Better omics-based predictions through the use of big co-data

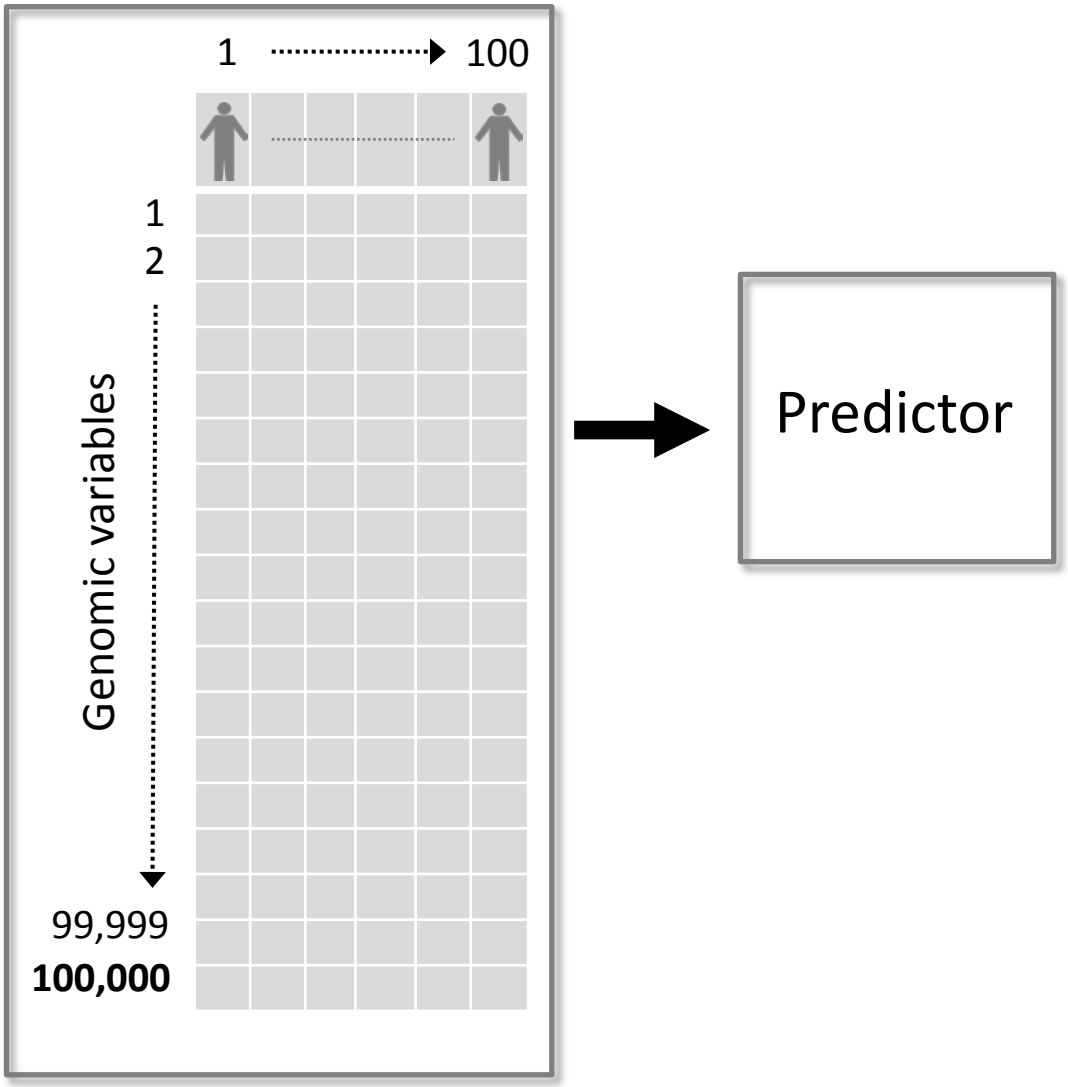
Mark van de Wiel

Chair Statistics for Genomics Unit
Dep Epidemiology & Biostatistics, VUmc
Dep Mathematics, VU

www.bigstatistics.nl

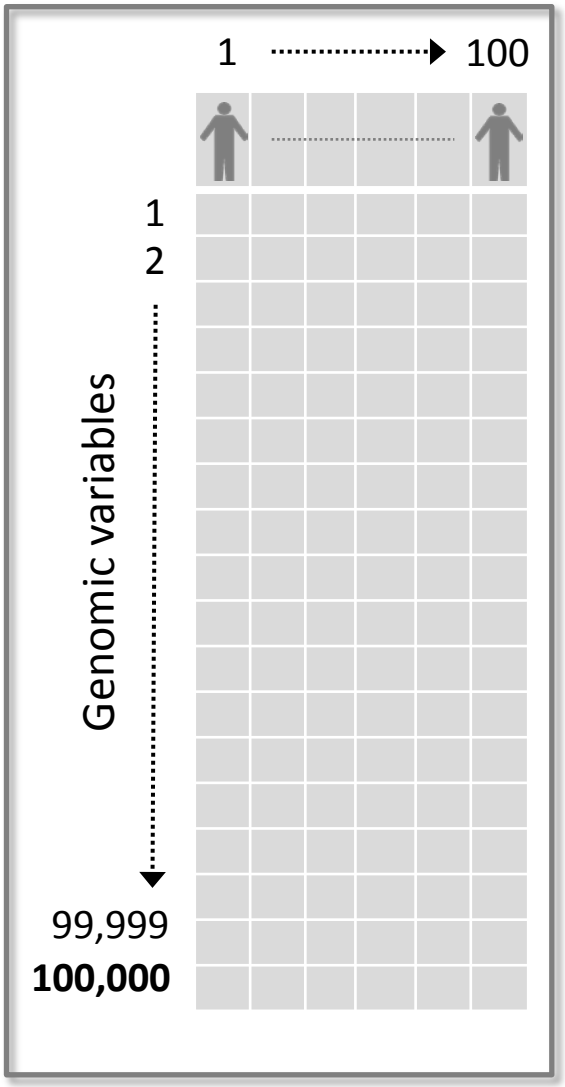
VUmc

Setting



VUmc

Setting



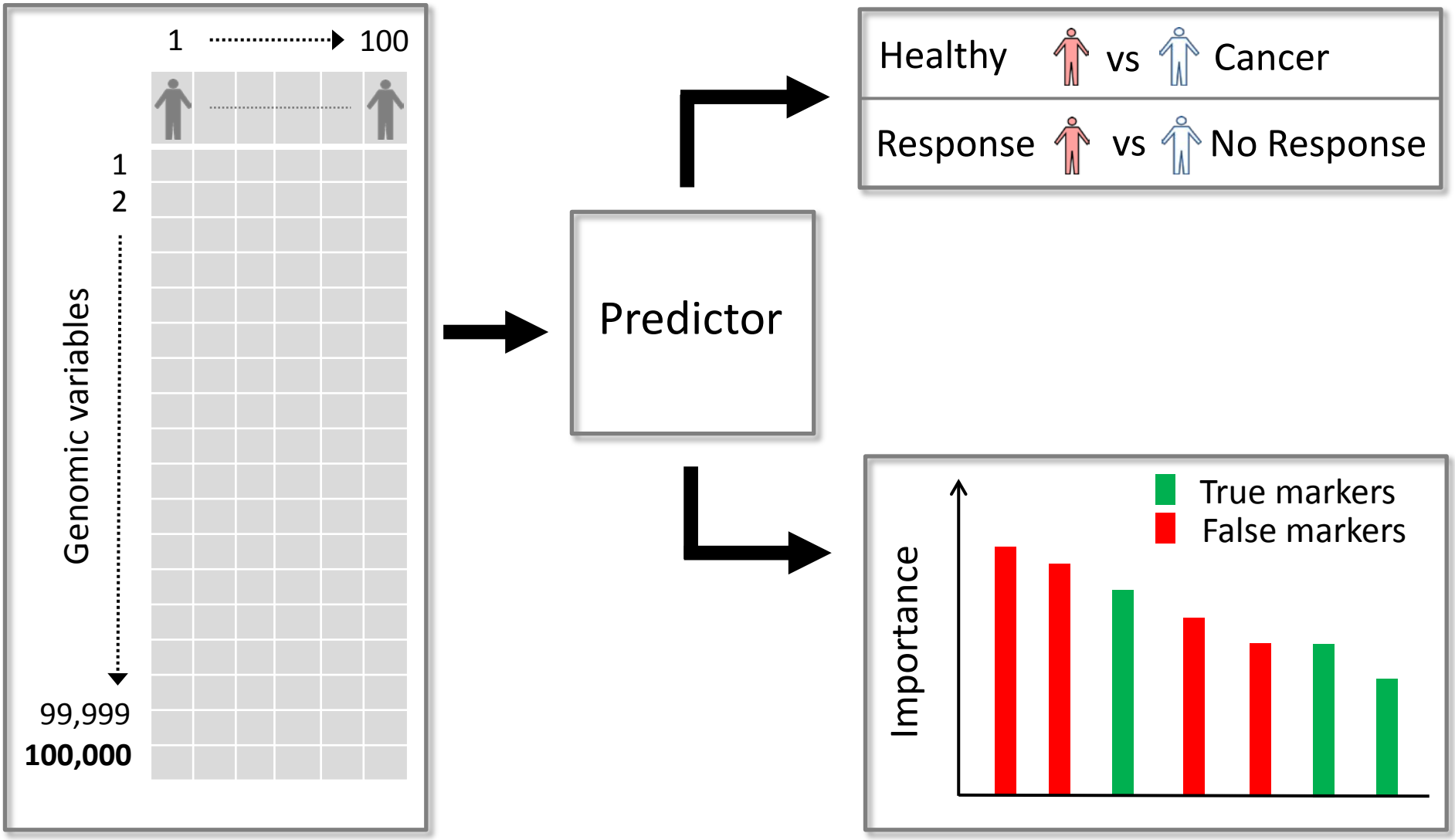
Predictor

Healthy vs Cancer

Response vs No Response

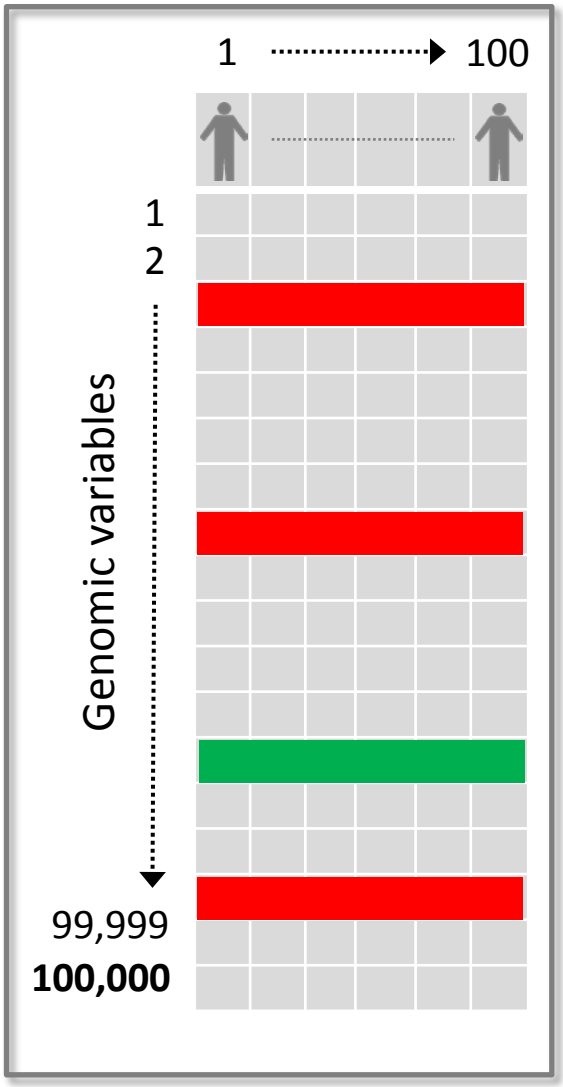
VUmc

Setting



VUmc

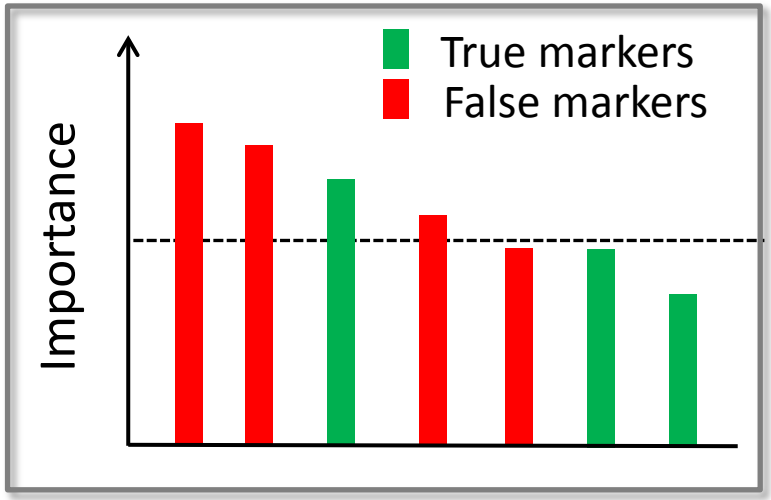
Setting



Predictor

Healthy vs Cancer

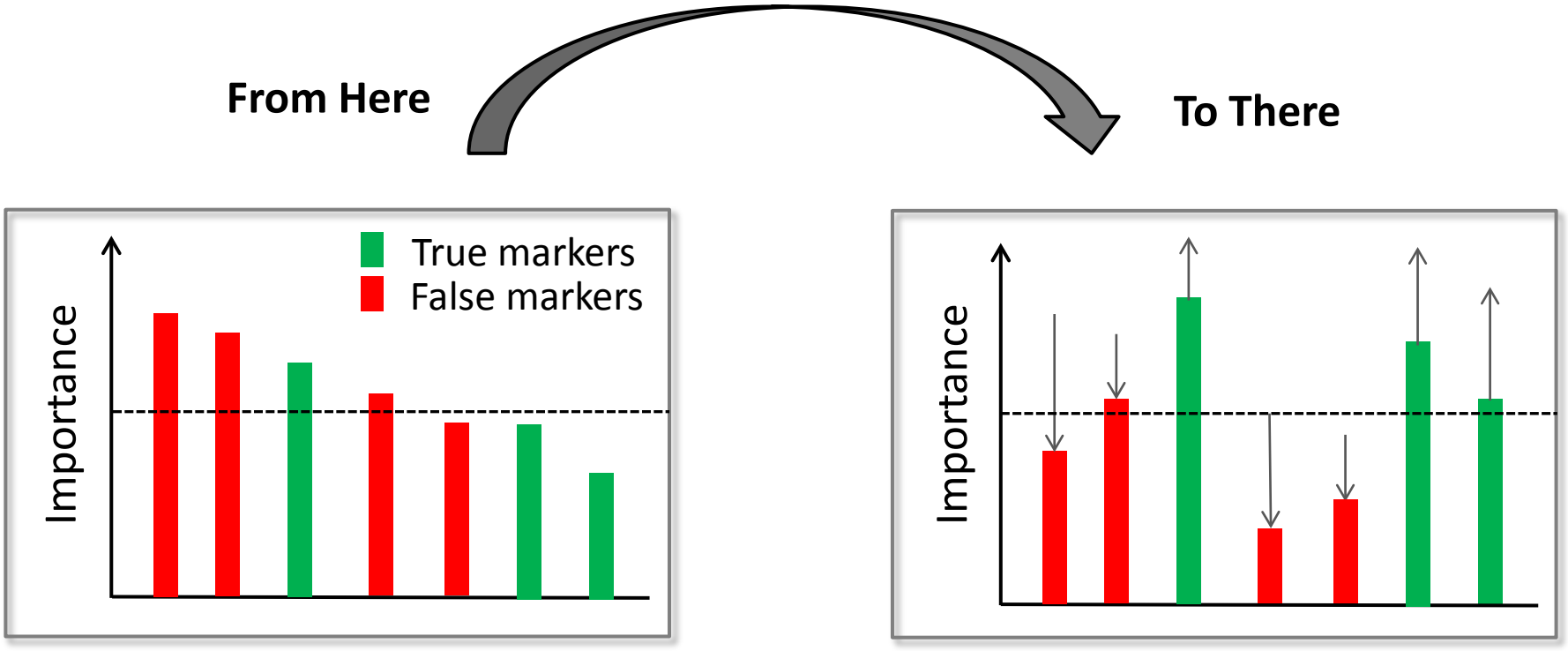
Response vs No Response



VUmc

The Aim

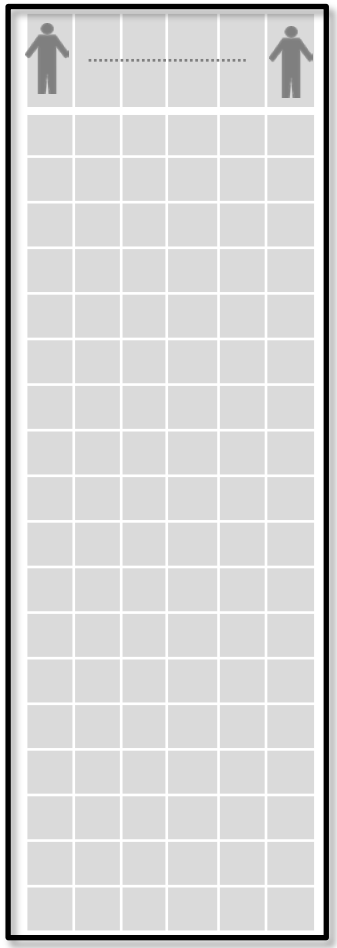
From Here



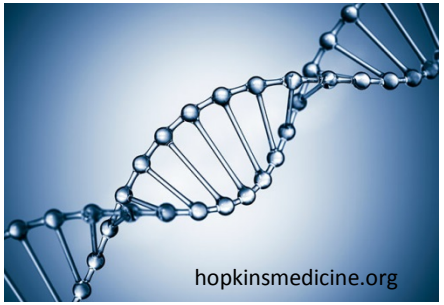
To There

How: Co-Data

Databases

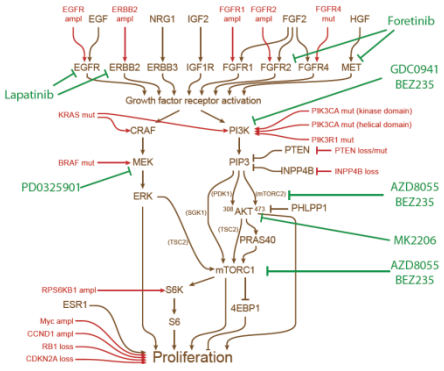


Primary Data

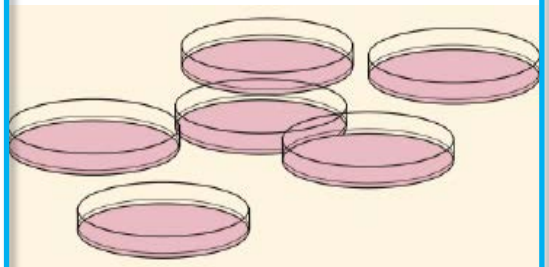


Related bio-molecules

Pathways



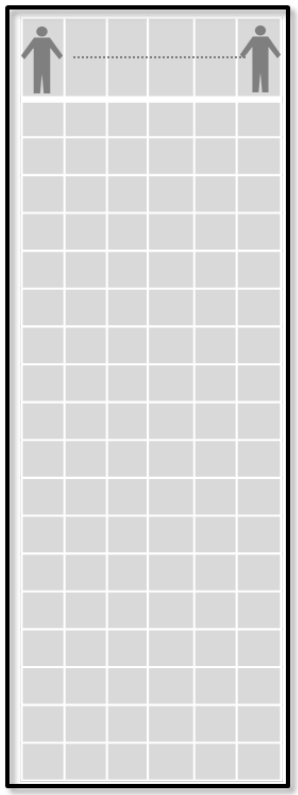
Cell lines



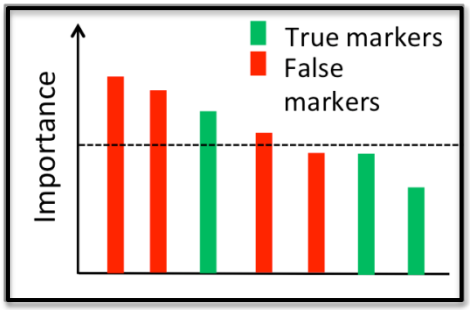
VUmc

The Methods

Primary Data



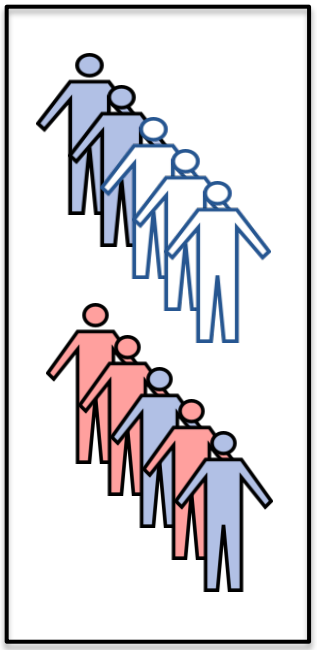
Initial prediction



Markers

Final prediction

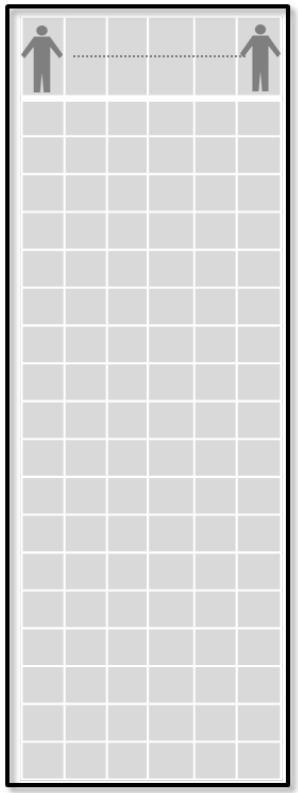
= error



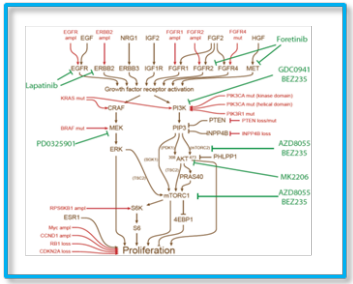
VUmc

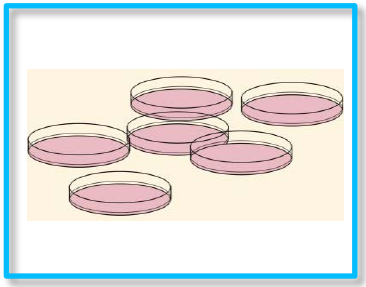
Add co-data

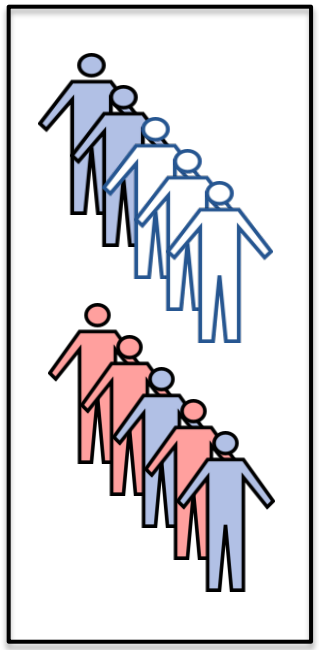
Primary Data



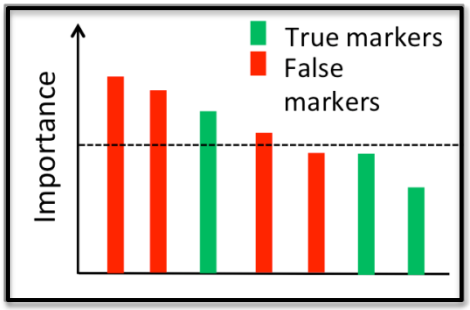
+







Initial prediction

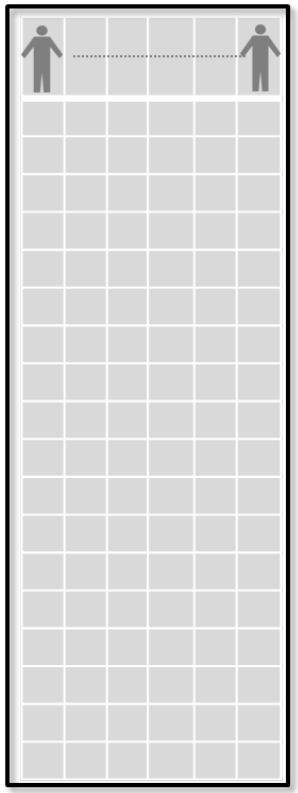


Markers

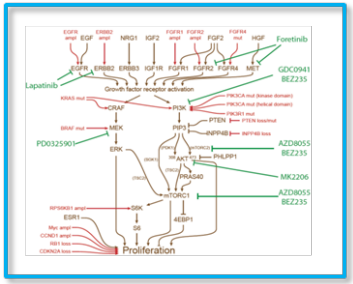
Final prediction

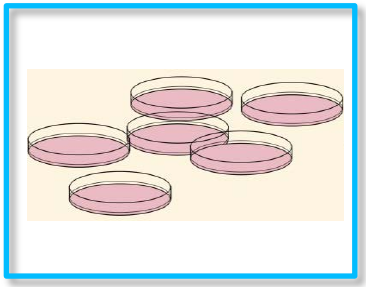
Perform weighted Learning

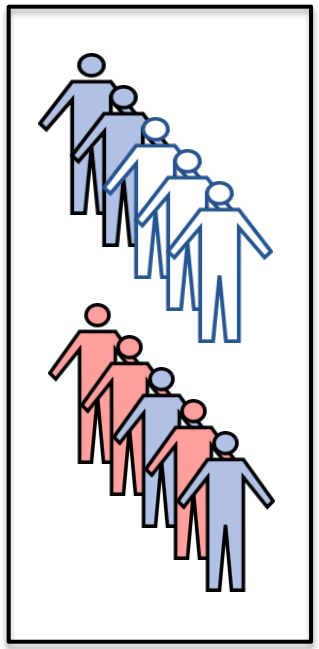
Primary Data



+

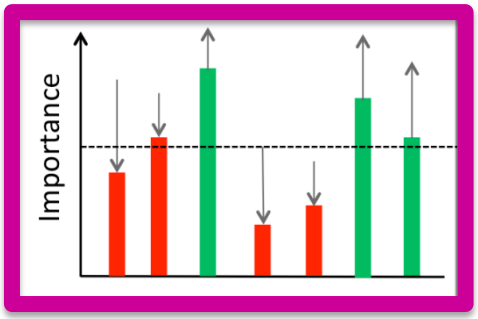






Initial prediction

Weighted learning



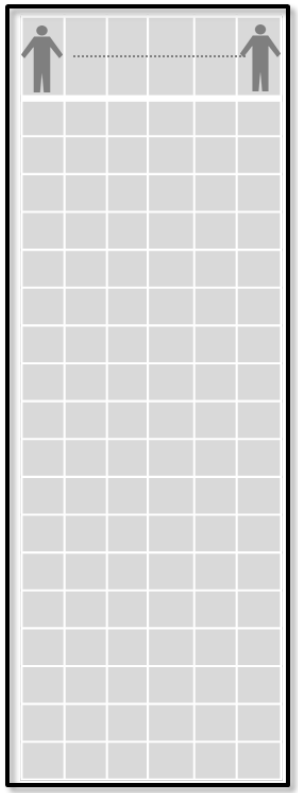
Better Markers

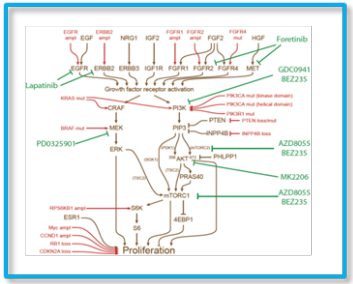
Final prediction

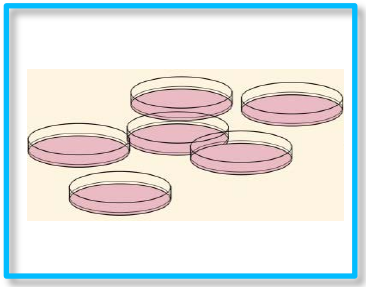
VUmc

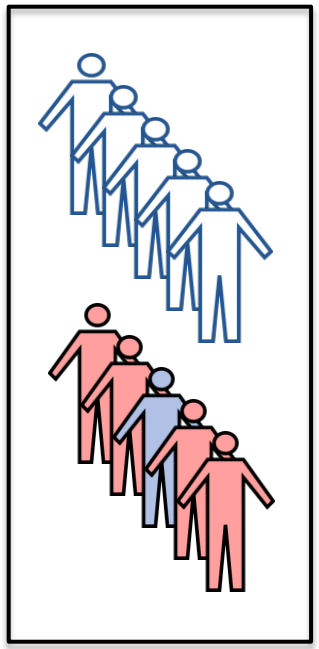
Improved markers and prediction

Primary Data



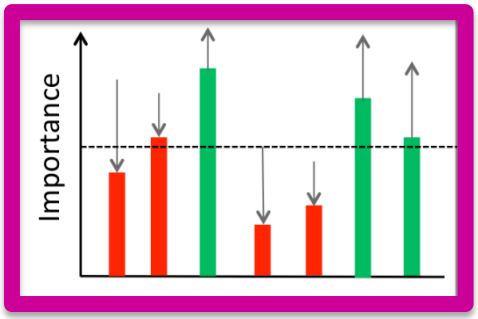






Initial prediction

Weighted learning



Better markers

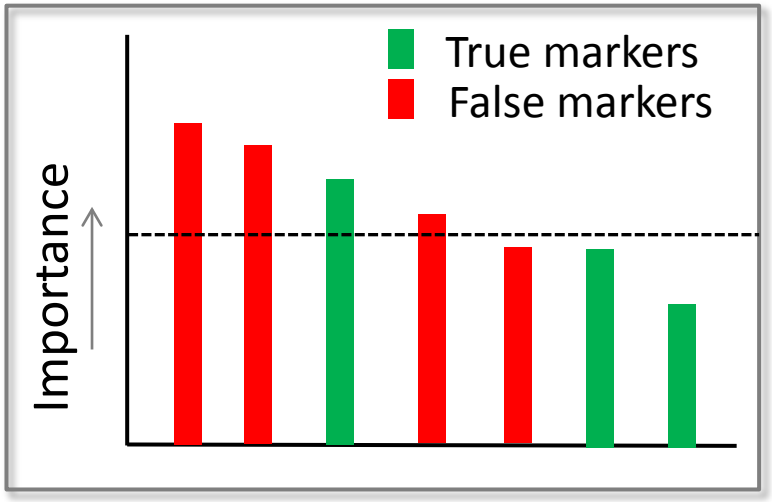
Better final prediction

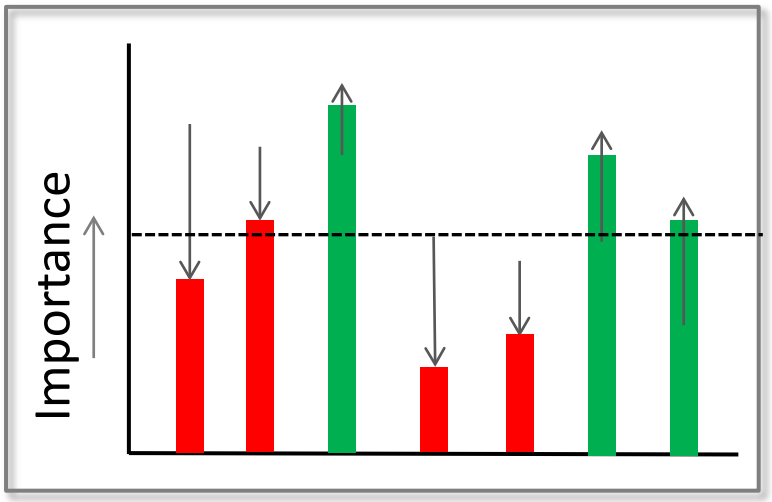
VUmc

The Effect

No Co-Data

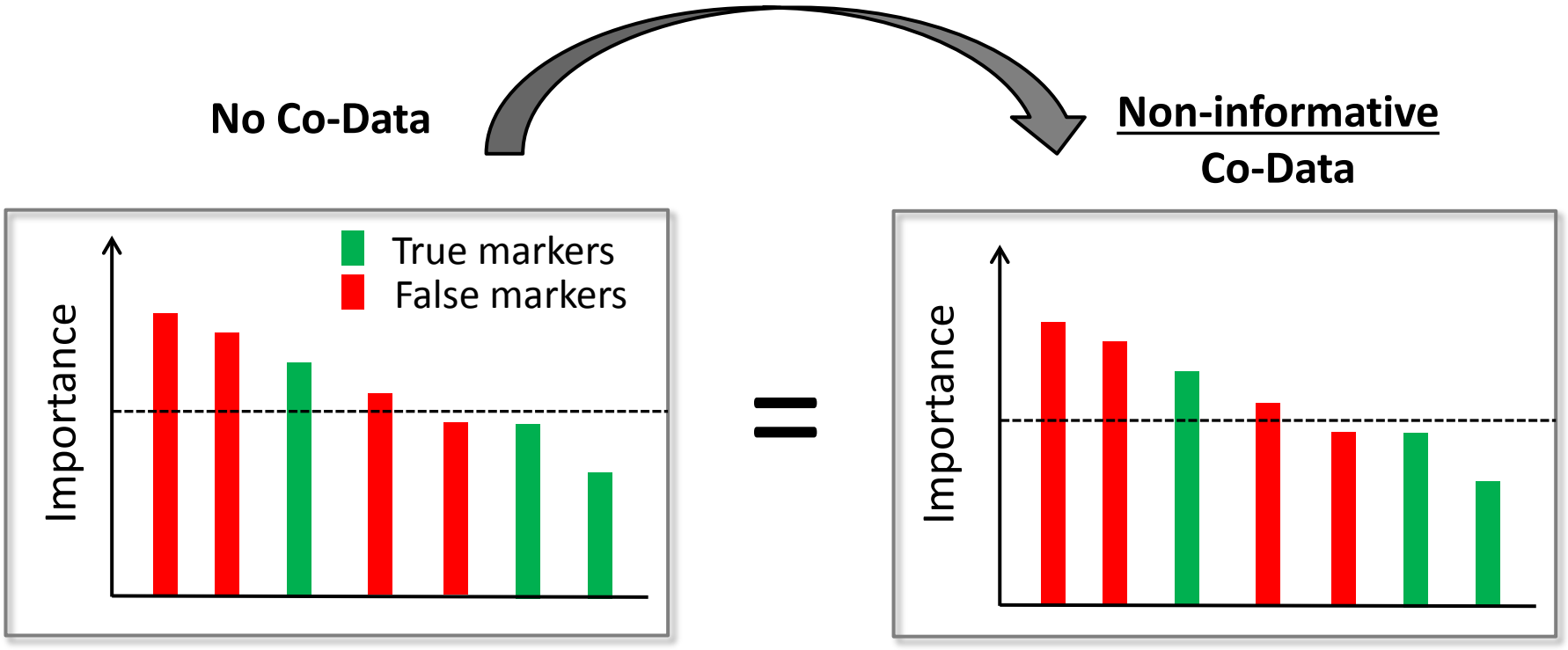
Informative Co-Data





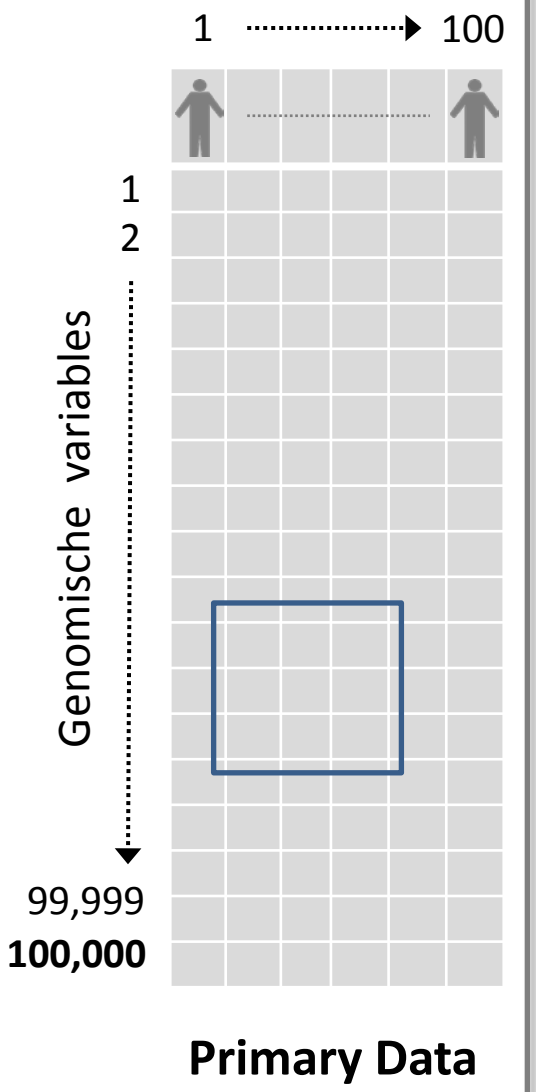
VUmc

Robustness



VUmc

Some Notation



Matrix $\mathbf{X} = (X_{ij})_{i=1,\ldots,p;j=1,\ldots,n}$

n = 100 = sample size

p = 100,000 = # variables

=

$X_{10,2}$	$X_{10,3}$	$X_{10,4}$
$X_{11,2}$	$X_{11,3}$	$X_{11,4}$
$X_{12,2}$	$X_{12,3}$	$X_{12,4}$

=

10	0	15
122	5	67
34	0	1209

Response vector $\mathbf{Y} = (Y_1,\ldots,Y_n)$

Binair (disease/healthy):

Y = 1 1 0 0 1 ...

Continuous (eg. BMI):

Y = 24.6 17.2 21.3 29.4 19.1 ...

Method: Ridge Regression

Response labels (n)

Regression Parameters (p)

Genomic variables (pxn)

$$\text{argmax}_{\beta}\mathcal{L}(\mathbf{Y};\beta,\mathbf{X}) - \lambda||\beta||_2$$

Fit Function (Likelihood)

Complexity Penalty

$$||\beta||_2 = \sum_{j=1}^{p} \beta_j^2$$

Grouped co-data, replace $\lambda||\beta||_2$ by $\sum_{g=1}^{G} \lambda_g||\beta_g||_2$

Group penalties are inverse weights → Weighted Learning

VUmc

Mathematical Challenge

Aim: estimate penalty parameters → CV slow and instable.

$\sum_{g=1}^{G} \lambda_g ||\beta_g||_2$ can be formulated as priors in a Bayesian framework

Empirical Bayes estimates prior parameters from the data by maximizing:

$$\int_{\beta} \mathcal{L}(\mathbf{Y}; \beta, \mathbf{X}) \pi_{\tau}(\beta) d\beta, \qquad \boldsymbol{\tau} = (\tau_1, \ldots, \tau_g)$$

Hard: approximate procedures needed, e.g. VdW et al. (2016), Stat Medicine

Once group-weights are known, fitting and prediction is straightforward

VUmc

Extension 1

Variable Selection: Important in clinical diagnostics: cheap molecular test

Current Method: Extended with post-hoc marker selection module

Weighted Elastic Net[1]

$$\text{argmax}_{\beta}\mathcal{L}(\mathbf{Y};\beta,\mathbf{X}) - \sum_{g=1}^{G}\lambda_g||\beta_g||_2 - \sum_{g=1}^{G}\sqrt{\lambda_g}||\beta_g||_1$$

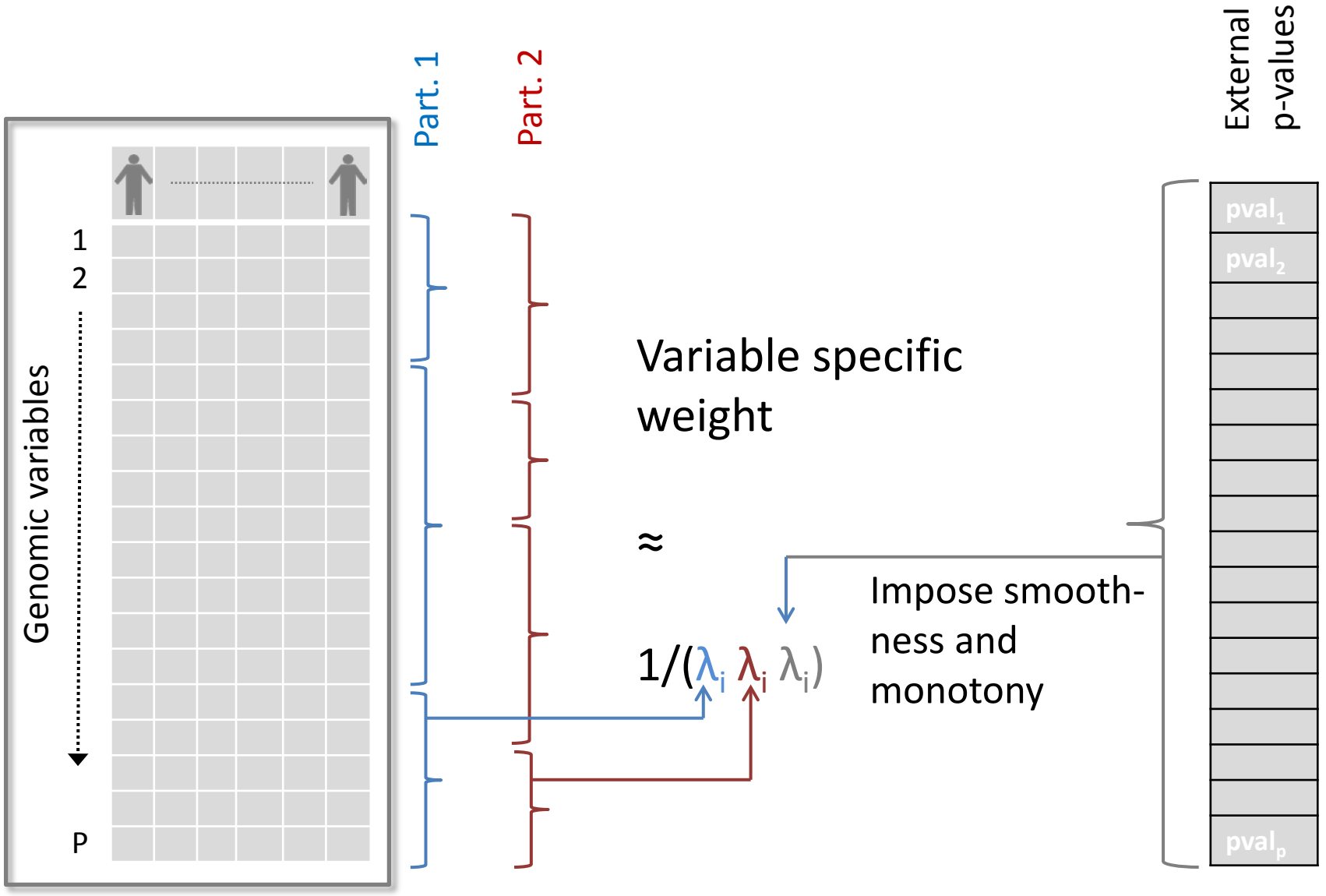
L1 (or lasso) penalty: explicitly sets betas to zero: variable selection

Group penalties estimated by Variational Bayes – Empirical Bayes

[1]Magnus Münch

VUmc

Extension 2: Multiple Partitions + Continuous Co-data



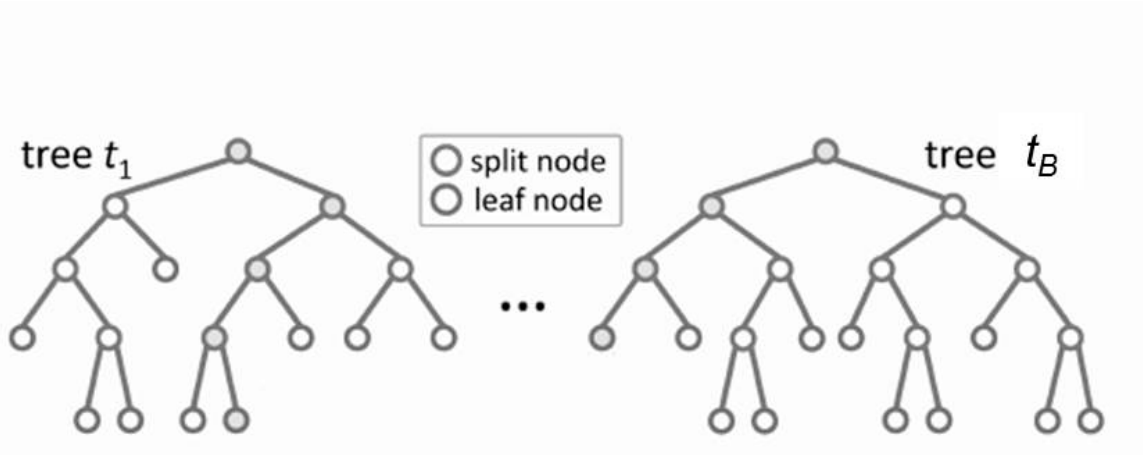
VUmc

Extension 3

Random Forest[1]

+ Computational scalability w.r.t. p
+ Scale invariant
+ Non-linearities

- Additive effects
- Variable importance / selection



Candidates for node splits sampled uniformly from 1,…, p by default
→ Main idea: moderate sampling weights by use of co-data model

[1]Dennis te Beest

VUmc

Software

- Group-regularized ridge regression: `GRridge`, R package. See Novianti et al. (2017), *Bioinformatics*
  - *By MvdW and Putri Novianti*
- Co-data Random Forest: `CoRF`, R package
  - *By Dennis te Beest*
- Group-regularized elastic net: `GRen`, R package, under development
  - *By Magnus Münch*

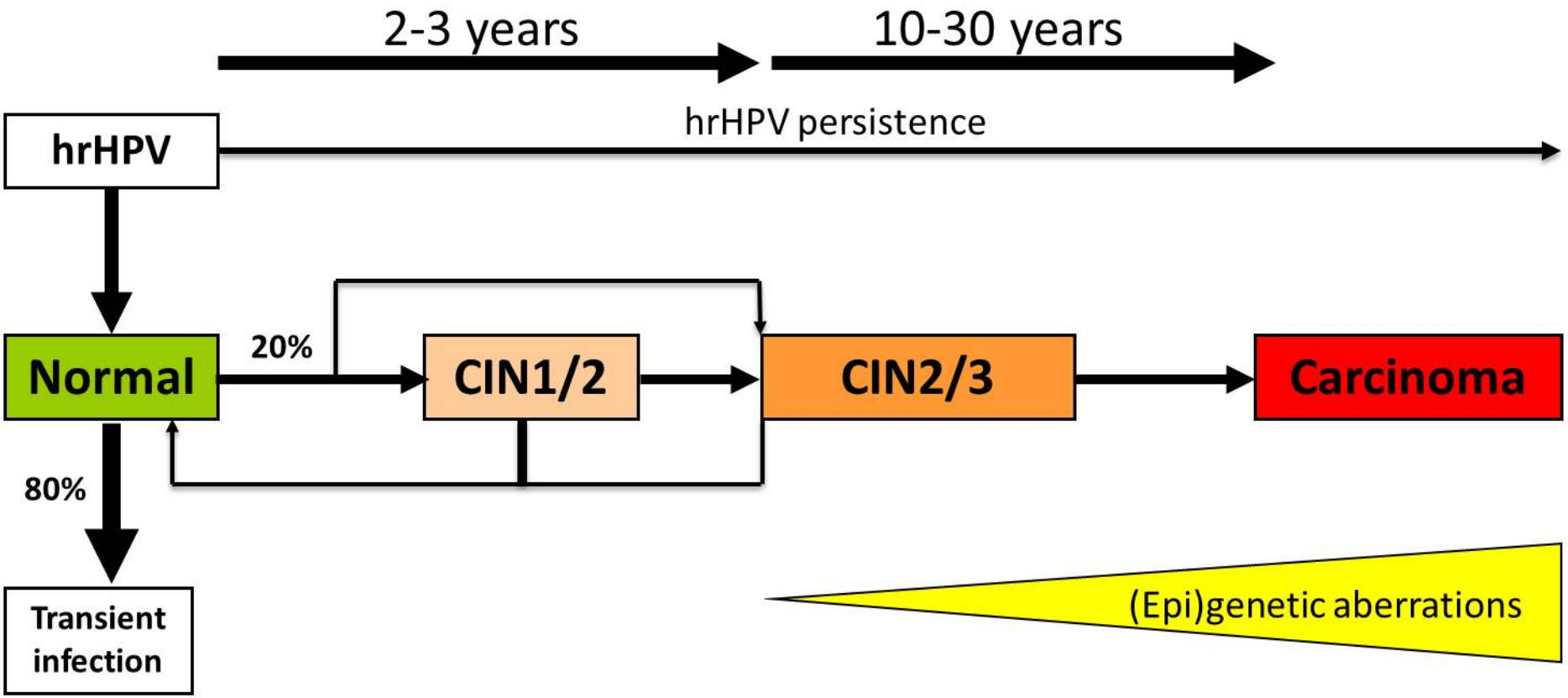
VUmc

Data Example: Cervical Cancer Diagnostics

- **Current tests**: Based on HPV (sometimes) i.c.w. cytology = accurate, but requiring high standards of cytological training
- **Additional problem**: Some women do not show up for screening
- **Molecular tests**:
  - Easy to implement,
  - Objective
  - Potentially cost-effective
  - Can be applied to self-samples.
- **Challenging**: Because self samples are of lower quality

VUmc

Cervical cancer progression



Aim: Detect CIN3 (generally leading to surgery)

VUmc

Molecular Self-Sample Test

Self-sampling at home

To laboratory

hrHPV testing

hrHPV-positive

Molecular testing

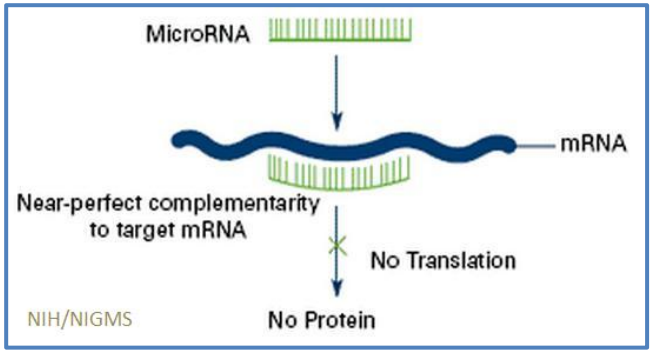
VUmc

Data Example

Goal: Select markers for classifying Normal vs CIN3
→ Final goal is a cheap PCR-based assay

Data:

- miRNA sequencing data
- n = 56 : 32 Normals, 24 CIN3s
- p = 772 (after filtering lowly abundant ones)
- Sqrt-transformed to quasi-Gaussian scale
- Standardized for penalty to have the same effect



Source: https://directorsblog.nih.gov/

VUmc

Data Example: Co-data

A. Conservation status:

1. Non-conserved (552)
2. Conserved across mammals (72)
3. Broadly conserved, across most vertebrates (148)

B. Standard deviation per variable:

10 groups of variables with decreasing s.d.

→ Allows natural variability to impact the classifier via penalty weights

VUmc

Data Example: Results

Conservation status:

1. Non-conserved (552): $\lambda_1 = 1.84$; $w_1 = 1/1.84 = 0.54$

2. Conserved across mammals (72): $\lambda_2 = 0.61$; $w_1 = 1.64$

3. Broadly conserved across vertebrates (148): $\lambda_3 = 0.30$; $w_3 = 3.33$

Standard deviation:

Penalties range from $\lambda_1 = 0.56$ (large sd) to $\lambda_{10} = 1.80$ (small sd)

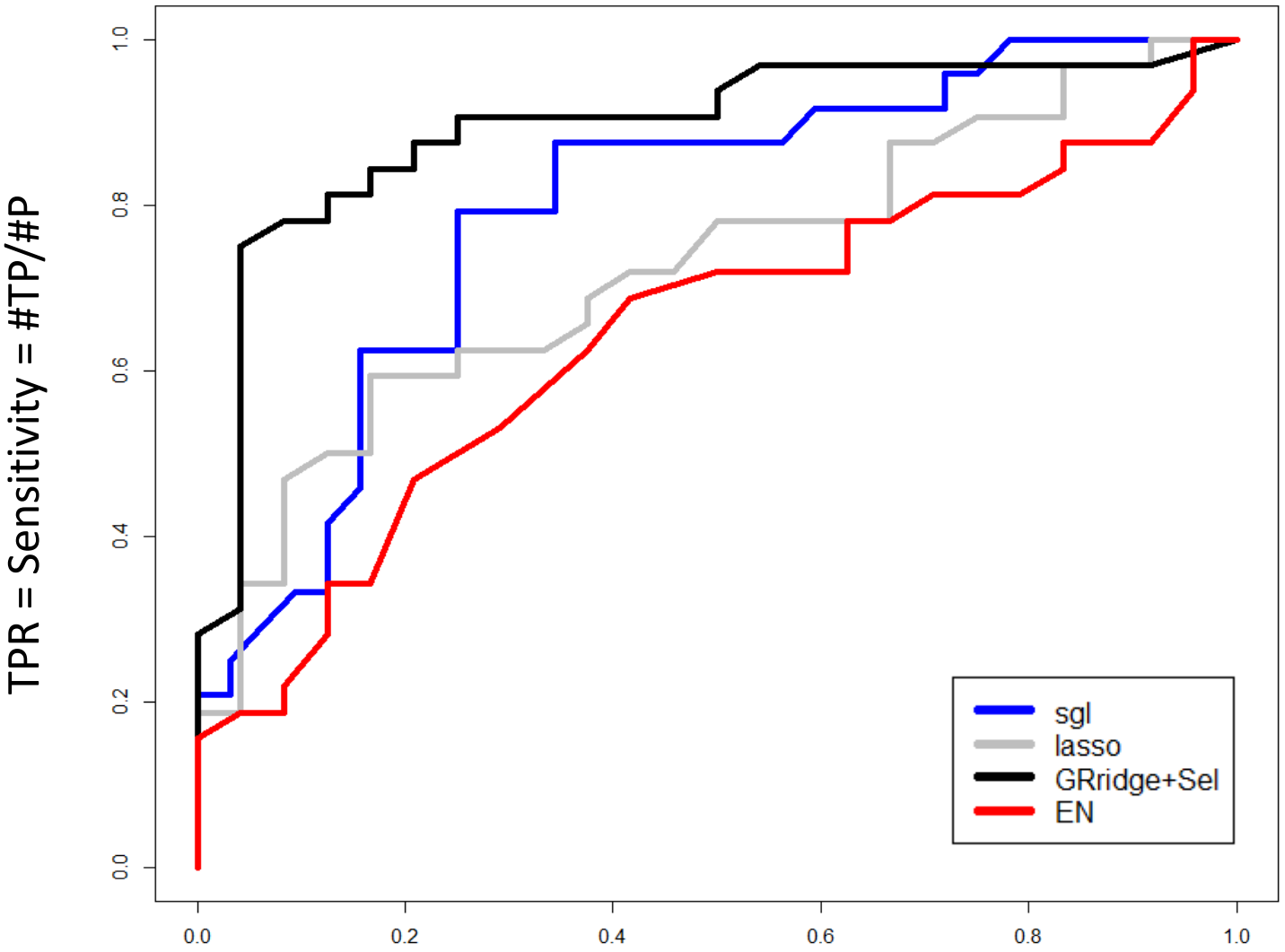
→ $w_1 = 1.79,\ldots,w_{10} = 0.56$

→ Indeed, partly 'undoes' the effect of standardization.

VUmc

Data Example: ROC curves

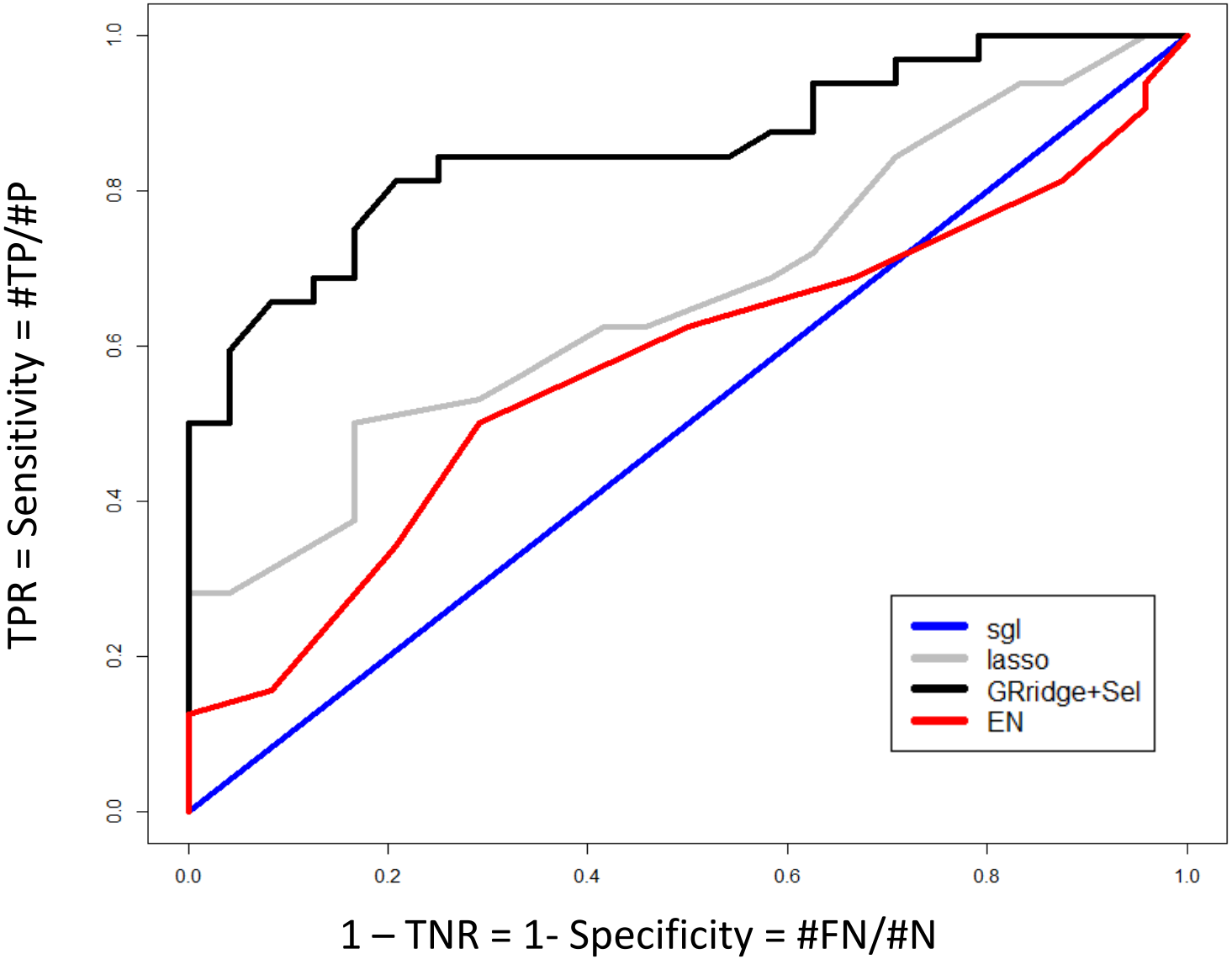
p_{sel} = 25 markers



VUmc

Data Example

p_{sel} = 3 markers



VUmc

Statement and Conclusion

Responsible data science → Every data scientist should aim to make maximal use of the available public data and information when developing a learner, thereby

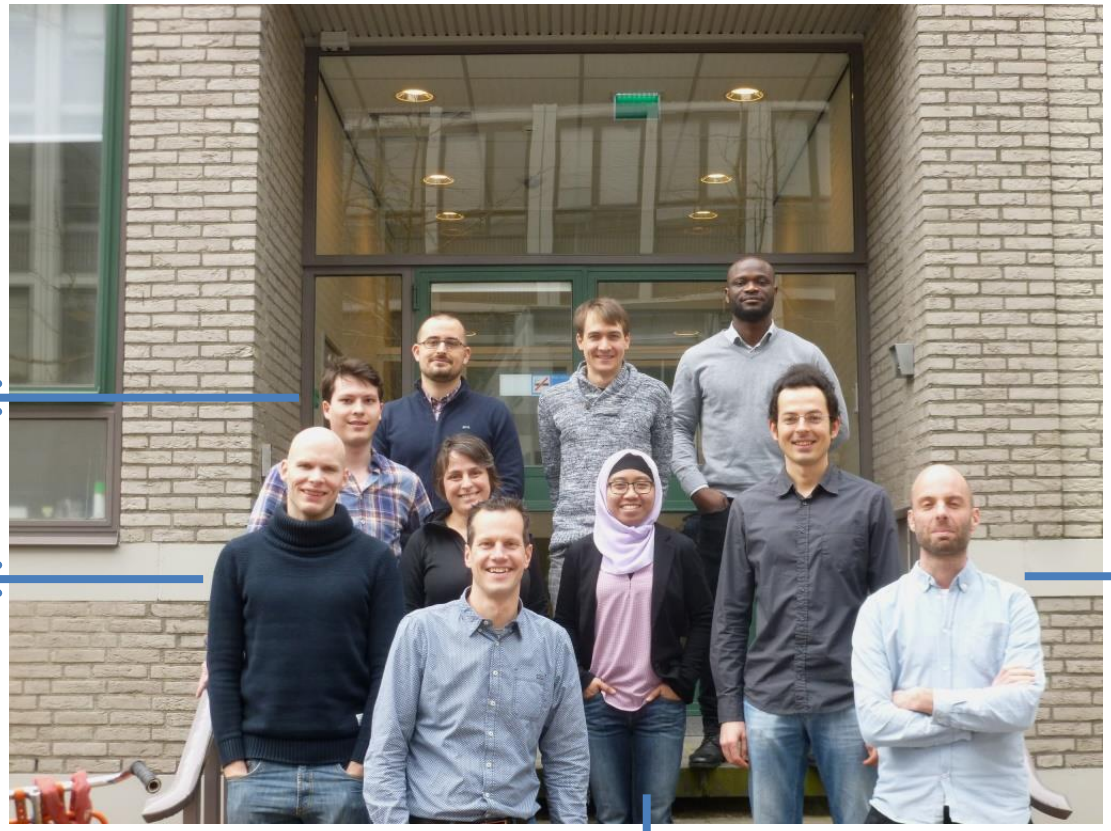
- **Acknowledging that the primary data set is the focus**
- **Guarding against overfitting**
- **Not decreasing performance when co-data is of bad quality**
- **Aiming for generalization to other settings**

The Co-Data approach fulfills these criteria, but **Improvements** are still needed:

- Explicitly accounting for (un)known relations between genes
- Large number of co-data parameters
- More sophisticated (deep) learners
- Merge with transfer learning

VUmc

Acknowledgements



Magnus Münch

Dennis te Beest

Wessel “Ridge” van Wieringen

Putri Novianti

+ Molecular Pathology, VUmc: Barbara Snoek, Saskia Wilting, Renske Steenbergen
+ Lodewyk Wessels (NKI, TU Delft)
+ ZONMW-TOP

VUmc