GenoMed4All & ERN-EuroBloodNet

Educational Program on Artificial Intelligence for public-at-large

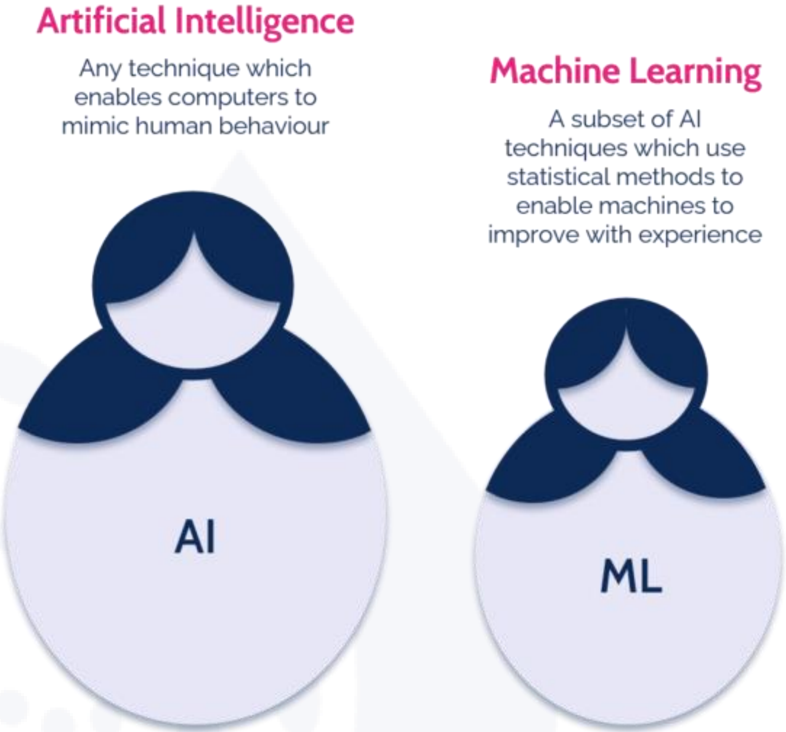
Data integration and analysis (Artificial Intelligence)

Disease Progression Prediction

Tiziana Sanavia

University of Torino (Italy)

Artificial Intelligence and Machine Learning



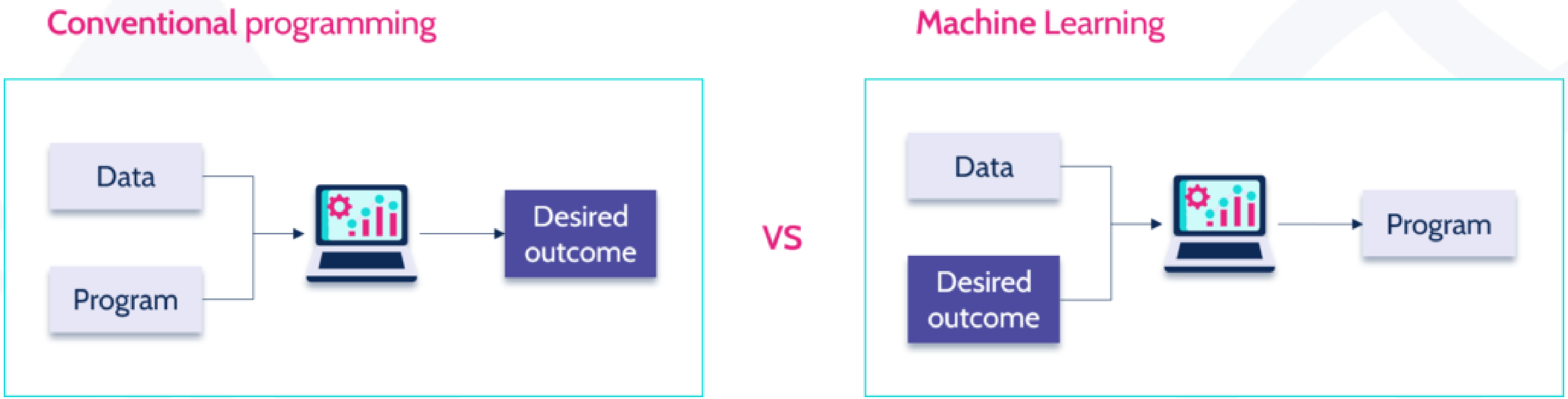
Artificial Intelligence (AI) refers to any technique that makes computers capable of mimicking human behavior to address and solve problems. In the 1980s, one group of AI methods became more common: Machine Learning.

Machine learning (ML) which describes the ability of an algorithm to ‘learn’ by finding patterns in large datasets. In other words, the ‘answers’ produced by ML algorithms are inferences made from statistical analysis of very large datasets.

WHAT DOES ‘LEARN’ MEAN?

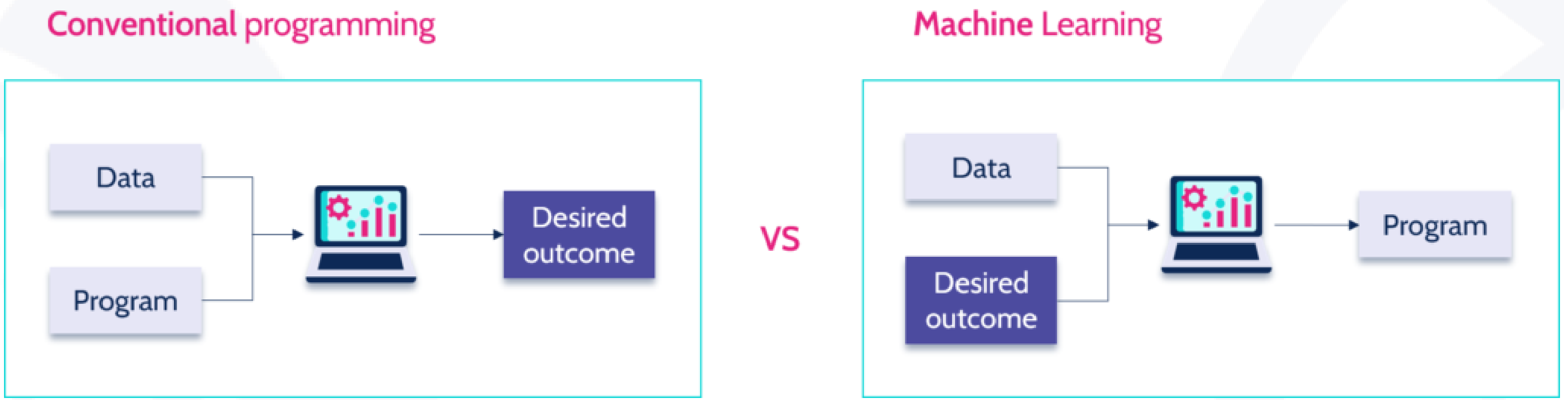
The concept of Learning

The key here is to step away from any preconceived notions on the concept of learning, as we humans understand it. Instead of combining a set of human-created rules with data to create answers to a problem, as is done also in conventional programming, machine learning uses **data and answers** to discover the rules behind a problem.



The concept of Learning

To learn the rules governing a phenomenon, machines must go through a **learning process**, trying different rules and learning from how well they perform. This is where reward and loss functions come into play: they allow the machine to automatically assess the rules it created. Thus, for a machine, ‘learning’ is better understood as the process of maximizing its reward function, limited to the context of that specific task and training data.



Learn from clinical data

Age | IHC Measure 'y' | Mutation 'x'

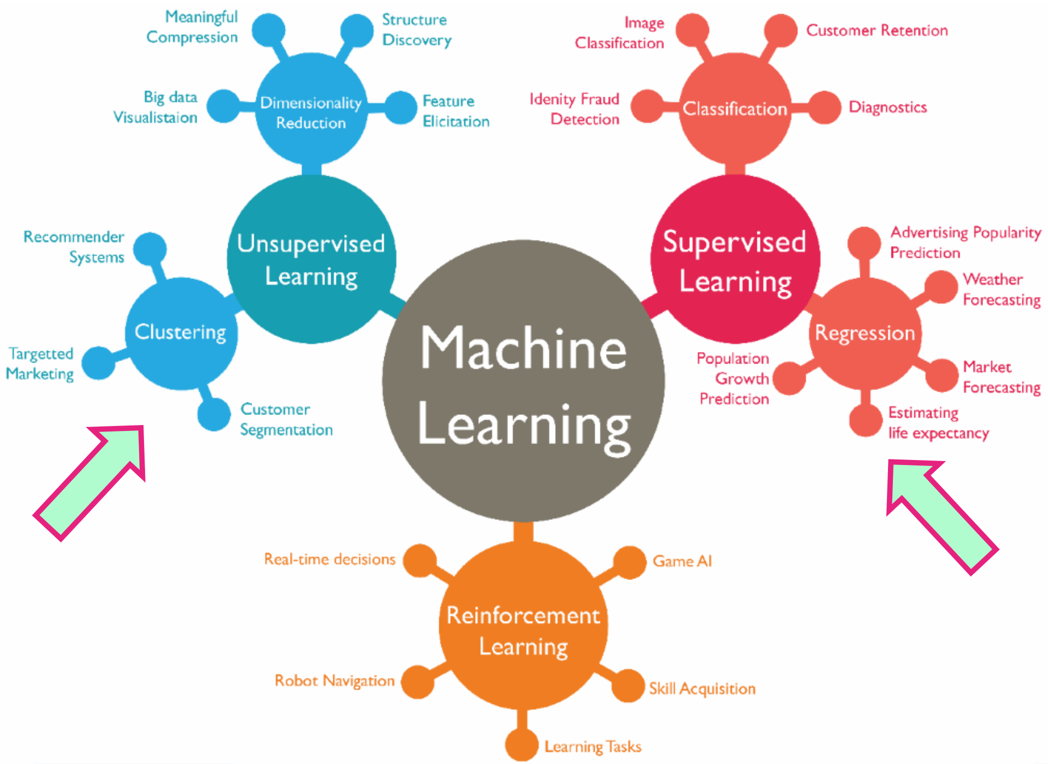
	Feature 1	Feature 2	Feature 3	…
Patient 1				
Patient 2				
Patient 3				
…				

Retinopathy | Microalbuminuria

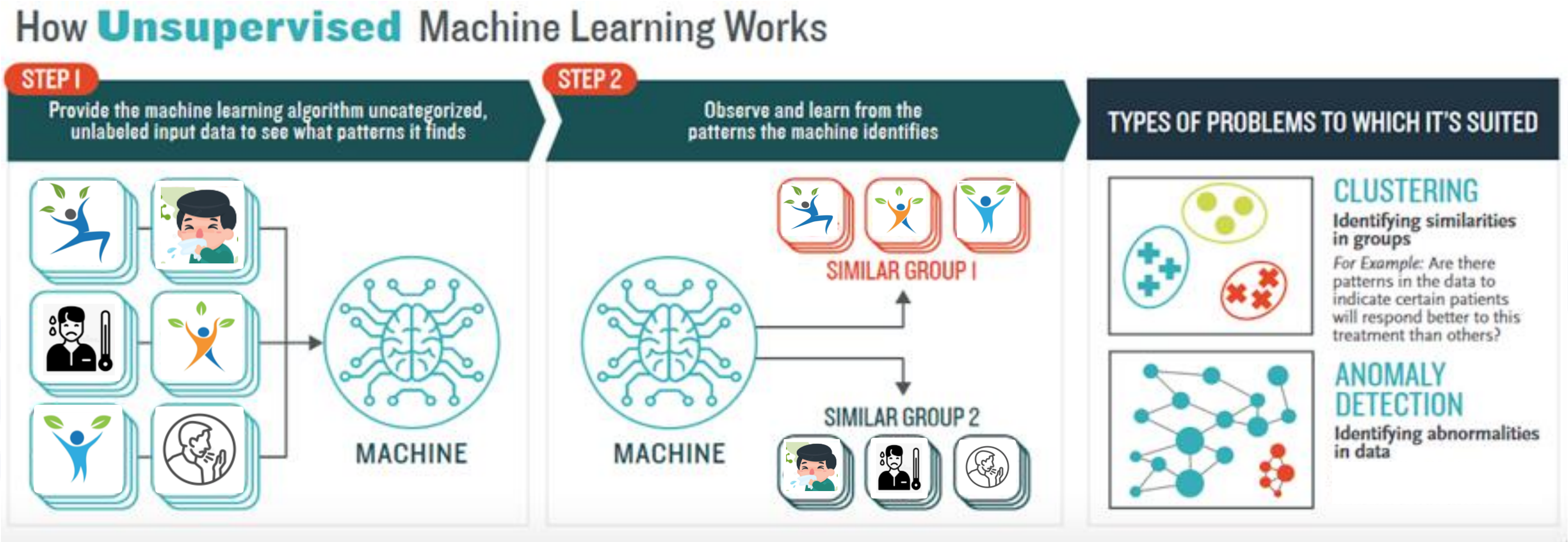
Outcome 1	Outcome 2	…
YES	YES	
YES	NO	
NO	YES	
NO	NO	

THESE ARE LABELS!

How can we learn: Machine Learning includes many algorithms!



Unsupervised approaches



Main AIM of Unsupervised Analysis → STRATIFICATION

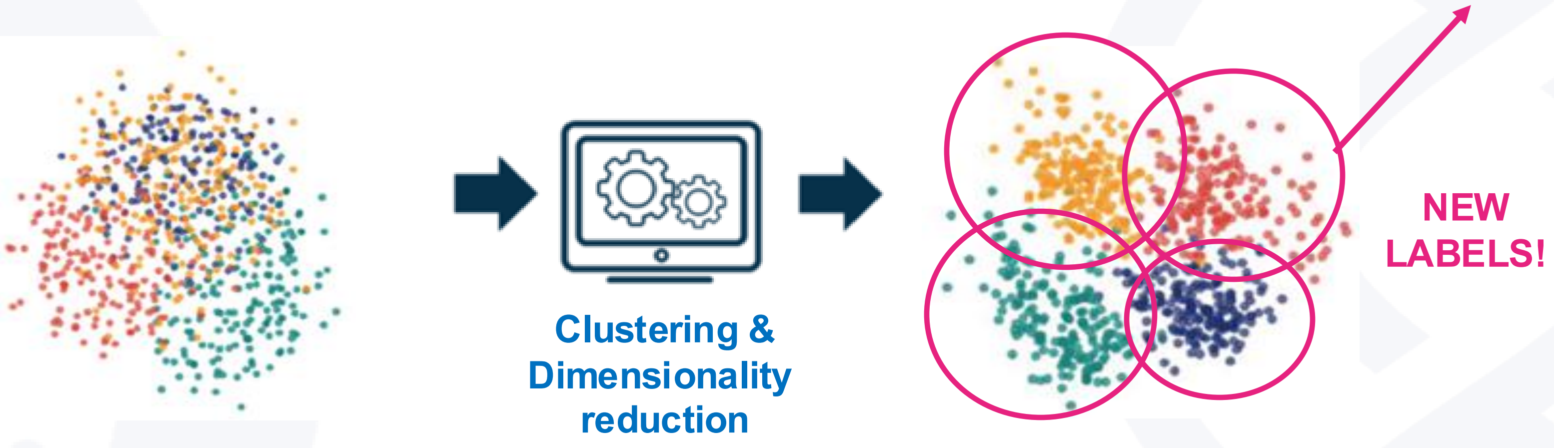
Dataset

Patients

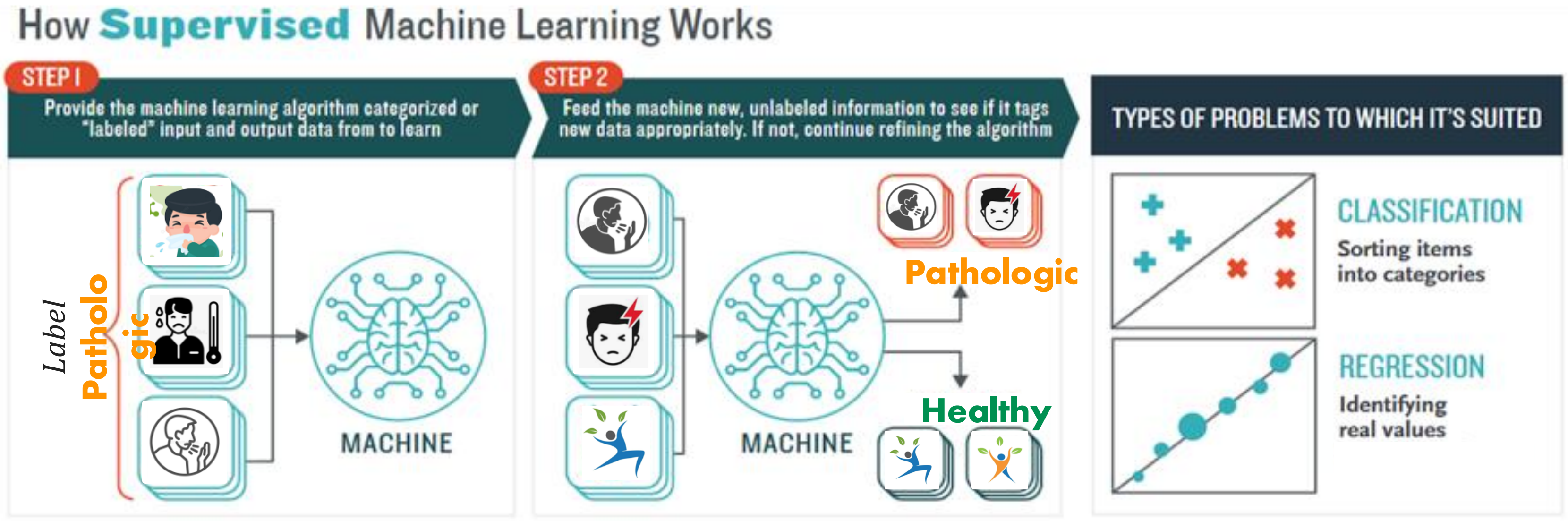
Feature 1	Feature 2	Feature 3	…

Dataset + NEW FEATURE/OUTCOME

Feature 1	Feature 2	Feature 3	…	Clusters

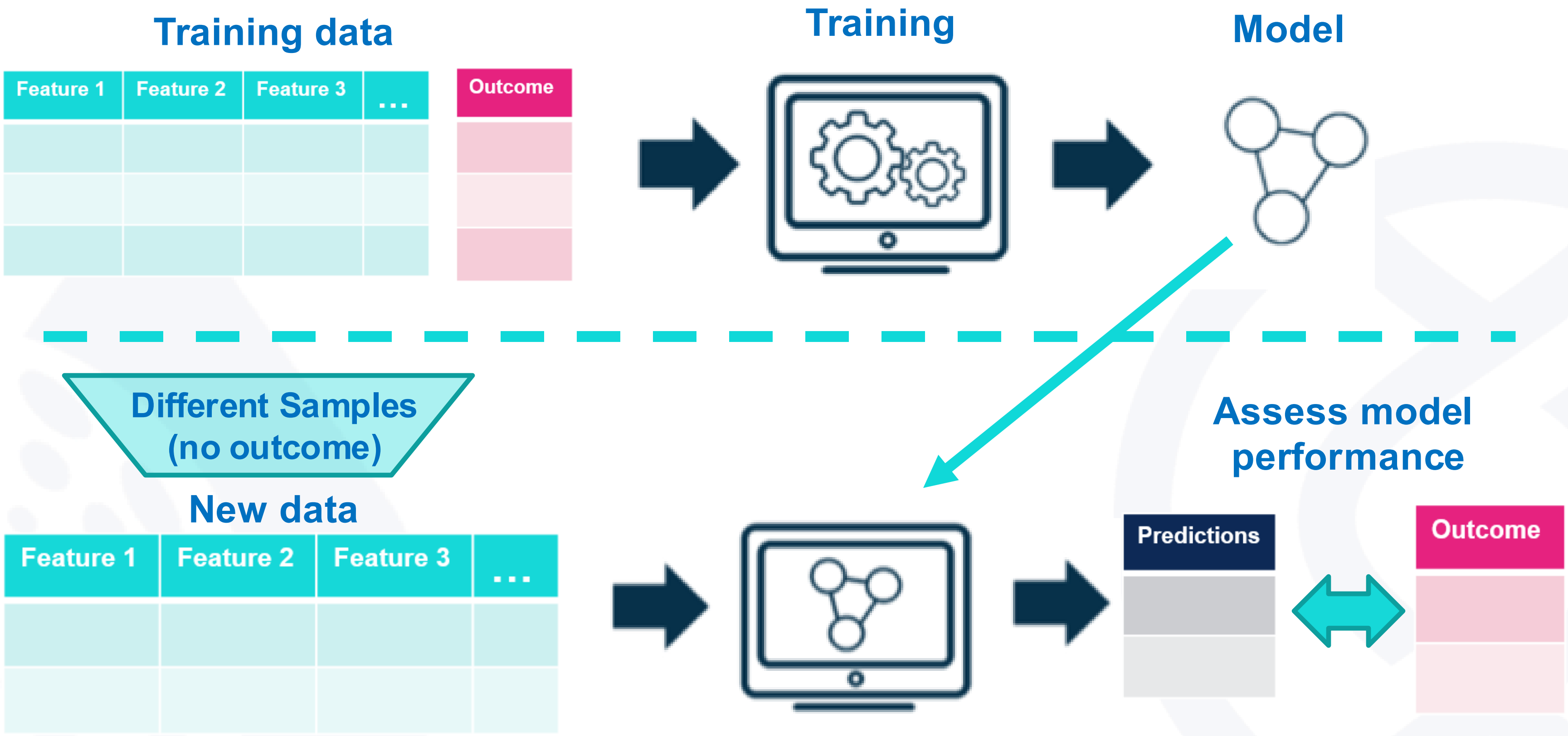


Supervised approaches

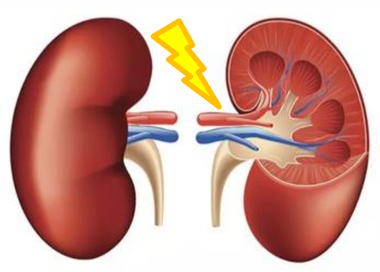


WHAT WE NEED? FEATURES, OUTCOMES AND LABELS!

Main AIM of Supervised Analysis → PREDICTION



Cross-sectional or Longitudinal?



Cross-sectional

Predict Microalbuminuria **events** in new patients?

Microalb.

Longitudinal

Predict across time

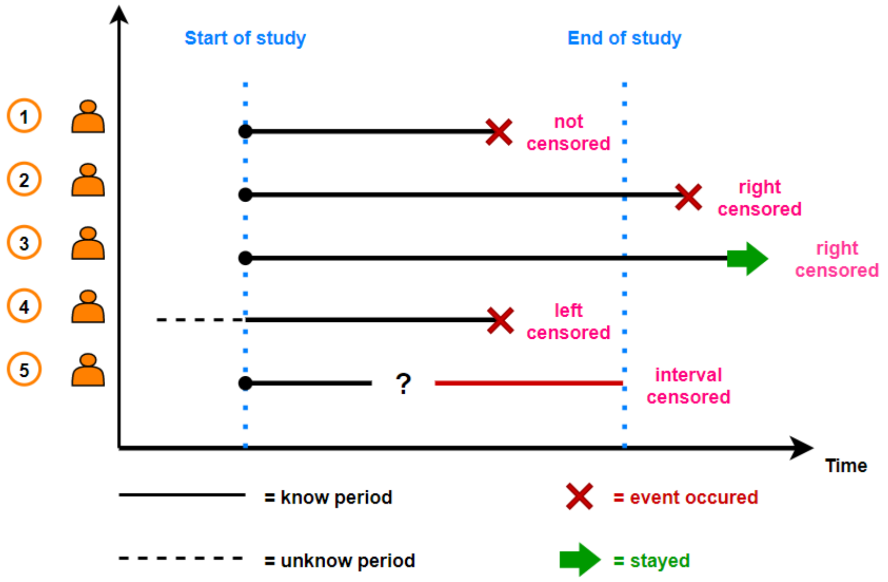
Microalbuminuria **events**

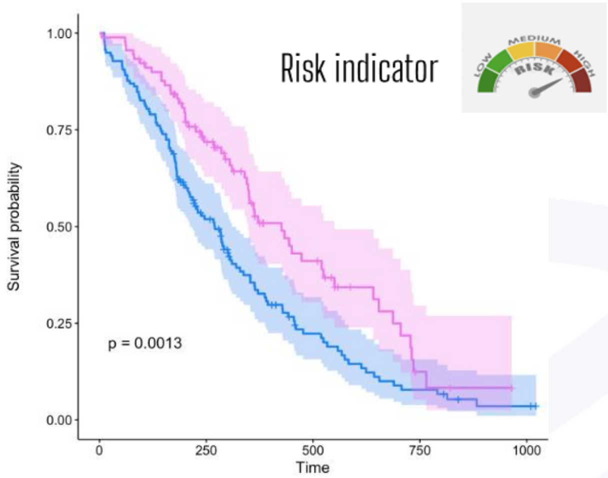
in new patients?

Microalb.	Time to Microalb.

Survival Analysis

Survival Analysis is a crucial tool that aims to predict the time to an event of clinical interest at an individual patient level Handles temporal censored data

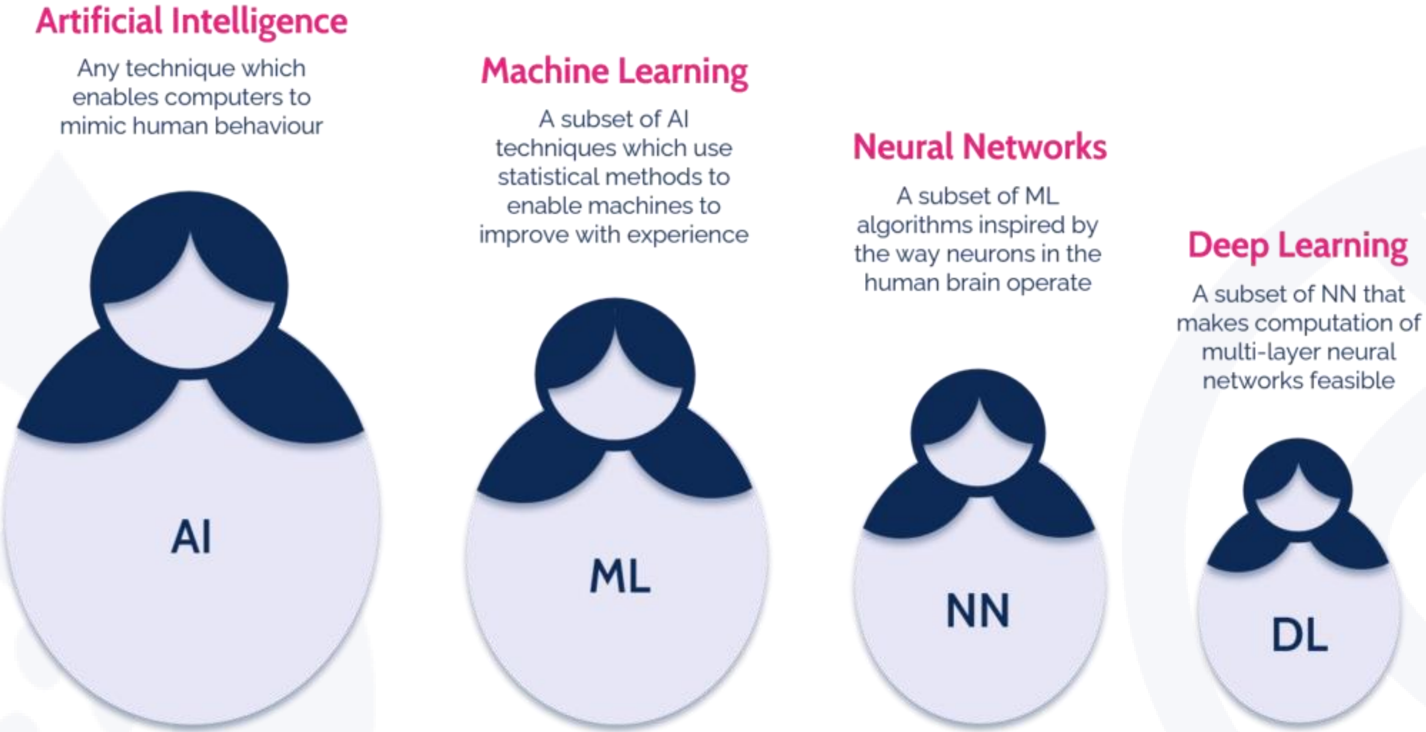




BASIC APPROACH:
COX REGRESSION MODEL

Linear combination
of the features

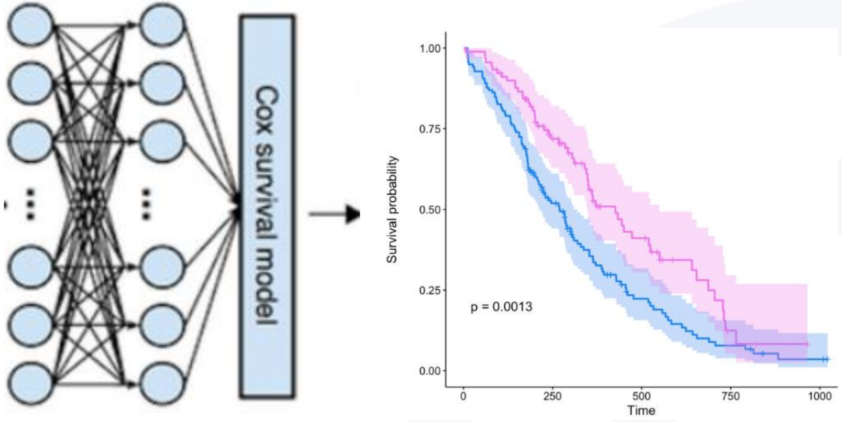
Survival Analysis and Machine Learning



Survival Analysis and Machine Learning

Methods

- Penalized Cox Regression
- Random Survival Forests
- DeepSurv, DeepHit and other deep learning approaches

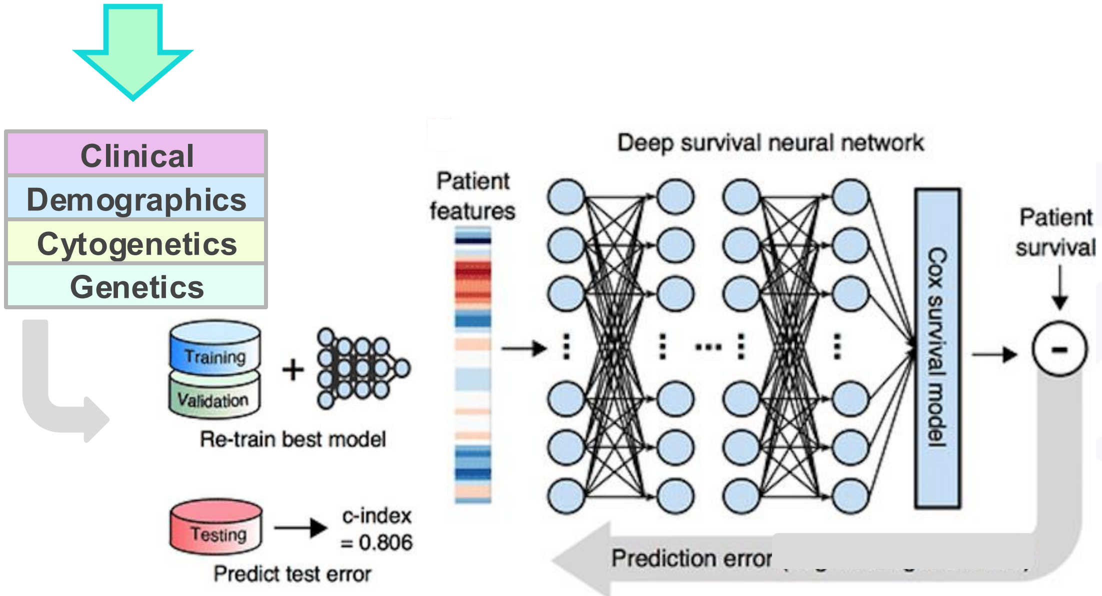


Advantages: NON-LINEAR combination of the features → new patterns among the clinical variables

Multi-modal Machine Learning survival approaches

Methods

- Penalized Cox Regression
- Random Survival Forests
- DeepSurv, DeepHit and other deep learning approaches



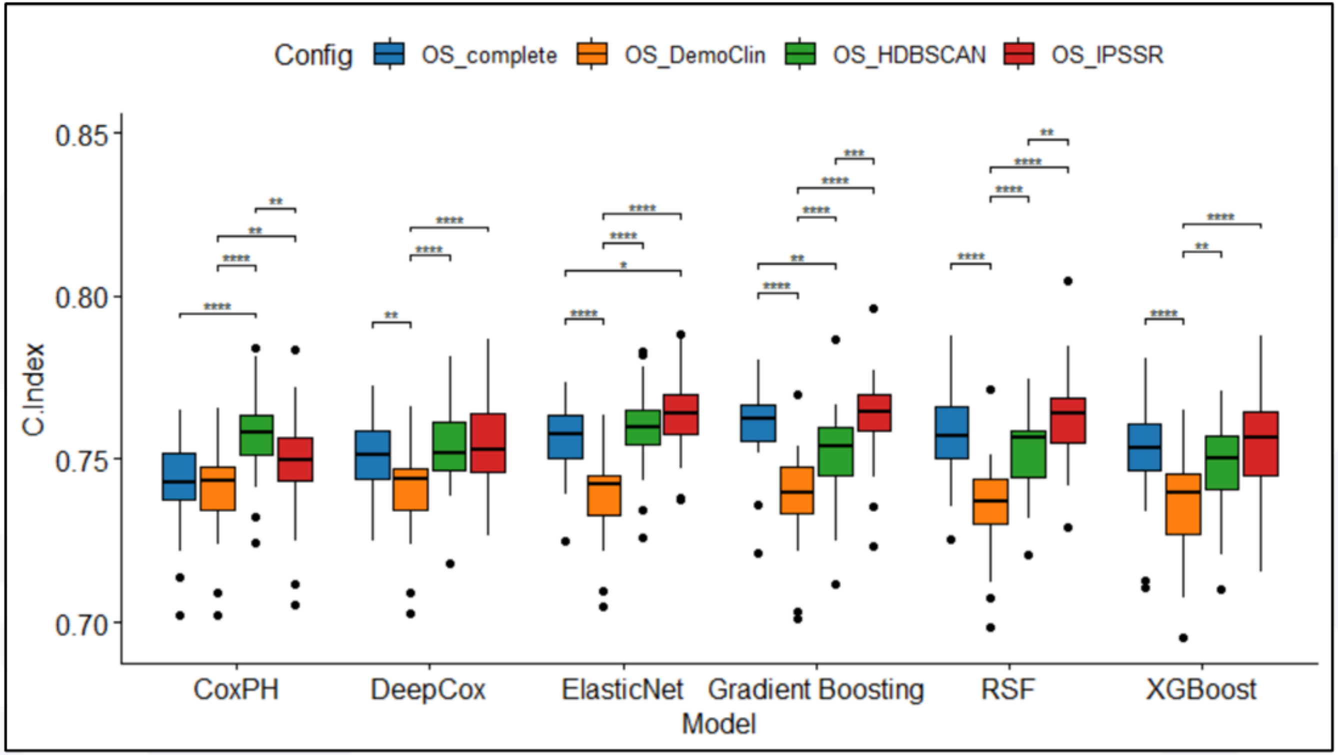
Critical points:
- **high-dimensional, heterogeneous**
- **containing missing information**

Application to use case MDS: 2,043 patients with 3 types of events: Overall, Leukemia-Free and Event-Free Survival

Integrative Analysis: data configurations

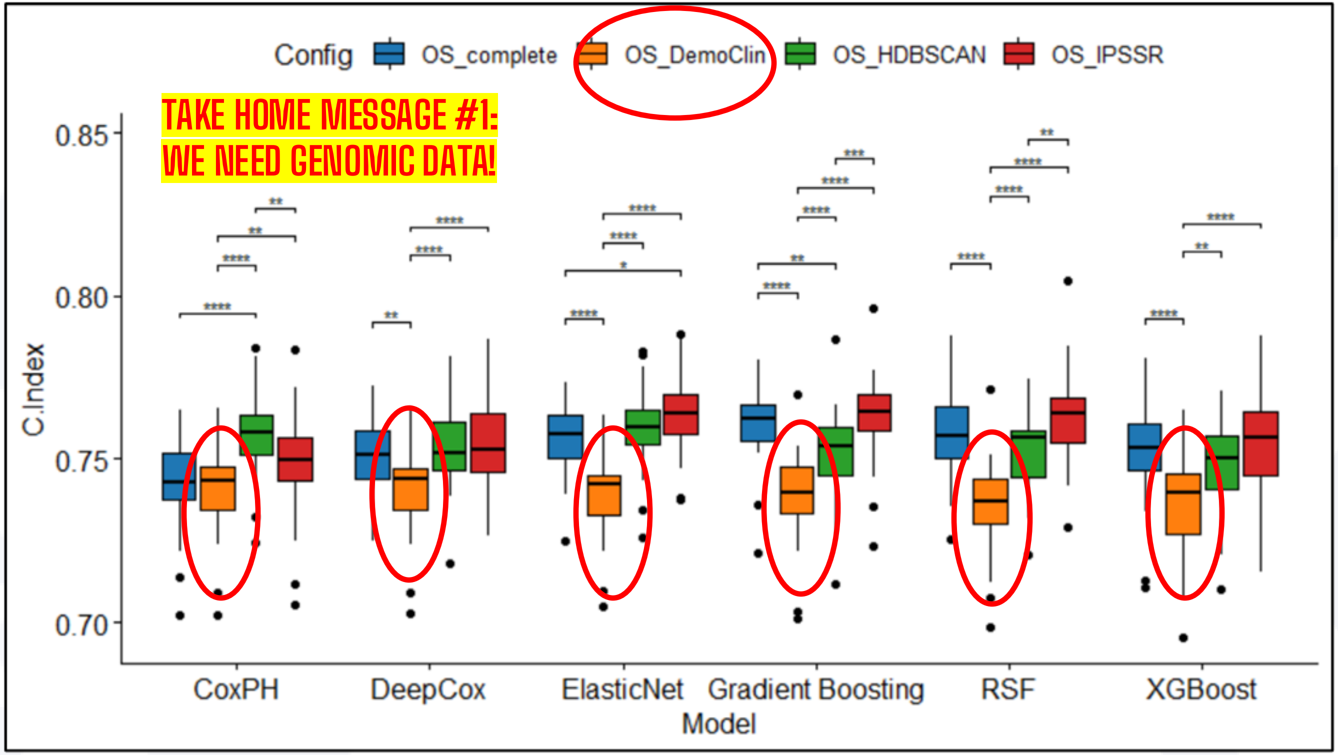
- "DemoClin": Demographic + Clinical variables
- "Complete": Demographic + Clinical + Genetic + Cytogenetic variables
- "IPSSR": Demographic + Clinical + Genetic variables + IPSSR score
- "HDBSCAN": Demographic + Clinical + Cytogenetic variables + HDBSCAN Clustering on genetic variables

Results



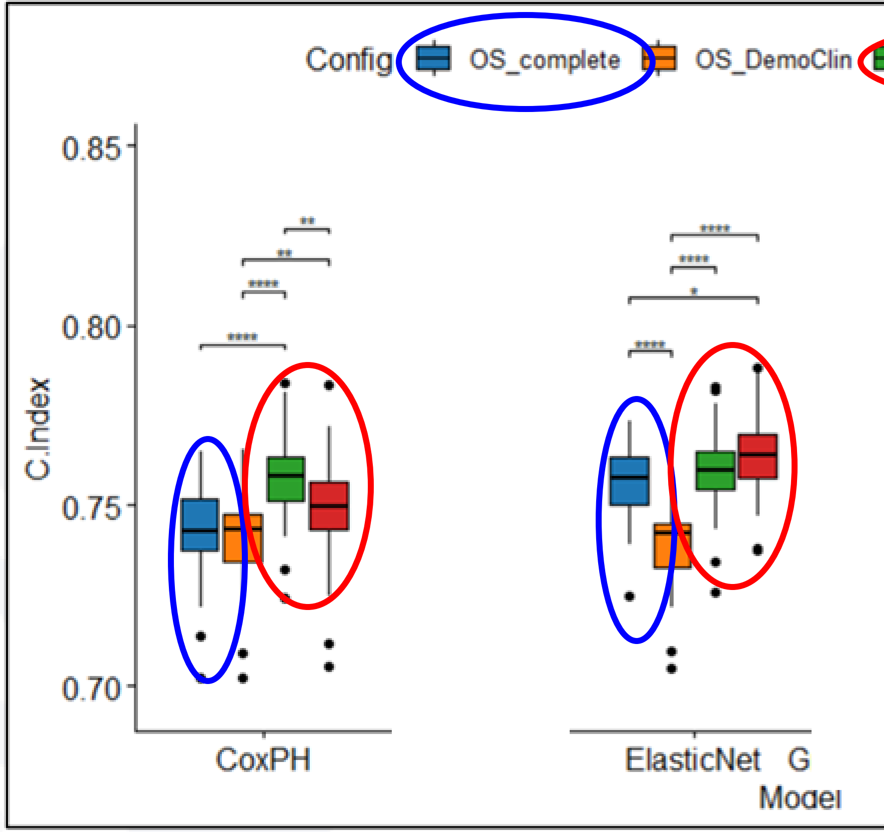
Config
OS_complete
OS_DemoClin
OS_HDBSCAN
OS_IPSSR
C.Index
0.85
0.80
0.75
0.70
CoxPH
DeepCox
ElasticNet
Gradient Boosting
RSF
XGBoost
Model

Results



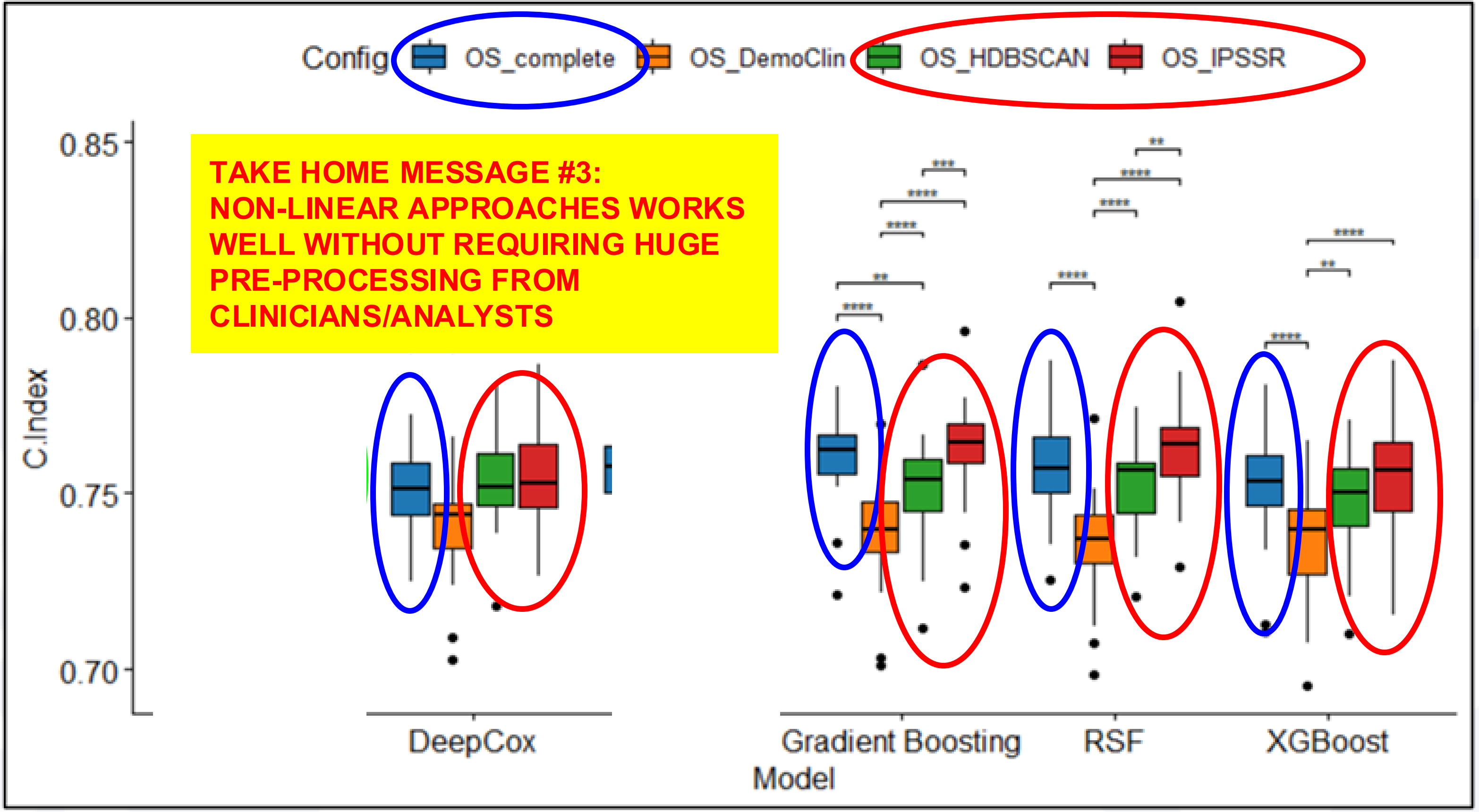
Config
OS_complete
OS_DemoClin
OS_HDBSCAN
OS_IPSSR
TAKE HOME MESSAGE #1:
WE NEED GENOMIC DATA!
C.Index
0.85
0.80
0.75
0.70
CoxPH
DeepCox
ElasticNet
Gradient Boosting
RSF
XGBoost
Model

Results



TAKE HOME MESSAGE #2: LINEAR APPROACHES REQUIRES HIGH-LEVEL PRE-PROCESSING TO IMPROVE THE PERFORMANCE

Results



Config
OS_complete
OS_DemoClin
OS_HDBSCAN
OS_IPSSR
TAKE HOME MESSAGE #3:
NON-LINEAR APPROACHES WORKS WELL WITHOUT REQUIRING HUGE PRE-PROCESSING FROM CLINICIANS/ANALYSTS
C.Index
0.85
0.80
0.75
0.70
DeepCox
Gradient Boosting
RSF
XGBoost
Model

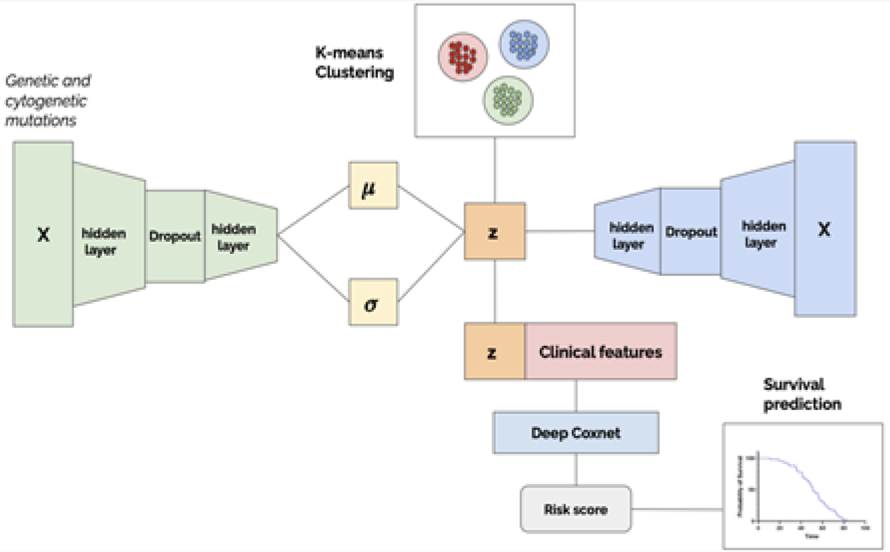
Combining unsupervised and supervised approaches

- Myelodysplastic syndromes complexity:
  - Heterogeneous group of hematopoietic stem cells disorders
  - Intricate genetic mutational landscape (point mutations, chromosomal deletions, translocations)
  - Possible progression into Amyotrophic lateral sclerosis (AML)
  - Heterogeneous survival rates

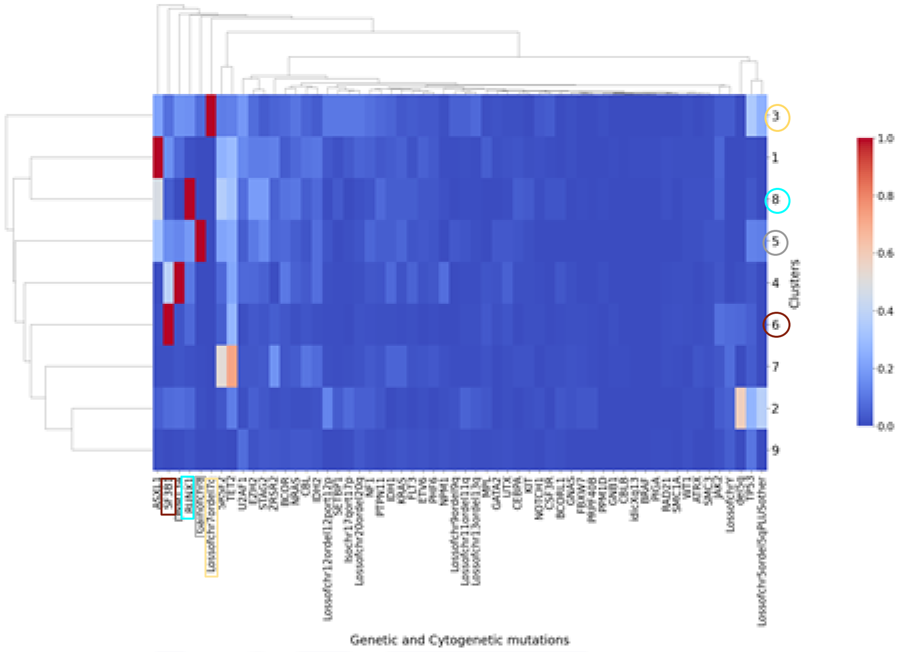
- A precise classification of MDS subtypes is required (WHO 2016)
- Prognosis prediction is challenging
- Risk stratification that includes genetic and cytogenetics information

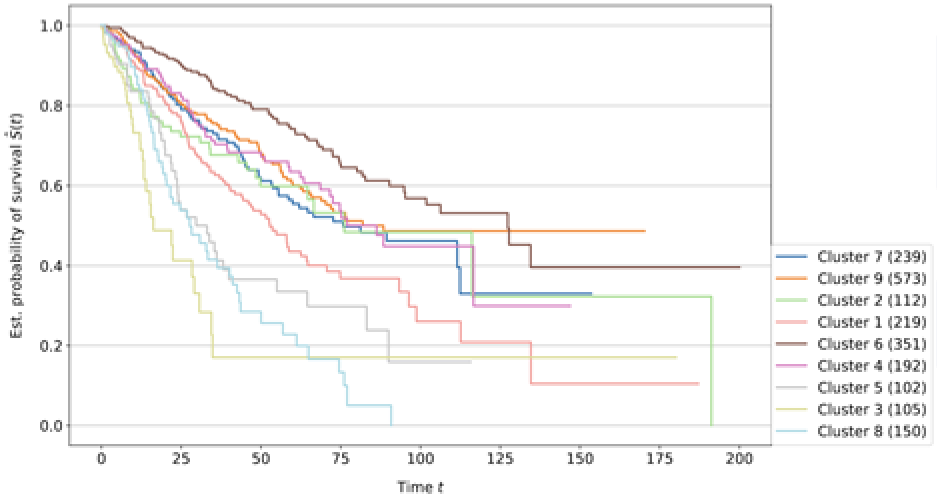
VAE-Surv Framework for MDS

- The aim of the proposed model is to build an end-to-end framework that allows the prediction of patients' prognosis and identify genetic-based clusters of the disease
- The VAE block is trained using only genetic and cytogenetic information
- The learned latent representation is concatenated with clinical features and used to feed a DeepSurv neural network in a separate step for survival analysis prediction



VAE-Surv Framework for MDS

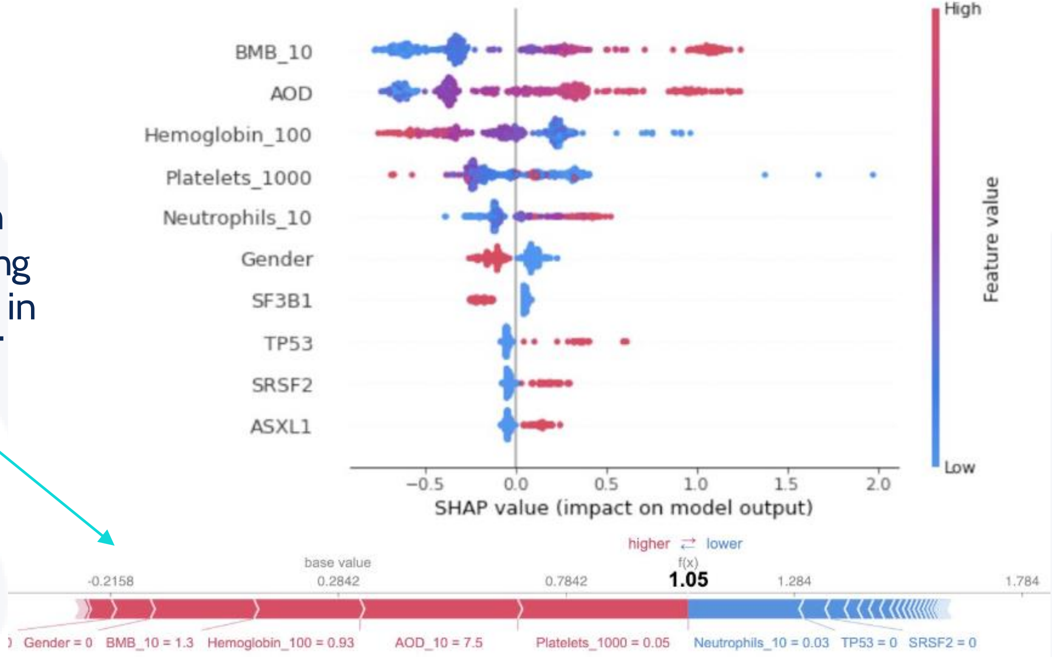




Making the results interpretable to clinicians

❖ Explainability

It is possible to detect the main features affecting the progression in **EACH PATIENT SEPARATELY**



Next Steps

❖ **Sickle cell disease**:

➢ A cohort of **1000 patients** (500 adults and 500 pediatric) from 5 different healthcare institutions has been collected → Data will be released by the end of the year

➢ It will include longitudinal clinical data, as well as genomic and metabolomic features

➢ One of the main clinical outcomes to predict will be the **Silent Cerebral Infarct** (SCI) event → MRI data available

❖ **Multiple Myeloma**:

➢ A supplementary cohort of **176 patients** (in addition to the 100 patients already collected) will be shared by the Genomed4All consortium

➢ Data will include clinical features and Copy Number Variation (**CNV**) information

➢ Radiomic data (PET derived) will be parallely analysed by University of Bologna

❖ **Federated learning** algorithm already validated on synthetic data will be applied to **real-world scenarios**

UNITO-AI team

Piero Fariselli

piero.fariselli@unito.it

Tiziana Sanavia

tiziana.sanavia@unito.it

Cesare Rollo

cesare.rollo@unito.it

Thanks!

Any questions?

GenoMed4All & ERN-EuroBloodNet

Educational Program on Artificial Intelligence for public-at-large

Follow us!

genomed4all.eu

@genomed4all

/genomed4all

eurobloodnet.eu

@ERNEuroBloodNet

/ERNEuroBloodNet

Acknowledgements

European Reference Network for rare or low prevalence complex diseases

Network Hematological Diseases (ERN EuroBloodNet)

Co-funded by the European Union

This project is supported by the European Reference Network on Rare Haematological Diseases (ERN-EuroBloodNet)-Project ID No 101085717. ERN-EuroBloodNet is partly co-funded by the European Union within the framework of the Fourth EU Health Programme.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

GENOMED4ALL

Funded by the European Union

GenoMed4All has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017549.

Federated Learning

Unleashing the Power of Collaborative Intelligence

How to achieve faster advances in medical research?

Introduction

Transforming Medicine with Federated Learning: An Indispensable Need

Data privacy

Global Collaboration

Treatment Personalization

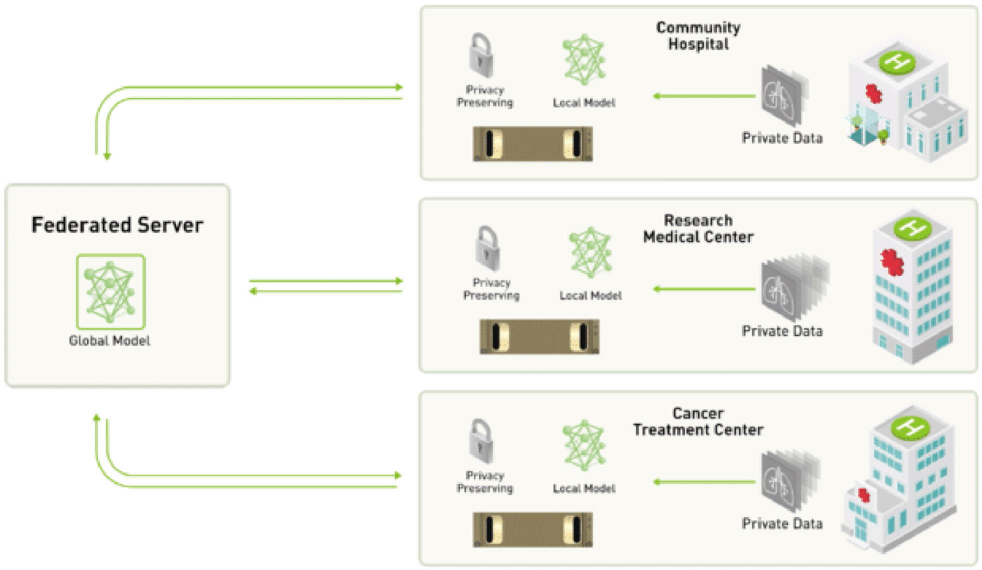
Research Acceleration

Costs Reduction

Definition

What is Federated Learning?

Federated Learning is a decentralized machine learning approach that enables training models on distributed data without the need for data sharing, bringing privacy and scalability benefits to the field of AI.



Process

1. Model and task selection
2. Local training with local data
3. Performance update
4. Global aggregation
5. Repeat steps 2, 3, 4

SOURCE: https://blogs.nvidia.com/blog/2019/10/13/what-is-federated-learning/

Challenges

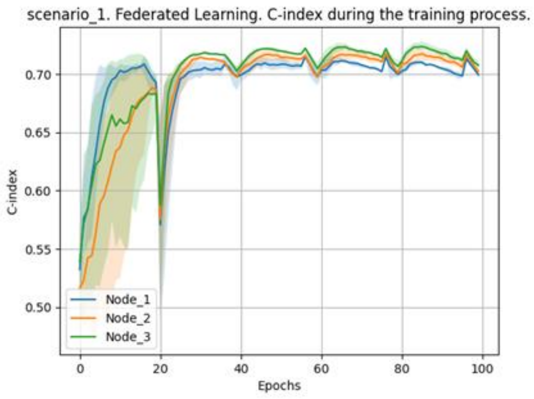
Implementation of Federated Learning environments

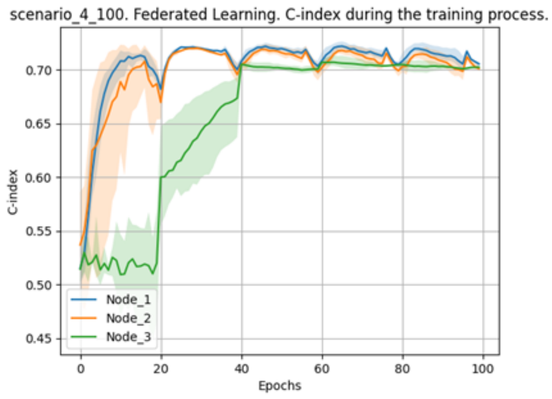
- Privacy and Security
- Communication and Bandwidth
- Model Coordination and Aggregation
- Data Heterogeneity
- Latency
- Refusal to participate
- Model evaluation

⚠ **A careful approach to the design and implementation of Federated Learning systems is essential**

Genomed4ALL results

MDS results in a FL environment





C-index value for each experiment. The higher value, the better performance.

Scenario_1: All participants have the same quantity and quality of data
Scenario_4: Participant 3 has less quantity (100 or 10 samples) and quality of data

Genomed4ALL results

MM results in a FL environment

Scenarios	Nodes	Therapy 1 (iso - fed)	Therapy 2 (iso - fed)	Therapy 3 (iso - fed)
Scenario 1	Node 1	0.28 – **0.40**	0.79 – **0.80**	0.82 – **0.86**
	Node 2	0.22 – **0.40**	0.76 – **0.80**	0.83 – **0.86**
	Node 3	0.36 – **0.40**	0.74 – **0.80**	0.80 – **0.86**
Scenario 3_10	Node 1	0.28 – **0.40**	0.79 – 0.75	0.82 – 0.81
	Node 2	0.22 – **0.40**	0.66 – **0.75**	0.79 – **0.81**
	Node 3	0.09 – **0.40**	0.77 – 0.75	0.78 – **0.81**
Scenario 3_50	Node 1	0.28 – **0.40**	0.79 – 0.75	0.82 – 0.80
	Node 2	0.22 – **0.40**	0.66 – **0.75**	0.79 – **0.80**
	Node 3	0.40 – **0.40**	0.67 – **0.75**	0.76 – **0.80**

F1-score for each experiment. The higher value, the better performance. Bold implies best result.

Scenario_1: All participants have the same quantity and quality of data
Scenario_3: Participant 3 has less quantity (10 or 50 samples) and quality of data

Conclusions

Federated Learning is moving medicine into a more collaborative, safe and efficient future.

Thanks!

Any questions?

GenoMed4All & ERN-EuroBloodNet

Educational Program on Artificial Intelligence for public-at-large

Follow us!

genomed4all.eu

@genomed4all

/genomed4all

eurobloodnet.eu

@ERNEuroBloodNet

/ERNEuroBloodNet

Acknowledgements

European Reference Network
for rare or low prevalence complex diseases

Network Hematological Diseases (ERN EuroBloodNet)

Co-funded by the European Union

This project is supported by the European Reference Network on Rare Haematological Diseases (ERN-EuroBloodNet)-Project ID No 101085717. ERN-EuroBloodNet is partly co-funded by the European Union within the framework of the Fourth EU Health Programme.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

GENOMED4ALL

Funded by the European Union

GenoMed4All has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017549.

Data integration and analysis (Artificial Intelligence)

Gastone.Castellani@unibo.it

Use cases

MM, MDS, SCD

In God we trust, all others bring data.

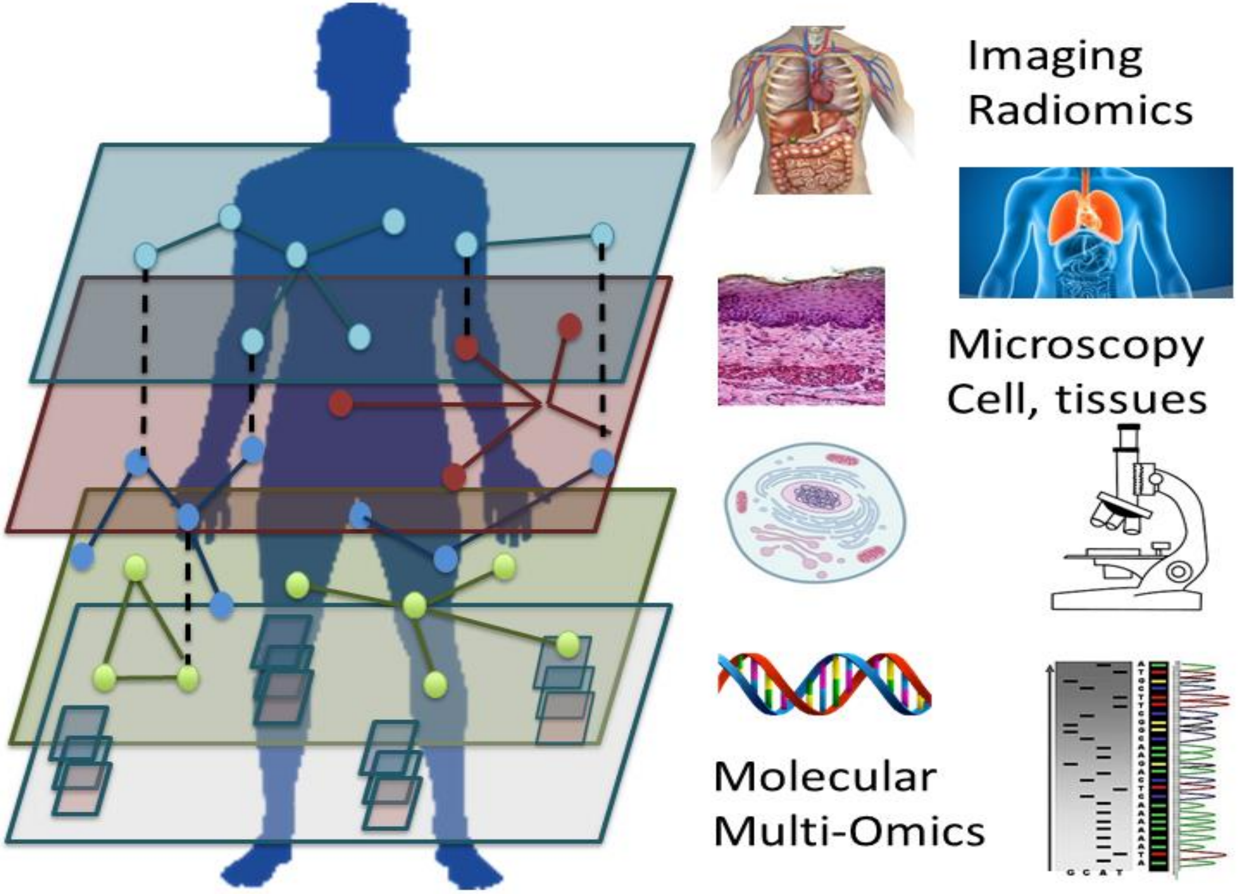
–William E. Deming

HARMONY https://www.harmony-alliance.eu/

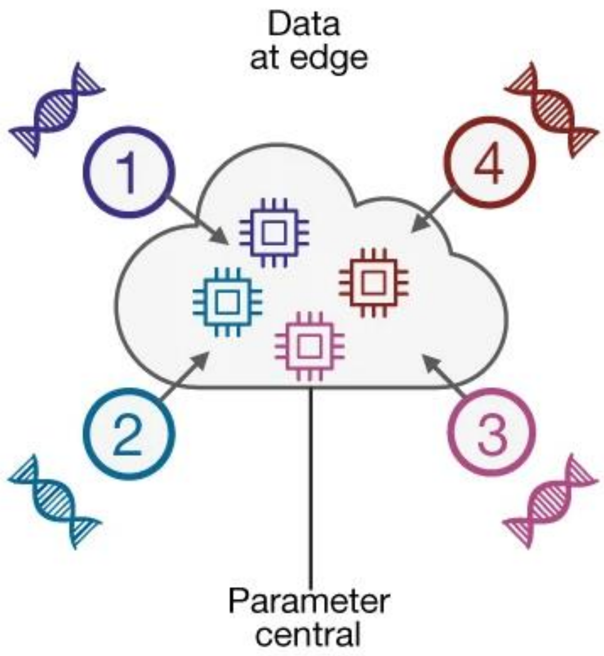
Login Subscribe search

Vision Alliance Partners Hematologic Malignancies Work Packages News Meetings Contact

Data Integration In Hematology



Federated learning applied to clinical data



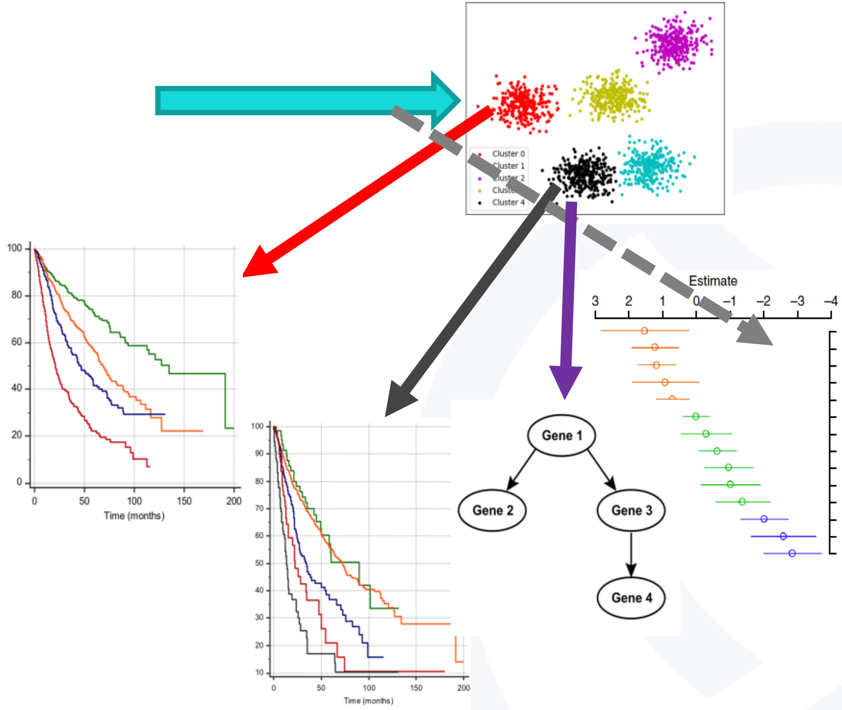
Federated learning:

- data are kept with the data contributor (“edges” or “nodes”)
- **<u>computing is performed at the site of local data storage</u>**
- parameter settings are orchestrated and “learned” by a central parameter server

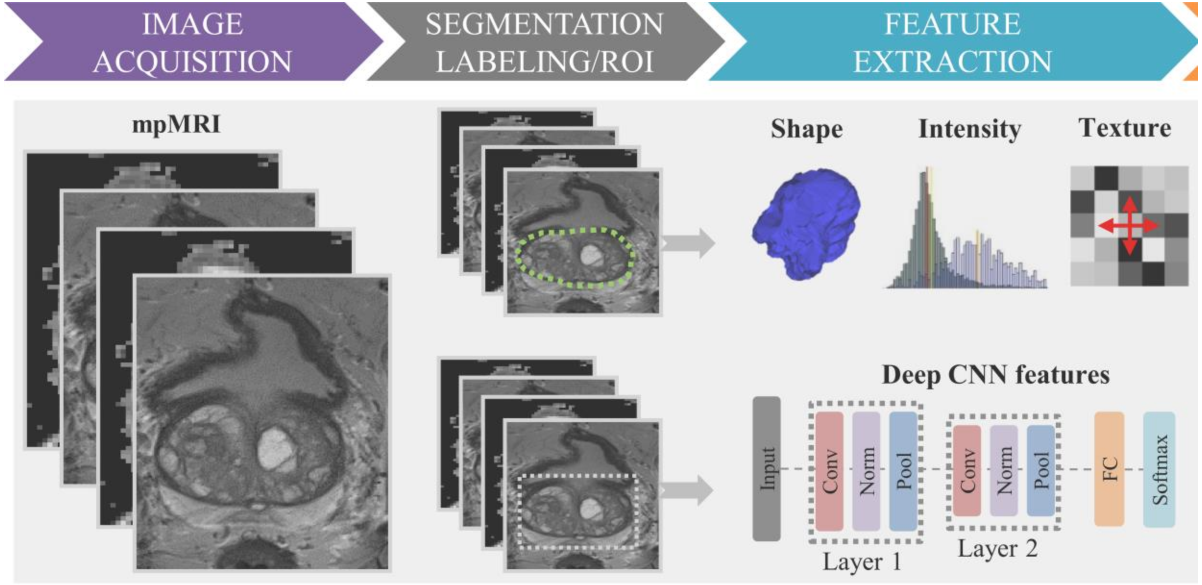
Cytogenetic, genetic and clinical data (VAF, TD, CNV etc)

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

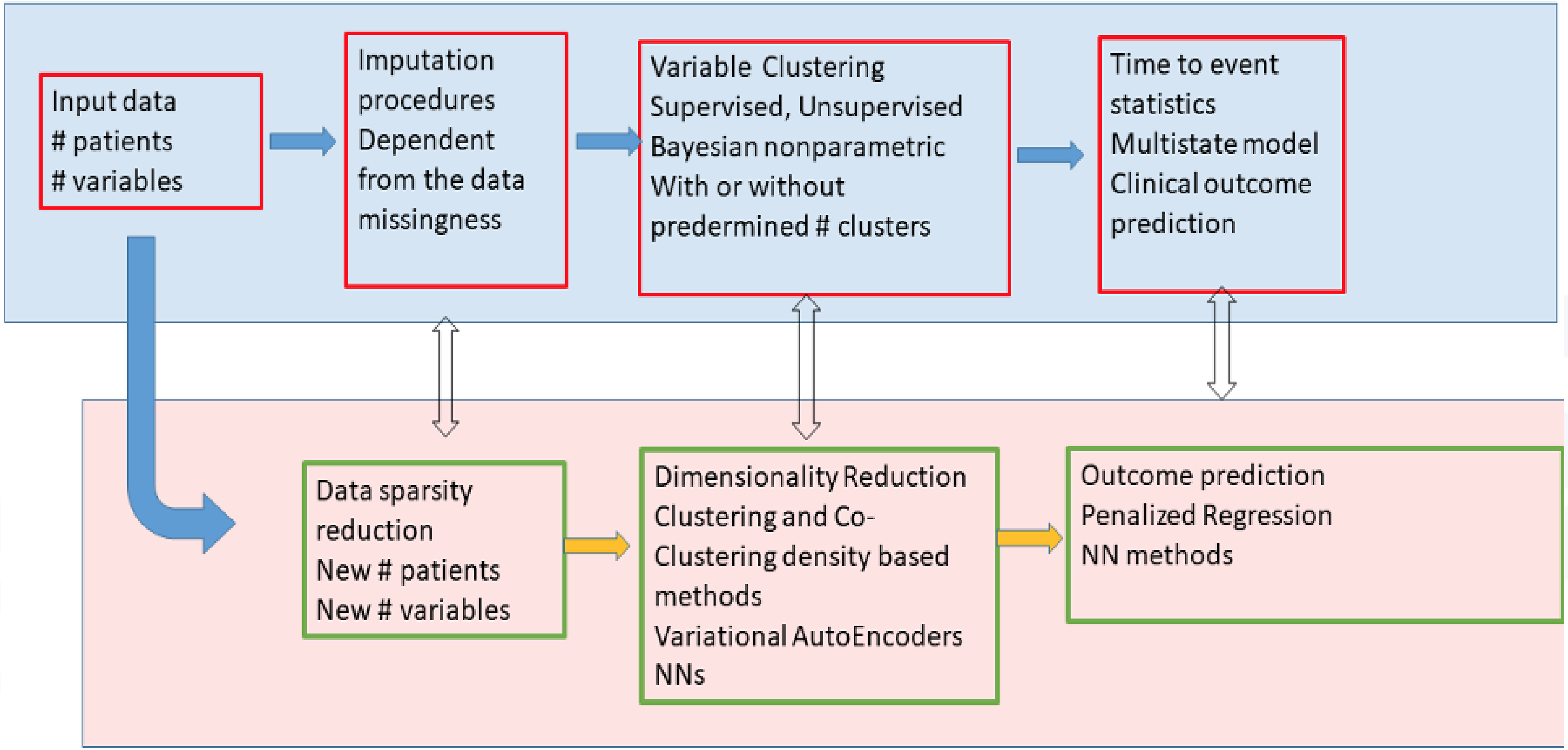
Driver Mutations (BN Causality)
Patient Stratification (HDP)
Survival analysis



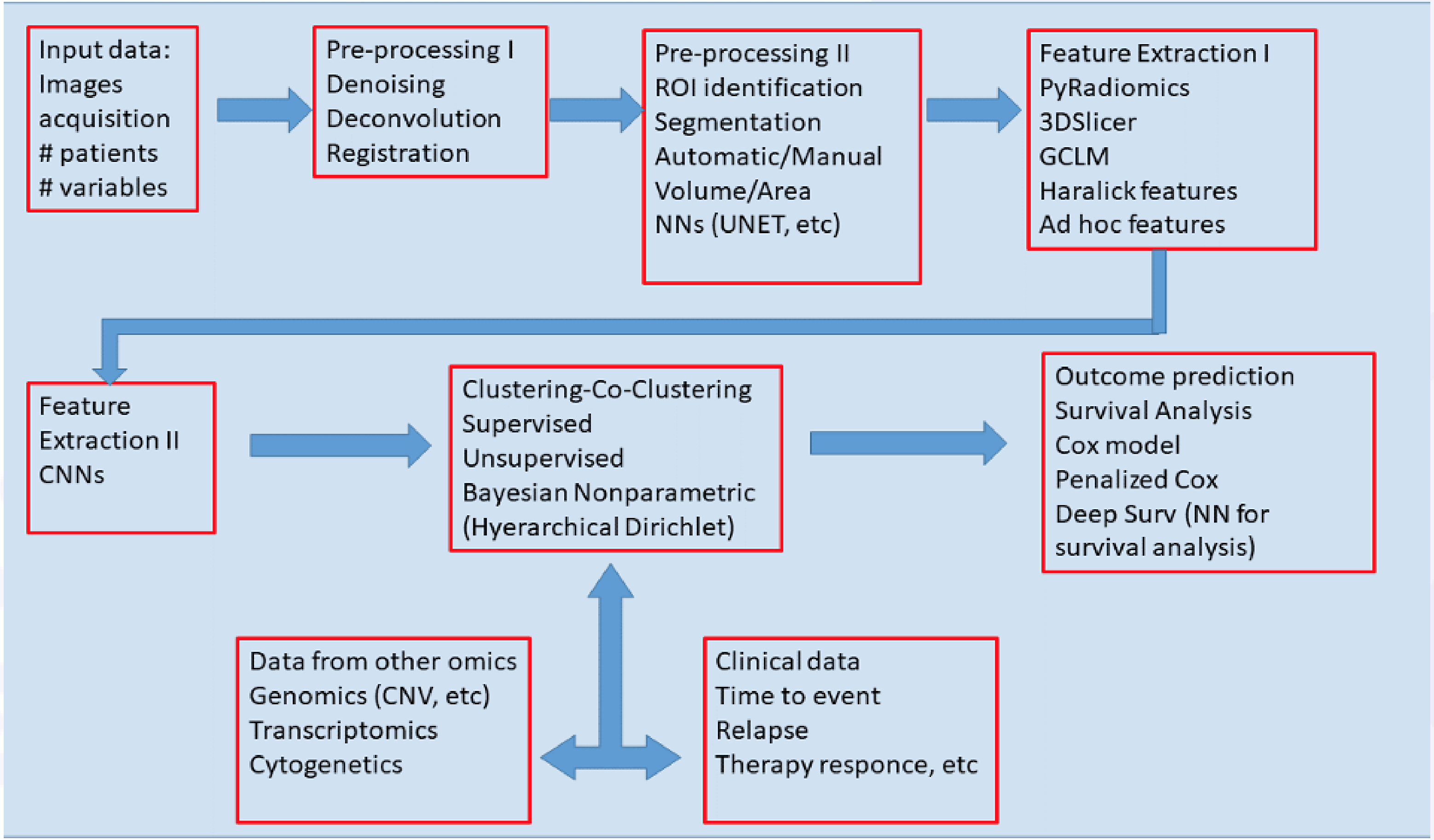
Radiomics



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

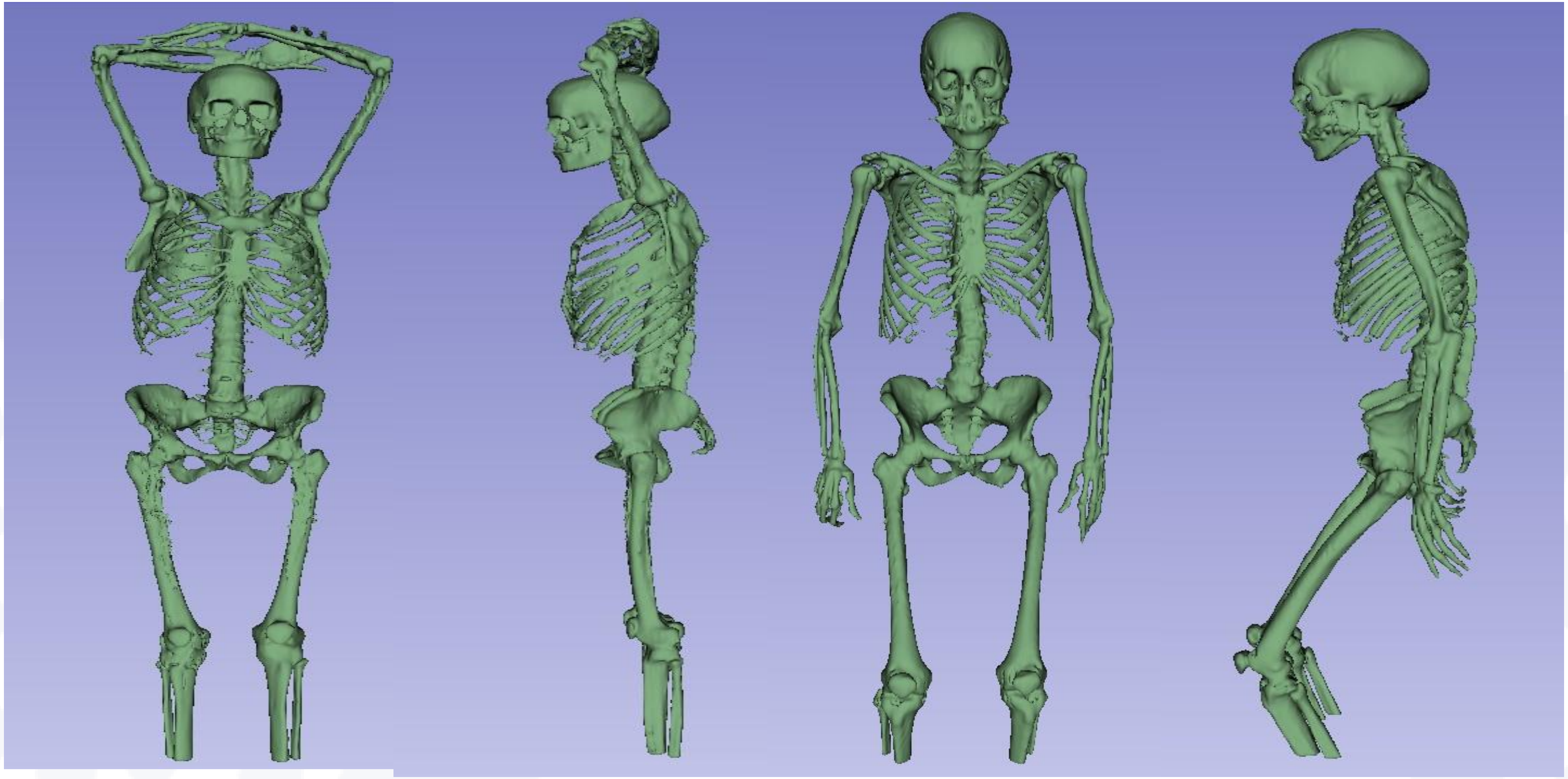


Pipeline for clinical and genomic data (MDS and MM). The choice between imputation data sparsity reduction depends from the mechanisms of missingness. After the imputation step, dimensionality reduction and/or data clustering are used to identify a putative patient stratification

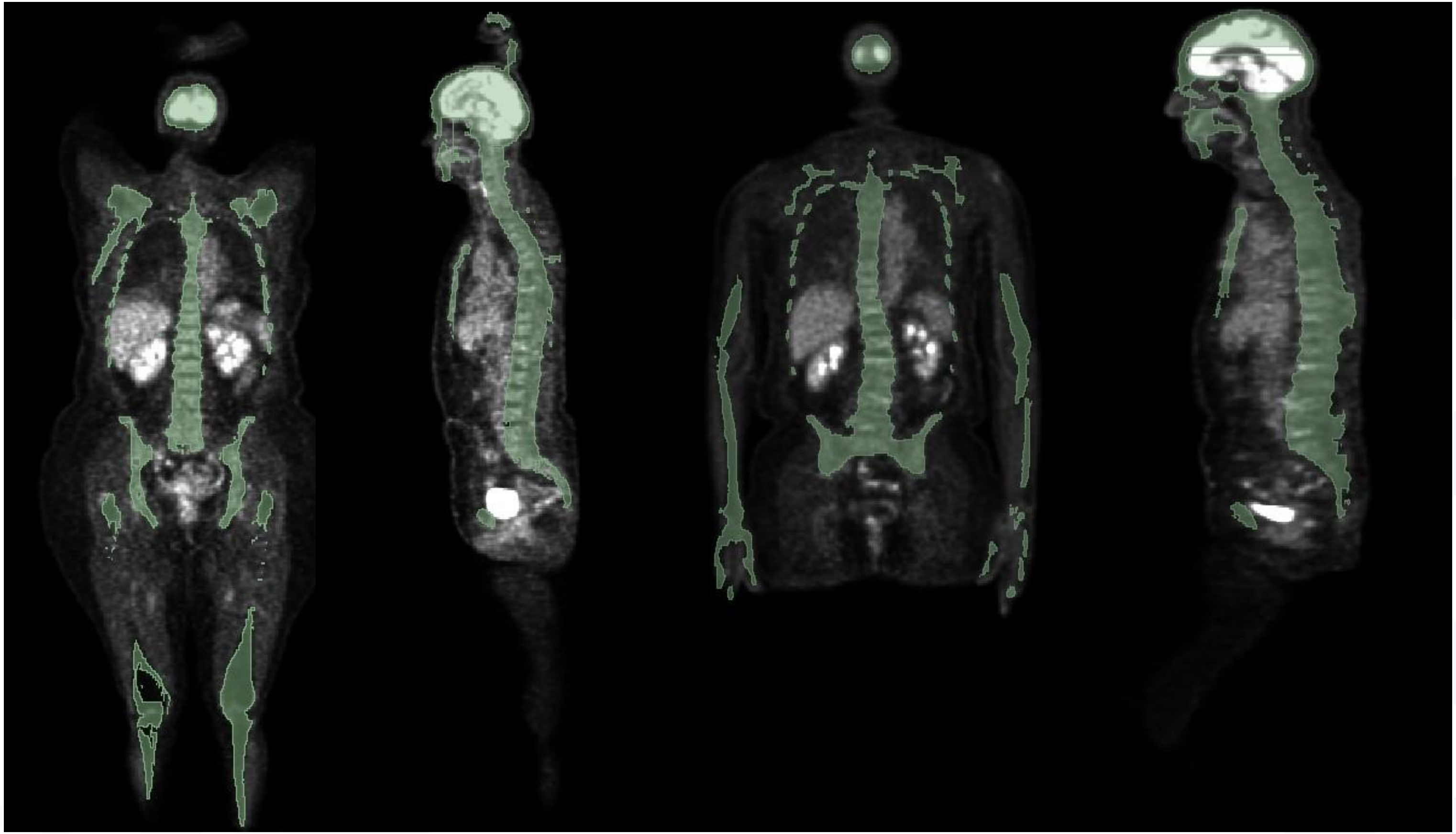


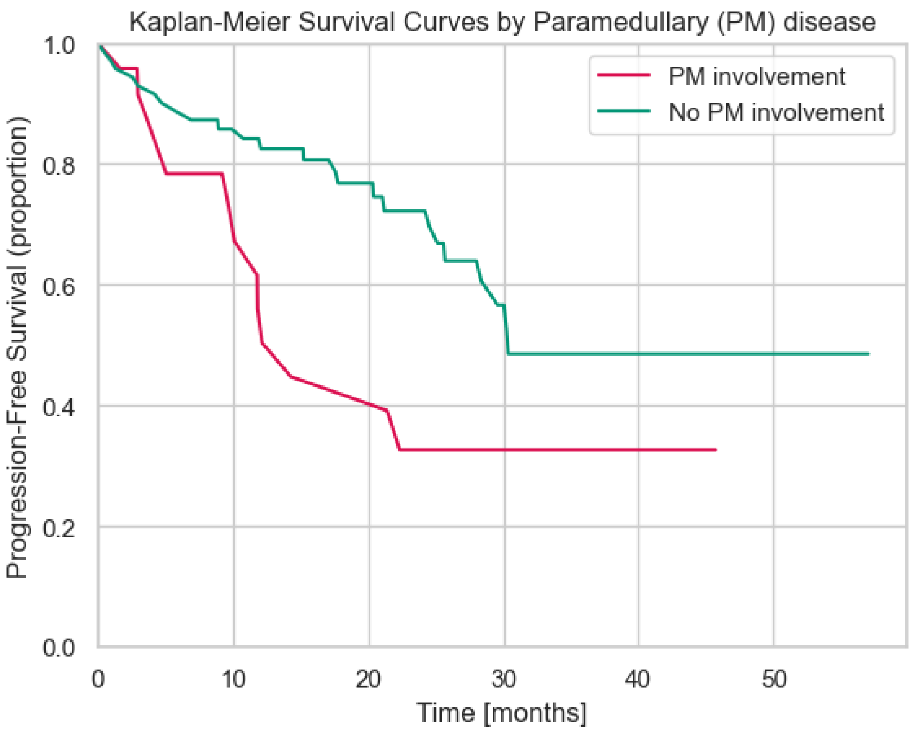
Pipeline for radiomic data (SCD, MM). Main steps for integrating imaging with clinical and genomic data

Two example of skeleton segmentation (CT)



CT segmentation supeimposed on PET

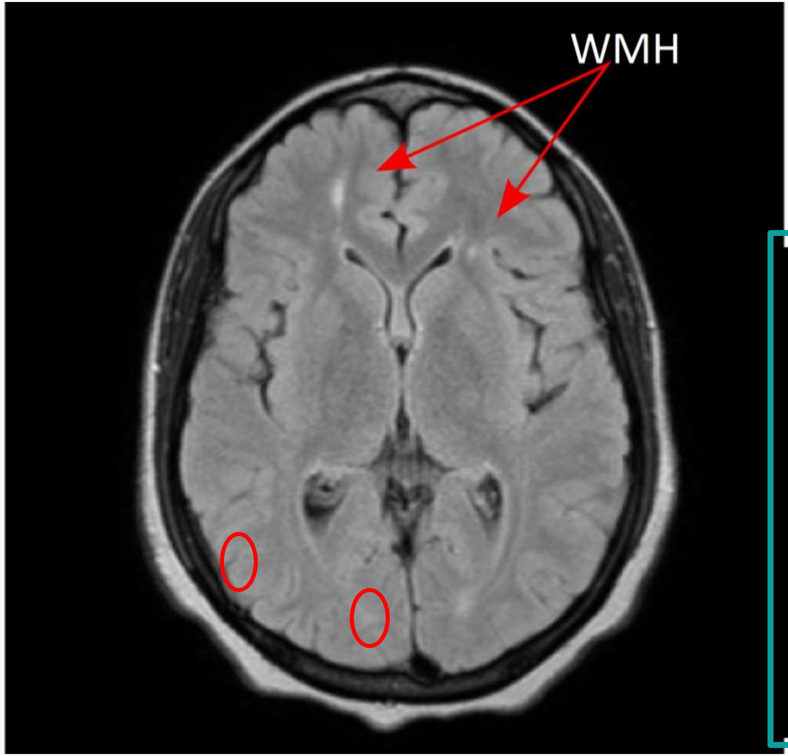




Kaplan-Meier curves for
PFS (patients stratified by
presence
of PM disease)

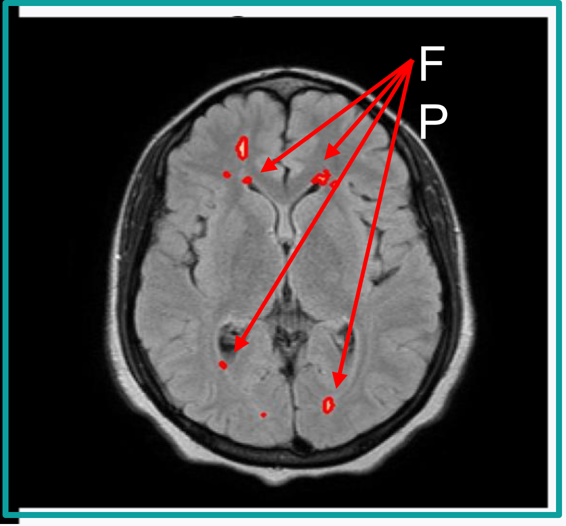
Segmentation of White Matter Hyperintensities (UNIBO and UNIPD)

(WMH on FLAIR images of patient affected by Sickle Cell Disease(SCD)



FLAIR image of a patient affected by SCD. We have highlight the WMH lesion areas

- Similar to the Multiple Sclerosis Lesions
- Lesions Can be very small
- Many tools for MS lesions segmentation
- Tool for MS can be adapted to SCD case



Pre-Trained UNet[1]:

- Winner of MICCAI WMH
- Segmentation Challenge
- Many False Positives(FP)

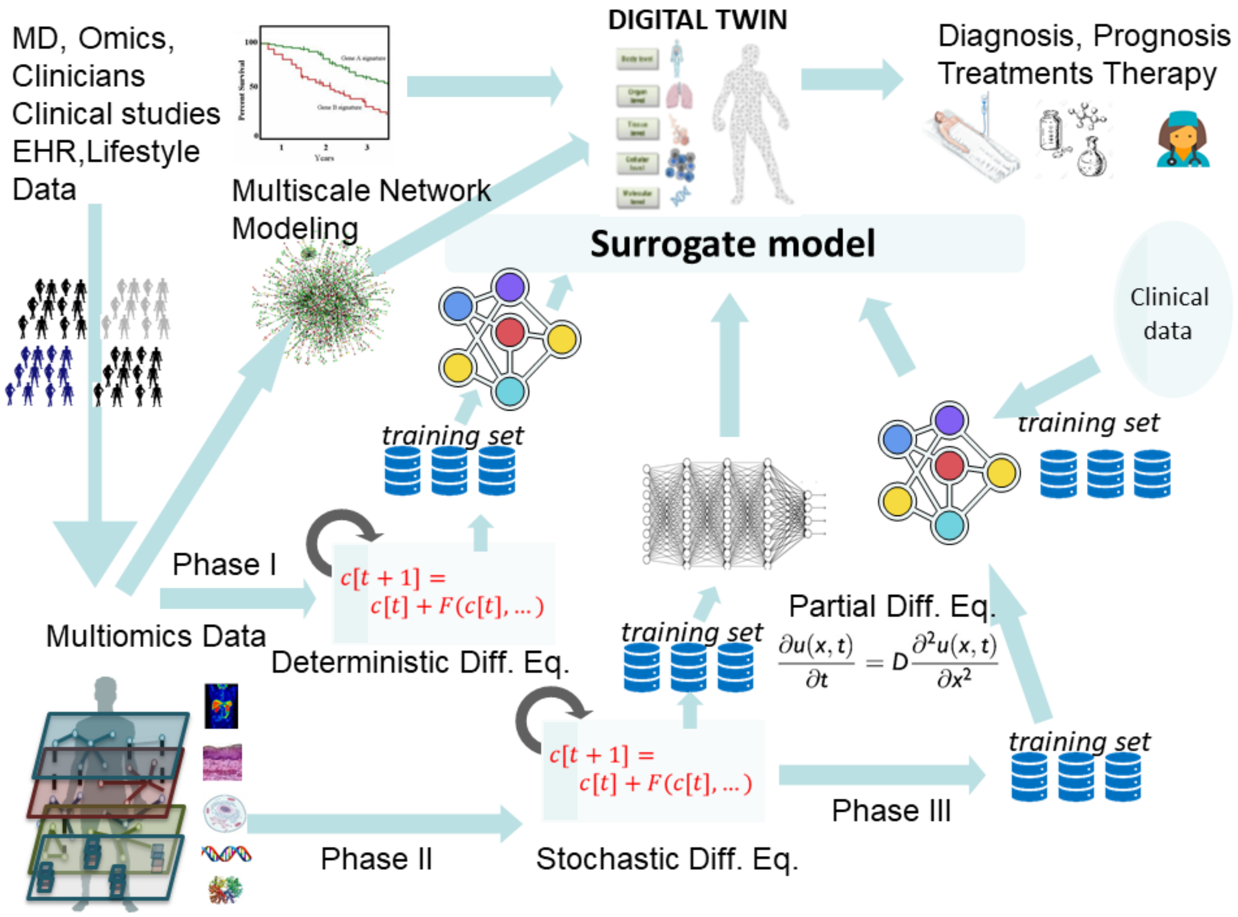
Adapt the Network:

- Removal of False Positives
- Lesion Area Enlargement

1 Fully Convolutional Network Ensembles for White Matter Hyperintensities Segmentation in MR Images, H.Li et al. NeuroImage, 2018

GENOMED4ALL

The long term vision



Thanks!

Any questions?

GenoMed4All & ERN-EuroBloodNet

Educational Program on Artificial Intelligence for public-at-large

Follow us!

genomed4all.eu

@genomed4all

/genomed4all

eurobloodnet.eu

@ERNEuroBloodNet

/ERNEuroBloodNet

Acknowledgements

European Reference Network
for rare or low prevalence complex diseases

Network Hematological Diseases (ERN EuroBloodNet)

Co-funded by the European Union

This project is supported by the European Reference Network on Rare Haematological Diseases (ERN-EuroBloodNet)-Project ID No 101085717. ERN-EuroBloodNet is partly co-funded by the European Union within the framework of the Fourth EU Health Programme.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

GENOMED4ALL

Funded by the European Union

GenoMed4All has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017549.