Kevin Wang

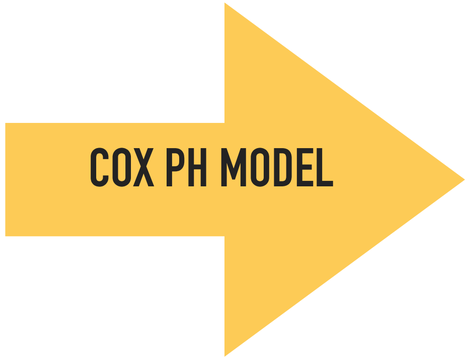
2019 Semester 2 Progress

CPOP

Cross-Platform Omics Prediction

The vision of a clinical risk score

- Framingham risk score: all predictors are on an absolute scale
  - Age (Years)
  - Cholesterol (mg/dL)
  - If smoker (Yes/No)
  - HDL cholestrol (mg/dL)
  - Systolic blood pressure (mm Hg)



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

Why we do not have an omics-based clinical risk score?

- Gene expression platforms are measured on a **relative scale**
  - Sequencing depth in RNA-Seq:

$$\hat{y} = X\hat{\beta} \qquad \hat{y} = (X - \mathbf{1})\hat{\beta}$$

  - Use of different reagents between experiments:

$$\hat{\beta} = (X^\top X)^{-1} X y$$

Why we do not have an omics-based clinical risk score?

- Most ML methods apply standardisation on both training and validation data

- But in clinical prediction, you might only have single samples to predict on

X	Sample 1	Sample 2	Sample 3
Gene 1	1.2	2.1	1.5
Gene 2	5.6	4.6	7.1
Gene 3	9.2	10.1	6.9
Gene 4	4.1	3.6	2.7

Why we do not have an omics-based clinical risk score?

- Most ML methods apply standardisation on both training and validation data

- But in clinical prediction, you might only have single samples to predict on

New X	Sample 4
Gene 1	4.2
Gene 2	3.8
Gene 3	8.4
Gene 4	3.1

But what about breast cancer?

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

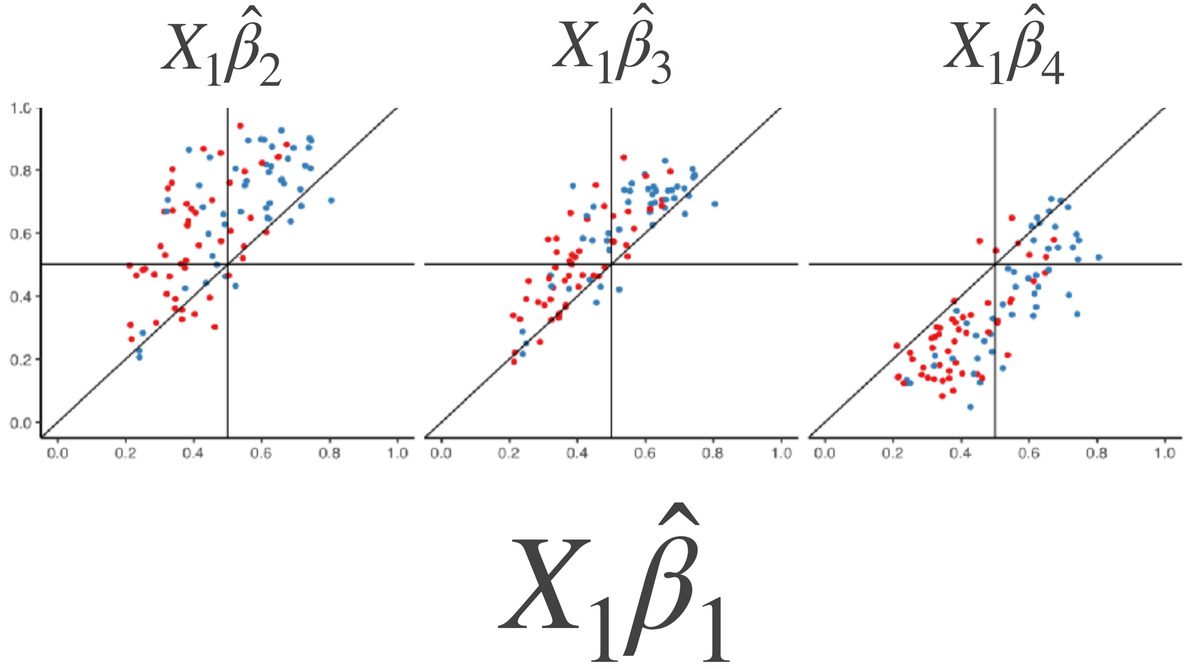
- Alvarado et. al. (2015) reported poor concordance in the prediction scores
- Hyeon et. al. (2017) considered NanoString as a viable alternative to RT-PCR

Statistical challenges

1. Single-patient prediction
2. Different scaling on genes between platforms
3. Concordance in feature selection/coefficient estimates

Transferability

The prediction on one gene expression platform should be equivalent to another platform



First component of CPOP: feature construction

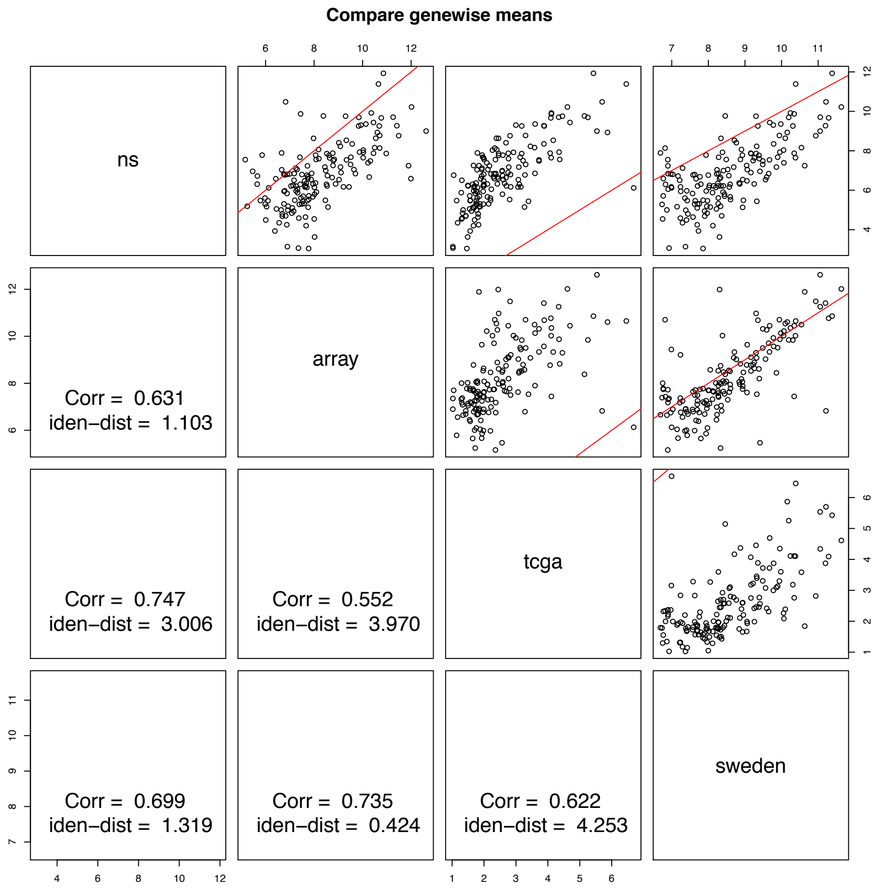
就让我 来次透彻心扉的痛
都拿走 让我再次两手空空
只有奋奋一息过
那个真正的我
他才能够诞生

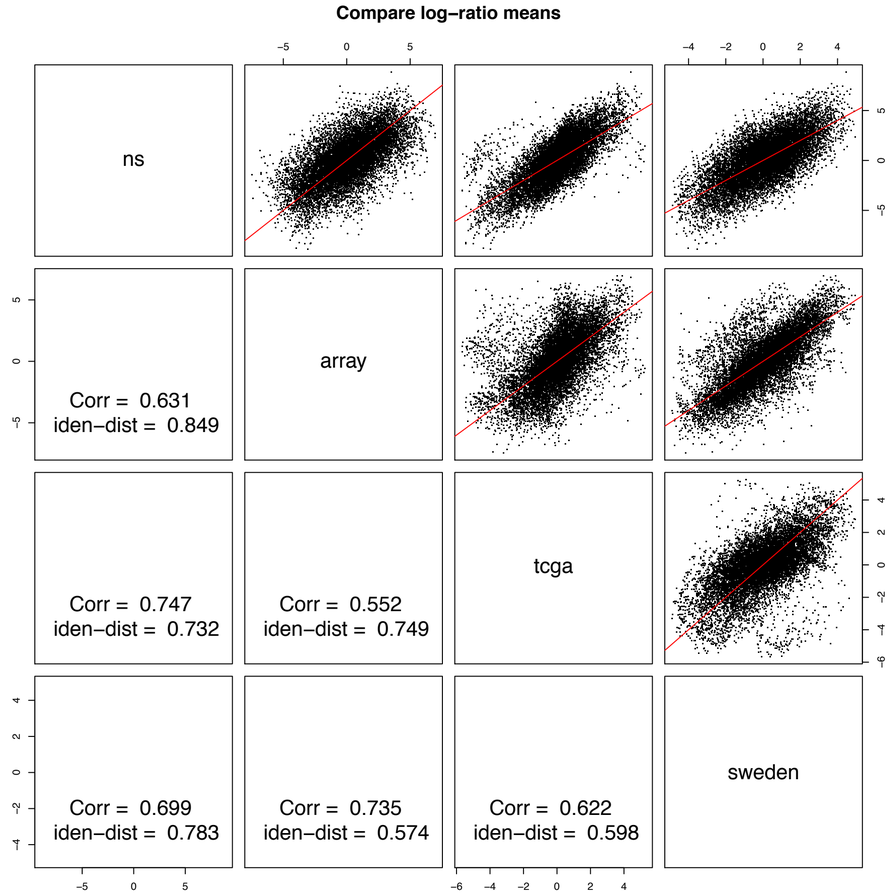
The solution is trivial.

- We have “standardised features” within every patient to build models

Log-ratio

log(gene A) – log(gene B)



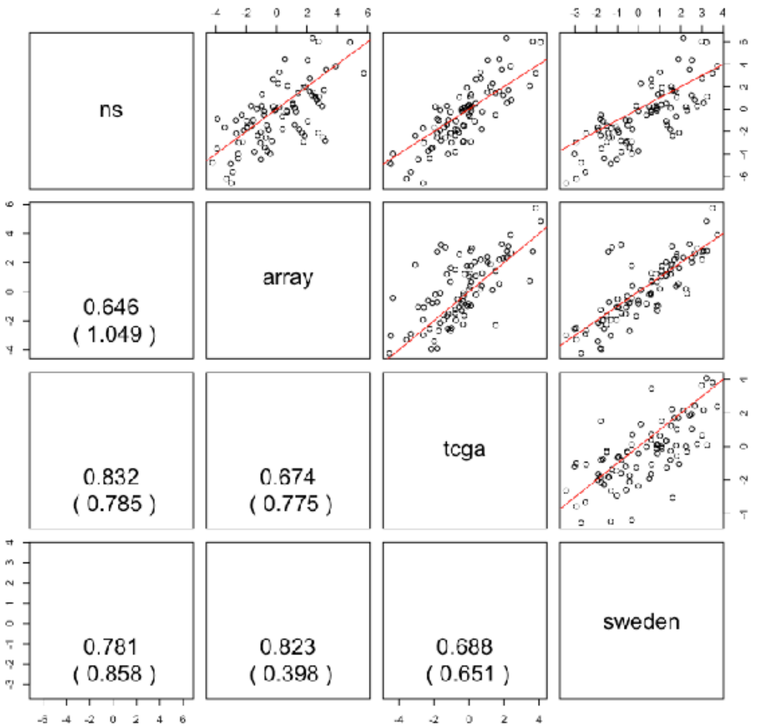


The solution is trivial?

1. Single-patient prediction
2. Different scaling on genes between datasets
3. Concordance in feature selection/coefficient estimates

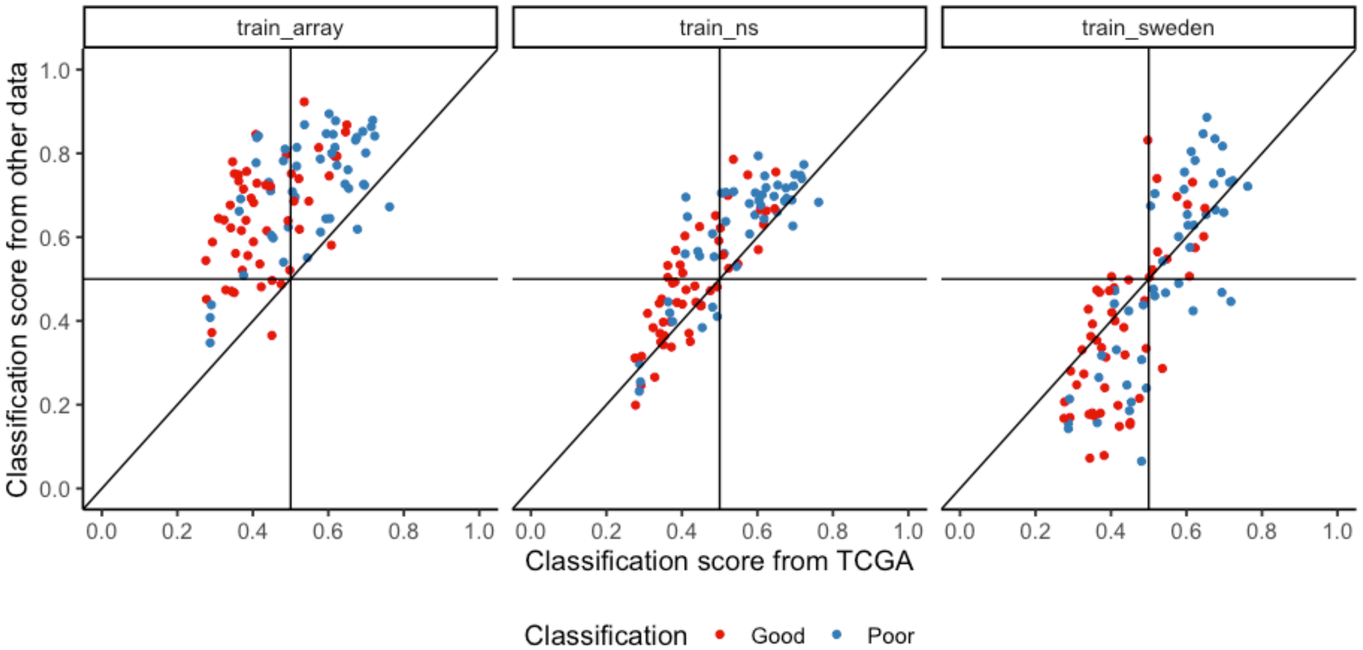
The solution is trivial?

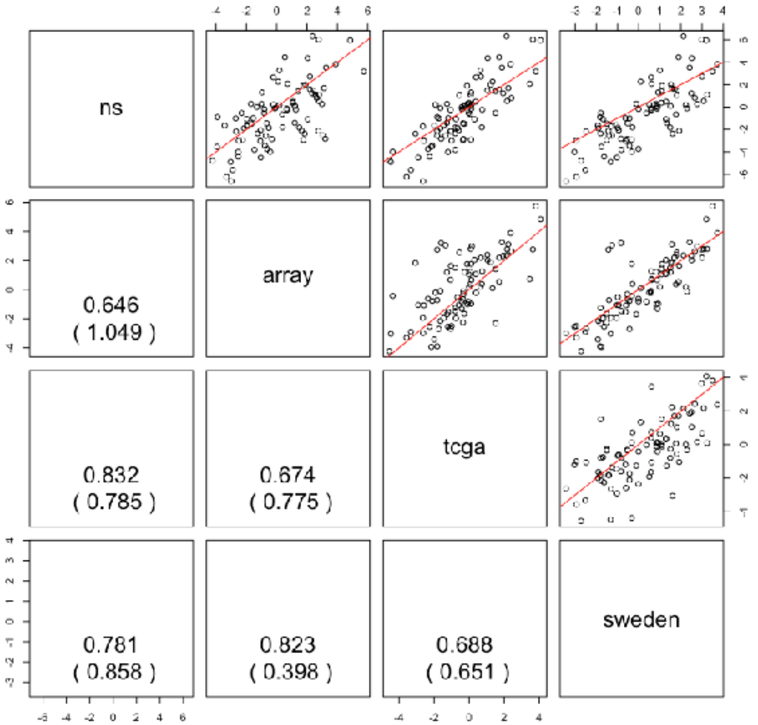
1. ~~Single patient prediction~~
2. Different scaling on log-ratio features between datasets
3. Concordance in feature selection/coefficient estimates



The solution is trivial?

1. ~~Single patient prediction~~
2. Different scaling on log-ratio features between datasets
3. Concordance in feature selection/coefficient estimates

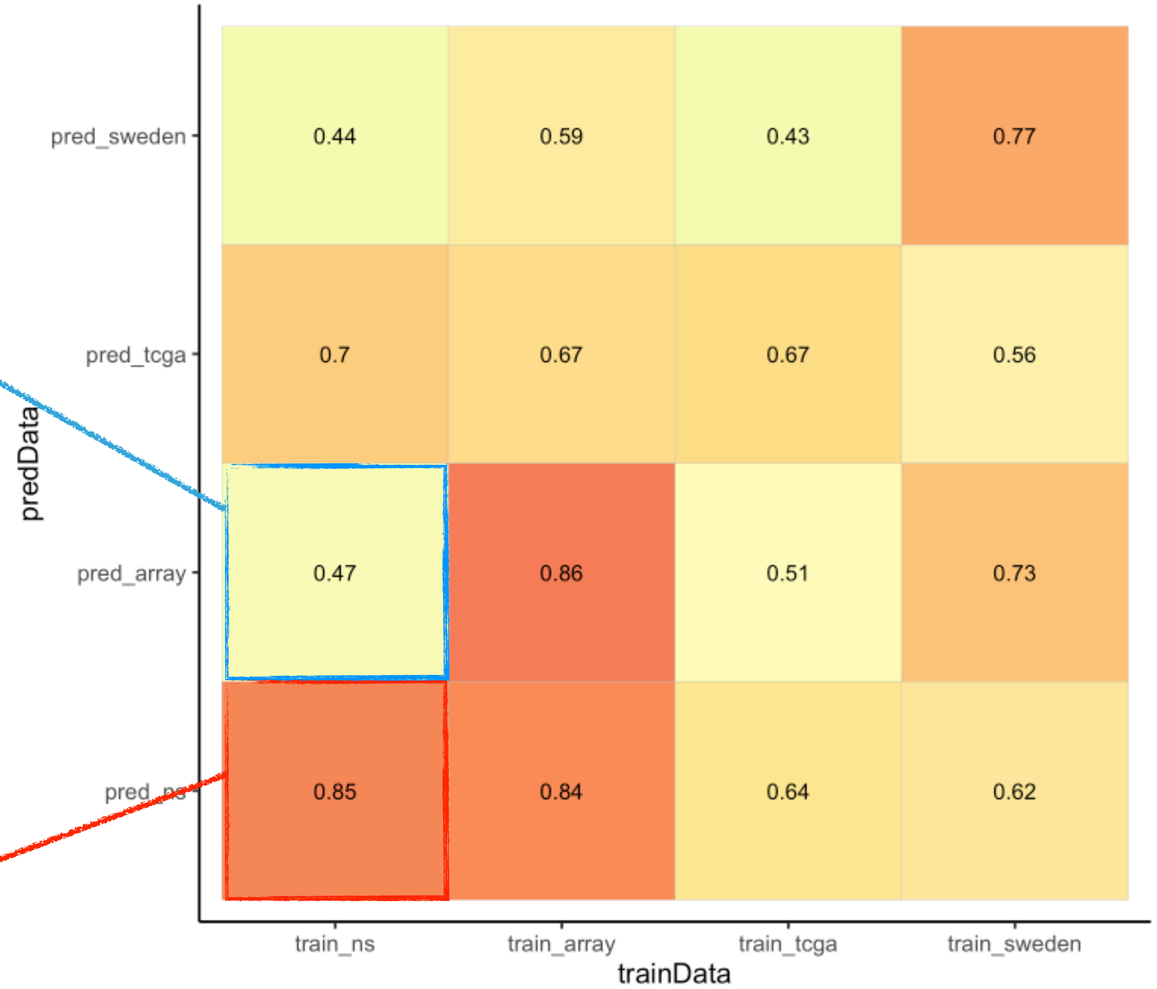




The solution is not so trivial

Poor validation accuracy

Good re-substitution accuracy



The solution is not so trivial

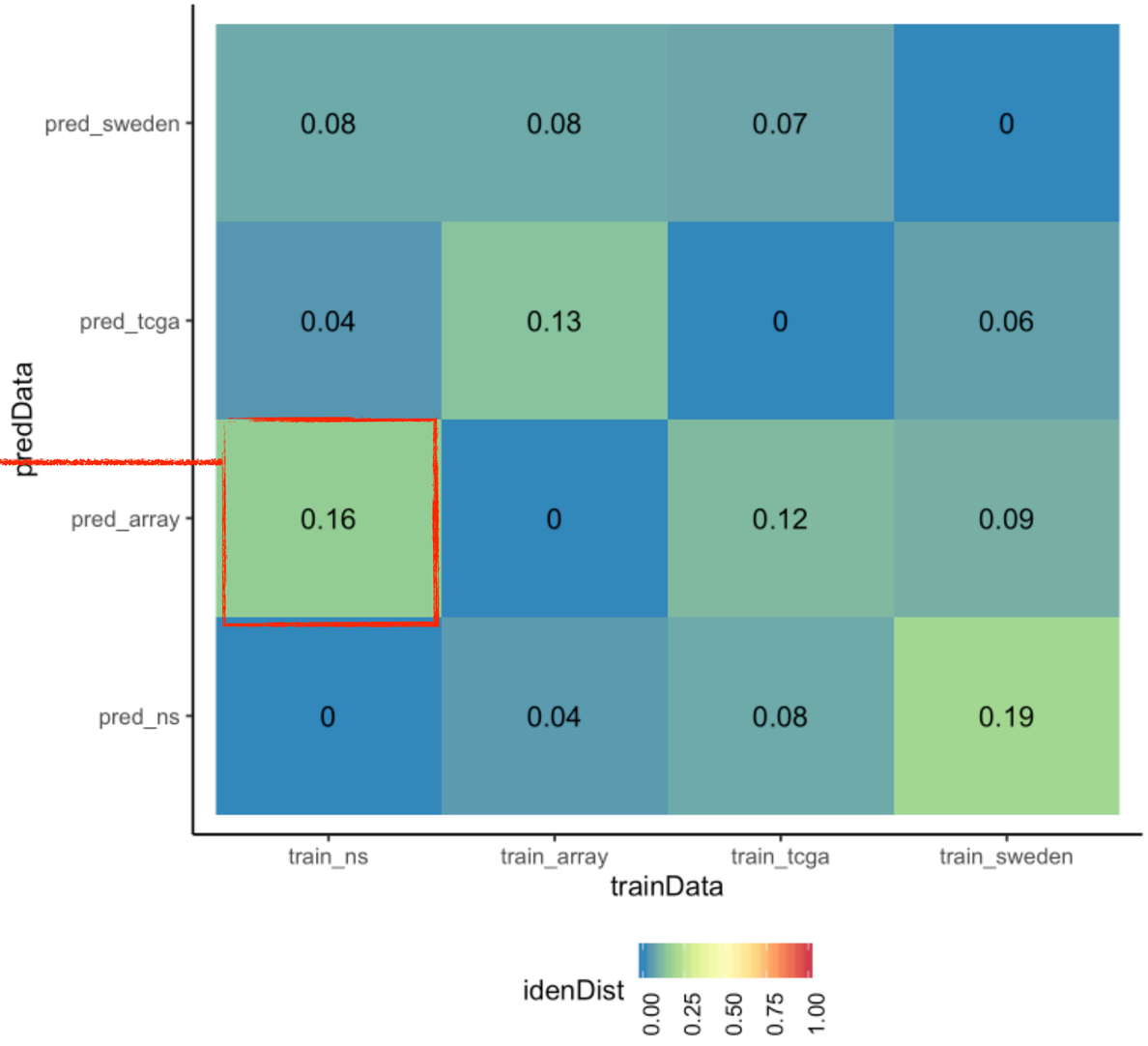
Estimated prognosis probabilities from

training data

vs

validation data

differ by 0.16 on average



And as you know, I always put jokes in my slides.

The past 9 months when I had to mentally deal with this problem was the most stressful time I had.

In the past 9 months, for every presentation that I did, I have put in subtle references to my favourite music.

This is my way to make the work that I do sufferable.

So, for the younger students: the PhD process can be a bit painful at times, not everyone can have a fun time sailing through it.

So do try to find your way to deal with it and talk to someone about it.

Second component of CPOP: feature selection

我曾经毁了我的一切
只想永远地离开
我曾经堕入无边黑暗
想挣扎无法自拔
我曾经像你像他像那野草野花
绝望着也渴望着
也哭也笑平凡着

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

$$Z_1\hat{\beta}_1 \approx Z_2\hat{\beta}_2$$

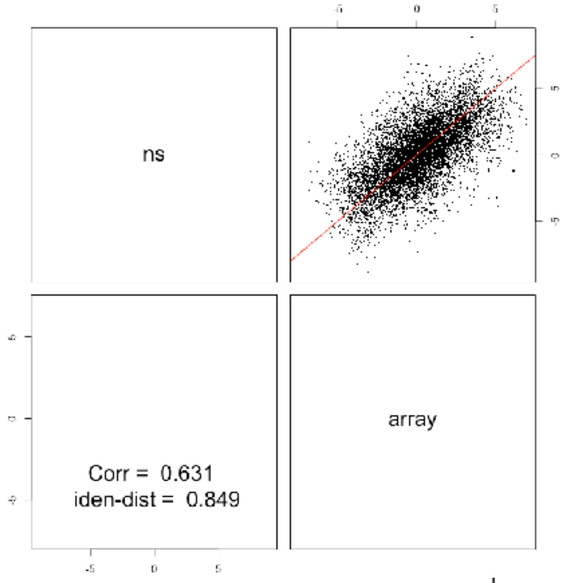
means

$Z_1 \approx Z_2$ column-wise

$\hat{\beta}_1 \approx \hat{\beta}_2$ element-wise

CPOP weighted variable selection

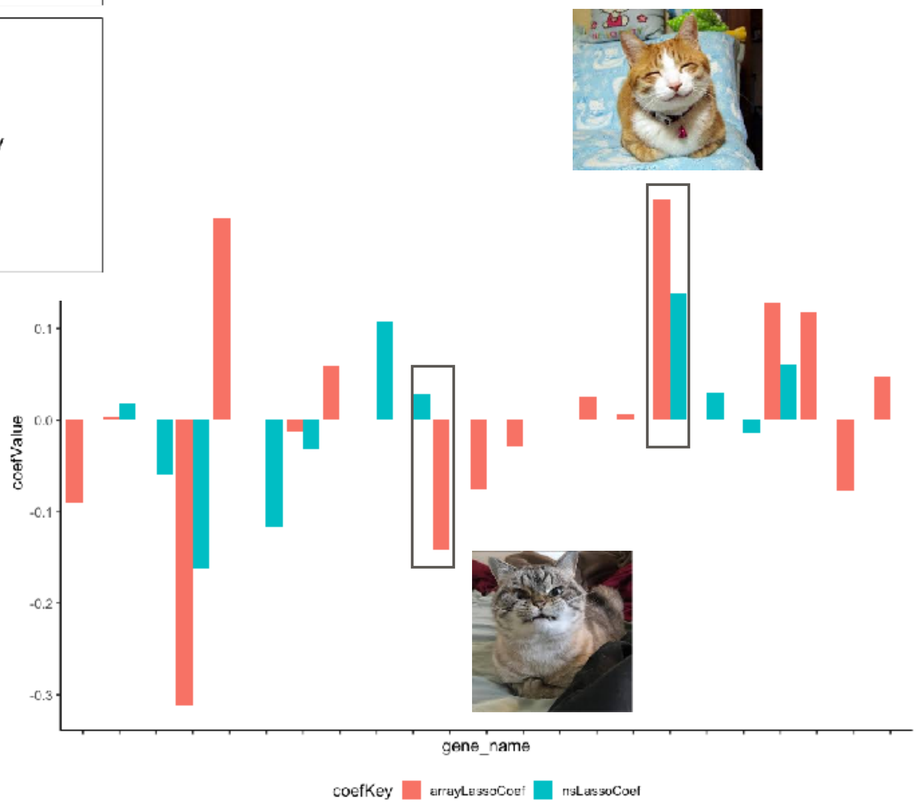
1. Perform a **weighted Lasso** by placing higher weights on features closer to the identity line



$$Z_1 \approx Z_2$$

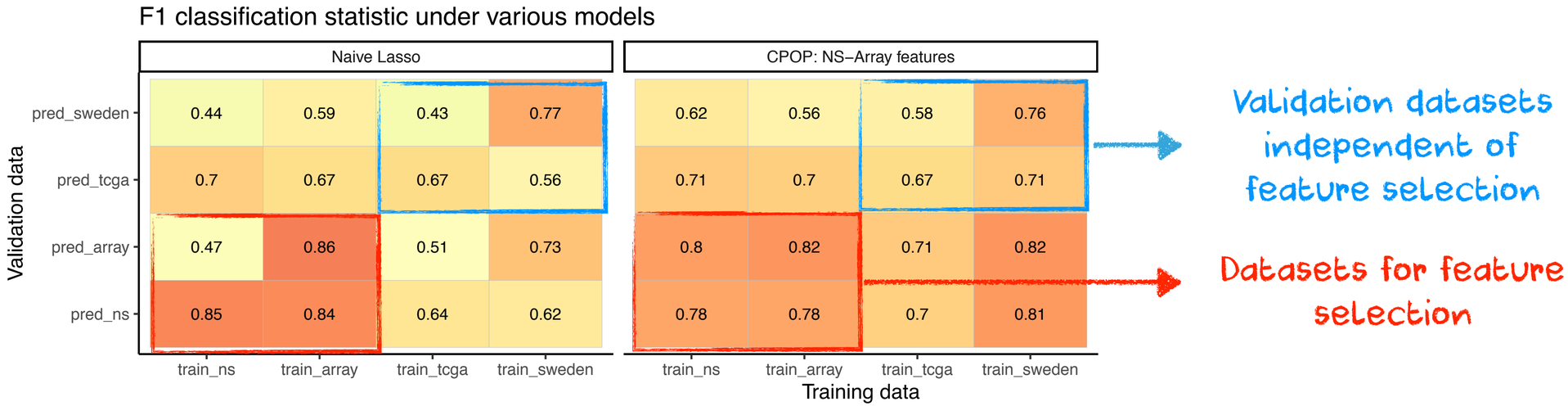
2. Perform a **Ridge regression** and only retain those features with coefficients similar to each other

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

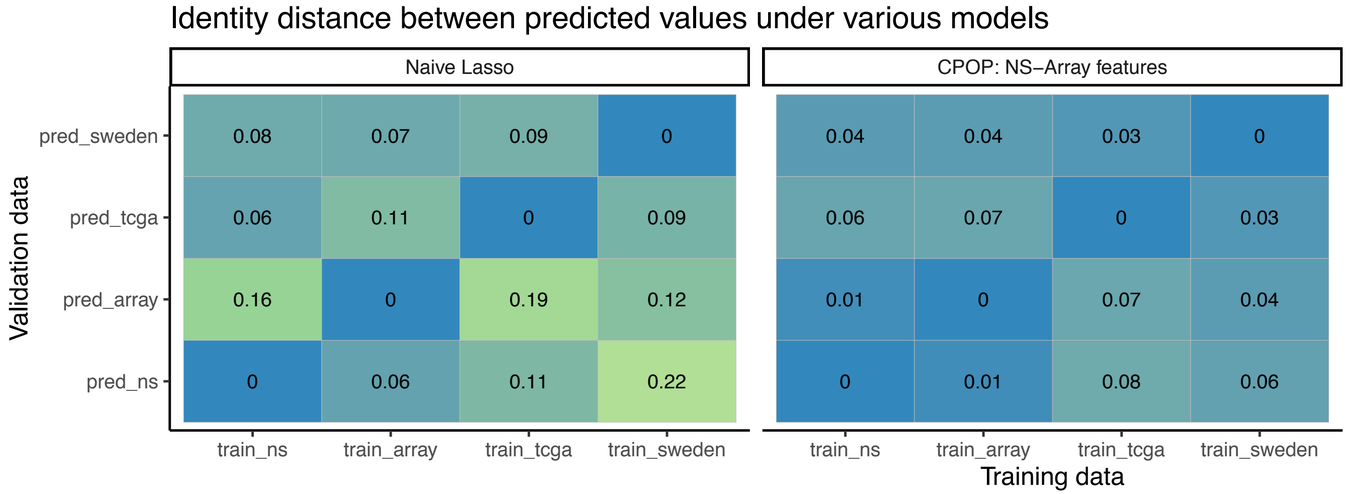


CPOP results 1: four melanoma data

1. Predictive performance of CPOP matches that of re-substitution

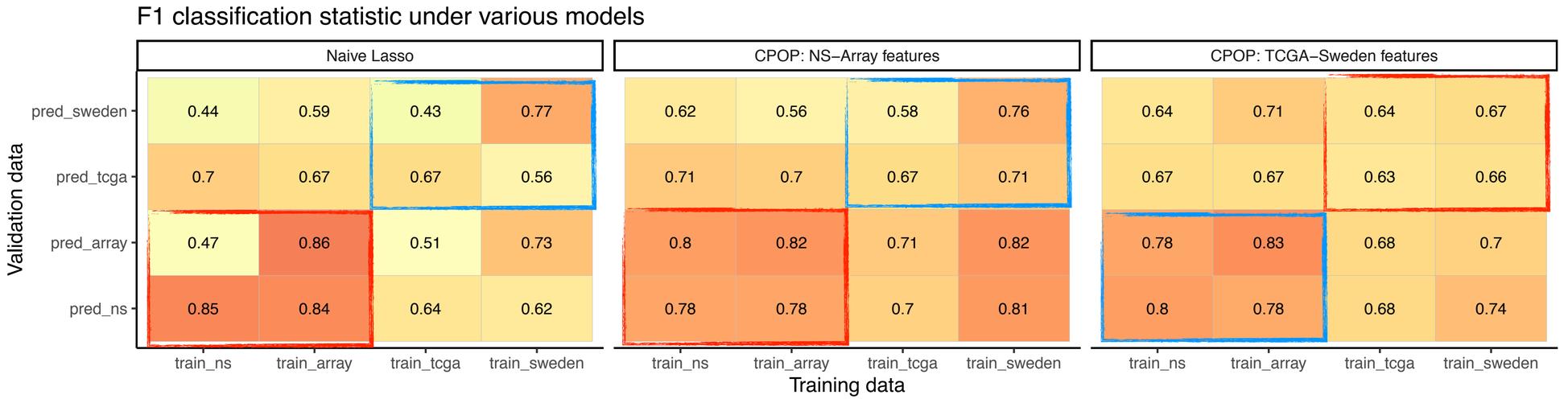


2. Smaller identity distance between predicted values



CPOP results 1: four melanoma data

1. Predictive performance of CPOP matches that of re-substitution

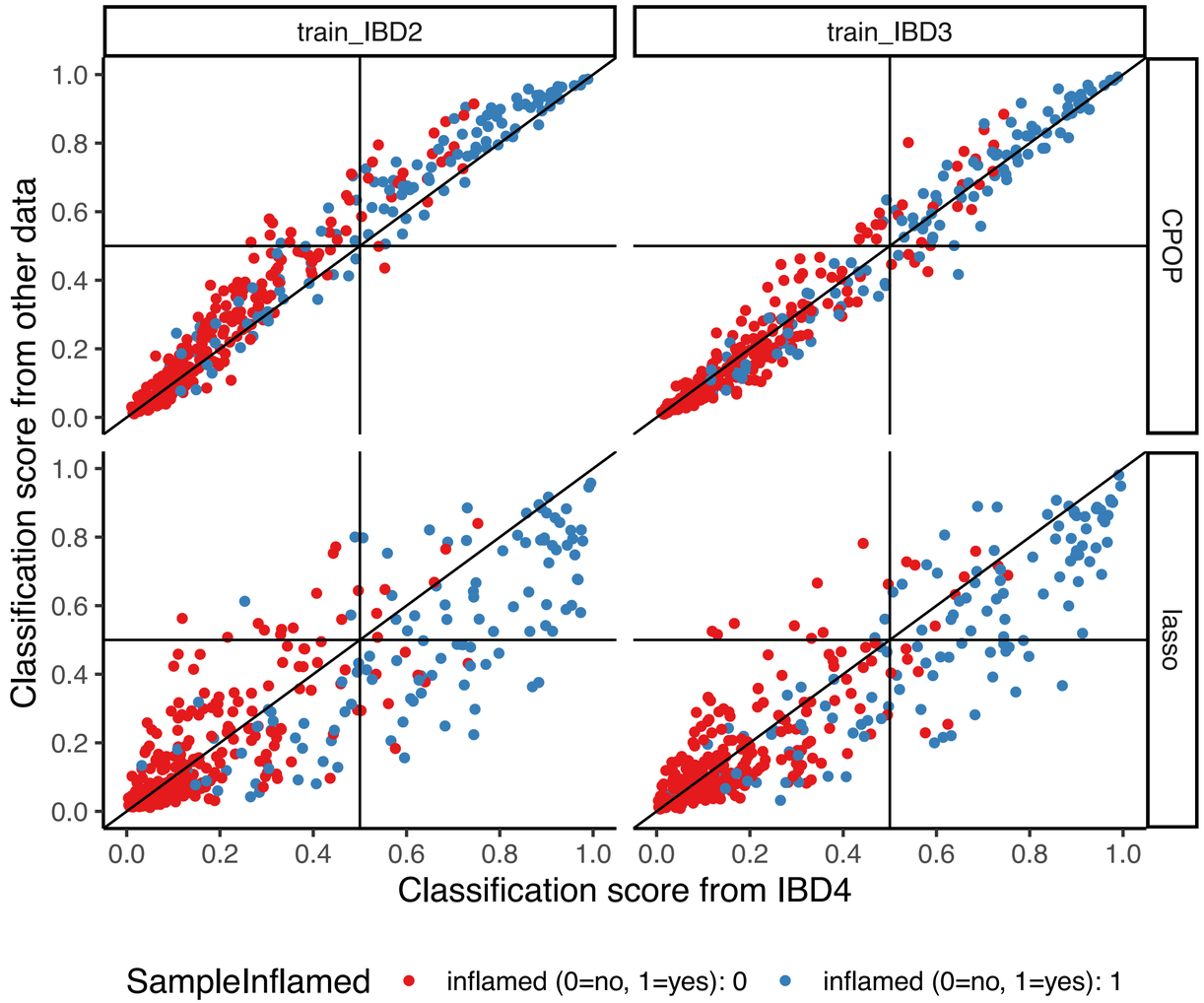


2. Smaller identity distance between predicted values

CPOP results 2: prospective prediction

- CPOP on IBD NanoString data demonstrated improvements on stability

- We are planning to exploring other data of higher relevance to precision medicine (e.g. drug sensitivity)



Concluding remarks

- CPOP is a flexible procedure that allows for
  - cross-platform omics prediction
  - stable single-patient prediction
- Not everyone is going to smooth-sailing through PhD, find your own way to deal with it