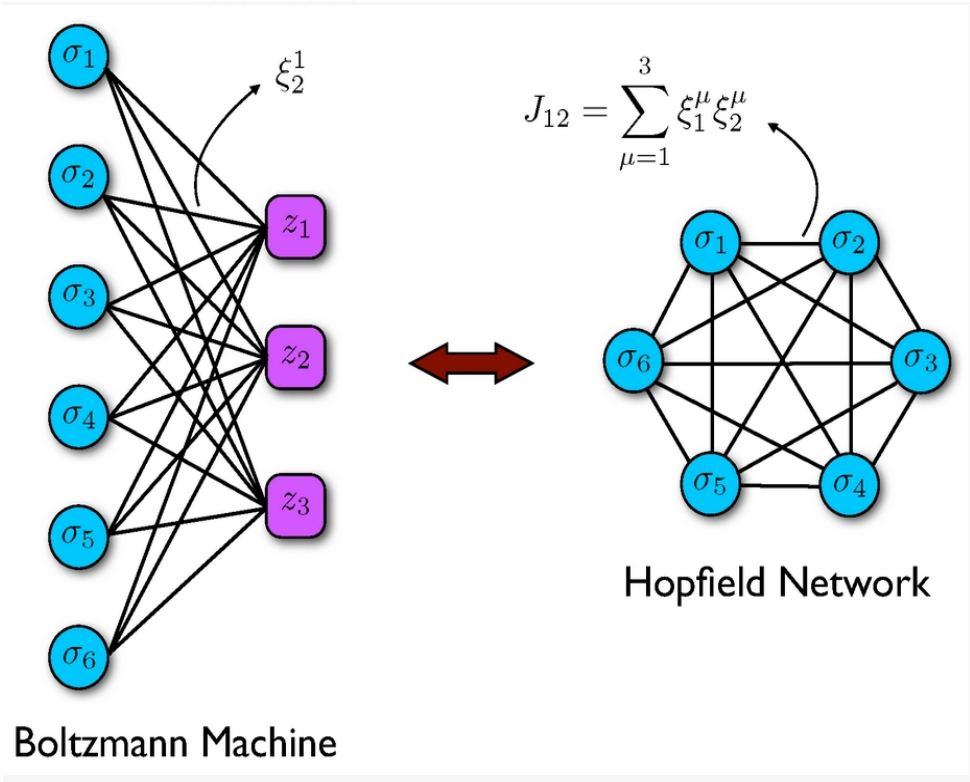




G. Parisi. Phys.Rev.Lett. 43, 1754 (1979)
“Infinite number of order parameters for spin-glasses”

Un’idea con infinite applicazioni



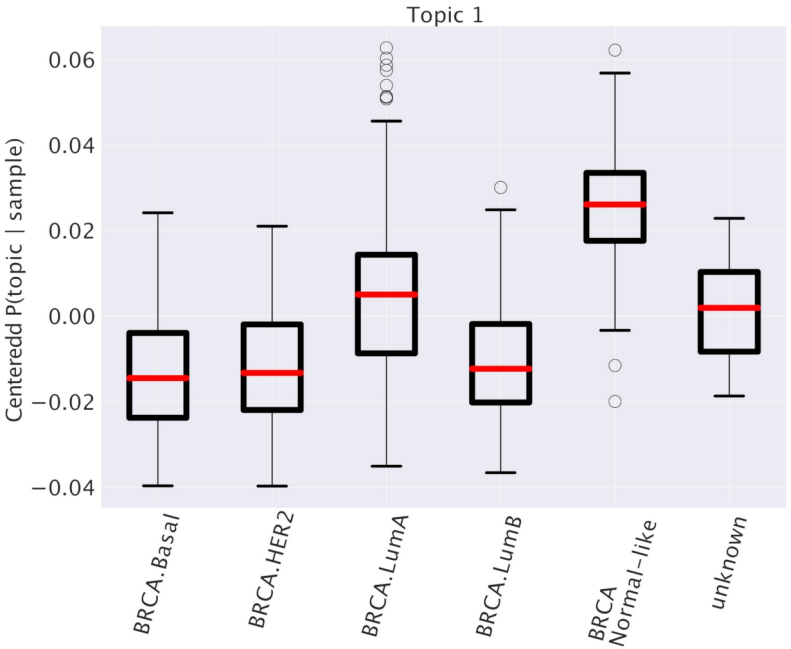


Esempio: Breast Cancer

I “topic” sono gruppi di geni.
L’algoritmo ci dice se in un particolare tipo di tumore alcuni di questi gruppi sono particolaremente espressi, più di quanto sarebbe normale aspettarsi.

Questi topic rappresentano “l’impronta digitale” di quel particolare sottotipo tumorale.

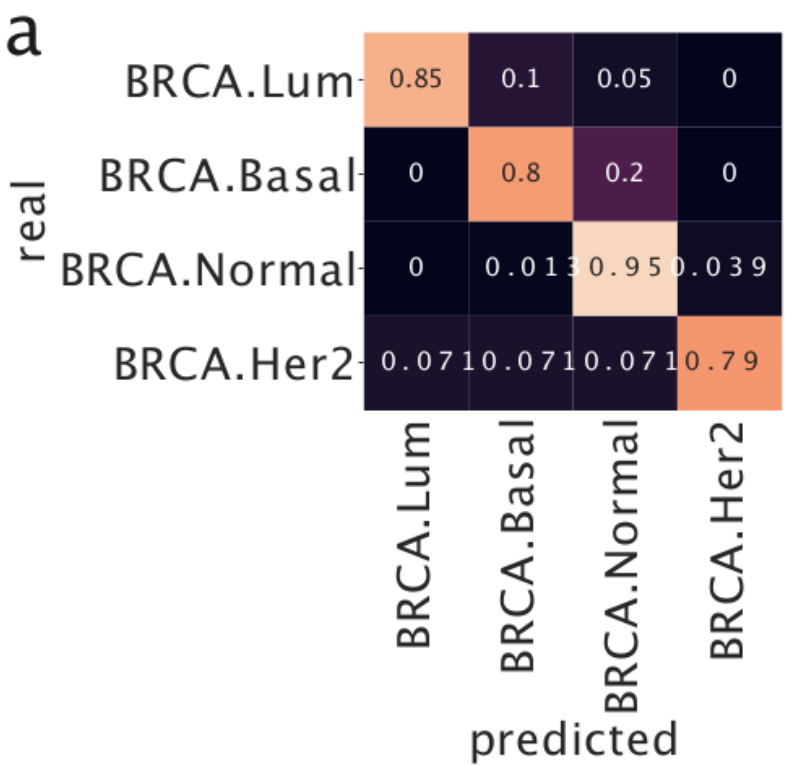
Ogni paziente ha la sua particolare collezione di topic arricchiti che possiamo usare per capire qual è il suo sottotipo tumorale



A topic modeling analysis of TCGA breast and lung cancer transcriptomic data
F Valle, M Osella, M. Caselle Cancers 12 (2020), 3799

Esempio: Breast Cancer

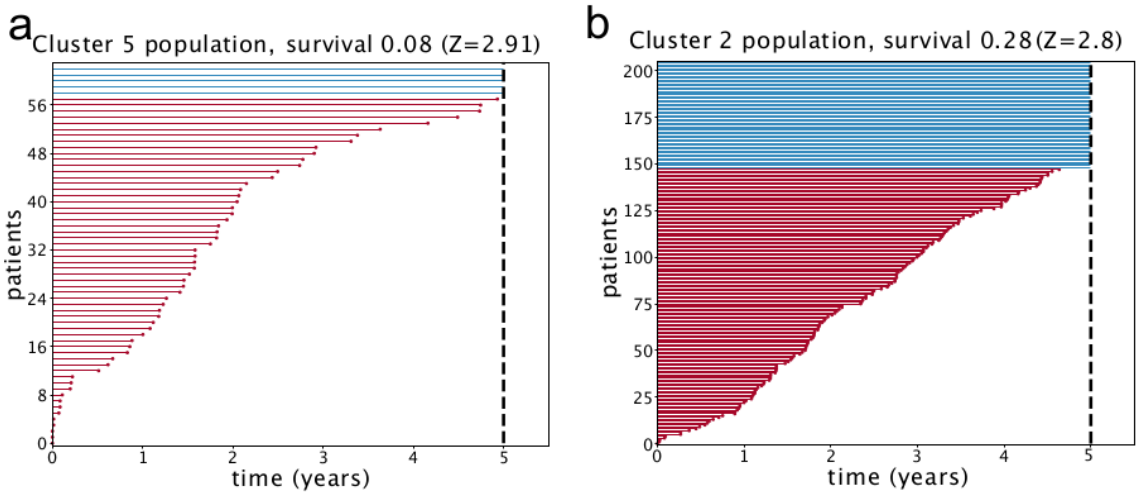
Usando questi “topic” come base di partenza è possibile costruire un classificatore (una *rete neurale supervisionata*) che partendo dalla biopsia di un paziente , usando la sua particolare espressione genica può **identificare il particolare sottotipo tumorale**, predire il rischio di recidiva, la probabilità di sopravvivenza ottimizzare, se possibile, il trattamento terapeutico.

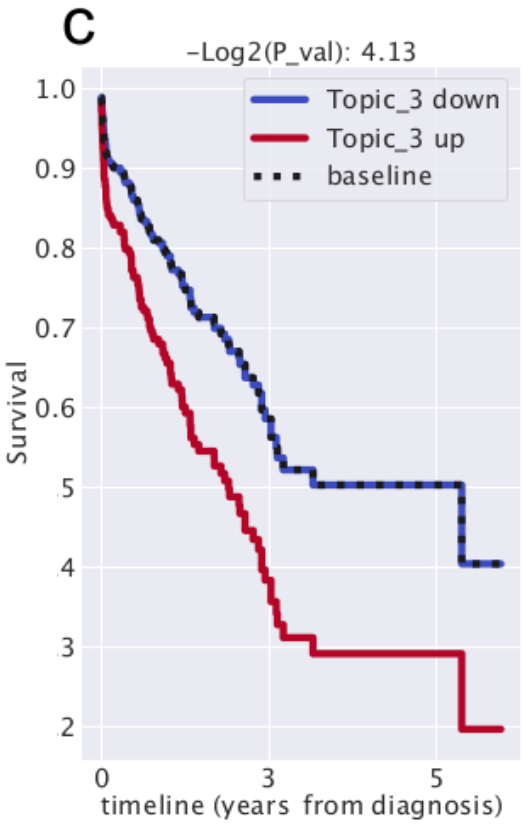


A topic modeling analysis of TCGA breast and lung cancer transcriptomic data
F Valle, M Osella, M. Caselle Cancers 12 (2020), 3799

Esempio: Breast Cancer

Guardando i livelli di espressione di questi gruppi di geni nei vari pazienti è possibile **predire la probabilità di sopravvivenza e il tipo di prognosi.**



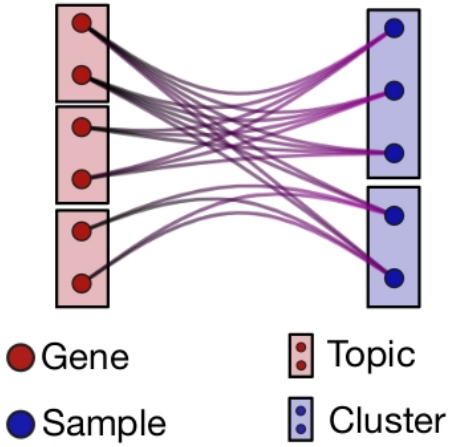


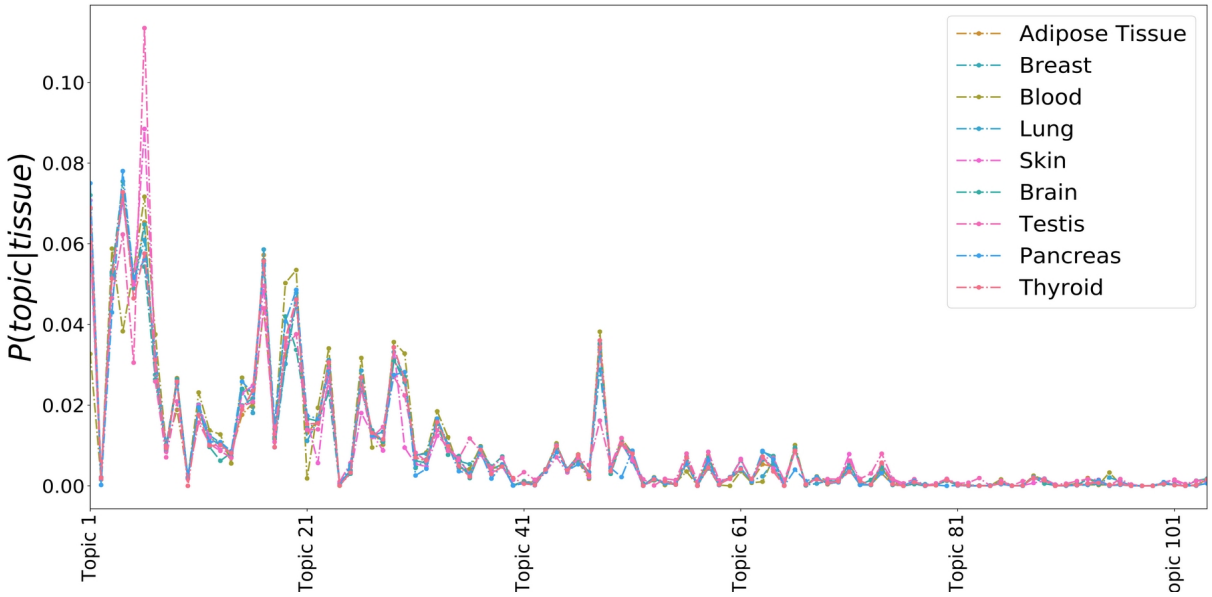
A topic modeling analysis of TCGA breast and lung cancer transcriptomic data
F Valle, M Osella, M. Caselle Cancers 12 (2020), 3799

samples

	P1	P2	P3	P4	P5
	100	10	0	36	21
	45	0	6	124	0
	163	6	0	72	0
	43	32	12	0	2
	0	0	0	0	80

Topic Modelling --->





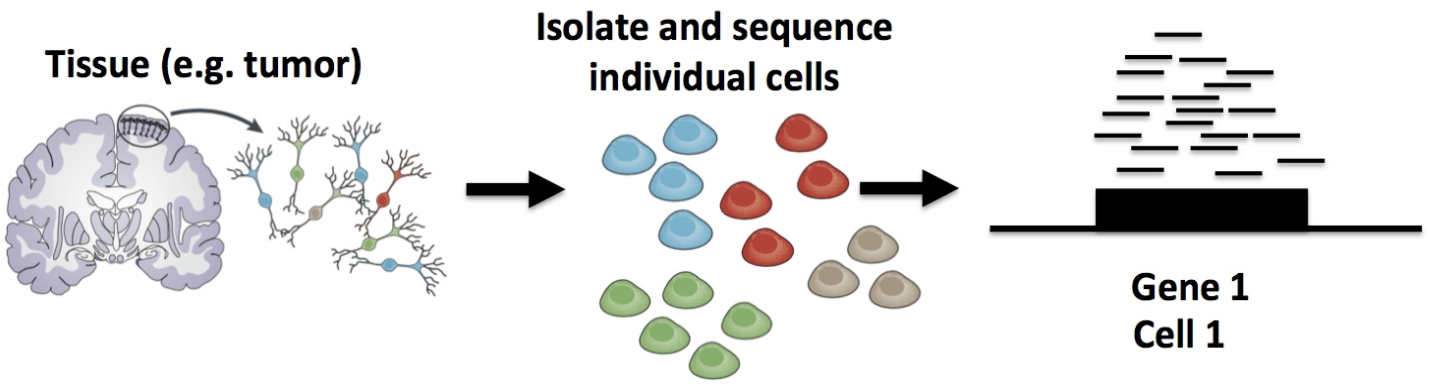
Analisi a livello di singola cellula

Le stesse tecniche si possono adattare anche allo studio di dati di espressione genica di singole cellule all'interno della biopsia.

L'obiettivo in questo caso è distinguere i diversi tipi cellulari, capire se ci sono cellule resistenti al farmaco o cellule "staminali" tumorali da cui l'intero tumore potrebbe ripartire.

Identification of Interpretable Clusters and Associated Signatures in Breast Cancer Single-Cell Data: A Topic Modeling Approach.
G. Malagoli et al. Cancers 12 (2024), 1350

Single Cell RNA-seq



Read Counts

	Cell 1	Cell 2	...
Gene 1	18	0	
Gene 2	1010	506	
Gene 3	0	49	
Gene 4	22	0	
...			

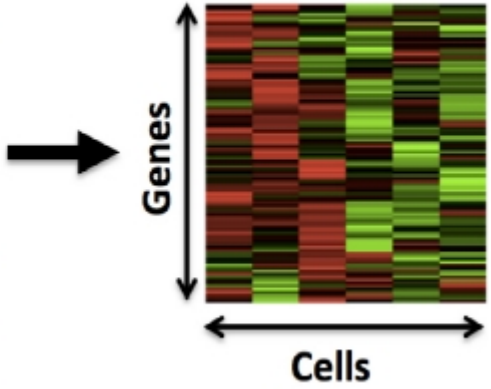
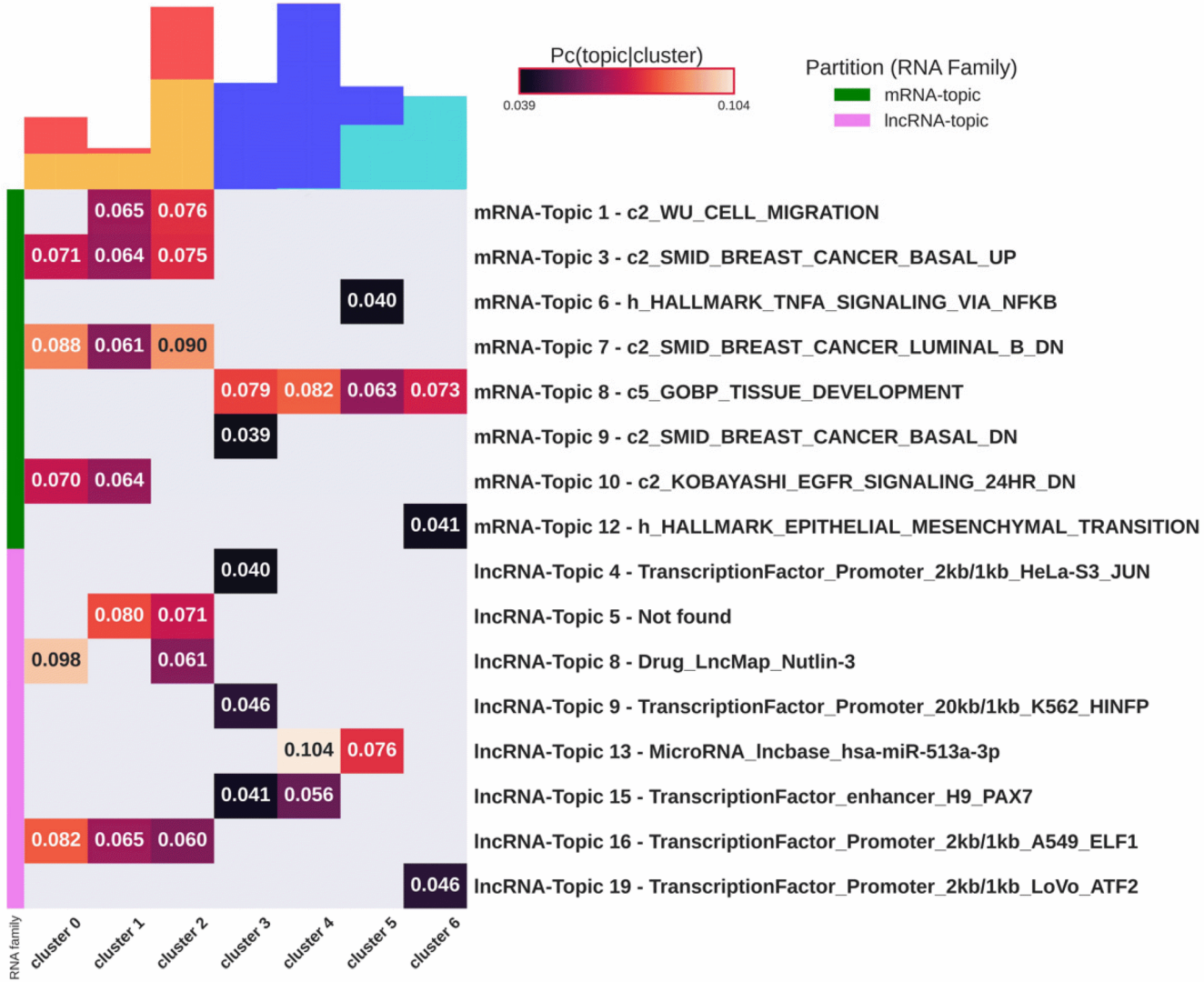


Image Credit: Stephanie Hicks, NYU



Identification of Interpretable Clusters and Associated Signatures in Breast Cancer Single-Cell Data: A Topic Modeling Approach.
G. Malagoli et al. Cancers 12 (2024), 1350

Reti Neurali e Medicina di Precisione

Questi algoritmi si basano su idee e metodi tipici delle reti neurali. Gli strumenti alla base del cosiddetto apprendimento automatico: il “machine learning”

Per capire come funzionano dobbiamo prima di tutto capire cosa sono gli algoritmi ad apprendimento automatico.

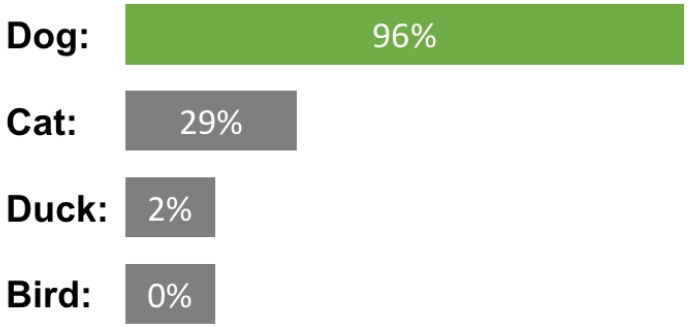
Reti Neurali

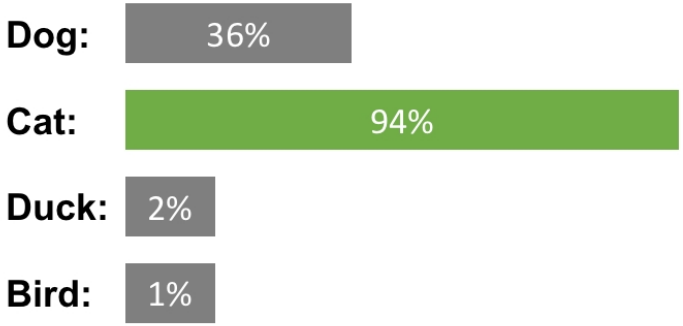
Gli algoritmi di intelligenza artificiale si dividono in tre grandi gruppi:

- Apprendimento Supervisionato:
  - Imita il funzionamento della retina
  - La rete apprende gradualmente ad assolvere al suo compito da una serie di esempi
- Apprendimento NON Supervisionato
  - Imita il funzionamento della corteccia cerebrale
  - La rete raggiunge il suo obiettivo scoprendo autonomamente le "simmetrie nascoste" del problema
- Apprendimento con Rinforzo
  - Imita il processo mediante cui il cervello impara a prendere decisioni
  - E' utile in presenza di processi sequenziali (ad esempio Chat-GPT)

Apprendimento Supervisionato

- La prima e più significativa applicazione è il riconoscimento automatico delle immagini





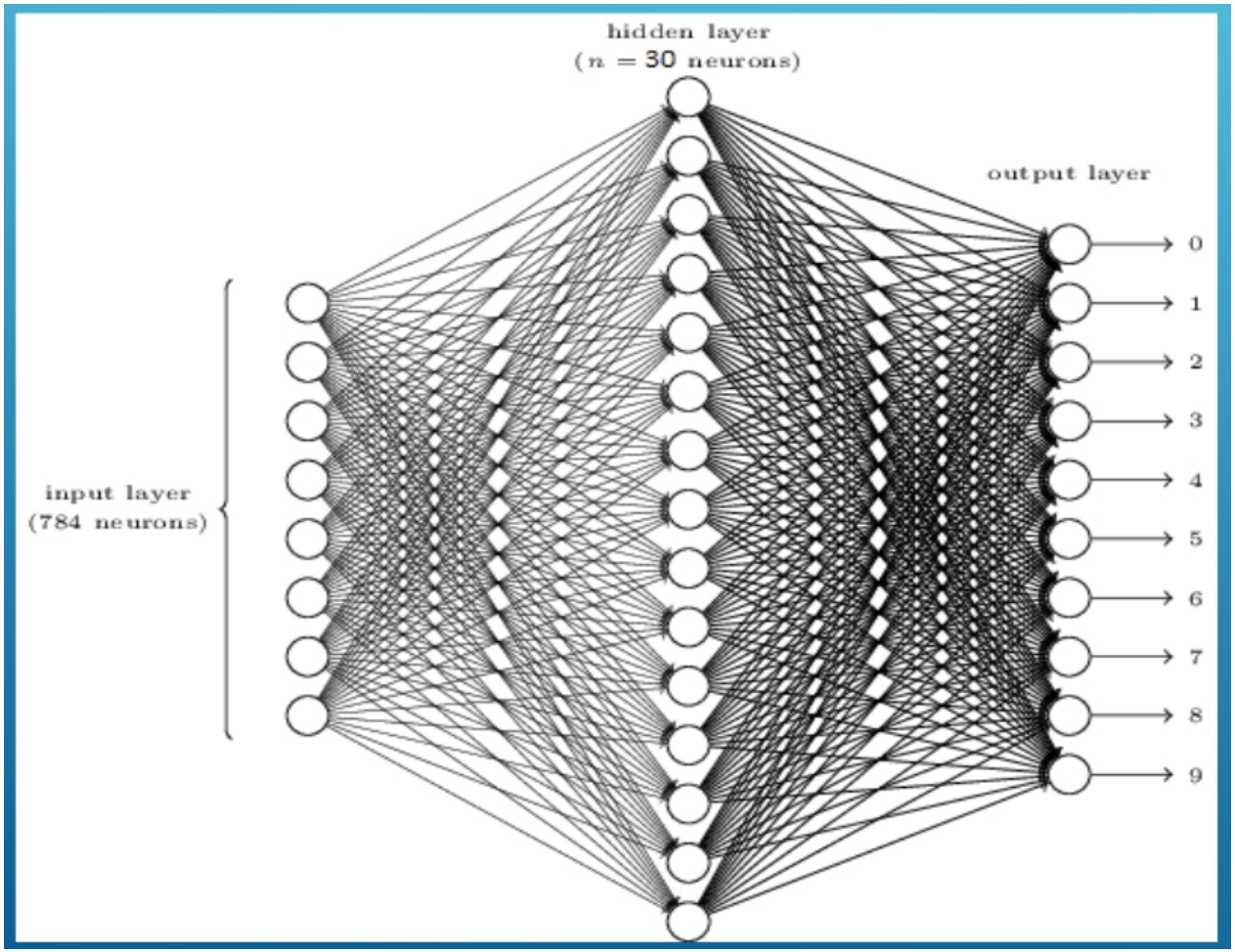
Ingredienti base di una I.A. supervisionata

- Un insieme di unità che processano l'informazione:
  i "Neuroni"

- Uno schema di connessioni tra questi neuroni:
  architettura "feed-forward"

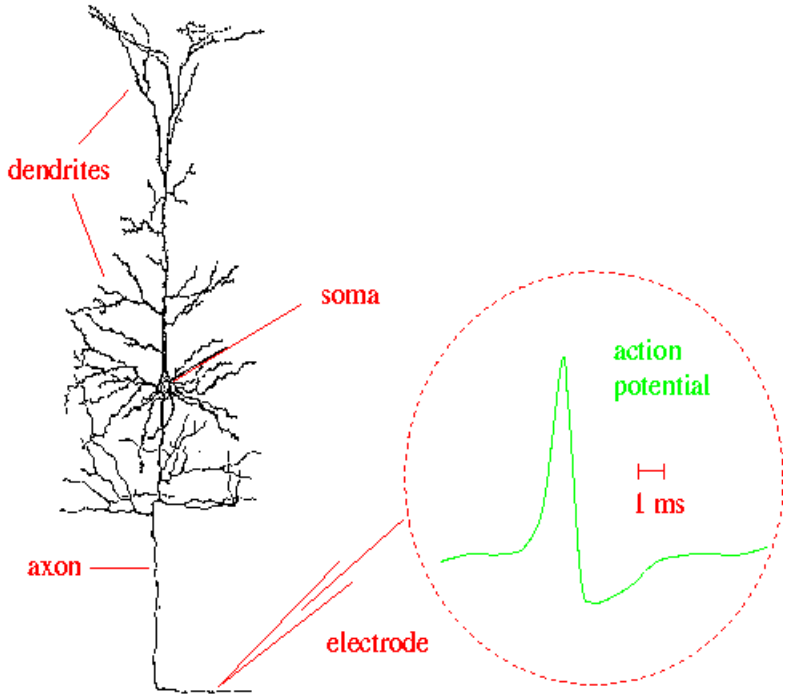
- Un insieme di regole per:
  - propagare i segnali lungo la rete
  - combinare i segnali in input
  - calcolare il segnale in output
  - modificare gradualmente i pesi sinaptici in modo che risolvano il problema: "back-propagation"

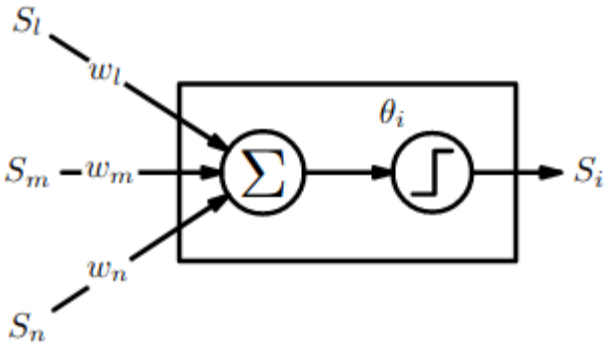
Esempio: classificatore di immagini (numeri scritti a mano su una griglia di 28x28=784 pixel):



Neuroni di input: 28x28 = 784 (le immagini sono 28x28 px)
Uno strato nascosto con 30 neuroni (scelta arbitraria)
Neuroni di output: 10 (10 sono le classi)

L’algoritmo imita il funzionamento dei neuroni

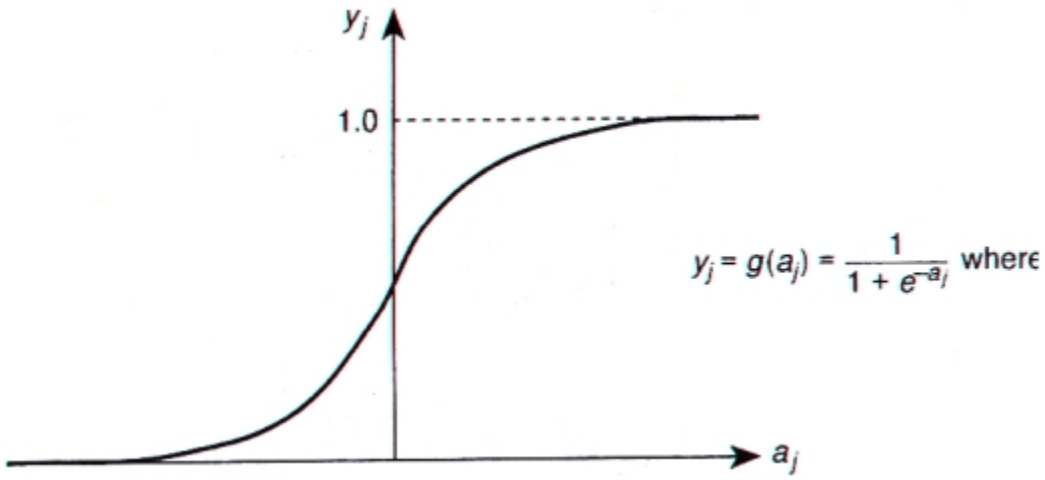




Come si combinano i segnali in input?

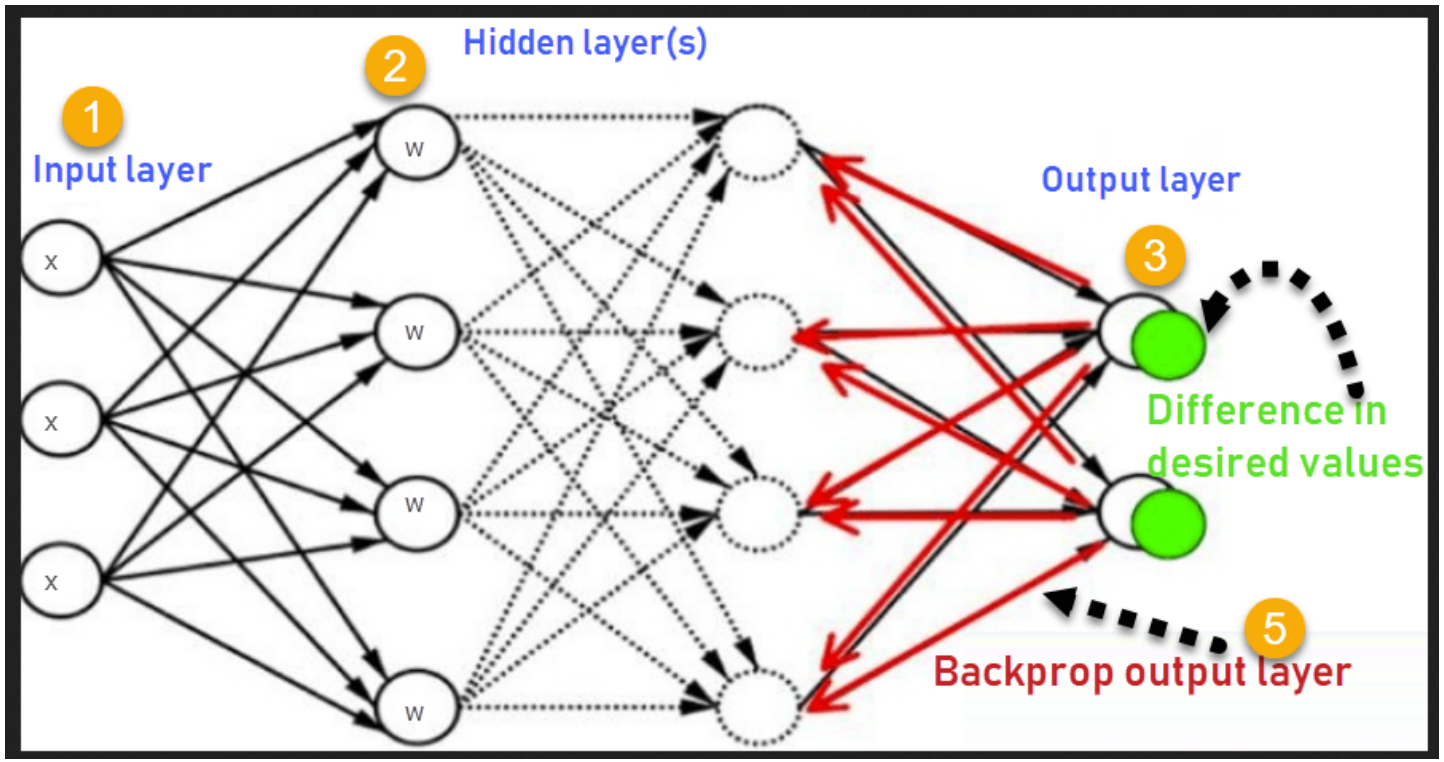
Se la somma degli input ricevuti dagli altri neuroni è positiva, il neurone si accende a sua volta, se è negativa rimane spento. Nel computer questa operazione è svolta dalla funzione “sigmoide”

$$y = g(x) = \frac{1}{1 + e^{-x}}$$



Come addestro la rete?

All'inizio i pesi sinaptici sono scelti a caso e la rete non sarà capace di riconoscere nulla. Misuro l'errore E che la rete compie quando le mostro un esempio e poi modifico i pesi in modo da diminuire l'errore: Backpropagation



Principali differenze tra un algoritmo tradizionale ed una I.A. basata sull'apprendimento supervisionato:

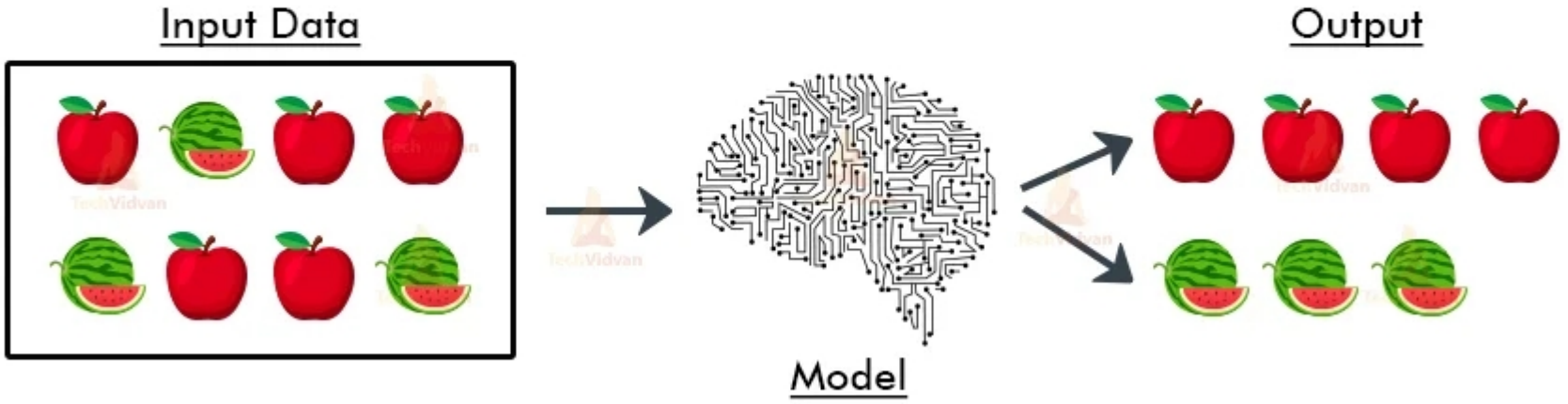
- Negli algoritmi tradizionali la soluzione del problema è codificata in modo esplicito nel programma mediante una serie di istruzoni di tipo IF-THEN-ELSE

- Negli algoritmi basati sulla I.A. la soluzione è codificata in modo implicito nei pesi sinaptici della rete che vengono modificati gradualmente durante il processo di apprendimento.

Apprendimento NON Supervisionato

L’algoritmo deve trovare da solo le “strutture interne” del dataset mettendo assieme gli oggetti che si somigliano.

E’ un processo che in gergo si chiama di “riduzione dimensionale” l’algoritmo trova le “variabili latenti” del problema



Apprendimento NON Supervisionato

L'algoritmo deve trovare da solo le "strutture interne" del dataset mettendo assieme gli oggetti che si somigliano.

E' un processo che in gergo si chiama di "riduzione dimensionale" l'algoritmo trova le "variabili latenti" del problema

Per farlo invece di minimizzare un'energia (l'errore degli algoritmi supevisionati) deve minimizzare una **ENTROPIA** che è chiamata in gergo **"minimum description length"**

L'entropia misura la quantità di disordine di un sistema. Nel nostro caso misura quanto è disordinata (poco informativa) la descrizione che ne facciamo.

Minimum Description Length

Se devo spedire a qualcuno la descrizione del mio dataset la cosa più ingenua che posso fare è elencare uno per uno i livelli di espressione dei geni per ognuno dei pazienti

Ma se mi accorgo che due colonne sono simili (e le metto vicine) invece di spedire tutti i dati della seconda colonna avviso l'interlocutore che le due colonne sono quasi uguali e gli mando solo i valori dei geni che differiscono (che saranno molti meno)

Ho diminuito la lunghezza della mia descrizione!

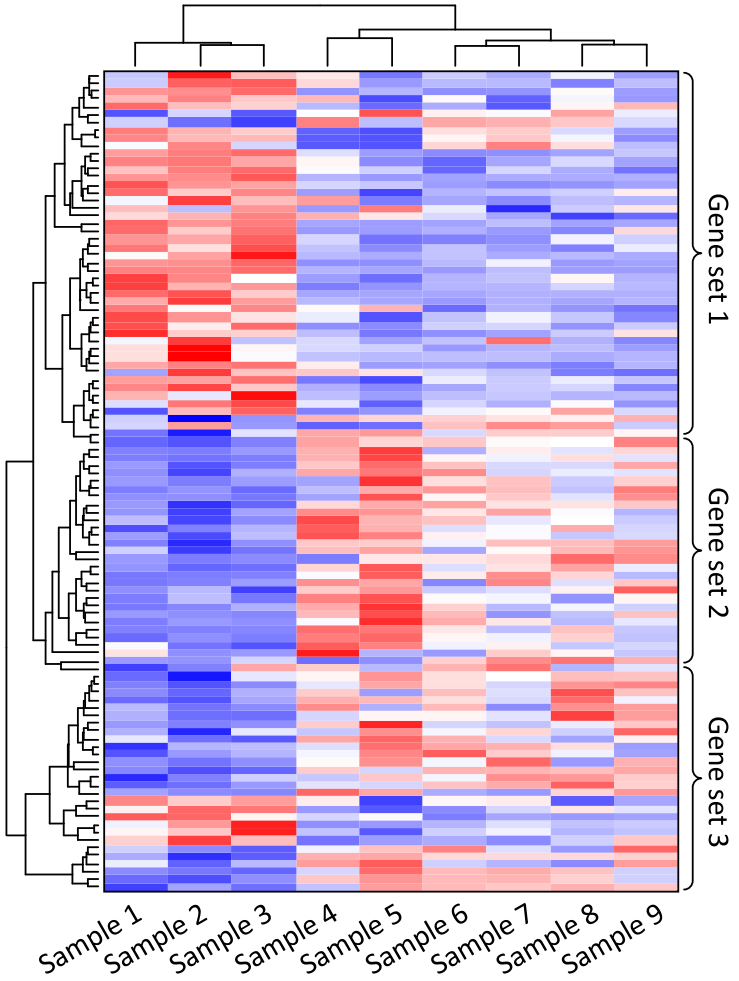


Image by Thomas Shafee - Own work, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=56957936

Conclusioni: Alcune idee guida

Gli sviluppi recenti della Medicina di Precisione sono basati su alcune “idee guida” dello studio dei Sistemi Complessi

- Riduzione dimensionale: il sistema può essere descritto da (poche) “variabli latenti”

- La presenza di molti processi in competizione tra loro:

  tutti questi processi vanno tenuti in conto quando si studia il sistema

- Minimizzazione dell’energia e dell’entropia!

- Approccio interdisciplinare, che combina Fisica, Biologia, Informatica e Matematica