Multi-voxel pattern analysis:

Decoding Mental States from fMRI Activity Patterns

Artwork by Leon Zernitsky

Jesse Rissman

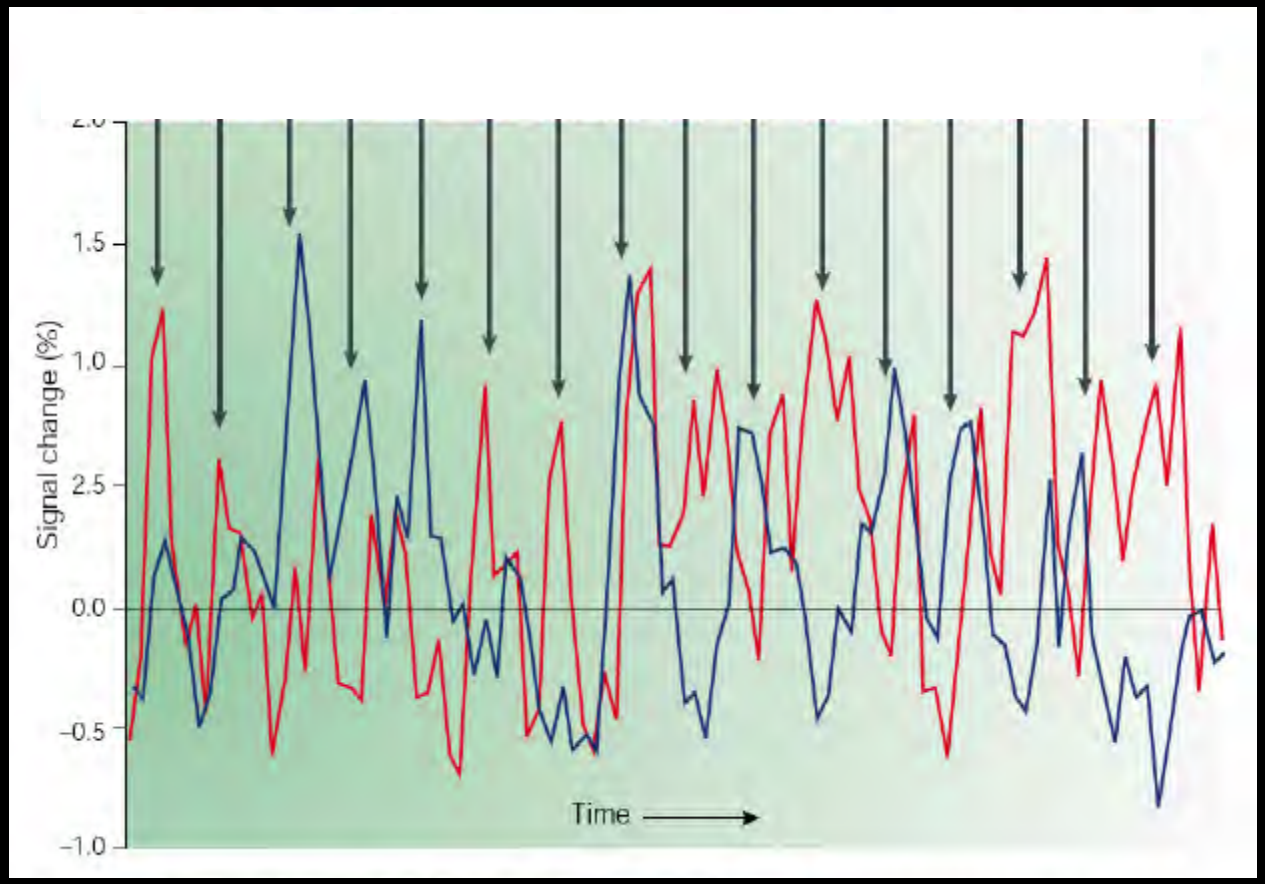
NITP Summer Program 2012

Part 1 of 2

Goals of Multi-voxel Pattern Analysis

- Decoding percepts or thoughts (a.k.a. “mind-reading”)
  - What is a person perceiving, imagining, planning, or remembering?
- Decoding brain patterns (activity or connectivity) that distinguish individuals
  - Useful for diagnosis
- Characterizing the distributed cortical representations mediating specific cognitive processes
  - How are the mental representations of stimuli parsed in the brain?
  - What features are extracted by which cortical areas or networks?
- Providing a index of the instantaneous activation level of specific mental representations
  - Can use this information to test psychological theories

An Early Attempt to Decode Individual Thoughts



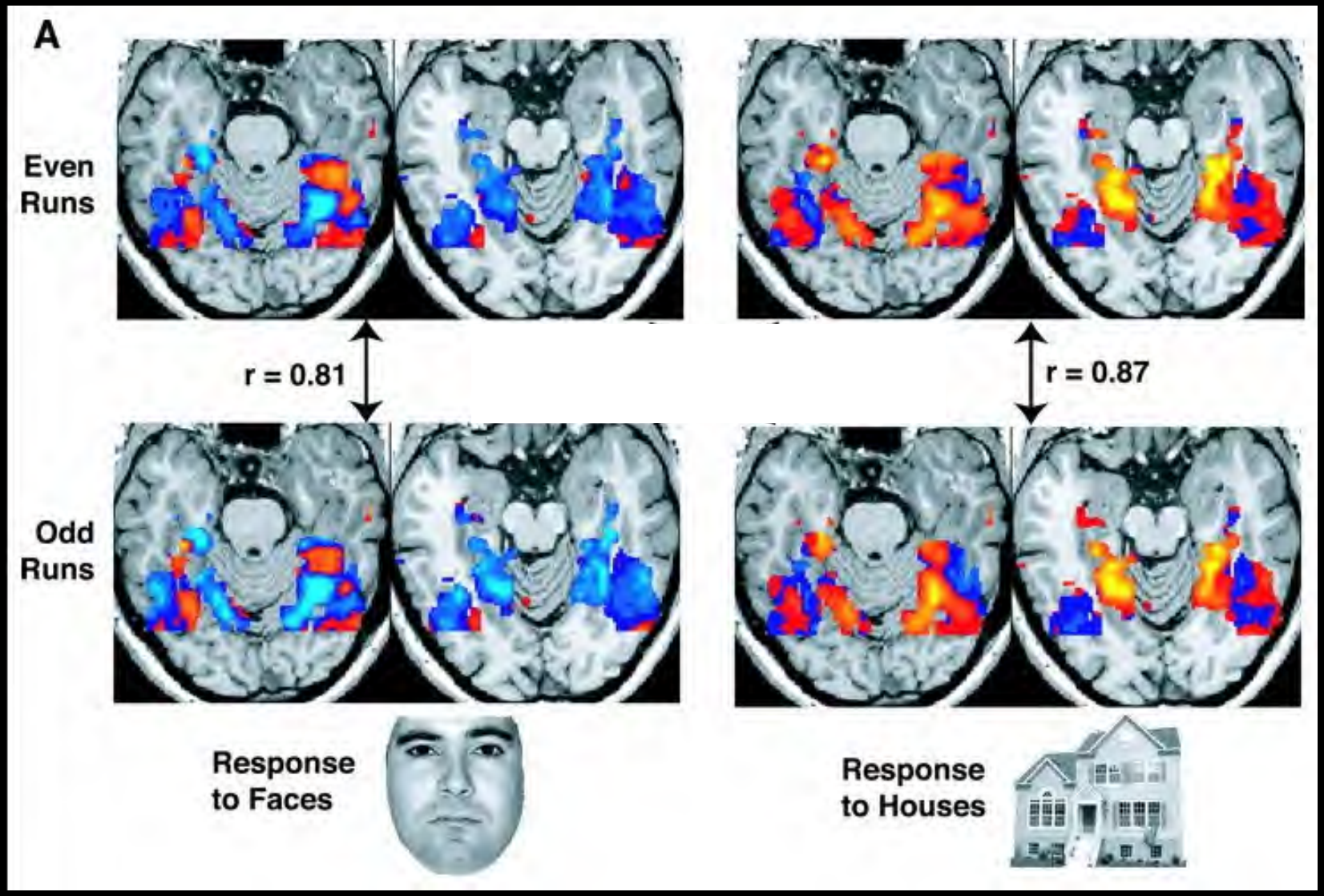
85% correct classification

O' Craven & Kanwisher, 2000

The Power of Patterns

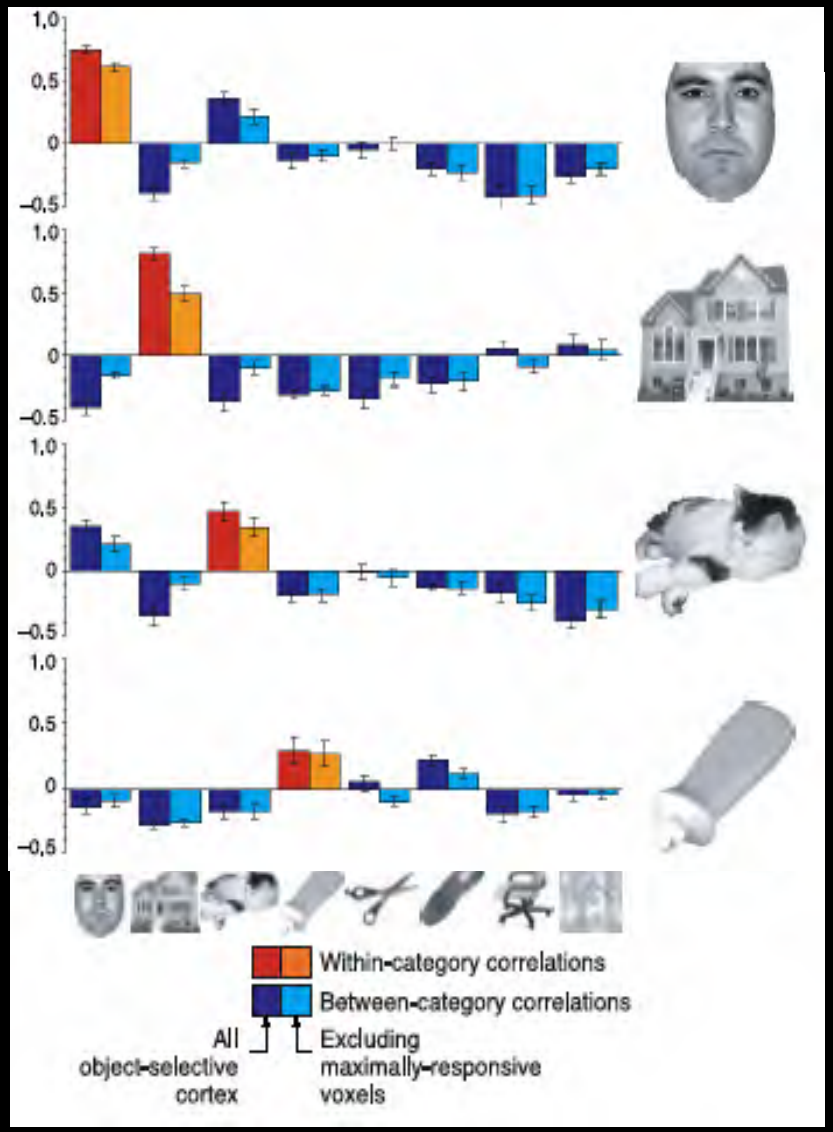
Haxby et al. (2001), *Science*

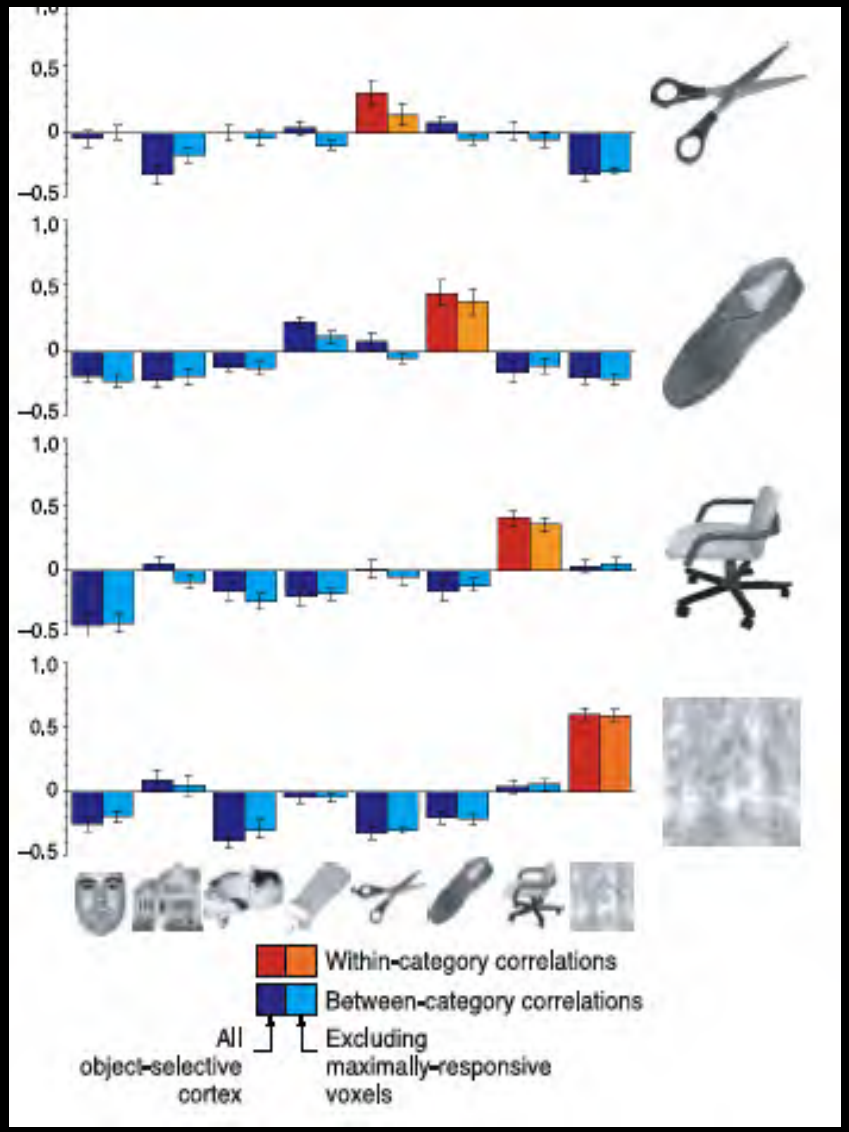
The Power of Patterns



Haxby et al. (2001), *Science*

The Power of Patterns





Haxby et al. (2001), *Science*

Adopting a Machine Learning Framework

ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

NeuroImage 19 (2003) 261–270

NeuroImage

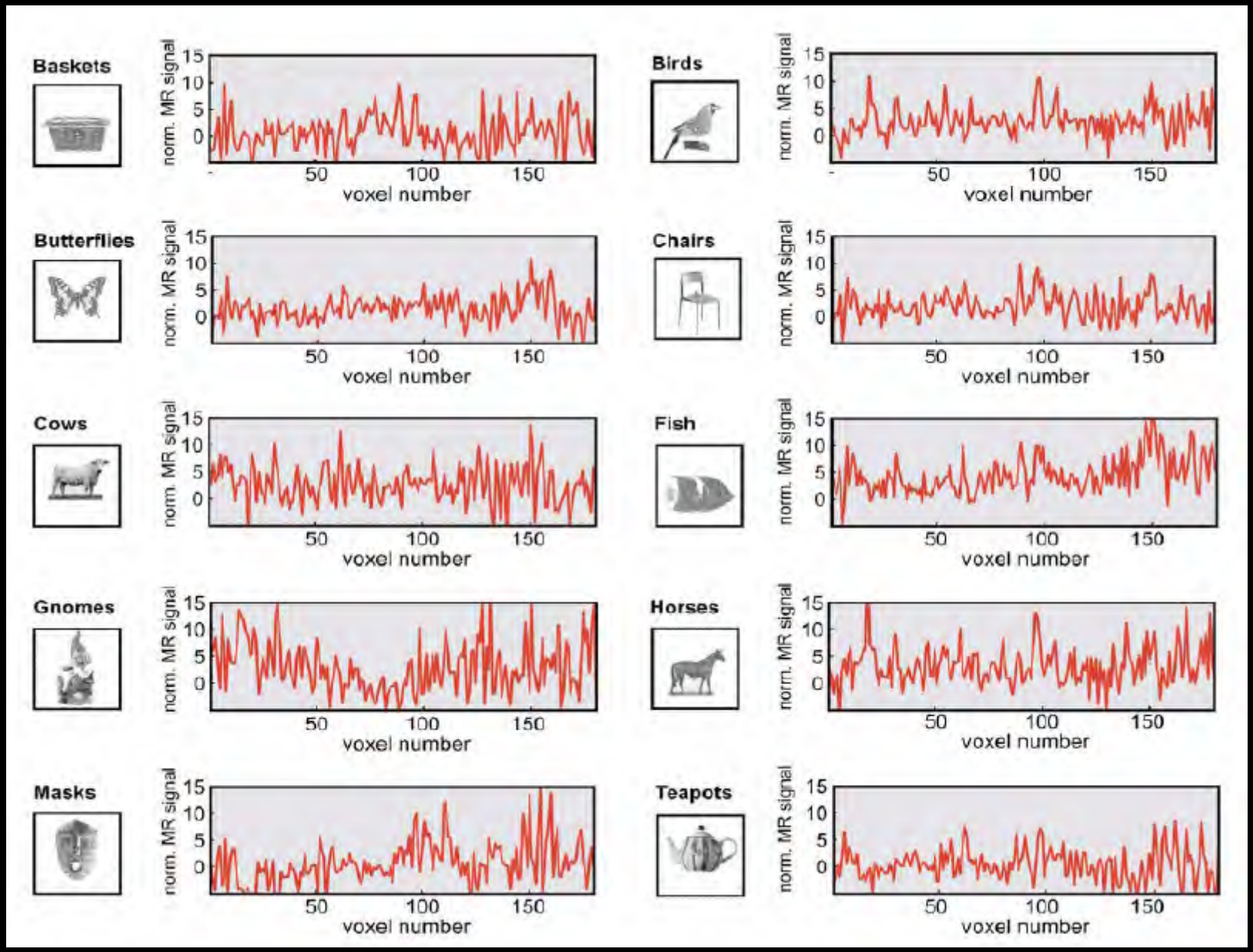
www.elsevier.com/locate/ynimg

Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex

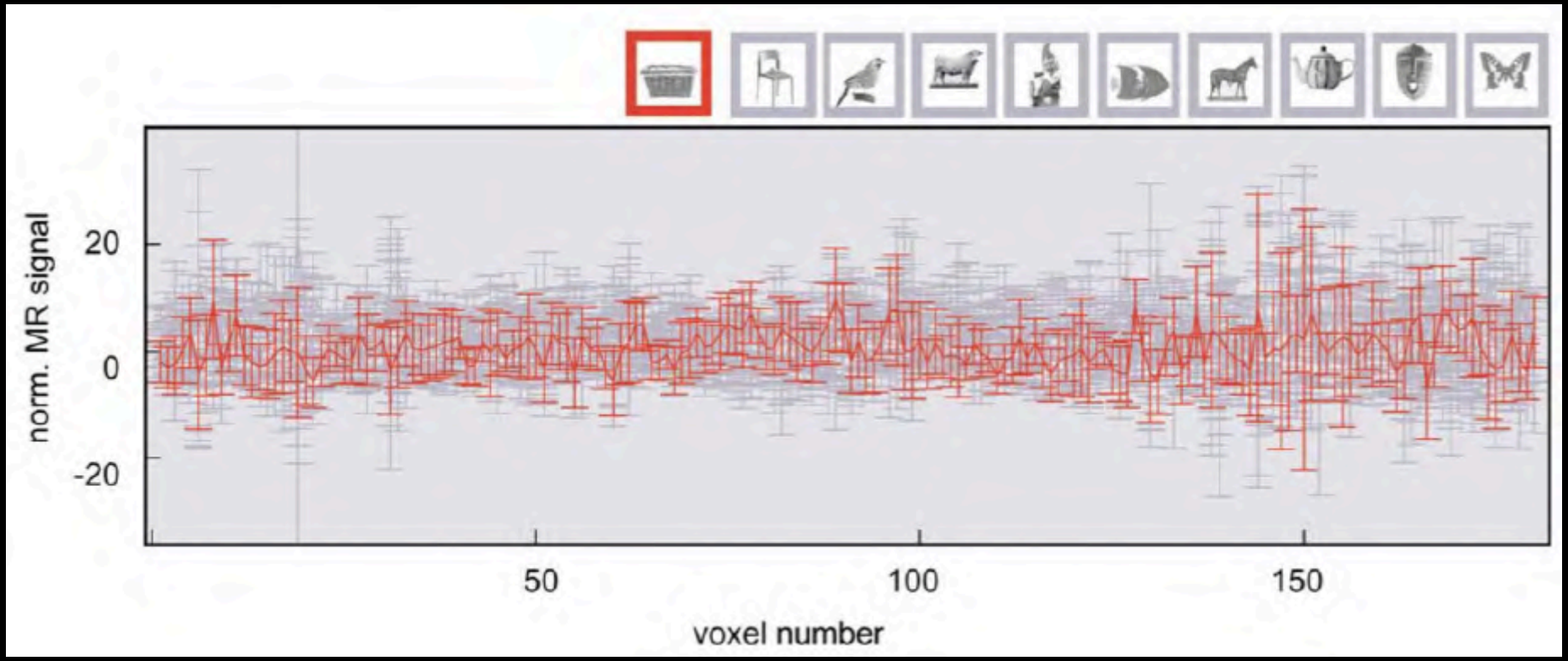
David D. Cox[a,b,*] and Robert L. Savoy[a,b,c]

[a] *Rowland Institute for Science, Cambridge, MA 02142, USA*
[b] *Athinoula A. Martinos Center for Structural and Functional Biomedical Imaging, Charlestown, MA 02129, USA*
[c] *HyperVision, Inc., P.O. Box 158, Lexington, MA 02420, USA*

Cox & Savoy (2003), *NeuroImage*

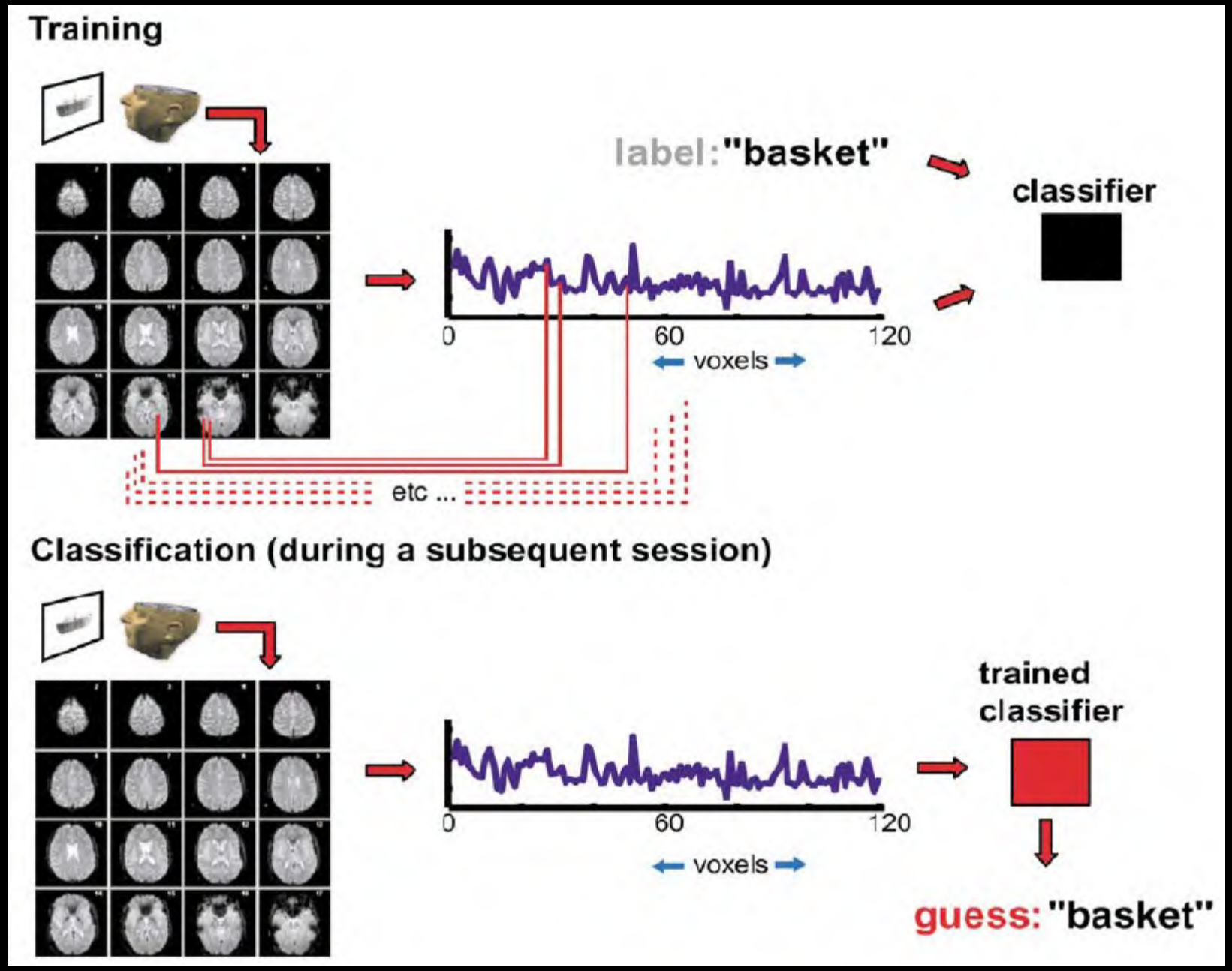


Cox & Savoy (2003)



NOTE: no individual voxels show strong basket-specific response

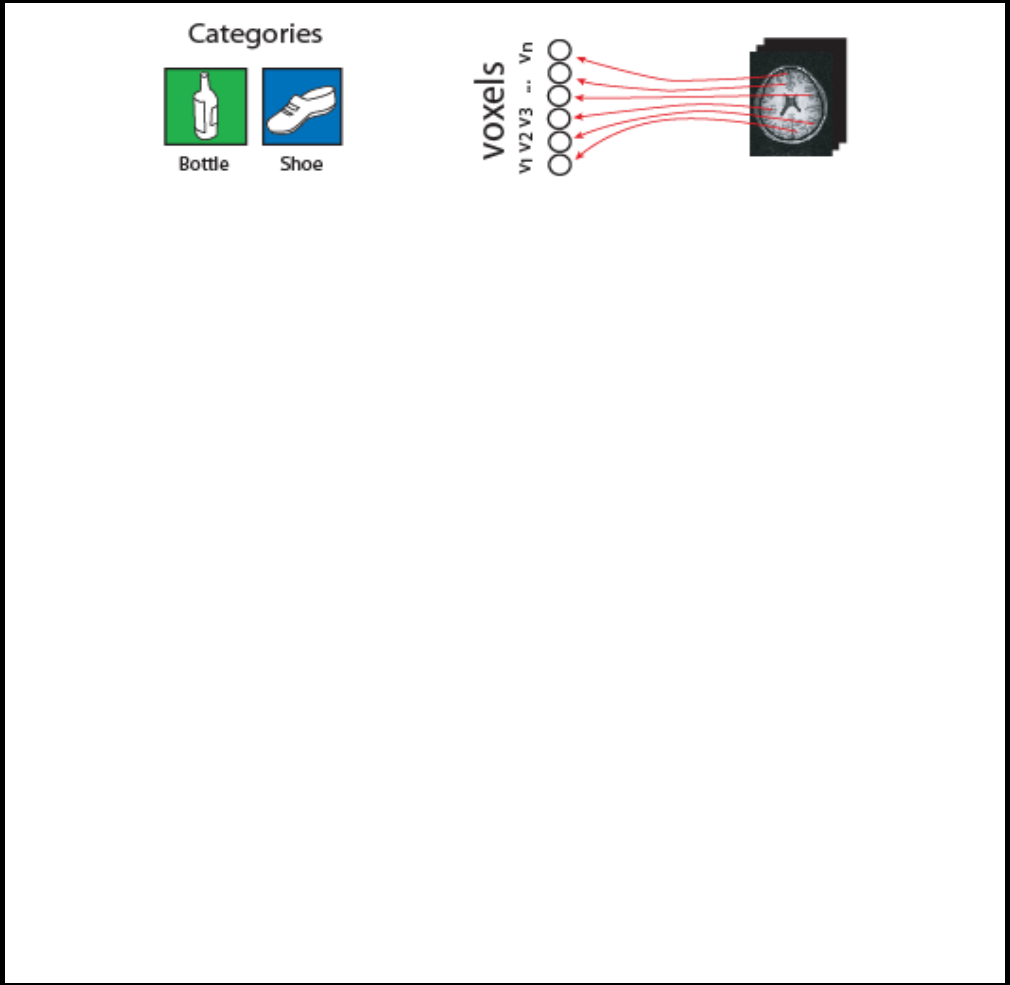
Cox & Savoy (2003)



Cox & Savoy (2003)

The Multi-Voxel Pattern Analysis Approach

1. Acquire brain data while the subject is viewing shoes or bottles



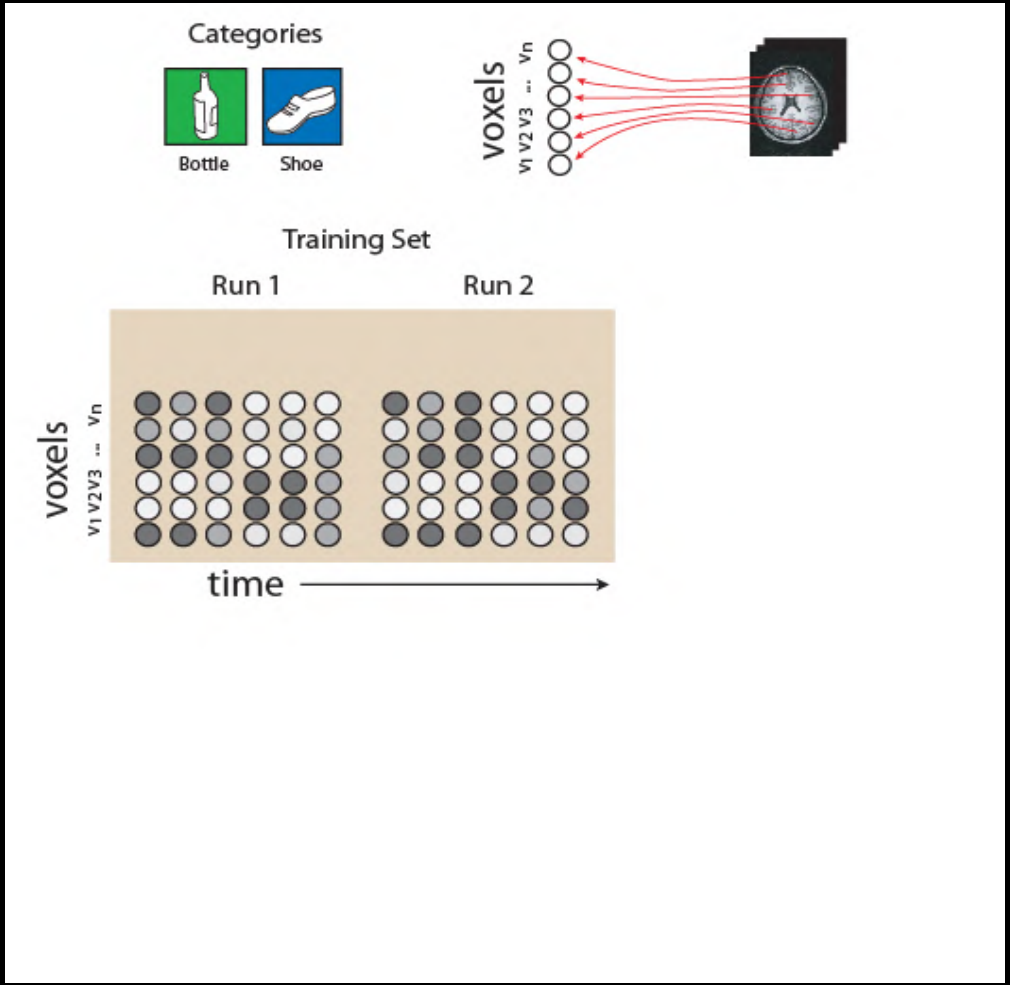
Norman et al. (2006), *TICS*

The Multi-Voxel Pattern Analysis Approach

1. Acquire brain data

2. Convert each functional brain volume (or trial) into a vector that reflects the *pattern of activity across voxels* at that point in time

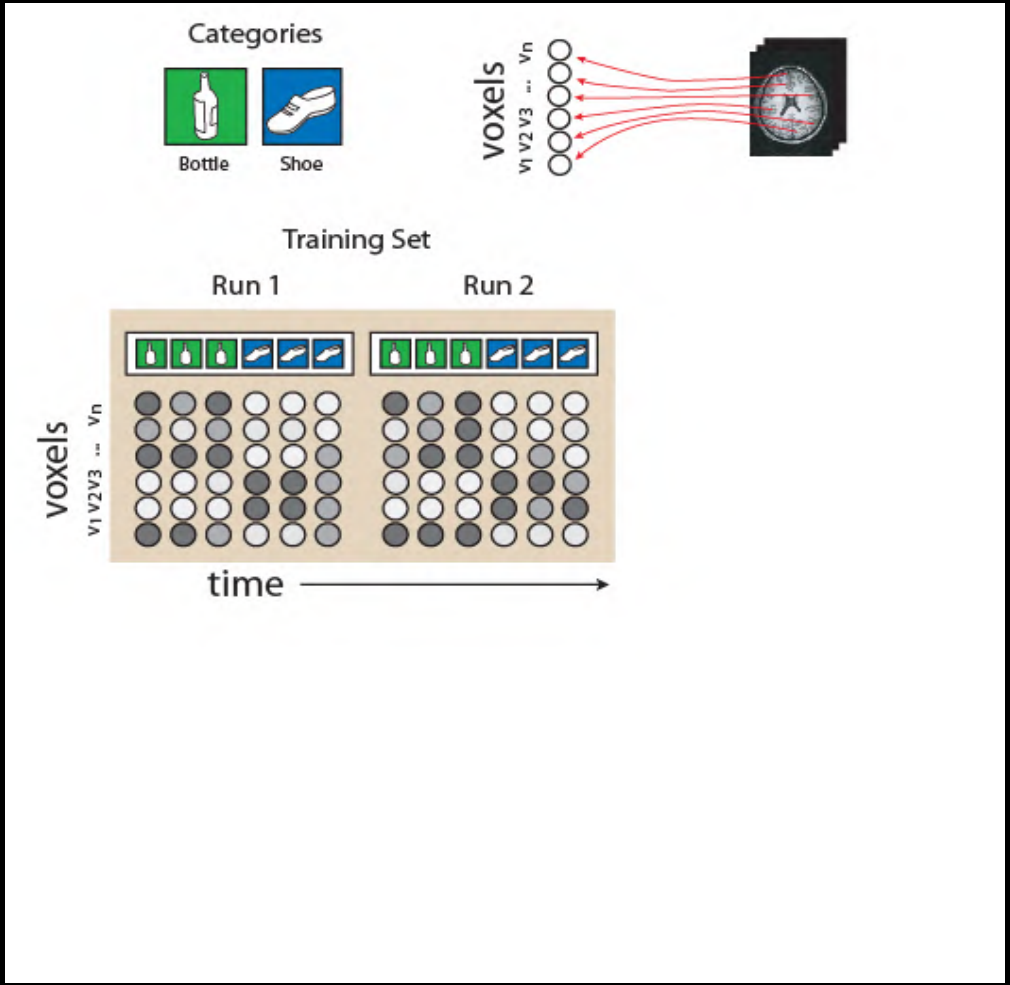
   We typically do some kind of *feature selection* to cut down on the number of voxels



Norman et al. (2006), *TICS*

The Multi-Voxel Pattern Analysis Approach

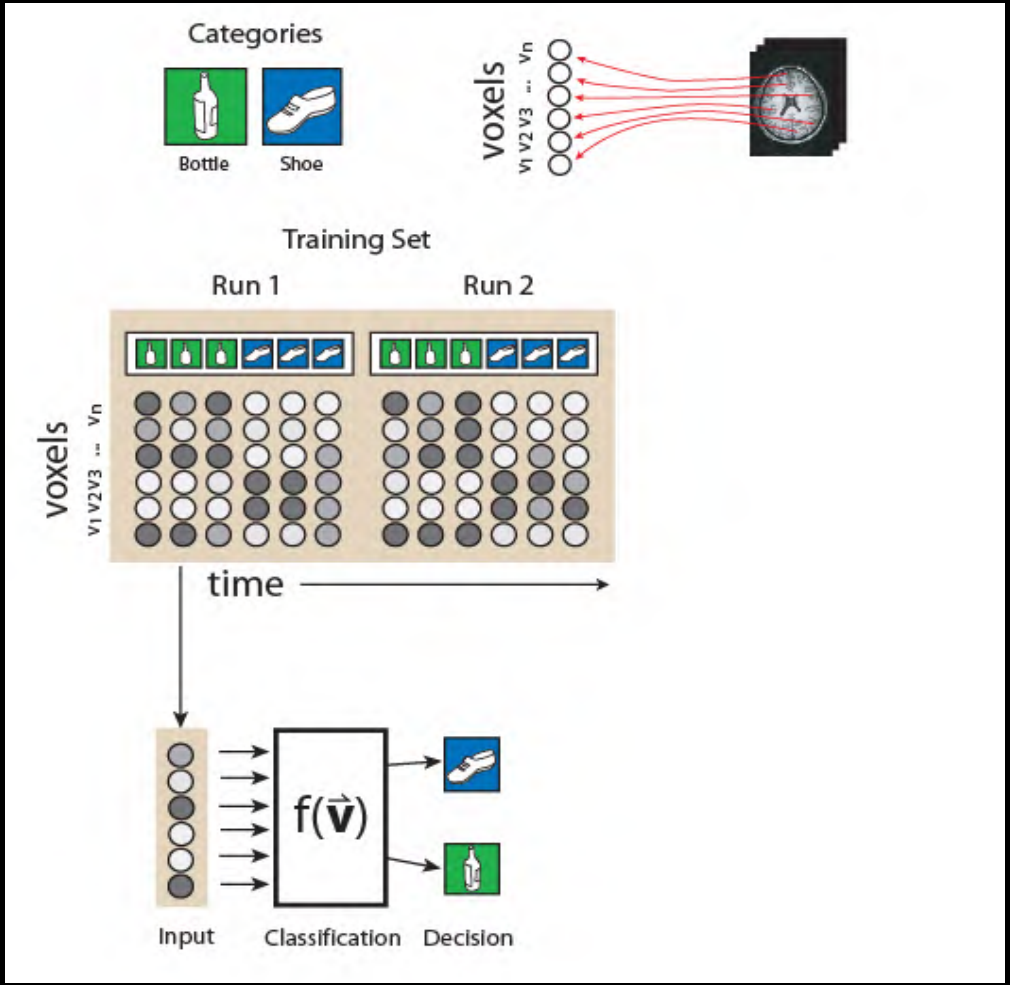
1. Acquire brain data
2. Generate brain patterns
3. Label brain patterns according to whether the subject was viewing shoes vs. bottles (adjusting for lag in the blood flow response)



Norman et al. (2006), *TICS*

The Multi-Voxel Pattern Analysis Approach

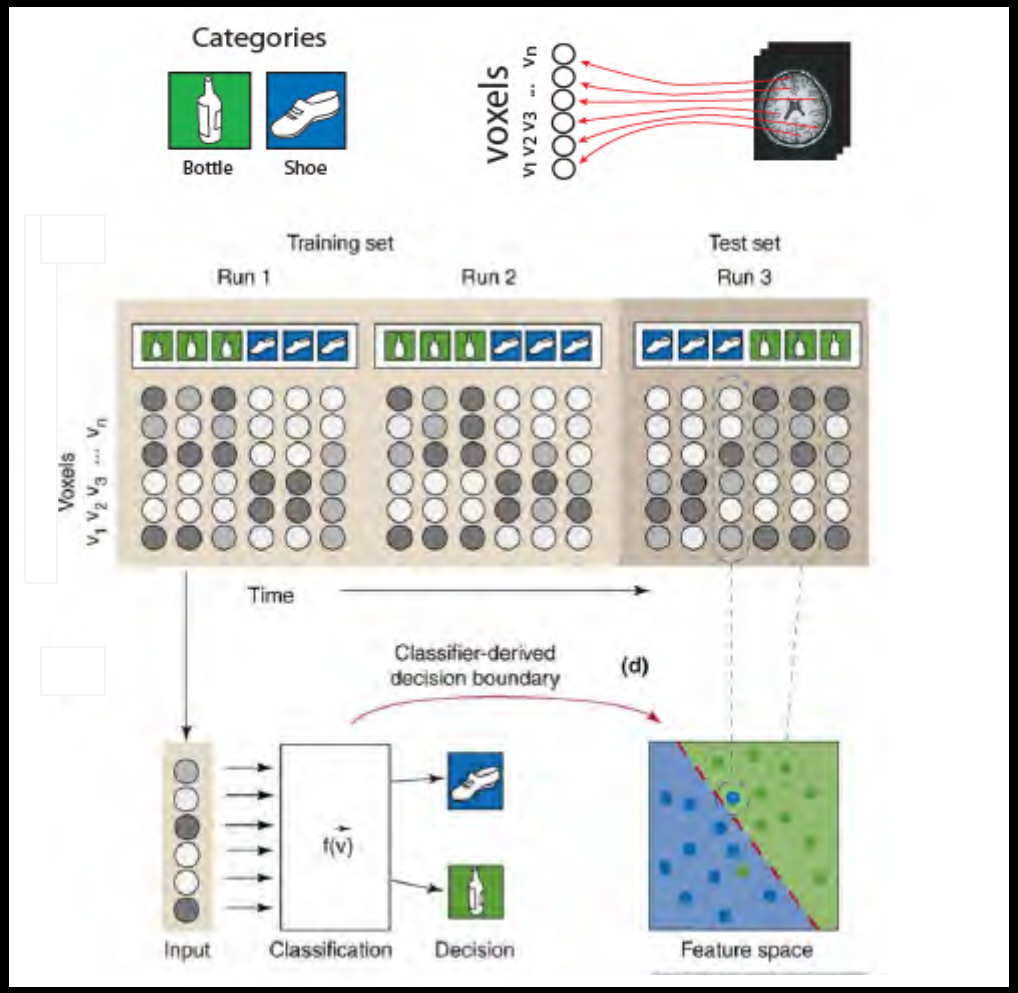
1. Acquire brain data
2. Generate brain patterns
3. Label brain patterns
4. Train a classifier to discriminate between bottle patterns and shoe patterns



Norman et al. (2006), *TICS*

The Multi-Voxel Pattern Analysis Approach

1. Acquire brain data
2. Generate brain patterns
3. Label brain patterns
4. Train a classifier
5. Apply the trained classifier to new brain patterns (i.e., not included in training set)



Norman et al. (2006), *TICS*

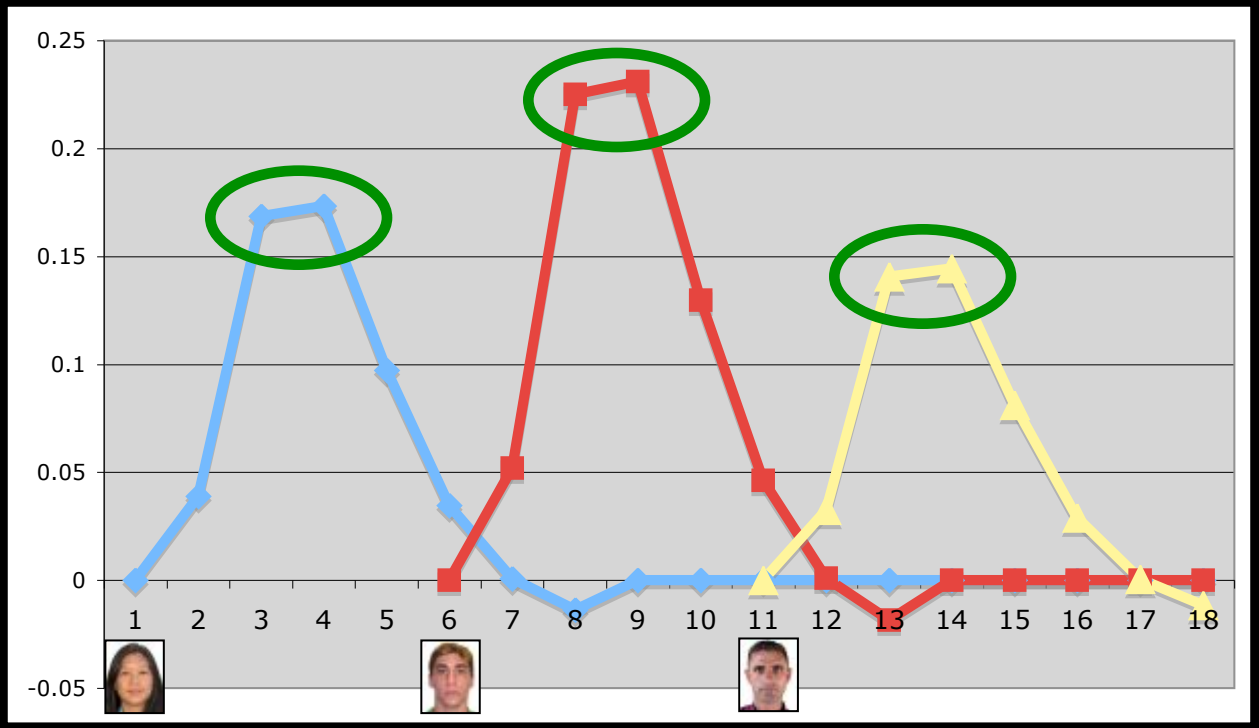
MVPA Data Processing Protocol

- Preparing the data (preprocessing)
  - Remove signal artifacts
  - Detrend each run
  - High-pass filter each run
  - Z-score data from each run
- Parsing the data into "examples"
  - Block designs
    - Average timepoints from each block (↑ signal stability; ↓ examples)
    - Or treat each timepoint as an independent example
  - Event-related designs
    - Choose single post-stimulus timepoint from each trial or average several
    - Need long inter-trial intervals to prevent hemodynamic overlap

*** Always important to balance the number of examples from each condition ***

Pattern Analysis Approach:
Temporal Selection

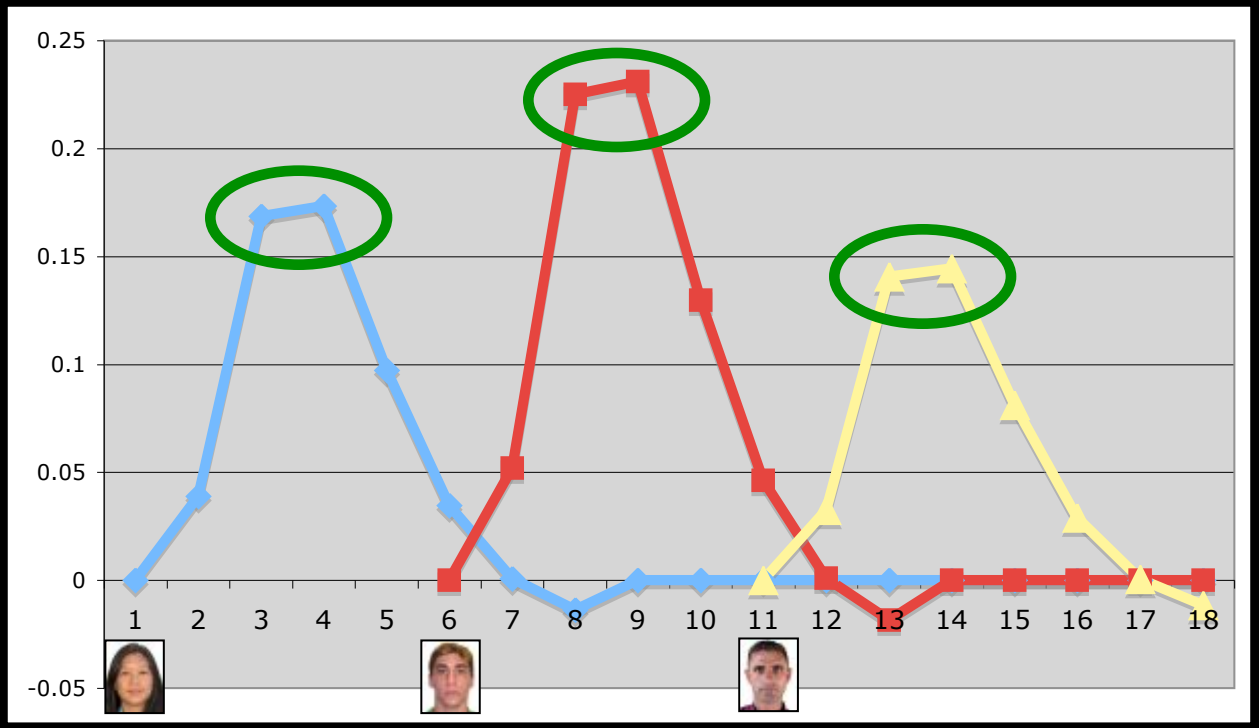
– Reduce full fMRI timeseries from having 5 values (TRs) per trial to having only 1 value per trial



– Average 3rd and 4th TR of each trial (e.g., 4-8 sec post-stimulus)

Pattern Analysis Approach:
Temporal Selection

– Reduce full fMRI timeseries from having 5 values (TRs) per trial to having only 1 value per trial

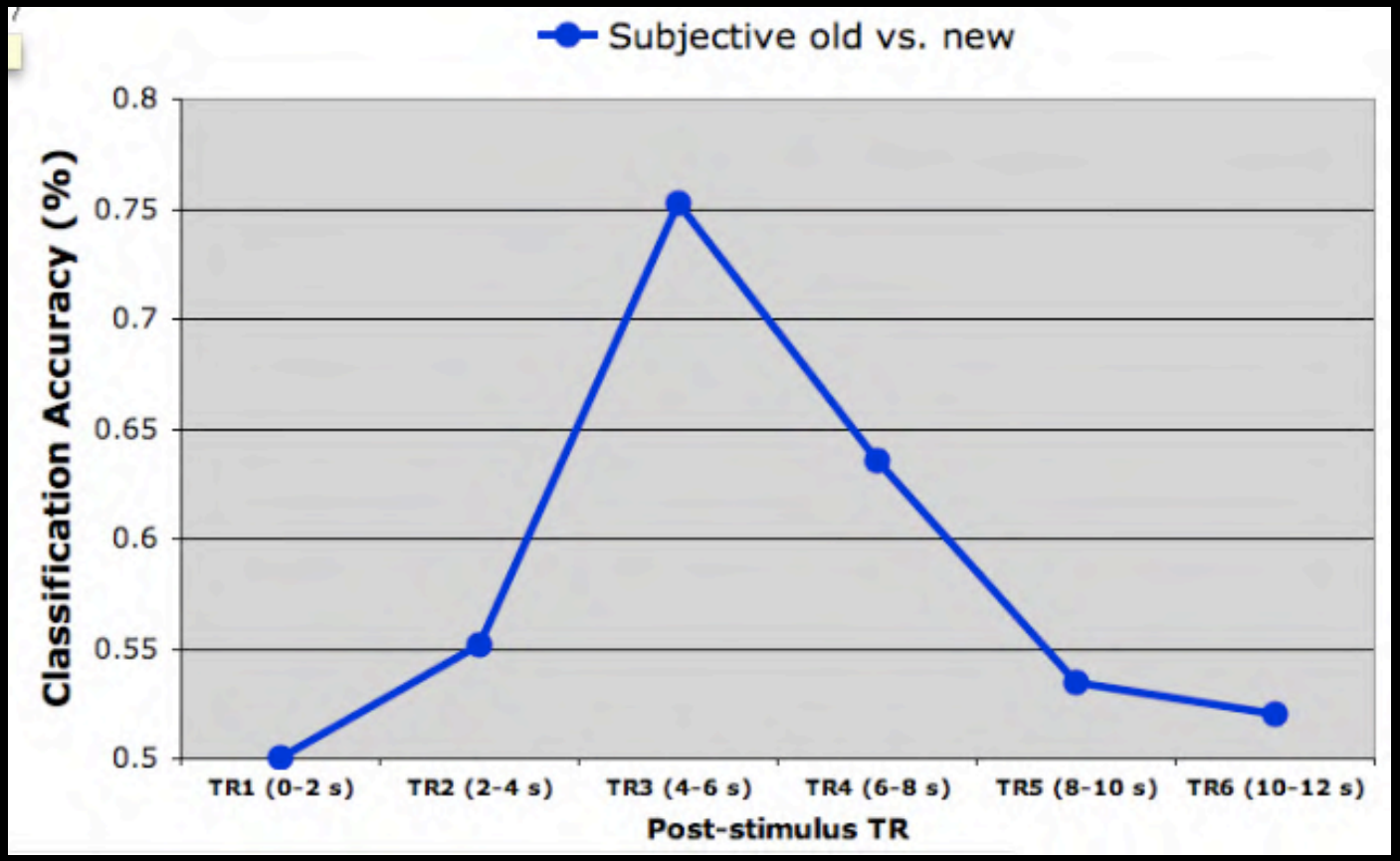


– Or run a new GLM that estimates a single parameter for each trial (i.e, beta-series approach; Rissman et al. (2004)

Pattern Analysis Approach:
Temporal Selection

Alternative strategy

– Train and test separate classifiers using data from each post-stimulus TR



Classification timecourse for one subject

Pattern Analysis Approach:
Temporal Selection

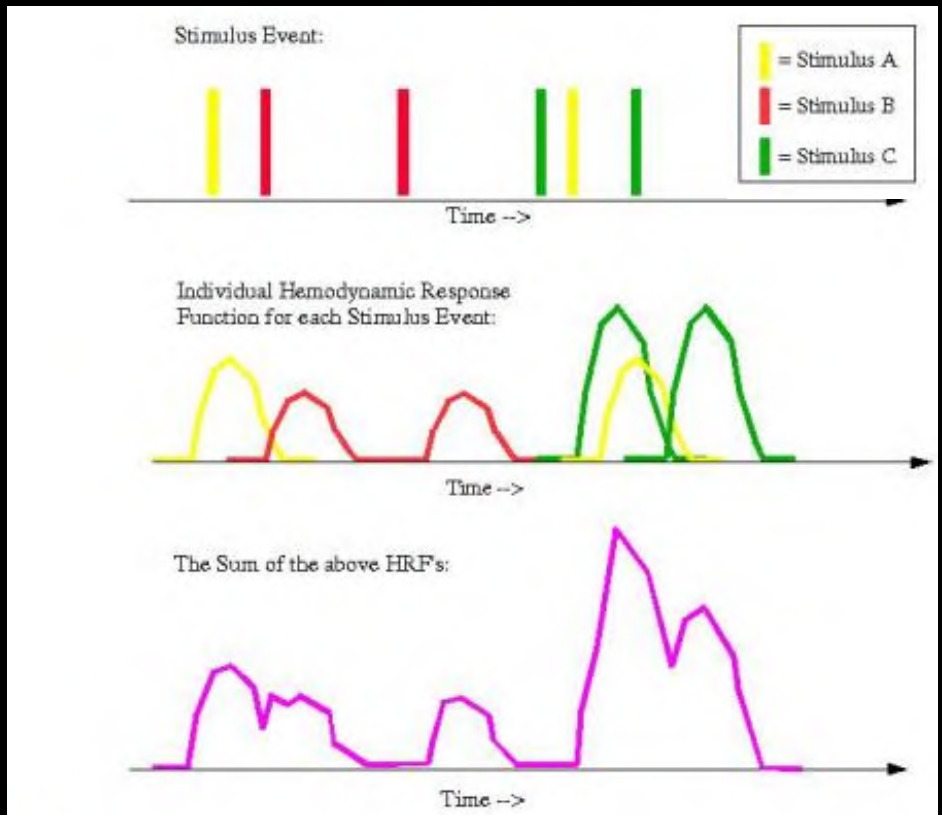
- NOTE: the previous examples assume that you are working with a slow event-related design (i.e., widely spaced trials with minimally-overlapping HRFs)

- What about rapid event-related designs?

Pattern Analysis Approach:
Temporal Selection

<u>One approach:</u>

- For each TR, examine your design matrix and determine which condition has the maximal predicted activity
- Specify a threshold to exclude to ambiguous TRs



Pattern Analysis Approach:
Temporal Selection

<u>Another approach (Mumford et al., 2012):</u>

- Estimate each trial's activity through a univariate GLM including one regressor for that trial as well as another regressor for all other trials.
  - Like Rissman et al. (2004) beta series estimation approach, but involves running **many separate GLMs** (# of GLMs = # of trials)

NeuroImage 59 (2012) 2636–2643

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses

Jeanette A. Mumford [a,*], Benjamin O. Turner [b], F. Gregory Ashby [b], Russell A. Poldrack [c]

[a] Department of Psychology, University of Texas, Austin, TX 78759, USA
[b] Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA 93016, USA
[c] Departments of Psychology and Neurobiology and Imaging Research Center, University of Texas, Austin, TX 78759, USA

ARTICLE INFO

Article history:
Received 27 May 2011
Revised 6 August 2011
Accepted 23 August 2011
Available online 5 September 2011

Keywords:
Functional magnetic resonance imaging
Classification analysis
MVPA
Beta series estimation
Rapid event-related design

ABSTRACT

Use of multivoxel pattern analysis (MVPA) to predict the cognitive state of a subject during task performance has become a popular focus of fMRI studies. The input to these analyses consists of activation patterns corresponding to different tasks or stimulus types. These activation patterns are fairly straightforward to calculate for blocked trials or slow event-related designs, but for rapid event-related designs the evoked BOLD signal for adjacent trials will overlap in time, complicating the identification of signal unique to specific trials. Rapid event-related designs are often preferred because they allow for more stimuli to be presented and subjects tend to be more focused on the task, and thus it would be beneficial to be able to use these types of designs in MVPA analyses. The present work compares 8 different models for estimating trial-by-trial activation patterns for a range of rapid event-related designs varying by interstimulus interval and signal-to-noise ratio. The most effective approach obtains each trial's estimate through a general linear model including a regressor for that trial as well as another regressor for all other trials. Through the analysis of both simulated and real data we have found that this model shows some improvement over the standard approaches for obtaining activation patterns. The resulting trial-by-trial estimates are more representative of the true activation magnitudes, leading to a boost in classification accuracy in fast event-related designs with higher signal-to-noise. This provides the potential for fMRI studies that allow simultaneous optimization of both univariate and MVPA approaches.

Pattern Analysis Approach:
Temporal Selection

<u>One more commonly used approach:</u>

- Run a standard univariate GLM analysis to derive condition-specific beta estimates for each scanning run
- Use these beta images as your patterns for classification
  - **Problem:**
    - If you only have 6 runs, then at best you'd only have 5 training examples for each condition
  - **Work-around:**
    - Subdivide each actual run into 2 or 3 mini-runs, and then run univariate GLM
    - More beta images to work with!

NeuroImage 61 (2012) 1113–1119

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs

Marc N. Coutanche *, Sharon L. Thompson-Schill

Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, USA

ARTICLE INFO

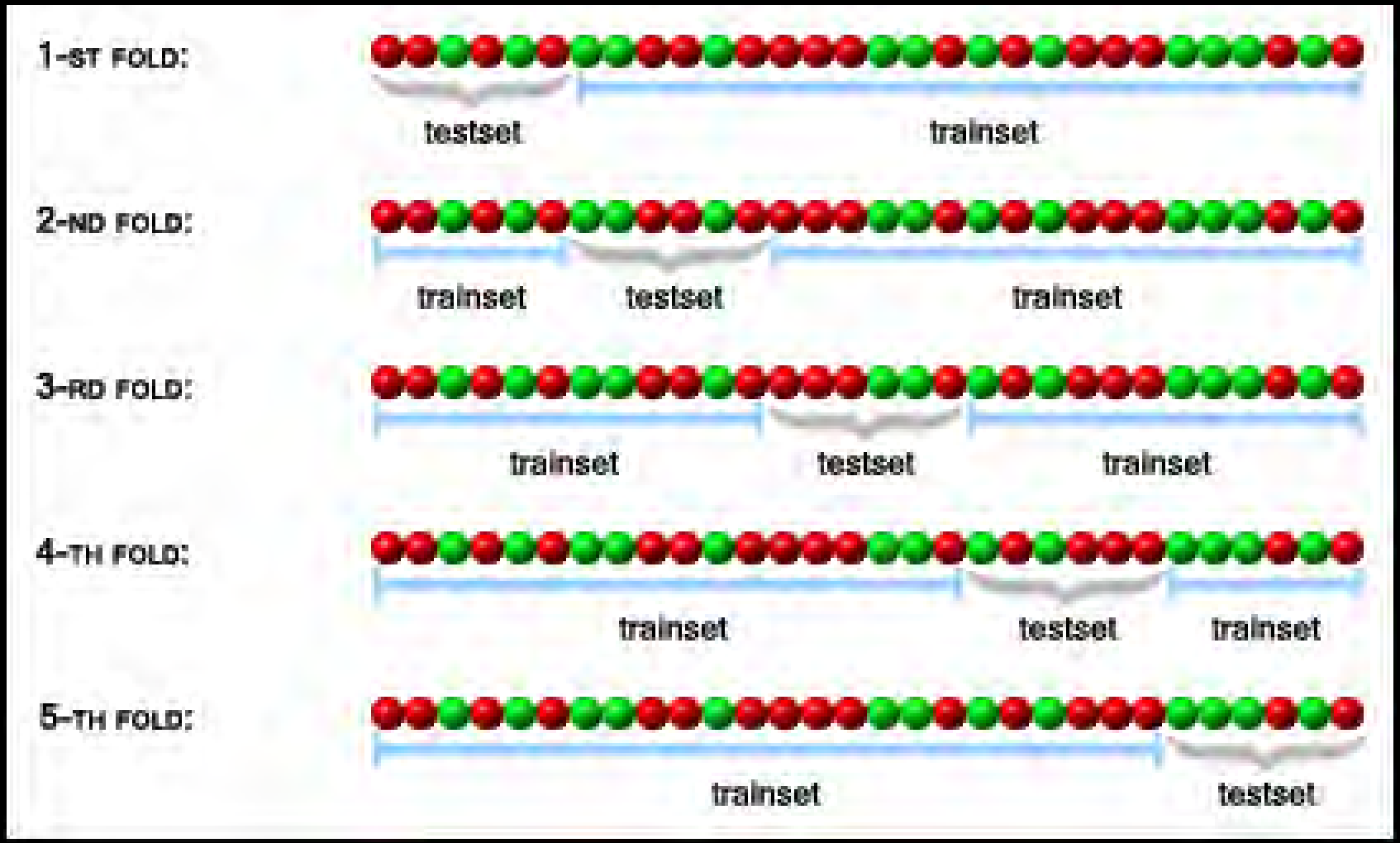
Article history:
Accepted 25 March 2012
Available online 3 April 2012

Keywords:
MVPA
fMRI
Multivariate
Runs
Design
Classification

ABSTRACT

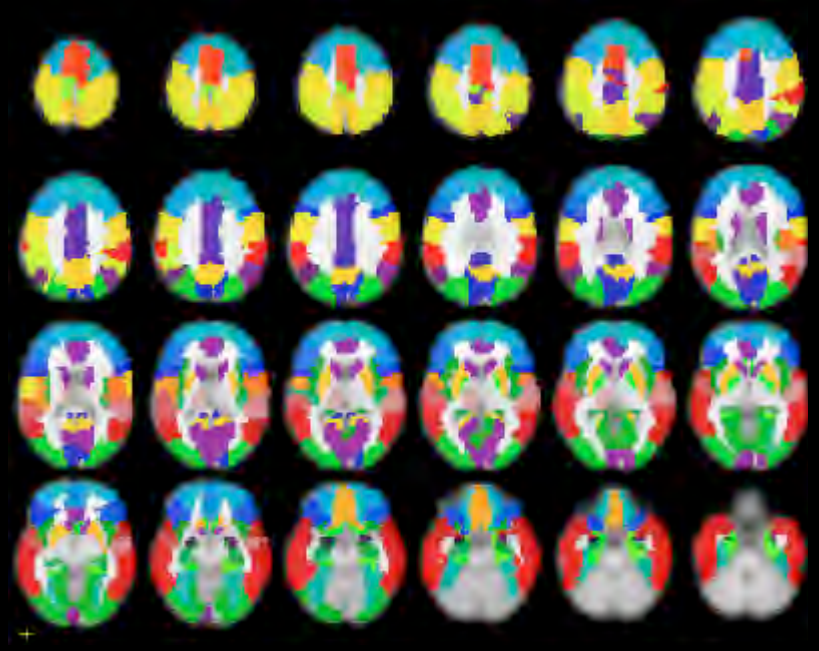
Functional magnetic resonance imaging (fMRI) studies are broken up into runs (or 'sessions'), frequently selected to be long to minimize across-run signal variations. For investigations that use multi-voxel pattern analysis (MVPA), however, employing many short runs might improve a classifier's ability to generalize across irrelevant pattern variations and detect condition-related activity patterns. We directly tested this hypothesis by scanning participants with both long and short runs and comparing MVPA performance using data from each set of runs. Every run included presentations of faces, places, man-made objects and fruit in a blocked 1-back design. MVPA performance significantly improved from using a large number of short runs, compared to several long runs, in across-run classifications with identical amounts of data. Superior classification was found across variations in the classifier employed, feature selection procedure and region of interest. Performance improvements also extended to an information brain mapping 'searchlight' procedure. These results suggest that investigators looking to maximize the detection of subtle multi-voxel patterns across runs might consider employing short fMRI runs.

Dividing up the data:
Cross-validation



Feature Selection

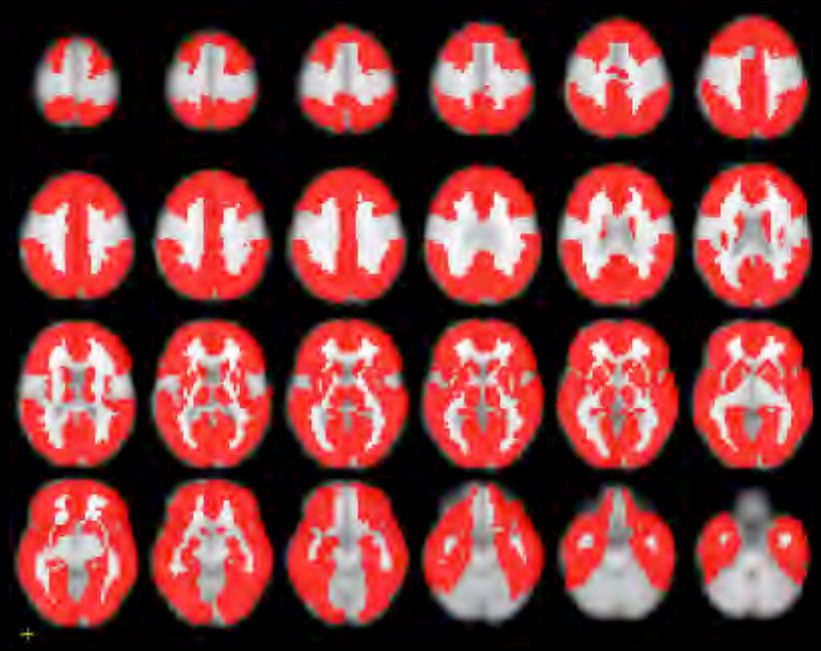
- Selecting which voxels to include in the analysis
  - Univariate GLM (a.k.a. conventional brain mapping)
    - Identify general task-responsive voxels (e.g., all conditions vs. baseline)
    - Identify task-selective voxels (e.g., Condition A vs. Condition B)
      - must be done without using "held-out" testing data
  - Independently-defined ROIs



ROIs from the AAL atlas

Feature Selection

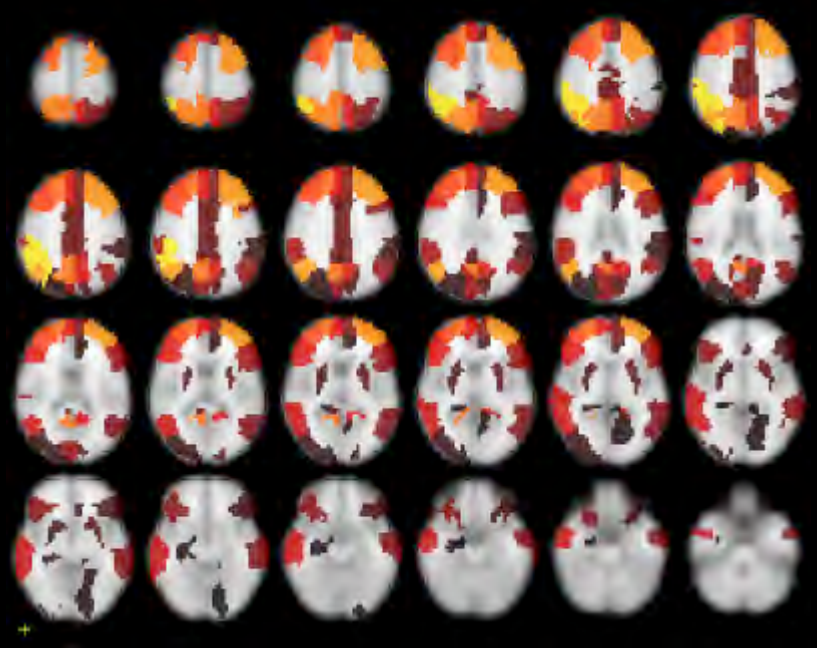
- Selecting which voxels to include in the analysis
  - Univariate GLM (a.k.a. conventional brain mapping)
    - Identify general task-responsive voxels (e.g., all conditions vs. baseline)
    - Identify task-selective voxels (e.g., Condition A vs. Condition B)
      - must be done without using "held-out" testing data
  - Independently-defined ROIs



You could combine all of these regions to make a large mask (in this case, excluding motor areas, white matter, and CSF)

Feature Selection

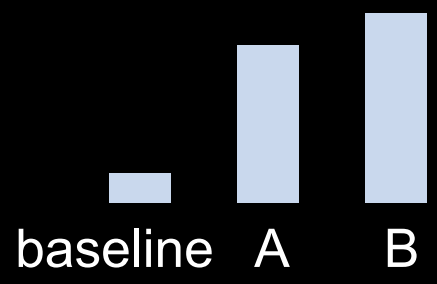
- Selecting which voxels to include in the analysis
  - Univariate GLM (a.k.a. conventional brain mapping)
    - Identify general task-responsive voxels (e.g., all conditions vs. baseline)
    - Identify task-selective voxels (e.g., Condition A vs. Condition B)
      - must be done without using "held-out" testing data
  - Independently-defined ROIs



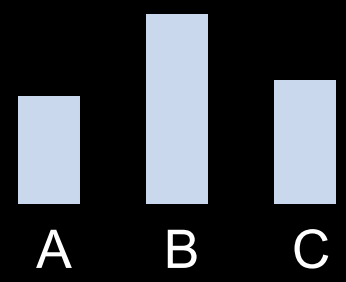
You also can compute classification performance within each ROI

Feature Selection

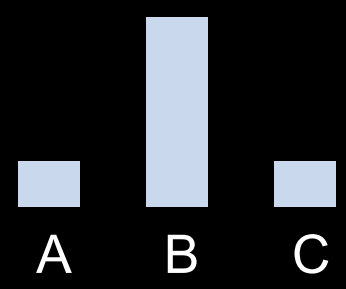
- What criteria should define important voxels?
  - difference from baseline



  - difference between classes (e.g. ANOVA)

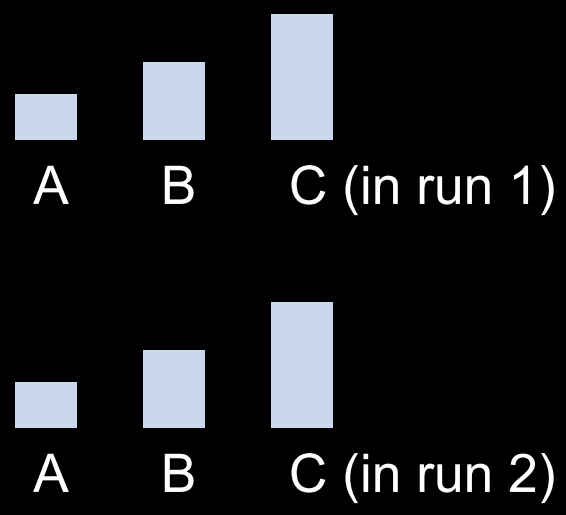


  - preferential response to one class



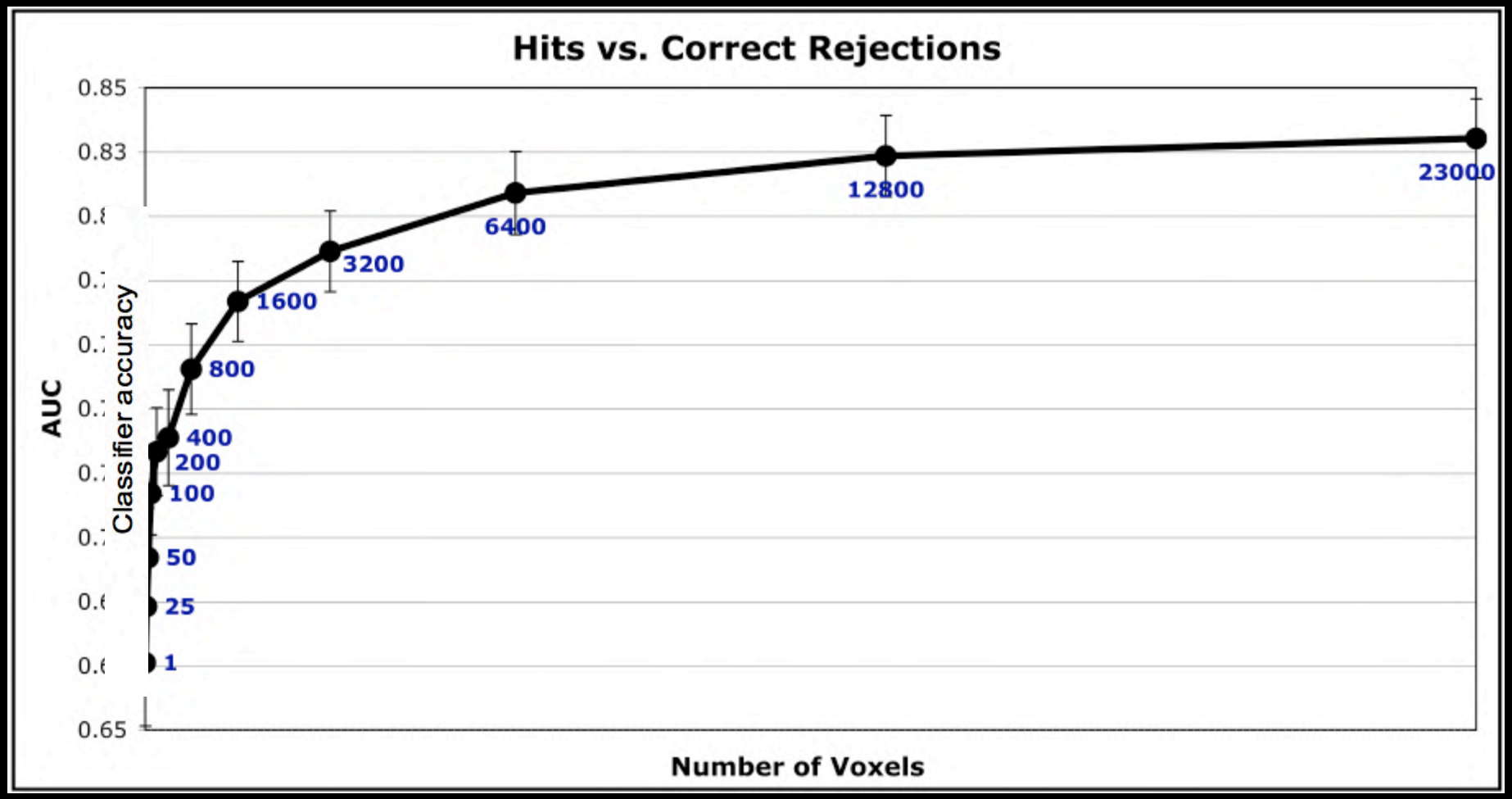
Feature Selection

- What criteria should define important voxels?
  - stability (i.e., across scanning runs)



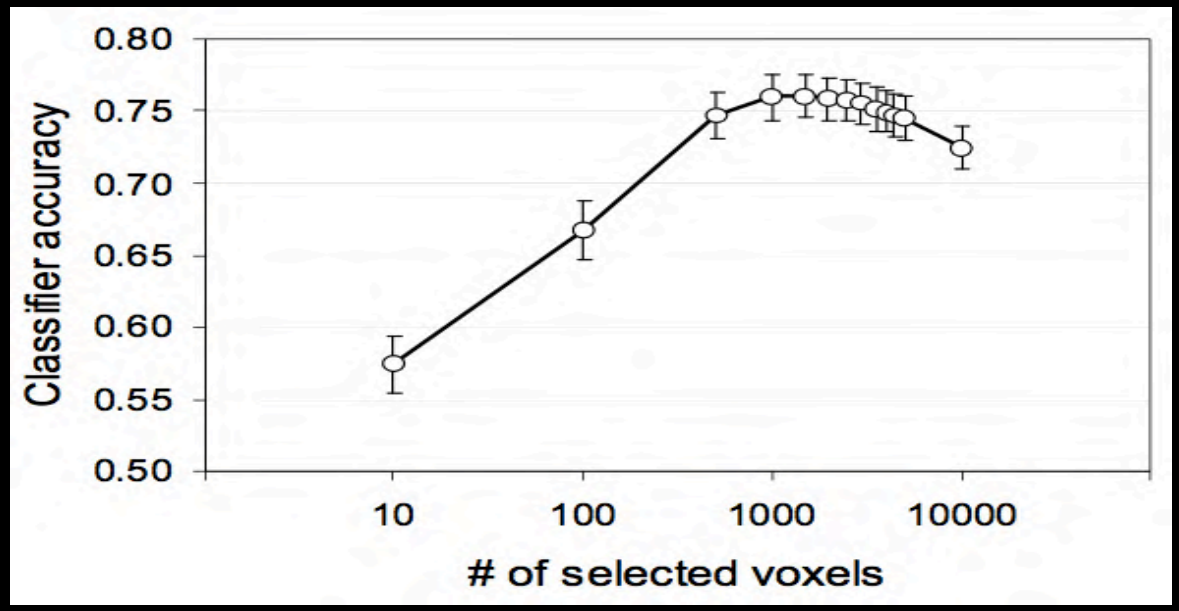
How many features (voxels) to use?

- Classification performance as a function of the number of voxels used by classifier (*ANOVA-based selection*)



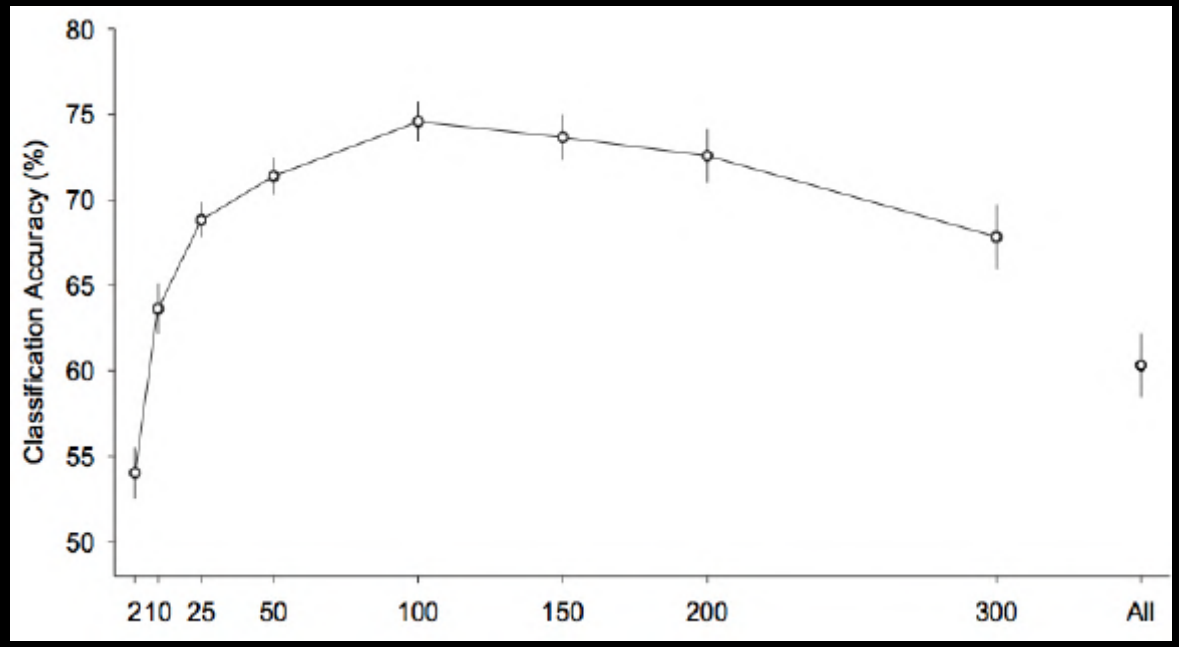
Peak performance with ~1000 voxels

Johnson et al. (2009)

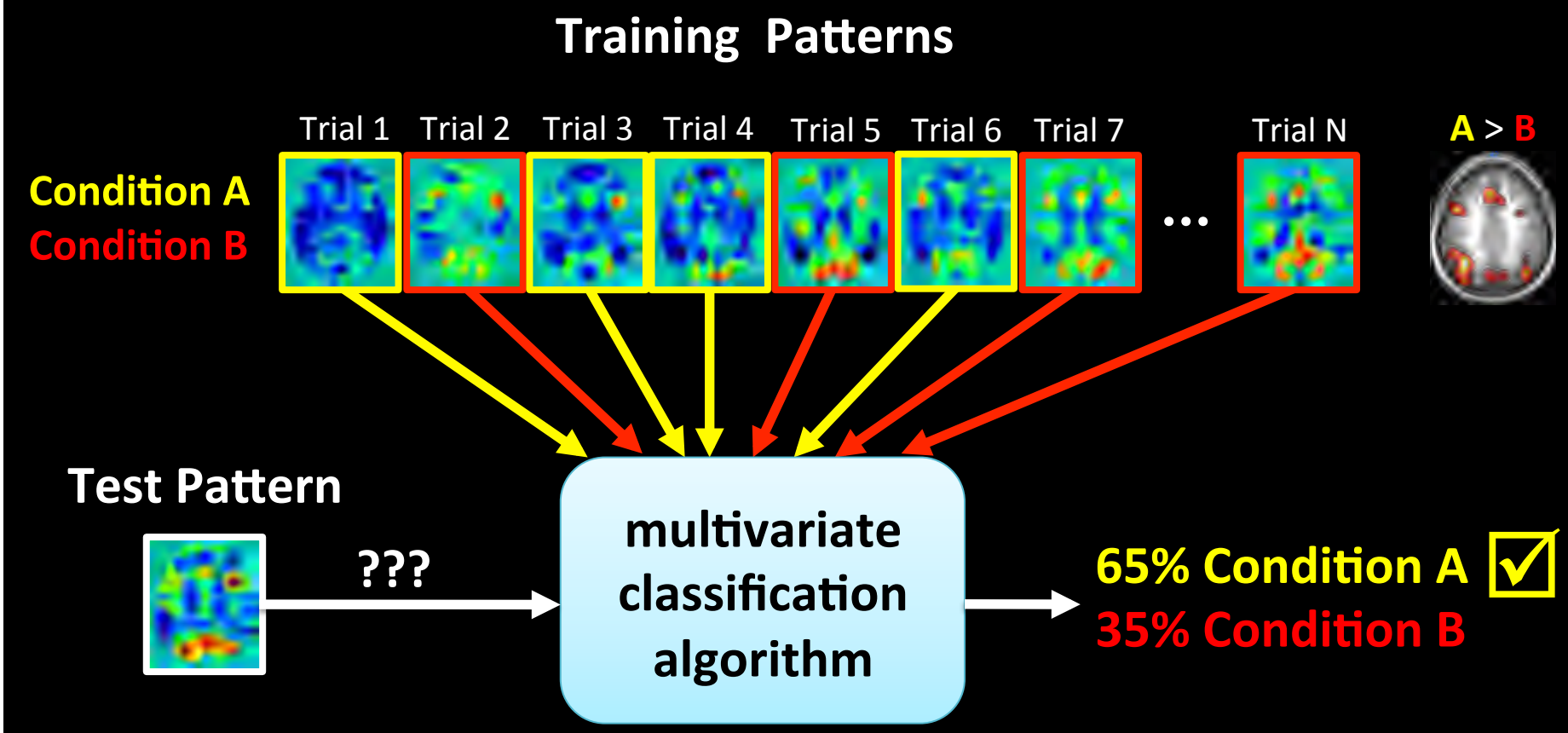


Peak performance with ~100 voxels

Esterman et al. (2009)

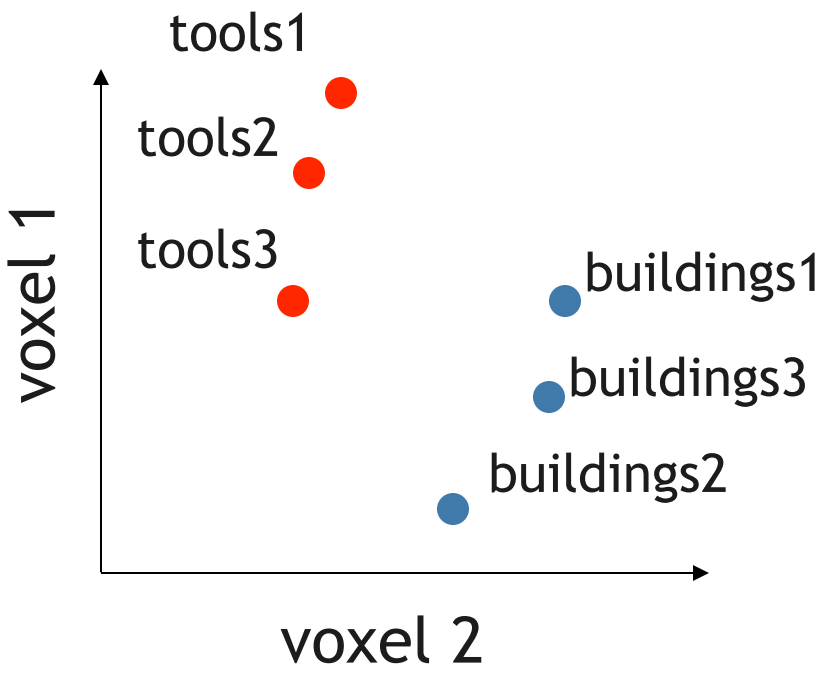


Training and testing the classifier



what is inside the box?

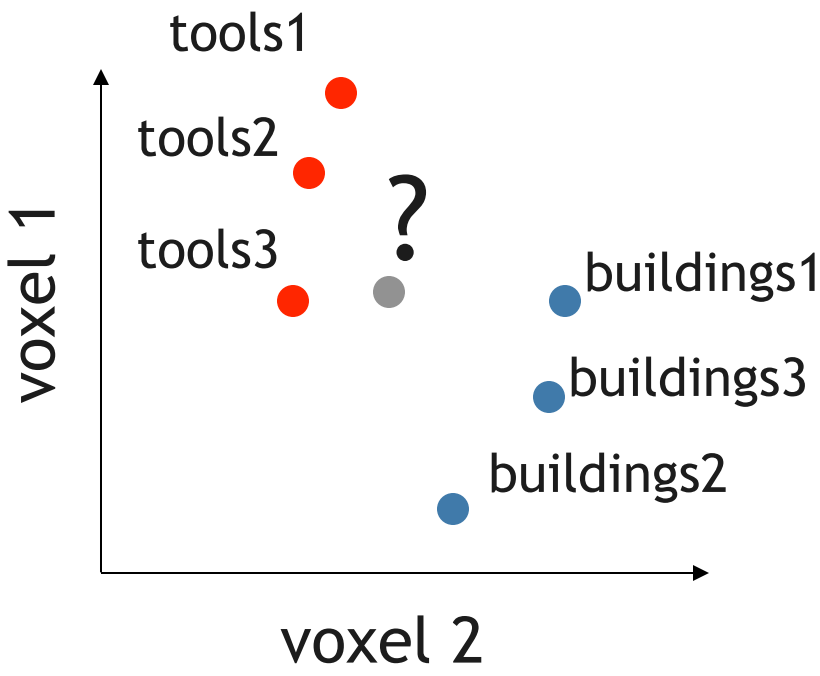
- simplest function is no function at all
- “nearest neighbour”



Slide from Francisco Pereira

what is inside the box?

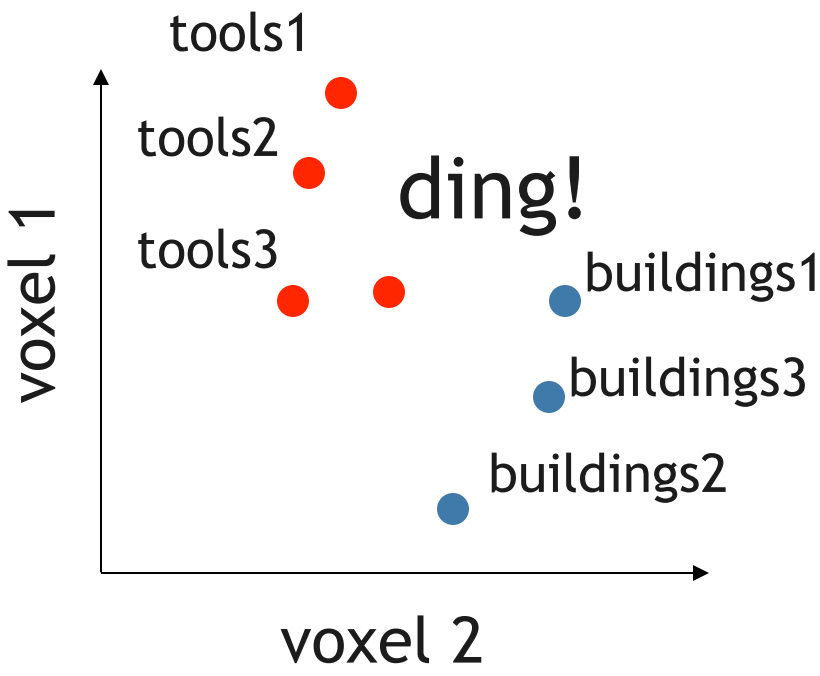
- simplest function is no function at all
- “nearest neighbour”



Slide from Francisco Pereira

what is inside the box?

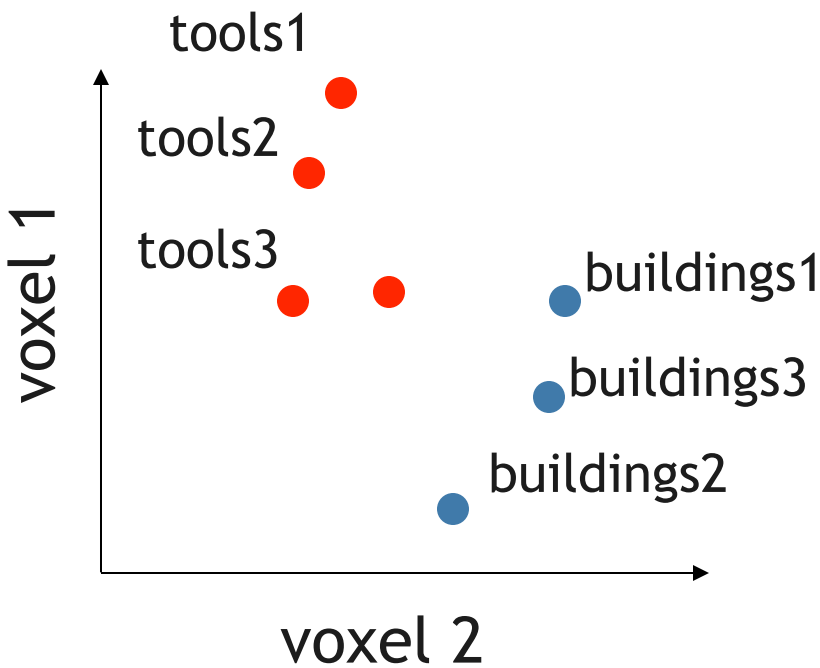
- simplest function is no function at all
- “nearest neighbour”



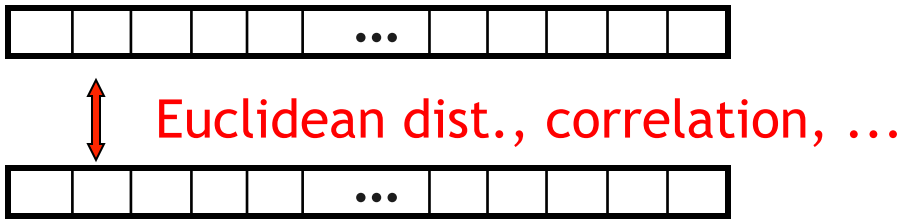
Slide from Francisco Pereira

what is inside the box?

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



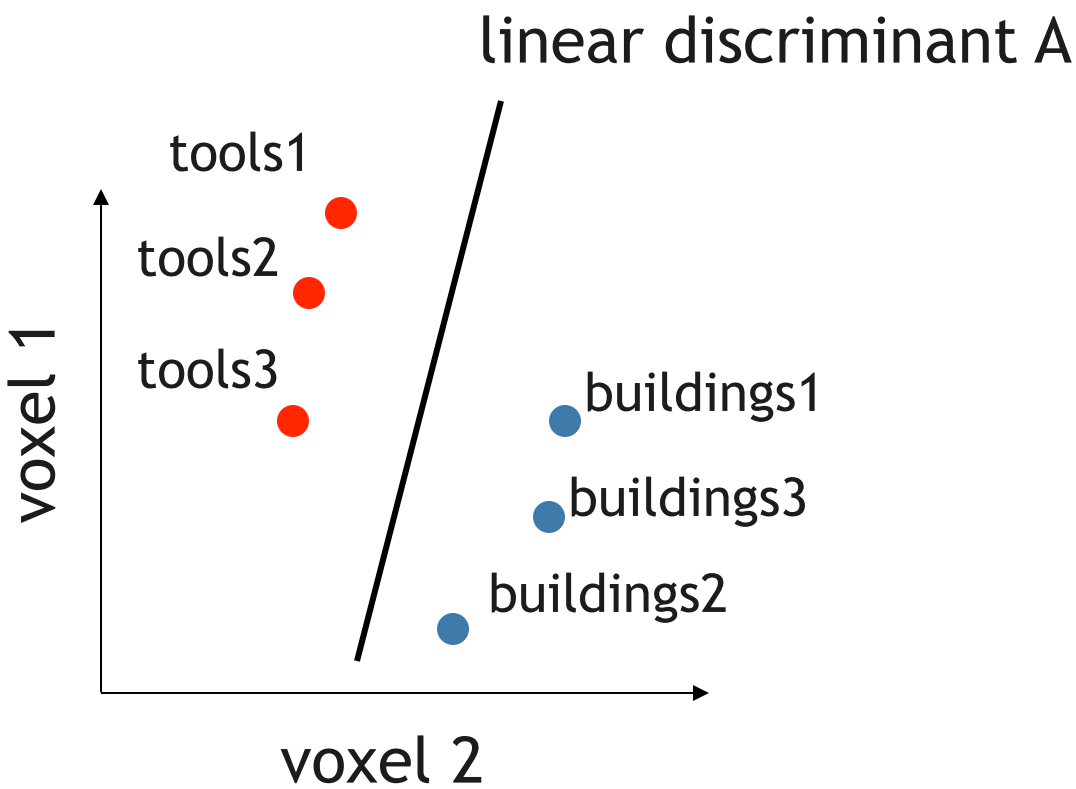
requires example similarity measure



Slide from Francisco Pereira

what is inside the box?

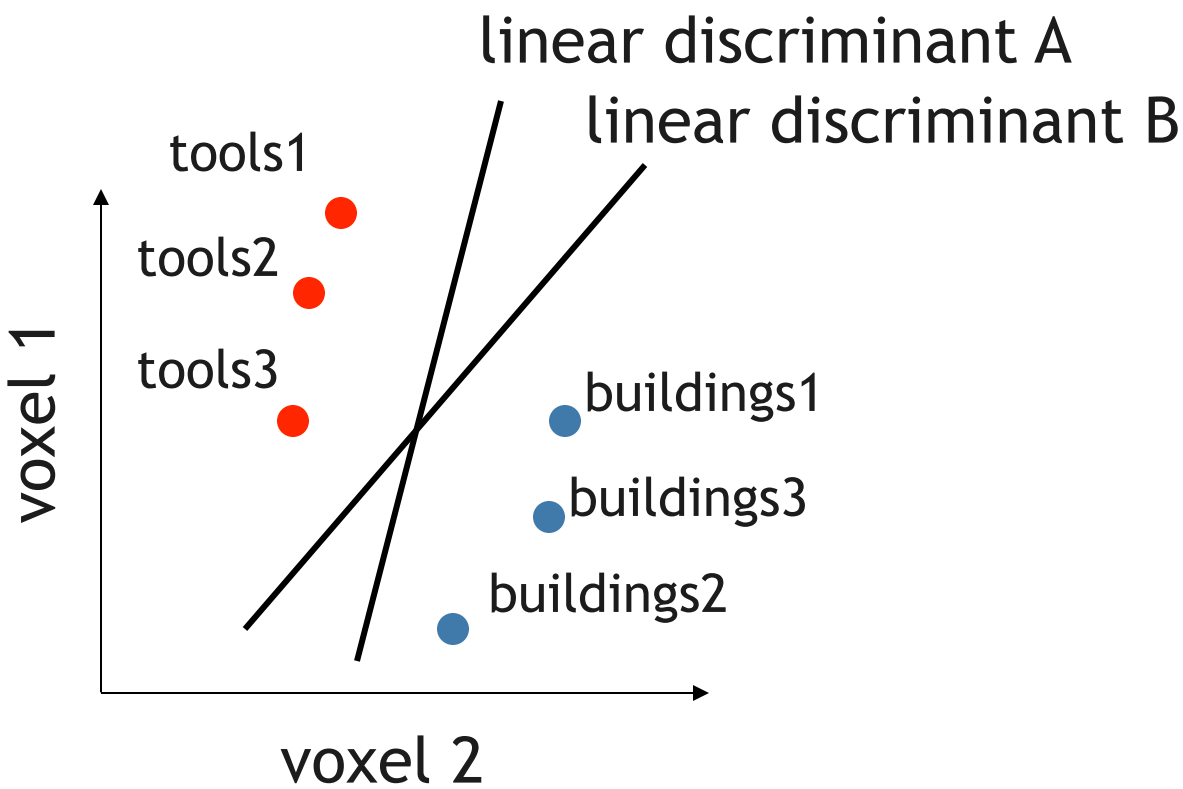
- next simplest: learn linear discriminant



Slide from Francisco Pereira

what is inside the box?

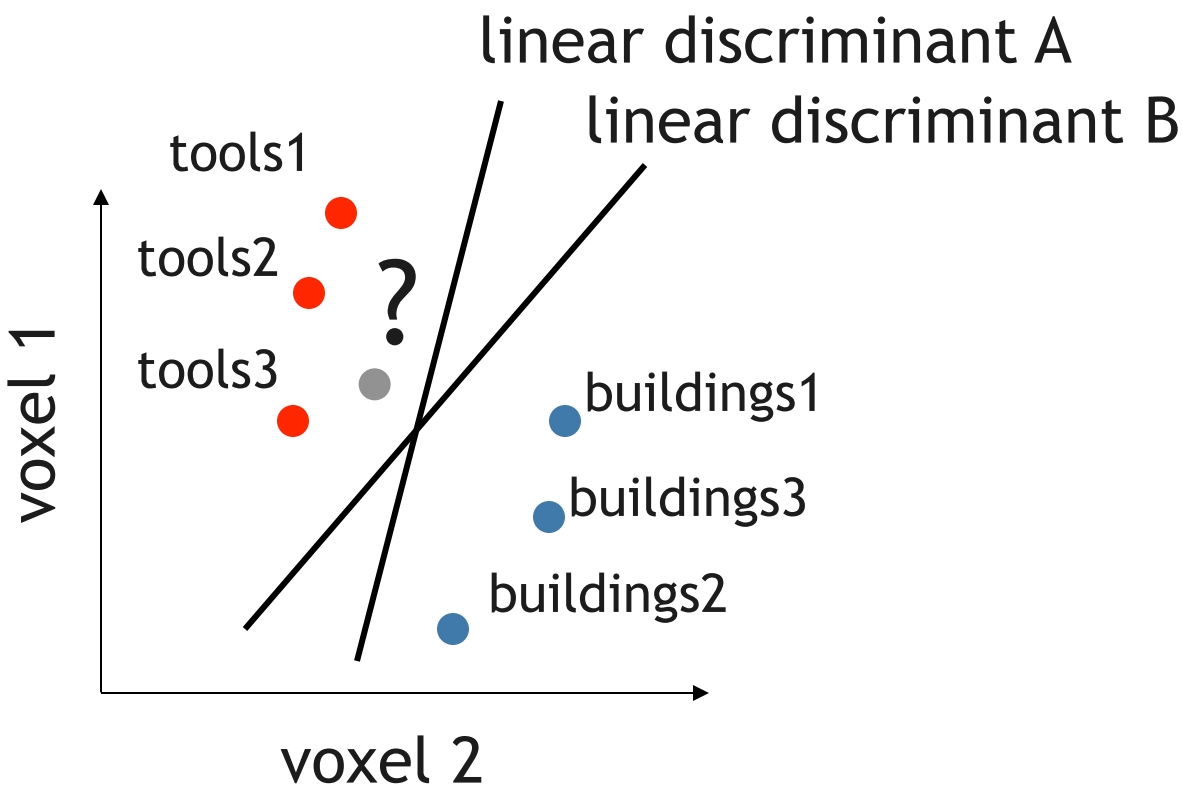
- next simplest: learn linear discriminant
- note that there are many solutions...



Slide from Francisco Pereira

what is inside the box?

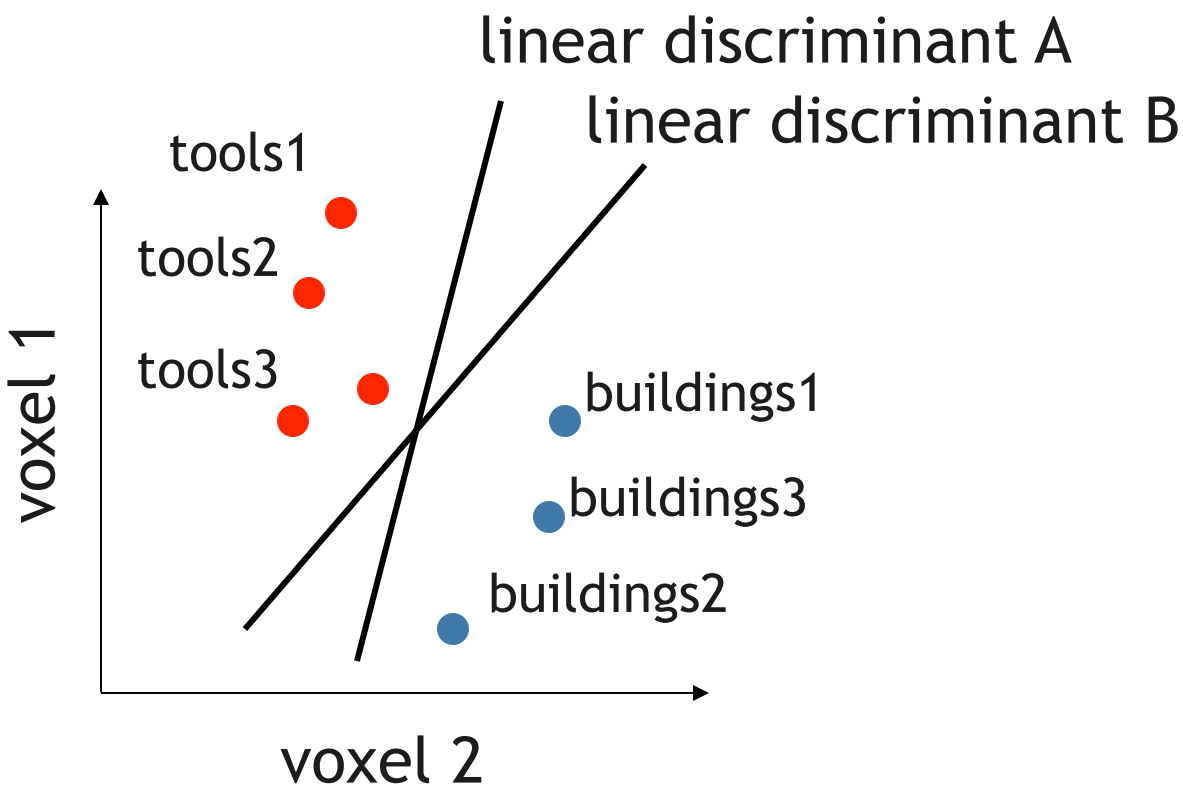
- next simplest: learn linear discriminant
- note that there are many solutions...



Slide from Francisco Pereira

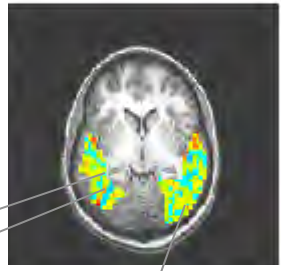
what is inside the box?

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



Slide from Francisco Pereira

linear classifiers



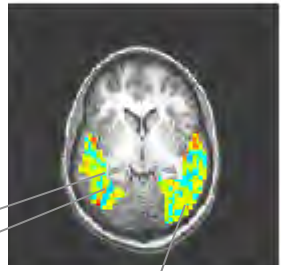
If weight0 + weight1 x + weight2 x + + + weight n x > 0 tools

voxel 1	voxel 2	...			voxel n

otherwise buildings

Slide from Francisco Pereira

linear classifiers



If weight0 + weight1 x + weight2 x + + + weight n x > 0 tools

voxel 1	voxel 2	...			voxel n

otherwise buildings

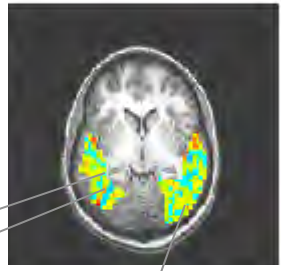
various kinds:

- Gaussian Naive Bayes
- Regularized Logistic Regression
- Linear Support Vector Machines (SVM)

differ on how weights are chosen

Slide from Francisco Pereira

linear classifiers

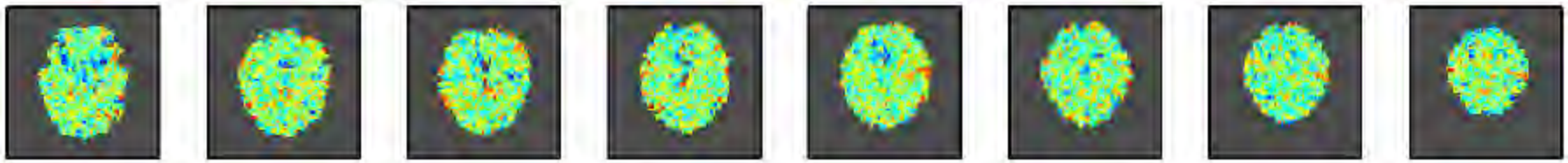


If weight0 + weight1 x + weight2 x + + + weight n x > 0 tools

voxel 1	voxel 2	...			voxel n

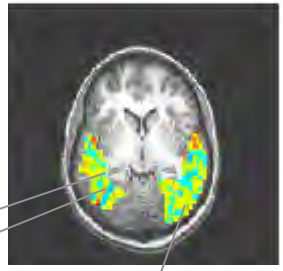
otherwise buildings

linear SVM weights:



Slide from Francisco Pereira

linear classifiers



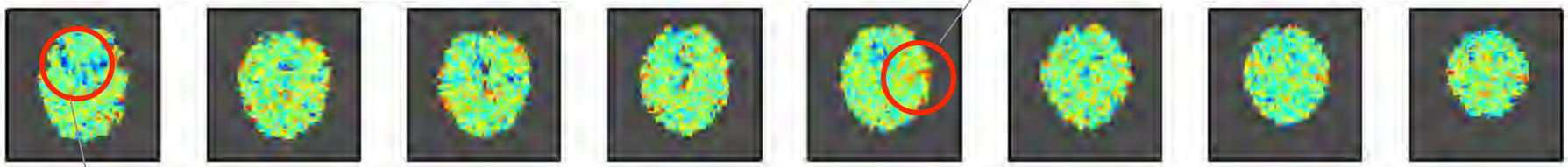
If weight0 + weight1 x + weight2 x + + + weight n x > 0 tools

voxel 1	voxel 2	...			voxel n

otherwise buildings

weights pull towards tools

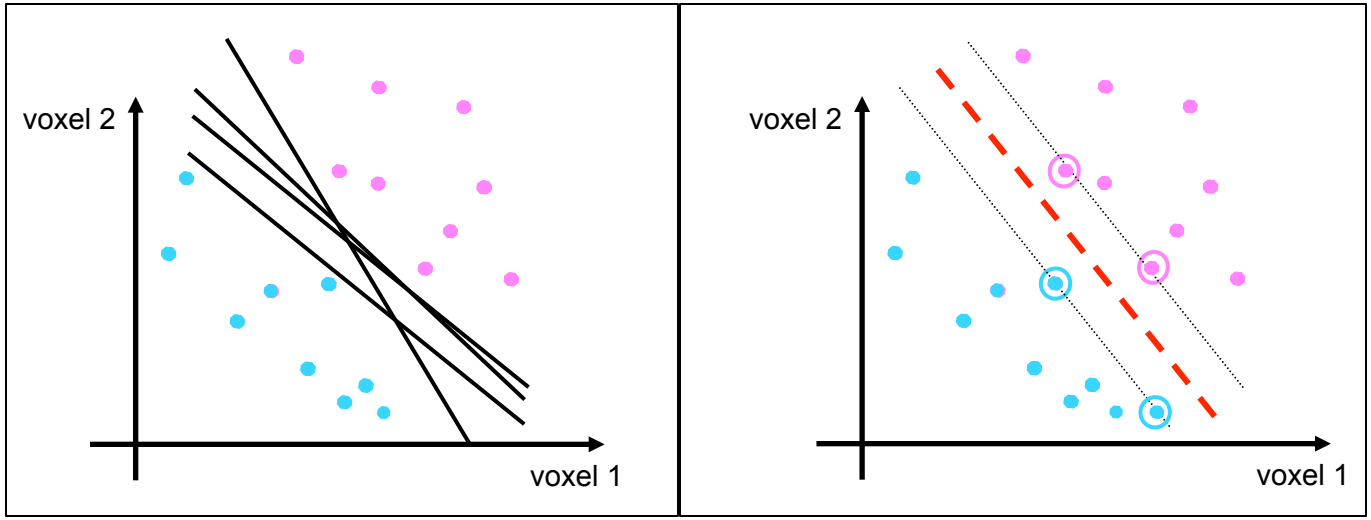
linear SVM weights:



weights pull towards buildings

Slide from Francisco Pereira

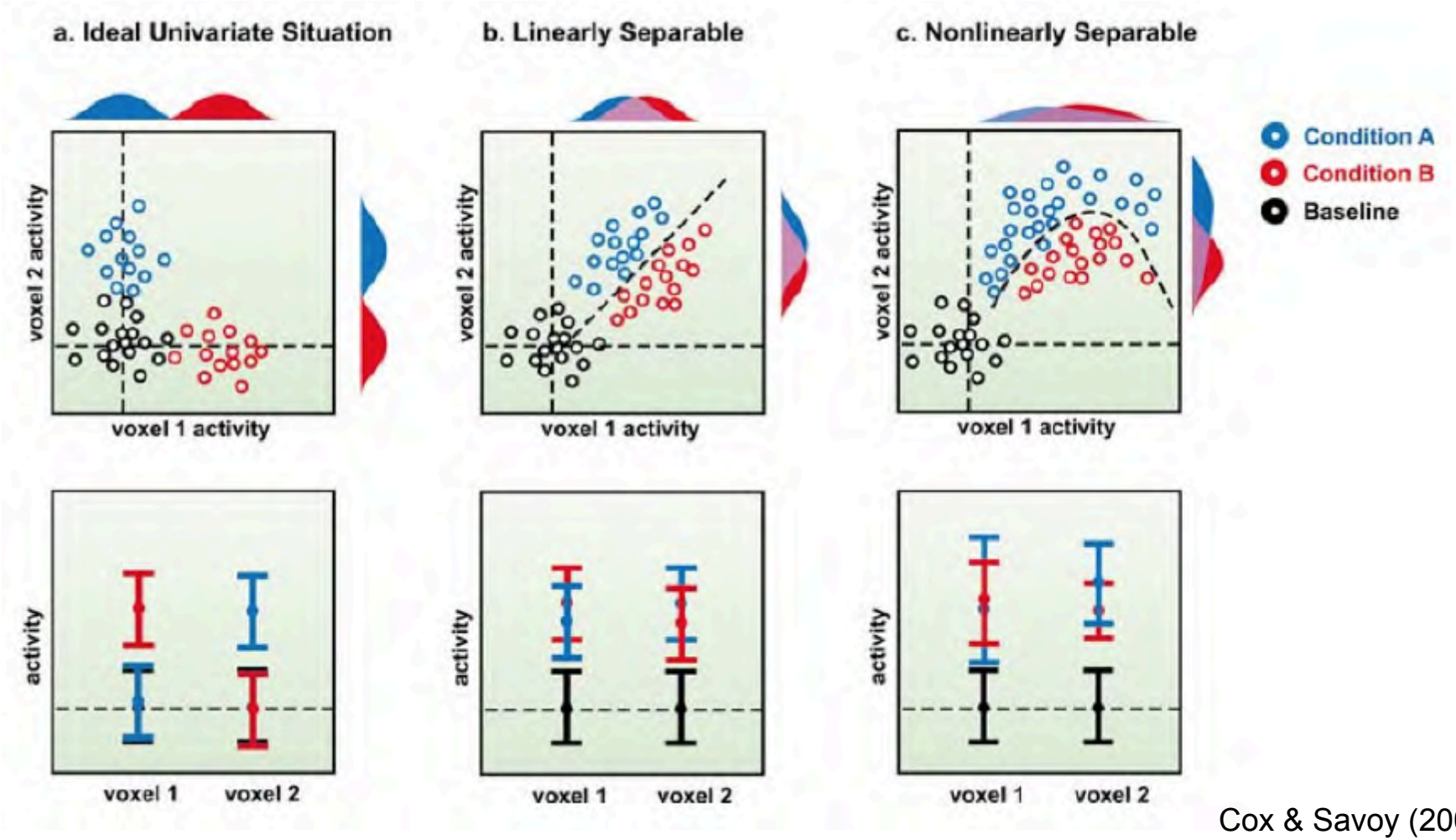
linear support vector machines



- Find linear decision boundary that maximizes the margin

nonlinear classifiers

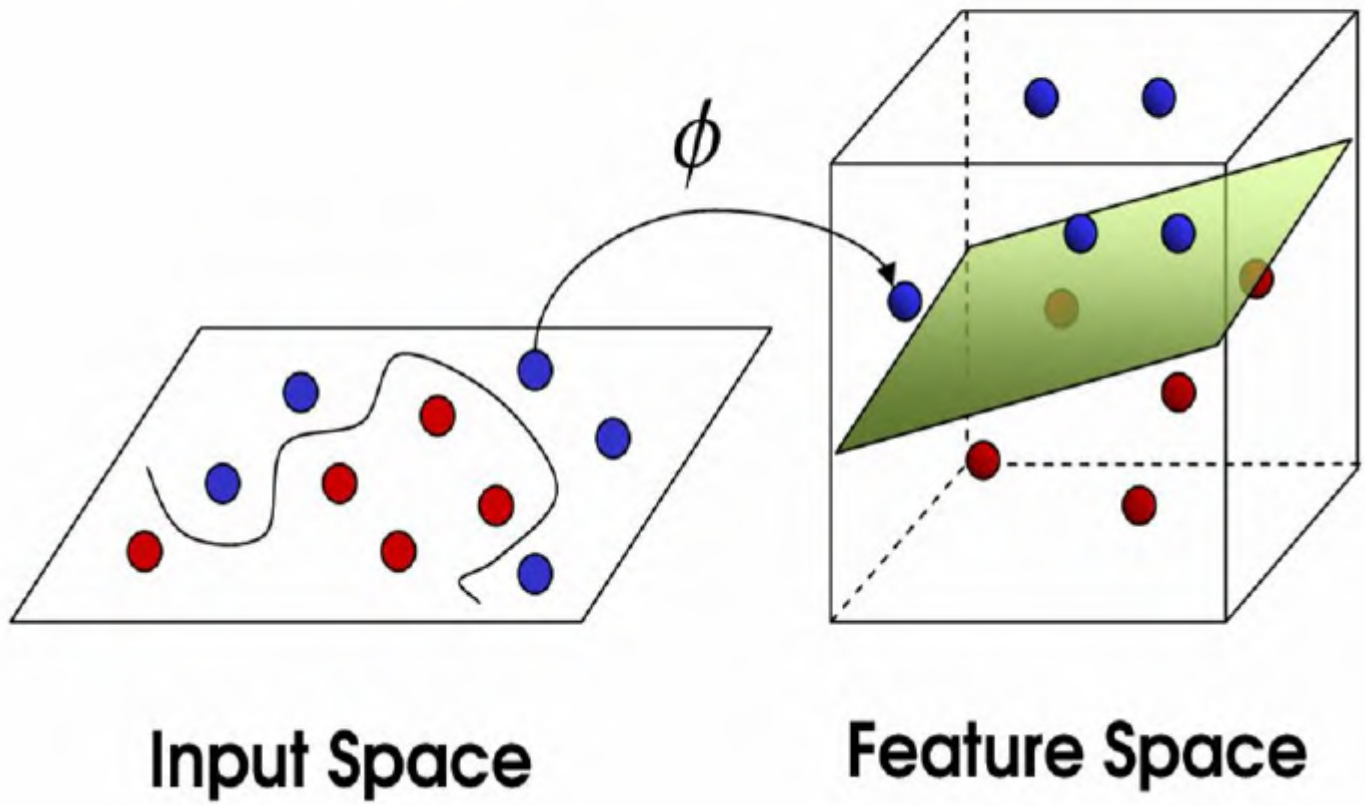
- sometimes, two classes will not be linearly separable



Cox & Savoy (2003)

nonlinear classifiers

- Nonlinear decision boundaries can be represented as linear boundaries on a transformed feature space



nonlinear classifiers

- Nonlinear decision boundaries can be represented as linear boundaries on a transformed feature space

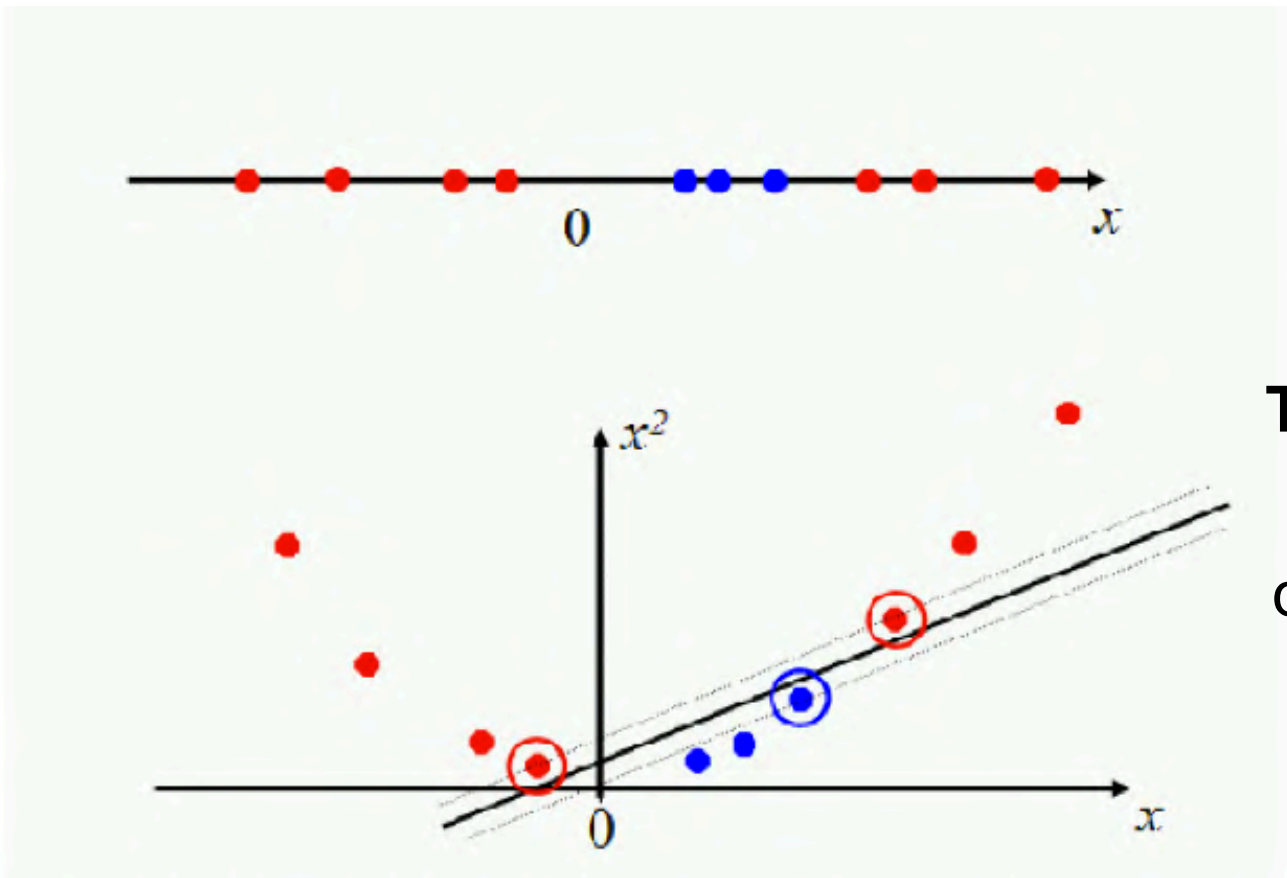


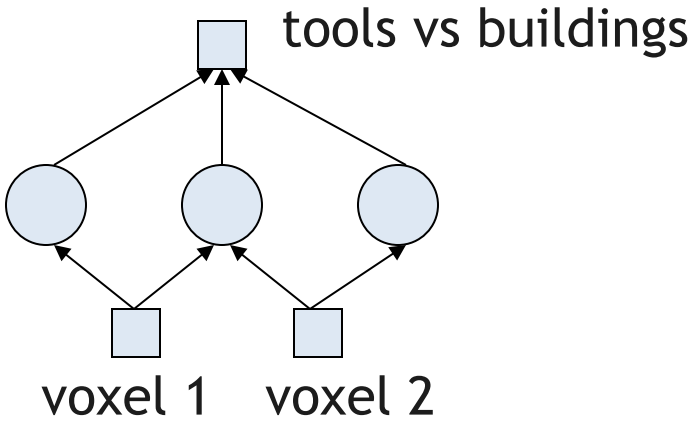
Figure 15.6: Projecting data that is not linearly separable into a higher dimensional space can make it linearly separable.

The "kernel trick"

(here a simple quadratic function creates the extra dimensionality)

nonlinear classifiers

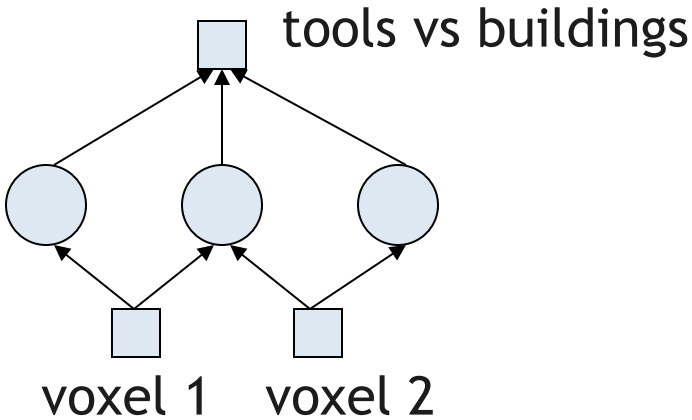
- **neural networks:**
  new features are learnt
  "hidden layer"



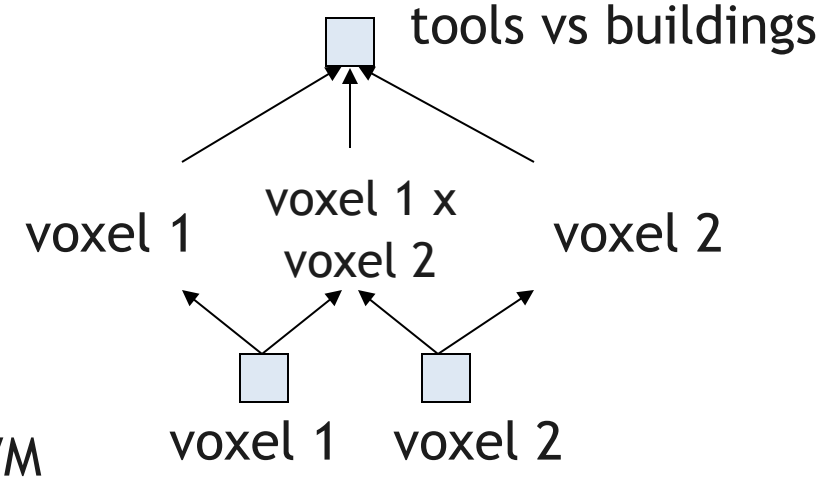
Slide from Francisco Pereira

nonlinear classifiers

- **neural networks:**
  new features are learnt
  "hidden layer"



- **SVMs:**
  new features are (implicitly)
  determined by a kernel



quadratic SVM

Slide from Francisco Pereira

nonlinear classifiers

reasons to be careful:

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

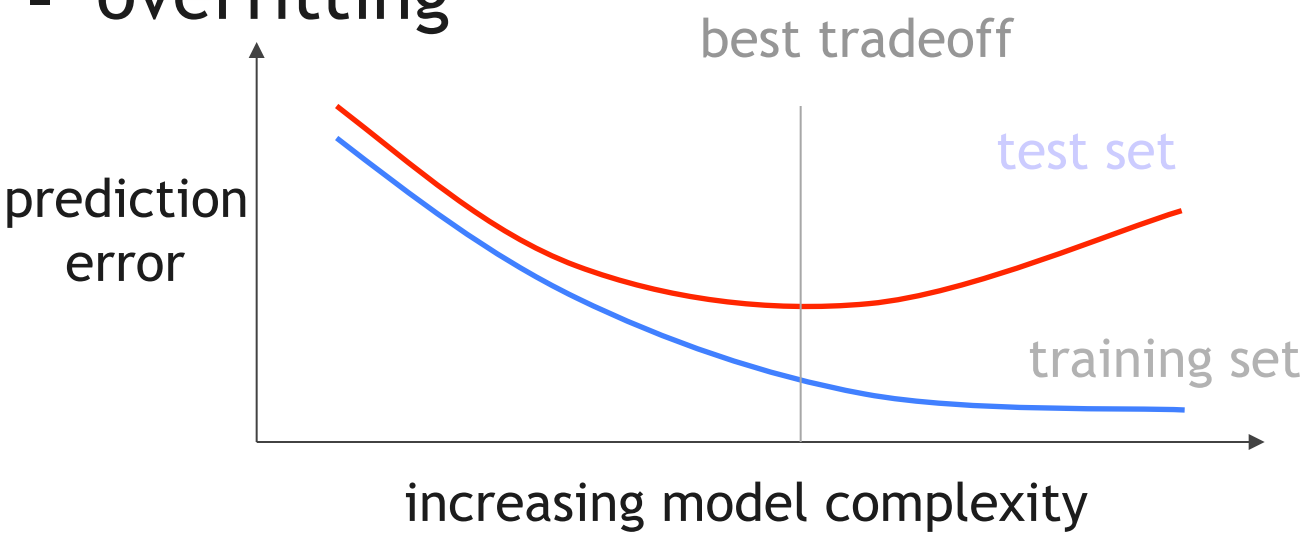
Slide from Francisco Pereira

nonlinear classifiers

reasons to be careful:

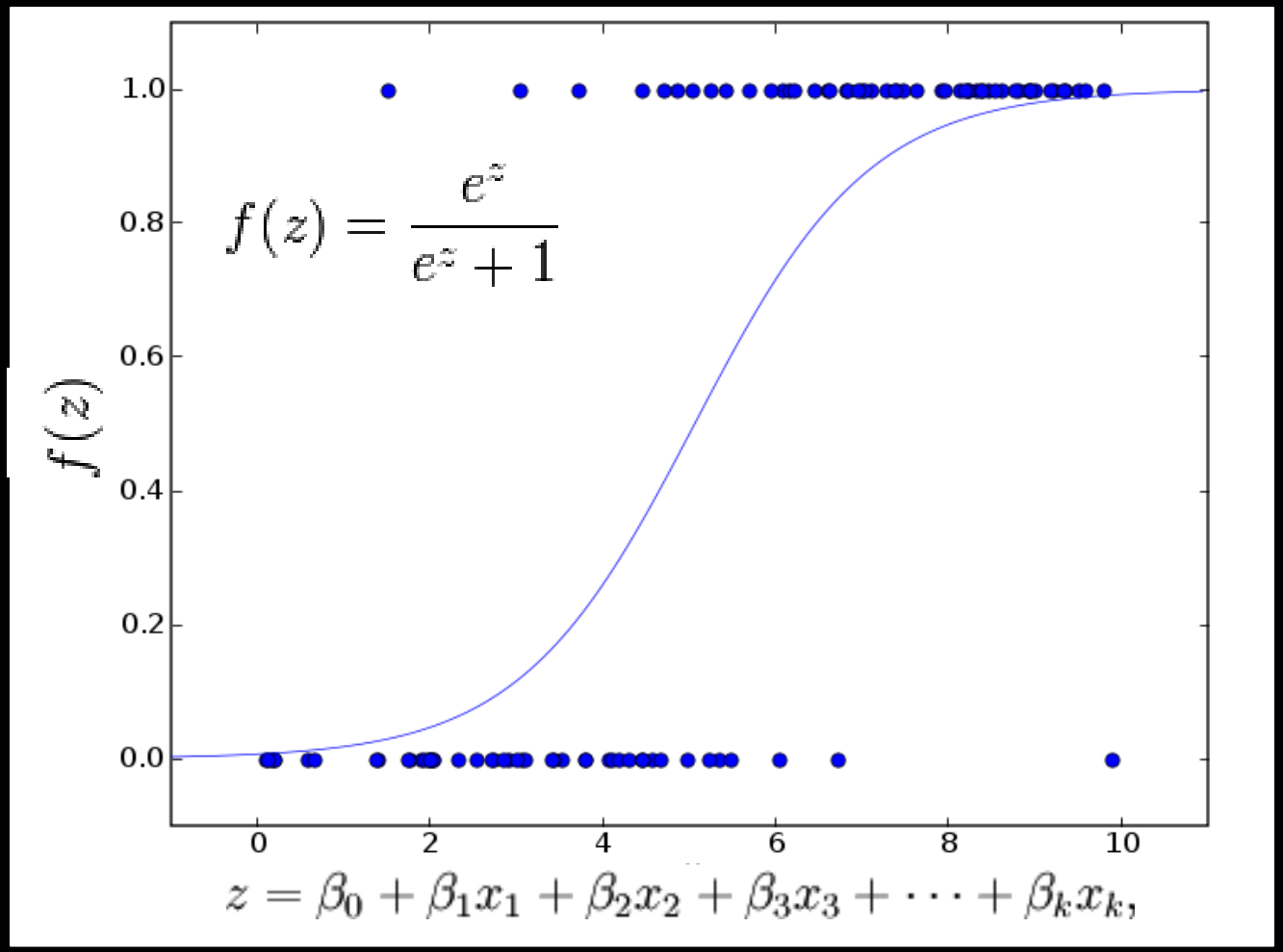
- too few examples,
  too many features
- harder to interpret

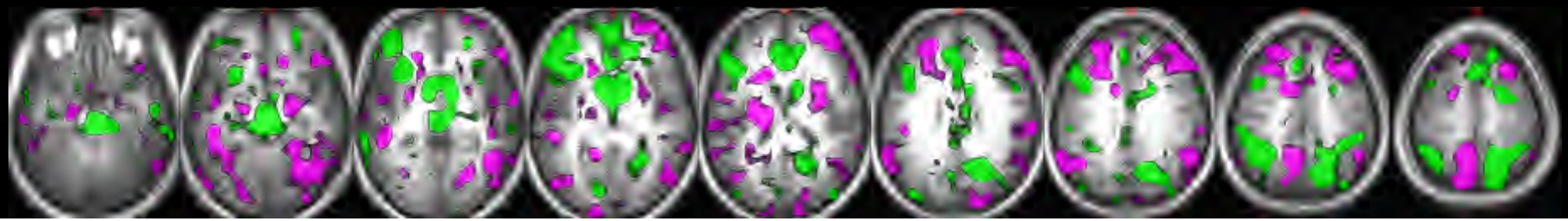
- overfitting



[from Hastie et al,2001]

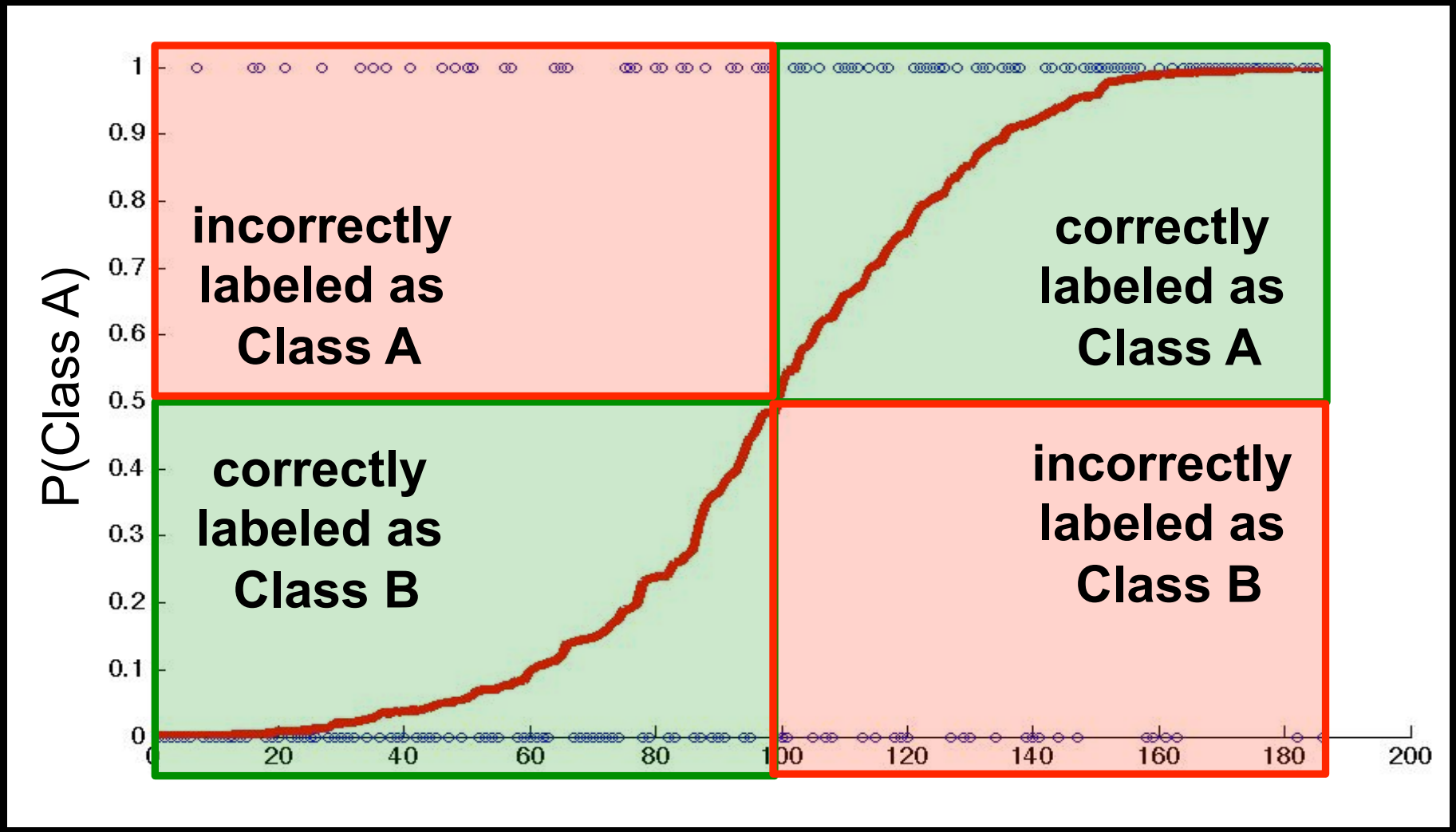
Logistic Regression Classifier





Testing the trained classifier:

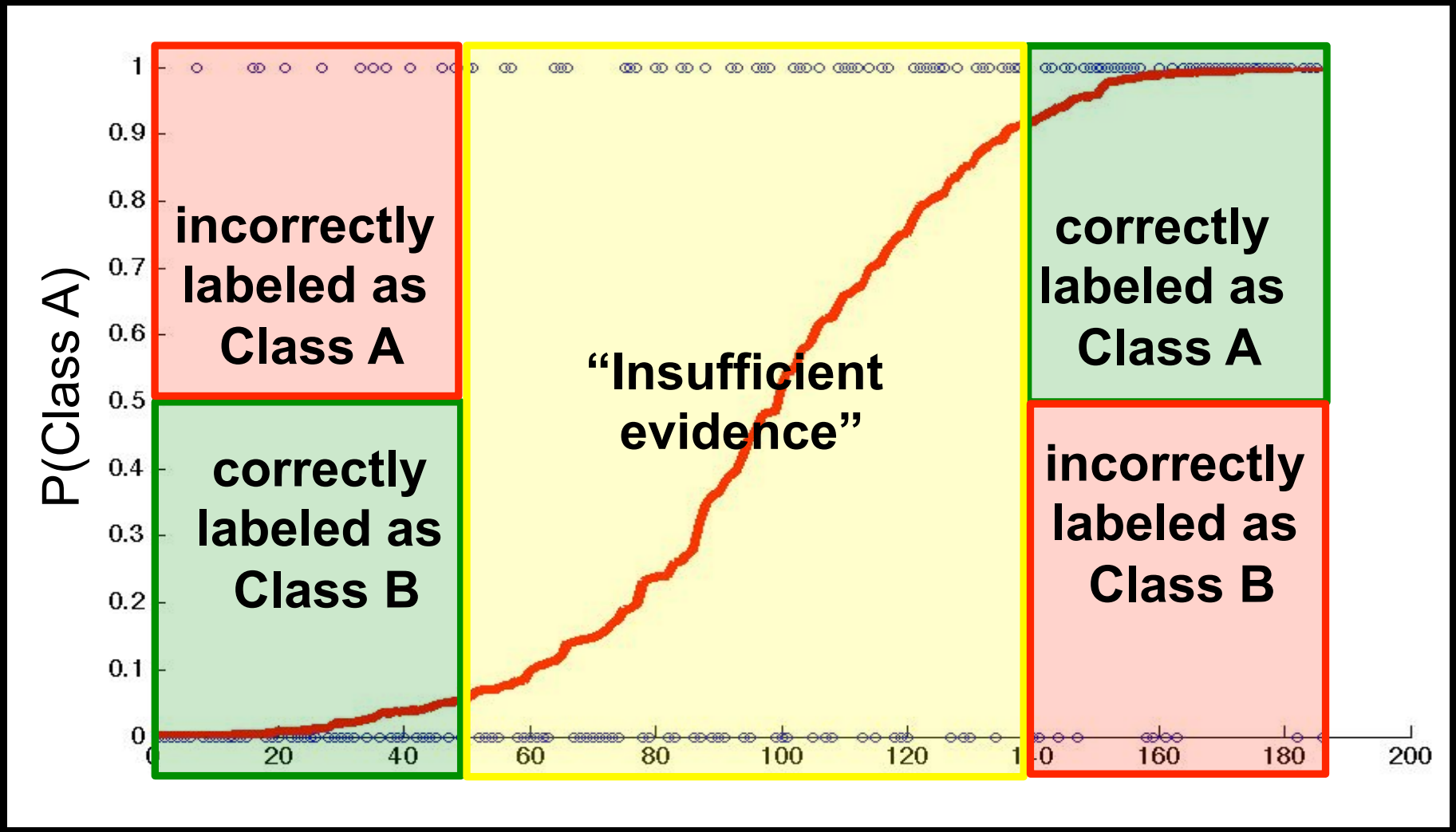
Ranked classifier predictions for individual test trials



70% accuracy

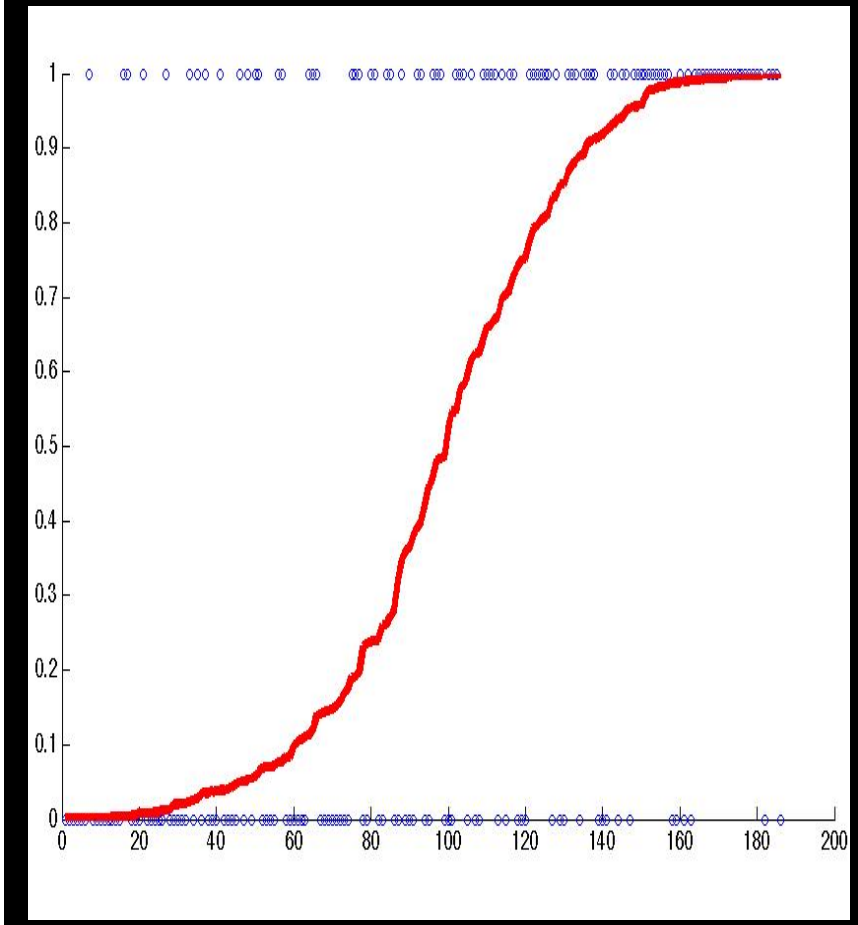
Testing the trained classifier:

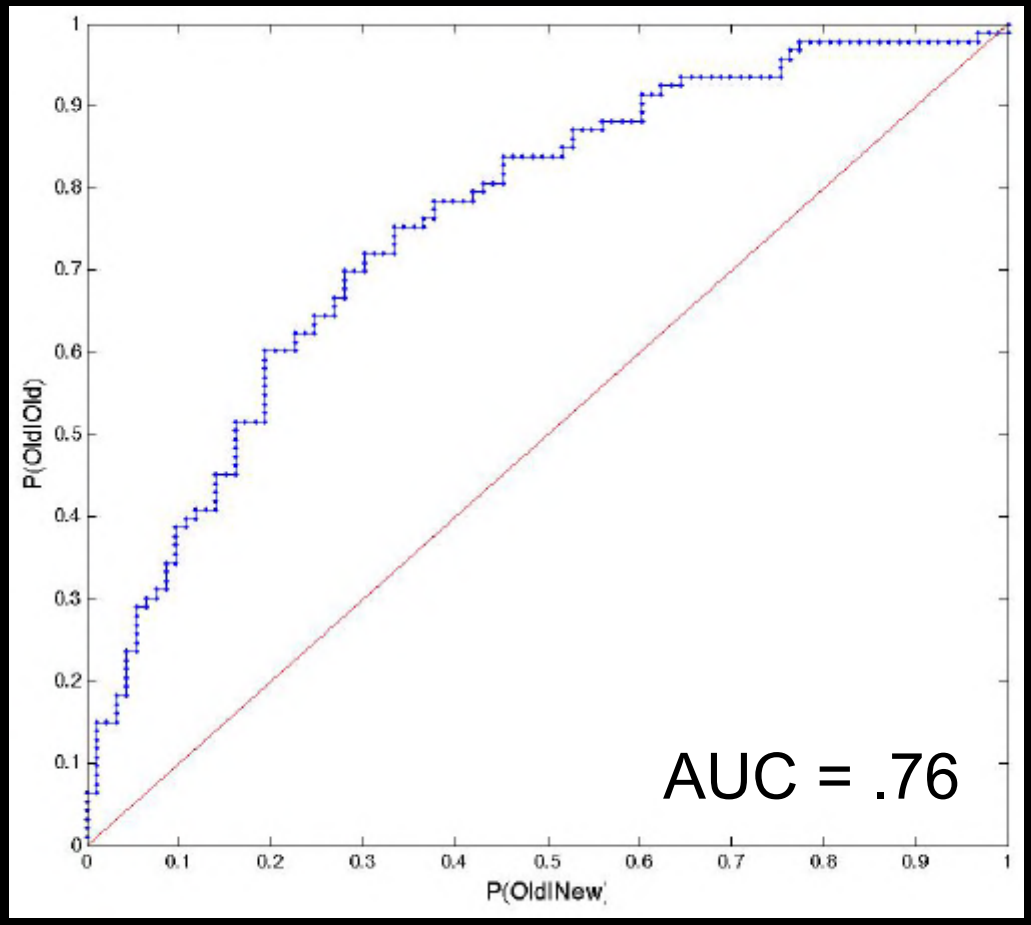
Ranked classifier predictions for individual test trials

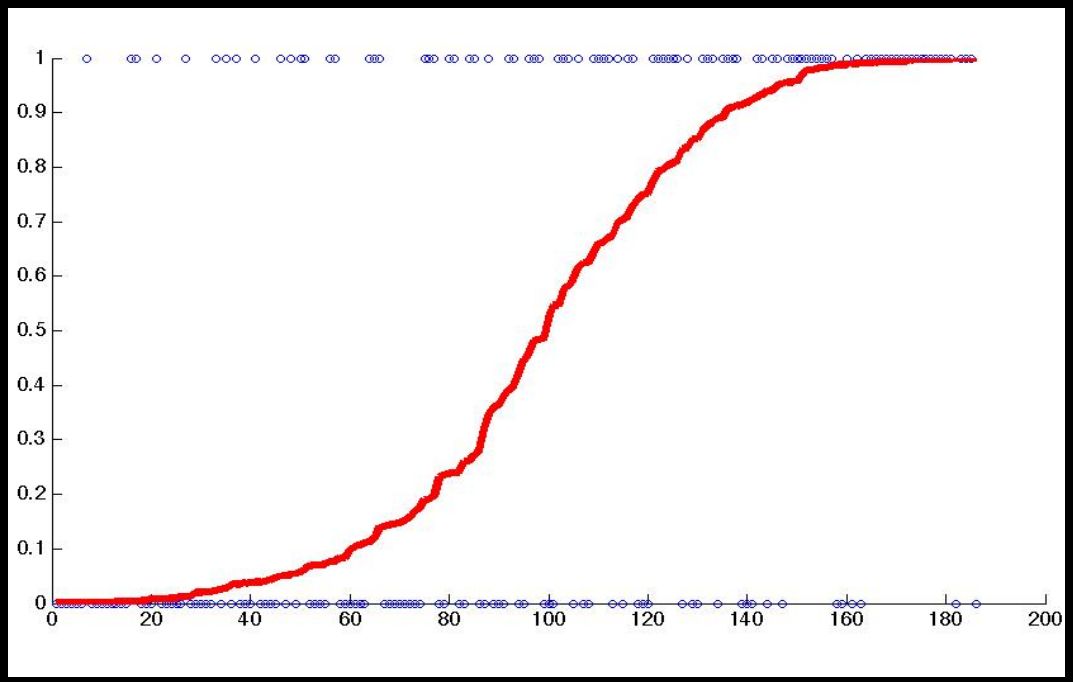


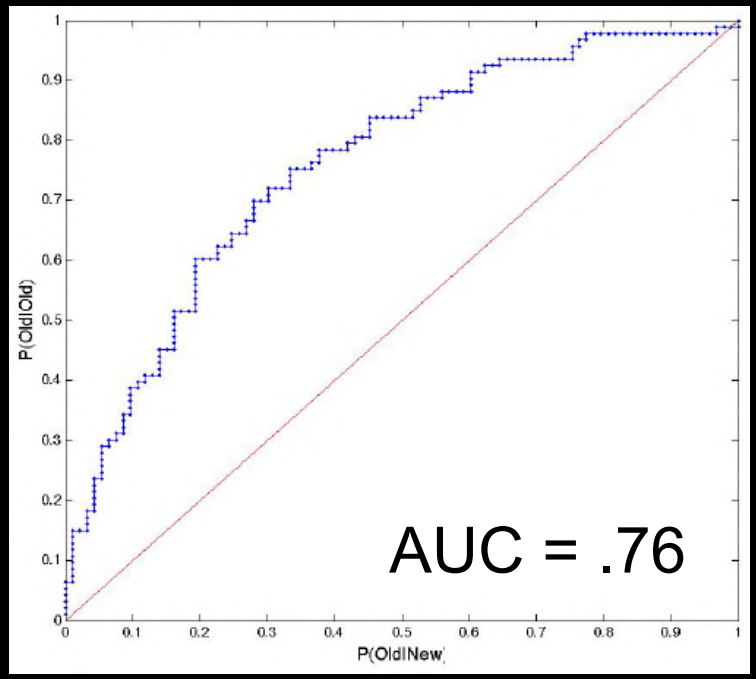
85% accuracy

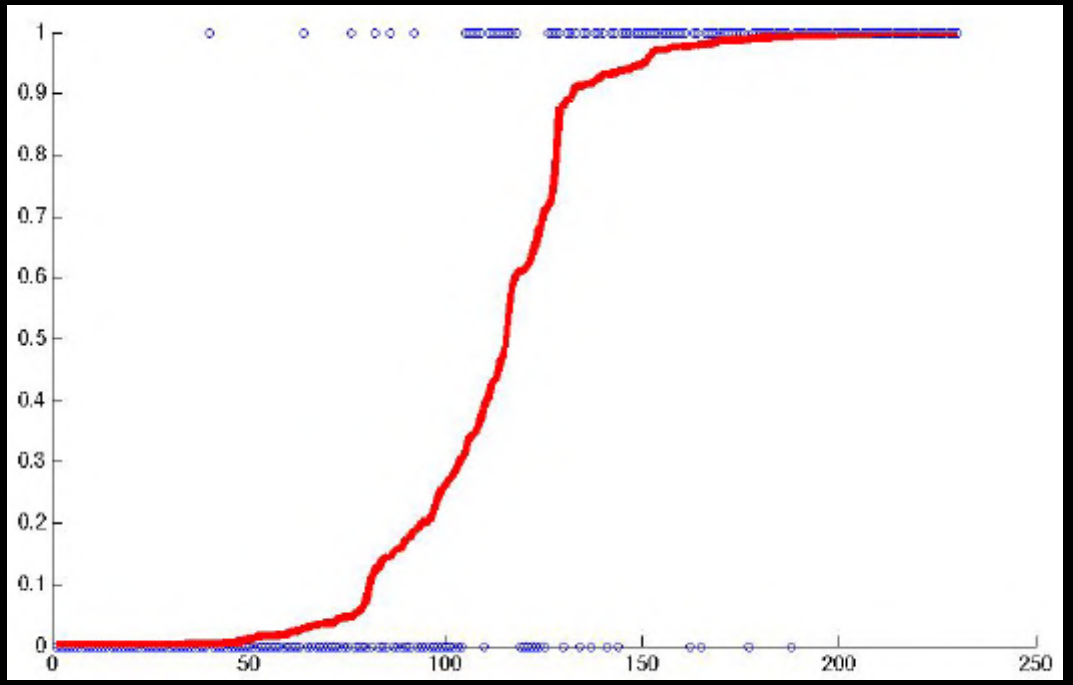
Representing classification performance with receiver operating characteristic (ROC) curves

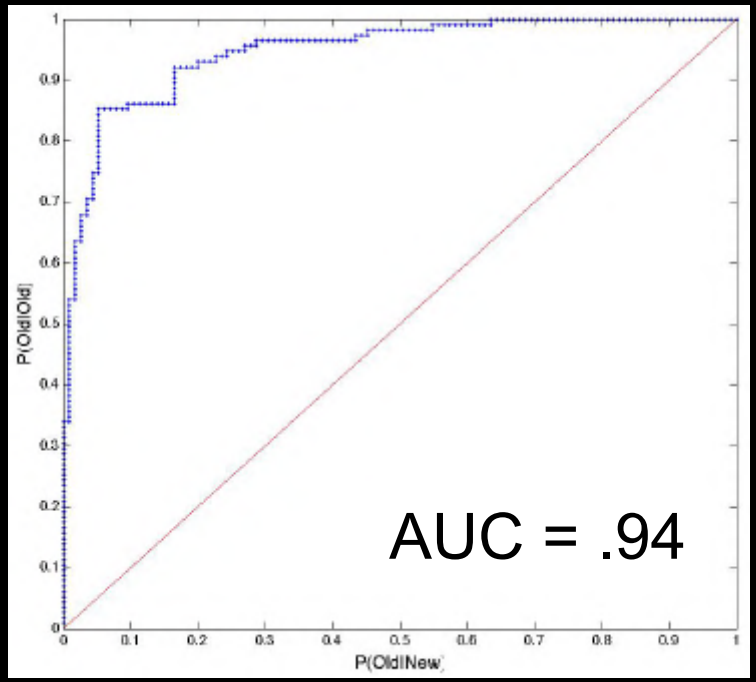


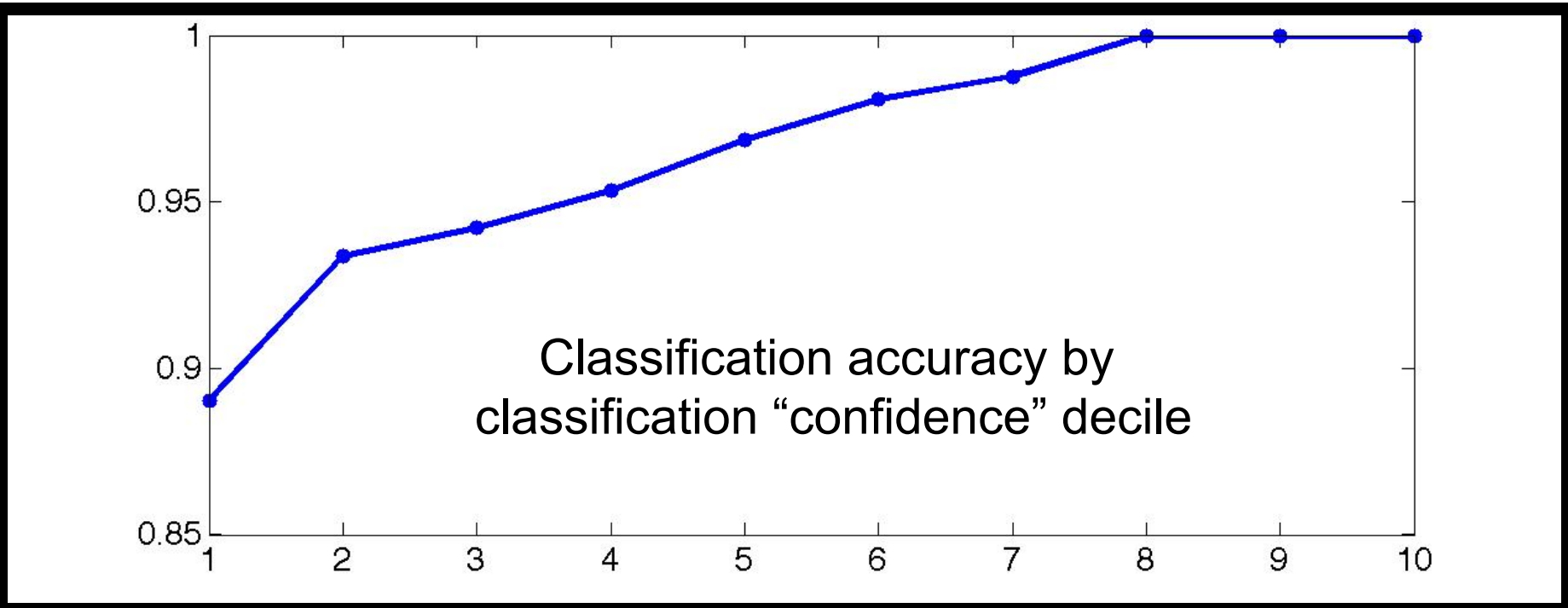


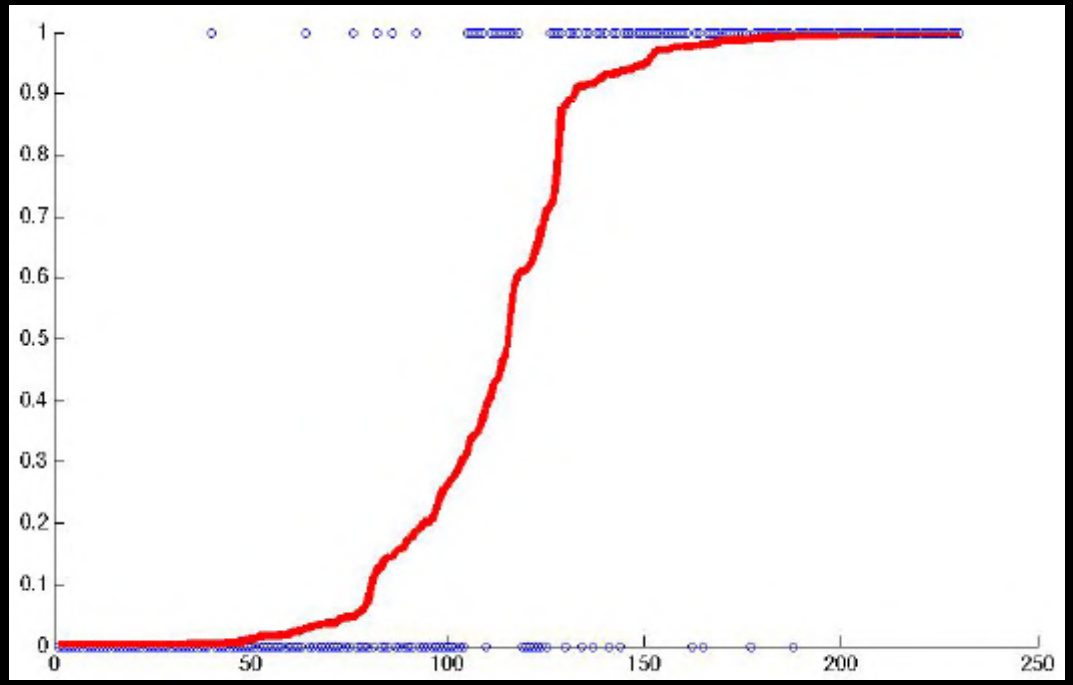

0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140
160
180
200

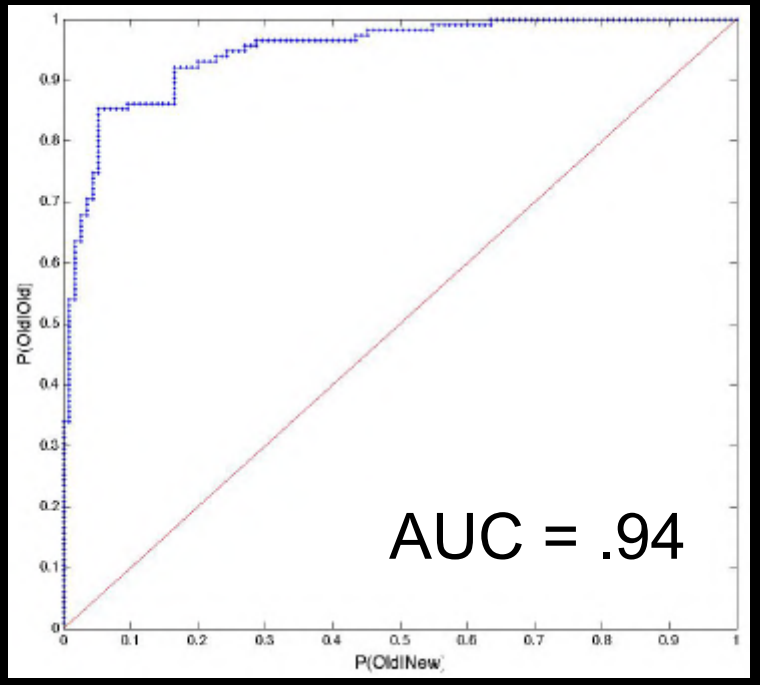

P(OldIOld)
P(OldINew)
AUC = .76


0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250

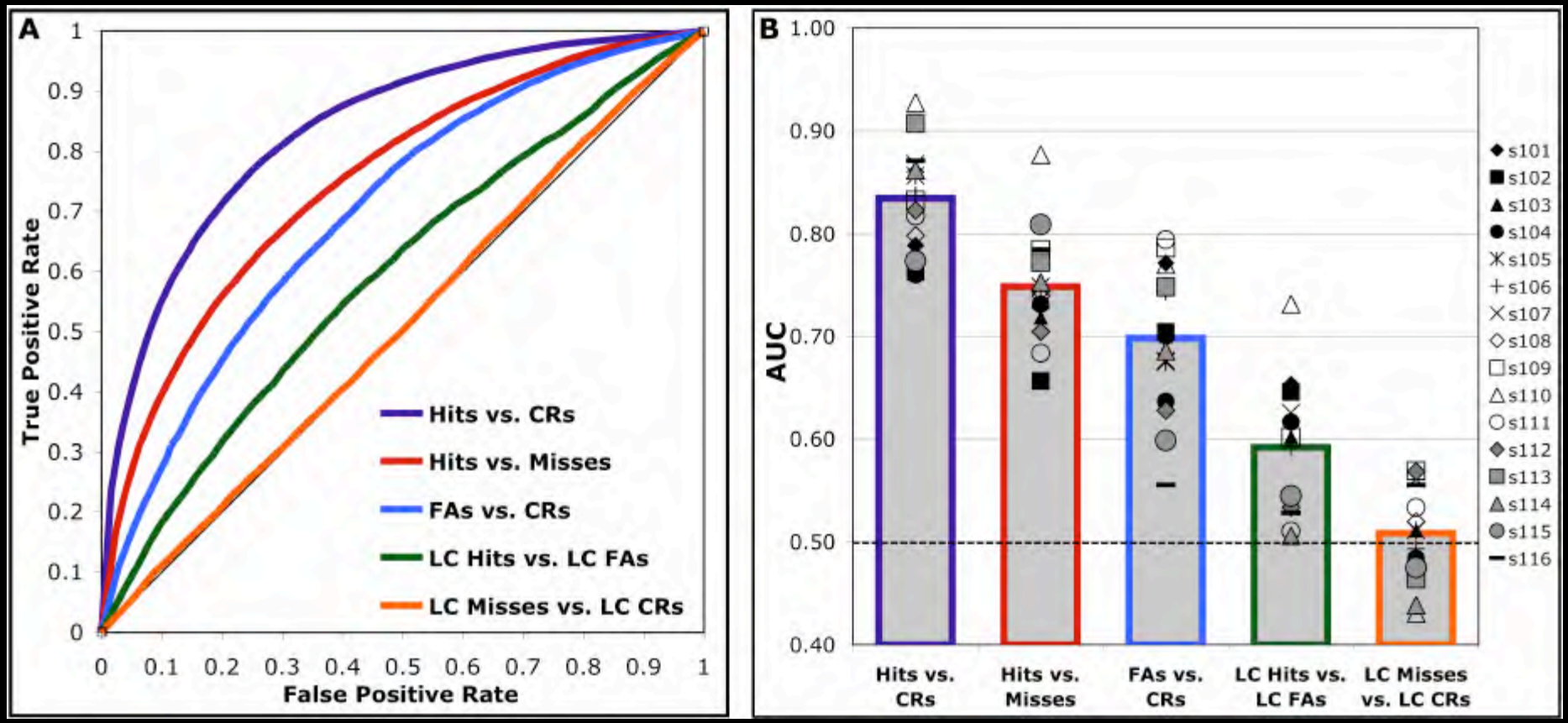

P(OldIOld)
P(OldINew)
AUC = .94


Classification accuracy by
classification “confidence” decile
1
0.95
0.9
0.85
1
2
3
4
5
6
7
8
9
10

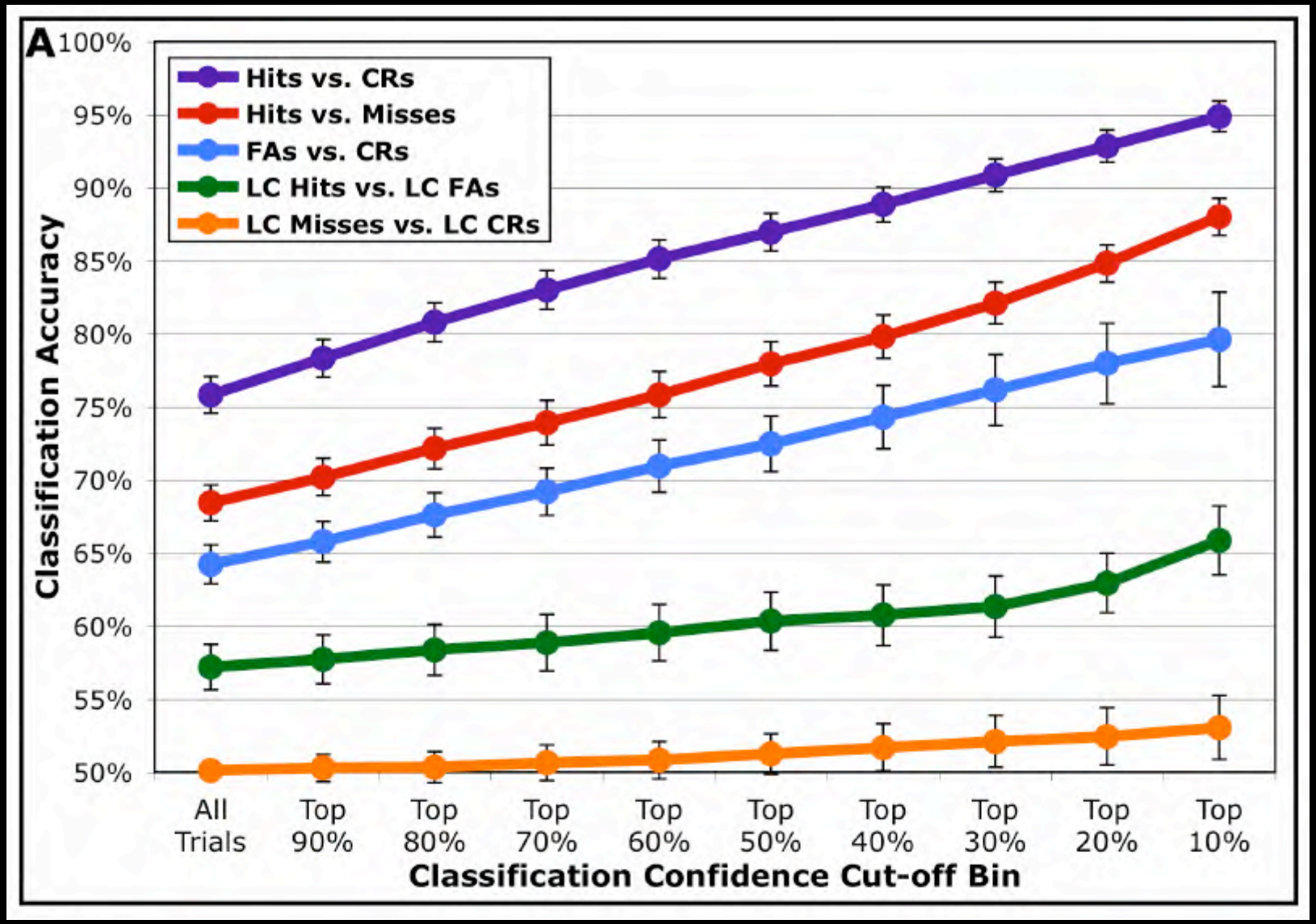

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250


AUC = .94
P(Old|Old)
P(Old|New)

Group summary of classification ROC data for a recognition memory experiment



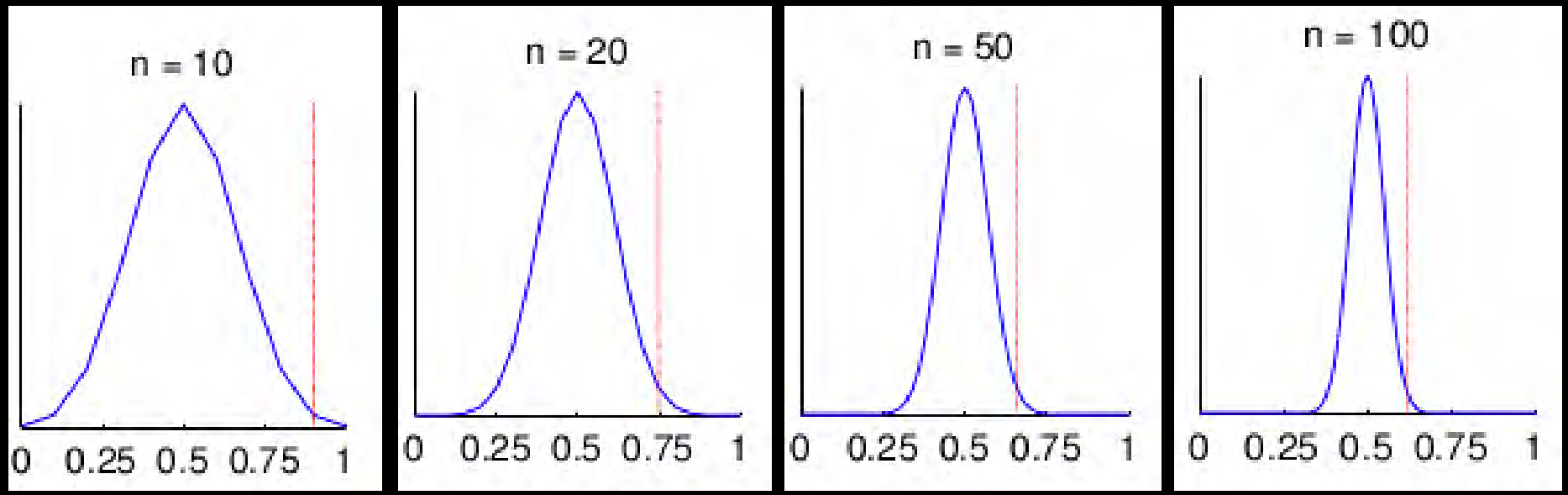
Rissman, Greely, & Wagner (2010) *Proceedings of the National Academy of Sciences*



Rissman, Greely, & Wagner (2010) *Proceedings of the National Academy of Sciences*

Determining whether classification accuracy is significantly better than chance

- **Within-subjects**
  - Can compare to chance using binomial distribution
    - *e.g.,* $p_{chance} = .5$, *trials = 120, successful = 75,* $p_{observed} = .0039$

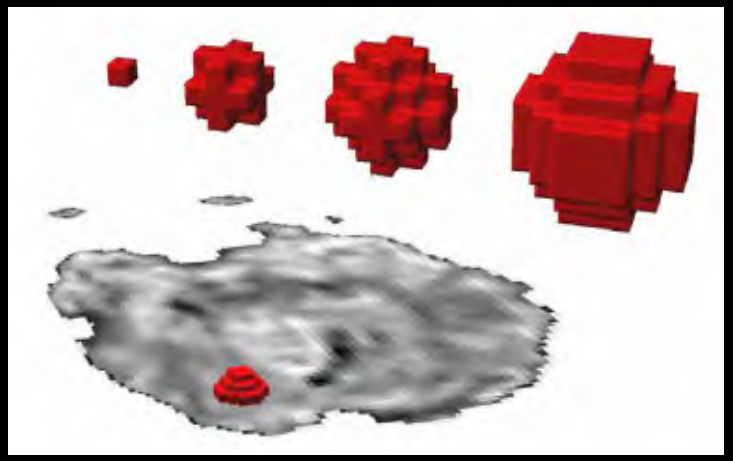


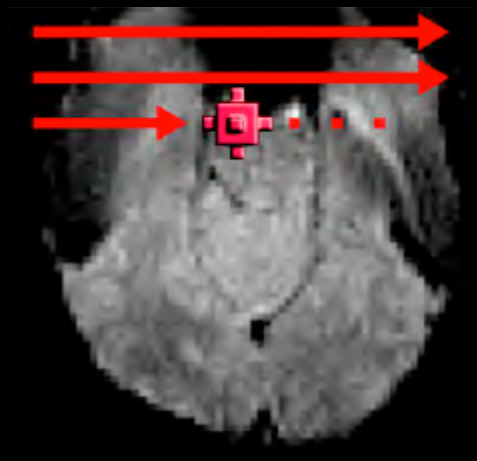
Determining whether classification accuracy is significantly better than chance

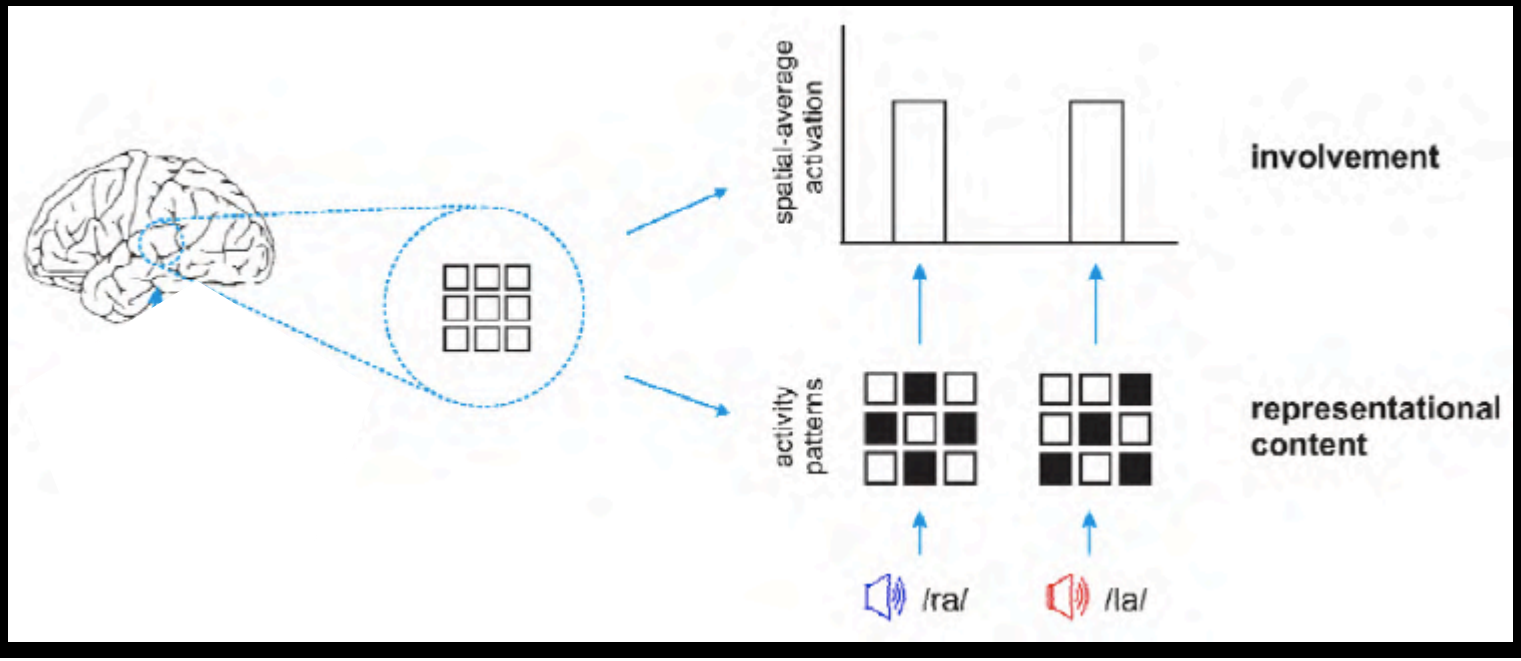
- **Within-subjects**
  - Can compare to chance using binomial distribution
    - *e.g.,* $p_{chance} = .5$, *trials = 120, successful = 75,* $p_{observed} = .0039$
  - Or use permutation test to compare observed accuracy to distribution of performance generated by shuffling regressors many times (e.g., 1000 shuffles)
- **Across-subjects**
  - Use one-sample t-test to compare subjects' mean classification accuracy to chance
  - Commonly used, but places no requirement on mean accuracy (i.e., 53% correct could be highly significant)

Information-based brain mapping

– Spherical searchlight mapping approach (Kriegeskorte et al. (2006), *PNAS*)

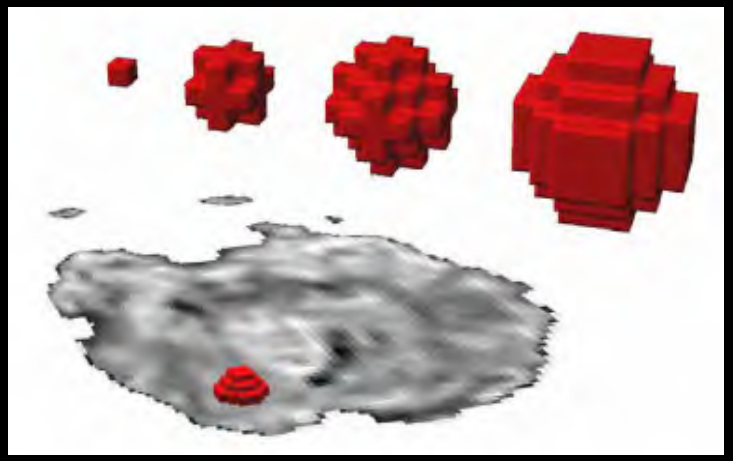


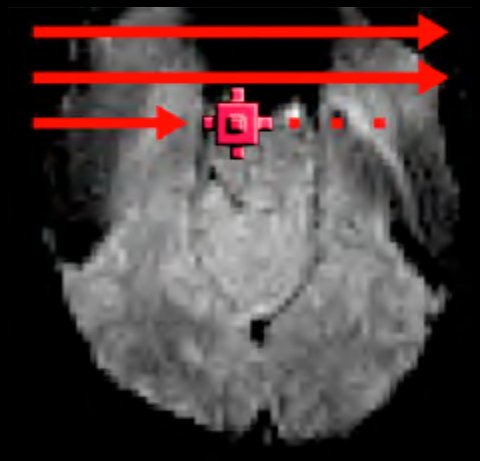




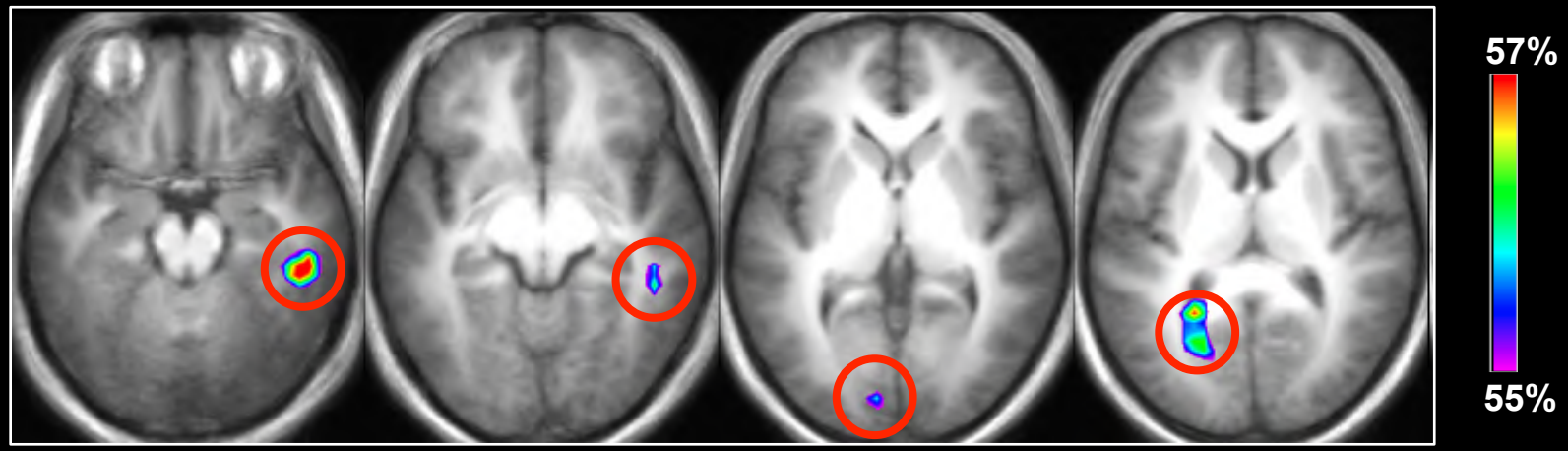
Information-based brain mapping

– Spherical searchlight mapping approach (Kriegeskorte et al. (2006), *PNAS*)



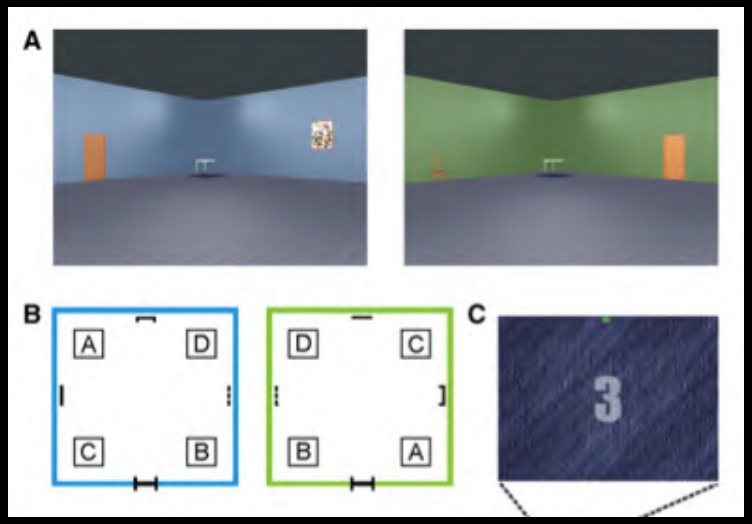


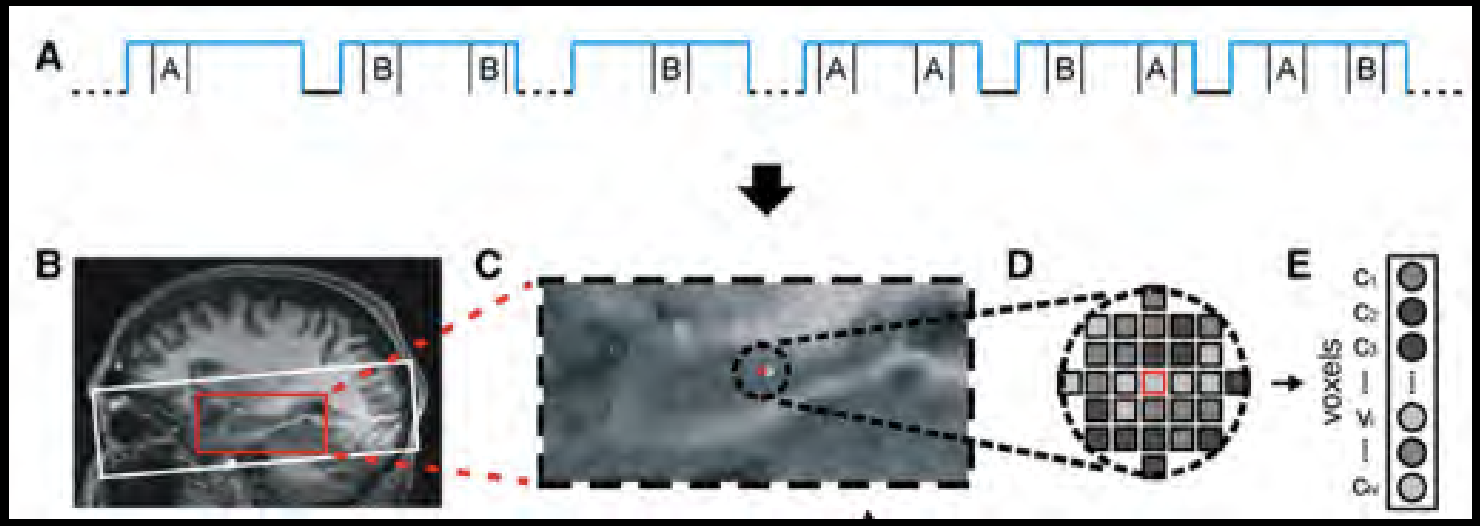
Decoding Memory for Faces: True vs. False Recognition



Rissman, Greely, & Wagner (2010) *Proceedings of the National Academy of Sciences*

Decoding the representational content of medial temporal lobe activity patterns with high resolution fMRI data

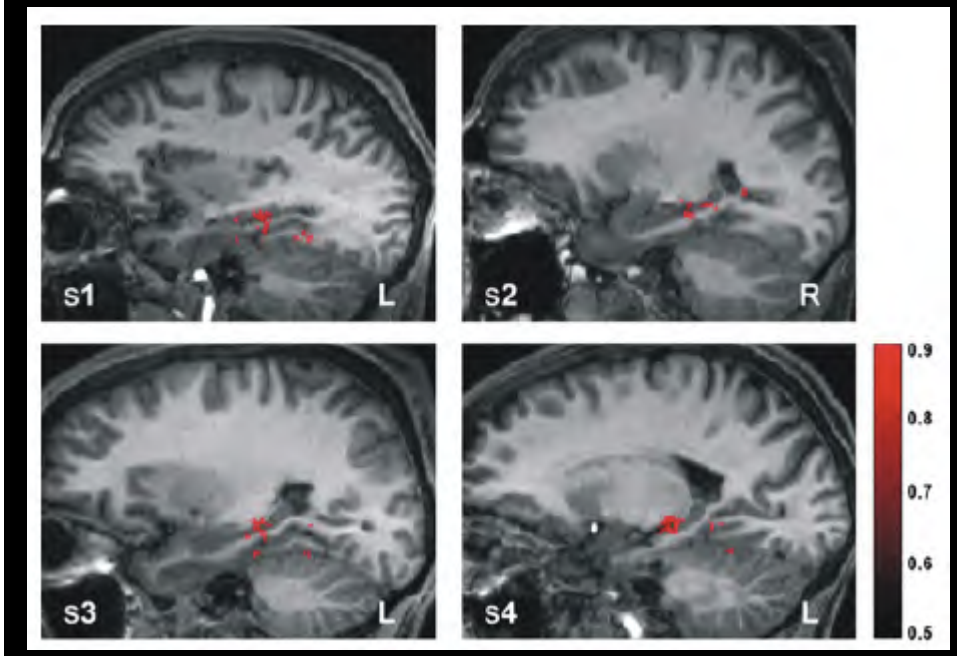




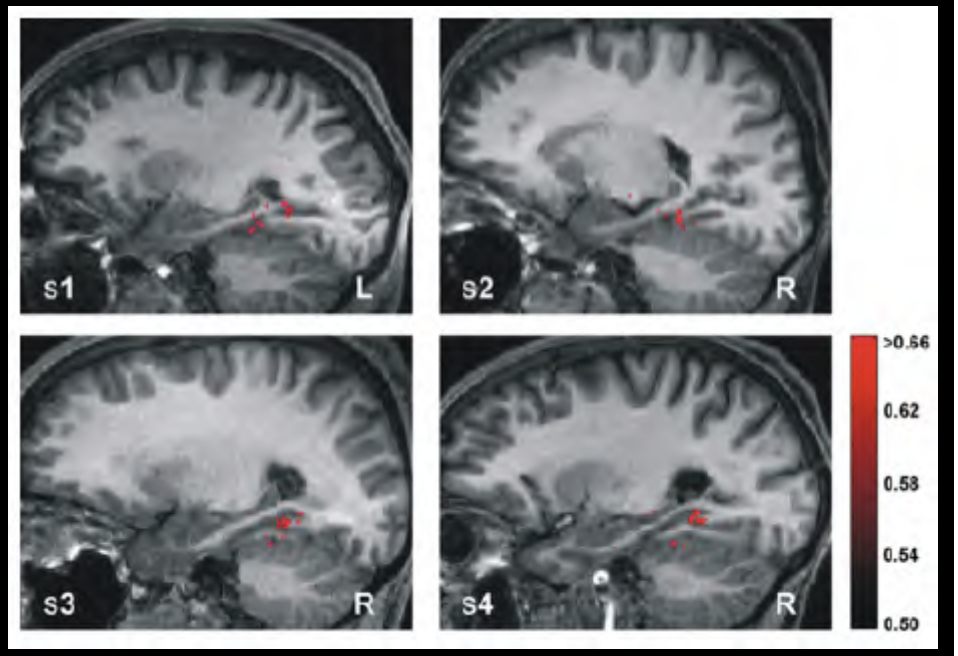
Hassabis et al. (2009) *Neuron*

Decoding the representational content of local activity patterns within the MTL

Which spatial quadrant?

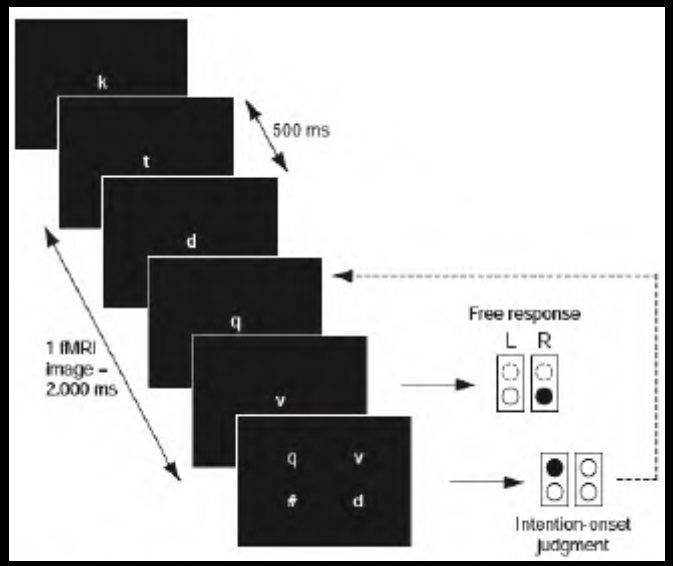


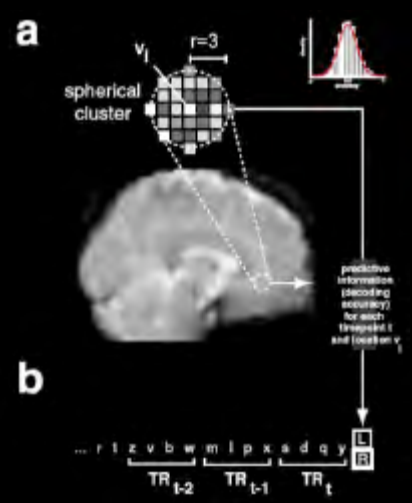
Which room?

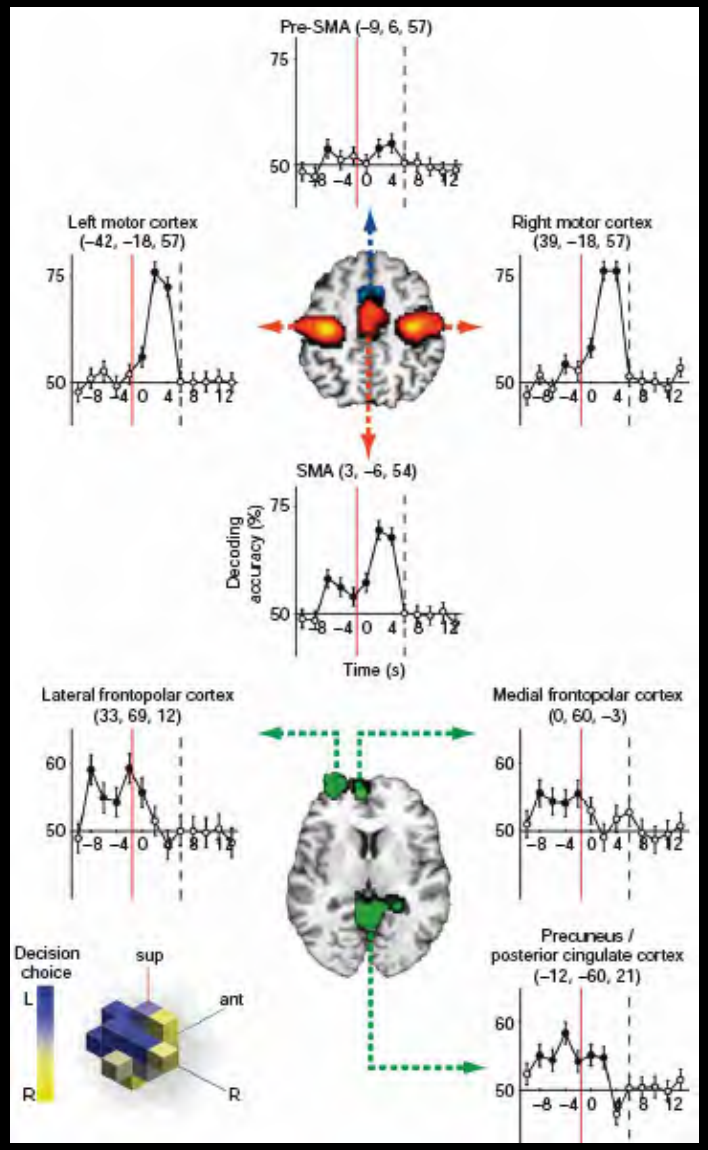


Hassabis et al. (2009) *Neuron*

Unraveling the Notion of Free Will?

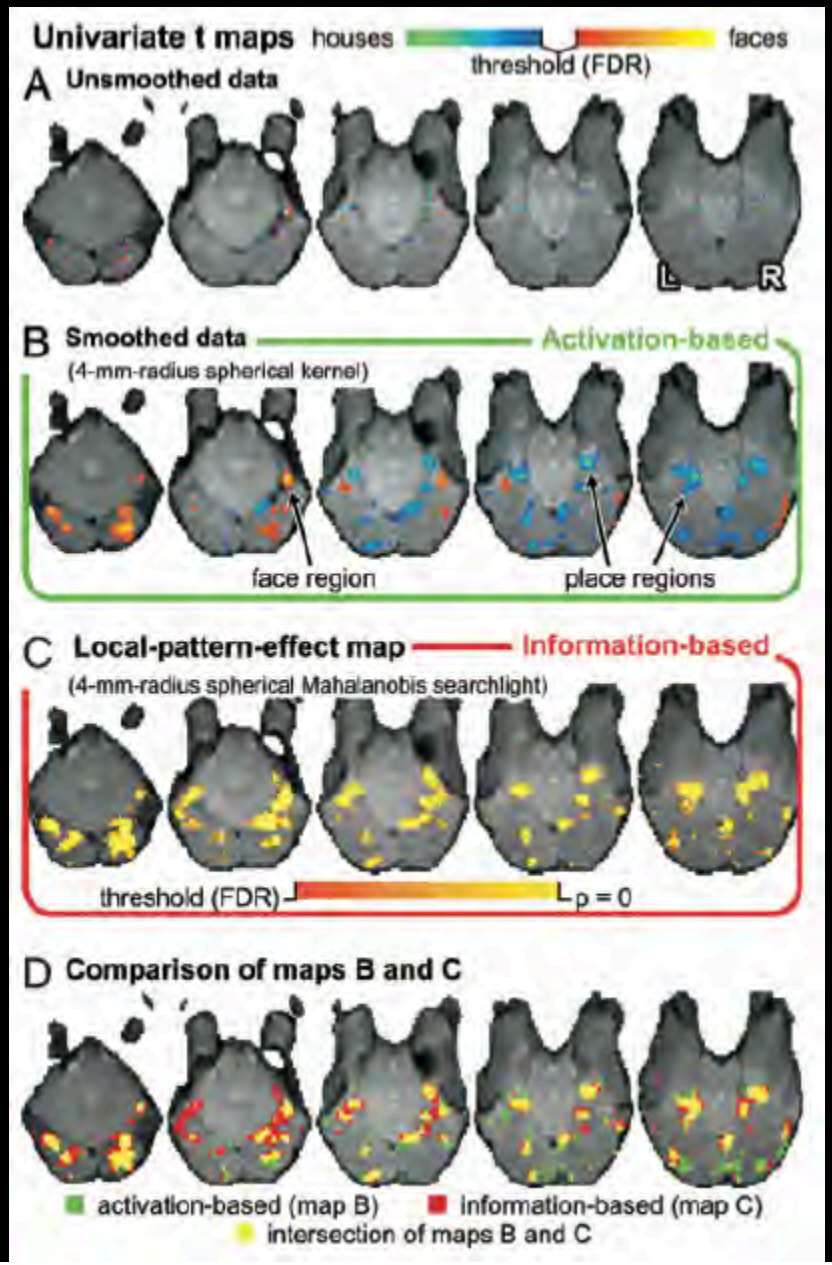






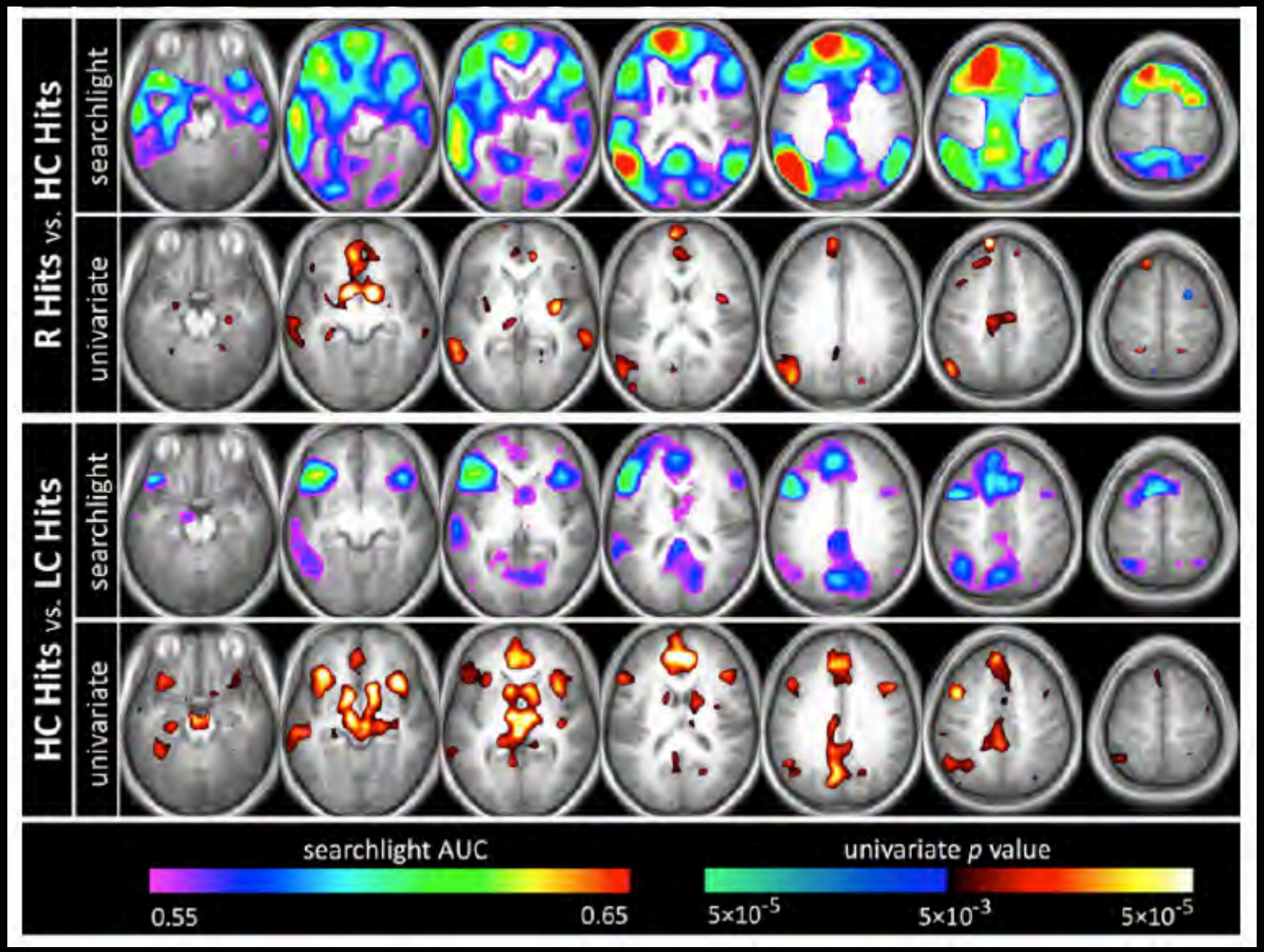
Soon et al. (2008) *Nature Neuroscience*

Univariate Mapping vs. MVPA Searchlight Mapping



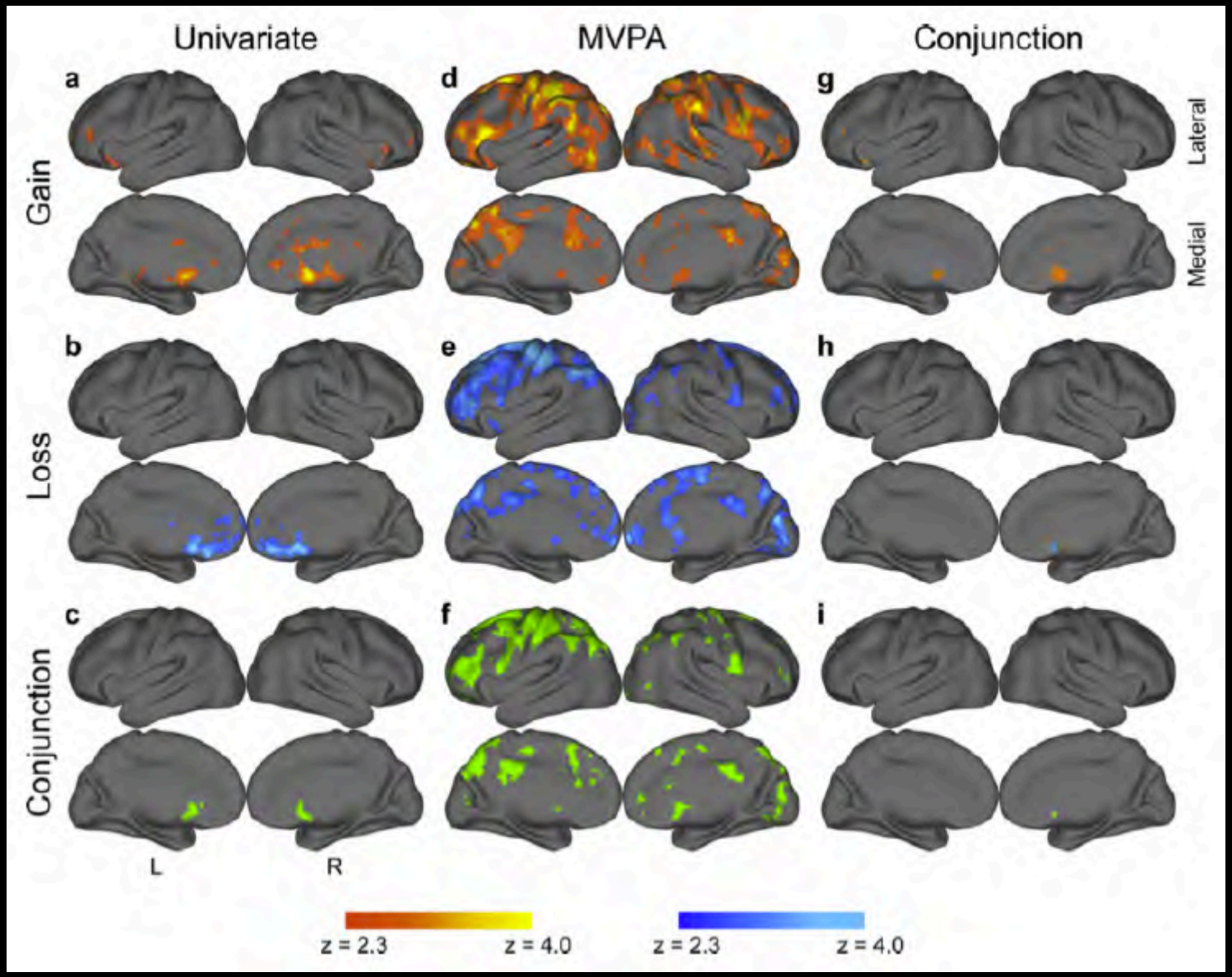
Kriegeskorte et al. (2006), *PNAS*

Univariate Mapping vs. MVPA Searchlight Mapping



Rissman, Greely, & Wagner (2010) *PNAS*

Univariate Mapping vs. MVPA Searchlight Mapping



Jimura & Poldrack (2012) *Neuroimage*

Lots more to cover tomorrow…