open data in biomedicina: opportunità e rischi nell'era dell'intelligenza artificiale

michele piana

dipartimento di matematica, università di genova

osservatorio astrofisico di torino, istituto nazionale di astrofisica

ospedale policlinico san martino IRCCS genova

genOAweek 2025 - etica e governance dei dati biomedicali nell'era dell'AI

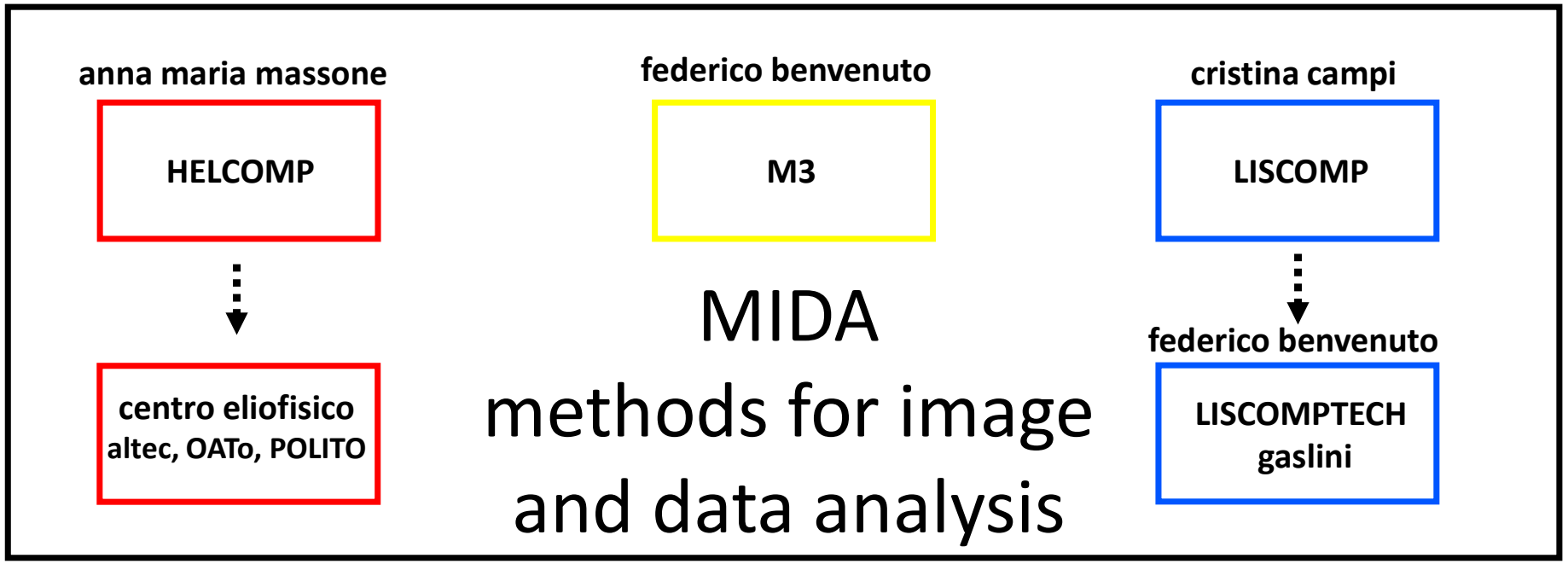
18 novembre 2025

introduzione

chi sono

- laurea e dottorato in fisica a UNIGE
- post-doc, department of mathematical sciences, university of delaware
- professore associato, università di verona
- professore ordinario @UNIGE
- co-fondatore e coordinatore di MIDA @UNIGE (https://mida.unige.it)
- co-fondatore di LISCOMP @UNIGE e @HSM
- temi di ricerca: metodi computazionali per l'analisi dati in biomedicina e fisica solare

MIDA in un guscio di noce



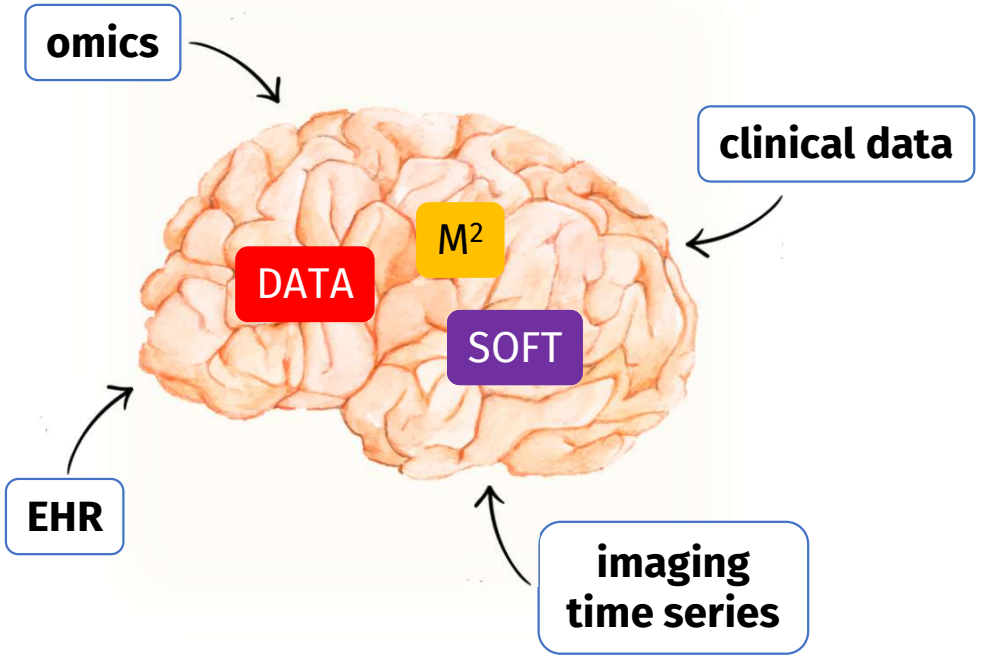
life science computational lab (LISCOMP)

- iniziativa congiunta tra università di genova e il policlinico san martino (2021)
- laboratorio di ricerca e innovazione per la modellizzazione matematica e l'analisi computazionale di dati biomedicali
- facility di scienziati computazionali e di risorse di calcolo a disposizione di medici e biologi all'interno dell'ospedale

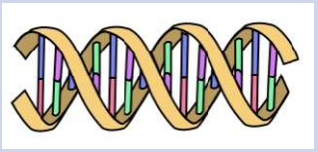
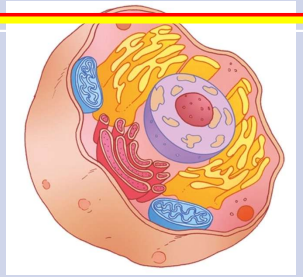
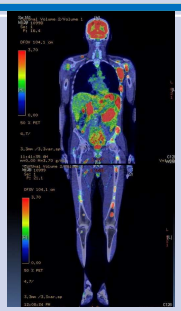
DATAunit
- T1: data collection
- T2: data infrastructure
- T3: data protection

M2unit
- T1: computational imaging
- T2: systems biology and medicine
- T3: computational genomics
- T4: biostatistics

SOFTunit



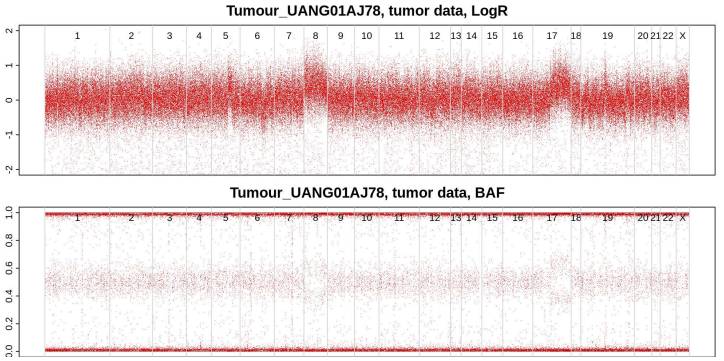
metodi e scale @LISCOMP

attività	scala	matematica
pipeline genomiche		machine learning ottimizzazione
modelli in-silico di proteomica		sistemi dinamici optimization
serie temporali in neurofisiologia		problemi inversi statistica avanzata
organoidi per malattie degenerative		machine learning pattern recognition
biomarkers in imaging medico		machine learning problemi inversi elaborazione di immagini
radiomica/radiogenomica		elaborazione di immagini pattern recognition machine learning

analisi dati: molte scale, molti modi

scala DNA: aberrazioni della struttura genomica

le aberrazioni di copy number sono variazioni strutturali del genoma dove pezzi di DNA sono acquisiti oppure cancellati



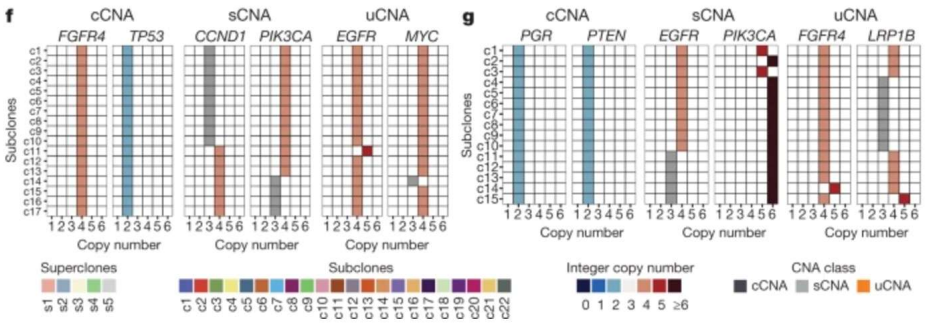
$$\mathcal{L}(\mathbf{r},\mathbf{b}) = \sum_{j=1}^{Q} \sum_{i \in I_j} (r_i - \bar{r}_{I_j})^2 + (b_i - \bar{b}_{I_j})^2 + \lambda Q$$

Article | Published: 24 March 2021

Breast tumours maint diversity during expa

Darlan C. Minussi, Michael D. Nicholson, Tarabichi, Emi Sei, Haowei Du, Mashiat R Schalck, Asha Multani, Jin Ma, Thomas C Lim, Banu Arun, Funda Meric-Bernstam,

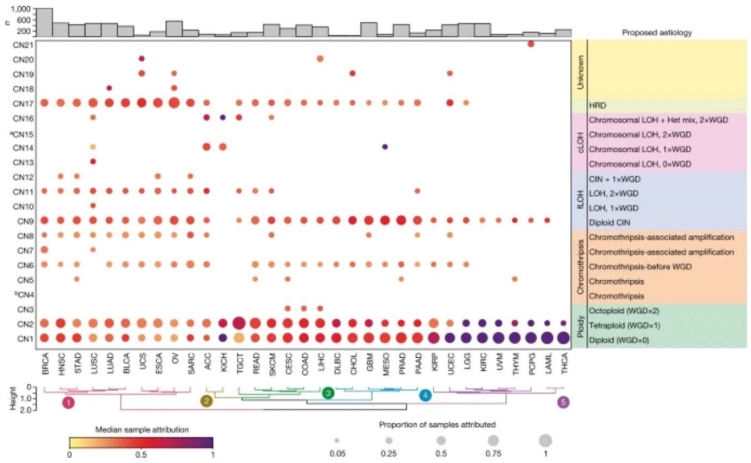
Nature **592**, 302–308 (2021) | Cite this



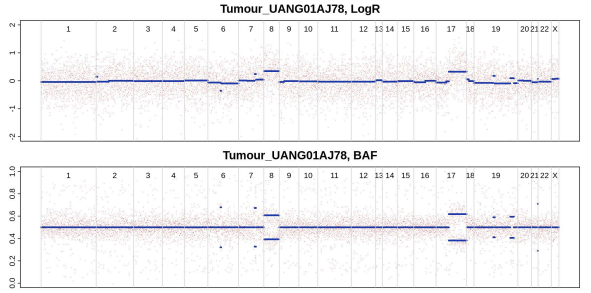
Signatures of copy numbe cancer

Christopher D. Steele, Ammal Abbasi, S. M. Ashiqu Haase, Shadi Hames-Fathi, Dolapo Ajayi, Annelien Lechner, Nicholas Light, Adam Shlien, David Malki Fredrik Mertens, Adrienne M. Flanagan, Maxime Ta Nischalan Pillay

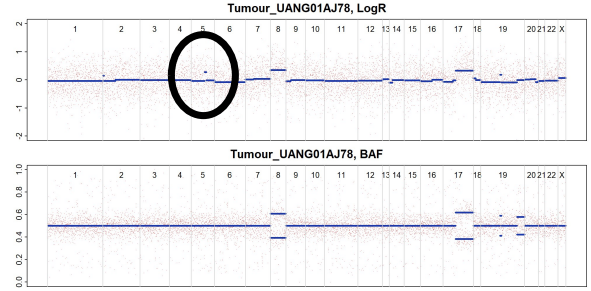
Nature **606**, 984–991 (2022) | Cite this article



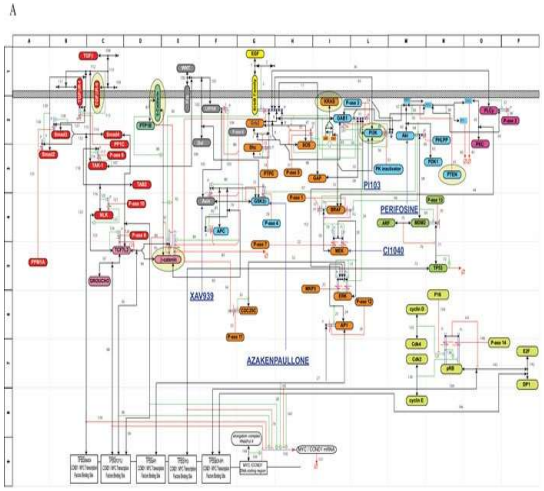
ASCAT

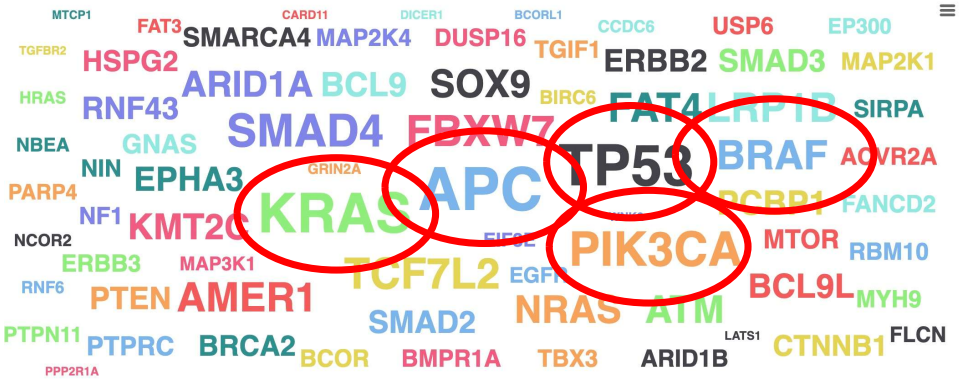


LISCOMP

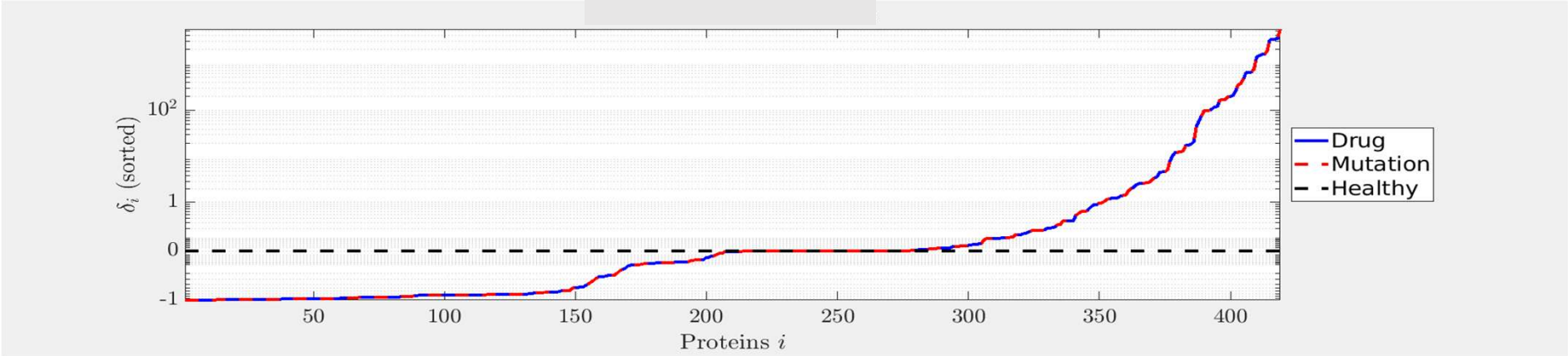


scala proteine: network di reazioni chimiche nel cancro





$$\begin{cases} \dot{\boldsymbol{x}}(\tau) = \mathbf{S}\,\boldsymbol{v}(\boldsymbol{x}(\tau), \boldsymbol{k}) \\ \boldsymbol{x}(0) = \boldsymbol{x}_0 \end{cases}$$



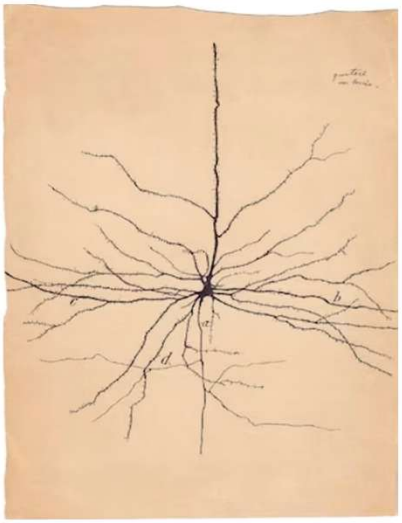
1. LLM per la generazione delle reti di reazioni

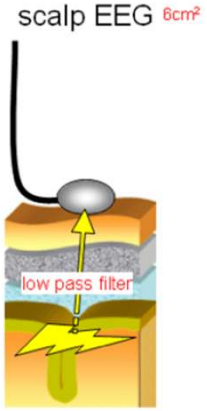
2. algoritmo per la generazione di sistemi dinamici

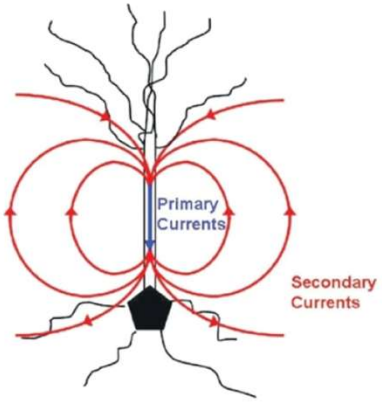
3. metodo di ottimizzazione per il calcolo degli equilibri

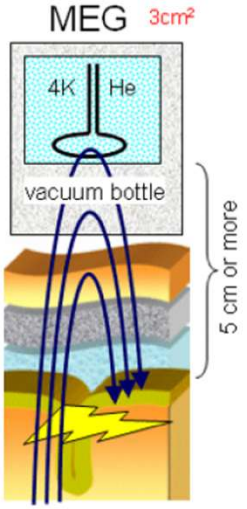
4. validazione con dati sperimentali

scala cellula: connettività









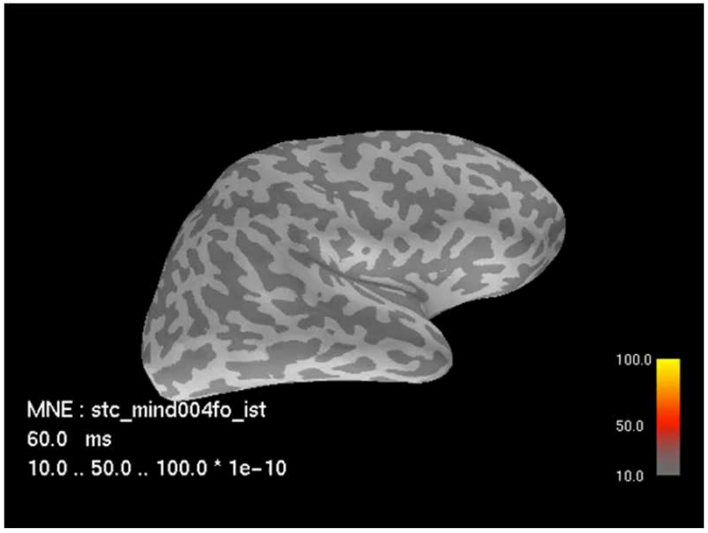
$$j = j_p + j_s \qquad j_s = \sigma e$$

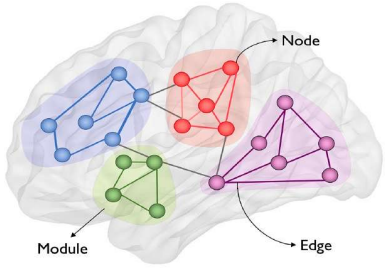
$$\nabla \times e = 0 \Rightarrow j_s = -\sigma \nabla v$$

$$\nabla \cdot j = 0 \Rightarrow \nabla \cdot j_p = \nabla \cdot (\sigma \nabla V)$$

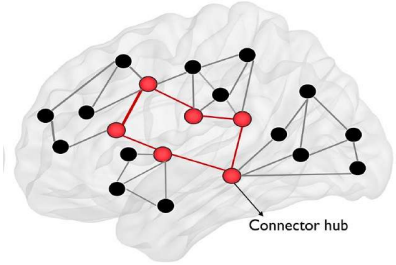
$$\frac{\partial e}{\partial t} = 0, \frac{\partial b}{\partial t} = 0$$

$$b(r,t) = \frac{\mu_0}{4\pi} \int_\Omega j(r',t) \times \frac{r - r'}{|r - r'|^3} dr'$$





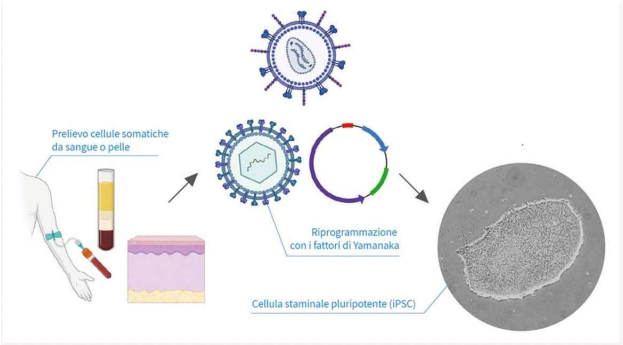
(a) Network Segregation

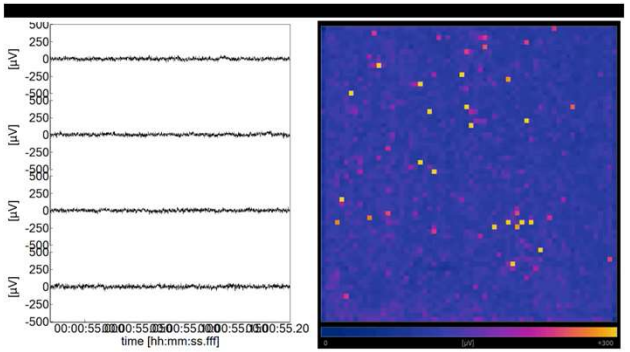


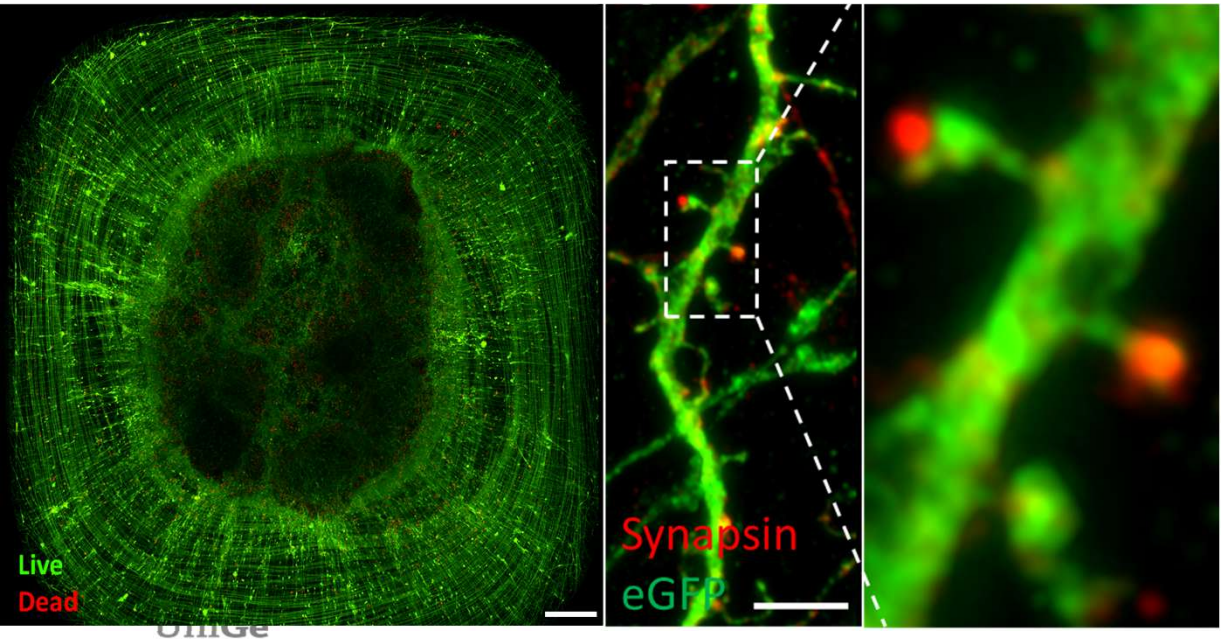
(b) Network Integration

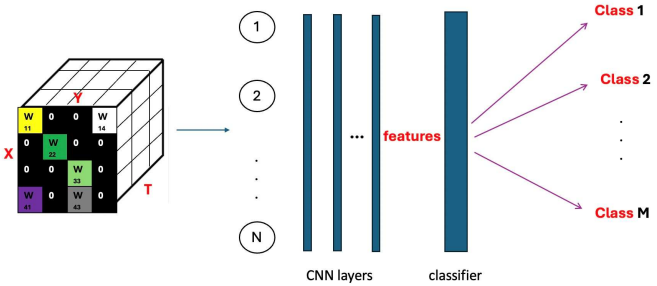
3D-BrAIn

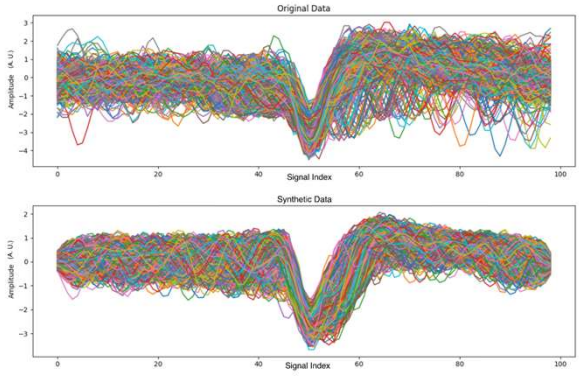
scala organoidi: gemelli bio-digitali











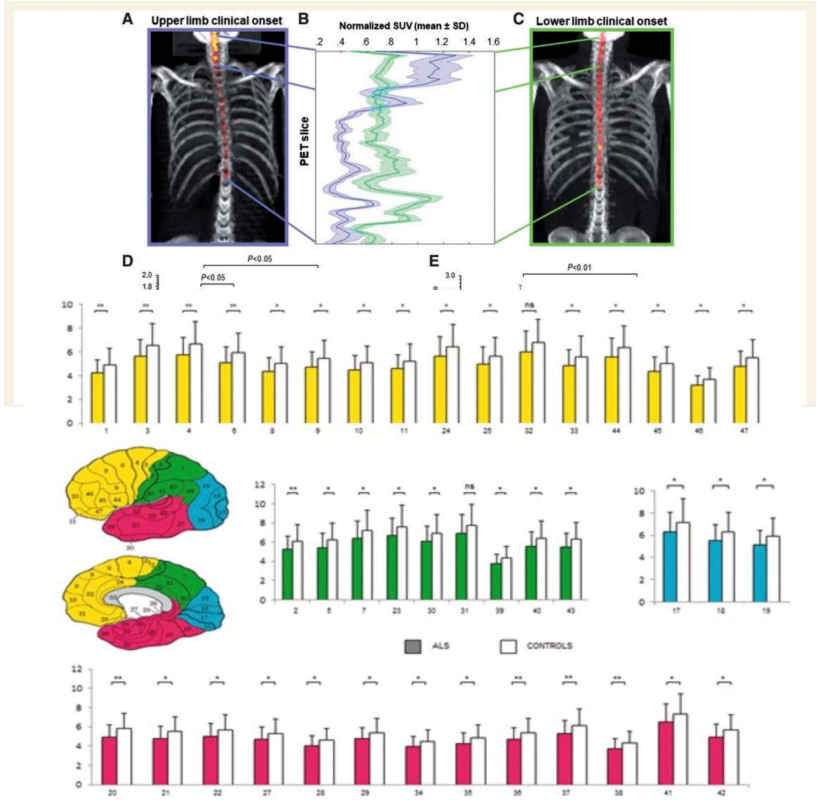
scala organi: biomarker da immagini

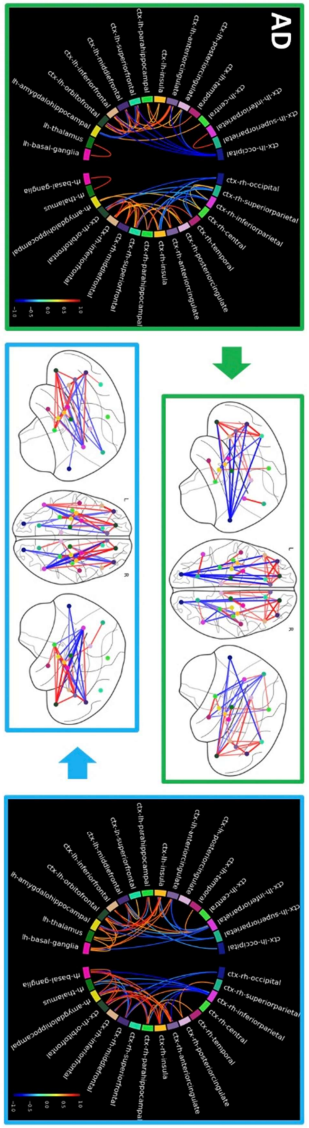
doi:10.1093/brain/awy152 BRAIN 2018: 141; 2272–2279 | 2272

BRAIN
A JOURNAL OF NEUROLOGY

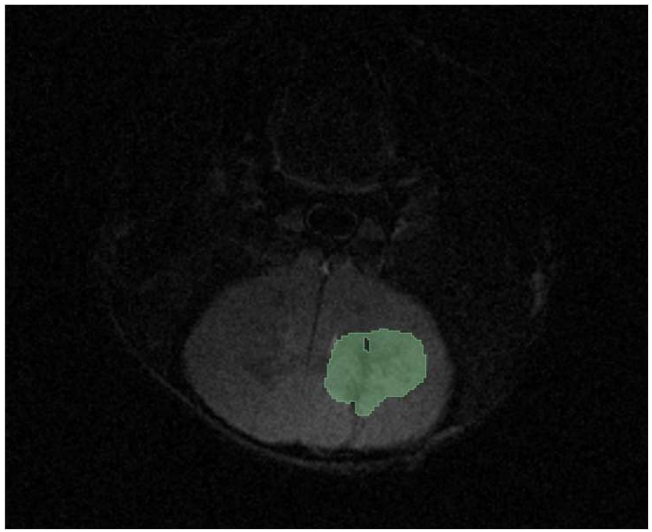
REPORT

Interplay between spinal cord and cerebral cortex metabolism in amyotrophic



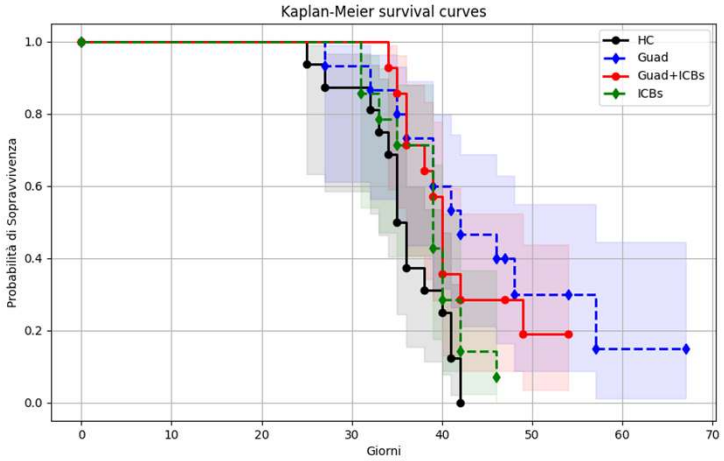


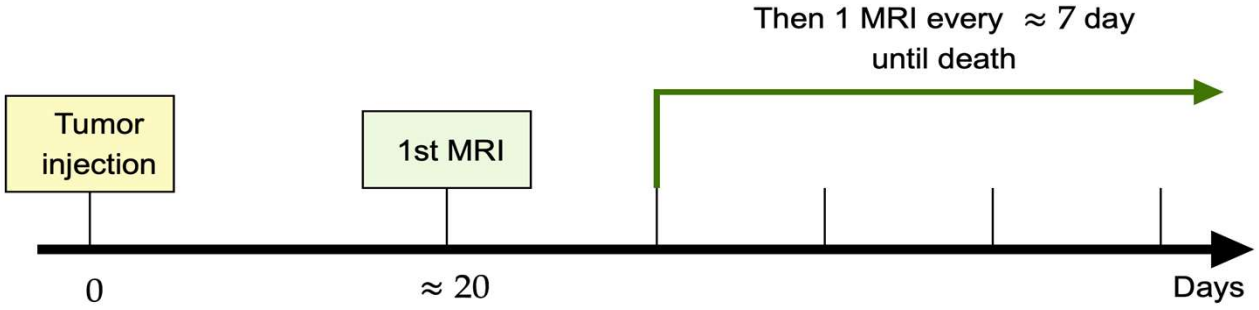
dalla scala organi alla scala DNA: modelli murini di glioblastoma



4 gruppi:

- controlli
- tre approcci terapeutici





l'obiettivo è quello di individuare possibili correlazioni tra le proprietà
delle immagini, le caratteristiche genetiche del tumore e l'efficacia della terapia

alcune questioni delicate (secondo me)

la questione dei big data

"information consumes the attention of its recipients.
hence a wealth of information creates a poverty of attention
and a need to allocate that attention efficiently
among the overabundance of information sources that might consume it"
(herbert simon, premio nobel per l'economia)

la questione dei big data è a volte fuorviante

a contare davvero sono
- la qualità
- la comprensione
- l'interpretazione
- lo sfruttamento dei dati

la scienza dei dati è multi-forme e comprende
- simulazione
- teoria dei problemi inversi
- pattern recognition
- machine/deep learning

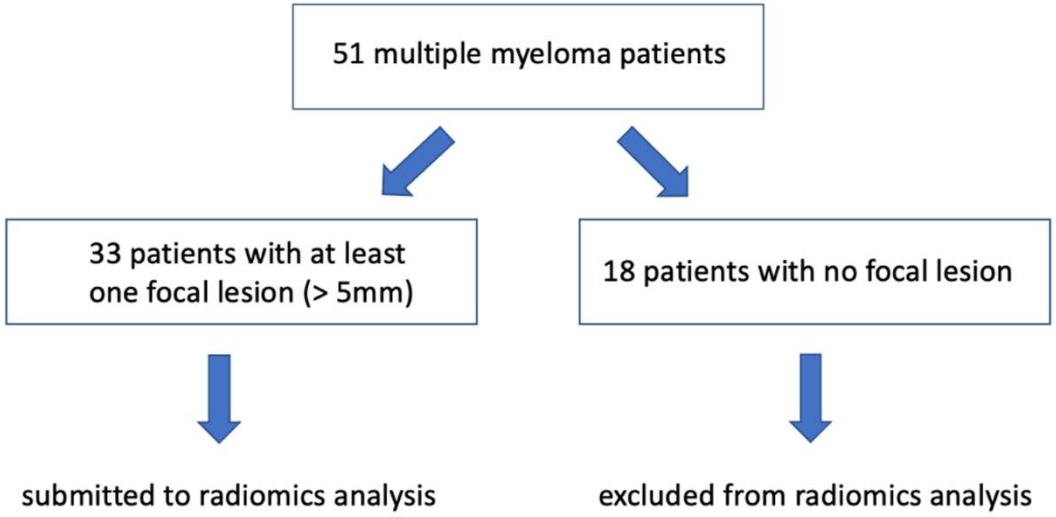
dati: quantità e qualità

- chat-GPT funziona benissimo perché internet è un archivio letteralmente impareggiabile di pdf

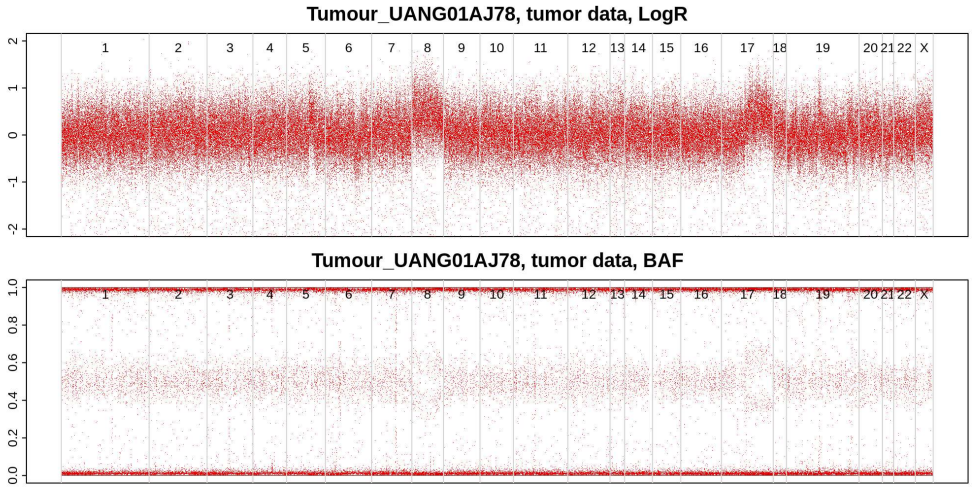
- chat-GPT riesce a generare immagini perché internet è un archivio letteralmente impareggiabile di immagini di buona qualità

- ma non è così in tutte le applicazioni:
  - ✓i dati possono essere intrinsecamente pochi
  - ✓i dati possono essere di scarsa qualità

la questione dei dati: due esempi

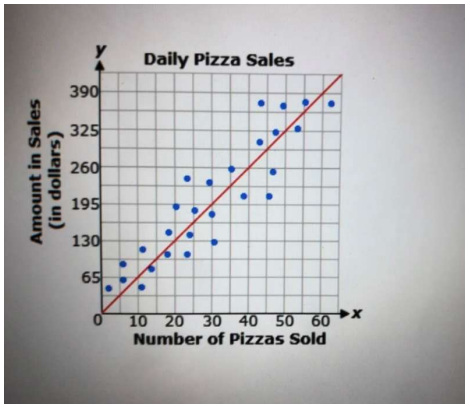


radiomica per la prognosi del mieloma multiplo

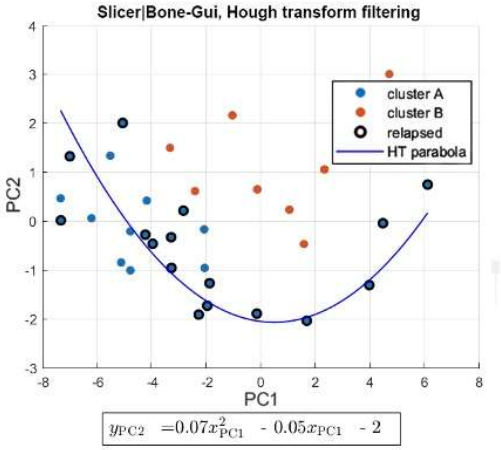


analisi di copy number

la questione energetica: cosa è una rete neurale



y=ax+b



$y=ax^2+bx+c$

una rete neurale è una funzione con molti parametri

nella fase di addestramento i parametri vanno calcolati in modo da

- descrivere un archivio storico molto grande di cui si possiede un'informazione completa
- stimare probabilisticamente l'informazione mancante in corrispondenza di nuovi dati

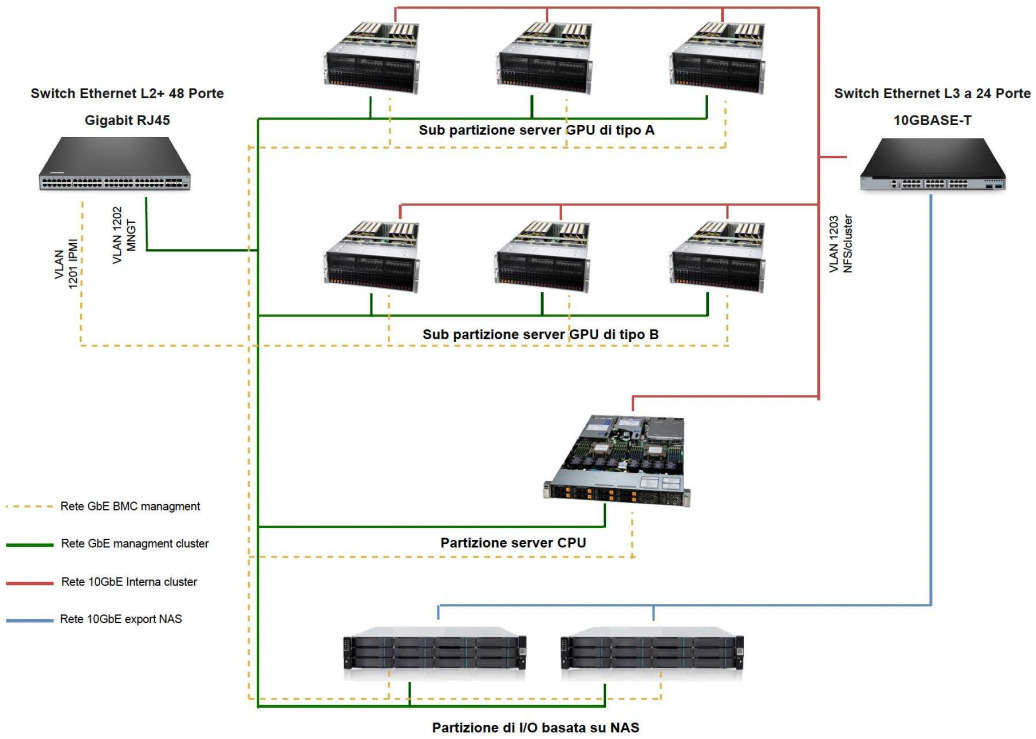
però:

- se ogni dato dell'archivio ha molte dimensioni (per esempio: immagini) i parametri da ottimizzare sono molti: reti profonde
- l'addestramento richiede la soluzione di un problema di minimo complesso

$$min_w L(\theta(w,x), y) + R(w)$$

- il costo computazionale della fase di addestramento è grande

la questione energetica: high performance computing (HPC)



e comunque non va dimenticato l'enorme consumo di corrente...

- la parola è uno strumento che trasmette il pensiero e la conoscenza
- la penna è uno strumento che infrange i limiti locali e temporali della comunicazione e permette l'accumulazione della conoscenza
- il computer è uno strumento che ci permette di risolvere problemi prima irrisolvibili
- non confondiamo l'aumentare della capacità di calcolo con l'aumentare della conoscenza matematica: l'AI si basa su matematica classica

"Considerate la vostra semenza:
fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza"
(Dante, XXVI - 118-120)

la questione del software

linguaggi di programmazione:

- python
- R
- matlab
- C (e derivati)

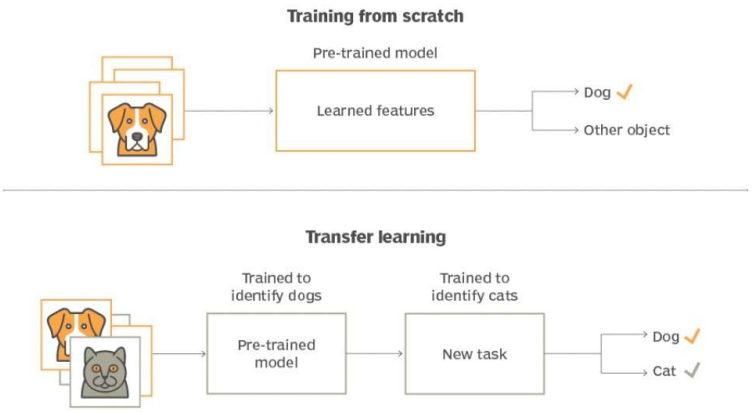
frameworks per deep learning:

- scikit_learn
- PyTorch
- TensorFlow
- Keras
- Theano
- Caffe
- DL4J
- CHAINER
- CNTK

anche l'open source pone domande difficili:

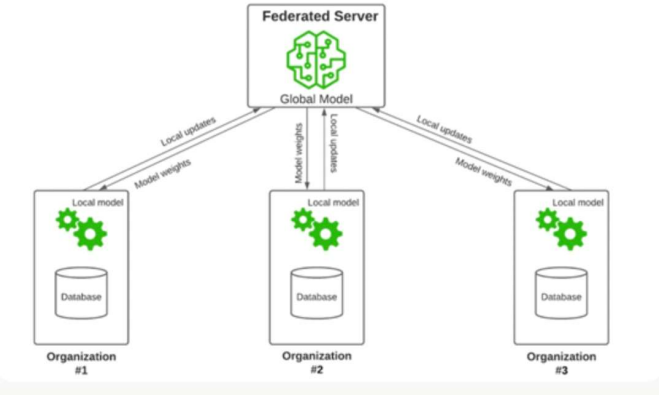
- chi è il padrone del software che realizza una soluzione AI?
- come si brevetta un software AI?
- è una buona idea brevettare un software AI?
- come si affronta la questione 'medical device'?

dati e software: nuove tendenze



transfer learning:

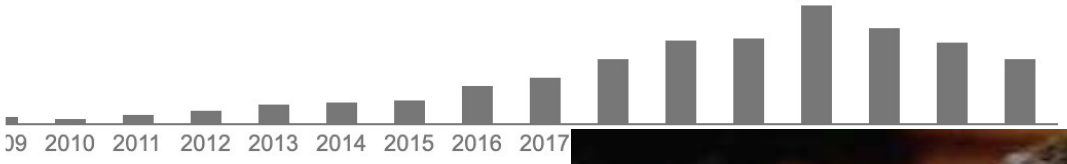
- una rete neurale è addestrata per un certo compito utilizzando un archivio storico adeguato
- la rete neurale ottimizzata viene messa in rete
- nel caso di un compito simile a quello per cui è stata addestrata, la rete neurale viene scaricata e addestrata con i dati relativi al nuovo compito
- modelli fondazionali: transfer learning estremizzato



federated learning:

- ogni nodo periferico ha a disposizione una copia pre-addestrata di una rete neurale, messa a disposizione dal noto centrale
- ogni nodo periferico addestra la propria copia con i propri dati
- i parametri ottimizzati da ogni nodo periferico sono trasferiti al nodo centrale che li utilizza per migliorare l'ottimizzazione della prima copia
- la rete neurale aggiornata è messa a disposizione dei nodi periferici

la questione giuridica: europa vs USA



- 1-2/10/2020: il consiglio
  intelligenza artificiale
- 21/04/2021: la commiss
  un regolamento sull'AI
- 6/12/2022: il consiglio e
  propria posizione sul pro
- 9/12/2023: consiglio e p
  raggiungono un accordo
- 21/5/2024: il consiglio e
  definitivo al regolament
- 1/8/2024: entra in vigore il regolamento sull'AI
  (ma con due anni di embargo)

nel frattempo, in USA:

- 2018: prima release di alphaphold
- 2020: seconda release di alphaphold
- 2022: pubblicazione di un
  se di 200 milioni di proteine
  te da alphaphold
- bre 2022: prima release di
  PT
- 2023: chat-GPT plus
- 24: versione di alphaphold che
  e proteine, DNA, RNA, e ligandi
- 2024: chat-GPT omni
- re 2024: i premi nobel per la
  due americani "per scoperte
  ionali nell'ambito" dell'AI
- re 2024: i premi nobel per la
  fisica a tre americani per alphaphold

gli USA non hanno un AI-act

la questione della conoscenza

- 1450: invenzione della stampa
- 1500: almeno 9 milioni di libri in europa

- la stampa trasmette nuova conoscenza
- la stampa ispira nuova conoscenza che verrà generata dall'intelligenza naturale
- l'AI al momento connette vecchia conoscenza con un'originalità mai vista prima

ma la domanda è:

l'AI è in grado di generare nuova conoscenza in modo autonomo?

nota bene:

- alphazero gioca a scacchi in modo completamente eccentrico rispetto all'umano e con risultati migliori. **ma alphazero non inventa un nuovo gioco con nuove regole**
- alphafold genera nuove proteine la cui struttura l'umano non è stato in grado di scoprire. **ma alphafold non genera un modello generale che spiega perché le mutazioni generano proteine in modo sbagliato**

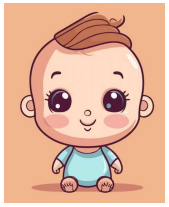
conclusioni (sempre secondo me)

una questione culturale...

- la scienza dei dati non è solo chat-GPT
- l'AI non è solo chat-GPT
- machine learning, problemi inversi, simulazione, pattern recognition sono solo cacciaviti: l'obiettivo rimane capire e riparare il motore
- c'è un bisogno urgente di laboratori interdisciplinari dove studiare tutti gli aspetti relativi alla scienza dei dati
- c'è un bisogno ancora più urgente di programmi formativi completamente nuovi, genuinamente interdisciplinari, e orientati ai problemi

"non esistono le discipline; esistono solo i problemi e i metodi per risolverli"
(k. popper)

il cervello di churchill funzionava come il machine learning



il cervello di un bimbo funziona con la simulazione

il cervello di un esploratore funziona con il pattern recognition

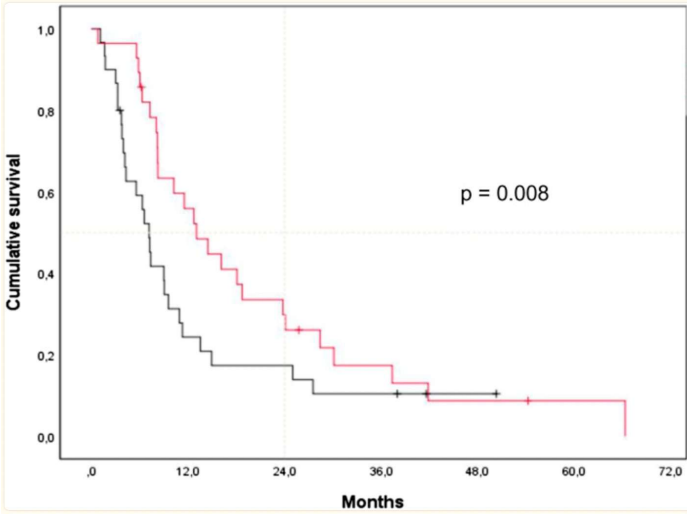
il cervello di sherlock holmes risolve problemi inversi

...e un po’ di ansietà

Breast cancer stage	*Localized*	*Regional*	*Distant*
% of people with this stage of cancer at diagnosis	61%	32%	6%
5-year relative survival rate	99%	85%	26%

Colorectal cancer stage	*Localized*	*Regional*	*Distant*
% of people with this stage of cancer at diagnosis	39%	36%	20%
5-year relative survival rate	90%	71%	13%

miglioramento della sopravvivenza grazie alla chemoterapia



miglioramento della sopravvivenza grazie alle terapie cellulari

a quando un miglioramento della sopravvivenza guidato dalla matematica?