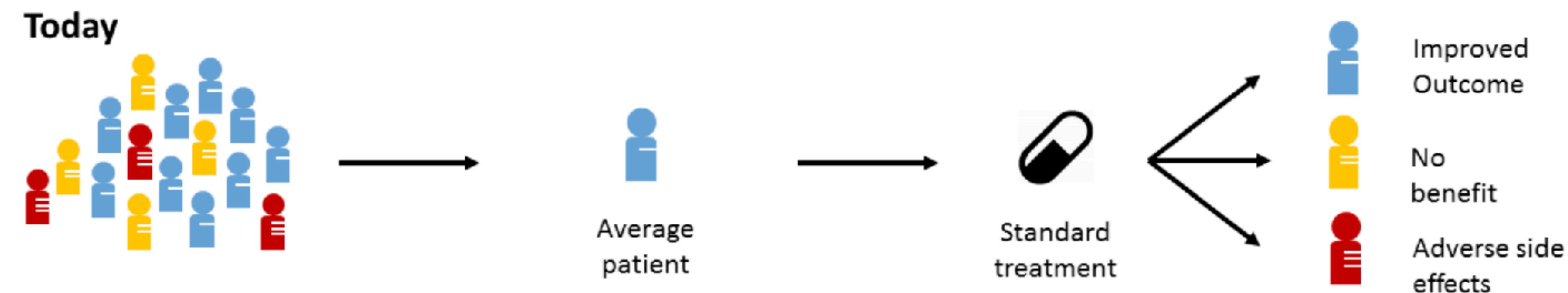


Kevin Wang

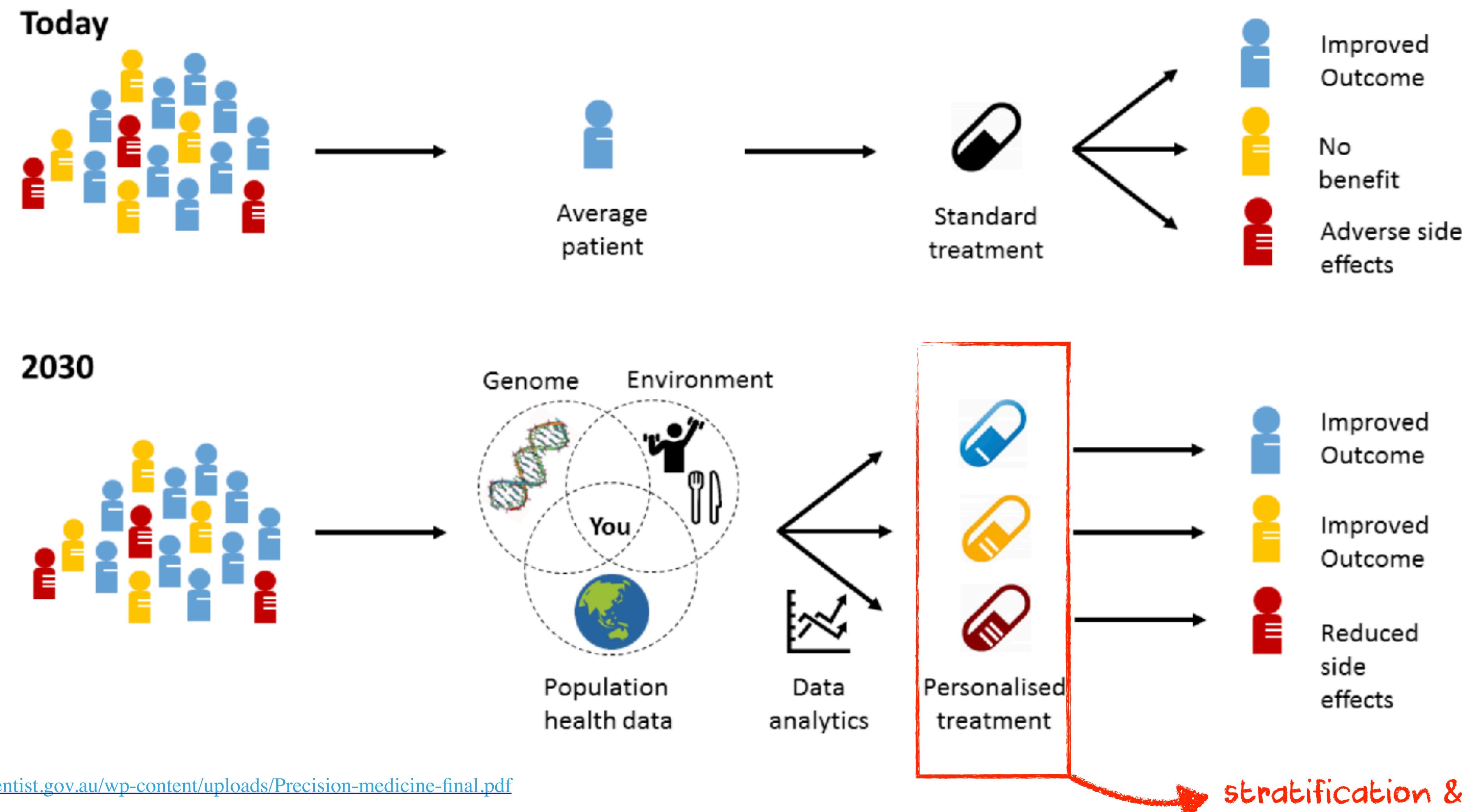
School of Mathematics and Statistics, University of Sydney

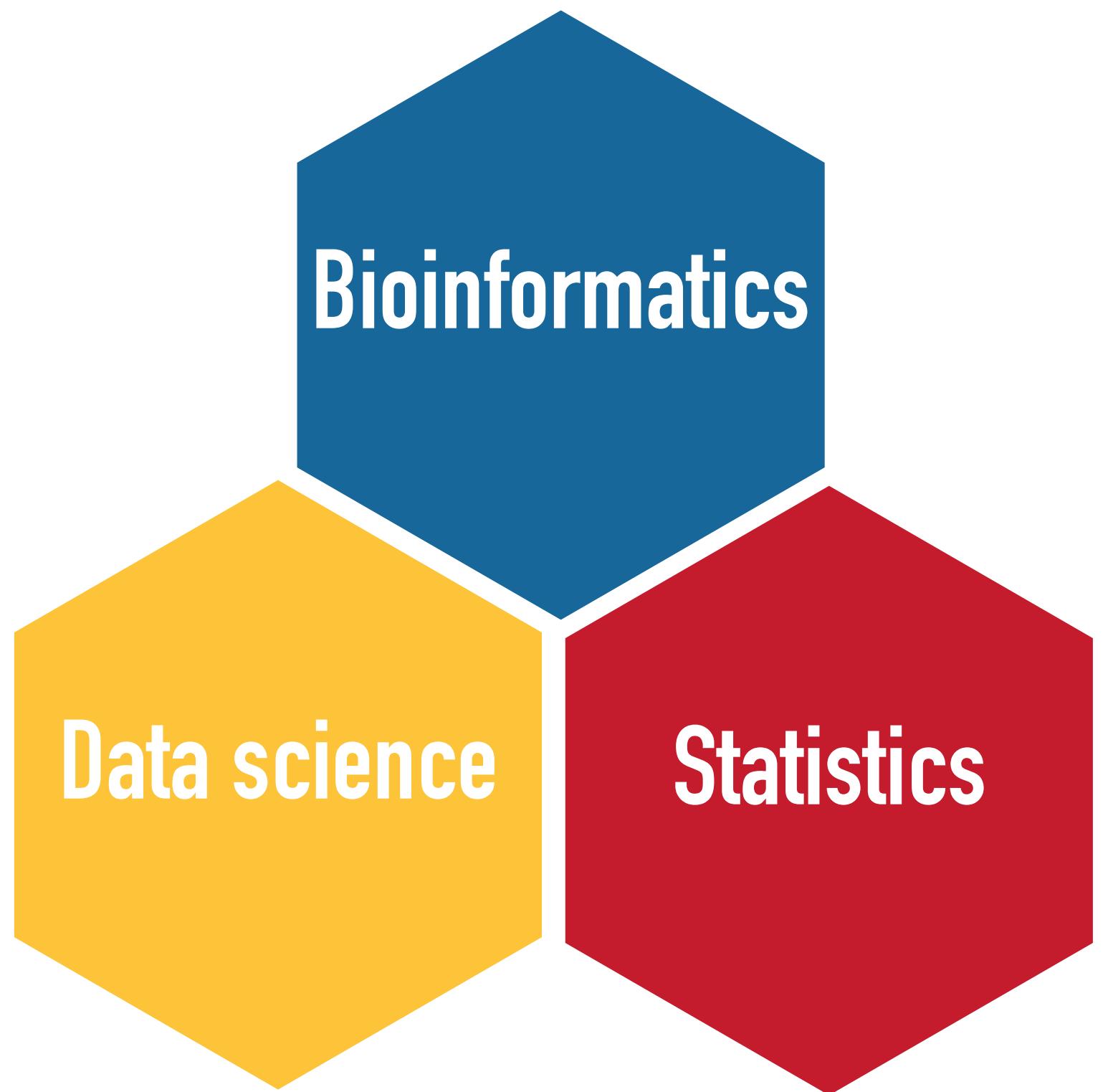
Cross-Platform Omics Prediction: a step towards precision medicine

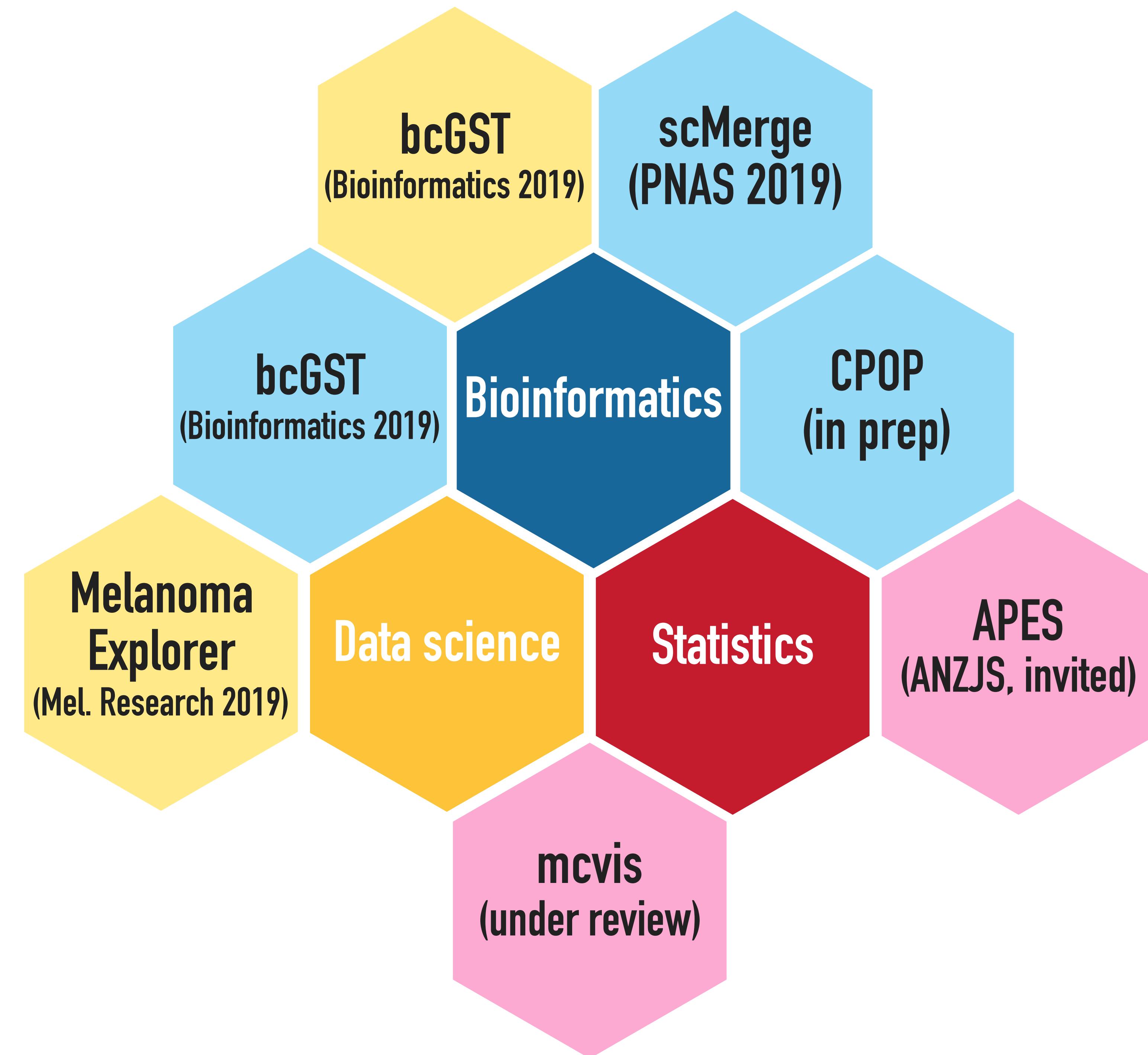
Precision medicine: predicting best cause of action using omics data



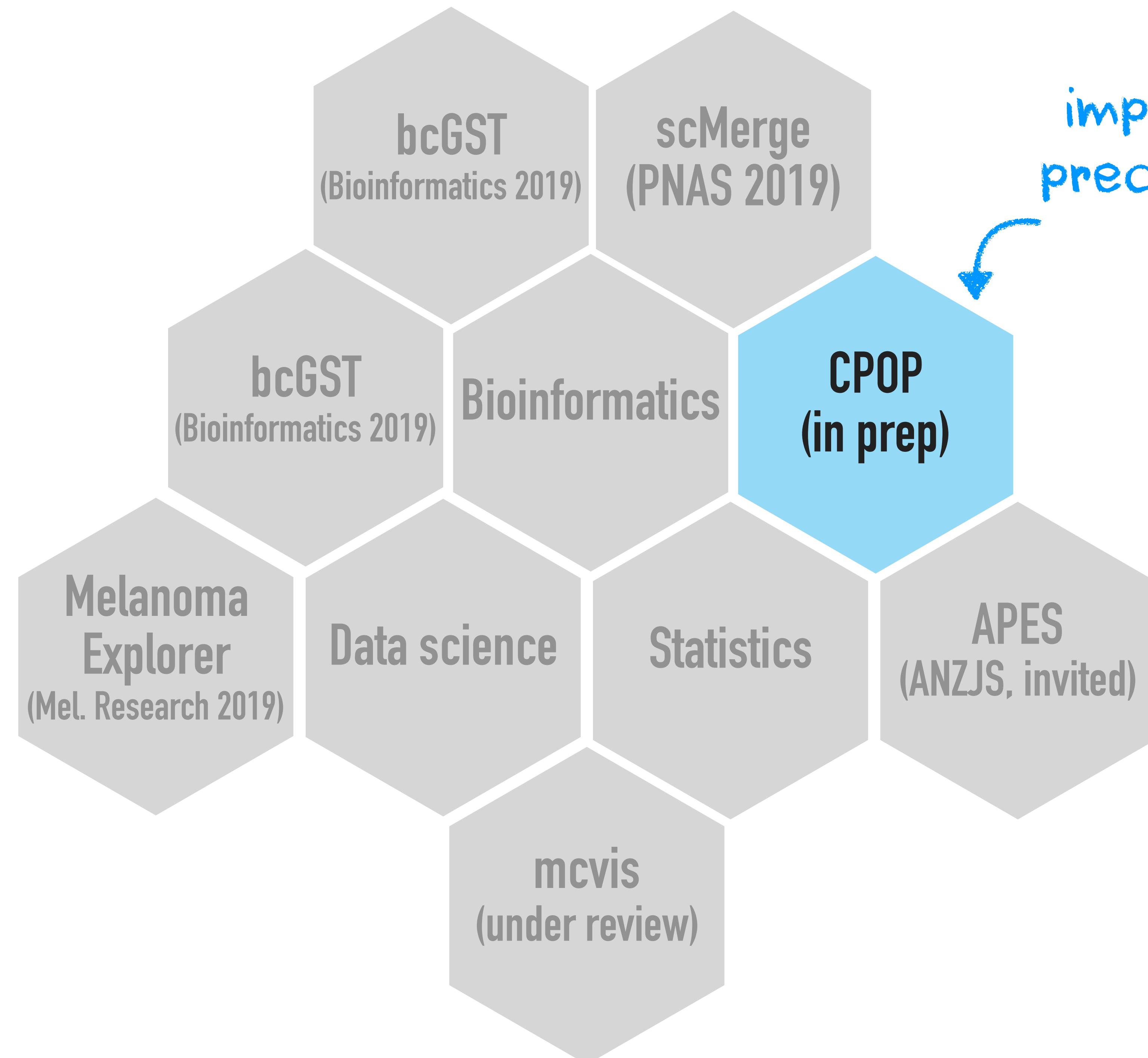
Precision medicine: predicting best cause of action using omics data







Towards
implementation
precision medicine

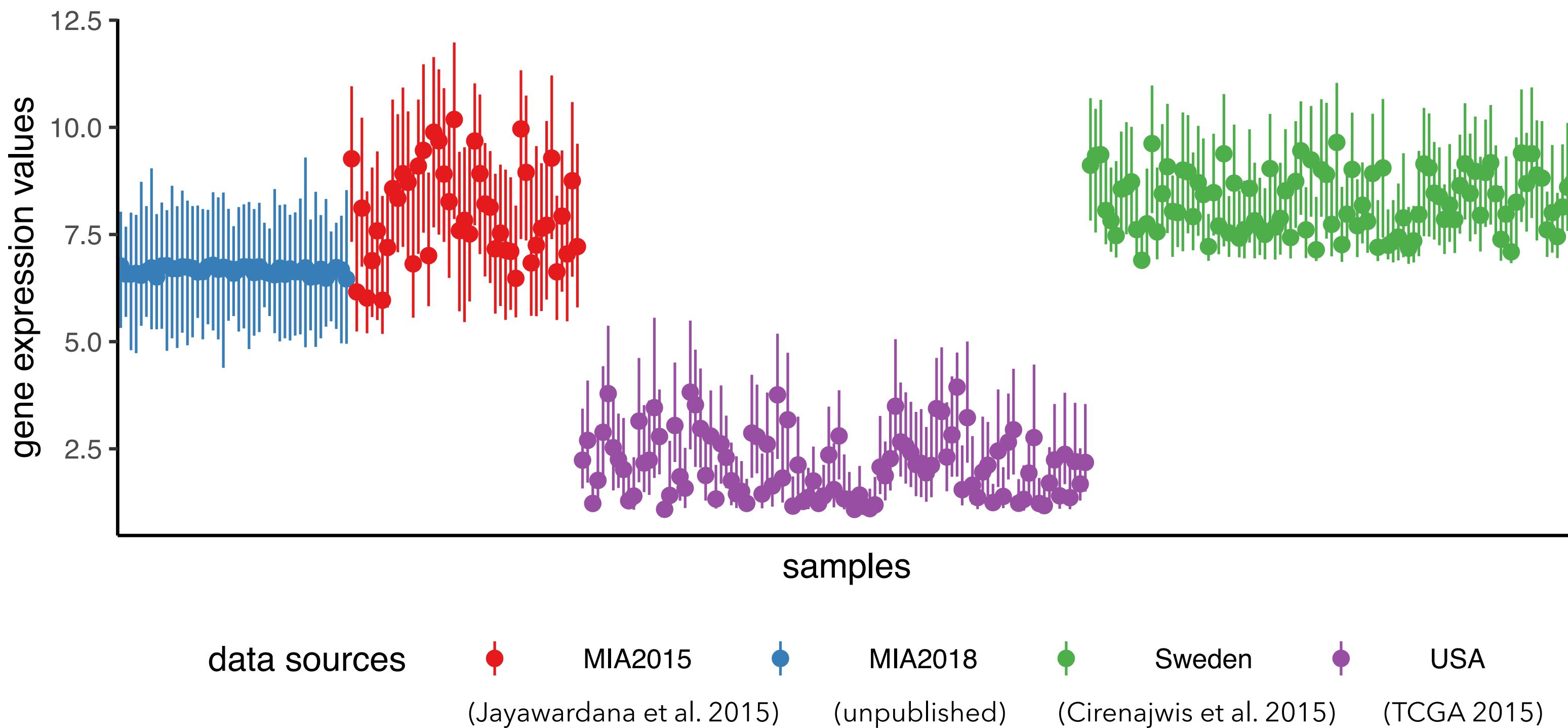


Melanoma Institute Australia

- ▶ Risk score using gene expression

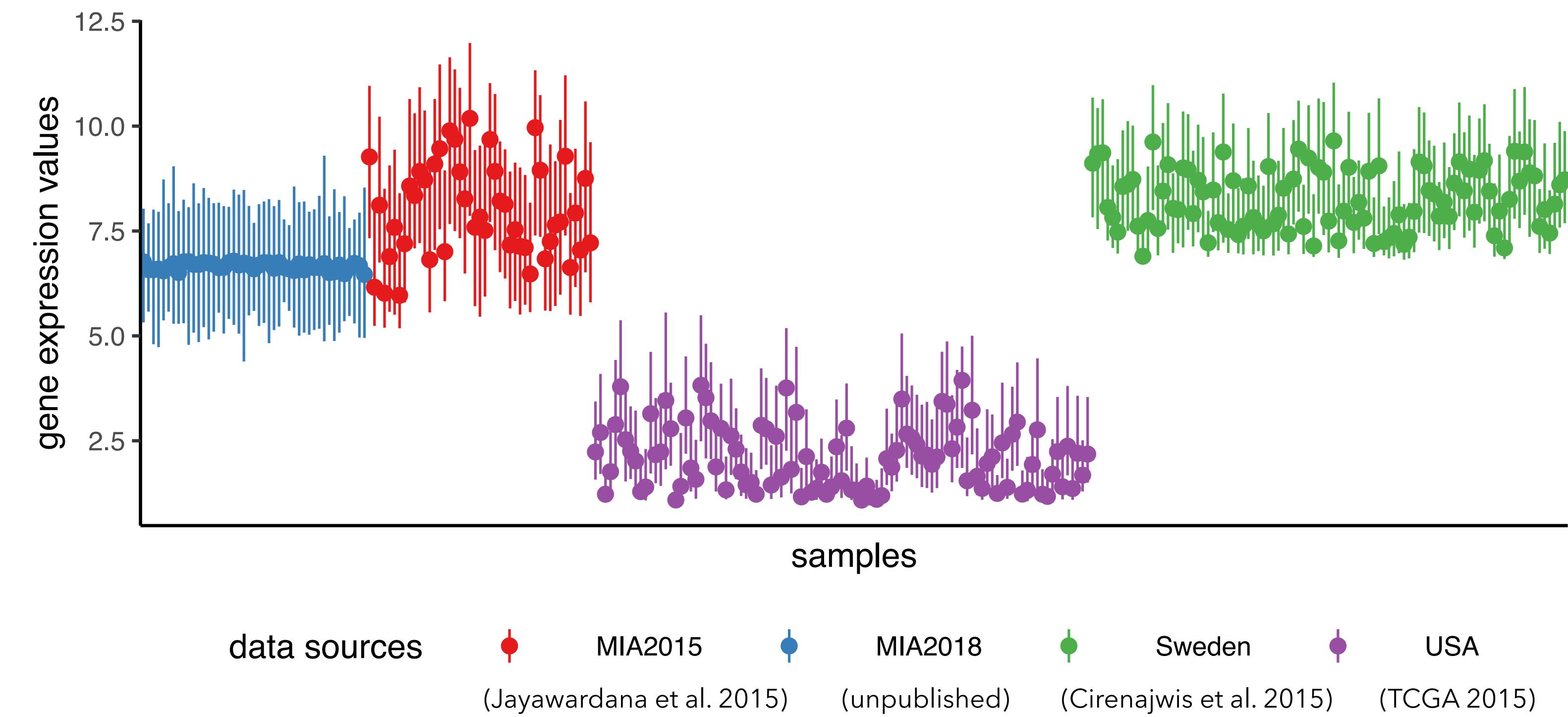
$$\hat{y} = X \hat{\beta}$$

- ▶ CRE grant will support implementation
 - ▶ prospective
 - ▶ **multi-centres**



Prof. Graham Mann

Omics-based clinical risk score: what is so difficult?



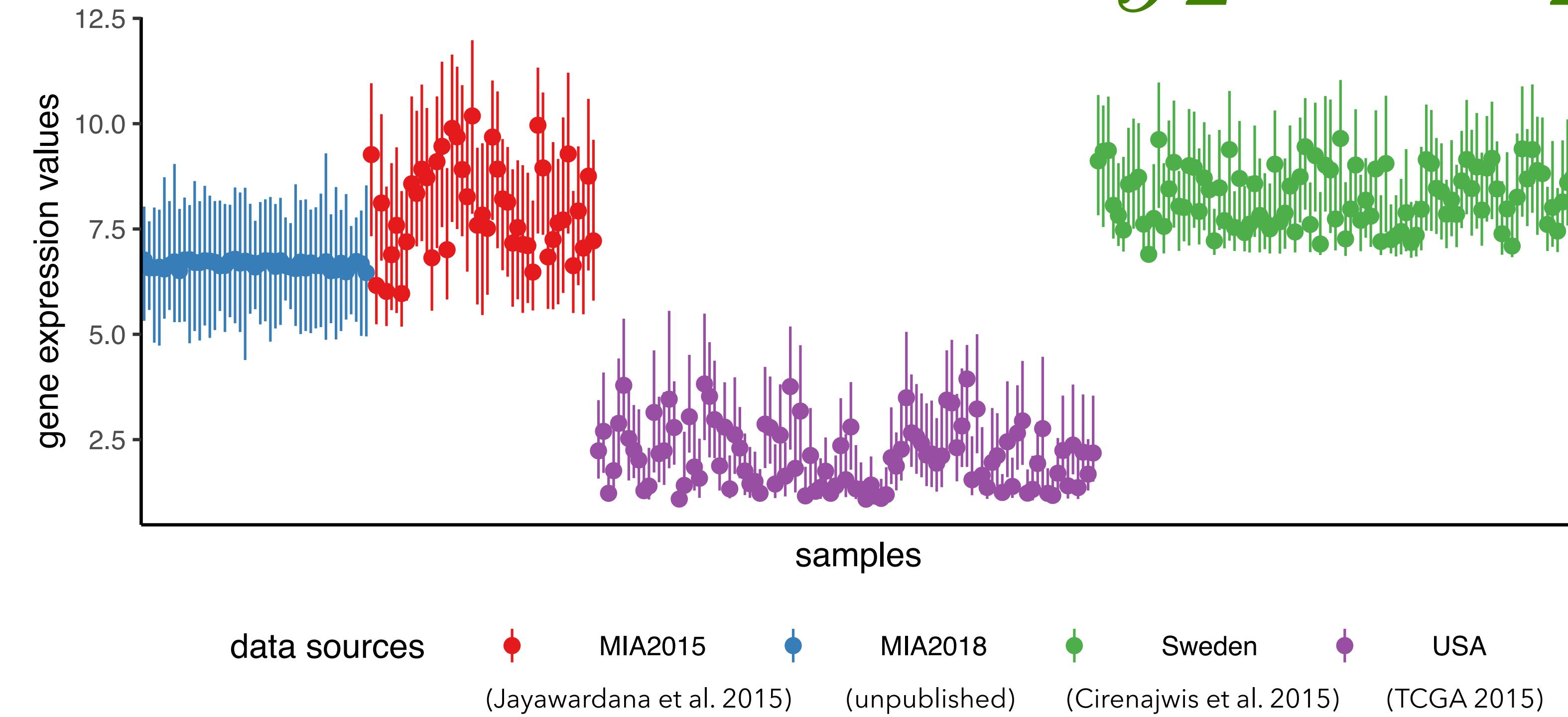
Gene expression features on a unit-less scale.

A typical value in one data can be an impossible value on another.

Omics-based clinical risk score: what is so difficult?

$$\hat{y}_1 = X_1 \hat{\beta}_1$$

$$\hat{y}_2 = X_2 \hat{\beta}_1$$



Gene expression features on a unit-less scale.

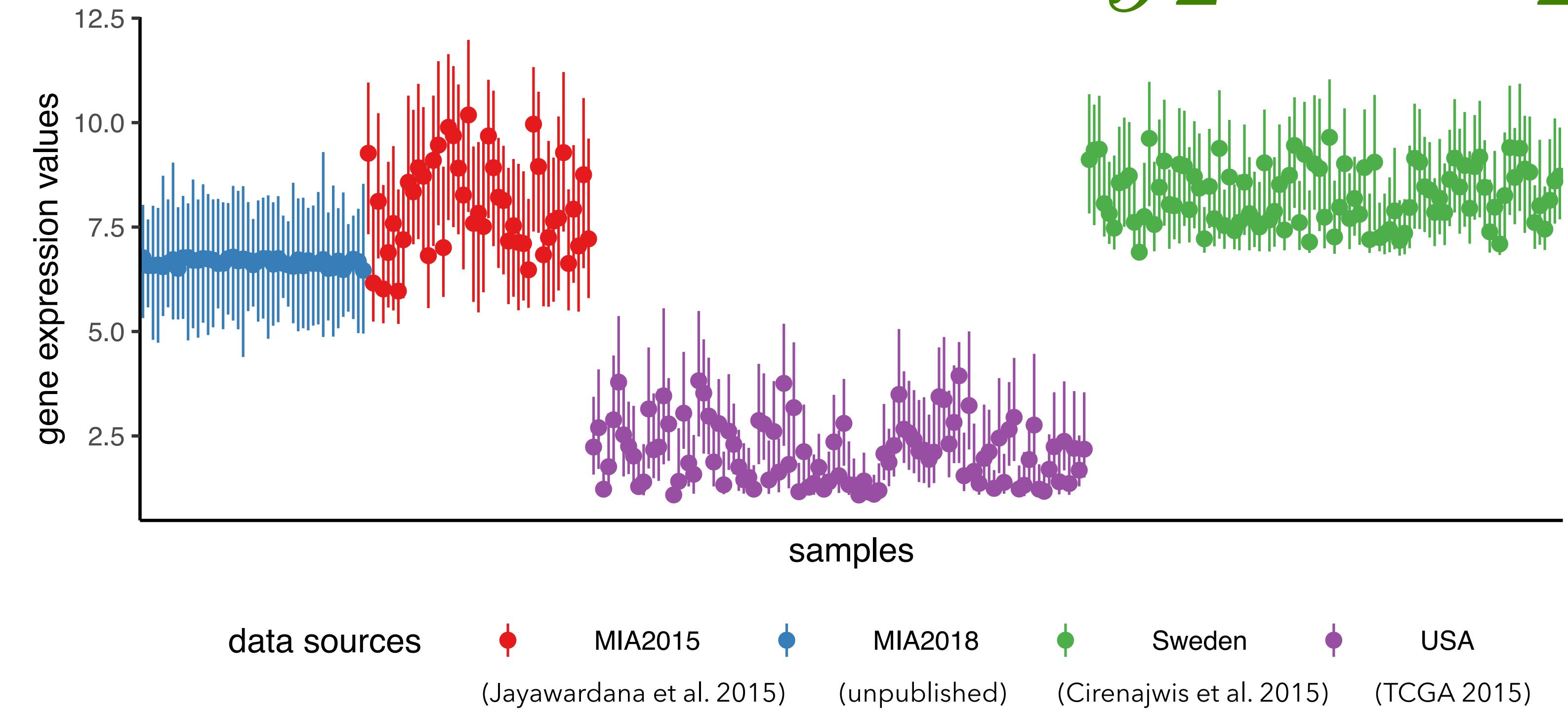
A typical value in one data can be an impossible value on another.

Omics-based clinical risk score: what is so difficult?

$$\hat{y}_1 = X_1 \hat{\beta}_1$$

$$\hat{y}_2 = X_2 \hat{\beta}_1 = (X_1 + 1) \hat{\beta}_1$$

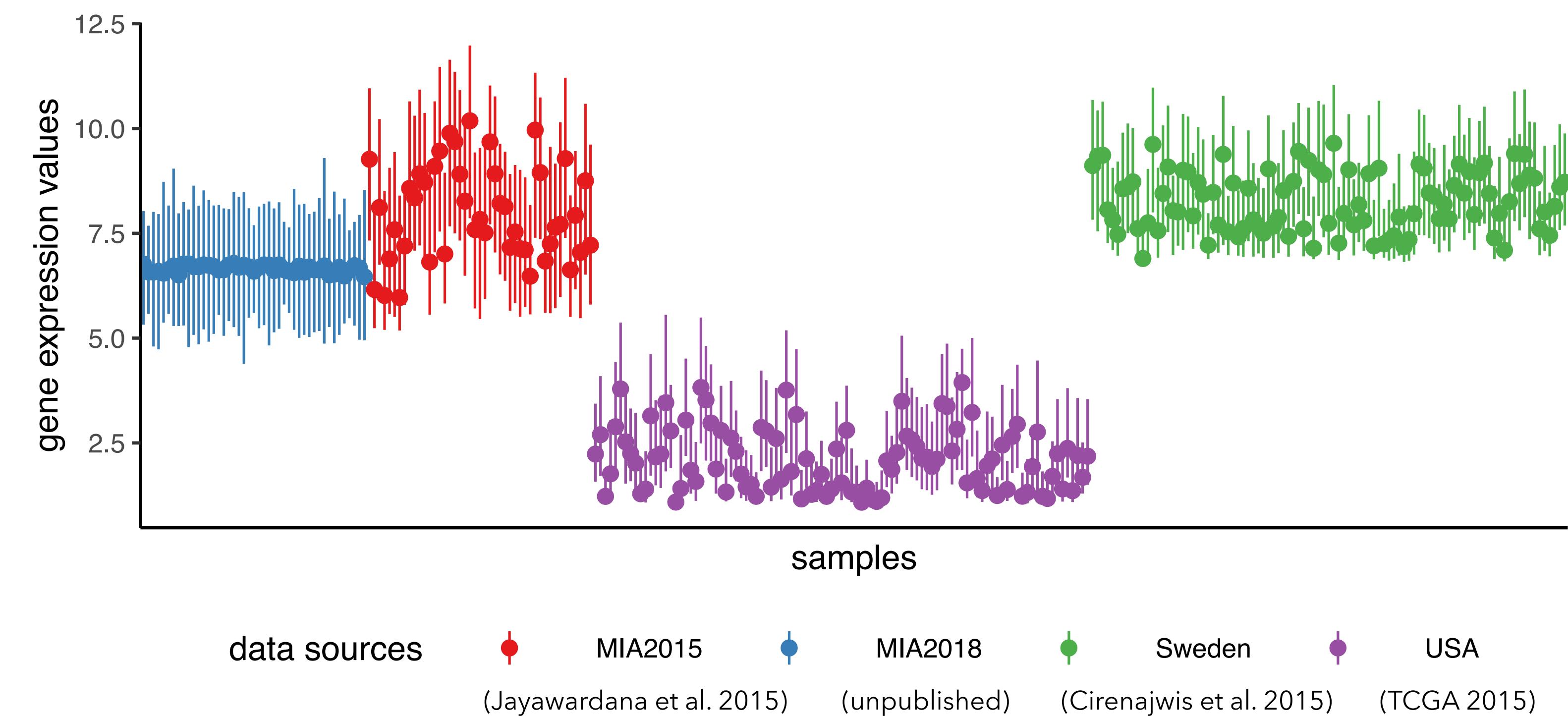
Assuming a
noiseless shift



Gene expression features on a
unit-less scale.

A typical value in one data can
be an **impossible** value on
another.

Omics-based clinical risk score: what is so difficult?



Transferability
A model trained from one
data and should be predictive
on another

Existing approaches

Components of a risk score

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

Prediction

$$\hat{y}_1 = X_1 \hat{\beta}_1$$

$$\hat{y}_2 = X_2 \hat{\beta}_1$$

Data-harmonisation

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

Prediction

$$\hat{y}_1 = X_1 \hat{\beta}_1$$

$$\hat{y}_2 = X_2 \hat{\beta}_1$$

Harmonisation

Classical estimation
methods

Classical model prediction

Data-harmonisation: normalisation or standardisation

Data

(X_1, y_1)

(X_2, y_2)

Harmonisation

1. Prospective: **re-normalisation** upon new single-samples
2. Multi-centres: **re-training** of model upon new populations

Prevented by ethics, privacy and data security

Clinical constraints in implementation

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

Model

$$\hat{\beta}_1$$

$$\hat{y}_1 = X_1 \hat{\beta}_1$$

$$\hat{y}_2 = X_2 \hat{\beta}_1$$

No re-normalisation

No model re-training

Scale-equivalent prediction

**CPOP: predictions that
respect clinical constraints**

CPOP flowchart

Data

$$(X_1, y_1) \rightarrow (\textcolor{red}{Z}_1, y_1)$$

$$(X_2, y_2) \rightarrow (\textcolor{red}{Z}_2, y_2)$$

Model

$$\hat{\beta}_1 \approx \hat{\beta}_2$$

Prediction

$$\textcolor{red}{Z}_1 \hat{\beta}_1 \approx Z_1 \hat{\beta}_2$$

$$Z_2 \hat{\beta}_1 \approx \textcolor{red}{Z}_2 \hat{\beta}_2$$

Feature transform

Stable estimation

Stable prediction

First component of CPOP: feature transform



Log-ratio transformation: modelling relative gene expression

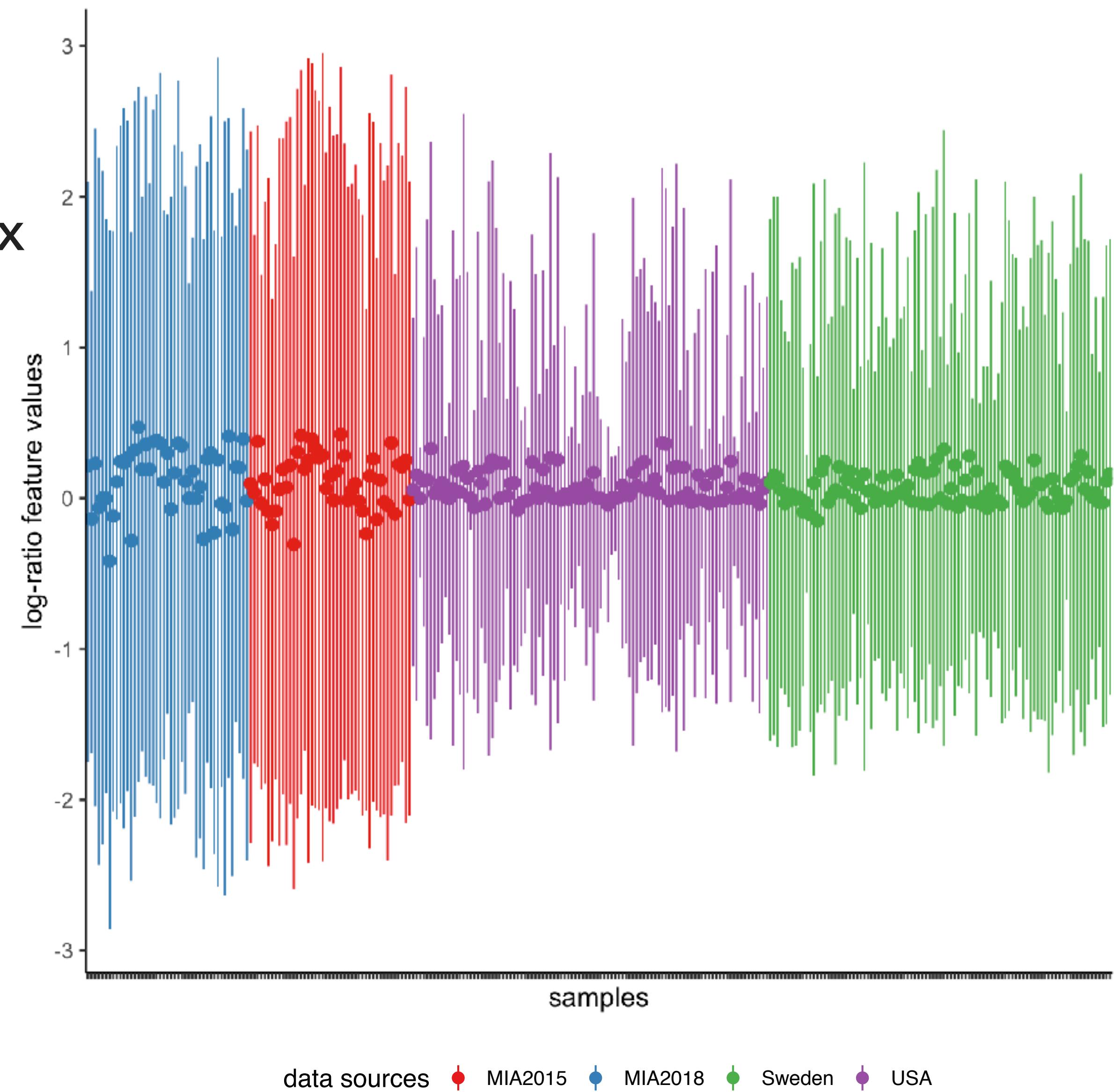
For each column in the gene expression matrix

$$X = \{x_1, \dots, x_p\} \in \mathbb{R}^{n \times p}$$

Construct $Z \in \mathbb{R}^{n \times \binom{p}{2}}$ column-wise:

$$z_j = \log\left(\frac{x_l}{x_m}\right)$$

$$\text{for } 1 \leq l < m \leq p$$

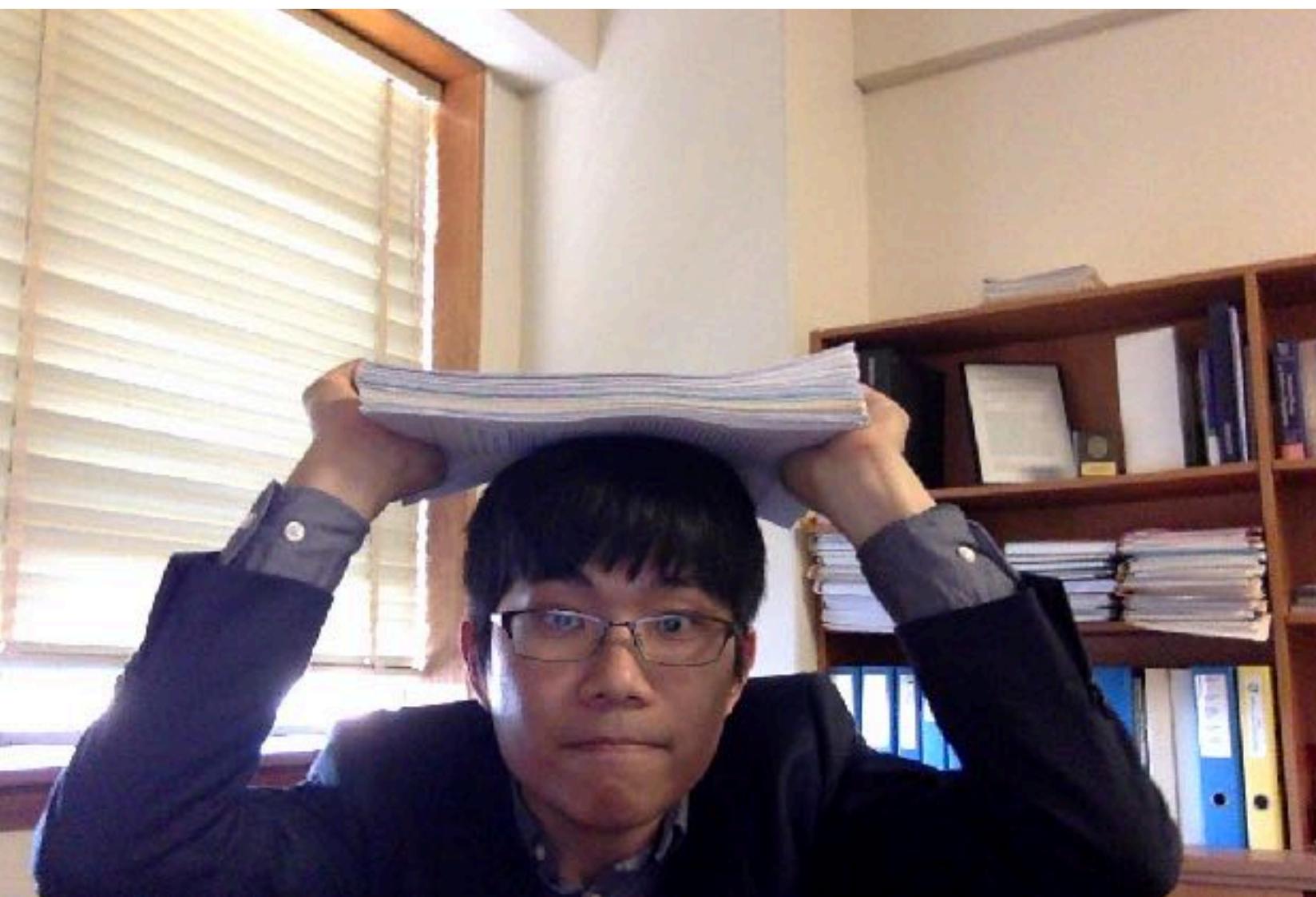


Why log-ratios?

- ▶ Within-sample standardisation avoids re-normalisation and model re-training

Why log-ratios?

- ▶ Within-sample standardisation avoids re-normalisation and model re-training
- ▶ Under-used in clinical implementation
- ▶ Potential for further method developments

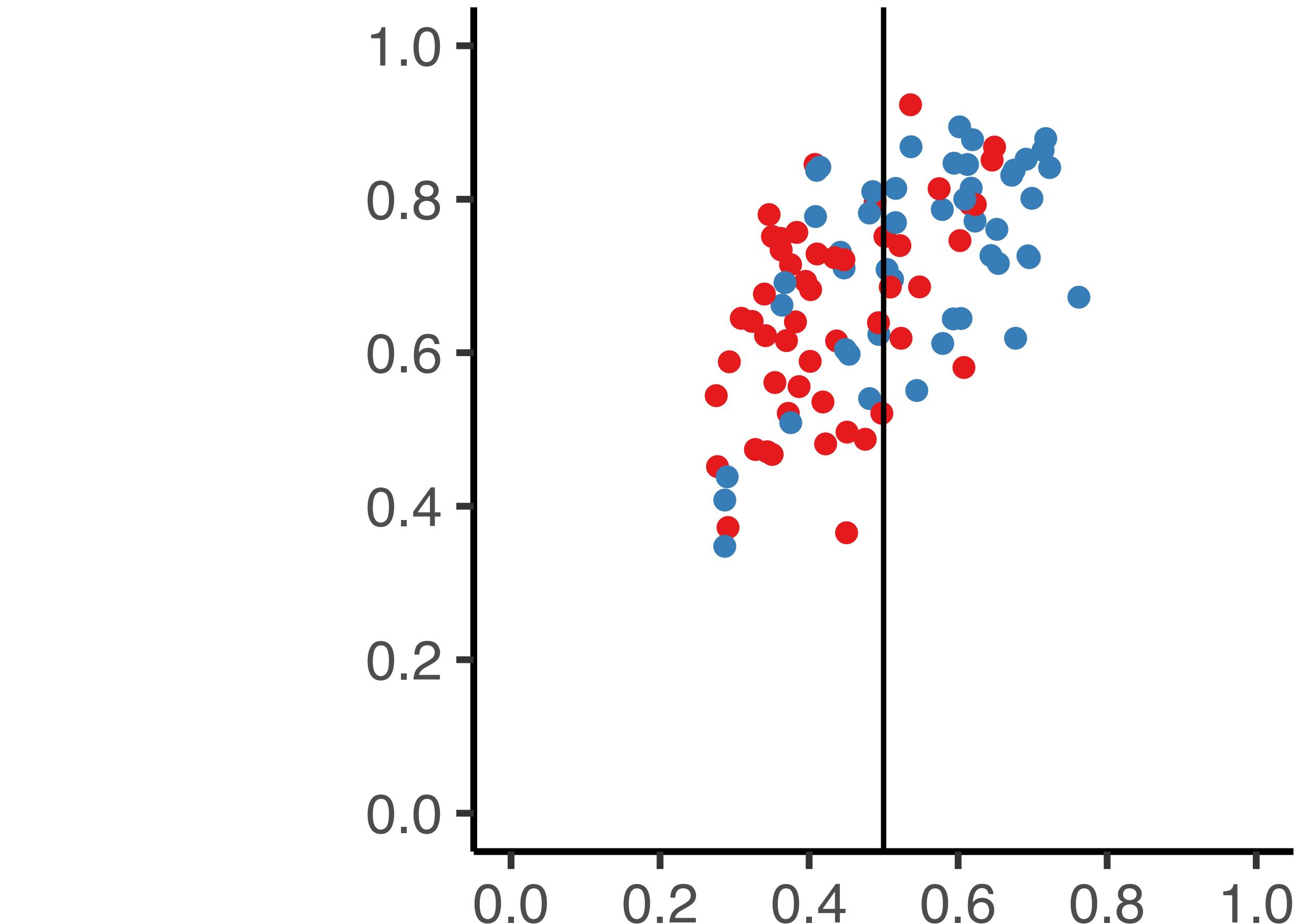


Papers that use gene expression



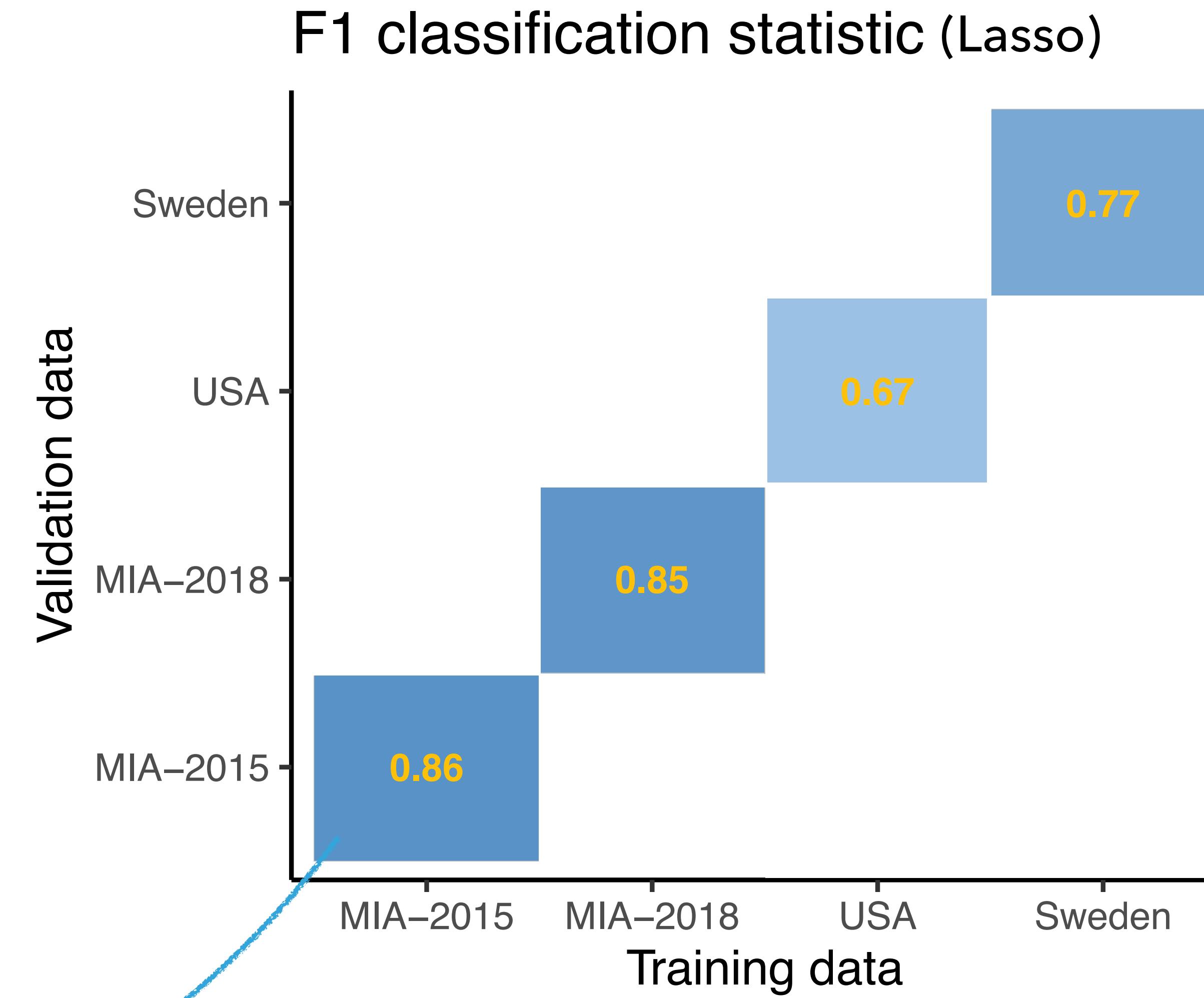
Papers that use ratios

Is log-ratio enough?

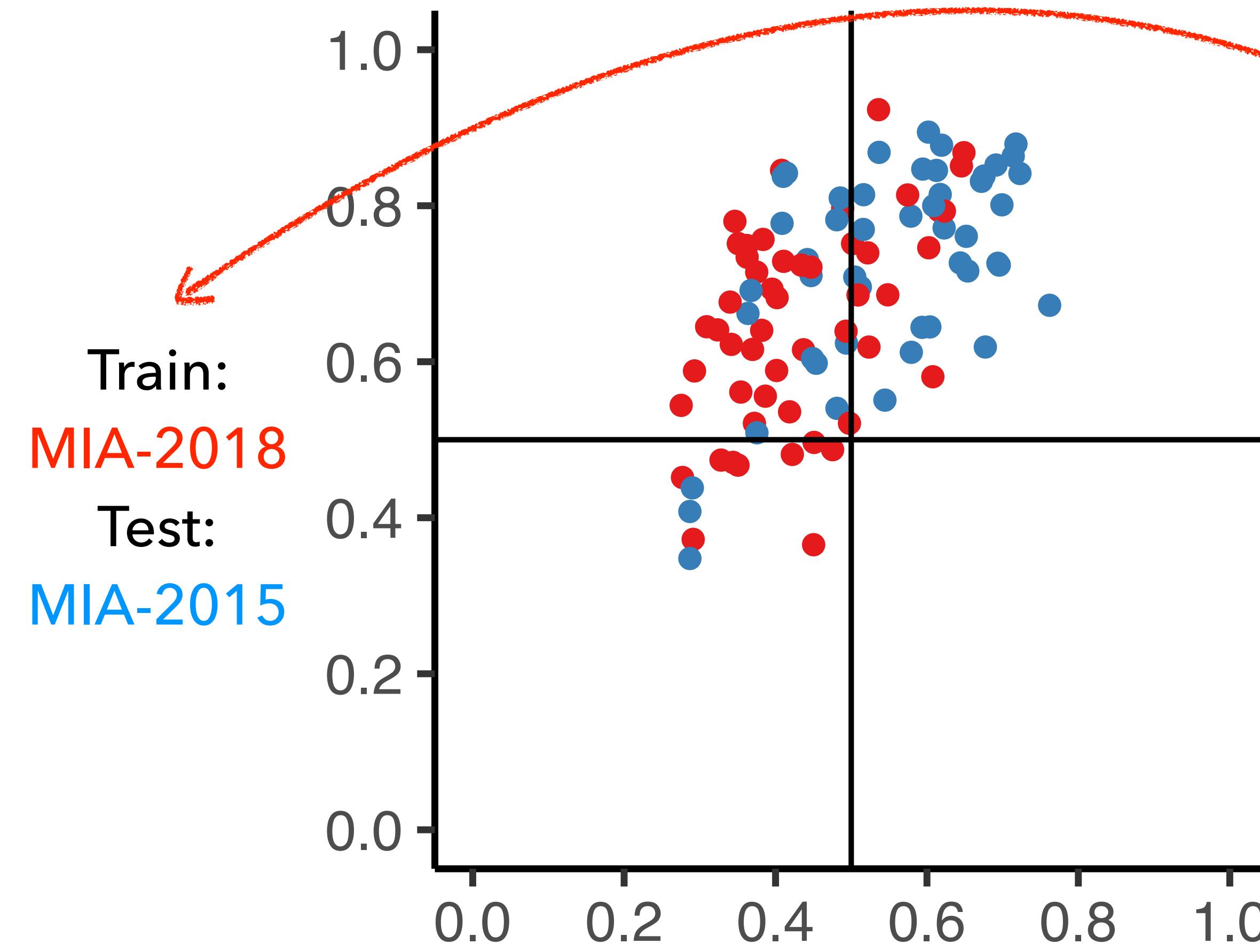


Sample class ● Good ● Poor

*Predicted values are shown as classification probability

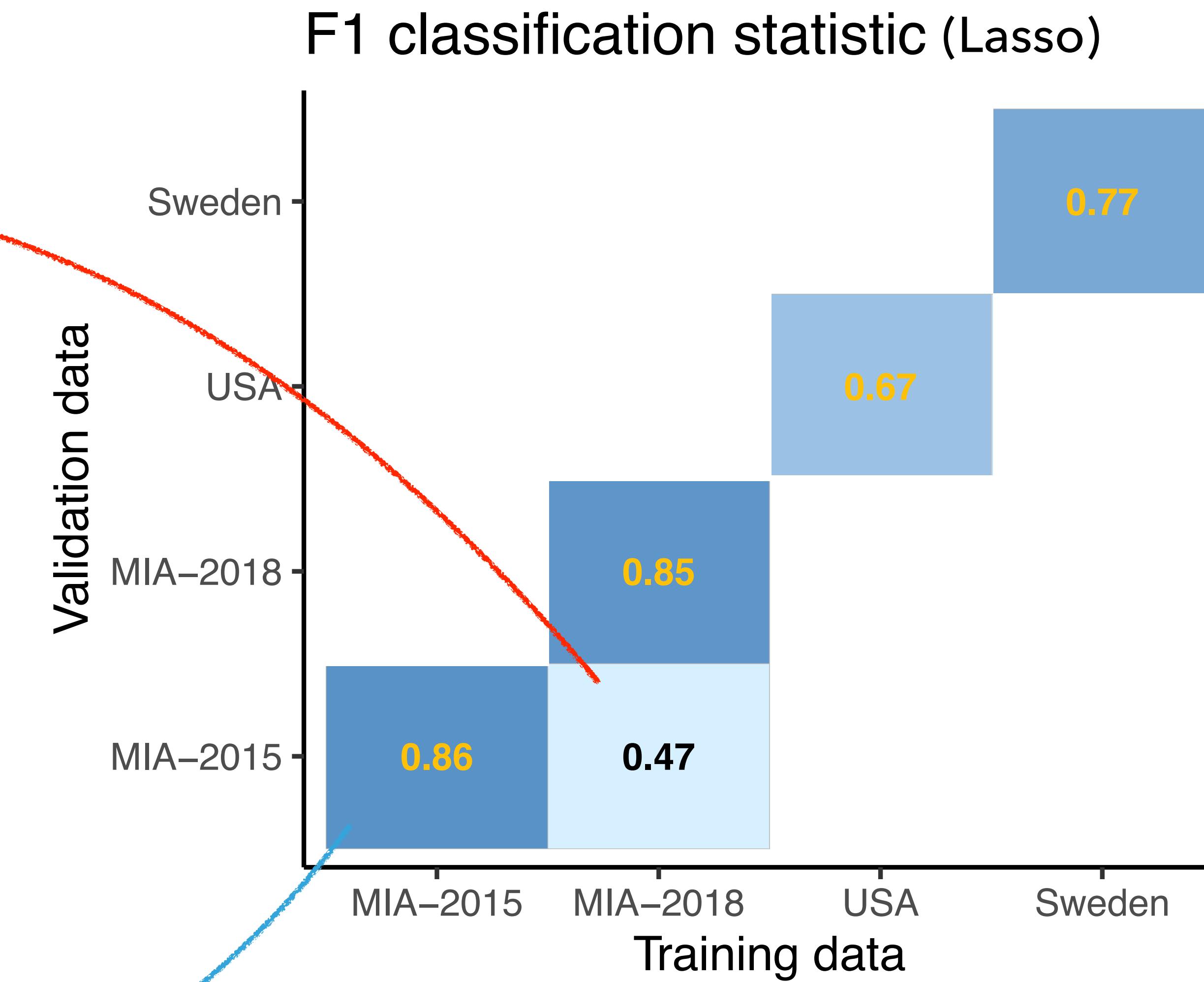


Is log-ratio enough?

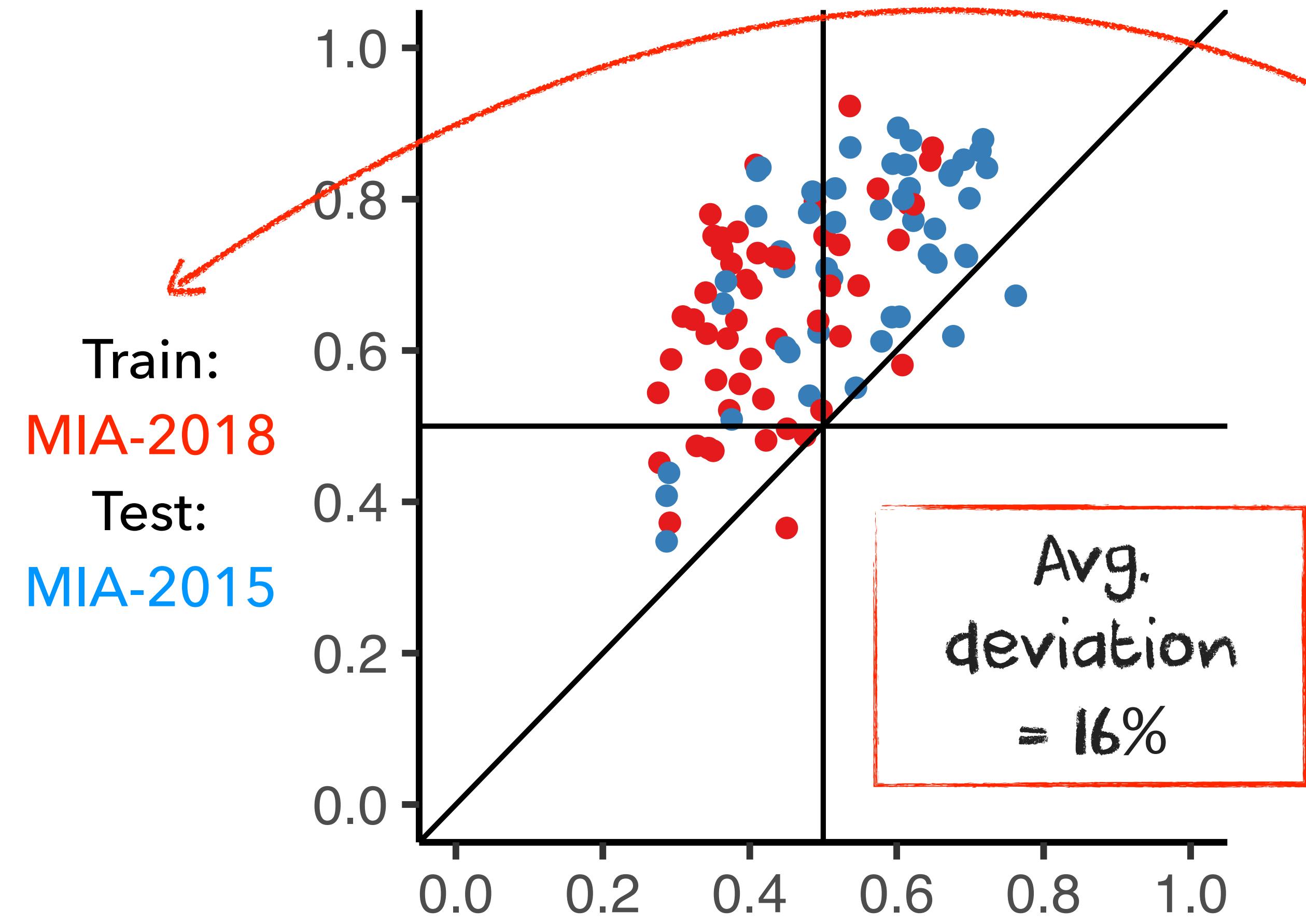


*Predicted values are shown as
classification probability

Train: MIA-2015
Test: MIA-2015

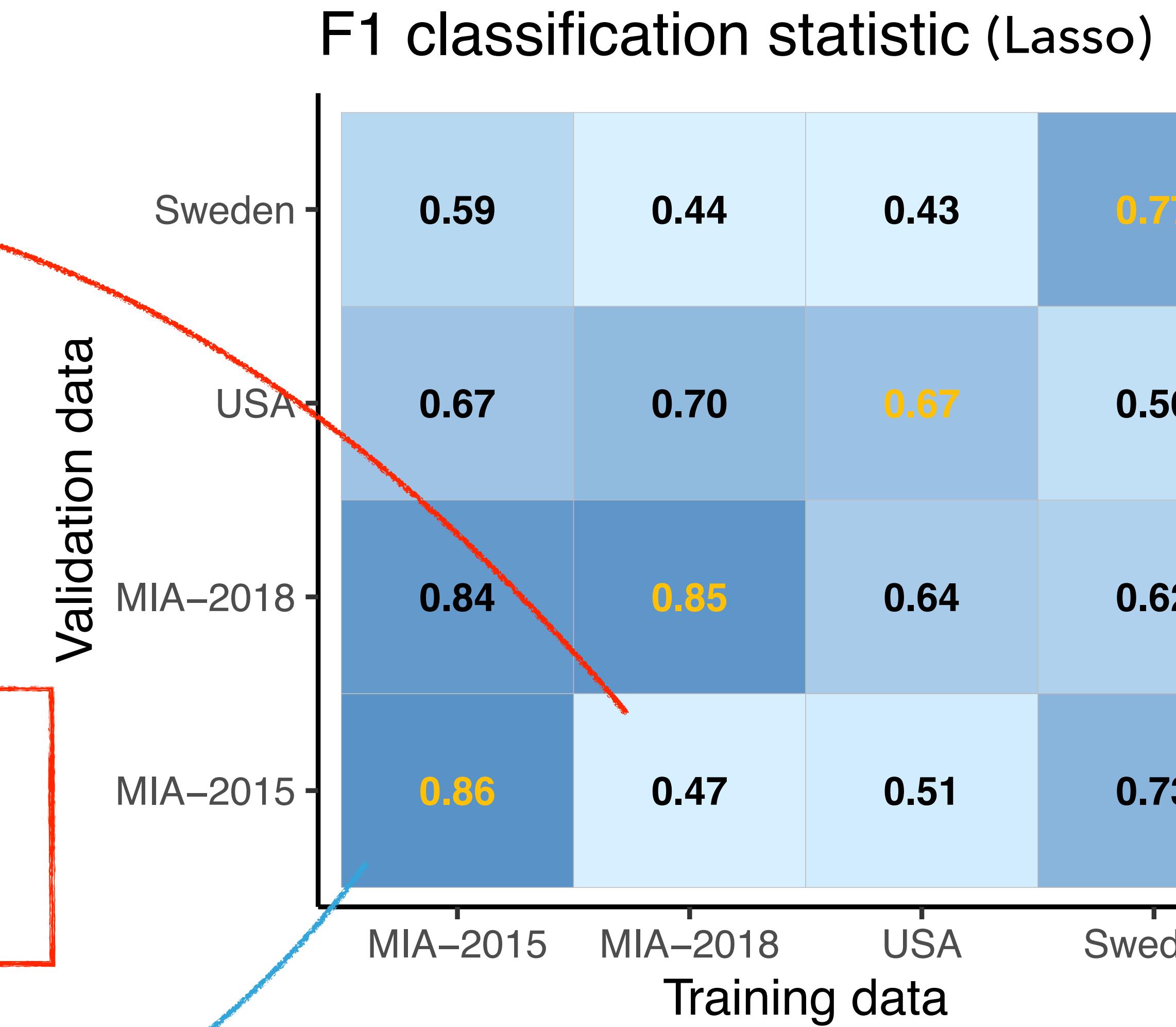


Is log-ratio enough?



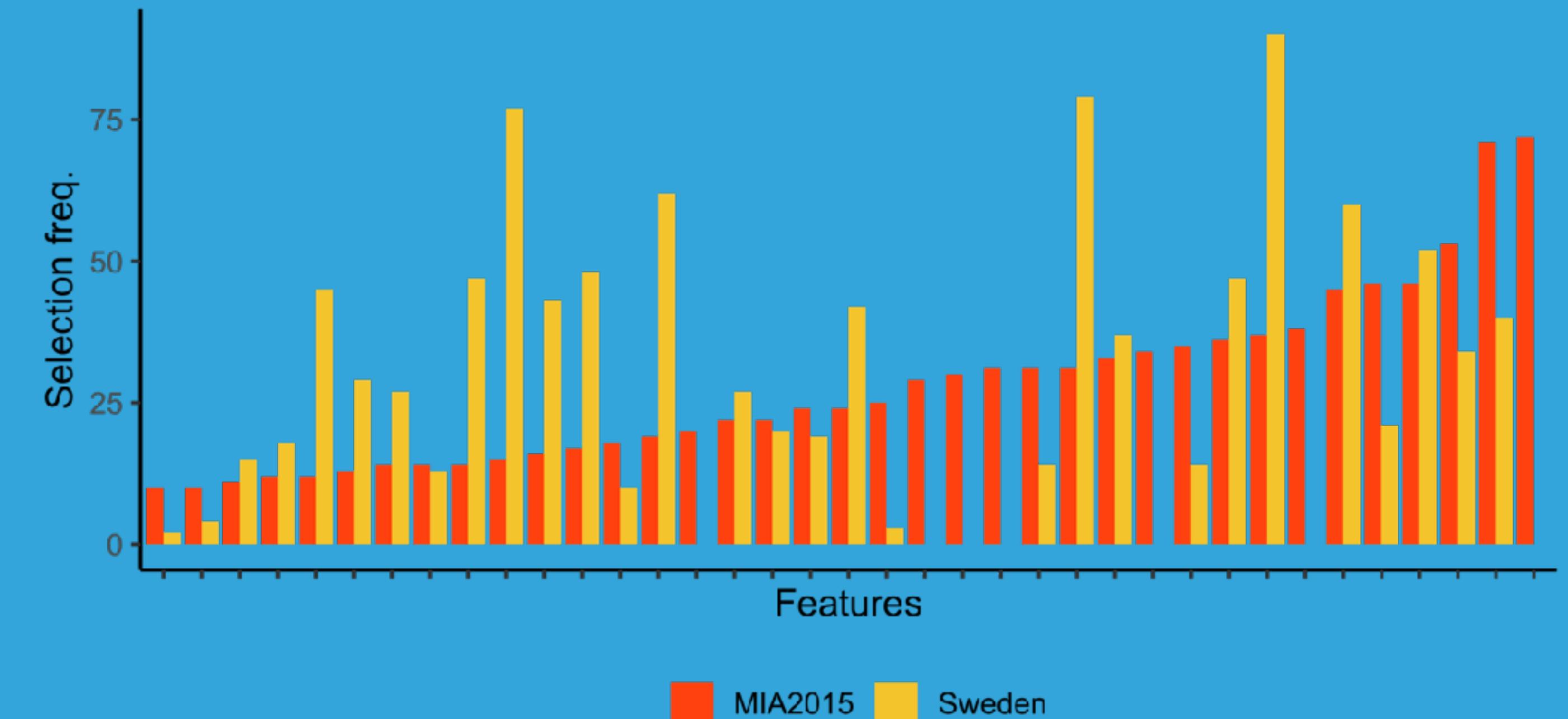
*Predicted values are shown as classification probability

Train: MIA-2015
Test: MIA-2015

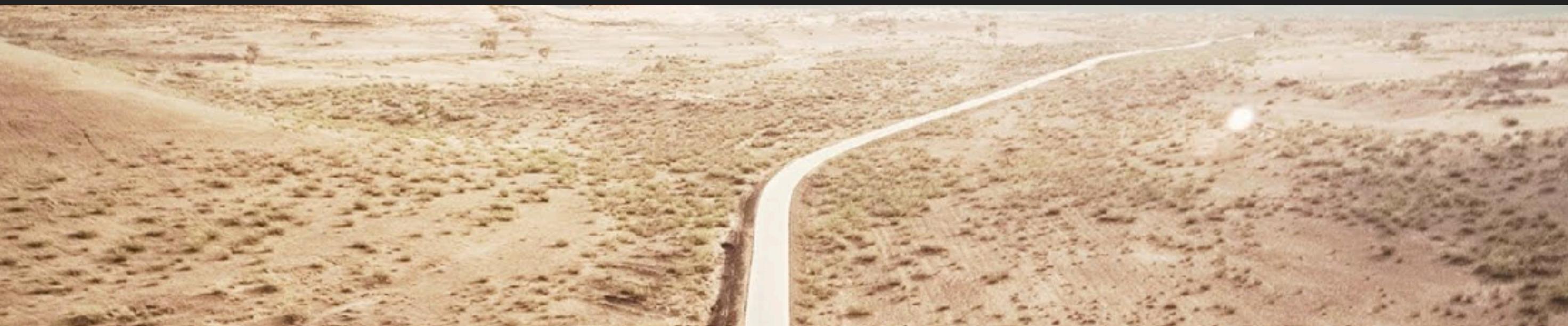


Transferability implies that predicted values should be evenly scattered around the identity line!

Lasso variable selection is not stable!



Second component of CPOP: stable feature selection and estimation



CPOP flowchart

Data

$$(X_1, y_1) \rightarrow (Z_1, y_1)$$

$$(X_2, y_2) \rightarrow (Z_2, y_2)$$

Model

$$\hat{\beta}_1 \approx \hat{\beta}_2$$

Prediction

$$Z_1 \hat{\beta}_1 \approx Z_1 \hat{\beta}_2$$

$$Z_2 \hat{\beta}_1 \approx Z_2 \hat{\beta}_2$$

Feature transform

Stable estimation

Stable prediction

Weighted Elastic Net

logistic loss
function

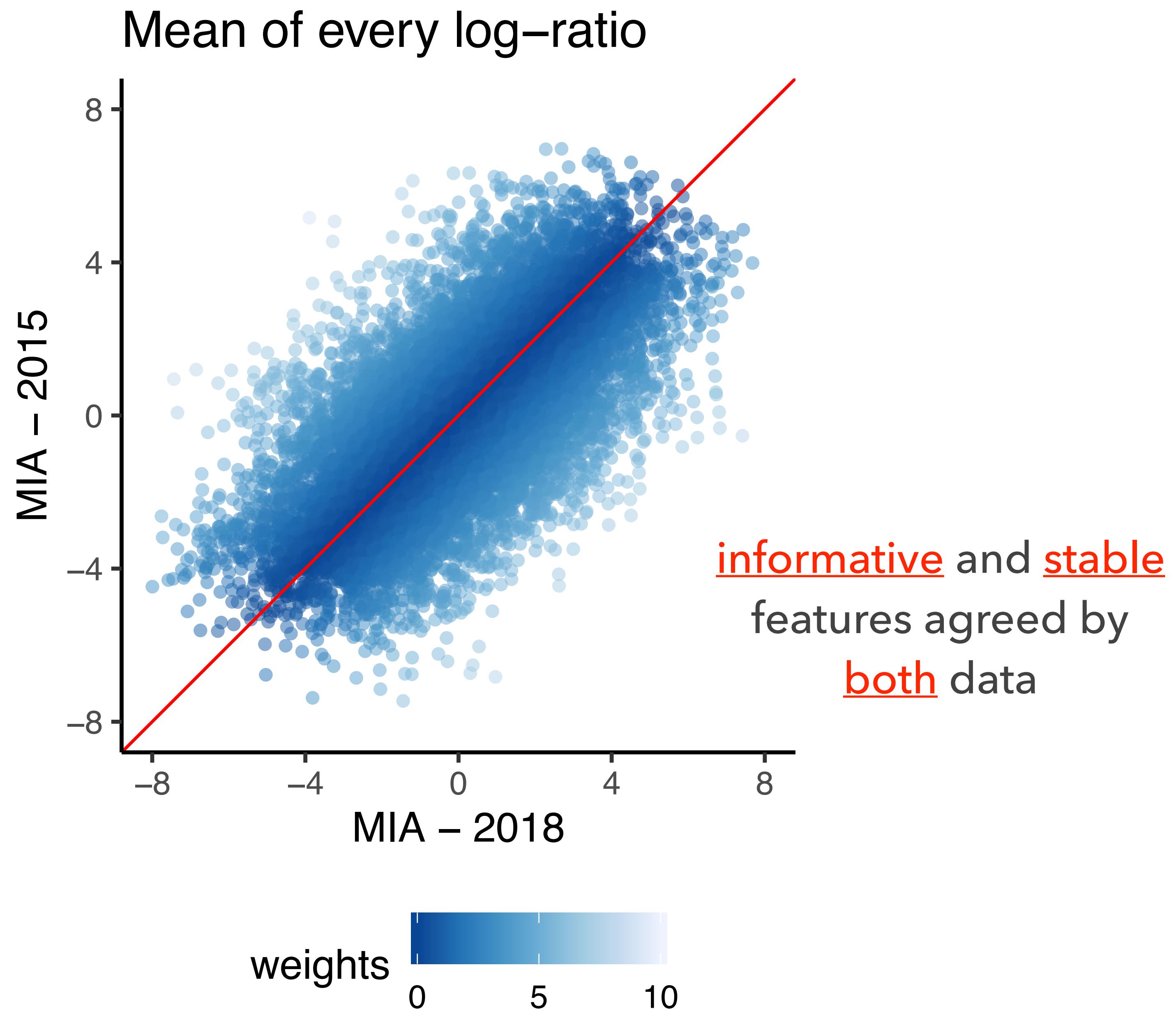
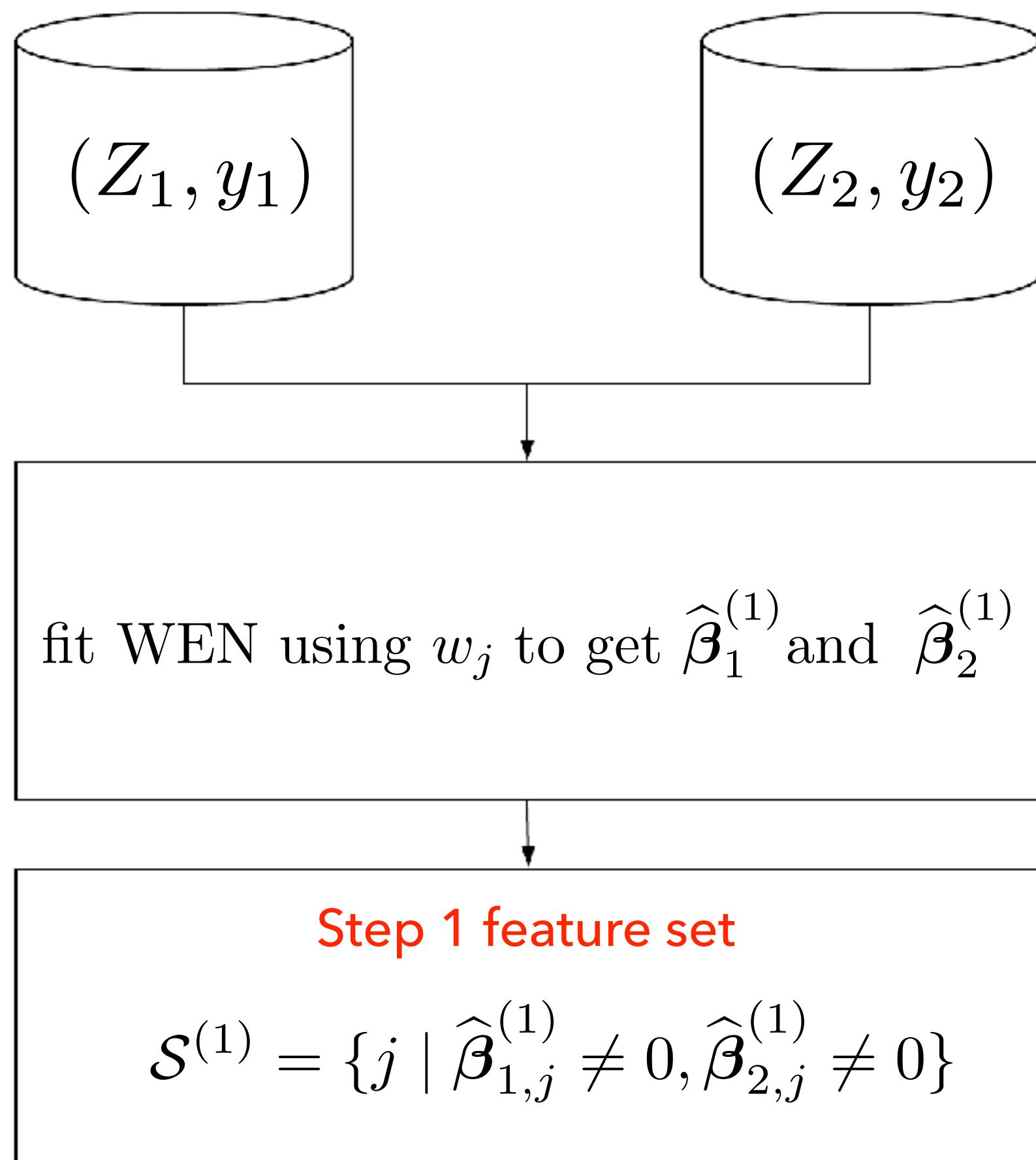
$$\hat{\beta}(y, Z) = \underset{\beta \in \mathbb{R}^{\binom{p}{2}}}{\operatorname{argmin}} \sum_{i=1}^n l(y_i, z_i^\top \beta) + \lambda \sum_{j=1}^q w_j \left[\frac{(1-\alpha)}{2} \beta_j^2 + \alpha |\beta_j| \right]$$

Modelling on log-
ratio features

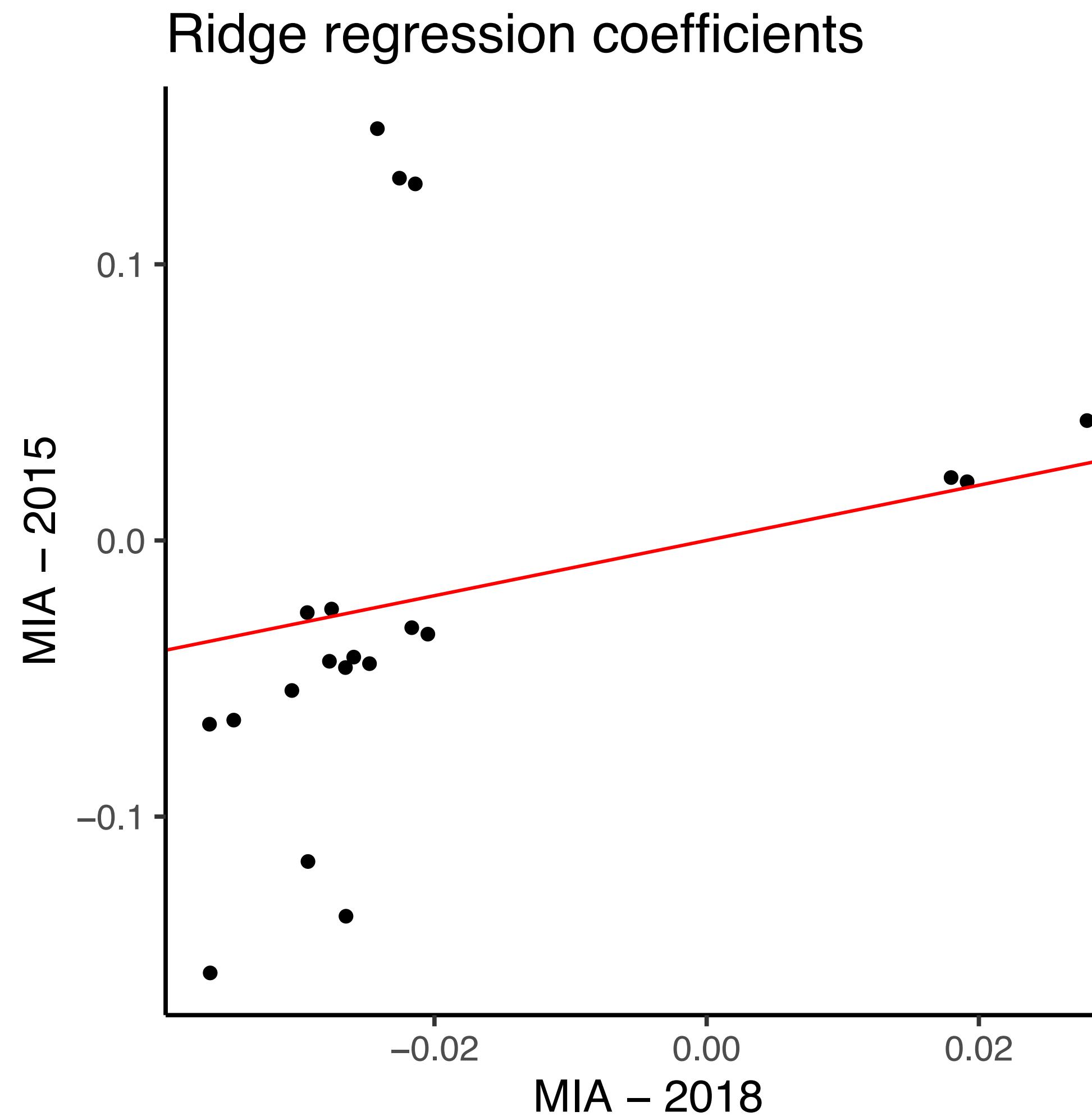
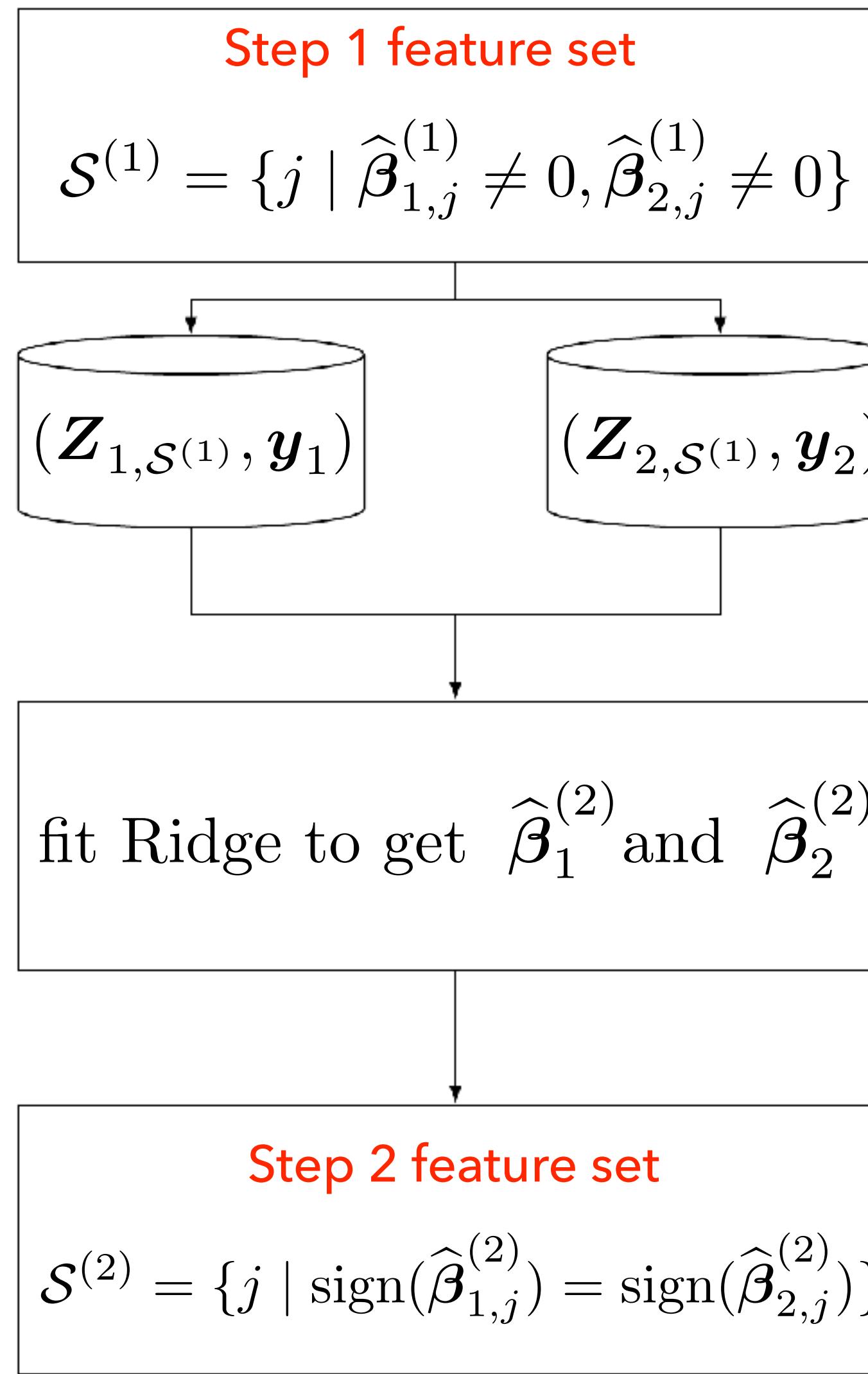
A mix of L1 and
L2 penalties

Weights on each feature,
proportional to the stability of
features

Step 1: feature selection stability

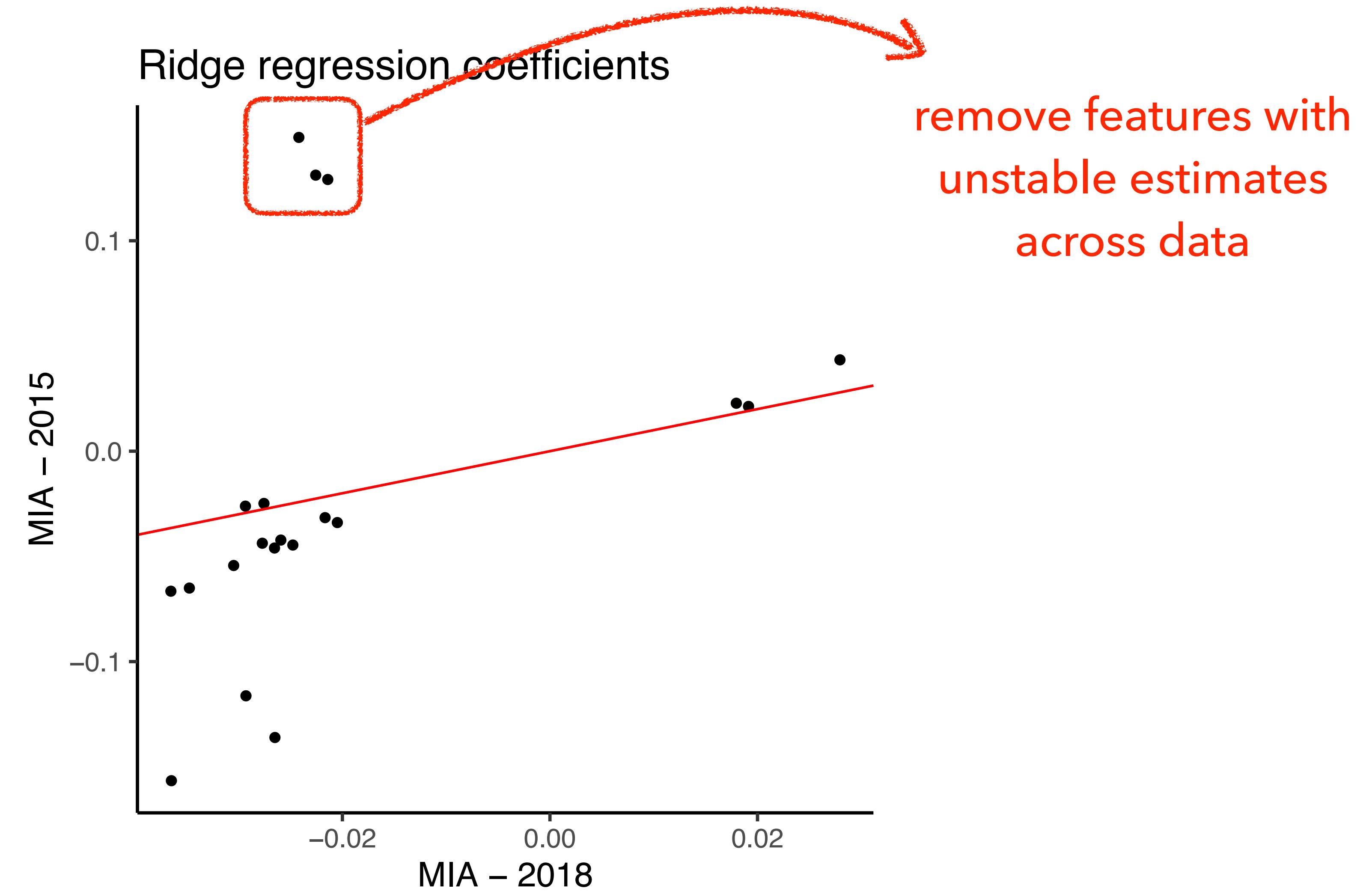
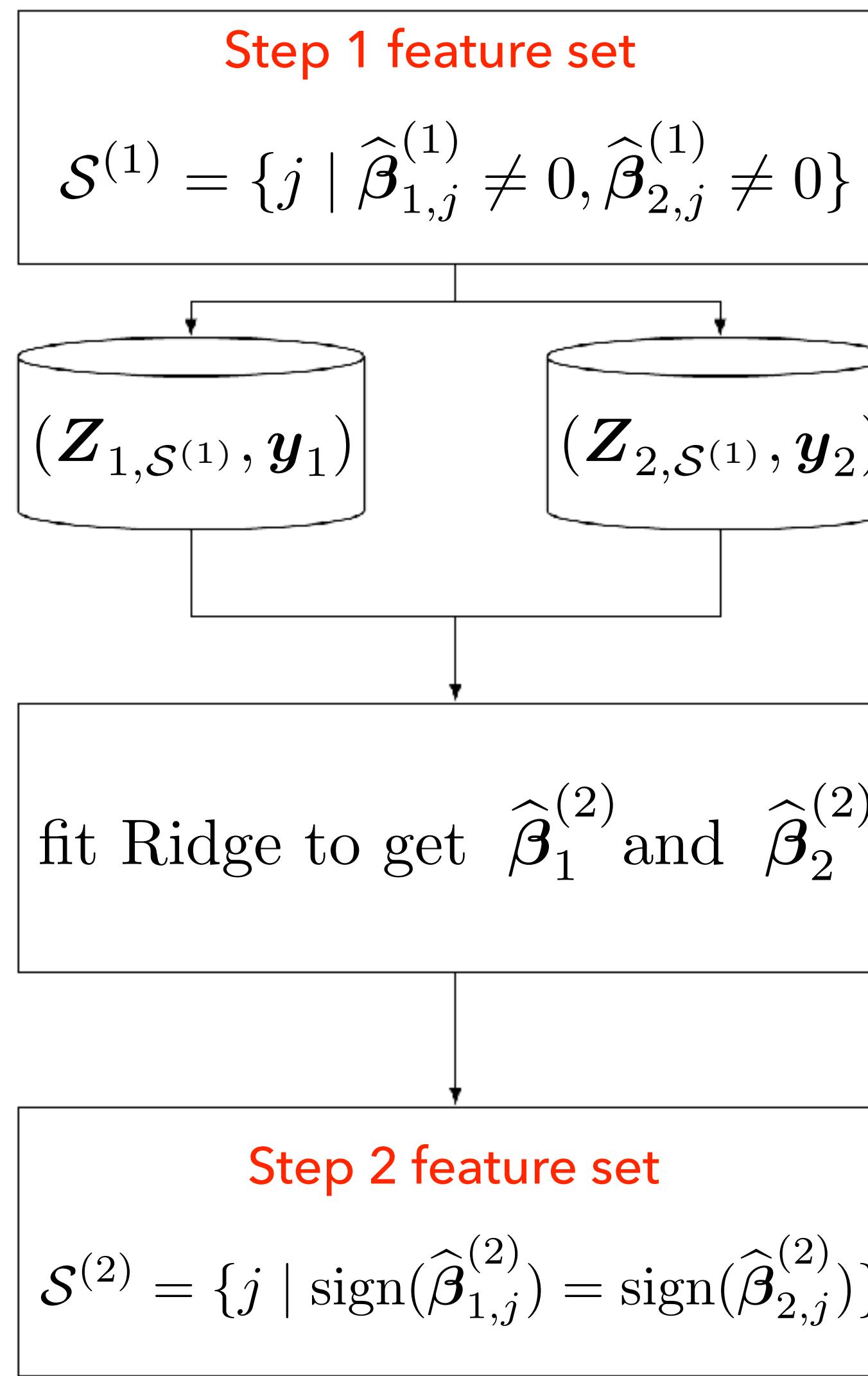


Step 2: feature estimation stability

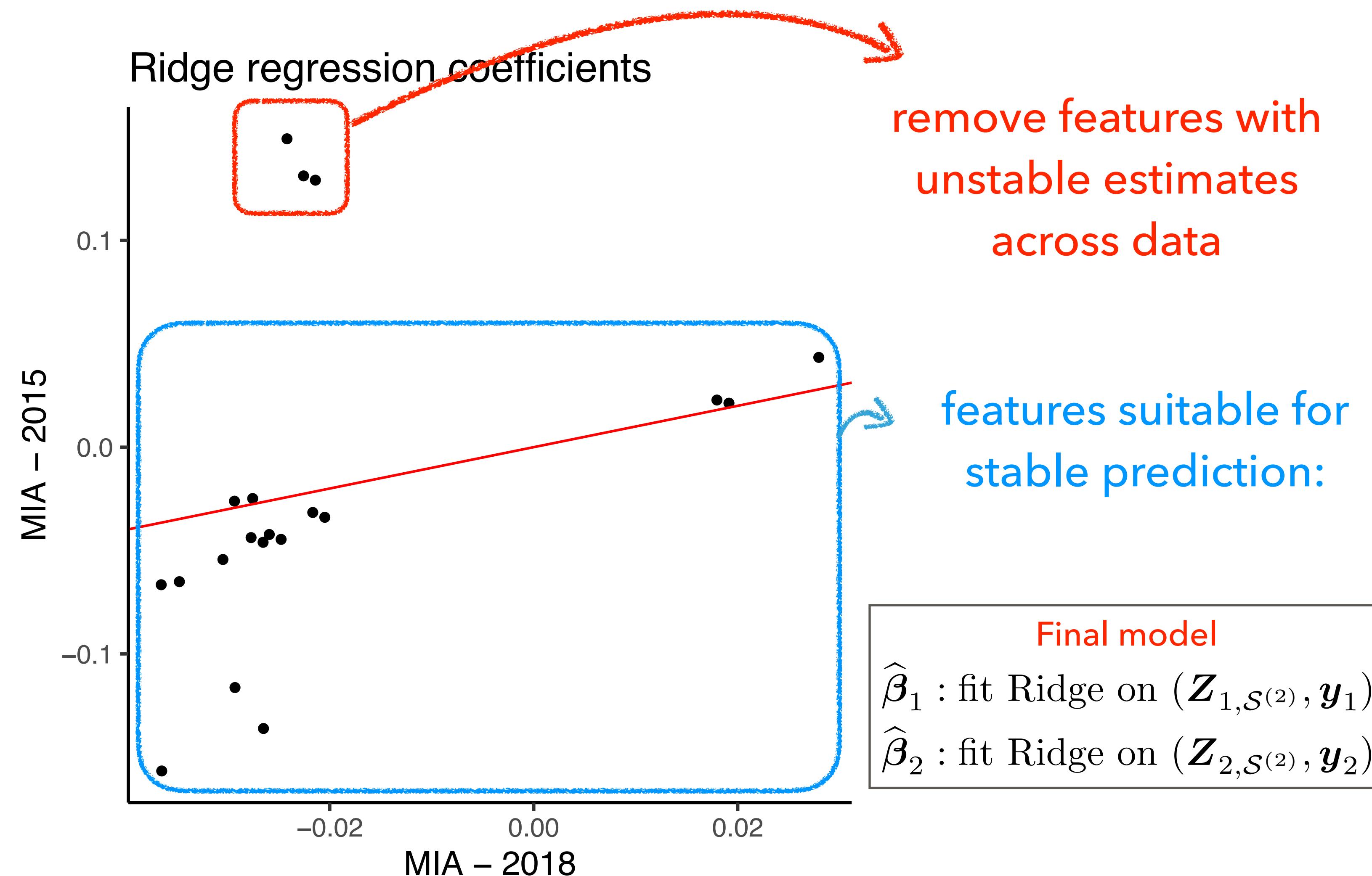
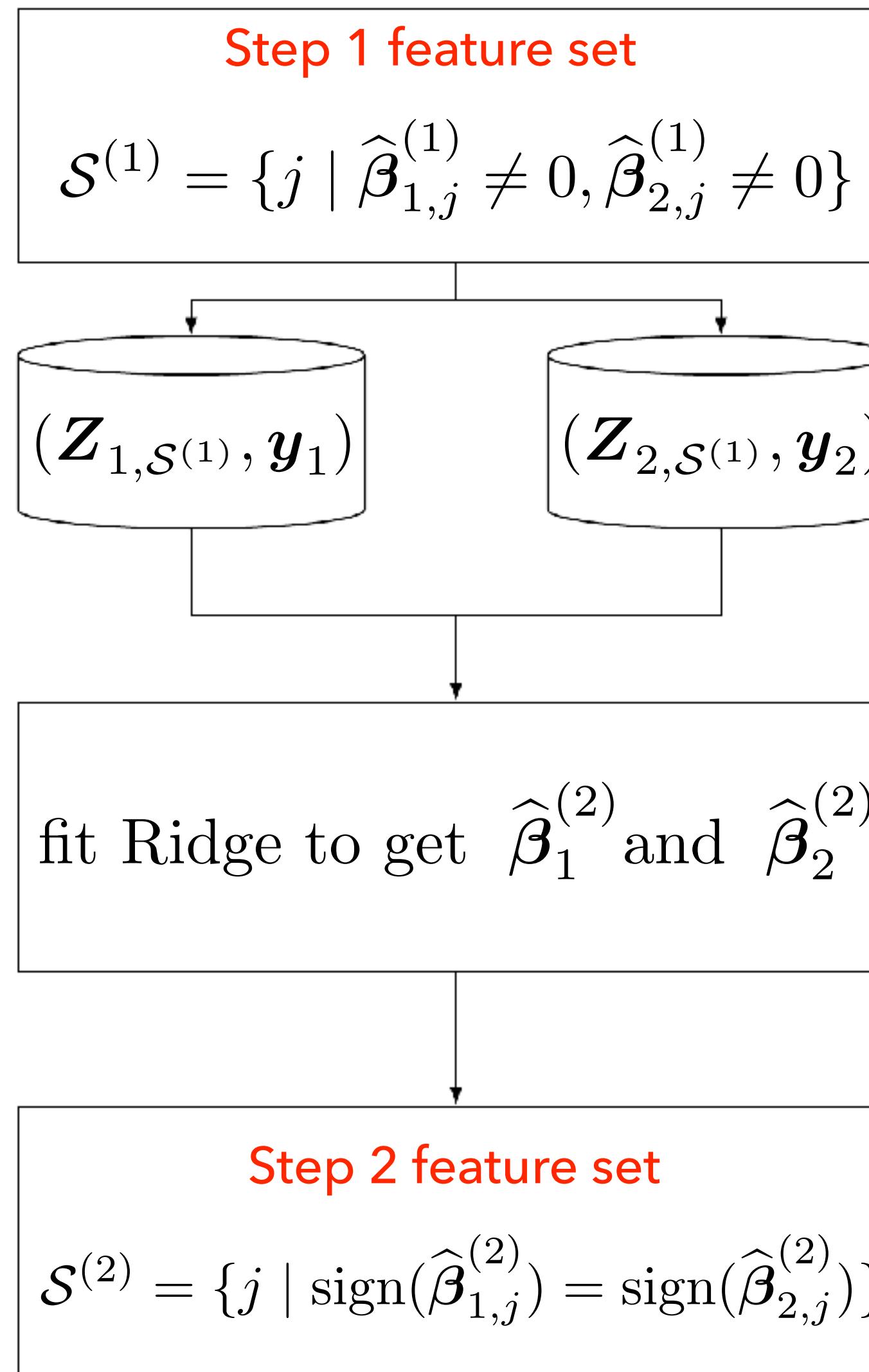


*Not all features are shown

Step 2: feature estimation stability



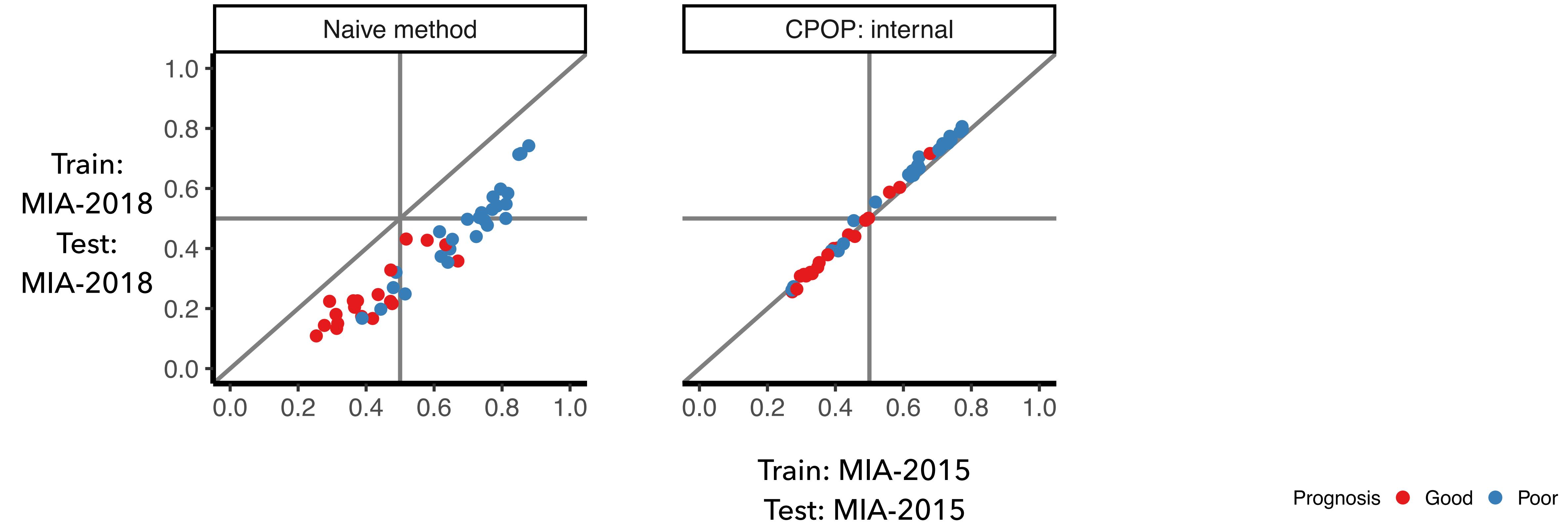
Step 2: feature estimation stability



*Not all features are shown

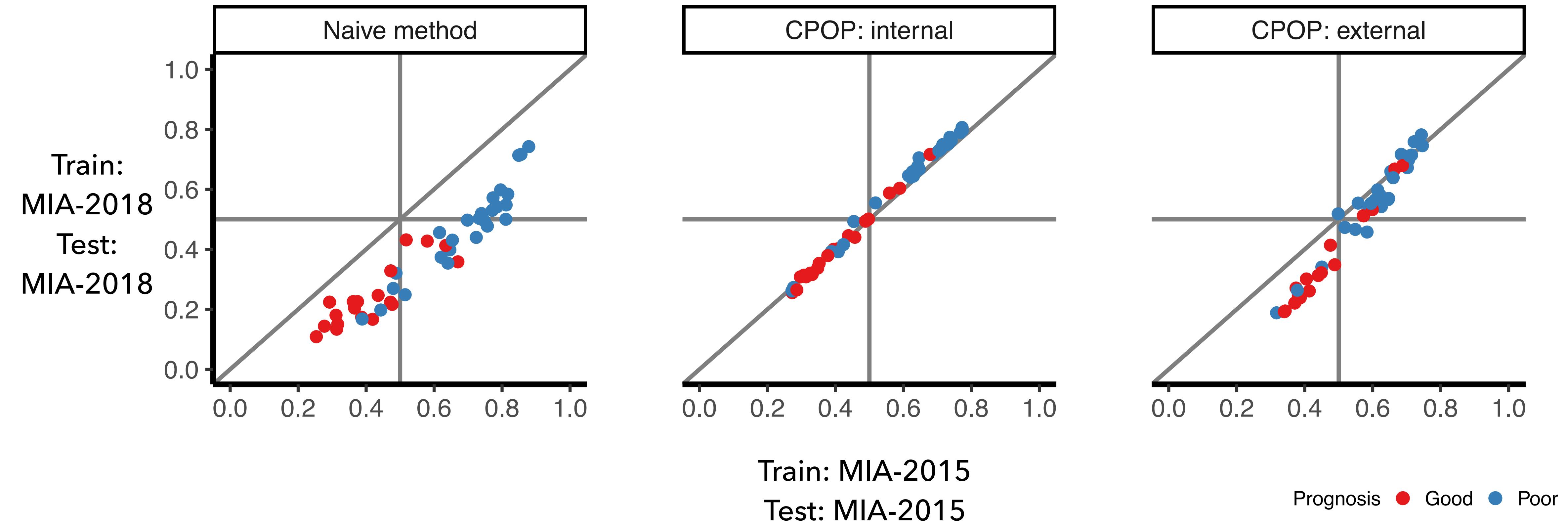
Results

Result 1: MIA 2015 vs 2018 data



Small deviation in **predicted values** across datasets

Result 1: MIA 2015 vs 2018 data

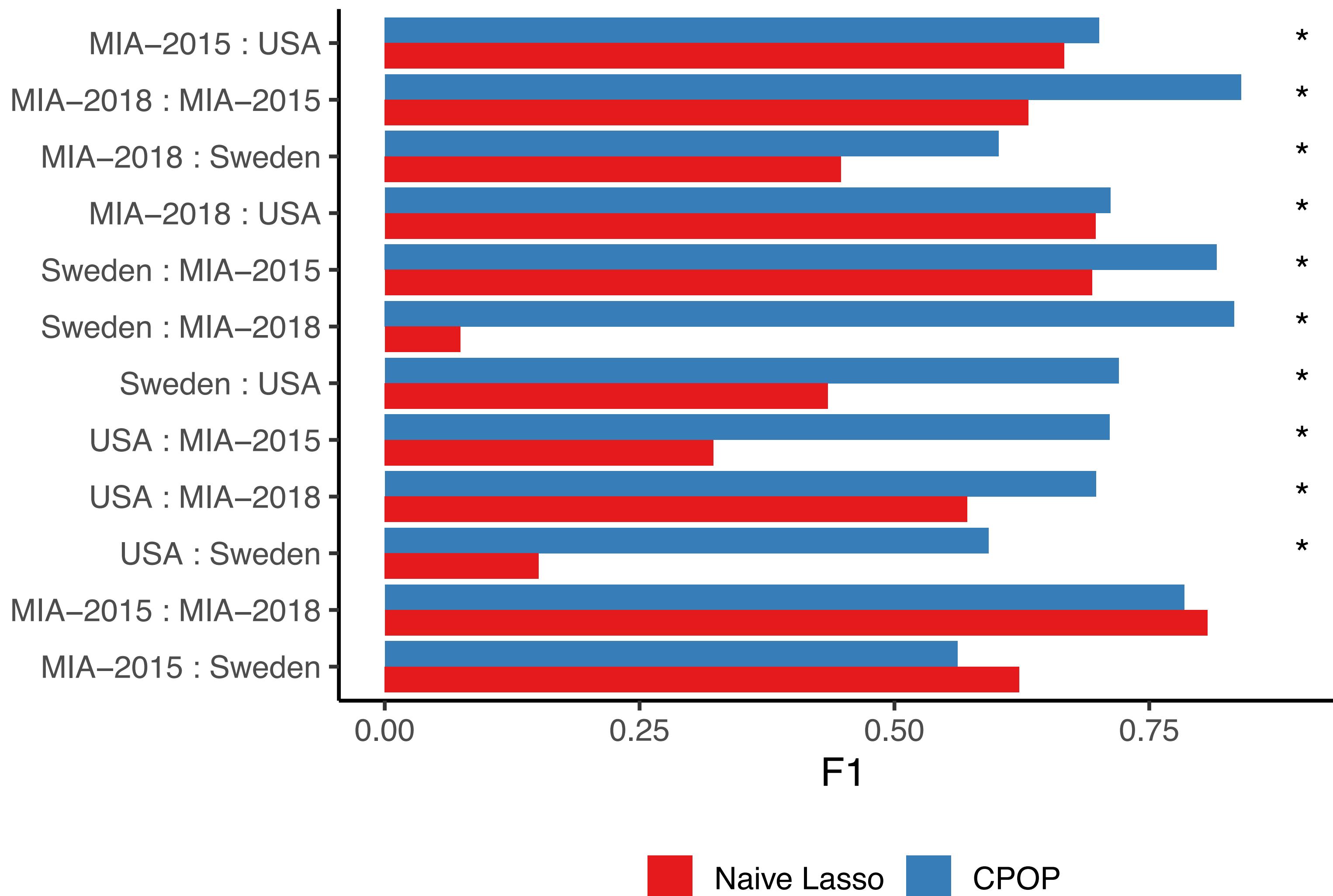


Small deviation in **predicted values** across datasets

Result 2: four melanoma data

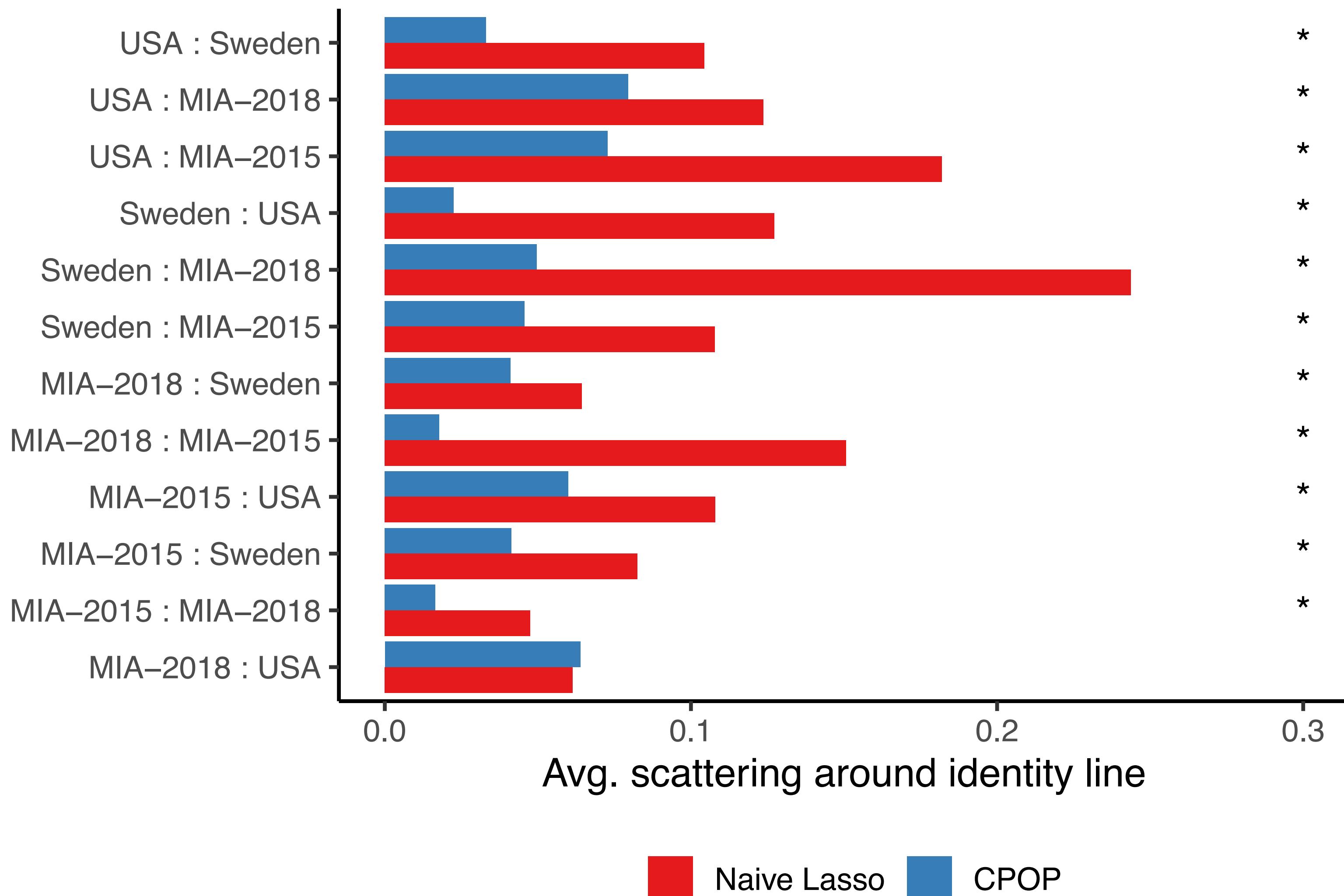
Compare F1 statistic (larger is better)

CPOP is highly predictive



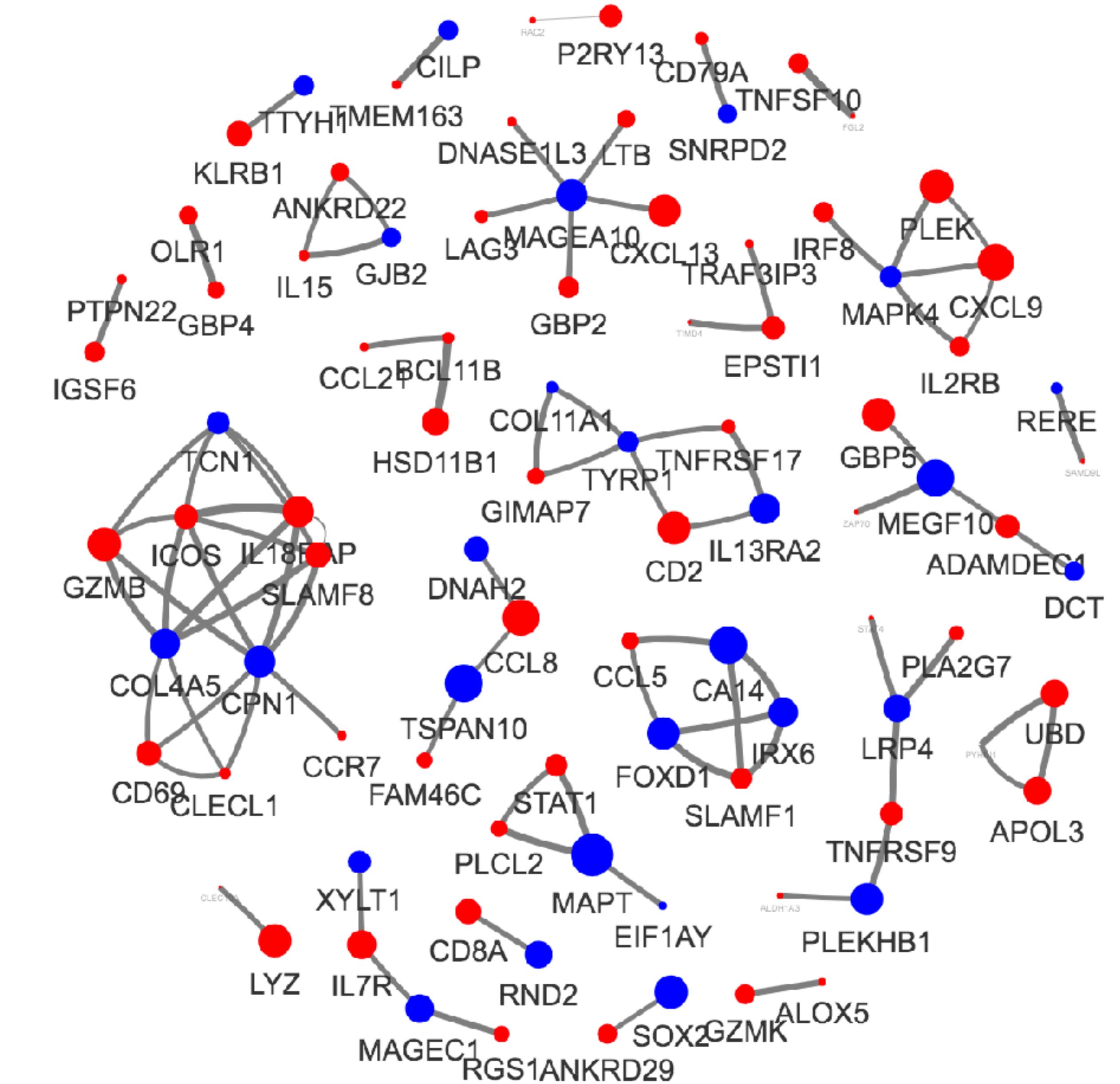
Result 2: four melanoma data

Compare avg. deviation (smaller is better)



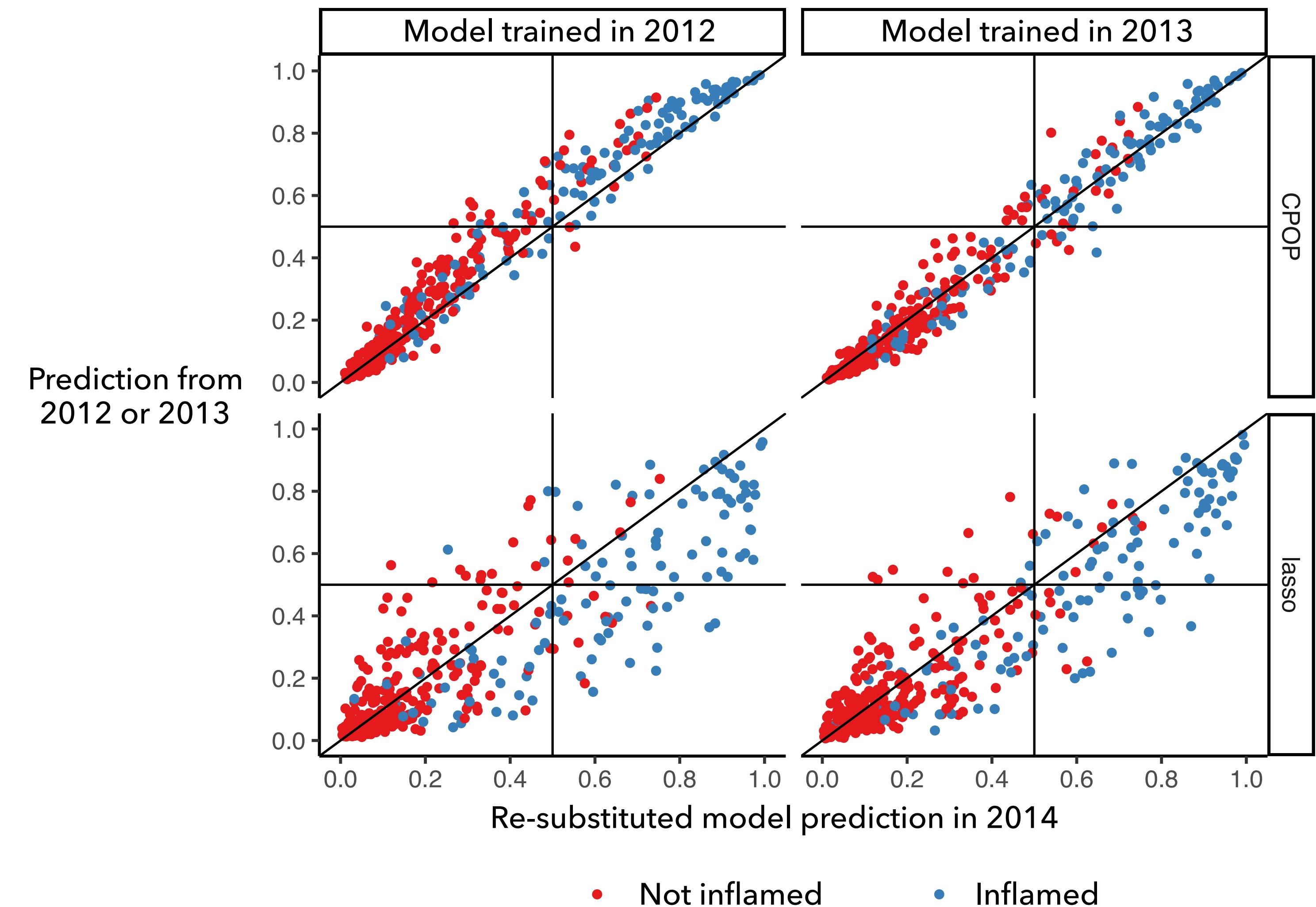
Result 3: melanoma feature set

CPOP features offer new biological interpretations



Result 4: prospective prediction on inflammatory bowel disease

CPOP works on
prospective experiments



Concluding remarks

1. Integrates clinical implementation constraints into the model
 2. Log-ratios enable prospective and multi-centres prediction
 3. Stable variable selection and estimation components
- ▶ A flexible and adaptable framework focuses on transferability
 - ▶ Potential to handle data with higher relevance to precision medicine (e.g. drug sensitivity)

Acknowledgement

- ▶ Supervision:
 - ▶ Jean Yang
 - ▶ Samuel Mueller
 - ▶ Garth Tarr
- ▶ Melanoma Institute Australia
- ▶ Sydney Precision Bioinformatics Group

Clinical constraints

Data

$$(X_1, y_1)$$

$$(X_2, y_2)$$

No re-normalisation

Model

$$\hat{\beta}_1$$

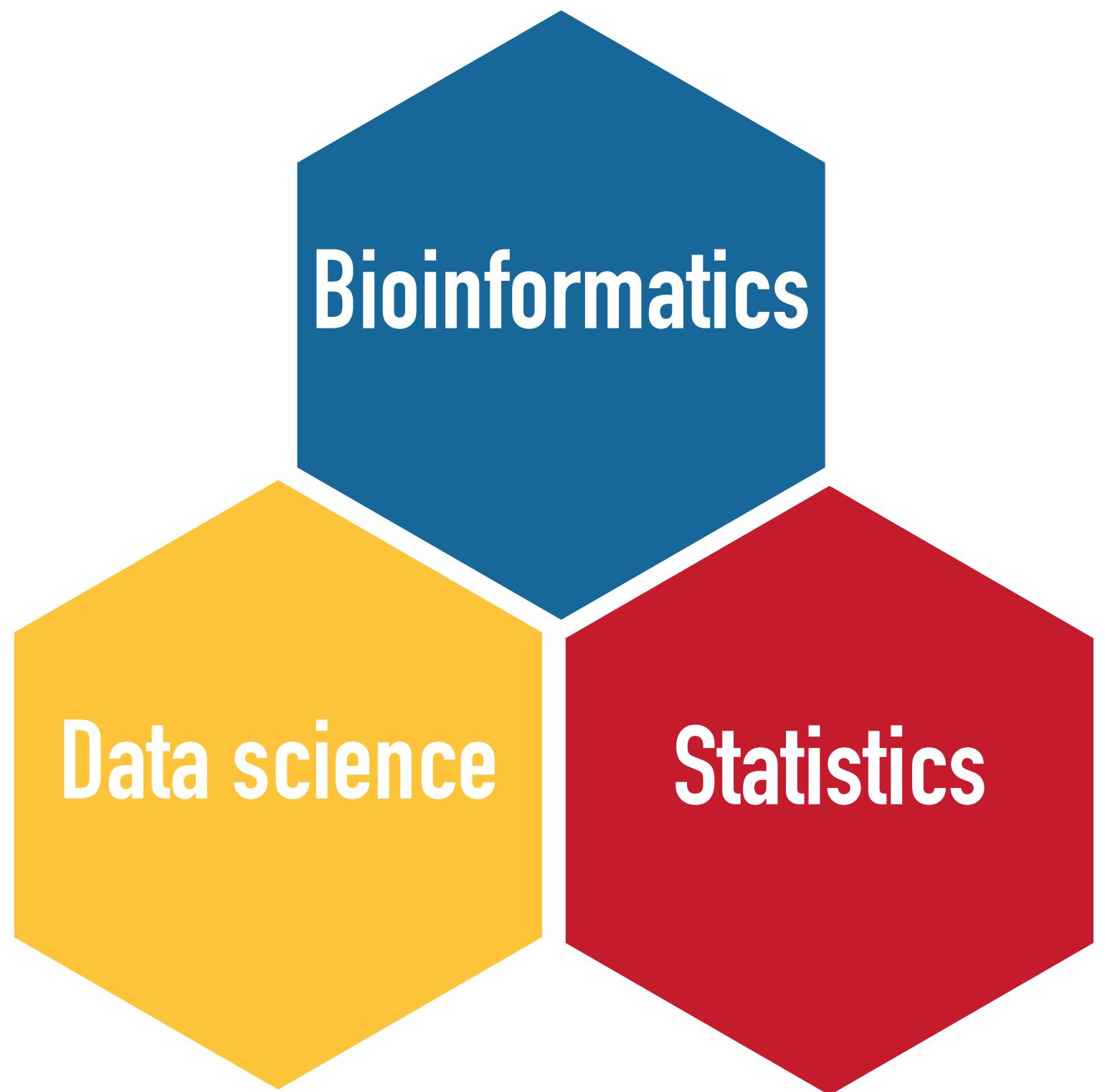
No model re-training

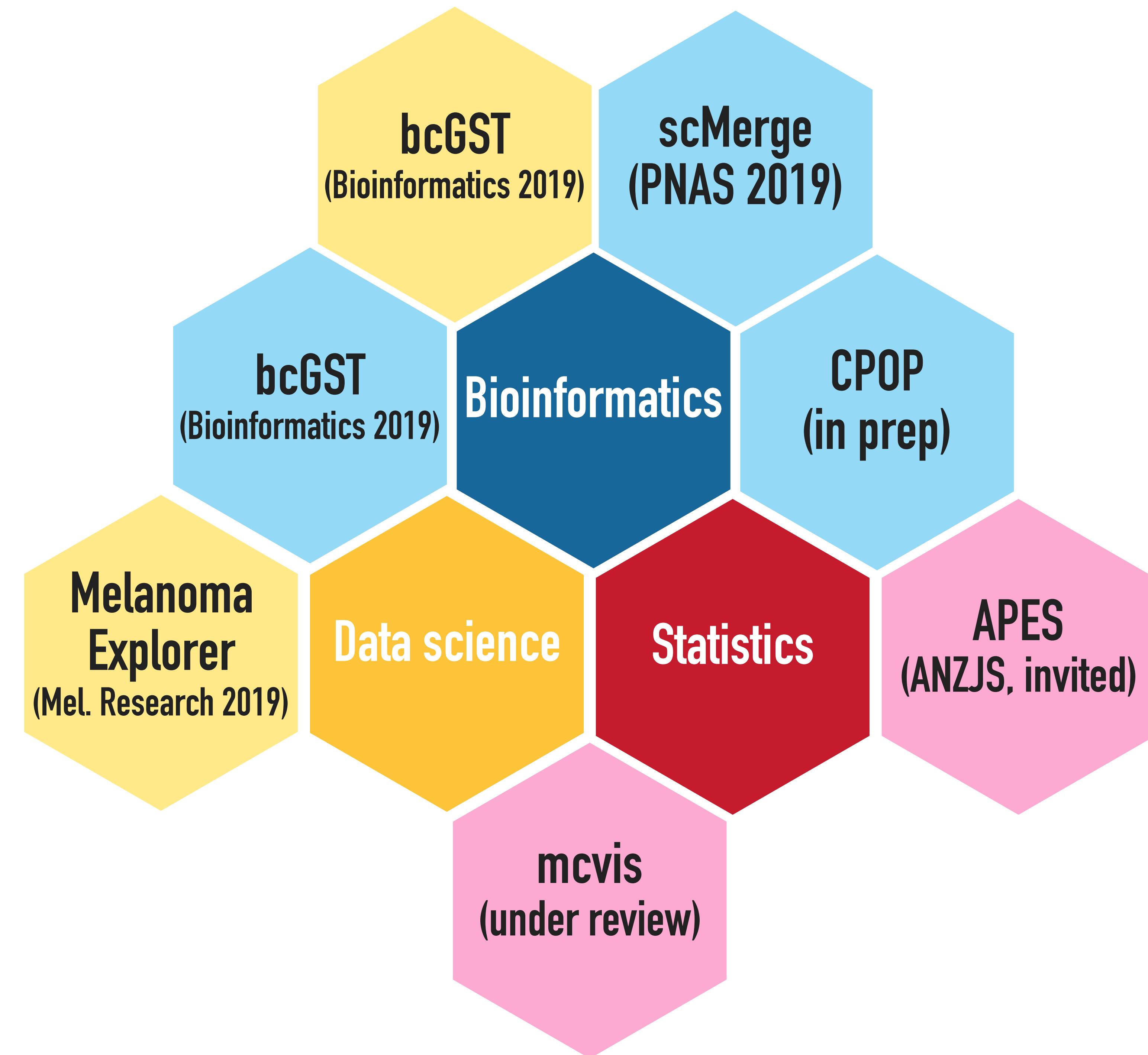
Prediction

$$\hat{y}_1 = X_1 \hat{\beta}_1$$

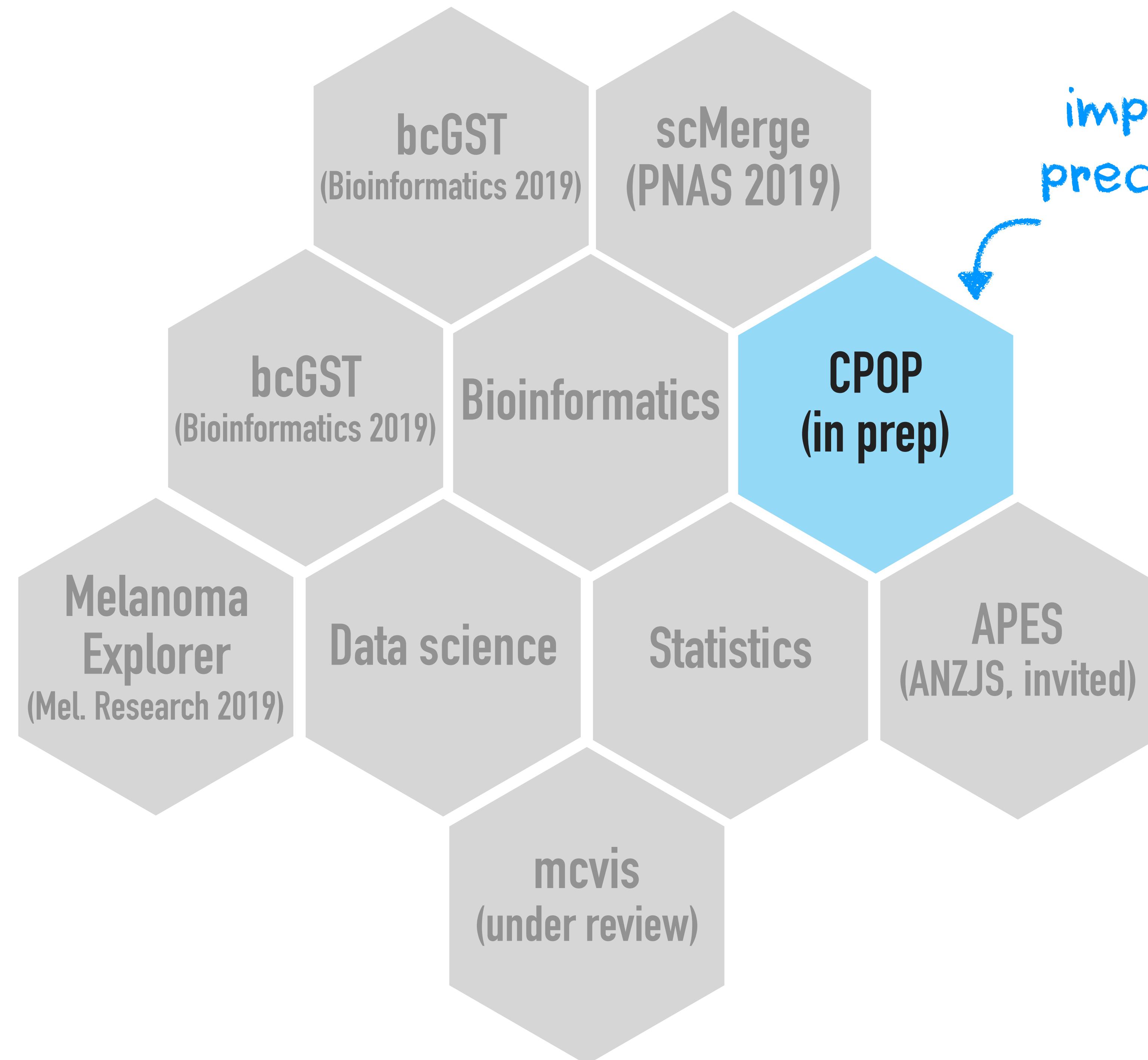
$$\hat{y}_2 = X_2 \hat{\beta}_1$$

Scale-equivalent prediction

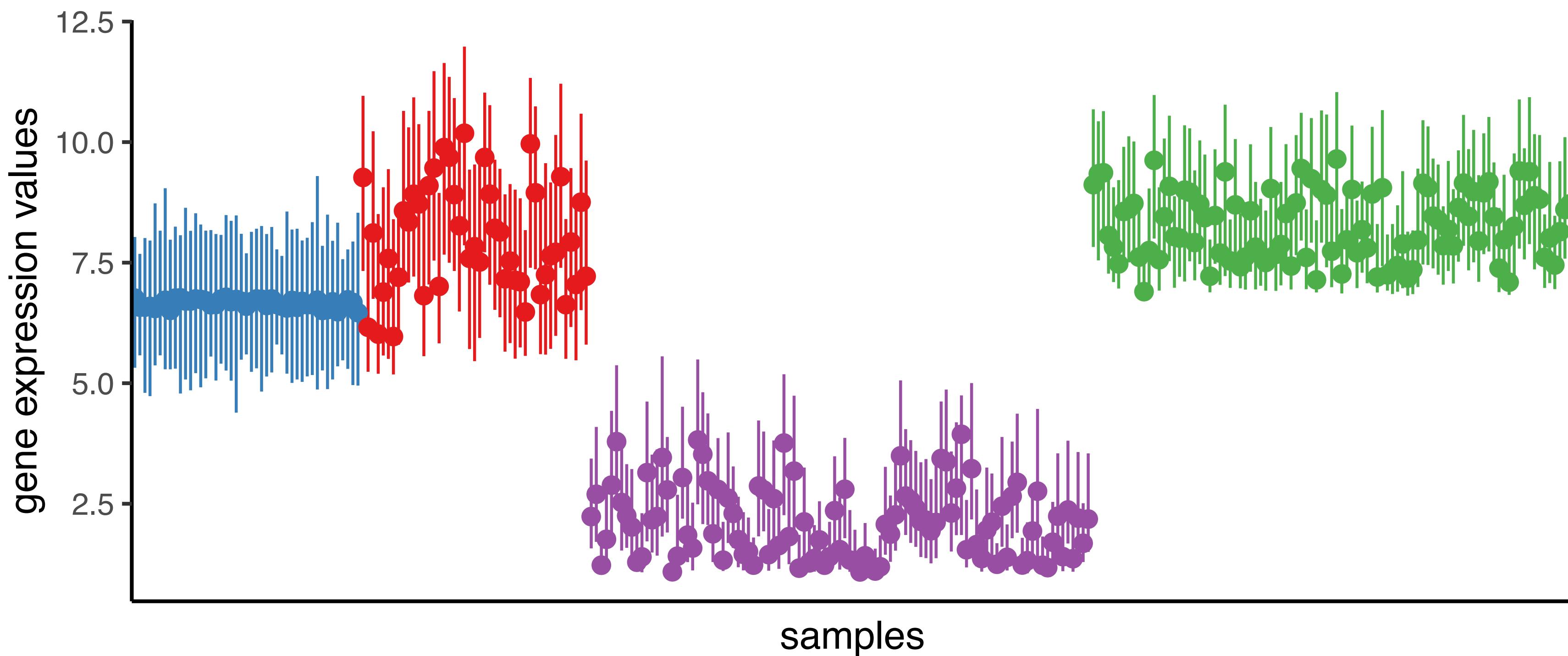




Towards
implementation
precision medicine



Vision of a risk score



(Jayawardana et al. 2015)

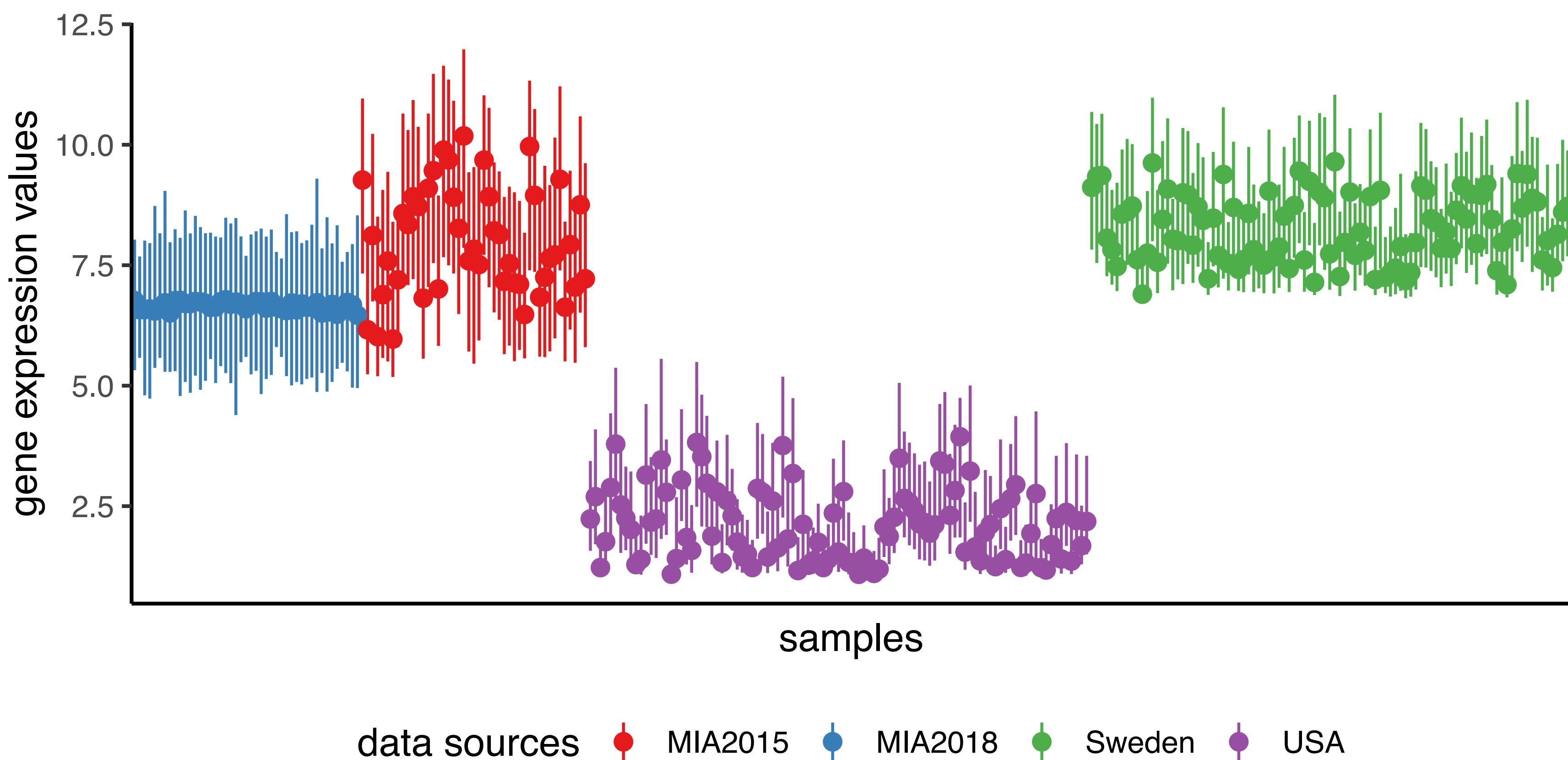
(Cirena jw is et al. 2015)



$$\hat{y} = X \hat{\beta}$$

regression-based
risk score

Melanoma Institute Australia

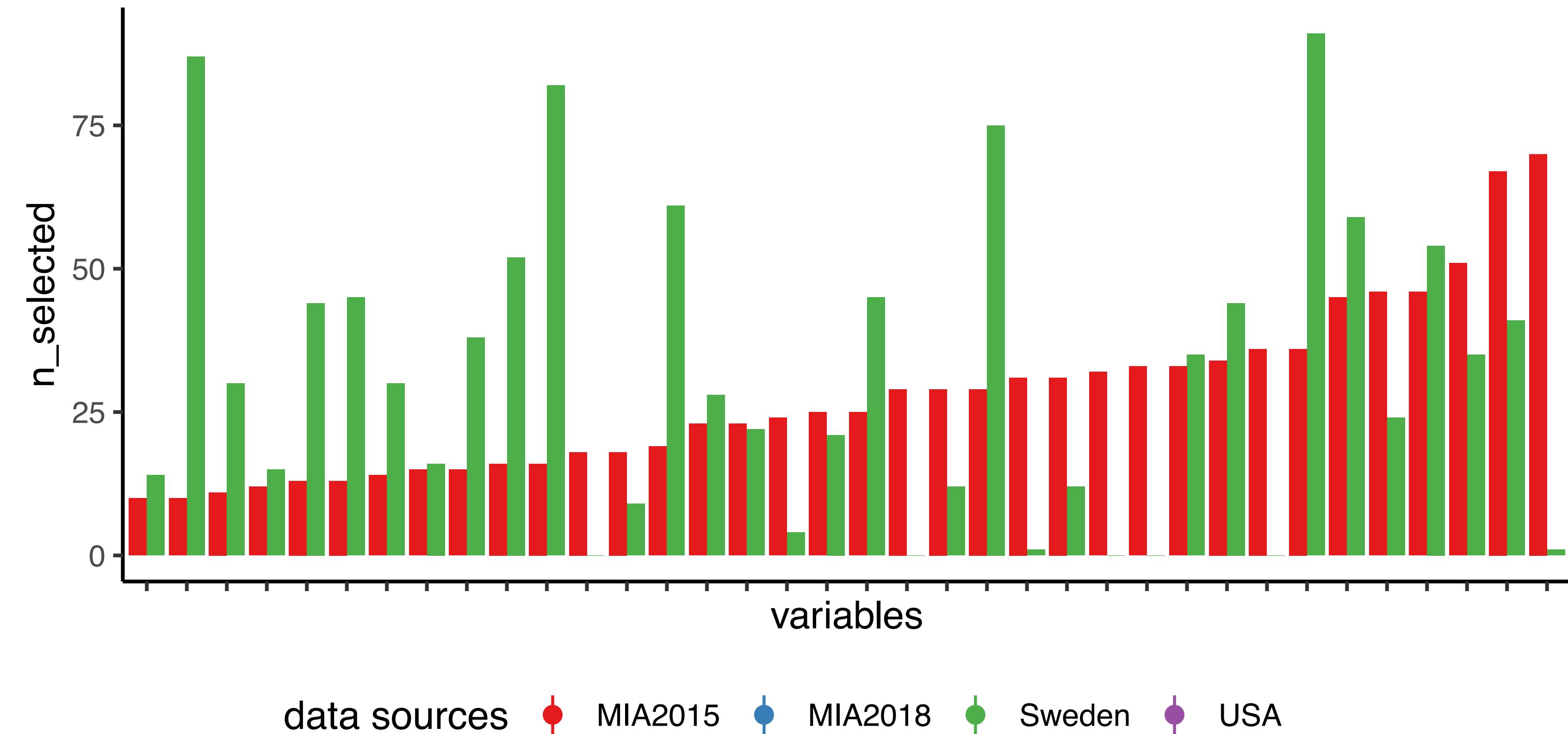


- ▶ Predict patient outcomes using gene expression:
 - ▶ prospectively
 - ▶ multi-centres
- ▶ CRE grant for implementation



Lasso variable selection is not stable!

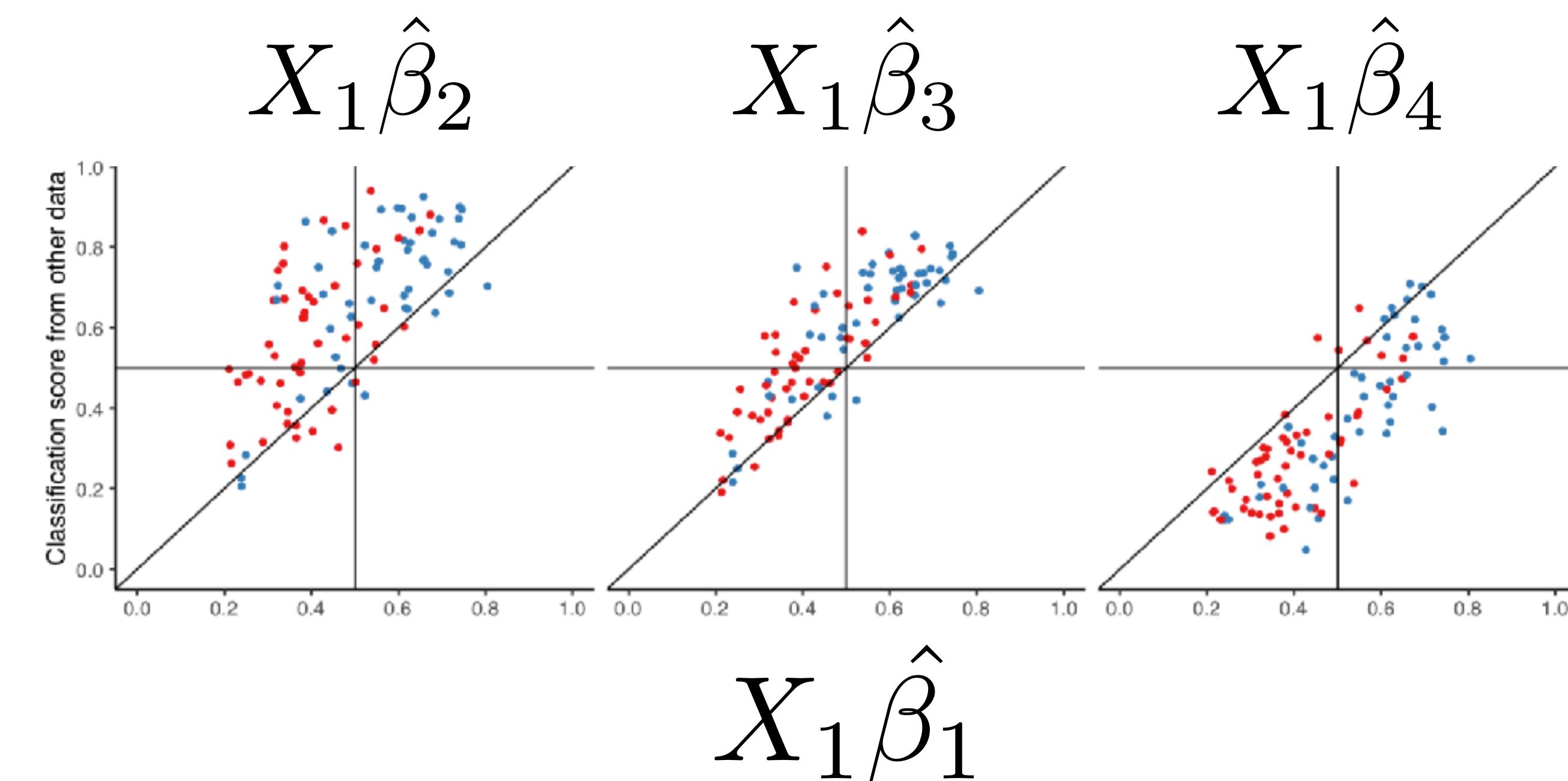
- ▶ Bach 2008
 - ▶ Meinshausen & Bühlmann 2010
 - ▶ Lim & Yu 2016
- all pointed out that Lasso is unstable under cross-validation



Statistical challenges

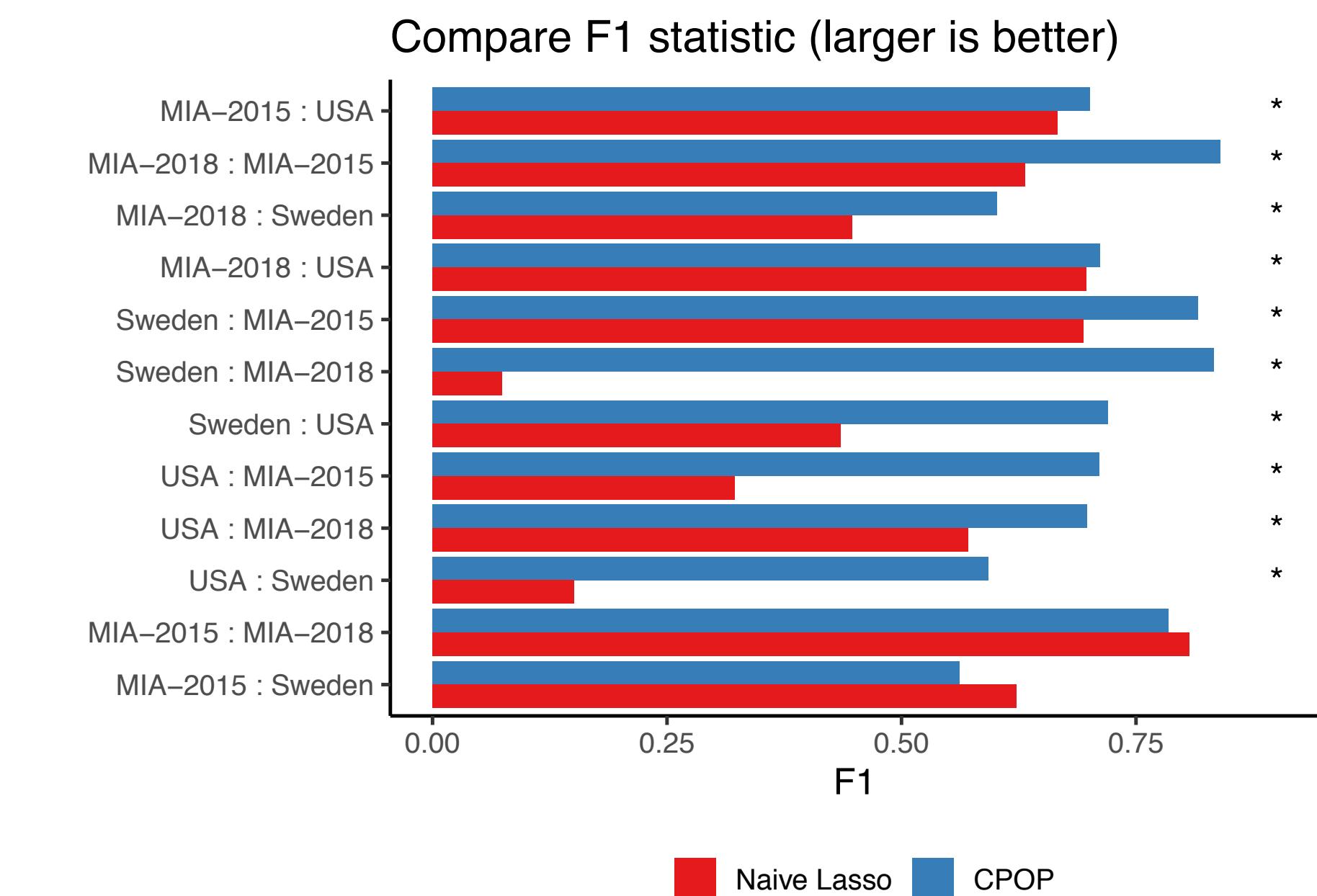
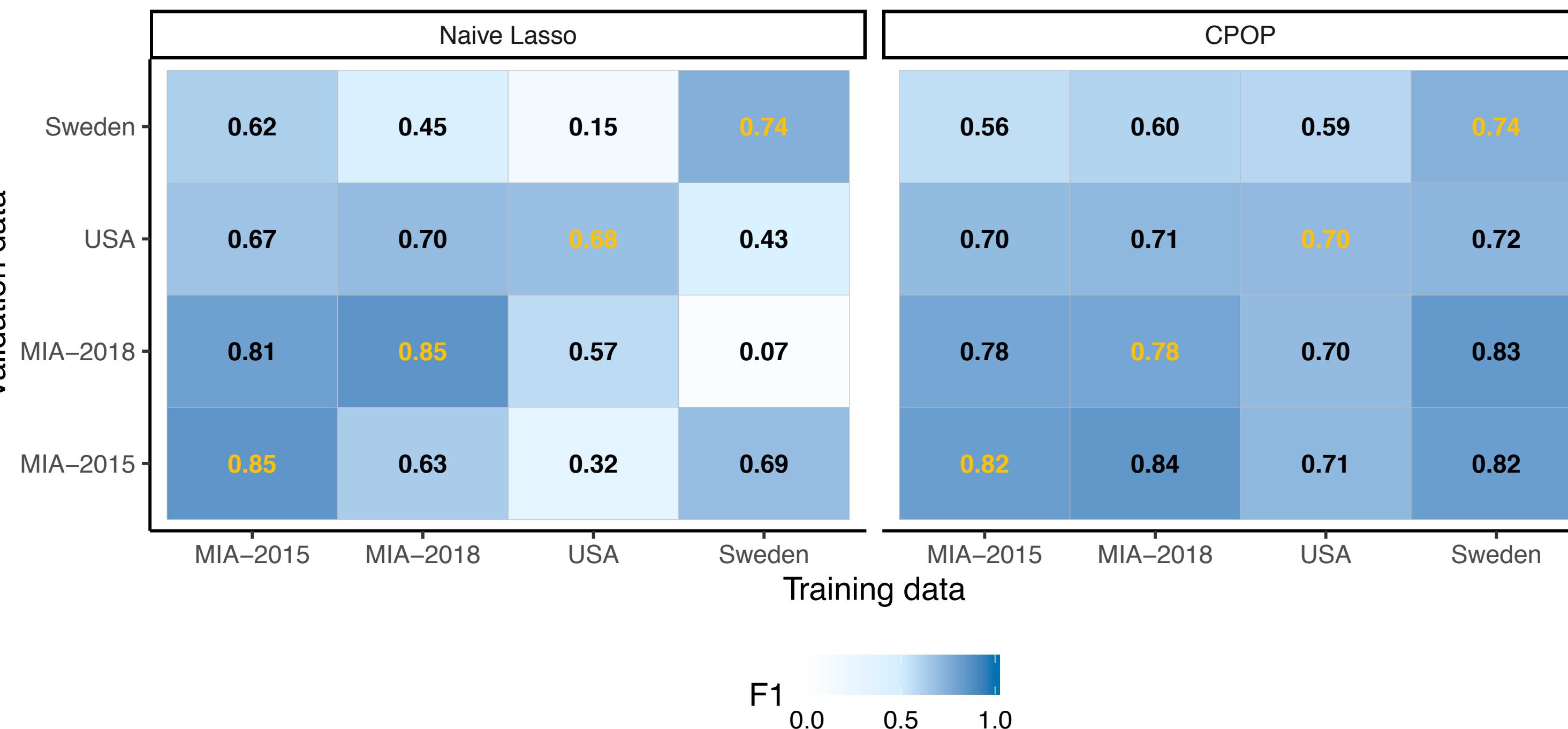
1. Concordance in features scaling across datasets
2. Concordance in feature selection and estimates
3. Single-patient prediction

Transferability implies that patients should be evenly scattered around the identity line



CPOP results 2: four melanoma data

F1 classification statistic



CPOP is highly predictive

Motivation for CPOP: one patient cohort, two gene expression data

$$X_1 \hat{\beta}_1 \approx X_2 \hat{\beta}_2$$

loosely translate to

$$X_1 \approx X_2 \quad \hat{\beta}_1 \approx \hat{\beta}_2$$

column-wise

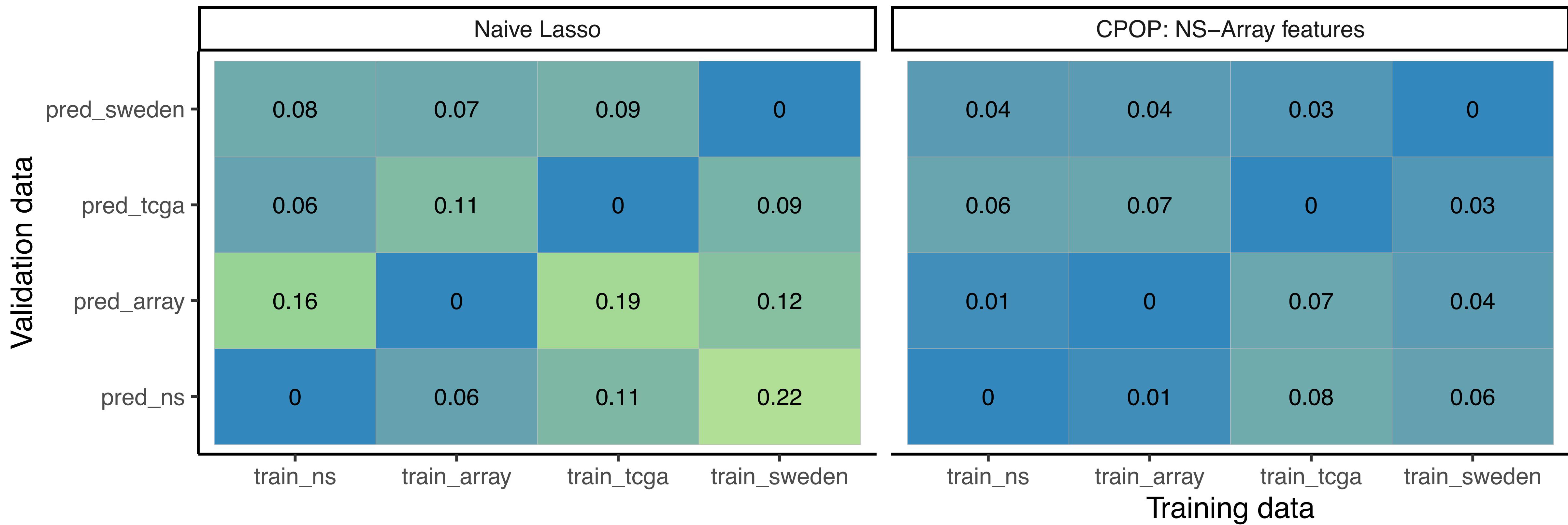
(feature distribution stability)

element-wise

(mode estimation stability)

CPOP results 1: four melanoma data

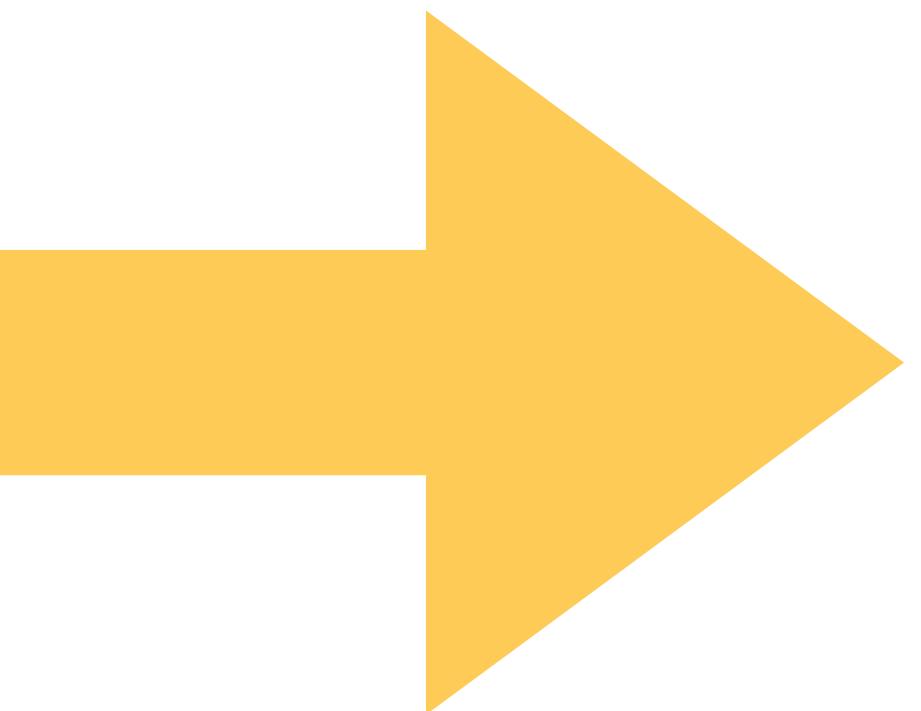
Identity distance between predicted values under various models



Small deviation in **predicted values** across datasets

Framingham heart disease risk score:

- ▶ Age (Years)
- ▶ Cholesterol (mg/dL)
- ▶ If smoker (Yes/No)
- ▶ HDL cholesterol (mg/dL)
- ▶ Systolic blood pressure (mm Hg)



$$\hat{y} = X \hat{\beta}$$

20 points model

Concluding remarks

- ▶ CPOP is a flexible procedure that allows for:
 - ▶ cross-platform omics prediction
 - ▶ stable single-patient prediction
- ▶ Not everyone can smooth-sailing through the PhD process, find your own way to deal with it (e.g. insert random pictures into your slides)

But what about breast cancer?

- ▶ Alvarado et. al. (2015) reported poor concordance in the prediction scores
- ▶ Hyeon et. al. (2017) considered NanoString as a viable alternative to RT-PCR

Name	Predictors	Targets	Prediction	Technology	Legit?
Oncotype DX	21 genes	ER +	Score	qRT-PCR	ASCO, NCCN
Prosigna	50 genes	Hormone receptor +	Score	NanoString	FDA 510k
MammaPrint	70 genes	Any ER status	Binary	DNA microarray	FDA

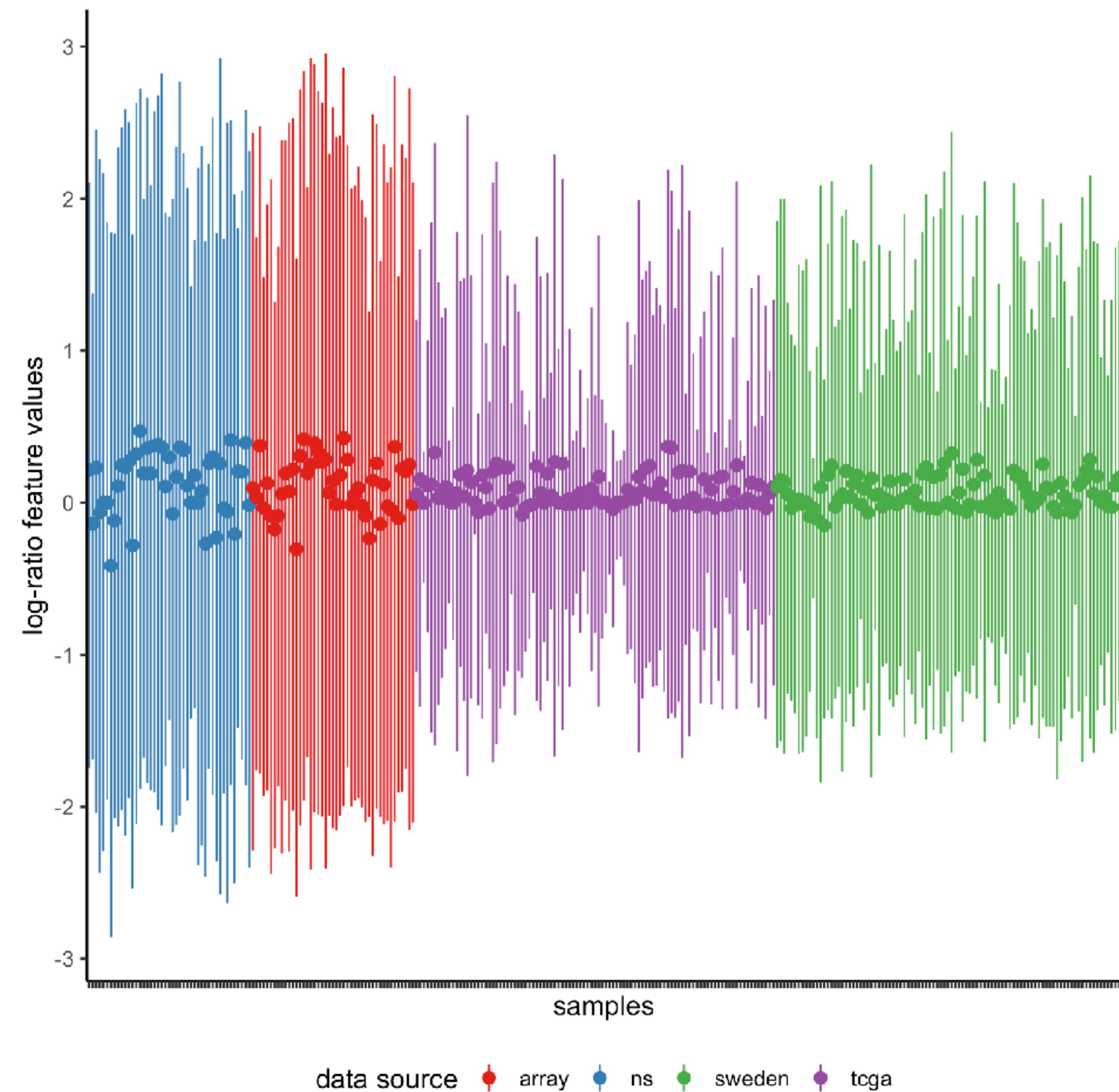
Is log-ratio really a new innovation?

Here is a list of papers that uses genes as predictors

Here is a list of papers that uses a single ratio for prediction

Our contribution is the advocacy using a whole collection of ratios for prediction

This has extra implications in terms of the statistics, but we are happy to tackle these.



Within-sample feature standardisation

