Machine Learning for Complex Biomedical Data

Andrea Cappozzo

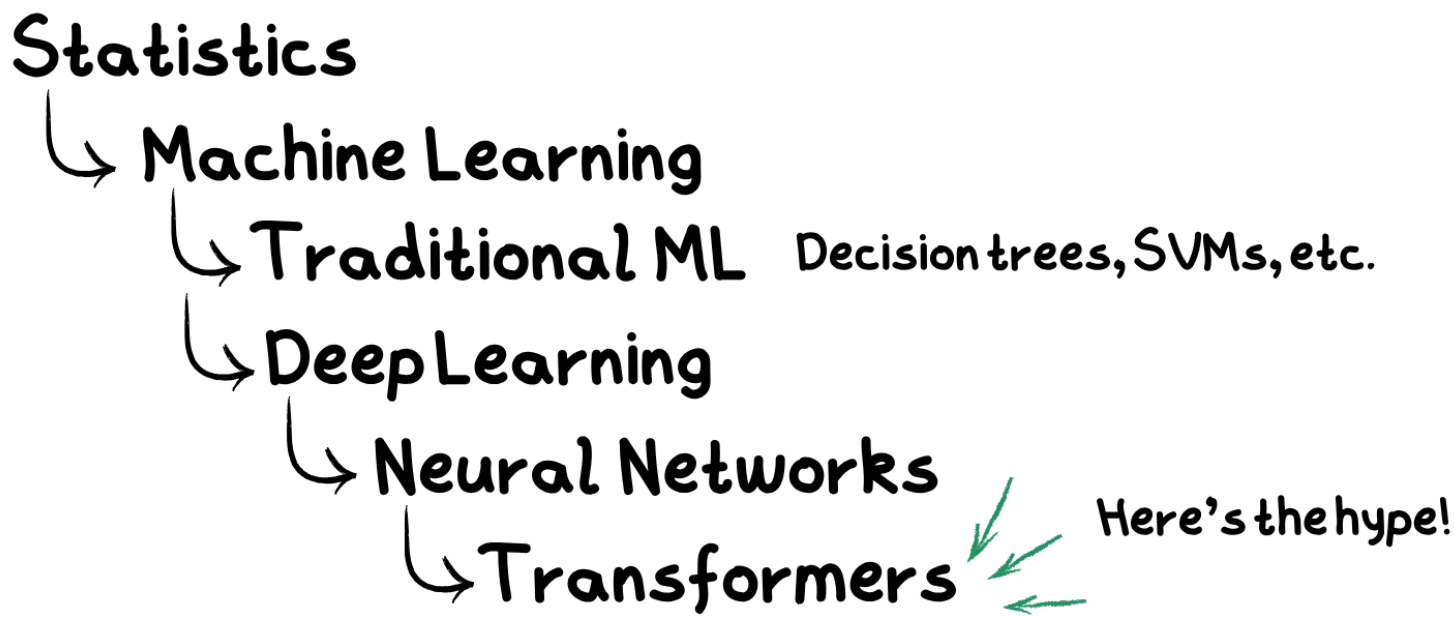
andrea.cappozzo@unicatt.it
@AndreaCappozzo
andreacappozzo.rbind.io

UNIVERSITÀ CATTOLICA del Sacro Cuore

Machine Learning Approaches for Hematological Disorder Diagnostics and Prognostics

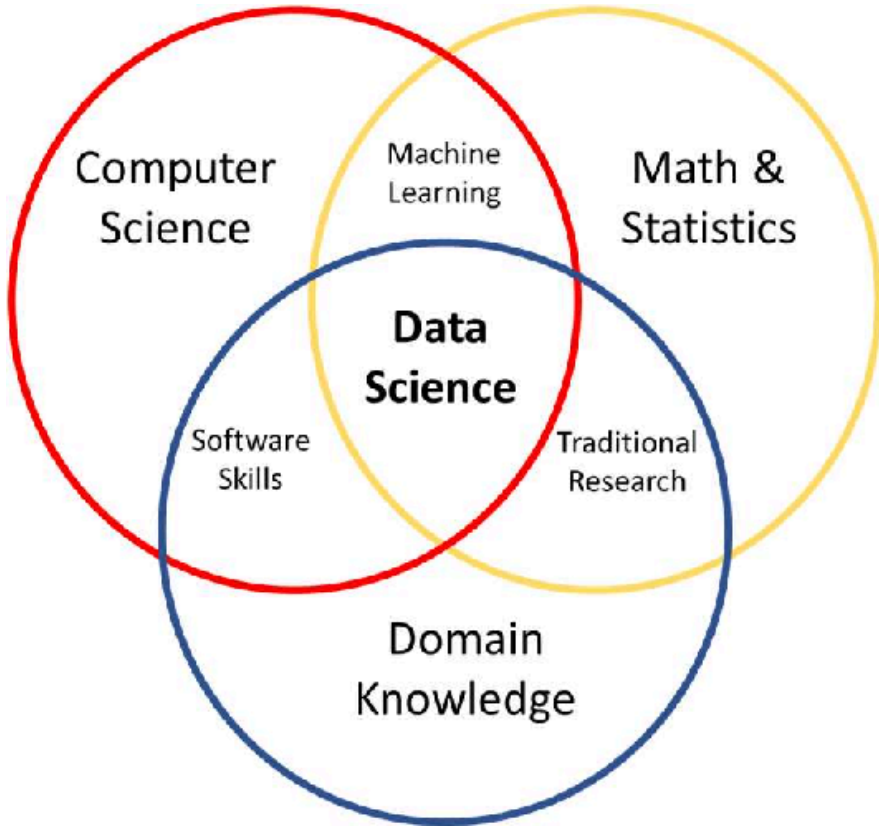
Introduction

It all starts with statistics

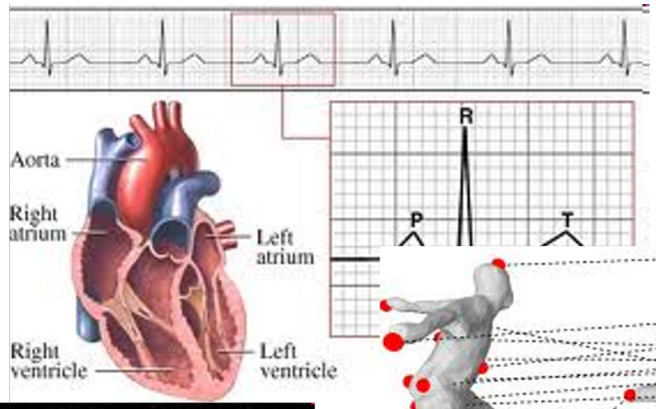


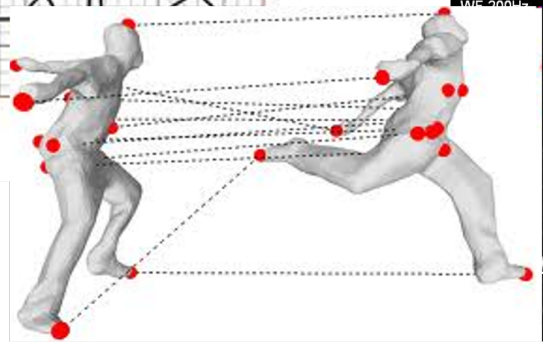
Source:
https://shiny.posit.co/blog/posts/shiny-side-of-llms-part-1/

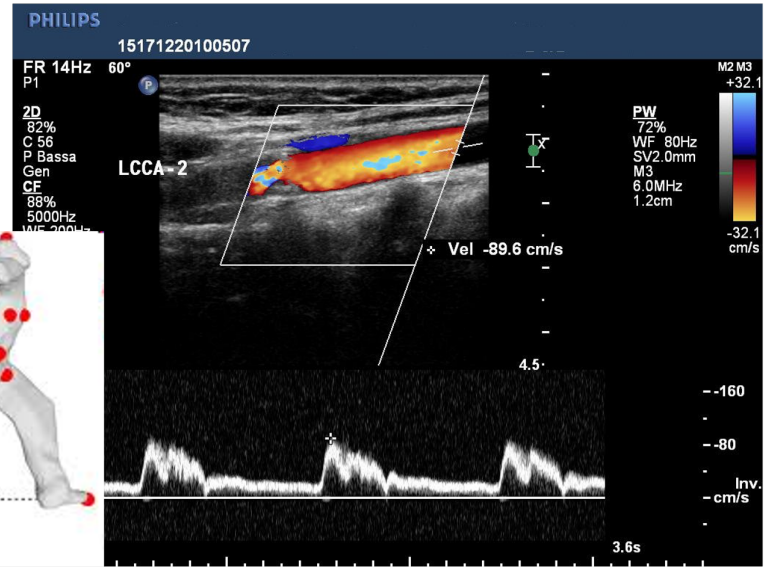
Interdisciplinary field

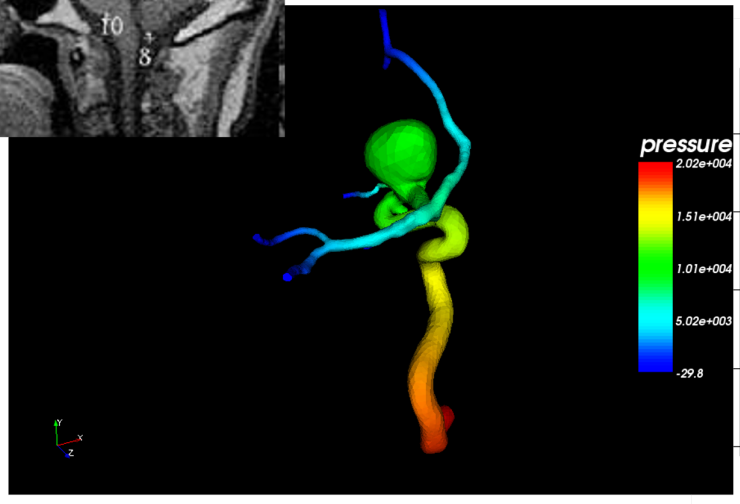


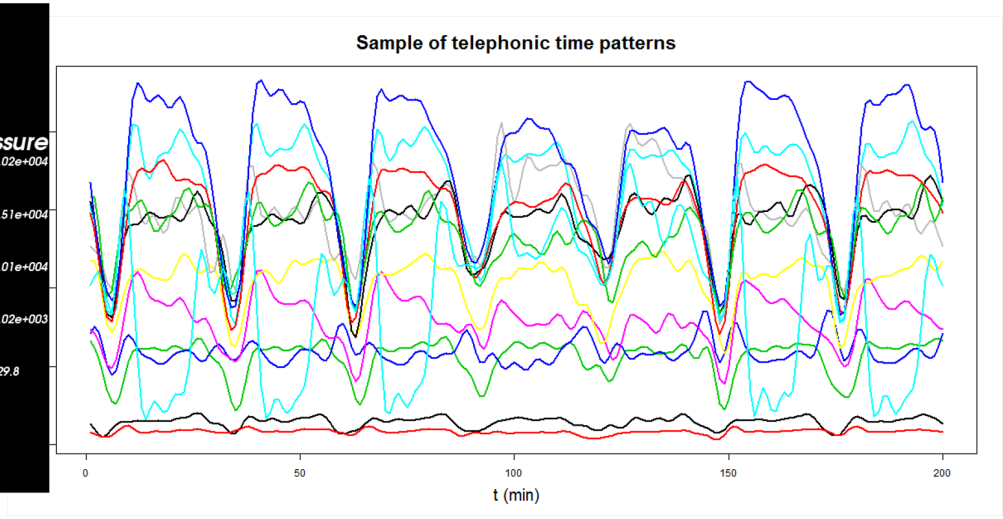
Big complex data



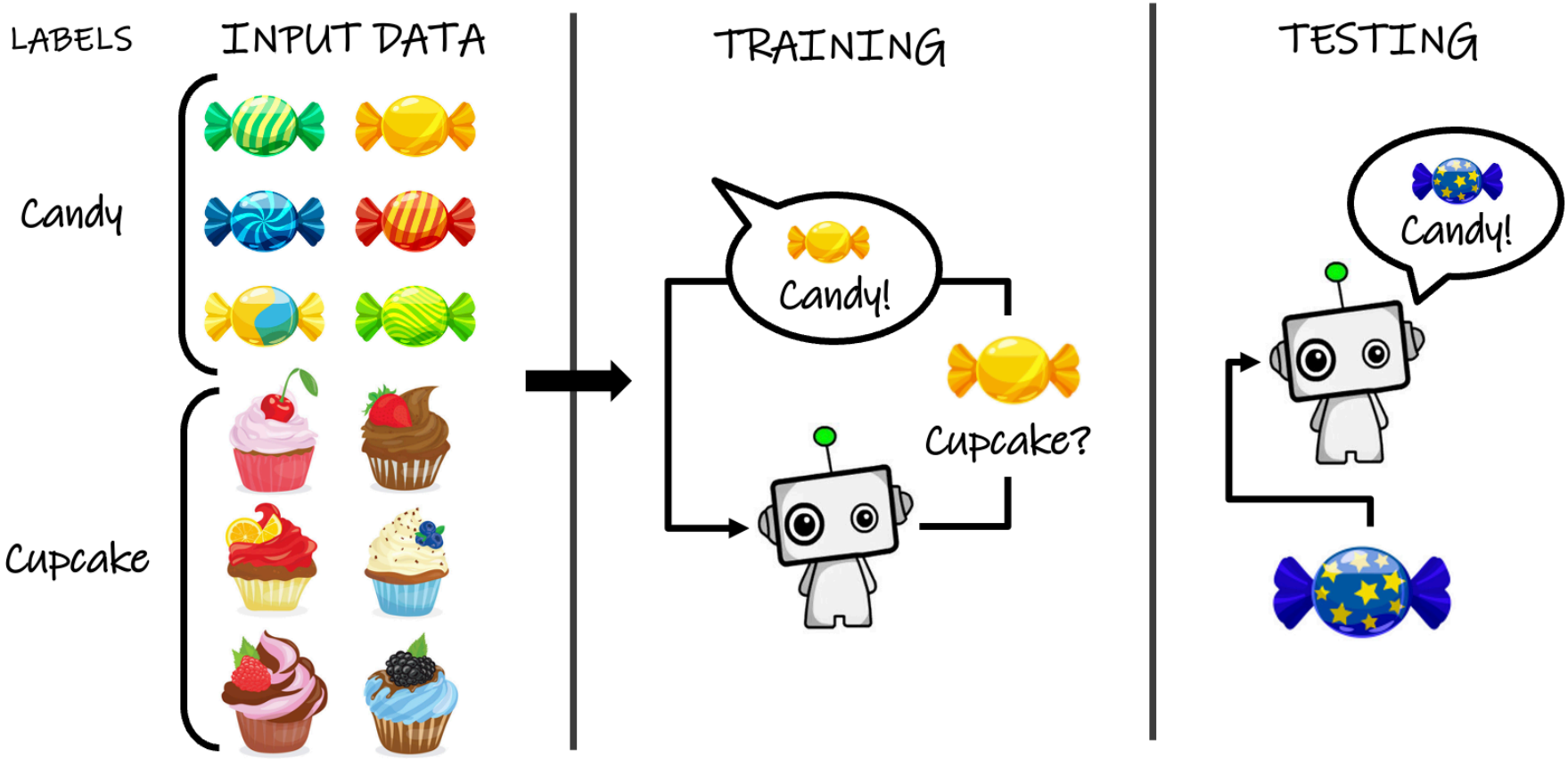








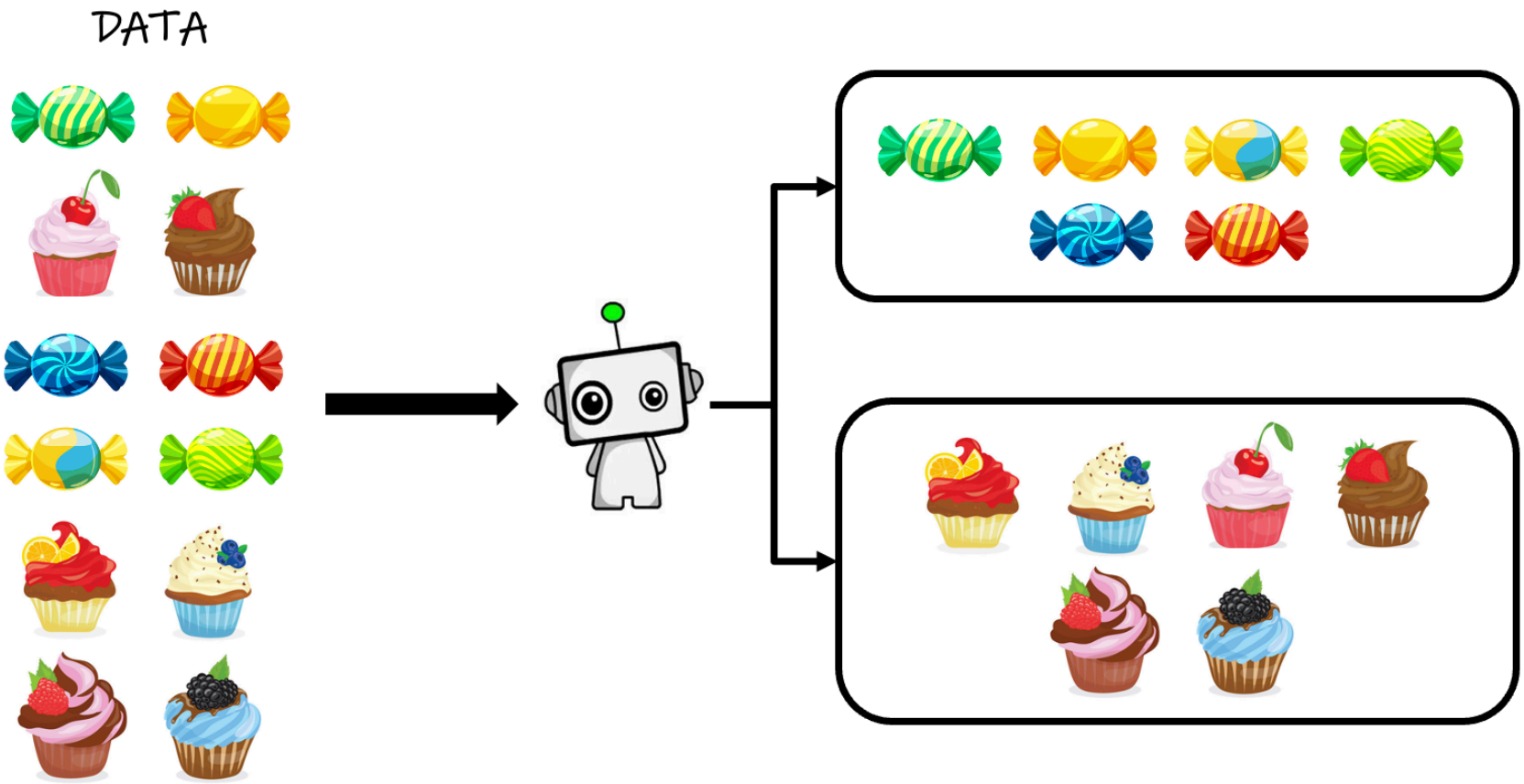
Supervised learning



Supervised learning

- The output variable is called **response variable** (in statistics also **dependent variable**).
- The input variables are called **predictors** (in statistics also **independent or explanatory variables**, computer scientists call them **features**).
- There is therefore a ”direction” of dependence from predictors to the response variable (or variables).

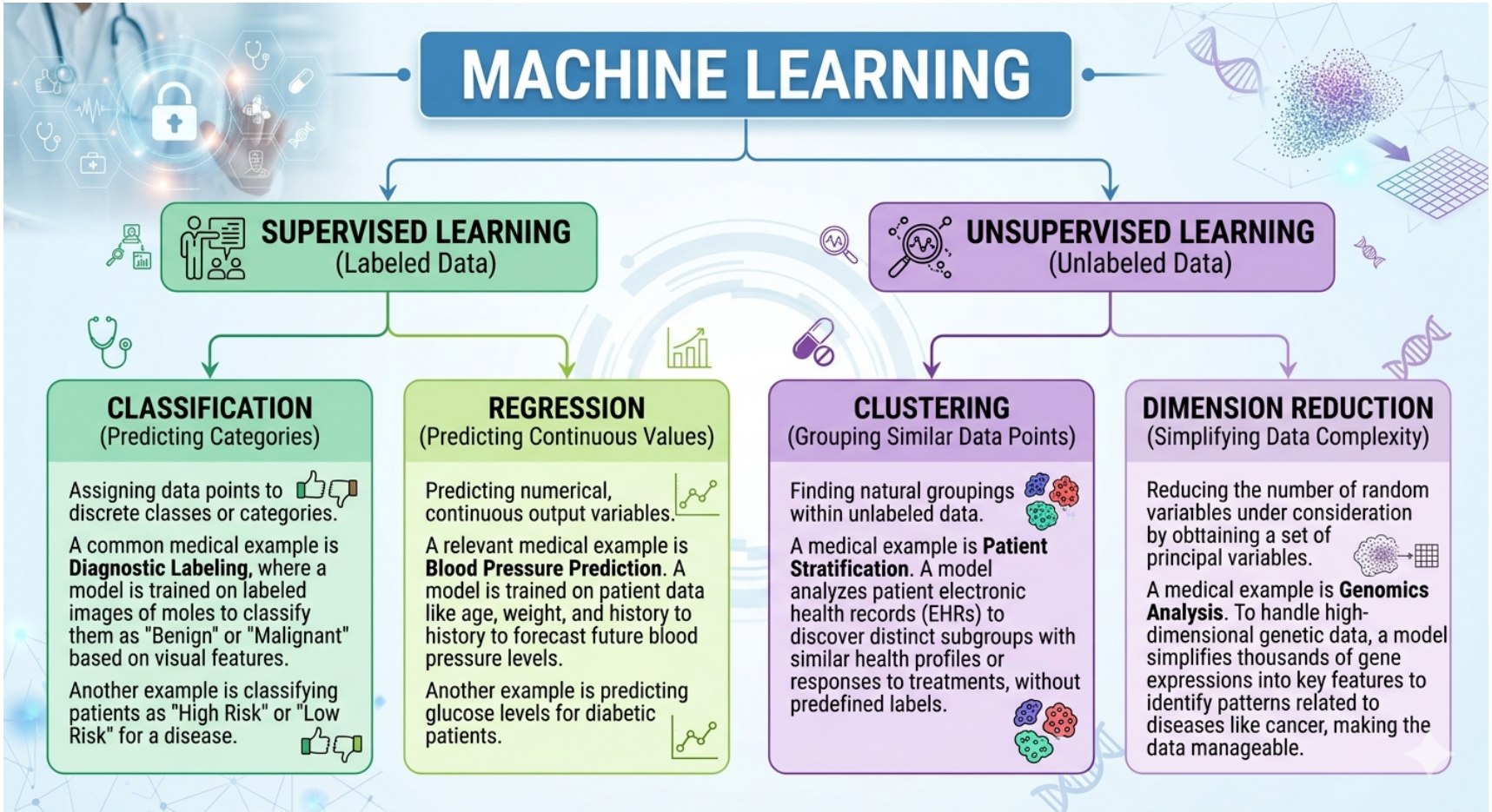
Unsupervised learning



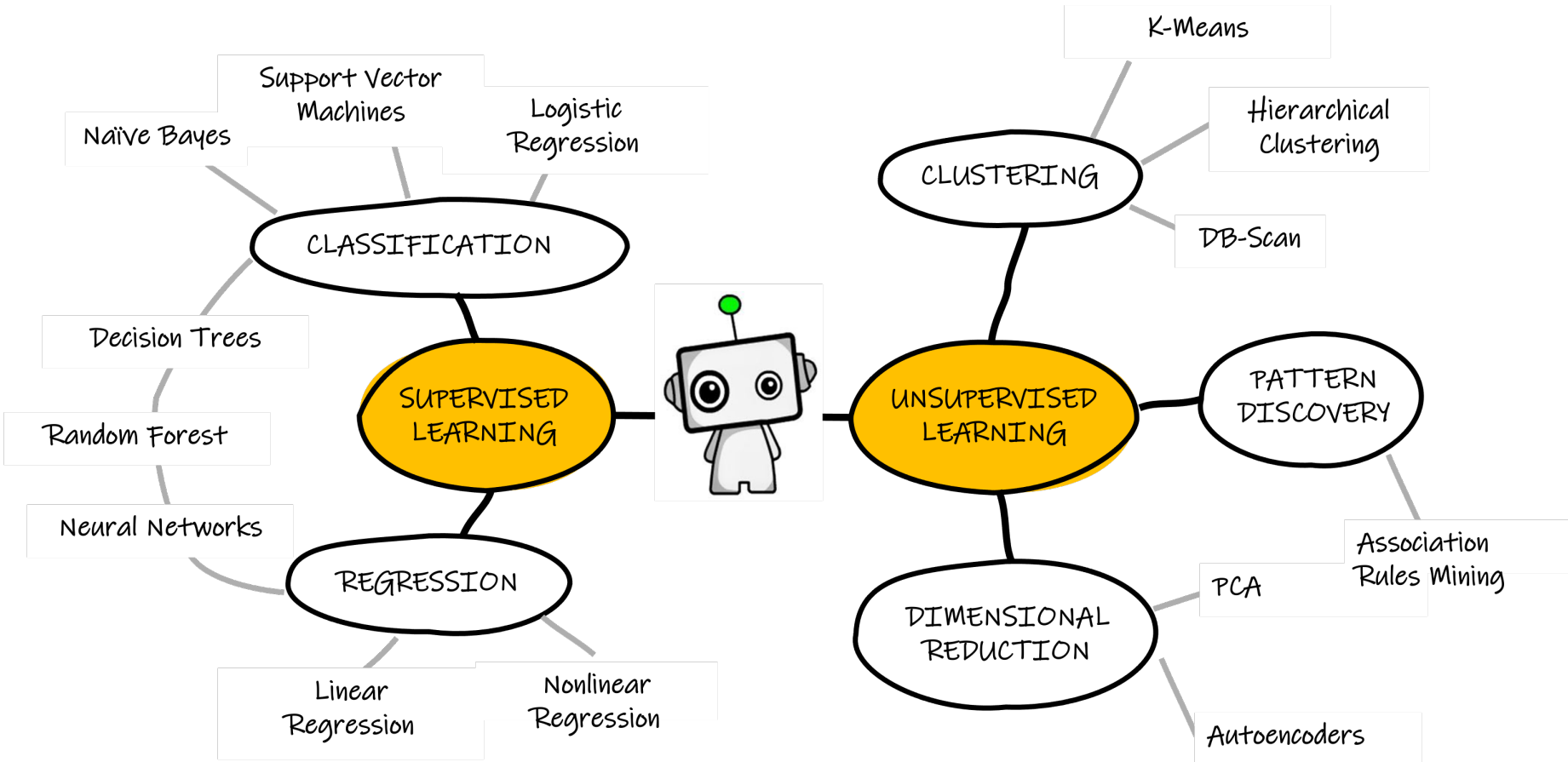
Unsupervised learning

- In unsupervised learning we do not focus on predicting a particular phenomenon, but seek to find interesting structures and relationships in the data.
- There is no response variable (target)
- The goal is to find *interesting patterns* in the data.
- So we try to find **similar groups or profiles** (clusters), or to **reduce the dimensionality** of the data.

Supervised and unsupervised learning



Supervised and unsupervised models: methods



Focus of the day

- A full tour of ML is impossible in 90 minutes
- Instead: a **curated set of modern approaches** especially relevant in medical research

Focus of the day

- A full tour of ML is impossible in 90 minutes
- Instead: a **curated set of modern approaches** especially relevant in medical research

Today's agenda:

- **Regression:** the core building block of supervised learning
  - **Hierarchical/multicentric data:** patients nested within hospitals
  - **High-dimensional data:** more predictors than we need
  - **Multivariate response:** predicting several outcomes at once (case study)
- All the aforementioned methods are particularly sensible in **small-data scenarios** (e.g., rare autoimmune neutropenia)

Linear regression

Predicting hemoglobin in CKD

A patient with **chronic kidney disease (CKD)** is admitted for anemia treatment.

- The kidney produces **erythropoietin (EPO)**, the hormone that drives red blood cell production — impaired kidneys make less of it.
- Hemoglobin (Hb, g/dL) is the key measure of anemia severity.
- Eight nephrology centres have collected 30 routine lab and clinical variables on 320 patients.

Predicting hemoglobin in CKD

A patient with **chronic kidney disease (CKD)** is admitted for anemia treatment.

- The kidney produces **erythropoietin (EPO)**, the hormone that drives red blood cell production — impaired kidneys make less of it.
- Hemoglobin (Hb, g/dL) is the key measure of anemia severity.
- Eight nephrology centres have collected 30 routine lab and clinical variables on 320 patients.

Goal:
Build a prediction model for Hb so that clinicians can anticipate the need for Erythropoiesis-stimulating agent (EDA) dose adjustment **before** the next blood test.

Linear regression: the first building block

Core idea: express the response as a **weighted sum** of predictors:

$$\hat{\text{Hb}}_i = \sum_j \hat{\beta}_j x_j$$

- Each predictor j gets a **coefficient** $\hat{\beta}_j$: positive if it raises Hb, negative if it lowers it.
- The model chooses weights to minimise the total squared error

$$\text{minimise} \quad \sum_i (\text{Hb}_i - \hat{\text{Hb}}_i)^2$$

- Result: **estimated coefficients** we can use to predict Hb for any new patient.

$$\widehat{\text{Hb}} = \hat{\beta}_0 + \hat{\beta}_1 \,\text{eGFR} + \hat{\beta}_2 \,\text{Ferritin} + \cdots$$

Linear regression: results on the CKD dataset

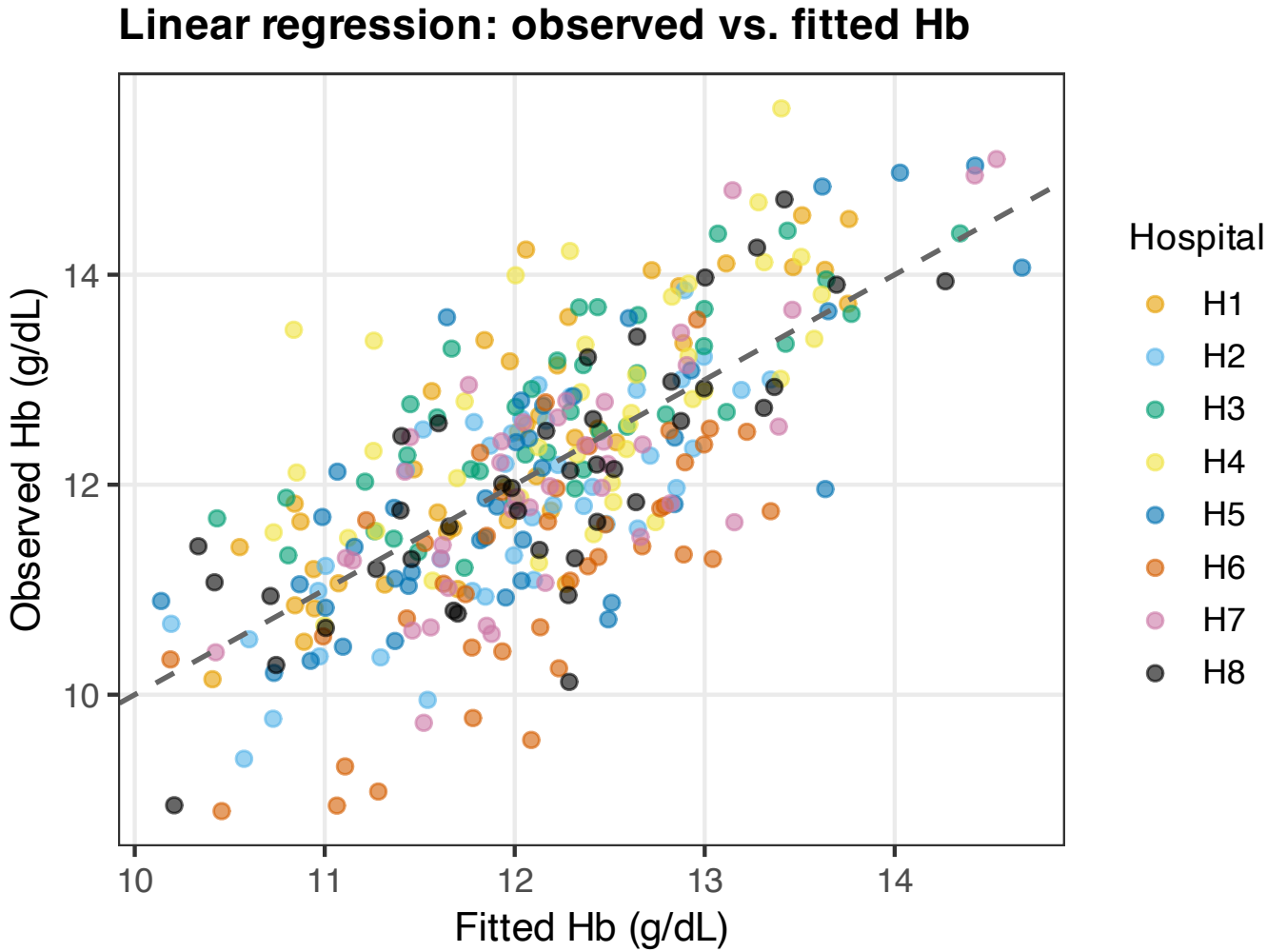
Fitted on 320 patients, 30 predictors.

Predictor	Coefficient	Direction
eGFR (mL/min/1.73m^2)	$+0.040$	↑ kidney fn. ⇒ ↑ Hb
Ferritin (µg/L)	$+0.005$	↑ iron stores
TSAT (%)	$+0.060$	↑ iron saturation
CRP (mg/L)	-0.040	↑ inflammation ⇒ ↓ Hb
EPO dose (IU/wk)	-0.080^*	dose given to sicker pts
Age (years)	-0.020	ageing reduces marrow

- $R^2 = 0.505$: predictors explain 50% of Hb variability.
- RMSE $= 0.86$ g/dL: typical prediction error $\approx \pm 1$ g/dL.
- Problem spotted: residuals cluster by hospital ⇒ *next slide*.

* Negative because higher doses are prescribed precisely to sicker, lower-Hb patients.

Linear regression: observed vs. fitted Hb

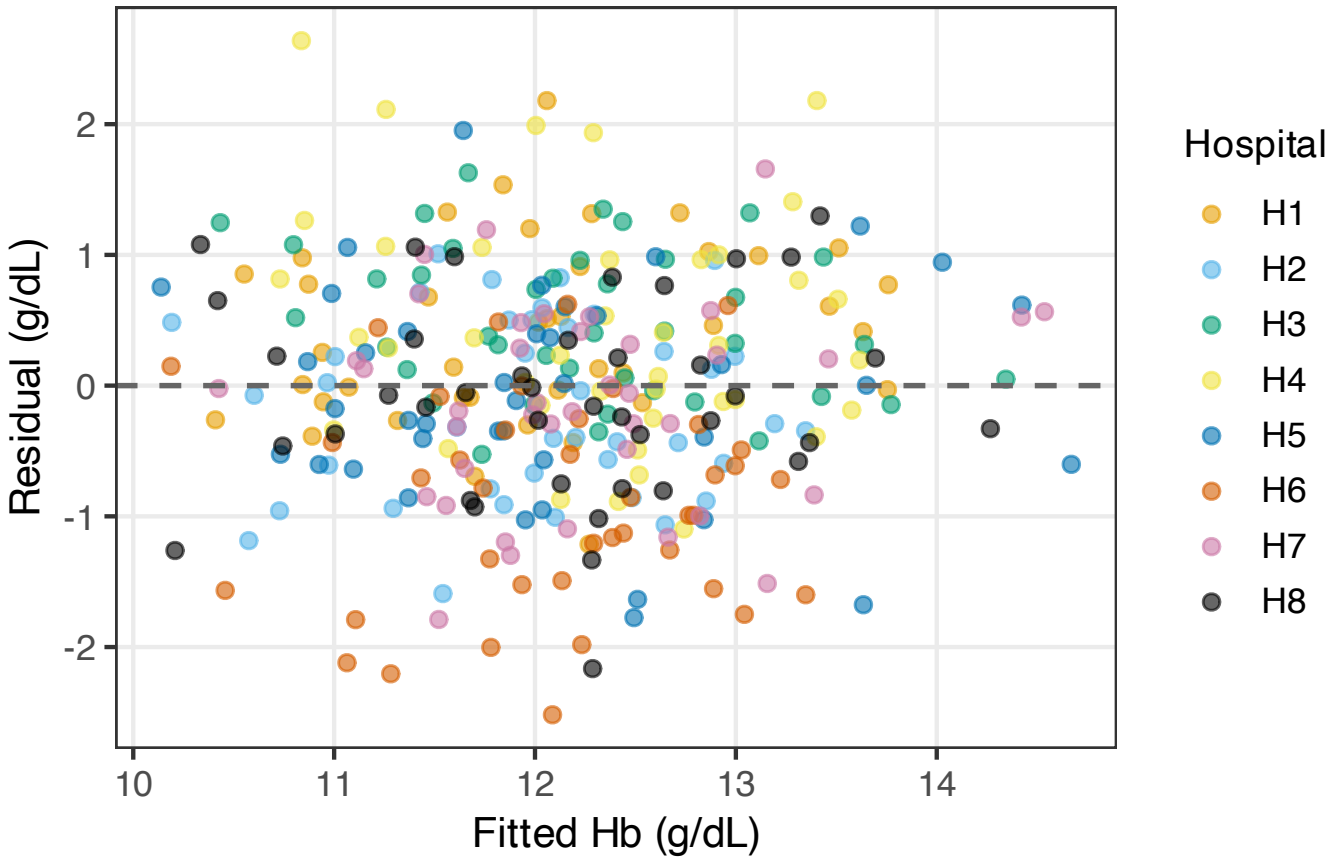


Each colour is a recruiting hospital; points on the dashed line would be perfect predictions.

Linear regression: a hidden problem

Linear regression: residuals by hospital

Systematic offsets reveal unmodelled hospital-level variation



Residuals are colour-coded by hospital. Systematic **positive or negative offsets by centre** reveal that the plain regression *ignores the hierarchical structure* of the data.

Handling complexity: multicentric data

Multicentric data: why hospitals differ

The problem:

In a multi-centre study, patients from the same hospital are *more similar to each other* than to patients at other sites.

- Centres recruit different CKD stages, use different protocols, and have lab calibration differences.
- Ignoring this leads to **over-confident** coefficient estimates and wrong standard errors.

Solution: give each hospital its own **random intercept**: a centre-specific Hb shift estimated from the data itself.

This approach is called a **linear mixed-effects model**: "mixed" because it combines *fixed* effects (shared coefficients for all patients) with *random* effects (hospital-specific adjustments).

Mixed-effects model: how it works

- **Fixed effects:** same coefficients for eGFR, ferritin, etc. for every patient (the biologically meaningful part).
- **Random intercept:** each hospital gets an extra term u_j (positive = centre tends to have higher Hb baseline, negative = lower).
- The model *learns* how much hospitals vary by estimating the spread of these shifts.

$$\widehat{\text{Hb}}_{ij} = \underbrace{\beta_0 + \beta_1\,\text{eGFR}_{ij} + \cdots}_{\text{same for all patients}} + \underbrace{u_j}_{\text{hospital } j\text{'s offset}}$$

- **ICC** (intraclass correlation): the fraction of total Hb variability *explained by hospital membership* — if high, ignoring centres is dangerous.

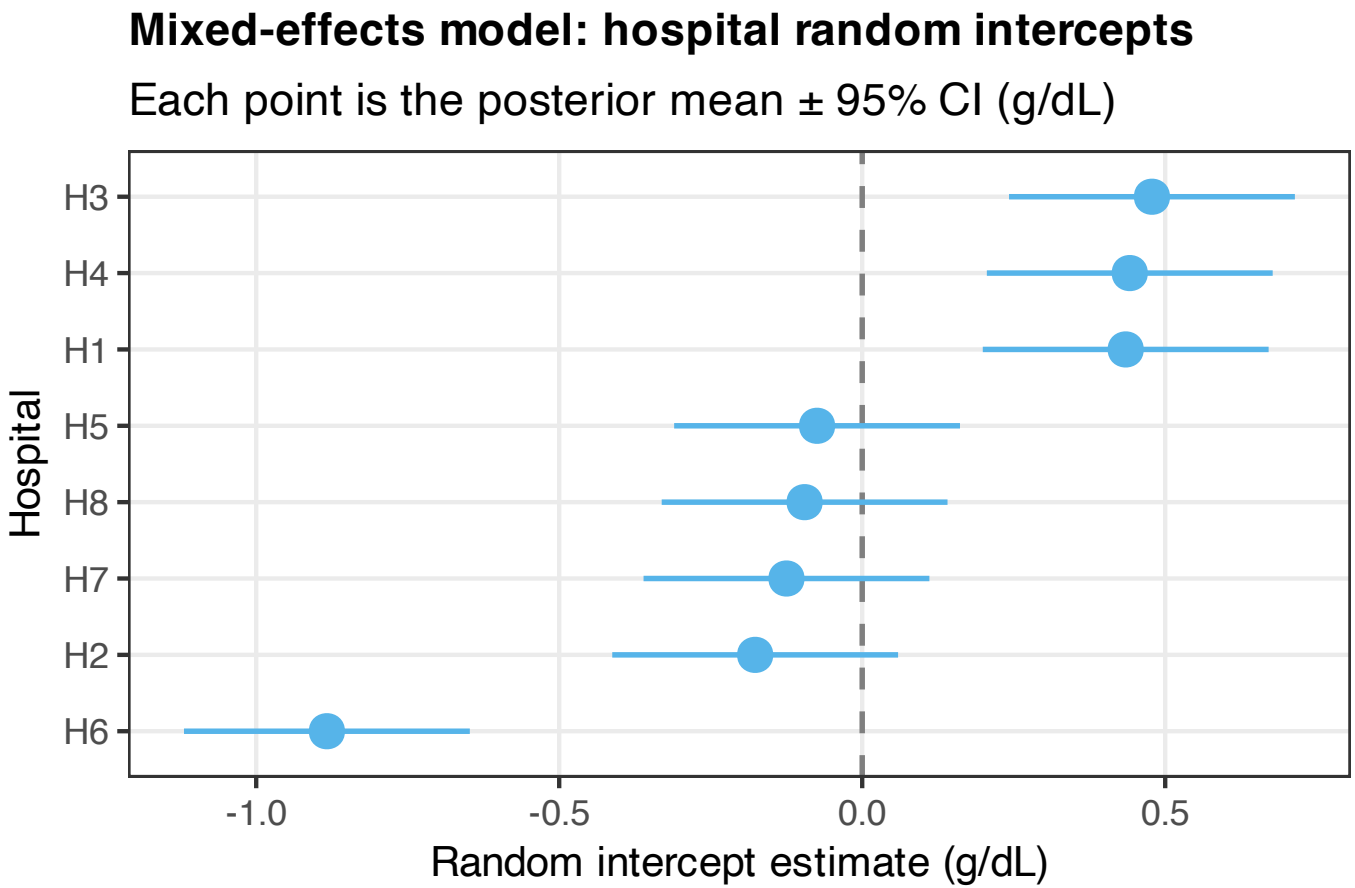
Mixed-effects model: results

Fitted on the same 320 CKD patients.

- **RMSE** drops from 0.86 to **0.74 g/dL**: hospital structure was masking real signal.
- **Random intercept SD** $= 0.47$ g/dL: hospitals differ on average by ± 0.5 g/dL in their typical patient's Hb — clinically meaningful.
- **ICC** $= 0.26$: **26%** of total Hb variation is attributable to hospital membership, *not* to the measured predictors.
- Fixed-effect coefficients for eGFR, Ferritin, TSAT, CRP, EPO dose and Age remain stable in direction but are more precisely estimated.

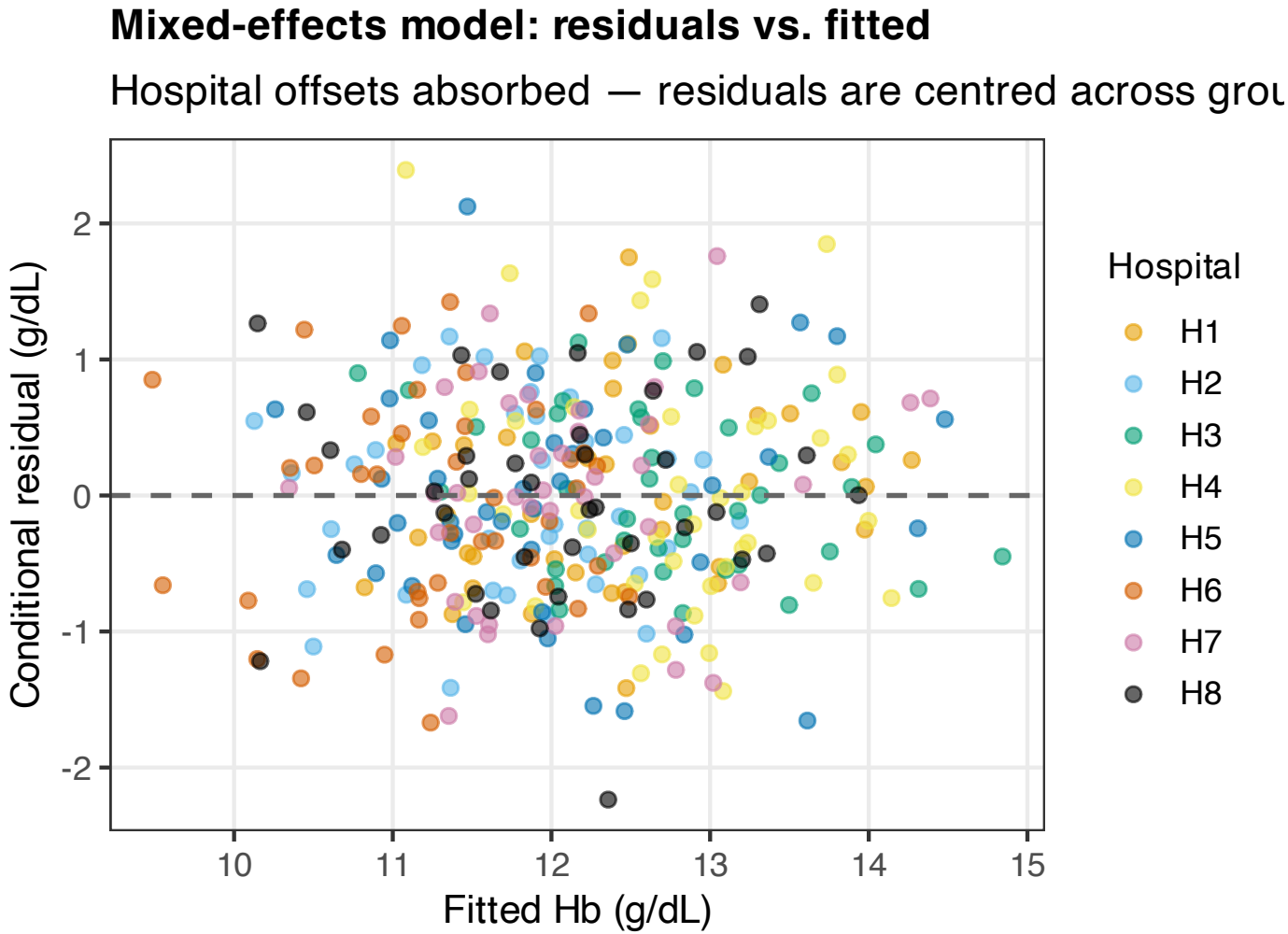
Take-home: ignoring the multilevel design inflates residual variance and hides how much of the unexplained variation is a *centre effect*, not patient-level noise.

Mixed-effects model: hospital random intercepts



Each point shows a hospital's estimated Hb offset ± 95% confidence interval. Centres above zero tend to have higher Hb patients; those below are lower.

Mixed-effects model residuals



After absorbing the hospital random intercept the per-centre offsets disappear — residuals are now centred around zero for all groups.

Handling complexity: high-dimensional data

The curse of too many predictors

The problem:

Our CKD dataset has **30 lab and clinical variables**. Do all of them genuinely predict Hb?

- Routine clinical panels measure everything available — not everything is useful.
- With many correlated predictors, ordinary regression *overfits*: it chases noise in the training data and performs poorly on new patients.
- We need a method that **automatically selects** the relevant variables.

Solution: LASSO (*Least Absolute Shrinkage and Selection Operator*): a penalized regression that **shrinks small, irrelevant coefficients exactly to zero**, leaving only the predictors that earn their place.

LASSO: how the penalty works

- Ordinary regression minimises prediction error.
- LASSO adds a **penalty** on the sum of absolute coefficient sizes:

$$\text{minimise} \quad \underbrace{\sum_i (\mathrm{Hb}_i - \hat{\mathrm{Hb}}_i)^2}_{\text{prediction error}} + \lambda \underbrace{\sum_j |\beta_j|}_{\text{penalty}}$$

- $\lambda > 0$ controls the strength of the penalty: large $\lambda \Rightarrow$ more coefficients shrunk to zero.
- **Cross-validation**: data are split into 10 folds; the λ that minimises out-of-fold error is chosen.
- The absolute-value penalty produces **exact zeros**: an automatic variable selector.

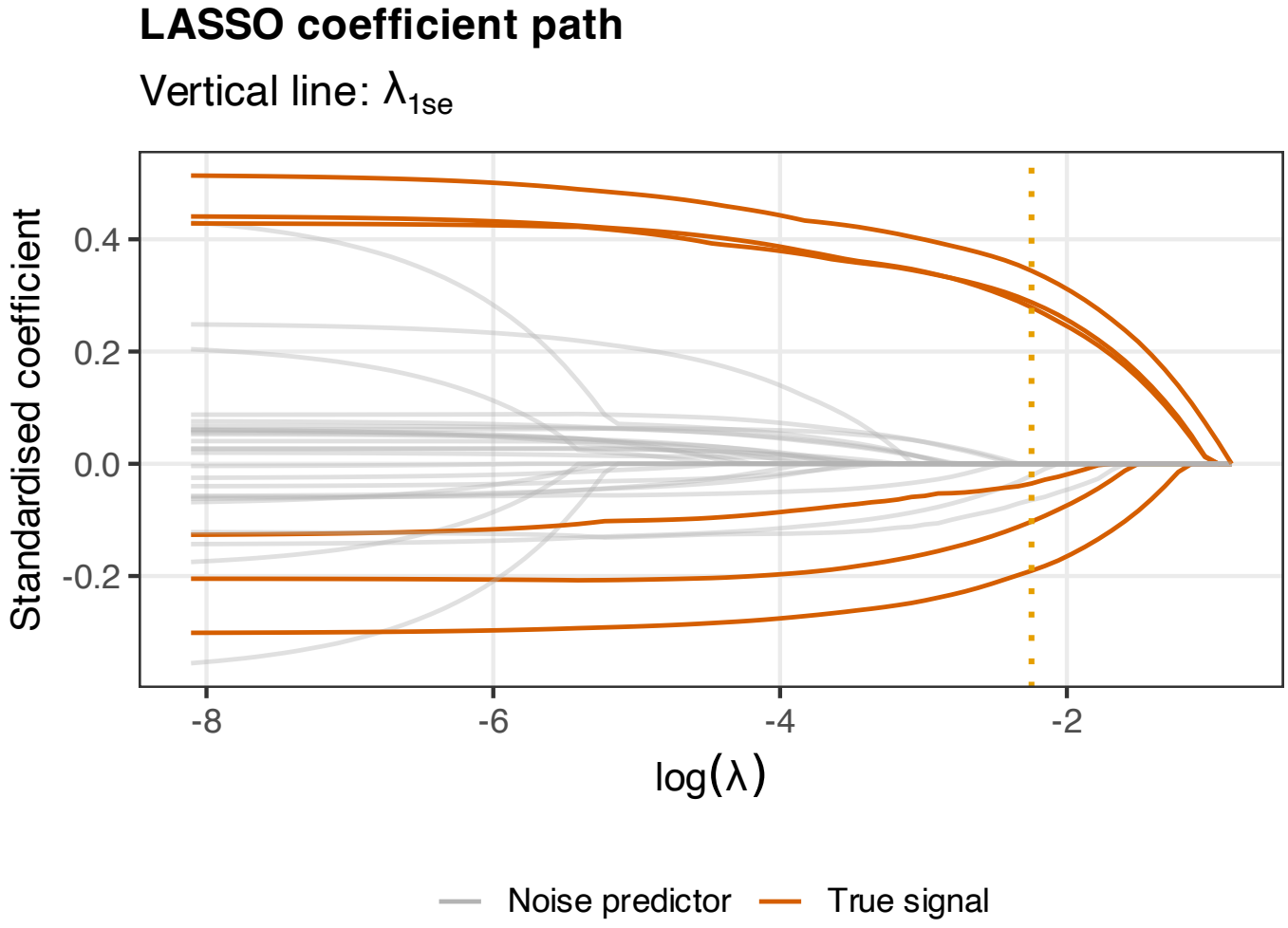
LASSO: results on the CKD dataset

8 predictors selected from 30 (26 set to exactly zero).

Variable	Clinical role	Selected?
eGFR	Kidney function	✓ True signal
Ferritin	Iron stores	✓ True signal
TSAT	Iron saturation	✓ True signal
EPO dose	ESA therapy	✓ True signal
CRP	Inflammation	✓ True signal
Age	Demographics	✓ True signal
Albumin	Nutrition	○ False positive
Reticulocytes	Correlated with EPO	○ False positive

- **Sensitivity:** 6/6 truly relevant predictors recovered (100%).
- **2 false positives** — both biologically plausible, correlated with true signals.

LASSO: coefficient shrinkage path



As λ increases (left to right), coefficients are shrunk toward zero.
Coloured lines are the six true predictors: they are the last to vanish.

Case study

Incredible coauthors

Francesca Ieva

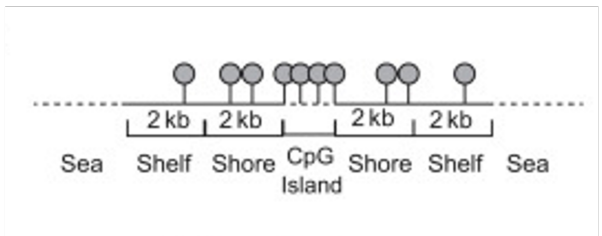
Politecnico di Milano
Human Technopole

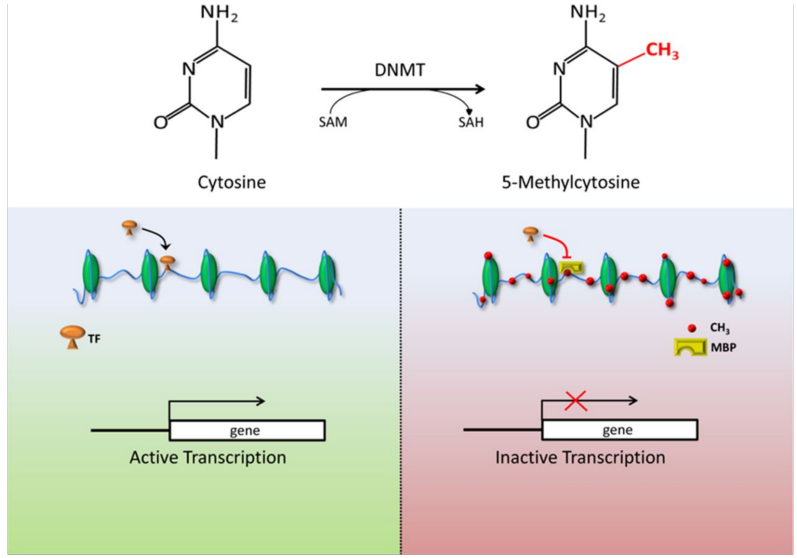
Giovanni Fiorito

Istituto Giannina Gaslini

Motivating problem: DNA methylation

- **DNA methylation (DNAm):** a methyl group is added to the DNA molecule converting cytosine to 5-methylcytosine.





- At each **CpG site**, DNA methylation is measured as a value in [0,1], indicating the proportion of methylated cells

Motivating problem: DNAm surrogates creation

- By regressing risk factors on methylation levels within CpG sites we obtain **DNAm surrogates of exposure**

- There are considerable advantages in employing DNAm surrogates for achieving epidemiological evidence
  - DNAm accounts for biases in self-reported (or measured) exposure (Zhang et al. 2016)
  - DNAm accounts for individual responses to risk factors (Conole et al. 2021)

Motivating problem: DNAm surrogates creation

- By regressing risk factors on methylation levels within CpG sites we obtain **DNAm surrogates of exposure**

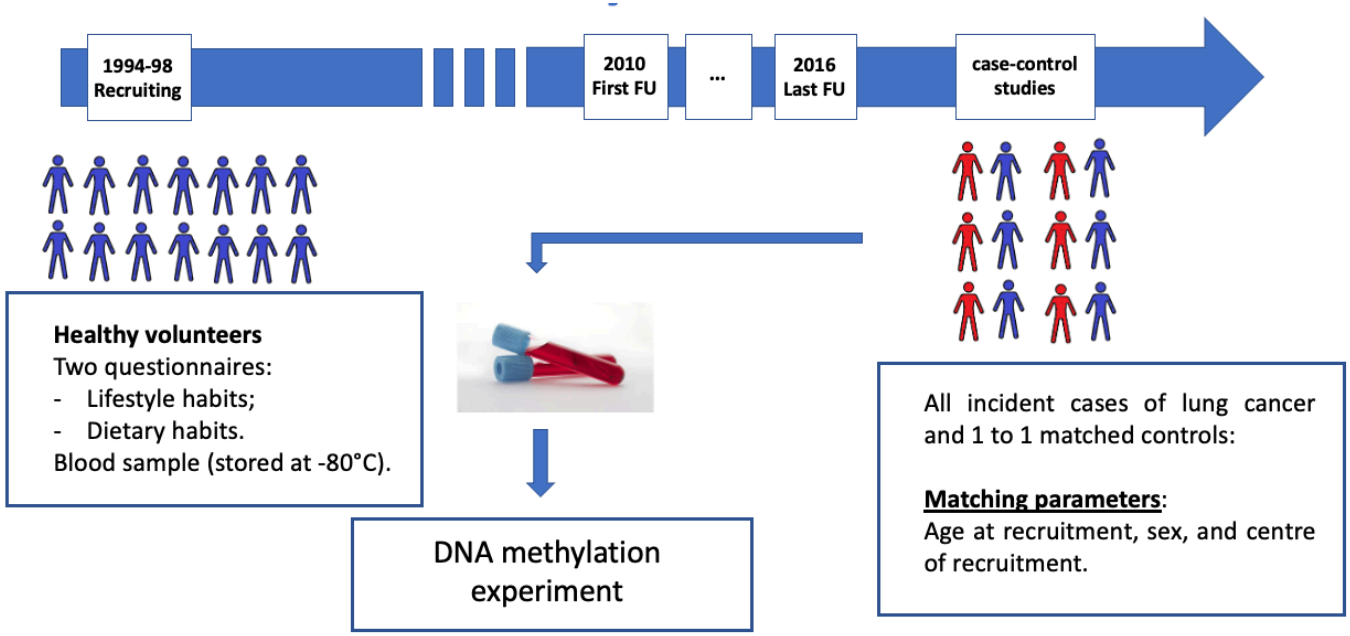
- There are considerable advantages in employing DNAm surrogates for achieving epidemiological evidence
  - DNAm accounts for biases in self-reported (or measured) exposure (Zhang et al. 2016)
  - DNAm accounts for individual responses to risk factors (Conole et al. 2021)

Aim:

Creating a **multi-dimensional** DNAm biomarker for cardiovascular and high blood pressure comorbidities from a **multi-center study**

EPIC Italy dataset

The European Prospective Investigation into Cancer and Nutrition (EPIC) study is one of the largest studies in the world



- **EPIC Italy:** four sub-cohorts identified by the centre of recruitment; Ragusa, Varese, Turin and Naples
- **5-dimensional response:** major risk factors for cardiovascular diseases; DBP, SBP, HDL, LDL and TG

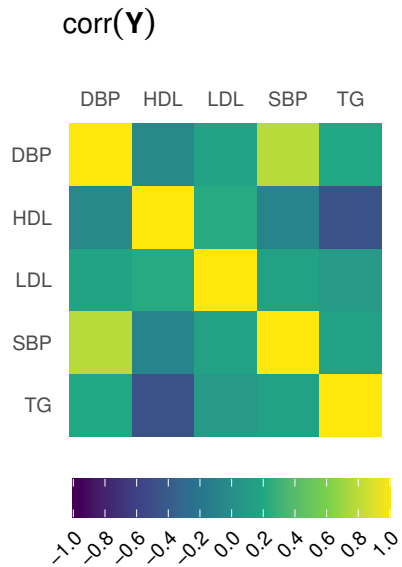
DNAm surrogate creation: modeling challenges

DNAm surrogate creation: modeling challenges

- **High-dimensionality:** 62128 DNA methylation features

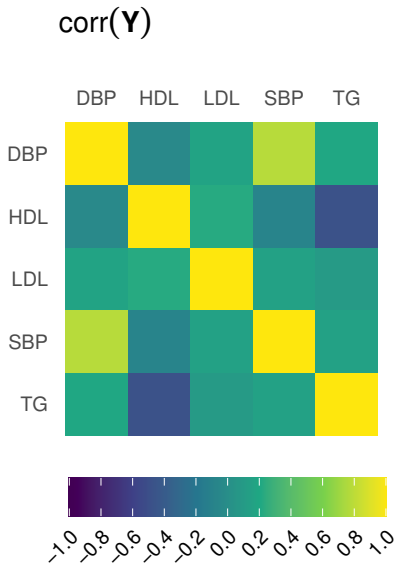
DNAm surrogate creation: modeling challenges

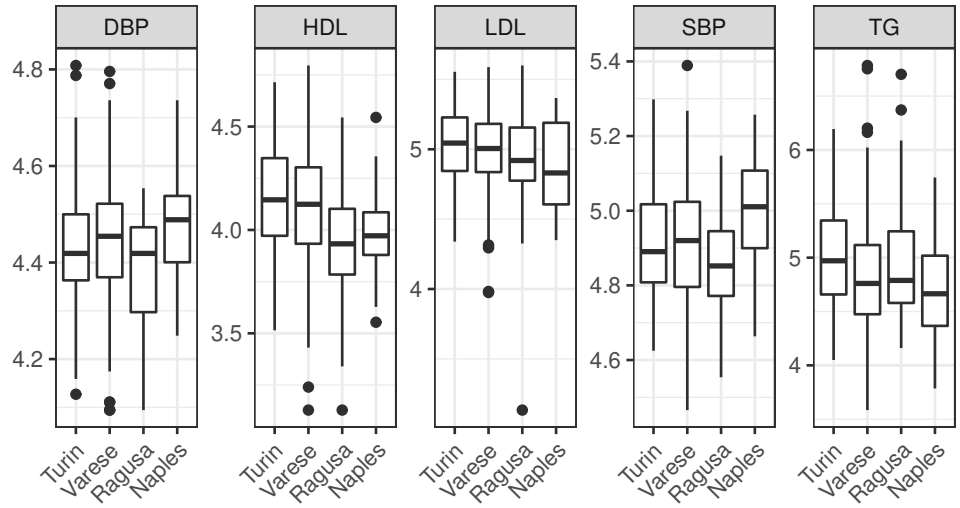
- **High-dimensionality:** 62128 DNA methylation features
- **Multivariate response**



DNAm surrogate creation: modeling challenges

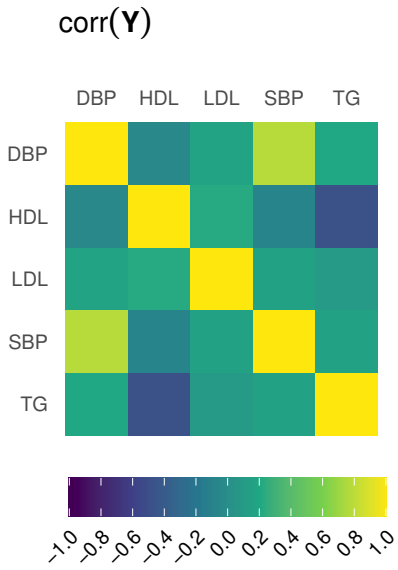
- **High-dimensionality:** 62128 DNA methylation features
- **Multivariate response**
- **Center-wise variability**

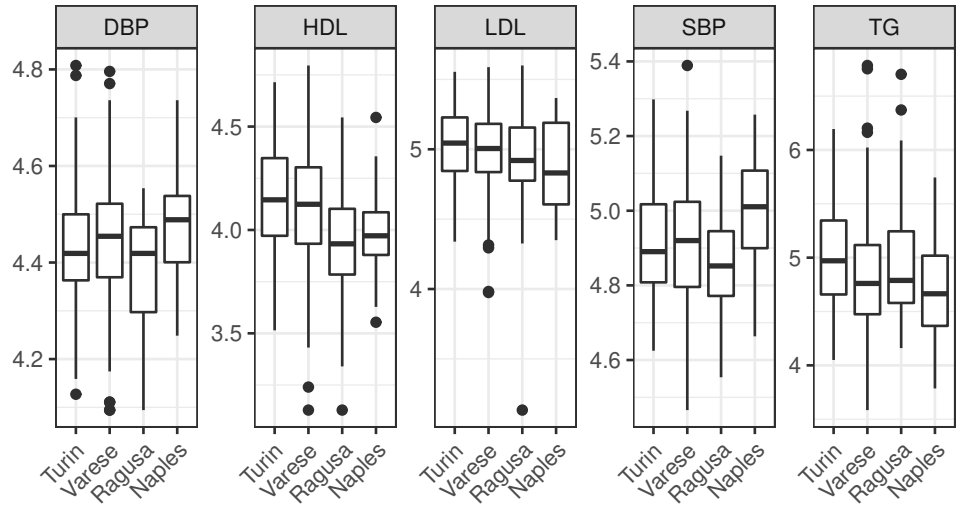




DNAm surrogate creation: modeling challenges

- **High-dimensionality:** 62128 DNA methylation features
- **Multivariate response**
- **Center-wise variability**





Proposed solution

Penalized mixed-effects multitask learning

Penalized MLMM: definition

Consider a Multivariate Mixed-Effects Model (MLMM)

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{B} + \boldsymbol{Z}_j \boldsymbol{\Lambda}_j + \boldsymbol{E}_j$$

- $j = 1, \ldots, J$, J total number of groups
- n_j units in group j, $\sum_{j=1}^{J} n_j = N$
- $\boldsymbol{B}$ is the $p \times r$ matrix of fixed-effects
- $\boldsymbol{\Lambda}_j$ is the $q \times r$ matrix of random-effects
- $\boldsymbol{X}_j$ is the $n_j \times p$ fixed-effects design matrix
- $\boldsymbol{Z}_j$ is the $n_j \times q$ random-effects design matrix
- $\boldsymbol{E}_j$ is the $n_j \times r$ within-group error matrix

We further assume that

$$\operatorname{vec}(\boldsymbol{\Lambda}_j) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \qquad \operatorname{vec}(\boldsymbol{E}_j) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j})$$

MLMM Group-lasso: objective function

We maximize the following penalized log-likelihood:

$$\ell_{pen}(\boldsymbol{\theta}) = \sum_{j=1}^{J} \log \phi \left(\text{vec}\left(\boldsymbol{Y}_j\right), \left(\boldsymbol{I}_r \otimes \boldsymbol{X}_j\right) \text{vec}\left(\boldsymbol{B}\right), \left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right) \boldsymbol{\Psi} \left(\boldsymbol{I}_r \otimes \boldsymbol{Z}_j\right)' + \boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_j} \right) +$$

$$- \lambda \left[(1-\alpha) \sum_{c=1}^{r} \sum_{l=2}^{p} b_{lc}^2 + \alpha \sum_{l=2}^{p} ||\boldsymbol{b}_{l.}||_2 \right]$$

- Log-likelihood of MLMM for a sample of $N = \sum_{j=1}^{J} n_j$
- Group-lasso penalty with shrinkage factor λ and α mixing ridge/group-lasso regularizers
- $\boldsymbol{\theta} = \{\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}\}$ set of model parameters
- b_{lc} and $\boldsymbol{b}_{l.}$ element in position (l, c) and l-th row of matrix $\boldsymbol{B}$, respectively
- With group-lasso penalty coefficients are **either all zero or none are zero** across all components of the response

DNAm surrogates creation

- EPIC Italy dataset split in training ($N_{tr} = 401$) and test ($N_{te} = 173$) sets
- Random-effects design matrix with $q = 1$ (random intercept model)
- λ tuned via 10-fold CV, $\alpha = 0.5$

DNAm surrogates creation

- EPIC Italy dataset split in training ($N_{tr} = 401$) and test ($N_{te} = 173$) sets
- Random-effects design matrix with $q = 1$ (random intercept model)
- λ tuned via 10-fold CV, $\alpha = 0.5$

Framework		Root Mean Squared Error					Active #
Model	Penalty type	DBP	HDL	LDL	SBP	TG	CpG sites
MLMM	Group-lasso	**0.1167**	**0.2442**	**0.3236**	0.1374	**0.4745**	417
MLMM	Elastic-net	0.1178	0.2480	0.3318	0.1379	0.4853	773
MLM	Group-lasso	0.1317	0.2602	0.3321	0.1364	0.4853	382
MLM	Elastic-net	0.1176	0.2513	0.3324	0.1362	0.4841	115
LM	Elastic-net	0.1179	0.2596	0.3383	**0.1359**	0.4996	1712

DNAm surrogates creation

- EPIC Italy dataset split in training ($N_{tr} = 401$) and test ($N_{te} = 173$) sets
- Random-effects design matrix with $q = 1$ (random intercept model)
- λ tuned via 10-fold CV, $\alpha = 0.5$

Framework		Root Mean Squared Error					Active #
Model	Penalty type	DBP	HDL	LDL	SBP	TG	CpG sites
MLMM	Group-lasso	**0.1167**	**0.2442**	**0.3236**	0.1374	**0.4745**	417
MLMM	Elastic-net	0.1178	0.2480	0.3318	0.1379	0.4853	773
MLM	Group-lasso	0.1317	0.2602	0.3321	0.1364	0.4853	382
MLM	Elastic-net	0.1176	0.2513	0.3324	0.1362	0.4841	115
LM	Elastic-net	0.1179	0.2596	0.3383	**0.1359**	0.4996	1712

- MLMM Group-lasso outperforms the state-of-the-art approach for 4 out of 5 dimensions of the response

DNAm surrogates creation

- EPIC Italy dataset split in training ($N_{tr} = 401$) and test ($N_{te} = 173$) sets
- Random-effects design matrix with $q = 1$ (random intercept model)
- λ tuned via 10-fold CV, $\alpha = 0.5$

Framework		Root Mean Squared Error					Active #
Model	Penalty type	DBP	HDL	LDL	SBP	TG	CpG sites
MLMM	Group-lasso	**0.1167**	**0.2442**	**0.3236**	0.1374	**0.4745**	417
MLMM	Elastic-net	0.1178	0.2480	0.3318	0.1379	0.4853	773
MLM	Group-lasso	0.1317	0.2602	0.3321	0.1364	0.4853	382
MLM	Elastic-net	0.1176	0.2513	0.3324	0.1362	0.4841	115
LM	Elastic-net	0.1179	0.2596	0.3383	**0.1359**	0.4996	1712

- MLMM Group-lasso outperforms the state-of-the-art approach for 4 out of 5 dimensions of the response
- The total number of active CpG sites is lower for MLMM Group-lasso than LM Elastic-net

Association of DNAm surrogates with CVD risk

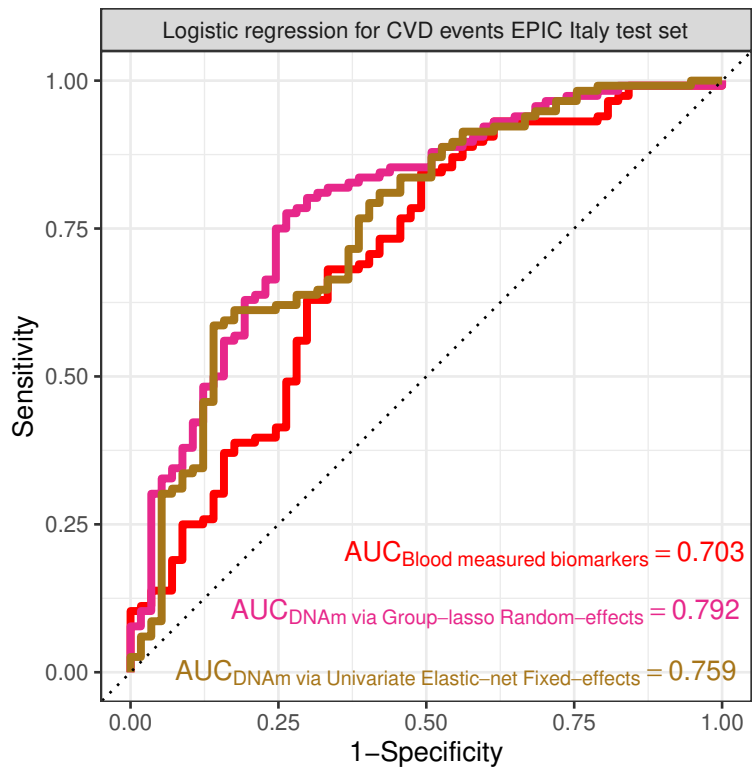
- DNAm surrogates provide reliable covariates for diseases prediction models

Association of DNAm surrogates with CVD risk

- DNAm surrogates provide reliable covariates for diseases prediction models
- Interest in predicting the probability of cardiovascular risk

Association of DNAm surrogates with CVD risk

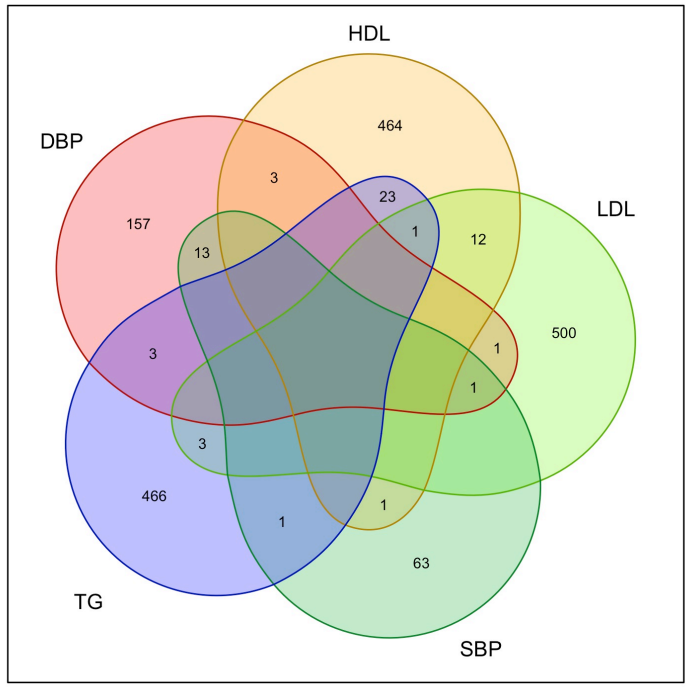
- DNAm surrogates provide reliable covariates for diseases prediction models
- Interest in predicting the probability of cardiovascular risk



- DNAm surrogates define more informative regressors than blood measured biomarkers

CpG sites selection and enrichment analyses

- **Pleiotropic effect**: multiple correlated phenotypes are affected by the same set of CpG sites (Tyler, Crawford, and Pendergrass 2013)
- Group-lasso automatically accounts for this effect
- The same is not achieved for univariate elastic-net models



- Significant enrichment for CpGs in inflammatory pathways associated with the dysregulation of the immune system

Conclusion

- **Linear regression** provides an interpretable baseline for predicting continuous outcomes, but assumes all patients are exchangeable (James et al. 2013)
- **Mixed-effects models** extend regression to *multicentric data* by learning a hierarchy-specific offset (Pinheiro and Bates 2006)
- **LASSO** addresses *high-dimensional data* by shrinking irrelevant predictors to zero (Hastie, Tibshirani, and Wainwright 2015)
- Each of these three methods handles an additional layer of **real-world complexity** present in medical data.
- The **case study** shows how all three ideas combine in a single framework (Cappozzo, Ieva, and Fiorito 2023)

Conclusion

- **Linear regression** provides an interpretable baseline for predicting continuous outcomes, but assumes all patients are exchangeable (James et al. 2013)
- **Mixed-effects models** extend regression to *multicentric data* by learning a hierarchy-specific offset (Pinheiro and Bates 2006)
- **LASSO** addresses *high-dimensional data* by shrinking irrelevant predictors to zero (Hastie, Tibshirani, and Wainwright 2015)
- Each of these three methods handles an additional layer of **real-world complexity** present in medical data.
- The **case study** shows how all three ideas combine in a single framework (Cappozzo, Ieva, and Fiorito 2023)

Thank you!

References

- Cappozzo, Andrea, Francesca Ieva, and Giovanni Fiorito (2023). "A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation". In: *The Annals of Applied Statistics* 17.4, pp. 3257–3282. arXiv: 2112.12719.
- Conole, Eleanor L.S. et al. (2021). "DNA Methylation and Protein Markers of Chronic Inflammation and Their Associations With Brain and Cognitive Aging". In: *Neurology* 97.23, e2340–e2352.
- Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical Learning with Sparsity*. Chapman and Hall/CRC, pp. 1–337.
- James, Gareth et al. (2013). *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York. arXiv: arXiv:1011.1669v3.
- Pinheiro, José and Douglas Bates (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.
- Tyler, Anna L., Dana C. Crawford, and Sarah A. Pendergrass (2013). "Detecting and characterizing pleiotropy: new methods for uncovering the connection between the complexity of genomic architecture and multiple phenotypes". In: *Biocomputing 2014*. WORLD SCIENTIFIC, pp. 183–187.
- Zhang, Yan et al. (2016). "Smoking-associated DNA methylation markers predict lung cancer incidence". In: *Clinical Epigenetics* 8.1, pp. 1–12.