An introduction to machine learning for fMRI

Francisco Pereira

Botvinick Lab
Princeton Neuroscience Institute
Princeton University

what is machine learning?

- how to build computer systems that automatically improve with experience

what is machine learning?

- how to build computer systems that automatically improve with experience

- building models of data for
  - predicting numeric variables (regression)
  - predicting categoric variables (classification)
  - grouping data points (clustering)
  - ...

what is machine learning?

- how to build computer systems that automatically improve with experience

- building models of data for
  - predicting numeric variables (regression)
  - predicting categoric variables (classification)
  - grouping data points (clustering)
  - ...

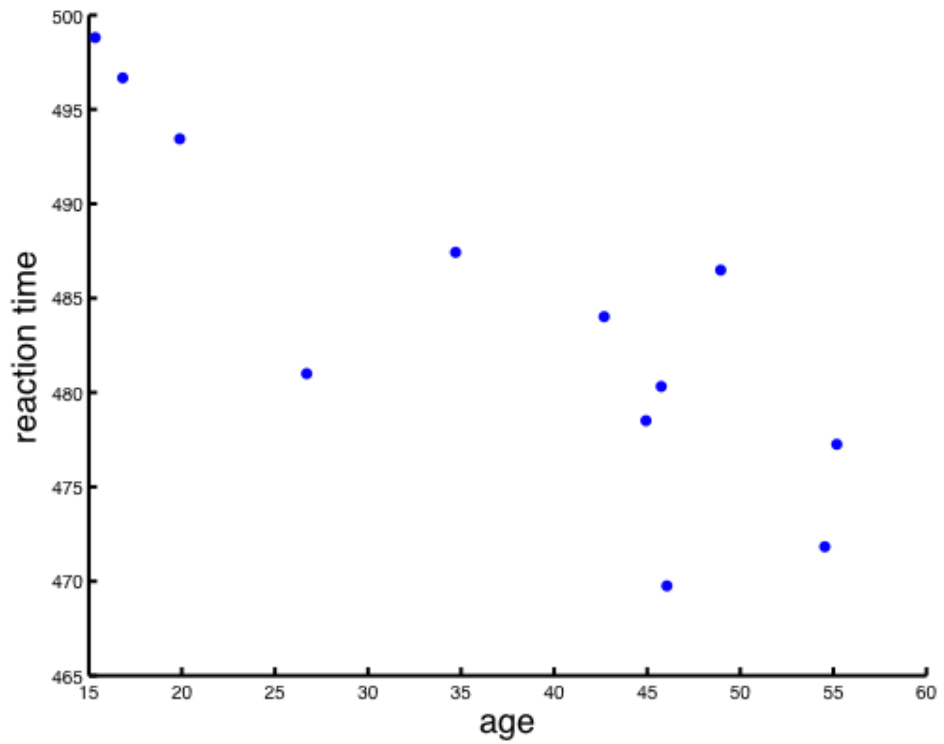
- overlaps with applied statistics

why use it at all?

to tell a story about data

[adapted from slides by Russ Poldrack]

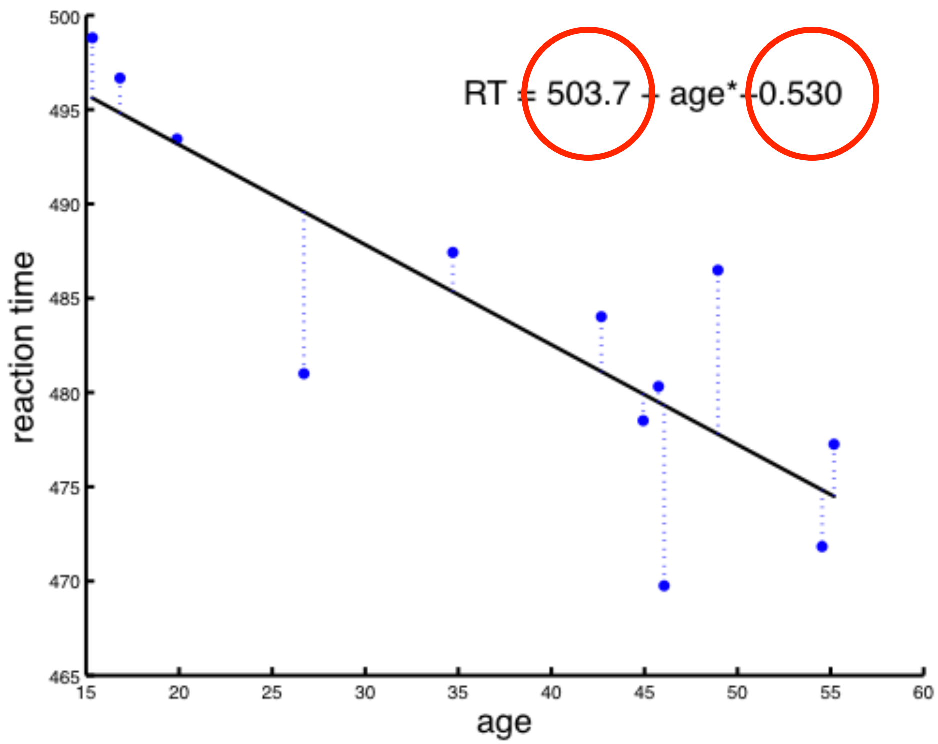
once upon a time there was a sample...



[adapted from slides by Russ Poldrack]

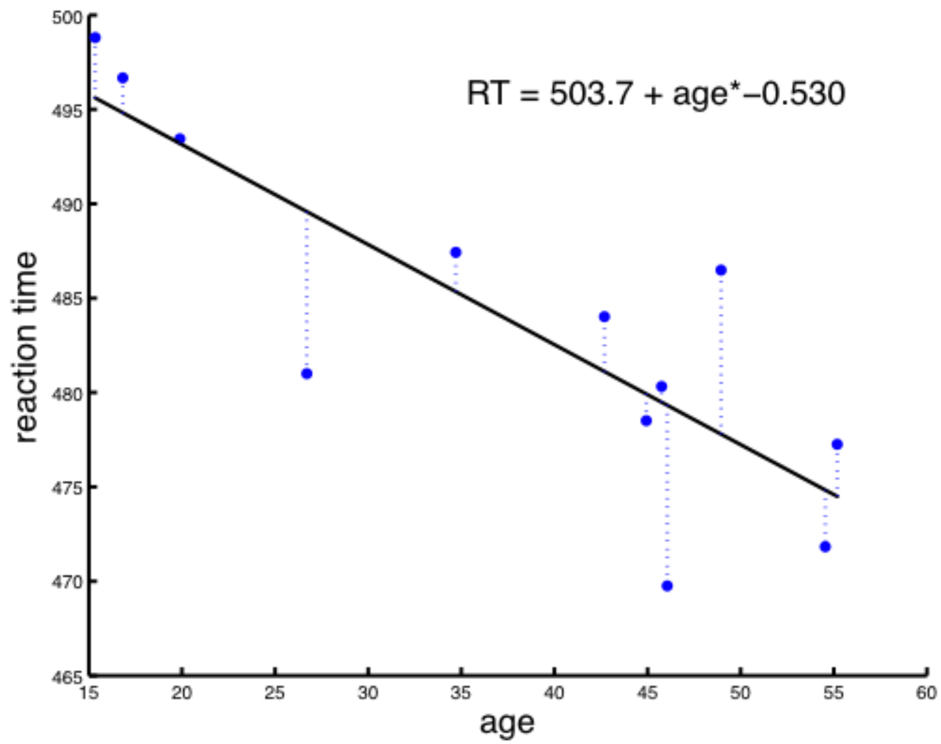
... and then came a beautiful model...

fit model by estimating parameters



[adapted from slides by Russ Poldrack]

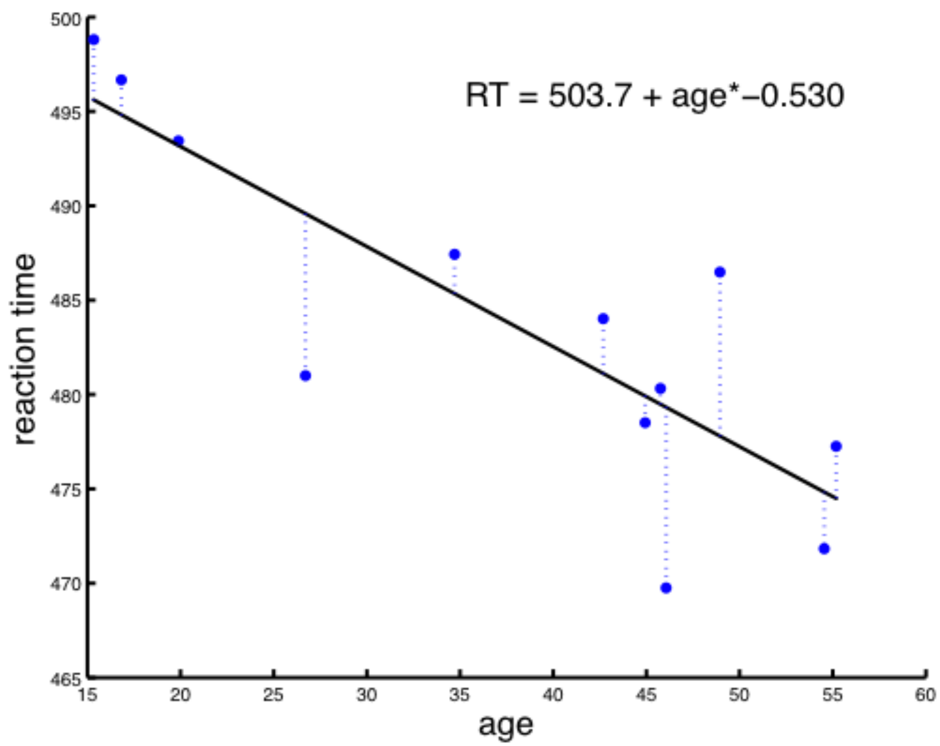
very suggestive, but...



Is RT really related to age?

[adapted from slides by Russ Poldrack]

is RT related to age?



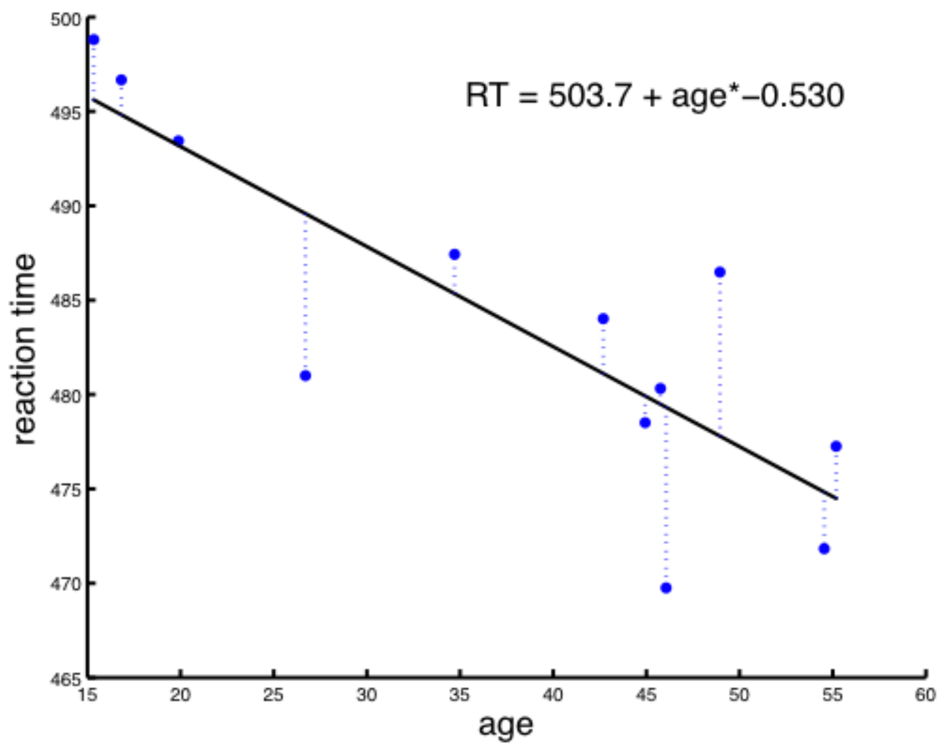
Model:

$RT = b_0 + b_1 * age + e$

parameters in the population

[adapted from slides by Russ Poldrack]

is RT related to age?



Model:

$RT = b_0 + b_1 * age + e$

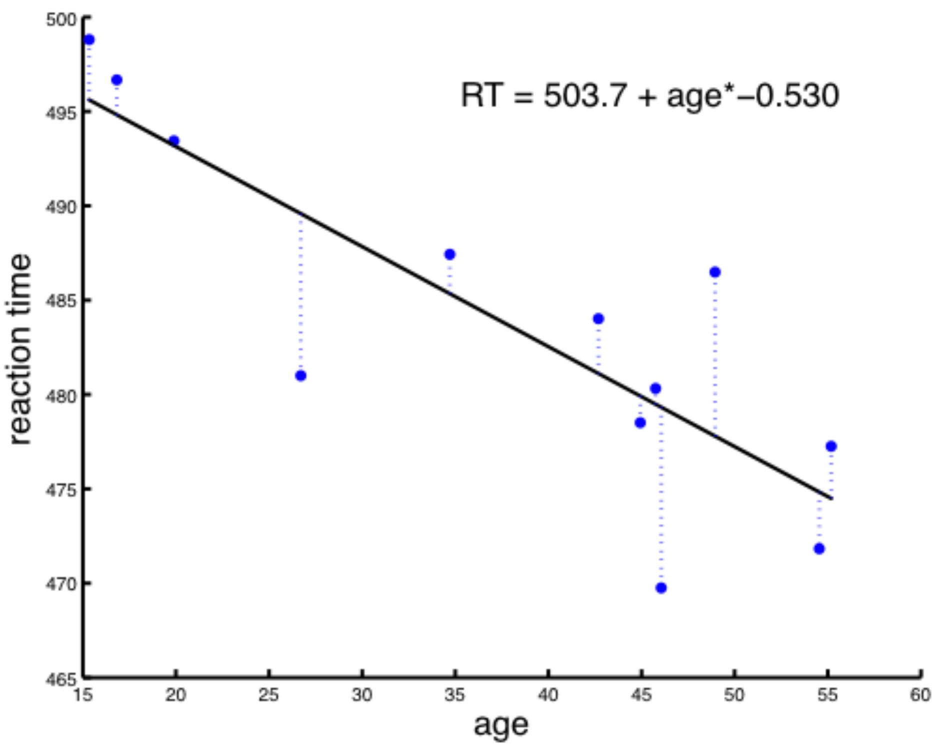
parameters in the population

parameters estimated from the sample

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

[adapted from slides by Russ Poldrack]

is RT related to age?



Model:

$RT = b_0 + b_1 * age + e$

parameters in the population

parameters estimated from the sample

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

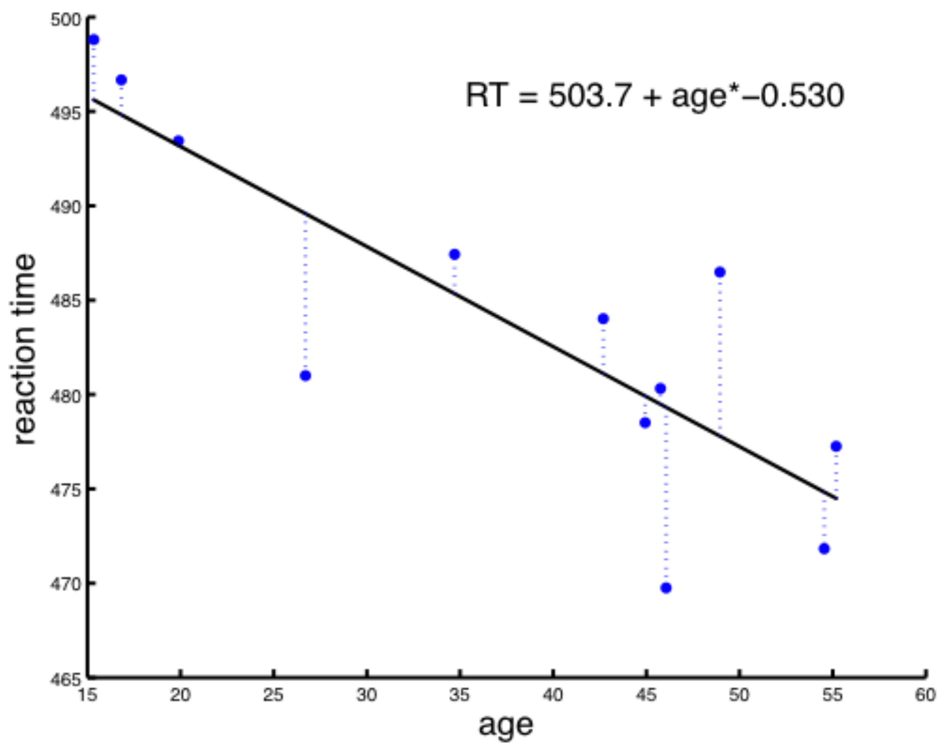
Hypothesis:

Is b_1 different from 0?

[adapted from slides by Russ Poldrack]

is RT related to age?

Null hypothesis: $b_1 = 0$
Alternative: $b_1 \sim= 0$

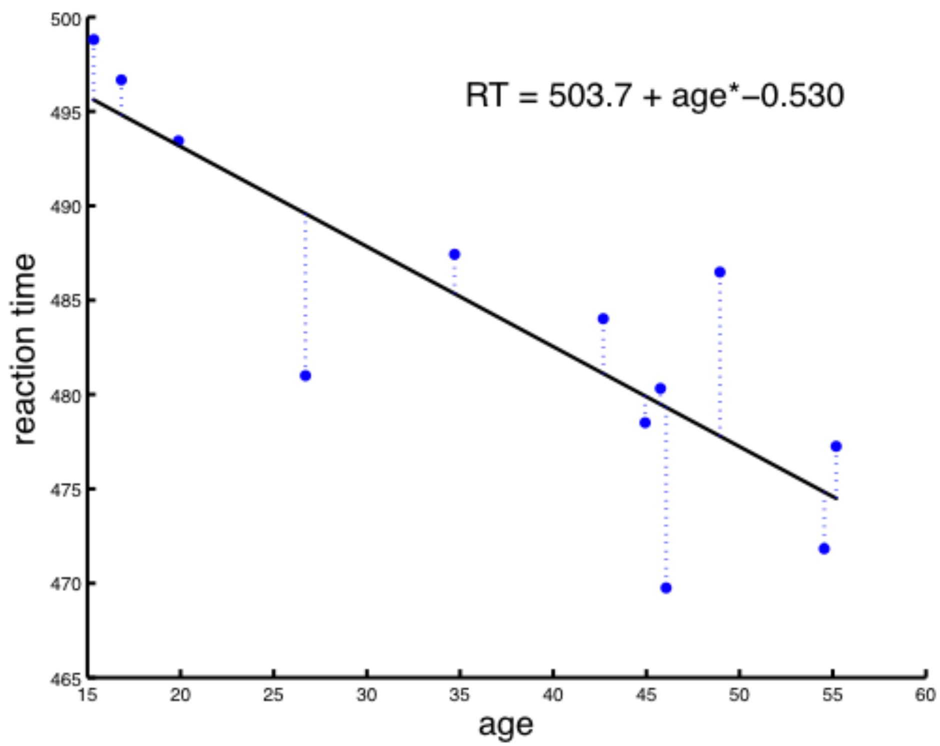


[adapted from slides by Russ Poldrack]

is RT related to age?

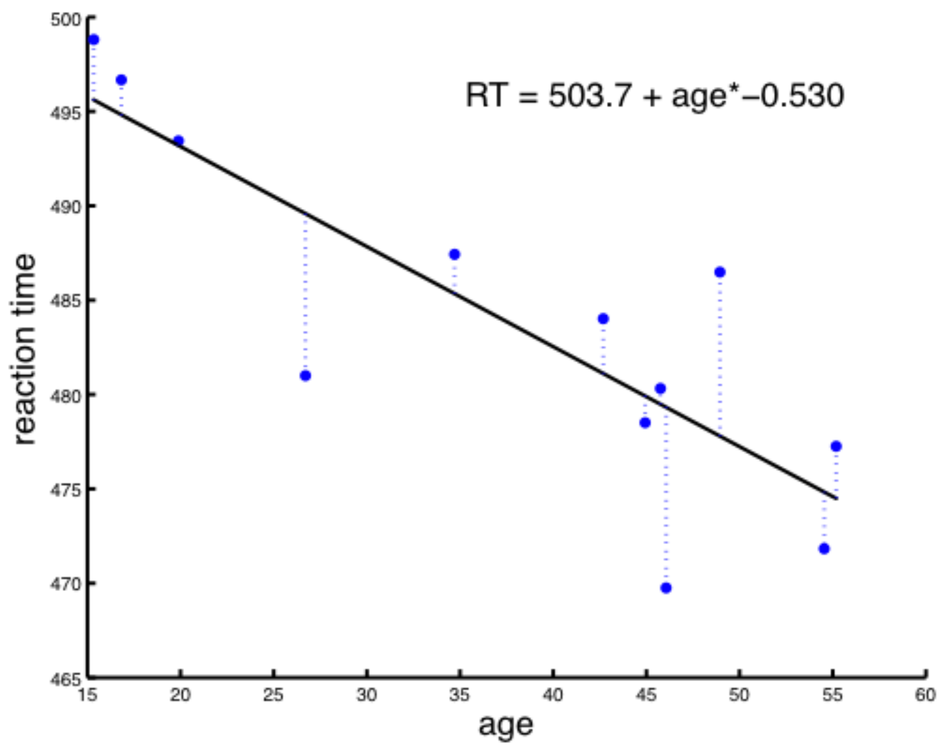
Null hypothesis: $b_1 = 0$
Alternative: b_1 ~= 0

How likely is the parameter estimate ($\hat{b}_1$ = -0.53) if the null hypothesis is true?



[adapted from slides by Russ Poldrack]

is RT related to age?



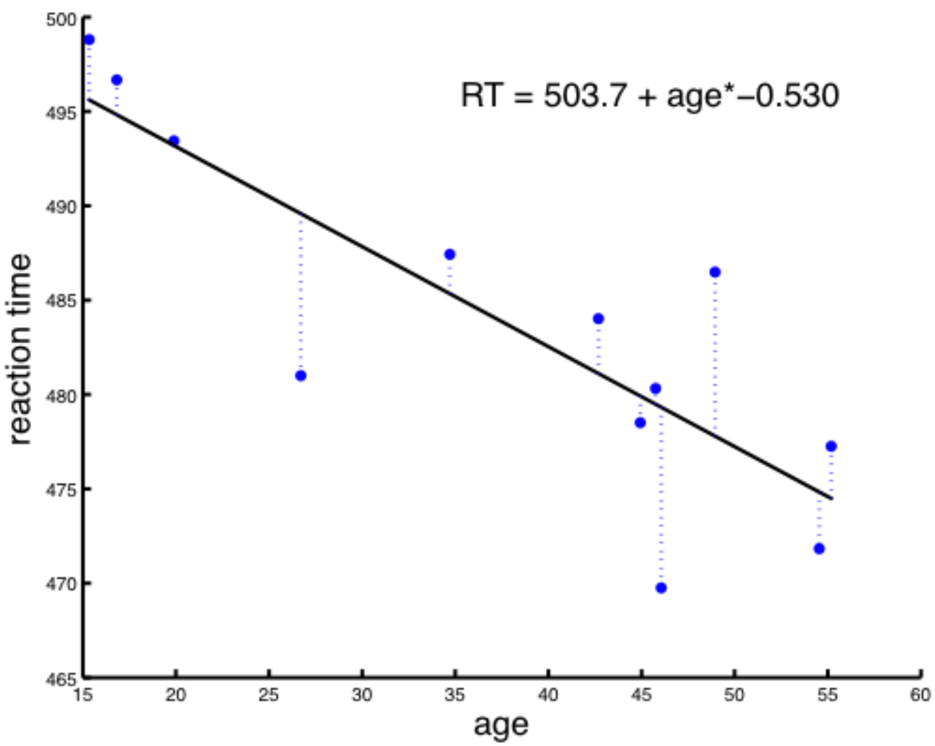
Null hypothesis: $b_1 = 0$
Alternative: b_1 ~= 0

How likely is the parameter estimate ($\hat{b}_1 = -0.53$) if the null hypothesis is true?

We need a statistic with a known distribution to determine this!

[adapted from slides by Russ Poldrack]

is RT related to age?



the CLT tells us

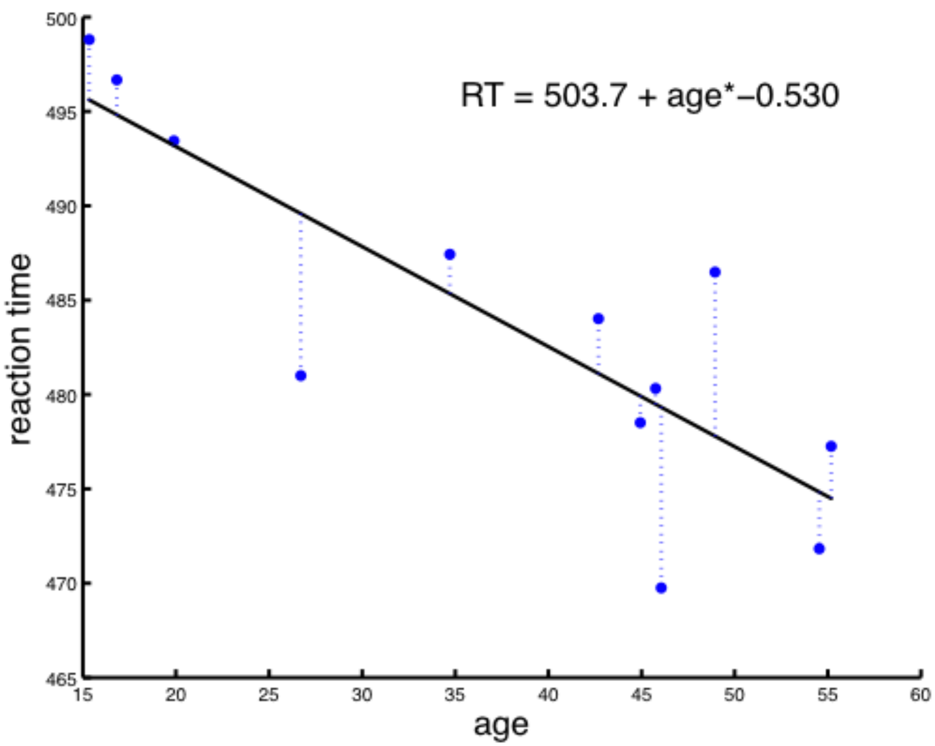
$$\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$$

but we don't know

$$Var(\hat{\beta}_1)$$

[adapted from slides by Russ Poldrack]

is RT related to age?



the CLT tells us

$$\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$$

but we don't know

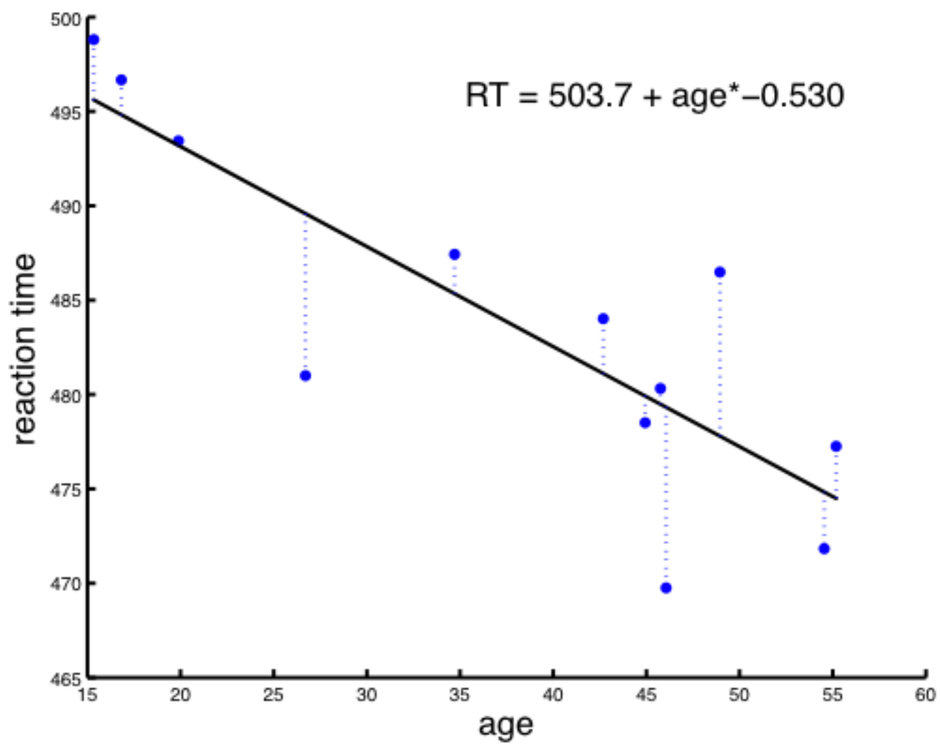
$$Var(\hat{\beta}_1)$$

we do know

$$t = \frac{\hat{\beta}_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} \sim T_{N-p}$$

[adapted from slides by Russ Poldrack]

is RT related to age?



$t(10) = -4.76$

How likely is this value in this t distribution ?

$p < .001$

[adapted from slides by Russ Poldrack]

what can we conclude?

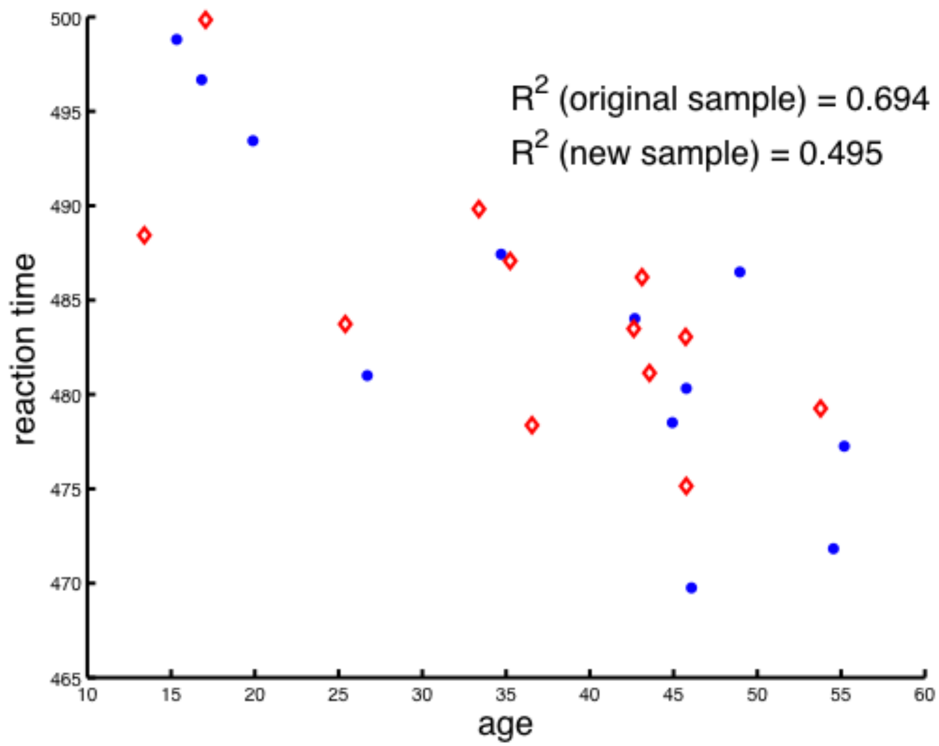
- in this sample
  - $p < 0.001$ - there is a relationship between age and RT
  - R2 - age accounts for 69% of variance in RT

- very unlikely if no relationship in the population

- the test does not tell us how well we can predict RT from age in the population

[adapted from slides by Russ Poldrack]

what happens with a new sample?

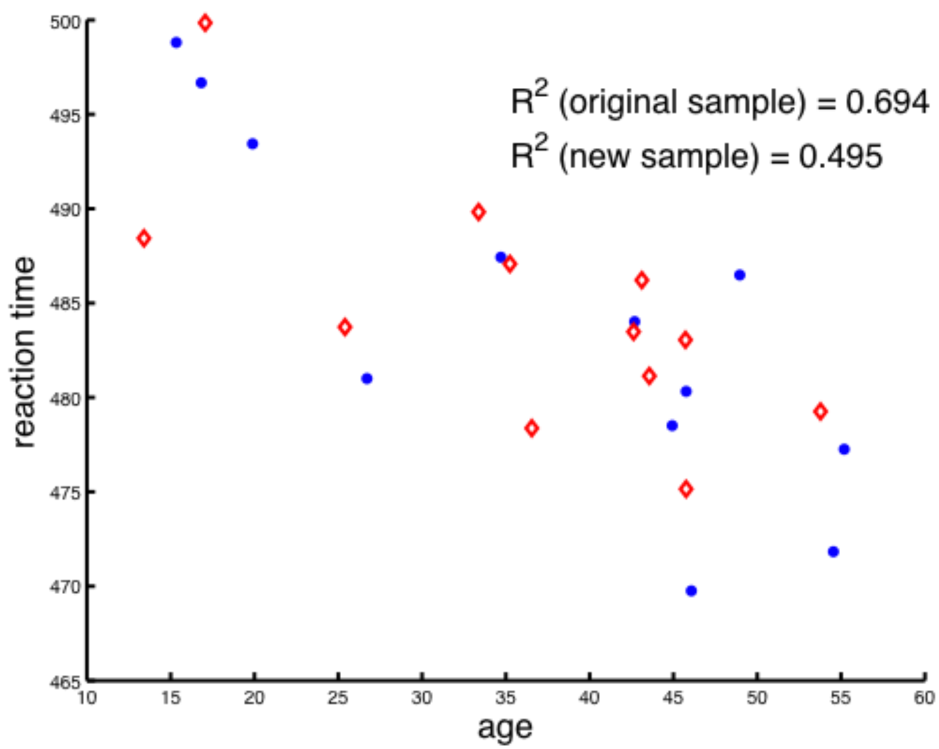


draw a new sample from the same population

compute the R2 using parameters estimated in the original sample

[adapted from slides by Russ Poldrack]

what happens with a new sample?



repeat this 100 times... using model parameters estimated from the original sample

average R2 = 0.578

a measure of how good the model learned from a single sample is

[adapted from slides by Russ Poldrack]

the learning perspective

When we estimate parameters from a sample, we are learning about the population from training data.

[adapted from slides by Russ Poldrack]

the learning perspective

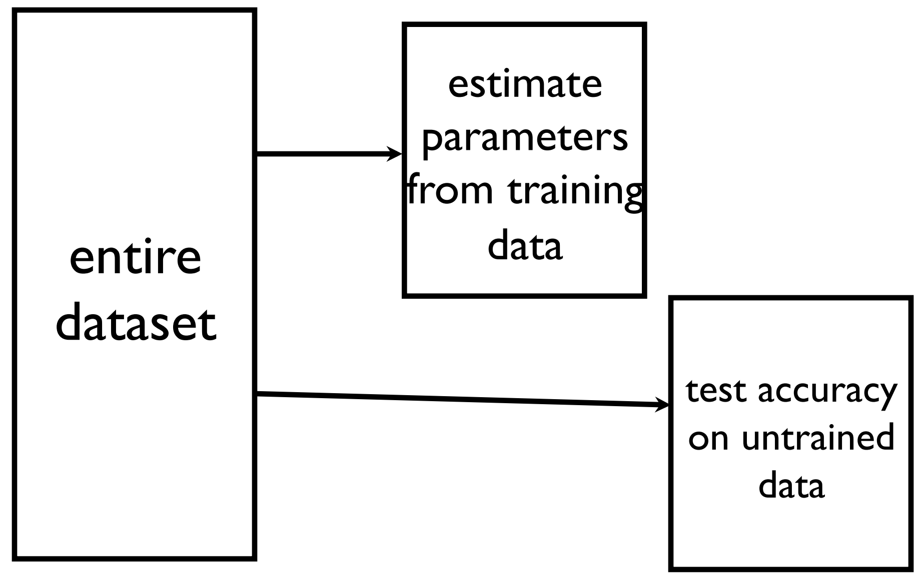
When we estimate parameters from a sample,
we are learning about the population
from training data.

How well can we measure the prediction
ability of our learned model?
Use a new sample as test data.

[adapted from slides by Russ Poldrack]

test data and cross-validation

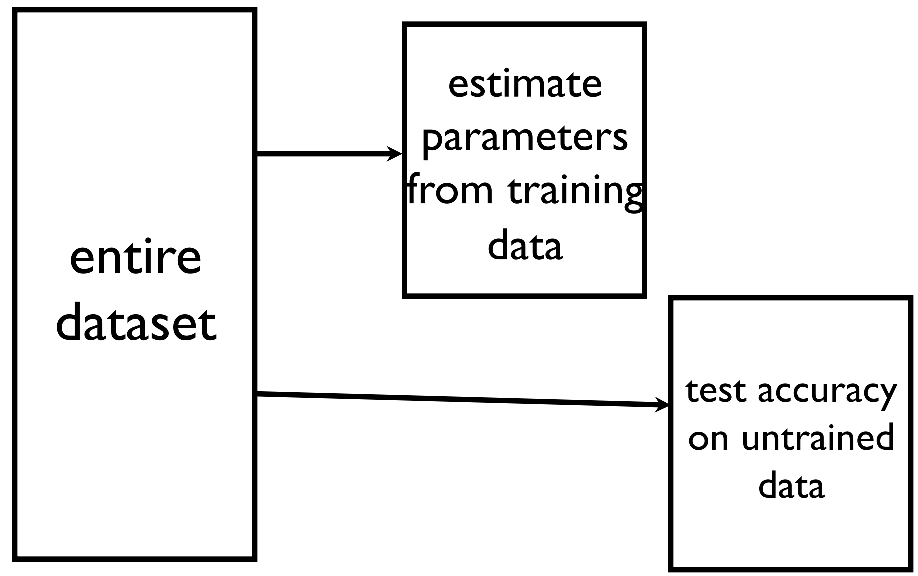
If you can't collect more split your sample in two…



[adapted from slides by Russ Poldrack]

test data and cross-validation

If you can't collect more
split your sample in two...



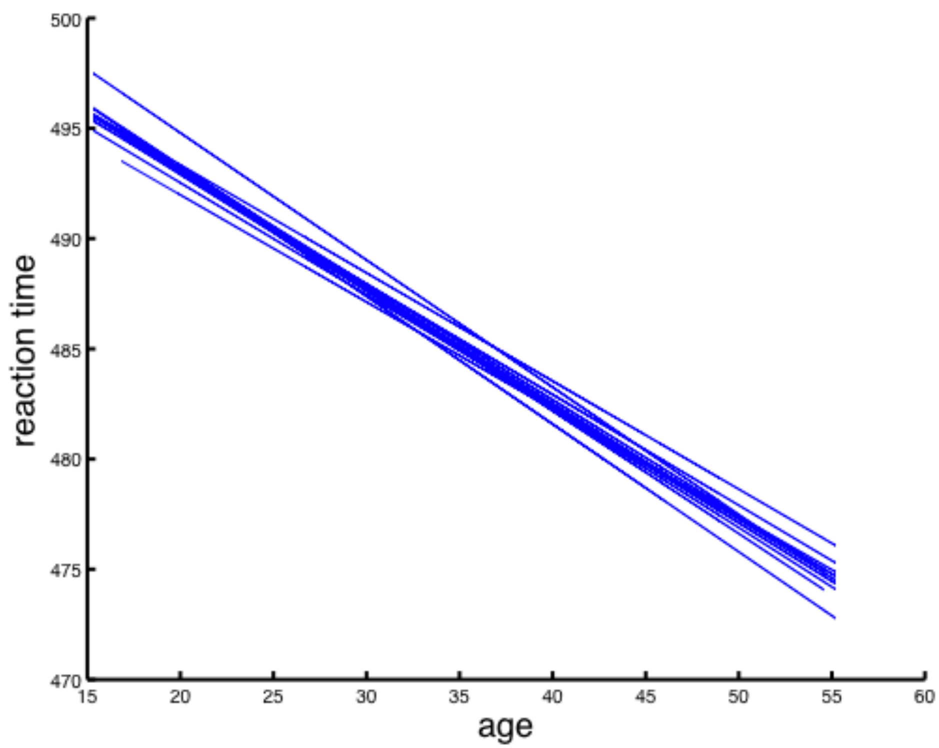
k-fold cross-validation:

- split into k folds
- train on k-1, test on the left out
- average prediction measure on all k folds
- several variants: all possible splits, leave-one-out

[adapted from slides by Russ Poldrack]

leave-one-out cross-validation

regression lines
on each training set



original sample
$R^2 = 0.694$

leave-one-out on original
$R^2 = 0.586$

mean over 100 new samples
$R^2 = 0.591$

[adapted from slides by Russ Poldrack]

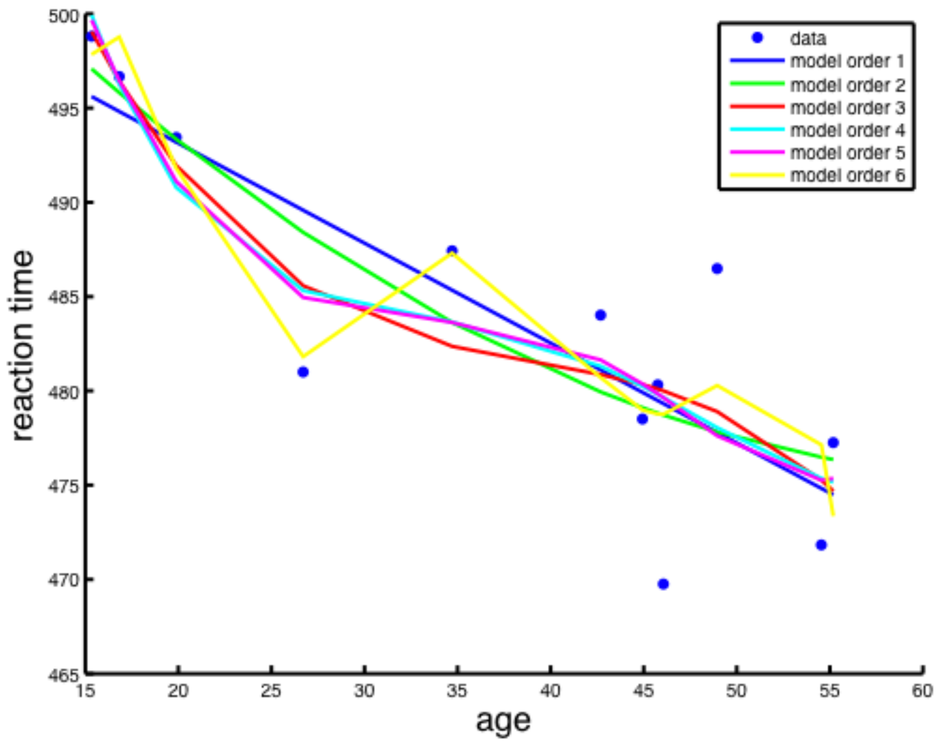
model complexity

As model complexity goes up,
we can always fit the training data better

What does this do to our predictive ability?

[adapted from slides by Russ Poldrack]

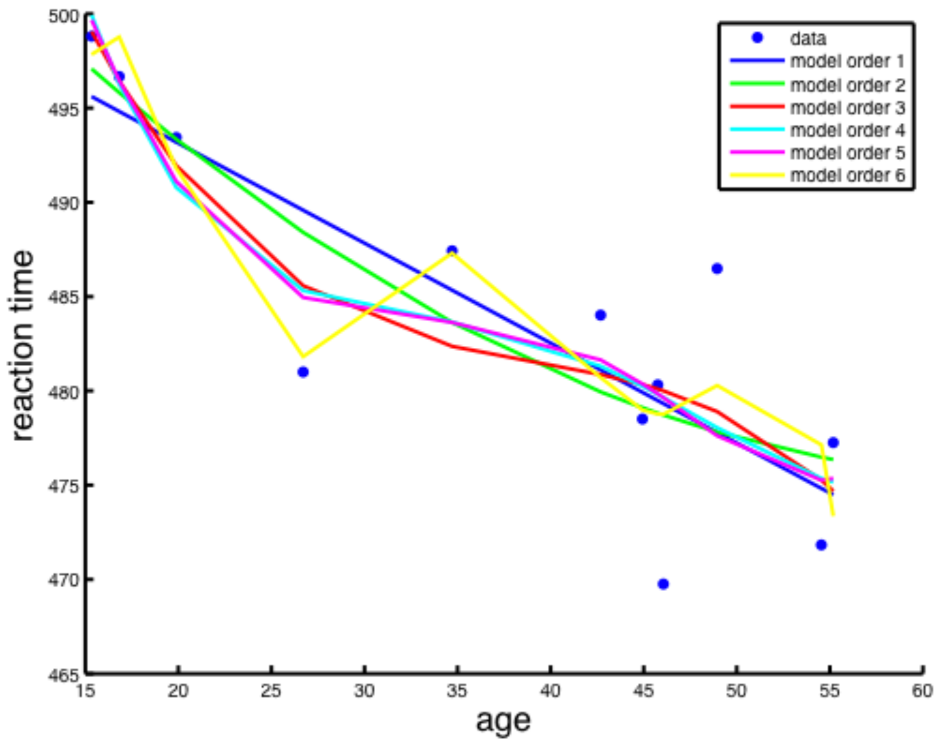
model complexity

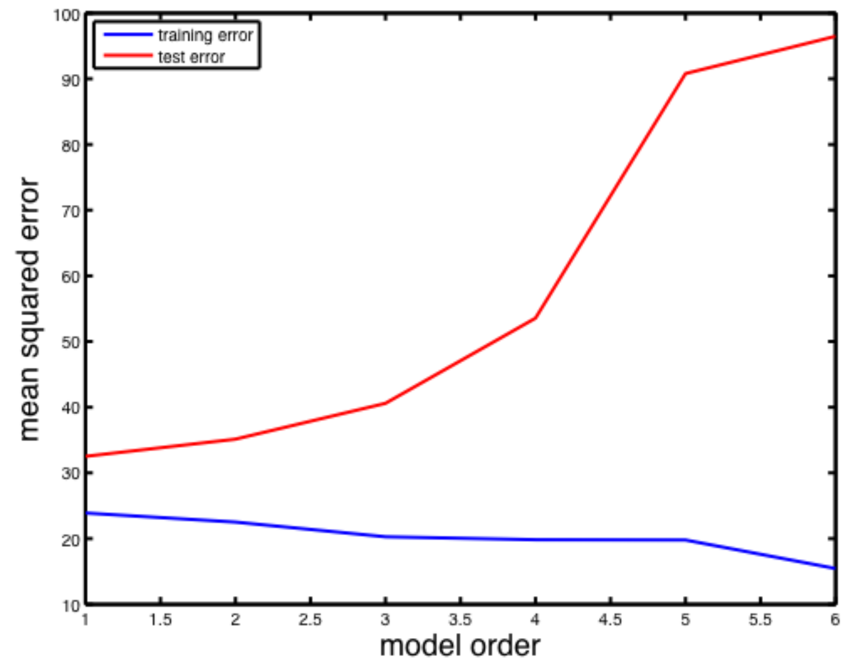


polynomials of higher degree
fit the training data better...

[adapted from slides by Russ Poldrack]

model complexity



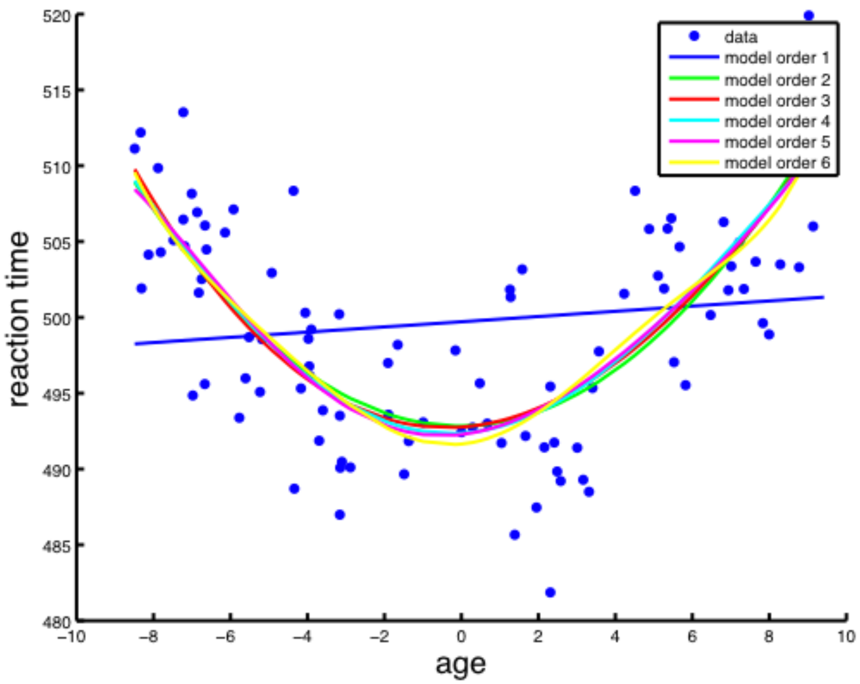


polynomials of higher degree fit the training data better...

... but they do worse on test data: overfitting

[adapted from slides by Russ Poldrack]

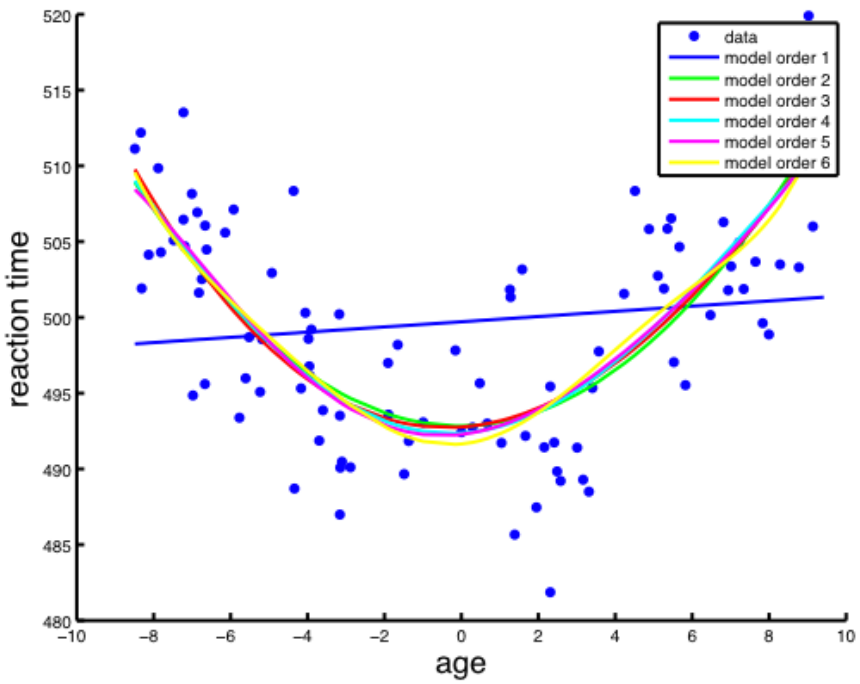
model complexity

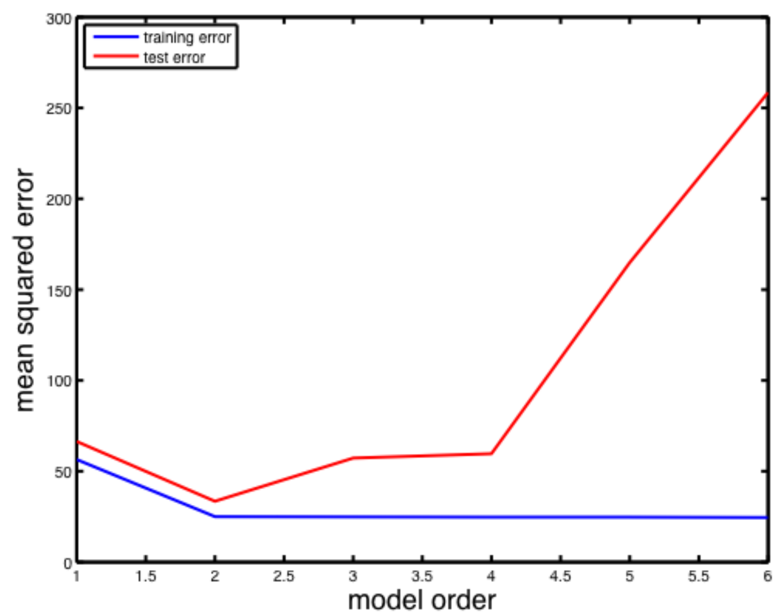


if the relationship in the
population were more complicated

[adapted from slides by Russ Poldrack]

model complexity





if the relationship in the population were more complicated

we could use CV to determine adequate model complexity

(this would need to be done with nested CV, CV inside the training set)

what is machine learning, redux

- generalization: make predictions about a new individual

- a model that generalizes captures the relationship between the individual and what we want to predict

- cross-validation is a good way of
  - measuring generalization
  - doing model selection (there are others)

what is machine learning, redux

- generalization: make predictions about a new individual
- a model that generalizes captures the relationship between the individual and what we want to predict
- cross-validation is a good way of
  - measuring generalization
  - doing model selection (there are others)
- “all models are wrong but some are useful”
  George Box

what does this have to do with fMRI?

In this talk:

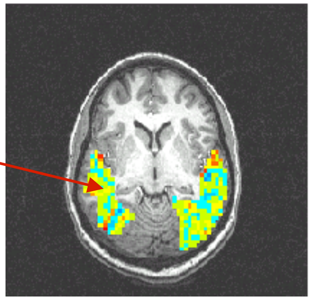
- prediction is classification
- generalization is within subject (population of trials)
- how to draw conclusions with statistical significance
- what has it been used for?

two questions

GLM: are there voxels that reflect the stimulus?

stimulus → fMRI activation (single voxel)

contrast of interest → GLM



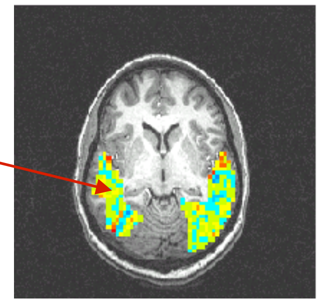
two questions

GLM: are there voxels that reflect the stimulus?

stimulus → fMRI activation

(single voxel)

contrast of interest → GLM

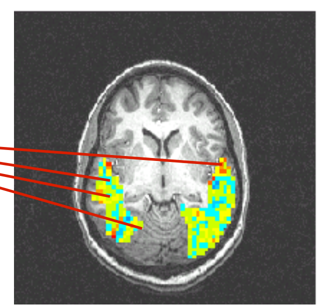


Classifier: do voxels contain information to predict?

stimulus ← fMRI activation

(multiple voxels)

what is the subject seeing?



case study: two categories

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

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

    see a word

    think about properties, use, visualize

    blank

3 seconds

8 seconds

case study: two categories

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

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

    see a word

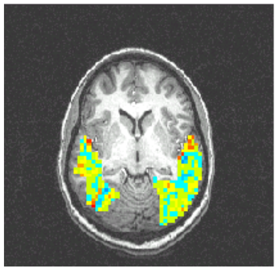
    think about properties, use, visualize

    blank

3 seconds

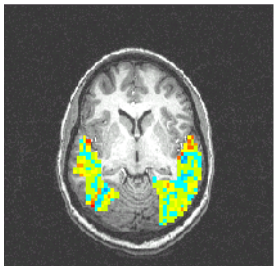
8 seconds

- goal: can the two categories be distinguished?
- average images around trial peak to get one labelled image



tools

the name(s) of the game



average trial image

example

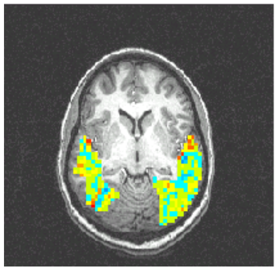
...

tools

voxels (features)

class label

the name(s) of the game



average trial image

example

...

tools

voxels (features)

class label

14 examples

...

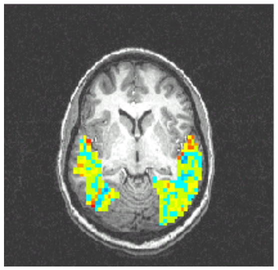
tools
labels

=

example group

...

the name(s) of the game



average trial image

example

...

tools

voxels (features)

class label

14 examples

...

tools labels = ...

example group

dataset

...

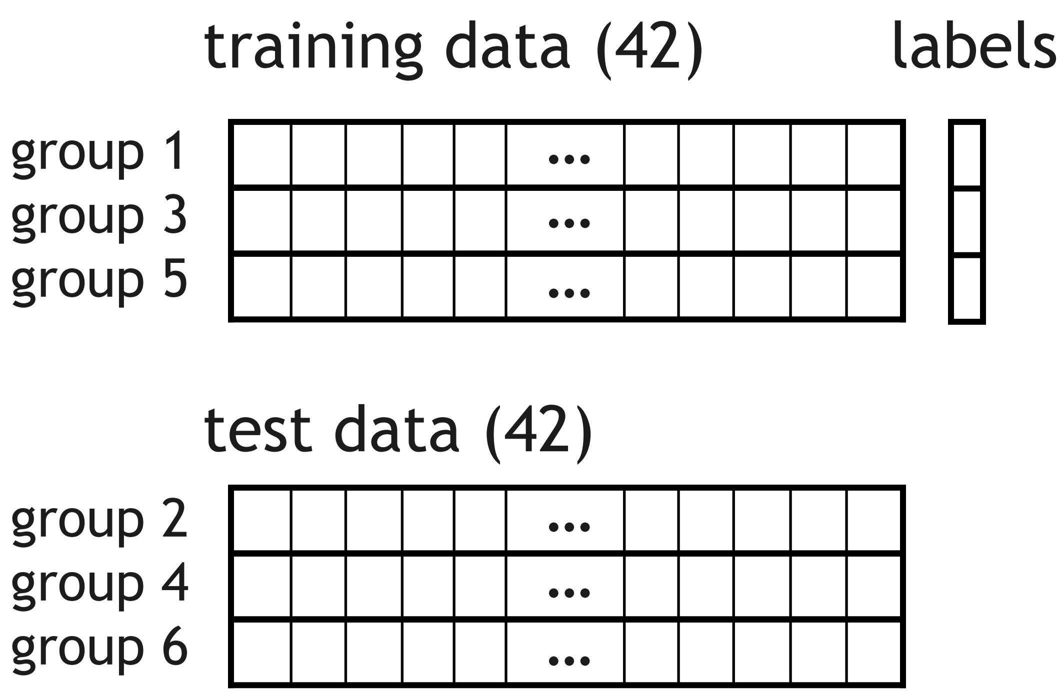
group 1

...

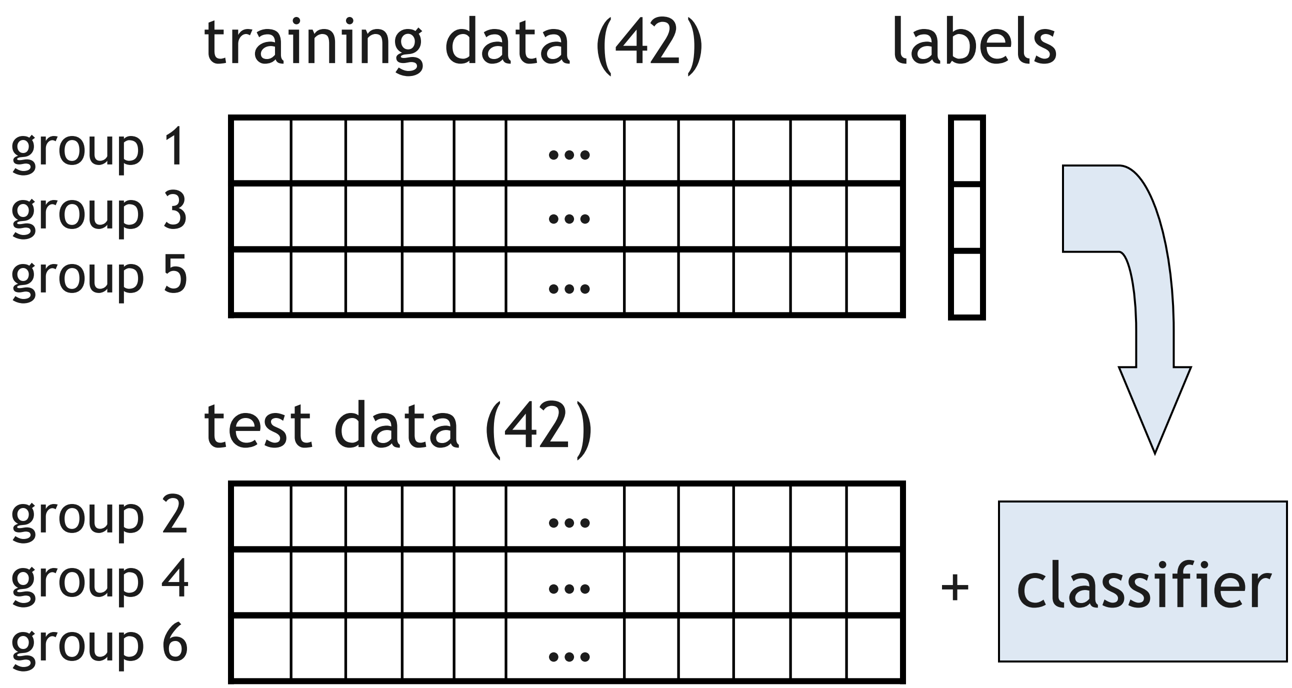
group 6

84 examples

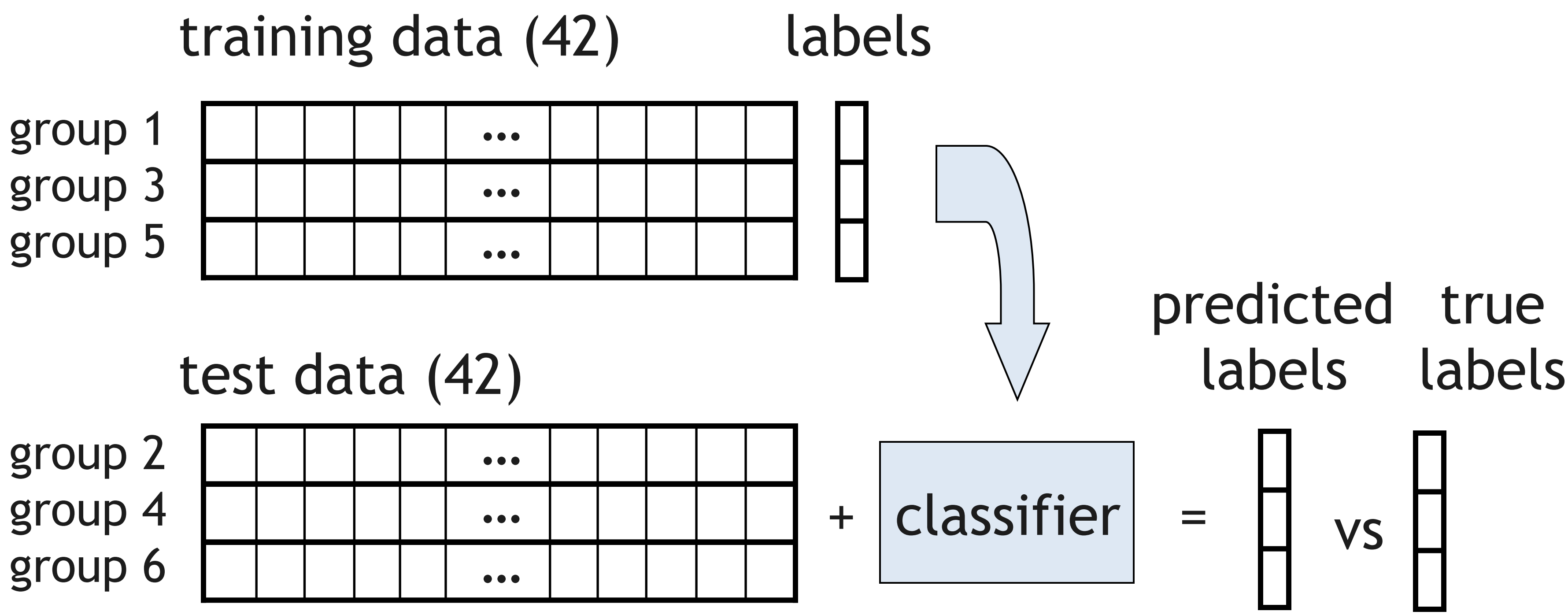
classifying two categories



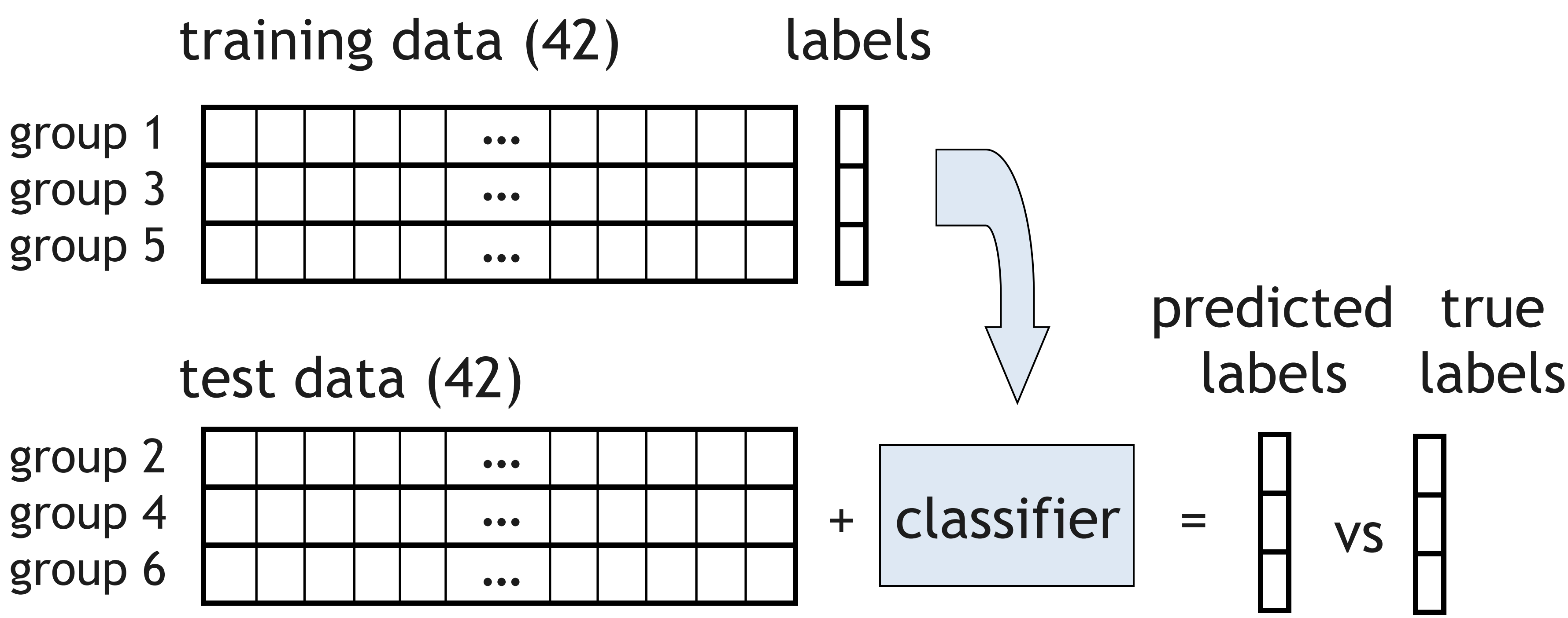
classifying two categories



classifying two categories



classifying two categories



accuracy estimate
= #correct/42
= 0.65

a classifier

- is a function from data to labels
- parameters learned from training data

a classifier

- is a function from data to labels
- parameters learned from training data

- want to estimate its true accuracy
  - "probability of labelling a new example correctly"
- estimation is done on the test data
  - it is finite, hence an estimate with uncertainty

a classifier

- is a function from data to labels
- parameters learned from training data

- want to estimate its true accuracy
  - "probability of labelling a new example correctly"
- estimation is done on the test data
  - it is finite, hence an estimate with uncertainty

- null hypothesis: "the classifier learned nothing"

what questions can be tackled?

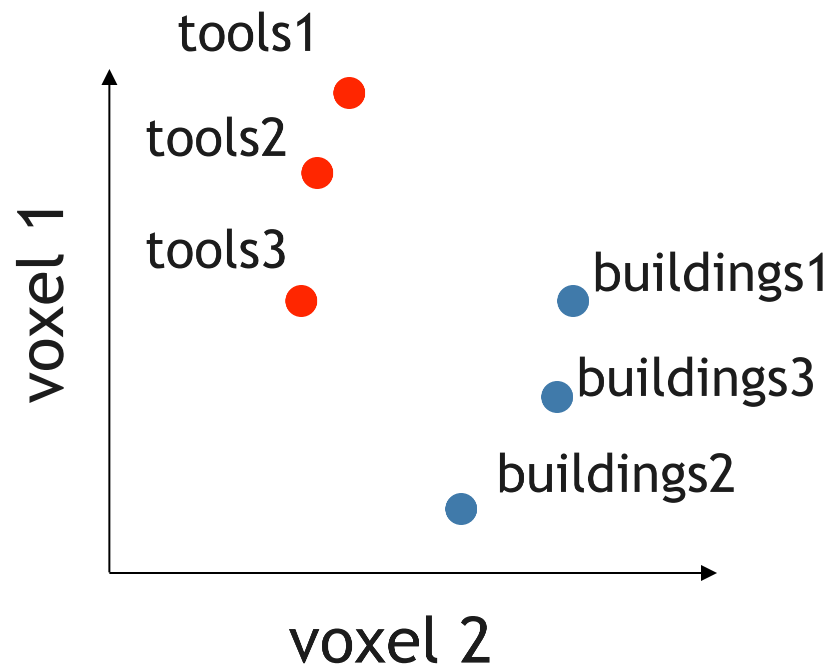
- is there information?
  (pattern discrimination)
- where/when is information present?
  (pattern localization)
- how is information encoded?
  (pattern characterization)

“is there information?”

- what is inside the black box?
- how to test results?
- from a study to examples

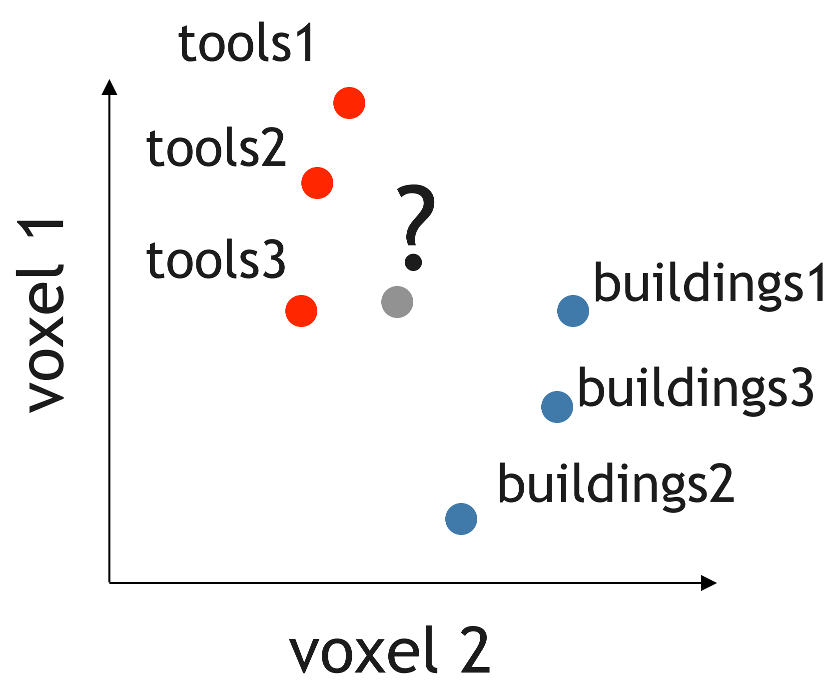
what is inside the box?

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



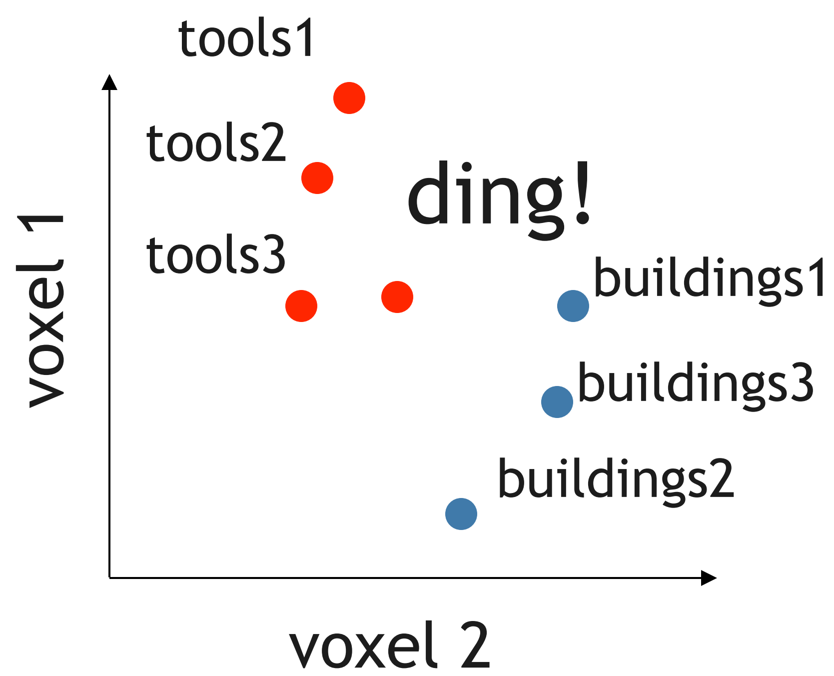
what is inside the box?

- simplest function is no function at all
- "nearest neighbour"



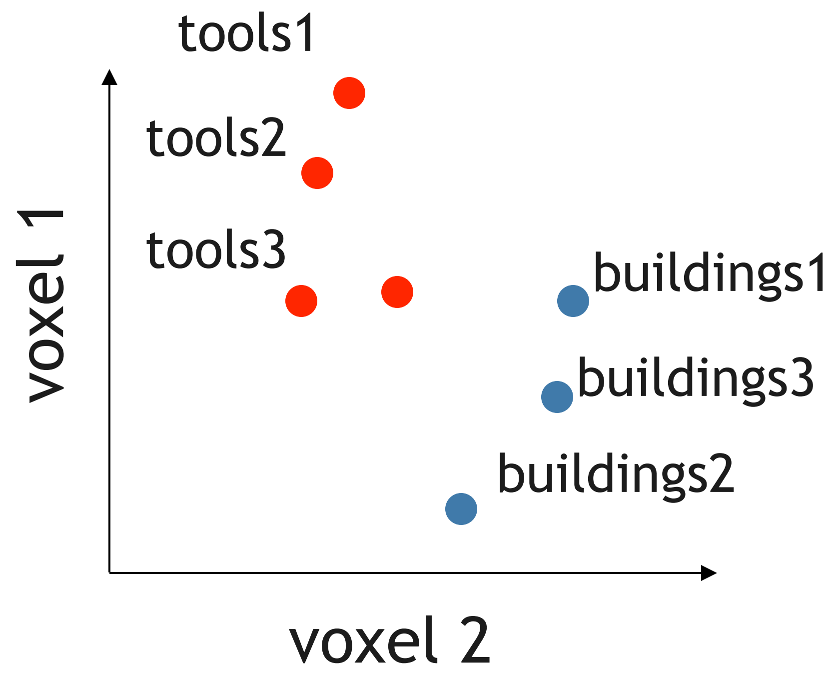
what is inside the box?

- simplest function is no function at all
- "nearest neighbour"

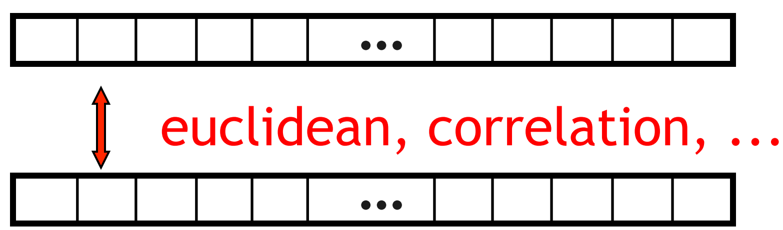


what is inside the box?

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

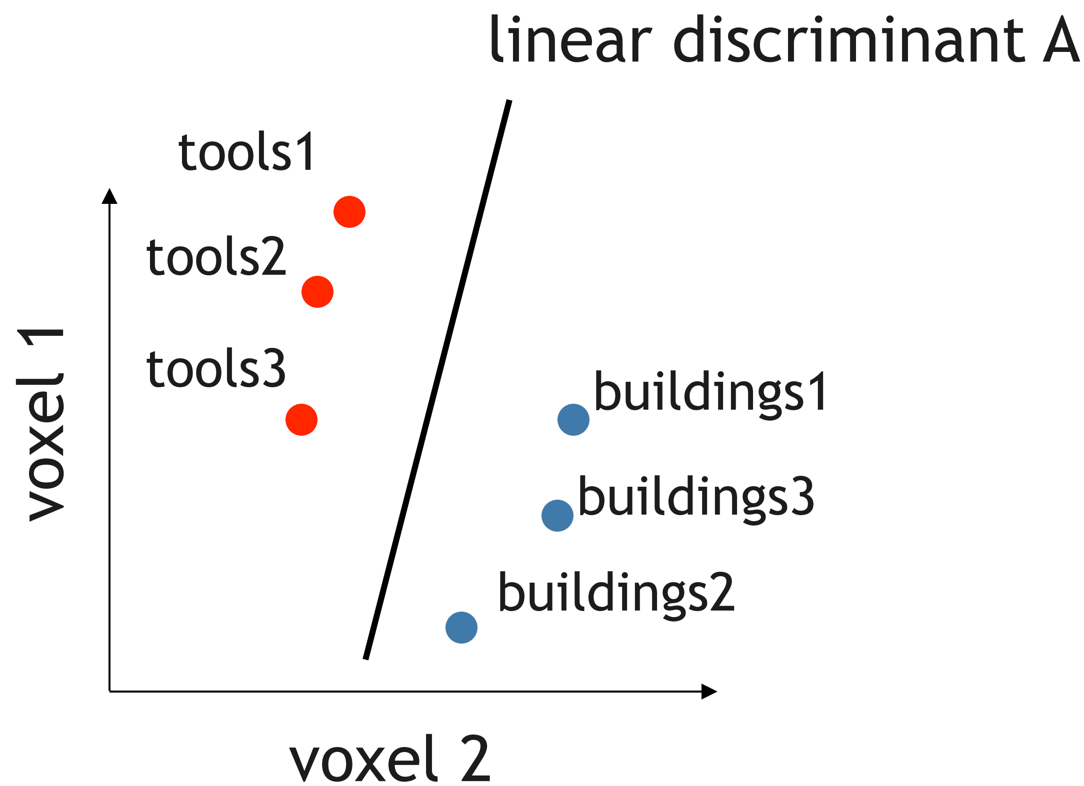


requires example similarity measure



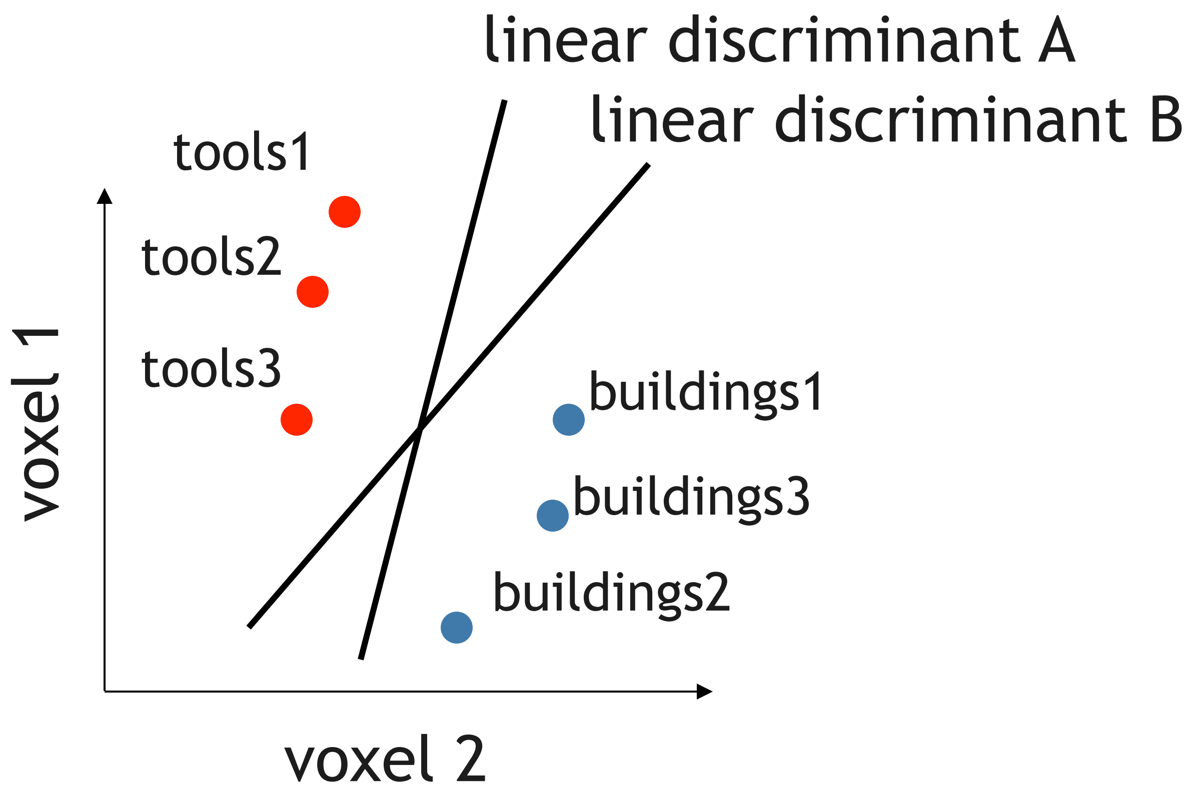
what is inside the box?

- next simplest: learn linear discriminant



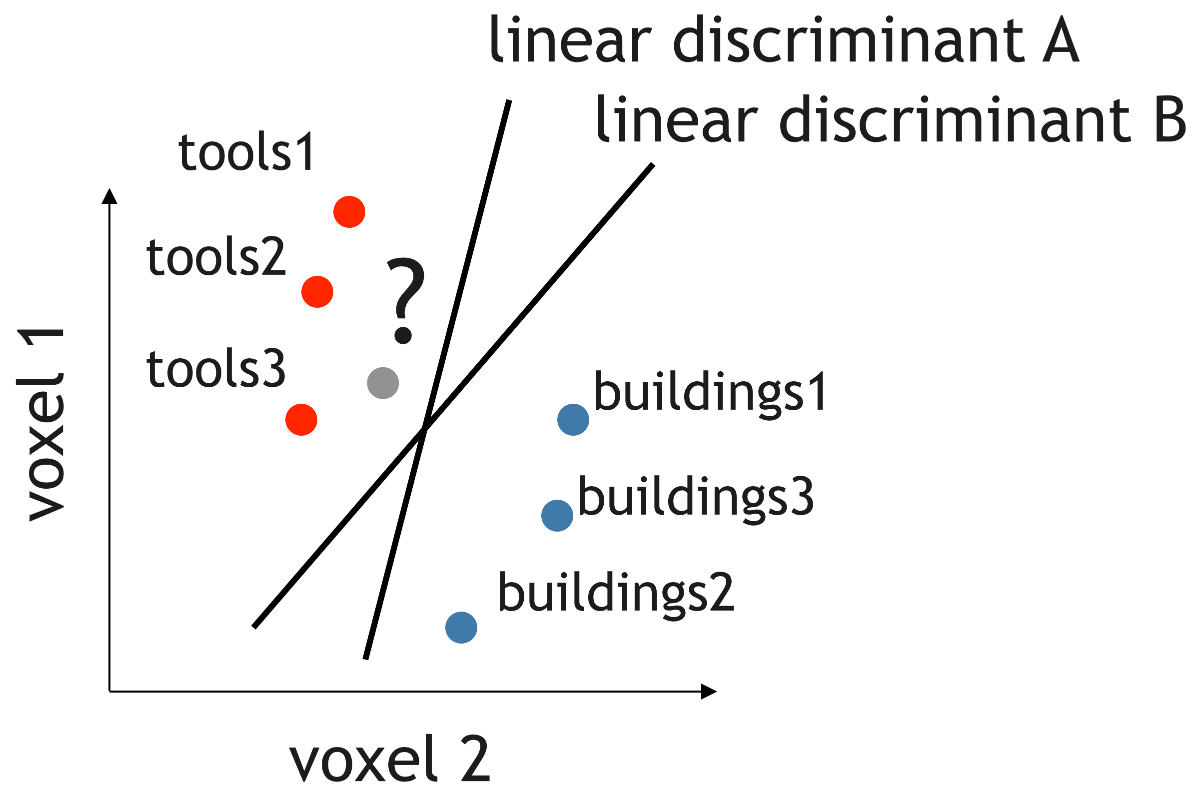
what is inside the box?

- next simplest: learn linear discriminant
- note that there are many solutions...



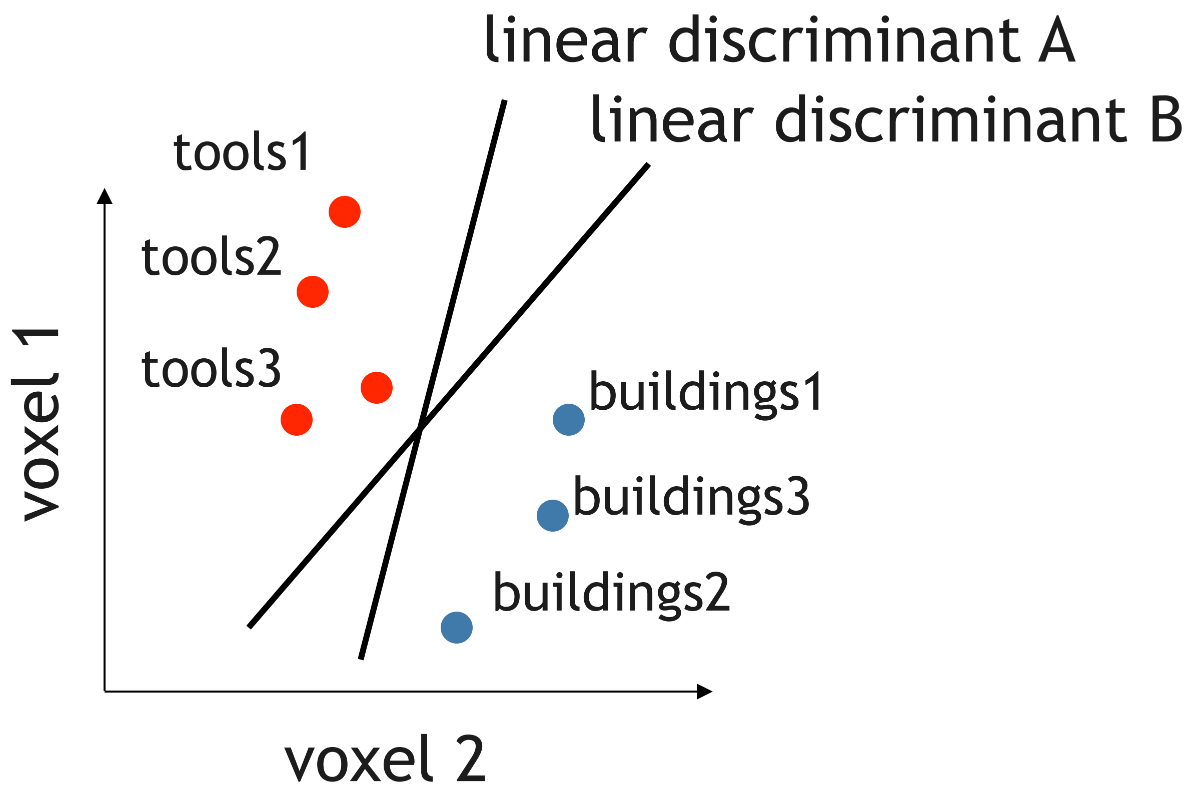
what is inside the box?

- next simplest: learn linear discriminant
- note that there are many solutions...

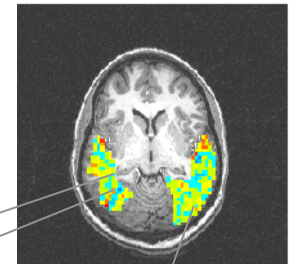


what is inside the box?

- next simplest: learn linear discriminant
- note that there are many solutions...



linear classifiers



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

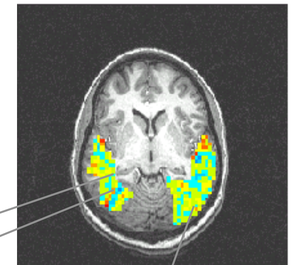
If

voxel 1	voxel 2	...			voxel n

> 0 tools

otherwise buildings

linear classifiers



If weight0 + weight1 x voxel 1 + weight2 x voxel 2 + ... + + + weight n x voxel n > 0 tools

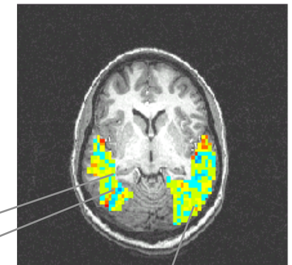
otherwise buildings

various kinds

- Gaussian Naive Bayes
- Logistic Regression
- Linear SVM

differ on how weights are chosen

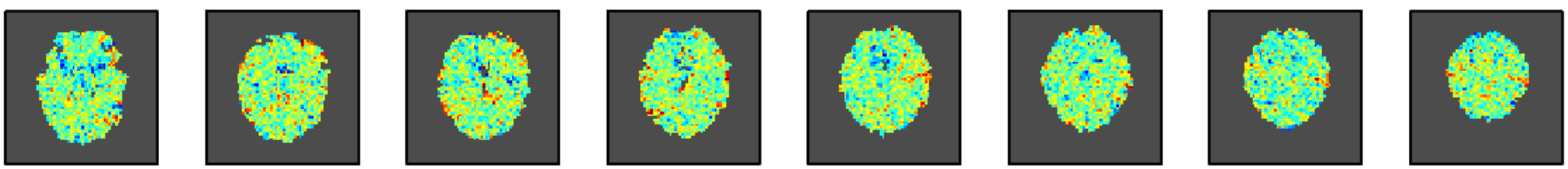
linear classifiers



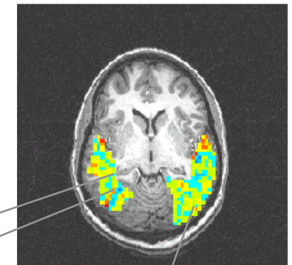
If weight0 + weight1 x voxel 1 + weight2 x voxel 2 + … + + weight n x voxel n > 0 tools

otherwise buildings

linear SVM weights:



linear classifiers

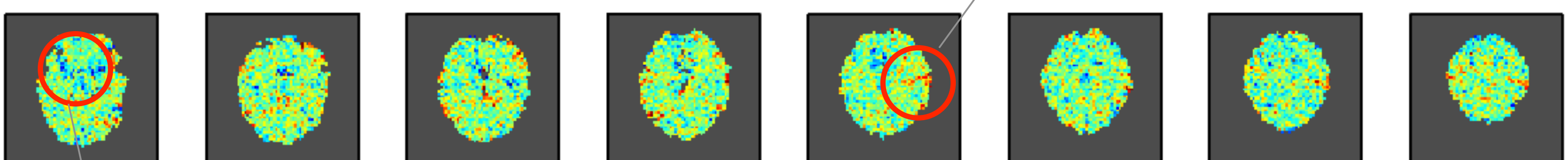


If weight0 + weight1 x voxel 1 + weight2 x voxel 2 + … + + weight n x voxel n > 0 tools

otherwise buildings

weights pull towards tools

linear SVM weights:



weights pull towards buildings

nonlinear classifiers

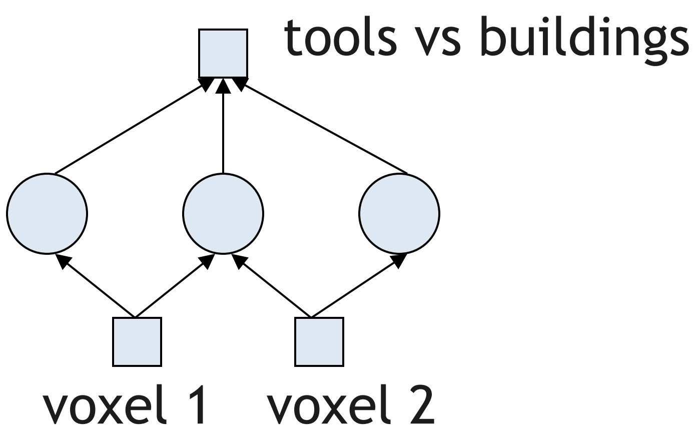
- linear on a transformed feature space

nonlinear classifiers

- linear on a transformed feature space

- neural networks:

  new features are learnt

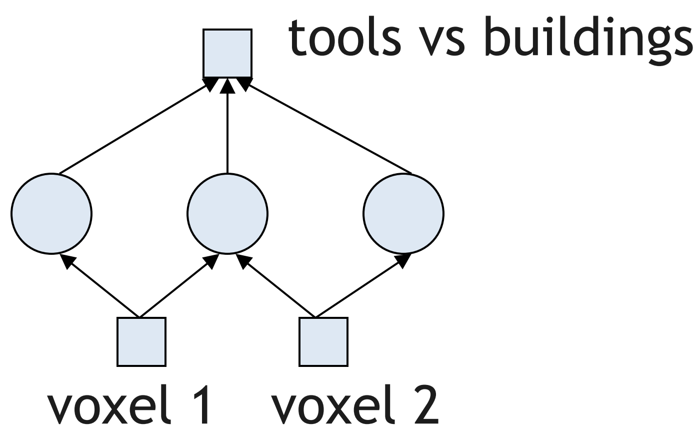


nonlinear classifiers

- linear on a transformed feature space!

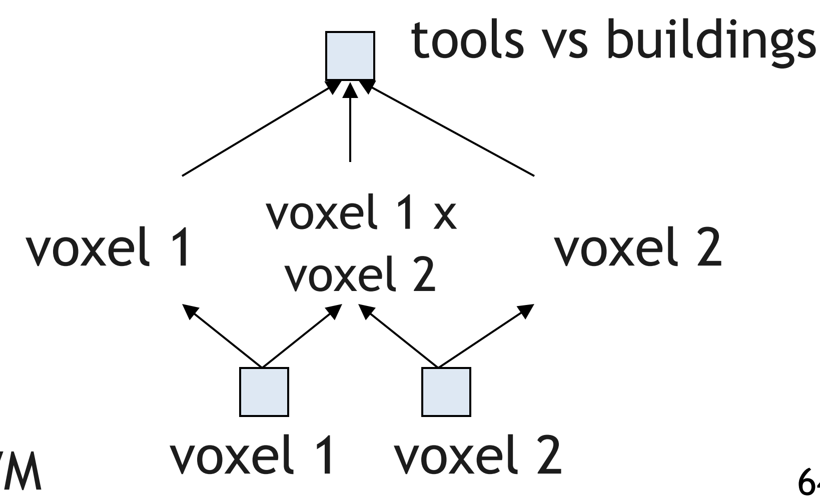
- neural networks:

  new features are learnt



- SVMs

  new features are (implicitly) determined by a kernel



quadratic SVM

nonlinear classifiers

reasons to be careful:

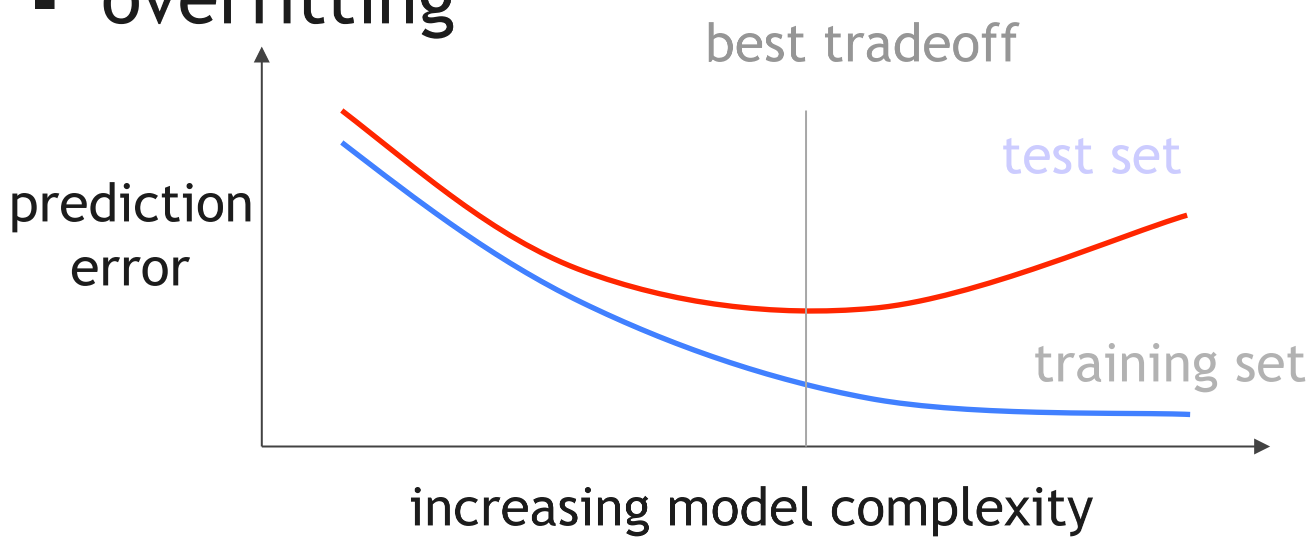
- too few examples,
  too many features
- harder to interpret

nonlinear classifiers

reasons to be careful:

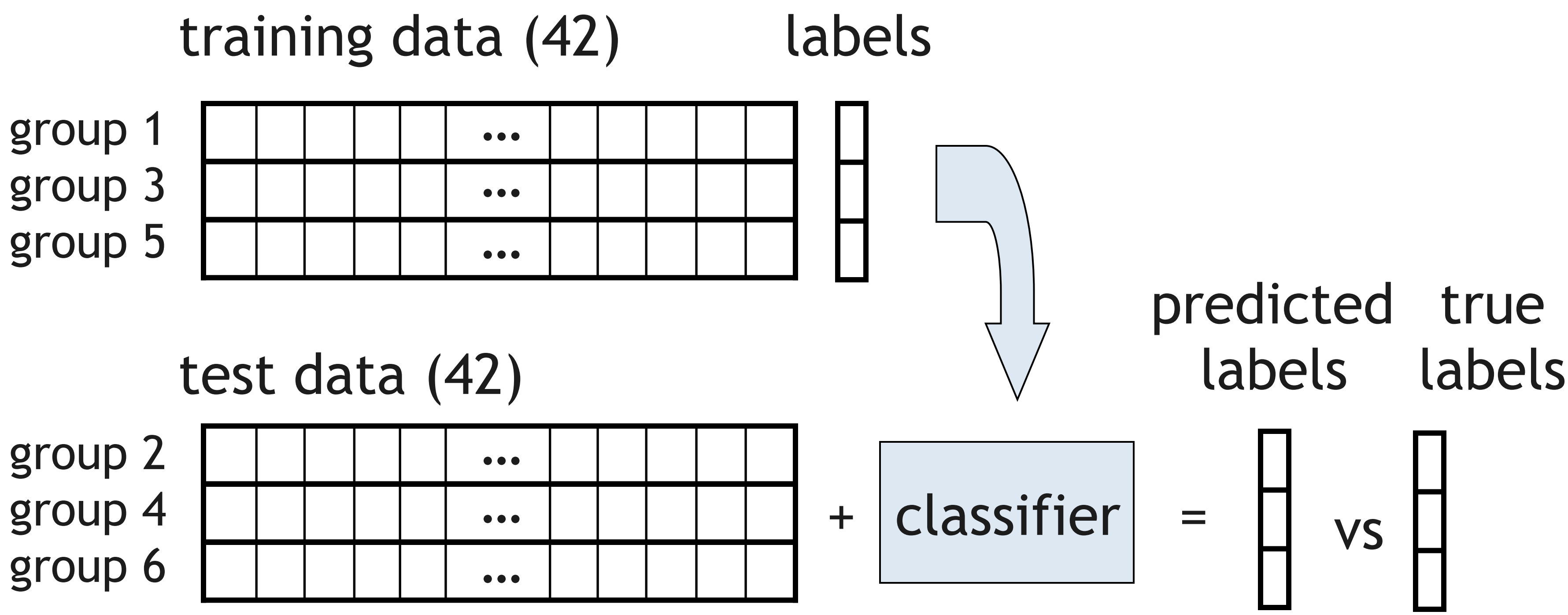
- too few examples,
  too many features
- harder to interpret

- overfitting



[from Hastie et al,2001]

how do we test predictions?



how do we test predictions?

Predicted labels

tools
buildings
buildings

...

tools
buildings
tools

how do we test predictions?

True labels	Predicted labels	
tools	tools	
tools	buildings	error
buildings	buildings	
...	...	
buildings	tools	error
buildings	buildings	
tools	tools	

#correct
out of
#test

how do we test predictions?

True labels	Predicted labels	
tools	tools	
tools	buildings	error
buildings	buildings	
...	...	
buildings	tools	error
buildings	buildings	
tools	tools	

#correct out of #test

- null hypothesis:

  “classifier learnt nothing” $\longrightarrow$ “predicts randomly”

how do we test predictions?

True labels	Predicted labels	
tools	tools	
tools	buildings	error
buildings	buildings	
…	…	
buildings	tools	error
buildings	buildings	
tools	tools	

#correct out of #test

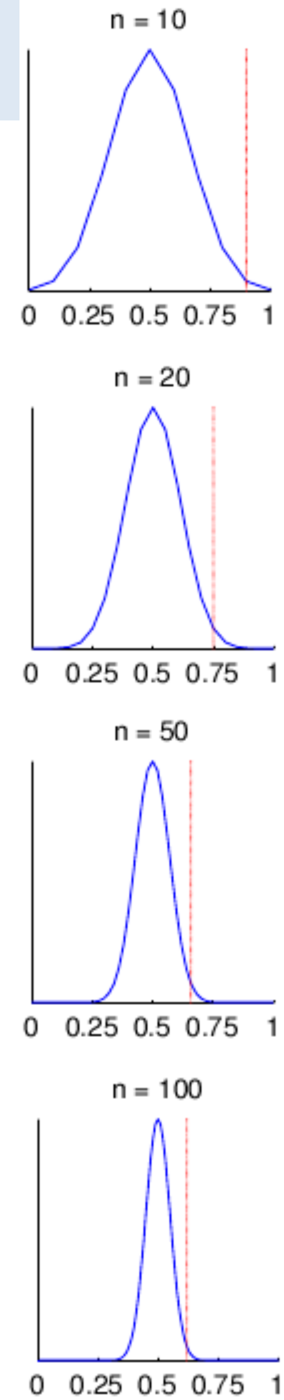
- null hypothesis:

  "classifier learnt nothing" → "predicts randomly"

- intuition:
  - a result is significant if very unlikely under null

how do we test predictions?

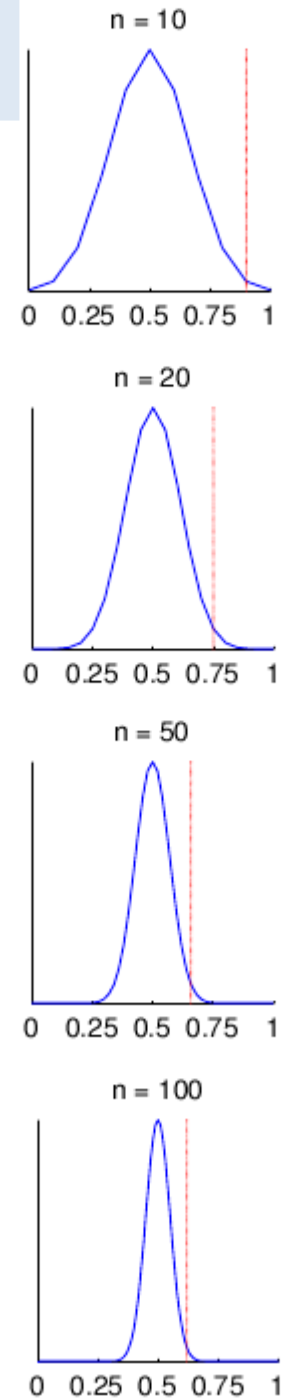
- X = #correct
- P(X|null) is binomial(#test,0.5)
- p-value is P(X>=result to test|null)



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

how do we test predictions?

- X = #correct
- P(X|null) is binomial(#test,0.5)
- p-value is P(X>=result to test|null)
- lots of caveats:
  - accuracy is an estimate
  - few examples ⟶ very uncertain
  - can get a confidence interval
  - must correct for multiple comparisons



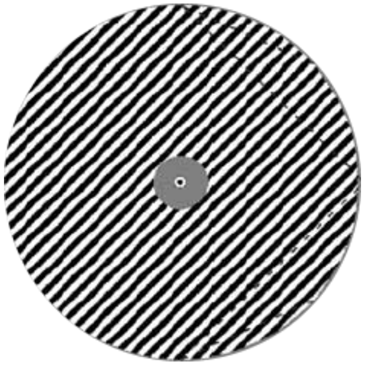
distribution under null
(0.05 p-value cut-off)

what questions can be tackled?

- is there information?
  (pattern discrimination)
- where/when is information present?
  (pattern localization)
- how is information encoded?
  (pattern characterization)

case study: orientation

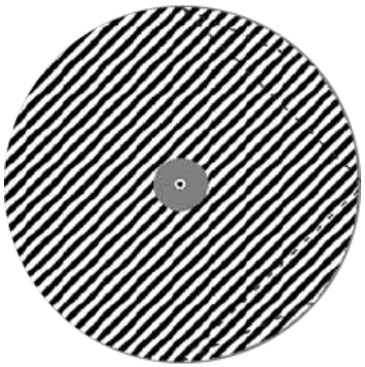
[Kamitani&Tong, 2005]



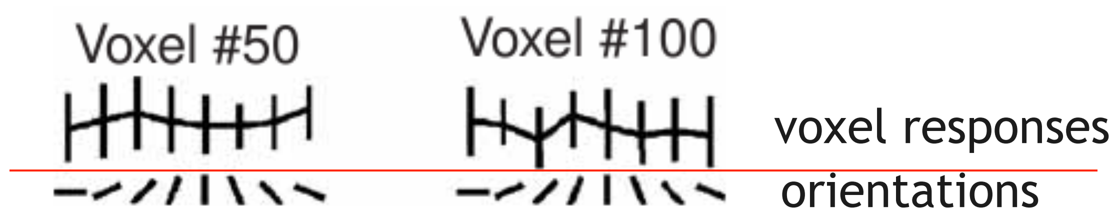
subjects see gratings in one of 8 orientations

case study: orientation

[Kamitani&Tong, 2005]



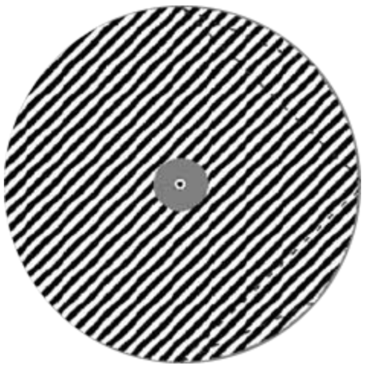
subjects see gratings in one of 8 orientations



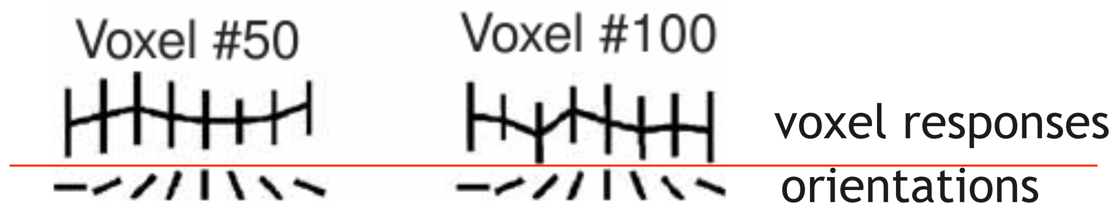
voxels in visual cortex respond similarly to different orientations

case study: orientation

[Kamitani&Tong, 2005]



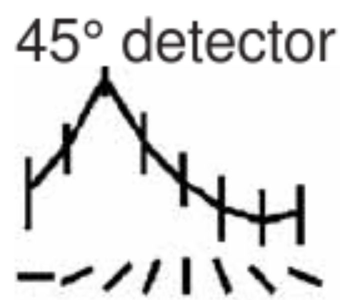
subjects see gratings in one of 8 orientations



...

voxels in visual cortex respond similarly to different orientations

linear SVM



yet, voxels can be combined to predict the orientation of the grating being seen!

features

- case study #1, features are voxels
- case study #2, features are voxels in visual cortex

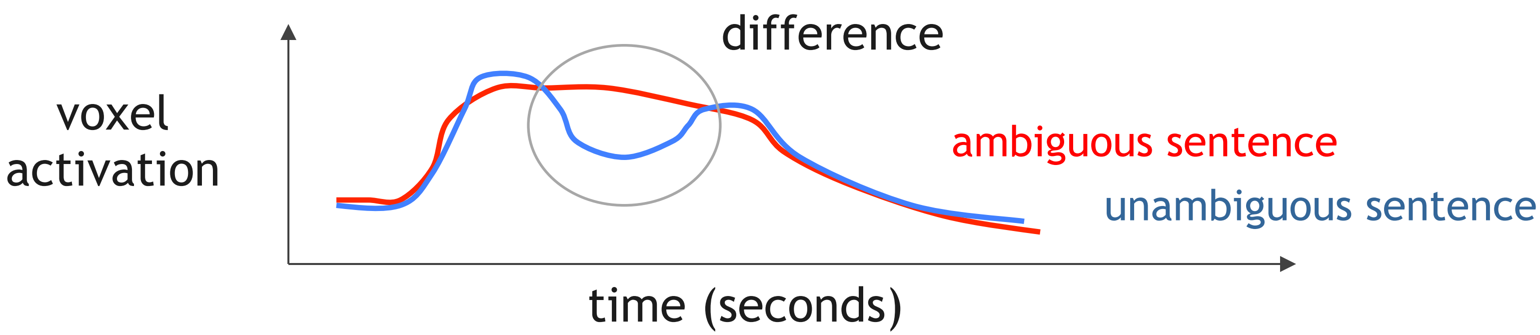
features

- case study #1, features are voxels
- case study #2, features are voxels in visual cortex

- what else could they be?

  voxels at particular times in a trial,

  syntactic ambiguity study



features

- you can also synthesize features
  - Singular Value Decomposition (SVD)
  - Independent Component Analysis (ICA)

examples | dataset | = | Z | basis images

voxels | | new features | | voxels

features

- you can also synthesize features
  - Singular Value Decomposition (SVD)
  - Independent Component Analysis (ICA)

examples	dataset	=	Z	basis images
	voxels		new features	voxels

- reduces to #features < #examples
- a feature has a spatial extent: basis image
- learn on the training set, convert the test set

example construction

- an example
  - can be created from one or more brain images
  - needs to be amenable to labelling

example construction

- an example
  - can be created from one or more brain images
  - needs to be amenable to labelling
- some possibilities
  - the brain image from a single TR
  - the average image in a trial or a block
  - the image of beta coefficients from deconvolution

example construction

- an example
  - can be created from one or more brain images
  - needs to be amenable to labelling
- some possibilities
  - the brain image from a single TR
  - the average image in a trial or a block
  - the image of beta coefficients from deconvolution
- caveats
  - remember the haemodynamic response time-to-peak
  - images for two examples not separate “enough”
    - in test set, lowers the effective #examples in statistical test
    - in between train and test set, “peeking”/”circularity”
  - read [Kriegeskorte et al. 2009] (“double dipping”)

localization

- key idea #1

  test conclusions pertain to whatever is fed to the classifier

localization

- key idea #1

  test conclusions pertain to whatever is fed to the classifier

- key idea #2

  one can predict anything that can be labelled:

  stimuli, subject percepts, behaviour, response, ...

localization

- key idea #1

  test conclusions pertain to whatever is fed to the classifier

- key idea #2

  one can predict anything that can be labelled:

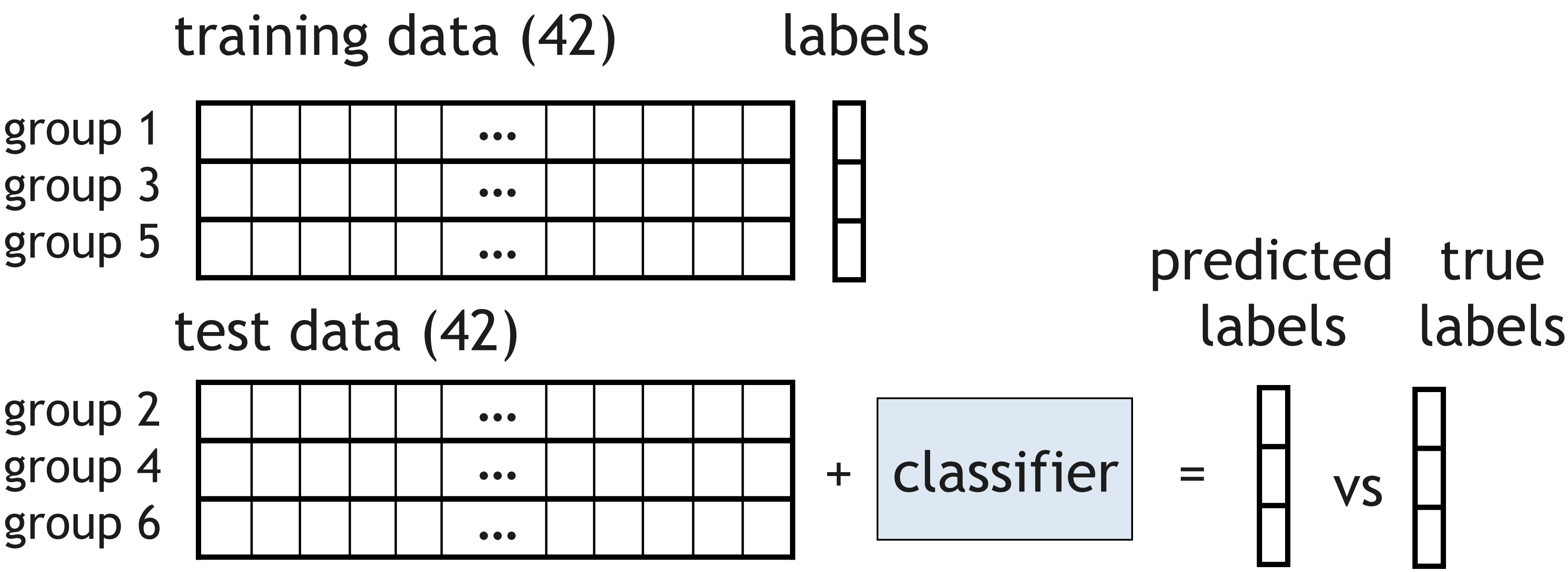
  stimuli, subject percepts, behaviour, response, ...

so what criteria can we use?

- location
- time
- voxel behaviour or relationship to label
  - aka “feature selection”

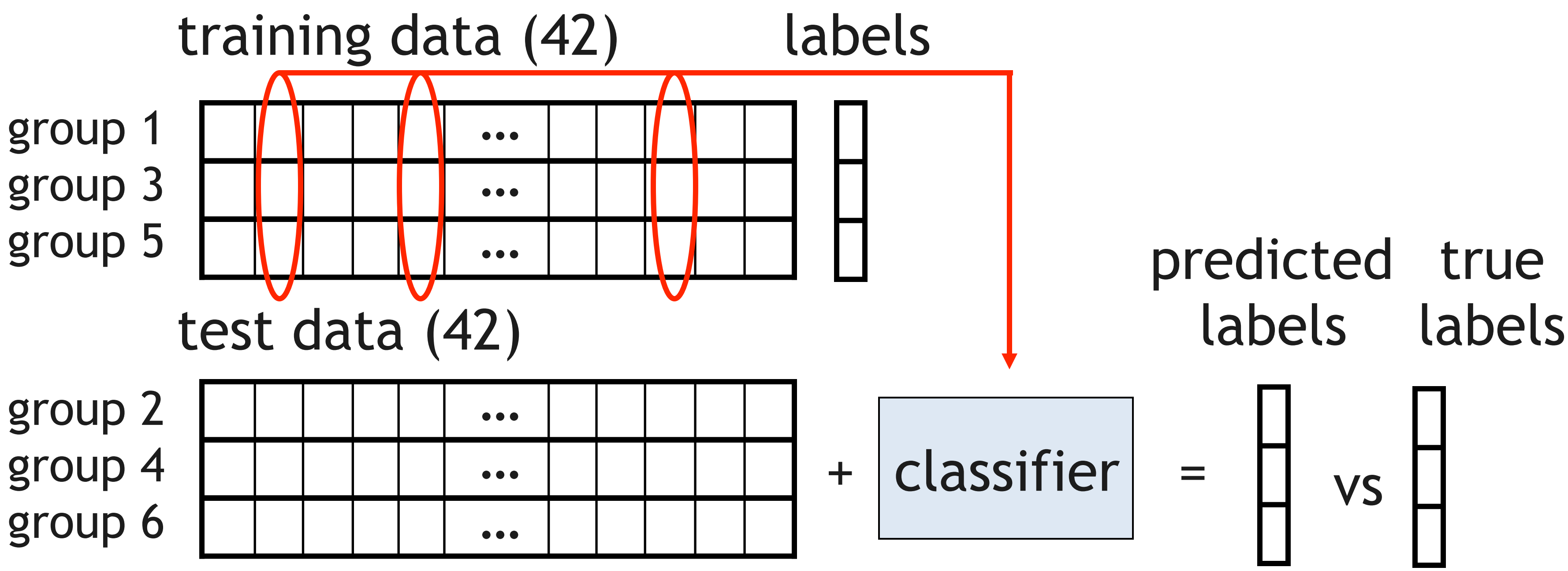
feature (voxel) selection

- what does it look like



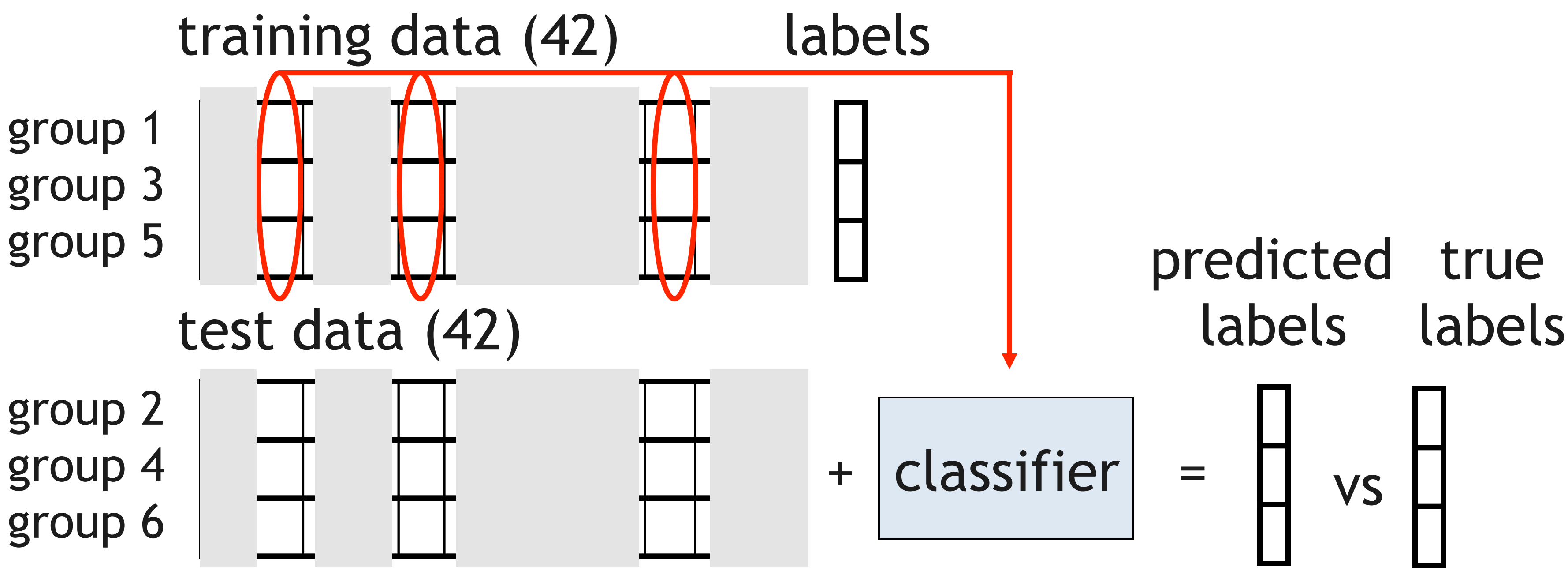
feature (voxel) selection

- what does it look like



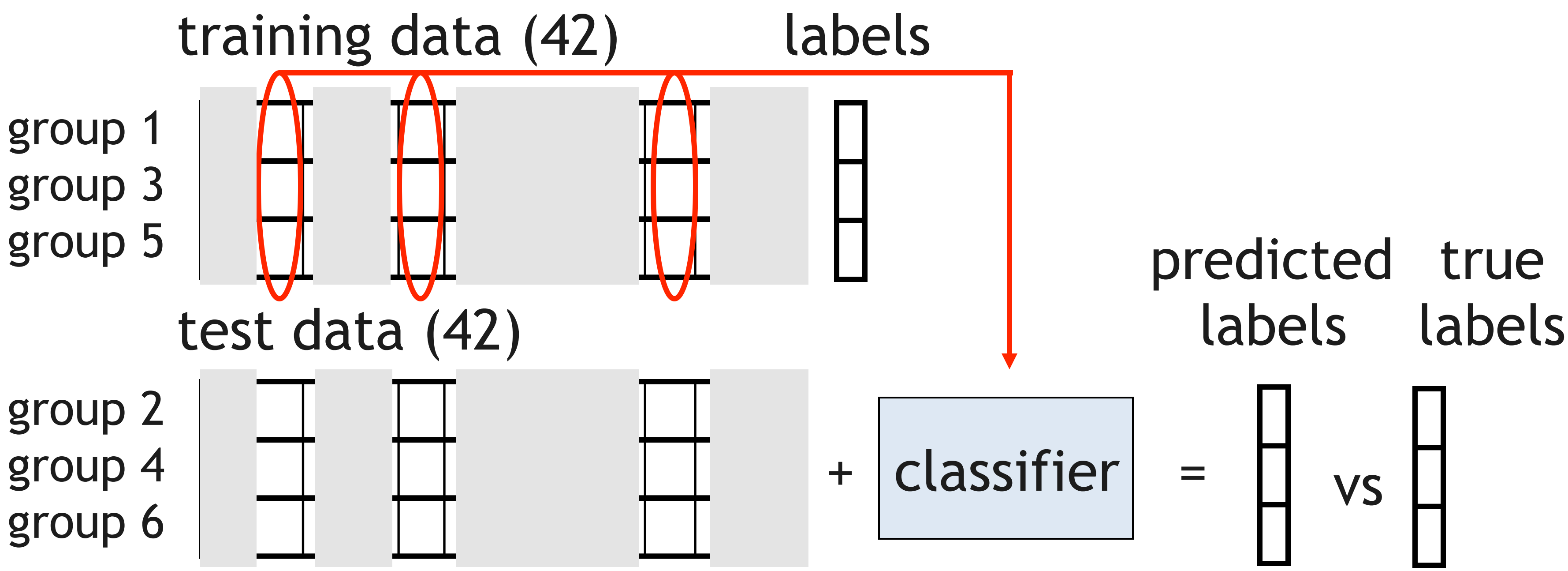
feature (voxel) selection

- what does it look like



feature (voxel) selection

- what does it look like



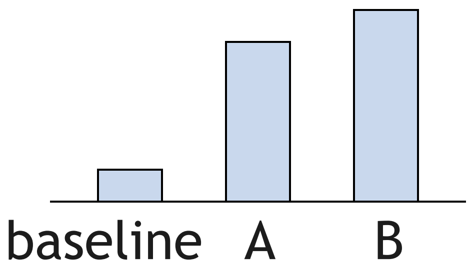
- great for improving prediction accuracy
- but
  - voxels often come from all over the place
  - very little overlap in selected across training sets

feature (voxel) selection

- look at the training data and labels

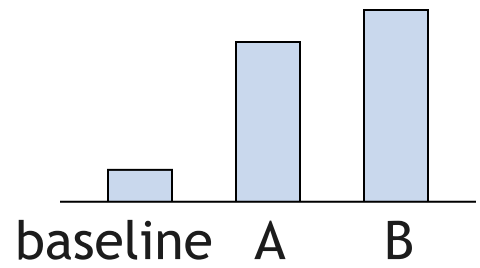
feature (voxel) selection

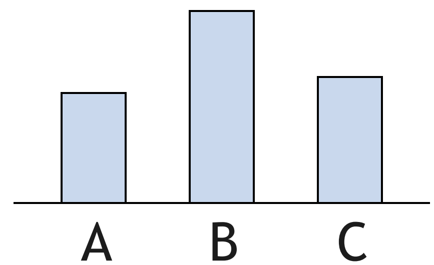
- look at the training data and labels
- a few criteria:
  - difference from baseline



feature (voxel) selection

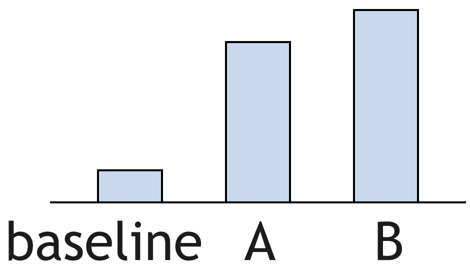
- look at the training data and labels
- a few criteria:
  - difference from baseline
  - difference between classes (e.g. ANOVA)

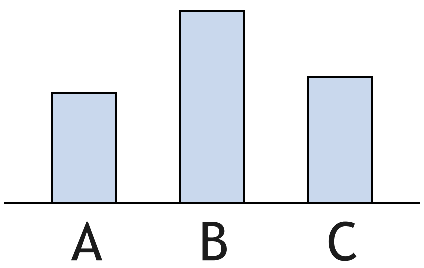


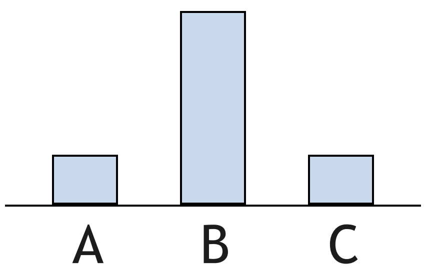


feature (voxel) selection

- look at the training data and labels
- a few criteria:
  - difference from baseline
  - difference between classes (e.g. ANOVA)
  - preferential response to one class

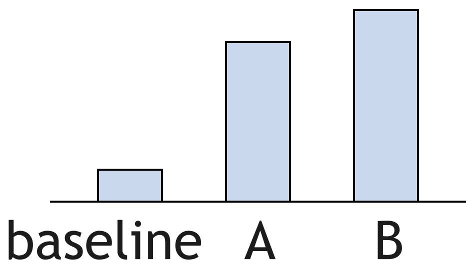


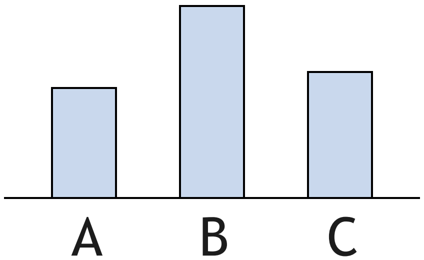


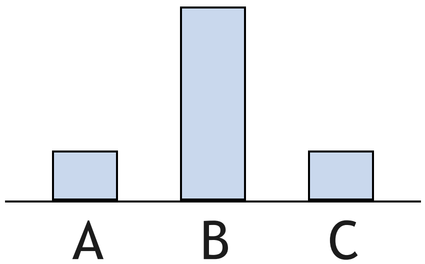


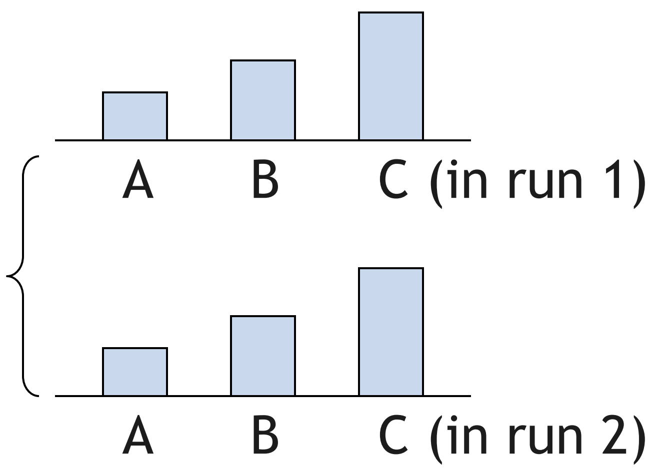
feature (voxel) selection

- look at the training data and labels
- a few criteria:
  - difference from baseline
  - difference between classes (e.g. ANOVA)
  - preferential response to one class
  - stability
  - ...





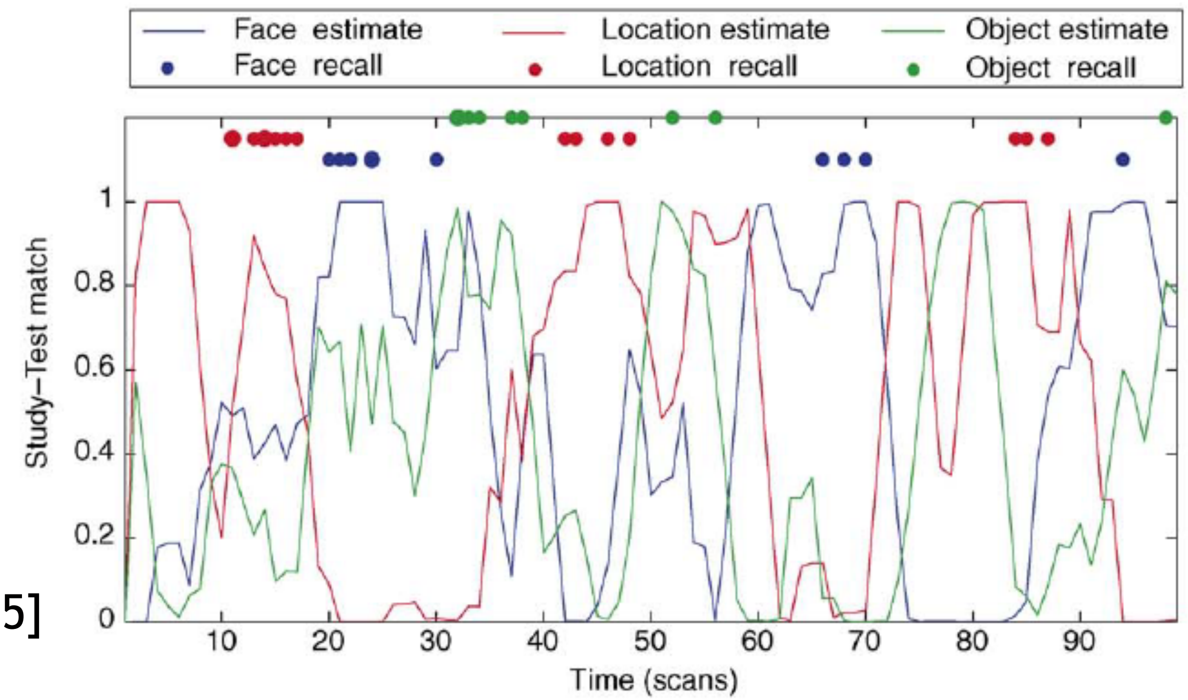




case study: category reinstatement

- items from 3 categories: faces, locations, objects
- classifier trained during study of the grouped items
- detect category reinstatement during free recall (test)

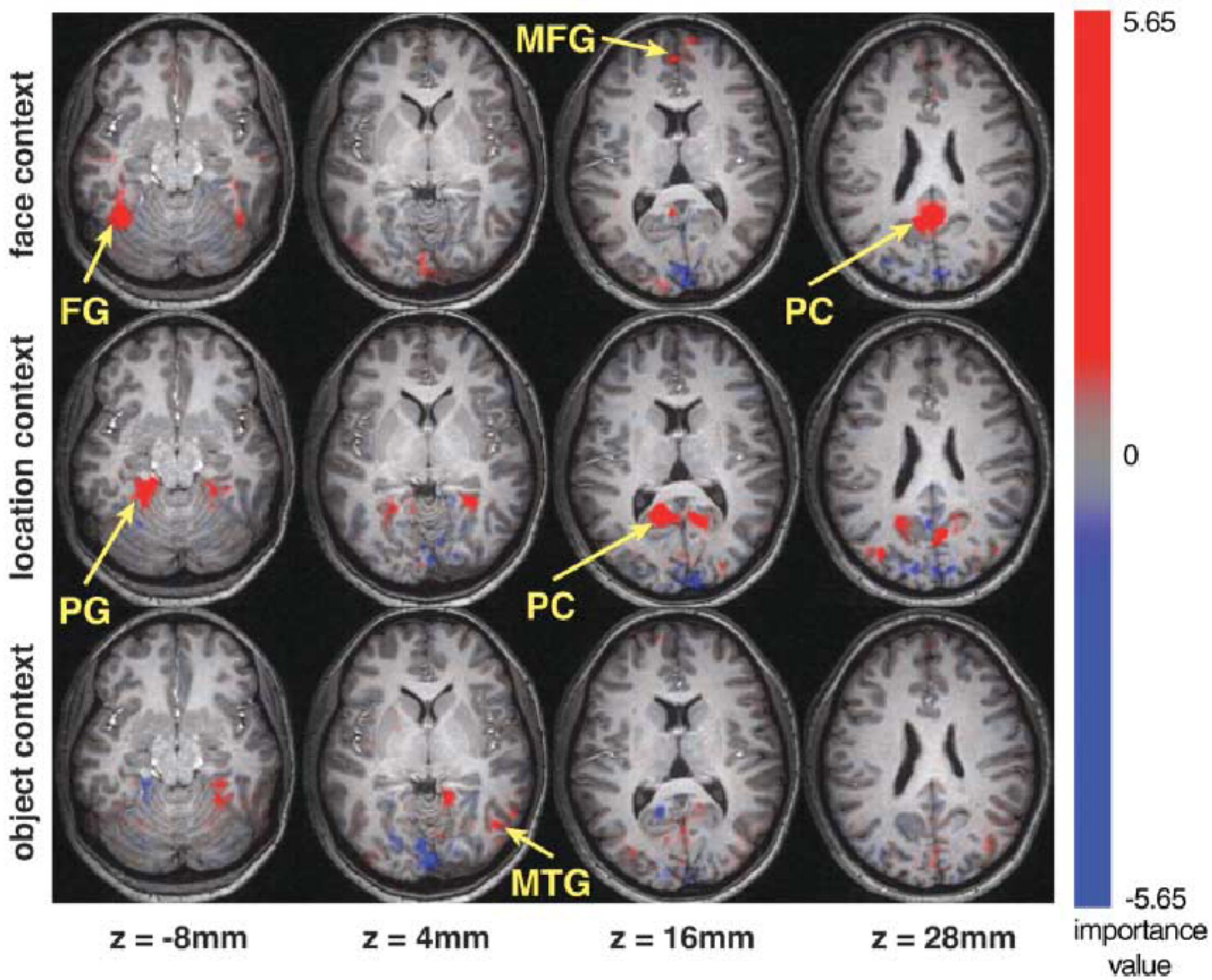
faces | locations | objects



[Polyn et al 2005]

temporal localization

voxel influence on reinstatement estimate



faces

locations

objects

[Polyn et al 2005]

temporal localization

key ideas:

- trained classifiers may be used as detectors
  to localize the time at which information is present

- test examples can be temporally overlapping

- it is feasible to decode endogenous events

what questions can be tackled?

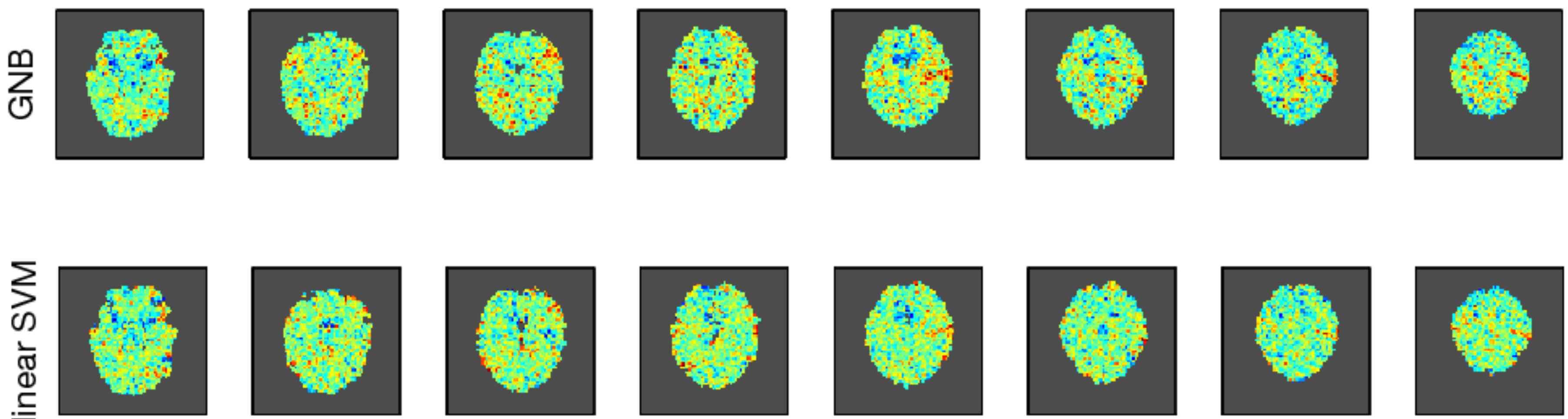
- is there information?
  (pattern discrimination)
- where/when is information present?
  (pattern localization)
- how is information encoded?
  (pattern characterization)

classifier dissection

- a classifier learns to relate features to labels
- we can consider not just what gets selected but also how the classifier uses it
- in linear classifiers, look at voxel weights

classifier dissection

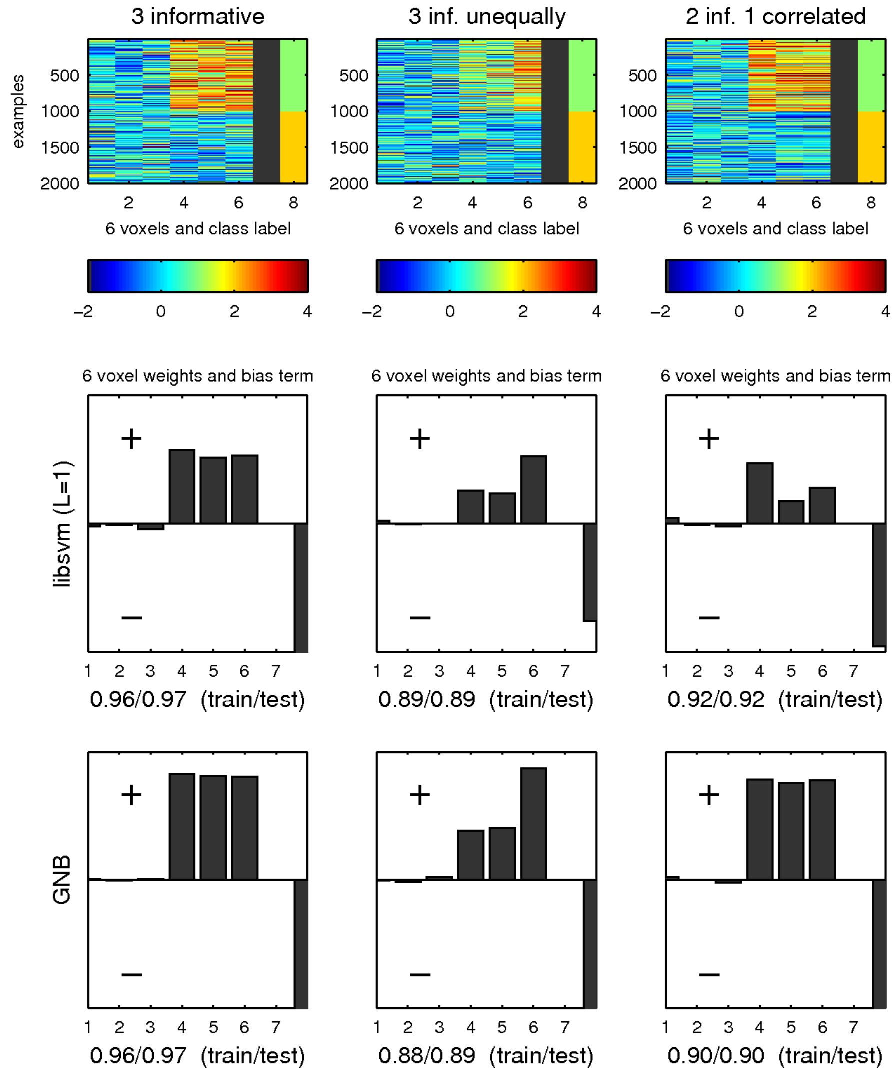
- weights depend on classifier assumptions
- less of an issue if feature selection is used



- weights are similar, but accuracy difference 15%

classifier dissection

- assumption effects on synthetic data



"voxels"

SVM

GNB

see [Pereira, Botvinick 2010]

case study: 8 categories

[Haxby et al., 2001]

subjects see photographs
of objects in 8 categories

- faces, houses, cats, bottles, scissors, shoes, chairs, scrambled
- block: series of photographs of the same category

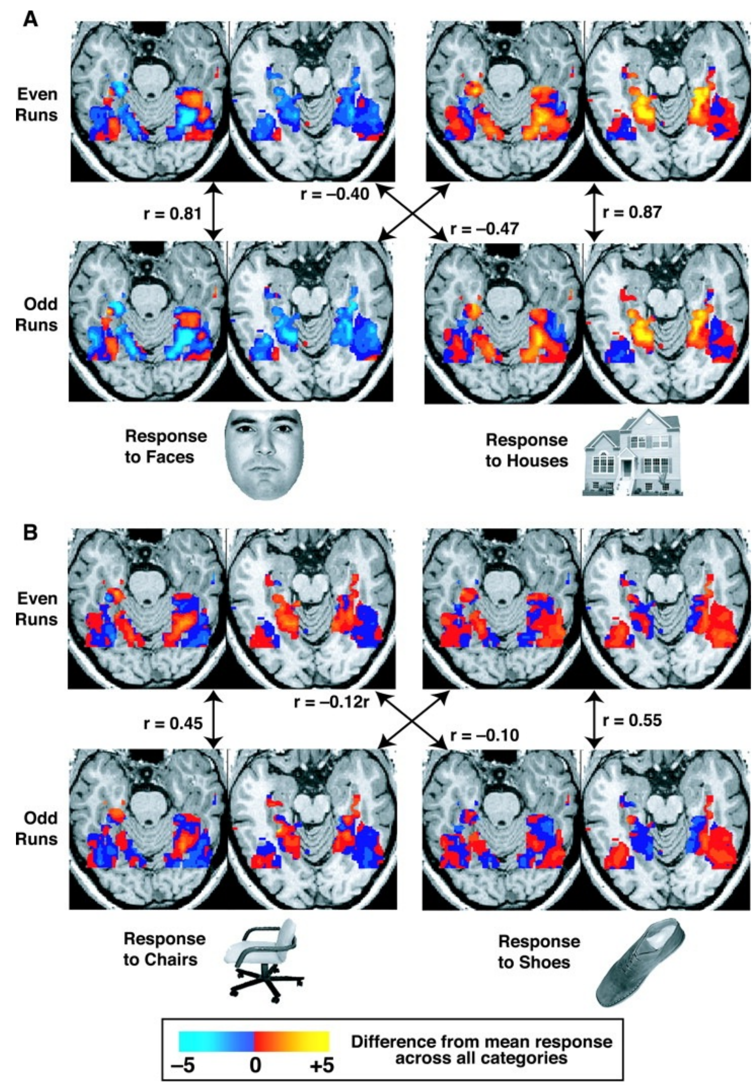
case study: 8 categories

[Haxby et al., 2001]

nearest neighbour classifier

- all category pair distinctions
- selects voxels by location
  - fusiform gyrus
  - rest of temporal cortex
- selects voxels by behaviour
  - responsive to single category
  - responsive to multiple

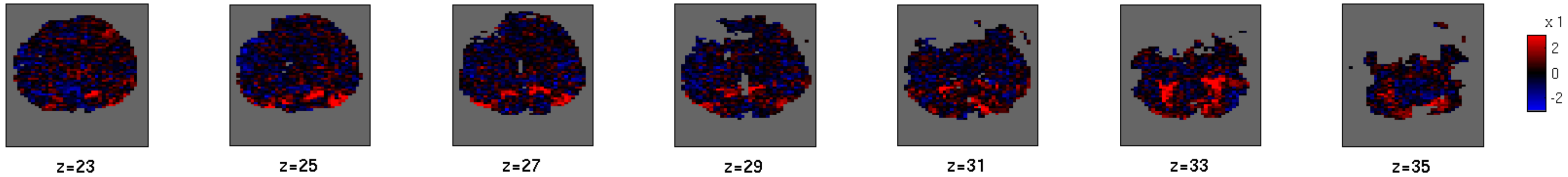
- logic:
  - restrict by location/behaviour
  - see if there is still information



classifier dissection

1) whole-brain logistic regression weights

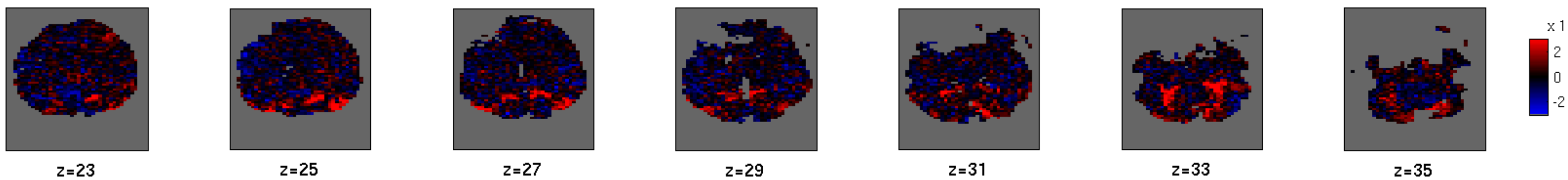
faces



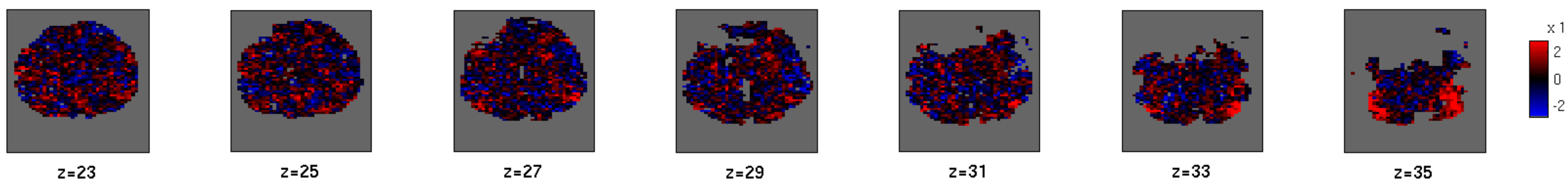
classifier dissection

1) whole-brain logistic regression weights

faces



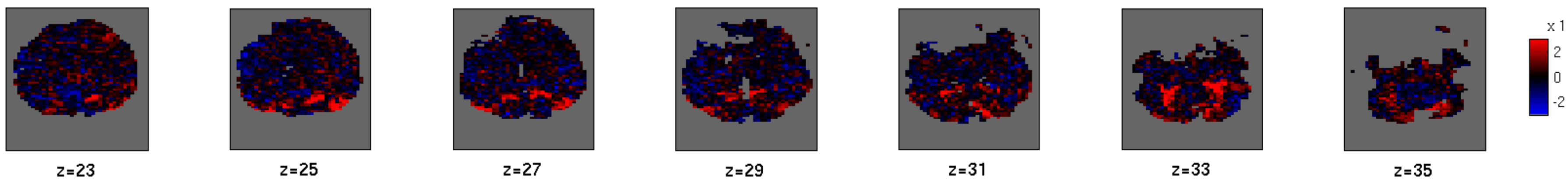
houses



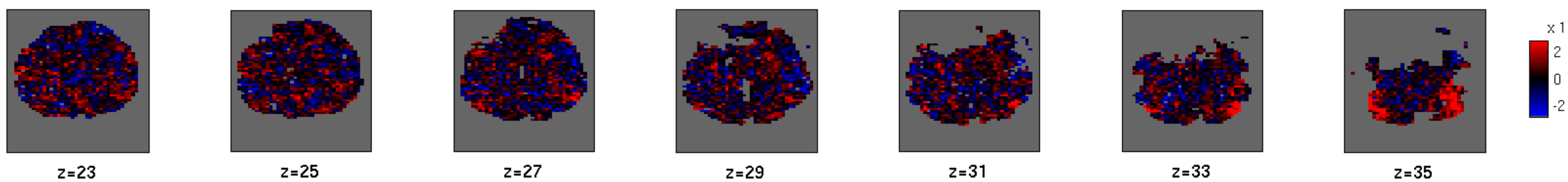
classifier dissection

1) whole-brain logistic regression weights

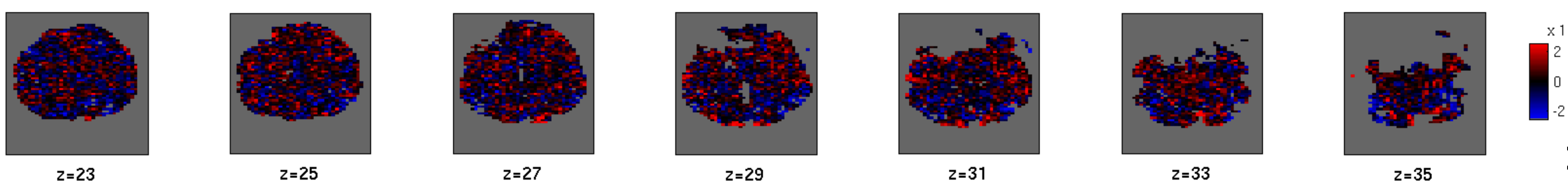
faces



houses



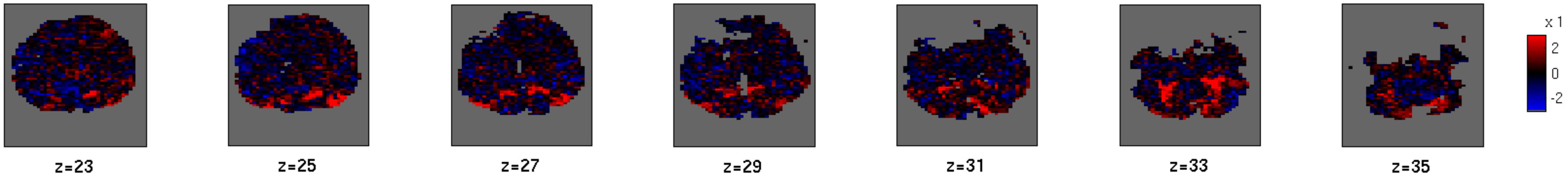
chairs



classifier dissection

2) feature selection

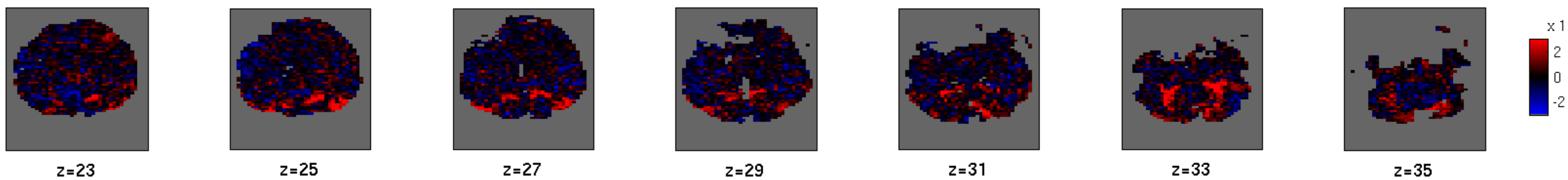
faces



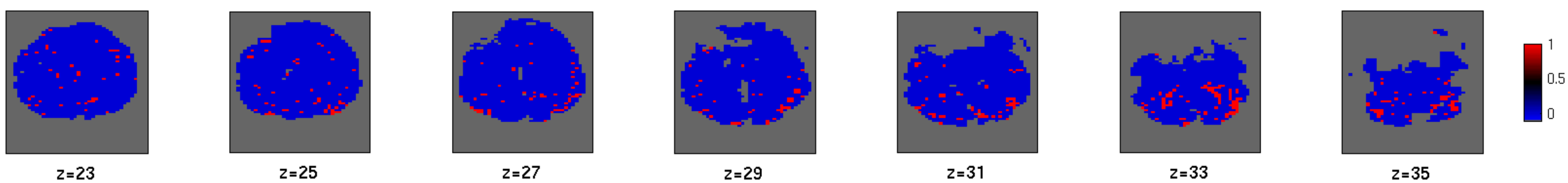
classifier dissection

2) feature selection

faces



top 1000 voxels



classifier dissection

whole brain classifier

- accuracy 40% in this case
- many more features than examples => simple classifier
- messy maps (can bootstrap to threshold)

classifier dissection

whole brain classifier

- accuracy 40% in this case
- many more features than examples => simple classifier
- messy maps (can bootstrap to threshold)

voxel selection

- accuracy 80% in this case
- sparse, non-reproducible maps
- different methods pick different voxels

classifier dissection

whole brain classifier

- accuracy 40% in this case
- many more features than examples => simple classifier
- messy maps (can bootstrap to threshold)

voxel selection

- accuracy 80% in this case
- sparse, non-reproducible maps
- different methods pick different voxels

a lot of work

[Mitchell et al 2004],[Norman et al 2006], [Haynes et al 2006], [Pereira 2007], [De Martino et al 2008], [Carrol et al 2009], [Pereira et al 2009]

classifier dissection

[Hanson et al., 2004]

neural network

- one-of-8 classifier
- temporal cortex

learned model

- hidden units
- activation patterns across units reflect category similarity

classifier dissection

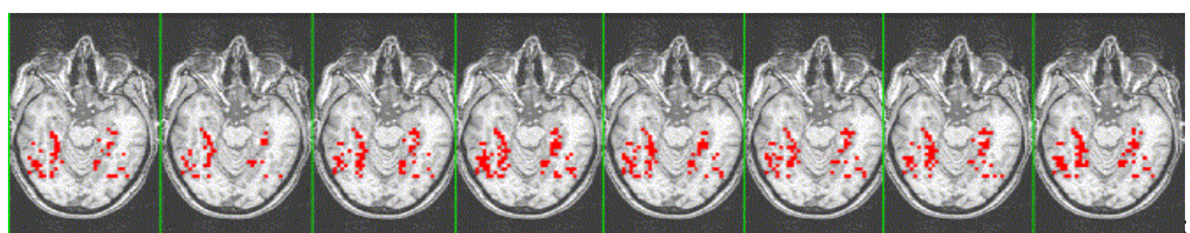
[Hanson et al., 2004]

neural network

- one-of-8 classifier
- temporal cortex

learned model

- hidden units
- activation patterns across units reflect category similarity
- sensitivity analysis
  - add noise to voxels
  - which ones lead to classification error?



classifier dissection conclusions

- if linear works, you can look at weights
  - but know what your classifier does (or try several)
  - read about bootstrapping (and [Strother et al. 2002])

classifier dissection conclusions

- if linear works, you can look at weights
  - but know what your classifier does (or try several)
  - read about bootstrapping (and [Strother et al. 2002])
- voxel selection
  - may be necessary in multiclass situations
  - try multiple methods and **look** at the voxels they pick
  - voxels picked may be a small subset of informative ones
  - report all #s of voxels selected or
    use cross-validation on training set to pick a # to use

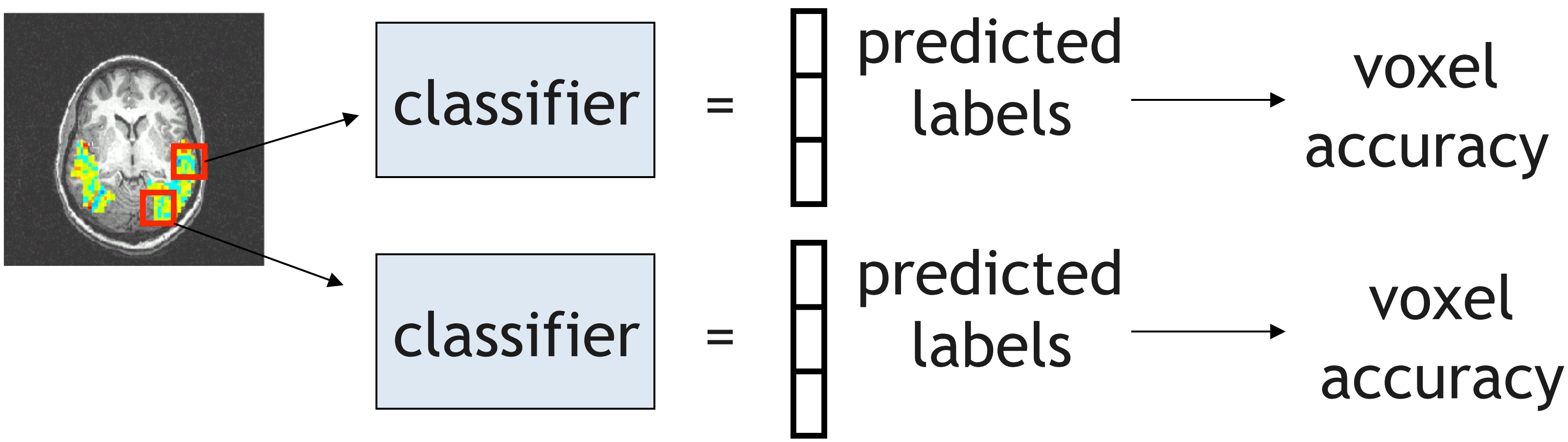
classifier dissection conclusions

- if linear works, you can look at weights
  - but know what your classifier does (or try several)
  - read about bootstrapping (and [Strother et al. 2002])
- voxel selection
  - may be necessary in multiclass situations
  - try multiple methods and **look** at the voxels they pick
  - voxels picked may be a small subset of informative ones
  - report all #s of voxels selected or
    use cross-validation on training set to pick a # to use
- nonlinear classifiers
  - worth trying, but try linear + voxel selection first
  - look at [Hanson et al. 2004] and [Rasmussen et al. 2011]
    for ways of gauging voxel influence on classifier

information-based mapping (searchlights)

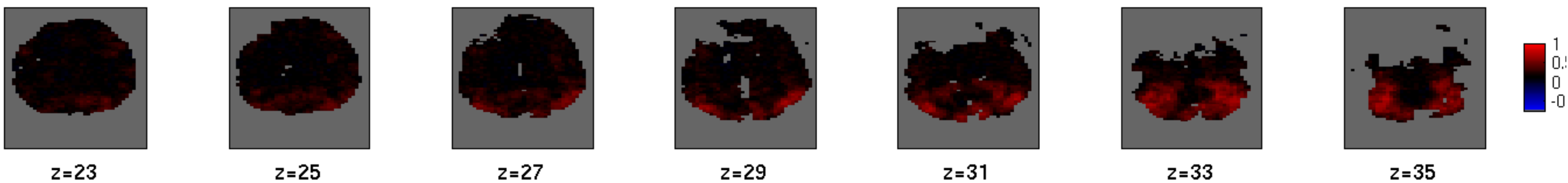
[Kriegeskorte 2006]

- focus the classifier on small voxel neighbourhoods
- more examples than features
- can learn voxel relationships (e.g. covariance matrix)
- can train nonlinear classifiers



information mapping

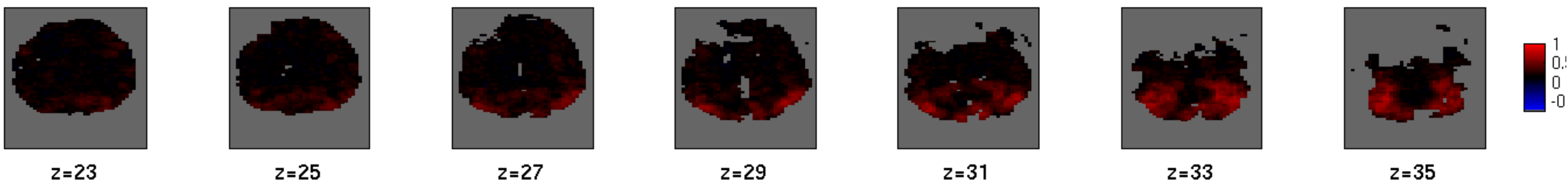
- on 8 categories, yields an accuracy map



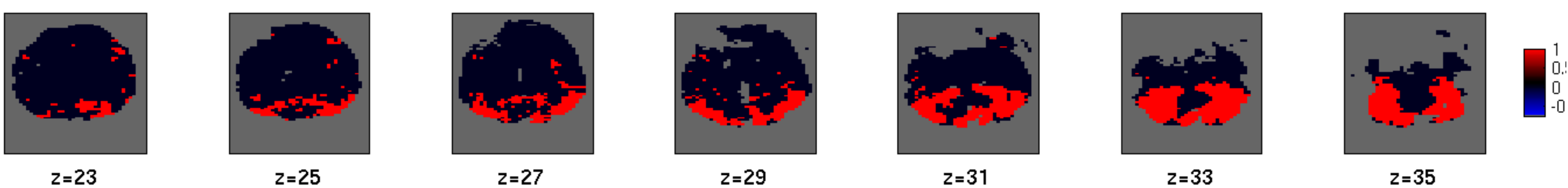
- also local information: covariance, voxel weights

information mapping

- on 8 categories, yields an accuracy map



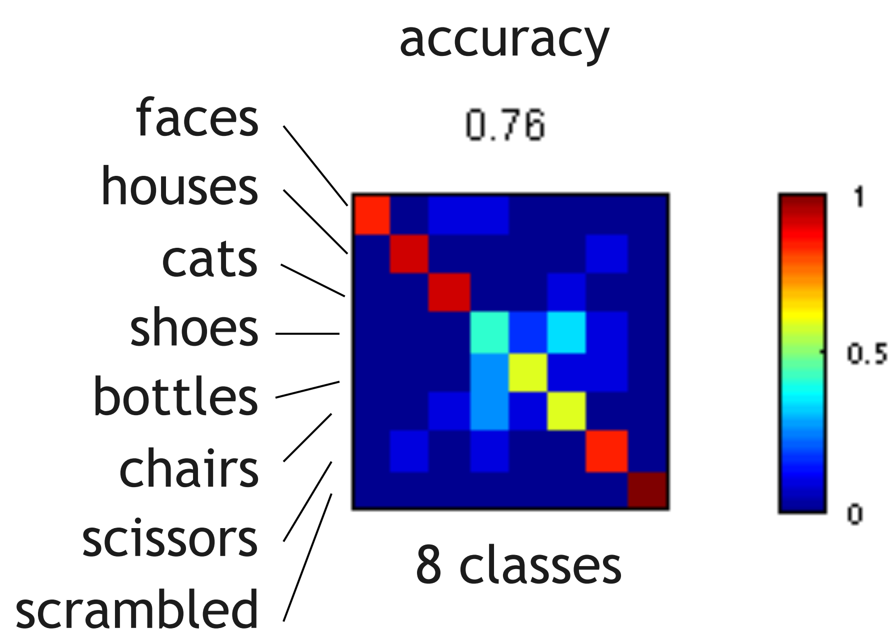
- also local information: covariance, voxel weights
- can be thresholded for statistical significance



information mapping

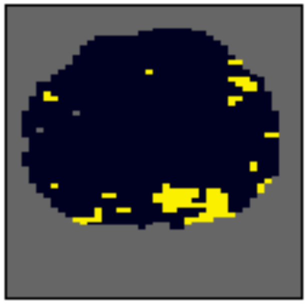
- “H0: chance level” deems many voxels significant

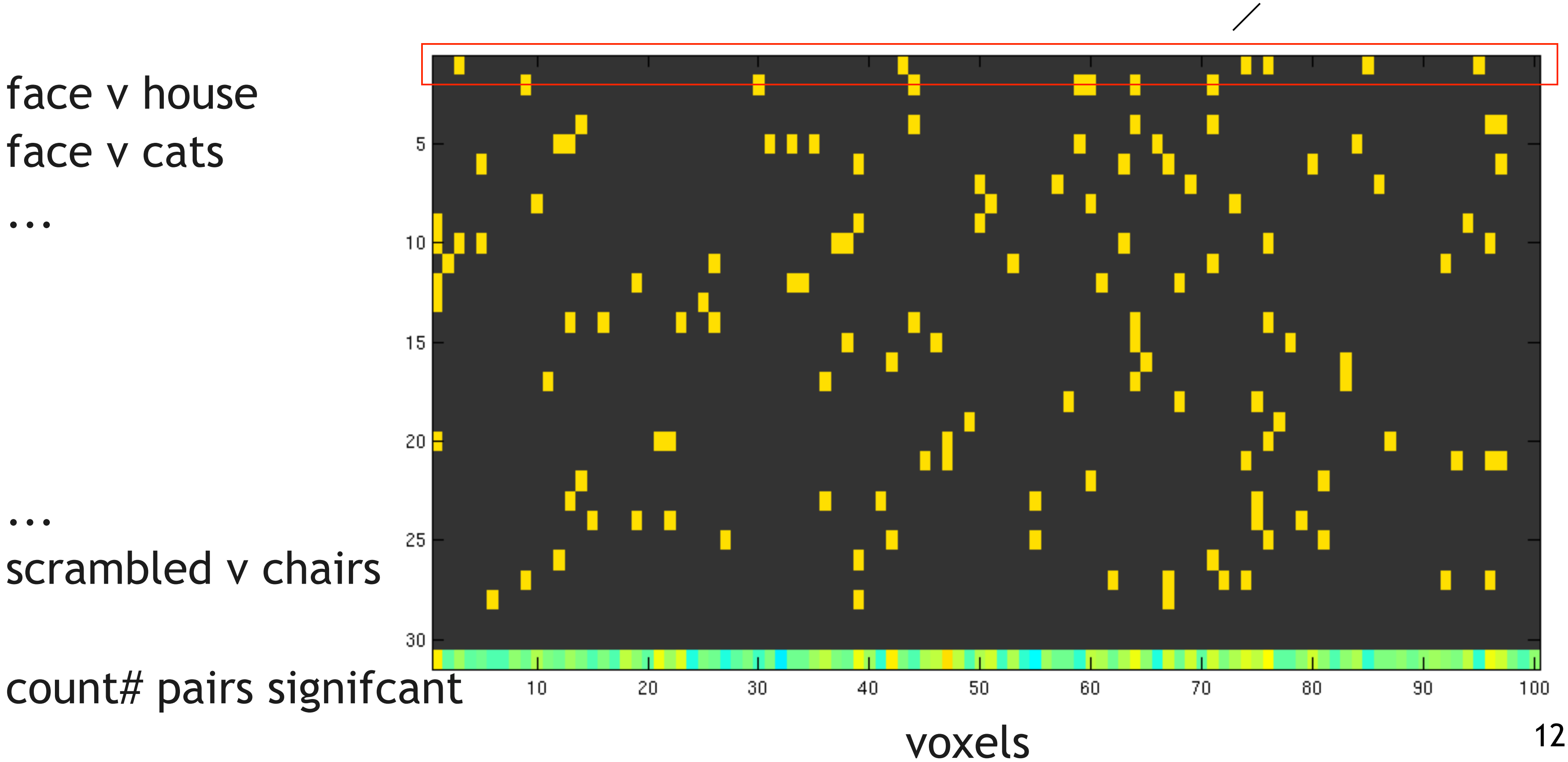
- what does accuracy mean in multi-class settings?
  - confusion matrix
  - for each class, what do examples belonging to it get labelled as?



information mapping

- contrast each pair of classes directly
- threshold accuracy to a binary image





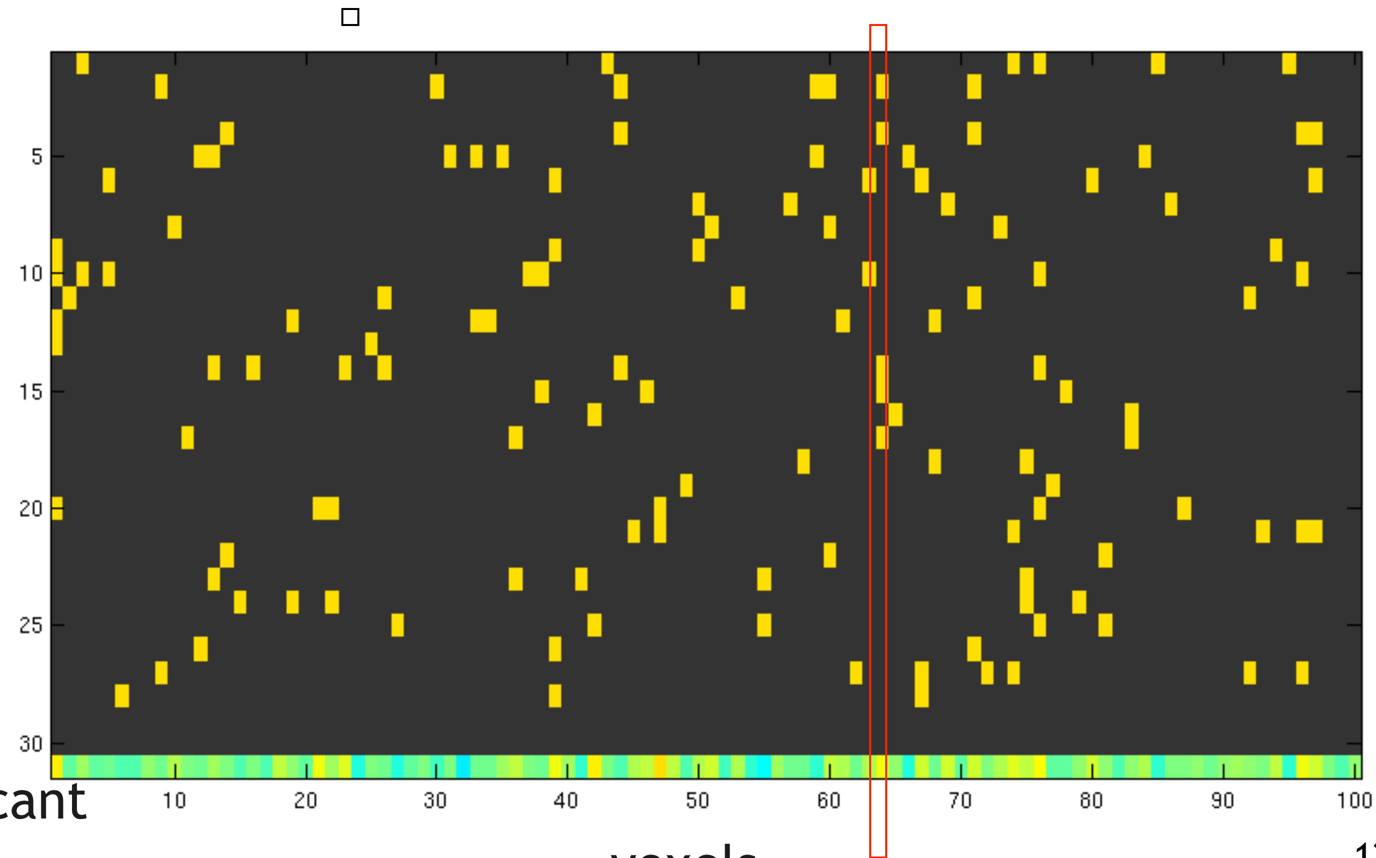
information mapping

- each voxel has a *binary profile* across pairs
- how many different ones?

face v house
face v cats
...

...
scrambled v chairs

count# pairs signifcant

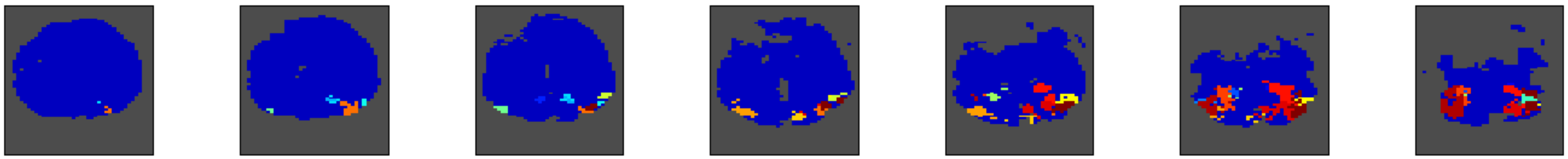


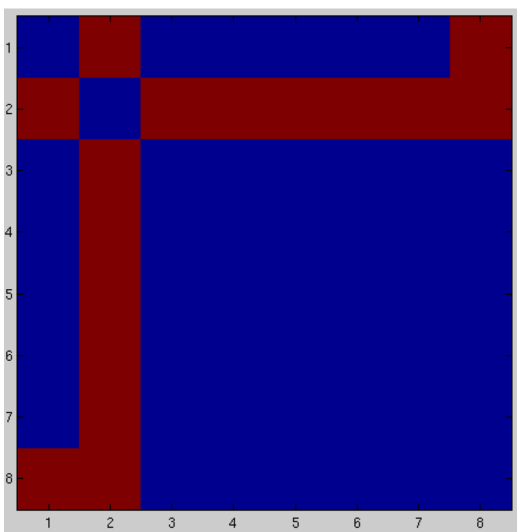
voxels

information mapping

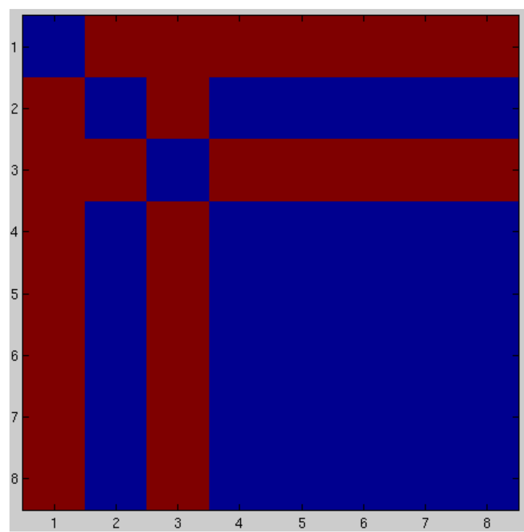
[Pereira&Botvinick, KDD 2011]

- a binary profile is a kind of confusion matrix
- only a few hundred profiles, many similar
- cluster them!





houses versus all else



faces and cat versus all else

1 face
2 house
3 cat
4 bottle
5 scissors
6 chairs
7 shoes
8 scrambled

information mapping

- a map of accuracy works well in 2 class situation
- some classifiers seem consistently better
  see [Pereira&Botvinick 2010] for details
- easy to get above chance with multiway prediction
  so reporting accuracy or #significant is not enough
- consider reporting common profiles or
  grouping classes into distinctions to test

what questions can be tackled?

- is there information?
  (pattern discrimination)
- where/when is information present?
  (pattern localization)
- how is information encoded?
  (pattern characterization)

to get started

- MVPA toolbox (in MATLAB)

  http://code.google.com/p/princeton-mvpa-toolbox
- PyMVPA (in Python)

  http://www.pymvpa.org
- support the whole workflow
  - cross-validation, voxel selection, multiple classifiers,...
  - helpful mailing lists (most people are on both)

to get started

- MVPA toolbox (in MATLAB)

  http://code.google.com/p/princeton-mvpa-toolbox
- PyMVPA (in Python)

  http://www.pymvpa.org
- support the whole workflow
  - cross-validation, voxel selection, multiple classifiers,...
  - helpful mailing lists (most people are on both)

- Searchmight (in MATLAB, shameless plug)
  - http://minerva.csbmb.princeton.edu/searchmight
  - special purpose toolbox for information mapping
  - can be used with MVPA toolbox

Thank you!

questions?