UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Predicting breast cancer metastasis from blood samples

“On variance and other problems”

Einar Holsbø

January, 2017

UNIVERSITY · OF TROMSØ ·

Q: can we predict metastasis from gene expression measurements in blood samples?

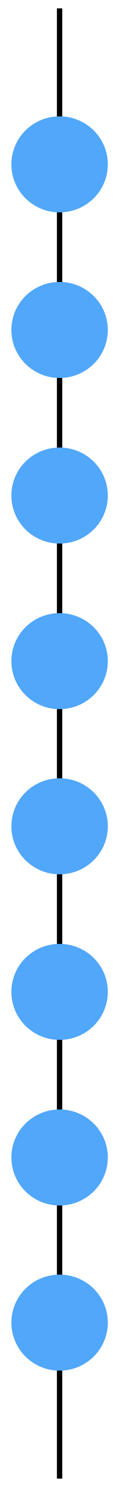
A: maybe

Norwegian Women and Cancer (NOWAC)

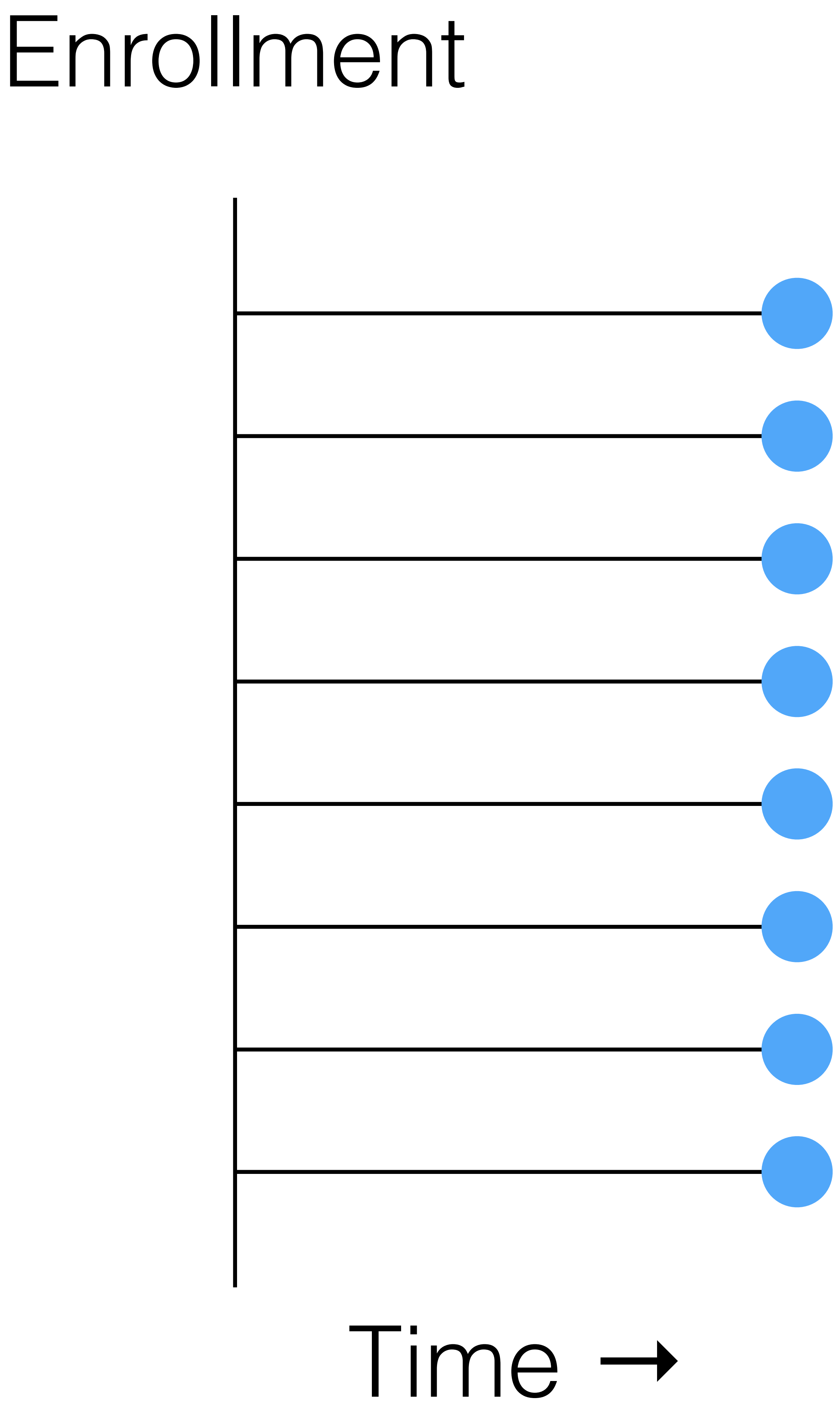
- Prospective case–control study
- Blood samples + questionnaires

Prospective

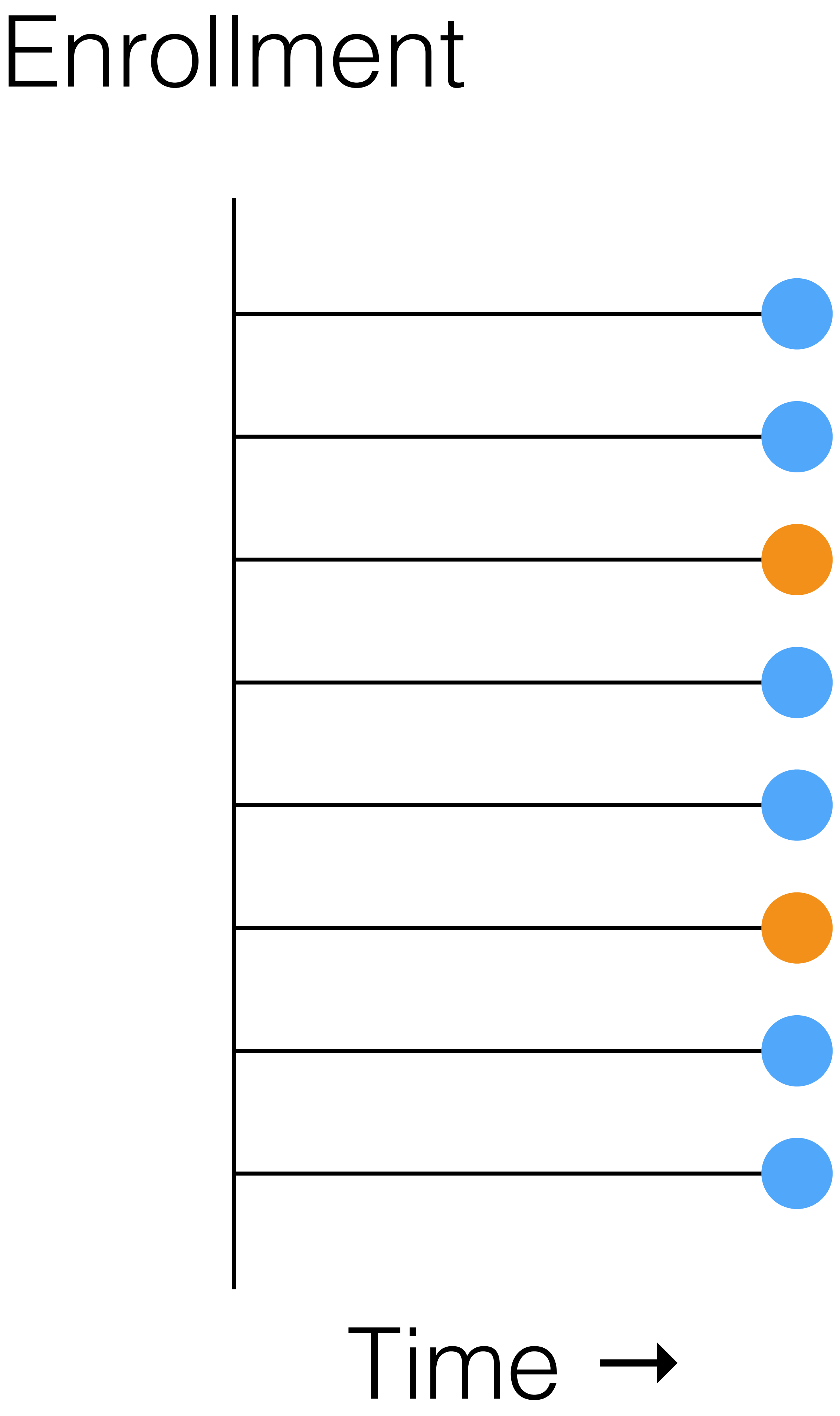
Enrollment



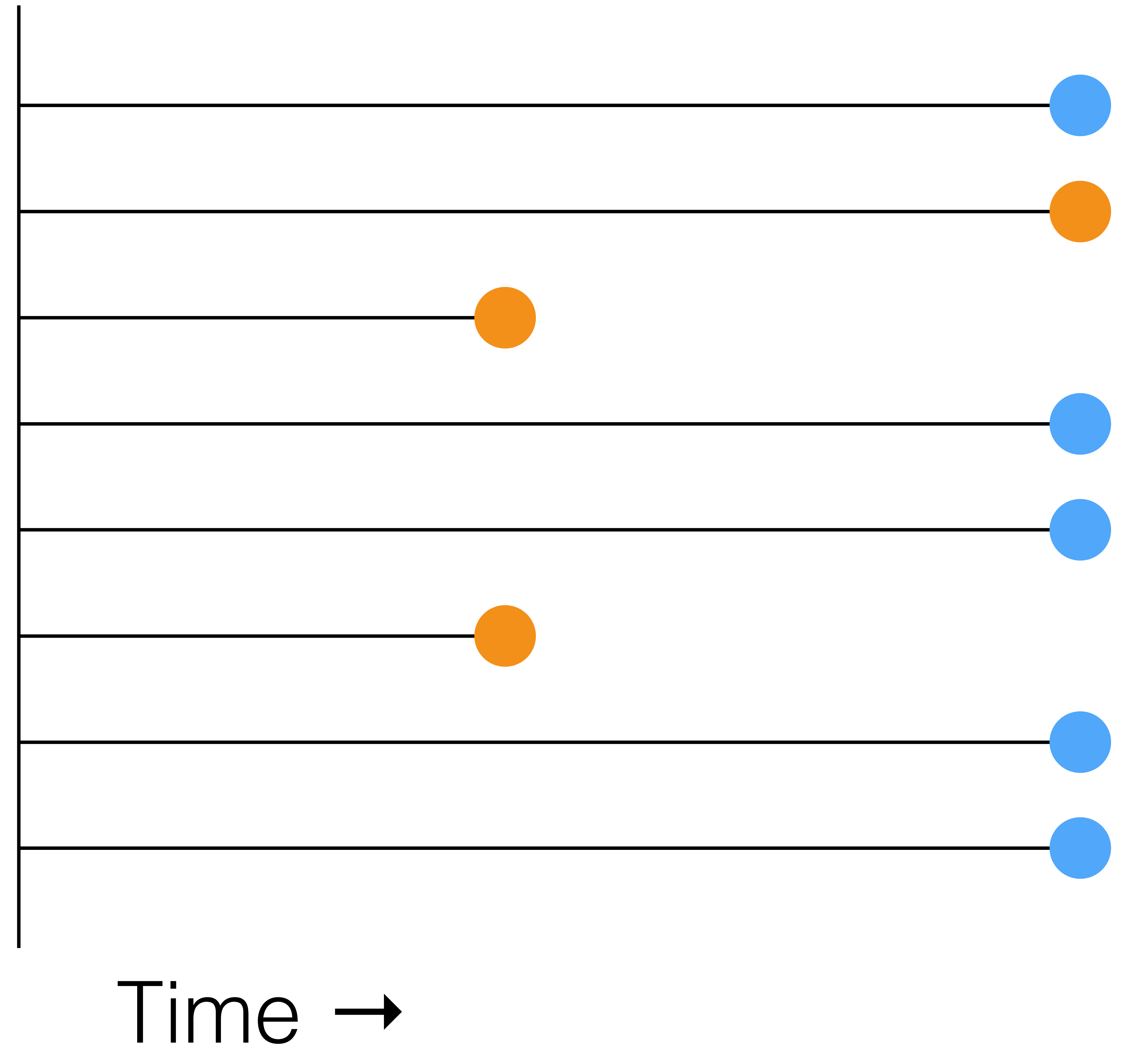
Prospective



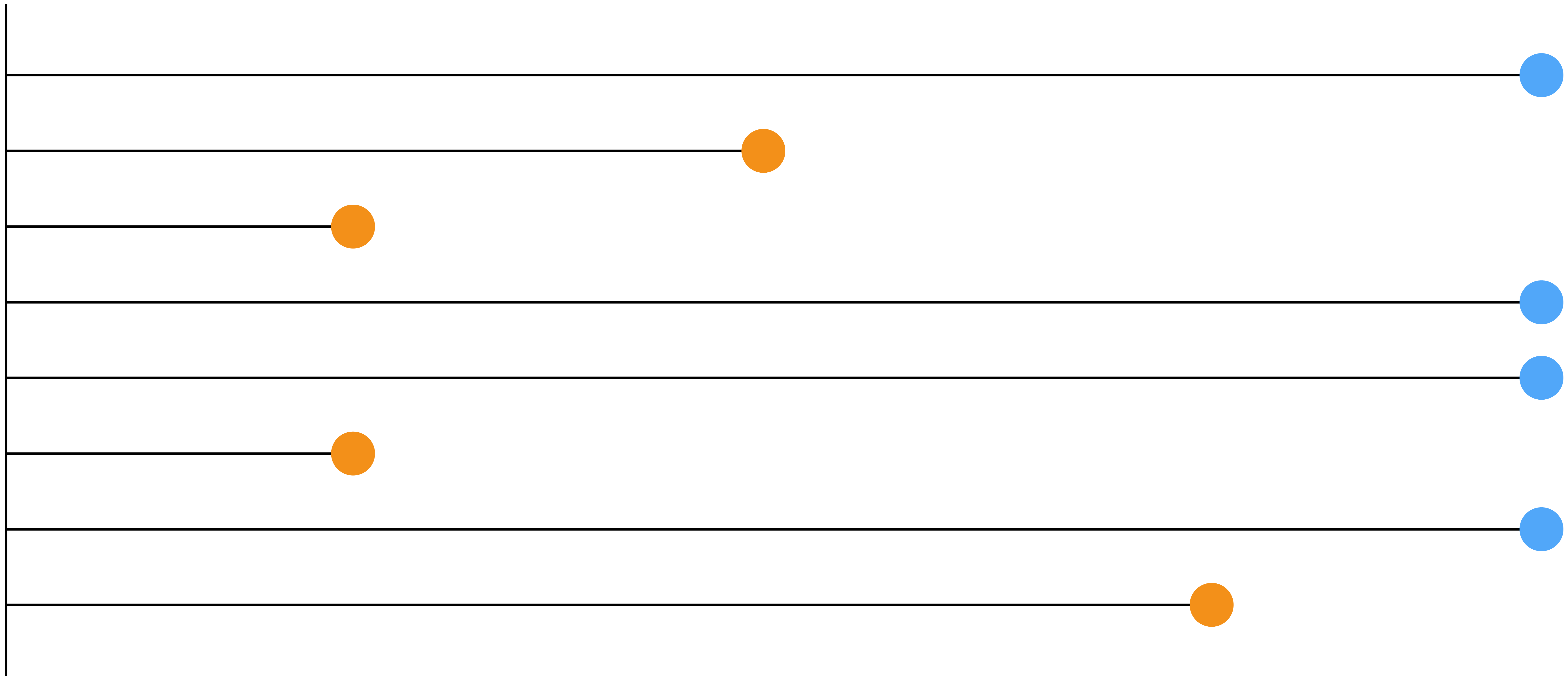
Prospective



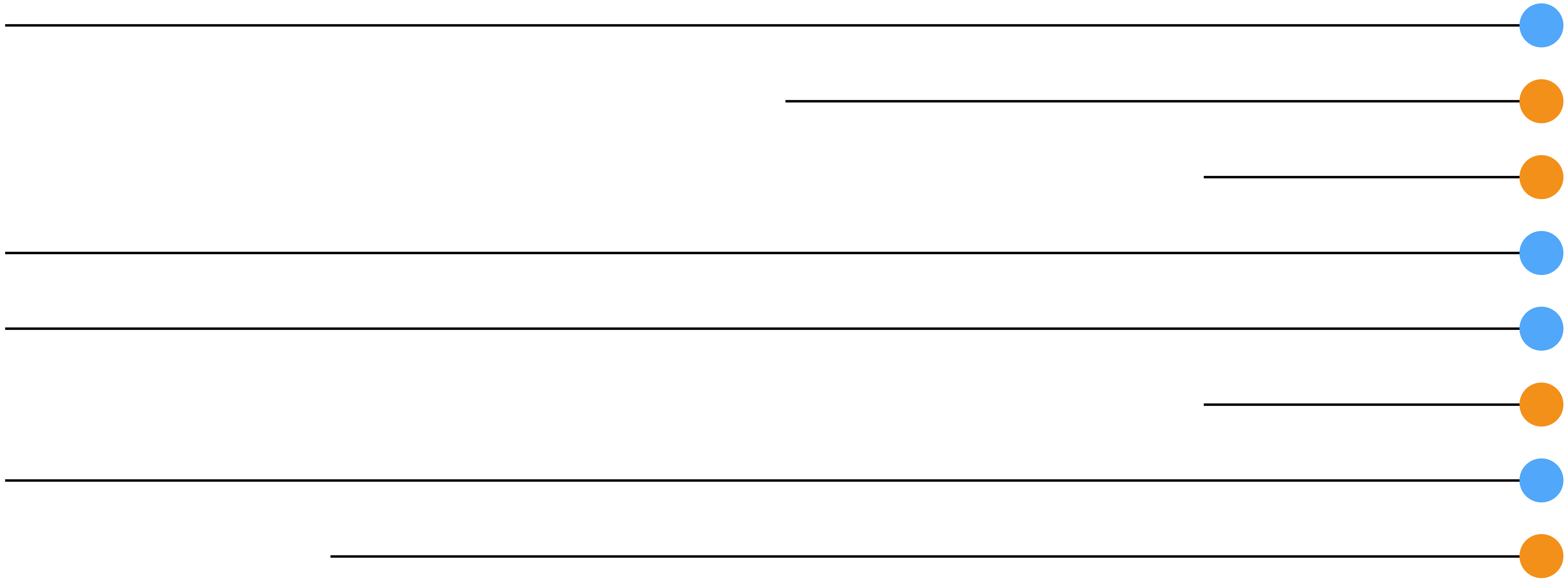
Prospective



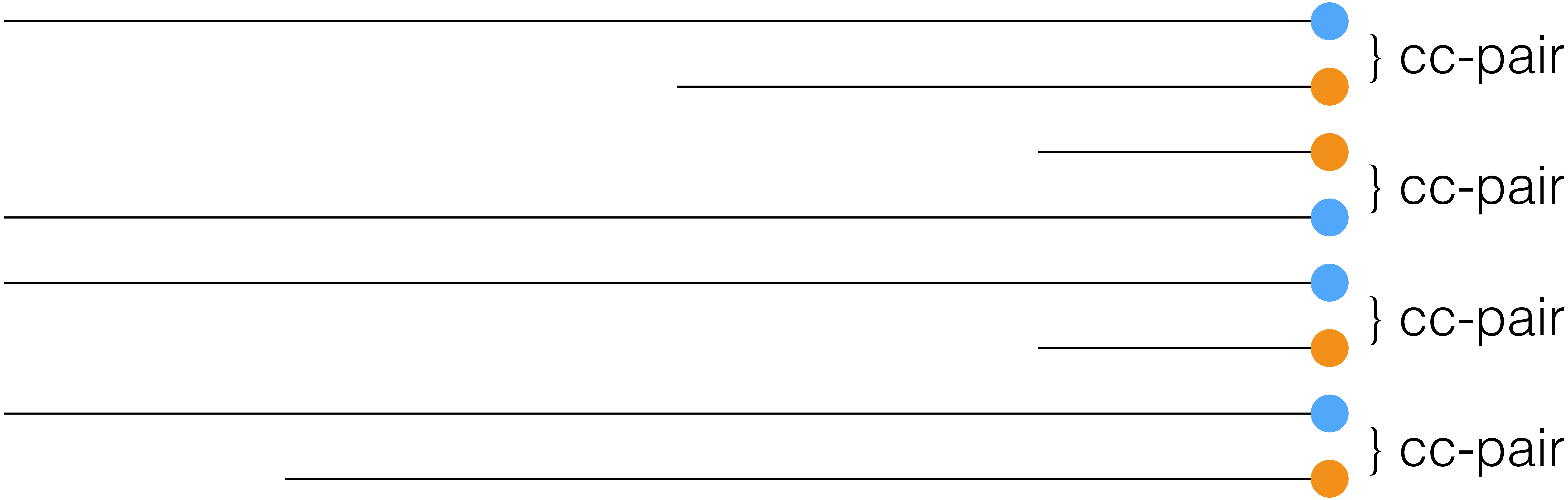
Prospective



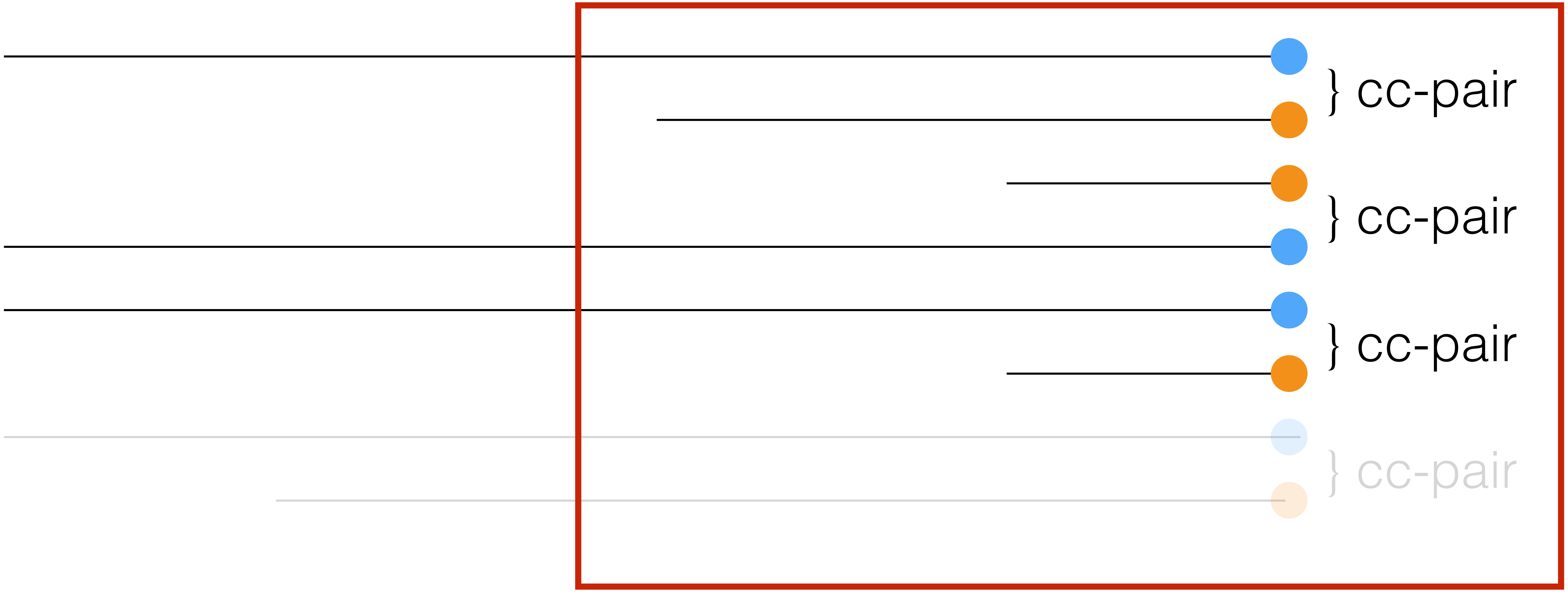
Prospective



Case–control



Case–control



1 year before diagnosis

Data at a glance

```
dim(gene_expression)
## [1]    88 12404

summary(days_to_diagnosis)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.0   117.8   189.5   186.8   269.2   358.0

summary(metastasis)
## FALSE  TRUE
##    66    22

table(metastasis, stratum)
##           stratum
## metastasis screening interval clinical
##      FALSE        43       10       13
##      TRUE          6        6       10
```

How to do predictive modelling

1. Pick some of your favorite models

2. Evaluate model performance by cross-validation

3. Fit tuning parameters by nested cross-validation

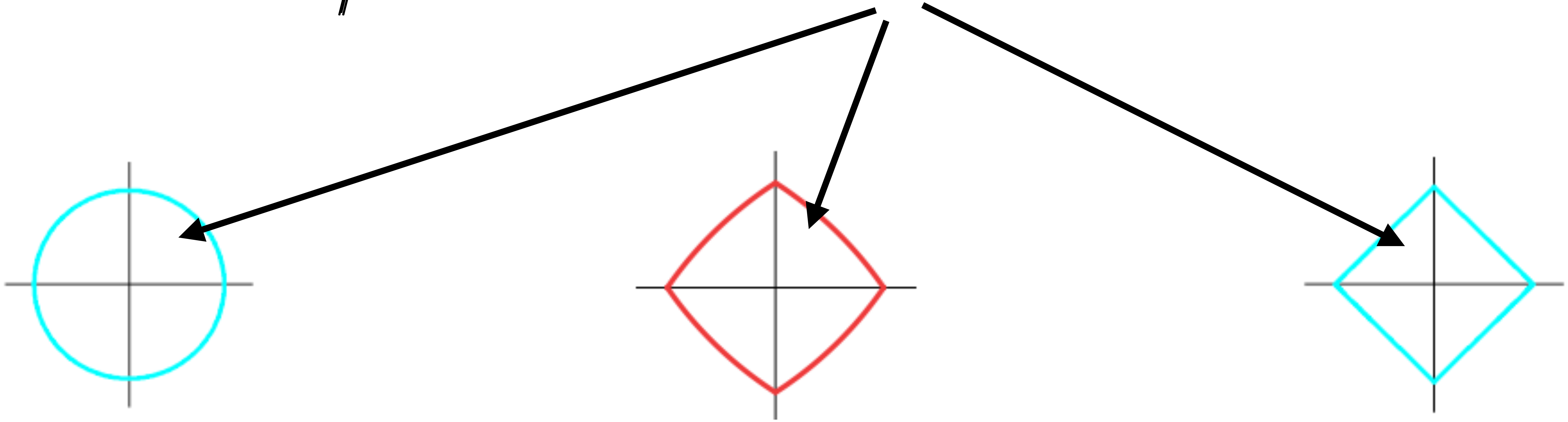
Some models

Penalized logistic regression

$$\text{find } \hat{\boldsymbol{\beta}} \text{ s.t. } \log \frac{p(Y|x)}{1-p(Y|x)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots$$

Some models

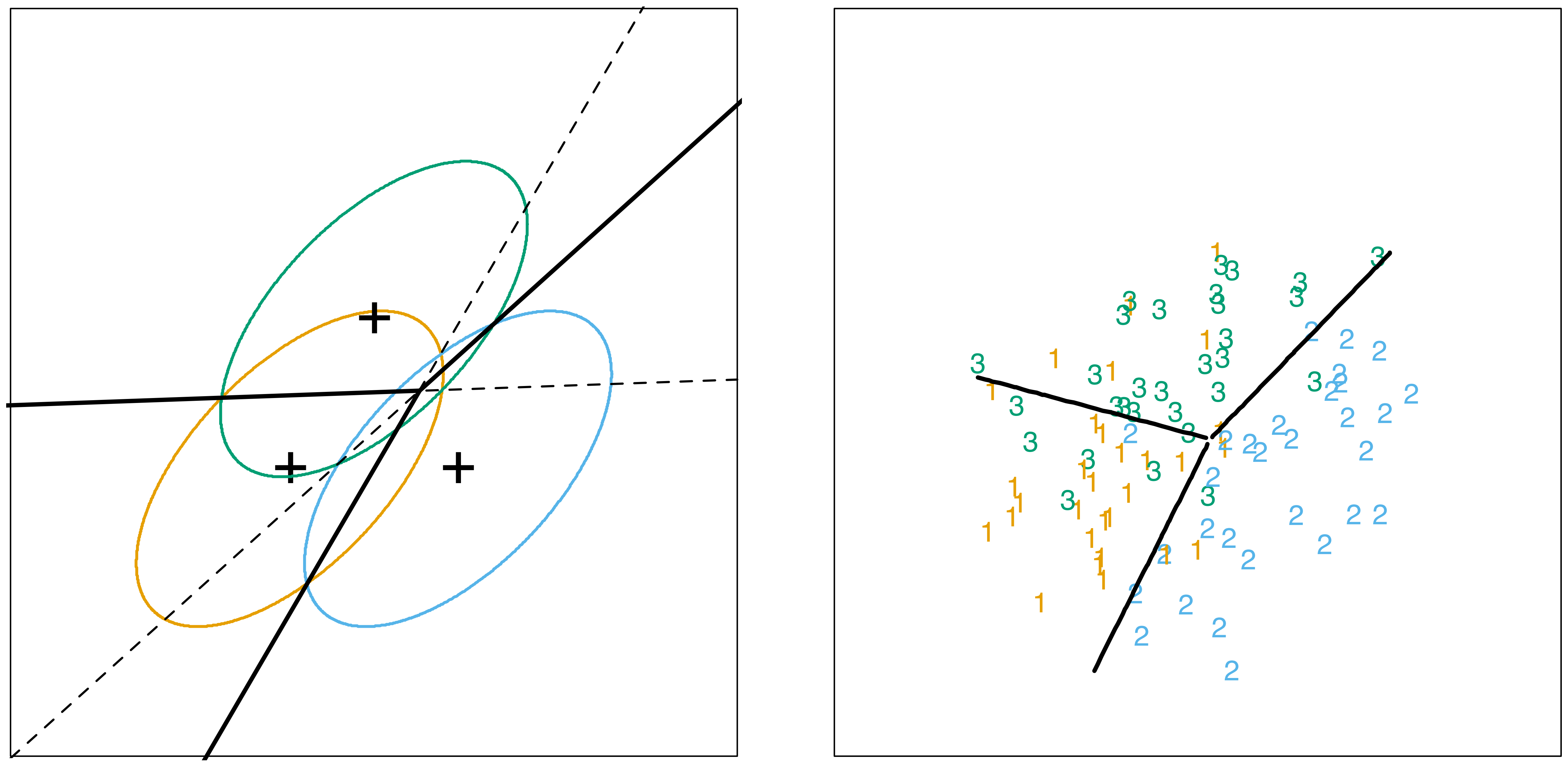
$\hat{\beta}$ can only be inside these shapes



From Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

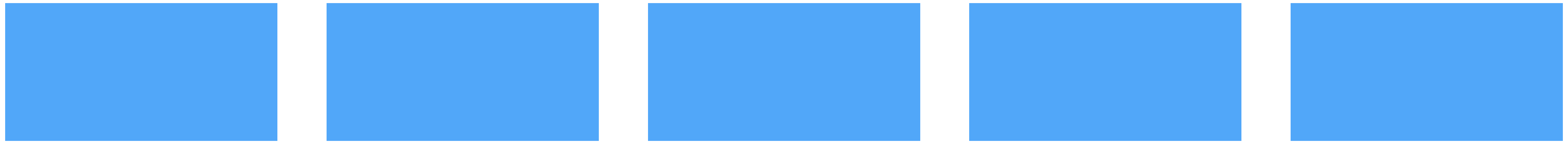
Some models

Nearest centroids

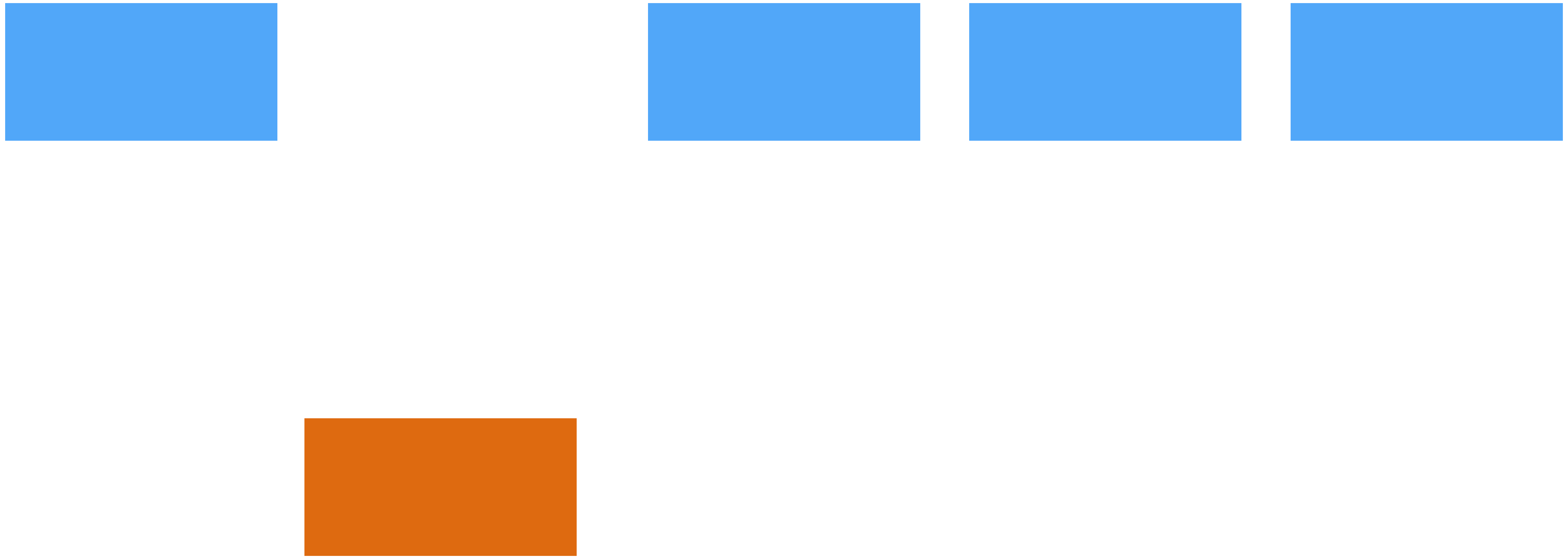


From Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

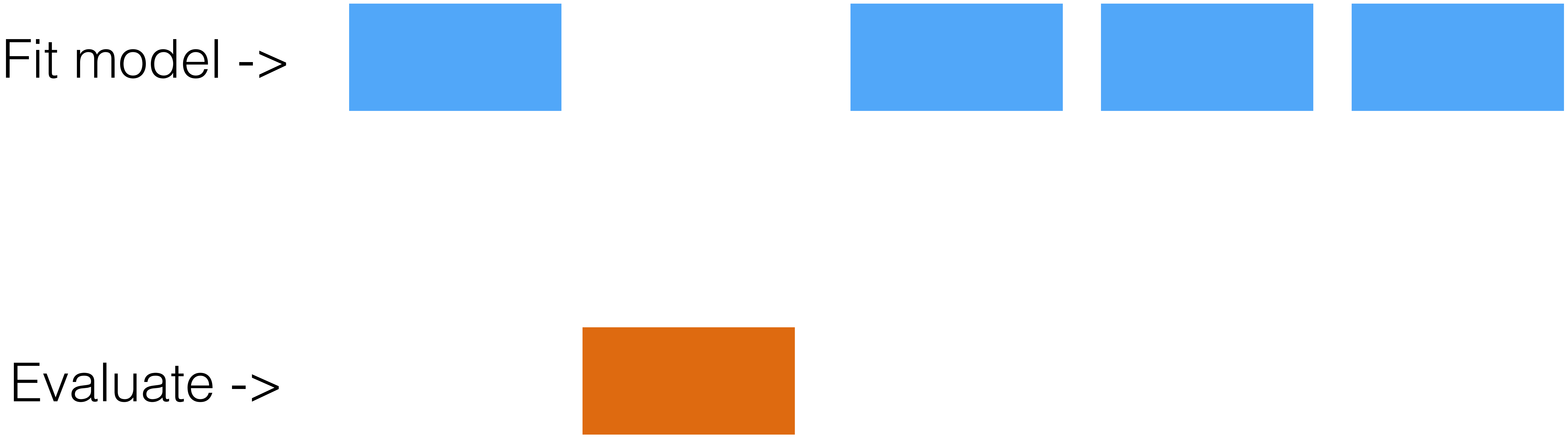
Cross validation



Cross validation



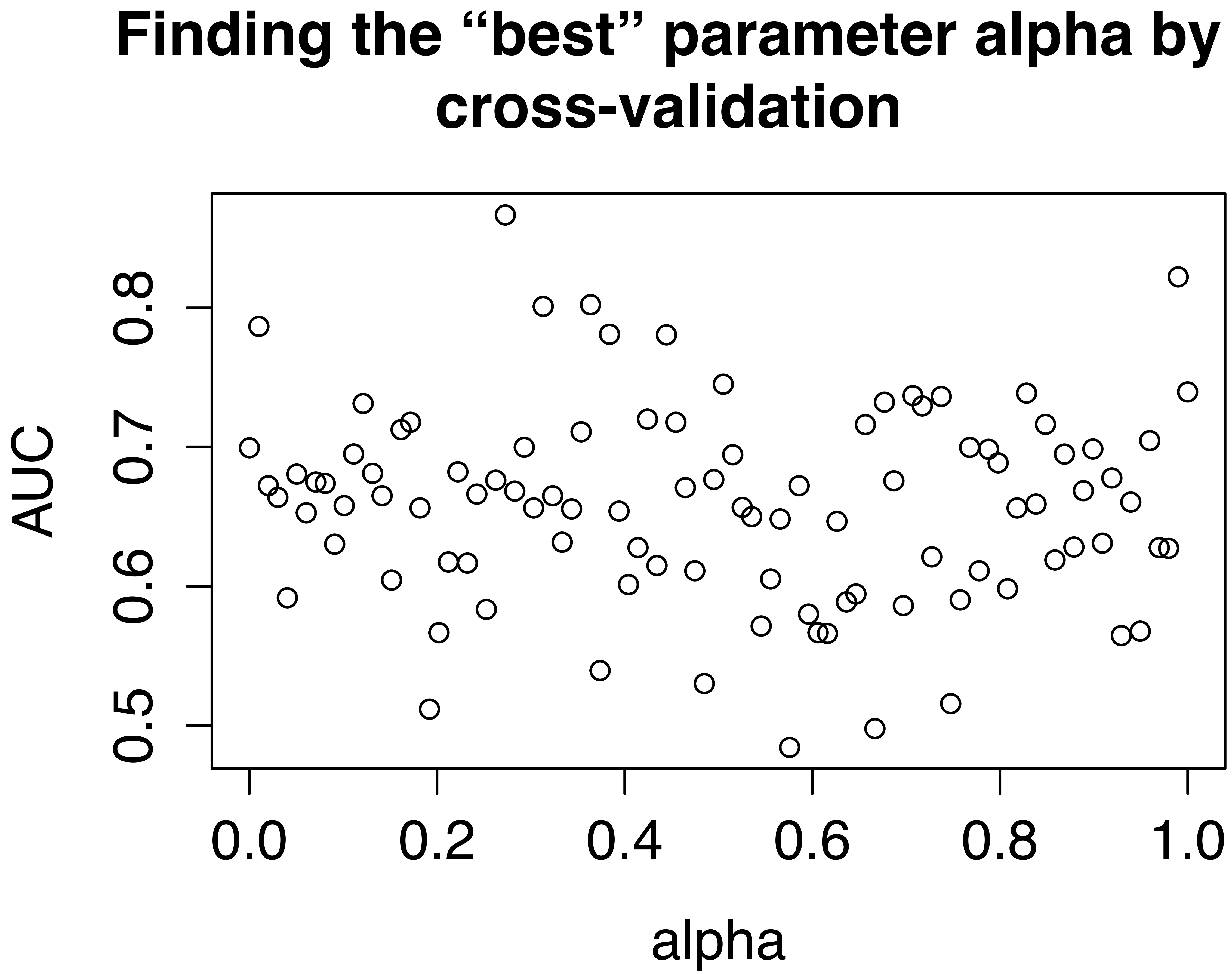
Cross validation



Fit model ->

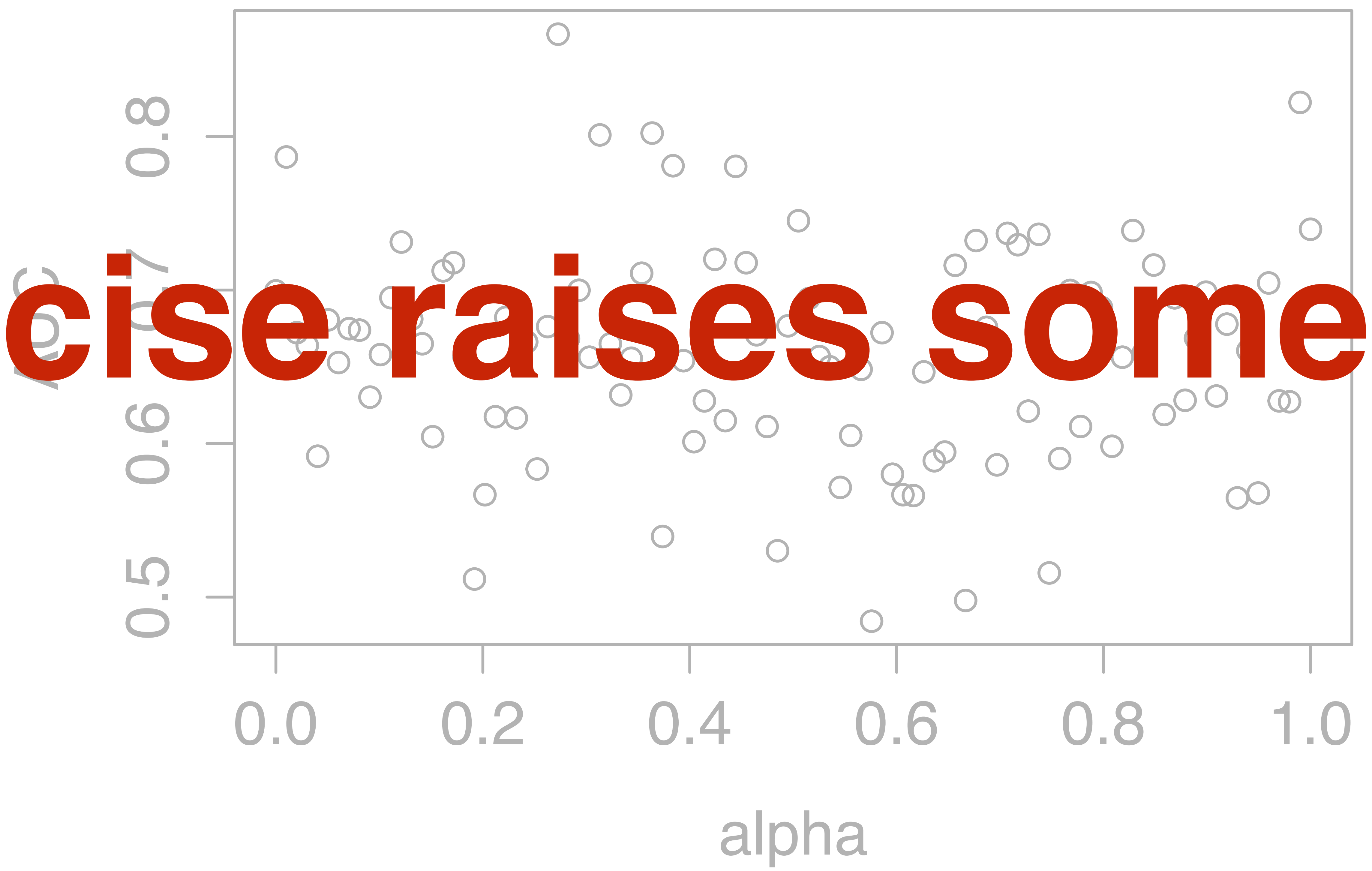
Evaluate ->

Fit tuning parameters????????



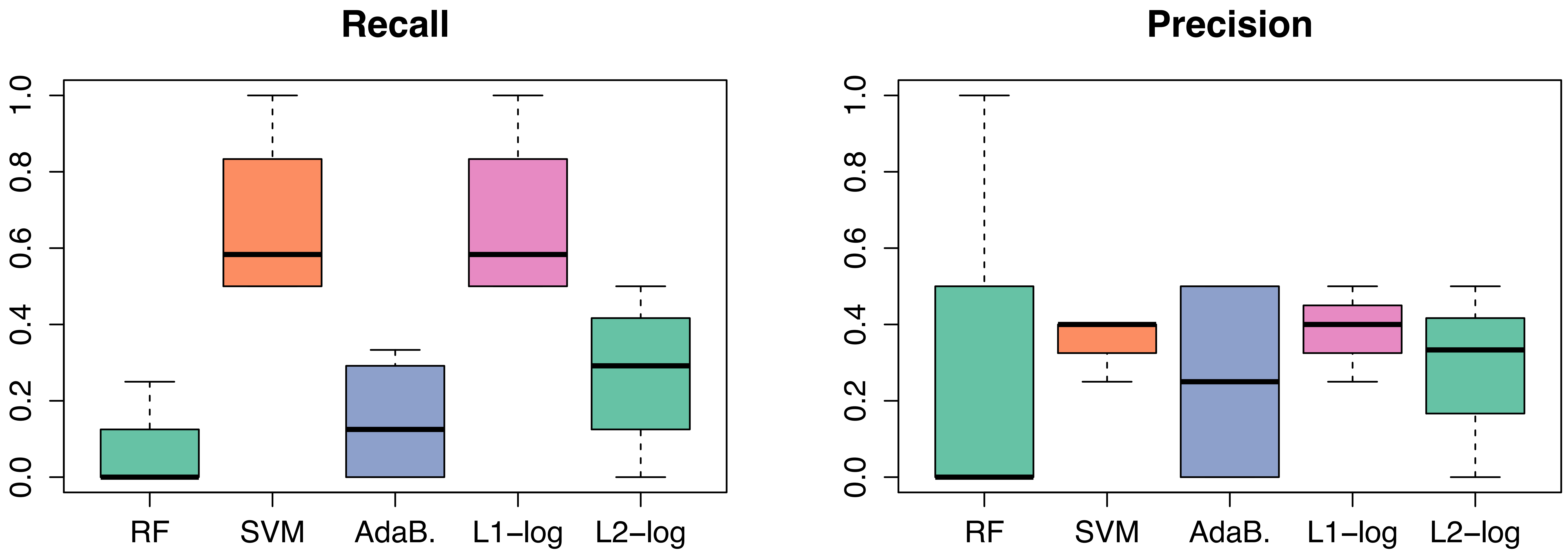
Fit tuning parameters????????

Finding the “best” parameter alpha by cross-validation

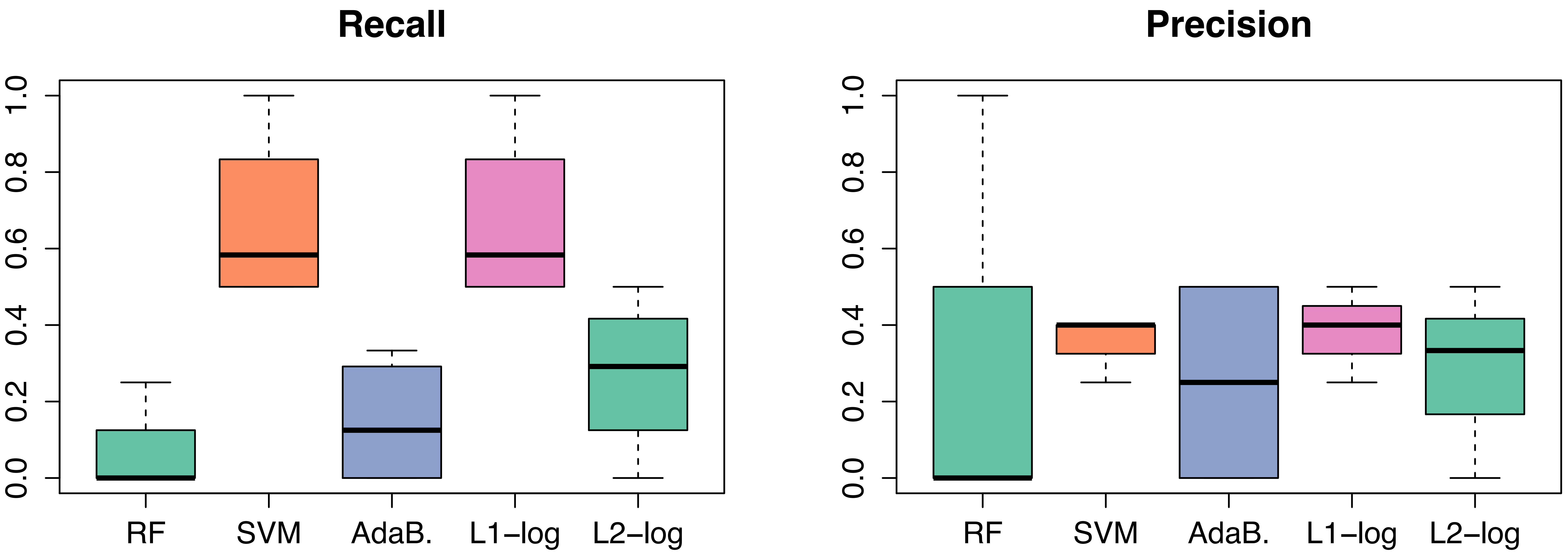


This exercise raises some questions

Cross validation is almost useless to me

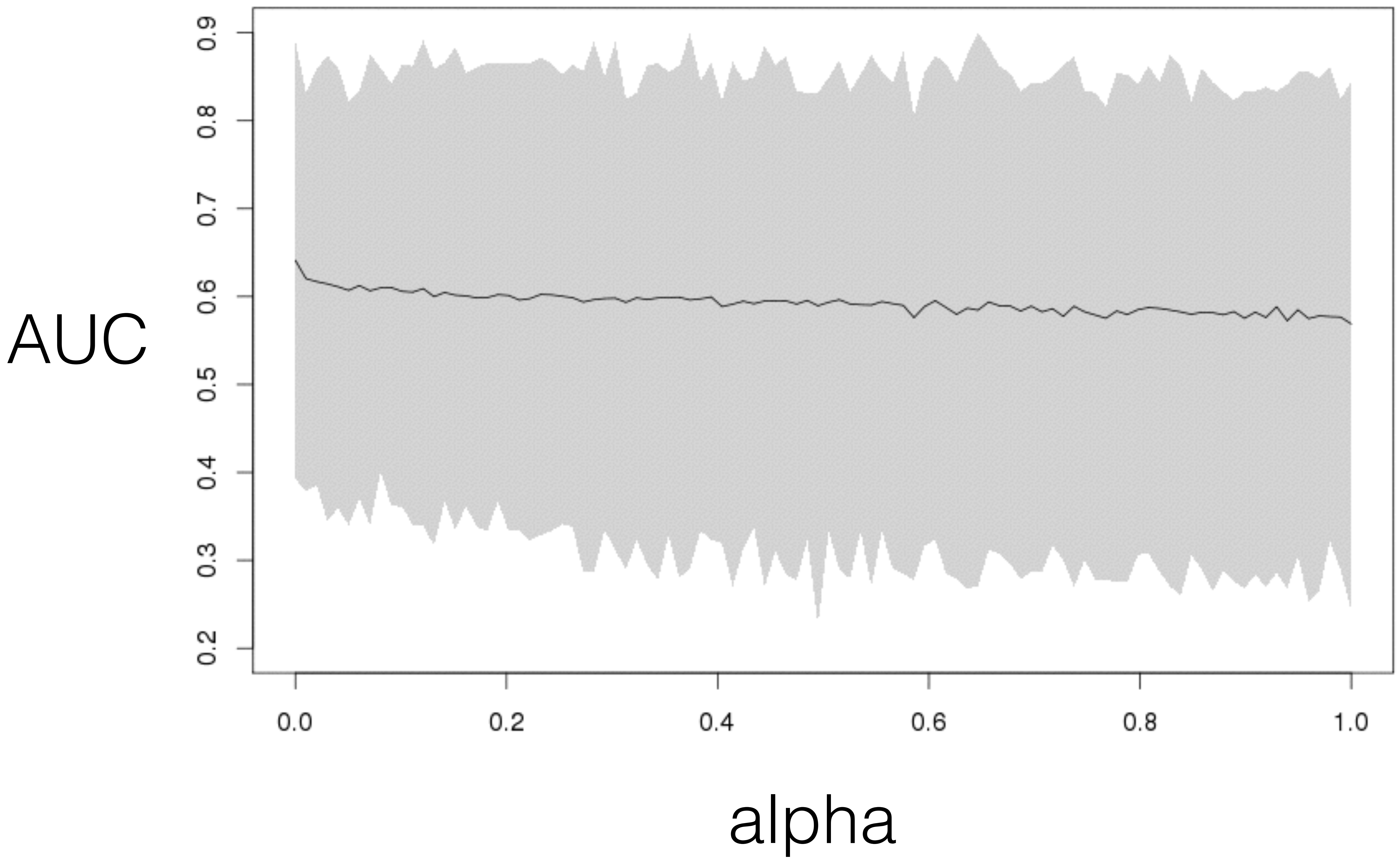


Cross validation is almost useless to me

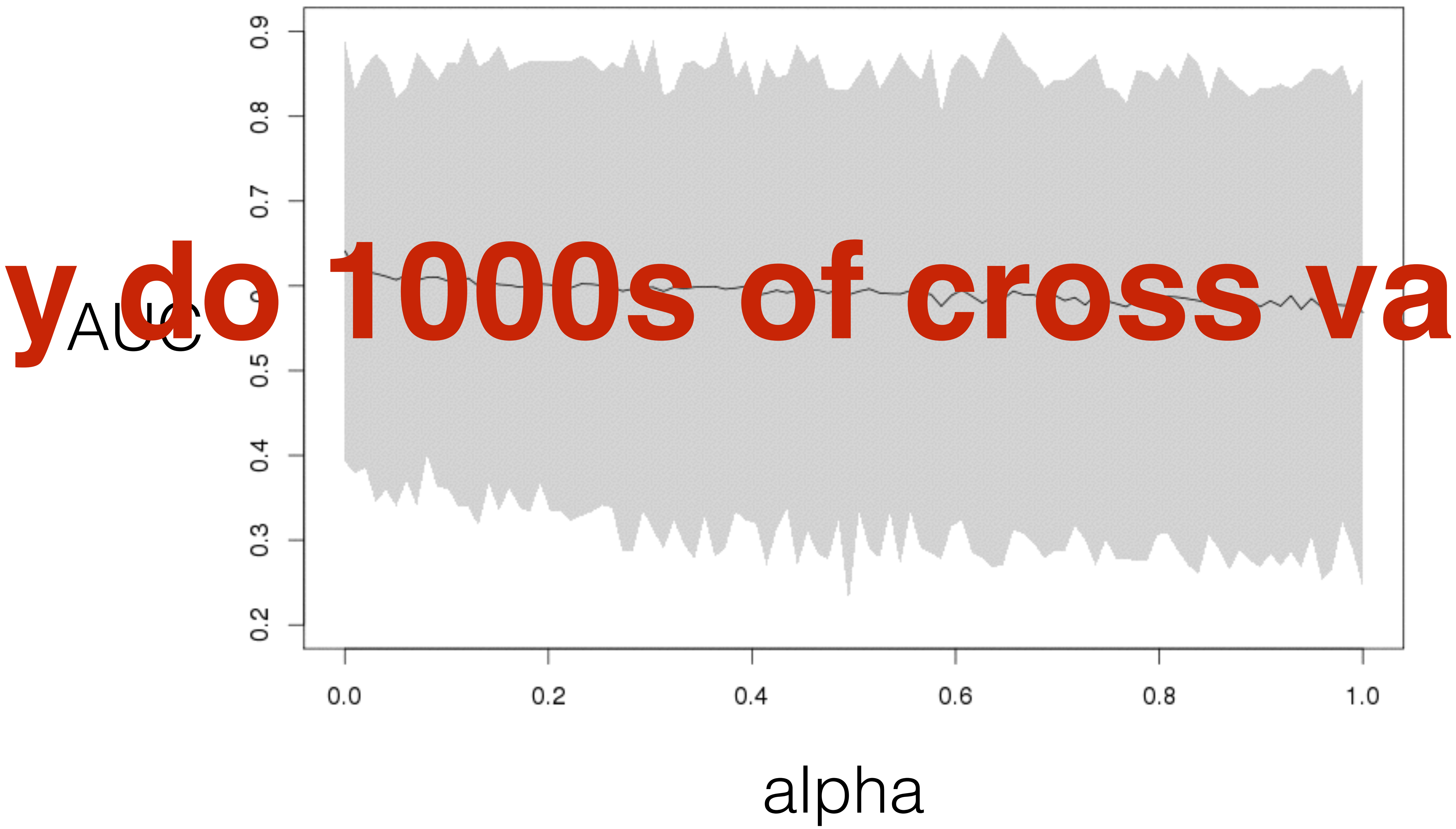


I spent actual time interpreting plots like these……………………

Solution: resampling

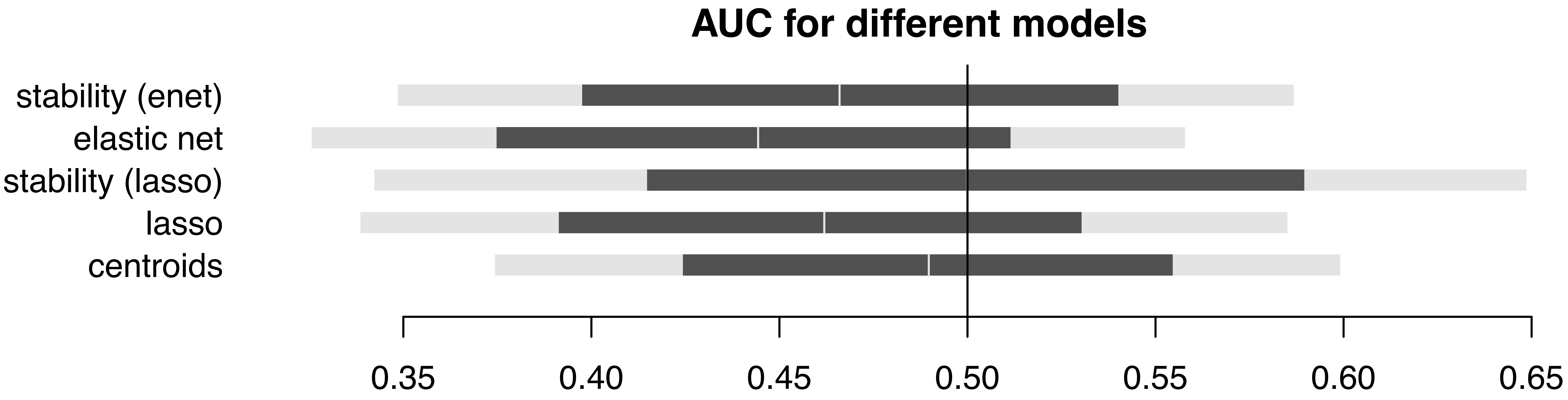


Solution: resampling

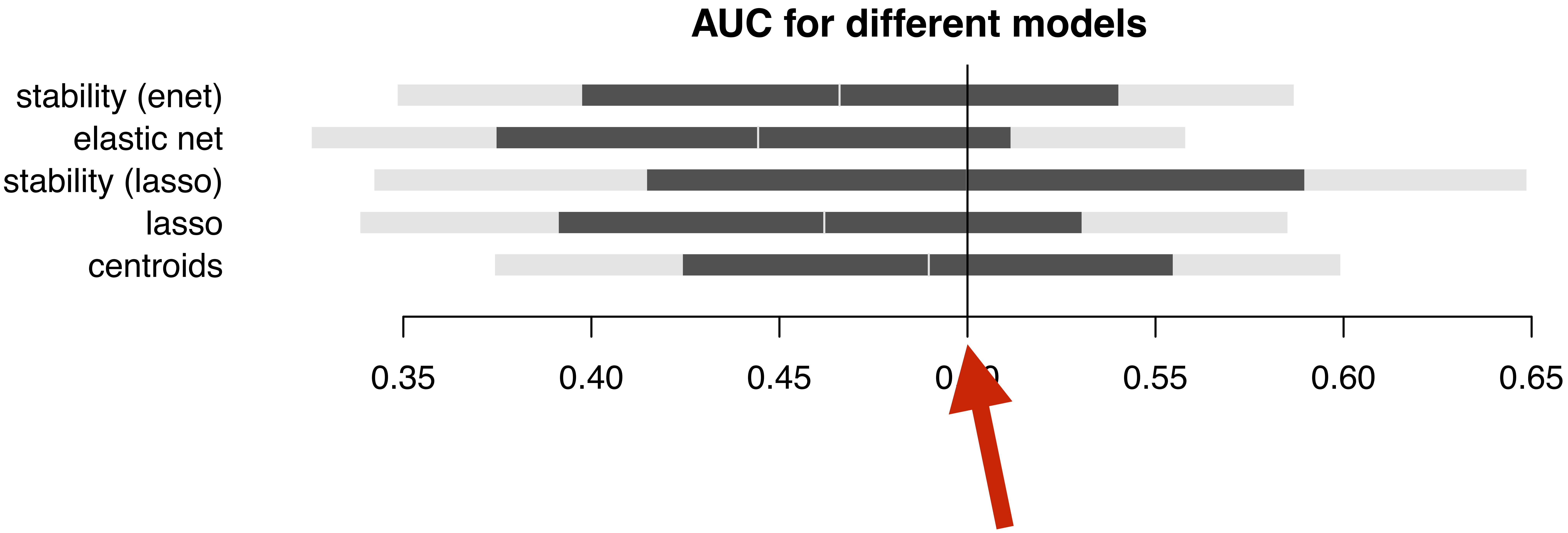


Simply do 1000s of cross validations

Another confusing thing



Another confusing thing



The line for random guess

2 ways to get AUROC < .5

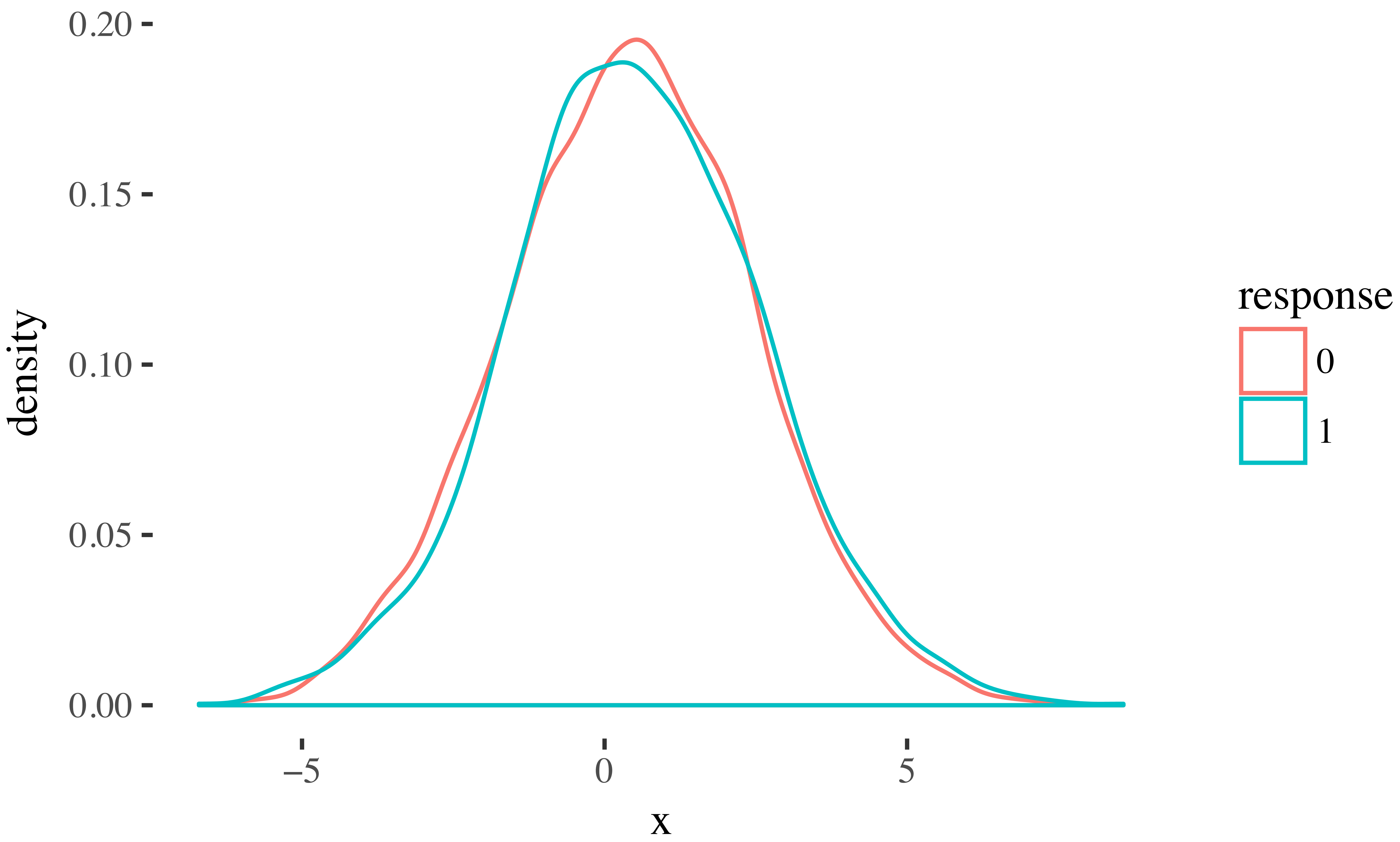
A. You made a mistake calculating AUC

B. There is something v. strange with the data

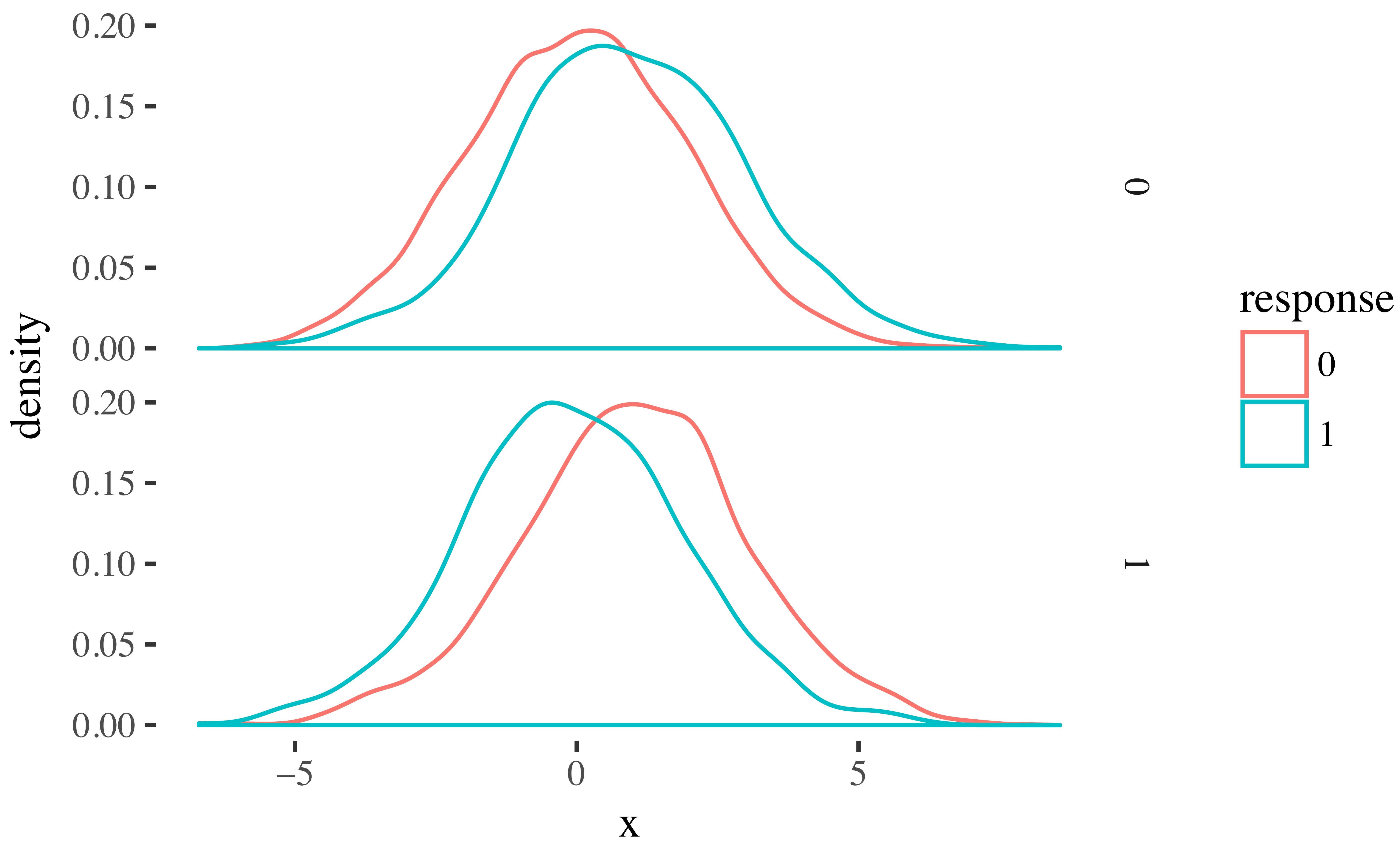
A simulated paradox

- One “gene,” x
- Response 1 or 0
- Two strata: 1 and 0
- If stratum == response, x ~ N(1, variance)
- Else, x ~ N(0, variance)

A simulated paradox



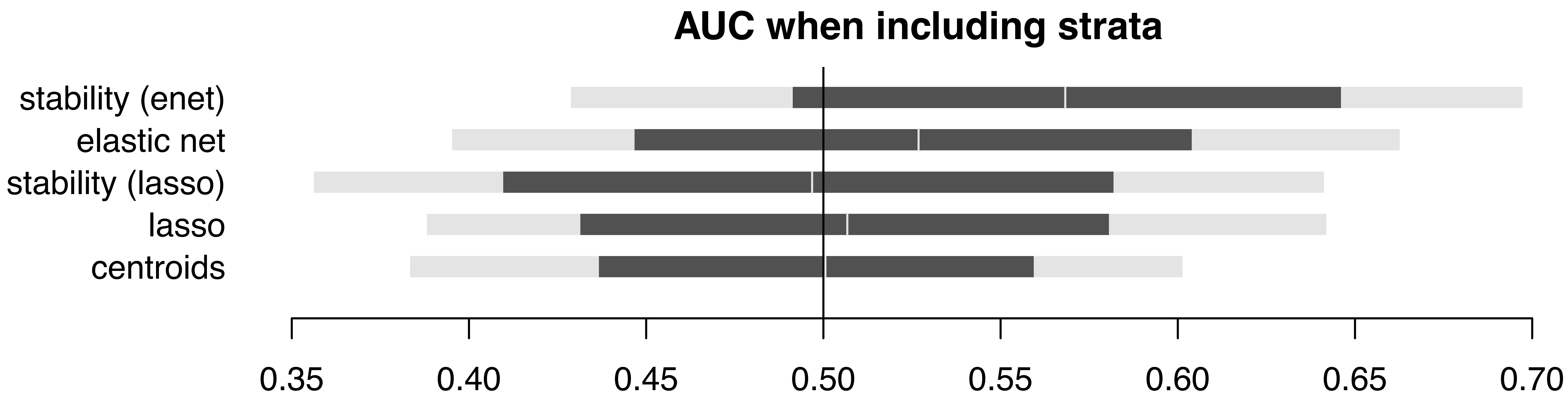
A simulated paradox



“You have to stratify.”

–Eiliv Lund to myself, like two-and-a-half years ago

Including stratum gives expected null behavior

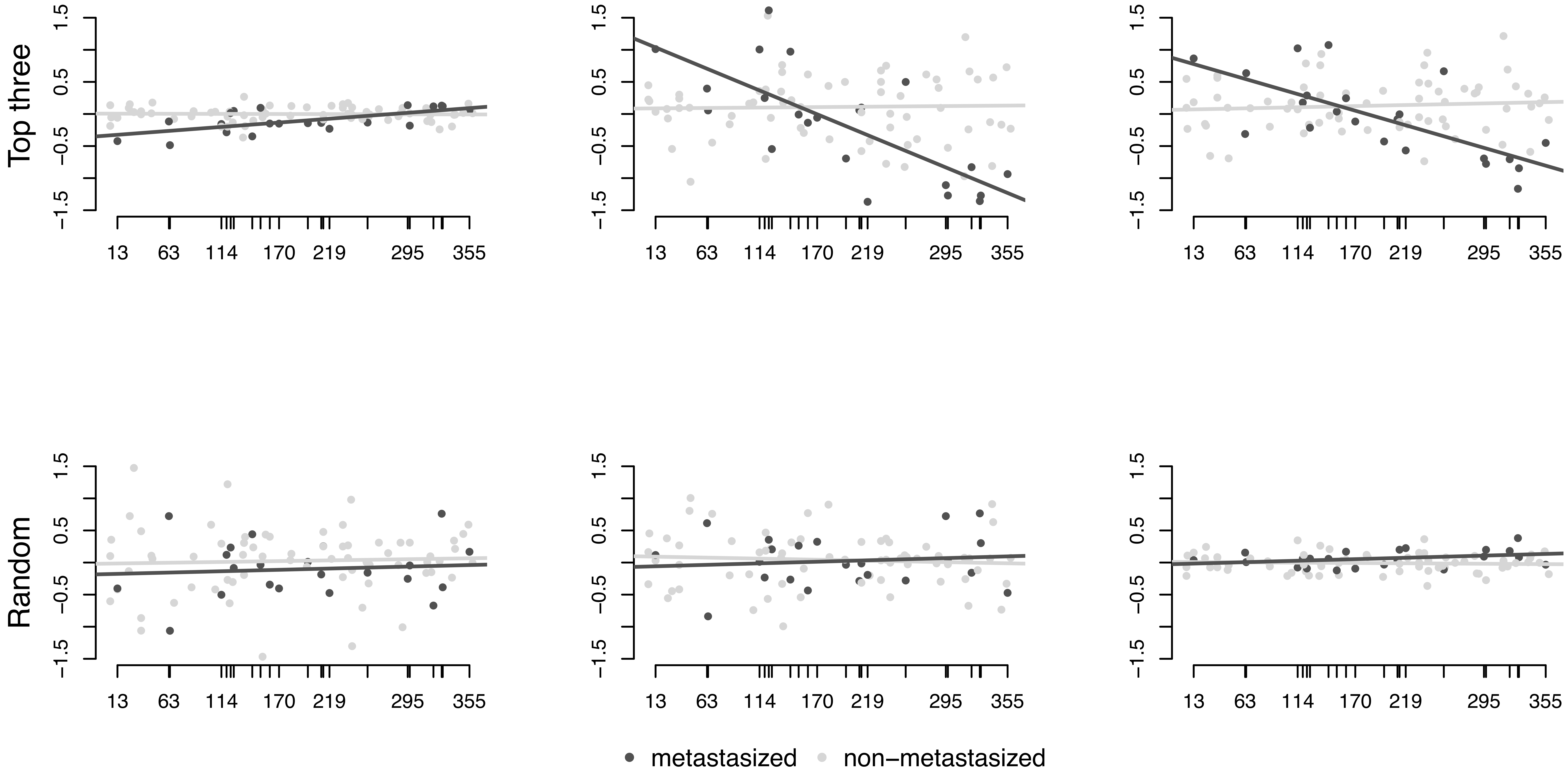


Introducing some bias: focus on a likely subspace

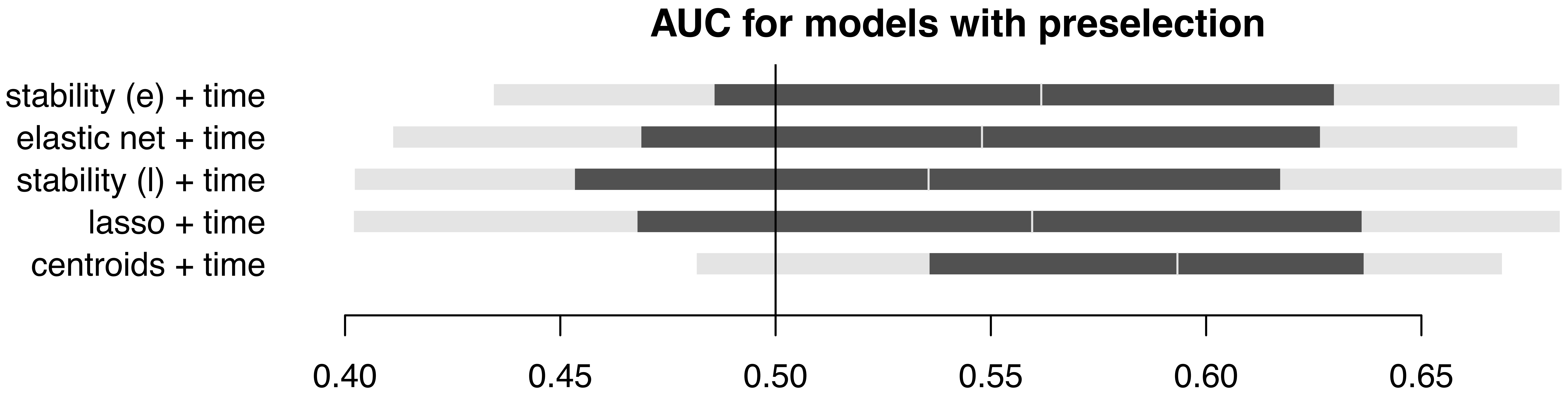
- In high dimensions, bias is your friend
- Theory: there is something going on in the gene expression as we get closer to diagnosis
- Rank by linear model:

$$\text{expression} = \beta_0 + \beta_1 \text{time} + \beta_2 \text{metastasis} + \beta_3 \text{time} \times \text{metastasis} + \text{error}$$

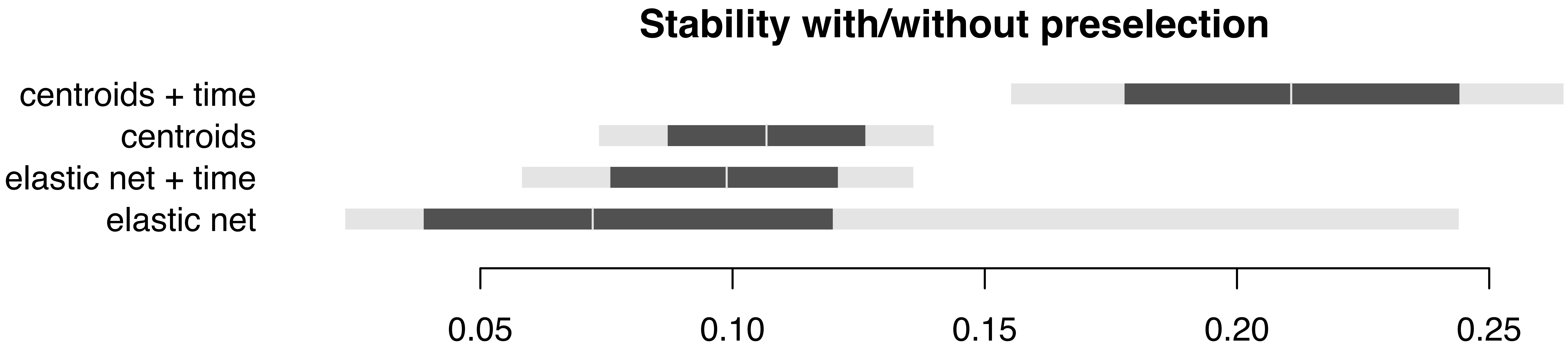
log(fold change) as linear function of time−to−diagnosis



Improved predictions



Lower variance/higher stability



Stability = set overlap between predictive genes across two resmaplings

Lessons/perspectives

- Cross validation can actually be super high in variance, be careful
- But be especially careful of holdout set validation
- Remember Simpson’s paradox, watch your strata
- Be critical of Signatures

Lessons/perspectives

- OTOH: There seems to be some weak signal here

These are my advisers

- **Lars Ailo Bongo**, BDPS group, University of Tromsø
- **Etienne Birmelé**, MAP5, Université Paris Descartes
- **Eiliv Lund**, Department of Community Medicine, University of Tromsø

Thank you!

email: einar@cs.uit.no
twitter: @0xeinar
github: github.com/3inar

Slides available online at 3inar.github.io/talks/