UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Small data: practical modeling issues in human-model -omic data

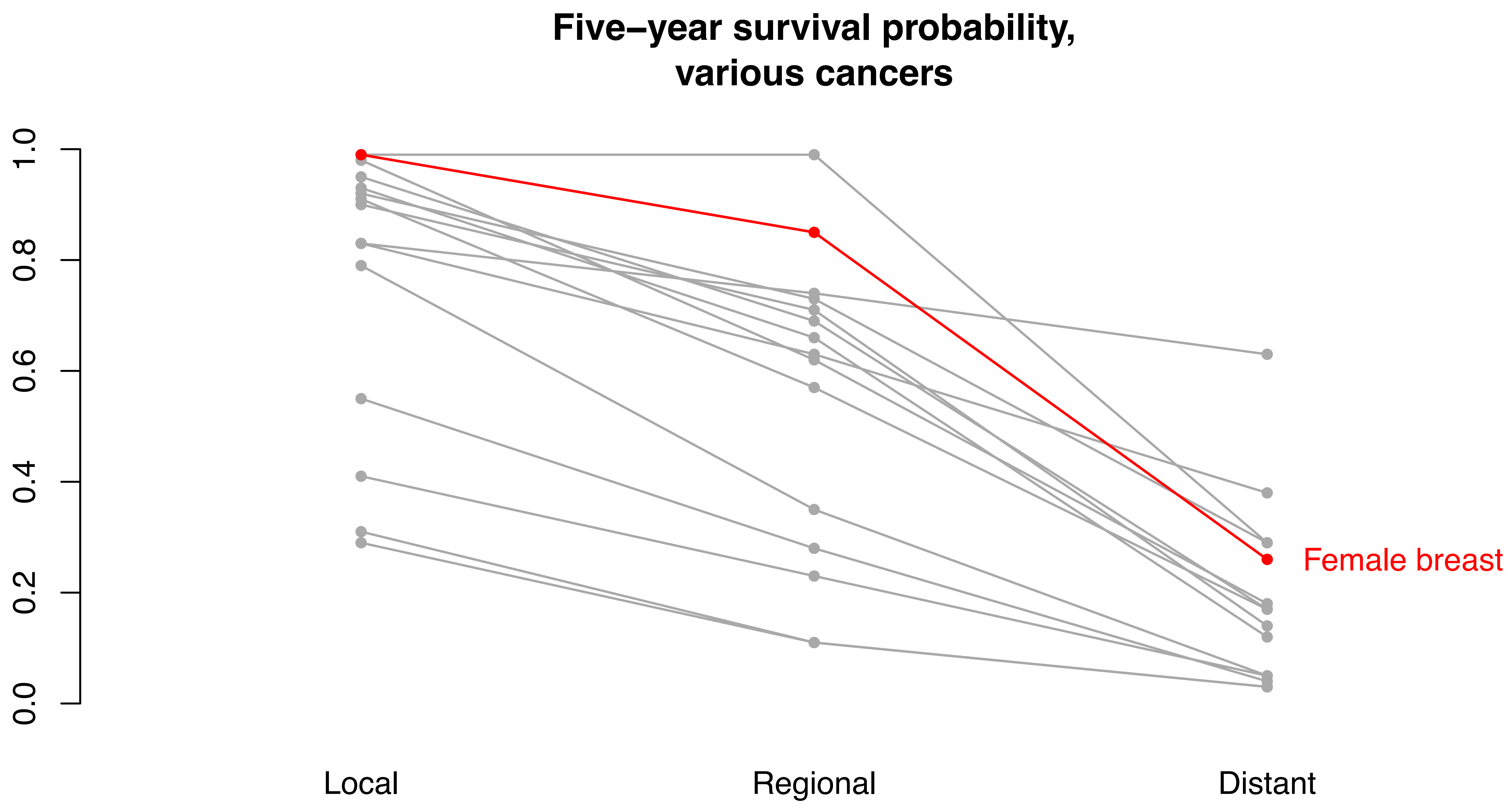
Defense for the degree of Ph. D.
Einar Holsbø
February 8th, 2019

Act I: “Boy Bitten by a Lizard” (1590s)

> Can we predict breast cancer metastasis from blood samples?
>
> –Eiliv Lund, 4.5 years ago, quote made up

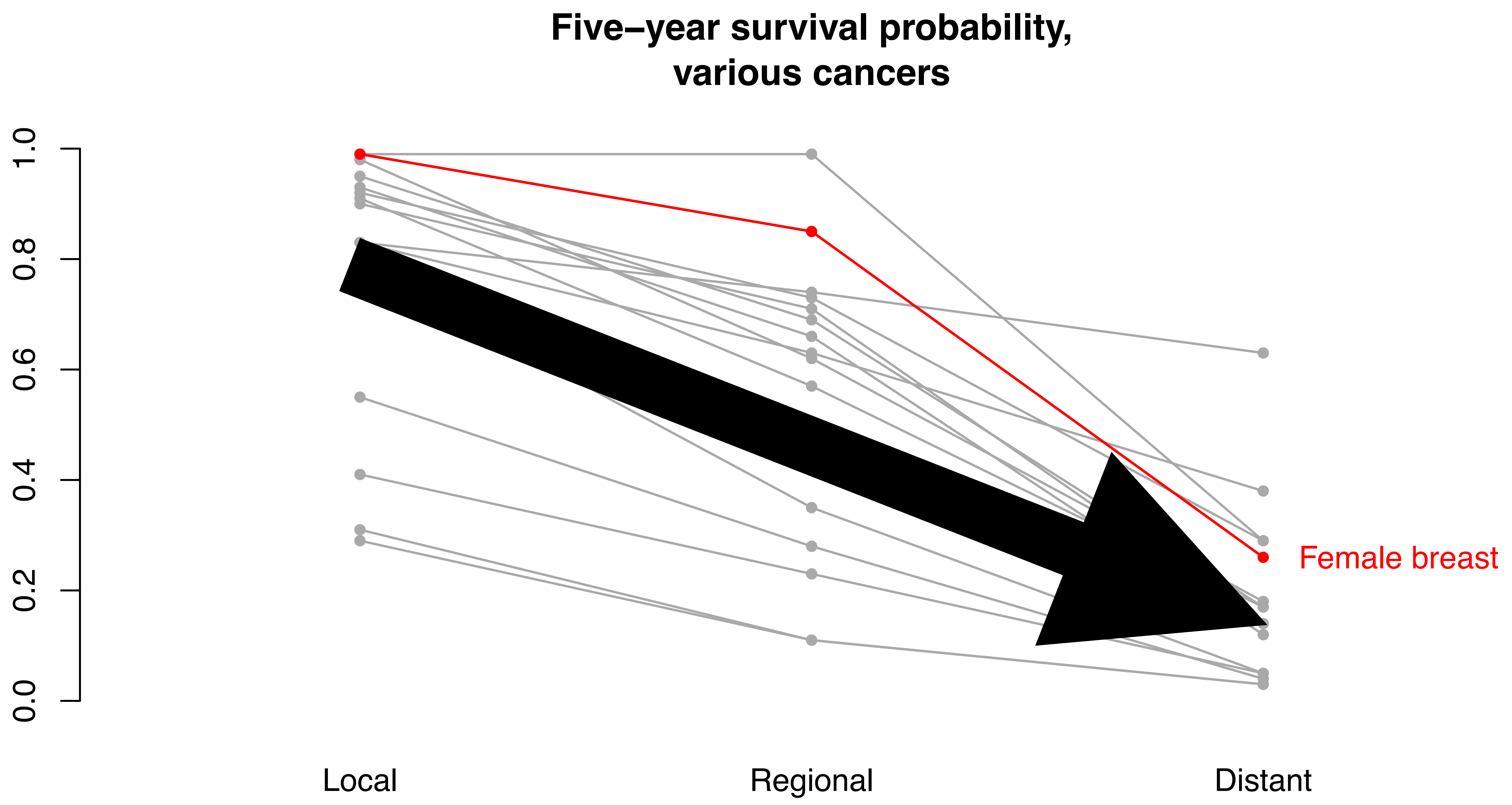
Metastasis is the spread of cancer in the body

Metastasis is the spread of cancer in the body



Data source: Siegel, R. L., Miller, K. D. and Jemal, A. (2017), Cancer statistics, 2017. CA: A Cancer Journal for Clinicians, 67: 7-30. doi:10.3322/caac.21387

Metastasis is the spread of cancer in the body



Data source: Siegel, R. L., Miller, K. D. and Jemal, A. (2017), Cancer statistics, 2017. CA: A Cancer Journal for Clinicians, 67: 7-30. doi:10.3322/caac.21387

Metastasis is the spread of cancer in the body

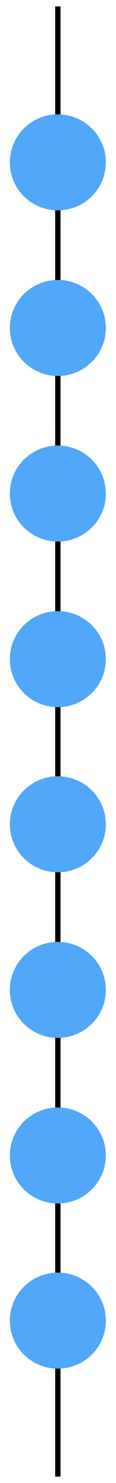
Goal: predict it, win the Nobel prize 🏅

Norwegian Women and Cancer

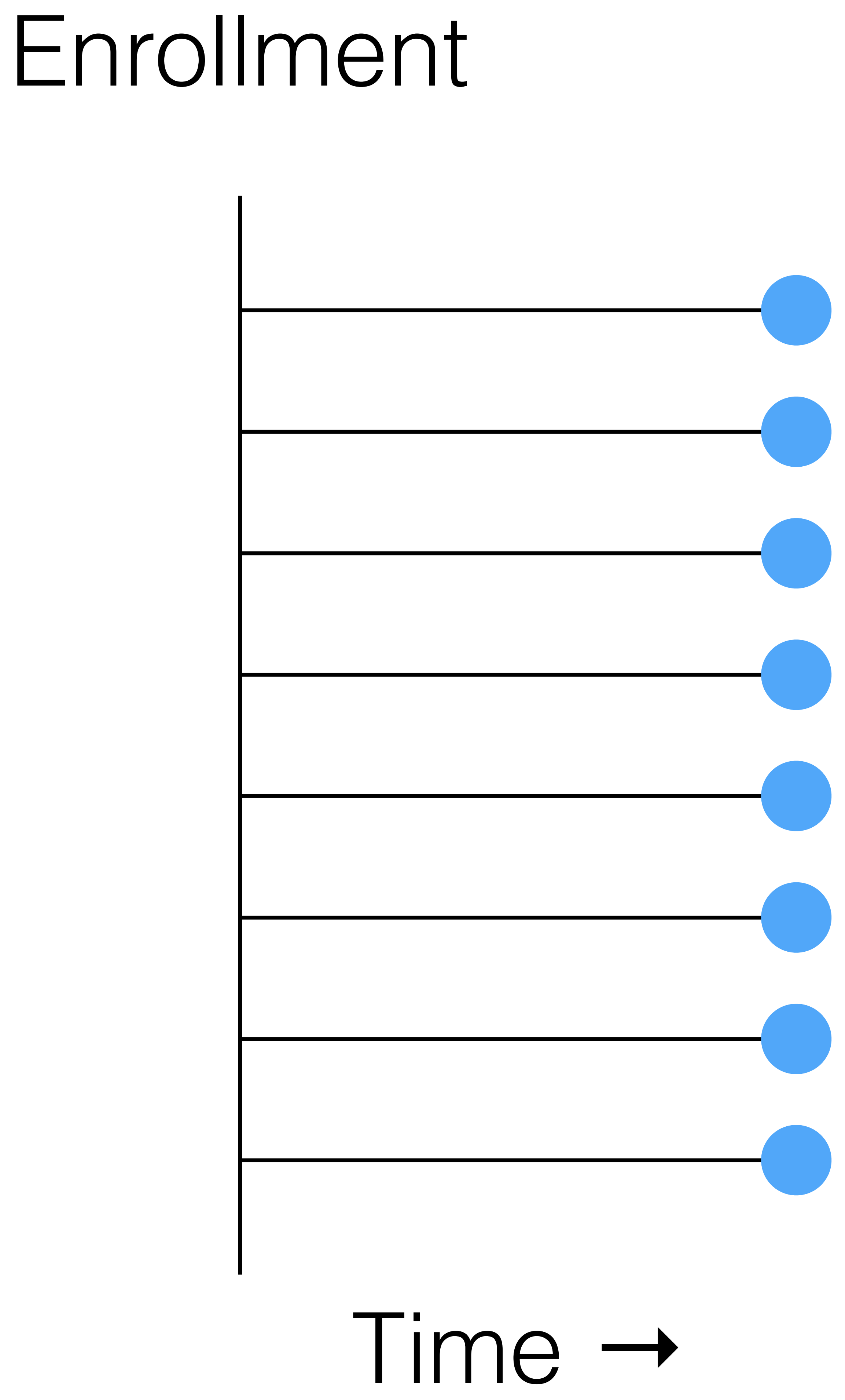
- Prospective population-based cohort that tracks 34% (170 000) of all Norwegian women born between 1943-57.
- The data collection started in NOWAC in 1991. Includes blood samples from 50.000 women, as well as more than 300 biopsies.
- Now contains various -omics material: microarray mRNA, miRNA, methylation, metabolomics, and RNA-seq.

Prospective

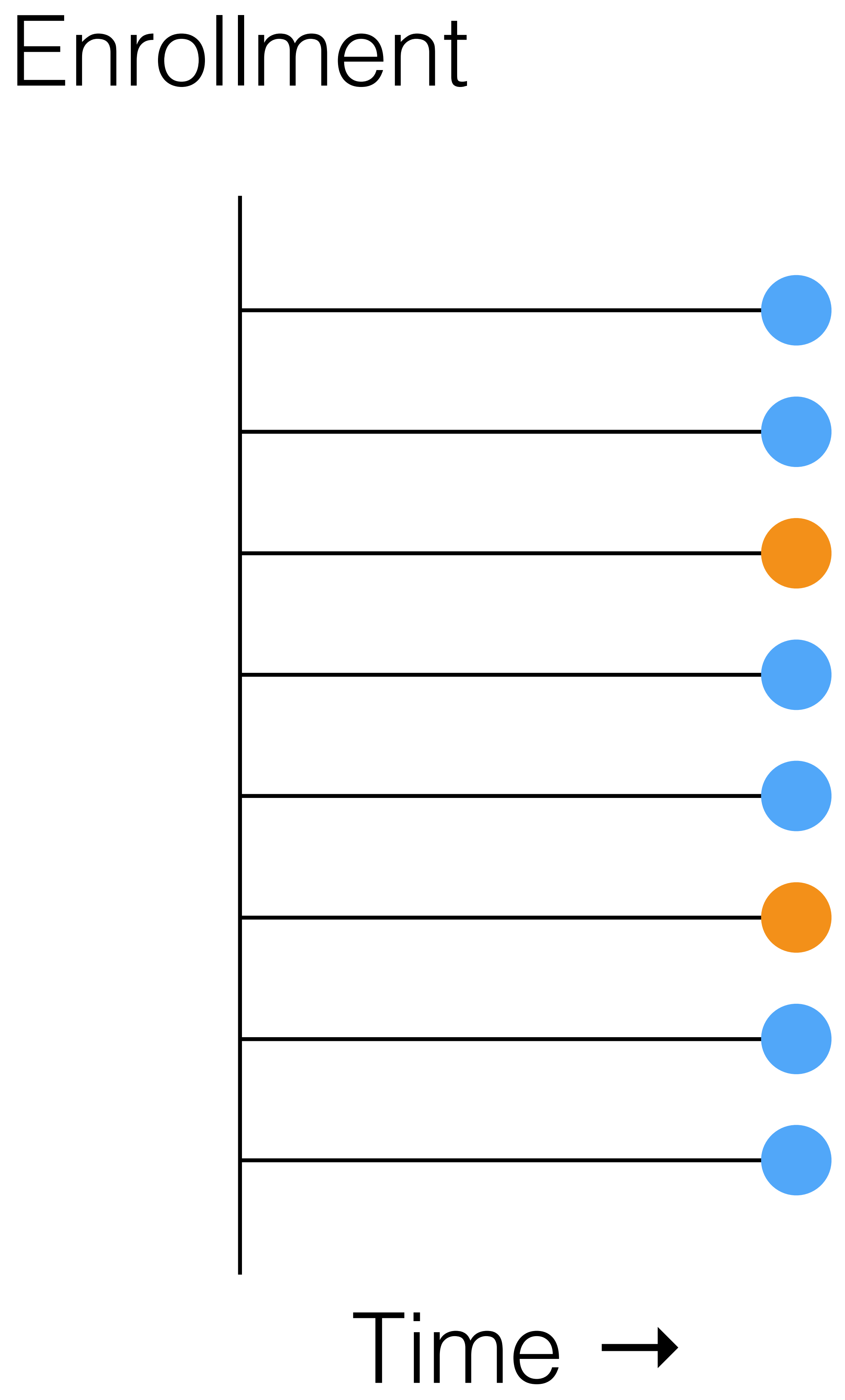
Enrollment



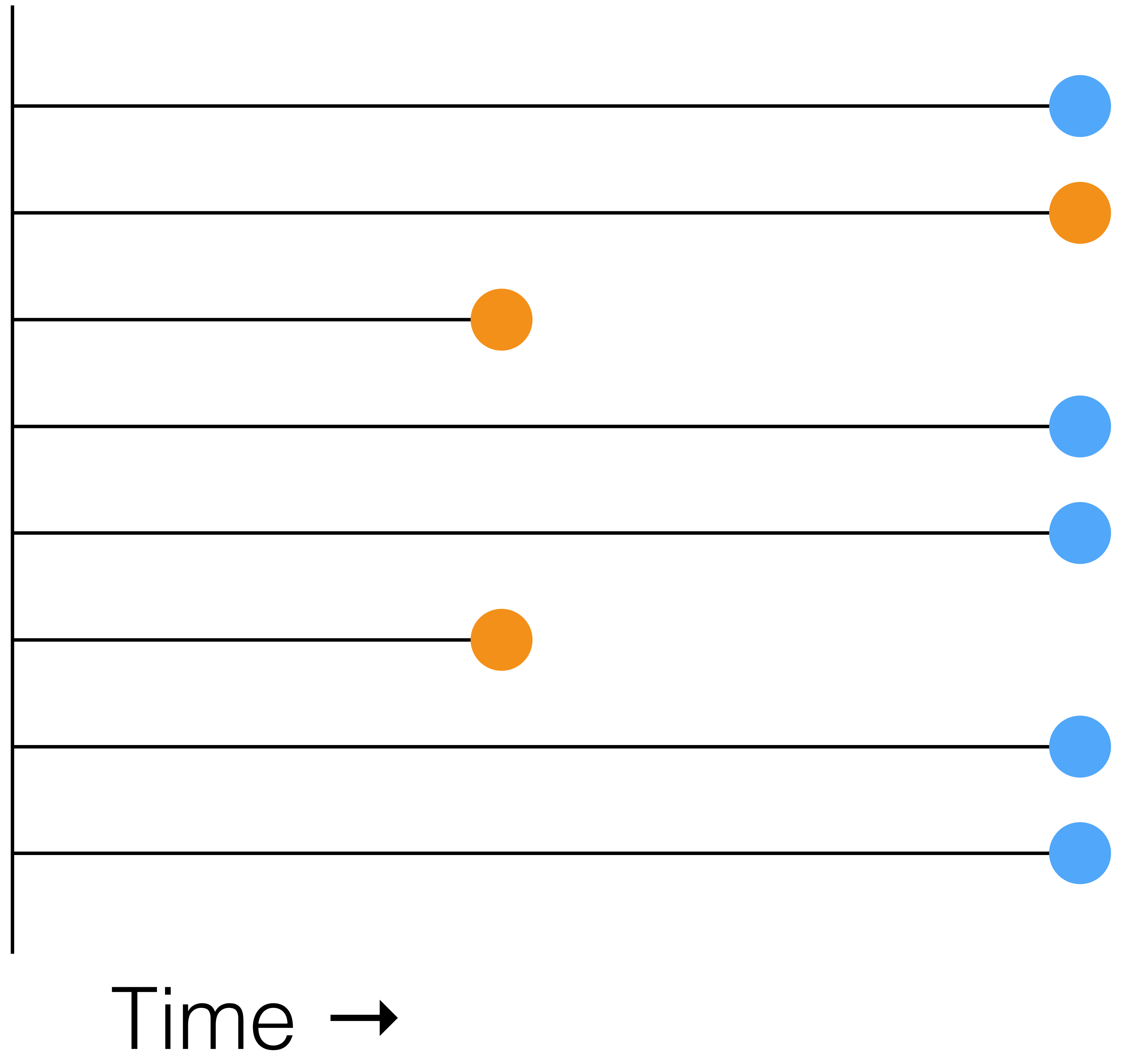
Prospective



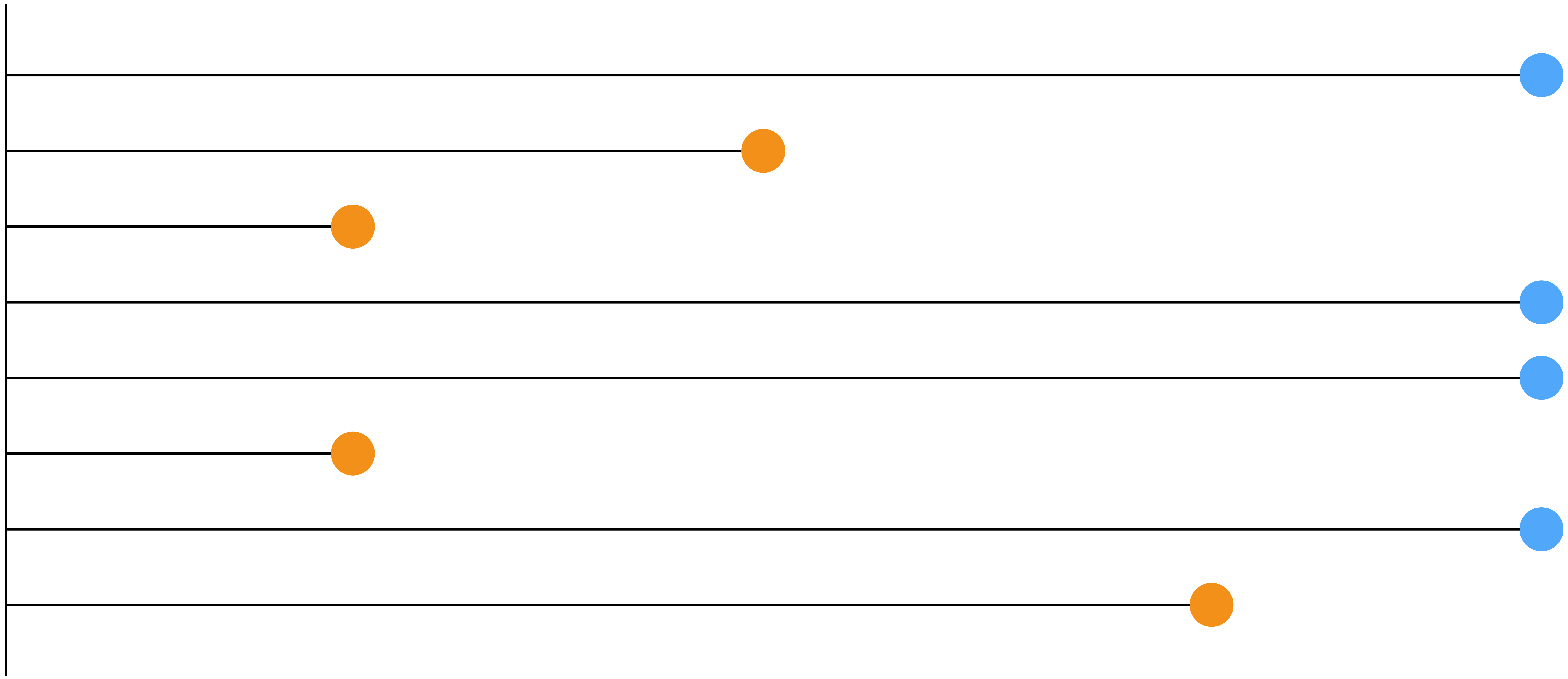
Prospective



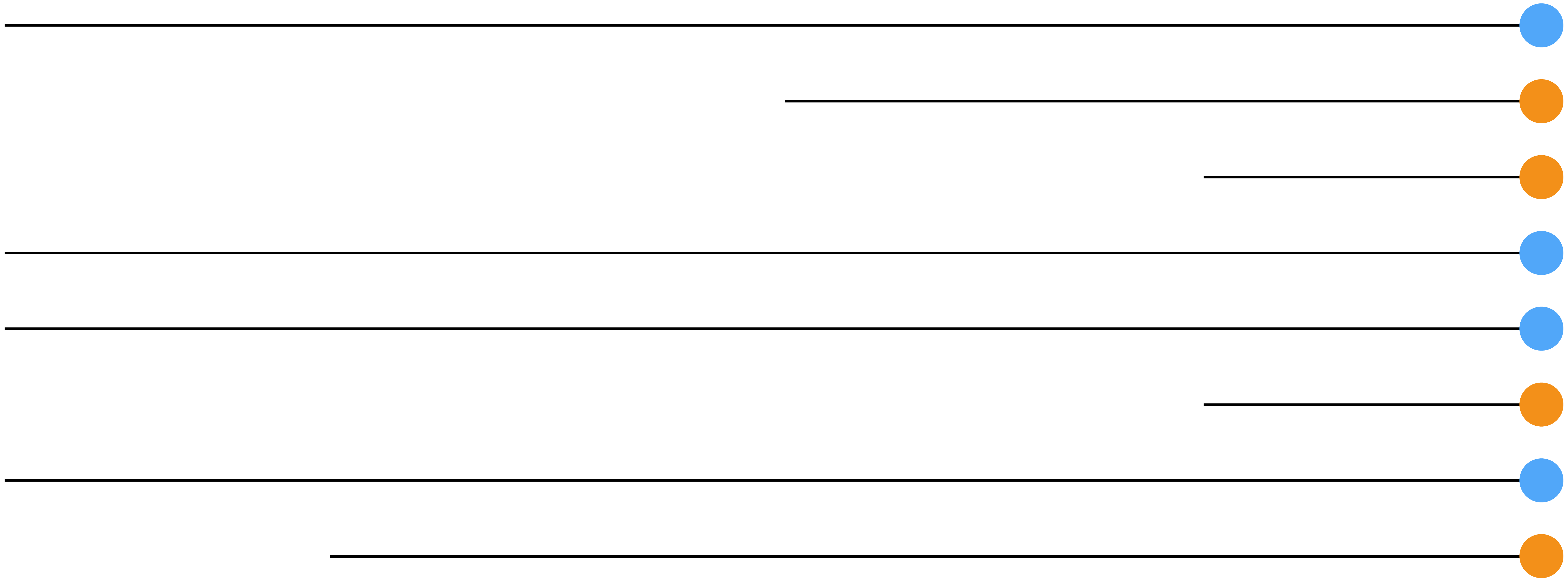
Prospective



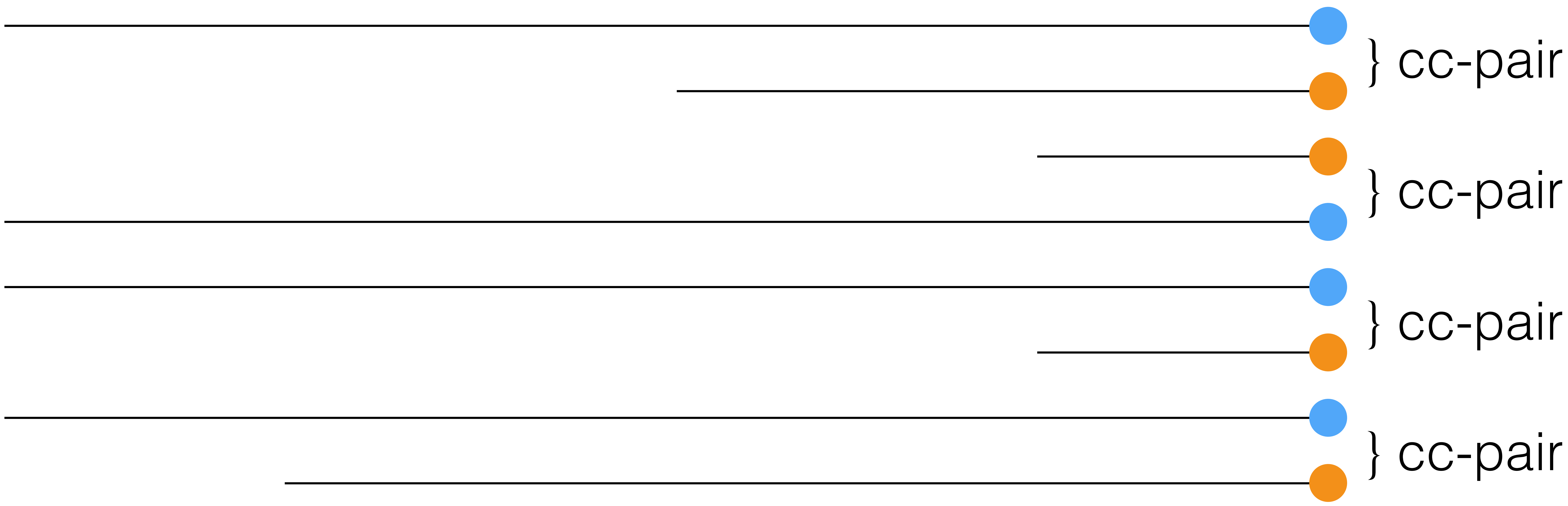
Prospective



Prospective



Nested case–control



Prospective design nice because recruitment is blinded to outcome and exposure

Prospective design nice because recruitment is blinded to outcome and exposure

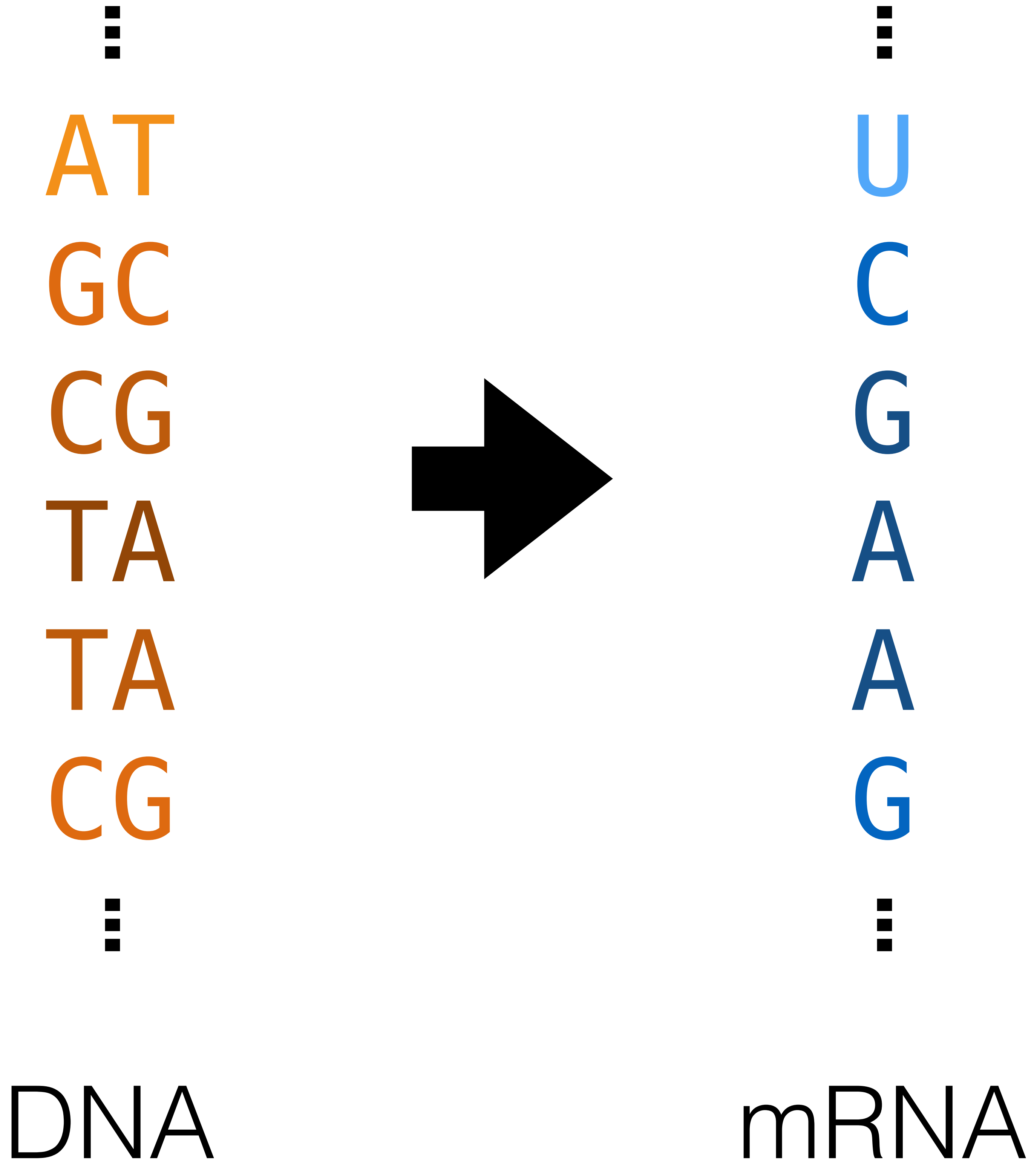
Low bias

Gene expression

⋮
AT
GC
CG
TA
TA
CG
⋮

DNA

Gene expression



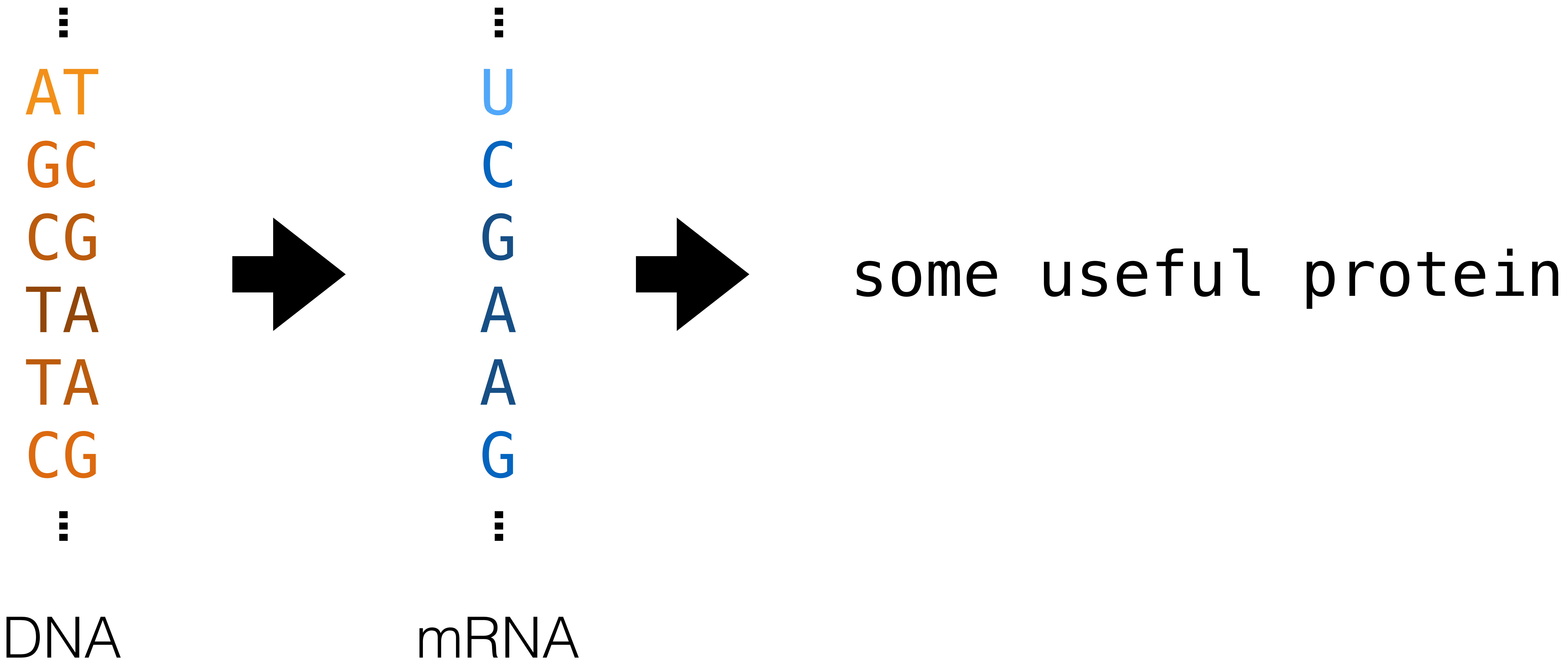
⋮
AT
GC
CG
TA
TA
CG
⋮

DNA

⋮
U
C
G
A
A
G
⋮

mRNA

Gene expression



⋮
AT
GC
CG
TA
TA
CG
⋮

DNA

⋮
U
C
G
A
A
G
⋮

mRNA

some useful protein

Gene expression

U
C
G
A
A
G

Gene expression

💡

U
C
G
A
A
G

Gene expression

💡
U
C
G
A
A
G

How much light
do we see?

Data at a glance

Data at a glance

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

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
```

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
```

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
```

These are “small data” & we should be careful with them

A computer scientist’s guide to precision medicine

- Step 1: pick some models
- Step 2: pick some scoring rules/performance metrics
- Step 3: “classification”

Scoring rule examples
(aka. loss functions, aka. metrics)

- Accuracy: how many did we get right?
- Precision: how many correct “success” predictions did we do
- Recall: how many of the true successes did we detect

Scoring rule examples
(aka. loss functions, aka. metrics)

$p > .5$? something else?

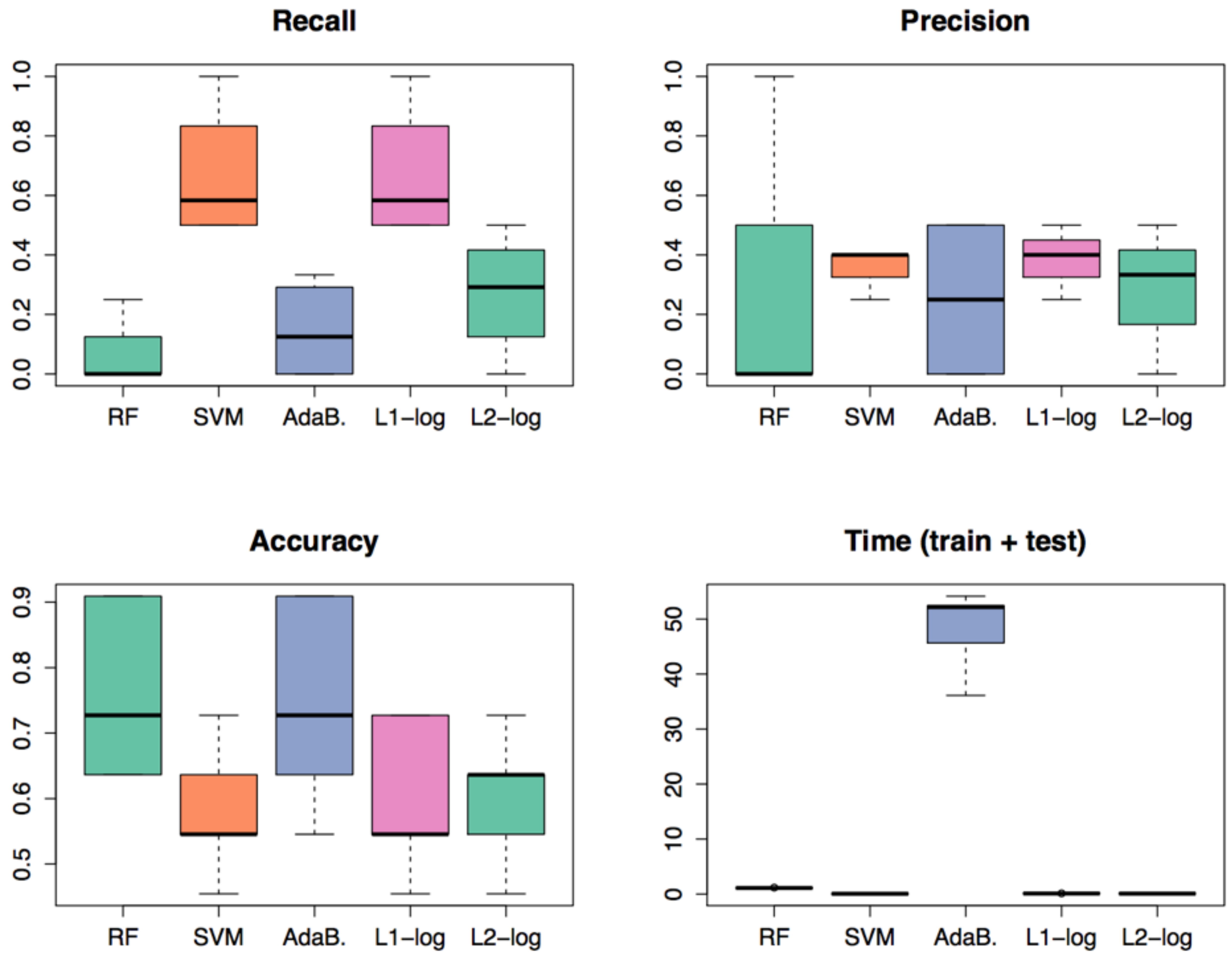
Decoupling score and decision threshold

- AUC: the probability of ranking success higher than failure

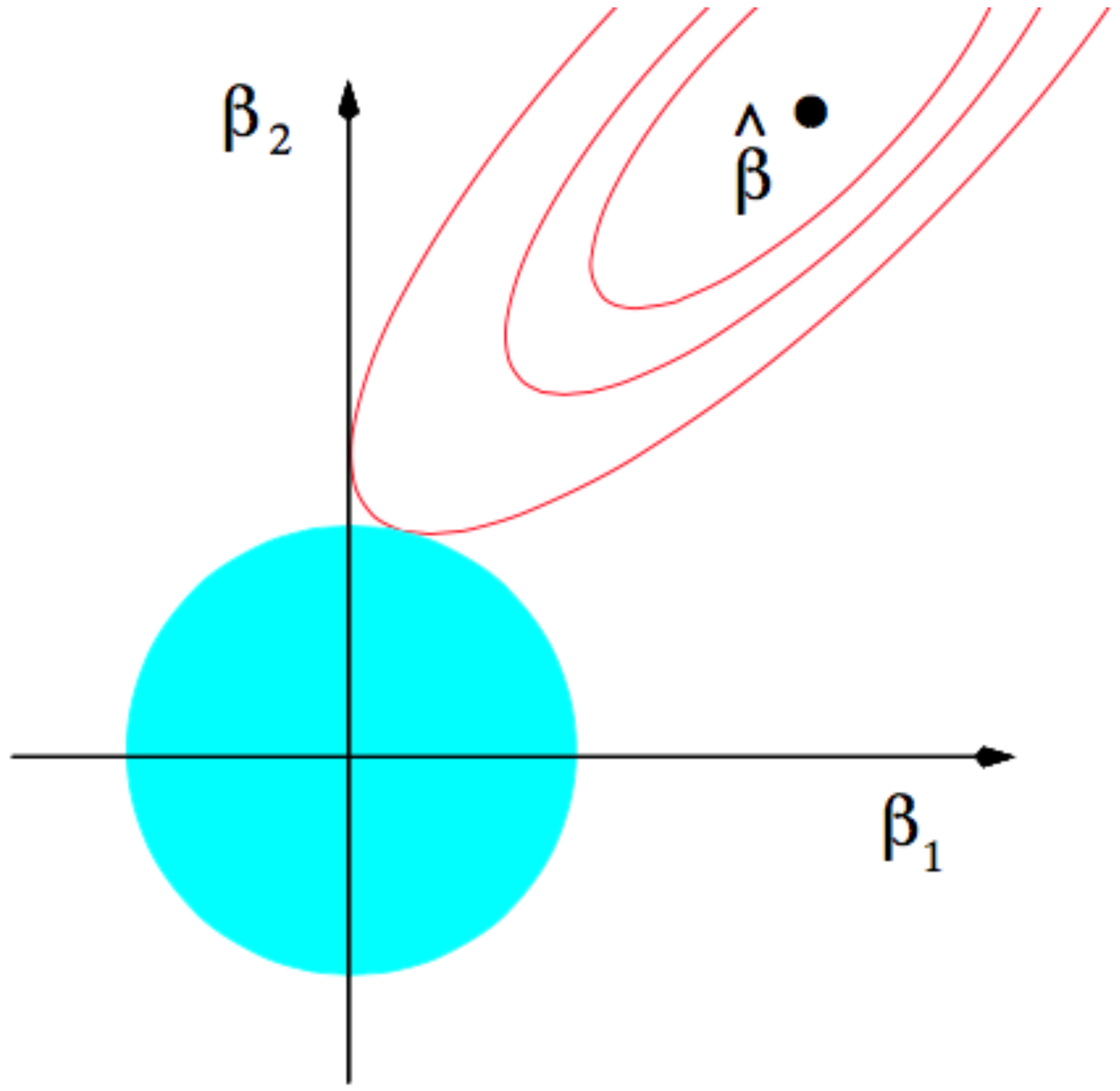
(aka. concordance probability)

Just trying some methods & scores

Just trying some methods & scores

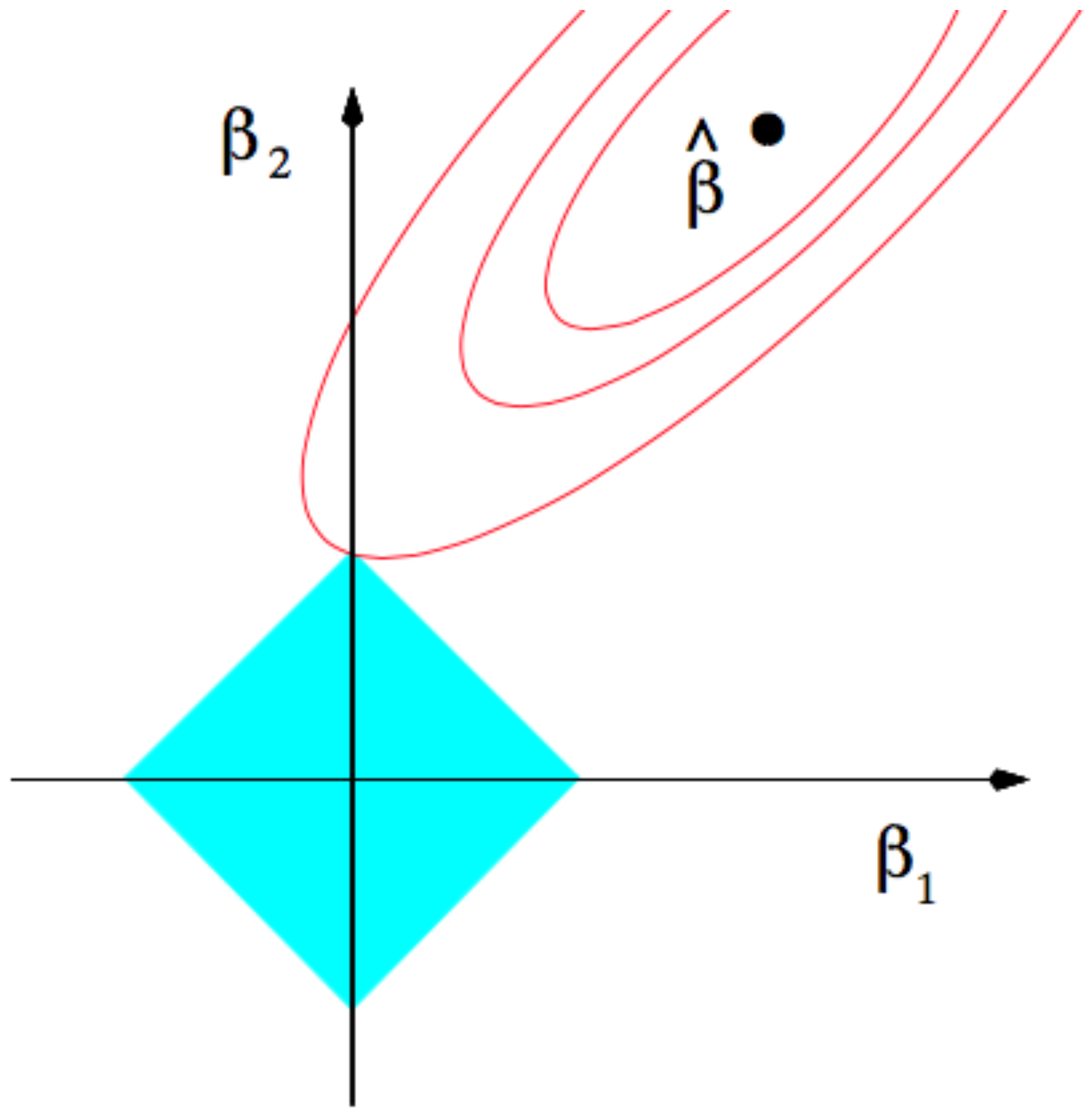


$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$



$$\sum \beta_i^2 \leq t$$

“ridge”

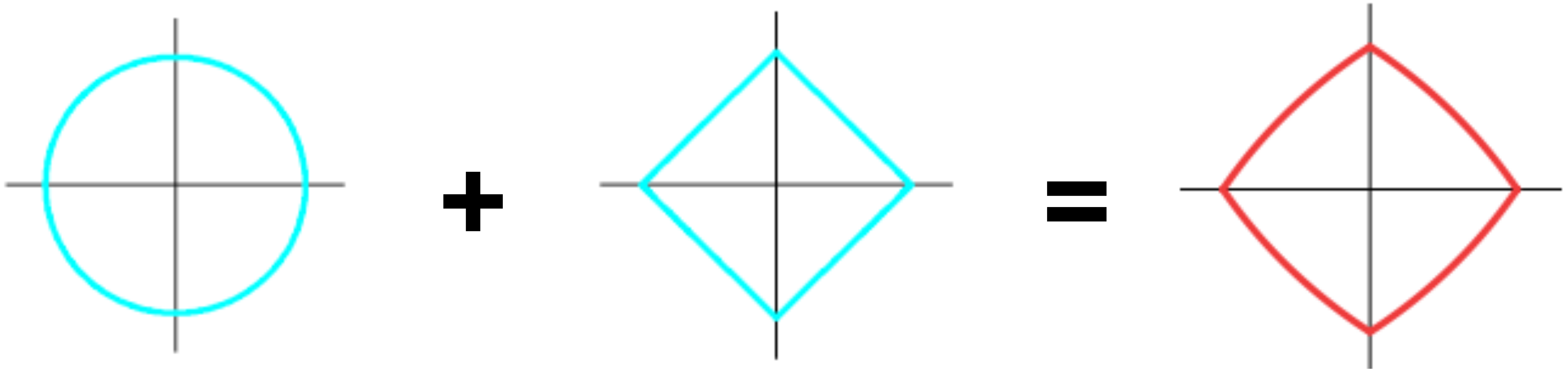


$$\sum |\beta_i| \leq t$$

“lasso”

Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

$$\sum \left[\alpha\beta_i^2 + (1-\alpha)|\beta_i|\right] \leq t$$

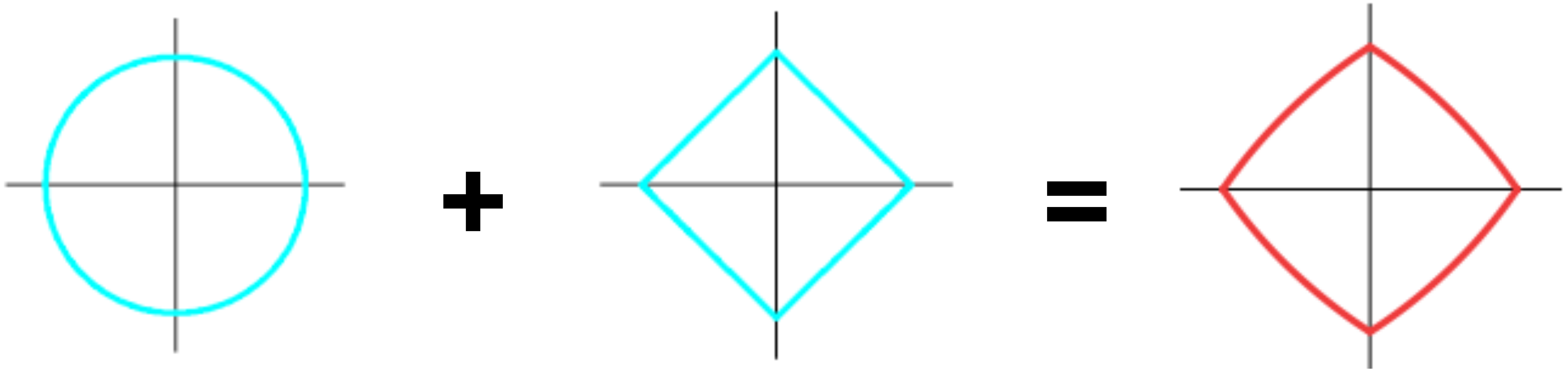


“ElasticNet”

Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

Tradeoff between penalty types, controls “roundness”

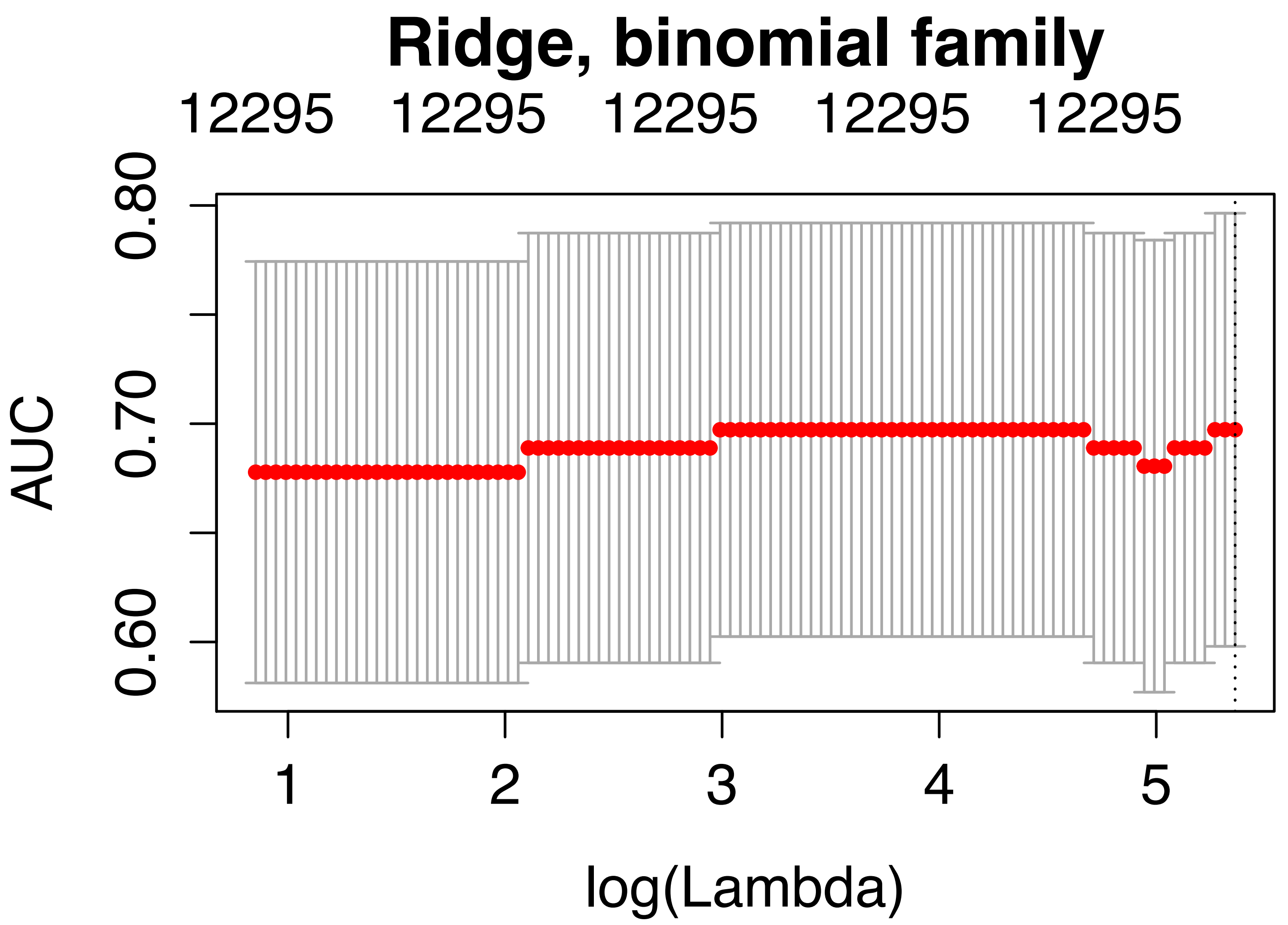
$$\sum \left[\alpha\beta_i^2 + (1-\alpha)|\beta_i|\right] \leq t$$



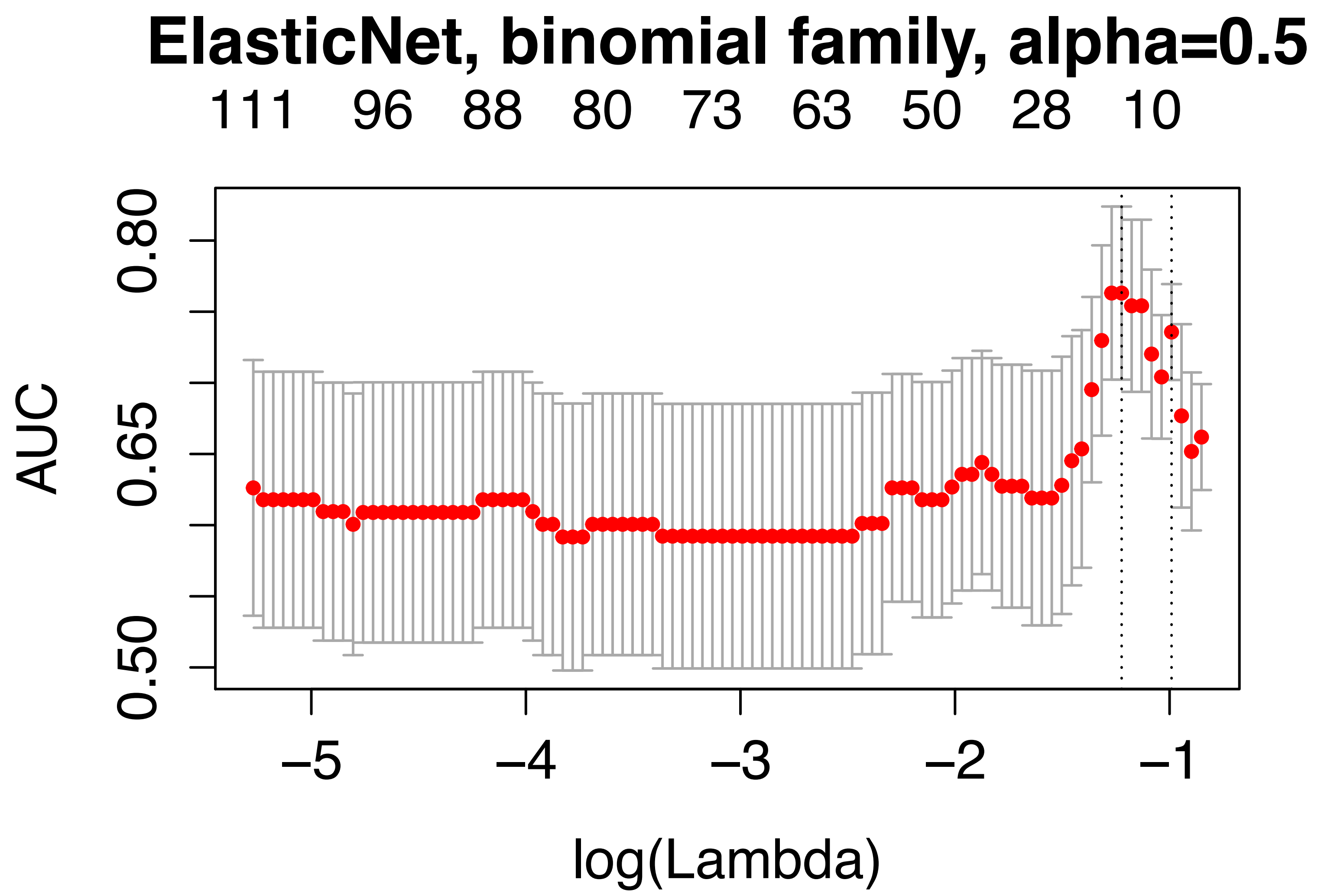
“ElasticNet”

Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

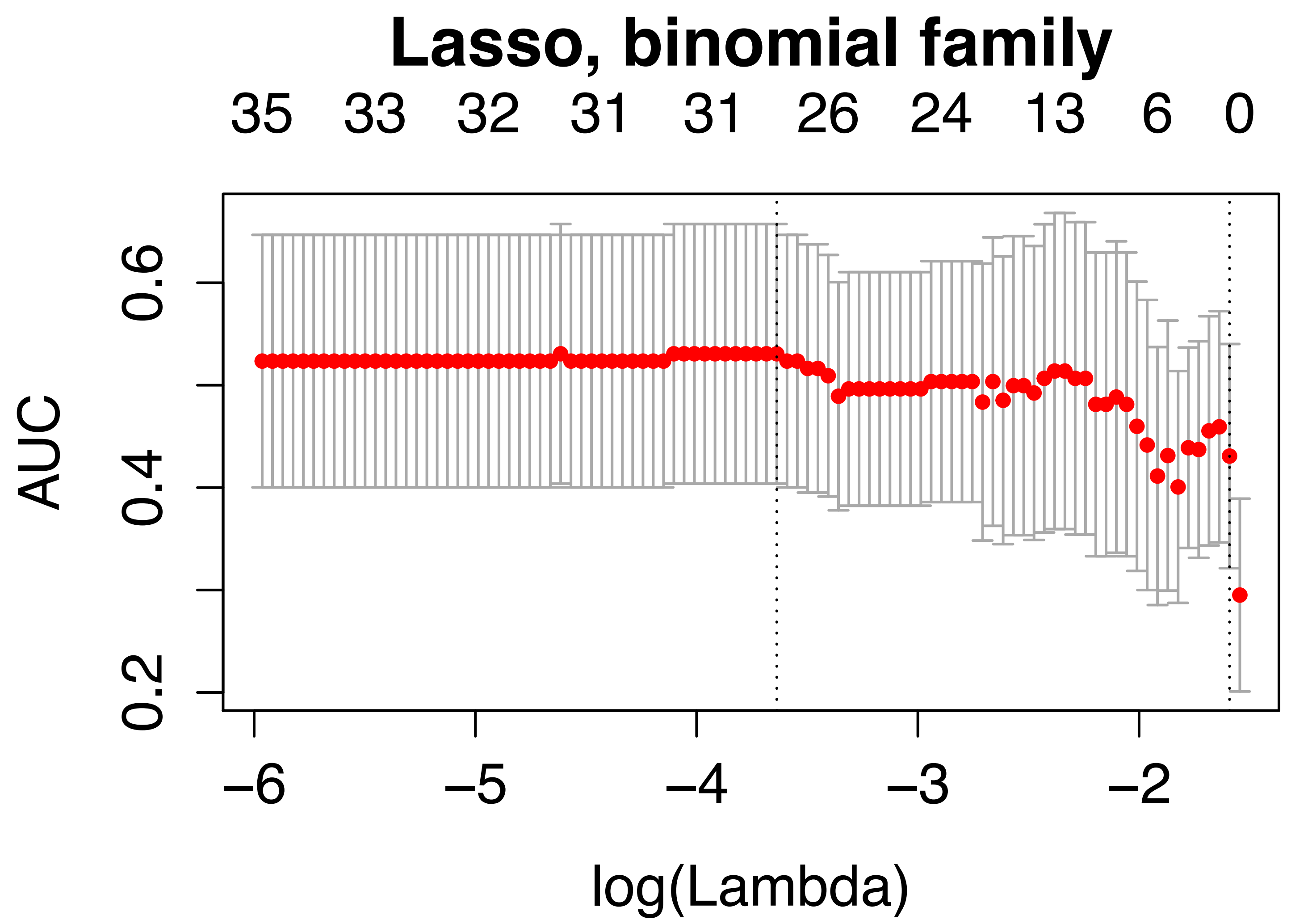
Trying different alphas



Alpha = 1



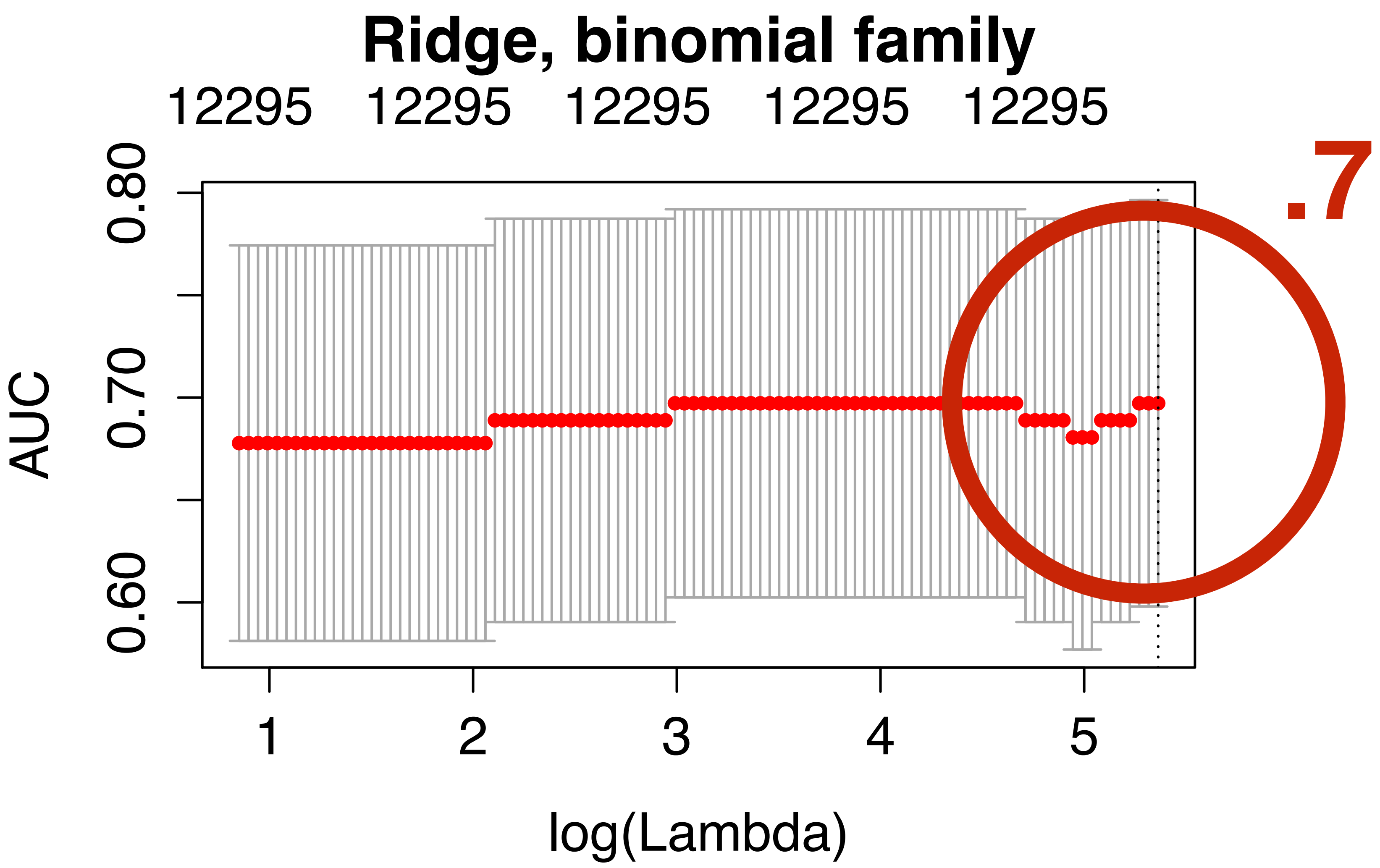
Alpha = .5



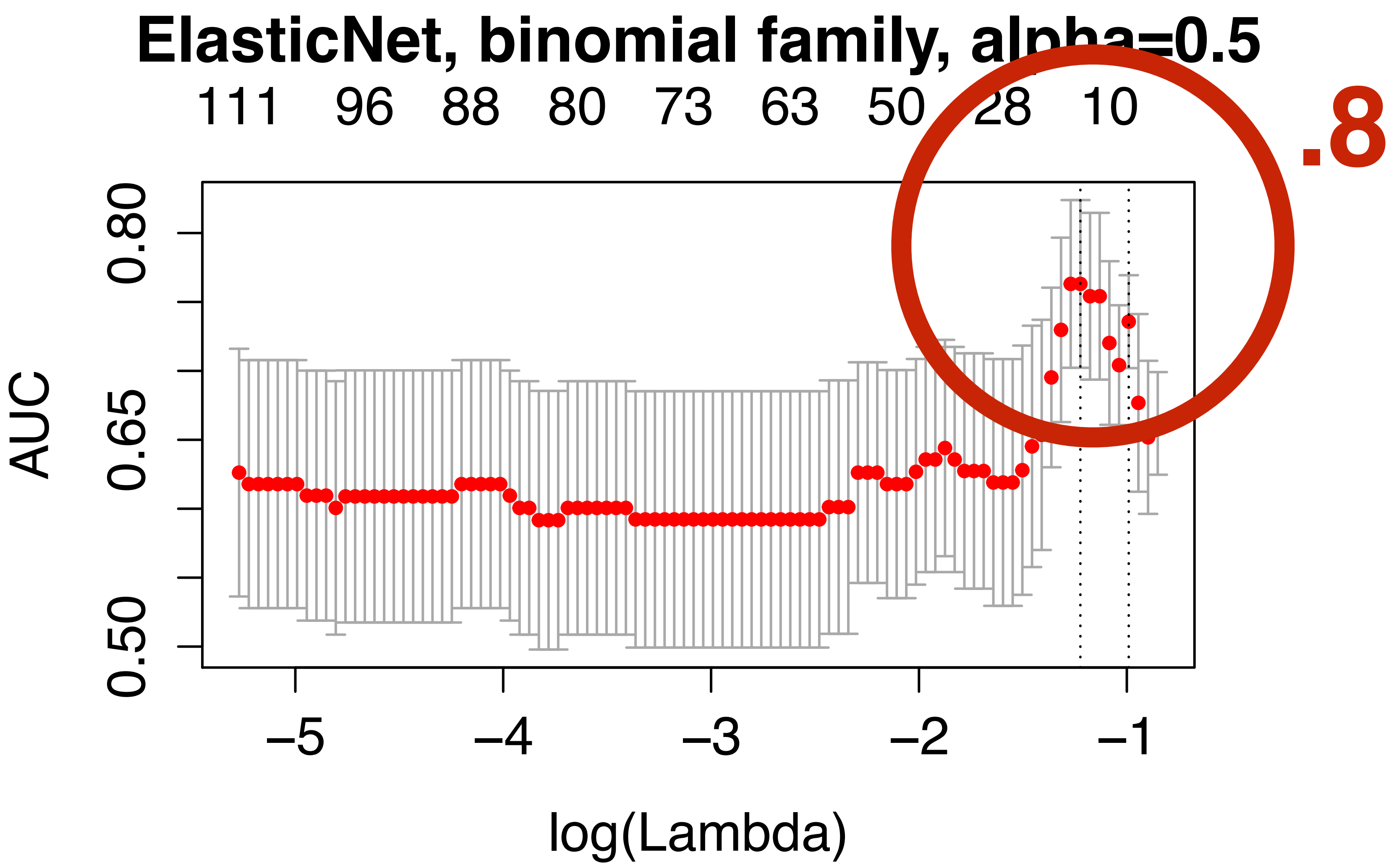
Alpha = 0

Figures show concordance (higher is better)

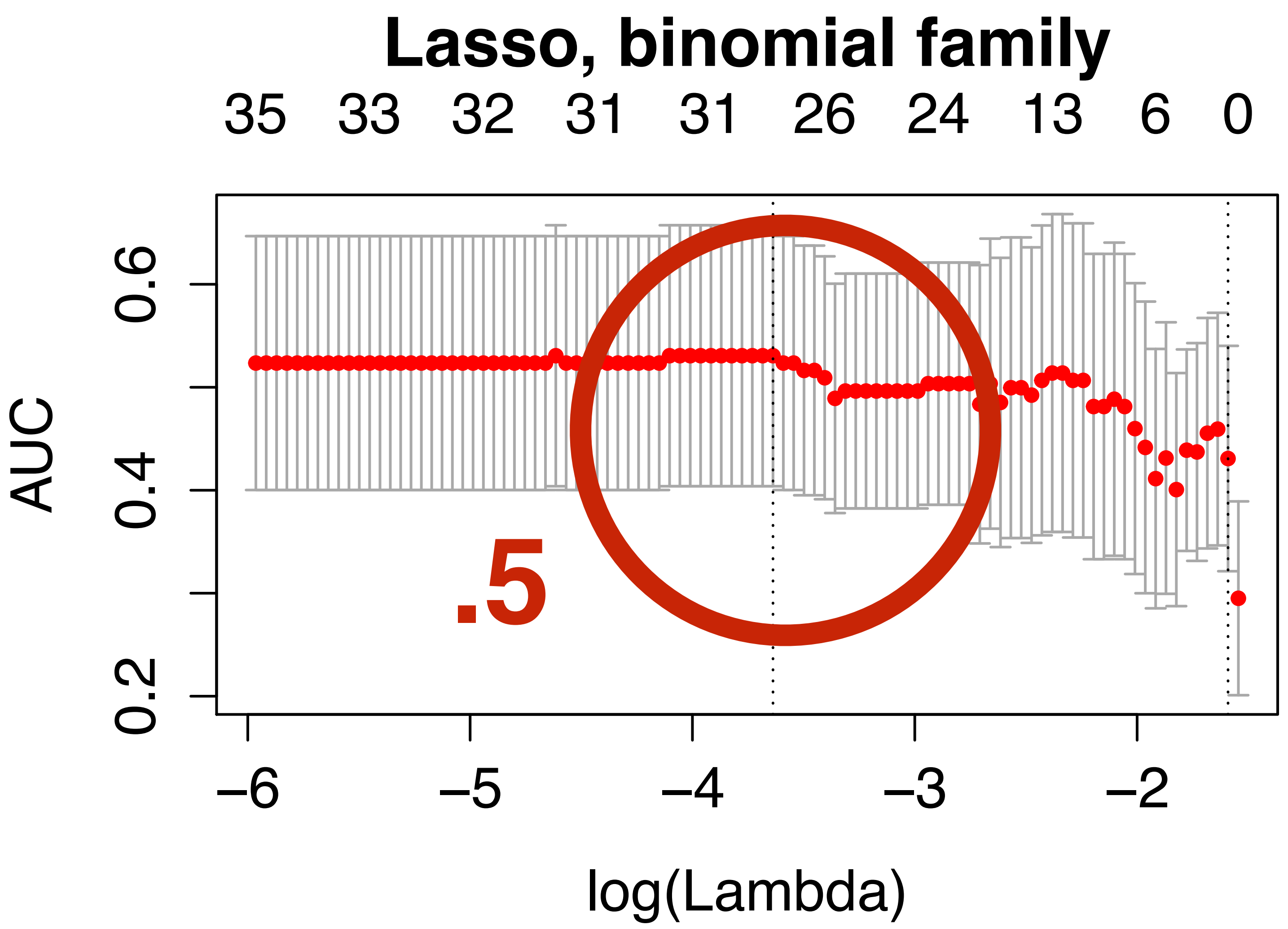
Trying different alphas



Alpha = 1



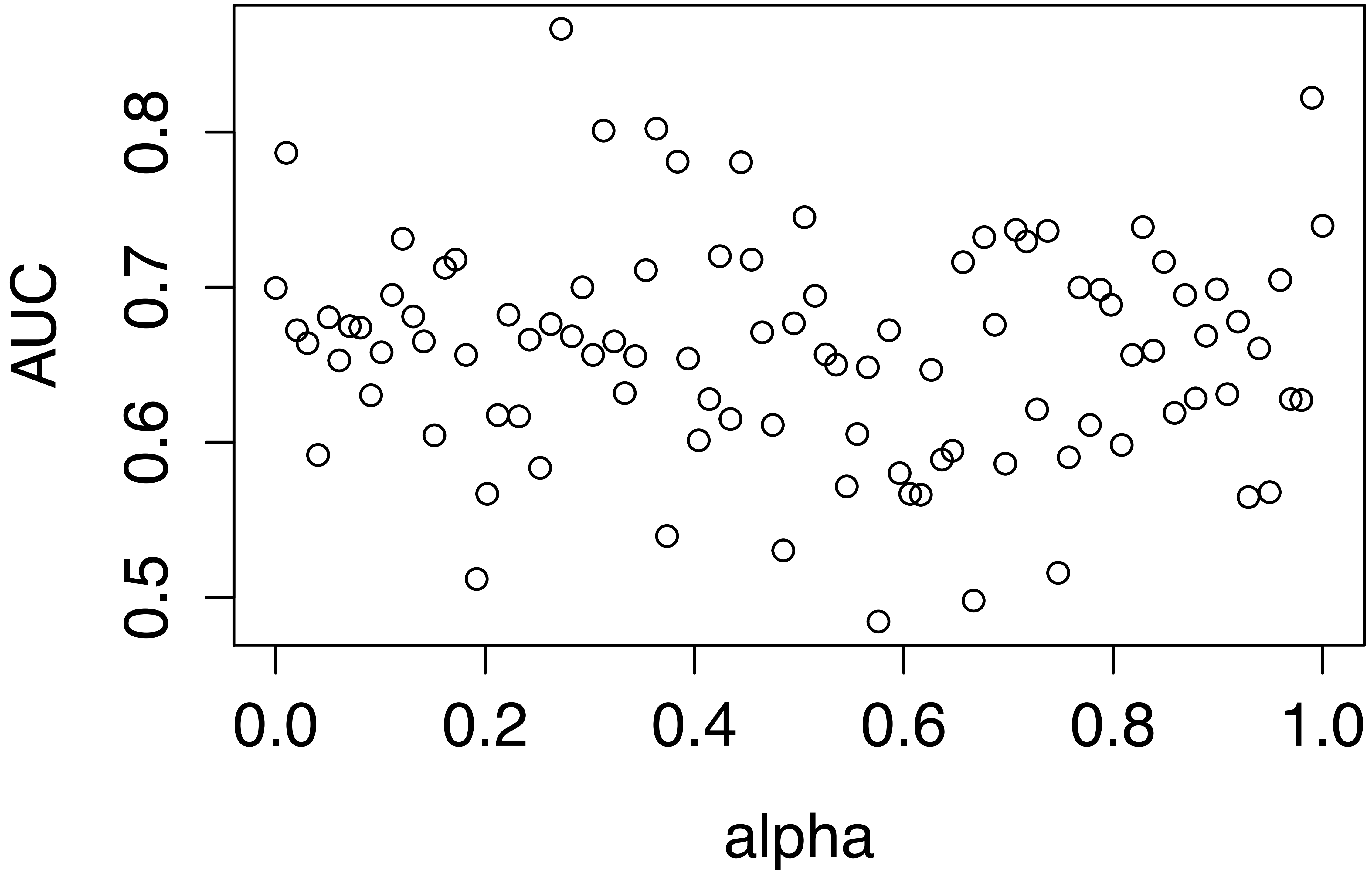
Alpha = .5

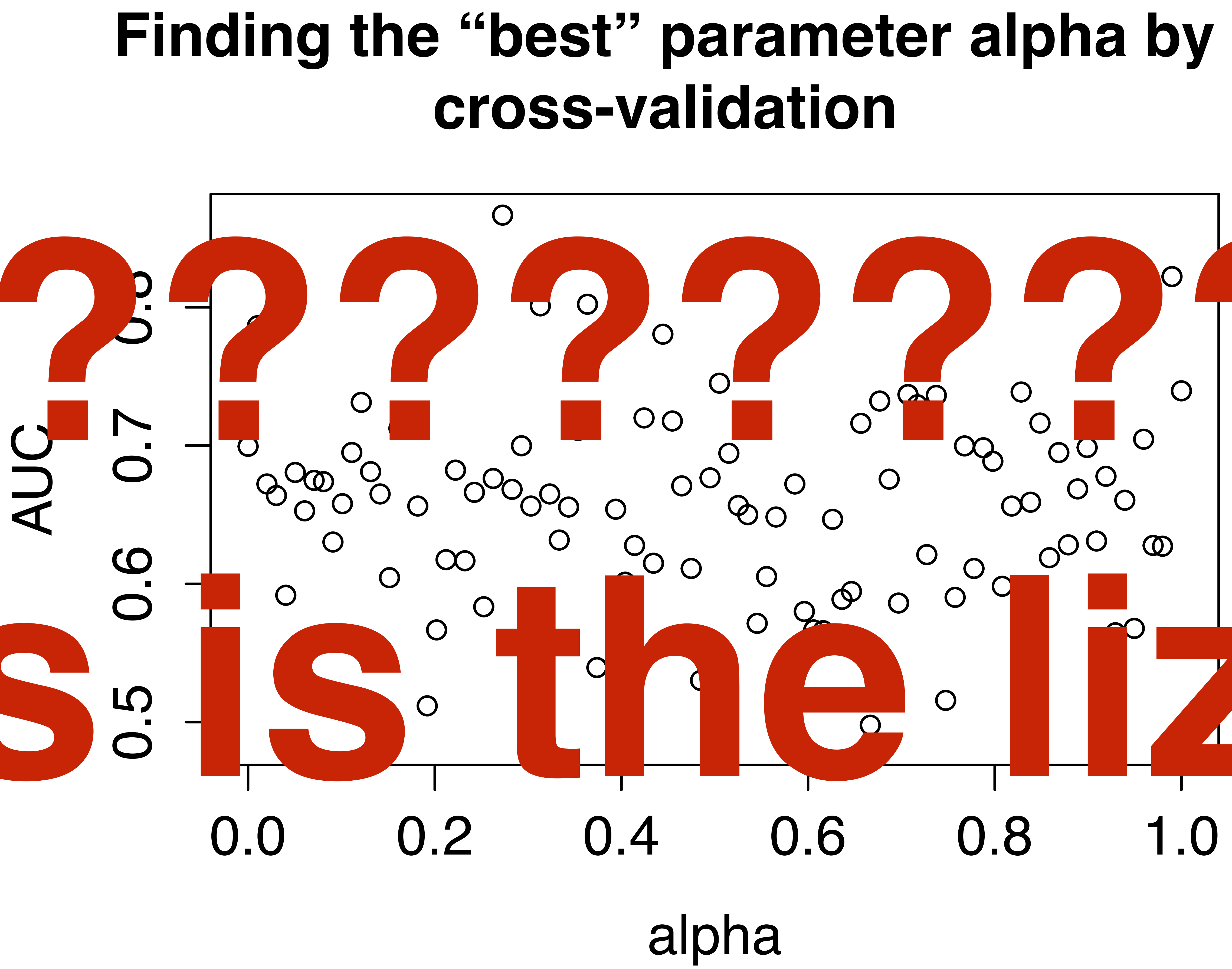


Alpha = 0

Figures show concordance (higher is better)

Finding the “best” parameter alpha by cross-validation





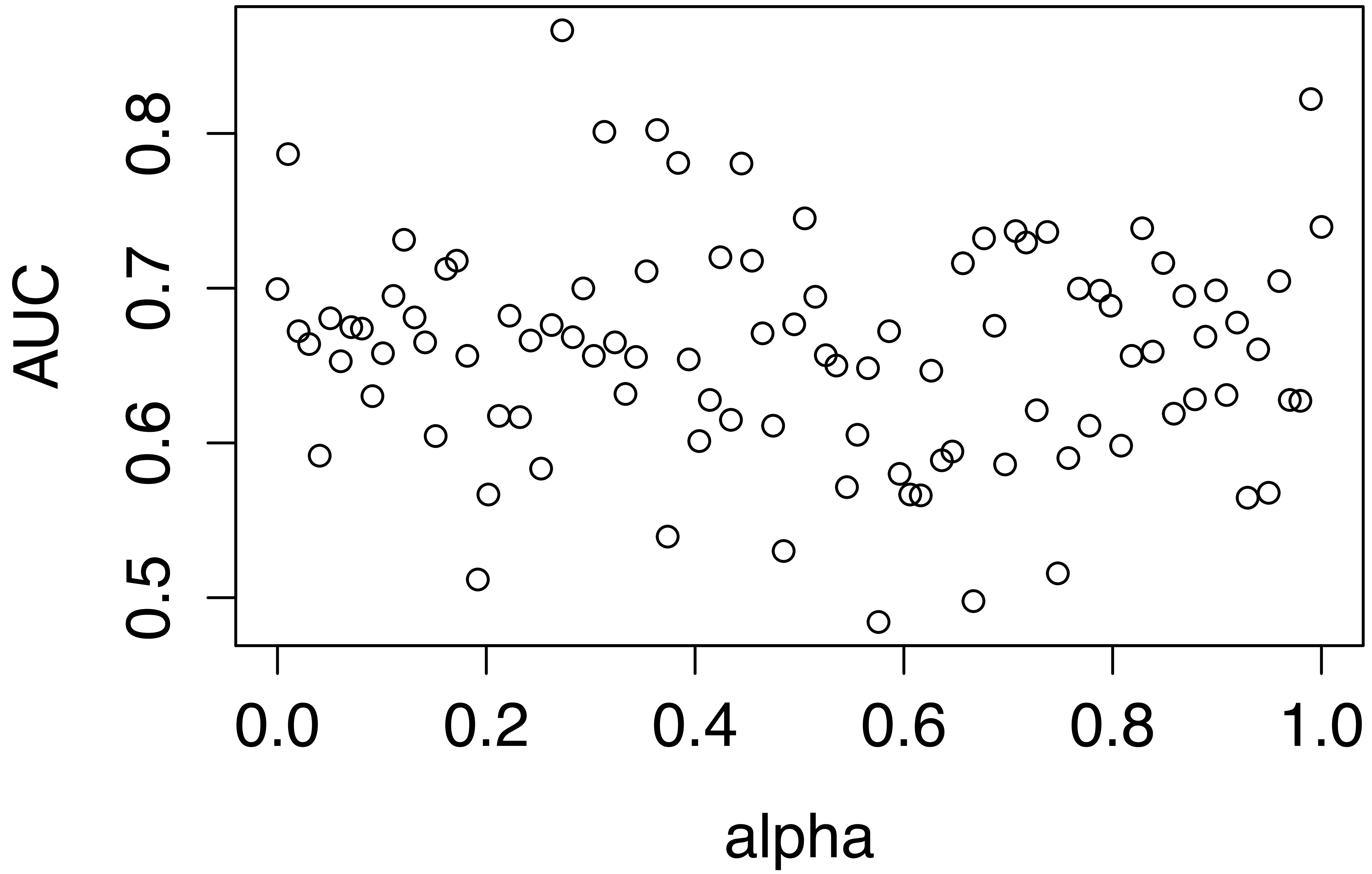
????????????

(this is the lizard)

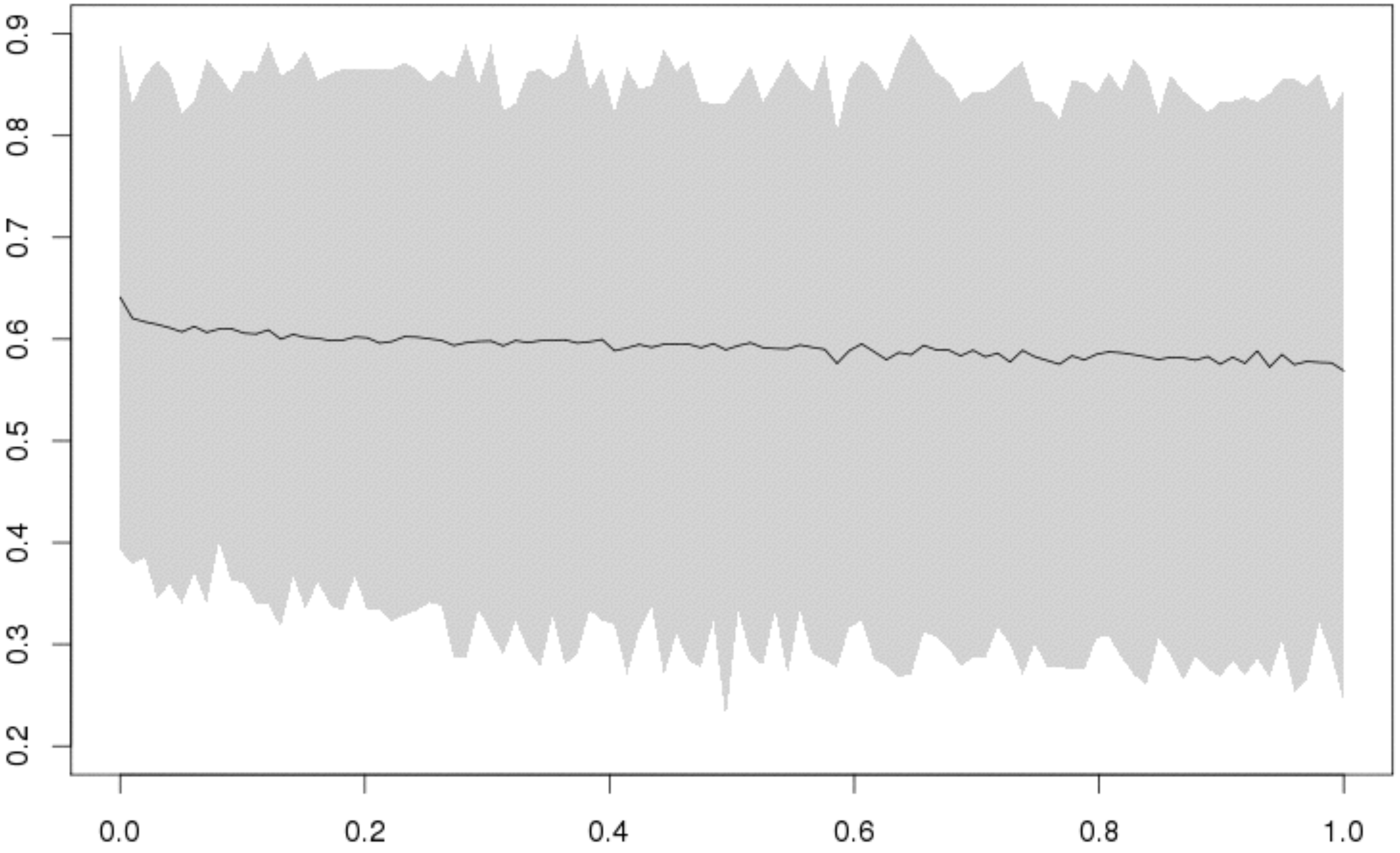
❧ intermission ☙

Act II: When you are engulfed in flames

Finding the “best” parameter alpha by cross-validation



AUC



alpha

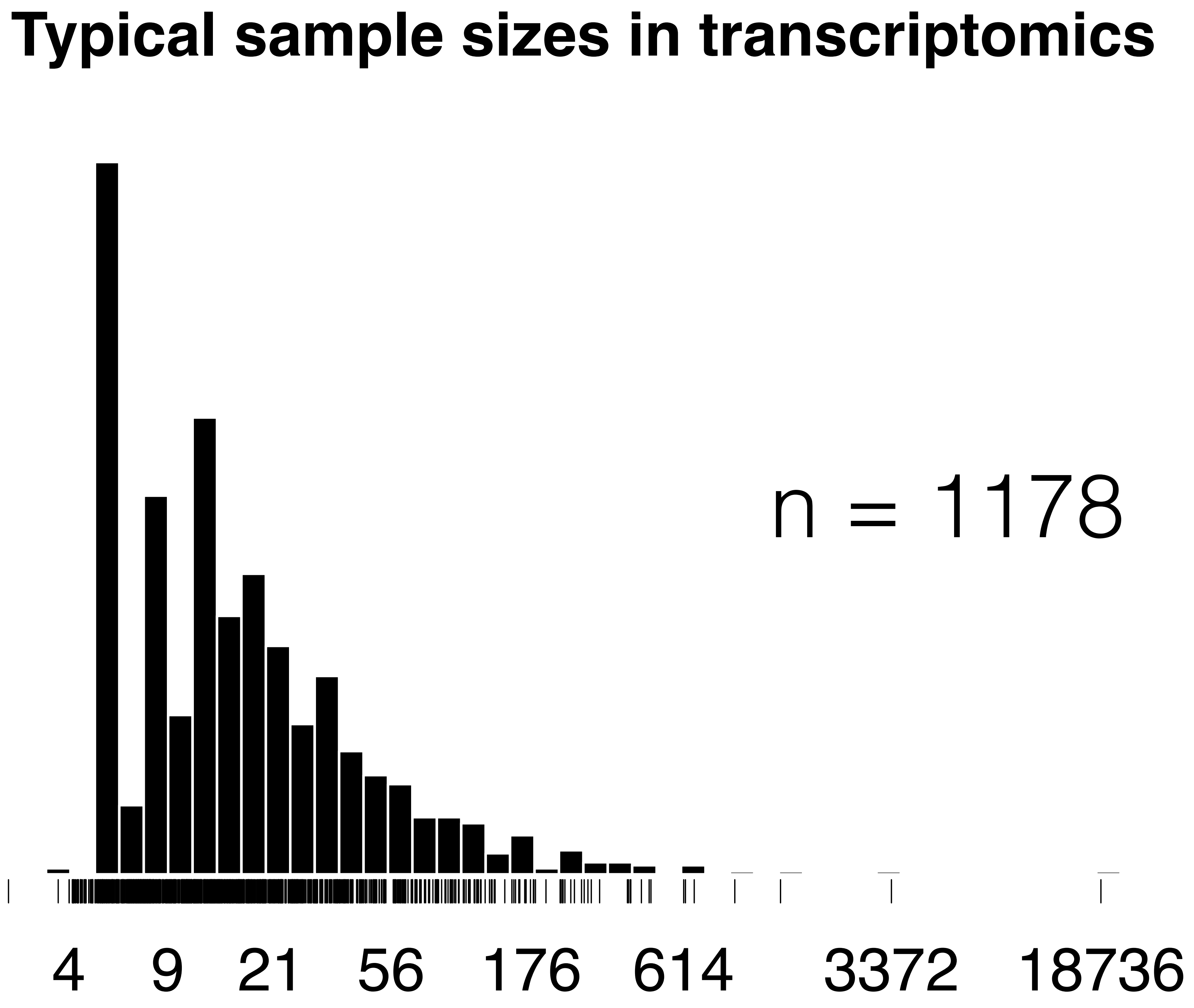
Some “technical” sources of variation

- The big classic one: sample size
- Scoring rule
- Validation procedure

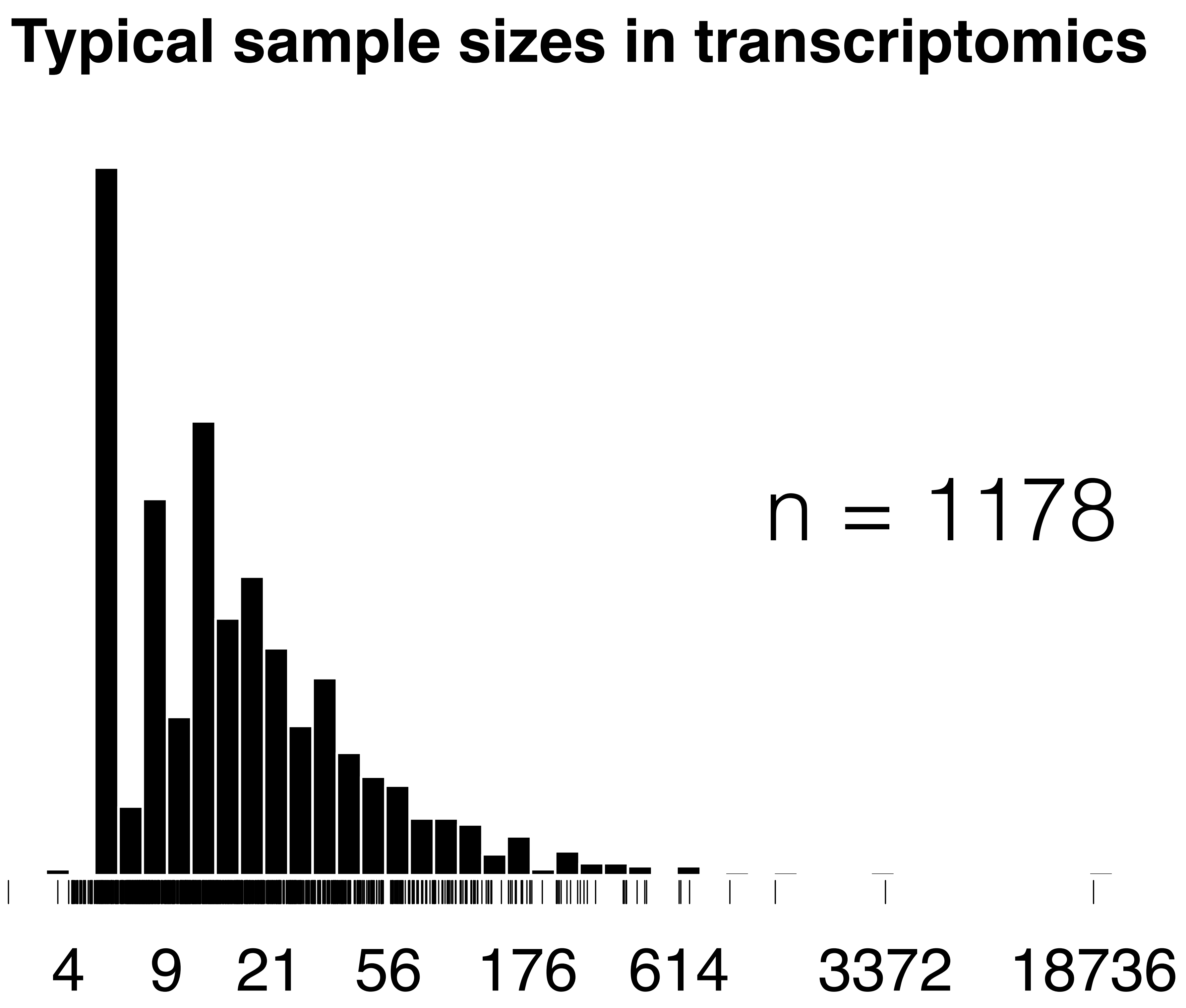
Some “technical” sources of variation

- The big classic one: sample size
- Scoring rule
- Validation procedure

Small data: sample size is more or less fixed in the human model



Small data: sample size is more or less fixed in the human model



Ethics, economy, logistics limit access to human obs.

Some “technical” sources of variation

- The big classic one: sample size
- Scoring rule
- Validation procedure

Yet another scoring rule

Brier's score is the mean squared errors of predicted probabilities

$$n^{-1} \sum (\hat{p}_i - p_i)^2$$

Some risk surfaces

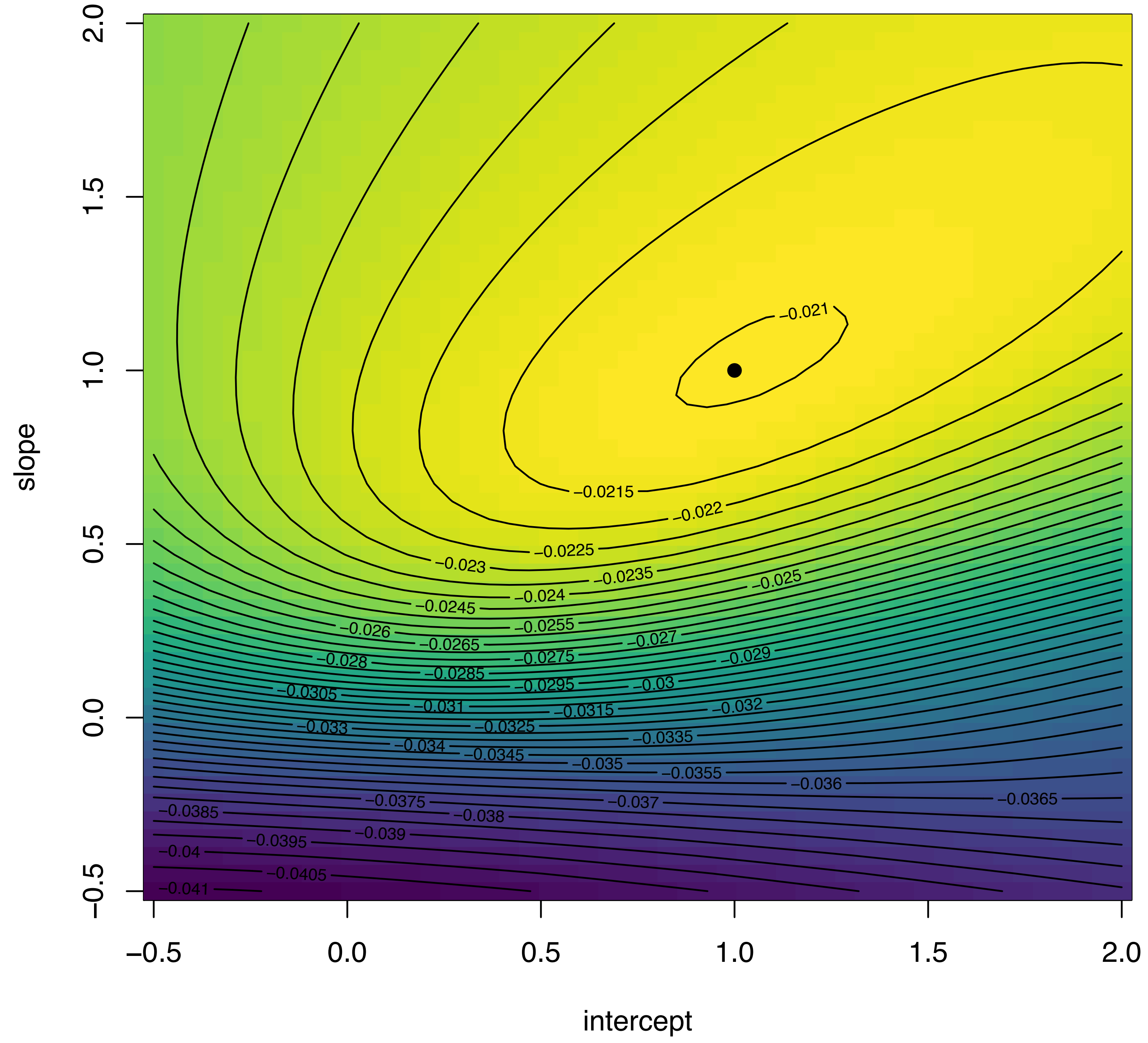
(risk = expected loss)

$$\log \frac{p}{1-p} = 1 + x,$$

$$x \sim U[-6, 6]$$

Some risk surfaces

Brier

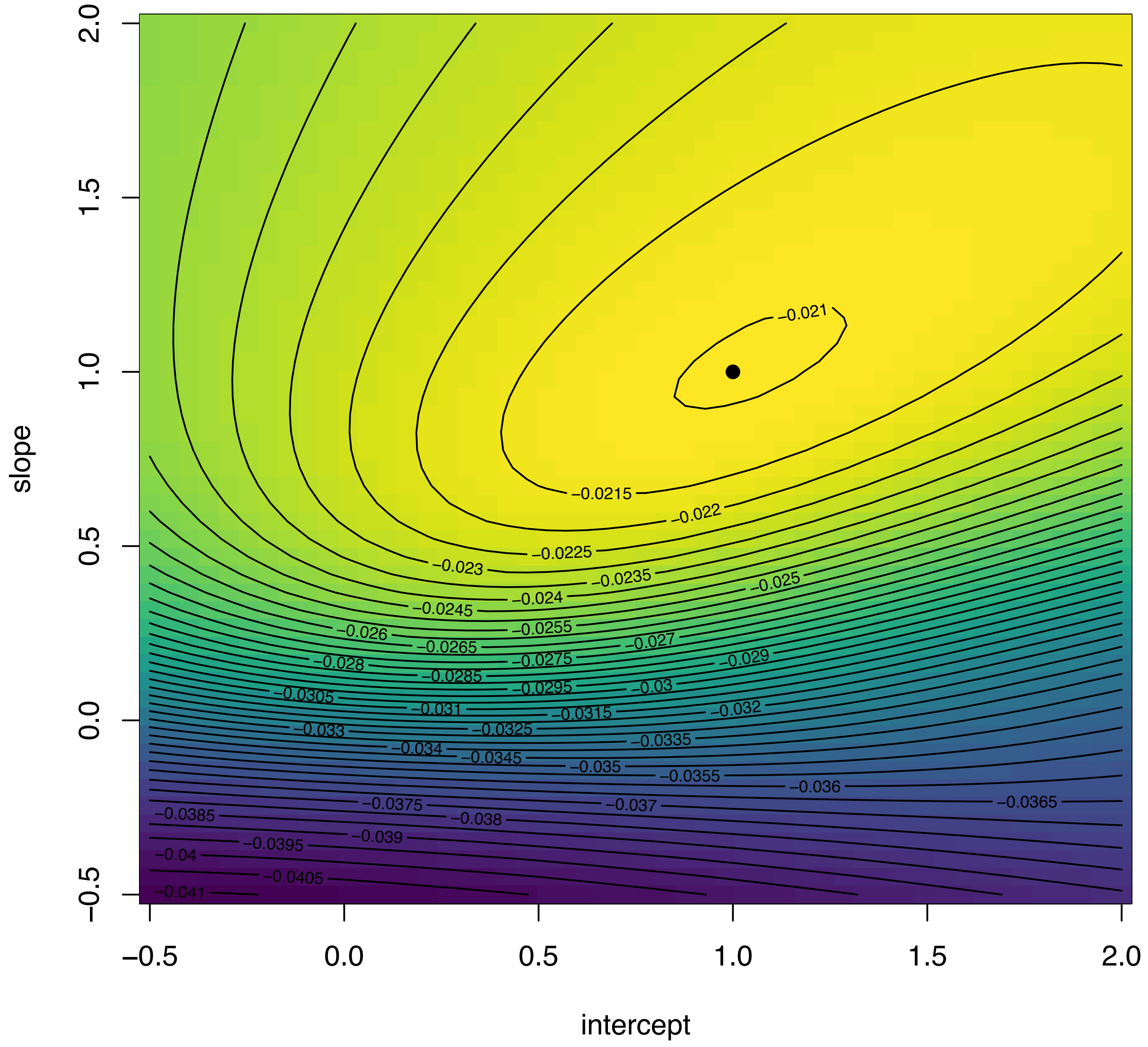


$$\log\frac{p}{1-p} = 1 + x,$$
$$x \sim U[-6,6]$$

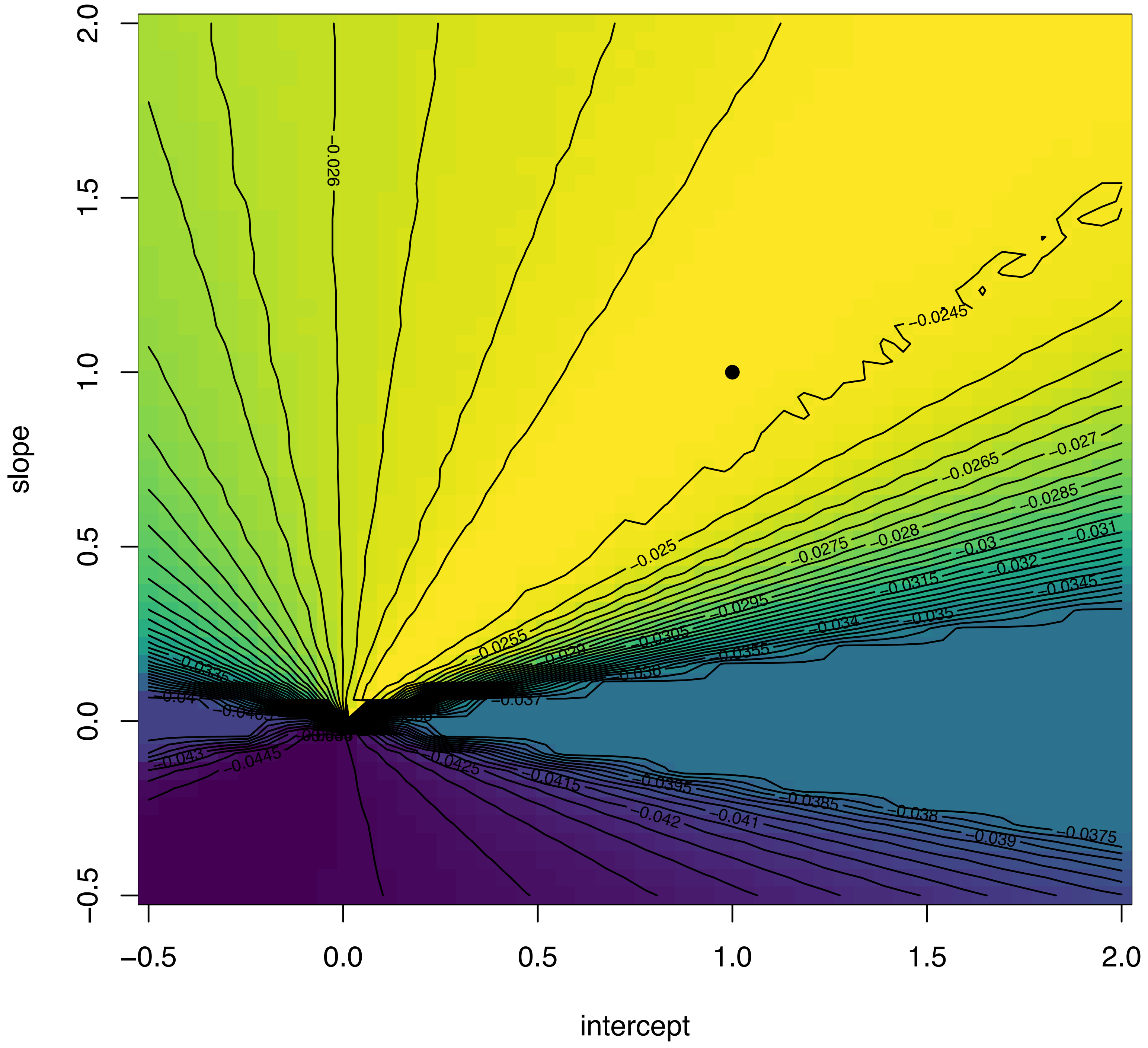
Brighter is better

Some risk surfaces

Brier



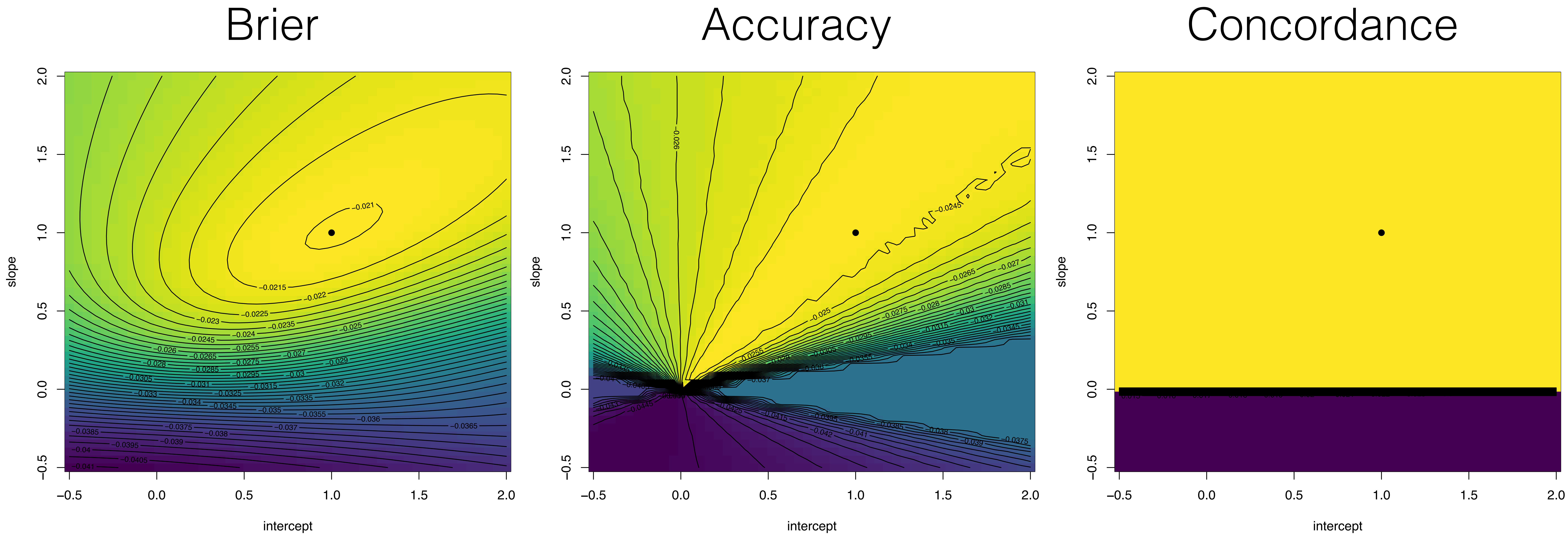
Accuracy



$$\log\frac{p}{1-p} = 1 + x,$$
$$x \sim U[-6,6]$$

Brighter is better

Some risk surfaces



Brier

Accuracy

Concordance

$$\log \frac{p}{1-p} = 1 + x,$$
$$x \sim U[-6, 6]$$

Brighter is better

Some “technical” sources of variation

- The big classic one: sample size
- Scoring rule
- Validation procedure

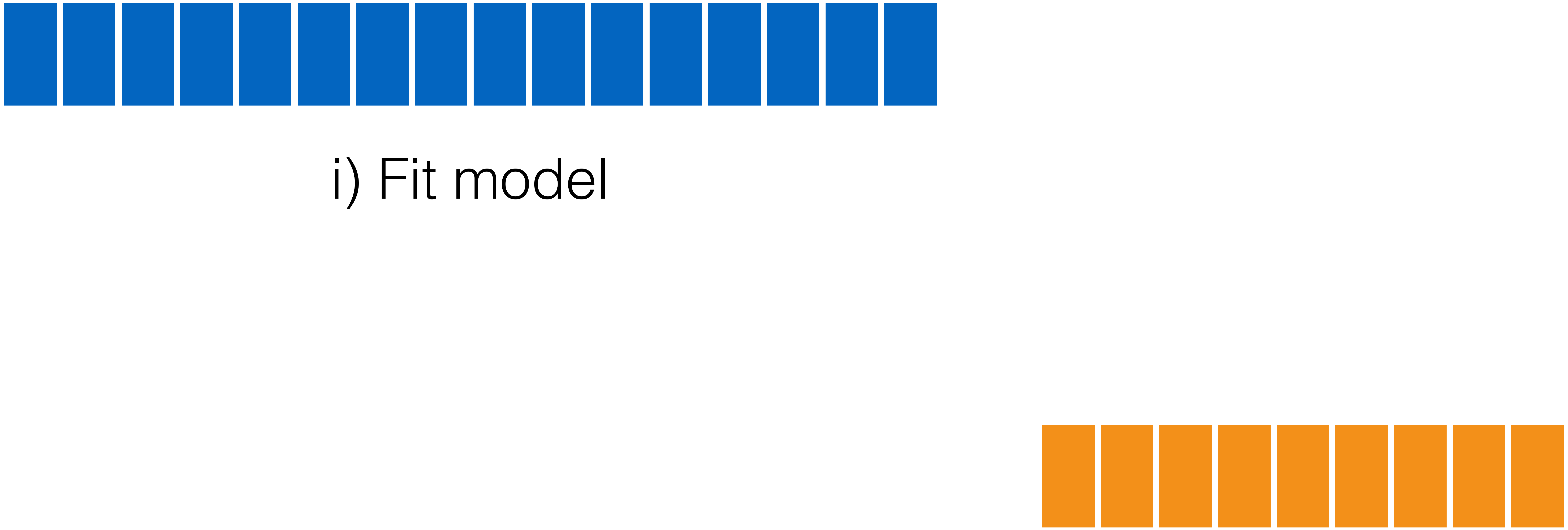
Validation

- Holdout data
- Cross-validation
- Repeat CV
- The Bootstrap

Holdout data

Holdout data

Holdout data



i) Fit model

ii) Calculate score

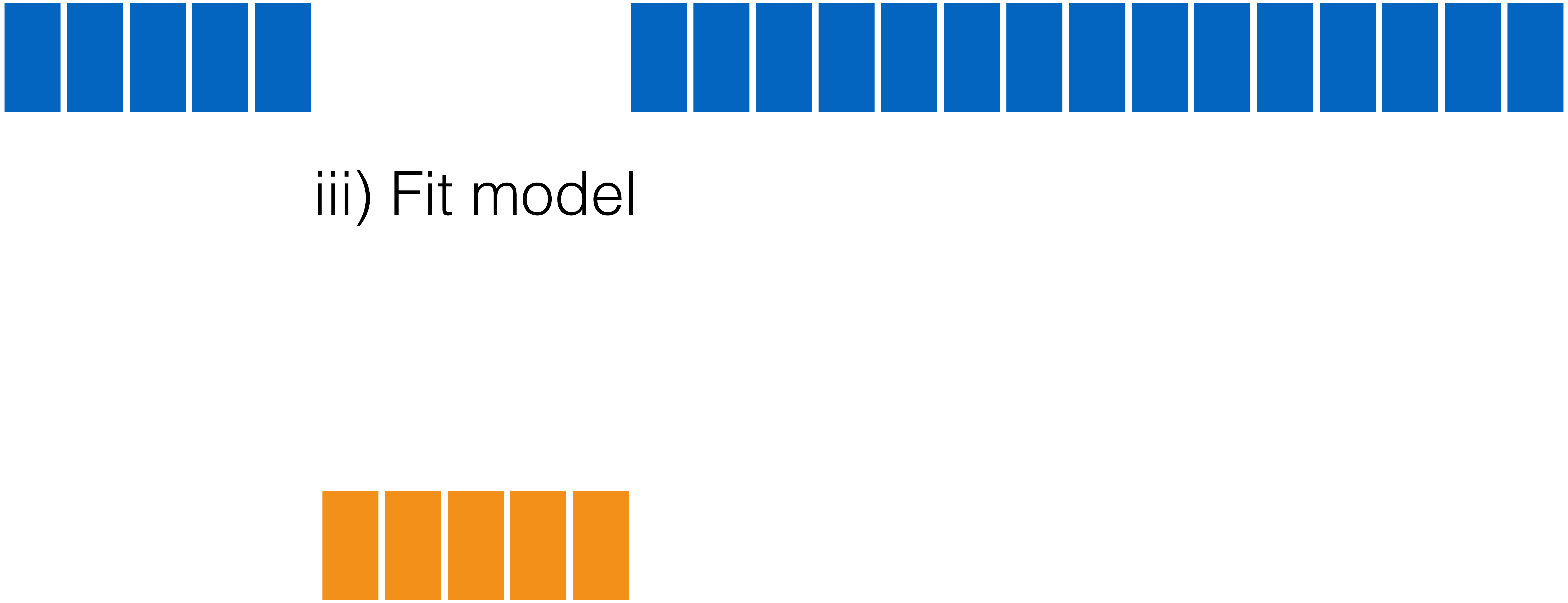
Cross validation

Cross validation

i) Fit model

ii) Score

Cross validation



iii) Fit model

iv) Score

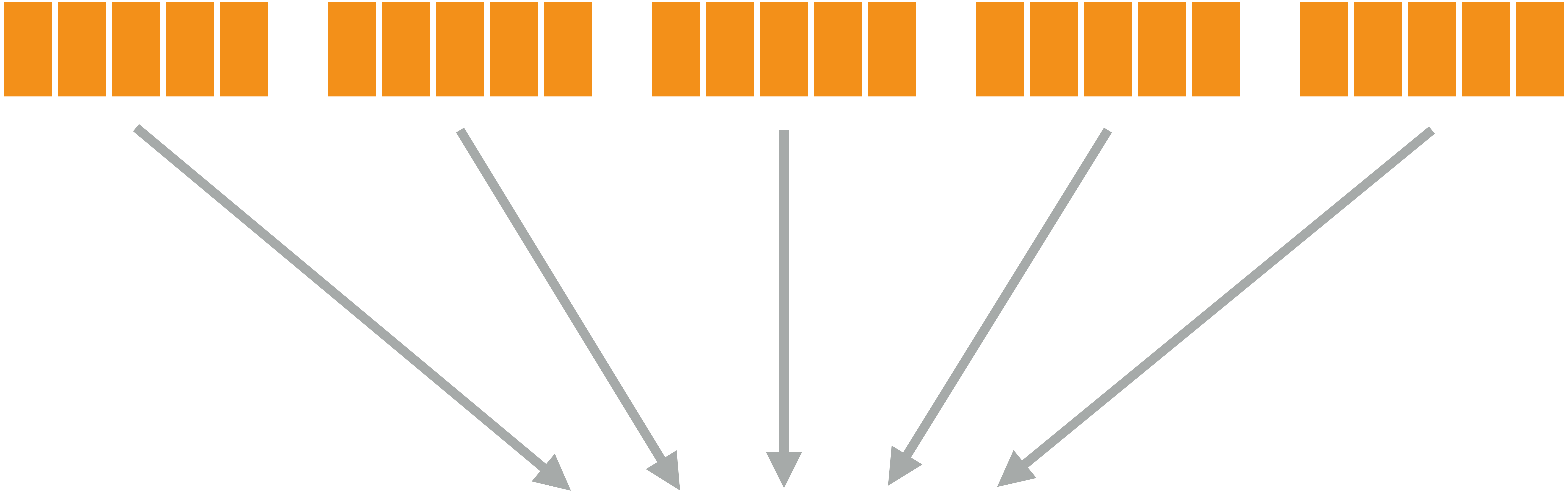
Cross validation

iii) Fit model

iv) Score

&c., &c.

Cross validation



xi) Summarize by mean, sd

Repeated cross validation

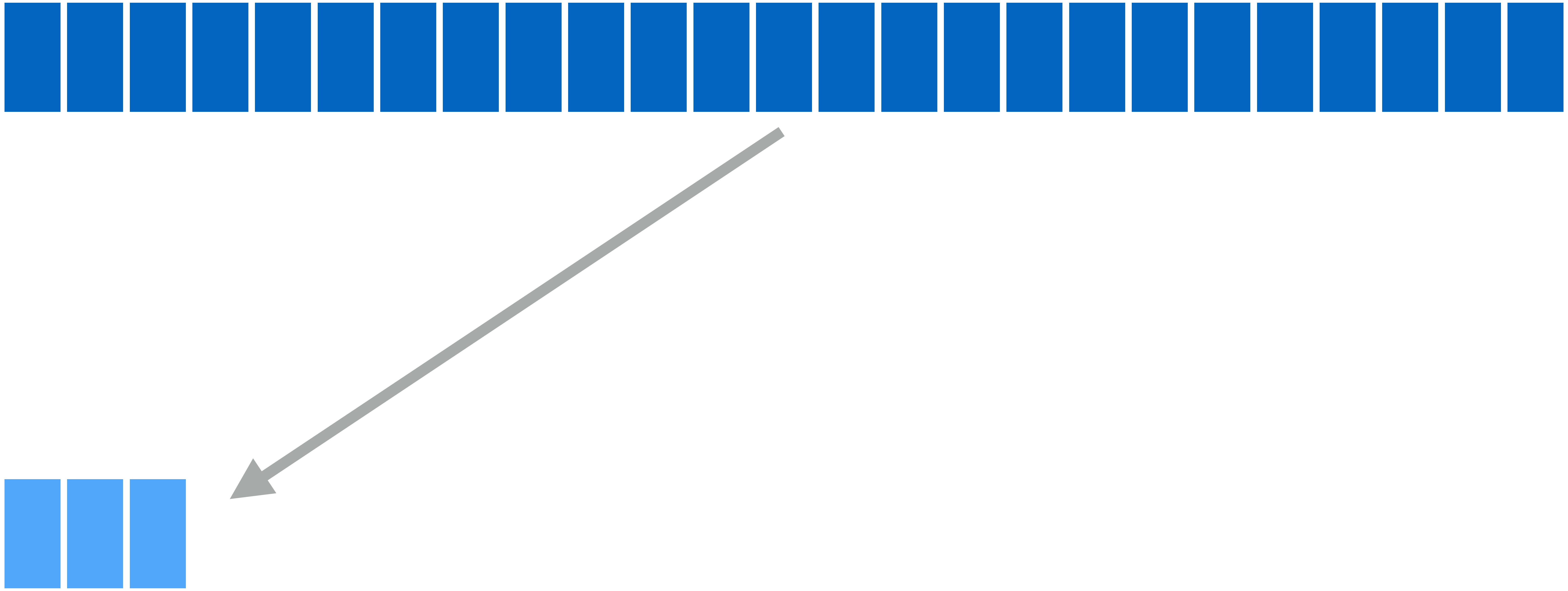
It’s exactly what you’d expect

Bootstrap

Bootstrap

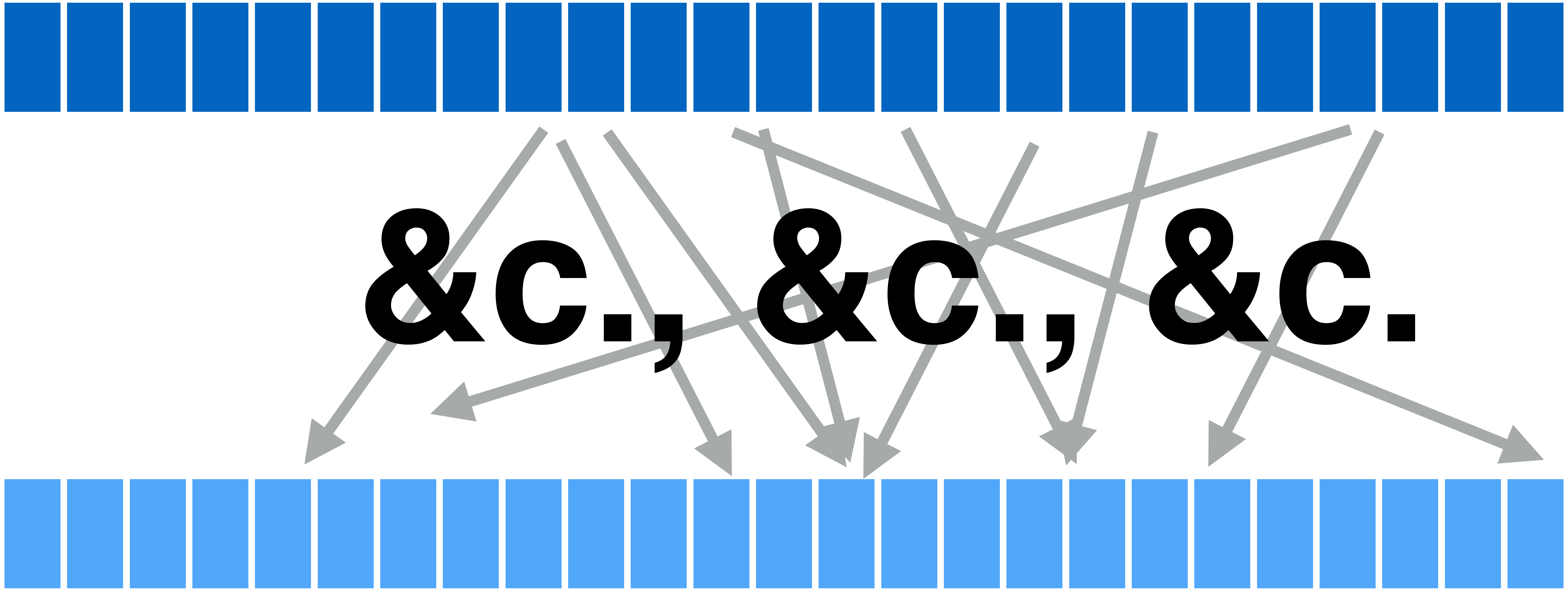
Bootstrap

Bootstrap



Bootstrap

Bootstrap



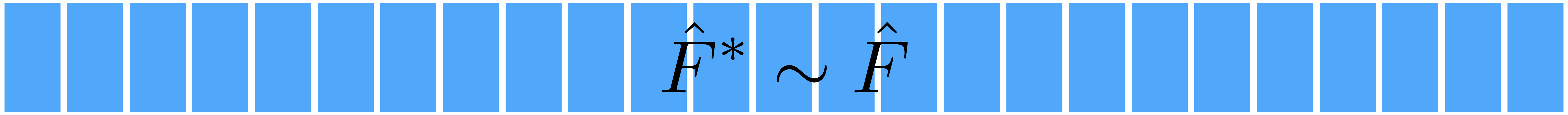
&c., &c., &c.

Bootstrap

$$\hat{F} \sim F$$

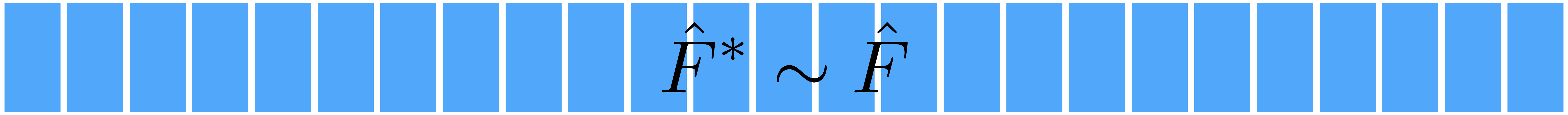
$$\hat{F}^* \sim \hat{F}$$

Bootstrap



$\hat{F}^* \sim \hat{F}$

Bootstrap



$\hat{F}^* \sim \hat{F}$

$$T(\hat{F}^*, \hat{F}) \sim T(\hat{F}, F)$$

“The bootstrap principle”

Relative efficiency of two estimators

For two estimators, $T_1, T_2,$ of the same quantity :

$$\frac{\mathrm{Var}(T_1)}{\mathrm{Var}(T_2)}$$

Relative efficiency of two estimators

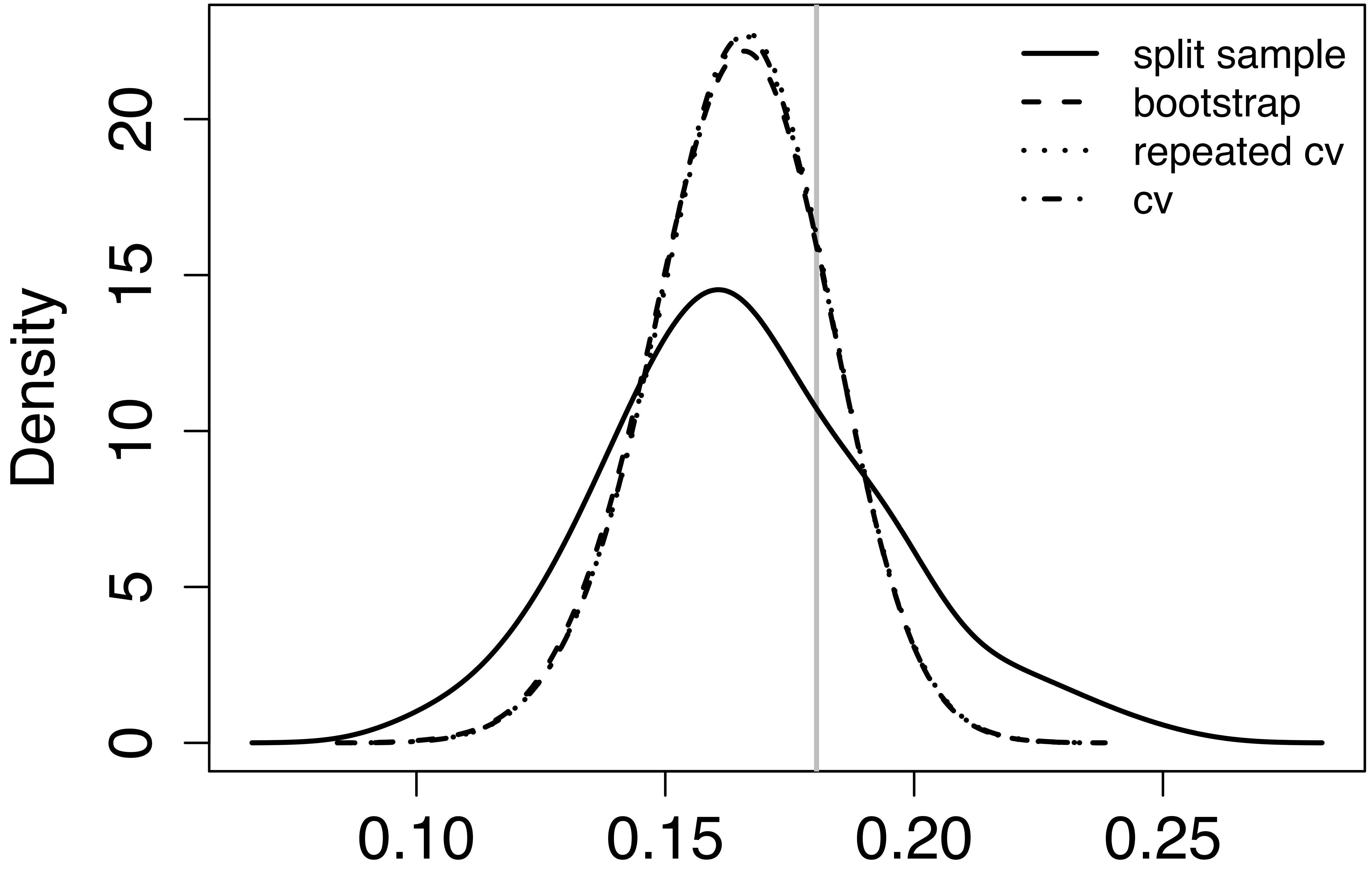
For two estimators, T_1, T_2, of the same quantity :

$$\frac{\mathrm{Var}(T_1)}{\mathrm{Var}(T_2)}$$

All else being equal, pick the less variable one

Relative efficiency of two estimators

Brier score estimated in different ways

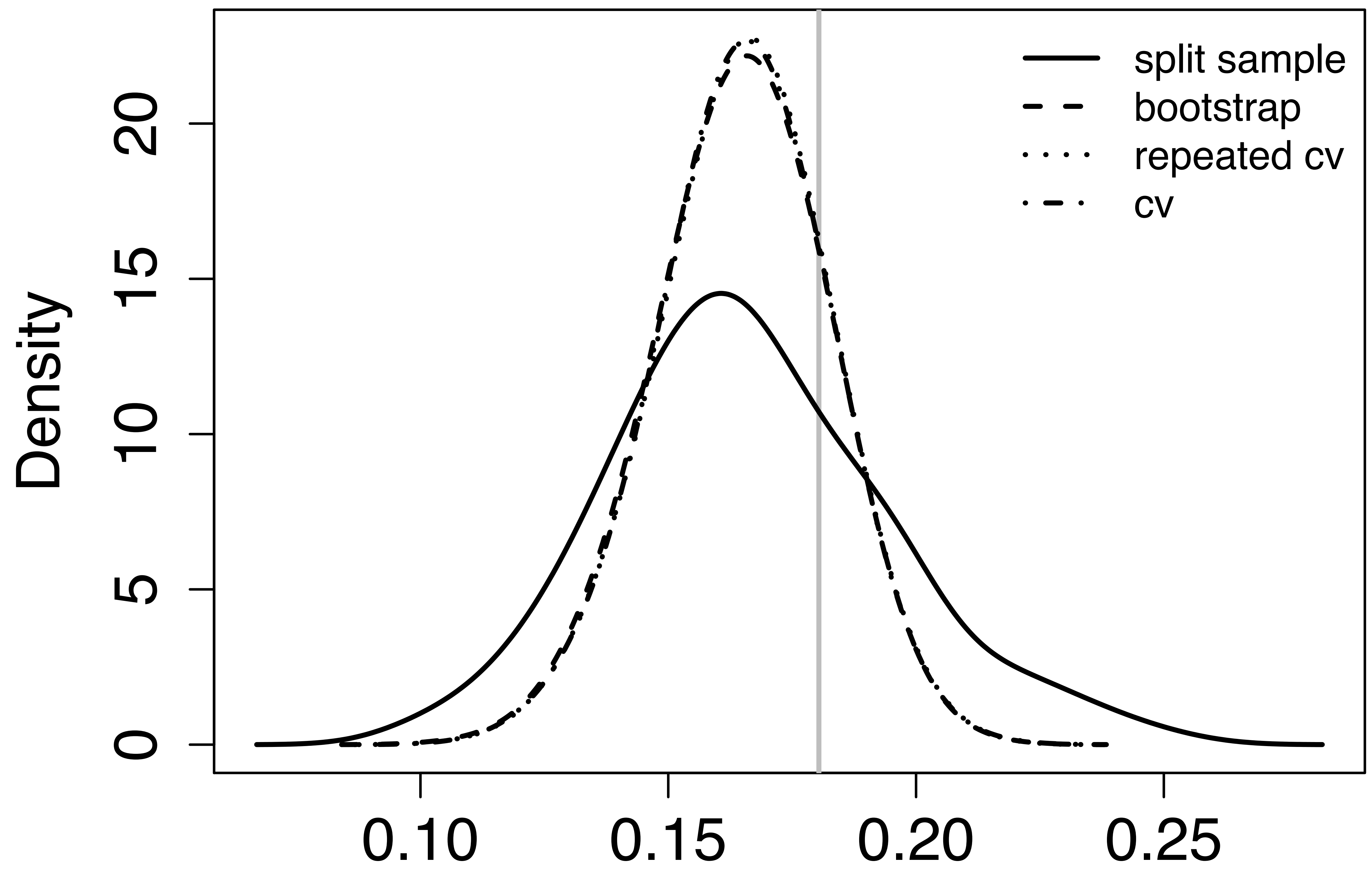


Relative efficiency to split sample:

Bootstrap:	3.5
CV:	3.6
Repeat CV:	3.6

Relative efficiency of two estimators

Brier score estimated in different ways



Relative efficiency to split sample:

Bootstrap:	3.5
CV:	3.6
Repeat CV:	3.6

Need 3–4 times as many obs. w/ split sample!

Some lessons

Some lessons

1. **Small data:** new observations are hard to get

Some lessons

1. **Small data:** new observations are hard to get
2. Optimize a less weird scoring rule

Some lessons

1. **Small data:** new observations are hard to get
2. Optimize a less weird scoring rule
3. Estimate with less variance

❧ intermission ☙

Act III: Hold Fast

Brier score + Bootstrap

	model 2.6	**model 2.7**
LIMMA-t	.44 ± .30	.76 ± .20
SAM	.46 ± .26	.75 ± .24
ANOVA-fs	.51 ± .29	.75 ± .16
ANOVA-s	.41 ± .57	.75 ± .38
t-test	.65 ± 1.5	.74 ± .71
ANOVA-f	.44 ± .25	.72 ± .21
intercept	.5	
stratum	.49 ± .055	
lasso	.36 ± 1.4	
ridge	.81 ± 3.3	

	model 2.6	**model 2.7**
t-test	.17 ± .45	.17 ± .33
ANOVA-fs	.27 ± .13	.18 ± .10
SAM	.34 ± .11	.20 ± .15
ANOVA-s	.33 ± .22	.20 ± .25
ANOVA-f	.31 ± .084	.21 ± .11
LIMMA-t	.35 ± .14	.20 ± .17
intercept	.19 ± .010	
stratum	.22 ± .029	
lasso	.27 ± .19	
ridge	.23 ± .30	

Concordance:
Higher better, random guess is .5

Brier score:
Lower better, null model is .19

Brier score + Bootstrap

Concordance

	model 2.6	model 2.7
LIMMA-t	.44 ± .30	.76 ± .20
SAM	.46 ± .26	.75 ± .24
ANOVA-fs	.51 ± .29	.75 ± .16
ANOVA-s	.41 ± .57	.75 ± .38
t-test	.65 ± 1.5	.74 ± .71
ANOVA-f	.44 ± .25	.72 ± .21
intercept	.5	
stratum	.49 ± .055	
lasso	.36 ± 1.4	
ridge	.81 ± 3.3	

Brier

	model 2.6	model 2.7
t-test	.17 ± .45	.17 ± .33
ANOVA-fs	.27 ± .13	.18 ± .10
SAM	.34 ± .11	.20 ± .15
ANOVA-s	.33 ± .22	.20 ± .25
ANOVA-f	.31 ± .084	.21 ± .11
LIMMA-t	.35 ± .14	.20 ± .17
intercept	.19 ± .010	
stratum	.22 ± .029	
lasso	.27 ± .19	
ridge	.23 ± .30	

Concordance:
Higher better, random guess is .5

Brier score:
Lower better, null model is .19

Brier score + Bootstrap

	model 2.6	**model 2.7**
LIMMA-t	.44 ± .30	.76 ± .20
SAM	.46 ± .26	.75 ± .24
ANOVA-fs	.51 ± .29	.75 ± .16
ANOVA-s	.41 ± .57	.75 ± .38
t-test	.65 ± 1.5	.74 ± .71
ANOVA-f	.44 ± .25	.72 ± .21
intercept	.5	
stratum	.49 ± .055	
lasso	.36 ± 1.4	
ridge	.81 ± 3.3	

	model 2.6	**model 2.7**
t-test	.17 ± .45	.17 ± .33
ANOVA-fs	.27 ± .13	.18 ± .10
SAM	.34 ± .11	.20 ± .15
ANOVA-s	.33 ± .22	.20 ± .25
ANOVA-f	.31 ± .084	.21 ± .11
LIMMA-t	.35 ± .14	.20 ± .17
intercept	.19 ± .010	
stratum	.22 ± .029	
lasso	.27 ± .19	
ridge	.23 ± .30	

Concordance:
Higher better, random guess is .5

Brier score:
Lower better, null model is .19

Brier score + Bootstrap

	model 2.6	model 2.7
LIMMA-t	.44 ± .30	.76 ± .20
SAM	.46 ± .26	.75 ± .24
ANOVA-fs	.51 ± .29	.75 ± .16
ANOVA-s	.41 ± .57	.75 ± .38
t-test	.65 ± 1.5	.74 ± .71
ANOVA-f	.44 ± .25	.72 ± .21
intercept	.5	
stratum	.49 ± .055	
lasso	.56 ± 1.4	
ridge	.81 ± 3.3	

	model 2.6	model 2.7
t-test	.17 ± .45	.17 ± .33
ANOVA-fs	.27 ± .13	.18 ± .10
SAM	.34 ± .11	.20 ± .15
ANOVA-s	.33 ± .22	.20 ± .25
ANOVA-f	.31 ± .084	.21 ± .11
LIMMA-t	.35 ± .14	.20 ± .17
intercept	.19 ± .010	
stratum	.22 ± .029	
lasso	.27 ± .19	
ridge	.23 ± .30	

Concordance:
Higher better, random guess is .5

Brier score:
Lower better, null model is .19

In short more lizards ahead

Reminder of likelihood penalties

$$\sum \beta_i^2 \leq t$$

$$\sum |\beta_i| \leq t$$

$$\sum \left[\alpha\beta_i^2 + (1-\alpha)|\beta_i|\right] \leq t$$

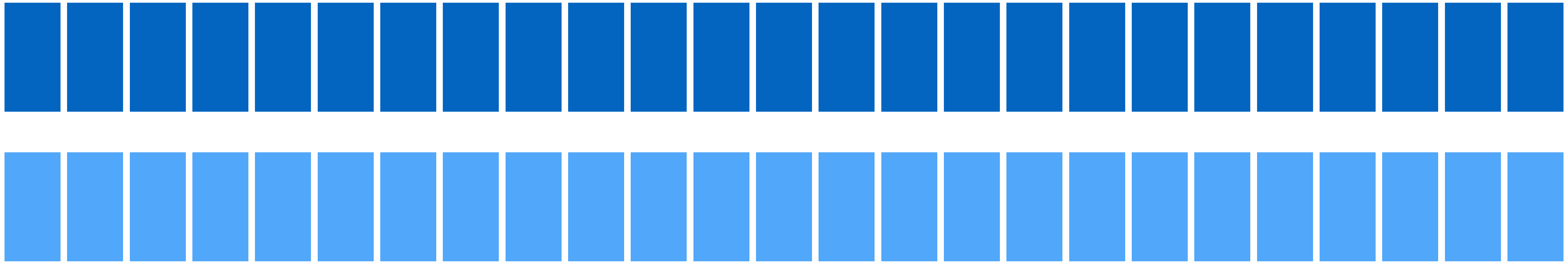
Need to choose t (aka lambda)

Risky procedure

```
# nested cv in bootstrap
boot <- boostrap_samples()
for (b in boot) {
  lambda <- cross_validate_glmnet(b)
}
```

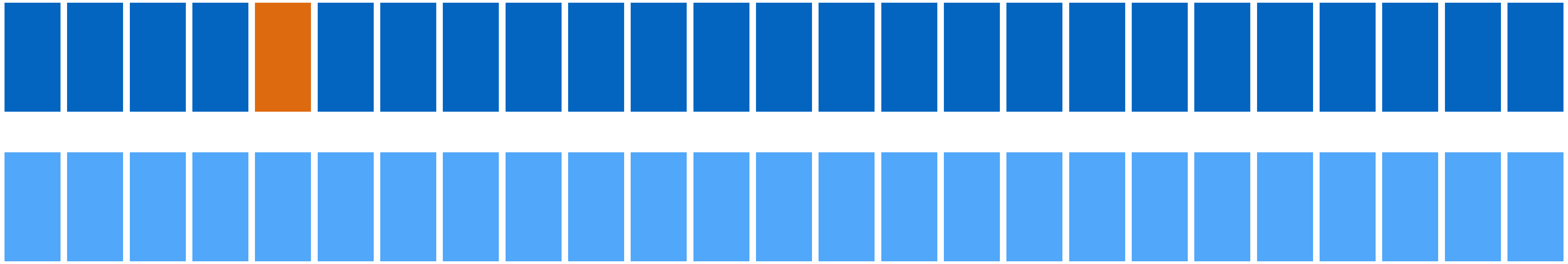
Risky procedure

```
# nested cv in bootstrap
boot <- boostrap_samples()
for (b in boot) {
  lambda <- cross_validate_glmnet(b)
}
```



Risky procedure

```
# nested cv in bootstrap
boot <- boostrap_samples()
for (b in boot) {
  lambda <- cross_validate_glmnet(b)
}
```

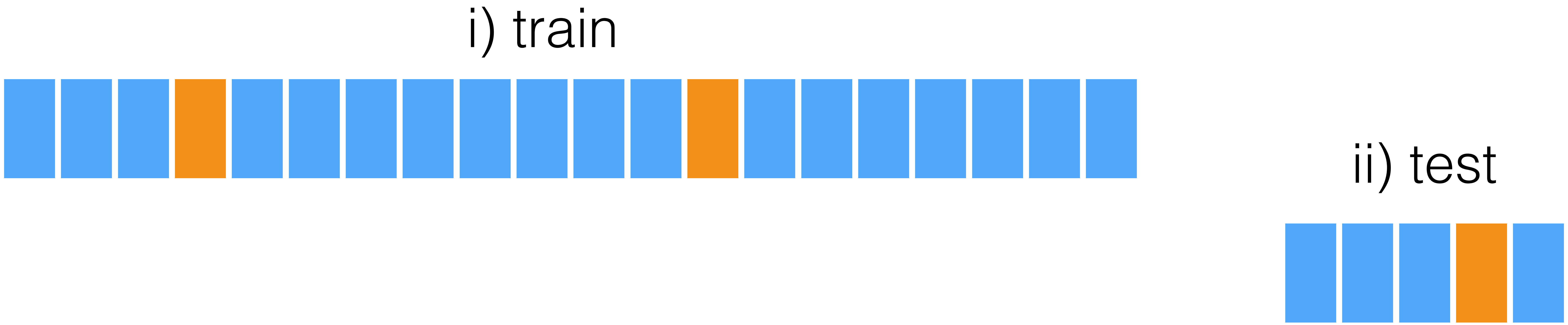


Risky procedure

```
# nested cv in bootstrap
boot <- boostrap_samples()
for (b in boot) {
  lambda <- cross_validate_glmnet(b)
}
```

Risky procedure

```r
# nested cv in bootstrap
boot <- boostrap_samples()
for (b in boot) {
  lambda <- cross_validate_glmnet(b)
}
```



Risky procedure

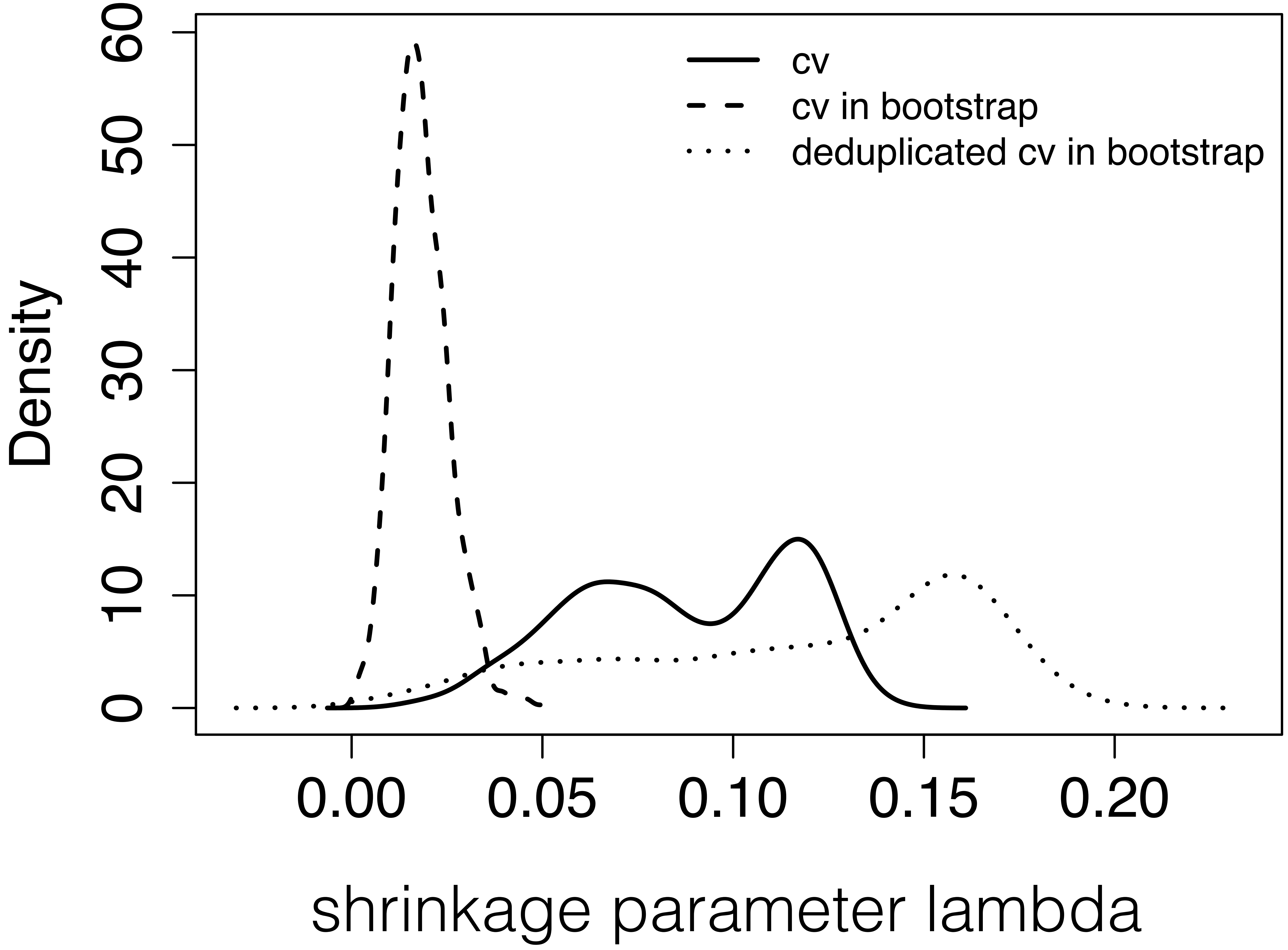
```r
# nested cv in bootstrap
boot <- boostrap_samples()
for (b in boot) {
  lambda <- cross_validate_glmnet(b)
}
```

i) train

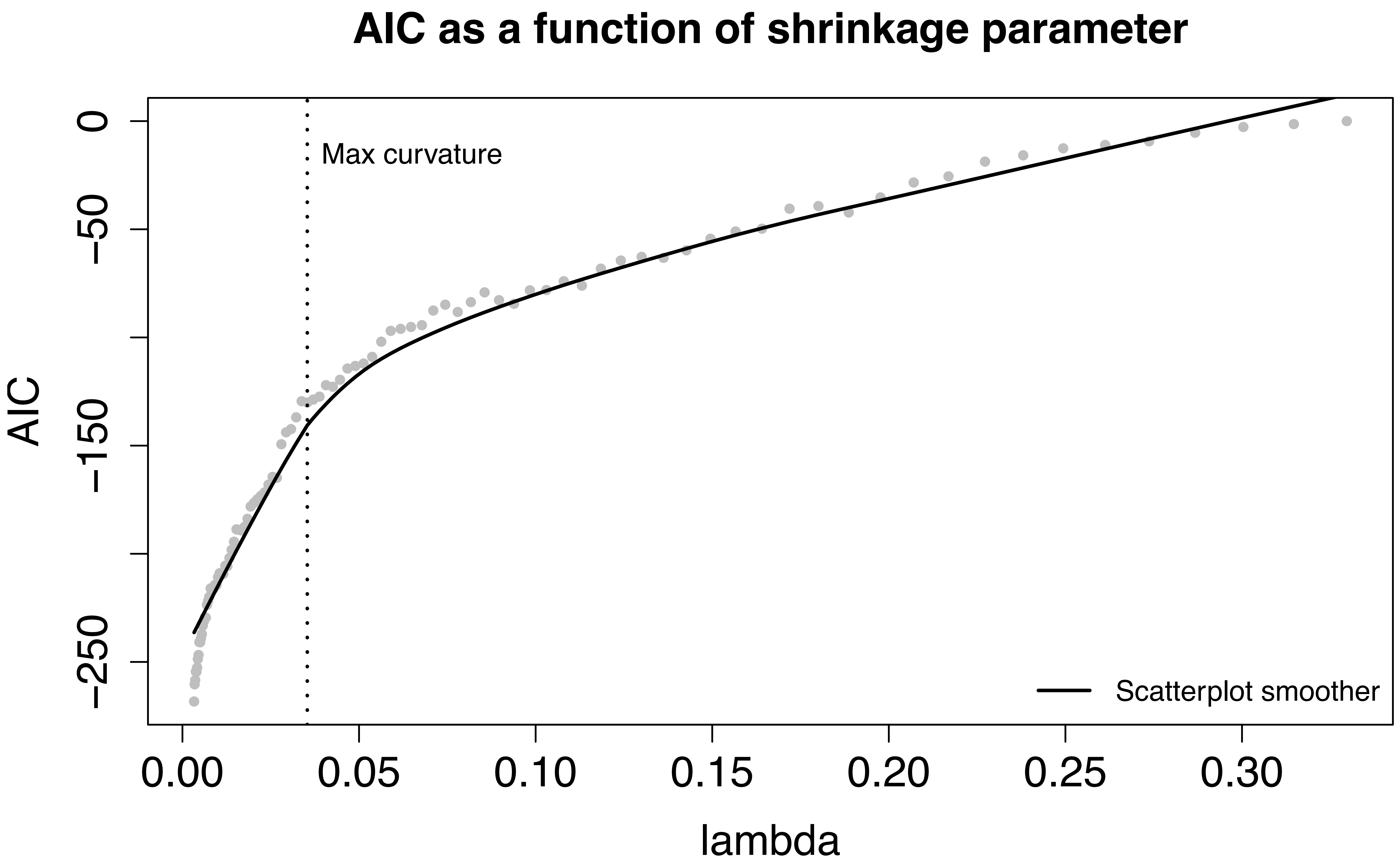
ii) test

Bias toward 🟧 **!!!!!!!!!**

Risky procedure

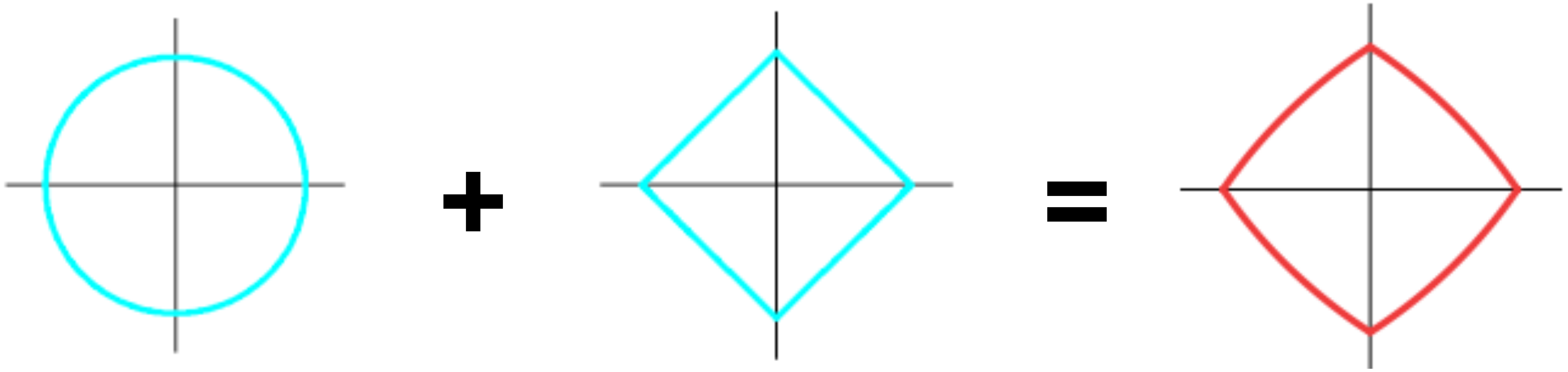


Instead choose lambda by AIC



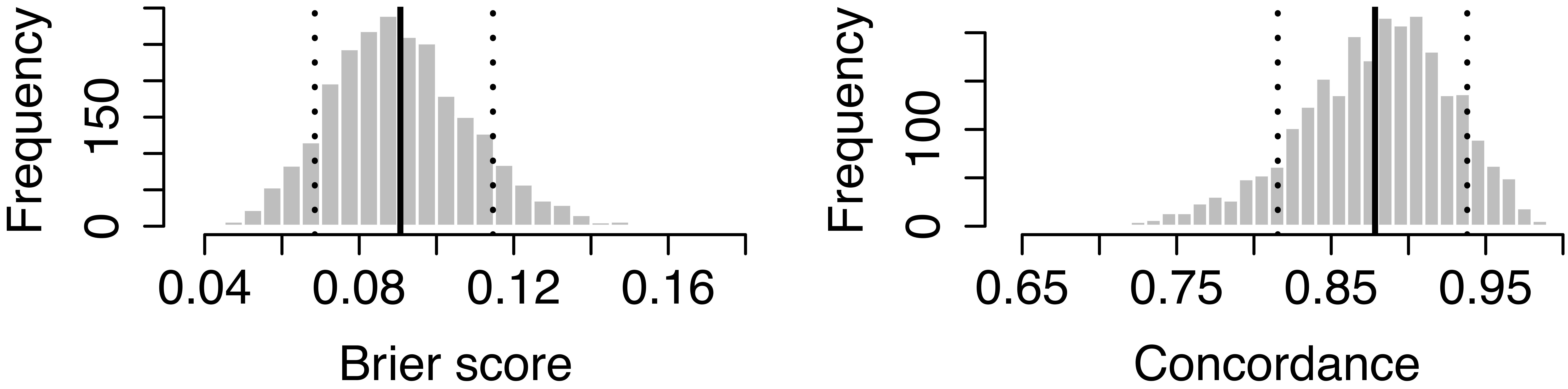
ElasticNet, alpha = .5

$$\sum \left[\alpha\beta_i^2 + (1-\alpha)|\beta_i|\right] \leq t$$



Figures from Hastie, Tibshirani, and Friedman: The Elements of Statistical Learning

ElasticNet, alpha = .5



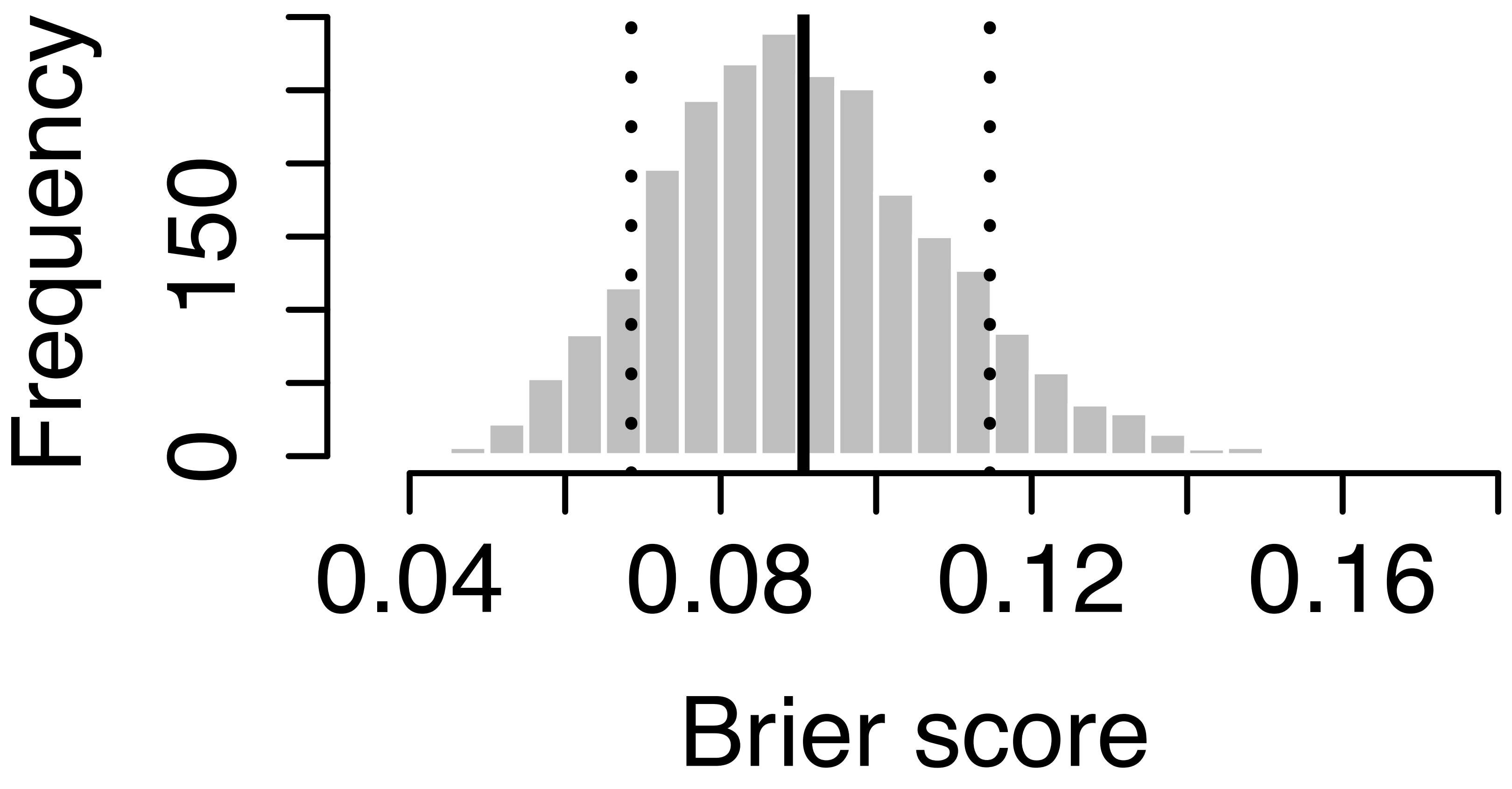
ElasticNet, alpha = .5

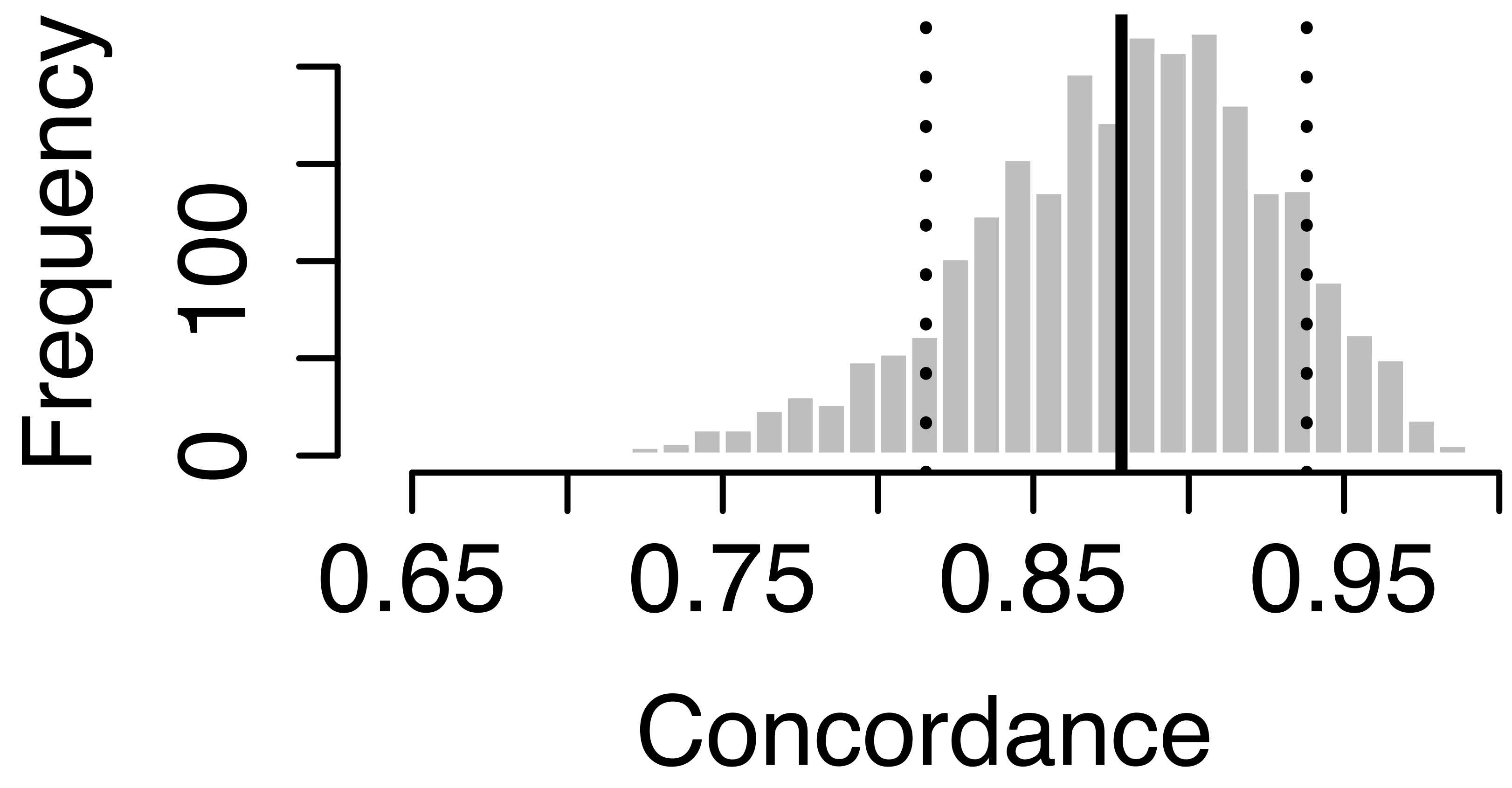
	model 2.6	**model 2.7**
LIMMA-t	.44 ± .30	.76 ± .20
SAM	.46 ± .26	.75 ± .24
ANOVA-fs	.51 ± .29	.75 ± .16
ANOVA-s	.41 ± .57	.75 ± .38
t-test	.65 ± 1.5	.74 ± .71
ANOVA-f	.44 ± .25	.72 ± .21

intercept	.5
stratum	.49 ± .055
lasso	.36 ± 1.4
ridge	.81 ± 3.3

	model 2.6	**model 2.7**
t-test	.17 ± .45	.17 ± .33
ANOVA-fs	.27 ± .13	.18 ± .10
SAM	.34 ± .11	.20 ± .15
ANOVA-s	.33 ± .22	.20 ± .25
ANOVA-f	.31 ± .084	.21 ± .11
LIMMA-t	.35 ± .14	.20 ± .17

intercept	.19 ± .010
stratum	.22 ± .029
lasso	.27 ± .19
ridge	.23 ± .30



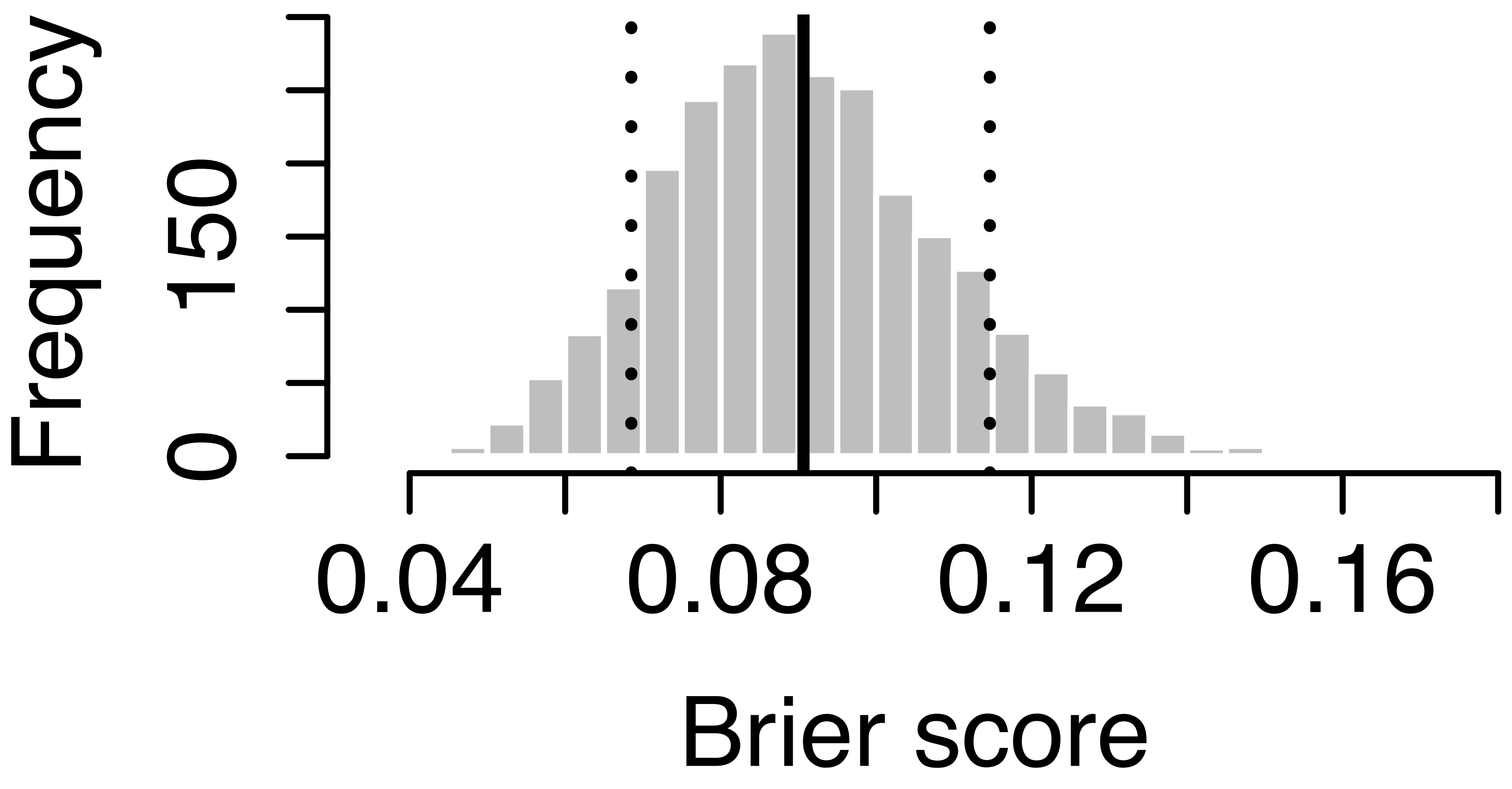


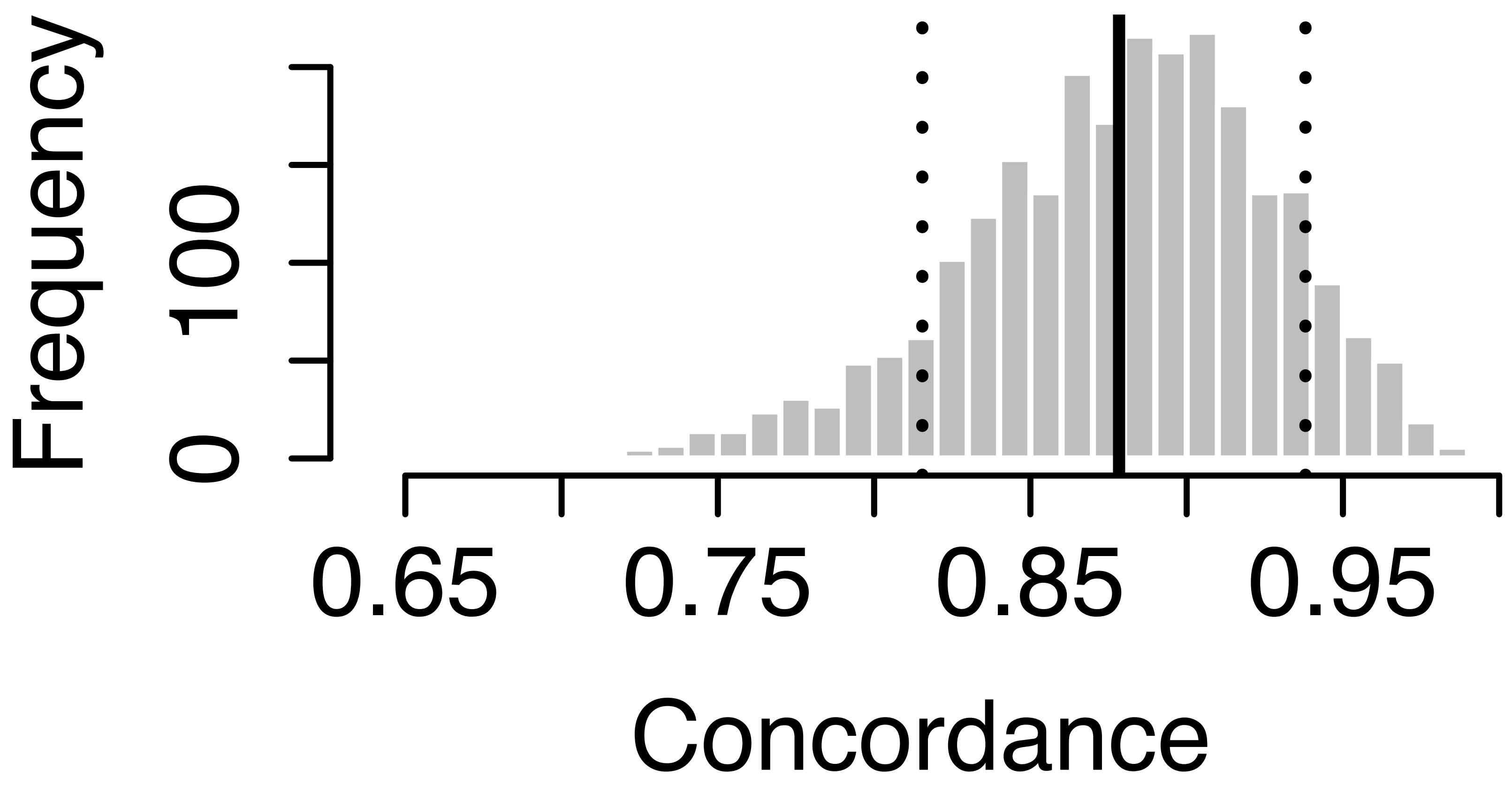
ElasticNet, alpha = .5

	model 2.6	**model 2.7**
LIMMA-t	.44 ± .30	.76 ± .20
SAM	.46 ± .26	.75 ± .24
ANOVA-fs	.51 ± .29	.75 ± .16
ANOVA-s	.41 ± .57	.75 ± .38
t-test	.65 ± 1.5	.74 ± .71
ANOVA-f	.44 ± .25	.72 ± .21
intercept	.5	
stratum	.49 ± .055	
lasso	.36 ± 1.4	
ridge	.81 ± 3.3	

	model 2.6	**model 2.7**
t-test	.17 ± .45	.17 ± .33
ANOVA-fs	.27 ± .13	.18 ± .10
SAM	.34 ± .11	.20 ± .15
ANOVA-s	.33 ± .22	.20 ± .25
ANOVA-f	.31 ± .084	.21 ± .11
LIMMA-t	.35 ± .14	.20 ± .17
intercept	.19 ± .010	
stratum	.22 ± .029	
lasso	.27 ± .19	
ridge	.23 ± .30	

Use stratum information



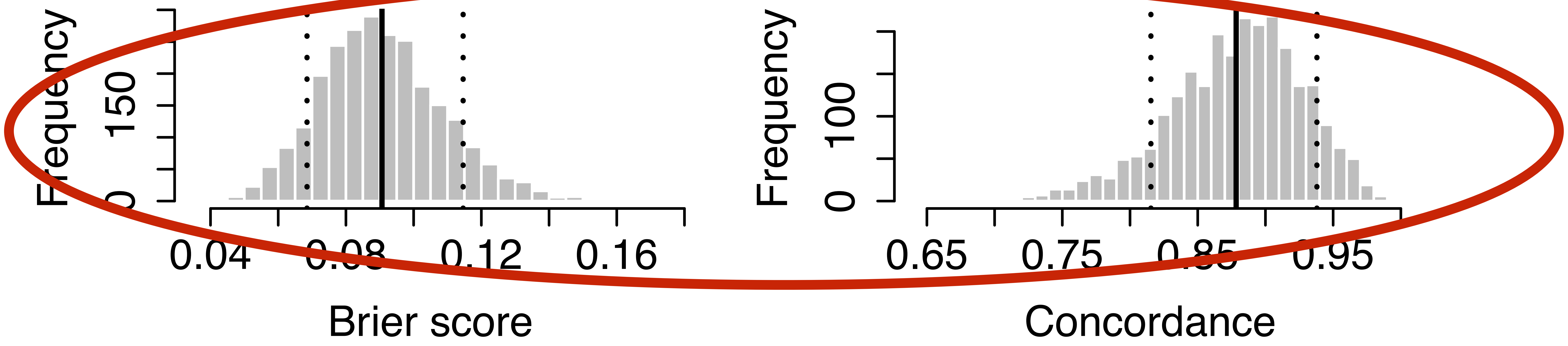


ElasticNet, alpha = .5

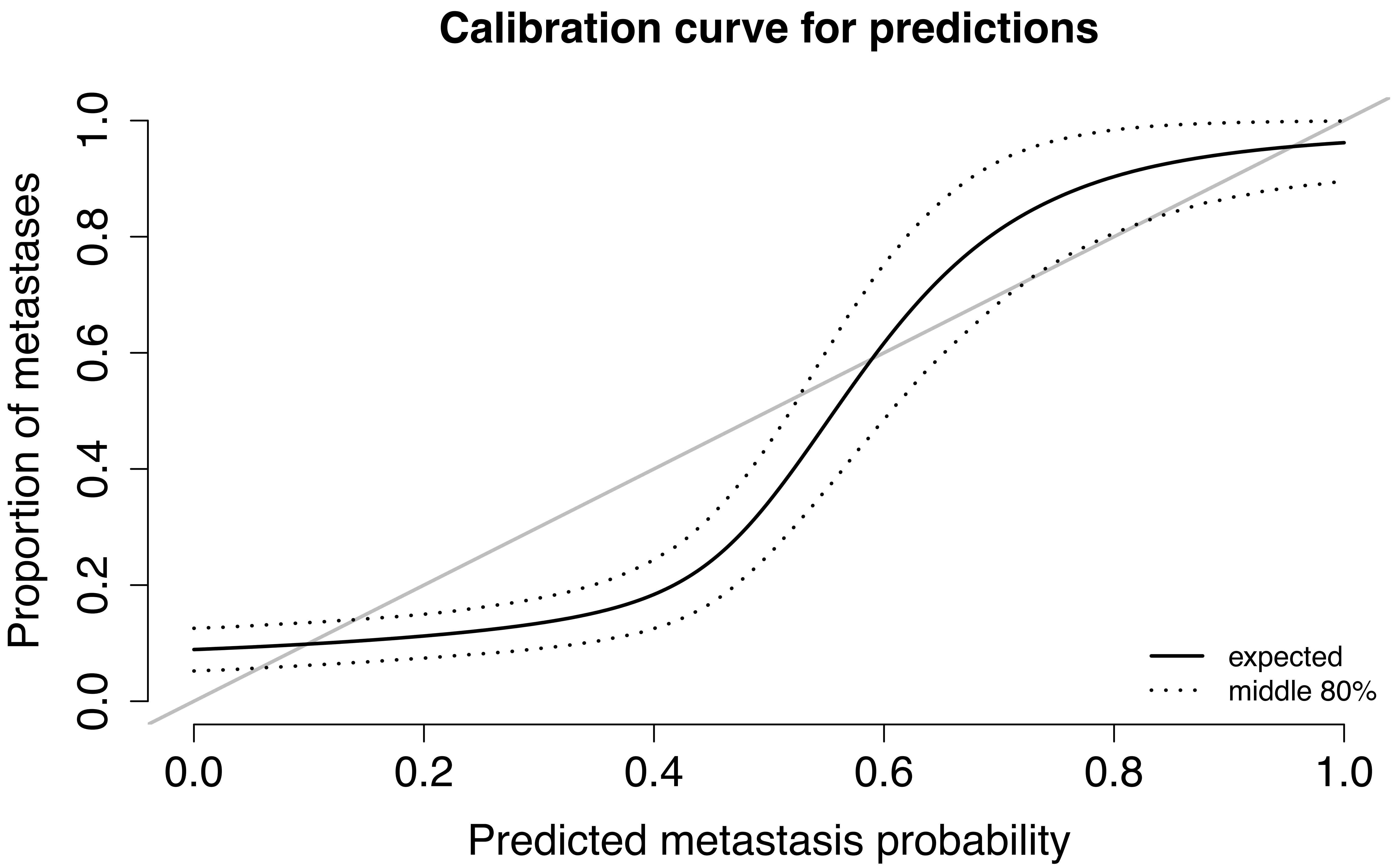
	model 2.6	model 2.7
LIMMA-t	.44 ± .30	.76 ± .20
SAM	.46 ± .26	.75 ± .24
ANOVA-fs	.51 ± .29	.75 ± .16
ANOVA-s	.41 ± .57	.75 ± .38
t-test	.65 ± 1.5	.74 ± .71
ANOVA-f	.44 ± .25	.72 ± .21
intercept	.5	
stratum	.49 ± .055	
lasso	.36 ± 1.4	
ridge	.81 ± 3.3	

	model 2.6	model 2.7
t-test	.17 ± .45	.17 ± .33
ANOVA-fs	.27 ± .13	.18 ± .10
SAM	.34 ± .11	.20 ± .15
ANOVA-s	.33 ± .22	.20 ± .25
ANOVA-f	.31 ± .084	.21 ± .11
LIMMA-t	.35 ± .14	.20 ± .17
intercept	.19 ± .010	
stratum	.22 ± .029	
lasso	.27 ± .19	
ridge	.23 ± .30	

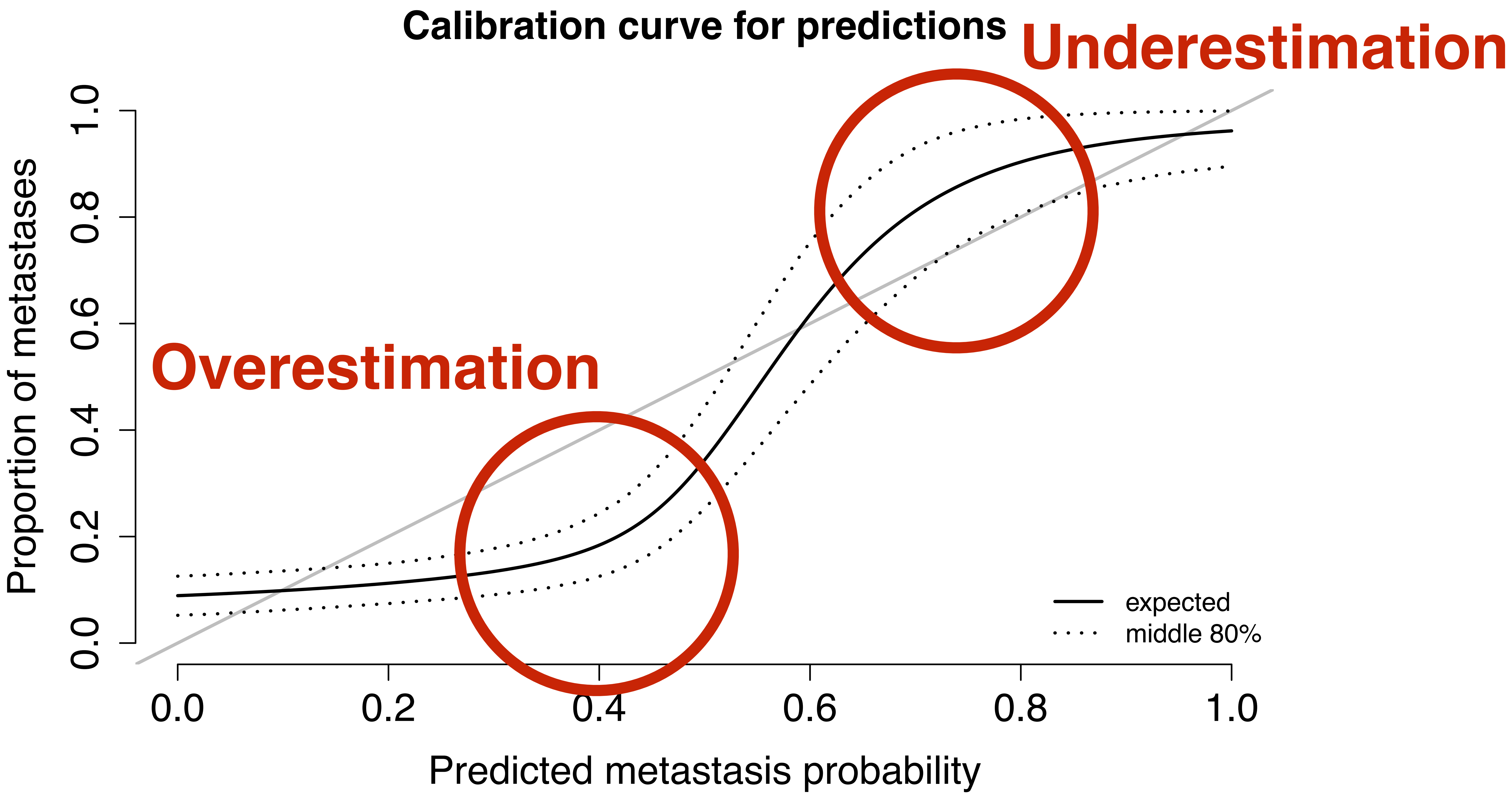
Does not



ElasticNet, alpha = .5



ElasticNet, alpha = .5



108 genes selected

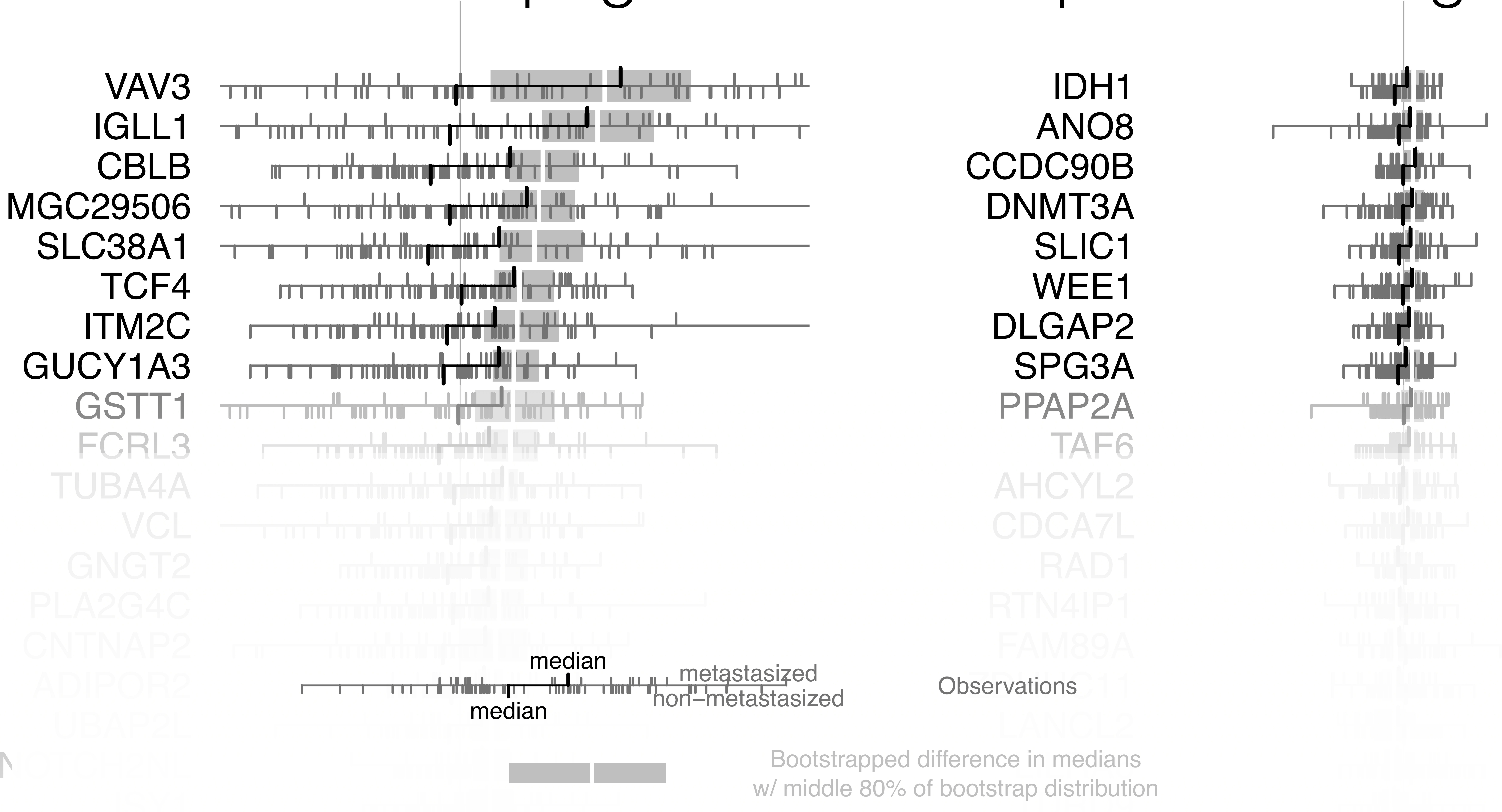
GRK5^a	0.853	C1orf115	0.290	ANO8	0.221	FBLN5	0.157
GPATCH4	0.682	LOC654055	0.287	PTTG1IP	0.219	BLMH	0.156
GNGT2	0.474	RNF214	0.280	3NDg8gVCd. . .^{b}	0.218	FCRL3	0.149
PDGFD^c	0.467	SULT1A1	0.278	USF1	0.216	TDRD9	0.143
FAM24B	0.457	ZNF365	0.271	BCCIP	0.210	ACY1	0.142
PTPRN2	0.442	USE1	0.267	MGC29506	0.209	ZFP57	0.142
CBLB	0.440	DNMT3A	0.267	GRK5^a	0.207	SLIC1	0.138
PDCL	0.410	LOC649210	0.266	WTIP	0.205	PICK1	0.135
RASA2	0.380	CNTNAP2	0.265	BCL10	0.204	RTN4IP1	0.134
C11orf48	0.376	IL2RA	0.265	DLGAP2	0.200	CDCA7L	0.132
TCEB1	0.374	CCT5	0.264	HRAS	0.199	BEX4	0.131
CAPN3	0.354	R3HDM1	0.263	RAD1	0.189	FCAR	0.130
STK19	0.351	MRPL43	0.260	PRKCE	0.187	ANKRD35	0.111
GUCY1A3	0.348	SLC38A1	0.256	UBAP2L	0.186	USP39	0.109
ZDHHC11	0.345	GNG8	0.255	BPI	0.186	KIAA0495	0.106
SULT1A3	0.336	PLA2G4C	0.251	DTX1	0.184	BRI3BP	0.106
Z6FIQGkeo. . .^{d}	0.335	TCF4	0.248	LASS5	0.182	TUBA4A	0.105
FAM89A	0.328	uX15cu4f_. . .^{e}	0.247	GSTT1	0.182	IDH1	0.102
rh13dQX04. . .^{f}	0.324	C20orf107	0.245	SPATA20	0.182	DDX52	0.100
LANCL2	0.323	VCL	0.242	IGLL1	0.172	ANKRD57	0.094

108 genes selected

GRK5[a]	0.853	C1orf115	0.290	ANO8	0.221	FBLN5	0.157
GPATCH4	0.682	LOC654055	0.287	PTTG1IP	0.219	BLMH	0.156
GNGT2	0.474	RNF214	0.280	3NDg8gVCd. . .[b]	0.218	FCRL3	0.149
PDGFD[c]	0.467	SULT1A1	0.278	USF1	0.216	TDRD9	0.143
FAM24B	0.457	ZNF365	0.271	ECCIP	0.210	ACY1	0.142
PTPRN2	0.442	USE1	0.267	MGC29506	0.209	ZFP57	0.142
CBLB	0.440	DNMT3A	0.267	GRK5[a]	0.207	SLIC1	0.138
PDCL	0.410	LOC649210	0.266	WTIP	0.205	PICK1	0.135
RASA2	0.380	CNTNAP2	0.265	ECL10	0.204	RTN4IP1	0.134
C11orf48	0.376	IL2RA	0.265	DLGAP2	0.200	CDCA7L	0.132
TCEB1	0.374	CCT5	0.264	HRAS	0.199	BEX4	0.131
CAPN3	0.354	R3HDM1	0.263	RAD1	0.189	FCAR	0.130
STK19	0.351	MRPL43	0.260	PRKCE	0.187	ANKRD35	0.111
GUCY1A3	0.348	SLC38A1	0.256	UBAP2L	0.186	USP39	0.109
ZDHHC11	0.345	GNG8	0.255	BPI	0.186	KIAA0495	0.106
SULT1A3	0.336	PLA2G4C	0.251	DTX1	0.184	BRI3BP	0.106
Z6FIQGkeo. . .[d]	0.335	TCF4	0.248	LASS5	0.182	TUBA4A	0.105
FAM89A	0.328	11X15c[illegible]4f. . .[e]	0.247	GSTT1	0.182	IDH1	0.102
rh13dQXo4. . .[f]	0.324	C20orf107	0.245	SPATA20	0.182	DDX52	0.100
LANCL2	0.323	VCL	0.242	IGLL1	0.172	[illegible]	0.094

Low selection frequencies: unstable signatures

Please turn to page 22 in the required reading



Some genes tend to be selected together

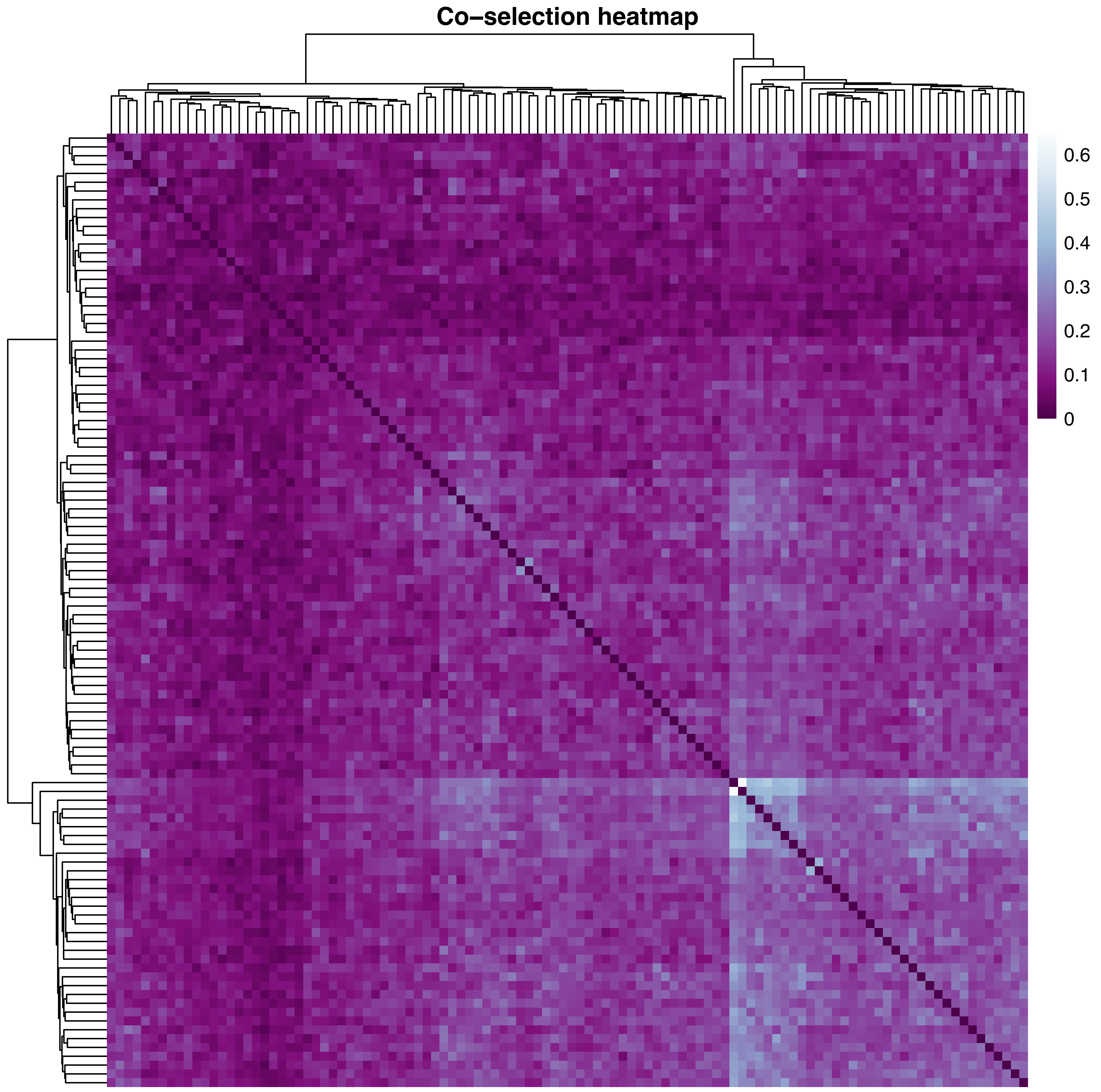


Table 2.3: Genes that tend to be selected together, ordered alphabetically.

- ADIPOR2
- C11orf48
- C1orf115
- CAPN3
- CBLB
- DNMT3A
- FAM24B
- FAM89A
- GNG8
- GNGT2
- GPATCH4
- GRK5
- GUCY1A3
- ITM2C
- LANCL2
- LOC647460
- LOC654055
- PDCL
- PDGFD
- PDGFD
- PRPSAP2
- PTPRN2
- R3HDM1
- RASA2
- rh13dQXo. . .
- SERPINE2
- STK19
- SULT1A1
- SULT1A3
- TCEB1
- TCF4
- WEE1
- Z6FIQGkeo. . .
- ZDHHC11
- ZNF365

https://commons.wikimedia.org/wiki/File:Biologist_Victoria_Achkasova_20150529.jpg

At this point maybe call a biologist

Thesis: Small data require particular care

Thesis: Small data require particular care

- 1000s of measurements, maybe 100 observations

Thesis: Small data require particular care

- 1000s of measurements, maybe 100 observations
- Validation matters more than you think

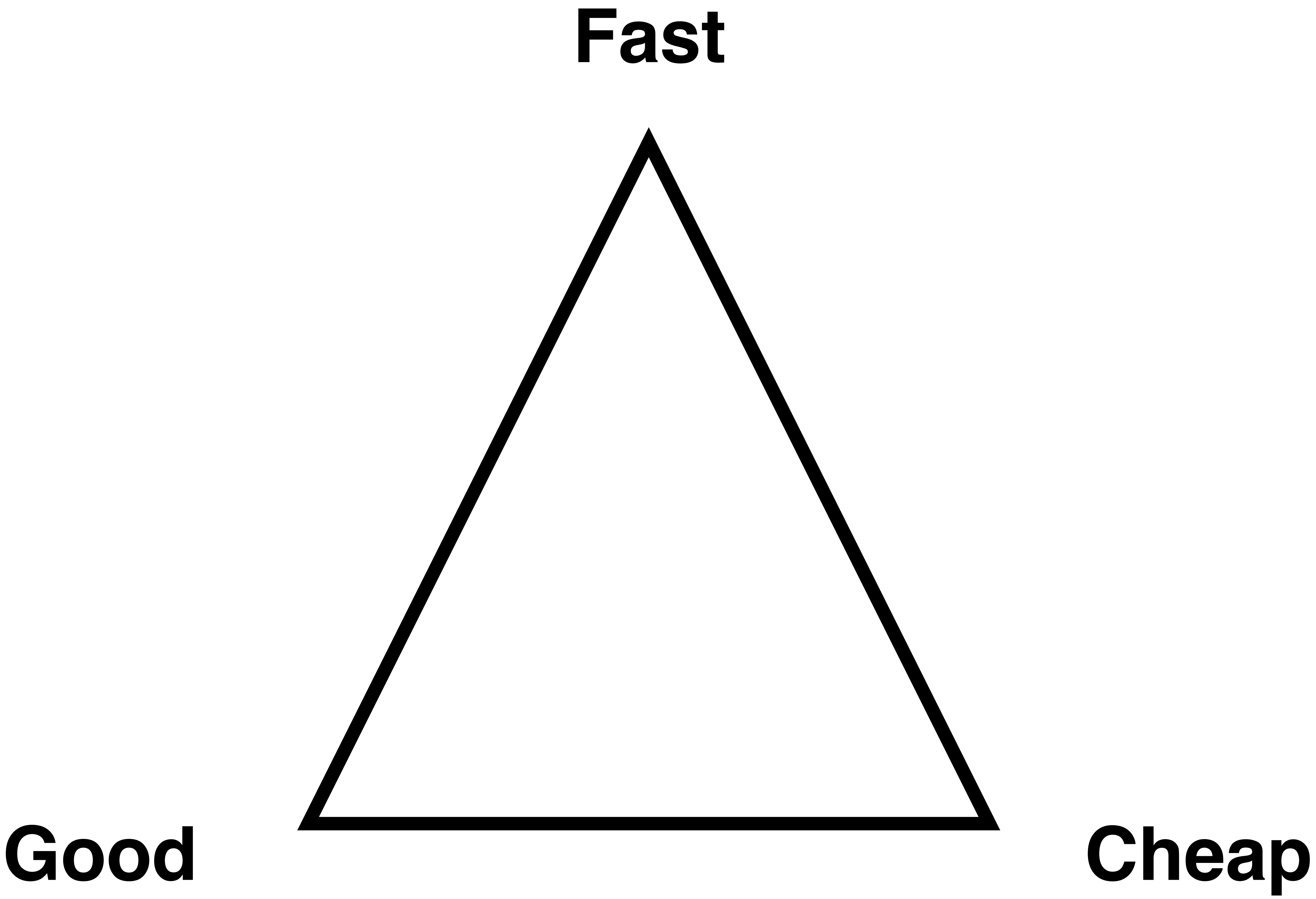
Thesis: Small data require particular care

- 1000s of measurements, maybe 100 observations
- Validation matters more than you think
- Model search difficult

Thesis: Small data require particular care

- 1000s of measurements, maybe 100 observations
- Validation matters more than you think
- Model search difficult
- I suggest to make more assumptions

Thesis: Small data require particular care

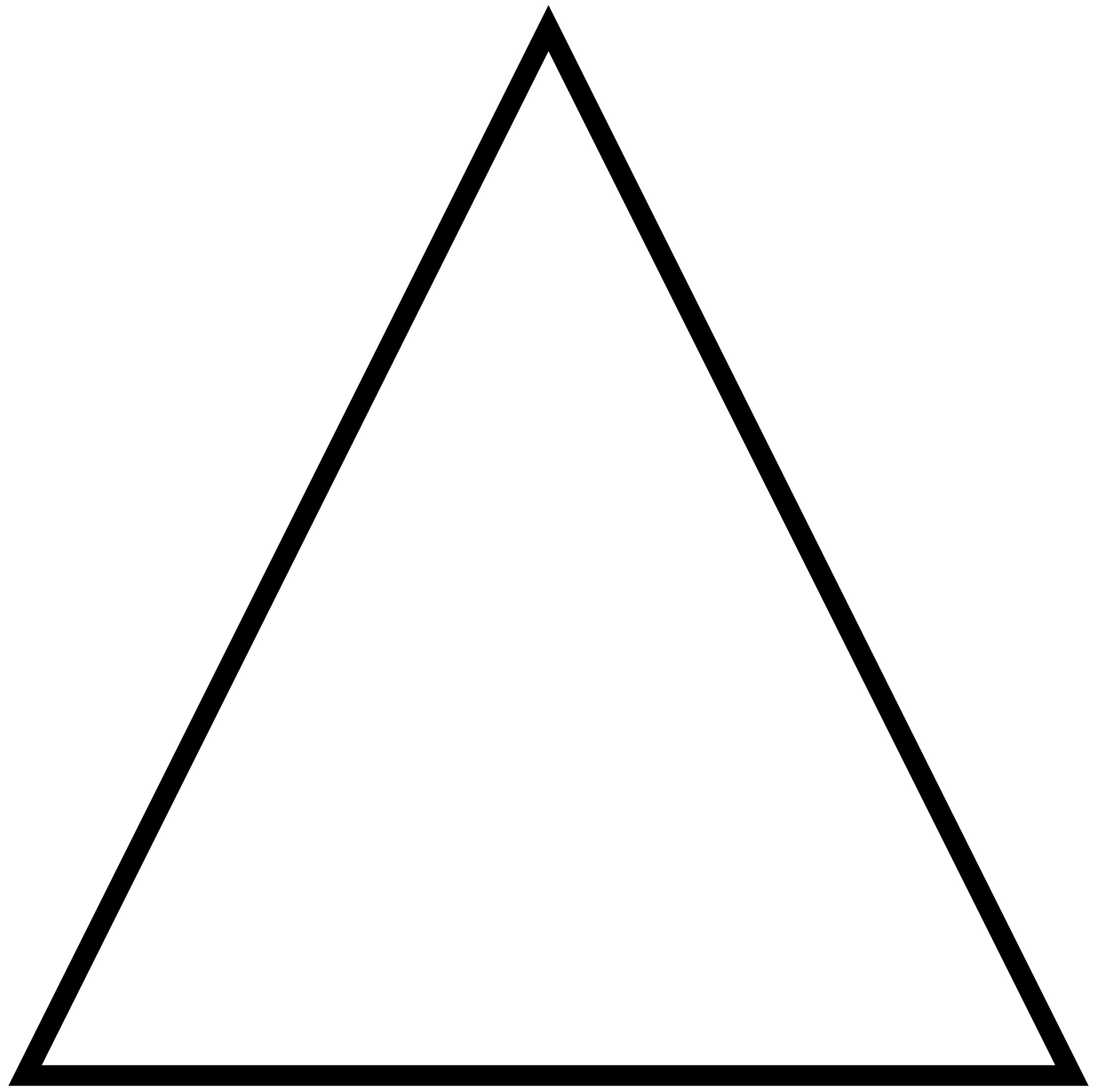


Fast

Good

Cheap

Thesis: Small data require particular care

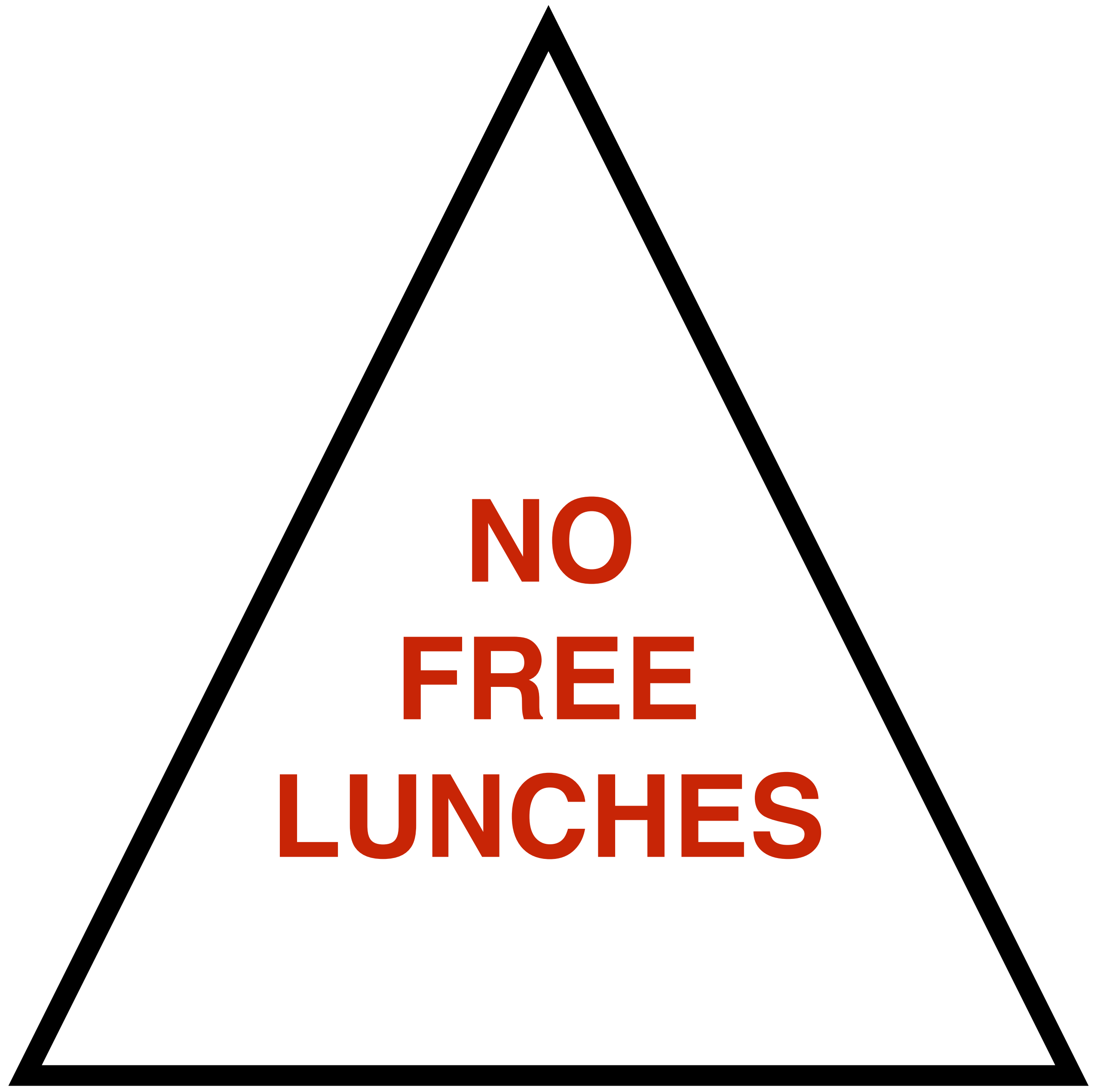


Agnostic modeling

Careful validation

Small data

Thesis: Small data require particular care



Agnostic modeling

NO FREE LUNCHES

Careful validation

Small data

> Naturally, we all desire an adequate assessment of both the indications and their uncertainties, but we shouldn’t refuse good cake only because we can’t have frosting too.

Mosteller & Tukey’s green book

Closing curtain: Thank you.