Lecture 1C – Model comparison

BAMB! 2024 Summer School

The plan for the next 120 minutes

- Model selection

⇒What makes a good model?

⇒Penalized likelihoods

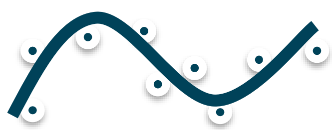
⇒Held-out data & Crossvalidation

- Model recovery (with confusion matrix)

A dual perspective on model evaluation

Capturing some *qualitative* properties of the data revealed by model-free analysis (e.g. psychometric curve)

Absolute criterion



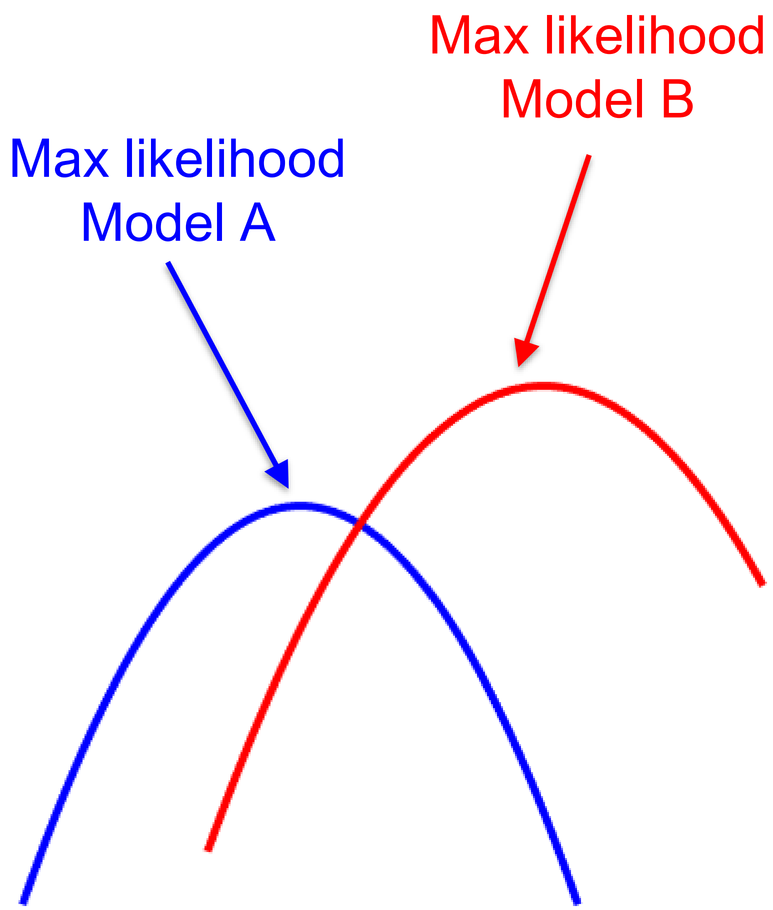
Outperforming other models on *quantitative* measures for how well the model explains the data (~ null hypothesis testing)

Relative criterion

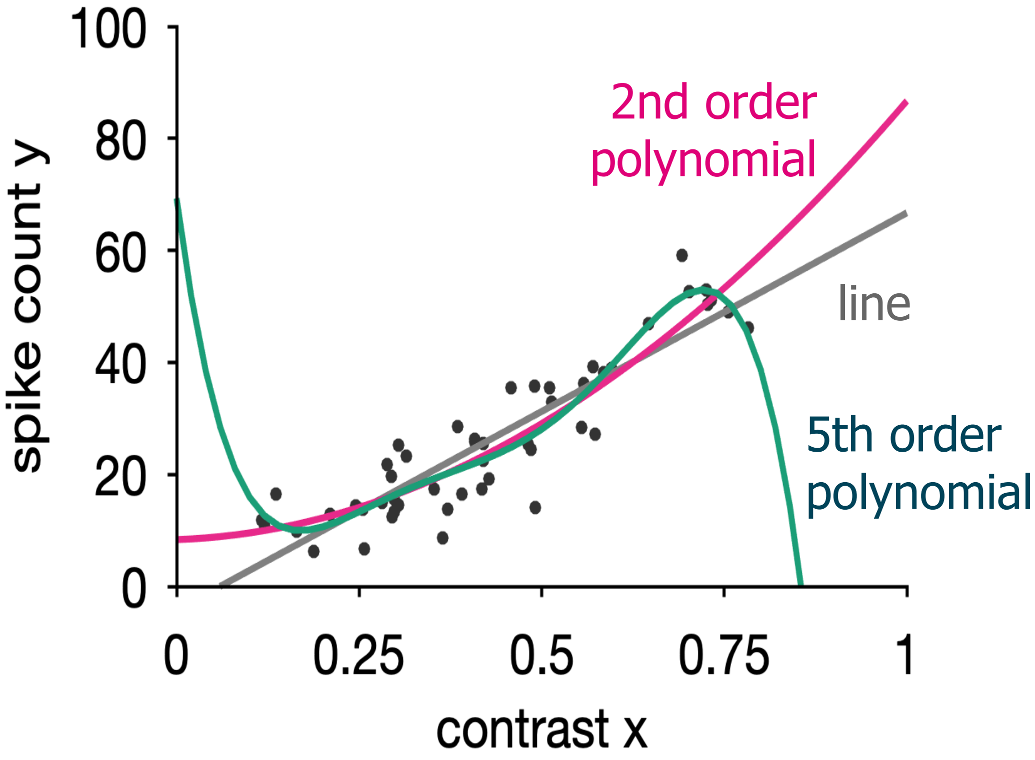
We want our model to pass **both** tests

What metric for model selection?

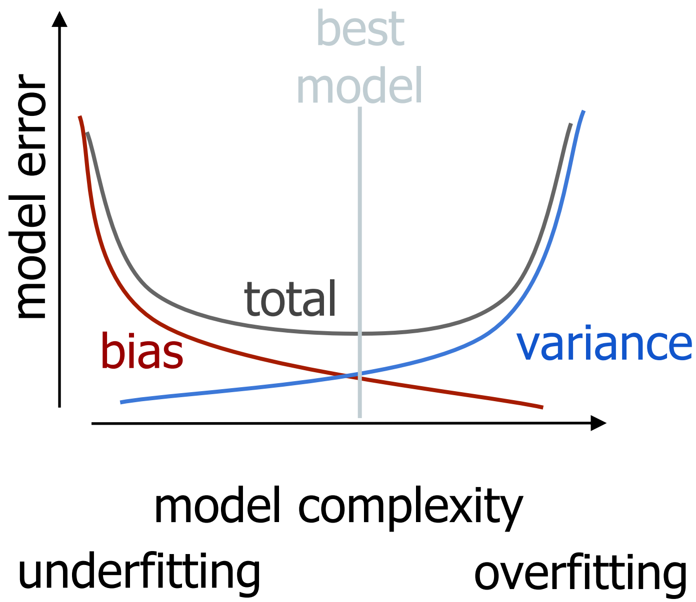
- We use likelihood to select best parameters *within* each model, so why not use likelihood to select *between* models?
- More parameters -> More flexibility. So comparison based on maximum likelihood is unfair.

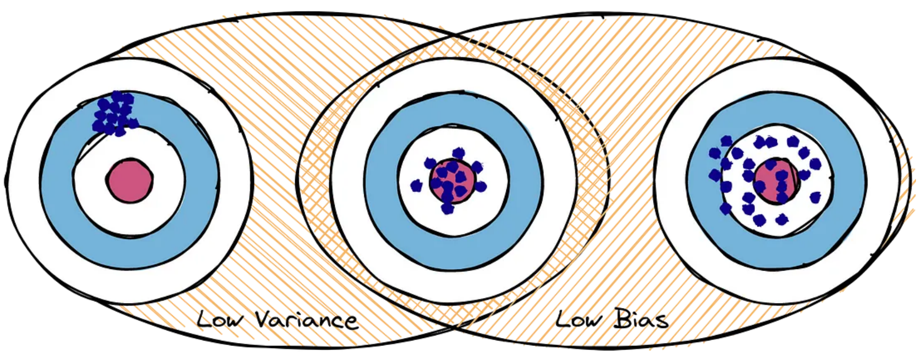


Comparing models



Bias-variance trade-off





Bias:
systematic deviation from structure underlying data
(high bias = underfitting)

Variance
Variability beyond the structure underlying data
(high variance = overfitting)

Total error = bias + variance

Best model 🏆
balances bias and variance

Taken from Ivan Reznikov @ Medium

Two alternative approaches to compensate for model complexity

(1) **Penalized likelihoods:** use maximum-likelihood and correct for the extra flexibility of complex models by penalizing the number of free parameters

(2) **Cross-validation:** fit on one part of the data and see how good the model is at predicting data that we have not used for fitting

Penalized likelihoods

Akaike Information Criterion

$$\mathrm{AIC} = -2 \log p(\boldsymbol{Y} | \boldsymbol{X}, \hat{\theta}) + 2k$$

Bayesian Information Criterion

$$\mathrm{BIC} = -2 \log p(\boldsymbol{Y} | \boldsymbol{X}, \hat{\theta}) + k \log n$$

maximum log-likelihood

penalty term

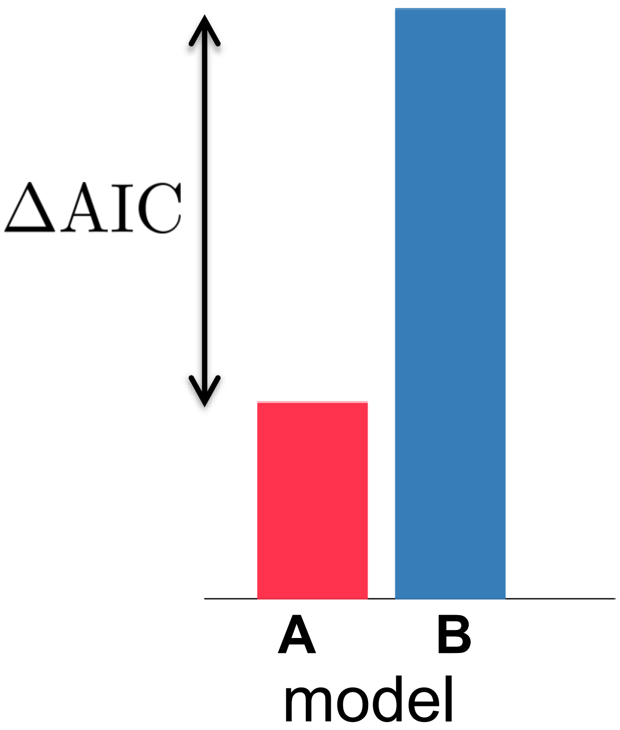
k: number of parameters

n: number of trials

- smaller = better
- easy to compute and transparent in how they quantify the trade-off between parsimony and goodness of fit. BIC is more conservative.
- relatively easy to compute assuming known estimates of parameters
- each criterion is correct for certain assumptions ONLY and they are widely employed (despite assumptions not always met in practice..)

Vandekerckhove et al. (2015)

Quality of evidence



$\Delta\mathrm{AIC}$	**Quality of evidence**
0 to 2	weak
2 to 6	positive
6 to 10	strong
>10	very strong

Raftery 1995

Other metrics

Likelihood ratio (for nested models)

$D = -2\log \frac{\text{max likelihood null model}}{\text{max likelihood alternative model}}$ follows a χ-distribution if null model is true

Bayes Factor

Approximation to model evidence

$$p(\boldsymbol{Y}|\boldsymbol{X}, \text{model A}) = \int_{\theta_A} p(\boldsymbol{Y}|\boldsymbol{X}, \theta_A)d\theta_A$$

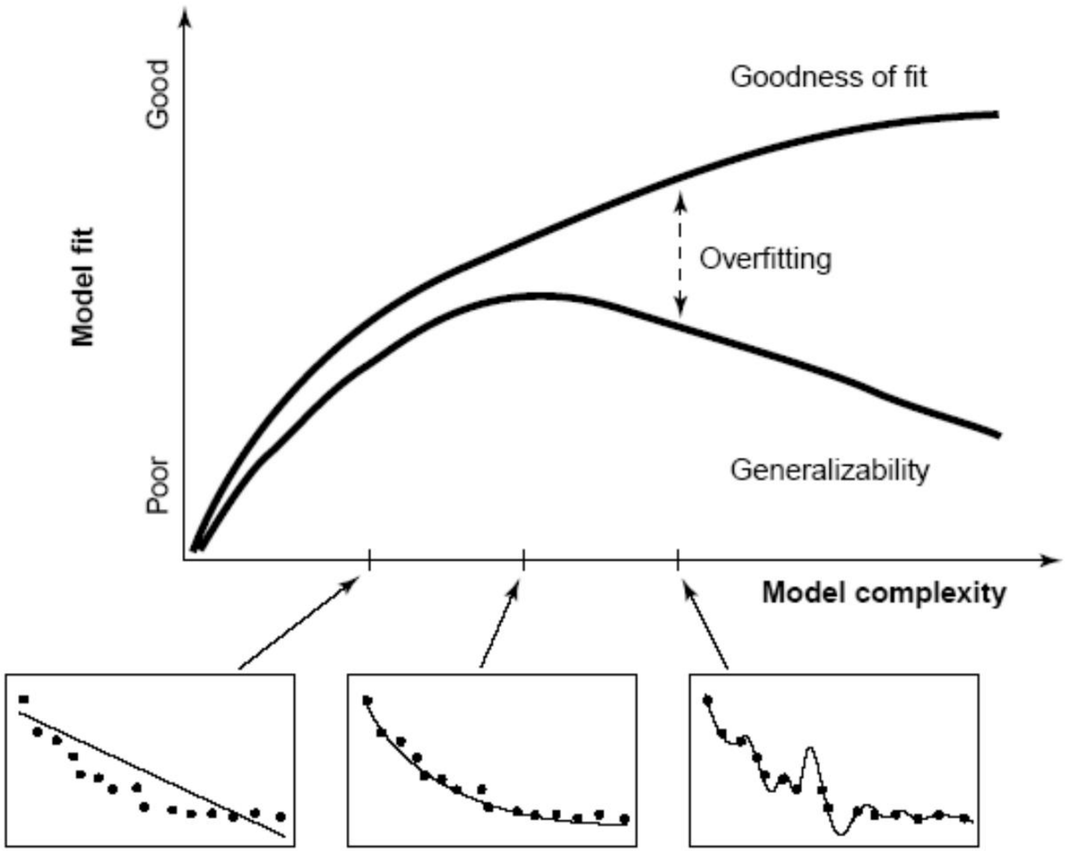
- **Laplace approximation** takes into account curvature of likelihood at MLE (BIC is derived from this)
- **Variational Bayes** (VB)

  negative free energy as lower bound approximation to the log evidence (“evidence lower bound” (ELBO))

– Sampling-based methods (DIC)

Cross-validation: testing generalizability of model fit

Which model represents the best trade-off between model fit and model complexity? Pitt & Miyung (2002) *TICS*



Testing on held-out data

- The preferred model is the one which **best predicts unseen data** from the same source

- Validation methods divide the observed data in a **training set** and a **test set** (there are many ways in which this can be done)

- We fit on one part of the data and see how good the model is at predicting data that we have not used for fitting

👍 minimal assumptions required about your data

👎 computationally expensive

Testing on held out data

Person presents at ER with a headache
Your goal is to predict whether person needs urgent further examination

Fever	Neck stiffness	Abrupt onset	Age	Urgent further exam
38	Yes	No	42	No
38	No	Yes	51	Yes
...	...	...	...	...

Then we see a new patient 🥴 ?

Testing on held out data

Person presents at ER with a headache
Your goal is to predict whether person needs urgent further examination

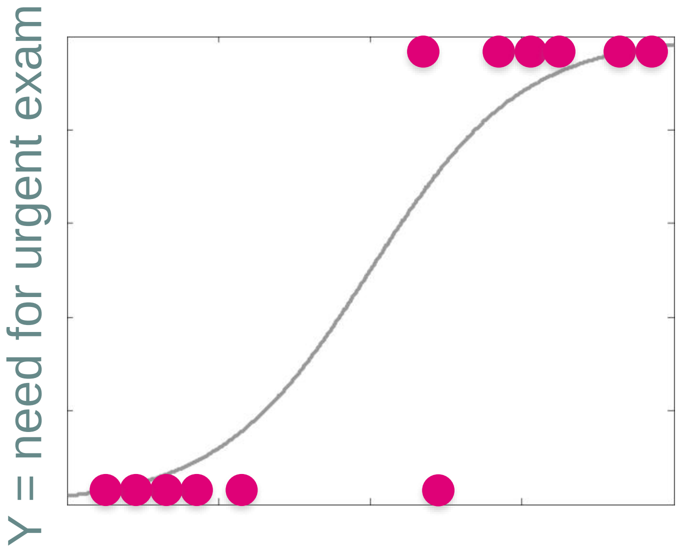
Fever	Neck stiffness	Abrupt onset	Age	Urgent further exam
38	Yes	No	42	No
38	No	Yes	51	Yes
...	...	...	...	...
Then we see a new patient 🥴				?

We want to use the variables: fever, etc to predict Y = urgent exam needed or not

Testing on held out data

We seek a model to relate [pink box] to [grey box]

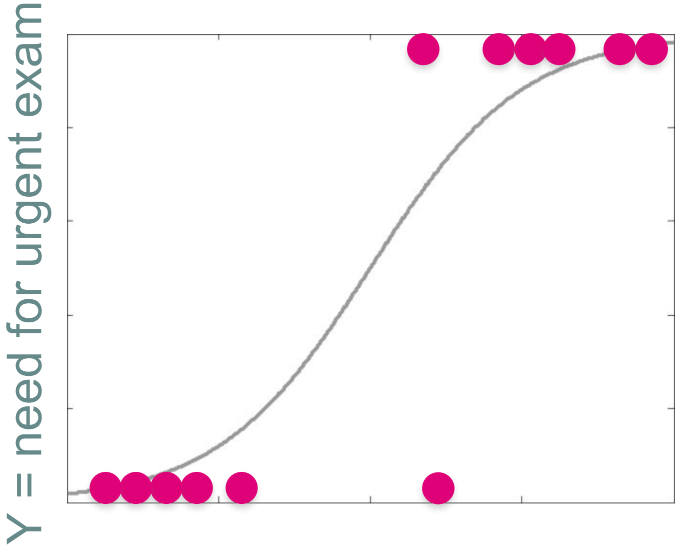
e.g. a logistic regression Y = Fever + Age + ...



Testing on held out data

We seek a model to relate  to 

e.g. a logistic regression Y = Fever + Age + ...



Cross-validation allows us to compare different models/methods and get a sense of how well they work in practice

Testing on held out data

This is all your data collected about people who needed or not needed urgent further exam

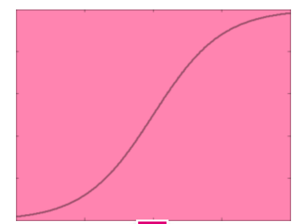
With your data you have to:

- Estimate model parameters (weights) of variables Fever, Age, etc. for the regression

$\Rightarrow$ i. e. TRAINING

- Evaluate how well your model (regression) works

$\Rightarrow$ i. e. TESTING



Does the obtained curve do a good job in categorizing new data?

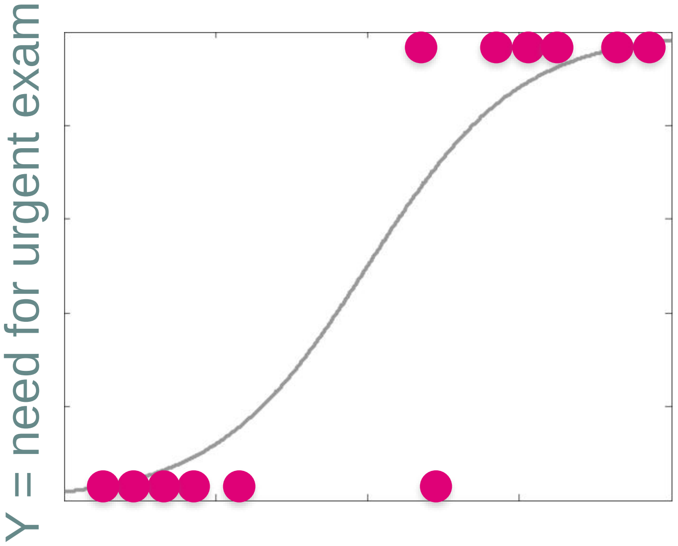
Testing on held out data

$\Rightarrow$ i. e. TRAINING

$\Rightarrow$ i. e. TESTING

Testing on held out data

⛔ A bad approach would be to use ALL our data to achieve this and estimate the parameters (slope):

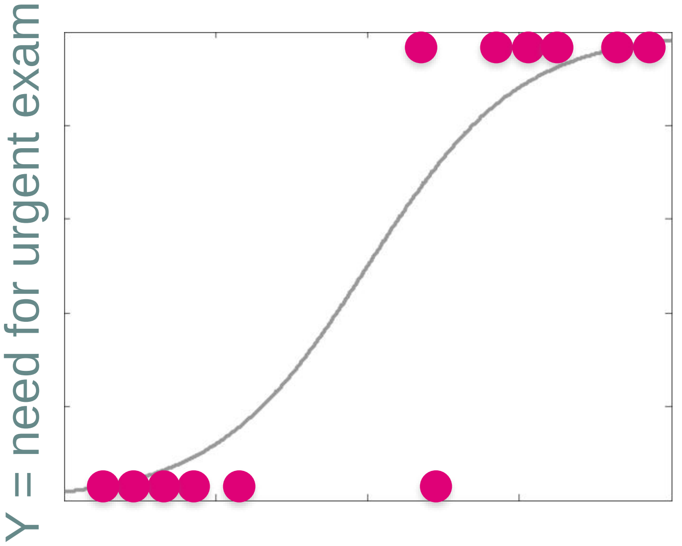


⇒ i. e. TRAINING

⇒ i. e. TESTING

Testing on held out data

⛔ A bad approach would be to use ALL our data to achieve this and estimate the parameters (slope):

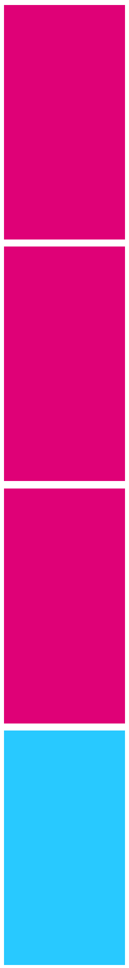


⇒ i. e. TRAINING

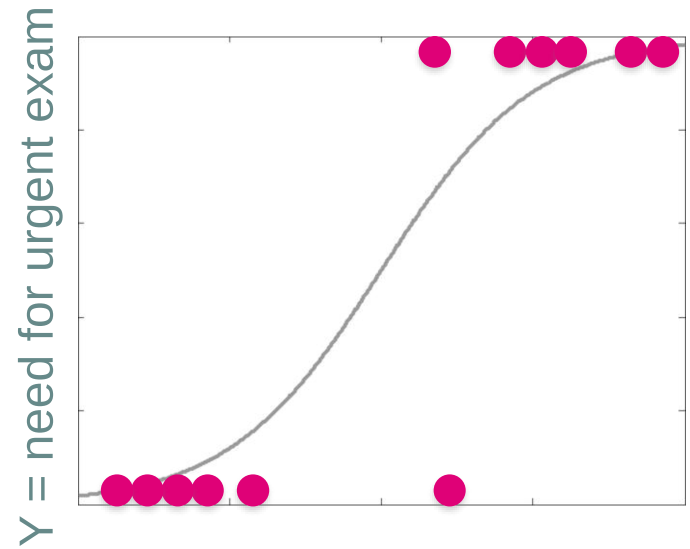
⇒ i. e. TESTING

Then you would have no data left to test your model 😰

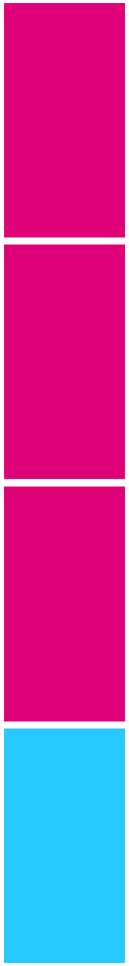
Testing on held out data



A slightly better approach would be to use 75% of your data to achieve the training and estimate the parameters (slope):

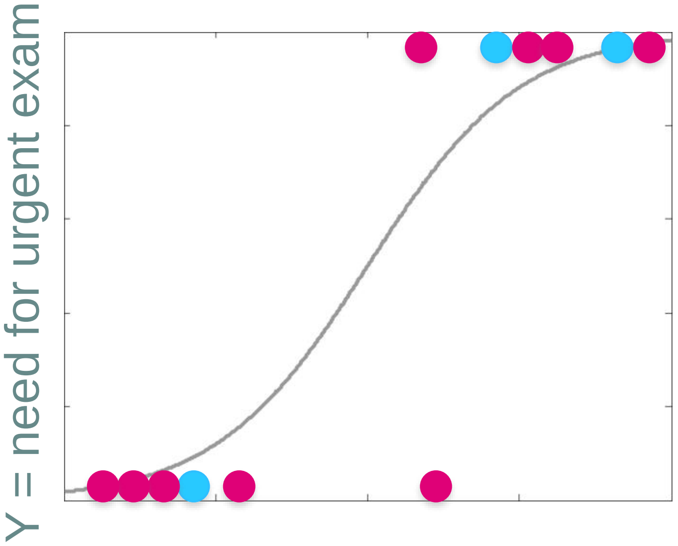


Testing on held out data

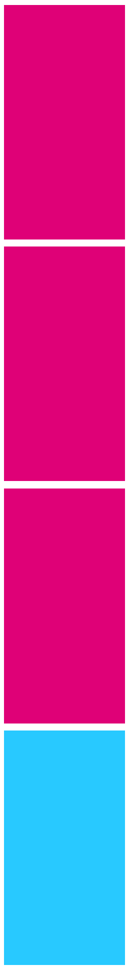


💡 A slightly better approach would be to use 75% of your data to achieve the training and estimate the parameters (slope):

💡 And the last 25% for testing:

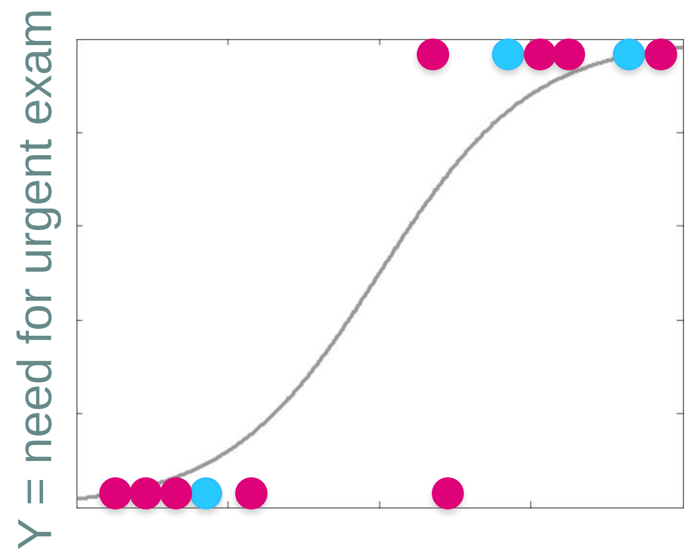


Testing on held out data



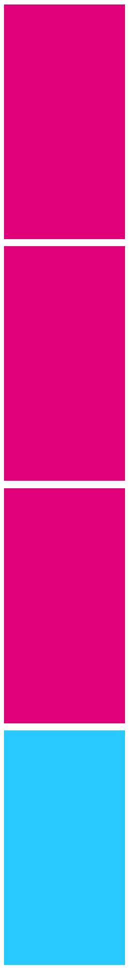
💡 A slightly better approach would be to use 75% of your data to achieve the training and estimate the parameters (slope):

💡 And the last 25% for testing:



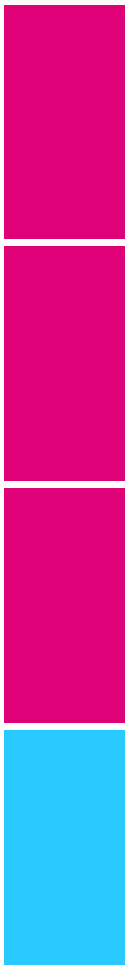
✅ We can then compare models by examining how well each one categorises the test data

Cross validation

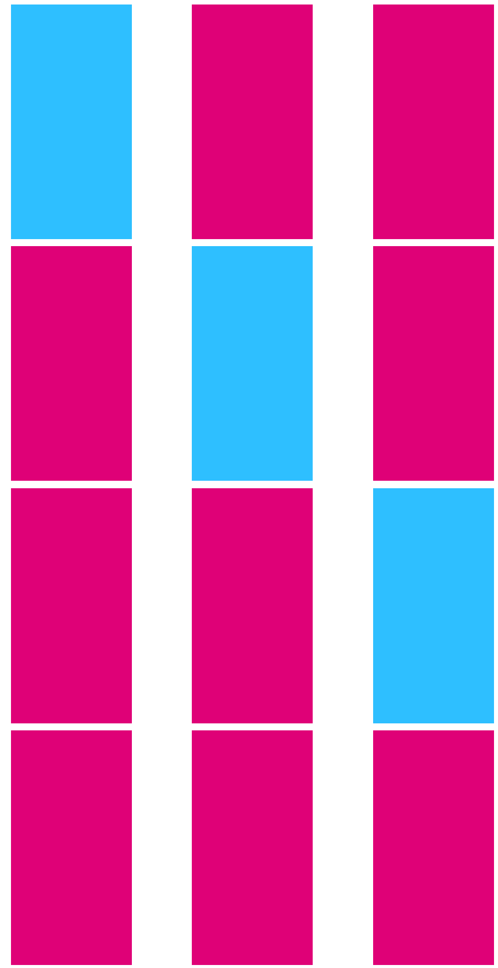


But how do we make sure results do not depend on which data is used for training for testing?

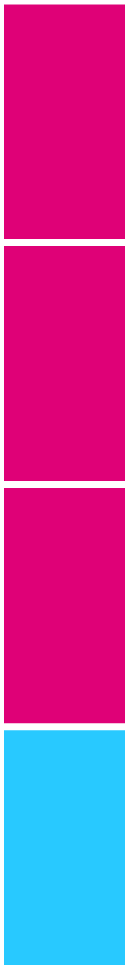
Cross validation



But how do we make sure results do not depend on which data is used for training for testing?

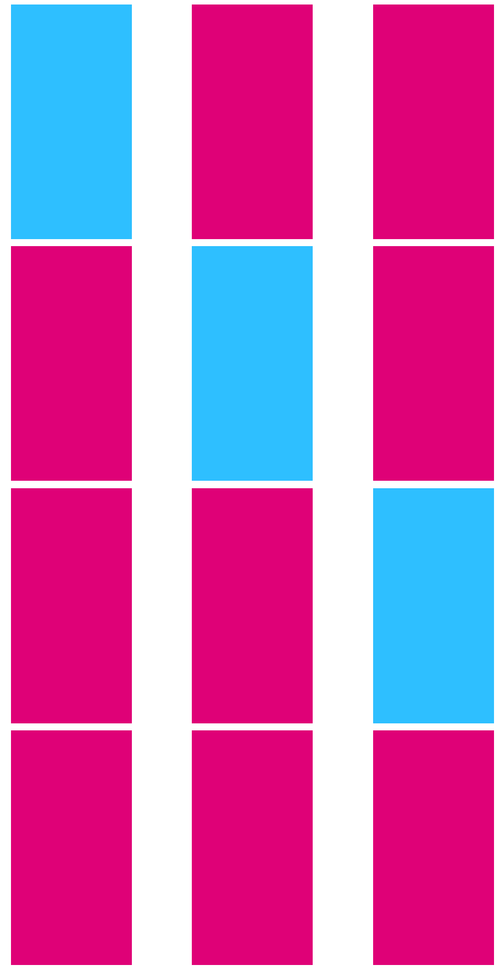


Cross validation

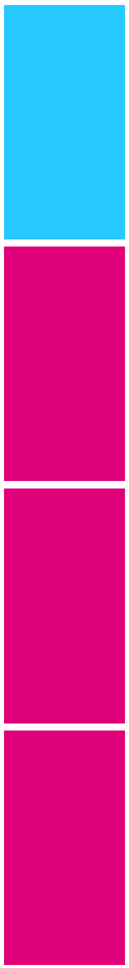


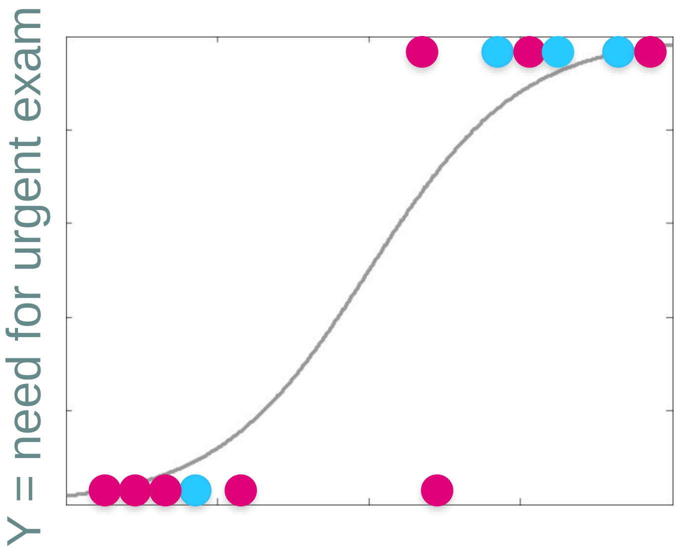
But how do we make sure results do not depend on which data is used for training for testing?

Cross-validation uses of all these blocks, one at a time, making sure each data point appears once and only once in the test set



Cross validation



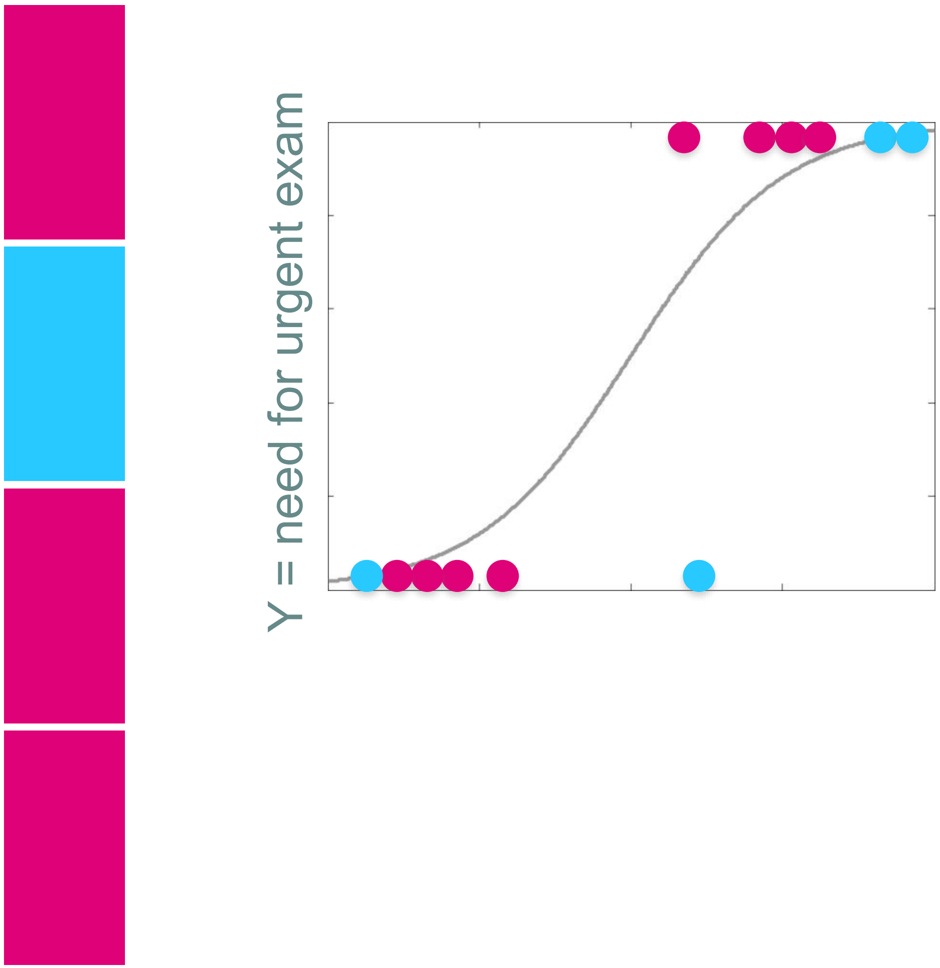


You keep track of how well the model does:

Test data categorization

Correct ✔	Incorrect ✘
3	1

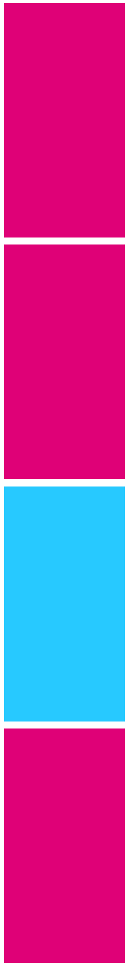
Cross validation

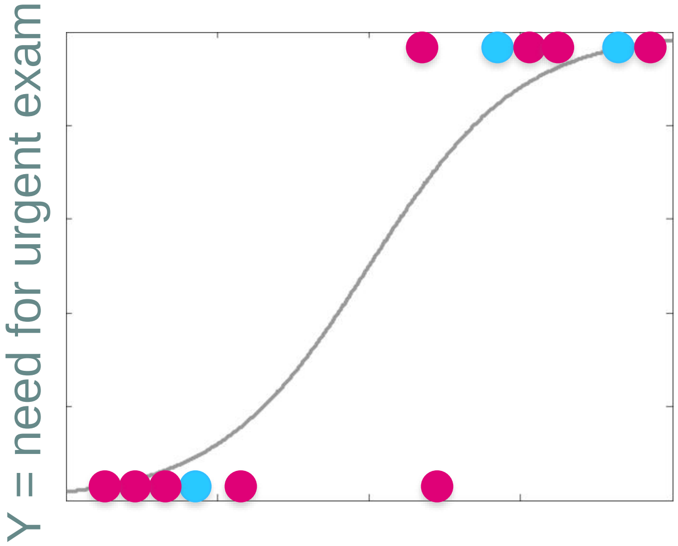


You keep track of how well the model does:

Test data categorization	
Correct ✔	Incorrect ✘
4	0

Cross validation





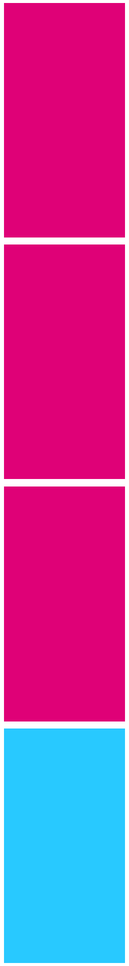
You keep track of how well the model does:

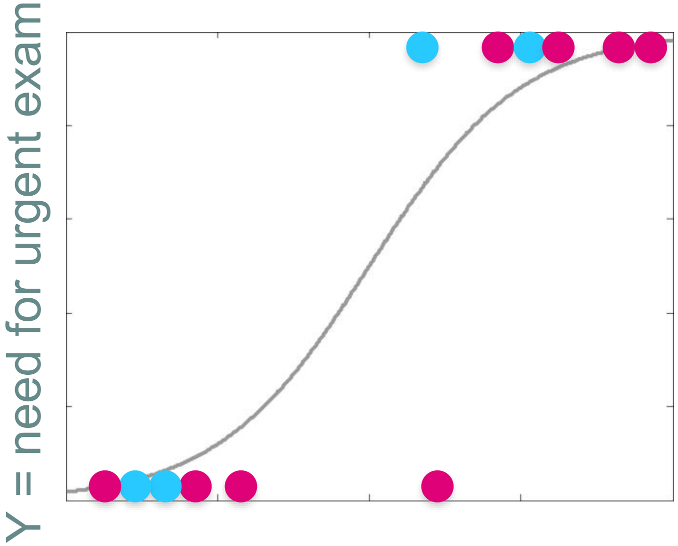
Test data categorization

Correct ✔ 2

Incorrect ✘ 2

Cross validation





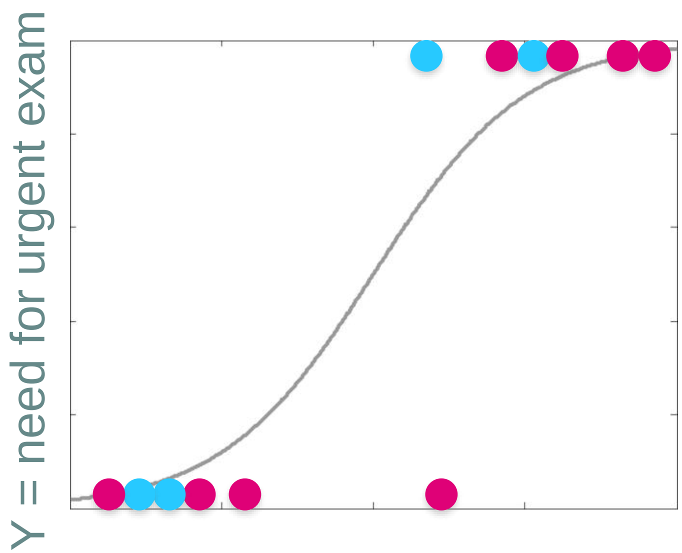
You keep track of how well the model does:

Test data categorization

Correct ✔	Incorrect ✖
3	1

Cross validation

In the end, every block of data has been used for testing



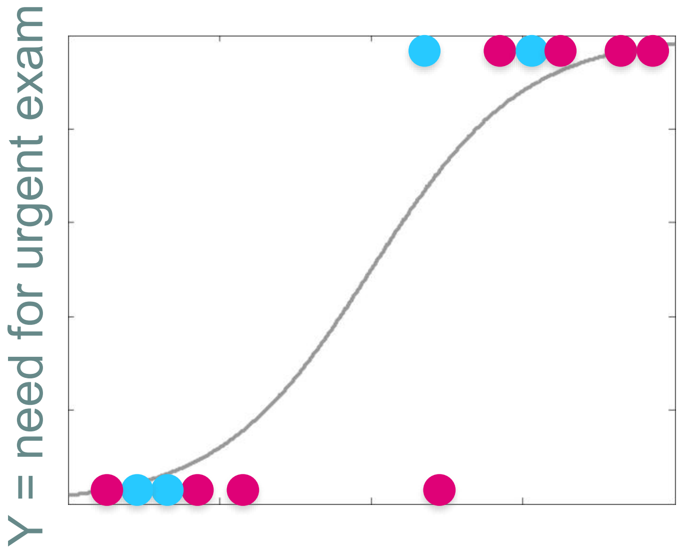
Hence you know how well the model does OVERALL:

Test data categorization

Correct ✔	Incorrect ✘
12	4

Cross validation

ModelA



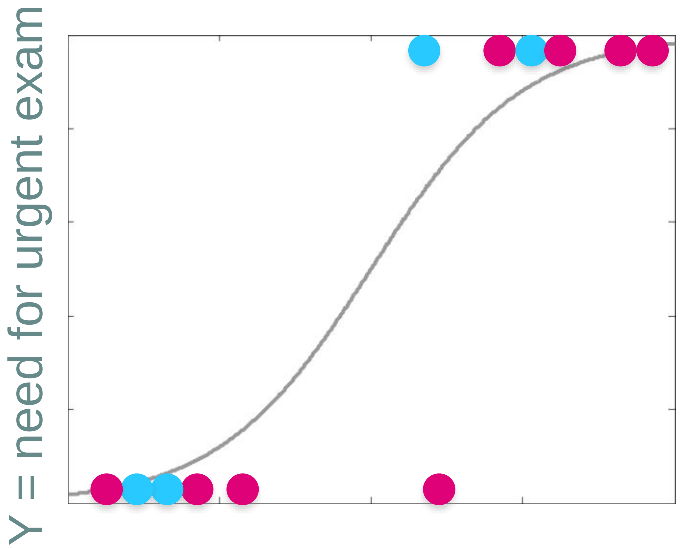
Hence you know how well the model does OVERALL:

Test data categorization

Correct ✔	Incorrect ✘
12	4

Cross validation

ModelA



Hence you know how well the model does OVERALL:

Test data categorization

Correct ✔	Incorrect ✘
12	4

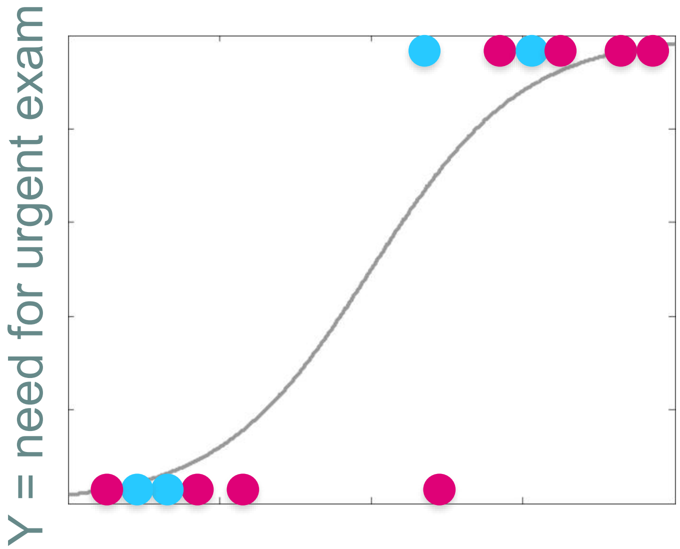
ModelB

Test data categorization

Correct ✔	Incorrect ✘
10	6

Cross validation

ModelA



Hence you know how well the model does OVERALL:

Test data categorization

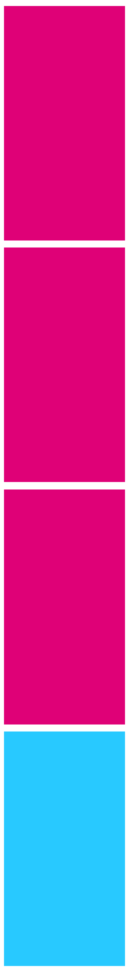
Correct ✔	Incorrect ✘
12	4

ModelB
ModelC
ModelD
ModelE

Test data categorization

Correct ✔	Incorrect ✘
10	6

Setting up the number of folds



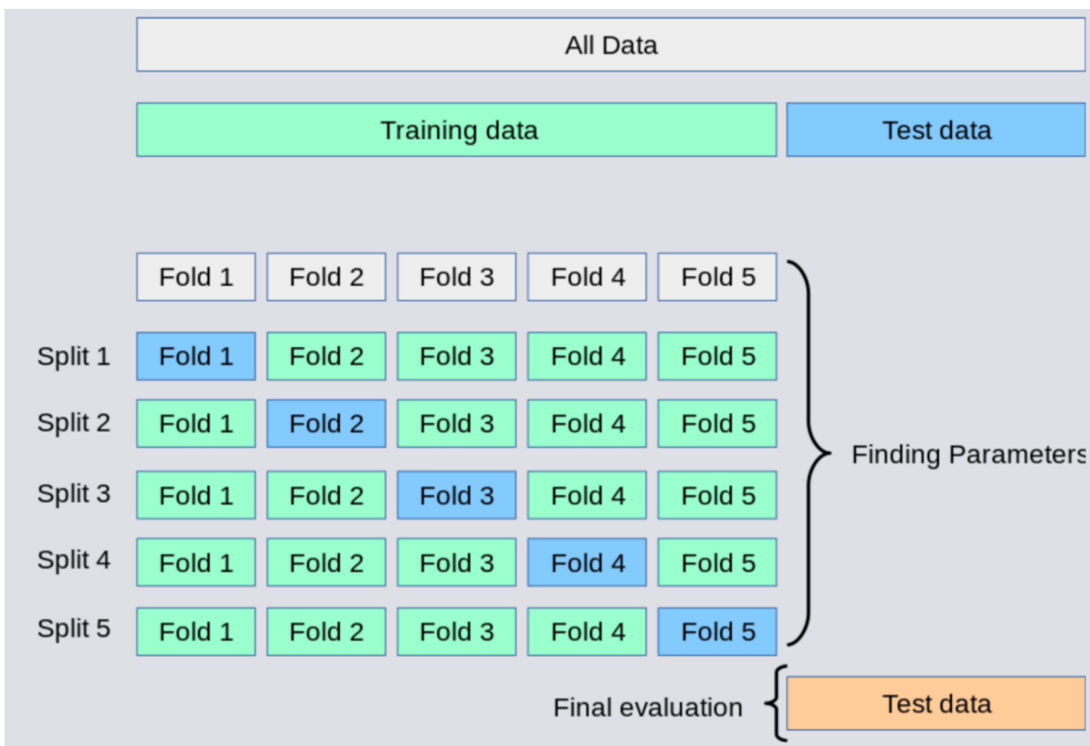
Number of divisions is arbitrary: larger numbers provide more robust results but are more computationally expensive 🤑

Ten-fold cross-validation is common

Extreme case: **leave-one-out** cross-validation (LOOCV): number of folds = number of data points (i.e. trials)
We fit the model on all trials but one and then test on this one trial, then repeat it for each trials separately.

Cross-validation with hyperparameters

1. We cannot use simple CV to select hyperparameters AND models at the same time
2. Usually: use some training data to select hyperparameters (using CV), then test different models on some held-out test data.
3. Even better: use *nested crossvalidation,* which rotates through inner and outer folds (a lot of work 💪)



- Model selection

⇒Goodness of fit: what makes a good model?

⇒Quantitative criteria (AIC, BIC, Bayes factor, elbo)

- Cross-validation

⇒For model comparison and generalization

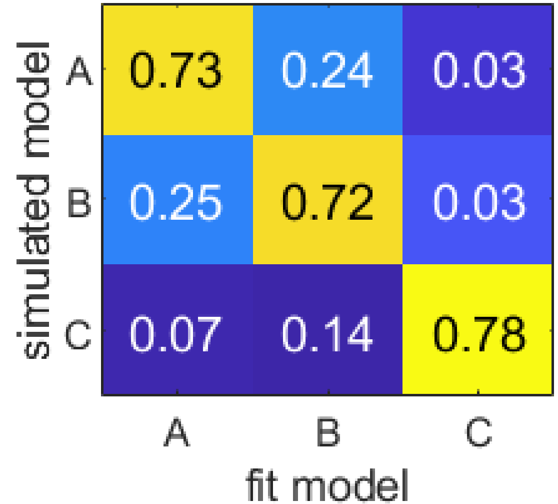
- Model recovery (with confusion matrix)

Model recovery

- ✓ Simulate data for your model space: only a limited number of models can be tested, carefully consider your choice
- ✓ Fit each model to all simulated data sets based on each model
- ✓ Estimate how often true generative model is identified and plot confusion matrix: we want each model to be ***identifiable***, so large values along the diagonal
- ✓ Can be used to **optimize paradigm and analysis pipeline** jointly (find the optimal experimental design)

Confusion matrix

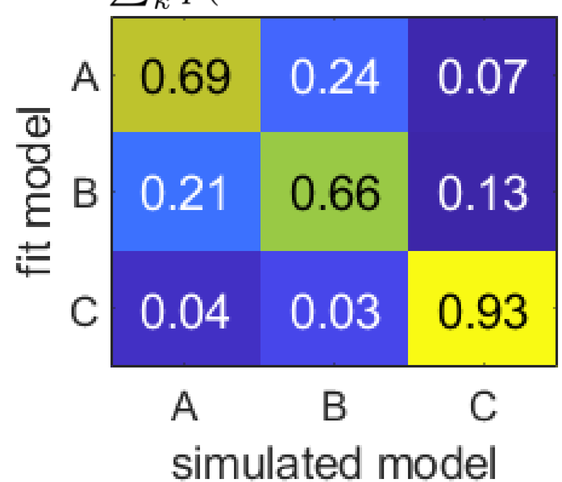
$$M_{ij} = p(\text{fit model}= i|\text{simulated model}= j)$$



Inverstion matrix

$$N_{ij} = p(\text{simulated model}= i|\text{fit model}= j)$$

$$N_{ij} = \frac{p(\text{simulated model}=i)M_{ji}}{\sum_k p(\text{simulated model}=k)M_{jk}}$$



Wilson & Collins (2019) *eLife*

Model recovery

😈 There are a number of choices to be made and the devil is in the details
Under what parameter regime do you perform the model recovery?

- Sample randomly between parameter bounds
- Sample randomly between reasonable parameter bounds
- Re-use best-fitting parameters for each model from participants
- At the boundaries, recovery may fail, but what matters most is that you can recover under the parameter space of relevance for your data

🚀 **What space of models do you choose to explore?**

- Strawman models will be easily set aside
- Models are never true in an absolute sense: identify *the best model among the set of models you have selected to compare*
- Parsimony applies to the model space ➡ carefully examine your hypotheses, keep the number of alternative models small without ignoring any potential hypotheses ⚖️
- If two or more models unidentifiable, you might need a new/better experimental design

Model selection conclusions

•Because of the *bias-variance trade-off*, we cannot compare models just based on their maximum likelihood. We need to compensate for the extra flexibility afforded to more complex models.

• Two alternative approaches: penalized likelihoods (AIC/BIC) and testing on held out data (cross-validation). If parameter fitting is not too costly, prefer cross-validation.

•Never lose the perspective of your modelling goals: you need an objective measure of which model best captures the data, but you want to make sure that this winning model captures the important part of the data. **Can you validate your best-fitting model AND *unvalidate* your alternative models?** (tutorial 1B)

Acknowledgements 🙏

- A. Wu, J. Drugowitsch: Neuromatch Academy

- K. Preuschoff (BAMB! 2019),
  M. Rouault (BAMB! 2023)

- Statquest

Tutorial 1C

- Model selection
- Cross validation
- Model recovery

Brief summary Tutorial 1C

Model selection compares quantitative criteria
such as AIC BIC model evidence etc
Each metric has pros and cons, no perfect recipe

Keep in mind that comparison is relative:
To the space of models that you have defined in the first place

Brief summary Tutorial 1C

Cross validation asks how well the model predicts new data that it hasn't seen yet.

This approach is to use held-out data which we call **testing data** or validation data: we do not fit the model with this data, but we use it to select our best model.

We often have a limited amount of data though (especially in neuroscience), so we do not want to further reduce our potential training data by reassigning some as validation.

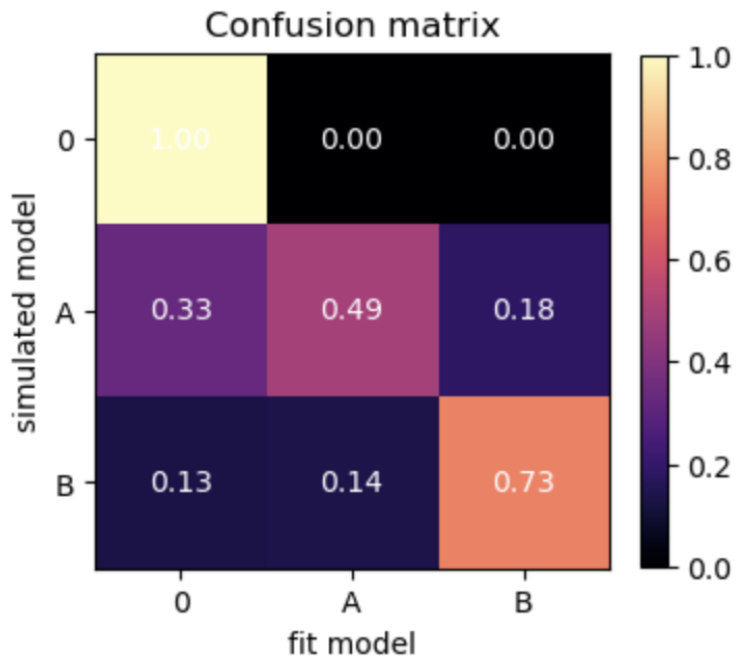
So we can use **k-fold cross-validation**!

- we divide up the training data into k subsets (called folds)
- train our model on the first k-1 folds
- then compute error on the last held-out fold

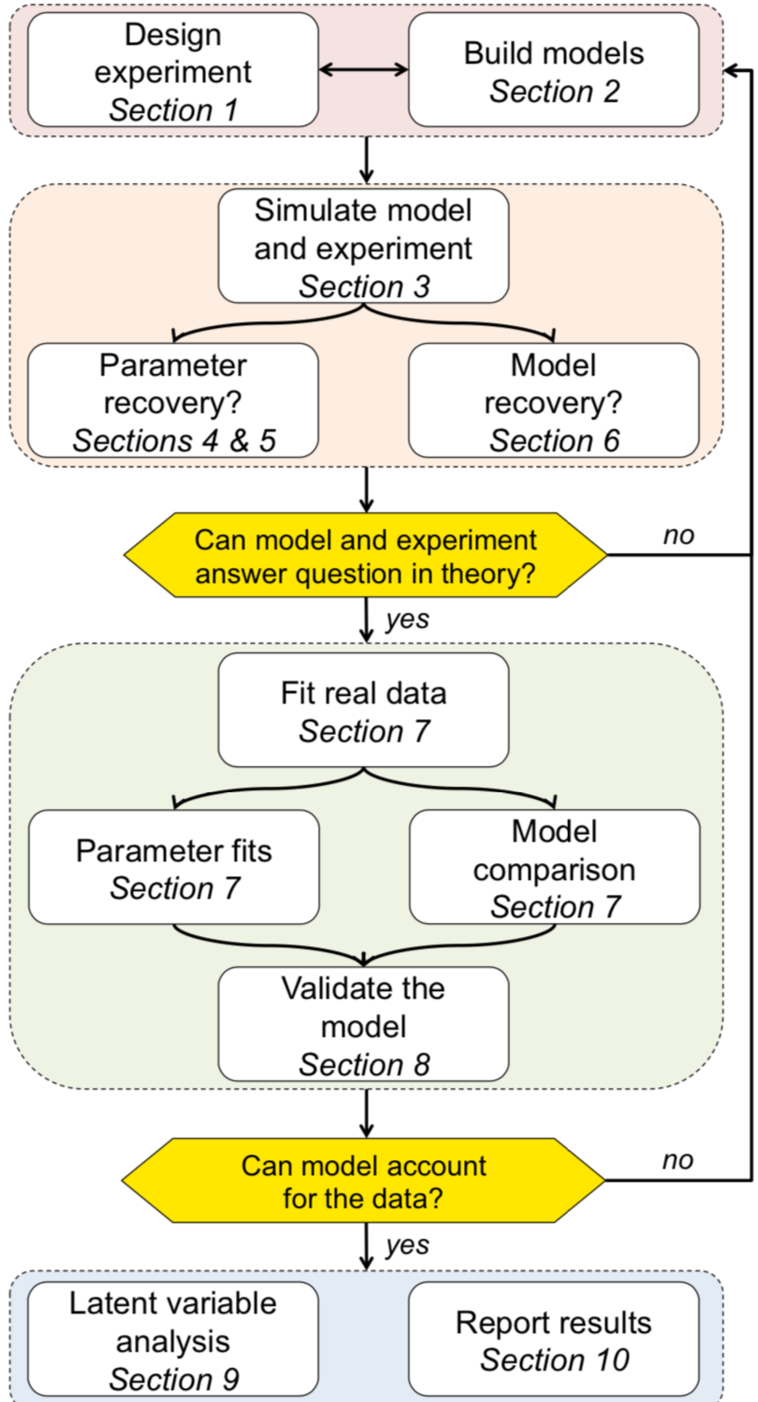
Brief summary Tutorial 1C

Model recovery analysis verifies that your model is **identifiable** from others.

The aim is to build an experimental paradigm that will allow you to identify a model distinctly from alternative accounts



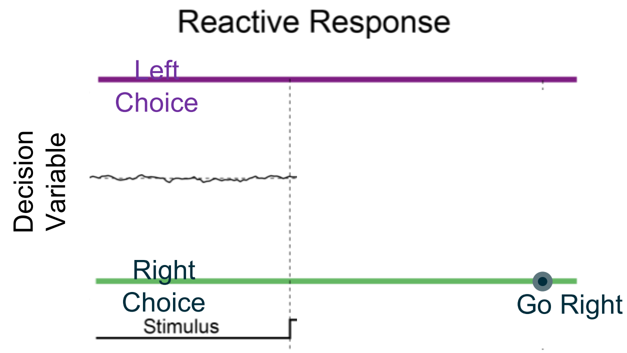
Wrap-up Day 1



Wilson & Collins, 2019 *eLife*

Case study: an alternative model for perceptual decision-making in rats

Decision variable:
accumulation-to-bound

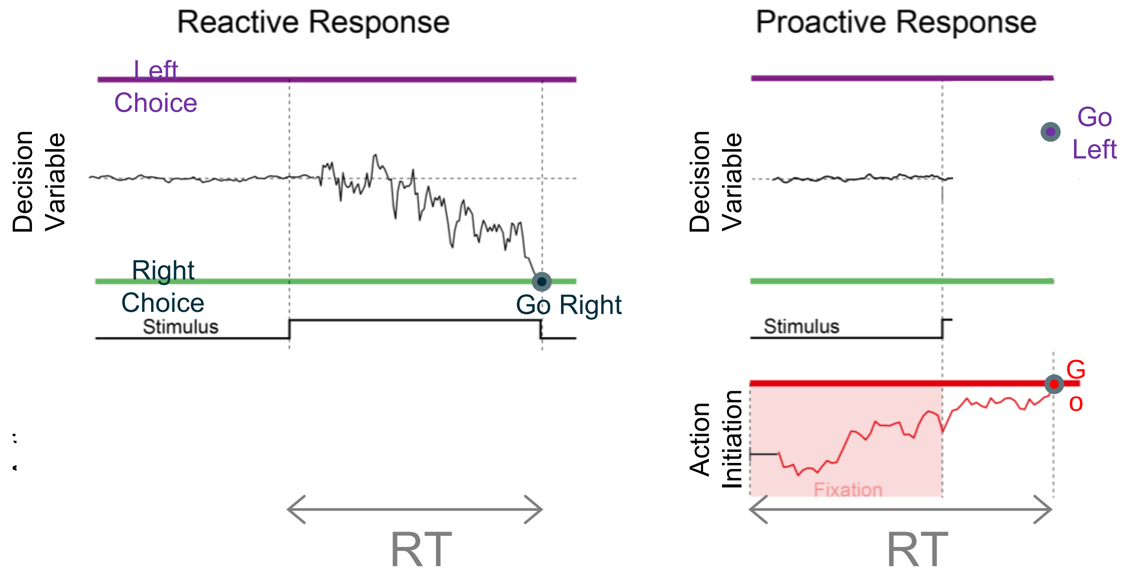


Hernández-Navarro, L., *Nat Comms* (2021).

Case study: an alternative model for perceptual decision-making in rats

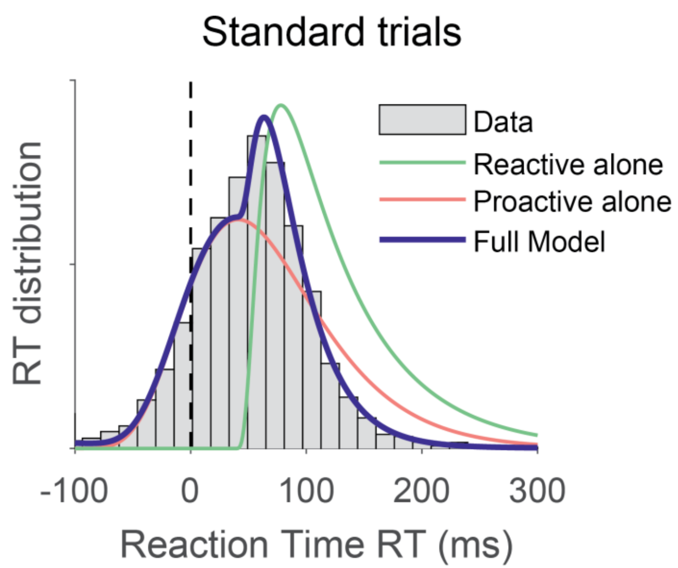
Decision variable:
accumulation-to-bound

Action initiation:
urgency signal



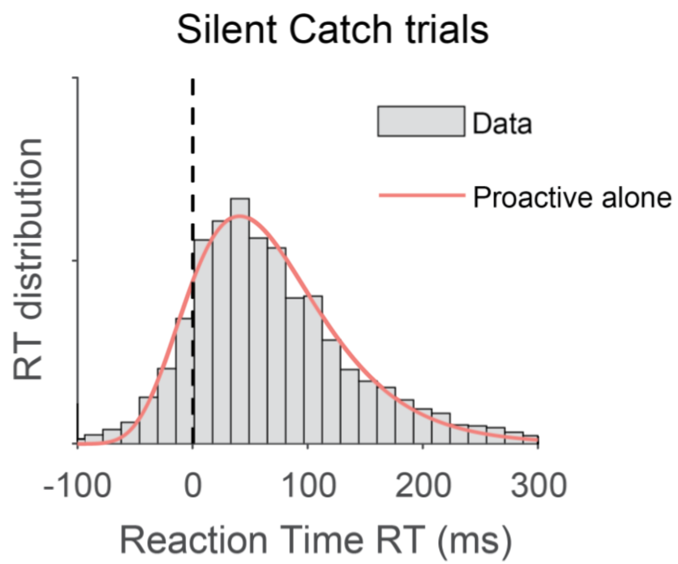
Hernández-Navarro, L., *Nat Comms* (2021).

Case study: model validation



Hernández-Navarro, L., *Nat Comms* (2021).

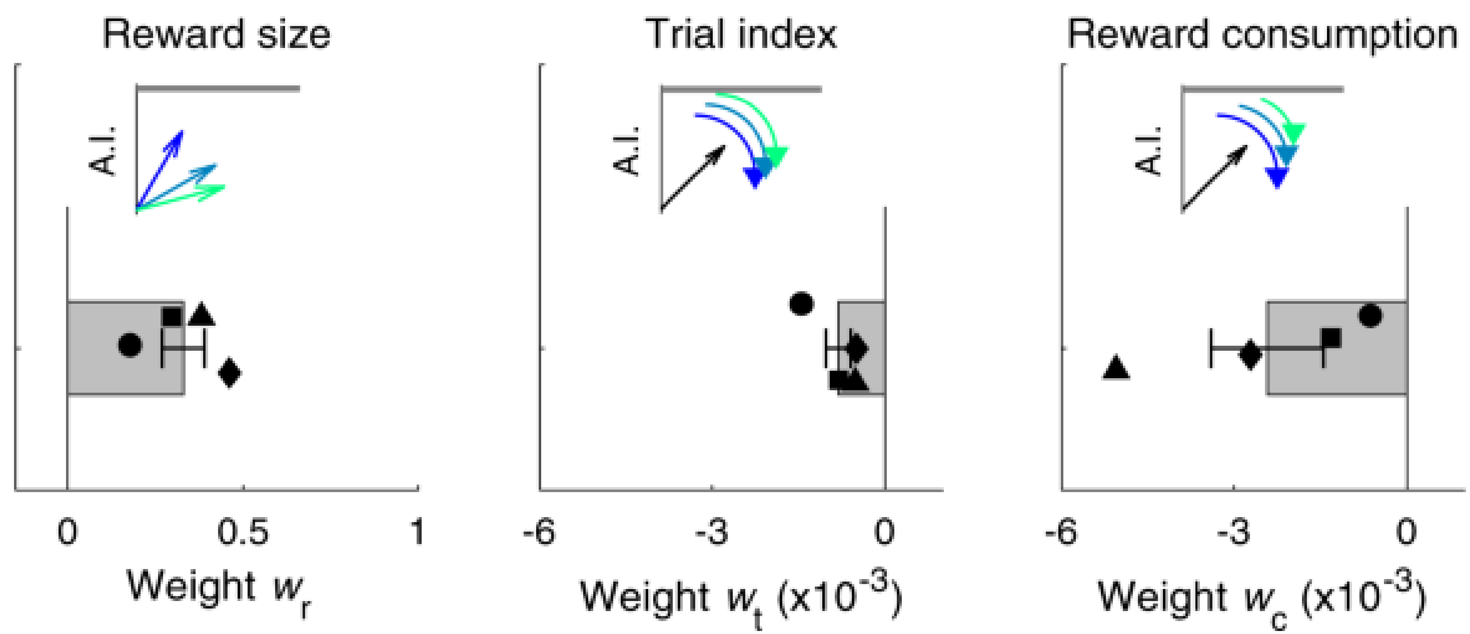
Case study: model predictions



Hernández-Navarro, L., *Nat Comms* (2021).

Case study: model estimates

Drift = w_r x Reward_size + w_t x Trial_index + w_c x Reward_consumption + w_0



Hernández-Navarro, L., *Nat Comms* (2021).

Take home messages

🏠

- ✅ Model-free analysis: check your raw data
  - Run model-independent data analysis first: what behavioral patterns/signatures do you expect?
- 😬 Experimental design: no amount of modeling can make up for a bad design! Does your design allow you to isolate the behavioral signatures you expect to see in the behavior?

Palminteri, Wyart & Koechlin (2017) *TICS*; Wilson & Collins (2019) *eLife*

Take home messages 🏠

✅ Model-free analysis: check your raw data

Run model-independent data analysis first: what behavioral patterns/signatures do you expect?

😬 Experimental design: no amount of modeling can make up for a bad design! Does your design allow you to isolate the behavioral signatures you expect to see in the behavior?

🤔 Parameter estimates: do the parameters take reasonable values? How are they related to each other (structure of the variance)? Are they in the range of what you expected?

🧽 Consider “sponge parameters” that will wash away unimportant variance: example: a leftward bias in choice. If not modelled, it may compromise the reliability of your other parameters of importance!

Palminteri, Wyart & Koechlin (2017) *TICS*; Wilson & Collins (2019) *eLife*

Cross-validation vs. bootstrapping

	Crossvalidation	Bootstrapping
Common	Both are resampling methods, computationally expensive (CPU hungry)	
Purpose	Good for estimating the model prediction errors	Good for estimating the confidence interval of model parameters.
Approach	Split the data into multiple sets, thus no overlapping between datasets.	Clone the data to create more sets, thus overlapping datasets.
Sample size	Needs a large sample size	Fine with small samples