When AI meets life sciences
当人工智能遇上生命科学

2025 清华大学 AI创新实践课程

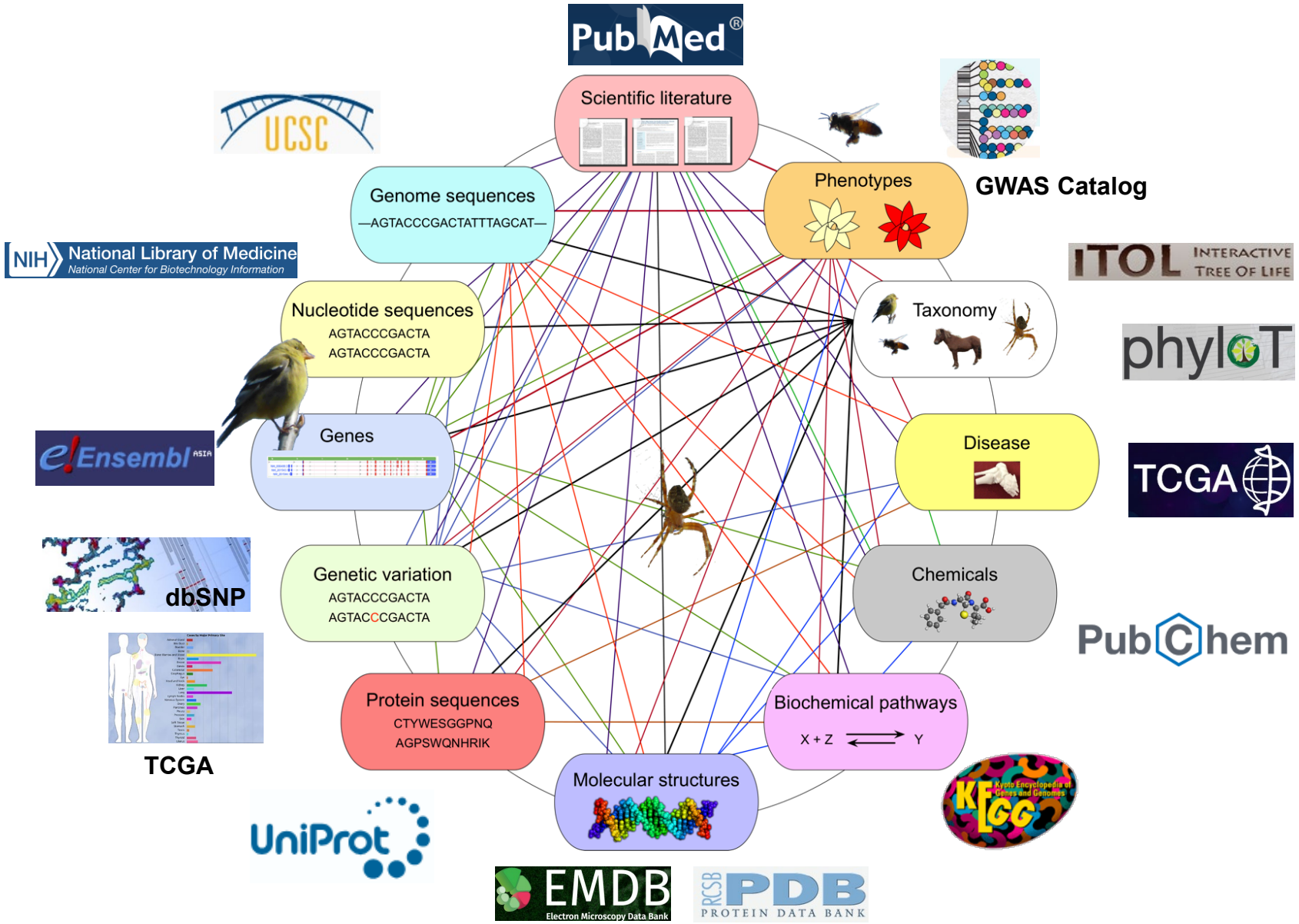
张强锋，清华大学

Nov 10, 2025

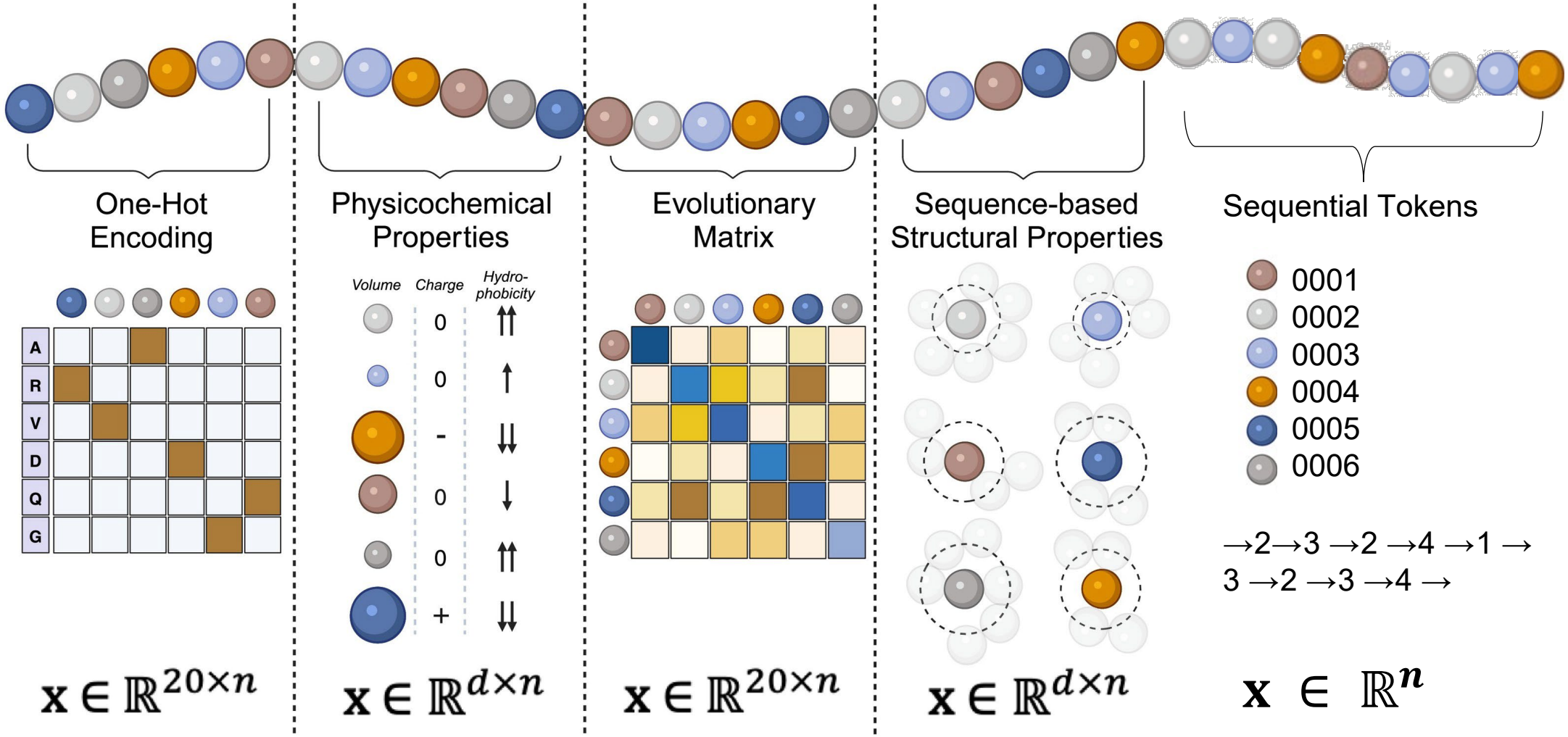
Outline

- **Biology data for AI**: sequence; structure; networks; omics; medical image; text;
- **The potential and boundaries of AI**: the potential and power of AI; challenges in AI biology; what truly defines AI model?
- **AI ethics**: privacy and security; technology misuse; transparency;

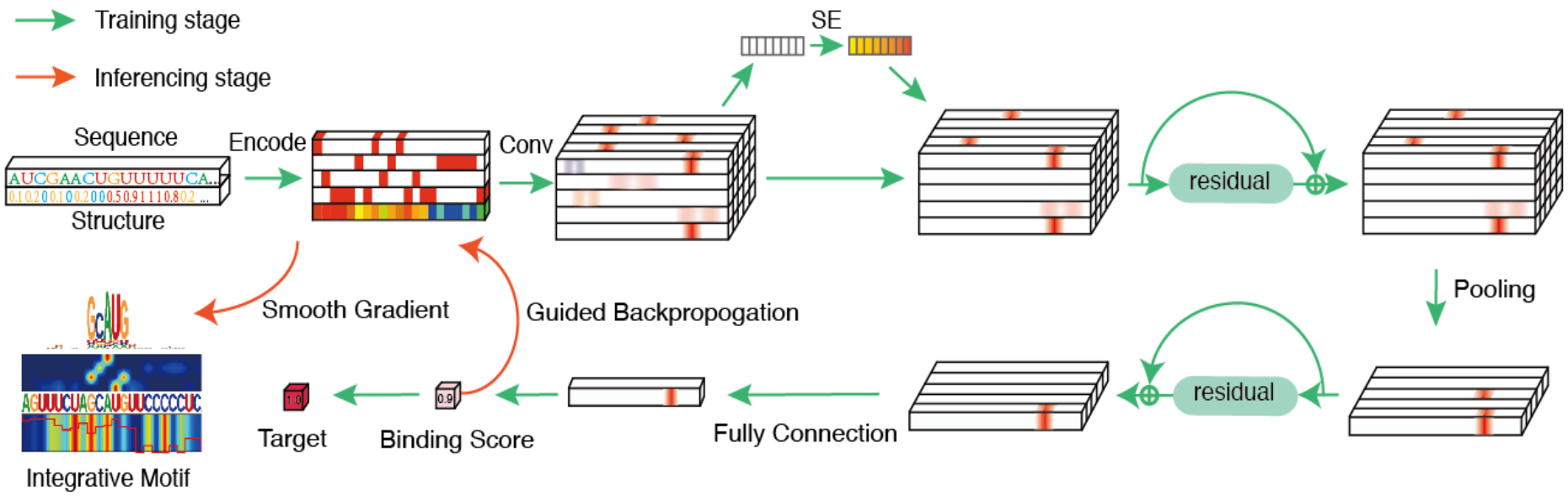
Biological data landscape

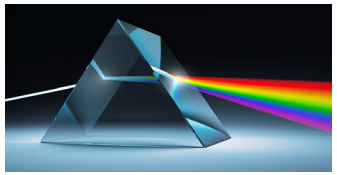


Sequence data (DNA, RNA, and protein sequence)



An example of sequence data learning





PrismNet: Protein-RNA Interaction by Structural Modeling using deep neural NETwork

Sun et al. Cell Research（2021）

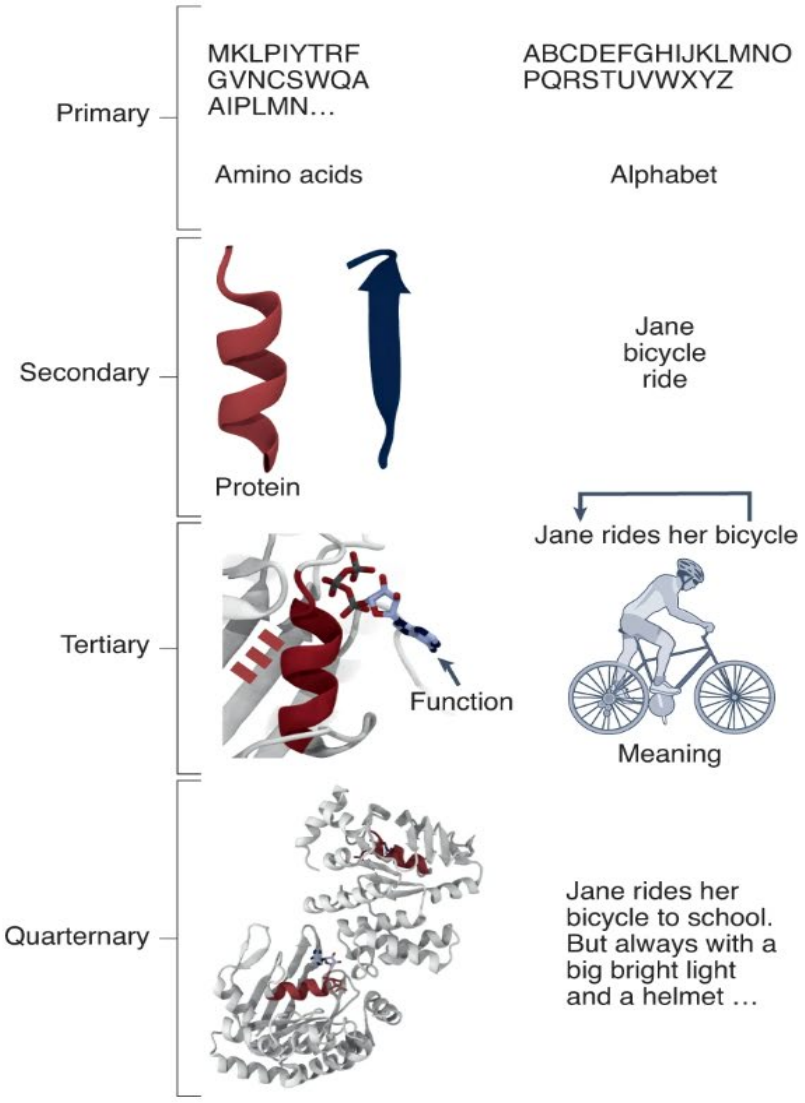
From sequence to function: structural data

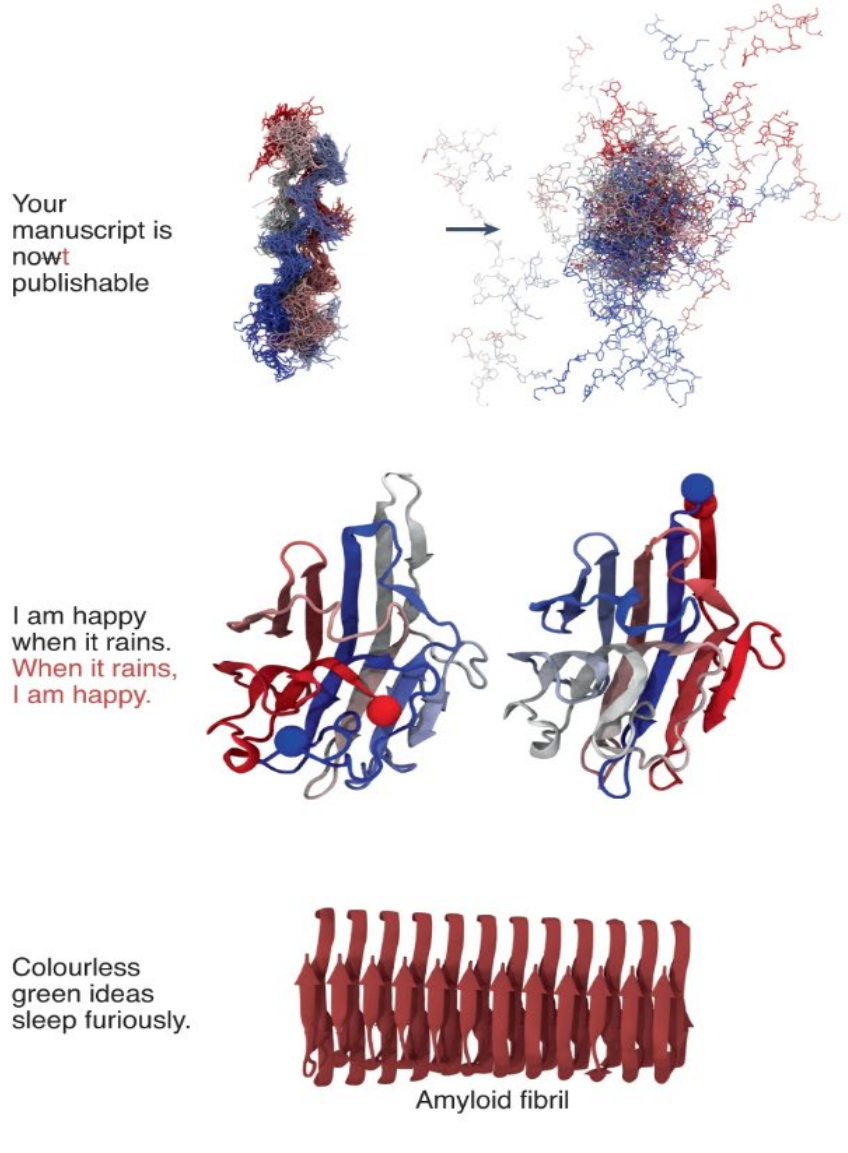
Sequence

Secondary structure

Tertiary structure

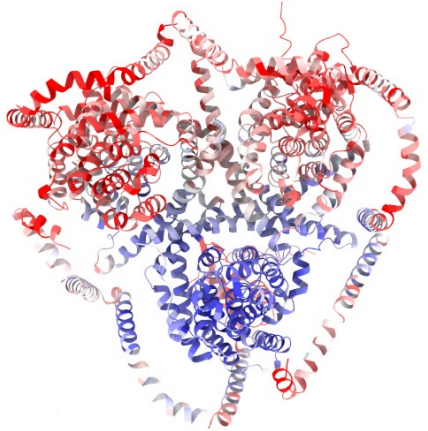
Quaternary structure



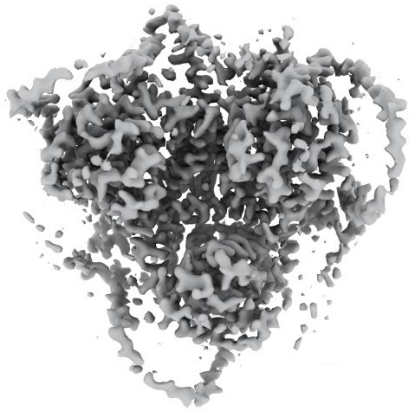


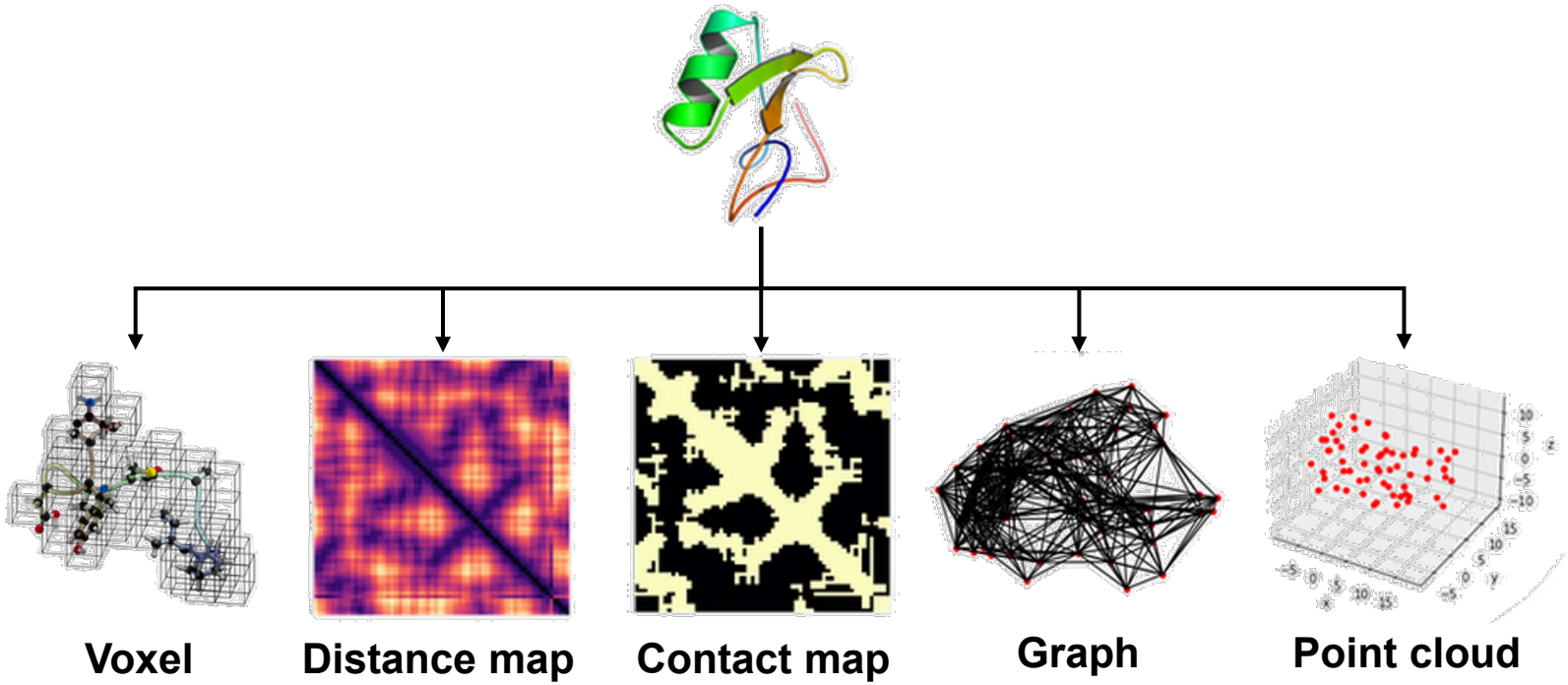
Structural data

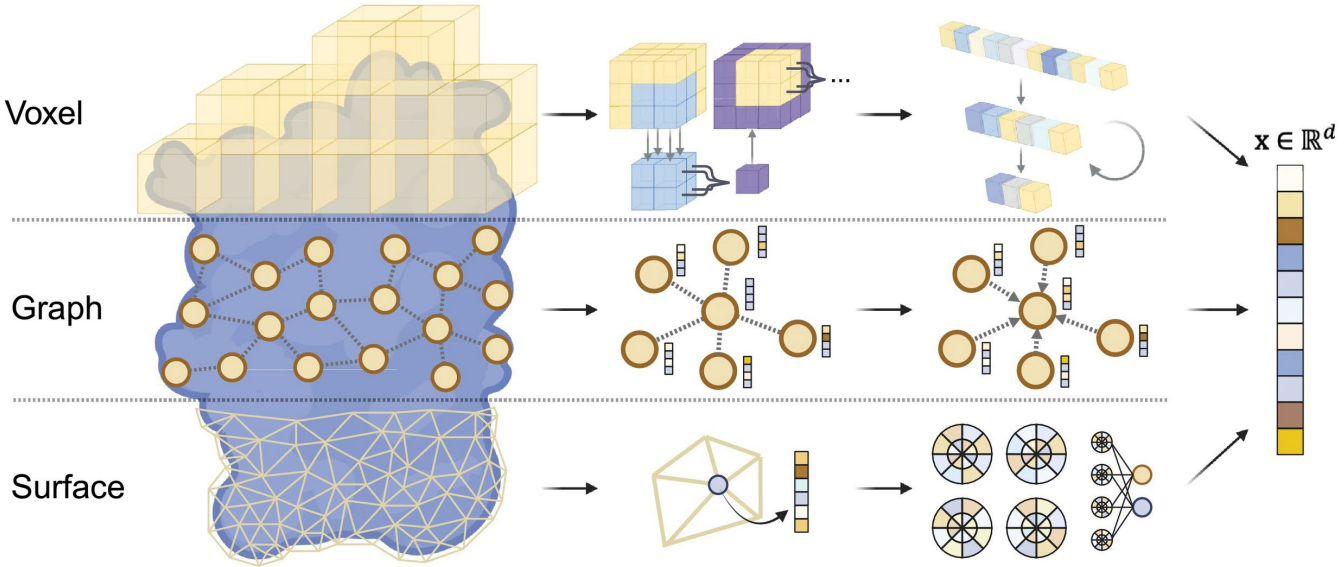
atomic model



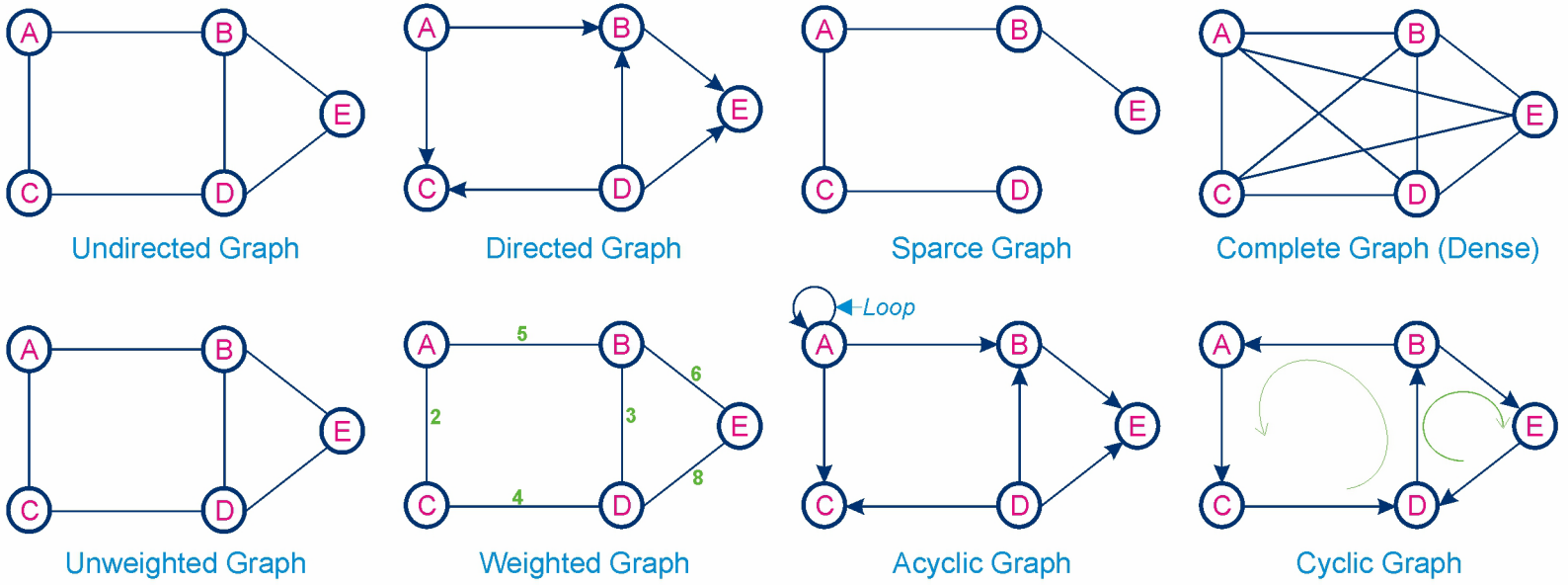
density map





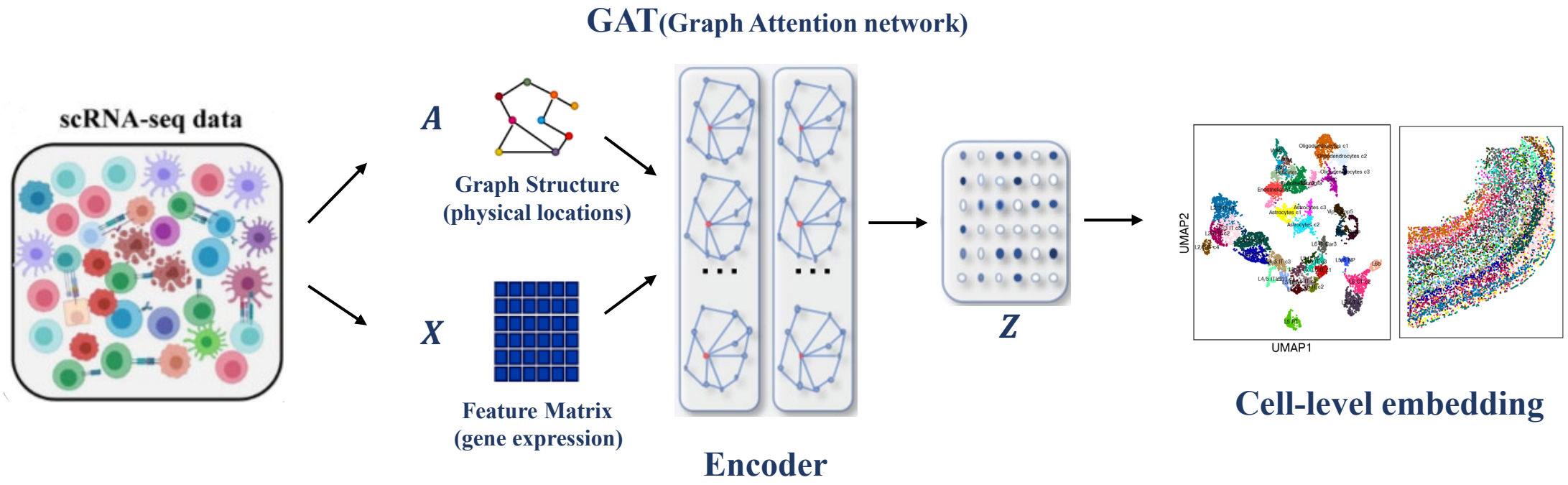


From interactions to graph



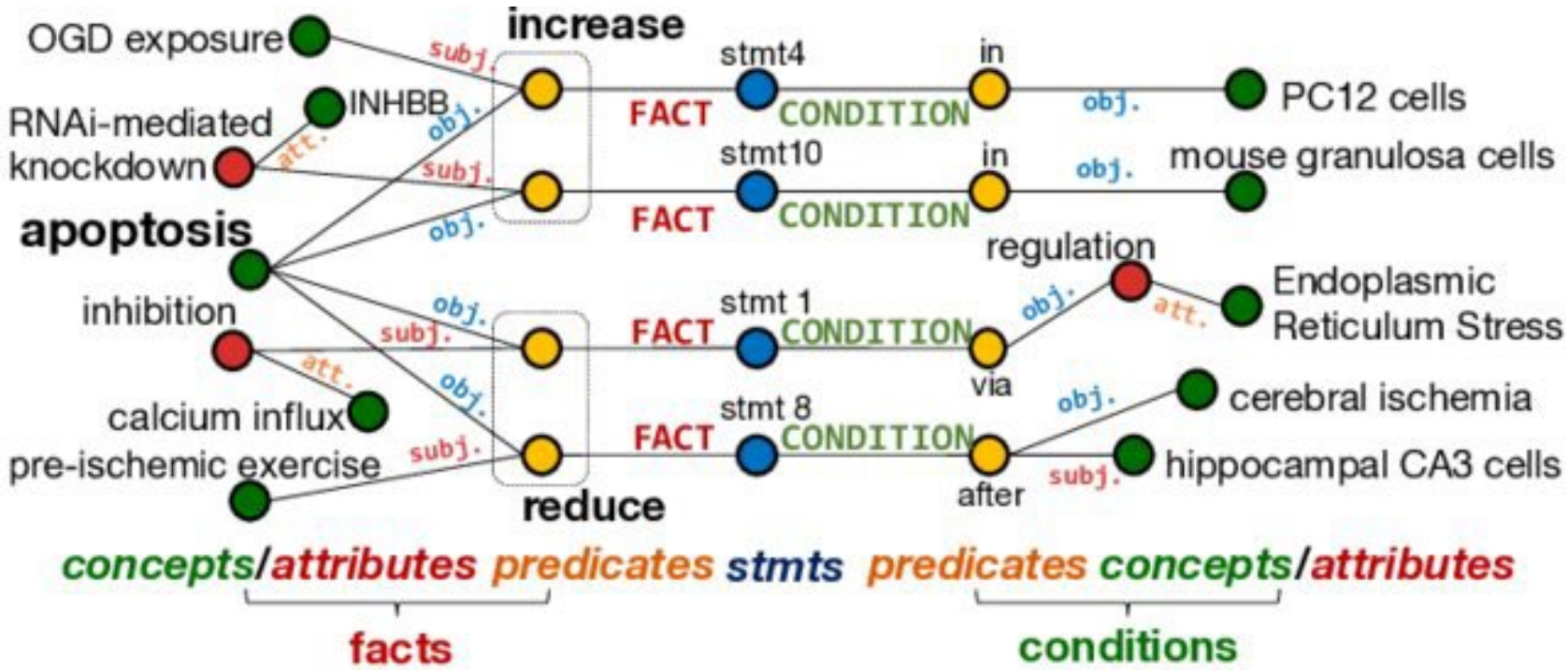
	The biological problem can be represented using a graph				
task	phylogenetic analysis	single cell data modeling	interaction network modeling	structure biology	structure biology
Graph	evolutionary tree	tissue	PPI network	backbone sequence	molecules
Edges	genetic relationship	cell-cell interactions	protein-protein interactions	peptide bond	bonds
Nodes	species	cells	proteins	residues	atoms

An example of graph data learning

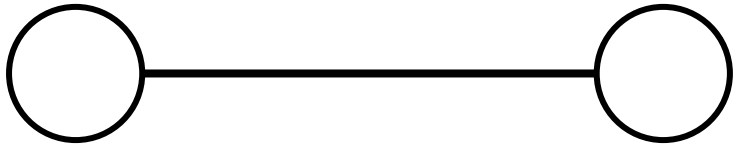


Li et al. Cell Systems 2024

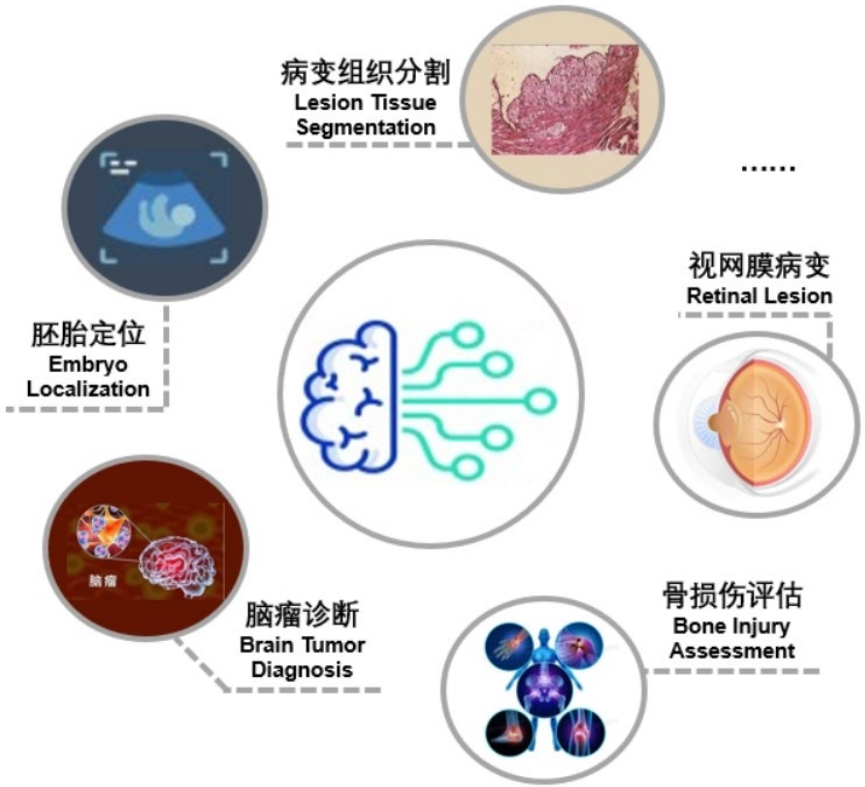
Knowledge graph

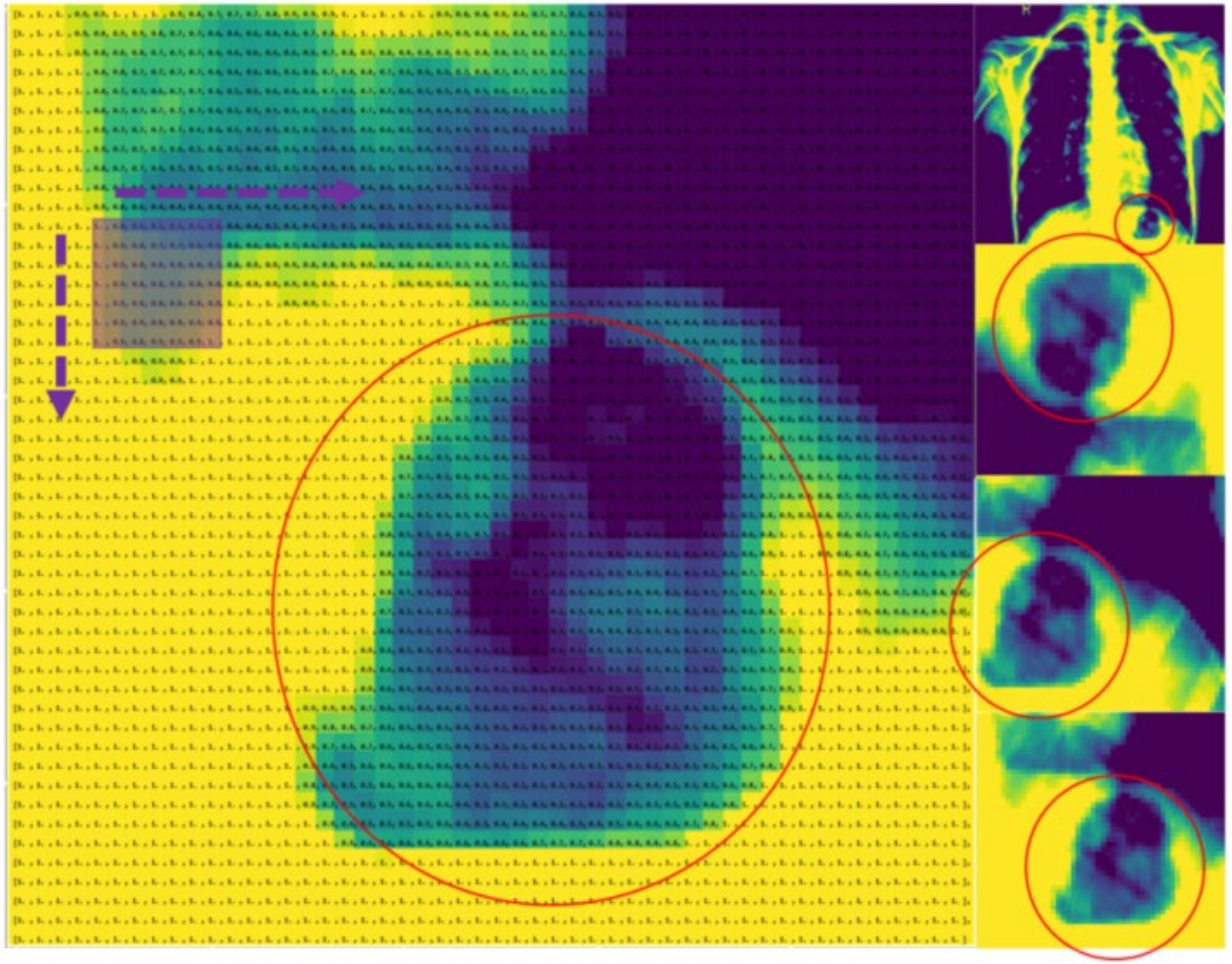


Entity - Relationship - Entity



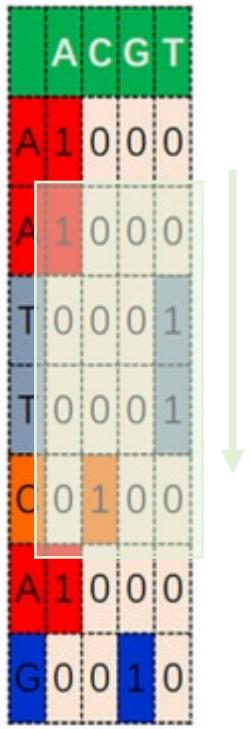
Medical imaging data



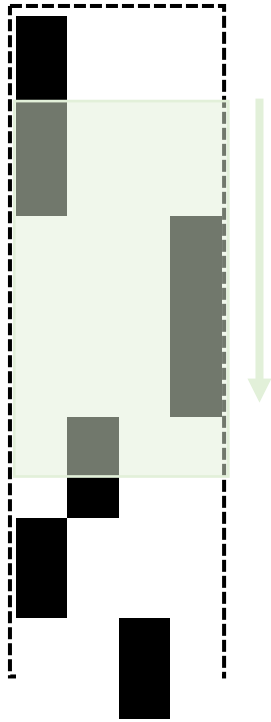


One-hot encoding is like an image

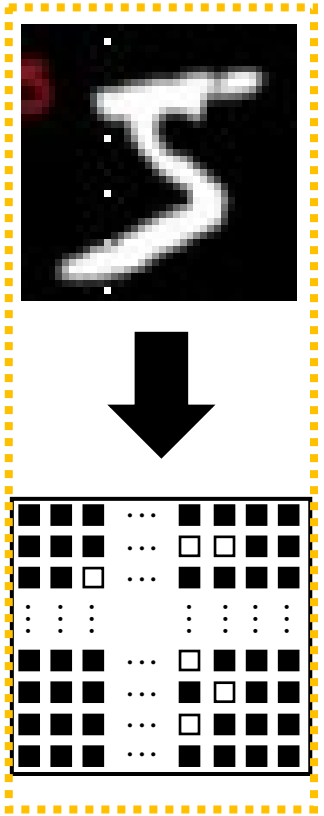
DNA sequence: AATTCAG



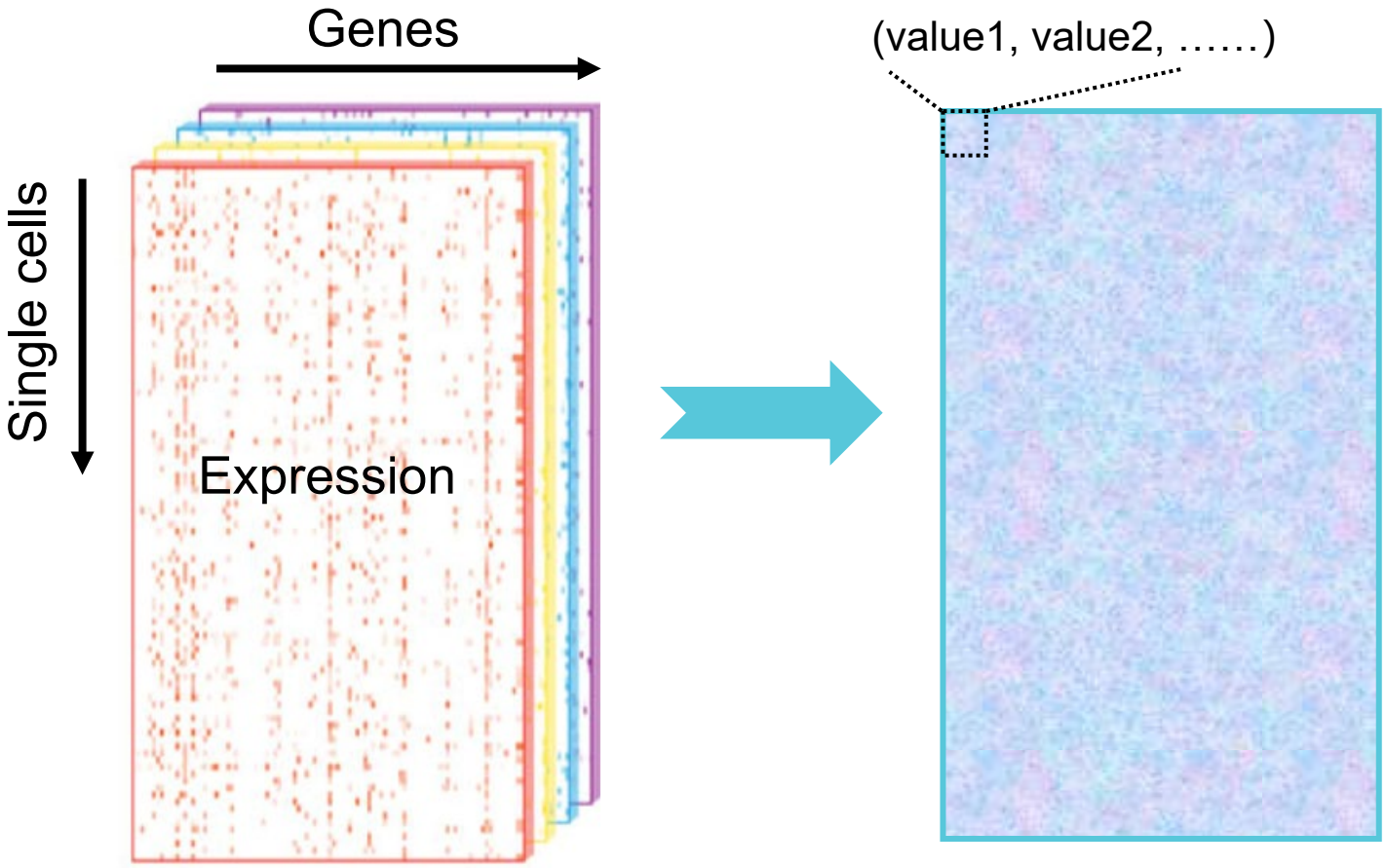
The one-hot encoding of an RNA sequence is also a type of gray image

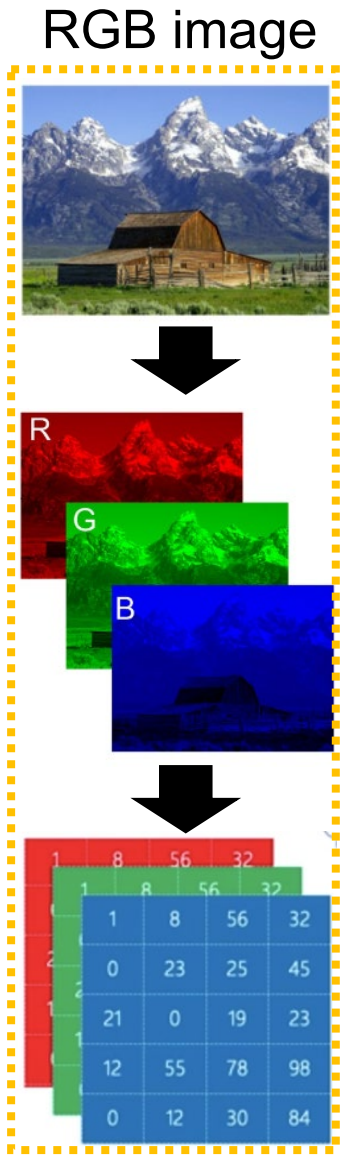


Gray image

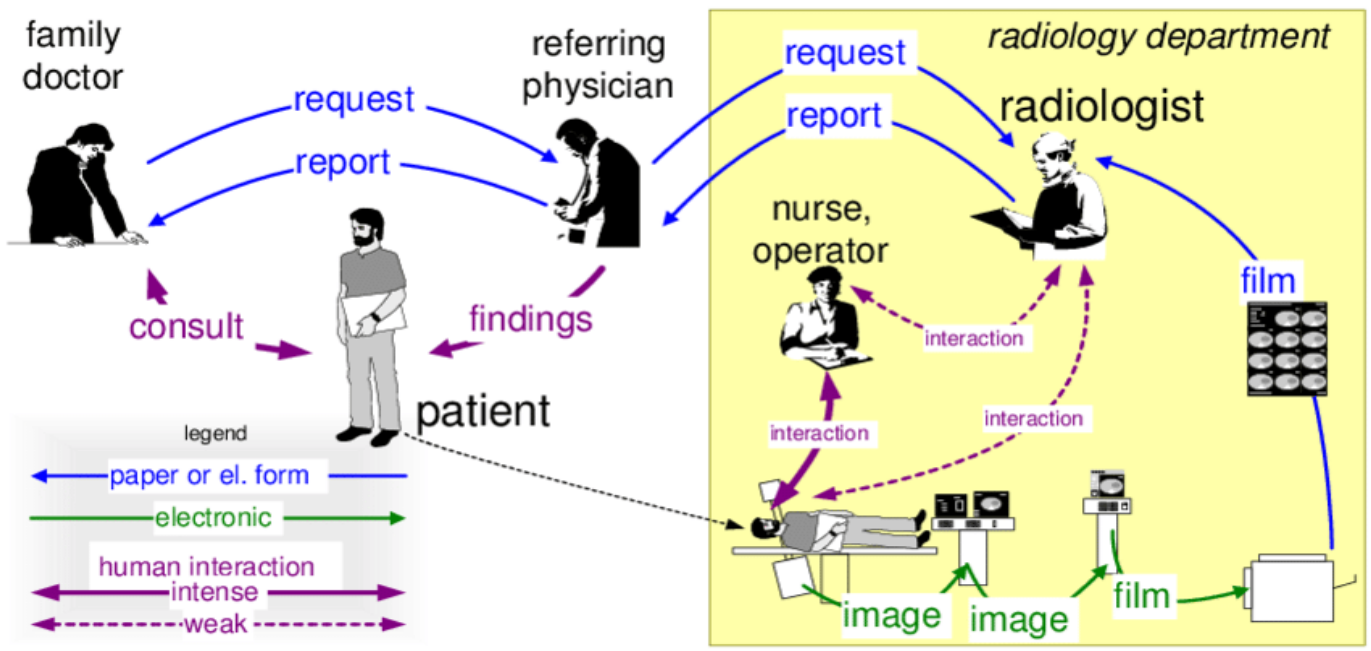


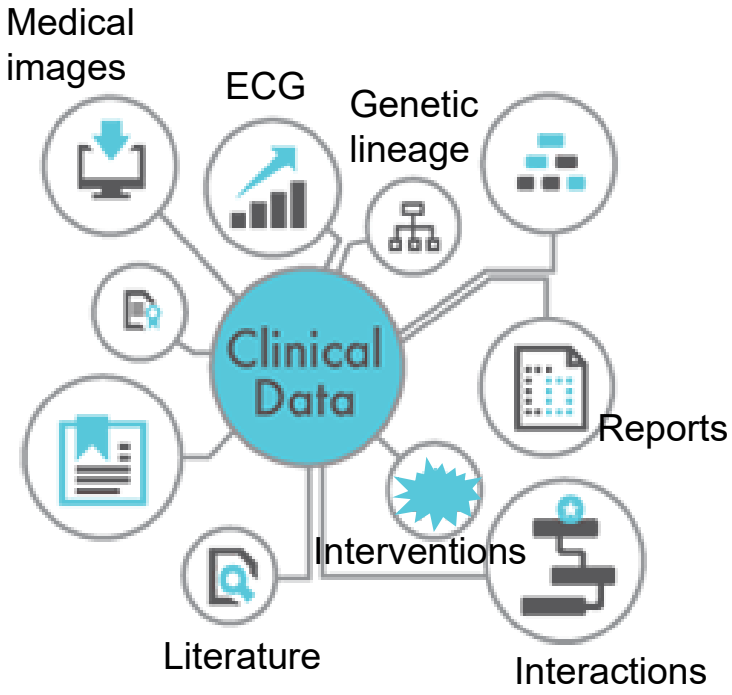
Gene expression profiles

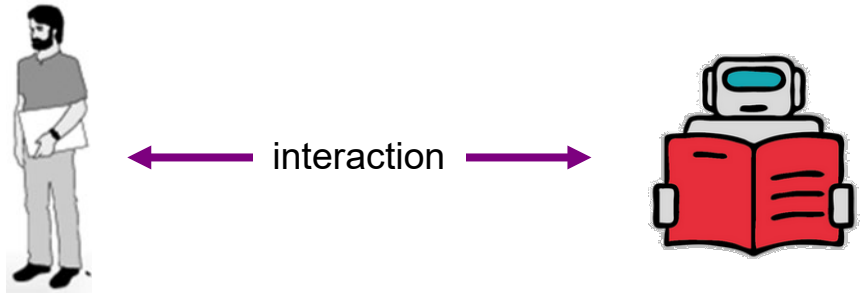




The clinical context





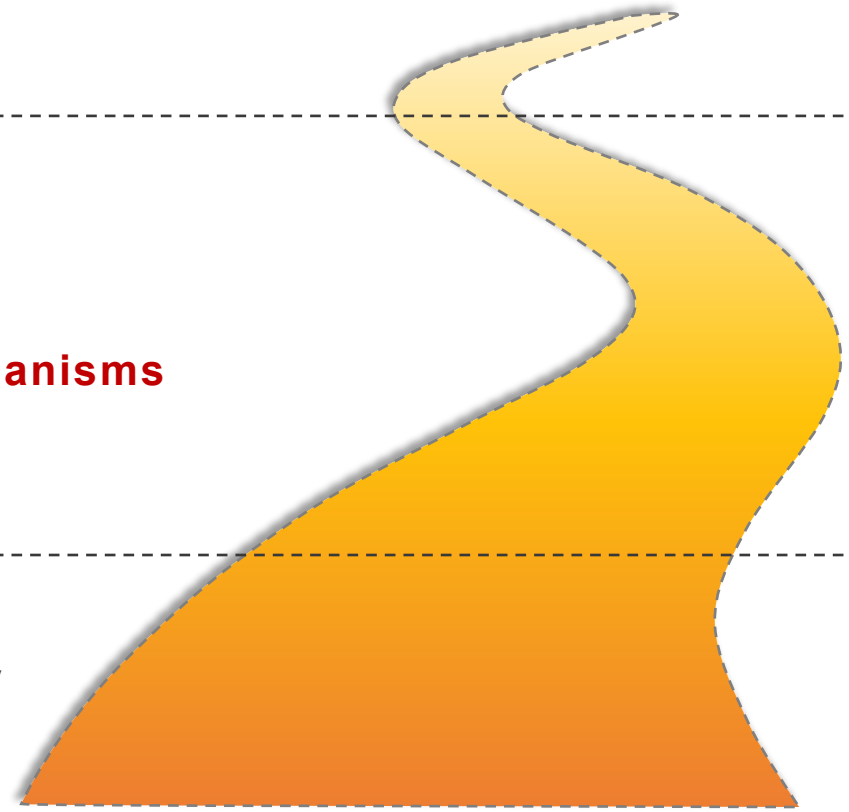


The life sciences

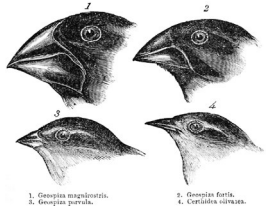
Phenotypes/
traits

Molecular
basis/**mechanisms**

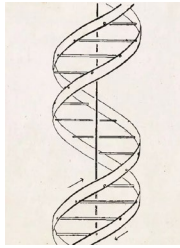
Complexity
Emergence



Life Sciences **1.0**:
Evolution



Life Sciences **2.0**:
Double helix



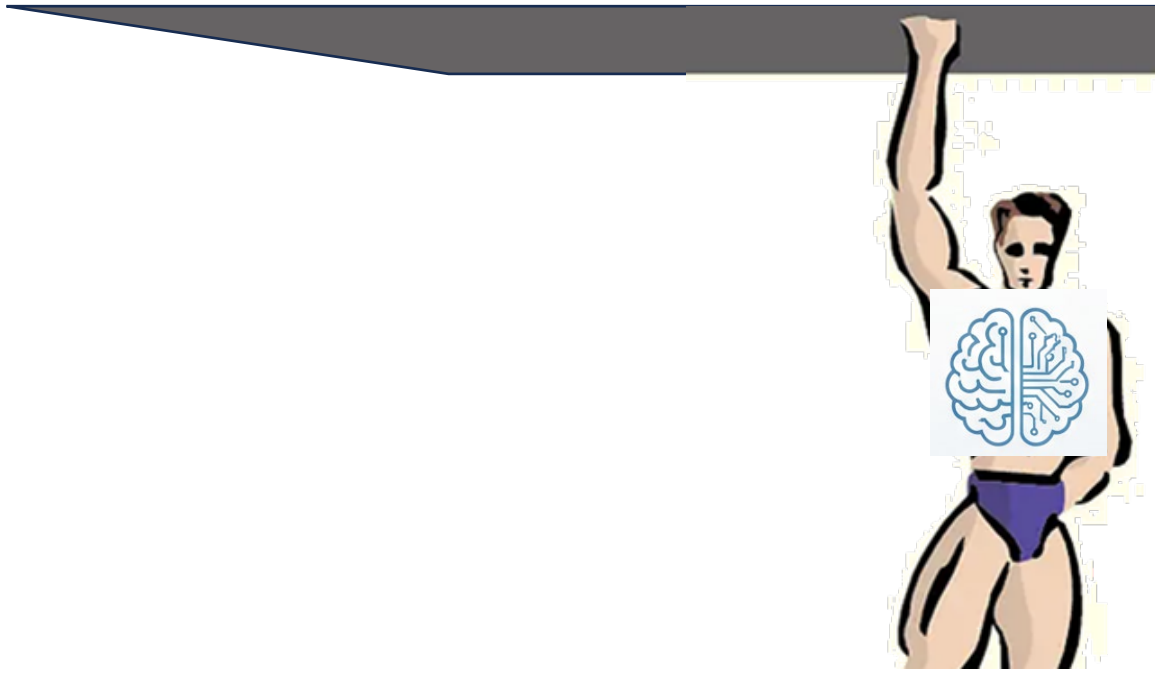
Life Sciences **3.0**:
AI?



Challenges and opportunities of AI biology

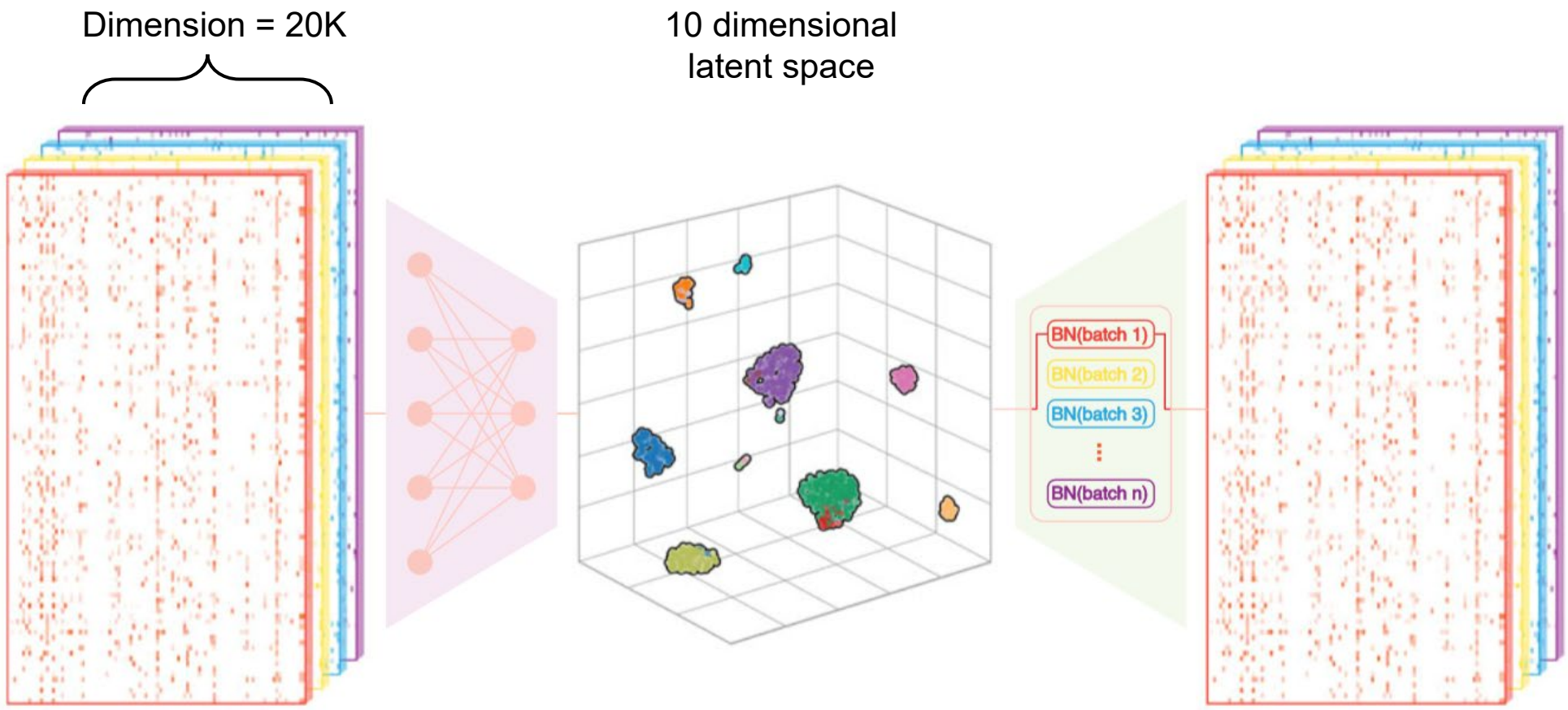


High-dimensional | Cross-scale & Multi-modal | Complex nonlinearity | Noise | Missing data | Limited samples



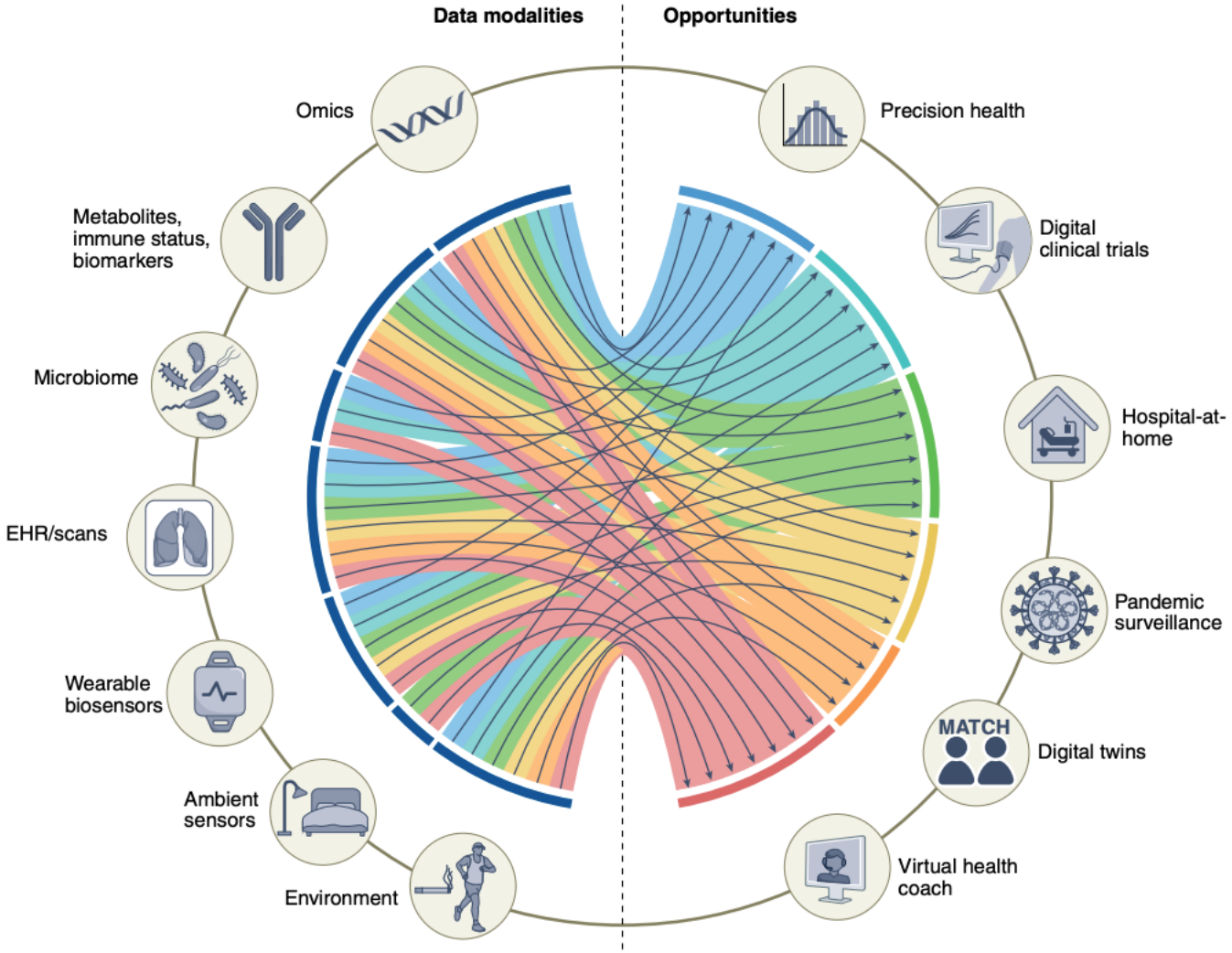
Biological problems and their associated data are highly complex, but AI possesses the capability to navigate these intricacies.

How to analyze high-dimensional data

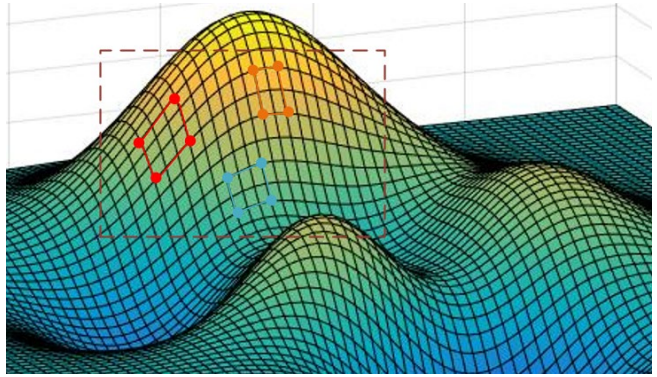


The expression data are extremely high-dimensional, where just the genes alone number up to 20k+.

Creating a unified view: multi-modal data fusion

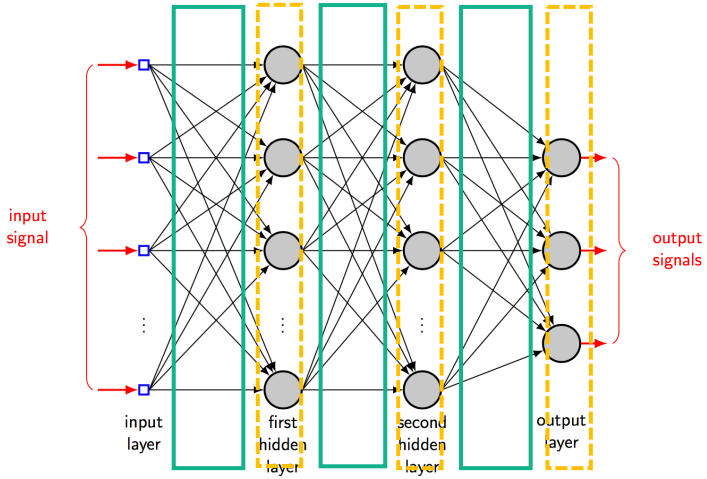


Learning from complex: the nonlinearity data problem

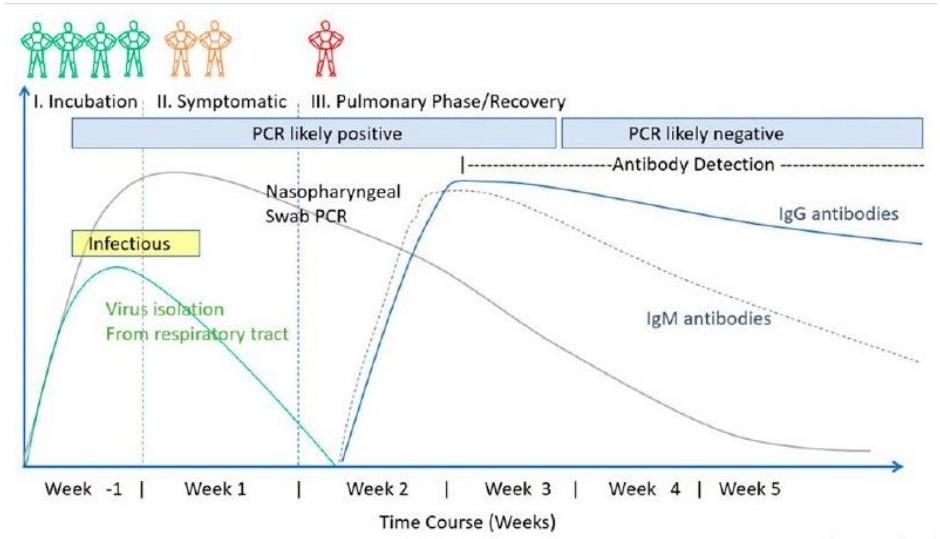


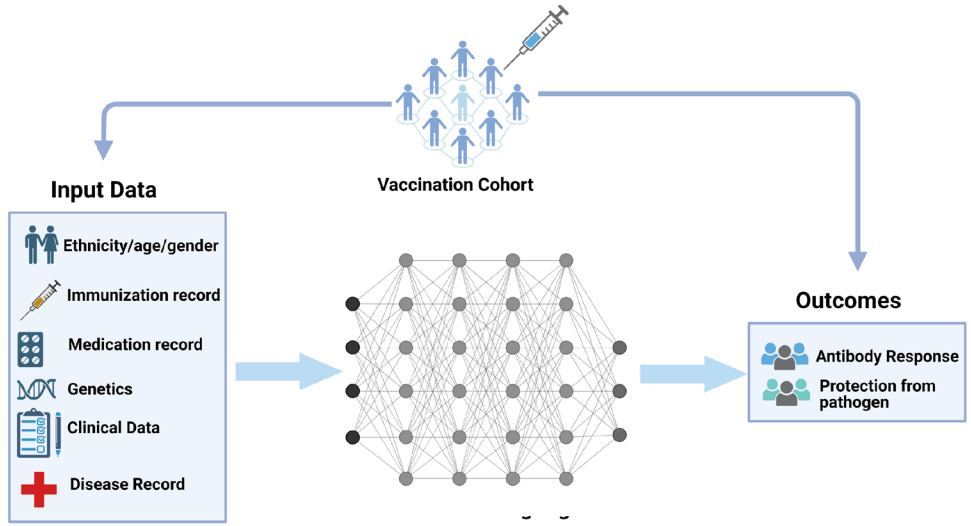
The local derivatives of a nonlinear system are linear. Therefore, such a system can be viewed as one that pieces together countless linear systems using simple nonlinear functions.

Nonlinear: the activation function of neurons

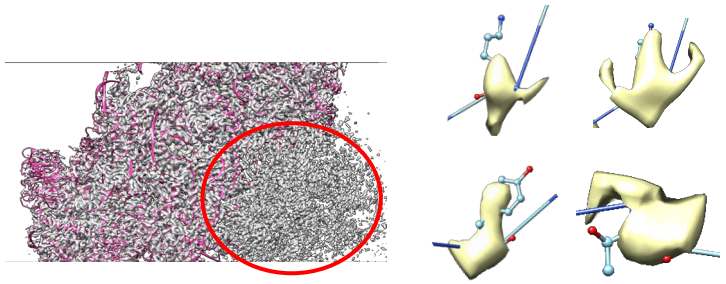


Linear: the connections between layers



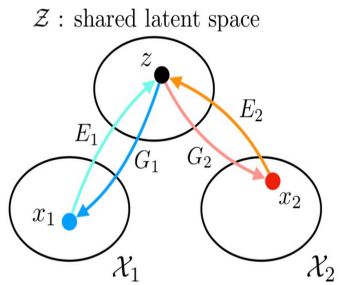


Learning from noise: the dirty data problem



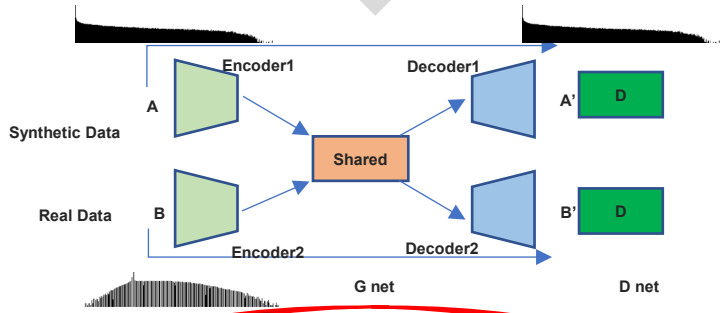
Real data are dirty Real data are noisy
(low resolution maps)



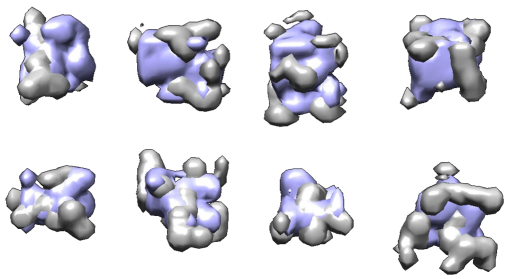


Data conversion using Generative Adversarial Networks (GAN)

Inspire

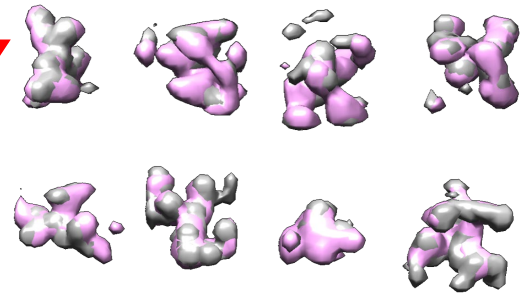


Accuracy on Raw
Real Maps: 25 %



Use GAN to transform between
real and synthetic data

Accuracy on Transferred
Real Maps: 56.3 %

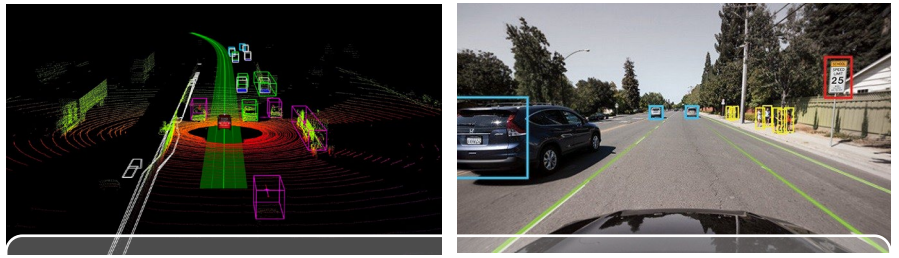


Synthetic Map Real Map Transferred Real Map

Learning from scarcity: the missing signal problem

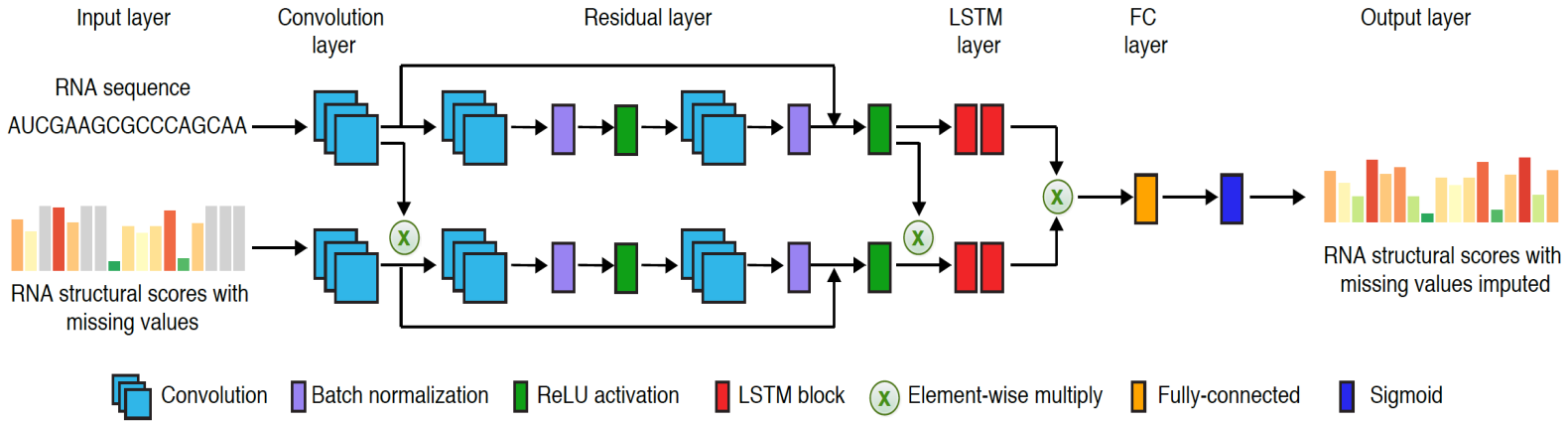


RNA structure probing data often suffers from signal dropout



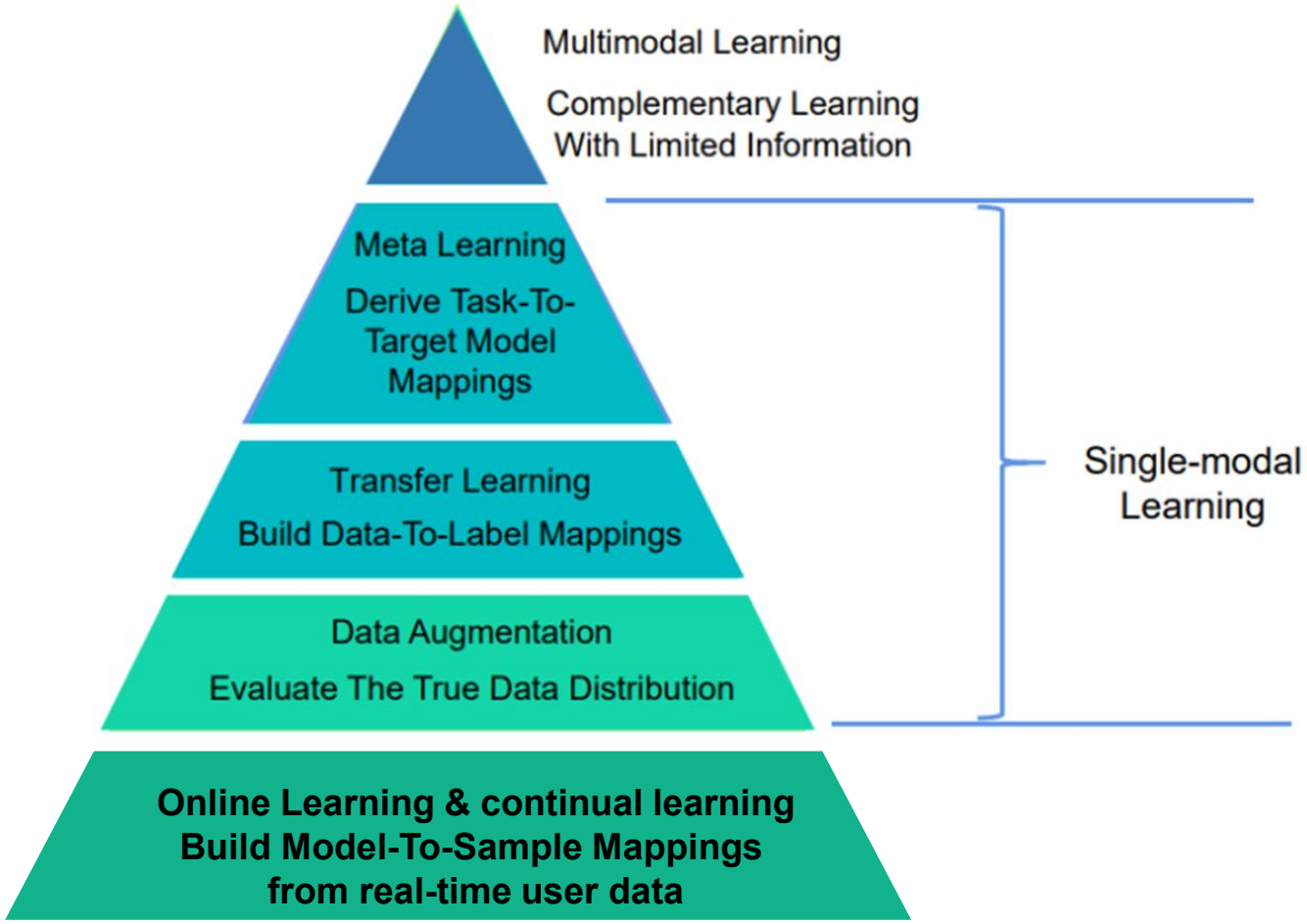
Augmenting sparse radar signals through 2D image fusion in autonomous vehicles

Inspire

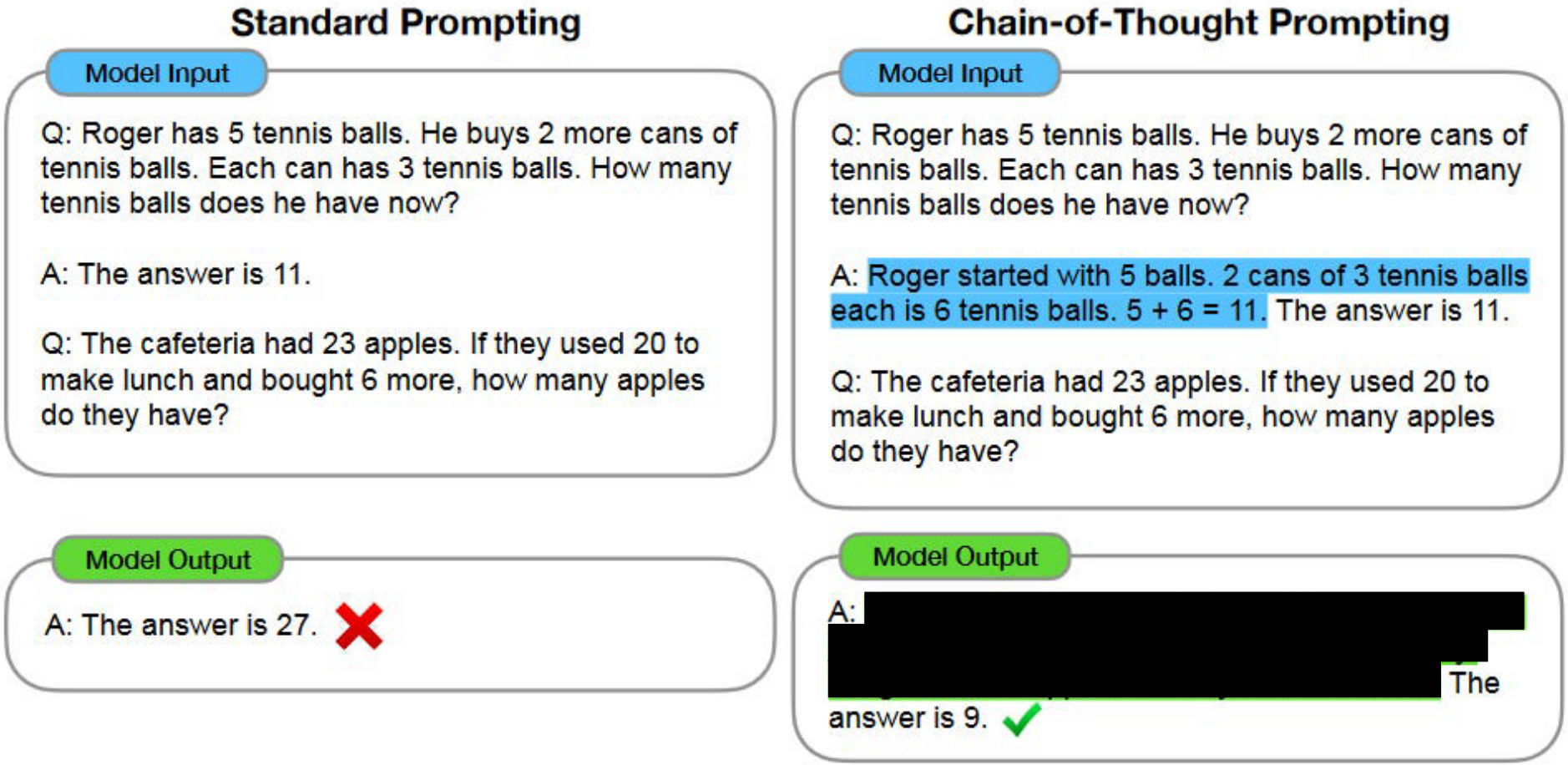


StructureImpute employs a self-supervised learning framework to recover and enhance missing signals in RNA structural profiling

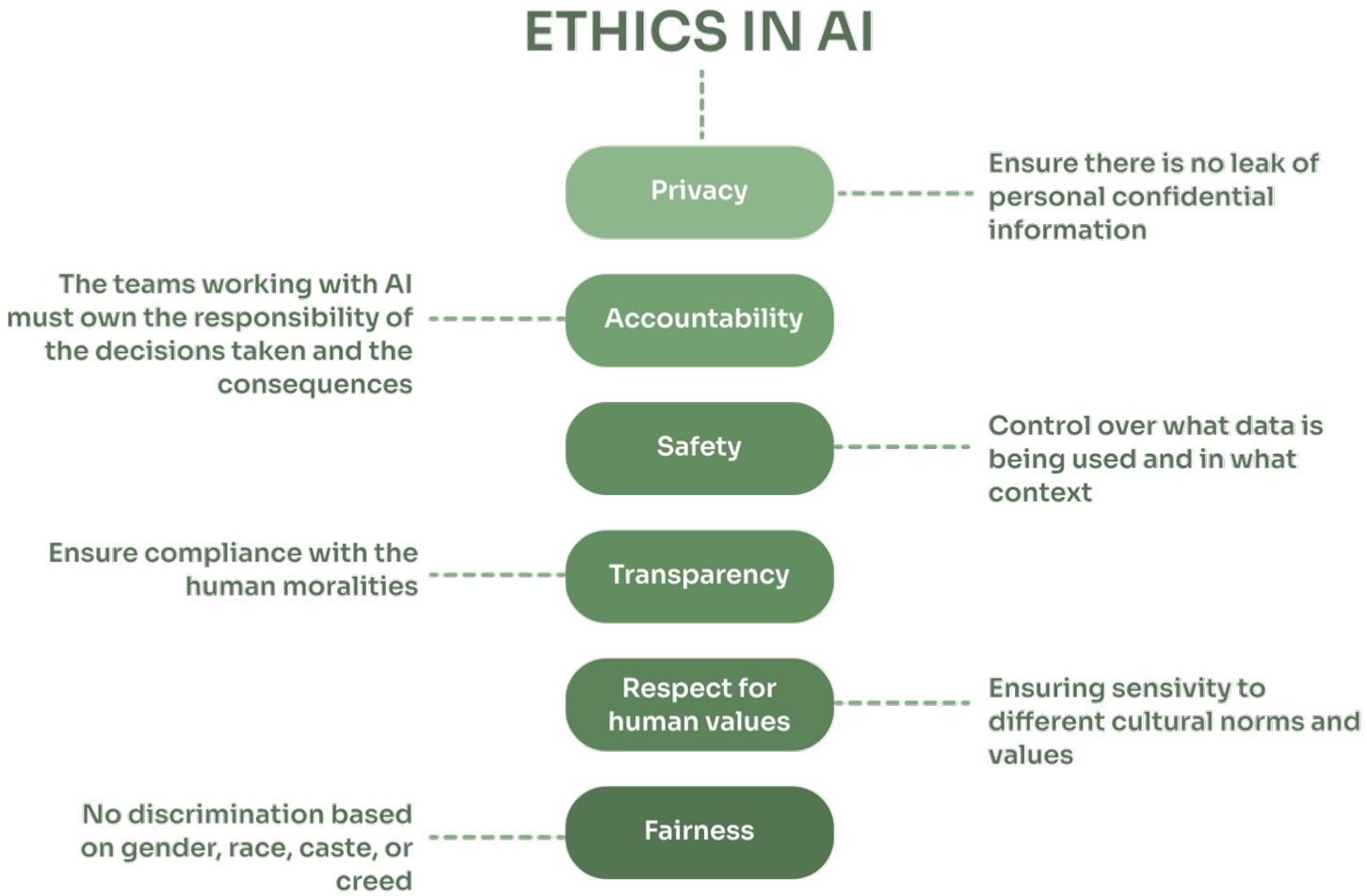
Few-shot learning: the small sample problem



The core challenge: the "Black Box" problem



AI ethics


ETHICS IN AI
Privacy
Ensure there is no leak of personal confidential information
The teams working with AI must own the responsibility of the decisions taken and the consequences
Accountability
Safety
Control over what data is being used and in what context
Ensure compliance with the human moralities
Transparency
Respect for human values
Ensuring sensivity to different cultural norms and values
No discrimination based on gender, race, caste, or creed
Fairness

Data privacy and privileges

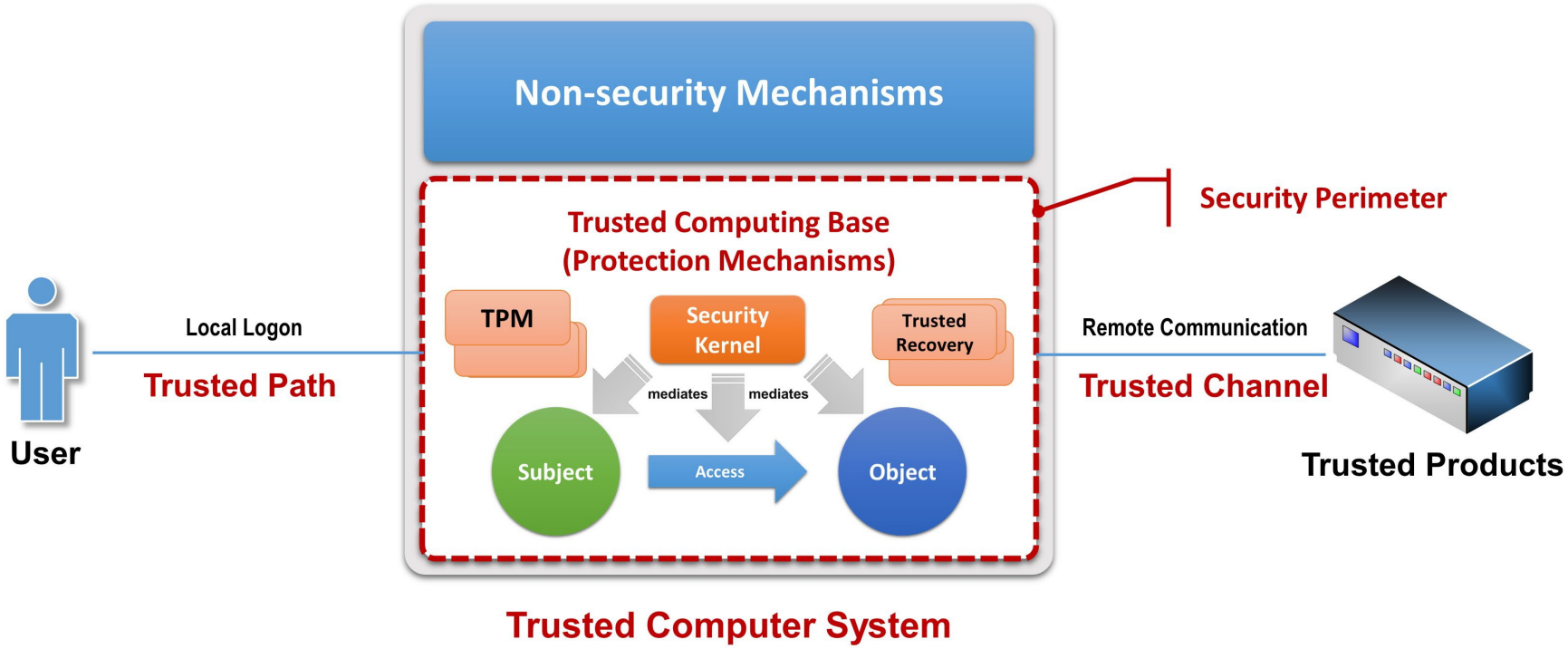
Data privacy refers to the protection and management of personal information that is collected from users. It's about ensuring that the data collected is handled in a way that respects the rights of individuals while also meeting business objectives.





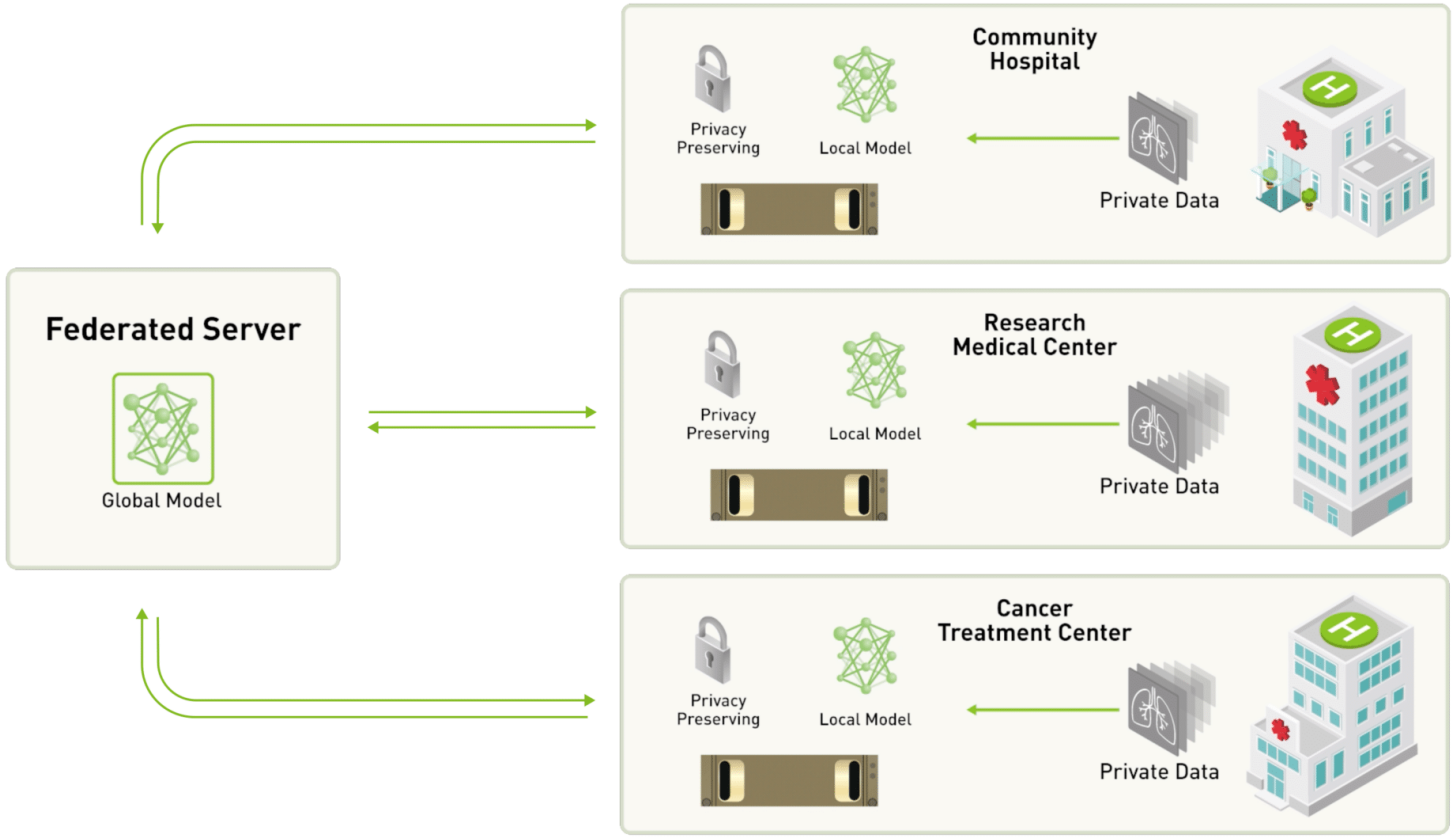
Trusted computing

- **Trusted computing** is a concept to protect and strengthen the computing platform against software-based attacks.

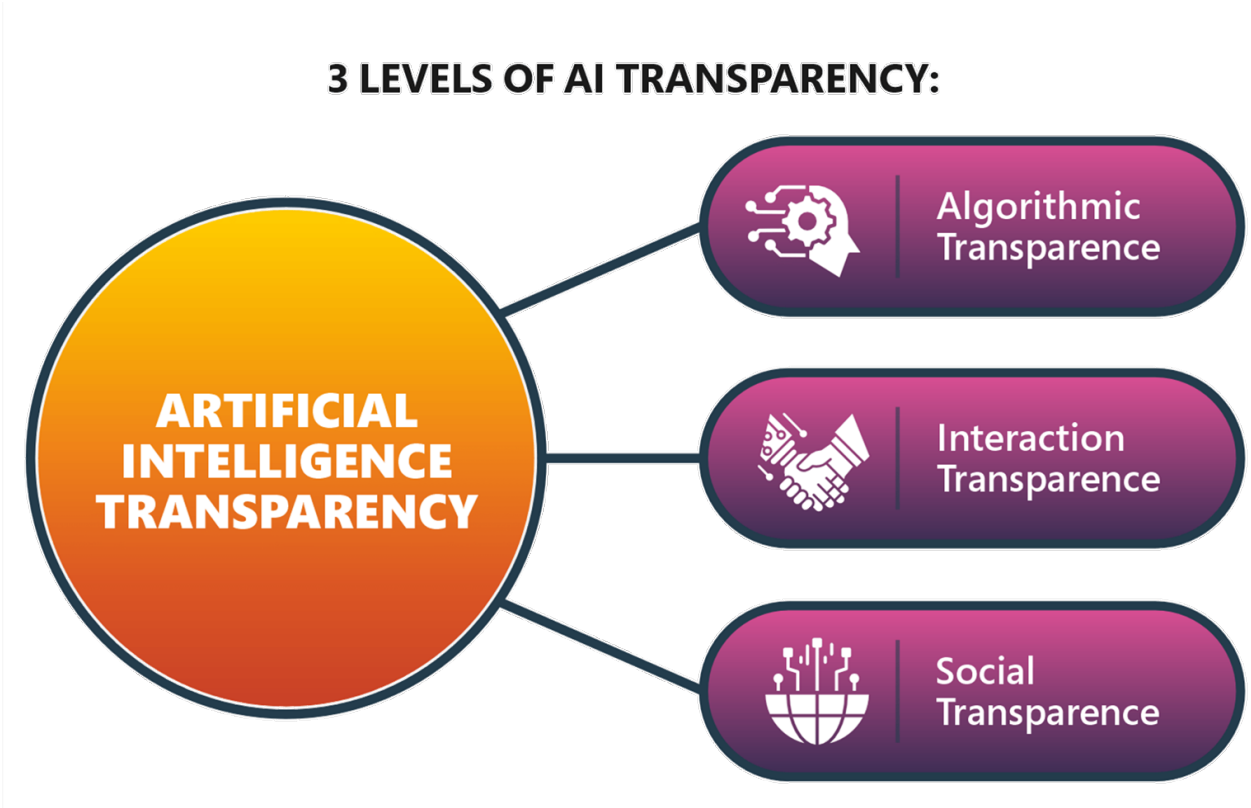


Federated learning

- **Federated learning** (also known as **collaborative learning**) is a machine learning technique in a setting where multiple entities (often called clients) collaboratively train a model while keeping their data decentralized, rather than centrally stored.



Transparency & interpretability

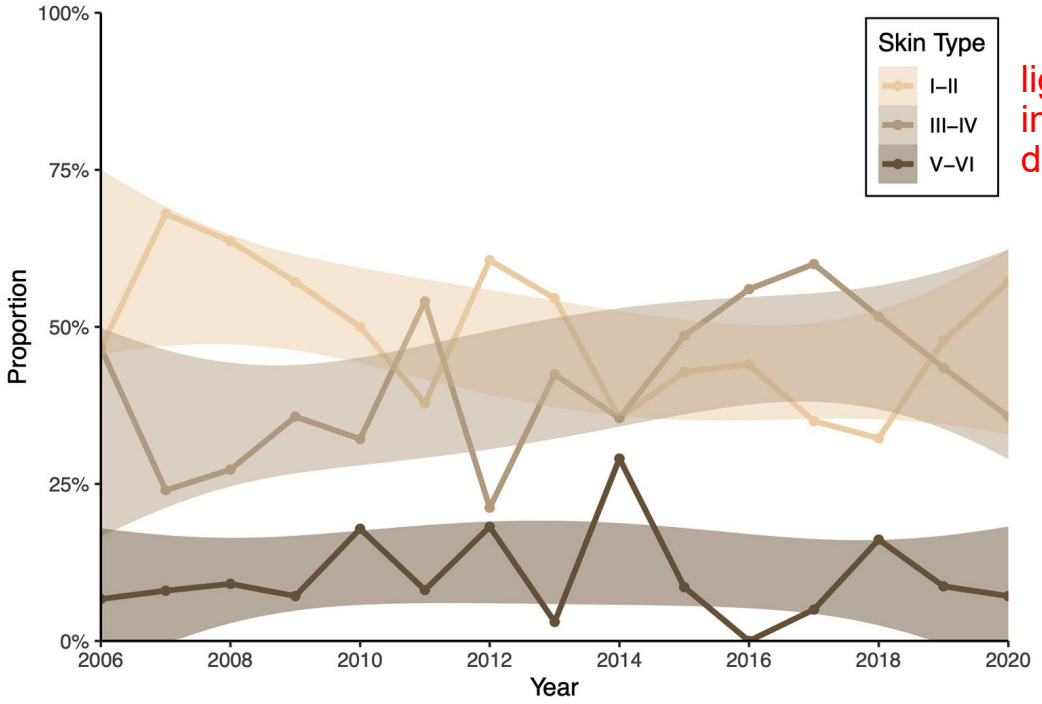


- **Algorithmic transparency** focuses on explaining the logic, interpretability, processes, and algorithms used in AI systems.
- **Interaction transparency** targets communication and interactions between users and AI systems.
- **Social transparency** goes beyond the technical aspects and aims at the broader impact of AI systems on society.

Algorithmic bias and social fairness

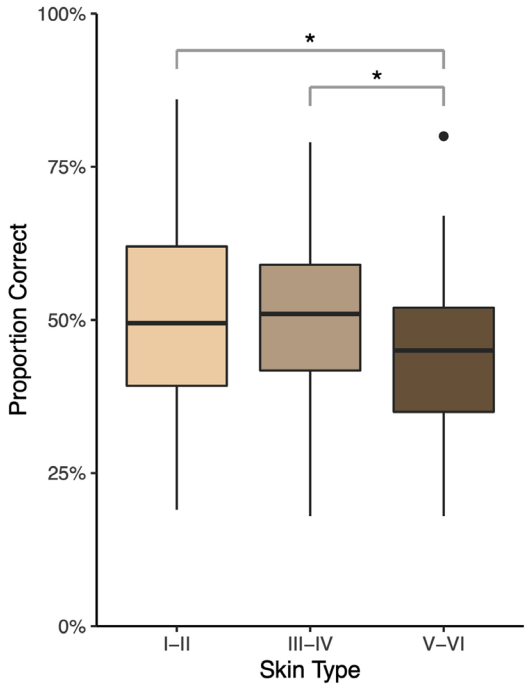
- Example: under-representation of darker skin types for skin cancer diagnosis
- AI could lead to better medical diagnosis for affluent regions and specific racial groups, thereby exacerbating existing social disparities.

Representation among skin types in the Image Challenge



light skin
intermediate skin
dark skin

Diagnostic accuracy among skin types in the Image Challenge



Data quality and equality

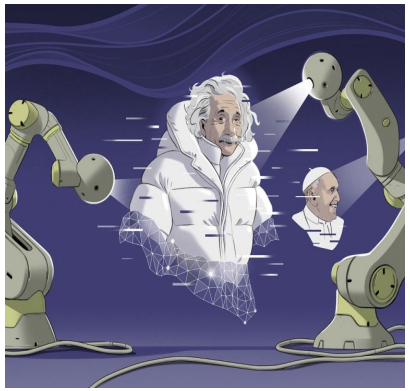
- If the data used to train an AI system contains inherent biases, the AI will learn and may even amplify these biases.







Pandora's box: technology misuse



Deepfake



Violence



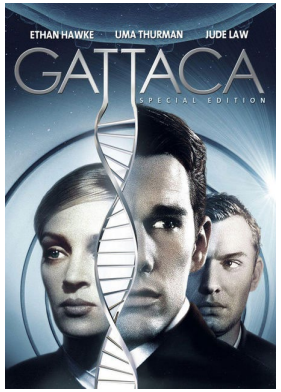
Internet fraud



Weapons & War



Image Nudification



Genes:
Privilege and Discrimination



Genomic security

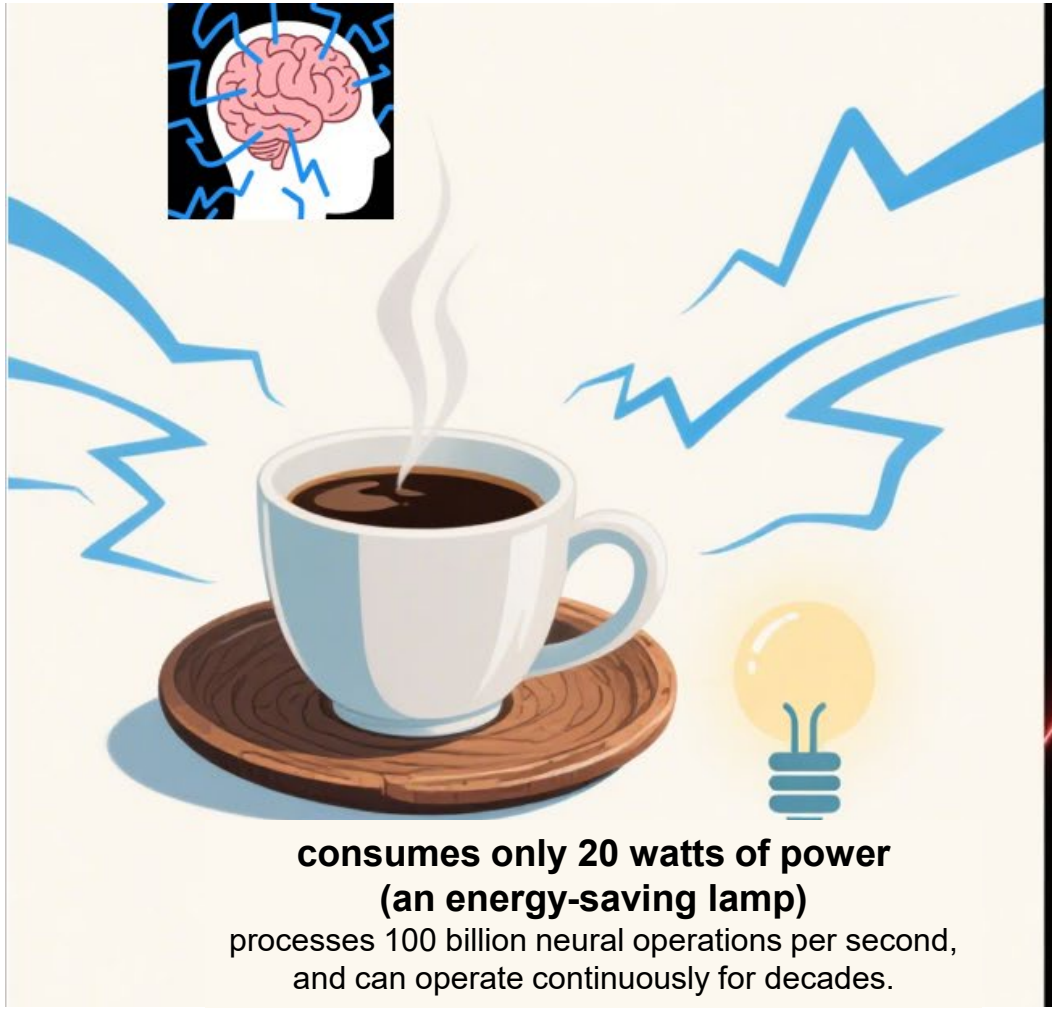


"Resident Evil"

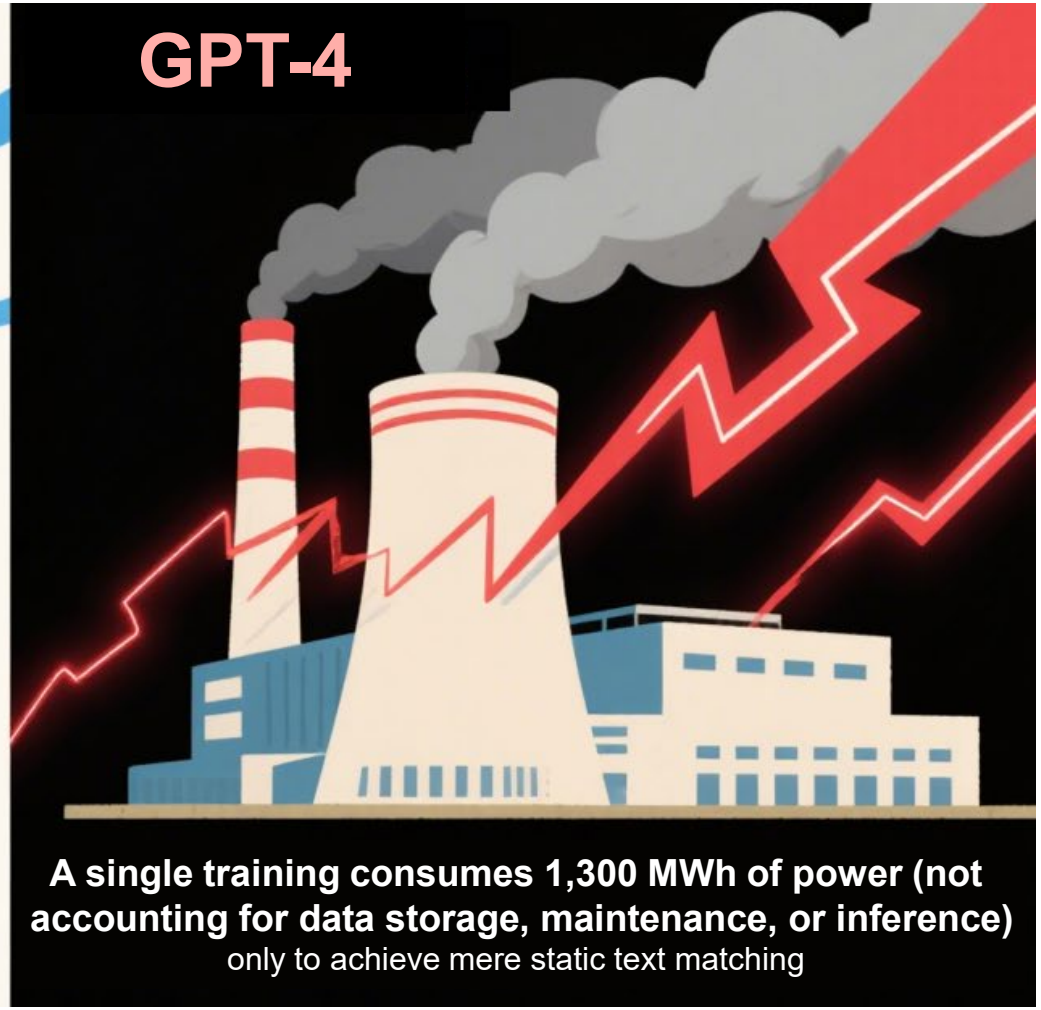
Does AI bear responsibility?



AI is a brain with higher energy consumption



consumes only 20 watts of power (an energy-saving lamp)
processes 100 billion neural operations per second, and can operate continuously for decades.



A single training consumes 1,300 MWh of power (not accounting for data storage, maintenance, or inference)
only to achieve mere static text matching

Q&A