



# 1C - Model comparison

Marion Rouault  
CNRS & Paris Brain Institute  
[marion.rouault@gmail.com](mailto:marion.rouault@gmail.com)

BAMB! 2025 Summer School

# The plan for the next 120 minutes

- Model selection

⇒ Goodness of fit: what makes a good model?

⇒ Quantitative criteria (penalized likelihoods)

⇒ Cross-validation

- Model recovery (with confusion matrix)

# The plan for the next 120 minutes

- Model selection

⇒ Goodness of fit: what makes a good model?

⇒ Quantitative criteria (penalized likelihoods)

⇒ Cross-validation

- Model recovery (with confusion matrix)

# A dual perspective on model selection

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



Outperform other  
models on  
*quantitative* measures  
of how well the model  
explains the data

Our winning model should pass **both** tests 🏆

# Quantitative criteria

Usually, there are several candidate models that potentially describe the behaviour and/or cognitive process you are interested in

The goal of **model selection** is to choose one of these models based on your data

In brief, the idea is:

- You define a minimum contrast estimator (e.g., log likelihood, least squares, ...) for each model
- Then you choose the model with the lowest estimator

The main problem: same data are used for training the algorithms and for choosing one

# Quantitative criteria

The number of models of even simple experiments grows very quickly

Suppose that you have two experimental design factors, A and B

Possible (linear) models include:

- single estimated value (constant)
- a main effect of A or B
- both main effects and their interaction

Hence  $2^f$  models to select from ( $f$  being the number of factors)

Plus, if you allow  $m$  types of model parameter this freedom, you get  $2^{f*m}$  number of variants

# Goodness of fit

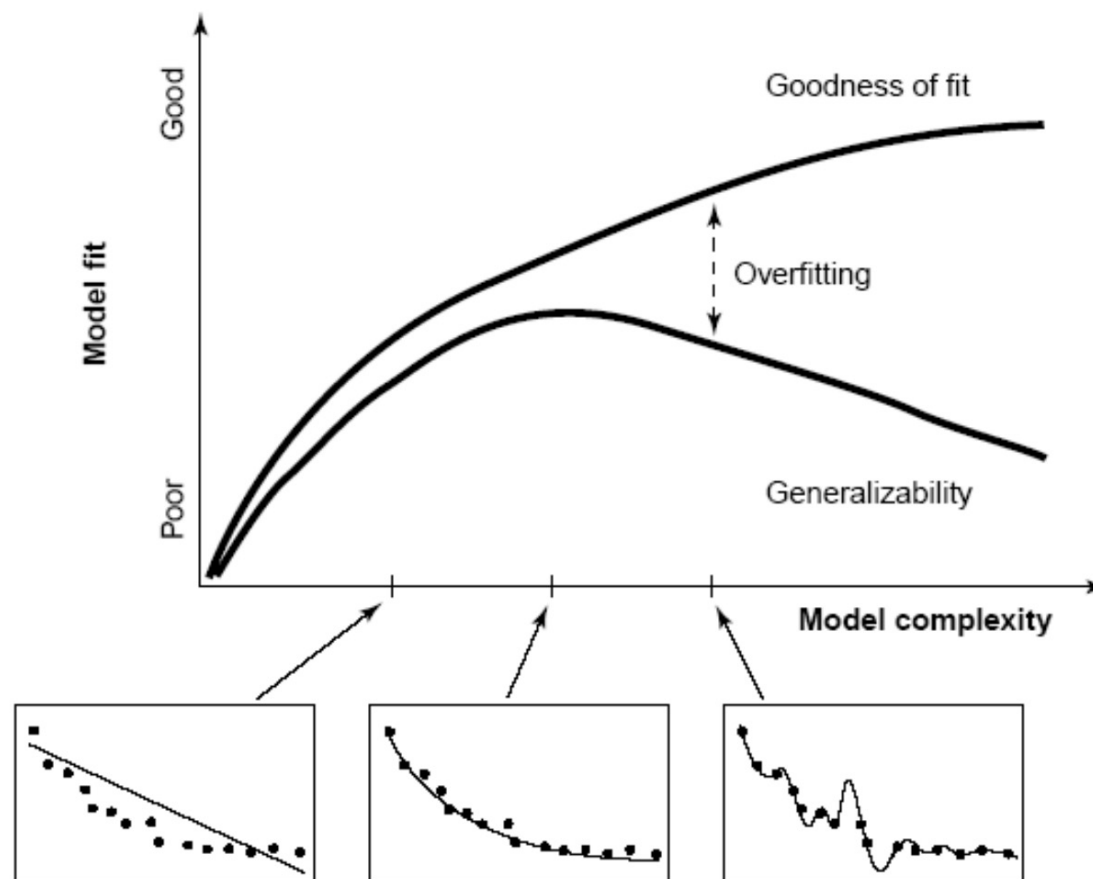
- (1) We can just take how good models are fitting and correct for the number of free parameters
- (2) We can fit on one part of the data and see how good the model is at predicting data that we have not used for fitting (**Cross-validation**)

- Measures of model evidence

- Akaike Information Criterion (AIC)
- Bayesian Information Criterion (BIC)
- Variational Bayes (VB)
  - negative free energy as lower bound approximation to the log evidence (“evidence lower bound” (ELBO), cf. Luigi Acerbi)
- Sampling-based methods

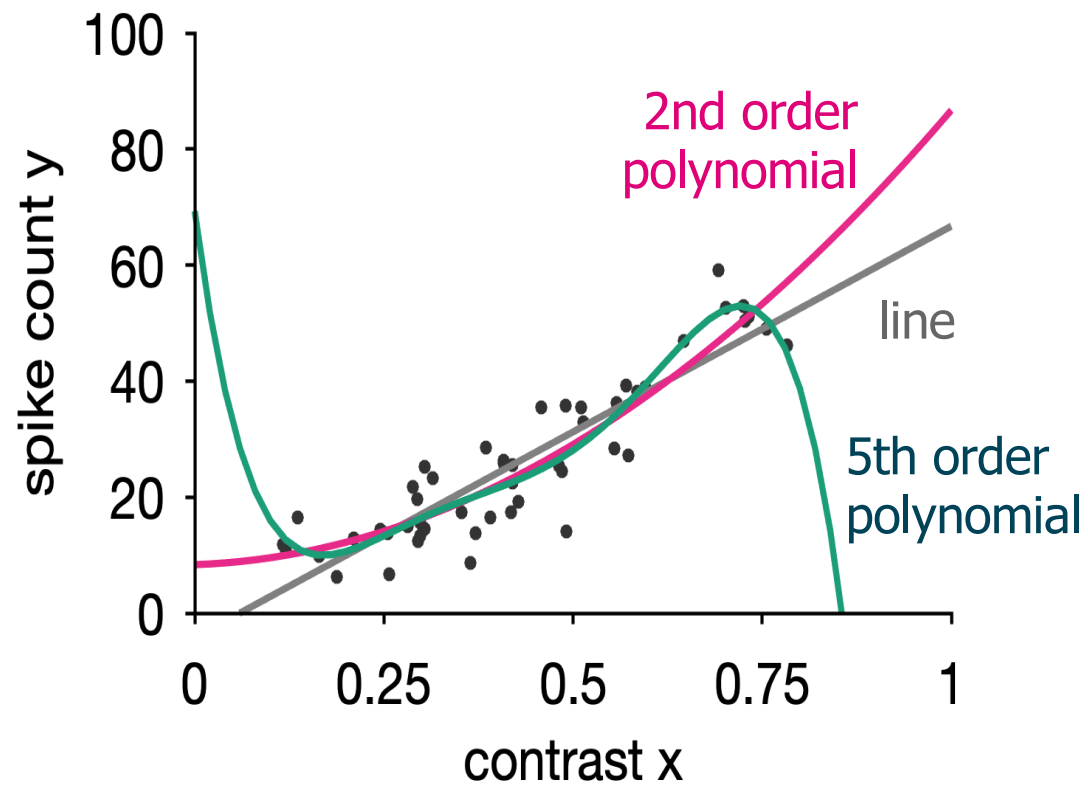
# Goodness of fit

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

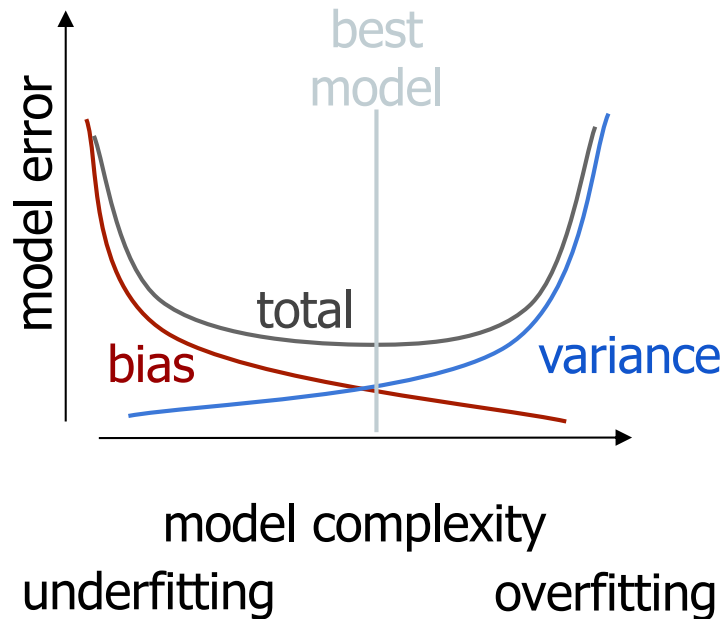




# Comparing models



# Bias-variance trade-off



## Bias

Low model complexity: systematic deviation from structure underlying data (underfitting)

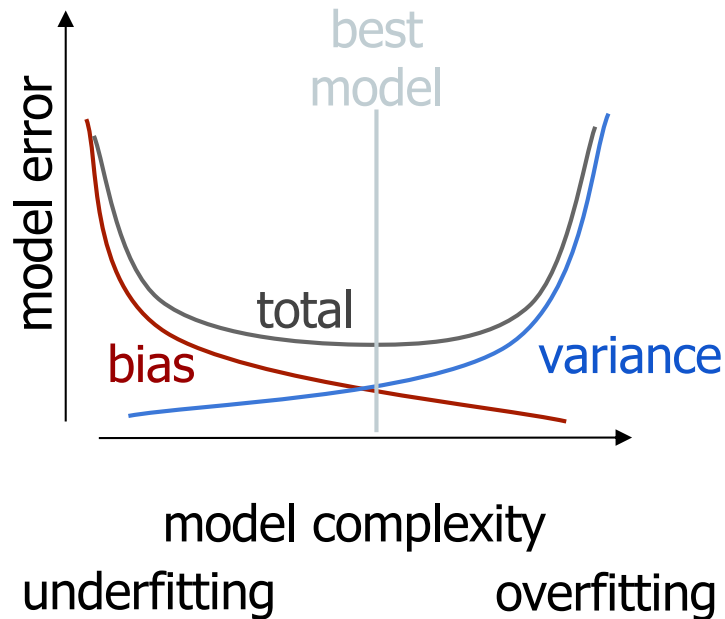
## Variance

High model complexity: capturing variability beyond the structure underlying data (i.e., noise; overfitting)

**Total error = bias + variance**

**Best model 🏆**  
balances bias and variance

# Bias-variance trade-off



## Bias

Low model complexity: systematic deviation from structure underlying data (underfitting)

## Variance

High model complexity: capturing variability beyond the structure underlying data (i.e., noise; overfitting)

**Total error = bias + variance**

**Best model** 🏆

**balances bias and variance**

To account for model complexity:

- Approach 1: Penalized likelihoods
- Approach 2: Cross-validation

# Quantitative criteria

- popular: Information criteria : AIC & BIC
- simple and transparent in how they quantify the trade-off between parsimony and goodness of fit
- **trade-off: criterion = goodness of fit - model complexity**
- small = good
- 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..)

# Comparing models while correcting for model complexity

When we do a regular ML fit of a model we get the likelihood

But if we have more parameters this always gets better (e.g. quadratic describes more variance as linear etc).

There is a first order way of correcting for this: Akaike Information Criterion (AIC) (k=number of parameters)

We prefer models with **lower AIC** - every parameter costs something

$$\text{AIC} = -2\ln \mathbf{p} \left( y | \hat{\theta} \right) + 2k$$

*Likelihood of best fit* *Penalty term*

# Quantitative criteria

Bayesian Information Criterion (**BIC**) penalizes model complexity more severely (k=number of parameters, n=number of data points)

$$\text{BIC} = -2 \ln \mathbf{p} \left( y | \hat{\theta} \right) + k \ln n$$

*Likelihood of best fit*


*Penalty term*

Everything else being equal, BIC tends to select models less complex than those selected via AIC

# Quantitative criteria

Bayesian Information Criterion (**BIC**) penalizes model complexity more severely (k=number of parameters, n=number of data points)

$$\text{BIC} = -2 \ln \mathbf{p} \left( y | \hat{\theta} \right) + k \ln n$$

In cognitive and behavioral science, often our data points are trials or decisions or responses 

BIC and AIC weigh equally all data points ...

... But not all responses matter equally 


e.g. at reversal, differences, but at plateau, all models perform well

# Other metrics

Likelihood ratio (for nested models)

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

null model is true



Bayes Factor

Approximation to model evidence

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

- Laplace approximation
- Variational Bayes (VB): negative free energy as lower bound approx to the log evidence (elbo, evidence lower bound)
- Sampling-based methods



# The plan

- Model selection
  - ⇒ Goodness of fit: what makes a good model?
  - ⇒ Quantitative criteria (penalized likelihoods)
  - ⇒ **Cross-validation**
- Model recovery (with confusion matrix)

# Cross validation – general principles

- The preferred model class is the one whose (weighted) parameterized models **best predict 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

# Cross validation – general principles

- The preferred model class is the one whose (weighted) parameterized models **best predict 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

# Cross validation

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
...	...	...	...	...

# Cross validation

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 🤔

?

# Cross validation

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

# Cross validation

We seek a model to relate



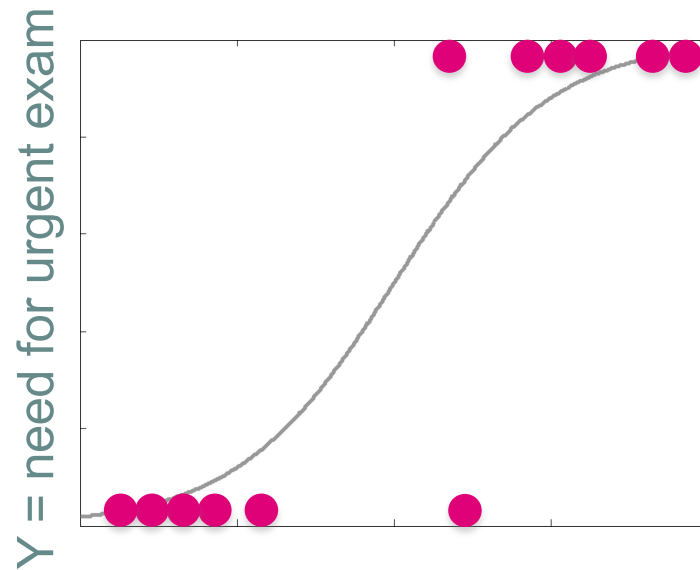
to



# Cross validation

We seek a model to relate  to

e.g. a logistic regression  $Y = \text{Fever} + \text{Age} + \dots$

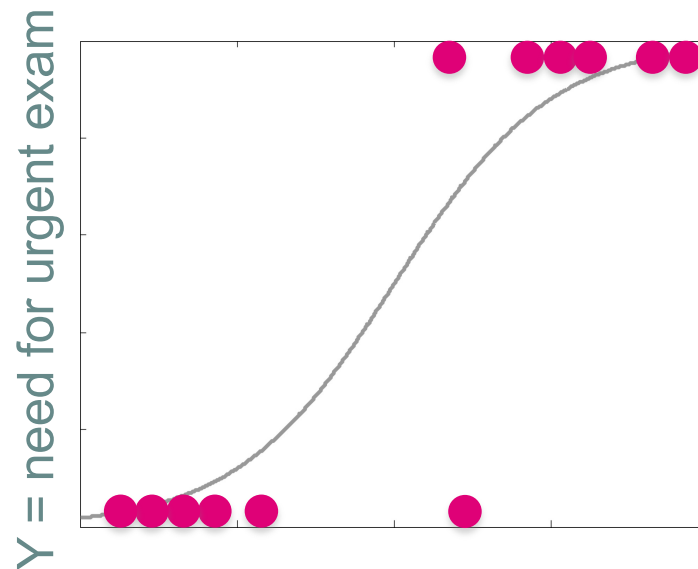




# Cross validation

We seek a model to relate  to

e.g. a logistic regression  $Y = \text{Fever} + \text{Age} + \dots$



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

# Cross validation



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

# Cross validation



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

⇒ i. e. TRAINING

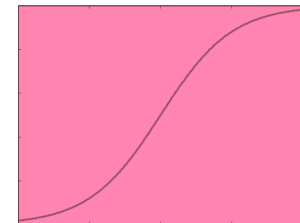
# Cross validation



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
- ⇒ i. e. TRAINING



# Cross validation



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

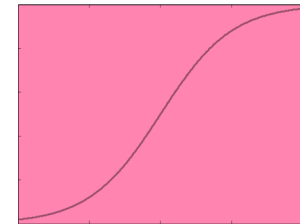
With your data you have to:

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

⇒ i. e. TRAINING

- Evaluate how well your model (regression) works

⇒ i. e. TESTING



# Cross validation



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  
⇒ i. e. TRAINING
- Evaluate how well your model (regression) works  
⇒ i. e. TESTING



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

# Cross validation



⇒ i. e. TRAINING

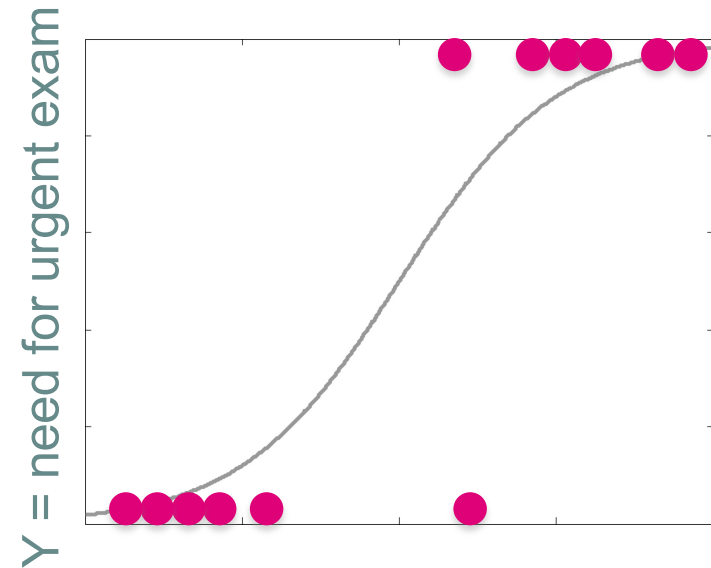
⇒ i. e. TESTING

# Cross validation

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

⇒ i. e. TRAINING

⇒ i. e. TESTING



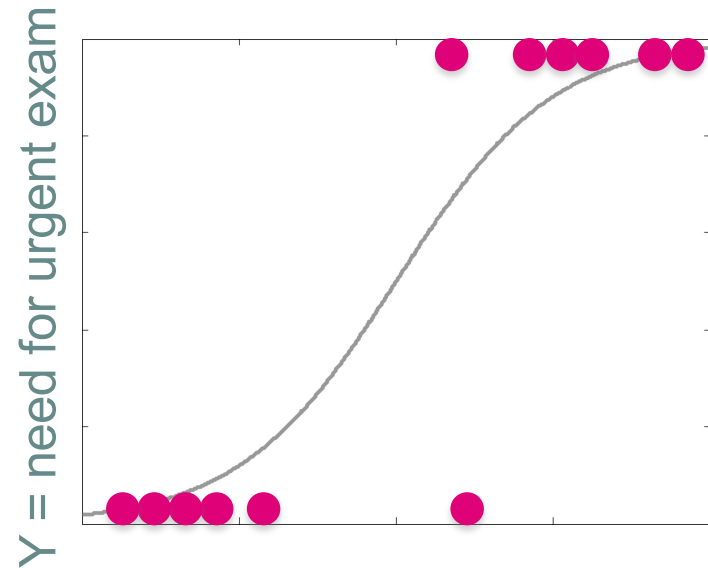


# Cross validation

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

⇒ i. e. TRAINING

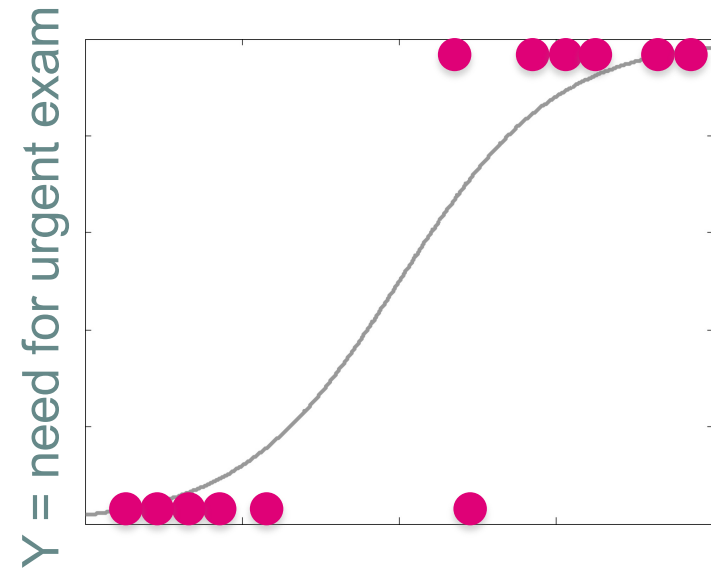
⇒ i. e. TESTING



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

# Cross validation

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

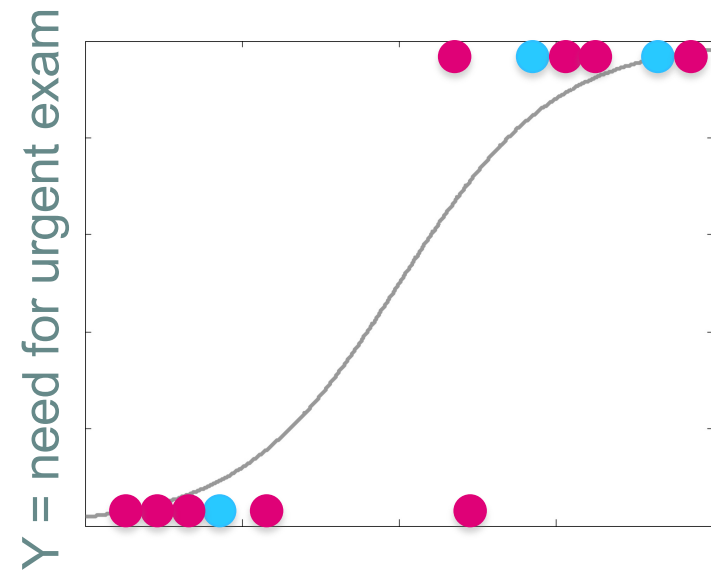


# Cross validation




💡 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:



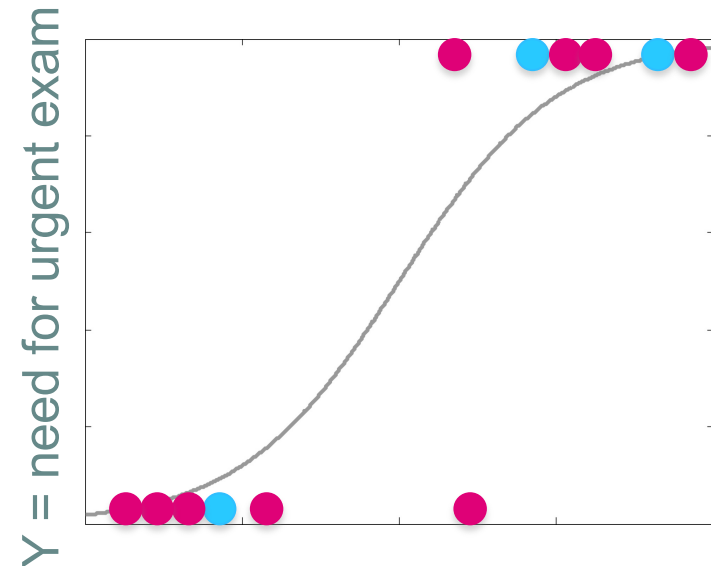
# Cross validation



💡 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

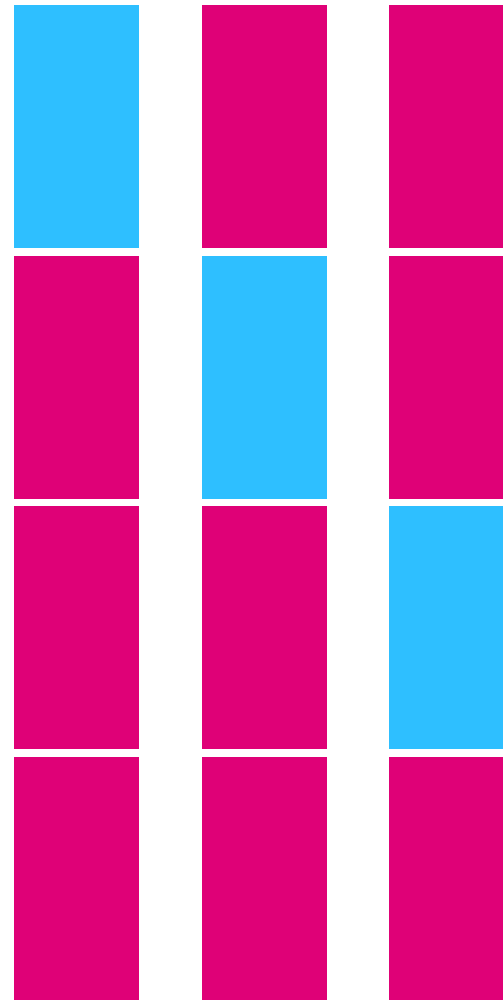


But how do you know that this is the best way to split your data into 75% for **training** and 25% for **testing**?

# Cross validation



But how do you know that this is the best way to split your data into 75% for **training** and 25% for **testing**?

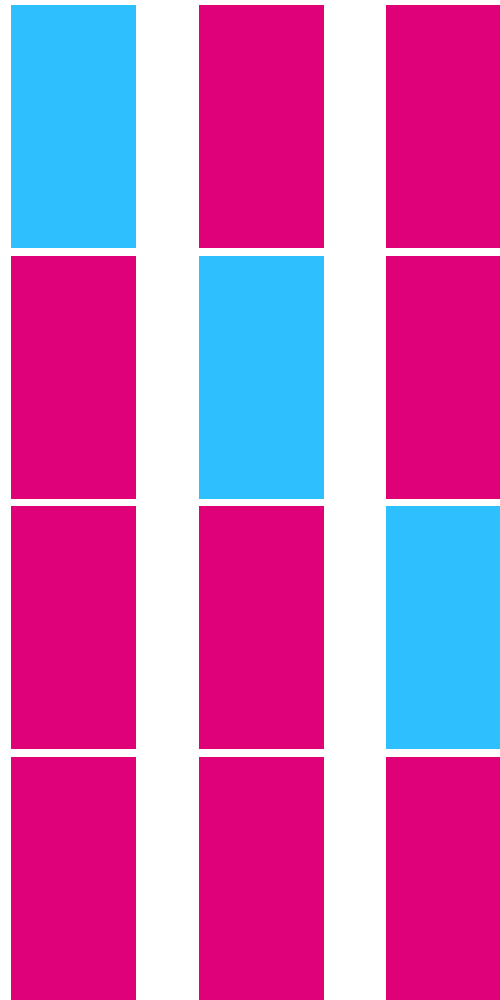


# Cross validation

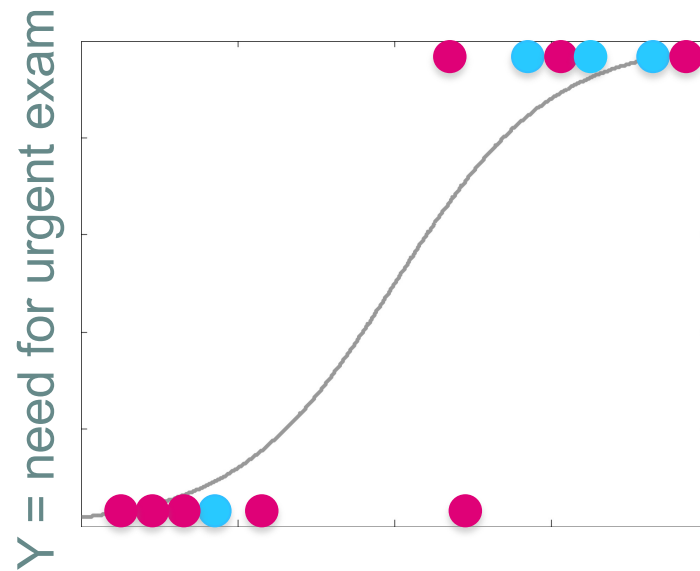


But how do you know that this is the best way to split your data into 75% for **training** and 25% for **testing**?

**Cross-validation uses all these blocks, one at a time, and summarizes the results at the end**



# Cross validation

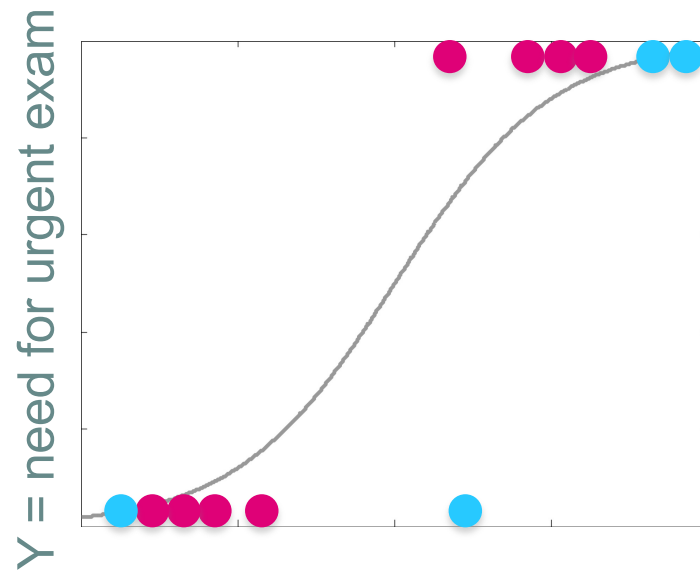
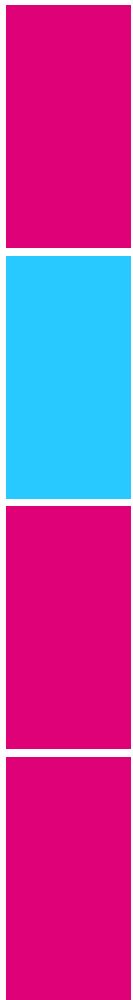


You keep track of how well the model does:

Test data categorization	
Correct ✓	Incorrect ✗
3	1



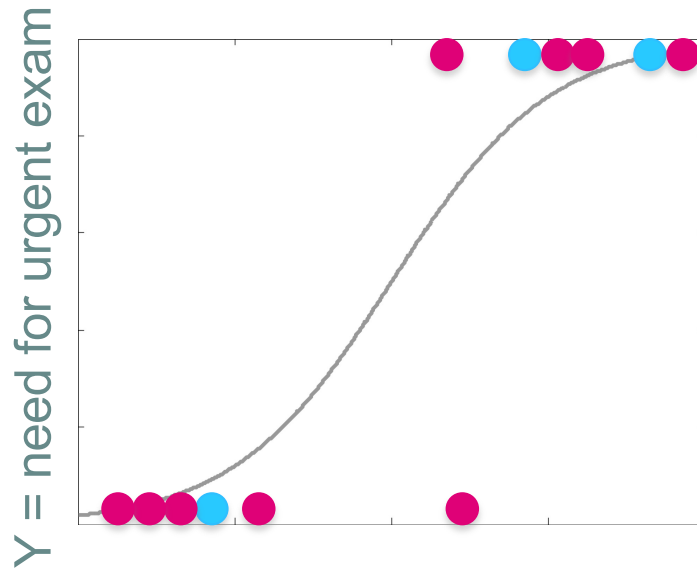
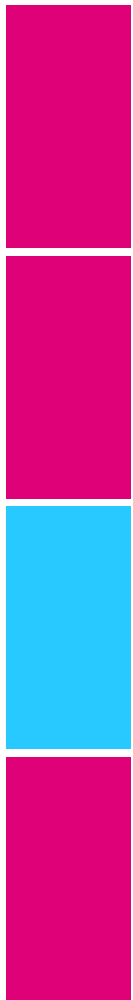
# Cross validation



You keep track of how well the model does:

Test data categorization	
Correct ✓	Incorrect ✗
4	0

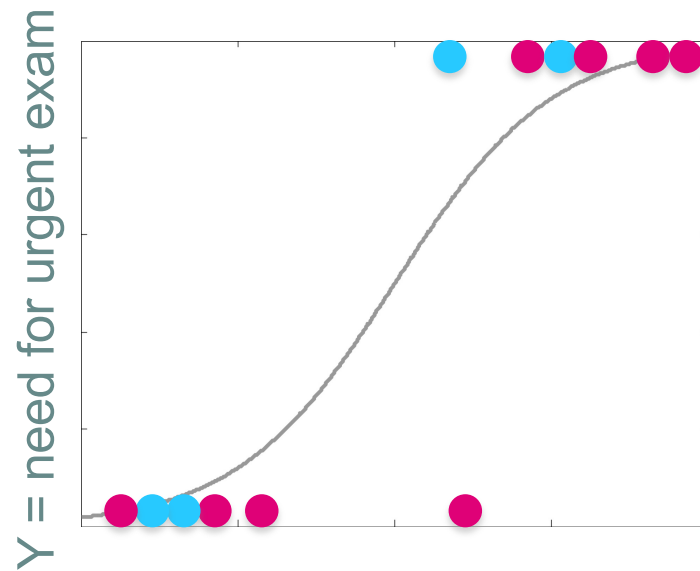
# Cross validation



You keep track of how well the model does:

Test data categorization	
Correct ✓	Incorrect ✗
2	2

# Cross validation

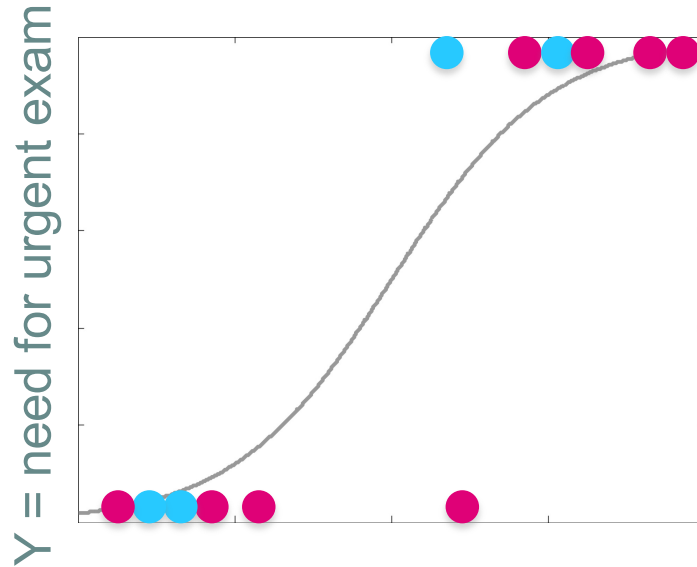


You keep track of how well the model does:

Test data categorization	
Correct ✓	Incorrect ✗
3	1

# Cross validation

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

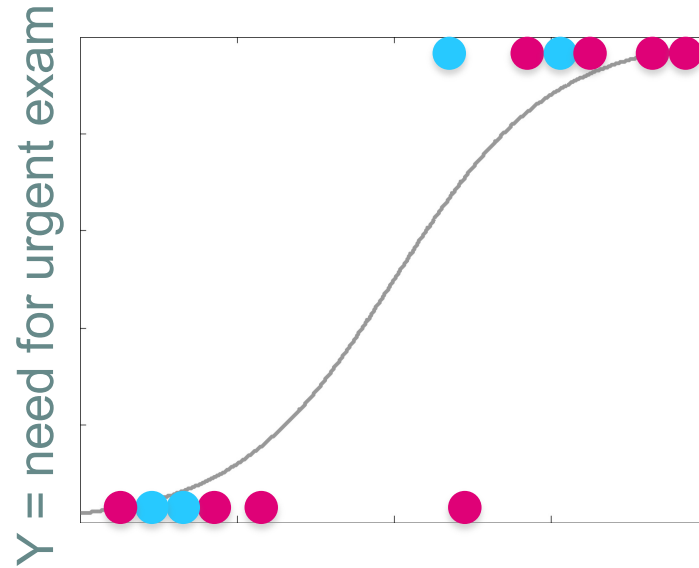


Hence you know how well the model does OVERALL:

Test data categorization	
Correct ✓	Incorrect ✗
12	4

# Cross validation

ModelA



Hence you know how well the model does OVERALL:

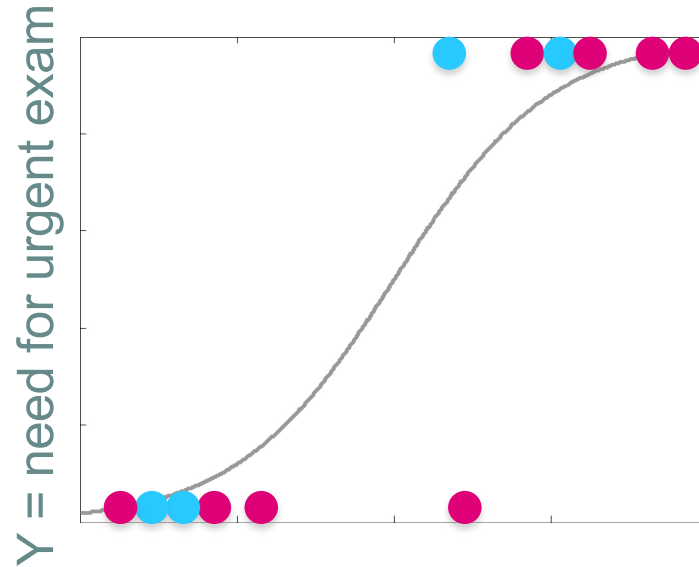
Test data categorization

Correct ✓  
12

Incorrect ✗  
4

# Cross validation

ModelA



Hence you know how well the model does OVERALL:

Test data categorization

Correct✓  
12

Incorrect✗  
4

ModelB

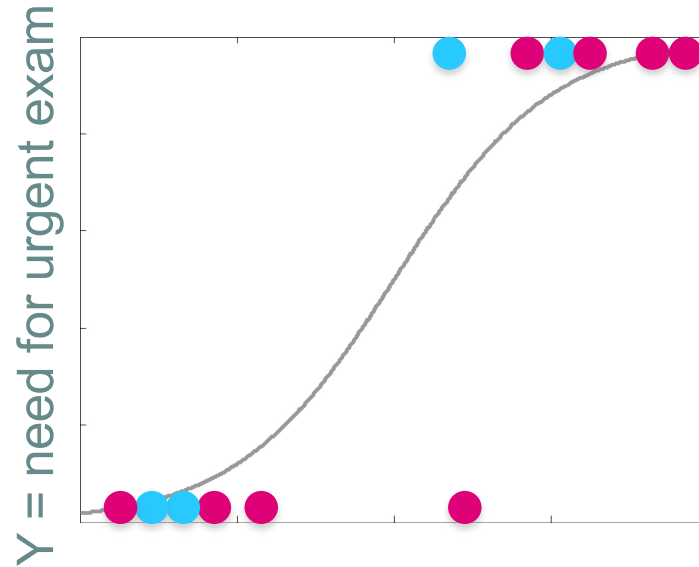
Test data categorization

Correct✓  
10

Incorrect✗  
6

# Cross validation

ModelA



Hence you know how well the model does OVERALL:

Test data categorization

Correct ✓  
12

Incorrect ✗  
4

ModelB  
ModelC  
ModelD  
ModelE

Test data categorization

Correct ✓  
10

Incorrect ✗  
6

# Cross validation



“Four-fold” cross-validation

Number of divisions is arbitrary

Ten-fold cross-validation is common



# Cross validation

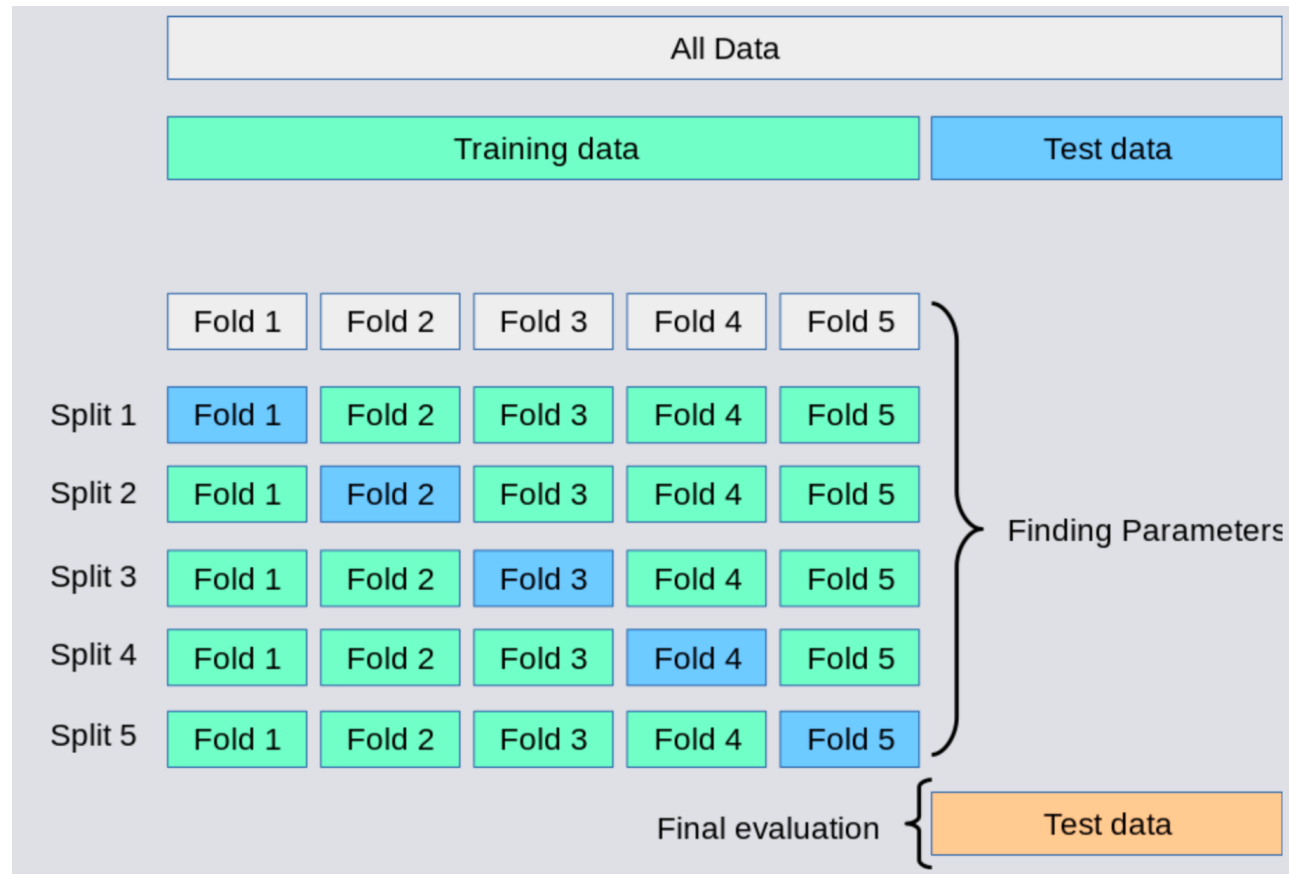


“Four-fold” cross-validation

Number of divisions is arbitrary

Ten-fold cross-validation is common

# Nested cross-validation



1. Evaluation of the parameters by 10-fold cross validation (“outer folds”)
2. The parameters themselves were tuned by 10-fold cross validation (“inner folds”)
3. Rotate through inner and outer folds: a lot of work 💪

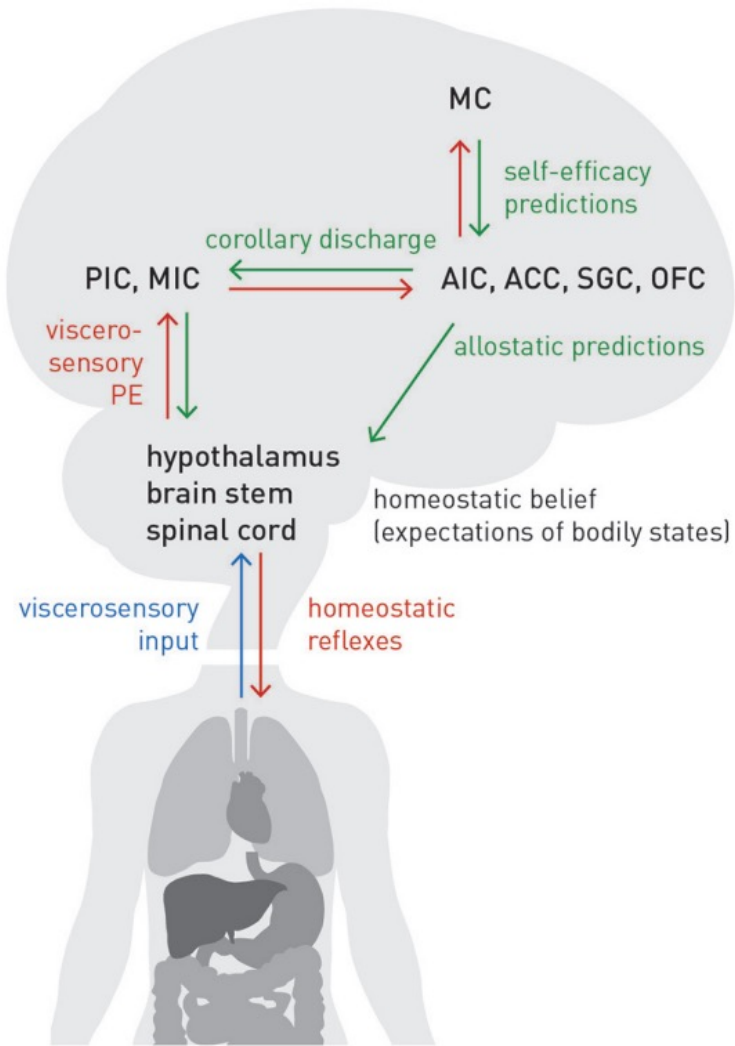
# Cross validation

- ✓ Extreme case of cross-validation: **leave-one-out cross-validation**
- ✓ Instead of 10-fold or 10 blocks division of your data, each division is one participant
- ✓ Popular in fMRI

# Cross validation

- ✓ Extreme case of cross-validation: **leave-one-out cross-validation**
- ✓ Instead of 10-fold or 10 blocks division of your data, each division is one participant
- ✓ Popular in fMRI
  - Remove one data point
  - Train on all but  $R-1$  data points, where  $R$  is the total number of data points
  - Test on remaining data point and record error
  - Repeat for all  $R$  data points
  - Report mean error
  - Computationally expensive 💰

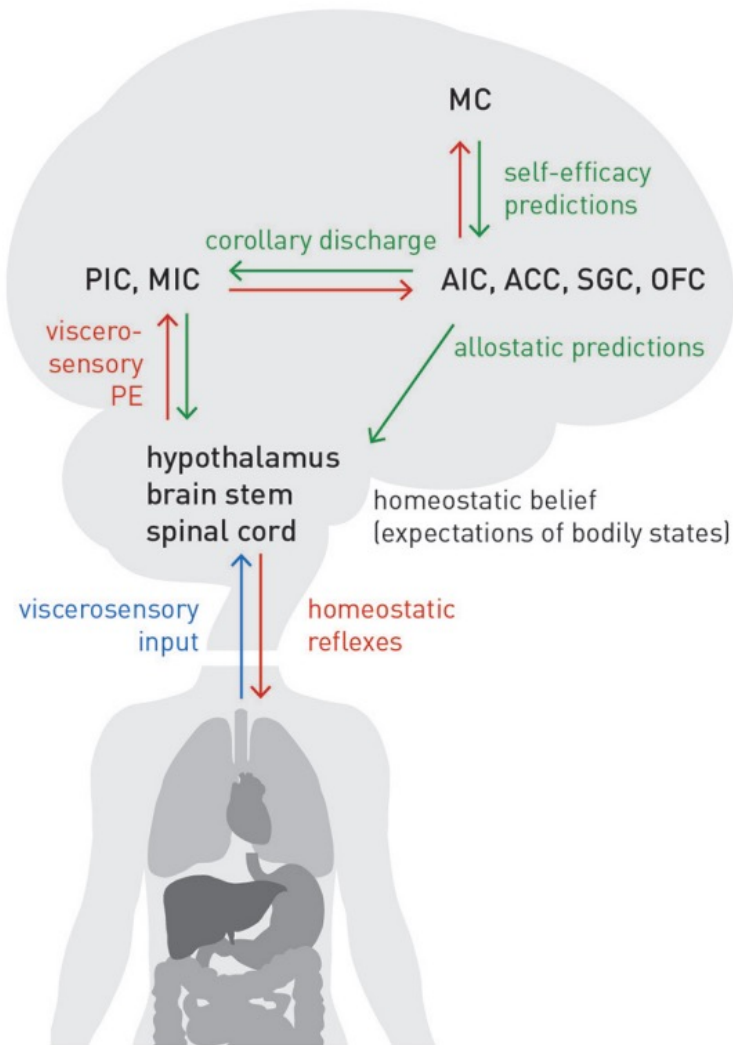
# Cross validation example



## Theoretical argument

Linking fatigue, interoception and metacognition

# Cross validation example



## Theoretical argument

Linking fatigue, interoception and metacognition

## Empirical work

- 71 participants with multiple sclerosis
- Y = degree of fatigue experienced
- 14 measurements: neuropsychological assessments, physiological assessments including interoceptive awareness and interoceptive beliefs, (meta)cognitive task assessments

# The plan for the next 120 minutes

- Model selection
  - ⇒ Goodness of fit: what makes a good model?
  - ⇒ Quantitative criteria (penalized likelihoods)
  - ⇒ Cross-validation
- Model recovery (with confusion matrix)

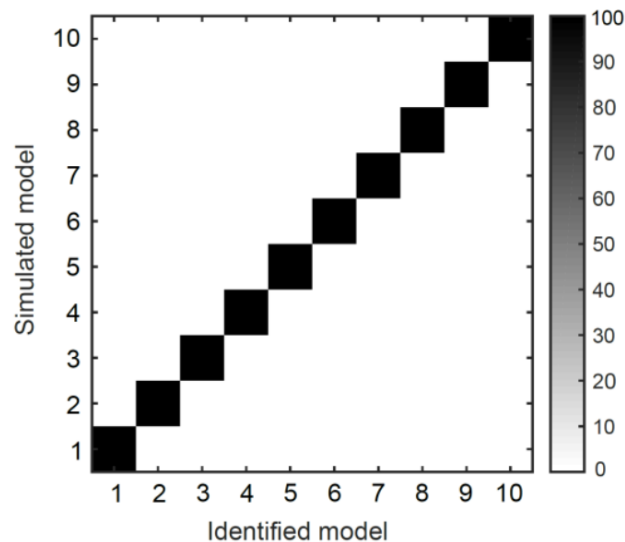
# Model recovery

- ✓ 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



# Model recovery

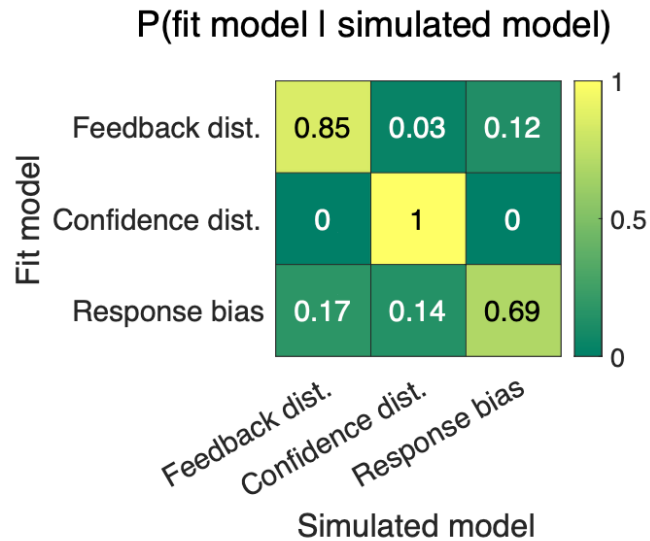
- ✓ Simulate data for your model space: only a limited number of models can be tested, carefully consider your choice
- ✓ Fit each model to all simulated data sets based on each model
- ✓ Estimate how often true generative model is identified and plot confusion matrix: we want each model to be **identifiable**



*i.e., large values along the diagonal*

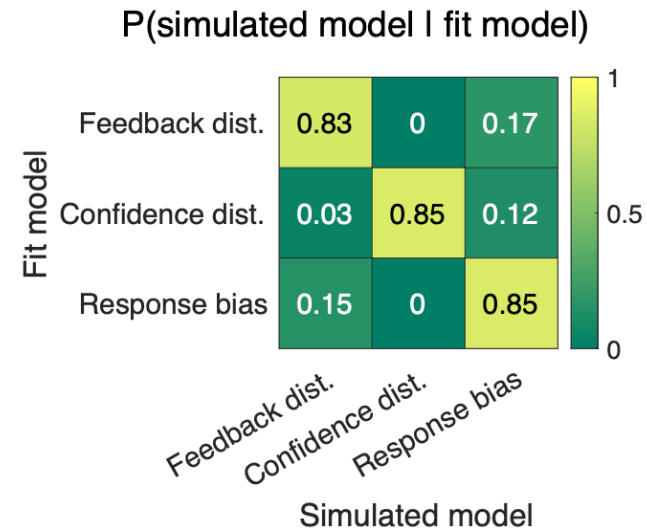
# Model recovery

## Confusion matrix



probabilities of the best-fitting model given which of the three models was simulated

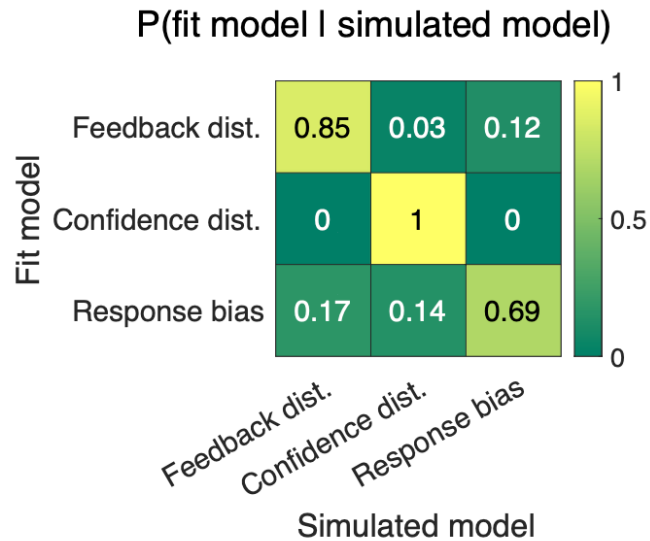
## Inversion matrix



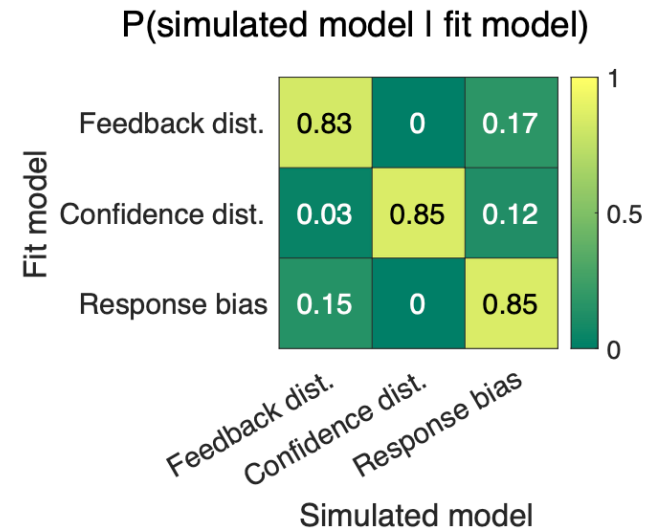
probabilities for which of the three models were simulated given a best-fitting distortion

# Model recovery

## Confusion matrix



## Inversion matrix



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

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

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

# 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
- Criterion on which to select: e.g., BIC

# 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
- Criterion on which to select: e.g., BIC



**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

- Popular quantitative information criteria: AIC & BIC
- Examine qualitative signatures of behaviour: model validation exercise (tutorial 1B) is key! **Can you validate your best-fitting model AND *unvalidate* your alternative models?**
- Cross-validation
  - Evaluate our models' performance
  - Evaluate a machine learning method
  - Make predictions for new data as accurately as possible

# Acknowledgements 🙏

- A. Wu, J. Drugowitsch, A. Hyafil: neuromatch academy
- K. Preuschoff, K. Wimmer (previous BAMB!s)
- 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.

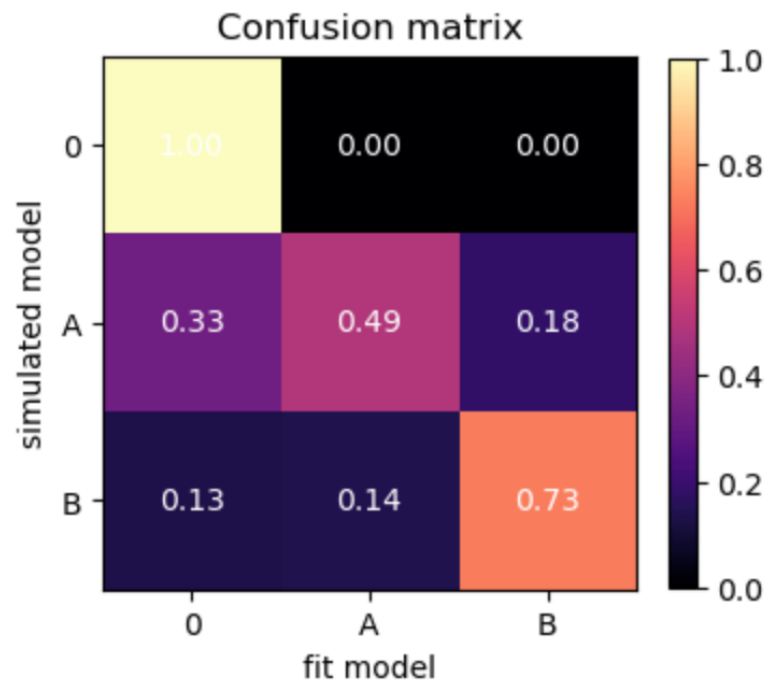
So we can use **k-fold cross-validation**!

- 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.

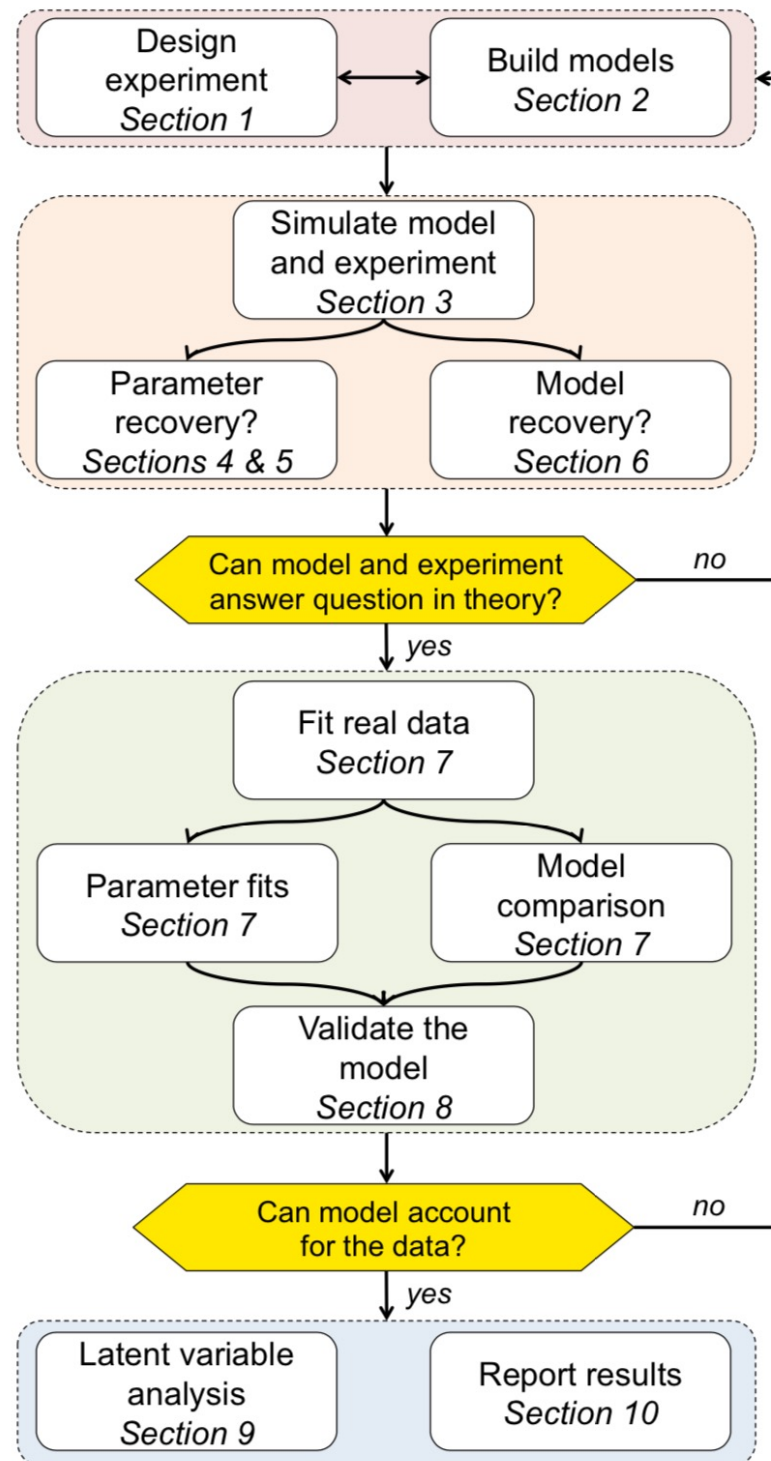
The aim is to build an experimental paradigm that will allow you to identify a model distinctly from alternative accounts

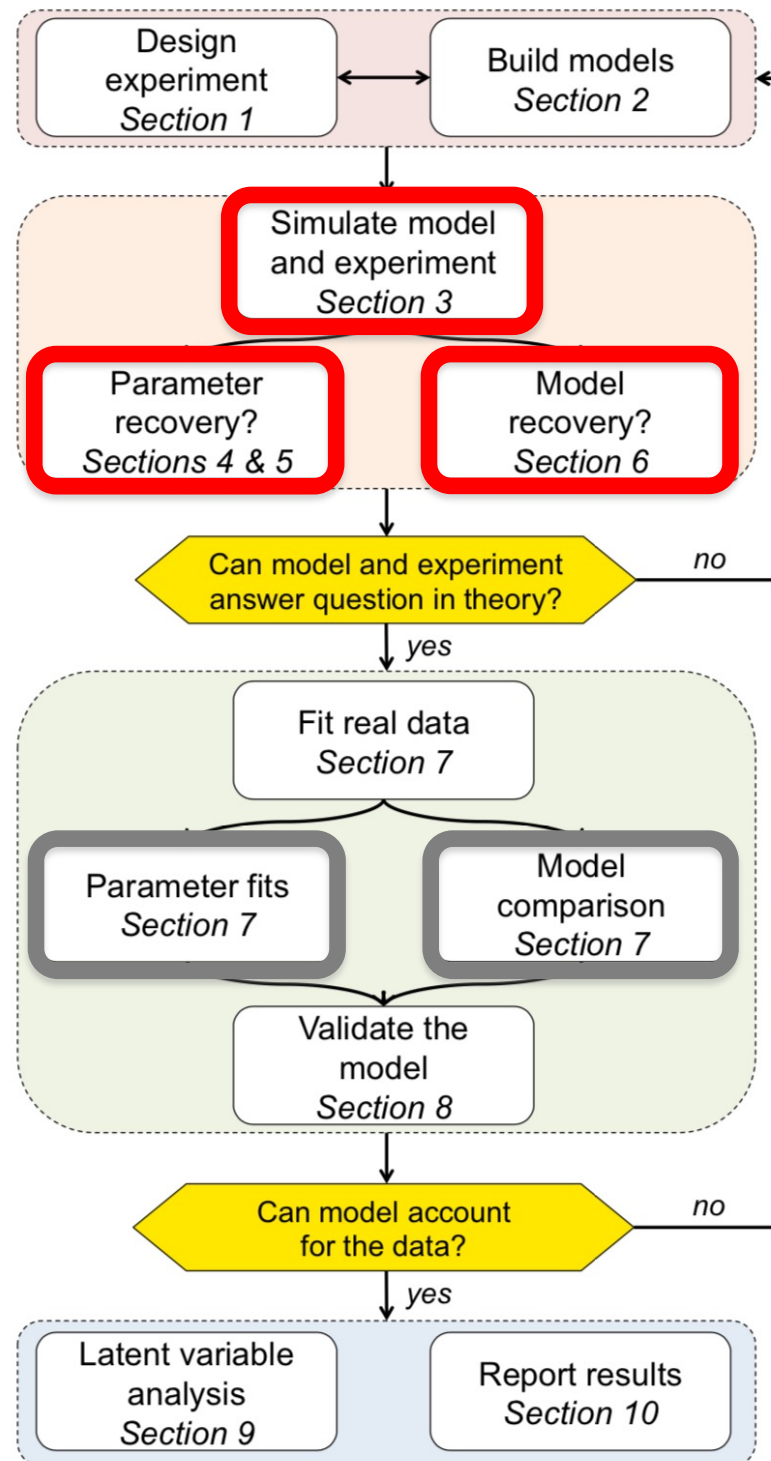


**Questions?**

# Wrap-up Day 1

# Wrap-up Day 1





# 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?



# 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!

# Cross-validation vs. bootstrapping

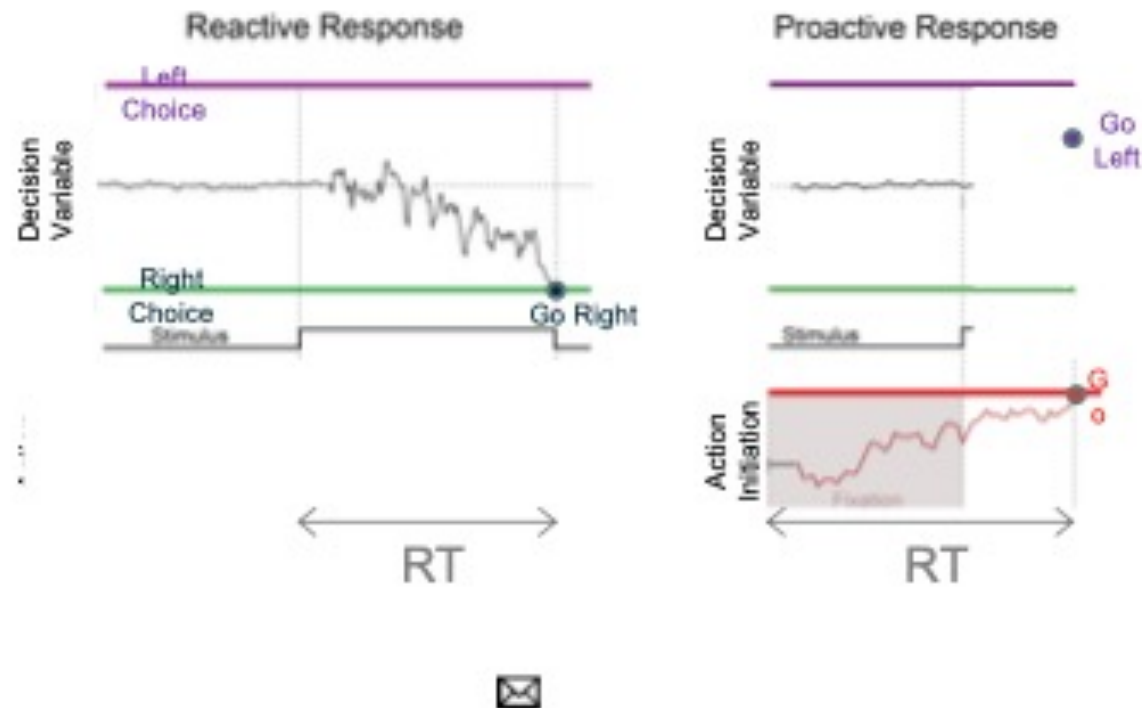
	Crossvalidation	Bootstrapping
<b>Common</b>	Both are resampling methods, computationally expensive (CPU hungry)	
<b>Purpose</b>	Good for estimating the model prediction errors	Good for estimating the confidence interval of model parameters.
<b>Approach</b>	Split the data into multiple sets, thus no overlapping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

**Questions?**

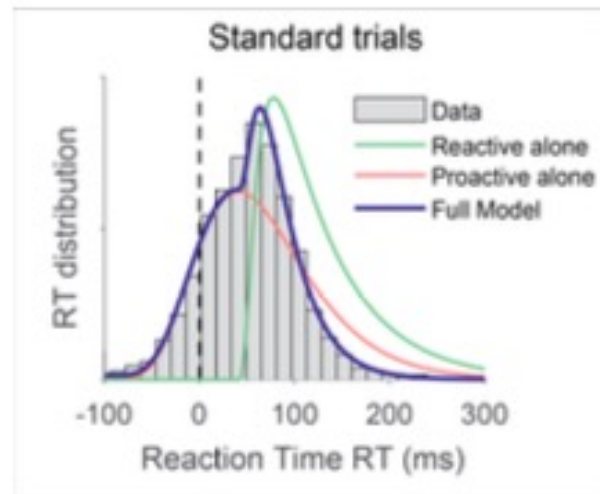
# Case study: an alternative model for perceptual decision-making in rats

Decision variable:  
accumulation-to-bound

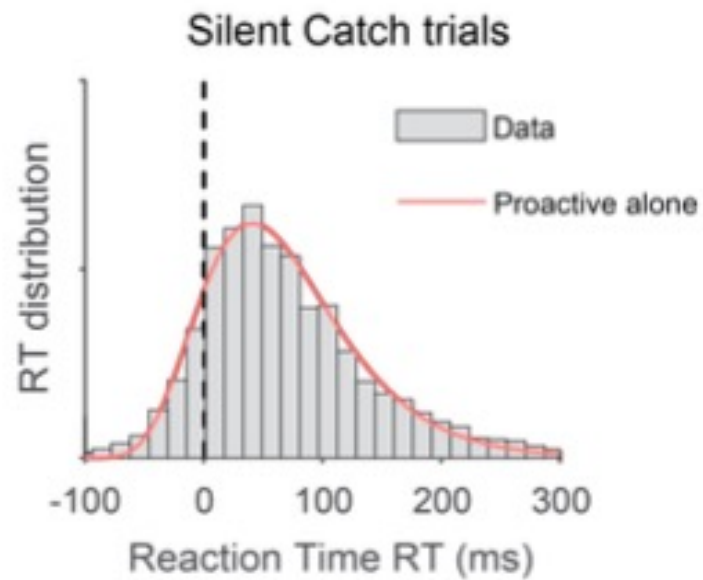
Action initiation:  
urgency signal



# Case study: model validation

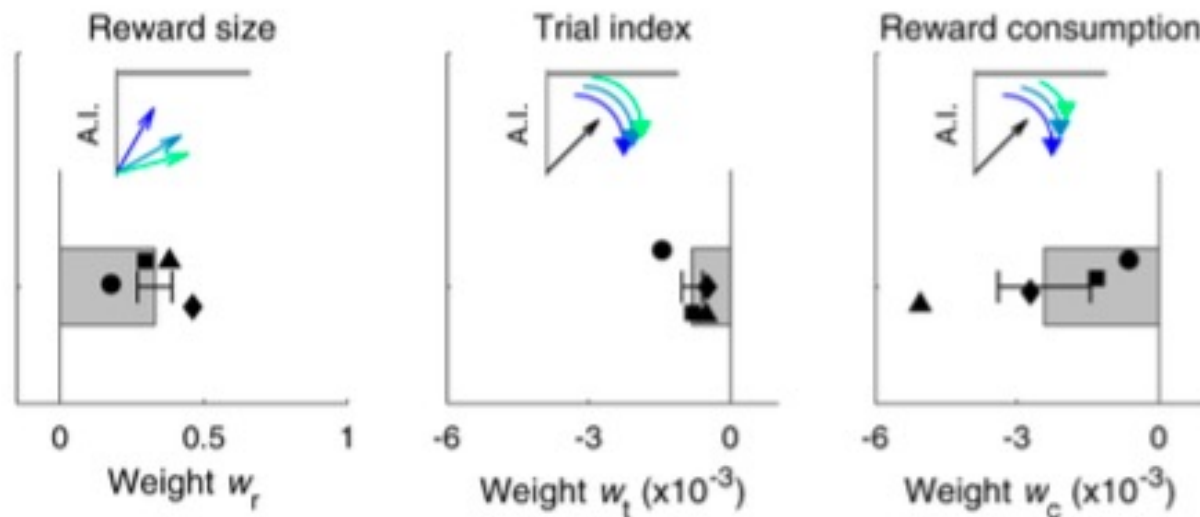


# Case study: model predictions



# Case study: model estimates

$$\text{Drift} = w_r \times \text{Reward\_size} + w_t \times \text{Trial\_index} + w_c \times \text{Reward\_consumption} + w_0$$



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

