machine learning for (f)MRI

Francisco Pereira

Machine Learning Team

Functional Magnetic Resonance Imaging Core

National Institute of Mental Health

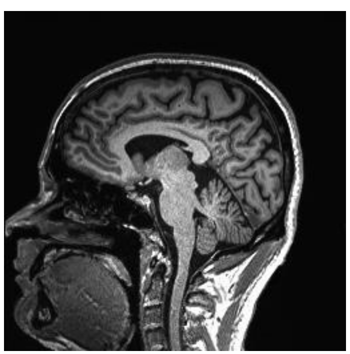
what is machine learning?

- study of computer programs that learn to predict something
- learn from data, without being told how to do it explicitly

[Arthur Samuel, 1959]

- "statistics, reinvented poorly from first principles"
- "statistics, but it works for real problems"

[assorted trash talking heard over the years]



structural MRI

classification — "is this structural image of a patient or a control? why?"

regression — "can we predict participant characteristics from the image?"

clustering — "are there subgroups of patients with similar images?"

science is about questions, not methods

- description

  “Are observations explainable in terms of a few (latent) variables?”

- prediction

  “Is the evolution of an outcome variable predictable from observations (or latent variables estimated from them)? How?”

- causality and control

  “How would intervening on some variables affect others?”

- mechanism or computation

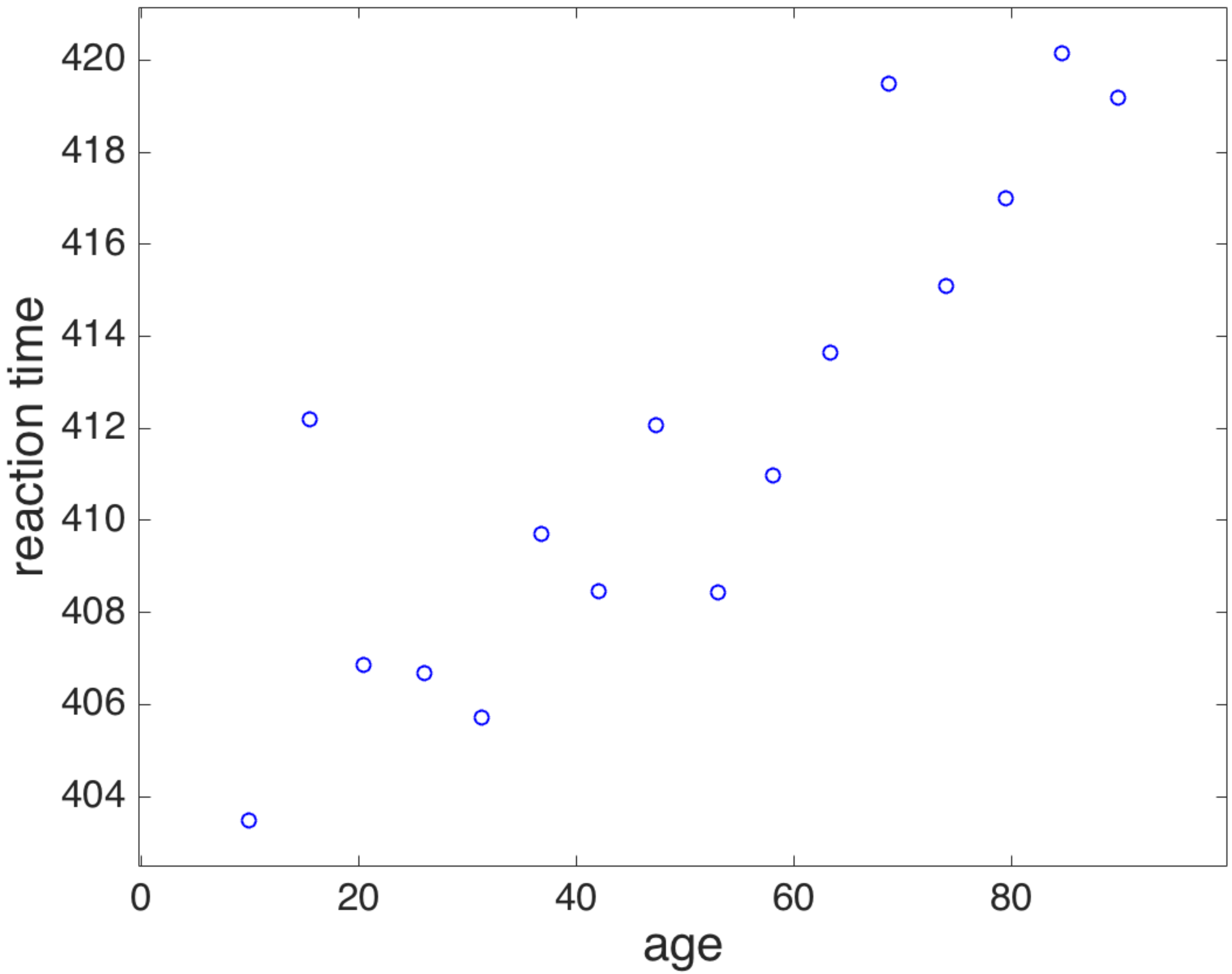
  “How does an input get transformed to produce the observations?"

linear regression from a machine learning viewpoint

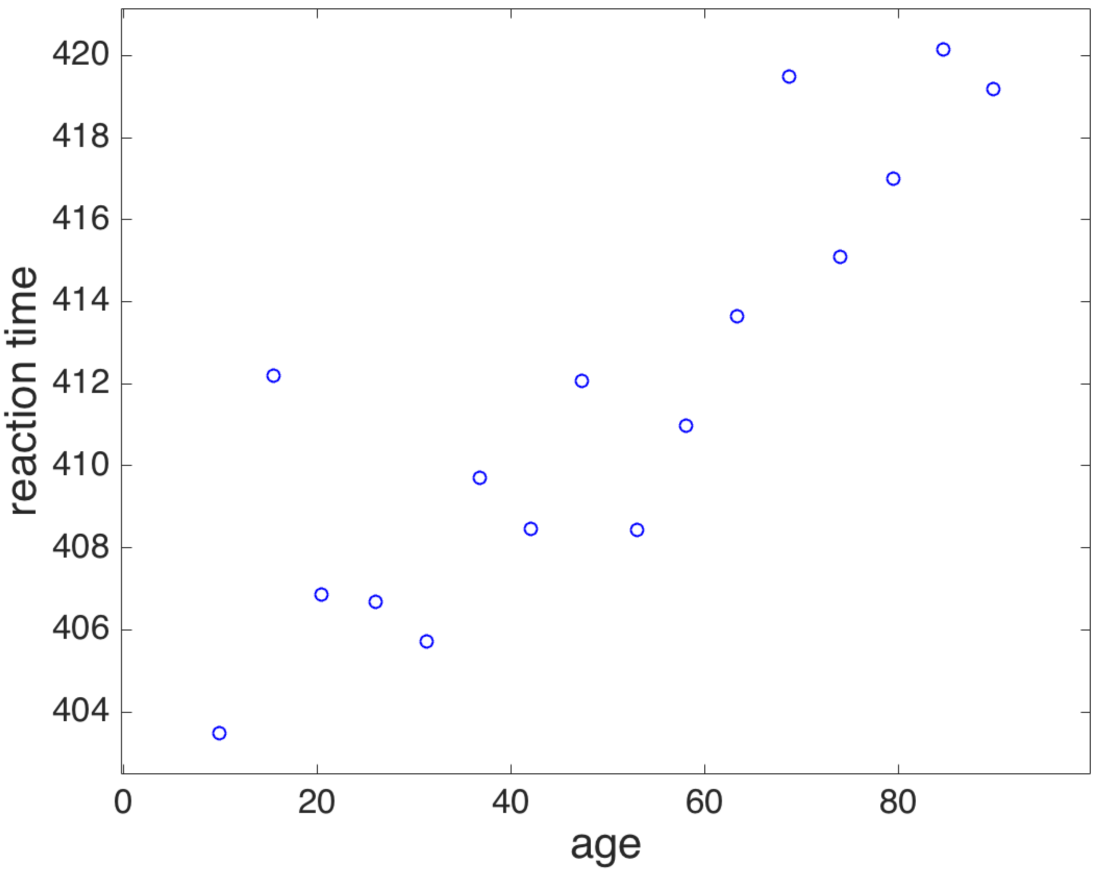
[adapted from slides by Russ Poldrack]

once upon a time there was a sample...

is reaction time (RT) related to age?

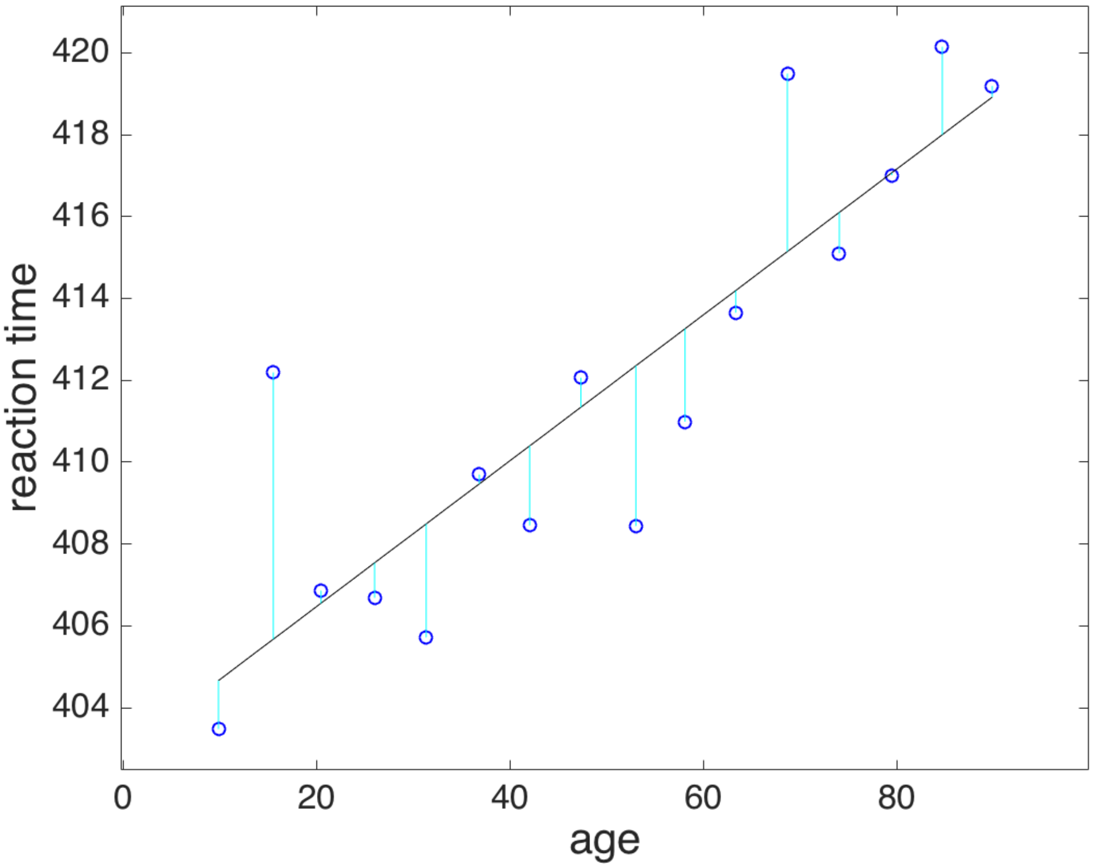


is RT related to age?



is RT related to age?

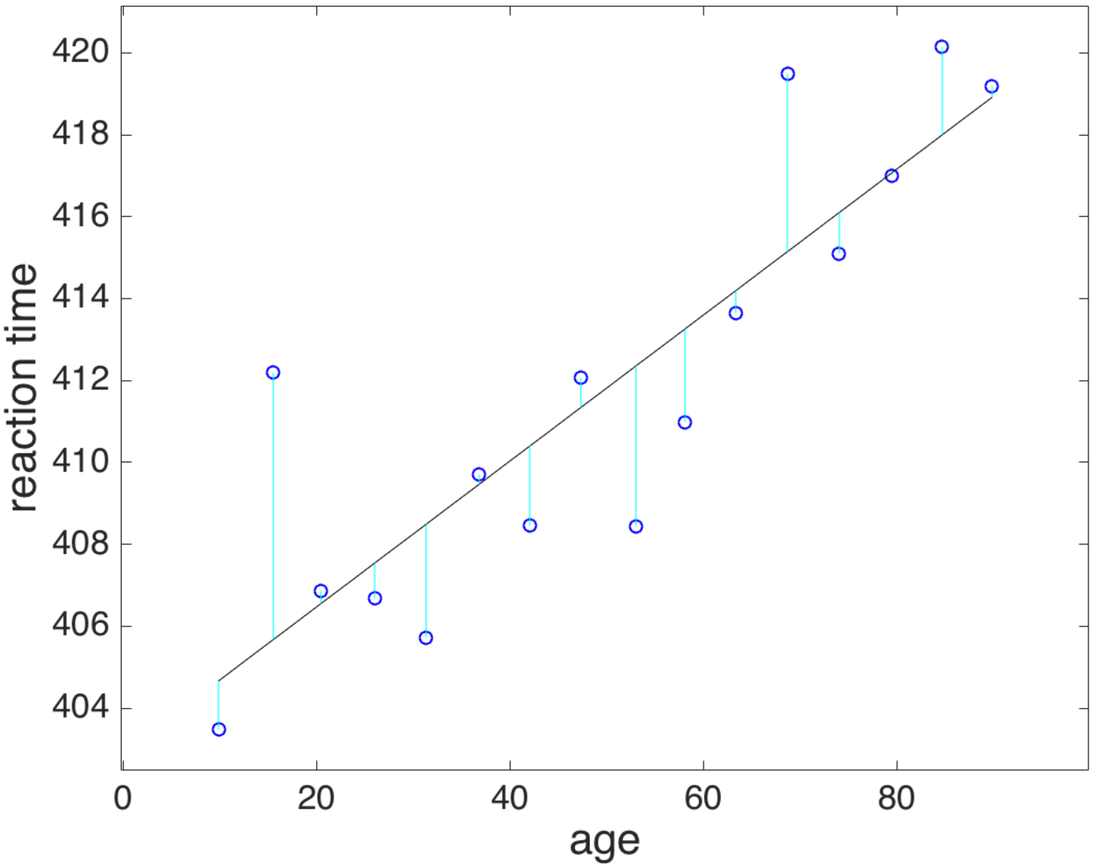
model: $RT = b_0 + b_1 * age + e$



is RT related to age?

parameters in the population

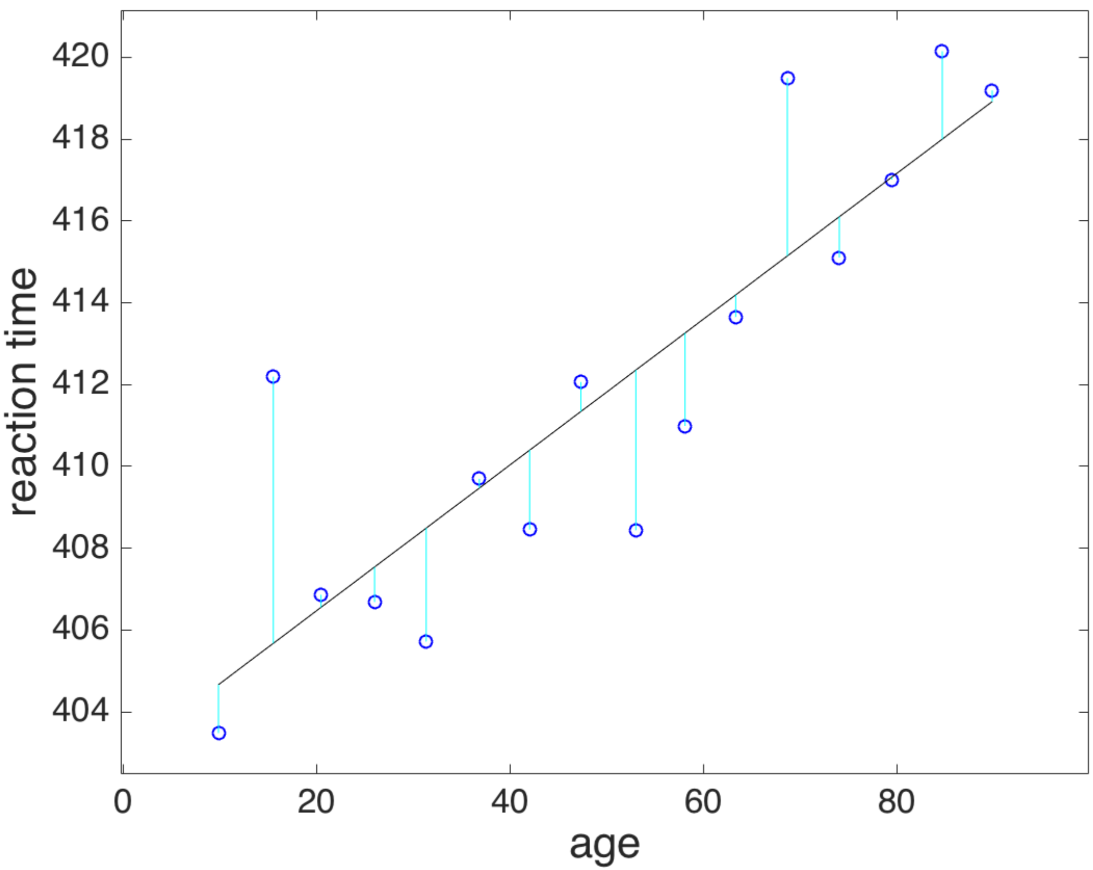
model: $RT = b_0 + b_1 * age + e$



is RT related to age?

parameters in the population

model: $RT = b_0 + b_1 * age + e$



parameters estimated from the sample with normal equations

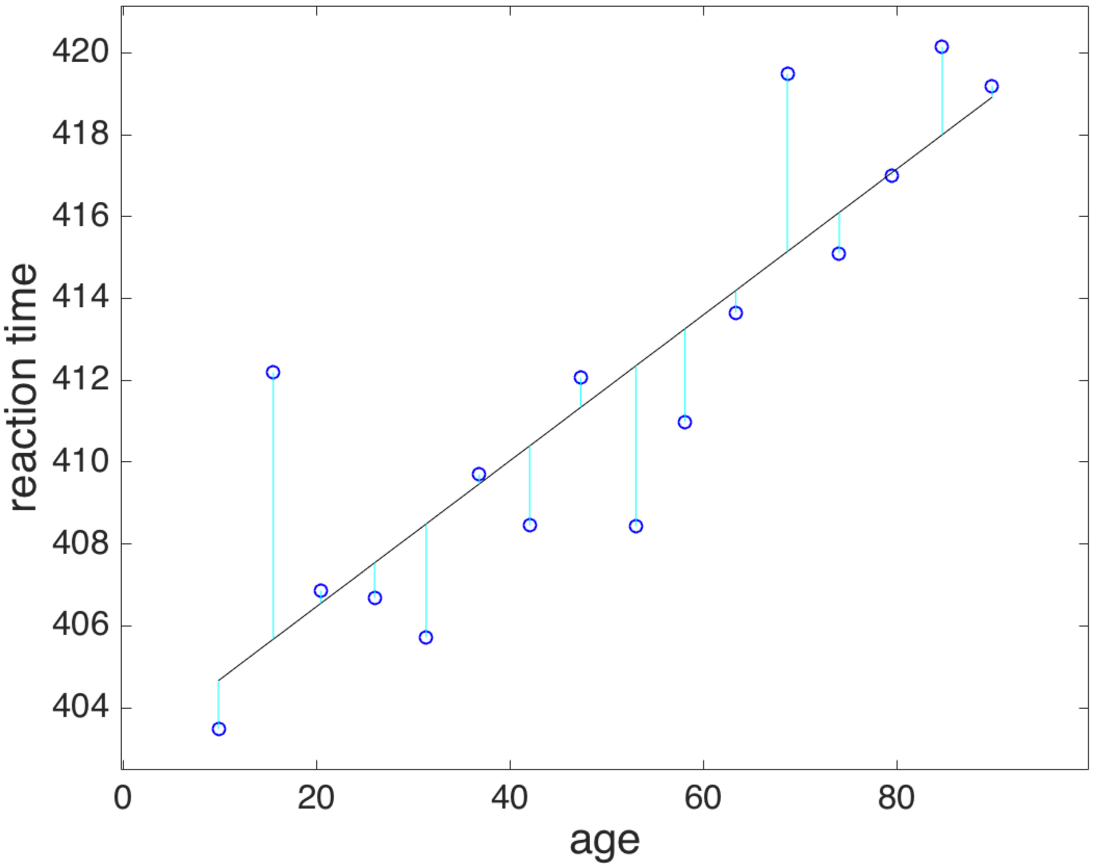
$$b_{est}=(X'X)^{-1}X'y$$

$\hat{b}_0 = 402.91$

$\hat{b}_1 = 0.18$

is RT related to age?

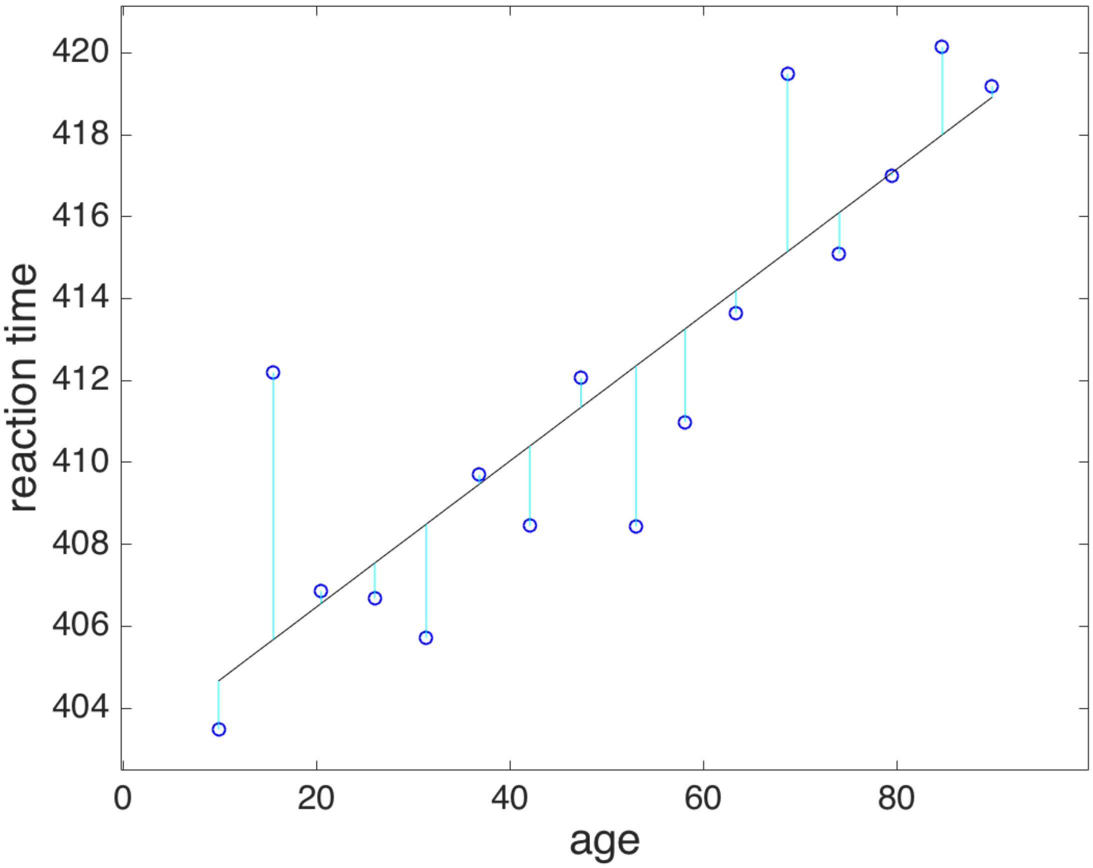
null hypothesis: $b_1 = 0$
alternative: $b_1 \neq 0$



is RT related to age?

null hypothesis: $b_1 = 0$
alternative: $b_1 \neq 0$

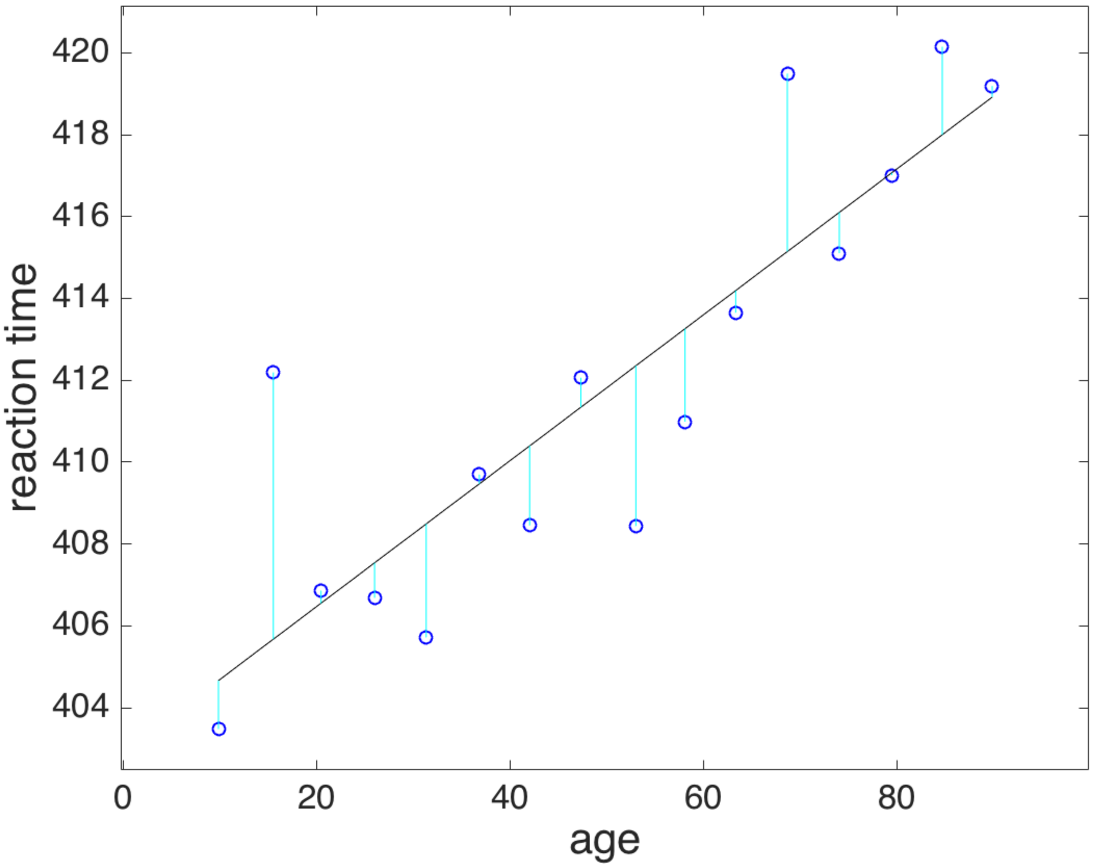
how likely is the parameter estimate ($\hat{b}_1 = 0.18$) if the null hypothesis is true?



is RT related to age?

null hypothesis: $b_1 = 0$
alternative: $b_1 \neq 0$

how likely is the parameter estimate ($\hat{b}_1 = 0.18$) if the null hypothesis is true?



$t = 6.49$

p-value < 0.001

$R^2 = 0.75$

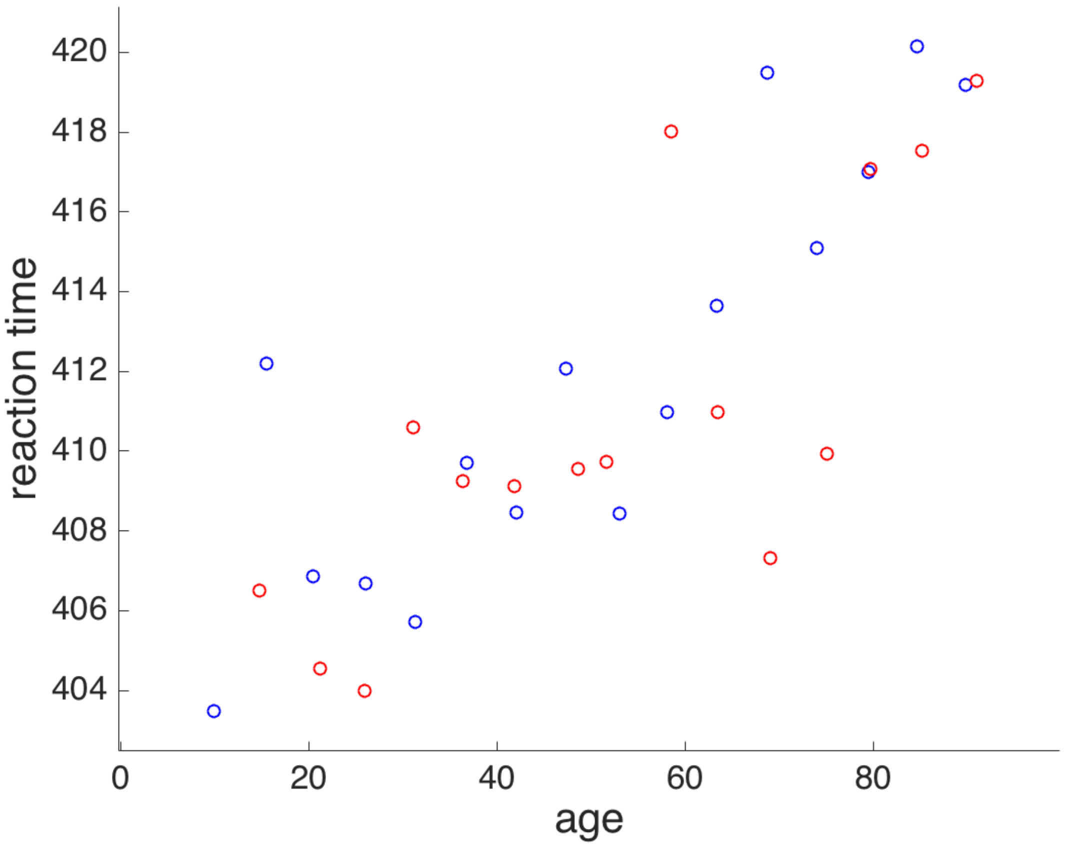
what can we conclude?

- from this sample
  - $p < 0.001$ – reject null hypothesis that ”RT is unrelated to age”
  - R^2 - age accounts for 75% of variance in RT
  - 95% confidence interval

$$\hat{b}_1 = 0.18 \quad \begin{bmatrix} 0.1193 \\ 0.2370 \end{bmatrix}$$

- the test does not tell us
  - how well we can predict RT from age in the population
  - whether this is the right model (or at least better than others)
  - whether or how age causes reaction time (or vice versa)

what happens with a new sample?



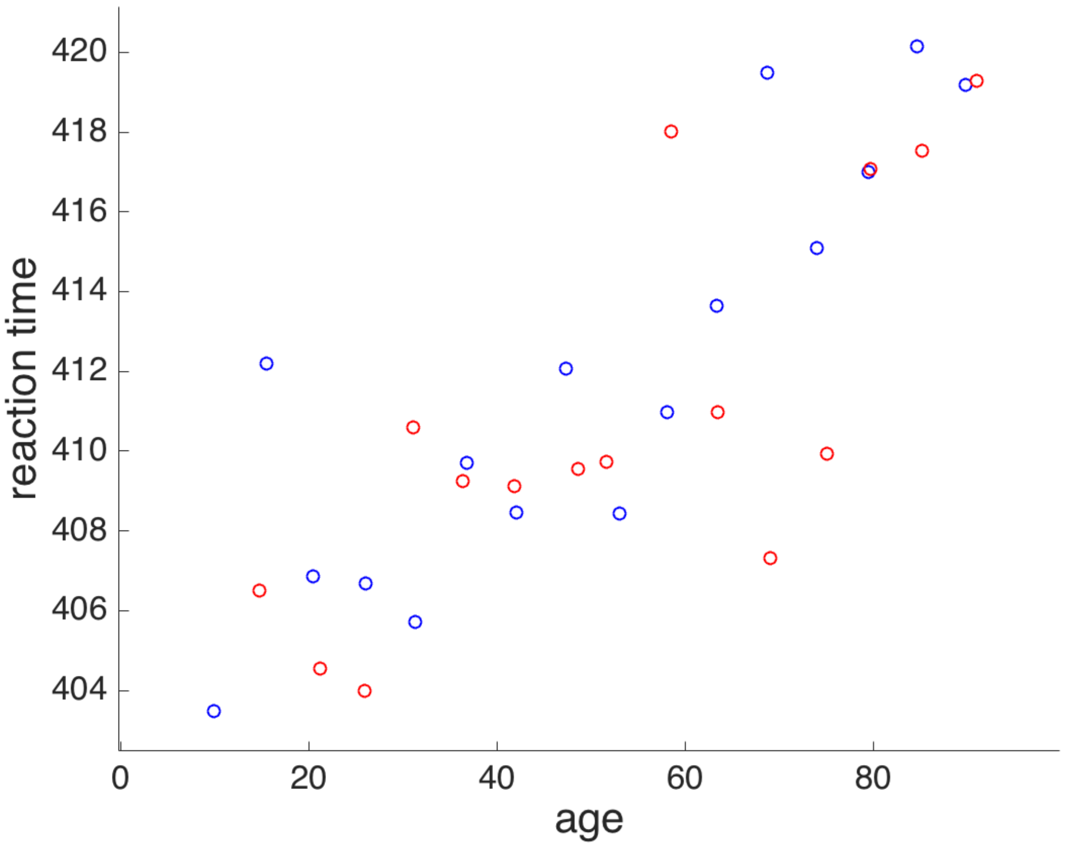
draw a new sample from the same population

compute the R^2 using parameters estimated in the original sample

R^2 = 0.65 (new sample)

R^2 = 0.75 (original sample)

what happens with a new sample?



draw 100 new samples

using model parameters estimated from the original sample,

average R^2 = 0.71

an estimate of how good the original model would be on any new sample

description vs prediction perspectives

estimating model parameters:

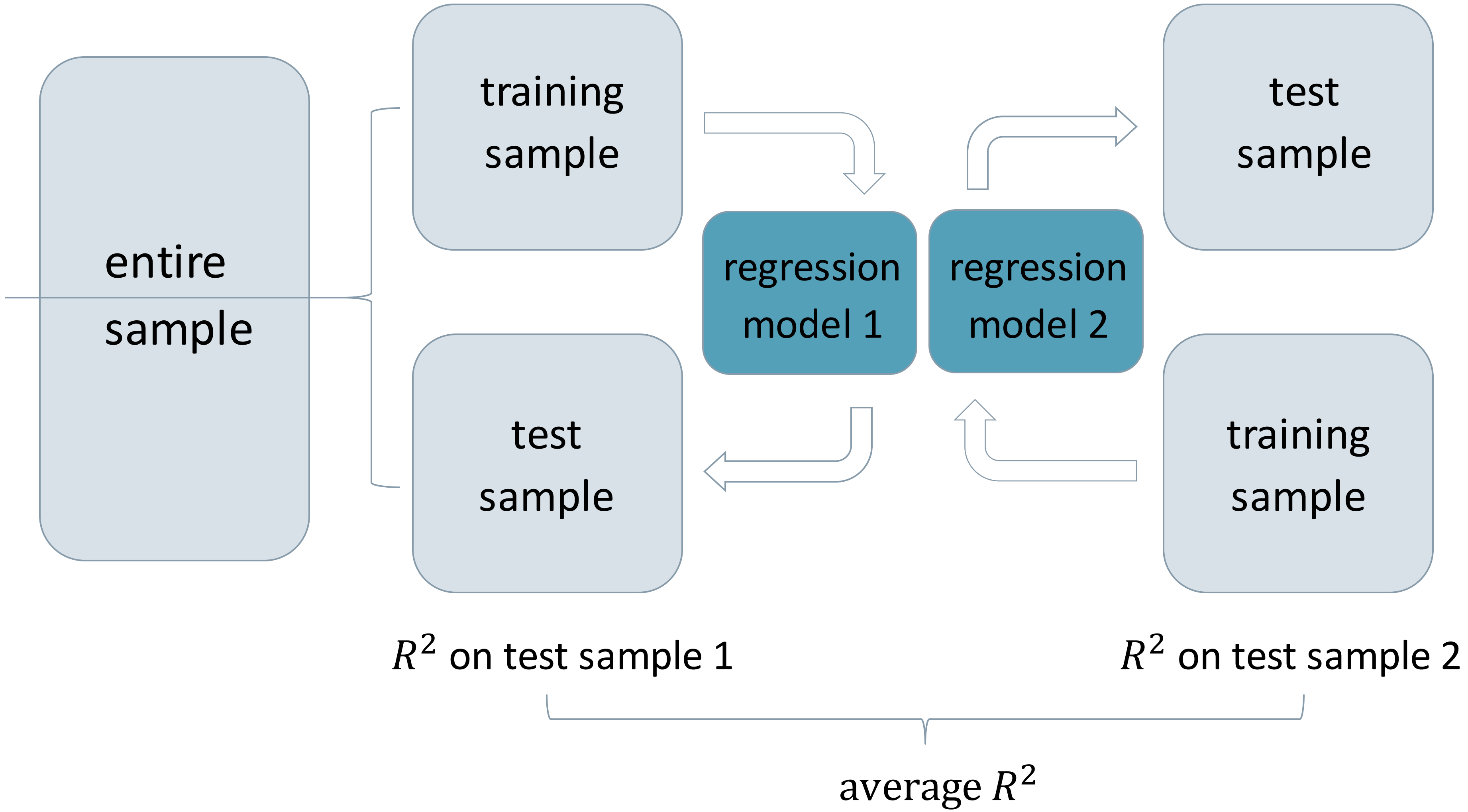
- description: describe variable relations in terms of few parameters
- prediction: learn about to model variable relation from training sample

evaluating the model

- description: goodness of fit, for limited model complexirty
- prediction: apply the model to a test sample not used in learning the model

but what if you cannot get more data?

there are two samples inside your sample...



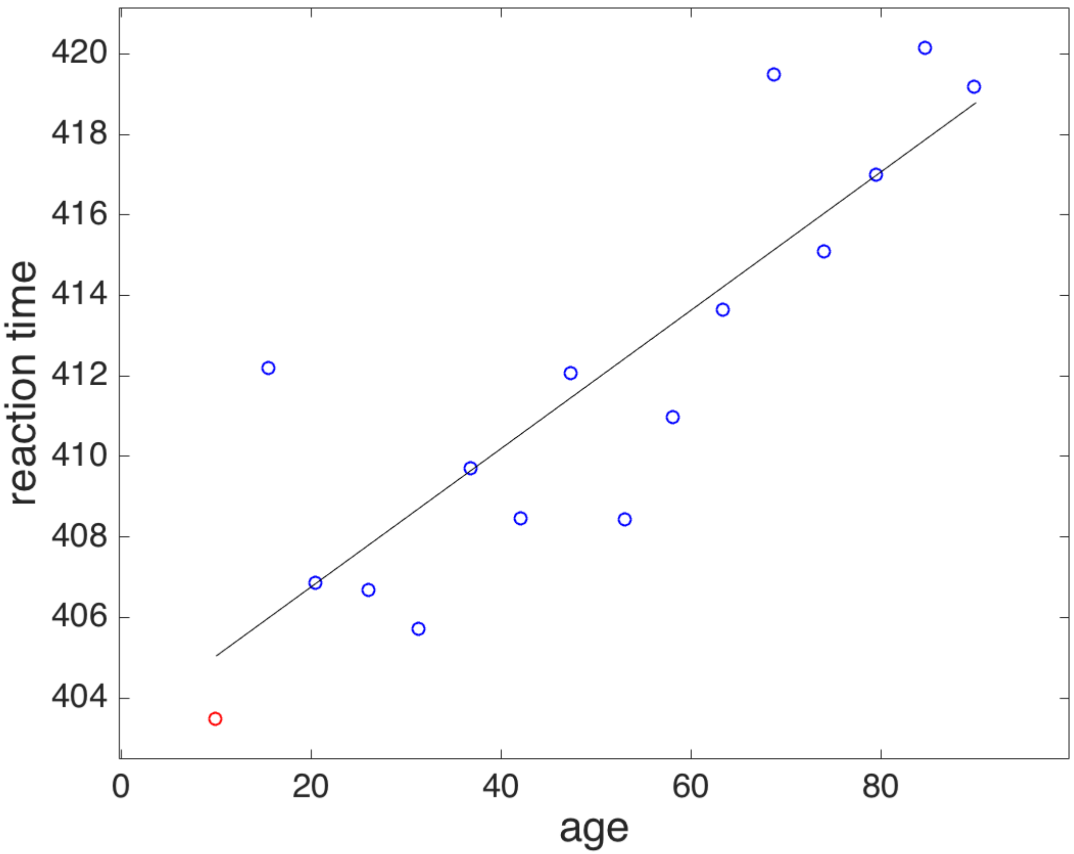
cross-validation

k-fold cross-validation:

- split into k folds
- train on k-1, test on the left out, iterate
- calculate average prediction measure across all k folds

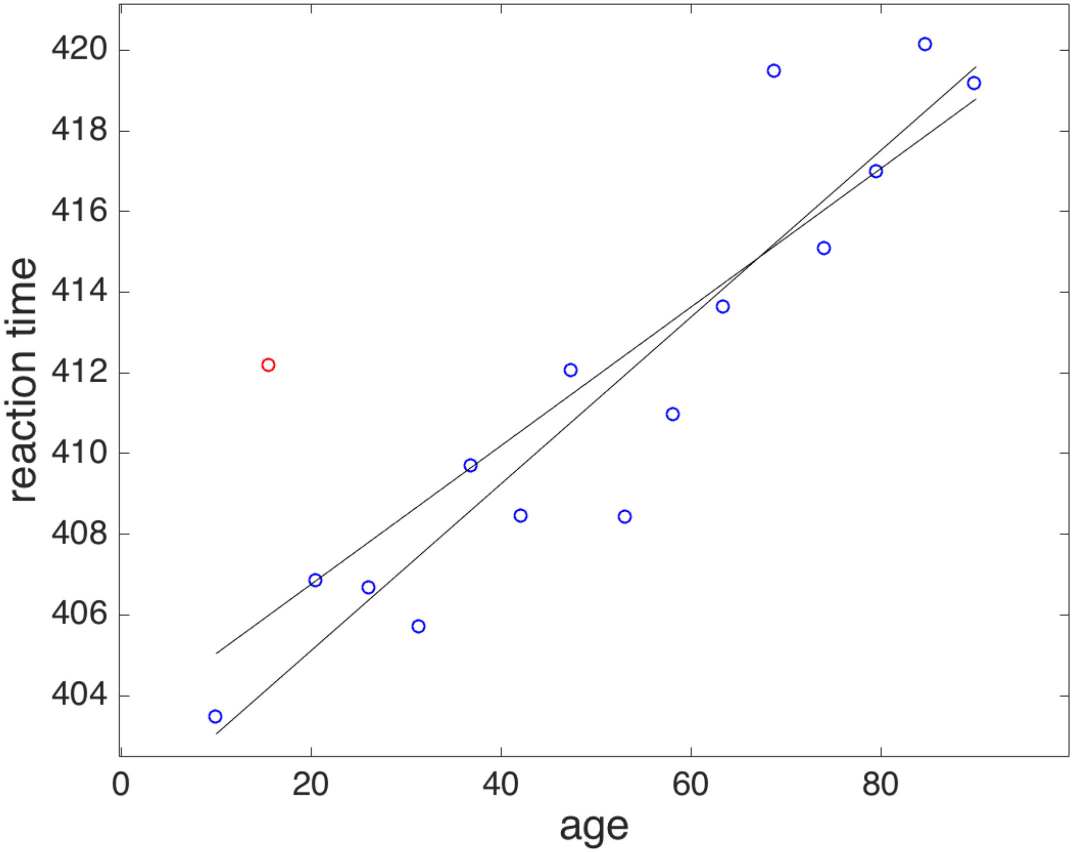
leave-one-out cross-validation

leave out the first data point, fit model to the others



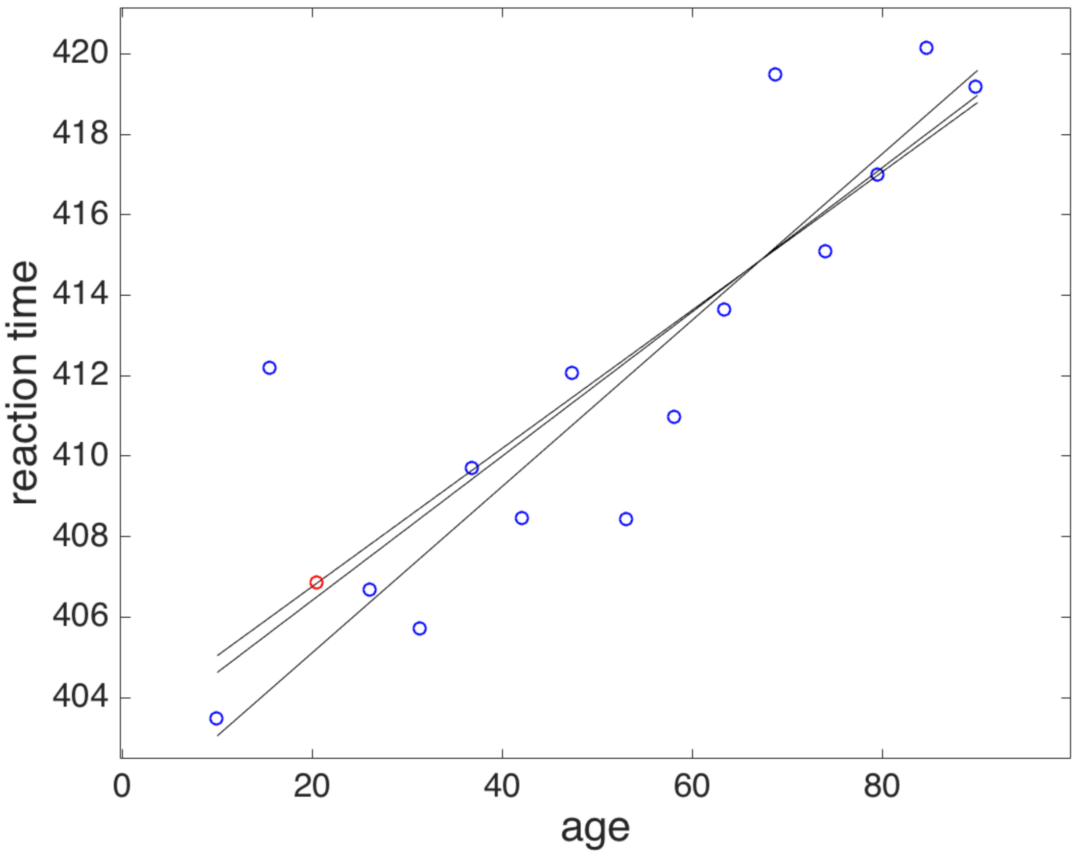
leave-one-out cross-validation

leave out the second data point, fit model to the others



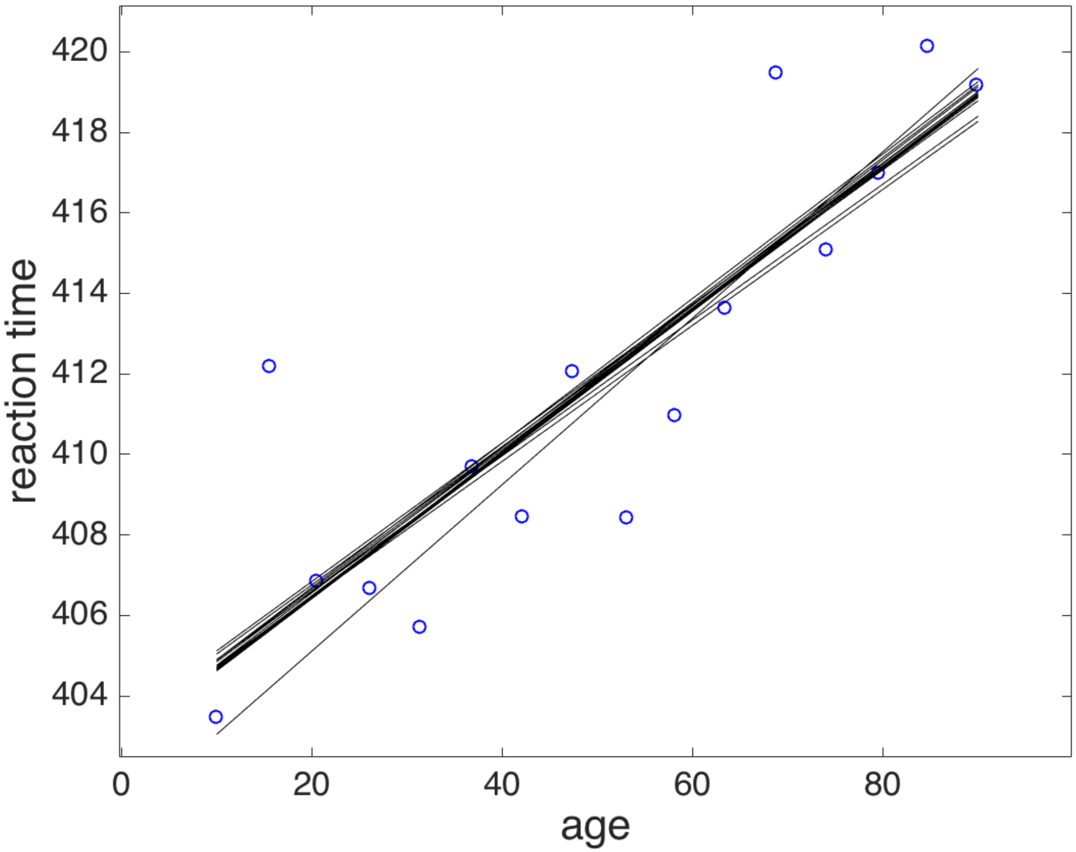
leave-one-out cross-validation

leave out the third data point, fit model to the others



leave-one-out cross-validation

all leave-one-out regression lines



leave-one-out
$R^2 = 0.67$

original sample
$R^2 = 0.75$

mean of 100
new samples
$R^2 = 0.71$

cross-validation

k-fold cross-validation:

- split into k folds
- train on k-1, test on the left out, iterate
- calculate average prediction measure across all k folds

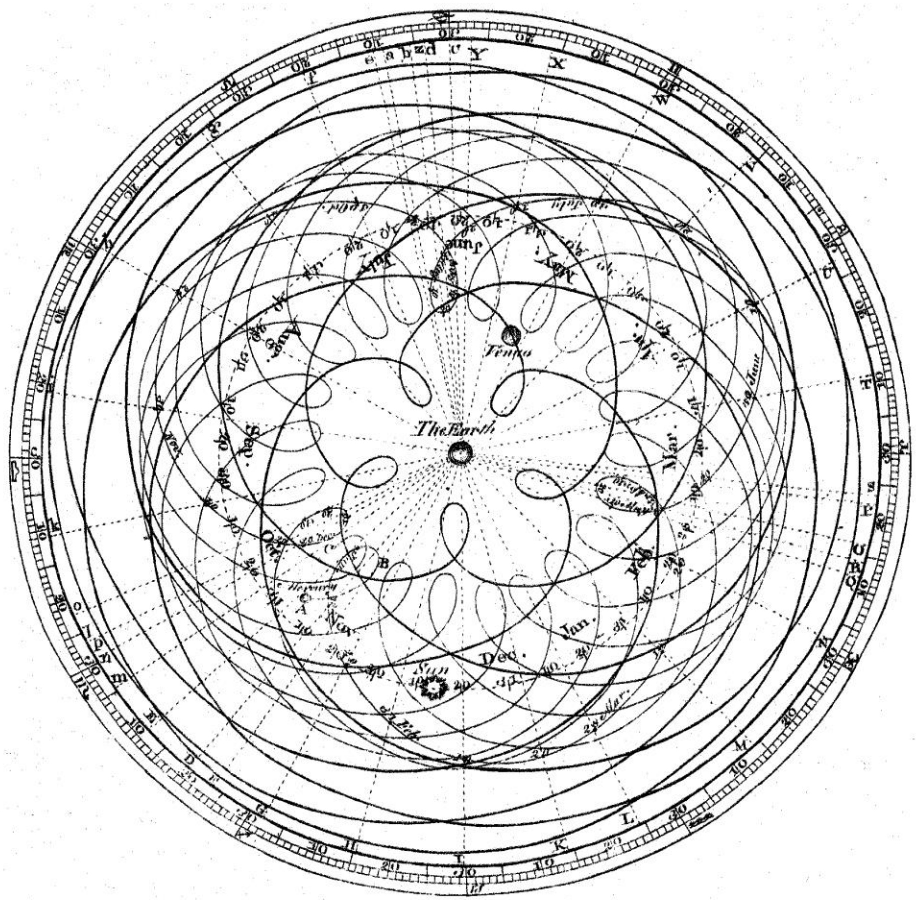
considerations:

- key assumption: models in different folds are very similar
- typical schemes are 10-fold or leave-one-out (more expensive, other issues)
- can be conservative and high variance, especially for small samples
- mistakes are easier to make than with separate train/test samples
- recommended reading:

  "Assessing and tuning brain decoders: cross-validation, caveats, and guidelines"

  Varoquaux et al. 2017

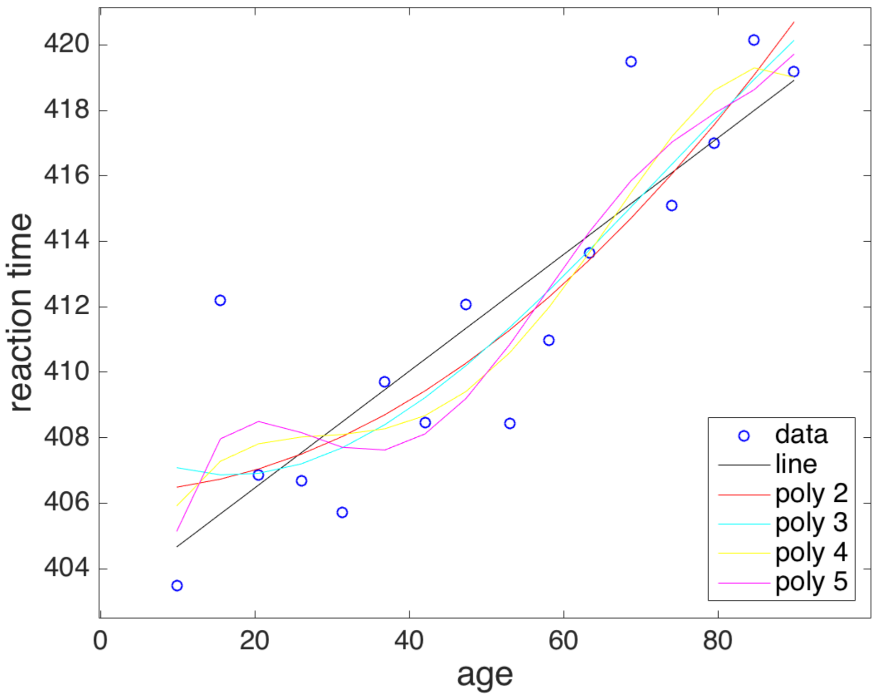
model complexity

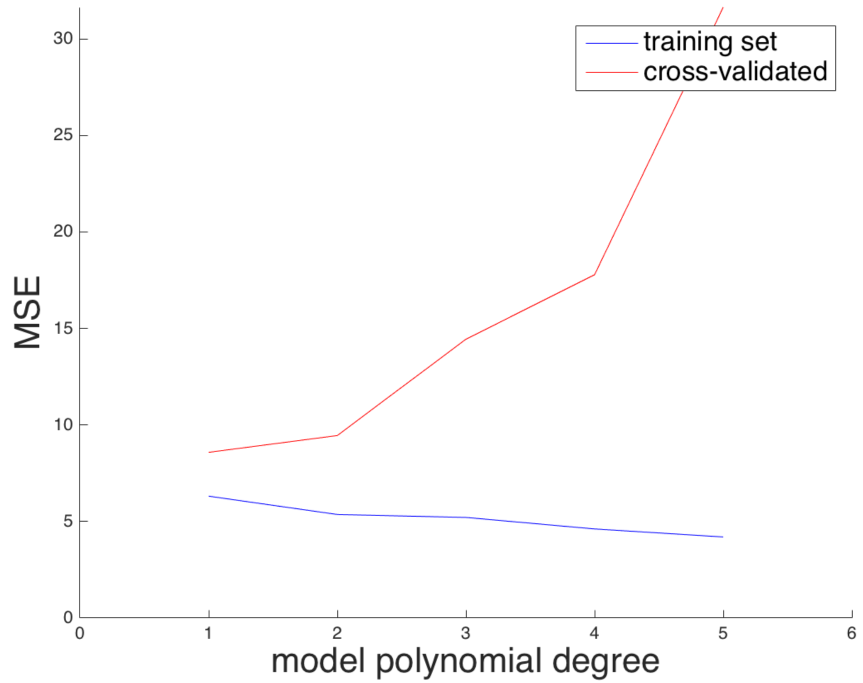


as model complexity goes up,

we can always fit the training data better

model complexity

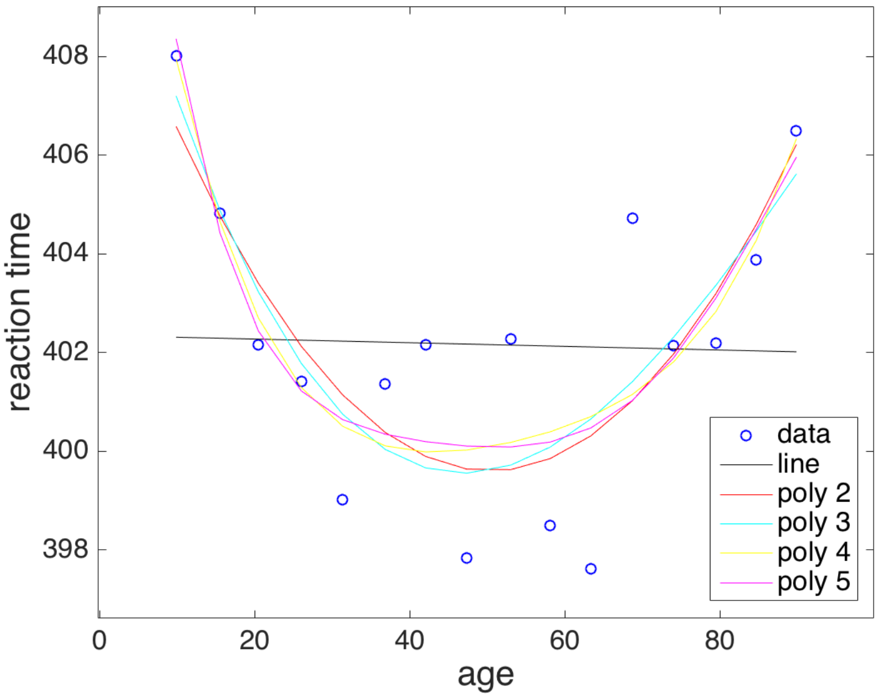


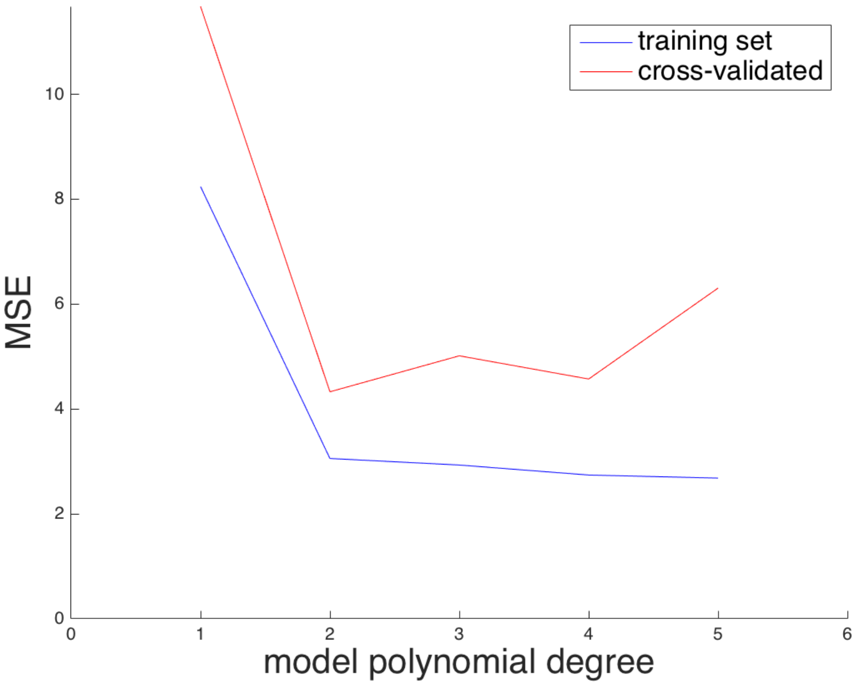


polynomials of higher degree fit the training data better...

... but they do worse on test data (overfitting)

model complexity





if the relationship in the population were more complicated, a line would be too simple (underfitting)...

... but cross-validation can show us a reasonable model complexity!

"All models are wrong, but some are useful.”

George Box

what is machine learning, redux

- generalization: ability to make predictions about new data

- a model that generalizes well
  - shows that there is information in the data about a prediction target
  - can be dissected to understand how the prediction can be made

but what does this have to do with brains?

case study: tools vs buildings

[data from Rob Mason and Marcel Just, CCBI, CMU]

- subjects read concrete nouns in 2 categories
  - words name either tool or building types
  - trial:

    see a word

    think about properties, use, visualize

    blank

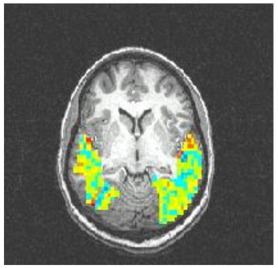
3 seconds

8 seconds

- average images around response peak

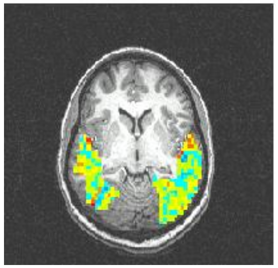
  to get one labelled image per trial

  (84 trials in 6 runs)



tools

case study: tools vs buildings



average trial image

example

...

tools

voxels (features)

class label

training data (42)

labels

run 1 ...

run 3 ...

run 5 ...

test data (42)

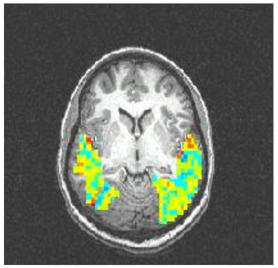
labels

run 2 ...

run 4 ...

run 6 ...

case study: tools vs buildings



average trial image

example

...

tools

voxels (features)

class label

training data (42)

labels

run 1 ...

run 3 ...

run 5 ...

test data (42)

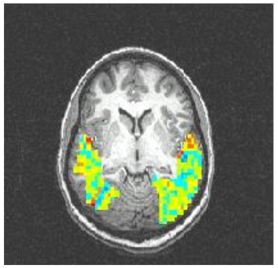
labels

run 2 ...

run 4 ...

run 6 ...

case study: tools vs buildings



average trial image

example

...

tools

voxels (features)

class label

training data (42)

labels

run 1
run 3
run 5

...
...
...

classifier

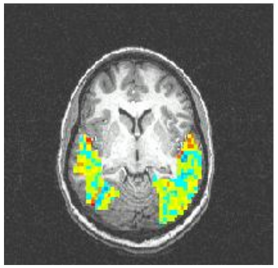
test data (42)

labels

run 2
run 4
run 6

...
...
...

case study: tools vs buildings



average trial image

example

...

tools

voxels (features)

class label

training data (42)

labels

run 1 ...

run 3 ...

run 5 ...

classifier

test data (42)

labels

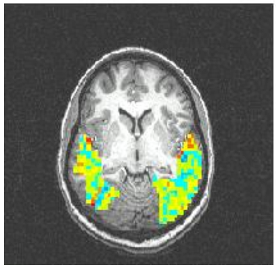
run 2 ...

run 4 ...

run 6 ...

classifier

case study: tools vs buildings



average trial image

example

...

tools

voxels (features)

class label

training data (42)

labels

run 1 ...
run 3 ...
run 5 ...

classifier

predicted labels

test data (42)

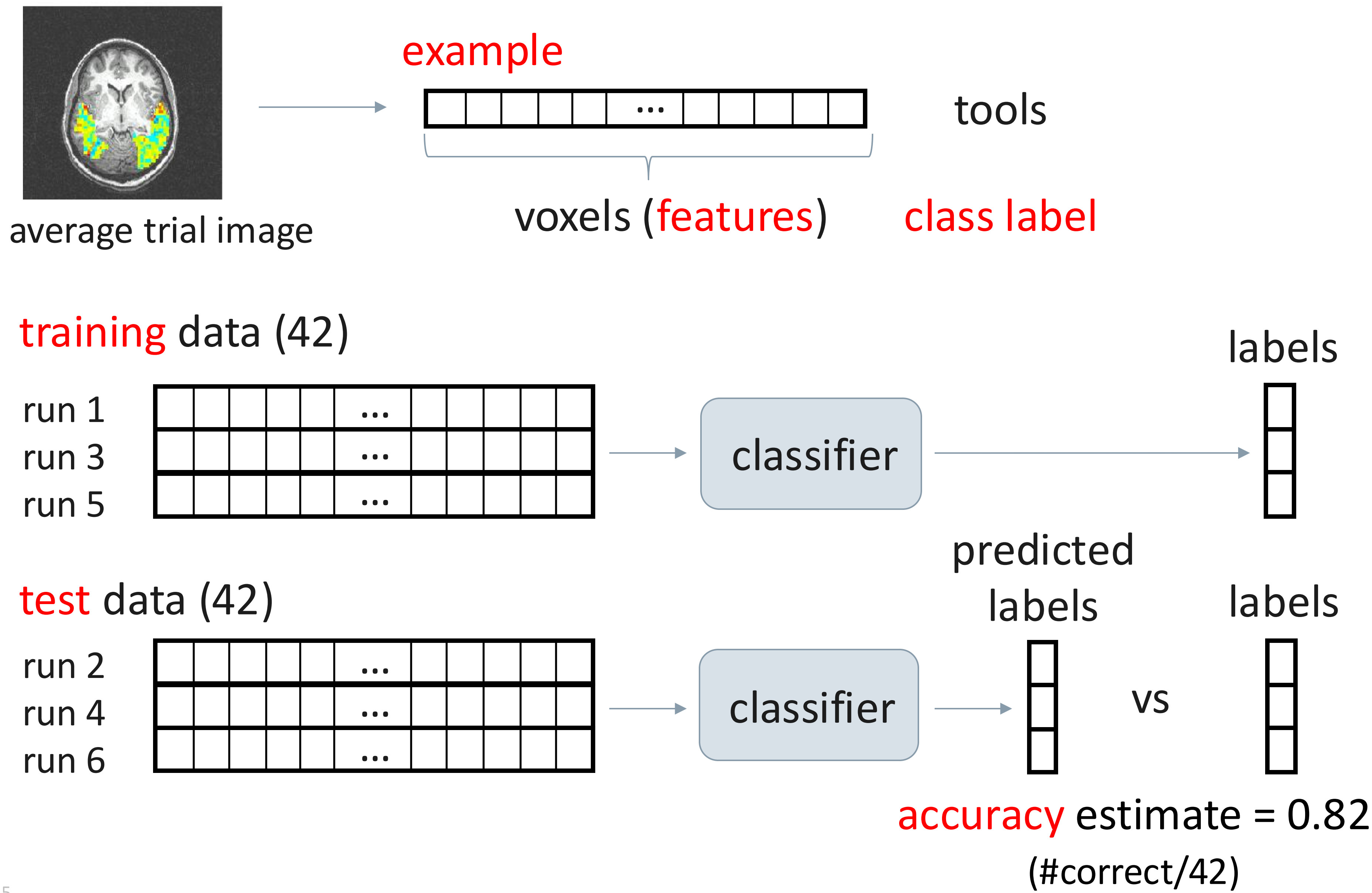
labels

run 2 ...
run 4 ...
run 6 ...

classifier

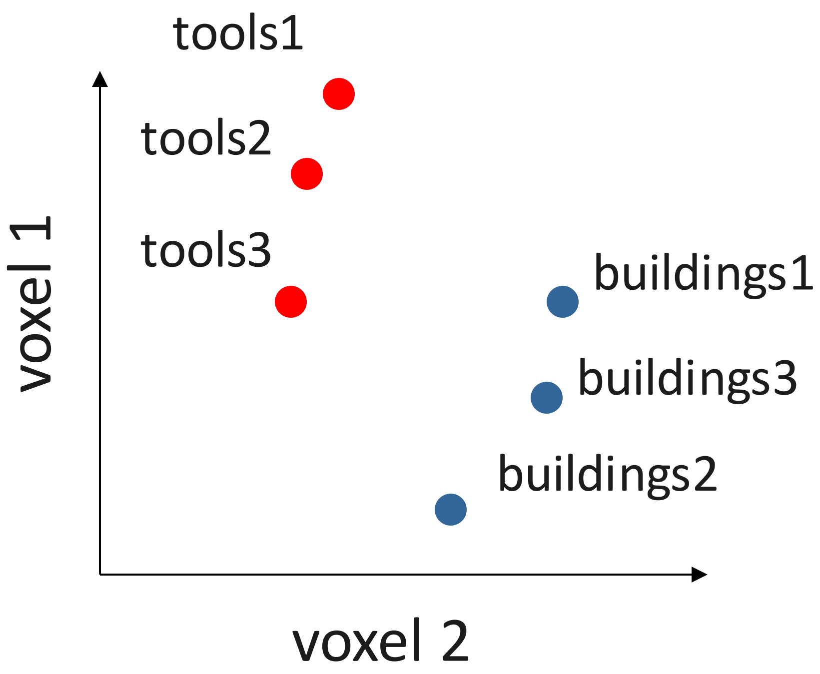
vs

case study: tools vs buildings

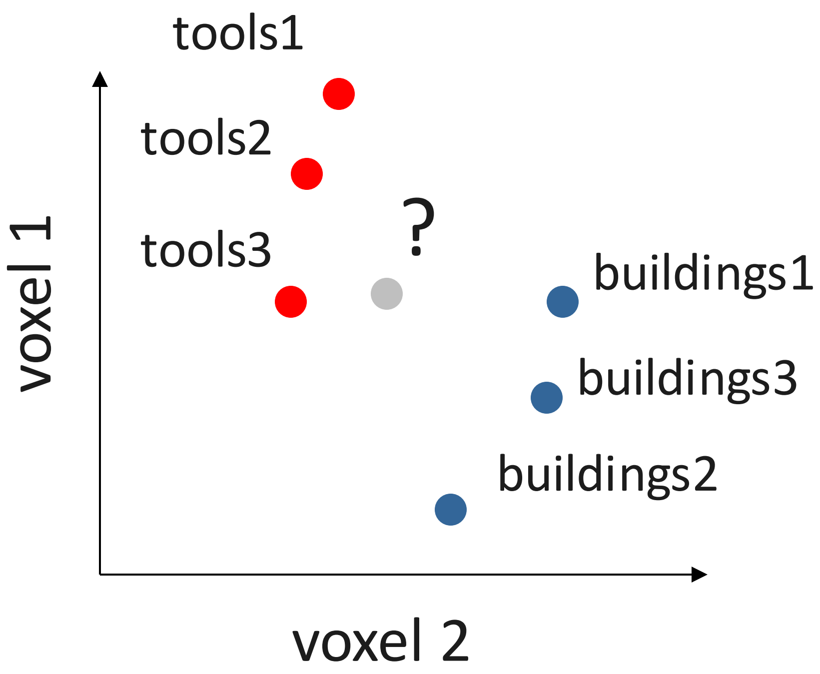


what is inside
the grey box?

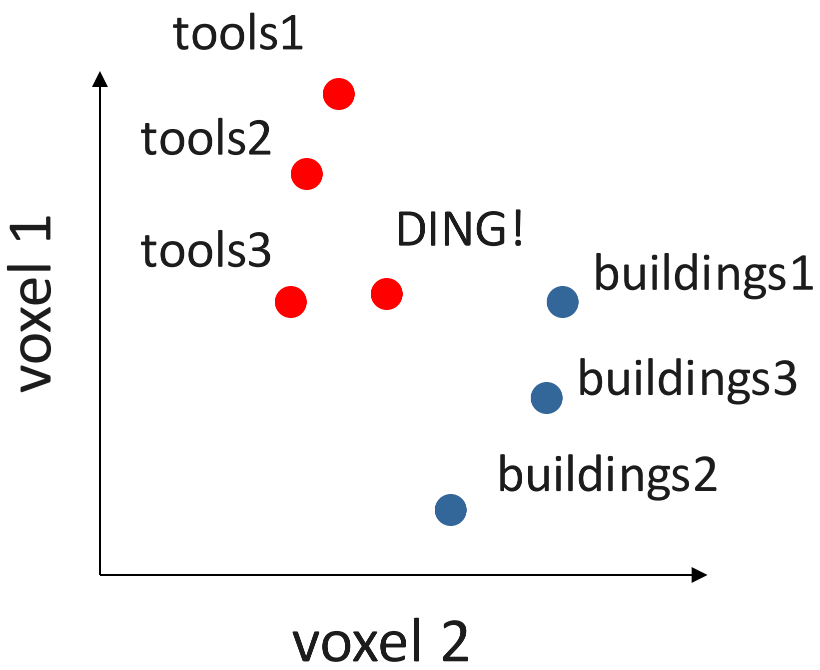
inside the grey box



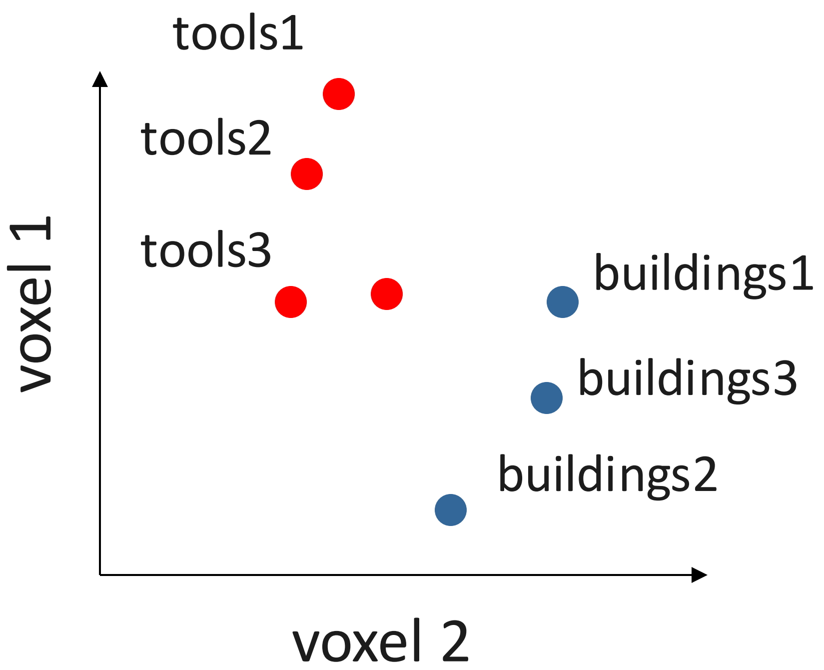
inside the grey box



inside the grey box



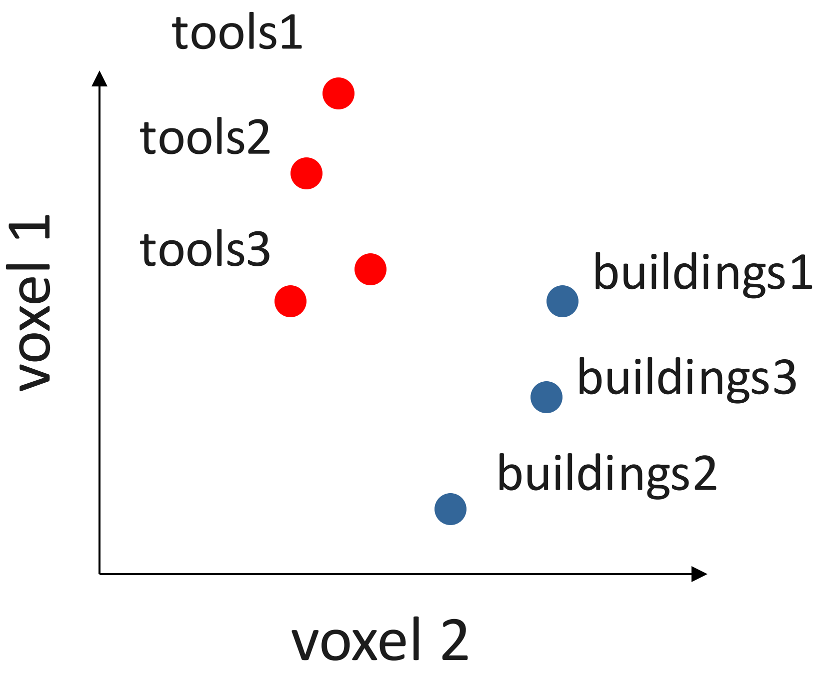
inside the grey box



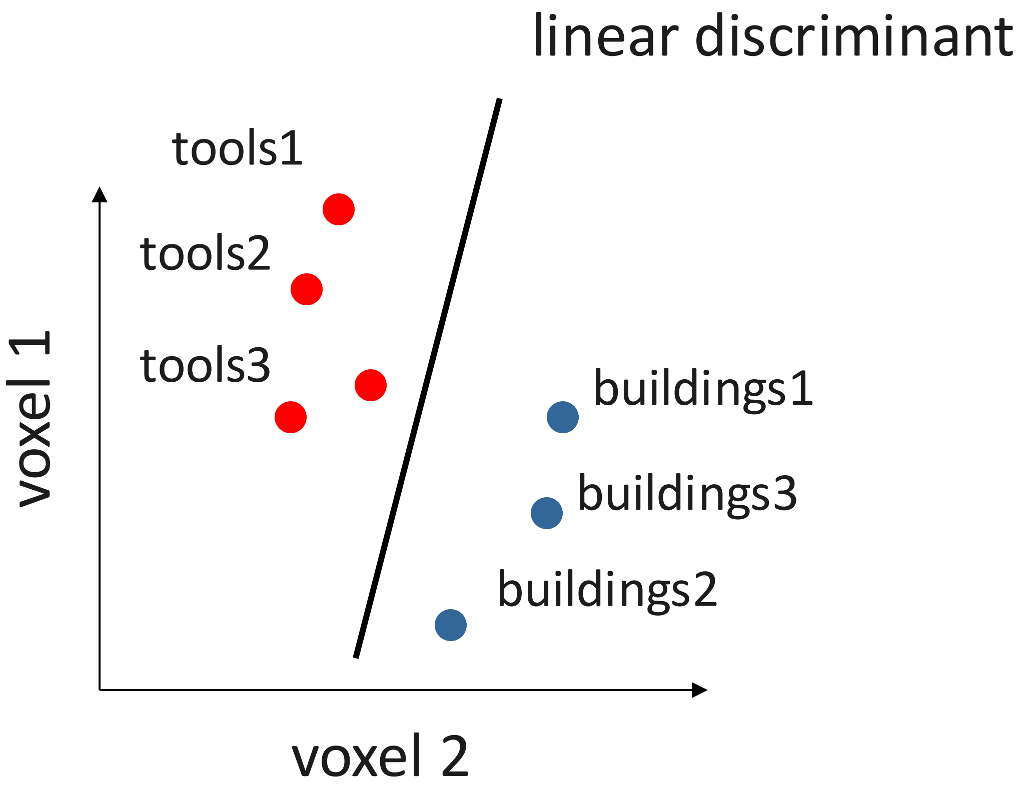
simplest function is no function at all: “nearest neighbour”

- implicit example similarity/distance measure
- can use more points in decision (k-nearest ...)

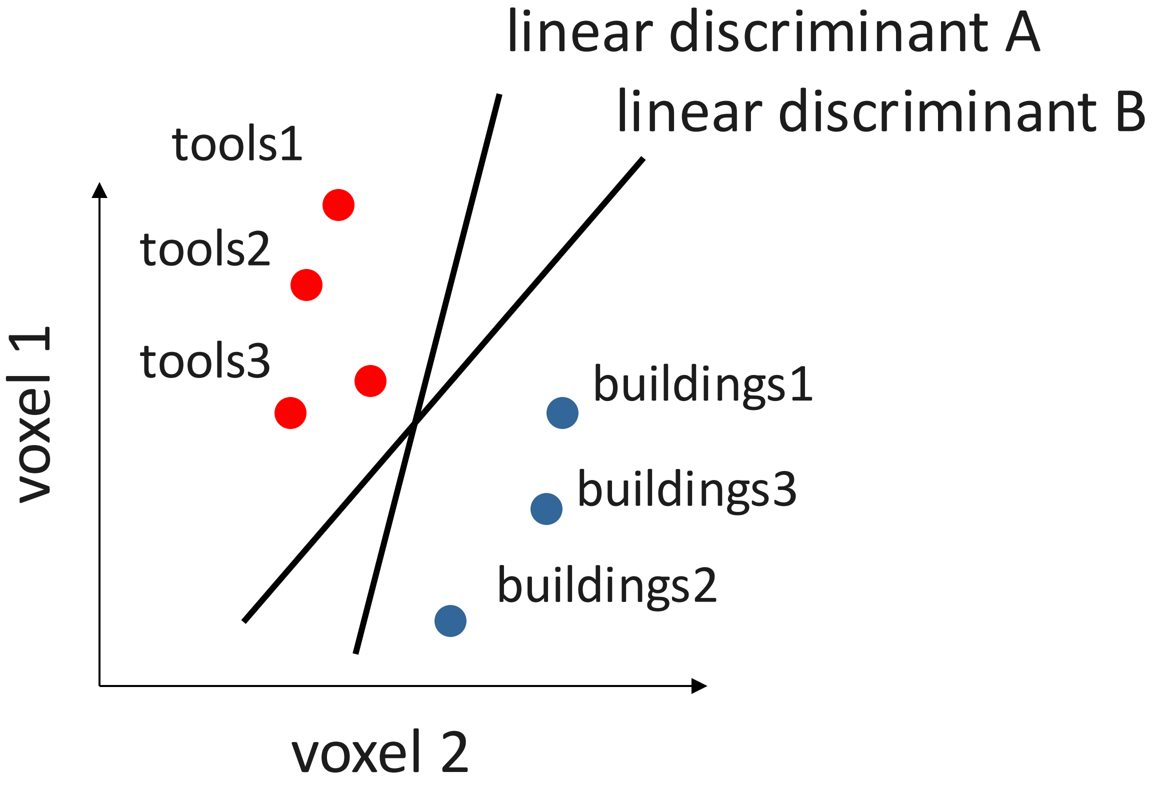
inside the grey box



inside the grey box

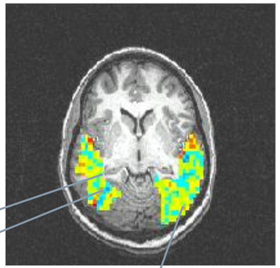


inside the grey box



- there are many possible linear discriminants
- LDA, logistic regression, linear SVM, ...

inside the grey box



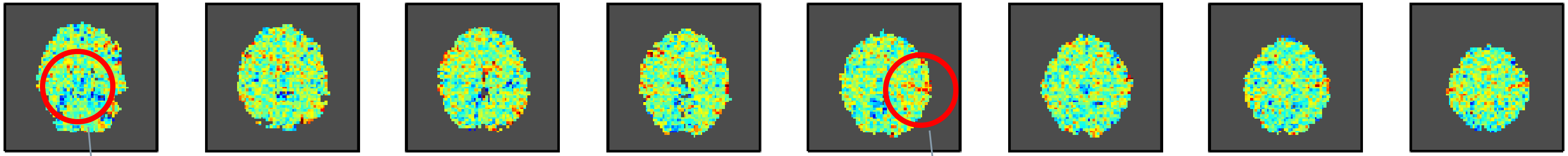
weight0 + weight1 x + weight2 x + + + weight n x

If [voxel 1 | voxel 2 | ... | | | voxel n] > 0 tools

otherwise buildings

decision value / label probability

classifier weights (linear Support Vector Machine)



- weights pull towards buildings

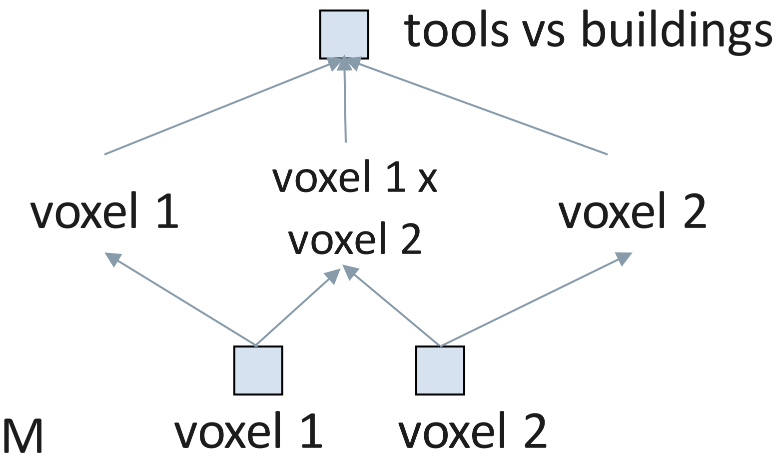
+ weights pull towards tools

inside the grey box – nonlinear classifiers

linear on a transformed feature space!

SVMs

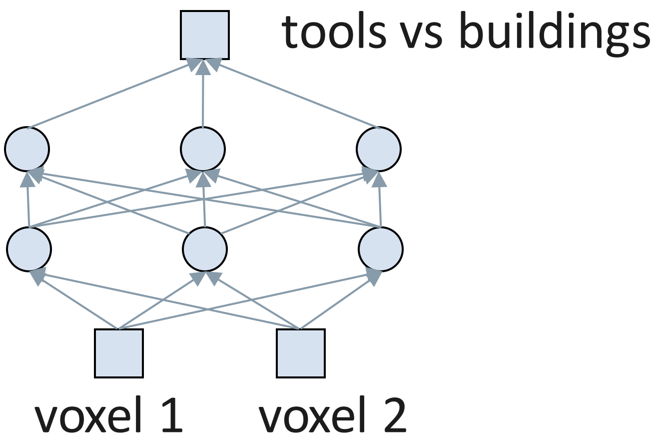
new features are (implicitly) determined by a kernel



quadratic SVM

neural networks:

new features are learned, and features of features,...



“Improving the Interpretability of fMRI Decoding using Deep Neural Networks and Adversarial Robustness” McClure et al. 2023 how do we test a classification result?

how do we test predictions?

true labels	predicted labels		
tools	tools		
tools	buildings		error
buildings	buildings		
…	…	⋮	
buildings	tools		error
buildings	buildings		
tools	tools		

accuracy:

#correct

out of

#test

null hypothesis:

"classifier learnt nothing" → "predicts randomly"

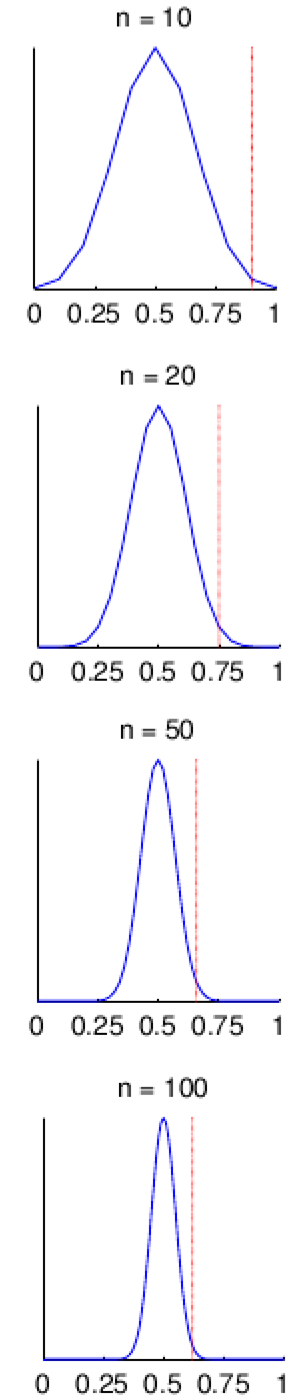
how do we test predictions?

null hypothesis: classifier learned nothing

- X = #correct
- P(X|null is true) is binomial(#test,0.5)
- p-value is P(X >= result to test|null is true)

many caveats:

- accuracy is an estimate
- few examples → very uncertain
- many examples → easy to be significant
- must correct for multiple comparisons



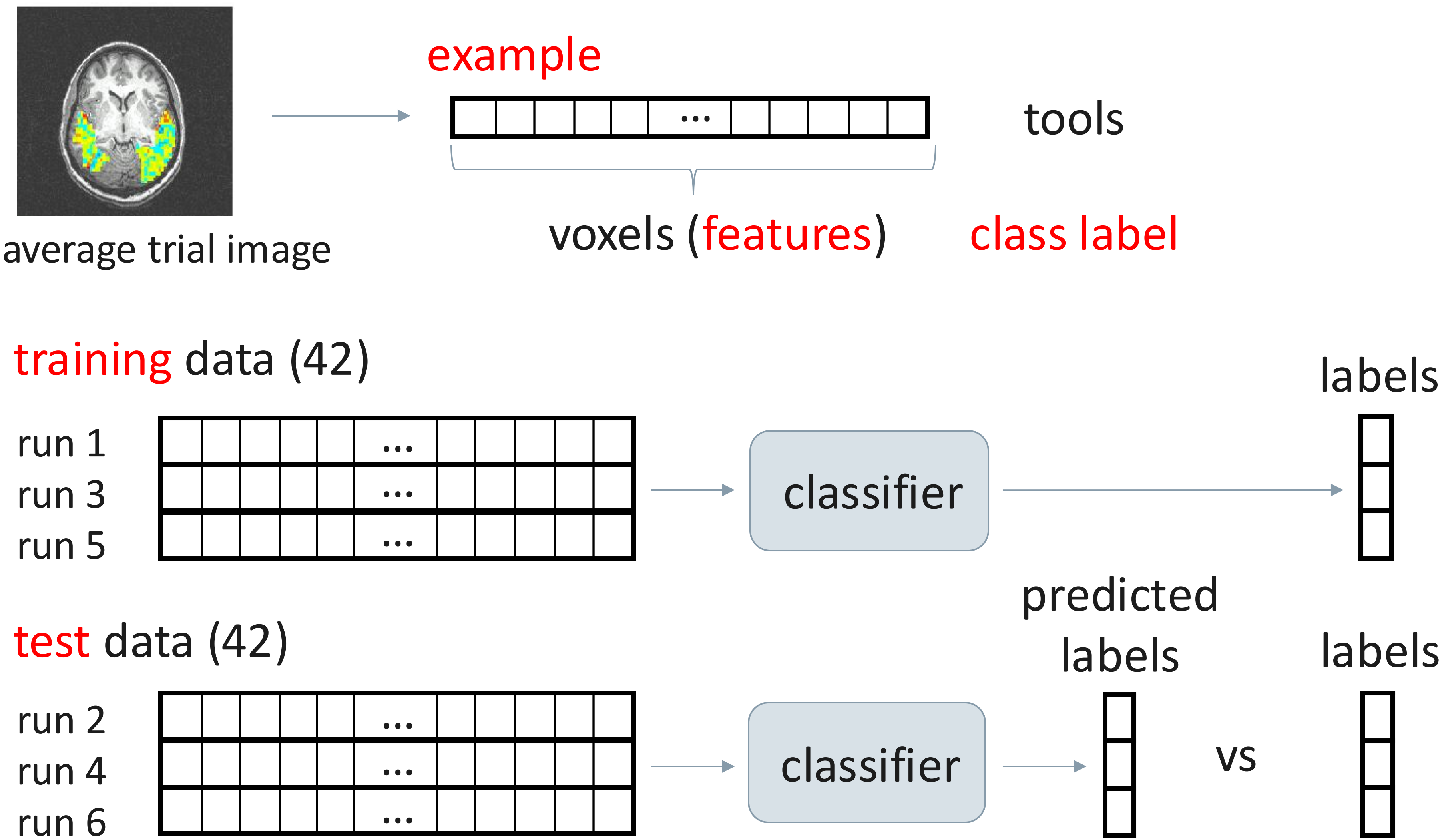
distribution under null
(0.05 p-value cut-off)

feature and example selection

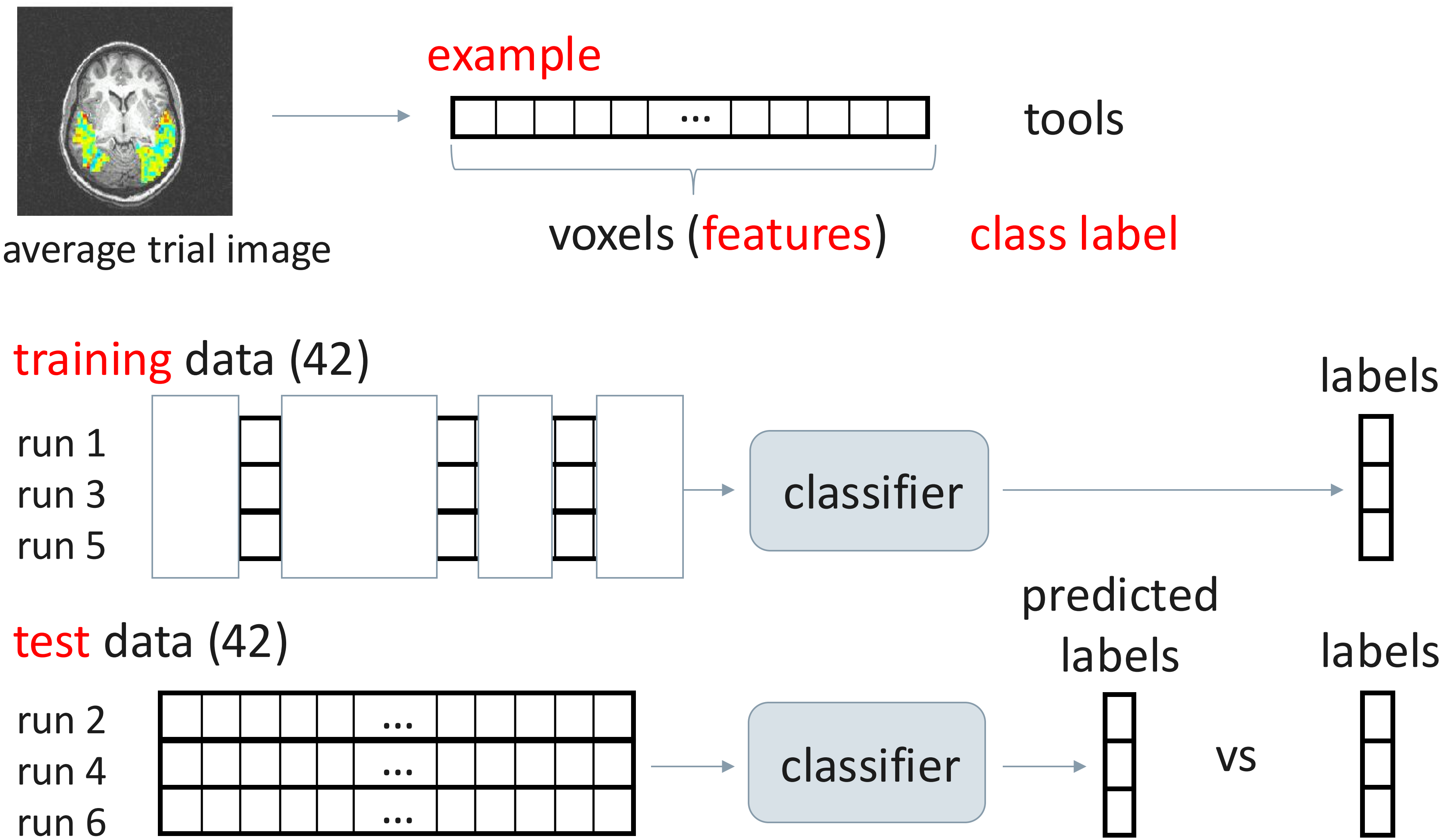
- a classifier answers one question, but often needs help...

- restrict voxels by
  - space (e.g. anatomical ROI, a priori ROI, etc)
  - time (e.g. different points in a trial)
  - behaviour (e.g. selective for a condition, consistent across them)

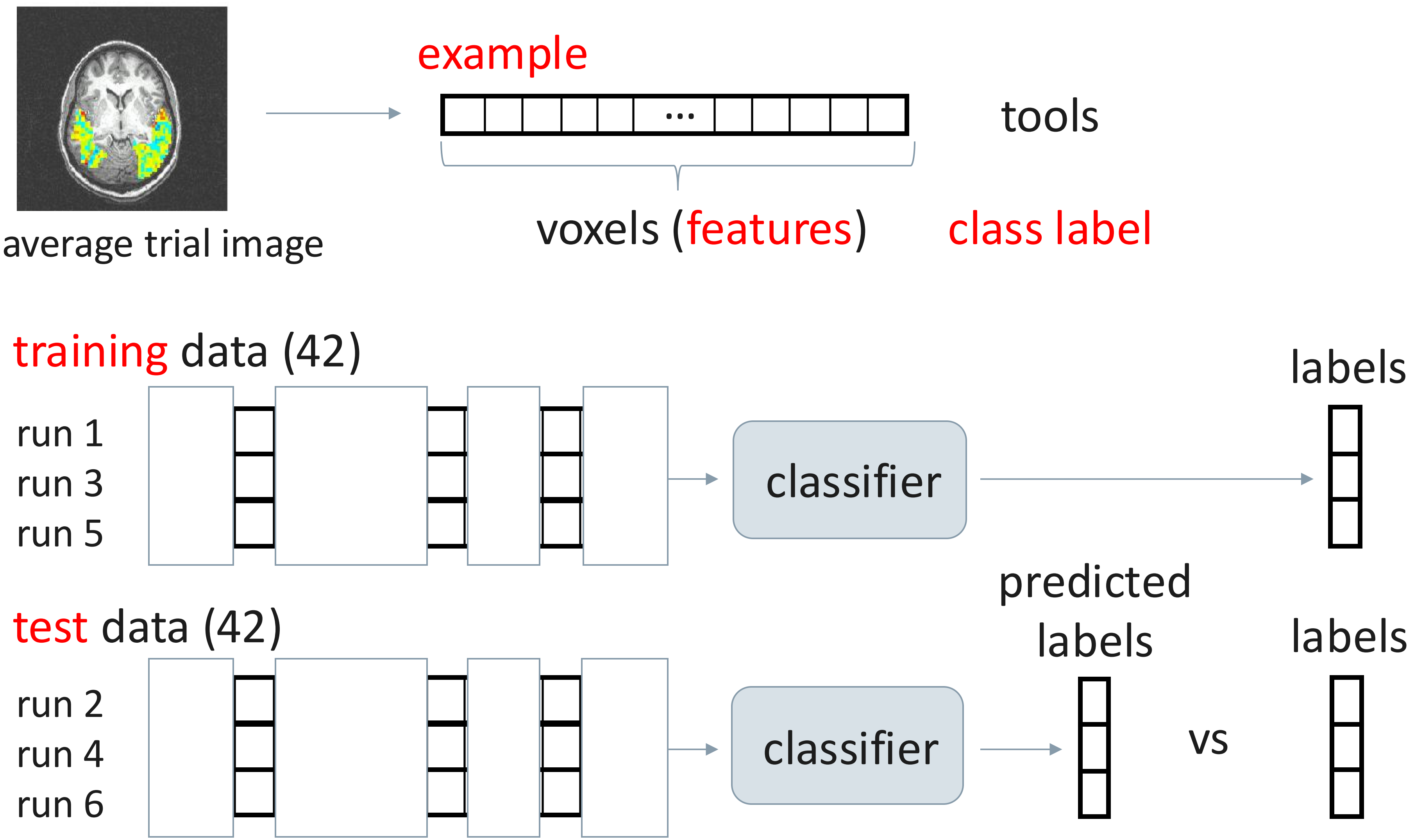
feature selection



feature selection



feature selection



feature and example selection

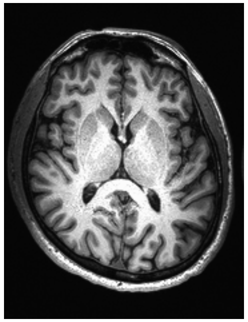
- a classifier answers one question, but often needs help...

- restrict voxels by
  - space (e.g. anatomical ROI, a priori ROI, etc)
  - time (e.g. different points in a trial)
  - behaviour (e.g. selective for a condition, consistent across them)

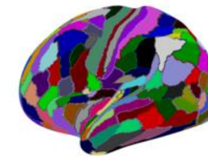
- restrict examples by
  - experiment phase (e.g. study versus free recall blocks)
  - trials (e.g. successful or not)

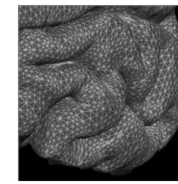
what about other modalities?

structural MRI

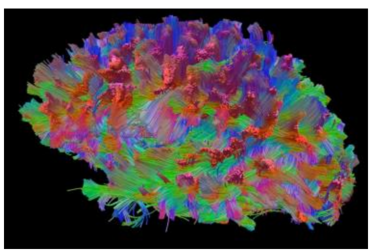


- group voxels into a brain region (parcellation)
- create a surface model (triangle mesh)



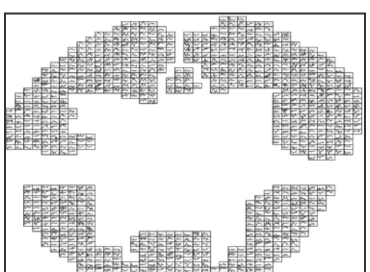


diffusion MRI



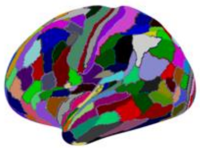
- count tracts passing through each region
- derive structural connectivity matrix

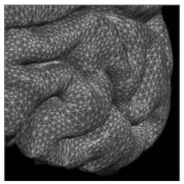
resting state fMRI



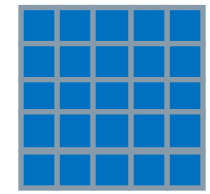
- create average time series per region
- calculate correlation between them
- derive functional connectivity matrix

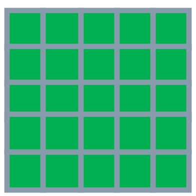
what about other modalities?



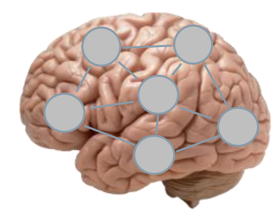


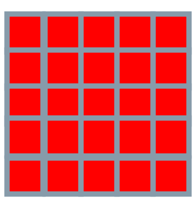
- gray matter volume
- cortical thickness
- surface area
- covariance of measures between regions





- reduce connectivity to region pairs or networks
- matrices => graphs => graph-theory measures
- dynamic versions (over time windows)





what about other modalities?

- problems to solve
  - classification: patients vs controls, treatment or disease outcome, ...
  - regression: symptom intensity, time to symptoms, subject characteristics
  - clustering: patient groups
- feature selection
  - region-of-interest or network restriction
  - t-test for individual matrix entries (within training set)
- other issues
  - interpreting classifier weights (aggregate by ROI is typical)
  - combining modalities (all together, meta-classifier, ...)

science is about questions, not methods

- description

  “Are observations explainable in terms of a few (latent) variables?”

- prediction

  “Is the evolution of an outcome variable predictable from observations (or latent variables estimated from them)? How?”

- causality and control

  “How would intervening on some variables affect others?”

- mechanism or computation

  “How does an input get transformed to produce the observations?"

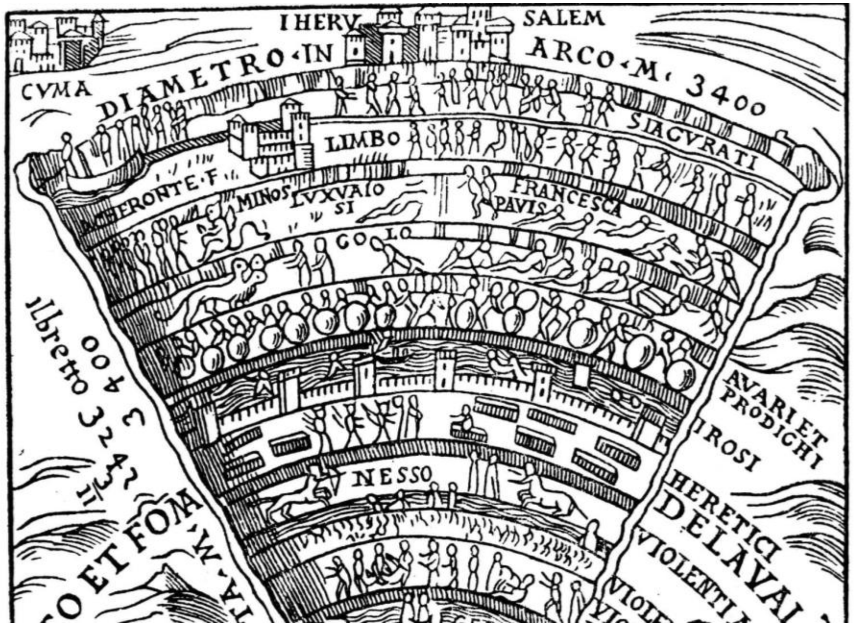
Thank you!

(questions?)

(or email francisco.pereira@nih.gov later)

potential issues

- small sample sizes
- significant but small effect
- class imbalance
- p-hacking
- circularity / double-dipping
- reporting training set results

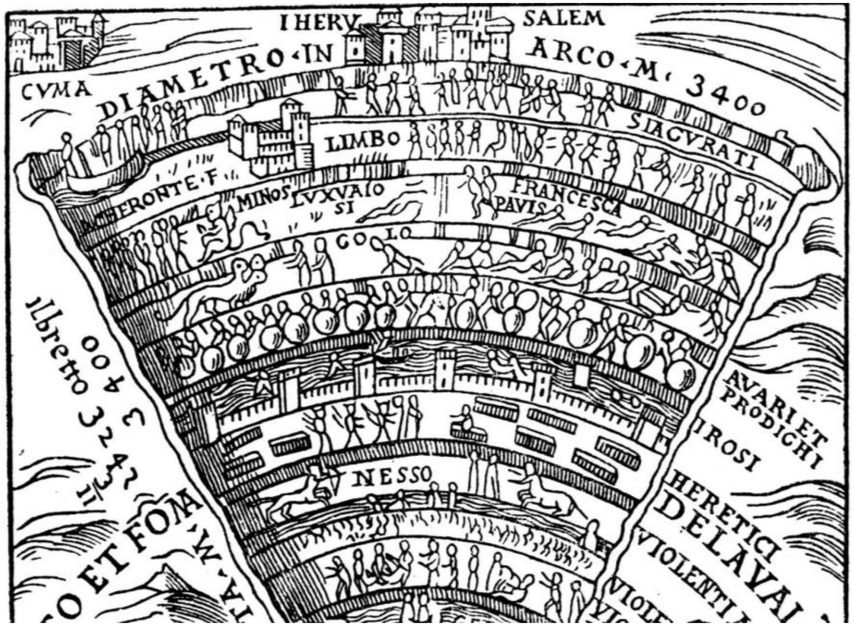


potential issues

- small sample sizes
  - low power is still an issue, even with a separate test set
  - suggestion: require power analysis (past effect sizes may be optimistic...)
- significant but small effect
  - what does 60% accuracy mean?
  - suggestion: error analysis (is there a pattern to errors?)
- class imbalance
  - if one class is more frequent than other, null model is not valid
  - suggestions: (under|over)sample class, nonparametric null

potential issues

- small sample sizes
- significant but small effect
- class imbalance
- p-hacking
- circularity / double-dipping
- reporting training set results



potential issues

p-hacking

- try many things, report a single one -> optimistic bias
- suggestion:
  - make method decisions on sample 1, test on sample 2
  - consider doing a pre-registration before sample 2

potential issues

circularity / double-dipping

- using train+test data to make decisions (e.g. feature selection)
- in the limit, can give you a result where there is none at all
- suggestion:

  always redo the analysis with permuted labels, a few times

  (if results are better than random, there is something wrong)

potential issues

reporting training set results

- vastly optimistic bias (especially for small datasets)
- suggestion: be wary of very high accuracy claims...

science is about questions, not methods

- description

  “Are observations explainable in terms of a few (latent) variables?”

- prediction

  “Is the evolution of an outcome variable predictable from observations (or latent variables estimated from them)? How?”

- causality and control

  “How would intervening on some variables affect others?”

- mechanism or computation

  “How does an input get transformed to produce the observations?"

Thank you!

(questions?)

(or email francisco.pereira@nih.gov later)