metagenopolis
mgps.eu

How AI is enhancing human microbiome research

MetaGenoPolis

Centre de recherche INRAE de Jouy-en-Josas
Domaine de Vilvert, Bât.325
78 350 Jouy-en-Josas France

by INRAE

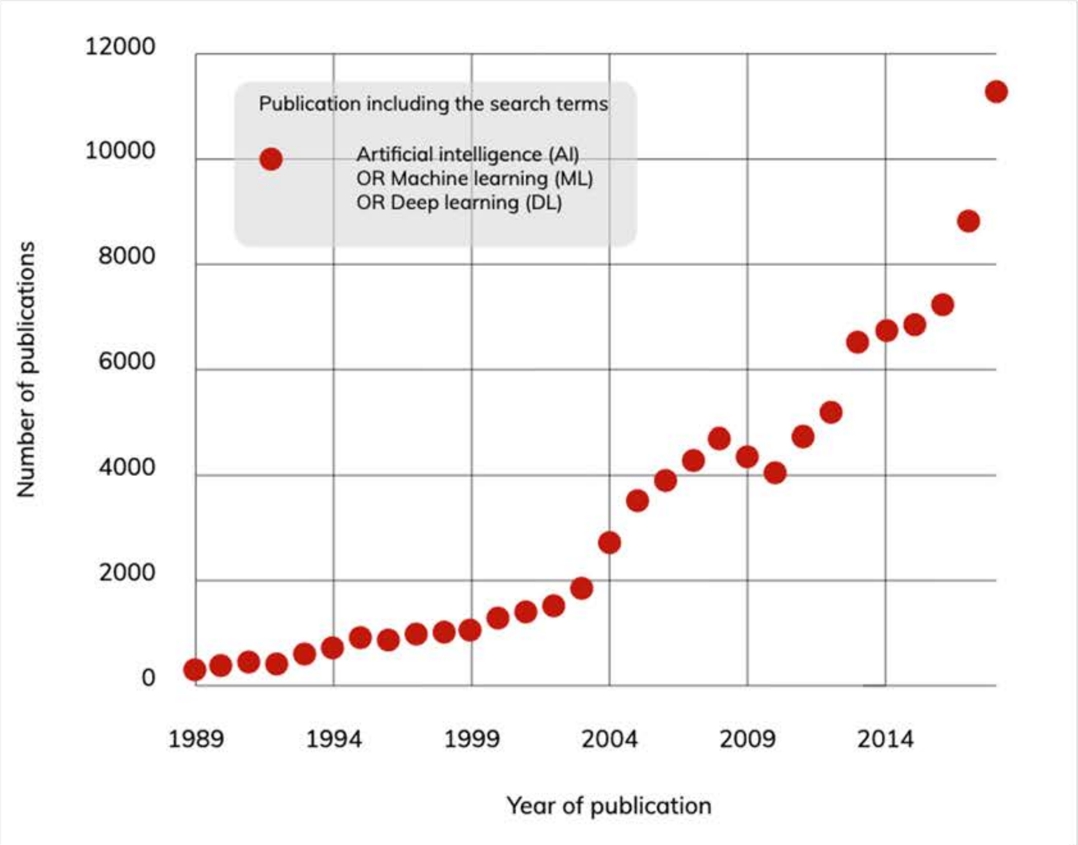
Outline

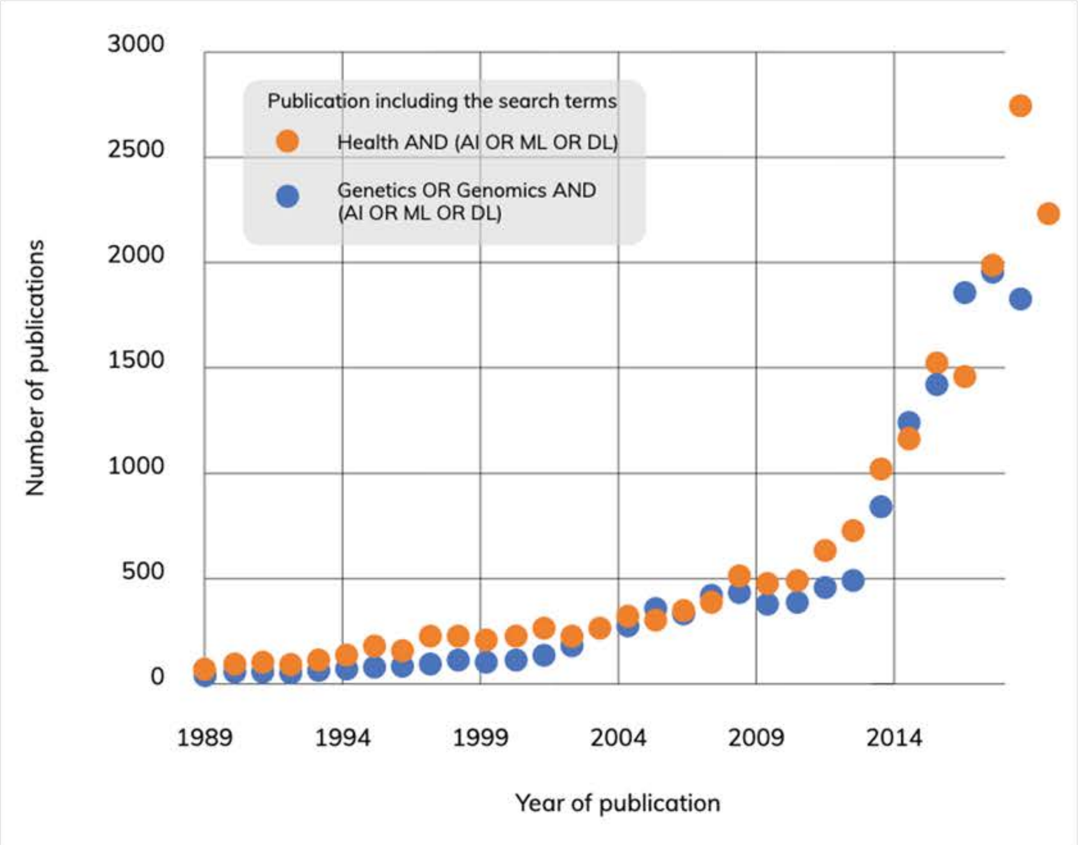
1. Defining Artificial intelligence (AI)
2. AI for microbiome data
3. Current AI challenges for microbiome data

Defining

Artificial intelligence (AI)

AI: a trendy topic



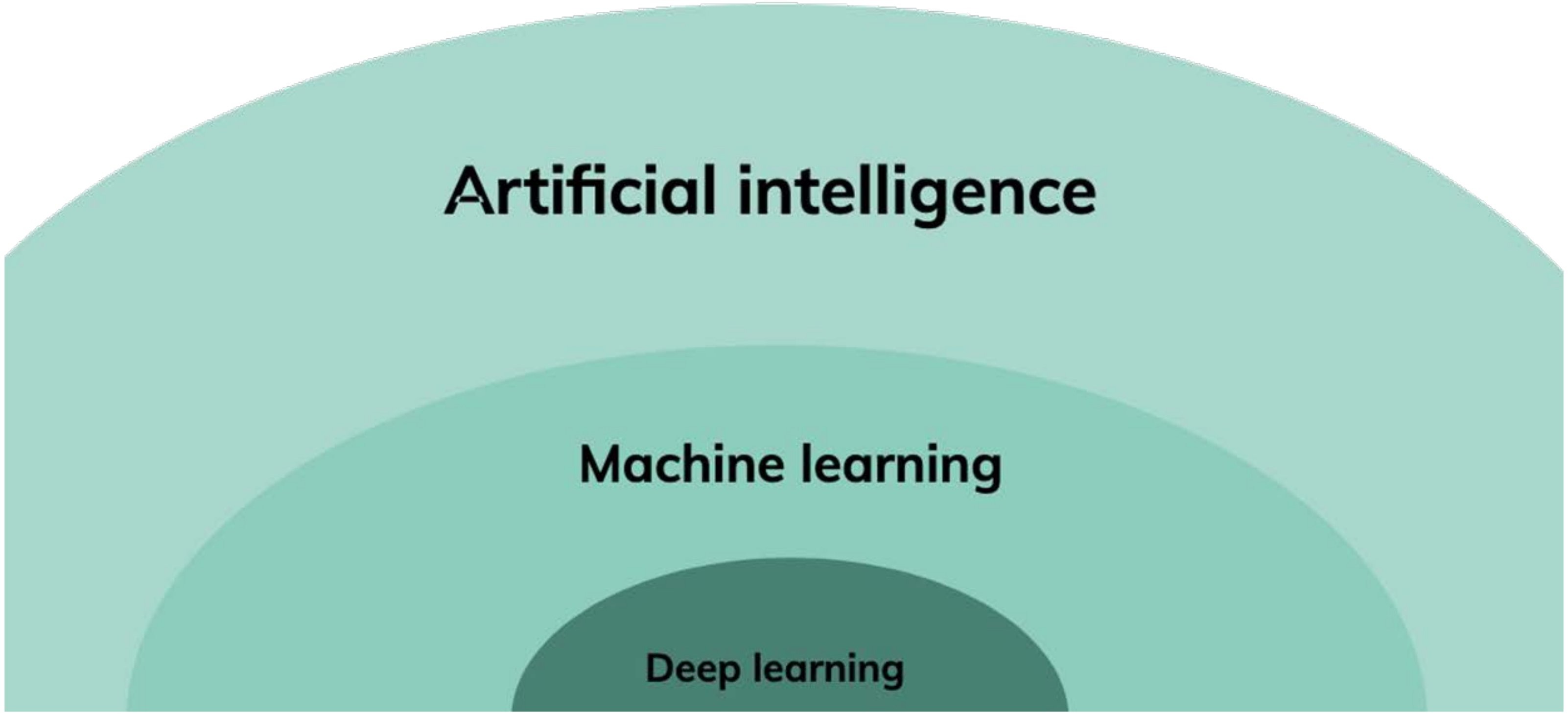


[Artificial intelligence for genomic medicine Report]

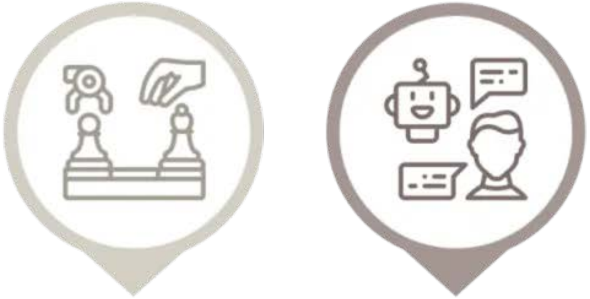
What do we mean by AI?

AI = development and use of **computing systems** concerned with making machines work in an **intelligent way**

[Artificial intelligence for genomic medicine Report]



Narrow AI



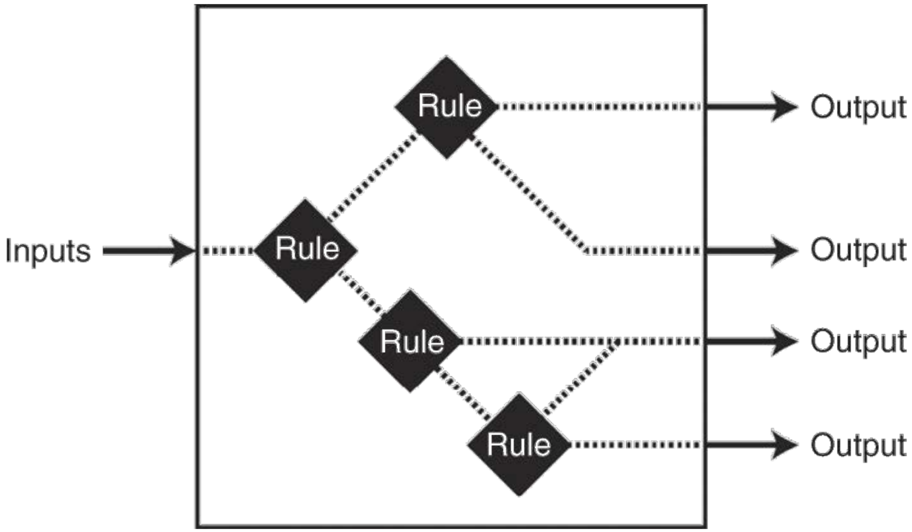
Focuses on performing one specific task

General AI

Sentient machine with 'human' reasoning

First generation of AI: rule based systems

Encoding the knowledge of a **human expert** in an **automated system**



Ex : Clinical decision support systems

Limitations

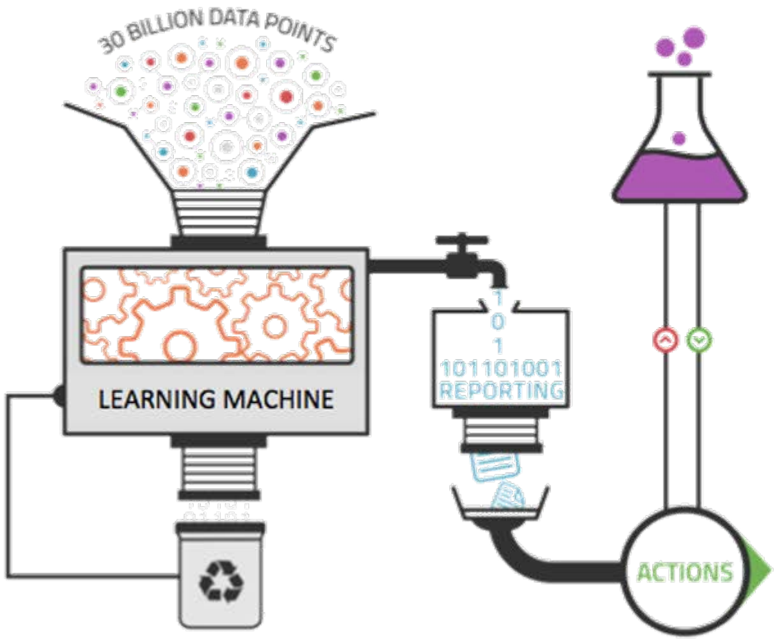
- Expensive and long to build (in-depth expertise)
- Difficult to develop (complex systems)
- Rigid (manual updates)

Advantages

- Highly interpretable

Machine Learning

Set of methods based on **algorithms** that use **mathematical procedures** to analyze data structuring



Machine learning algorithms 'learn' from data and can improve

Advantages

- Less demanding to build (data-driven learning)
- Less difficult to encode (rules established by the process)
- More flexible (integration of new data)

Limitations

- More difficult to interpret (especially deep learning)

"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed"

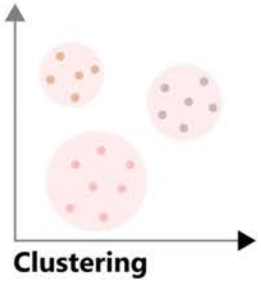
— Arthur L. Samuel, AI pioneer, 1959

The Three Types of Machine Learning Algorithms

metagenopolis mgps.eu

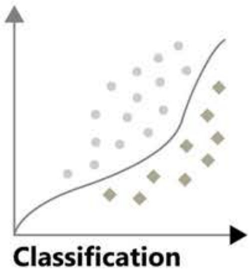
Unsupervised Learning

- ▷ No labels
- ▷ No feedback
- ▷ Find an underlying structure in the data



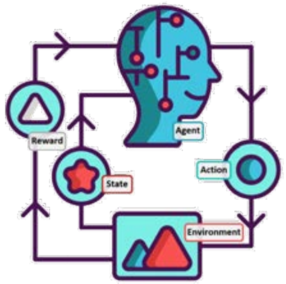
Supervised Learning

- ▷ Labelled data
- ▷ Direct feedback
- ▷ Prediction of an output



Reinforcement Learning

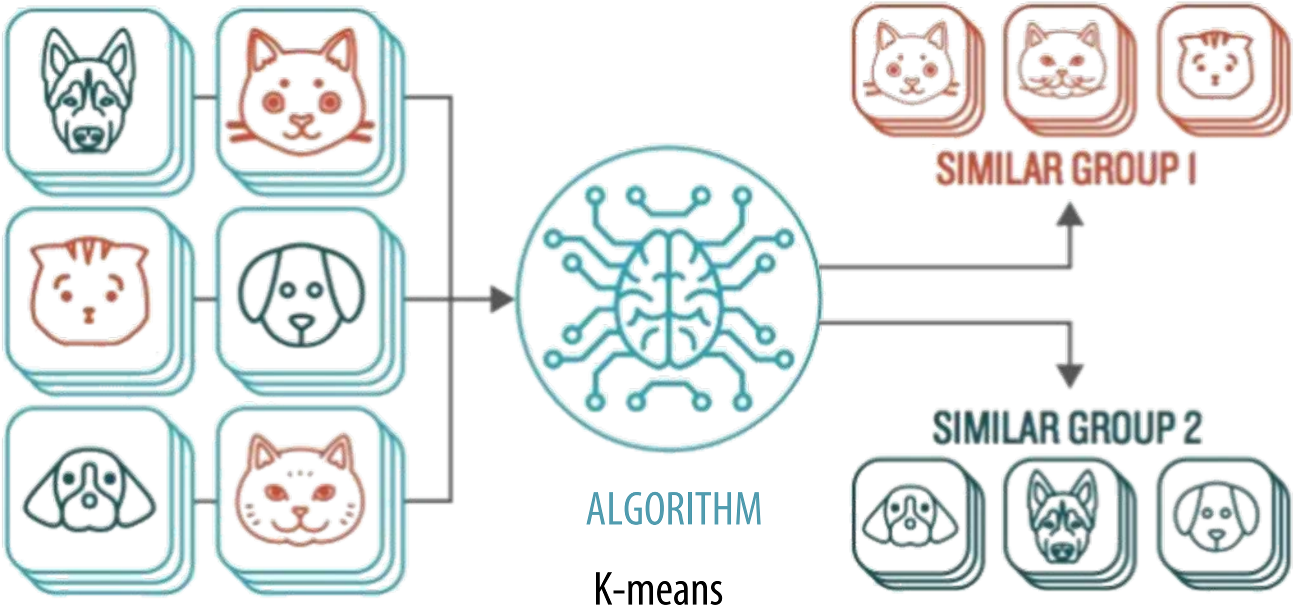
- ▷ A set of rules / No labels
- ▷ Reward system
- ▷ Iterative self-teaching



How unsupervised learning works

STEP 1

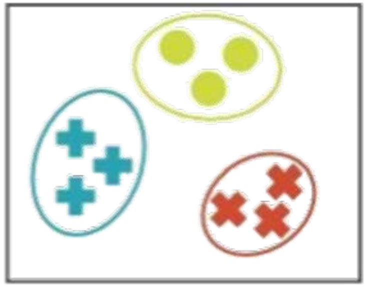
STEP 2



K-means
Hierarchical Clustering
Gaussian Mixte Model
Principal Component Analyses
Multidimensional scaling (MDS)
. . .

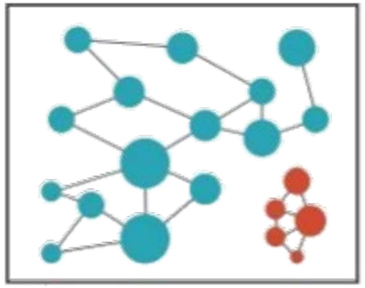
Source: adapted from Booz Allen Hamilton

TYPES OF PROBLEMS



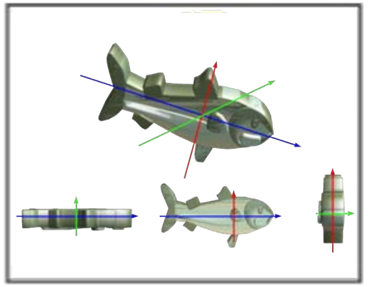
CLUSTERING

Identifying similarities in groups



ANOMALY DETECTION

Identifying abnormalities in data

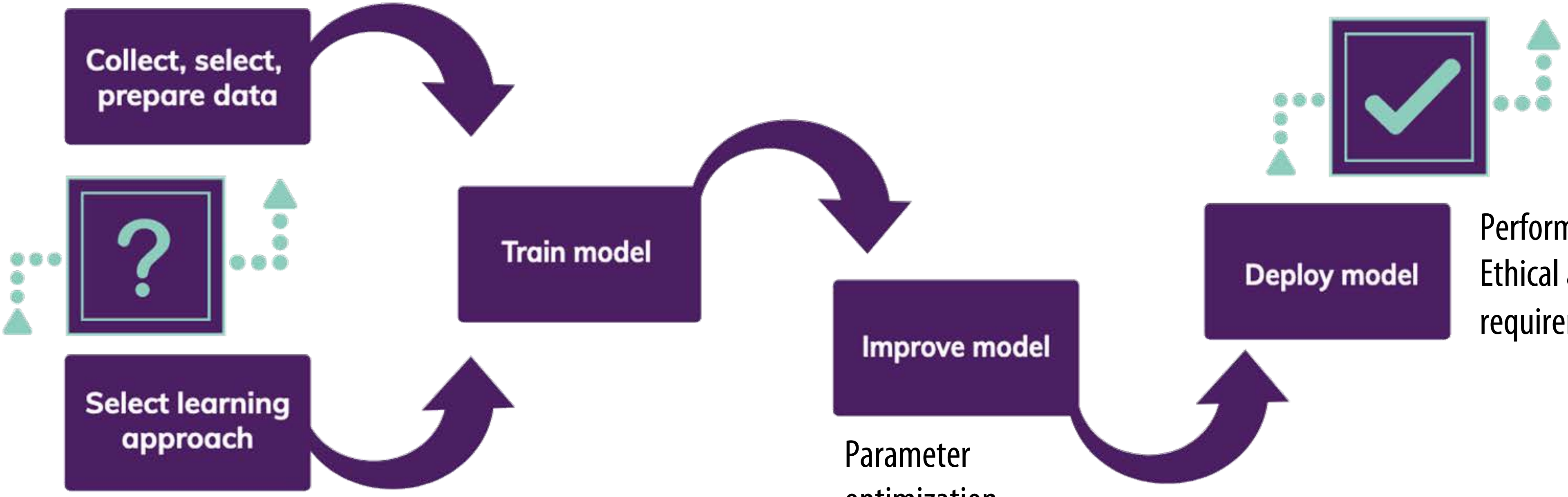


DIMENSIONALITY REDUCTION

Concise input for supervised learning

Building a supervised learning model

Data organization
and preparation



Performance criteria
Ethical and regulatory requirements

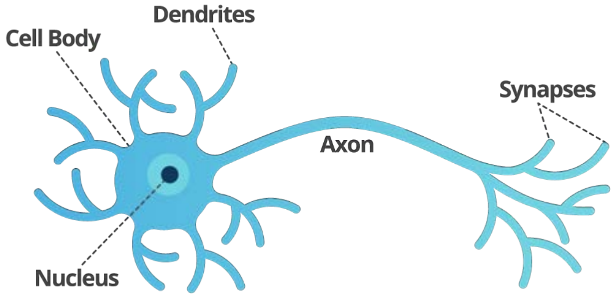
Parameter optimization

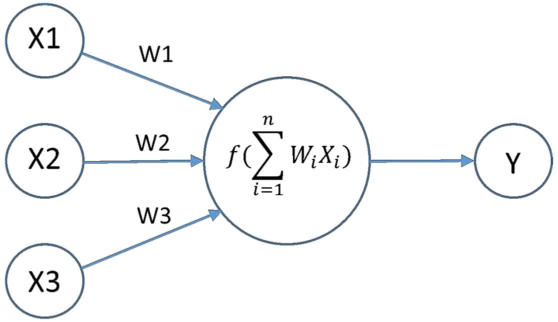
Lasso
Random forest
SVM
…

[Artificial intelligence for genomic medicine Report]

Artificial neural network

Collection of **connected units** (artificial neurons) whose functioning is inspired by **neurons** in the brain.



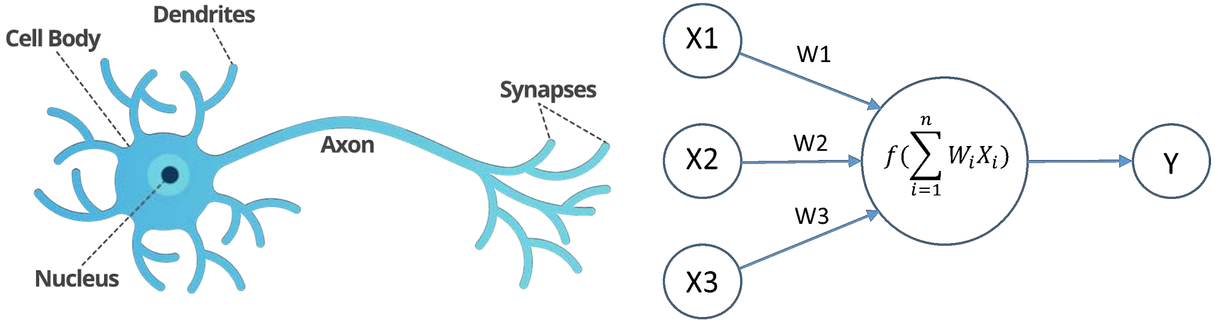


Deep learning

Deep Learning

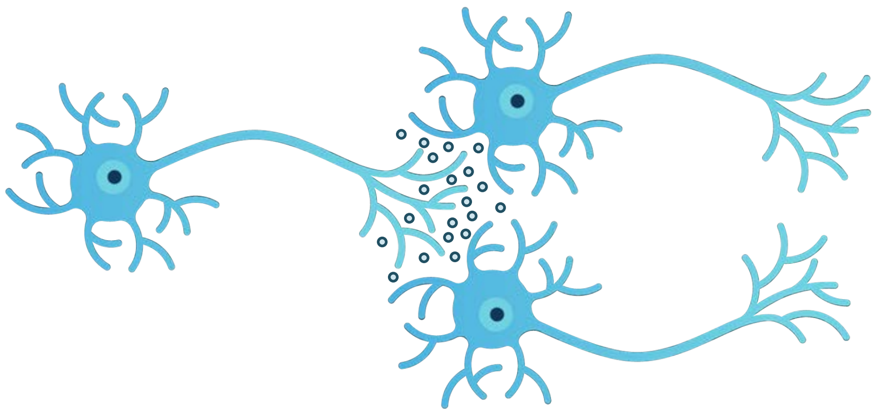
Artificial neural network

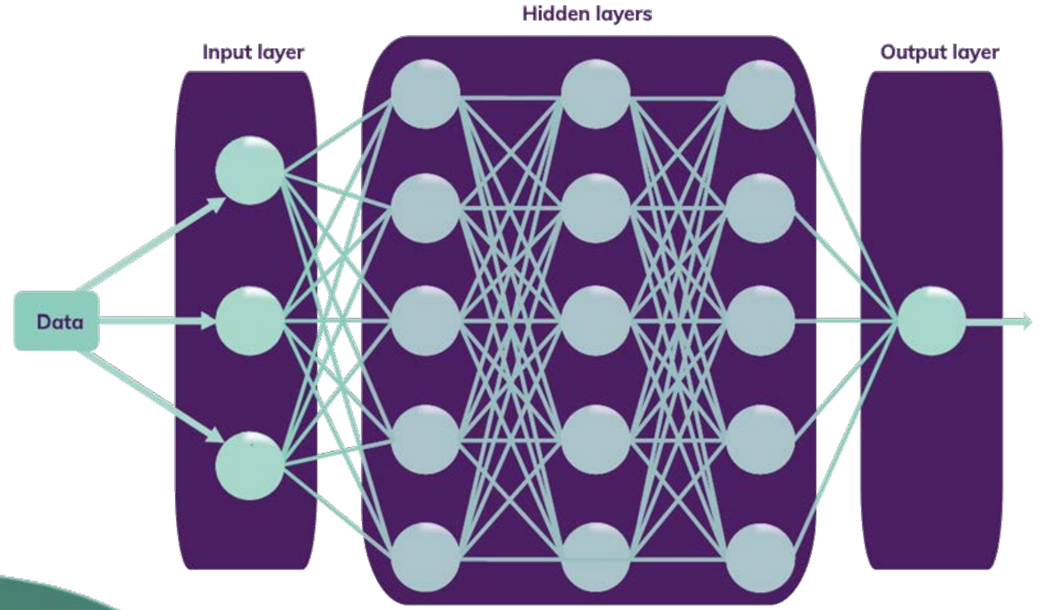
Collection of **connected units** (artificial neurons) whose functioning is inspired by **neurons** in the brain.



Deep Learning

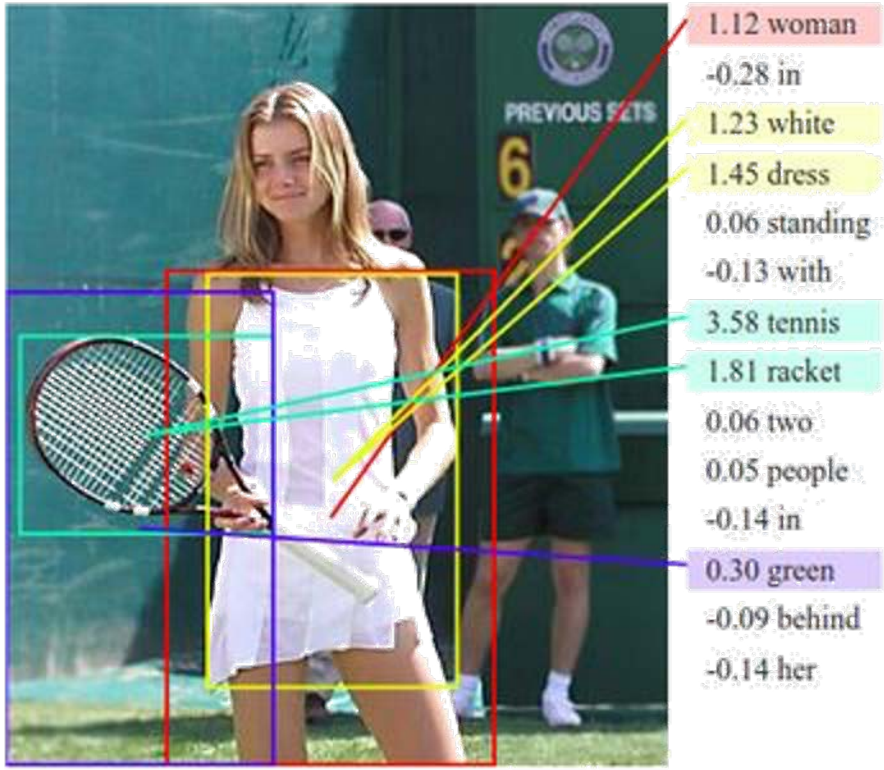
Learning process based on **large artificial neural networks** (many hidden layers)





[Artificial intelligence for genomic medicine Report]

Deep Learning

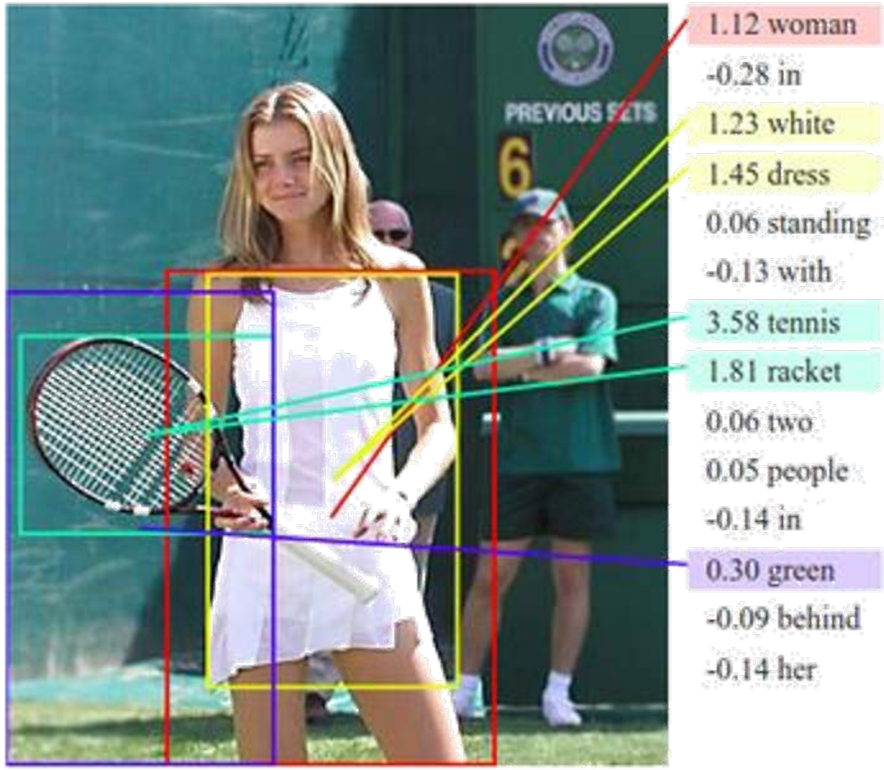


https://cs.stanford.edu/people/karpathy/deepimagesent/

Main applications

- Image recognition,
  facial recognition and object detection
- Natural language processing

Deep Learning



https://cs.stanford.edu/people/karpathy/deepimagesent/

Main applications

- Image recognition,
  facial recognition and object detection
- Natural language processing

Advantages

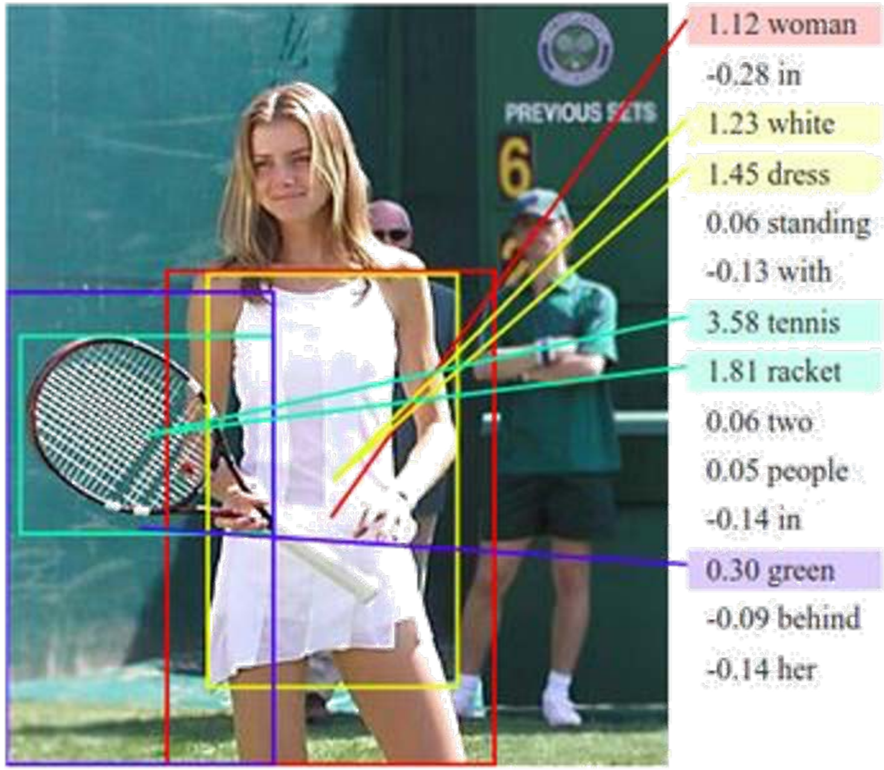
- More flexible (modeling very complex relationships)
- Less dependent on prior knowledge of the field

Limitations

- Require huge amount of data
- May be subject to overfitting (generalization to other data)
- Costly calculation (large number of operations)
- Difficult to interpret (extraction of biological knowledge)

Deep learning

Deep Learning



https://cs.stanford.edu/people/karpathy/deepimagesent/

Advantages

- More flexible (modeling very complex relationships)
- Less dependent on prior knowledge of the field

Limitations

- Require huge amount of data
- May be subject to overfitting (generalization to other data)
- Costly calculation (large number of operations)
- Difficult to interpret (extraction of biological knowledge)

Main applications

- Image recognition,
  facial recognition and object detection
- Natural language processing

a) Husky classified as wolf (b) Explanation

"The question of whether a computer can think is no more interesting than the question of whether a submarine can swim."

— Edsger W. Dijkstra

Application of Deep Learning for microbiome?

What works in other domains

- **Nature of the data**

  - Images (well known modelling)

  Challenge*: microbiome data are not deeply understood*

  - Large datasets (ImageNet: 14+ M images)
  - Transfer learning : it is possible to train a neural network on one image category to transfer it to another

  Challenge*: much less data available, large heterogeneity*

- **Nature of the *question***
  - Humans can solve the problem

  Challenge*: humans can't solve the problem*

from Chloé-Agathe Azencott

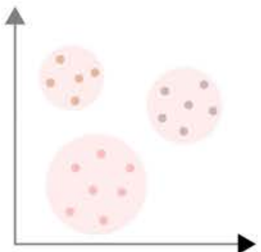
AI for microbiome data

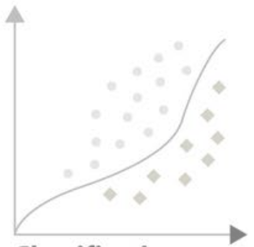
@MgpsLab

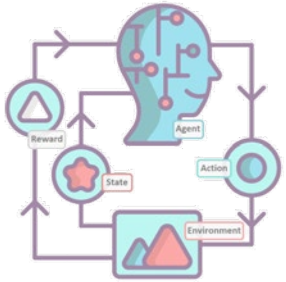
The Three Types of Machine Learning Algorithms

metagenopolis mgps.eu

Unsupervised Learning	▷ No labels ▷ No feedback ▷ Find an underlying structure in the data
Supervised Learning	▷ Labelled data ▷ Direct feedback ▷ Prediction of an output
Reinforcement Learning	▷ A set of rules / No labels ▷ Reward system ▷ Iterative self-teaching

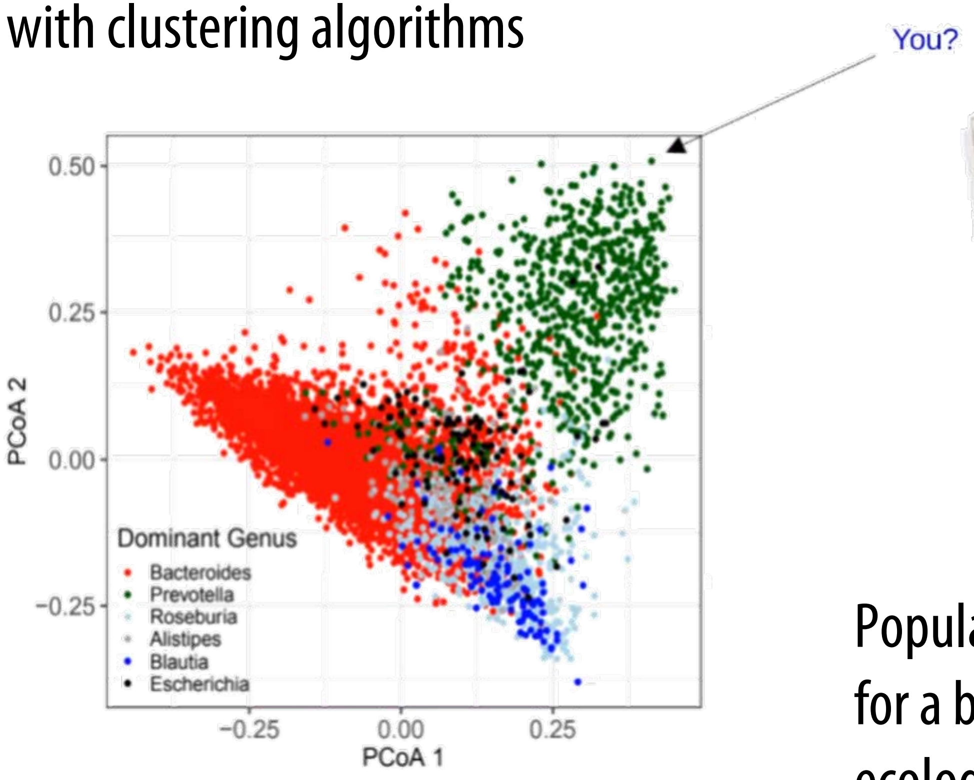


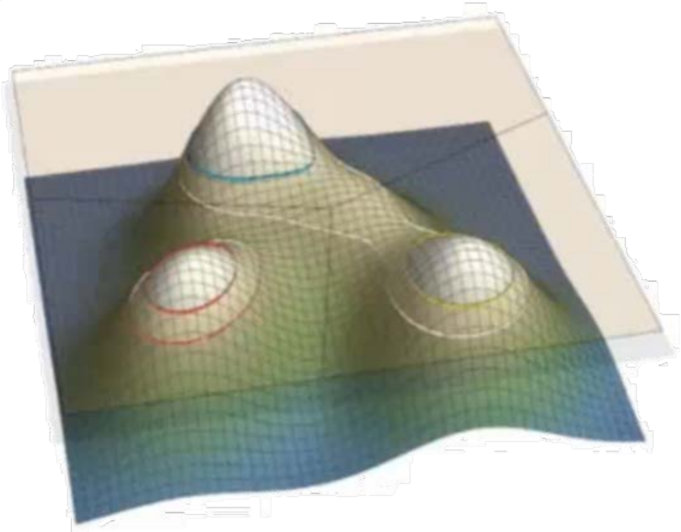




Unsupervised learning: enterotypes

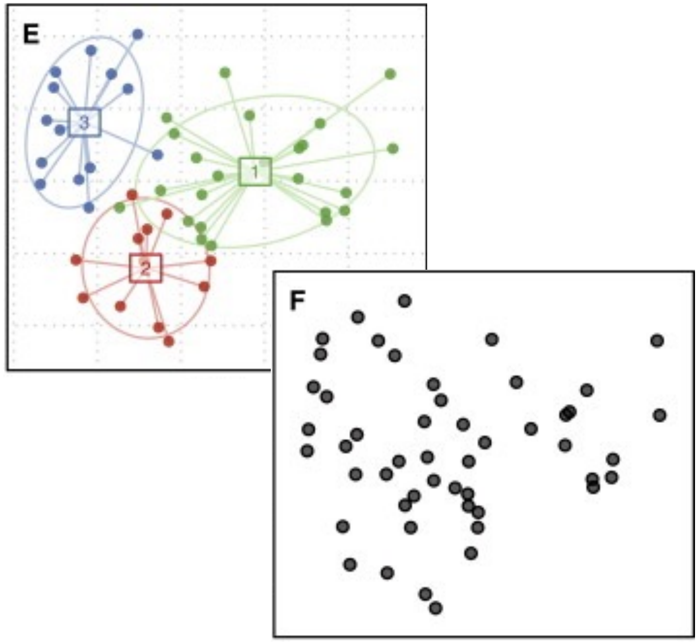
Identification of microbiome enterotypes with clustering algorithms





Population stratification is a useful approach for a better understanding of functional, ecological and medical information.

Certain visualizations can cause the eye to perceive discrete clusters to be stronger than they are



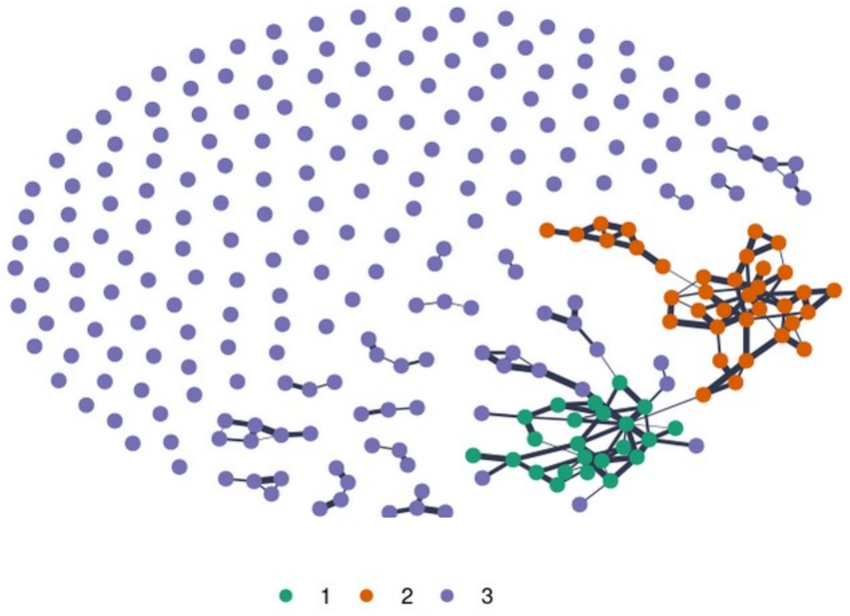
Salosensaari, Aaro, et al. Nature communications 12.1 (2021): 1-8.

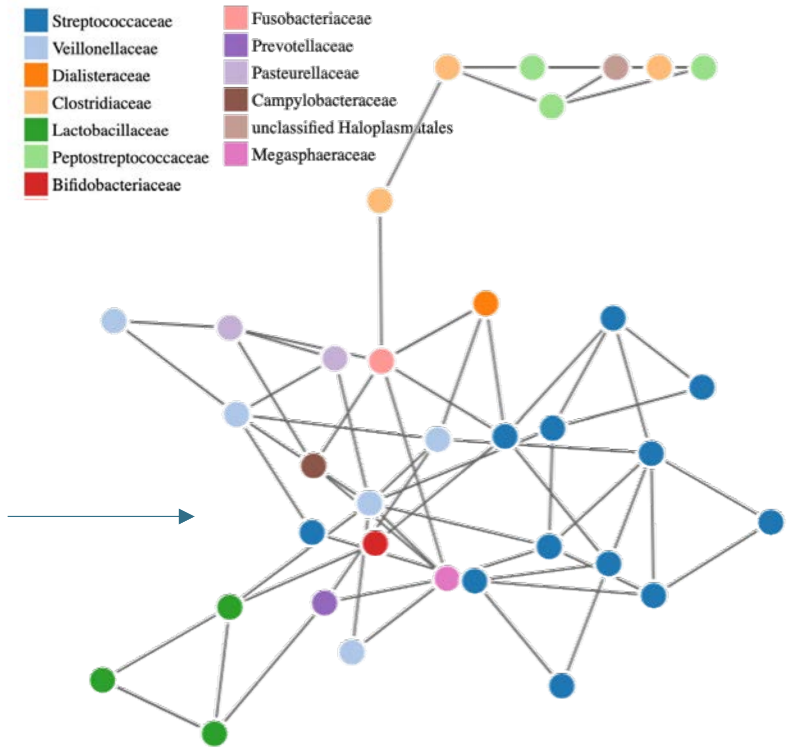
Costea, Paul I., et al. Nature microbiology 3.1 (2018): 8-16

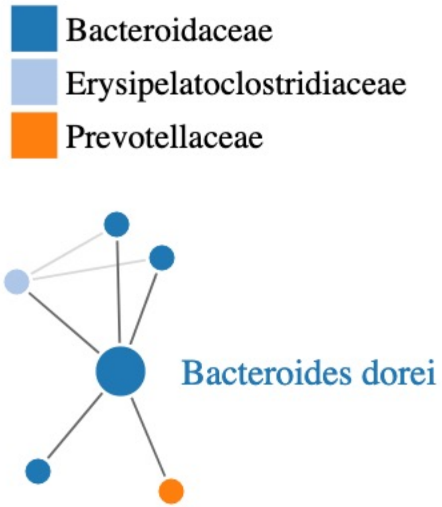
Knights, Dan, et al. Cell host & microbe 16.4 (2014): 433-437.

Unsupervised learning: microbial networks

Microbial network construction is a popular explorative data analysis technique…







… to identify taxa sharing a common role in an ecosystem

Faust, Karoline. "Open challenges for microbial network construction and analysis." The ISME Journal (2021): 1-8.

Unsupervised learning: dimensionality reduction

~10 million genes

~2 000 MGS

~20 guilds

MGS reconstruction

Network inference
from MGS co-abundances

Genes co-varying in abundance
as encoded on the same genome

Plaza Oñate, Florian, et al. Bioinformatics 35.9 (2019): 1544-1552.

The Three Types of Machine Learning Algorithms

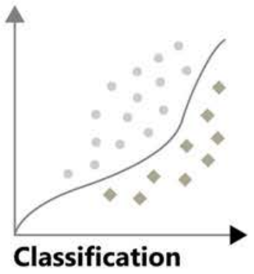
metagenopolis mgps.eu

Unsupervised Learning

- No labels
- No feedback
- Find an underlying structure in the data

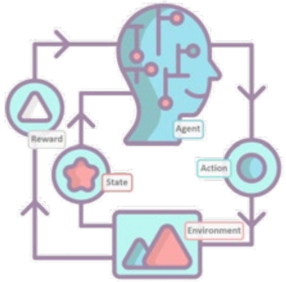
Supervised Learning

- Labelled data
- Direct feedback
- Prediction of an output



Reinforcement Learning

- A set of rules / No labels
- Reward system
- Iterative self-teaching



Supervised learning: diagnostic or prognostic

Common algorithms used for disease-prediction tasks :

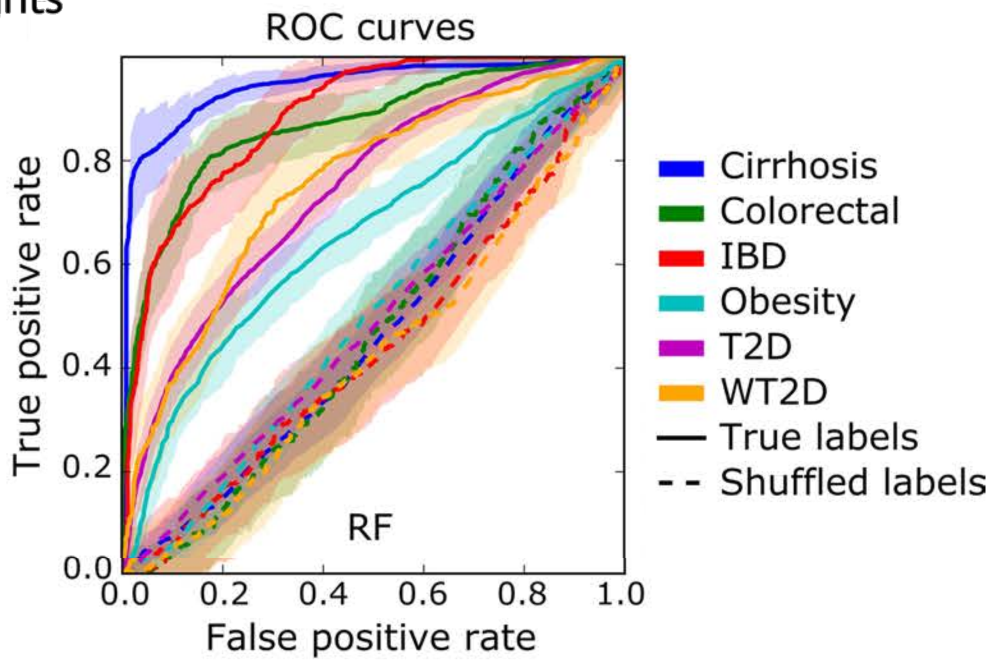
- Random forest (RF) / decision trees
- Support vector machines (SVM)
- Gradient boosting
- LASSO / ridge / elastic net regression
- Partial Least square regression (PLS)
- Neural networks
- K-nearest neighbors (KNN)
- …

Marcos-Zambrano, Laura Judith, et al. Frontiers in microbiology 12 (2021): 313

Moreno-Indias, Isabel, et al. Frontiers in Microbiology 12 (2021): 277.

Some popular Machine Learning tools

Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights

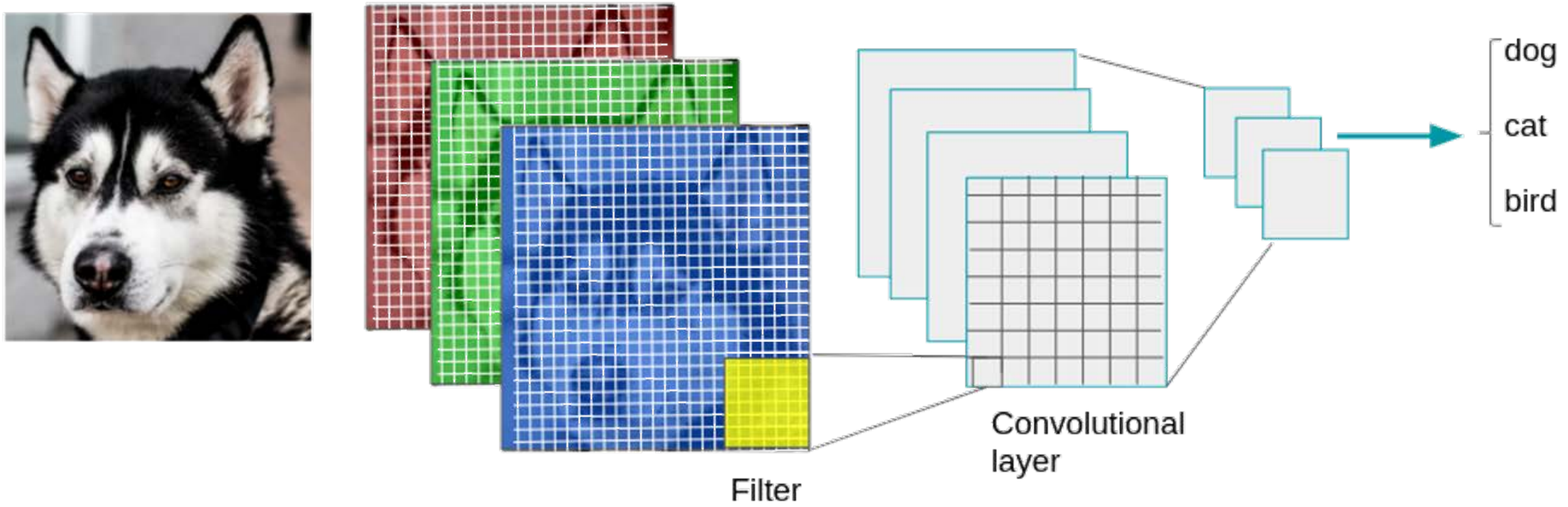


Pasolli, Edoardo, et al. PLoS computational biology 12.7 (2016): e1004977.

Deep learning for microbiome data

Architectures types

Convolutional neural networks (CNNs) | **Deep neural networks (DNNs)** | **Recurrent Neural Networks (RNNs)** | **Auto-encoders (AEs)**



LaPierre, Nathan, et al. Methods 166 (2019): 74-82.

https://datakeen.co/en/3-deep-learning-architectures-explained-in-human-language/

Deep learning for microbiome data

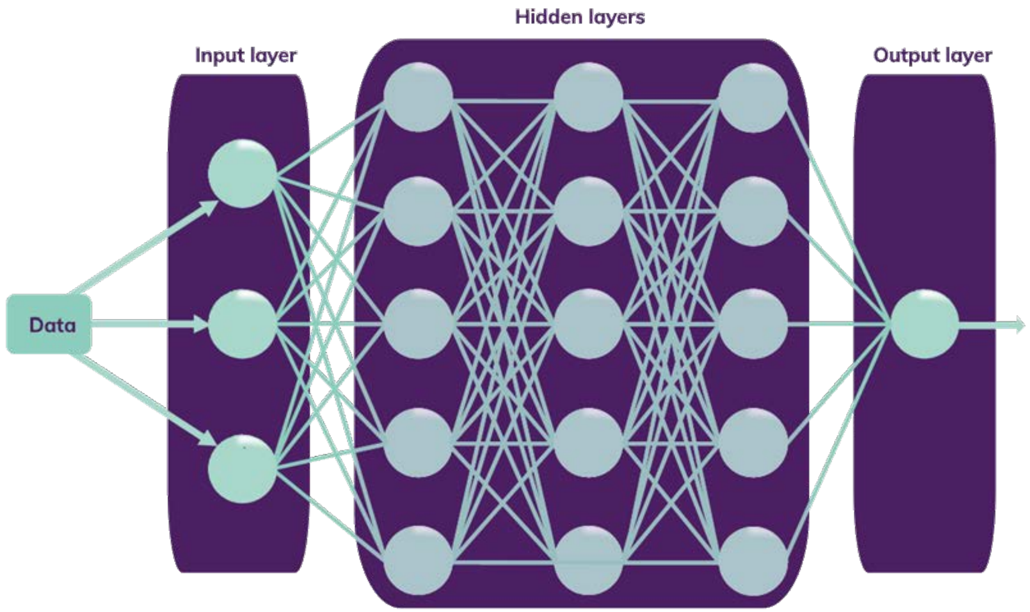
Architectures types

Convolutional neural networks (CNNs)

Deep neural networks (DNNs)

Recurrent Neural Networks (RNNs)

Auto-encoders (AEs)



[Artificial intelligence for genomic medicine Report]

Deep learning for microbiome data

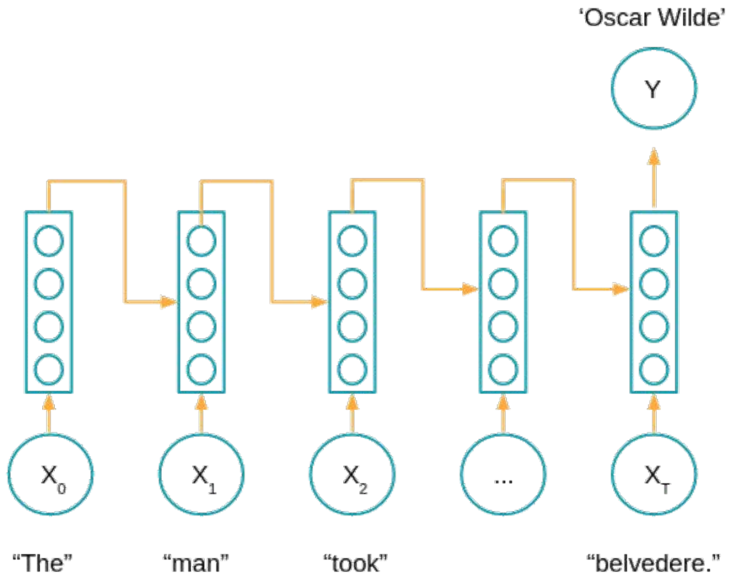
Architectures types

Convolutional neural networks (CNNs)

Deep neural networks (DNNs)

Recurrent Neural Networks (RNNs)

Auto-encoders (AEs)



https://datakeen.co/en/3-deep-learning-architectures-explained-in-human-language/

Deep learning for microbiome data

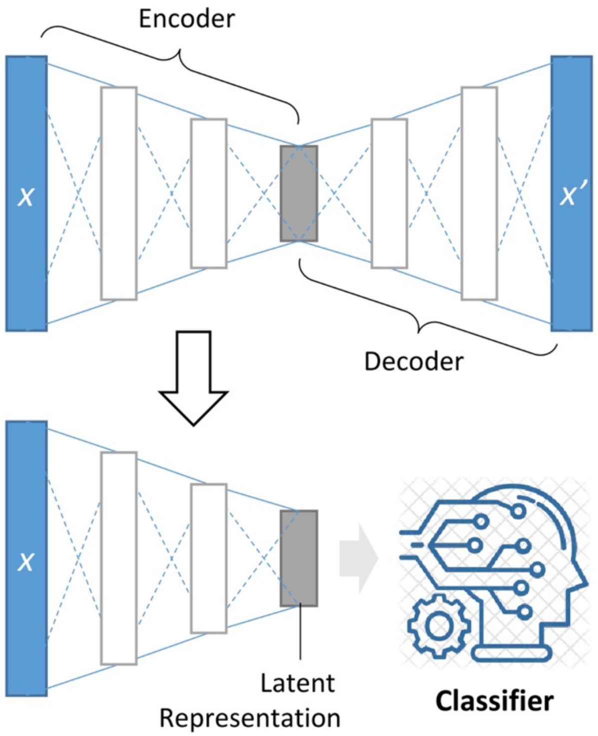
Architectures types

Convolutional neural networks (CNNs)

Deep neural networks (DNNs)

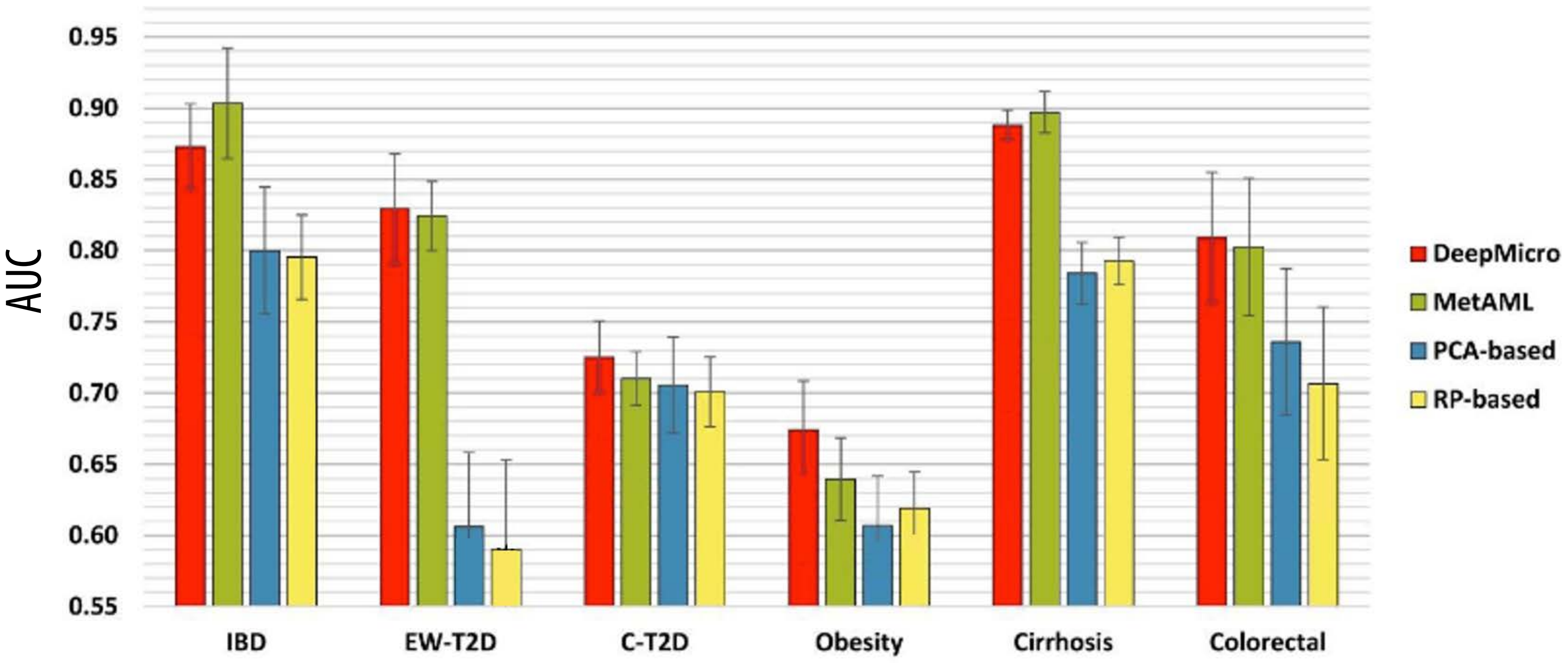
Recurrent Neural Networks (RNNs)

Auto-encoders (AEs)



Oh, Min, and Liqing Zhang. Scientific reports 10.1 (2020): 1-9.

Deep learning for microbiome data



Disease prediction performance for abundance profiles-based models

Oh, Min, and Liqing Zhang.
Scientific reports 10.1 (2020): 1-9.

Considerable effort has gone into increasingly powerful deep learning algorithms, but with only minor improvements in performance and modest changes in the ranking of the importance of features.

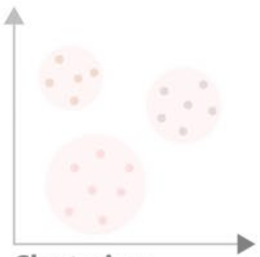
LaPierre, Nathan, et al.
Methods 166 (2019): 74-82.

The Three Types of Machine Learning Algorithms

metagenopolis mgps.eu

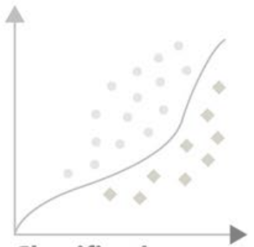
Unsupervised Learning

- No labels
- No feedback
- Find an underlying structure in the data



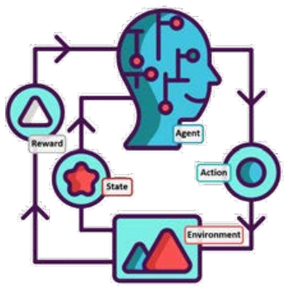
Supervised Learning

- Labelled data
- Direct feedback
- Prediction of an output

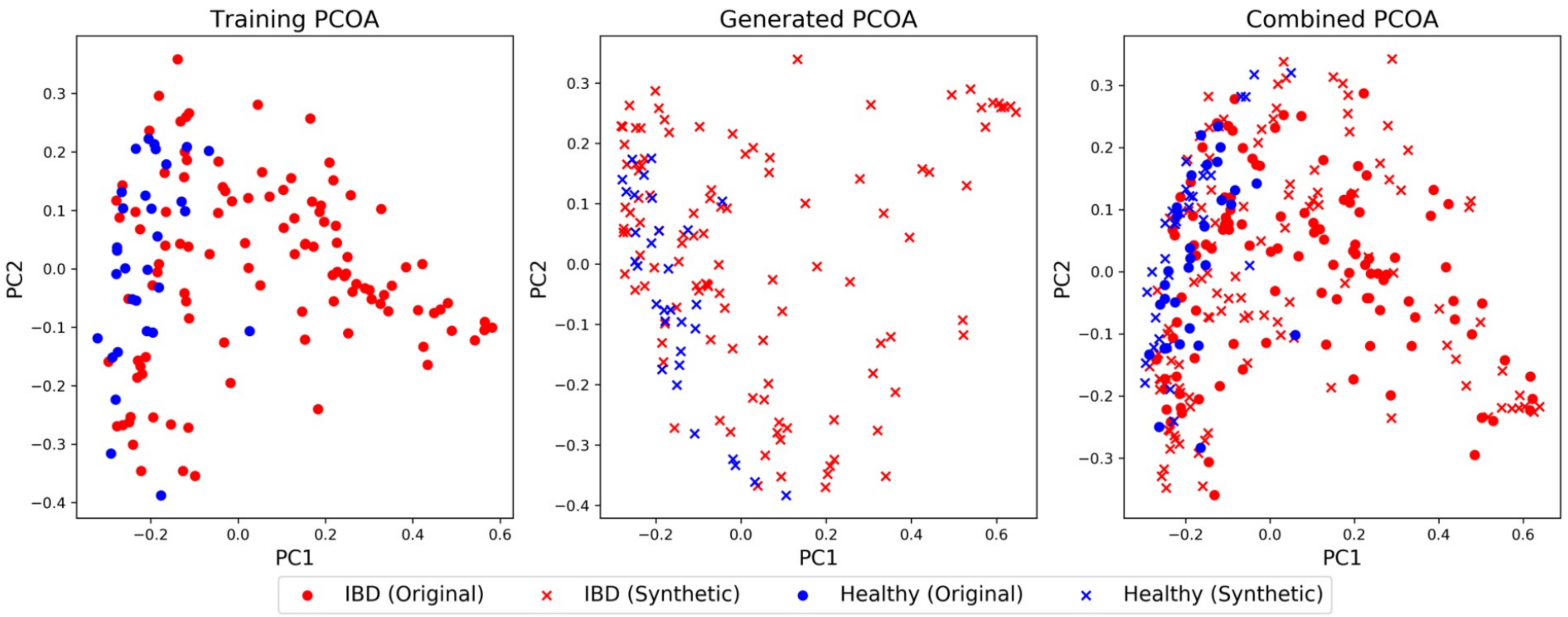


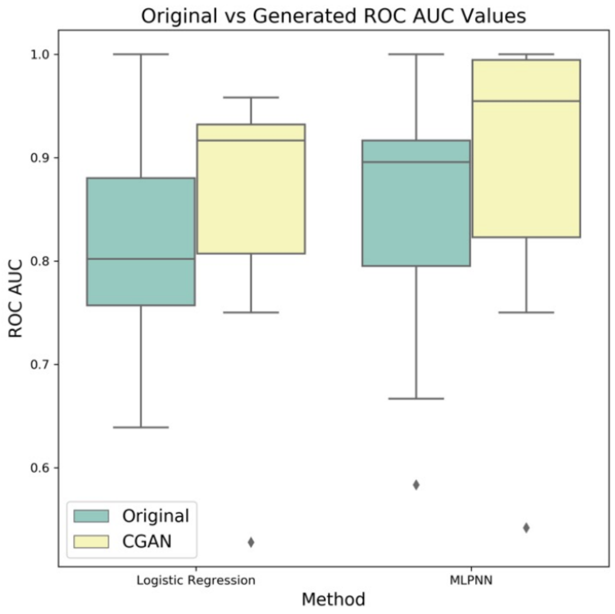
Reinforcement Learning

- A set of rules / No labels
- Reward system
- Iterative self-teaching



Data-driven simulation of microbiome data using a conditional generative adversarial network





Synthetic samples generated can boost disease prediction

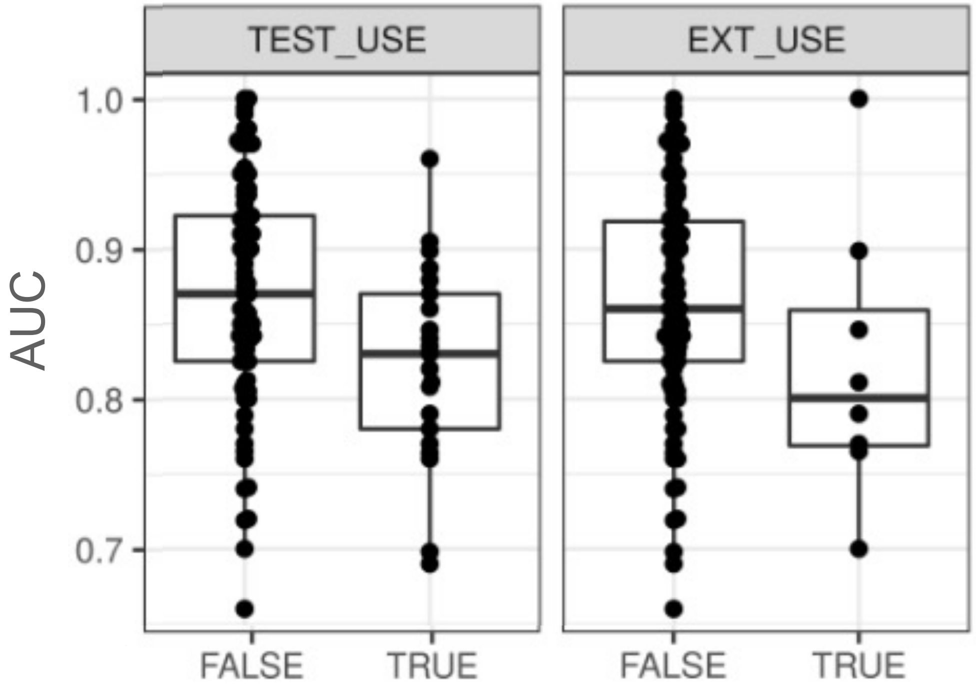
Reiman, Derek, and Yang Dai. "Using Conditional Generative Adversarial Networks to Boost the Performance of Machine Learning in Microbiome Datasets." bioRxiv (2020).

metagenopolis mgps.eu

Current AI challenges for microbiome data

Misuse of machine learning models

Failures in model verification make it impossible to know whether or not a trained model is fit for purpose



Among 102 articles **88% of the published AUCs** cannot be trusted at face value.

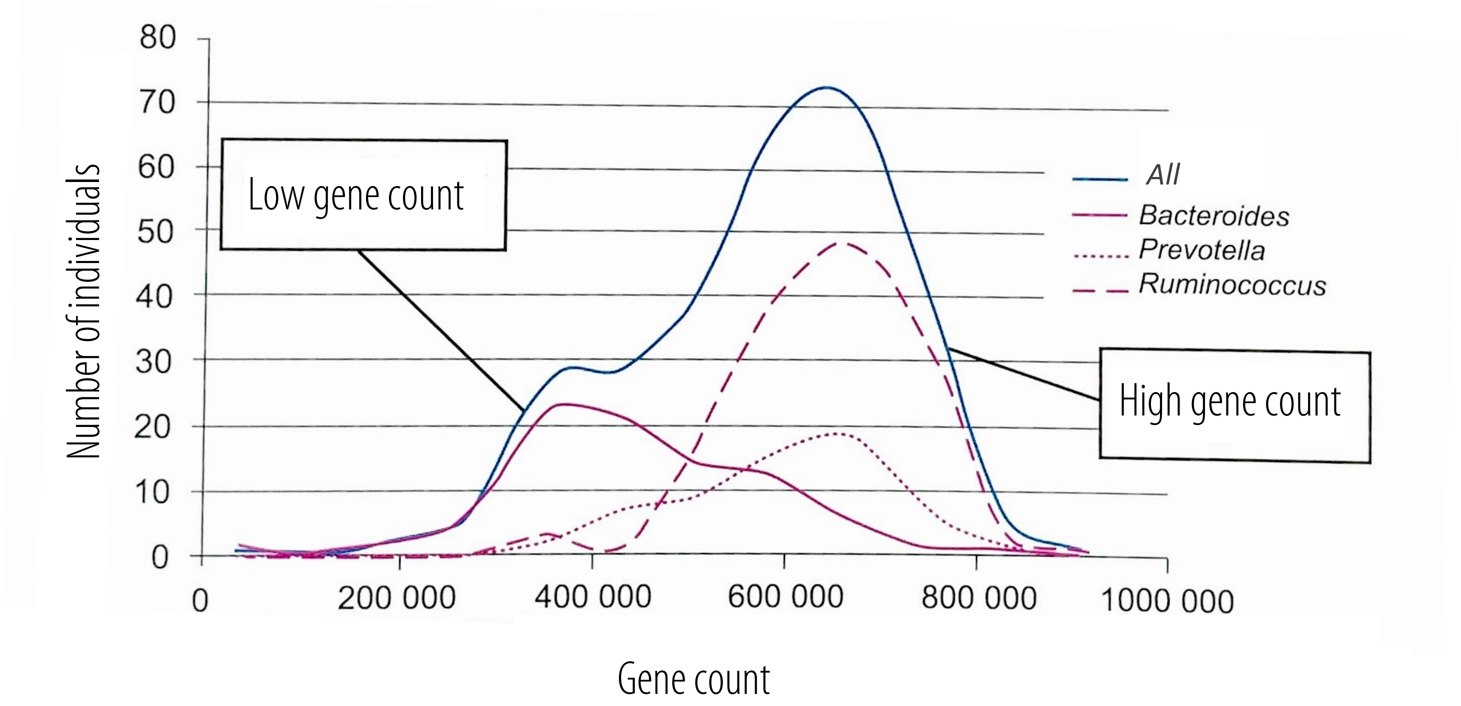
“These findings cast serious doubt on the general validity of research claiming that the gut microbiome has high diagnostic or prognostic potential in human disease.”

Quinn, Thomas P. "A Systematic Review of Human Gut Microbiome Research Suggests Widespread Misuse of Machine Learning." arXiv preprint arXiv:2107.03611 (2021).

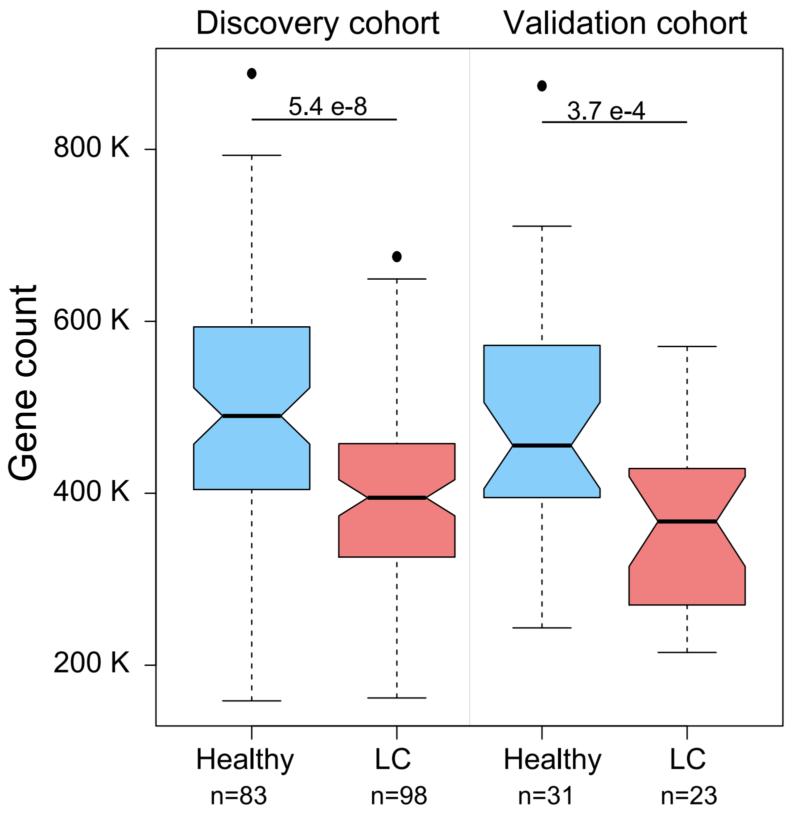
Inter-individual variability

High inter-individual variability

& limited data available



Marteau, Philippe, and Joël Doré. Ed John Libbey (2017).



Qin, Nan, et al. Nature 513.7516 (2014): 59-64.

Integration of French gut in an international project
(MMHP : Million Microbiomes from Humans Project)

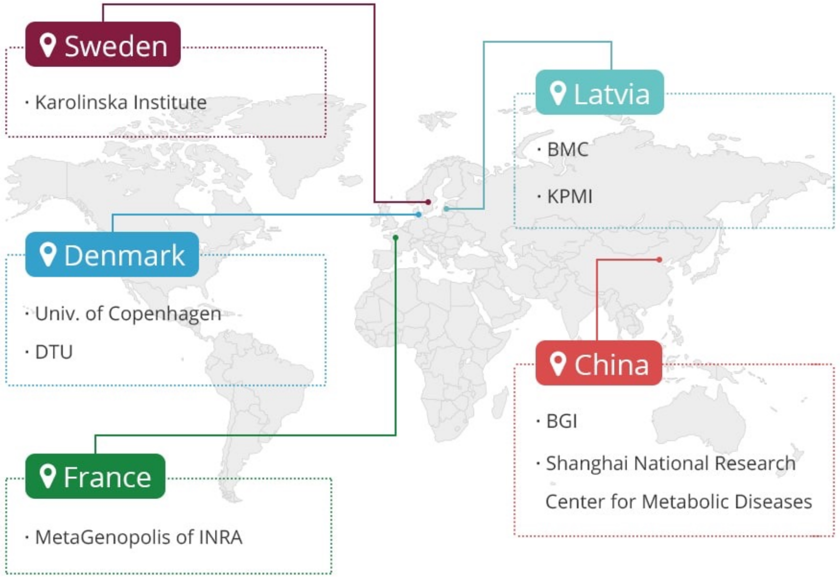
Vision and mission of MMHP

- Analyze 1 million microbial samples from intestines, mouth, skin, reproductive tract...
- Build the world's largest database of human microbiome
- Create solid data foundation for microbiome research
- Draw a microbiome map of the human body

MGP is a founding member of MMHP, officially launched on October 26th, 2019 at the 14th International Conference on Genomics (ICG-14)

MGP participates to MMHP by bringing 100,000 French gut metagenomes

Founding members of the project



https://db.cngb.org/mmhp/

With Partners (Open for collaboration) :
Germany , Italy, The Netherlands, Spain....

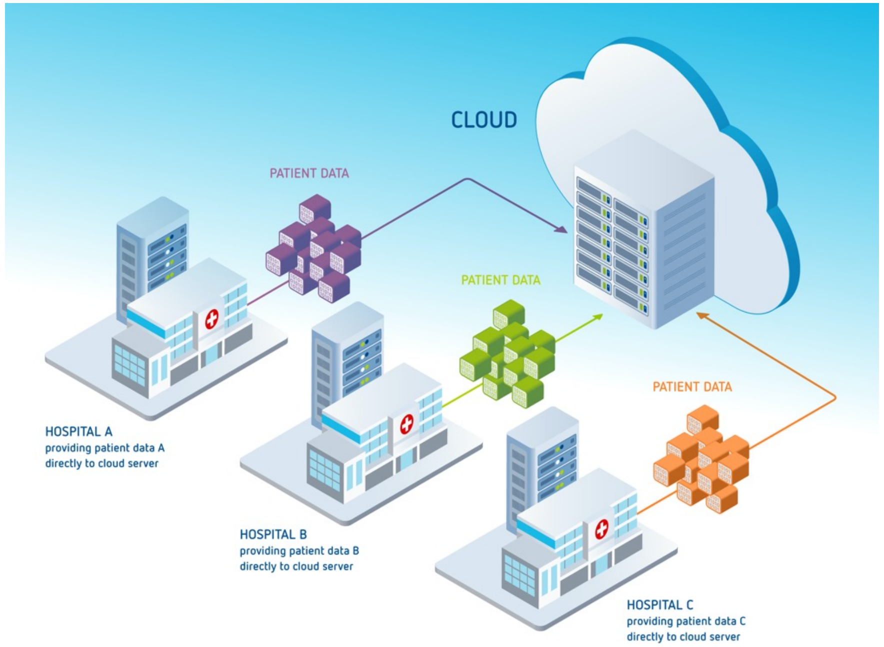
@MgpsLab

metagenopolis mgps.eu INRAE

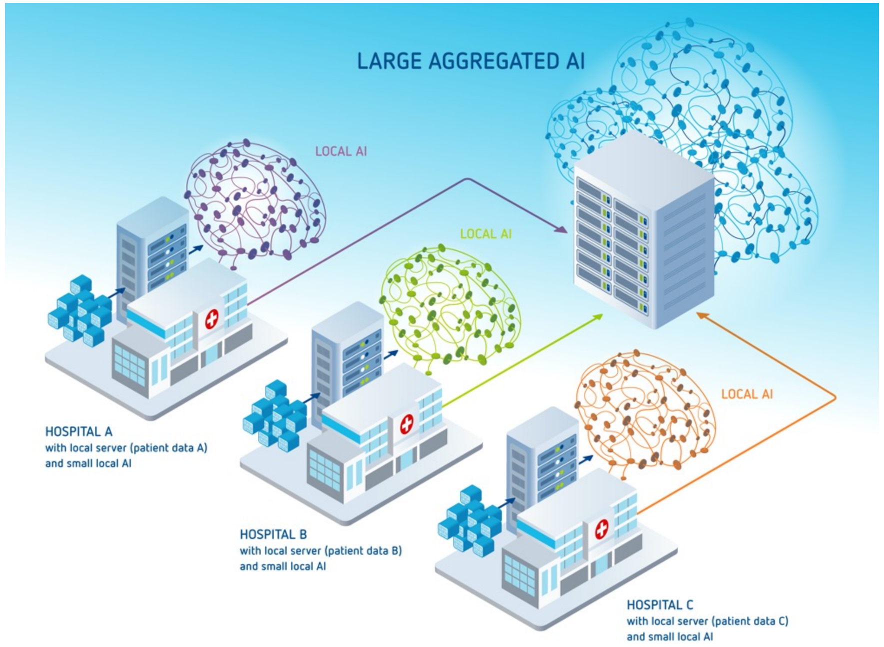
Learning from multiple datasets

Federated learning / differential privacy / domain adaptation

Current general approach to machine learning in medicine



Federated approach to machine learning in medicine

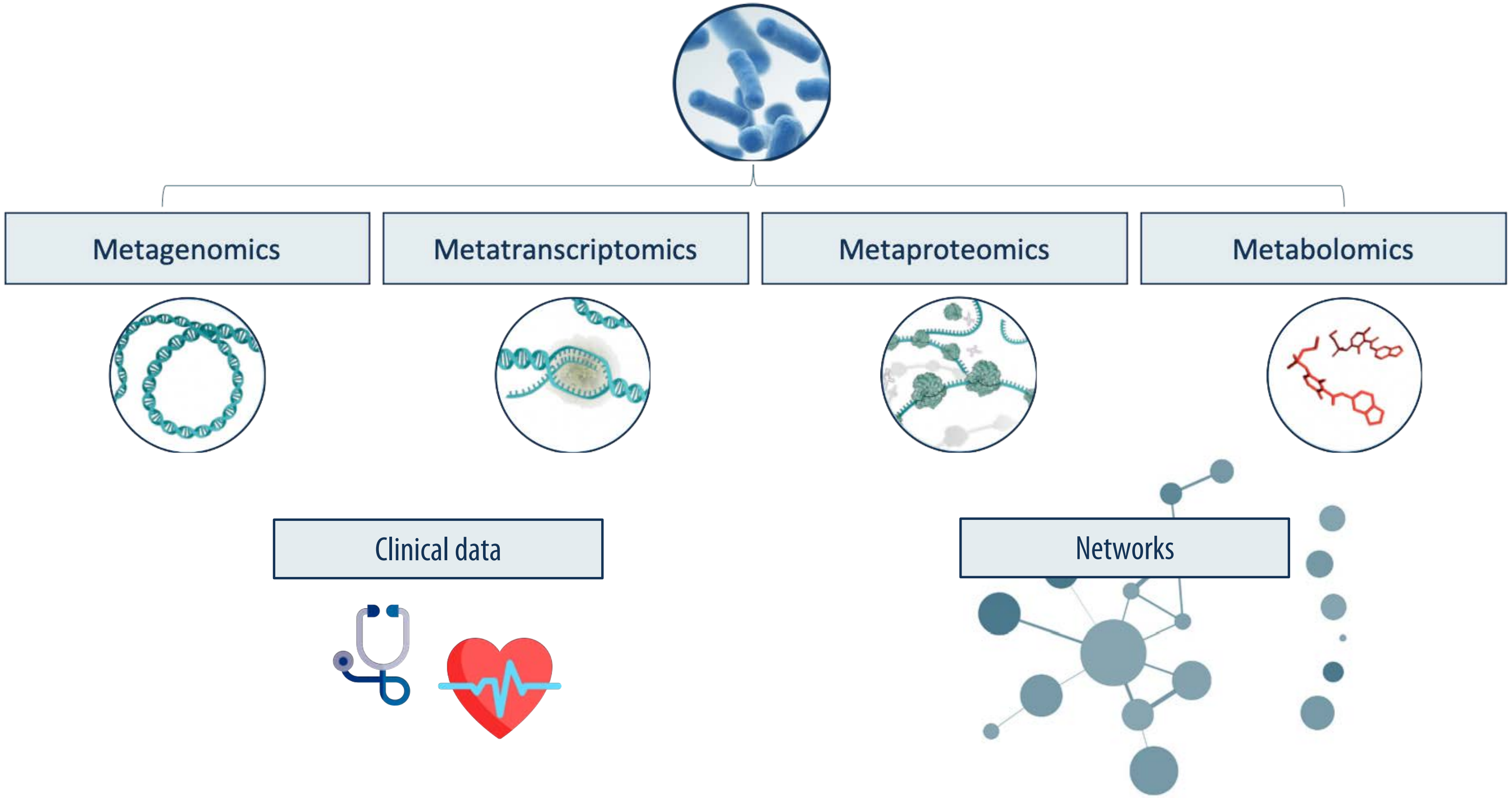


https://featurecloud.eu

Integration: learning from heterogeneous data sets

metagenopolis mgps.eu

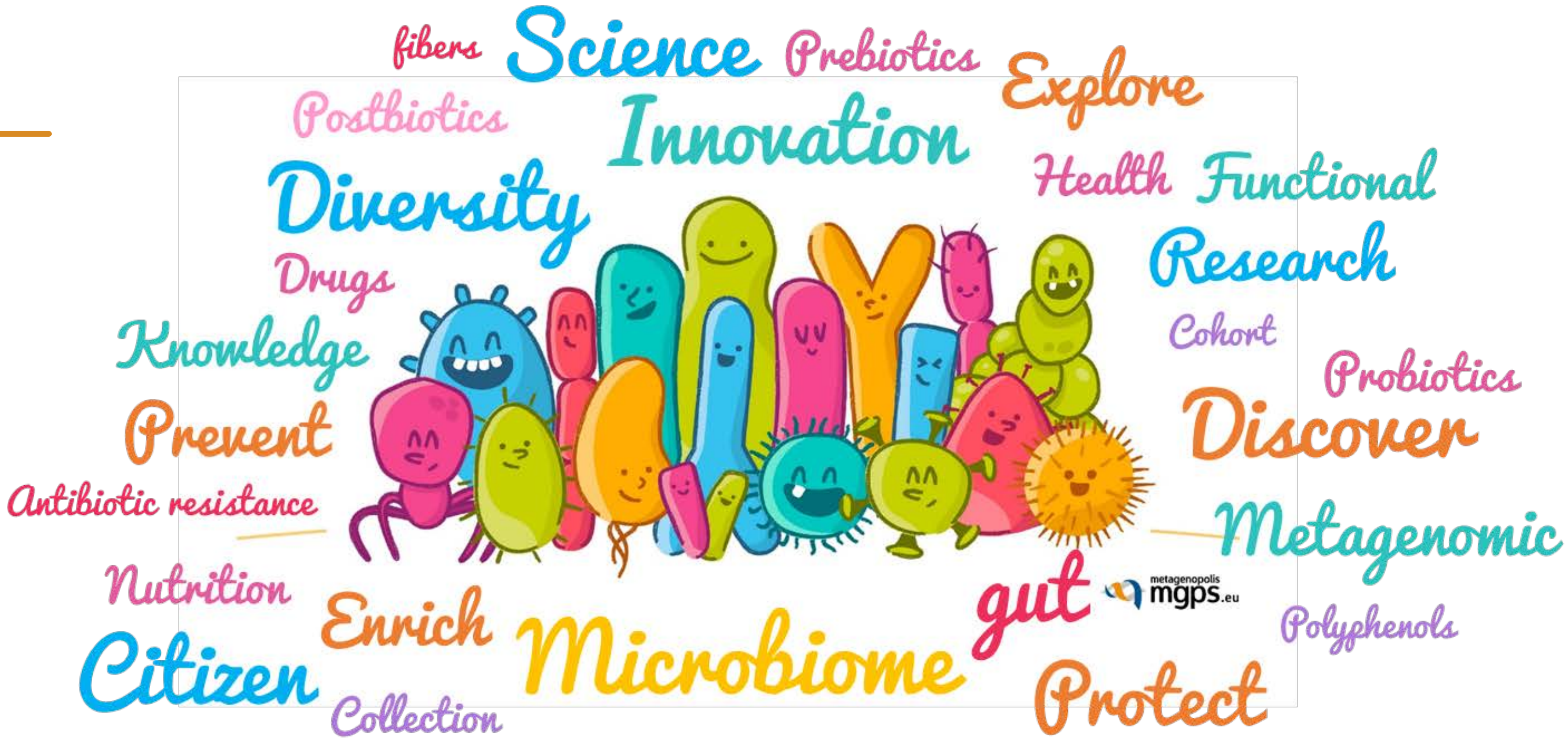
multi-view learning



Take home messages

- **Machine learning** (ML) is an important component of the growing field of artificial intelligence.
- **Deep learning** is a subtype of ML with promising result on some topics
- **Unsupervised**, **Supervised** and **Reinforcement learning** are the three types of ML algorithms successfully applied to microbiome data
- The **current challenges and active research areas** are the misuse of ML models, high inter-individual variability, federated learning and data integration

Thanks



EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

MICROBIOME