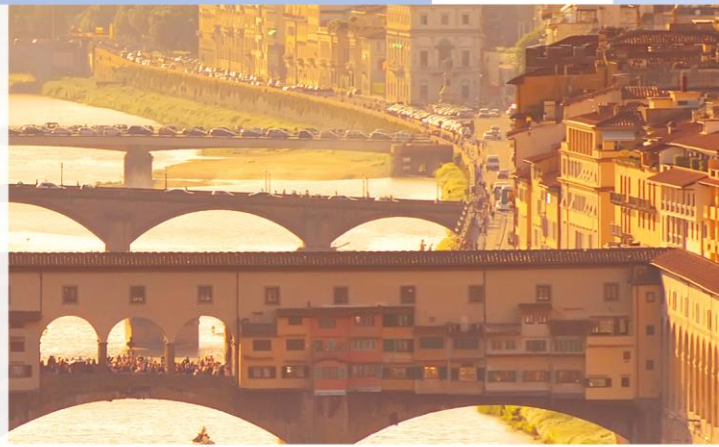
8° WORKSHOP IN EMATOLOGIA TRASLAZIONALE

DELLA SOCIETÀ ITALIANA DI EMATOLOGIA SPERIMENTALE

Firenze - Auditorium CTO - A.O.U. Careggi, 22-23 giugno 2023



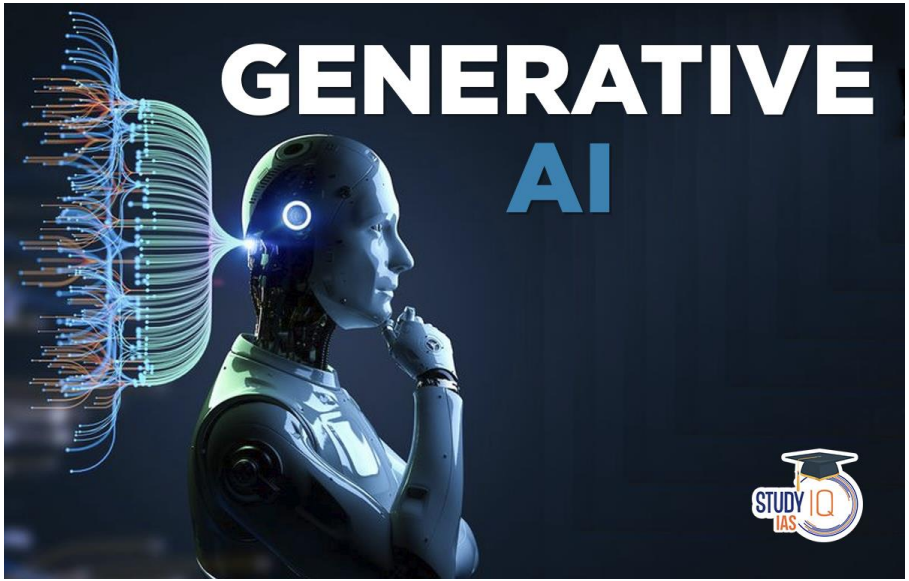
Machine learning per identificare fattori prognostici e predittivi

Gastone.Castellani@unibo.it

Dipartimento di Medicina Specialistica Diagnostica e Sperimentale
Università di Bologna

Disclosures of Name Surname

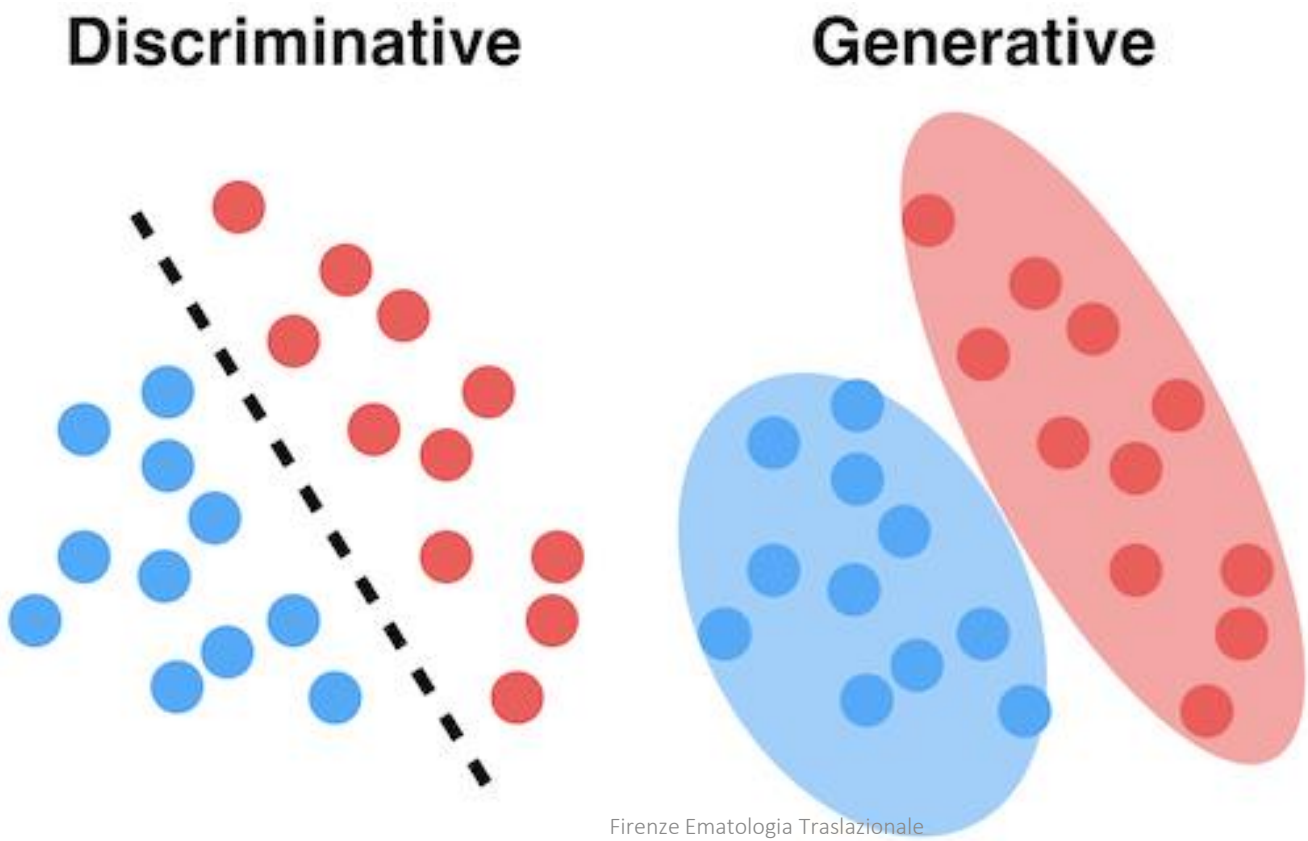
Company name	Research support	Employee	Consultant	Stockholder	Speakers bureau	Advisory board	Other



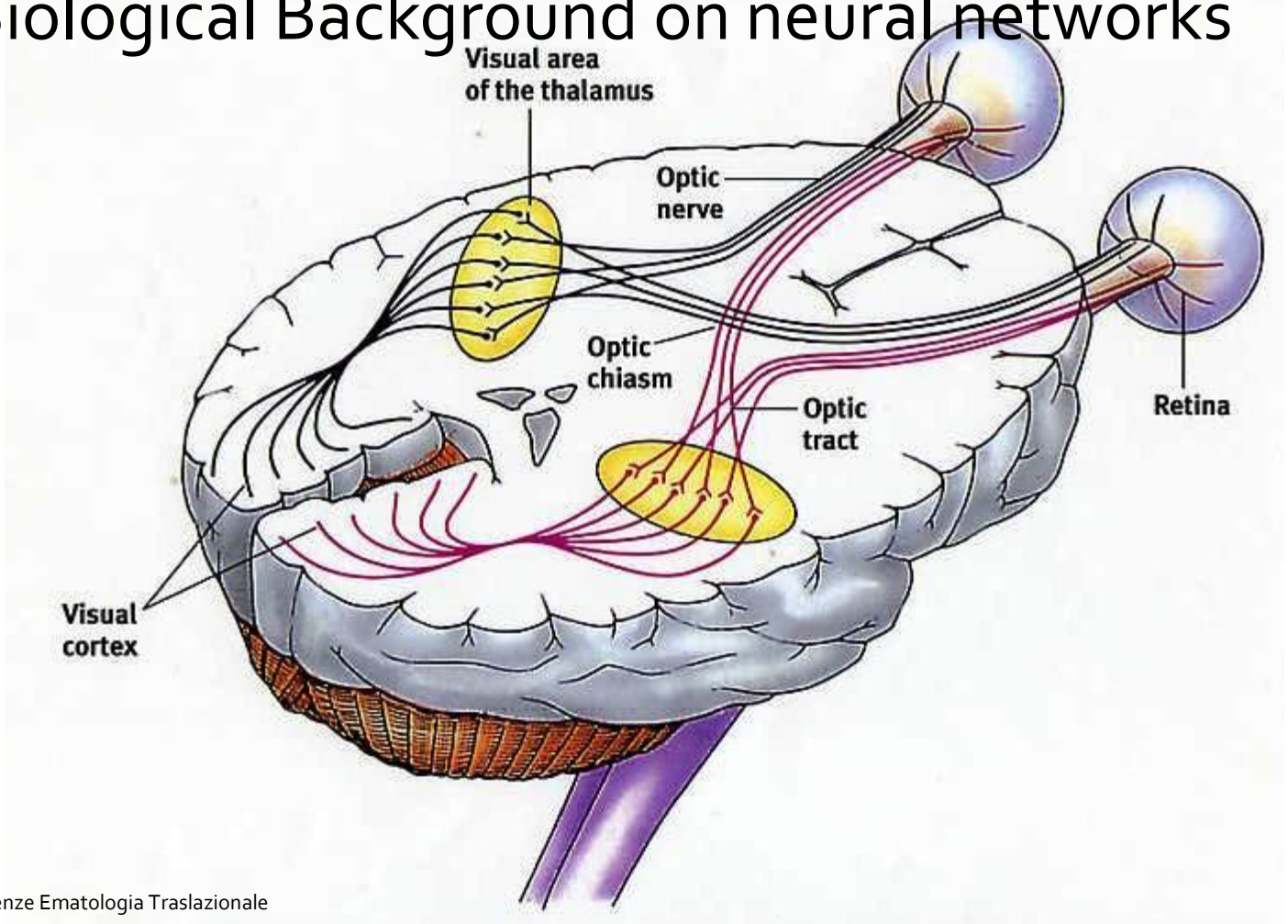
Chat Generative Pre-training Transformer

A transformer is a deep learning model. It is distinguished by its adoption of self-attention, differentially weighting the significance of each part of the input (which includes the recursive output) data. It is used primarily in the fields of natural language processing (NLP)[1] and computer vision (CV).[

Bayesian Discriminative Vs Bayesian Generative



Biological Background on neural networks



LEARNING PROCESSES

- learning with a teacher (supervised learning)
- learning without a teacher
  - unsupervised learning
  - reinforcement learning

- *supervised learning*: is based on the availability of examples or desired output (the training set) to assess a specific input–output mapping by minimizing a suitable cost function **(regression problem)**;
- *unsupervised learning*, The network to learn in a self-organized manner without the examples by finding the minima of an Energy Function **(clustering problem);**
- *reinforcement learning*: The learning of an input–output mapping is performed through continued interaction with the environment in order to minimize a scalar index of performance **(Markov Processes optimal policy assessment)**

"In God we trust.
All others must bring data."

- Dr. W. Edwards Deming

Multiomics

Radiomics

Genomics

Metabolomics

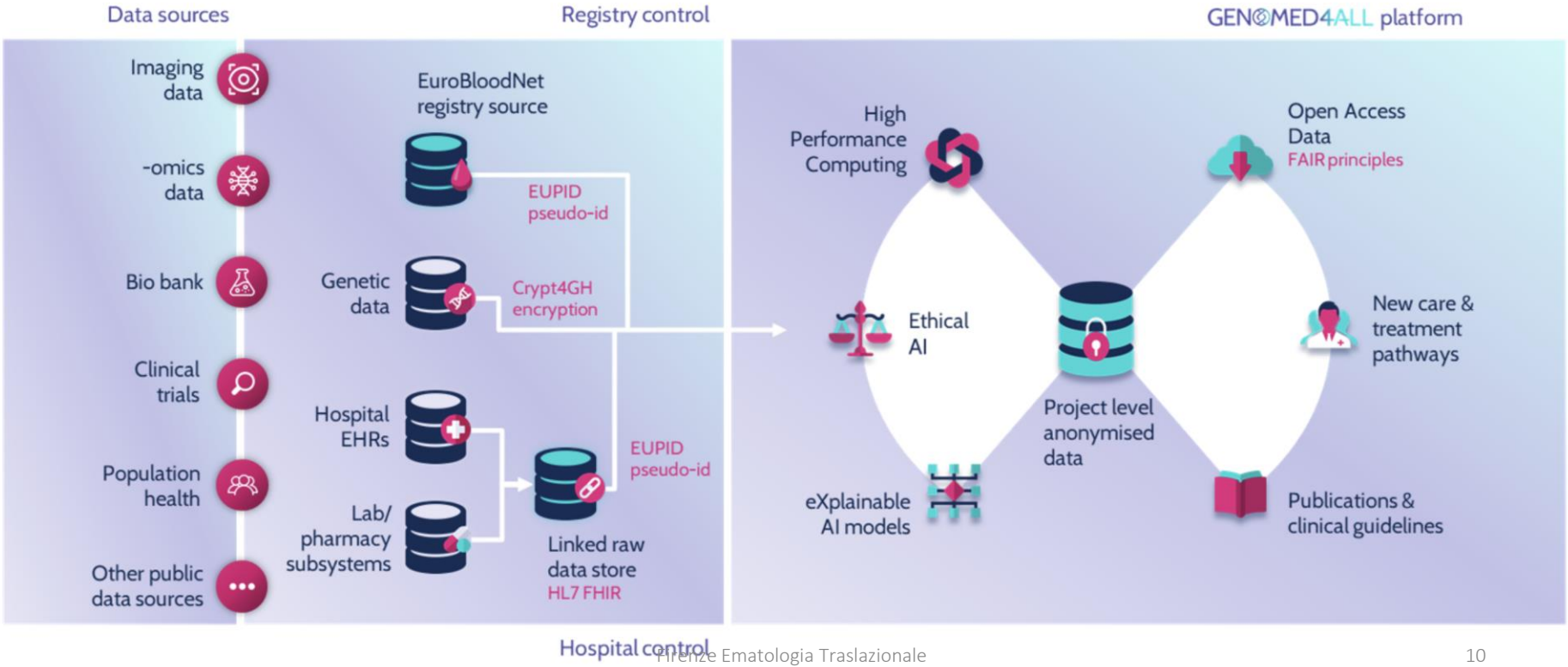
Radiogenomics

Pathomics

IMI2 Project 40M€ project 56 EU partners, including the pharma companies; 20M€ «real» and 20 «in kind»

GENOMED4ALL

Genomics For Next Generation Healthcare




POLITÉCNICA


ThermoFisher
SCIENTIFIC


Dedalus


Datawizard


AUSTRALO


CINECA


The European Institute For
Innovation Through Health Data


HUMANITAS
RESEARCH HOSPITAL


Vall
d'Hebron


ASSISTANCE
PUBLIQUE
HÔPITAUX
DE PARIS


ESIEE
PARIS


UMC Utrecht


IBSAL
Instituto de Investigación
Biomédica de Salamanca


THE CYPRUS INSTITUTE OF
NEUROLOGY & GENETICS


UNIVERSITÀ
DEGLI STUDI
DI PADOVA


UNIVERSITÄT
LEIPZIG


FORTH


HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA


KØBENHAVNS
UNIVERSITET


CRG
Centre
for Genomic
Regulation


UNIVERSITÀ
DEGLI STUDI
DI TORINO

Synthetic hematological data over federated computing frameworks

UPM
Spain

Datawizard
Italy

SBA
Austria

AUSTRALO
Spain

Università di Bologna
Italy

University of Southampton
United Kingdom

Vicomtech
Spain

i-HD
Belgium

Vall d'Hebron Institut de Recerca
Spain

Humanitas
Italy

UMC Utrecht
Netherlands

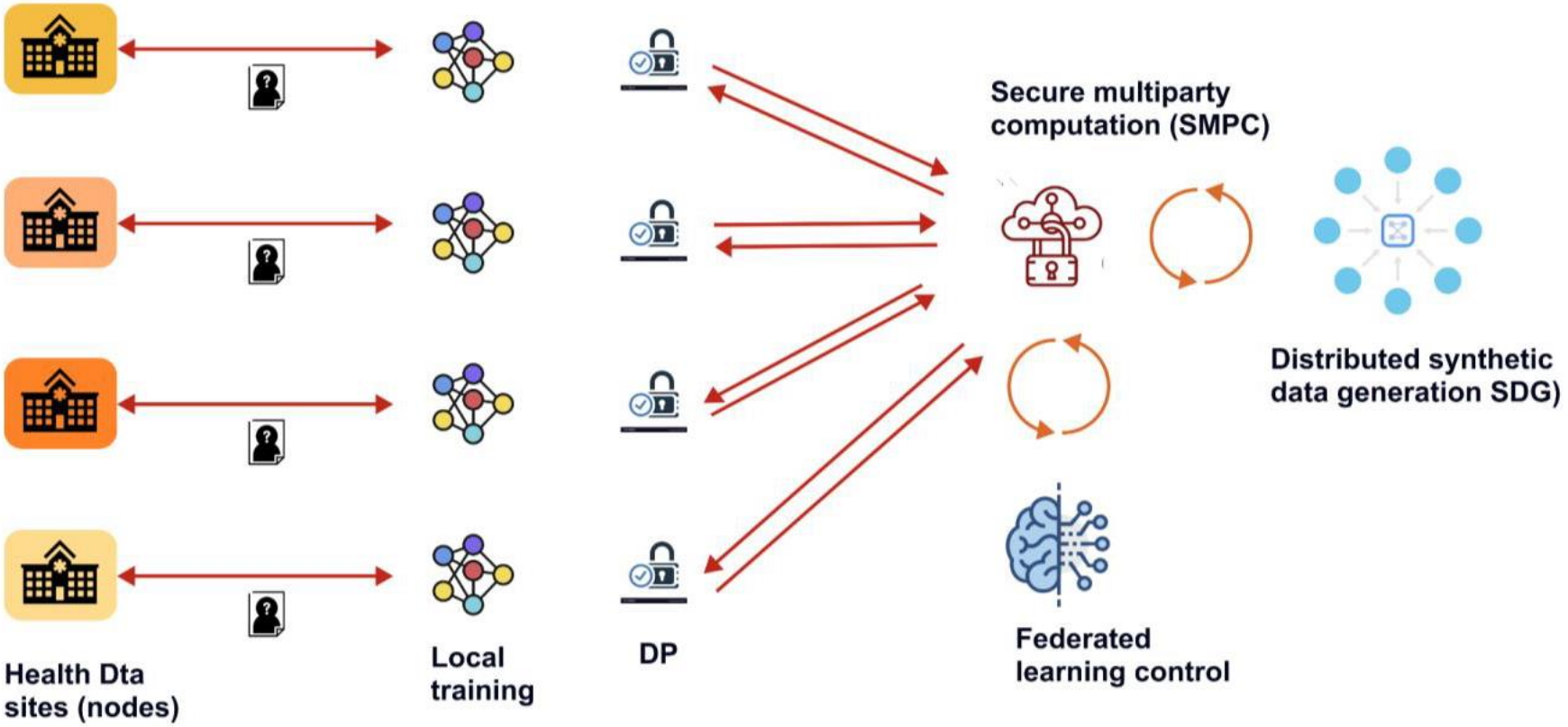
Assistance Publique - Hôpitaux de Paris
France

Charité Universitäts-medizin Berlin
Germany

GLSMED
Portugal

Università degli Studi di Padova
Italy

Netcompany-Intrasoft
Luxembourg

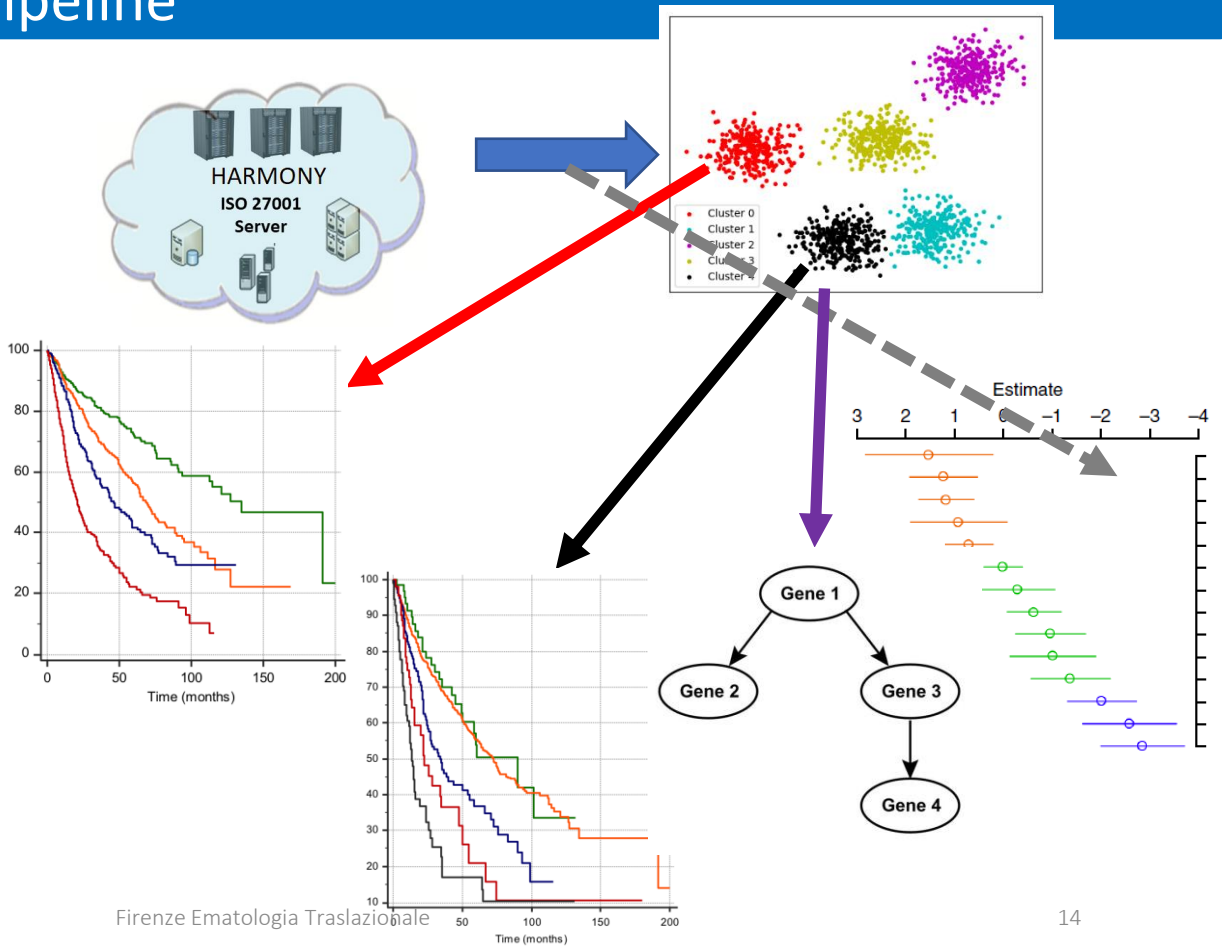

Secure multiparty computation (SMPC)
Distributed synthetic data generation SDG)
Health Dta sites (nodes)
Local training
DP
Federated learning control

A very «generic» pipeline

Cytogenetic, genetic and clinical data (VAF, TD, CNV etc)

Clustering with HDP & Clonality/Subclonality with B&T and mutation timing

Driver Mutations (BN Causality)
Patient Stratification (HDP)
Survival analysis



Some Data Analytics

Common Data Model OMOP Data Interface

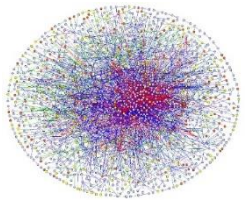
- All the data set are in OMOP format and they are accessible by script in Python and R (all the software is OPEN)
- All the libraries for the algorithms have been installed as well external data (PPI network) for subsequent analysis

Methods & Algorithms

- Clustering procedures, classical and advanced (Hierarchical Dirichlet Process)
- Mutation Co-Occurrence matrix
- Bradley Terry Method for detection of clonal/subclonal mutations and timing
- Bayesian network for mutation causality assessment
- Survival analysis (Penalized Cox Regression Model)

New algorithms, Complex Networks, ML, SL, AI

- Drug repurposing algorithm
- Network diffusion algorithm
- Graphs Neural Network



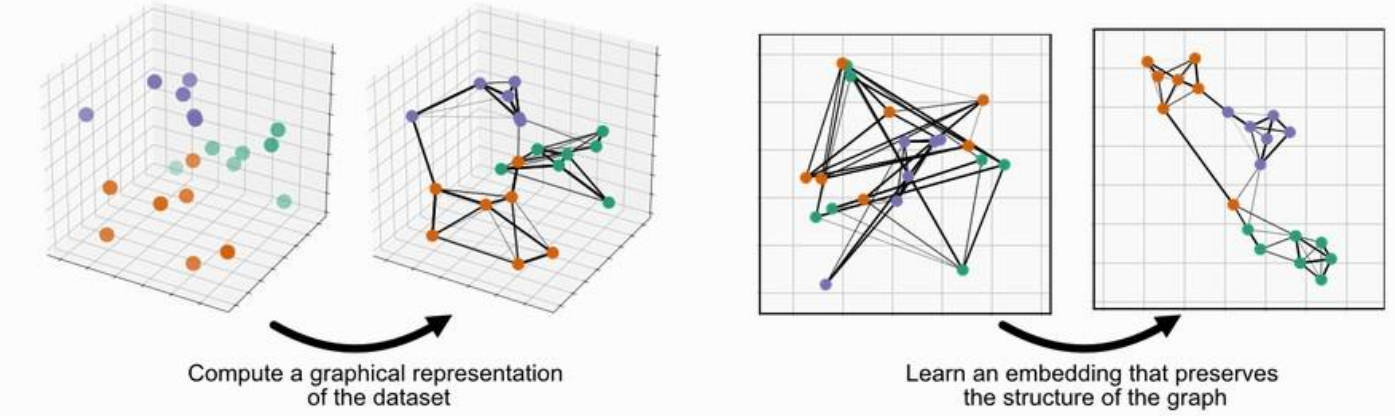
do Valle ÍF,.., Castellani G, Remondini D. Network integration of multi-tumour omics data suggests novel targeting strategies. Nat Commun. 2018

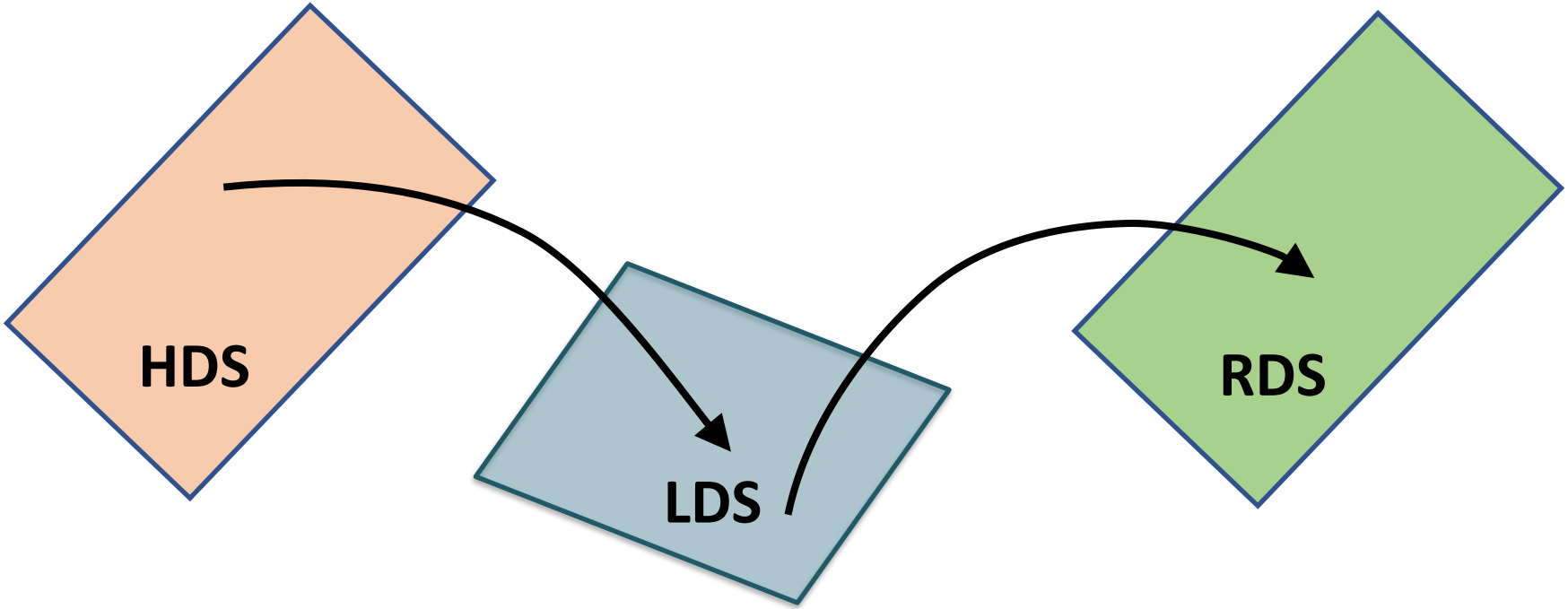
Dimensionality reduction step: UMAP

Faster than tSNE, better at wrapping global information than PCA.

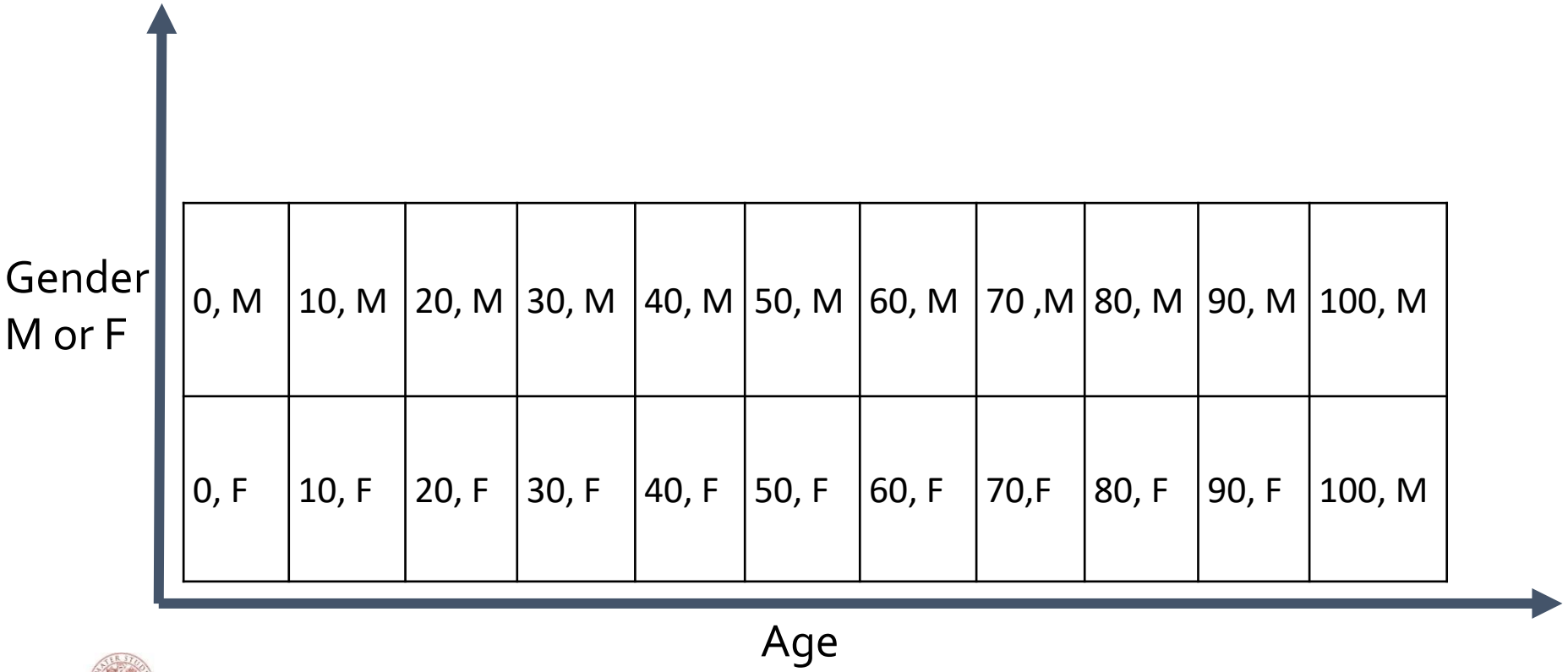
The concept is trying to learn the shape of the manifold where data lie on, and reproduce it as similarly as possible in a lower dimensional space: the **embedding**.

UMAP does it by moving data from the original space to a graph, optimizing this graph respresentation, and then moving data from the graph to the final embedding space.




HDS
LDS
RDS

Latent Space sampling and factorization for HMs



3. SD definition from EU community

EUROPEAN DATA PROTECTION SUPERVISOR

Home | About | Data Protection | Press & Publications

https://edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en

Synthetic data is artificial data that is **generated from original data** and a model that is trained to reproduce the characteristics and structure of the original data.
This means that **synthetic data and original data should deliver very similar results when undergoing the same statistical analysis**.
The degree to which synthetic data is an accurate proxy for the original data is a measure of the *utility* of the method and the model.

Analysis of copy number variation in multiple myeloma patients

Overview of the study

Data

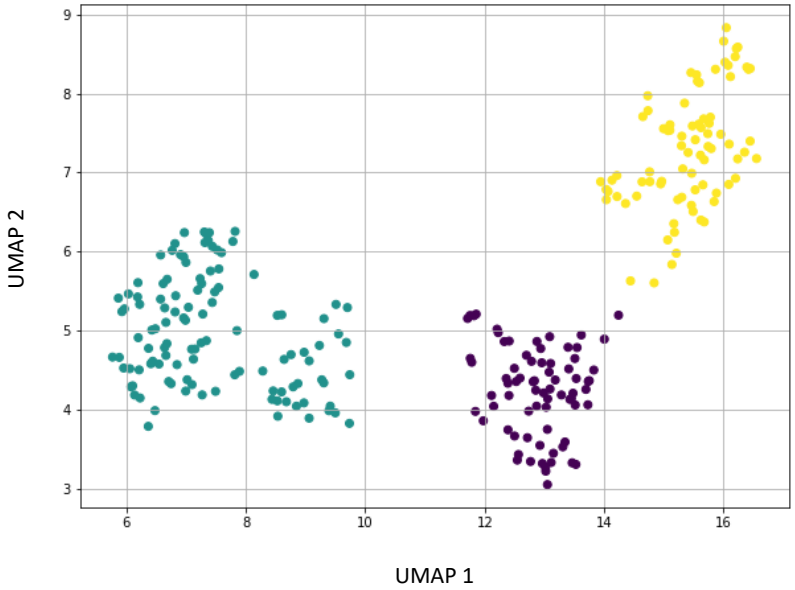
- **133 patients** - CNV data provided by the Institute of Hematology "L. and A. Seràgnoli“, Department of Experimental, Diagnostic and Specialty Medicine (DIMES), University of Bologna
- **883 patients** - CNV data downloaded by COMMPASS database

Analysis and results

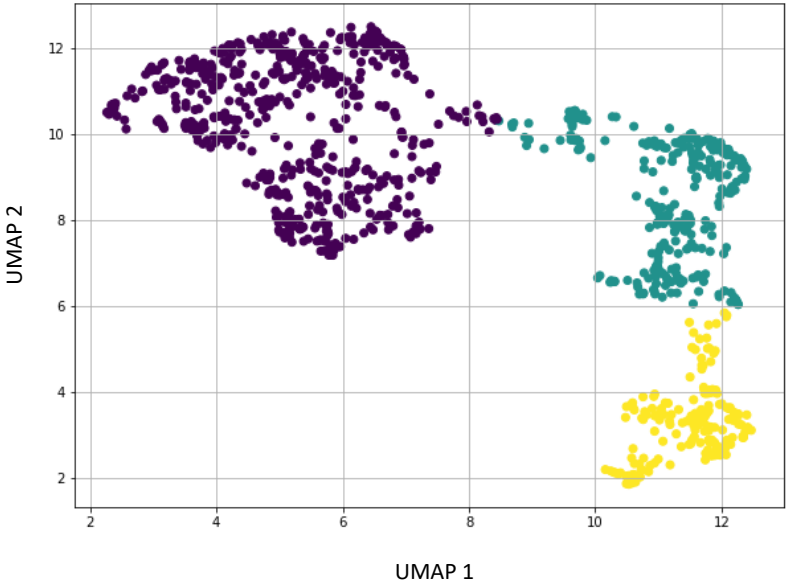
- By combining dimensionality reduction techniques and clustering methods, we obtained a stratification of the patients, at the time of **diagnosis**, into **3 groups**. Each one of these groups is characterized by a specific set of genomic aberrations.

Patient clustering – three different groups

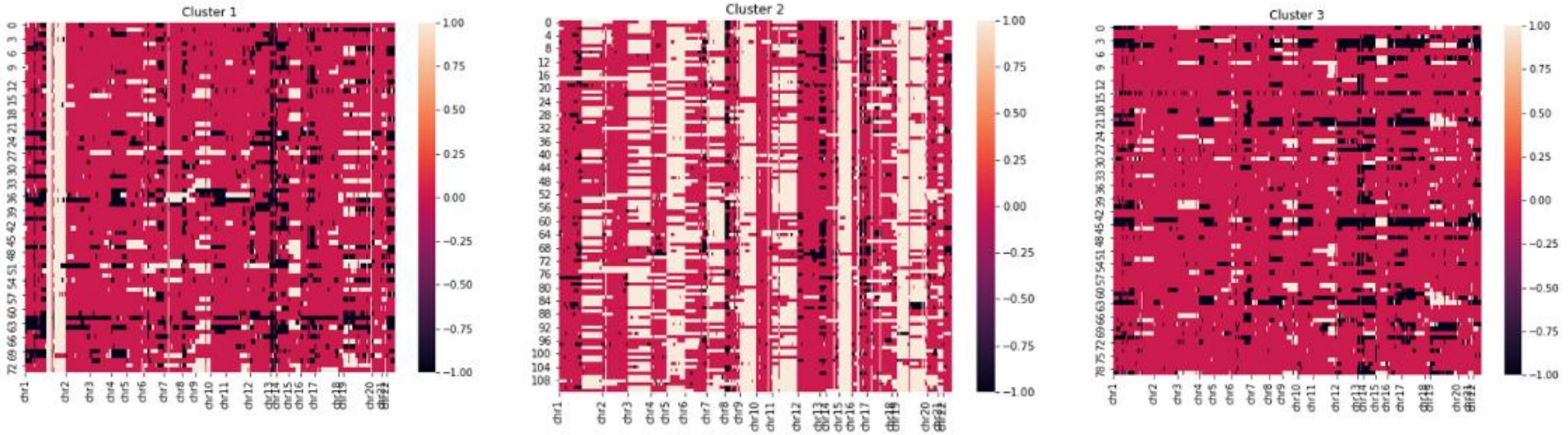
Internal dataset



Internal + CoMMpass dataset

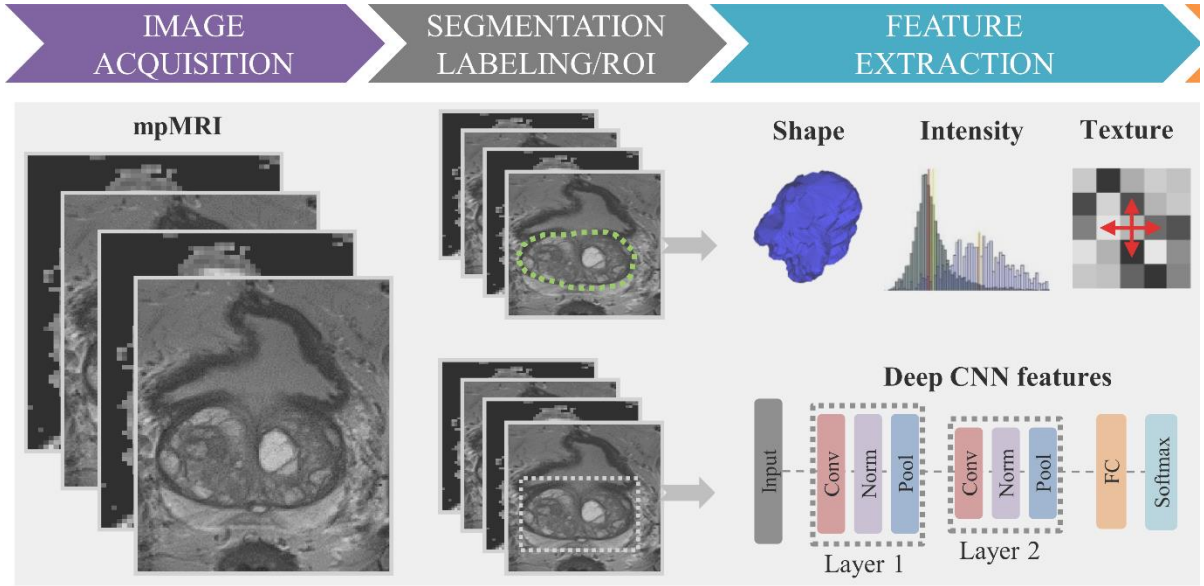


Rosso CNV ≈ 2, Bianco CNV > 2.1, Nero CNV<1.9



- Pazienti iperdiploidi;
- Pz. Con amplificazione del cromosoma 1 e delezione del cromosoma 13 (amp1 + del13);
- pazienti con delezione del cromosoma 13 (del13).

Radiomics



- Aims to extract quantitative, and ideally reproducible, information from diagnostic images.
- Includes complex pattern difficult to recognize and quantify by the human eye

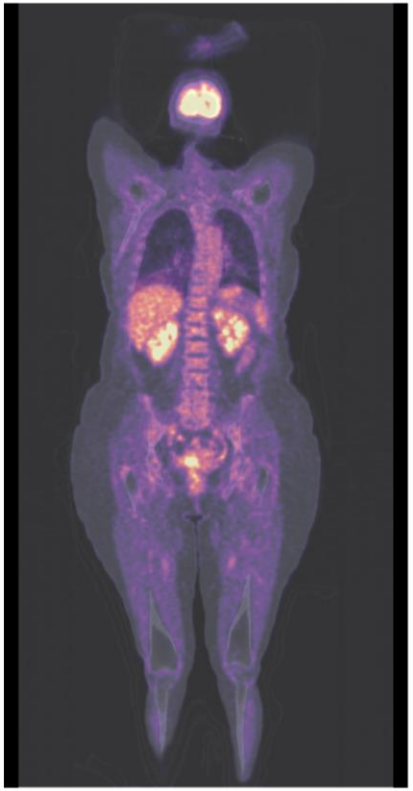
Multiple Myeloma dataset (Bologna)

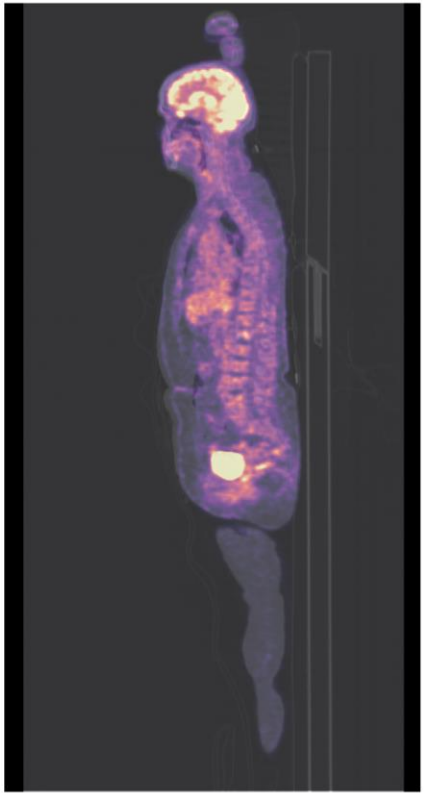
- All patients anonymised for privacy reasons through an identification number e.g. MPC_001
- Three available data types:
  1. **Clinical data** (clinical variables and outcomes) for 110 patients
  2. **Imaging data** (^{18}F-FDG-PET/CT images, sometimes at multiple time-points for the same patient) for 329 patients
  3. **Genomic data** for 154 patients
- 102 patients with both imaging and clinical data
- 89 patients with all the three data types

Operations performed on imaging dataset

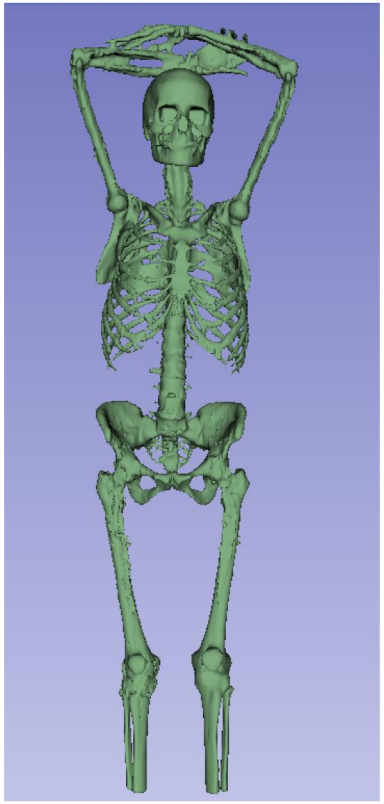
- Multimodal image registration
- Image segmentation to segment bones
- Selection of an interesting region to search for lesions: the **spine** - as suggested by clinicians
- Feature extraction, both from CT images and from PET images separately, using the segmentation as mask → *work in progress* ⚠
- Feature selection & analysis → *work in progress* ⚠

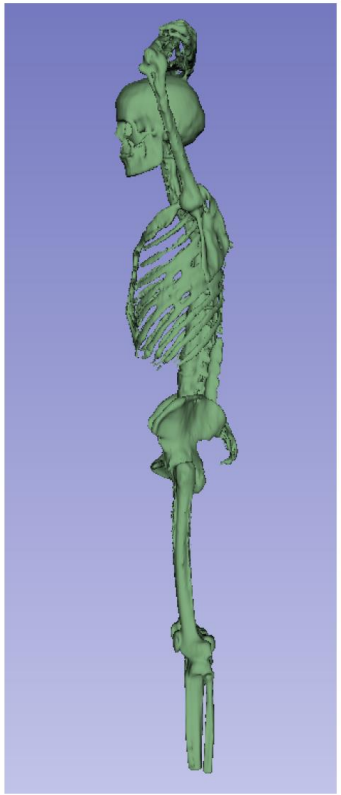
Image registration



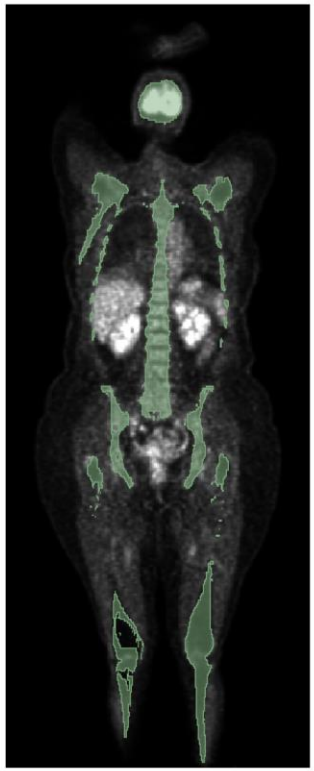


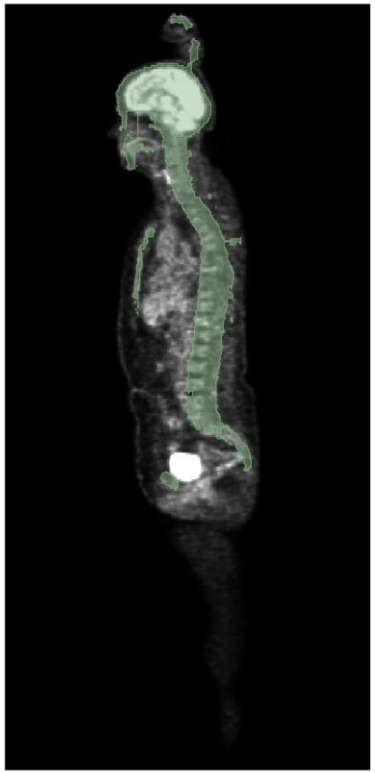
Skeleton segmentation





Segmentation superposed on PET





Volume cropping: spine region

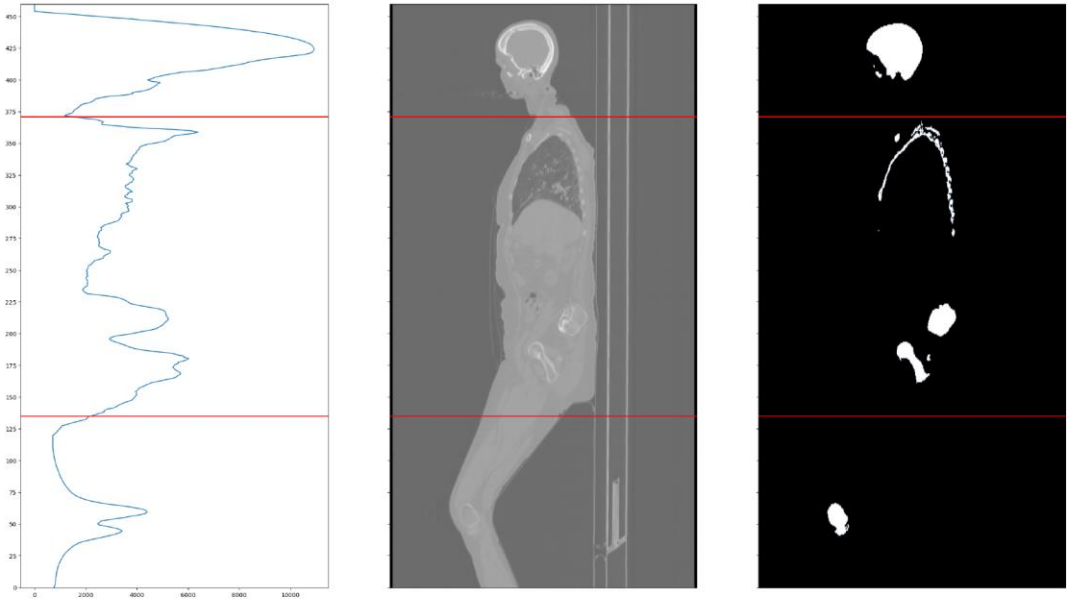


Figure 1: By plotting the CT (or segmentation) signal along the z axis, one may select the appropriate cut-points by taking the minimum to cut around the neck and the inflection point to cut below the femur head

Operations performed on clinical dataset

- Survival analysis (considering the Progression-Free Survival PFS) on the basis of PET data annotations made by clinicians:
  - BM = Bone Marrow,
  - FL = Focal Lesion,
  - EM = Extra-Medullary,
  - PM = Para-Medullary,

  each with an associated Deauville Score[1](DS) to quantify the radiopharmaceutical uptake.
- Used model: Cox's proportional hazards model

[1]C. Nanni, PET-FDG: Impetus, *Cancers* (2020), https://doi.org/10.3390/cancers12041030

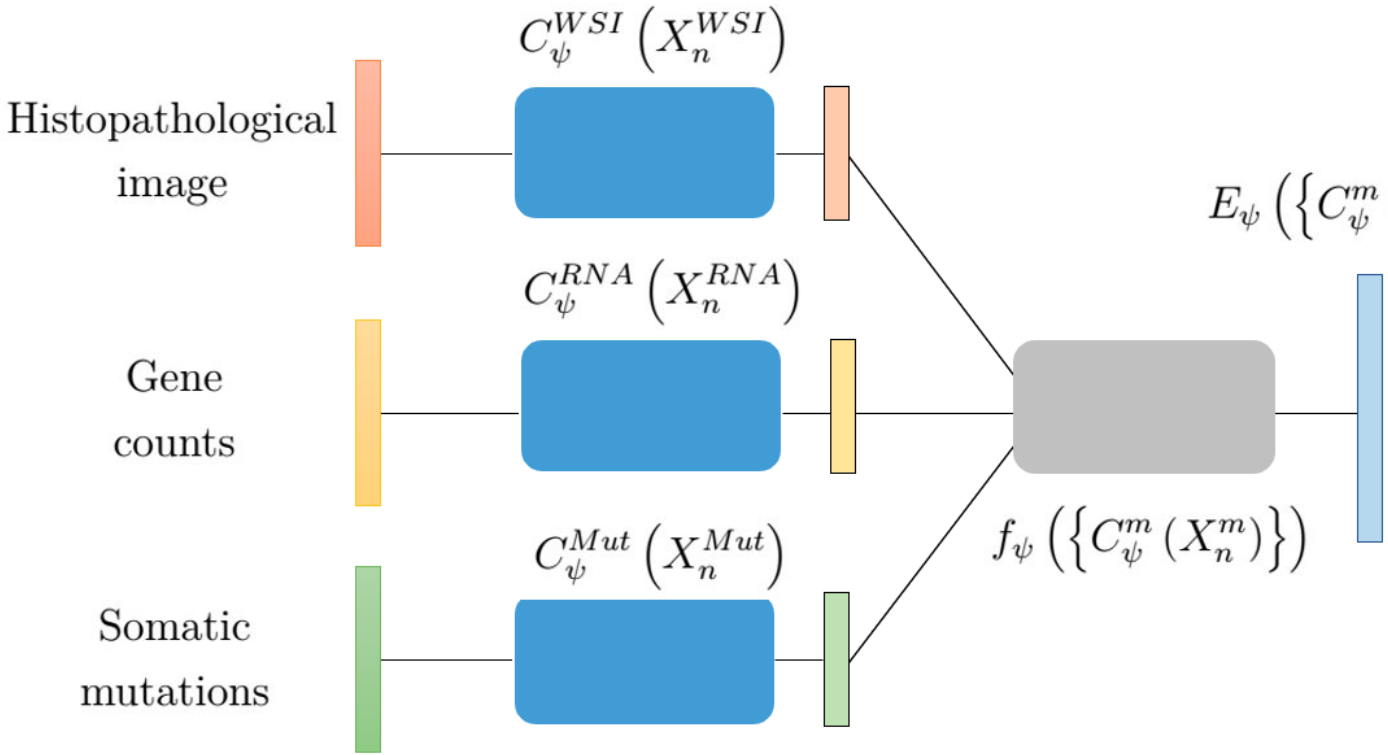
covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
BM DS	0.297	1.346	0.206	-0.107	0.701	0.898	2.016	0.0	1.439	0.150	2.736
FL DS	-0.147	0.864	0.083	-0.309	0.015	0.734	1.016	0.0	-1.773	0.076	3.714
PM DS	0.187	1.205	0.088	0.014	0.360	1.014	1.433	0.0	2.118	0.034	4.872
EM DS	0.255	1.290	0.089	0.081	0.429	1.084	1.535	0.0	2.870	0.004	7.930

Figure 2: Results for PFS_I

covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
BM DS	0.284	1.329	0.263	-0.232	0.801	0.793	2.227	0.0	1.078	0.281	1.832
FL DS	-0.359	0.699	0.139	-0.631	-0.086	0.532	0.917	0.0	-2.583	0.010	6.673
PM DS	0.361	1.434	0.133	0.100	0.622	1.105	1.862	0.0	2.711	0.007	7.221
EM DS	0.231	1.259	0.120	-0.005	0.467	0.995	1.595	0.0	1.915	0.055	4.172

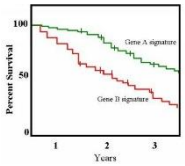
Figure 3: Results for PFS_II

- PFS_I event: 1^{st} (potential) event of disease progression
- PFS_II event: 2^{nd} (potential) event of disease progression



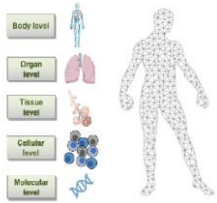
$$X = TW^{\mathrm{T}} + T_{Y\perp}P_{Y\perp}^{\mathrm{T}} + E$$
$$Y = UC^{\mathrm{T}} + U_{X\perp}P_{X\perp}^{\mathrm{T}} + F$$

MD, Omics,
Clinicians
Clinical studies
EHR,Lifestyle
Data



Multiscale Network
Modeling

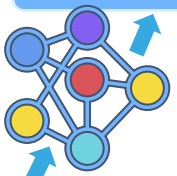
DIGITAL TWIN



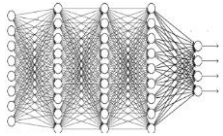
Diagnosis, Prognosis
Treatments Therapy

Surrogate model

Clinical data



training set



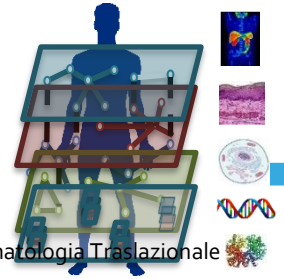
training set

Phase I

$c[t+1] = c[t] + F(c[t], \dots)$

Deterministic Diff. Eq.

Multiomics Data



training set

Partial Diff. Eq.

$$\frac{\partial u(x,t)}{\partial t} = D\frac{\partial^2 u(x,t)}{\partial x^2}$$

training set

$c[t+1] = c[t] + F(c[t], \dots)$

Phase II

Stochastic Diff. Eq.

Phase III

Related project **ongoing**

2023 PRIN-Personalized Medicine In Myeloid Neoplasms: Explainable Artificial Intelligence Solutions For Next Generation Classification And Management Of The Patients (Vannucchi, Della Porta)

2023 MAECI Science and Technology Cooperation Italy-South Korea Grant Years 2023–2025 by the Italian Ministry of Foreign Affairs and International Cooperation.

2023 PNRR on Antimicrobial Resistance (1M€)

2022 EU SYNTHEMA Synthetic generation of haematological data over federated computing frameworks 500 k€

2022- AIRC Individual Grant - IG 2021 Artificial intelligence for genomics and personalized medicine in myelodysplastic syndromes (MDS) 700 k€

2021 H2020 GENOMED4ALL Genomics and Personalized Medicine for all though Artificial Intelligence in Haematological Diseases . Federated Learning. 800 k€

ISW: (H2020)In Silico World Lowering the barriers to a universal adoption of In Silico Trials 200 k€

2019 EU Project Versatile Emerging infectious disease Observatory (VEO) 60 months Data analytics and modeling. Data Analytics and modeling. EU contribution to UNIBO 341378 € (the whole project is 15M€) Coordinator Marion Koopmans

2019 EU project HARMONY-PLUS: HEALTHCARE ALLIANCE FOR RESOURCEFUL MEDICINES OFFENSIVE AGAINST NEOPLASMS IN HEMATOLOGY – PLUS (HARMONY PLUS). 36 months . Data Analytics and Big Biomedical data integration for hematological malignancies, including the set-up of a pan European computing facility. Role WP Co-Leader. Coordinator J.M. Hernandez. EU contribution to UNIBO 339.000 € (the whole project is a 12 M€)

2017 EU project HARMONY: Alliance for Resourceful Medicines Offensive against Neoplasms in HematologY. 60 months . Data Analytics and Big Biomedical data integration for hematological malignancies, including the set-up of a pan European computing facility. Role WP Leader. J.M. Hernandez. EU contribution to UNIBO 800.000 € (the whole project is a 40 M€)