Intro to Statistics for fMRI Analysis

Jingyuan Chen
032119

1st-level Analysis:
Single Subject-level Analysis

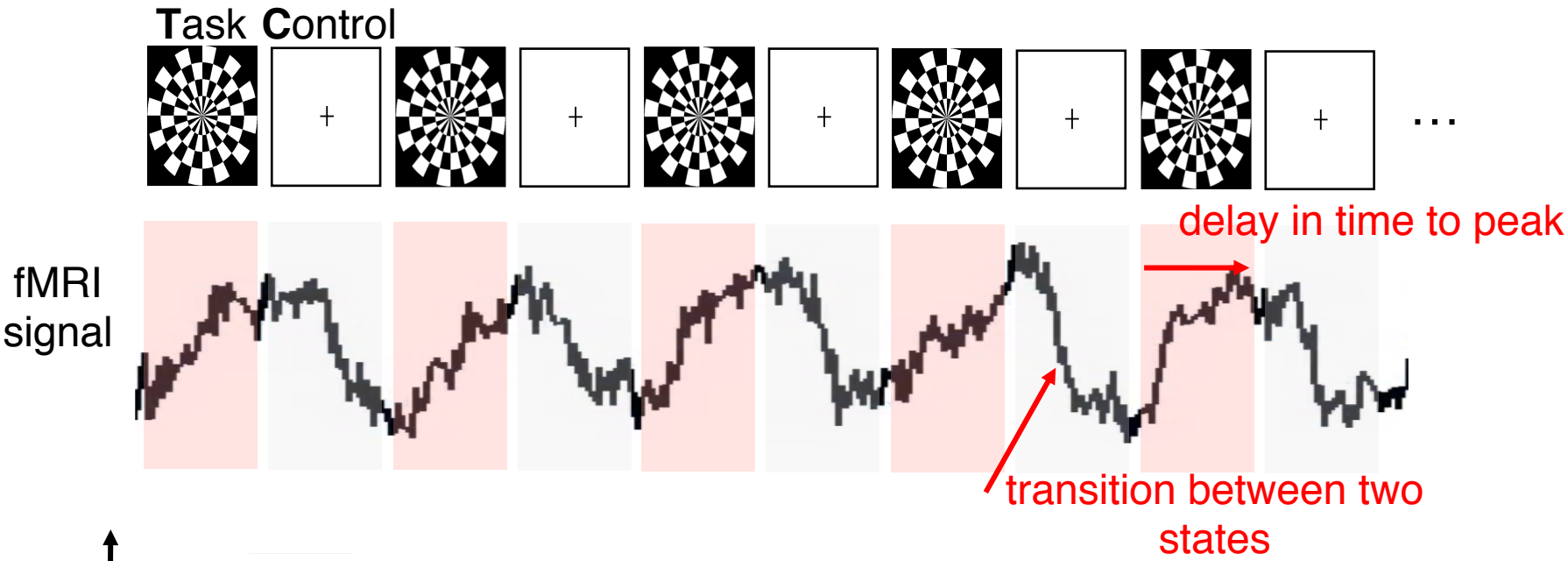
2nd-level Analysis:
Group-level Analysis

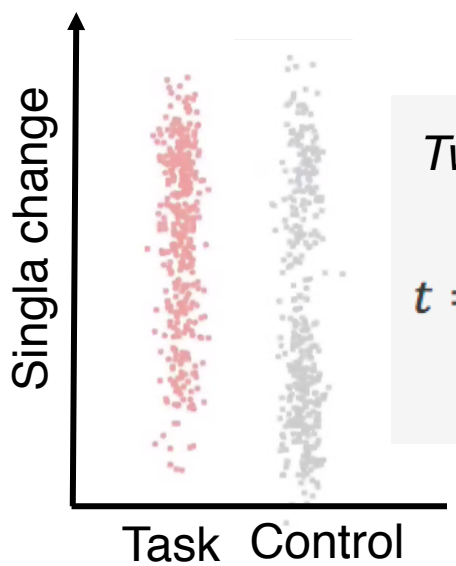
Correcting for Multiple Comparisons

Power Analysis

1st-level Analysis:
Single Subject-level Analysis

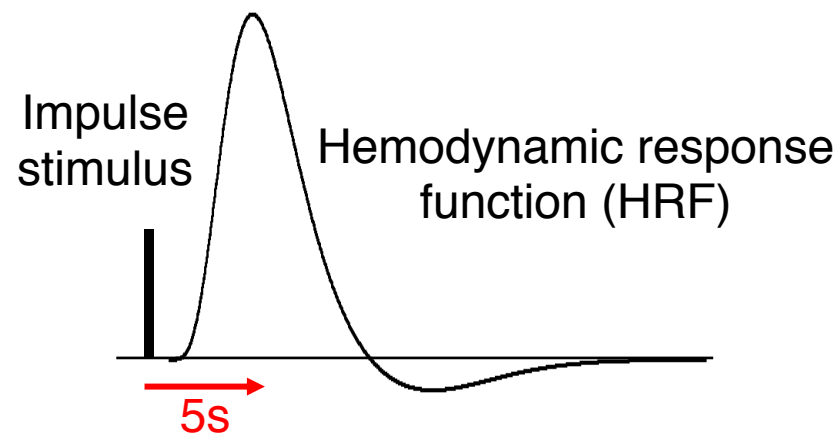
Simple t-tests are not so easy in fMRI





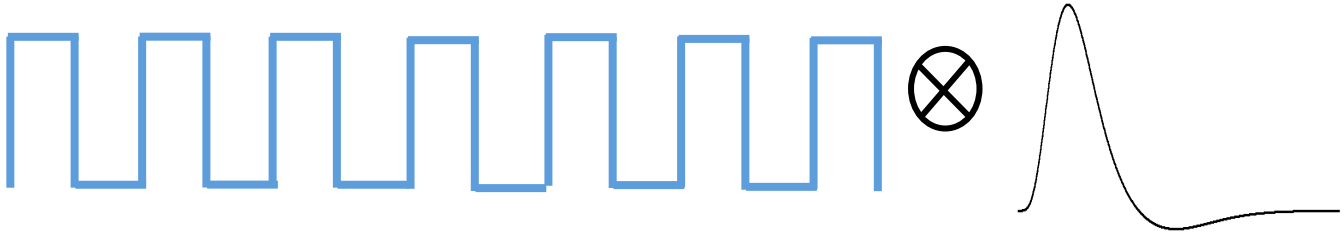
Two-sample t-test

$$t = \frac{\overline{S_T} - \overline{S_C}}{\sqrt{(\sigma_T^2 + \sigma_C^2)/n}}$$



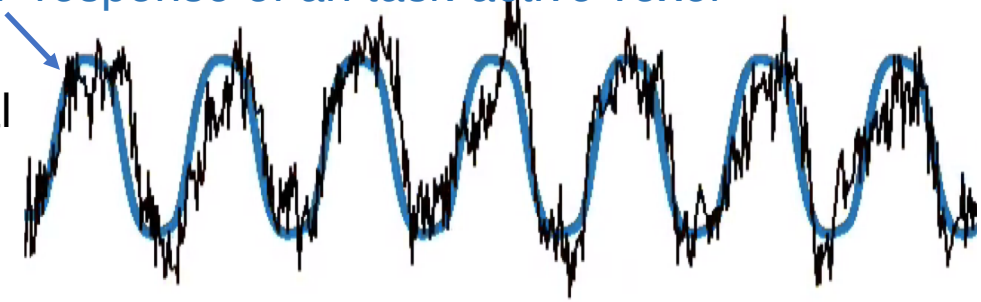
Correlation Analysis

Task paradigm



"Ideal" response of an task-active voxel

fMRI signal



linear Pearson correlation

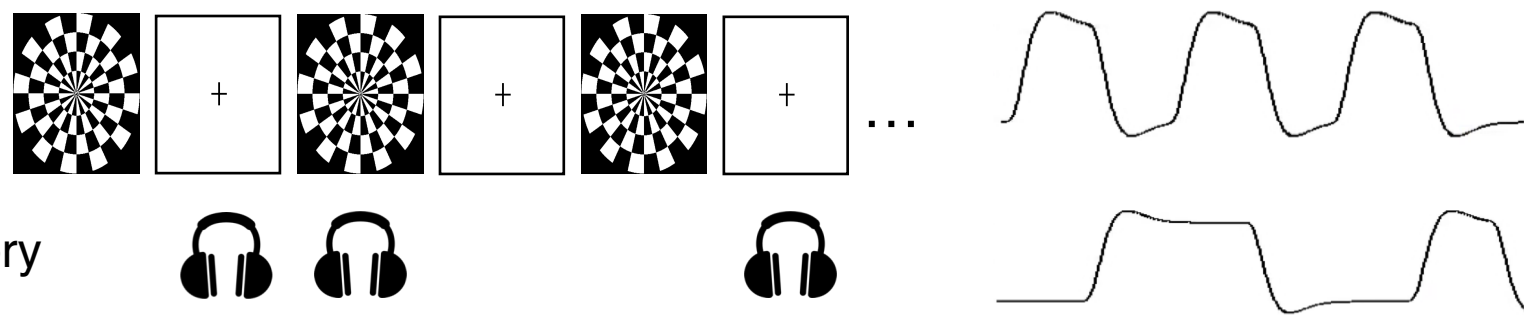
$$r_{x,y} = \frac{1}{n-1} \times \frac{\sum_n (x-\bar{x})(y-\bar{y})}{\sigma_x \sigma_y}$$

- Cannot be extended to scenarios with multiple task contrasts

concurrent

visual

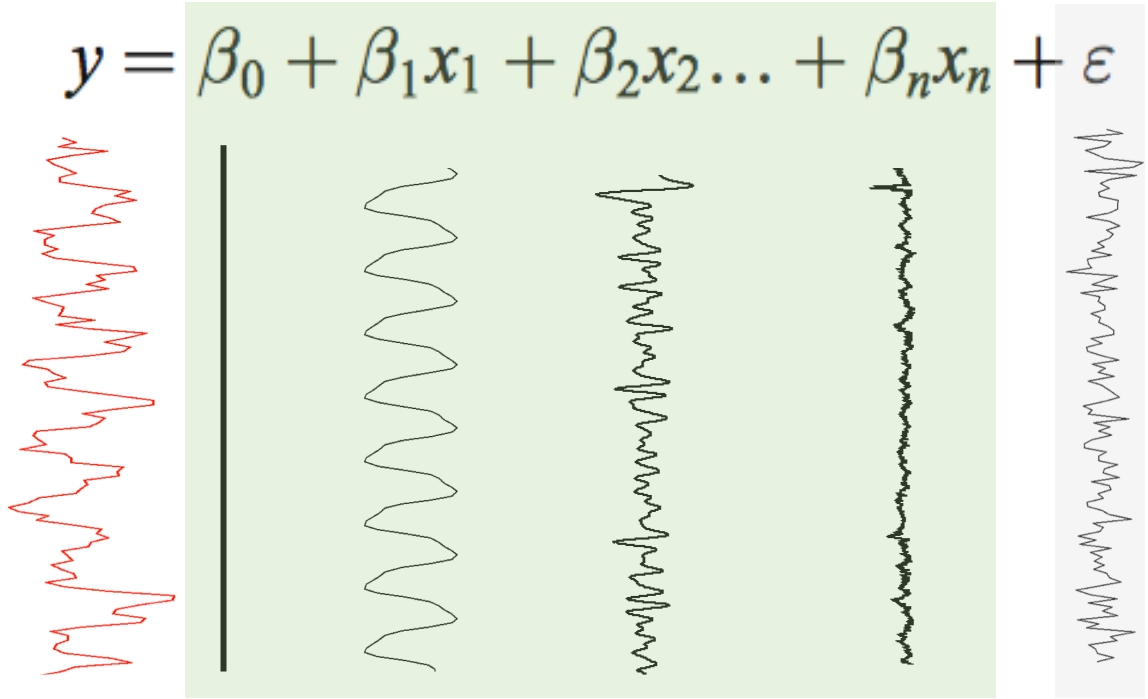
auditory



General Linear Model (GLM)

- fMRI signals as linearly additive mixtures of neural responses and “noise”

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n + \varepsilon$$



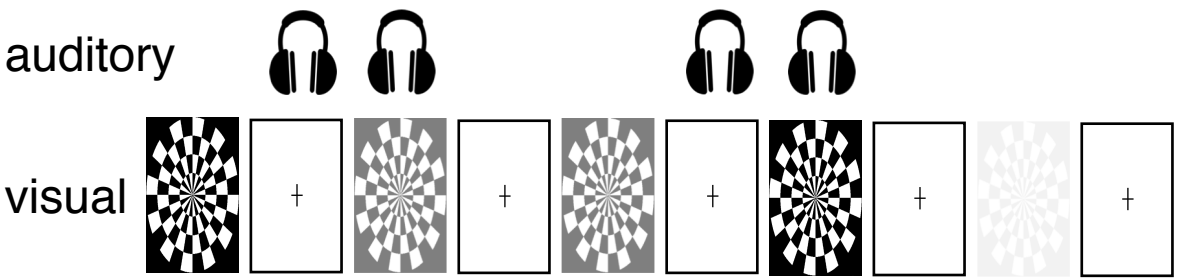
fMRI signal

modelled task responses,
behavioral data,
motion, physiology, etc.

un-modelled
“noise”

Design matrix

GLM applications - Example



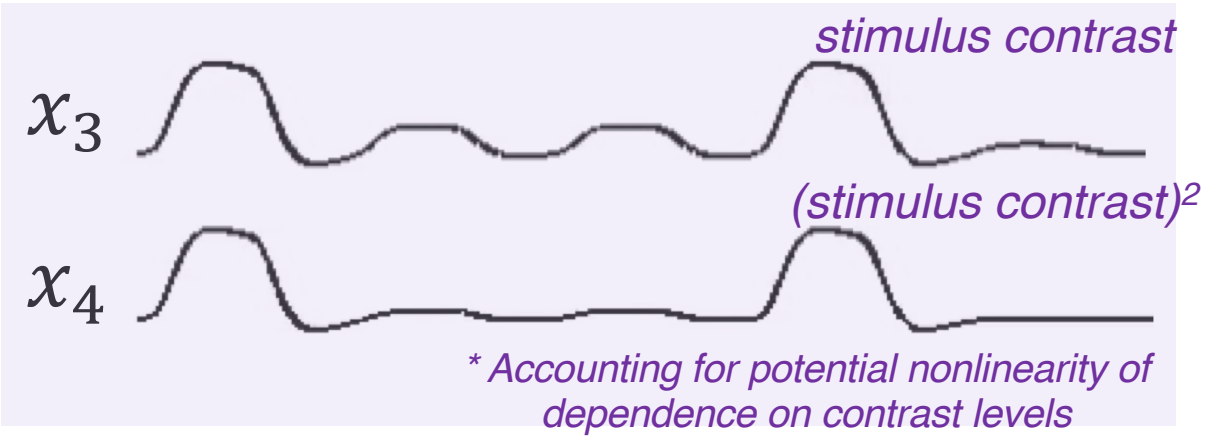
$\cdots$ Hypothesis testing

"not activated by the auditory task"

$H_0\colon\ \beta_1 = 0$

"not activated by the visual task"

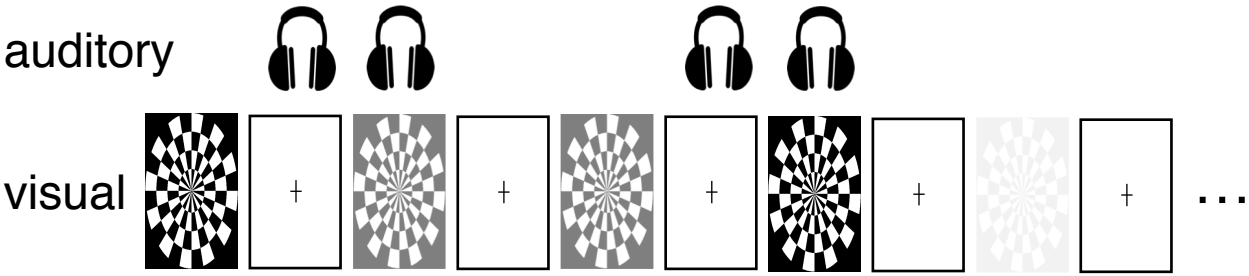
$H_0\colon\ \beta_2 = 0$

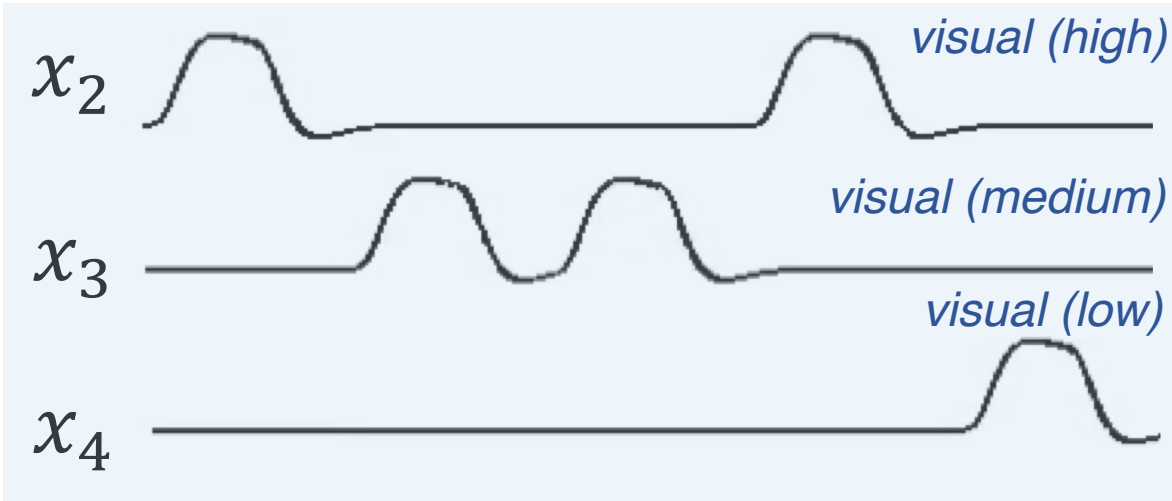


"no effects of visual contrasts"

$H_0\colon\ \beta_3 = \beta_4 = 0$

GLM applications – Design matrix is not unique





Hypothesis testing

"not activated by the auditory task"

$H_0\colon\ \beta_1 = 0$

"not activated by the flashing checkerboard"

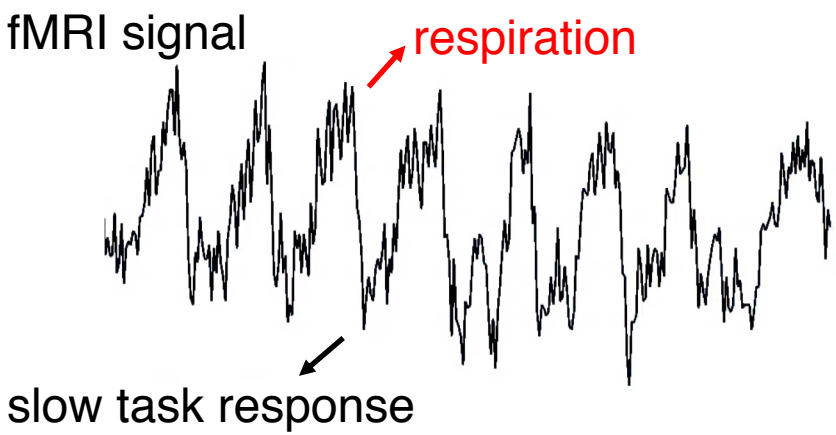
$H_0\colon\ (\beta_1 + \beta_2 + \beta_3)/3 = 0$

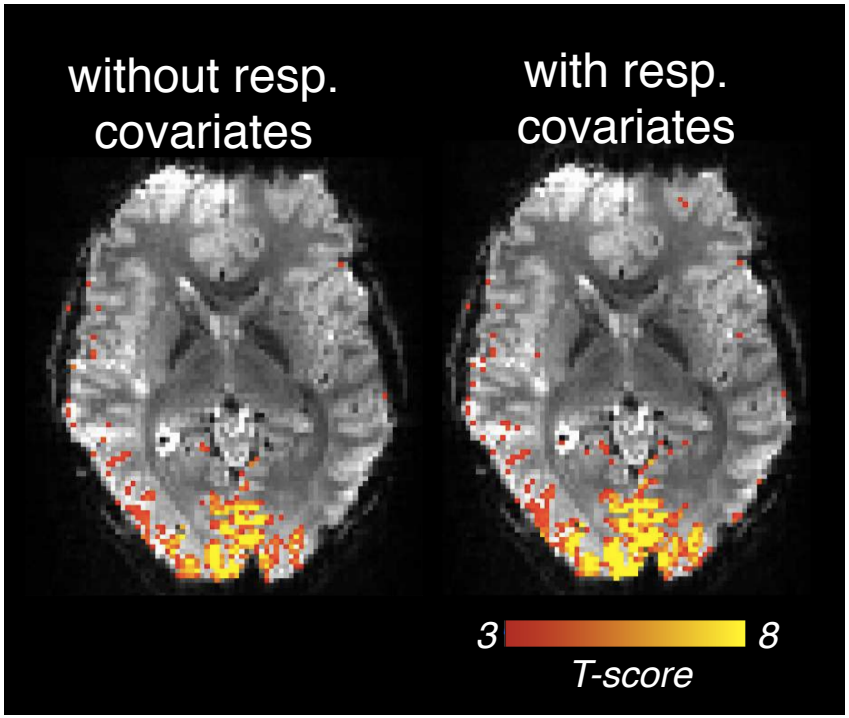
"not activated by any contrasts"

$H_0\colon\ \beta_1 = \beta_2 = \beta_3 = 0$

GLM cautions – incomplete design matrix

- Include all known covariates of interest if the scan is sufficiently long
  - reduce residual noise, and therefore the variability of summary statistics
  - increase *T or F* scores (if the reduction in degrees of freedom is minor)

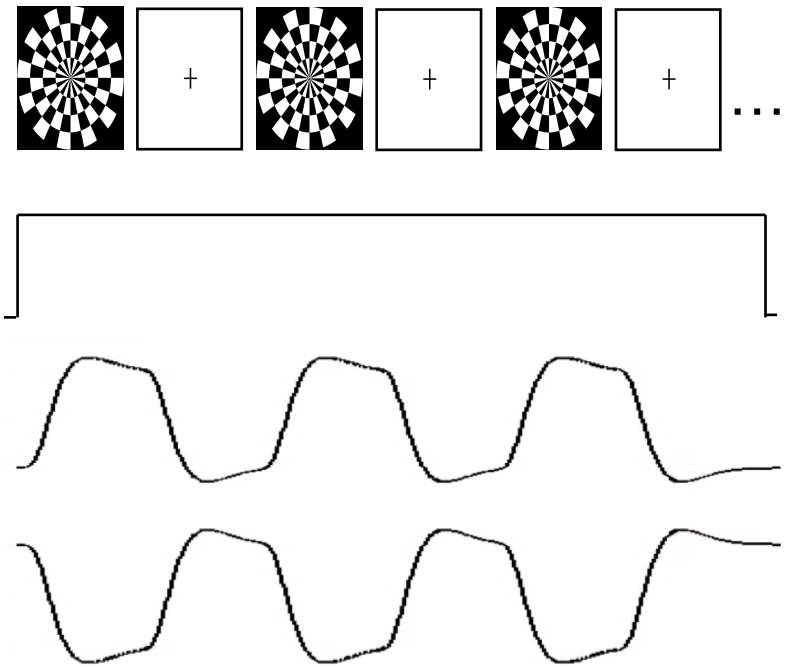




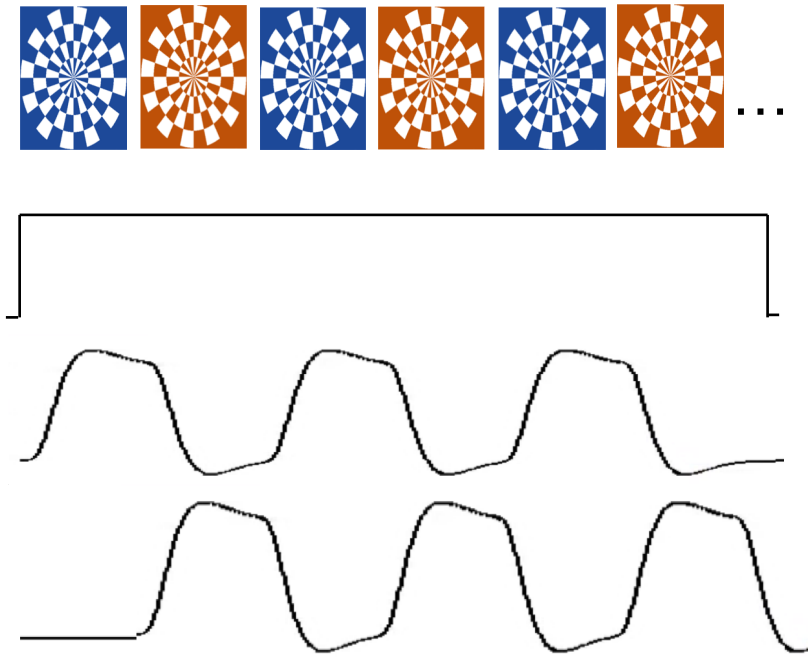
GLM cautions – over-modelling and collinearity

- Covariates should not depend on each other (i.e., the design matrix should be full rank)
- What is wrong with the following two designs?

Exp.1



Exp.2

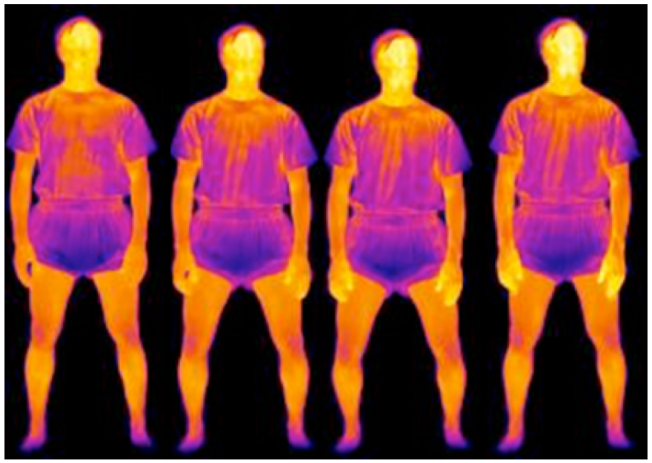


GLM cautions – over-modelling and collinearity

- If two covariates are highly correlated with each other, it is unlikely to identify “activation” associated with each particular covariate

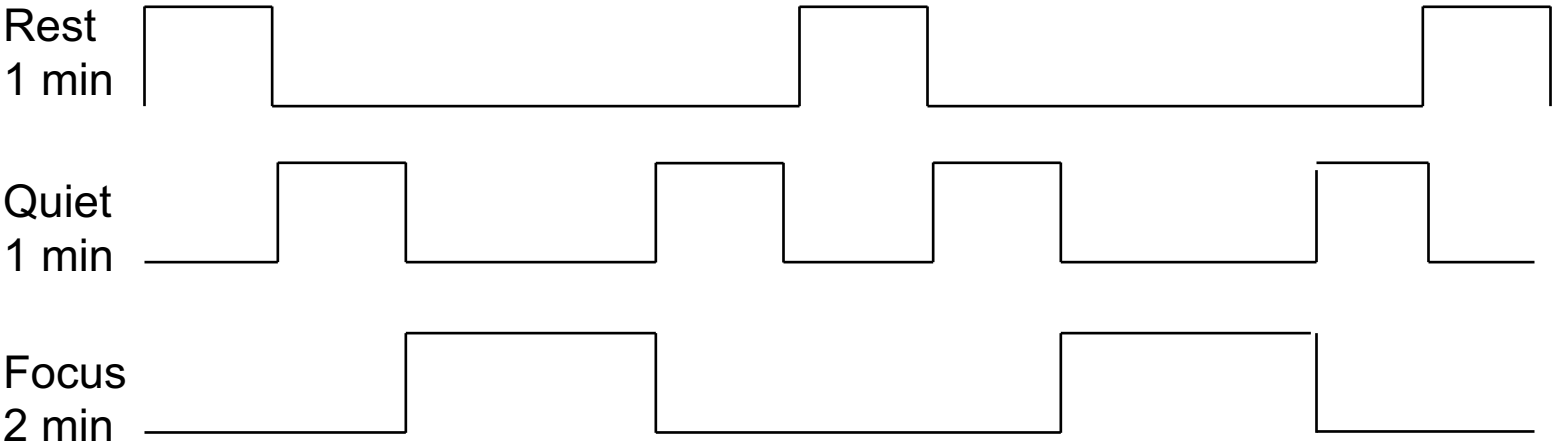
Tai Qi task: auditory cues

Temperature increase when focusing ‘Qi’ on hands



Time →

Cue: Rest - Quiet - Focus - Quiet - Rest - Quiet - Focus - Quiet - Rest

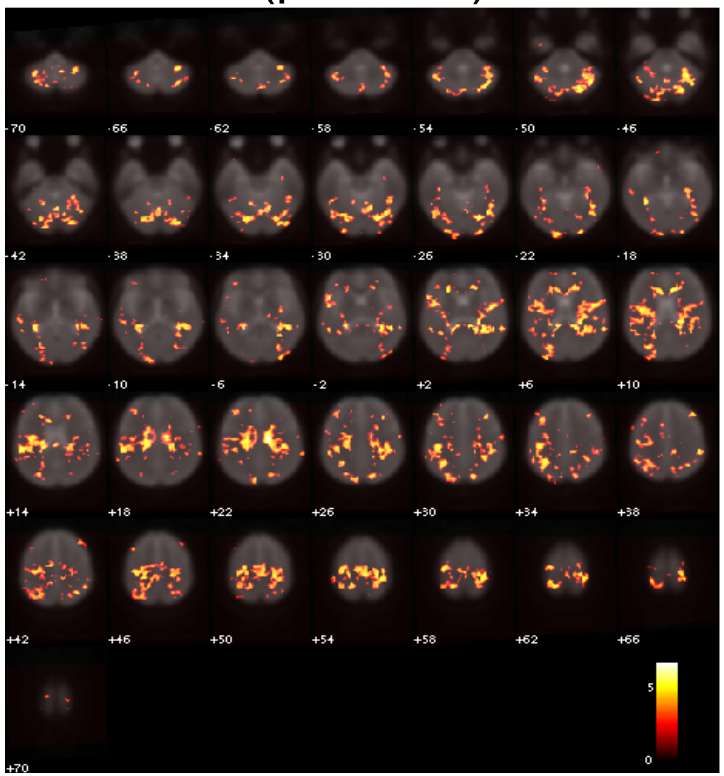


Slides courtesy by Catie Chang, Gary Glover (unpublished)

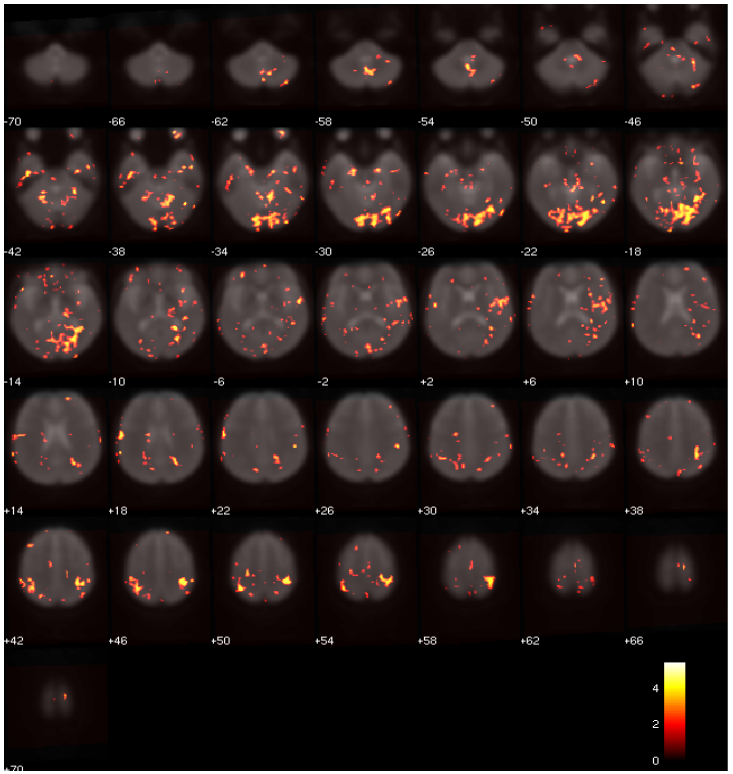
GLM cautions – over-modelling and collinearity

- If two covariates are highly correlated with each other, it is unlikely to identify “activation” associated with a particular covariate

Task: ‘Focus on Hands’ (p<0.001)



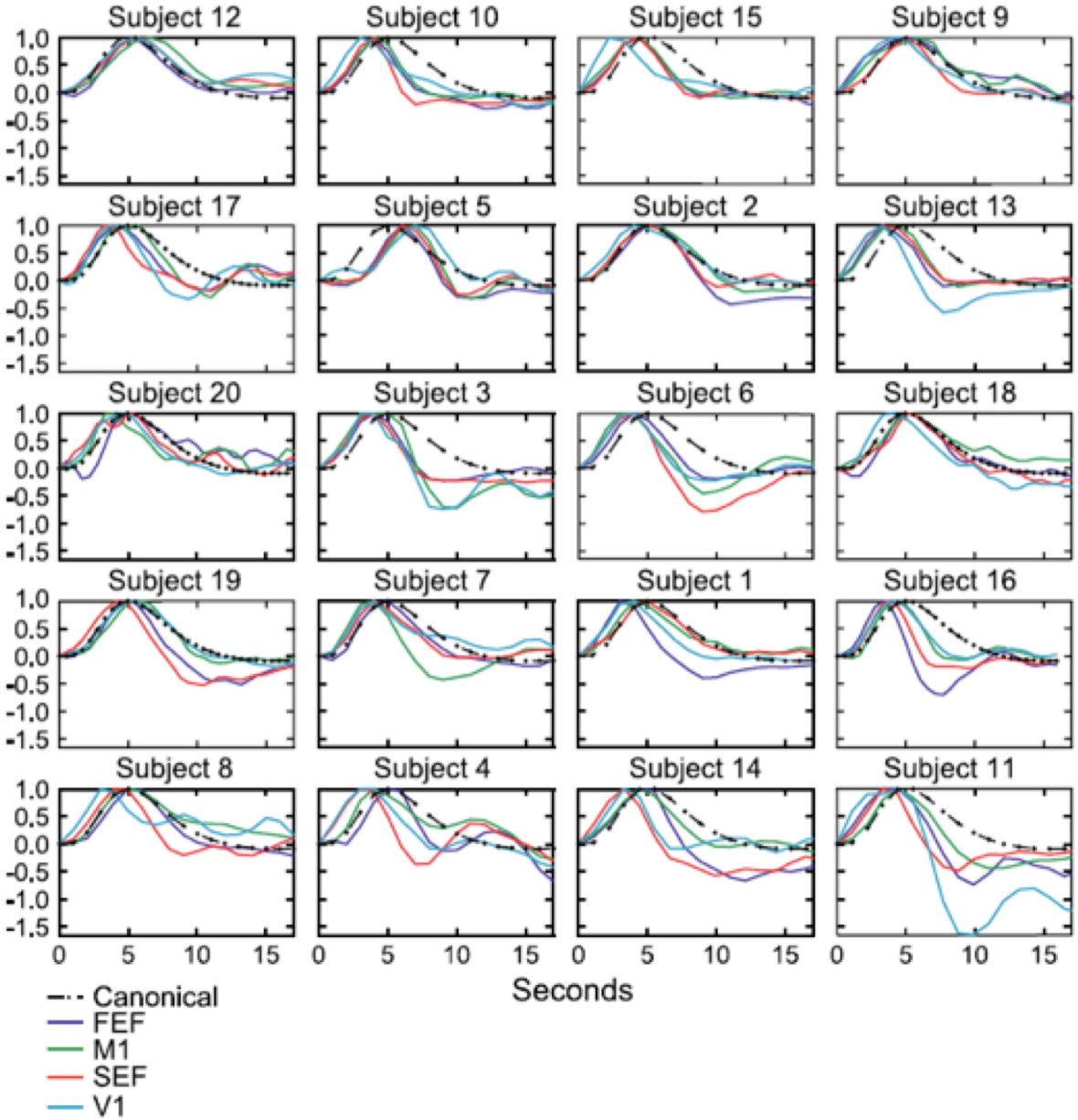
with respiration correction



Slides Courtesy by Catie Chang, Gary Glover (unpublished)

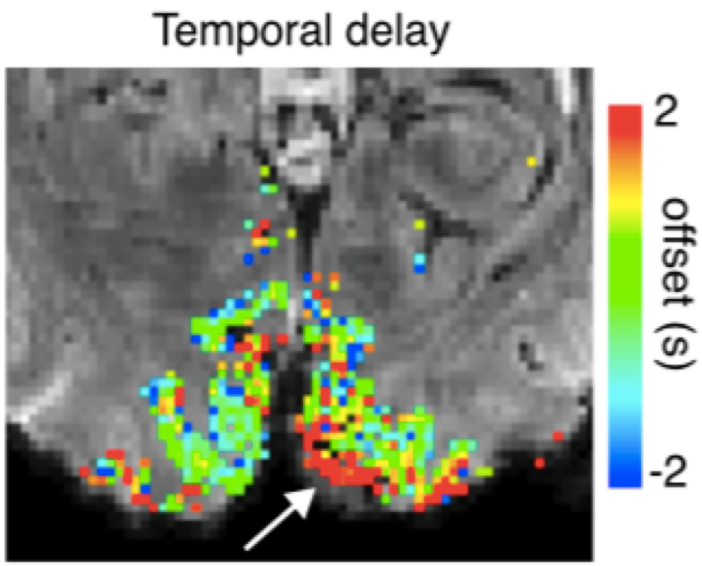
HRFs are highly variable across voxels/subjects

D.A. Handwerker et al. / NeuroImage 21 (2004) 1639–1651

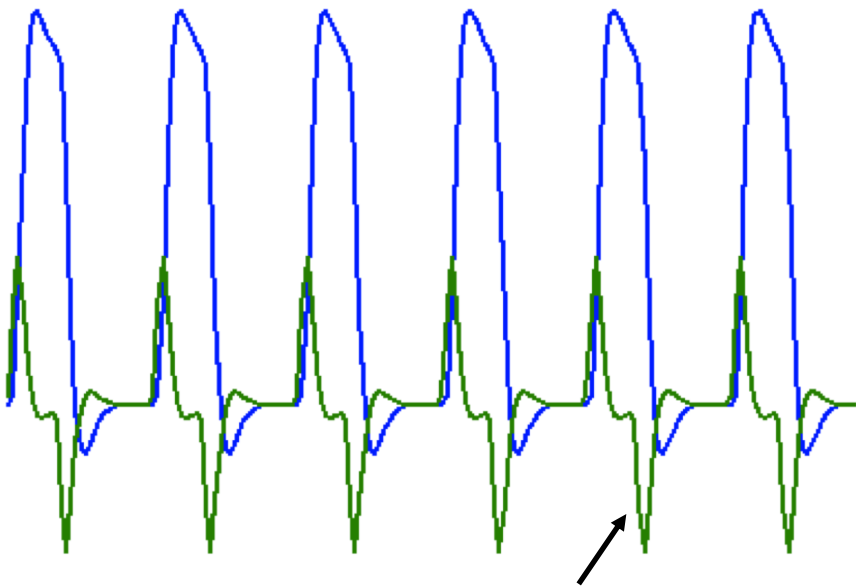


GLM cautions – accounting for HRF variability

- HRFs are highly variable across voxels/subjects
- Slow task paradigms filter out high-frequency HRF variability, so regional delays dominate



Lewis et al., 2018



Include an additional temporal derivative term

GLM cautions – noise serial correlations

- original assumptions of GLM

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathbb{N}\,(0,\, \sigma^2 I)$$

Ordinary least square (OLS) fitting yields the best linear unbiased estimator

$$\hat{\beta} = (X^T X)^{-1}(X^T y)$$

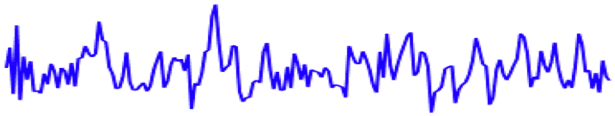
$$E(\hat{\beta}) = \beta$$

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

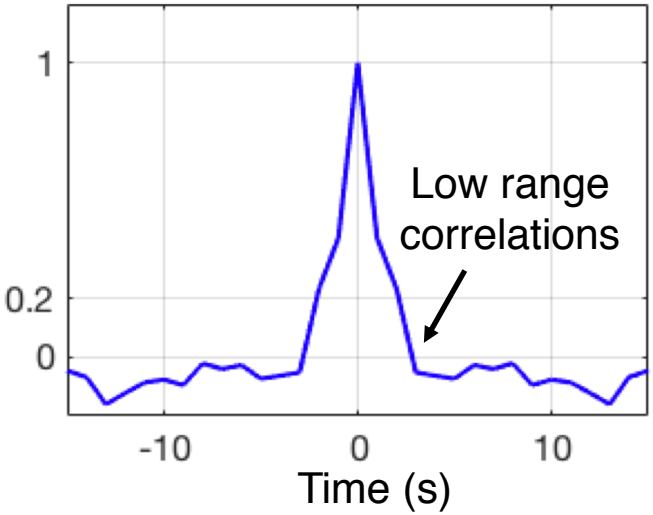
$$t(c\hat{\beta}) = \frac{c\hat{\beta}}{\sqrt{cVar(\hat{\beta})c^T}}$$

- fMRI noise are not white

Residuals post OLS



Autocorrelation



GLM cautions – accounting for noise serial correlations

- Pre-whitening: remove temporal autocorrelations in the residuals

$$y = X\beta + \varepsilon, \varepsilon \sim \mathbb{N}\,(0,\, V_i), K_i K_i^T = V_i$$

$$K_i^{-1} y = K_i^{-1} X\beta + K_i^{-1}\varepsilon$$
$$y_w = X_w \beta + \varepsilon_W$$

$$\varepsilon_W \sim \mathbb{N}\,(0,\, I)$$
$$E(\widehat{\beta_w}) = \beta$$
$$Var(\widehat{\boldsymbol{\beta}_W}) = (X^T V_i^{-1} X)^{-1}$$
$$t(c\widehat{\beta_w}) = \frac{c\widehat{\beta_w}}{\sqrt{c(X^T V_i^{-1} X)^{-1} c^T}}$$

- Alternatives

- e.g., pre-coloring (temporal filtering), S is a full rank matrix representation of the data preparation procedure

$$Sy = SX\beta + S\varepsilon$$
$$y_S = X_S \beta + \varepsilon_S$$

$$E(\widehat{\beta_S}) = \beta$$
$$Var(\widehat{\beta_S}) = pinv(SX)\, \mathrm{SV_iS^T}\, pinv(SX)^{\boldsymbol{T}}$$
$$t(c\widehat{\beta_S}) = \frac{c\widehat{\beta_S}}{\sqrt{cVar(\widehat{\beta_S})c^T}}$$

- Pre-whitening is the most efficient (highest t-scores) approach, but also more vulnerable to model bias (inaccurate modelling of serial correlations)[1]

[1] Friston et al., 2000

Summary I: single-subject-level analysis

- GLM provides a flexible framework to make statistical inferences for versatile task designs

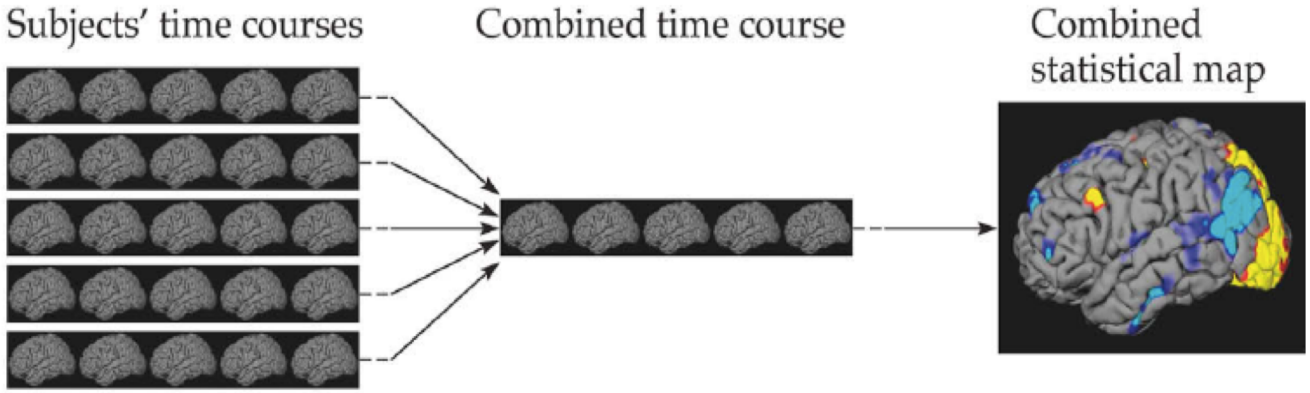
- GLM designs can be optimized by
  - Including all known covariates of interest if the scan is sufficiently long
  - avoiding over-modelling and collinearity
  - accounting for HRF variability

- Noise serial correlation should be considered in statistical inferences of GLM
  - Pre-whitening (most common and efficient approach)
  - Results of statistical tests may differ depending on how serial correlation is modelled and accommodated

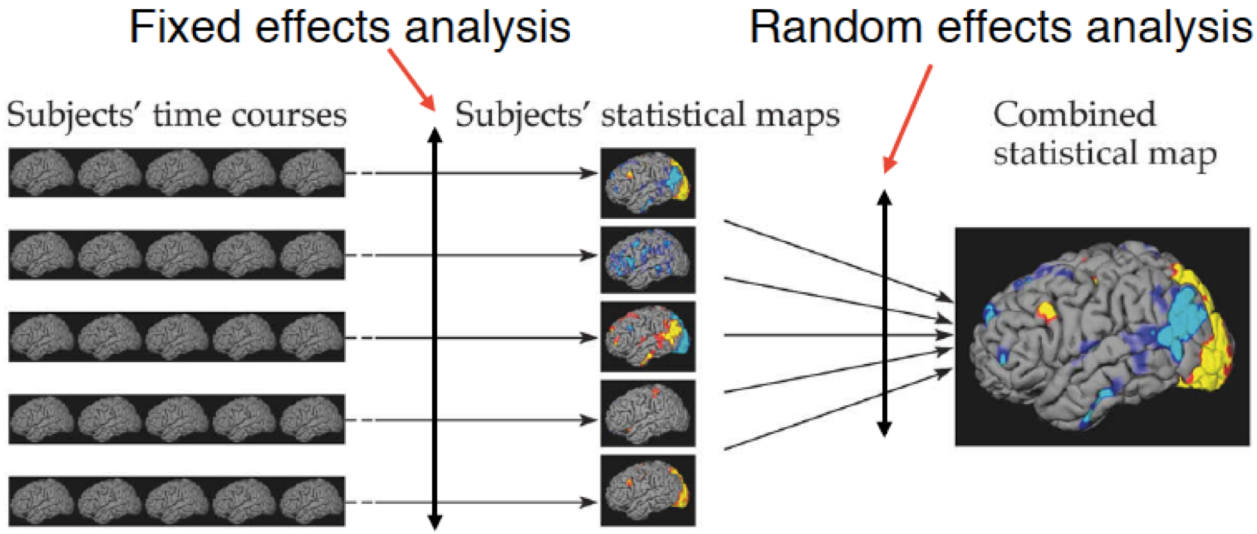
2nd-level Analysis:
Group-level Analysis

Integrating results across multiple sessions/subjects

- Fixed effects: do not consider inter-session variability of summary statistics



- Random effects: consider inter-session variability of summary statistics



Fixed vs. Random effects

Fixed effects

- a small number of sessions (repeated scans of a subject)
- interested in group effects (modelling the mean) but not statistical inferences

Random effects

- necessary for applying inference to total population (modelling the variance)
  - how significant an effect is present in the total population

Group-level inference under the GLM framework: **one-sample t-test**

- Example: identify voxels activated during a visual task (20 subjects)

GLM formation:

$$y = \beta_0 + \varepsilon$$

20 subjects

$$\begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ \vdots \\ G_{20} \end{bmatrix} \qquad \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

Inter-subject variability

1st-level summary statistics (e.g., percent signal change, contrast value)

Overall mean

Hypothesis testing:

"No voxels are robustly activated across all subjects"

$$H_0\colon\ \beta_0 = 0$$

Group-level inference under the GLM framework: **two-sample t-test**

- Gender differences in logistic reasoning (10 subjects per gender)

GLM formation:

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\begin{array}{r} \textit{Male} \\ \\ \\ \\ \textit{Female} \\ \\ \\ \\ \end{array} \begin{bmatrix} G_1^M \\ G_2^M \\ \vdots \\ G_{10}^M \\ G_1^F \\ G_2^F \\ \vdots \\ G_{10}^F \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad \varepsilon$$

1st-level summary statistics | *Mean across male subjects* | *Mean across female subjects* | *Inter-subject variability*

Hypothesis testing:

"No voxels exhibit significant gender differences"

$$H_0\colon\ \beta_1 - \beta_2 = 0$$

Group-level inference under the GLM framework: **paired t-test**

- Influence of a certain medical treatment on brain activity (5 subjects, pre- and post-treatment)

GLM formation:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

$$\begin{bmatrix} G_1^{pre} \\ G_1^{post} \\ G_2^{pre} \\ G_2^{post} \\ G_3^{pre} \\ G_3^{post} \\ G_4^{pre} \\ G_4^{post} \end{bmatrix} \quad \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

1st-level summary statistics

Mean of (Pre-Post)

Effects unique to each subject

Hypothesis testing:

"No voxels demonstrate the effects of treatments"

$$H_0\colon\ \beta_1 = 0$$

Group-level inference under the GLM framework: **ANOVA**

- **Age** (young vs. old) and **Gender** effects in working memory (WM) activation

GLM formation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

	1st-level summary statistics	Overall WM effect	Gender effect	Age effect	Interaction between Gender and Age
Male Young	G_y^M	1	1	1	1
	G_y^M	1	1	1	1
Male Old	G_o^M	1	1	−1	−1
	G_o^M	1	1	−1	−1
Female Young	G_y^F	1	−1	1	−1
	G_y^F	1	−1	1	−1
Female Old	G_o^F	1	−1	−1	1
	G_o^F	1	−1	−1	1

Hypothesis testing:

"No voxels are activated by the WM task"

$$H_0\colon\ \beta_0 = 0$$

"No voxels show gender effects"

$$H_0\colon\ \beta_1 = 0$$

"No voxels show age effects"

$$H_0\colon\ \beta_2 = 0$$

"Gender and age effects are independent"

$$H_0\colon\ \beta_3 = 0$$

Summary II: group-level analysis

- Fixed effects vs. random effects
  - Depend on the sample size and your interest (group effects or inference)
  - Two-level analyses assuming mixed effects: single subject-level summary statistics (fixed effects), group-level inference (random effects)
- Group-level inference can be flexibly implemented under the GLM framework
  - One-sample, two-sample, paired t tests, ANOVA

Correcting for Multiple Comparisons

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

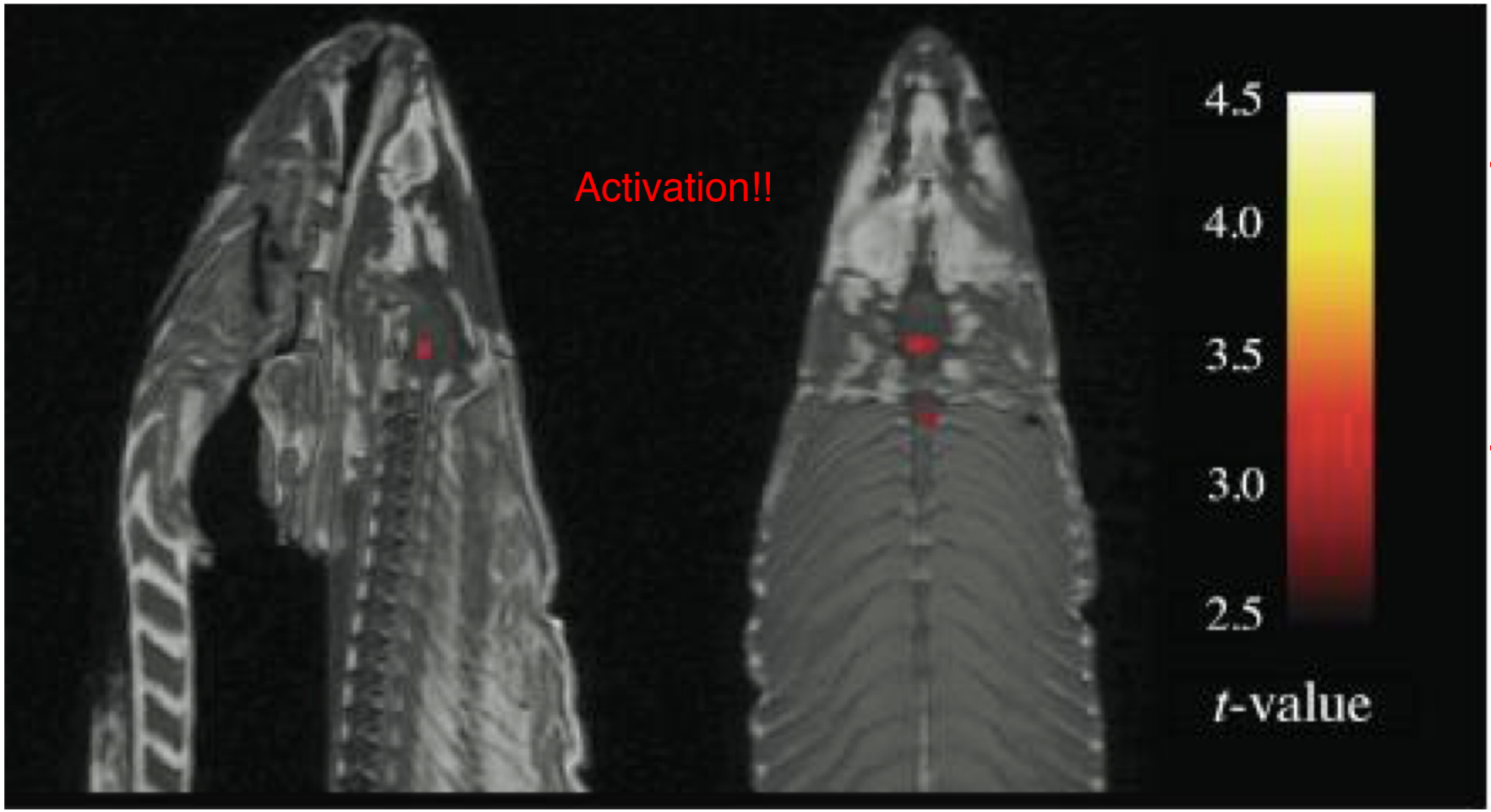
Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

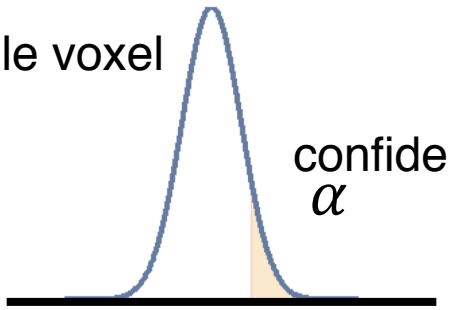
[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY; [3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH



Problem with multiple comparisons

Pr (at least a voxel is considered active I H_0)

single voxel



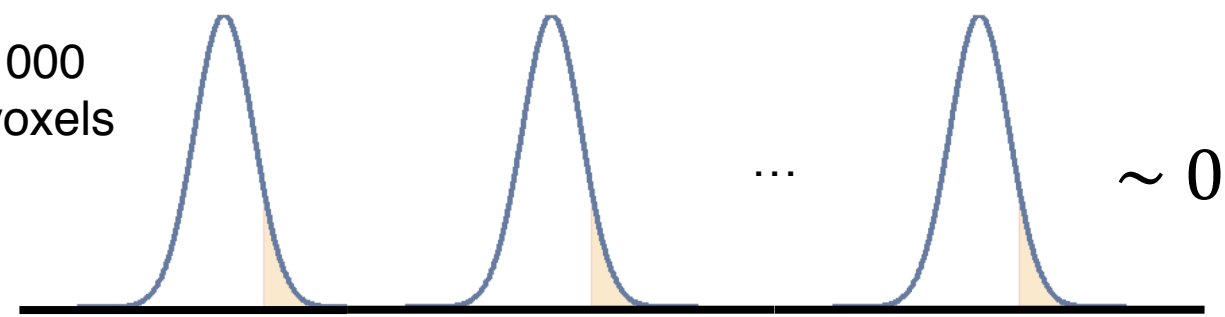
confidence level α

$0.01\ (\alpha)$

Two voxels

$1 - 0.99^2$

1000 voxels

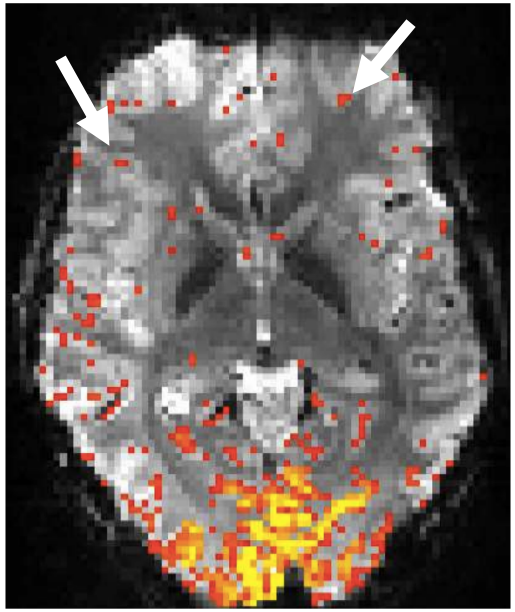


~ 0.999959

How to correct for multiple comparisons

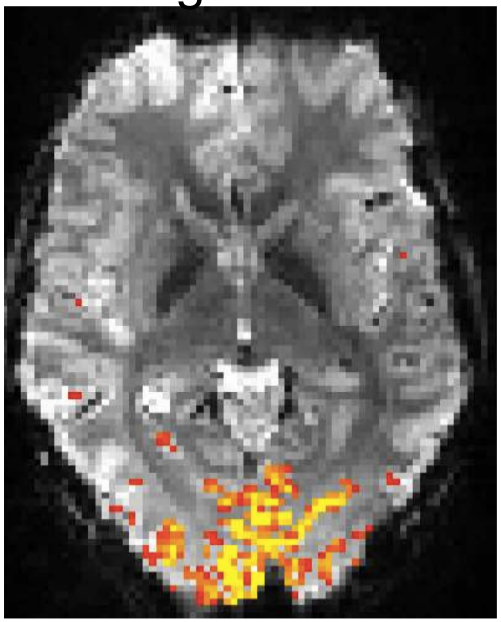
- Reducing Type I error (false positives)

No correction



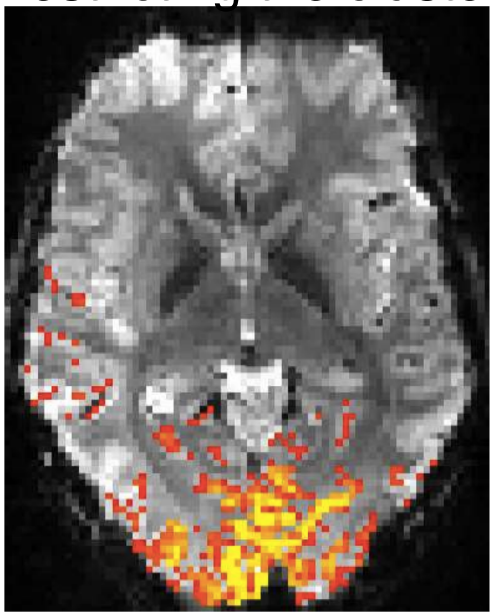
$p < 0.01$

Increasing the threshold



$p < 0.0001$

Restricting the cluster size



$p < 0.01$

- Family wise error (FWE): controlling the probability of occurrence of falsely rejected tests

Single-threshold method: Bonferroni correction

- Under the null condition, the probability of rejecting a single test is p, then for N independent tests:

$$Pr\{\text{at least one test is significant}\} = 1 - (1-p)^N \approx Np$$

$$p < \alpha_{bon} = \frac{\alpha_{FWE}}{N}$$

- Easy to implement (simply count the # of total voxels)
- No prior assumptions about data structures

- Assume that all tests are independent
  - Very conservative if the fMRI data is intrinsically smooth
  - More prone to type II errors (less powerful)

Single-threshold Method: Random Field Theory (RFT)

- Incorporating the degree of smoothness in the data to control FWE
  - Euler characteristics (EC): the number of clusters above a given threshold
  - Find the significance threshold so that $E(EC) < \alpha_{FWE}$

$$\mathrm{E(EC)} = \mathrm{R} \times \frac{(4\ln(2))^{\frac{3}{2}}}{(2\pi)^2} e^{-\frac{t^2}{2}} (t^2 - 1) < \alpha_{FWE}$$

  - an estimate of spatial smoothness: $FWHM_x \times FWHM_y \times FWHM_z$ (the virtual voxel size, resel)
  - ~ # of independent tests: $R = V / FWHM_x \times FWHM_y \times FWHM_z$

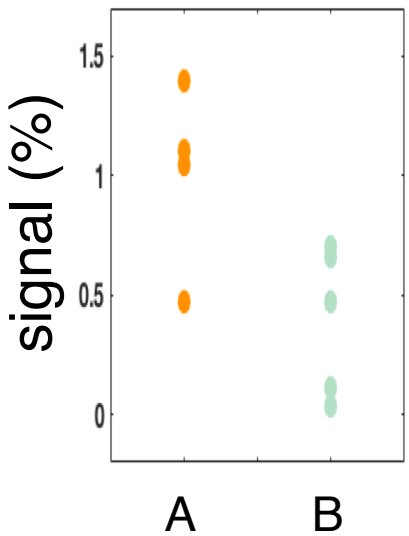
- Need prior knowledge of spatially-invariant smoothness

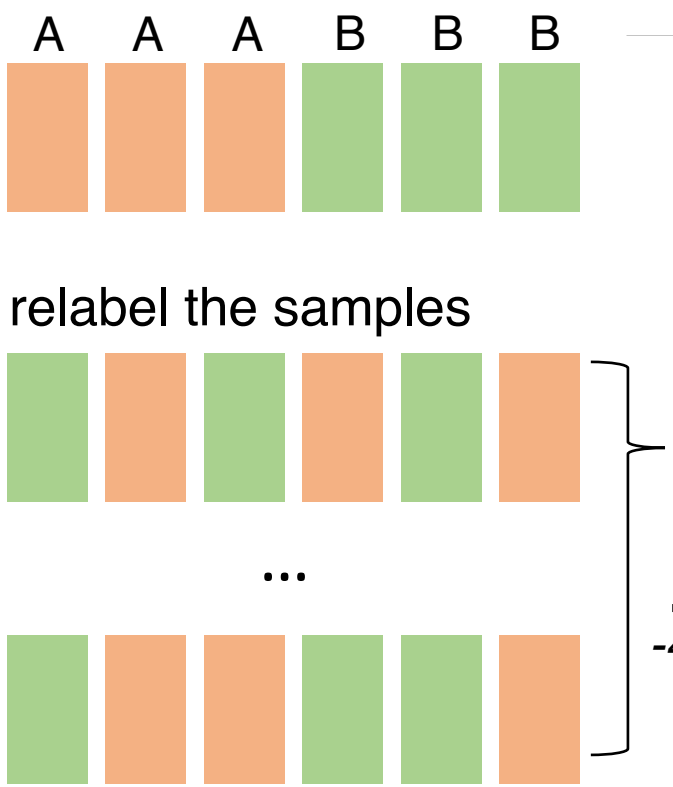
Single-threshold method: nonparametric permutation test

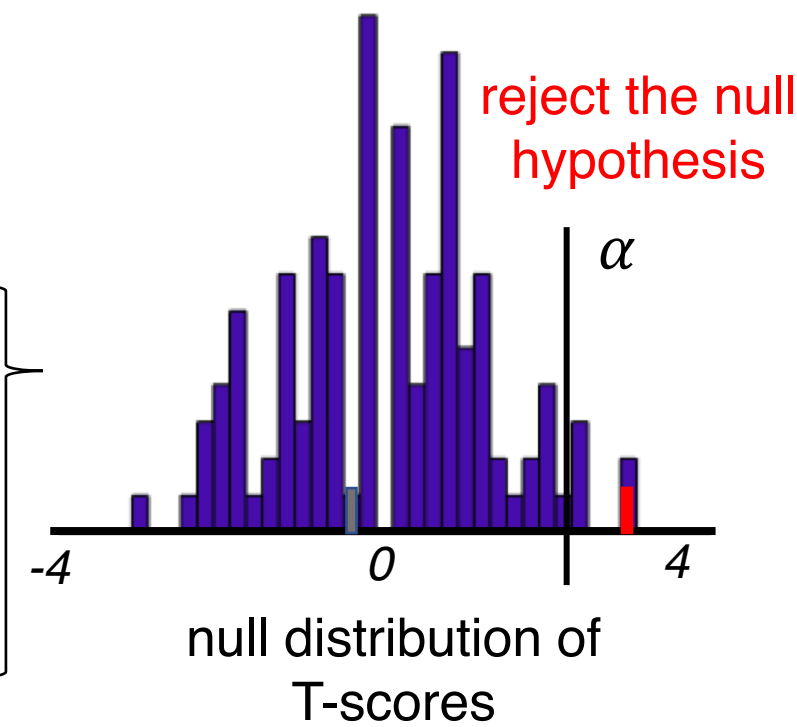
- No prior assumptions about the data structure
- Construct the null distribution by reshuffling the labels of samples

Single-voxel-level

Toy example

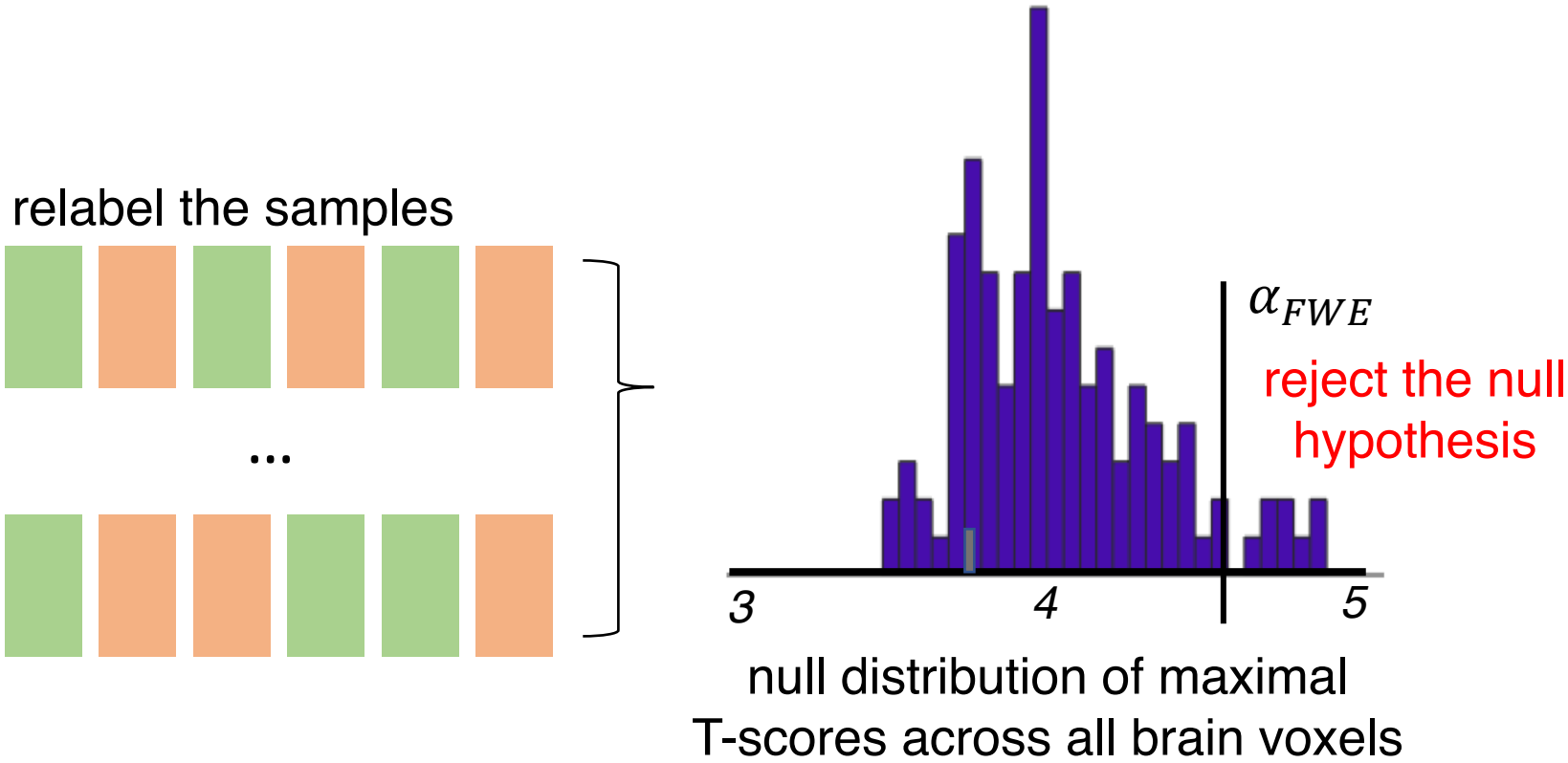




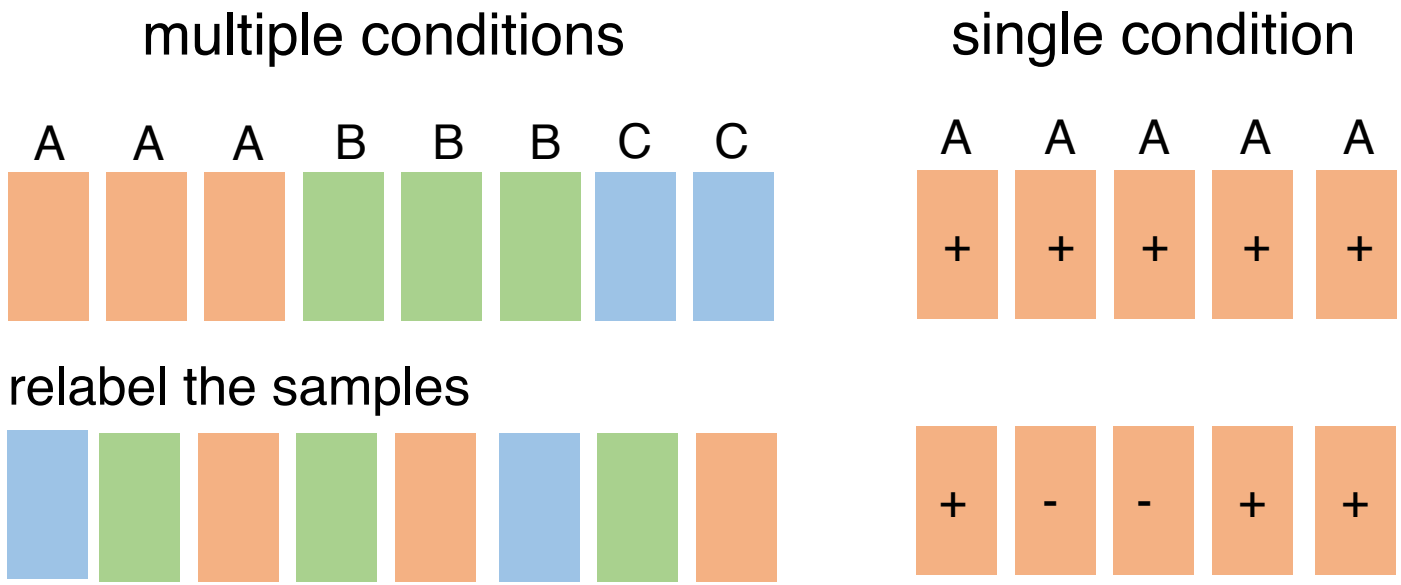


Single-threshold method: nonparametric permutation test

- Use *maximal statistic* to correct for multiple comparisons
- After each reshuffling procedure, take the maximal T-score across all brain voxels to establish the null distribution for FWE correction



Single-threshold method: nonparametric permutation test



Single-threshold method: nonparametric permutation test

- Not suitable for small sample size studies

  - Not enough permutations to establish the null distribution (e.g., at least 20 permutations to find the threshold for $\alpha_{FWE} = 0.05$)

- More conservative than equivalent parametric approaches

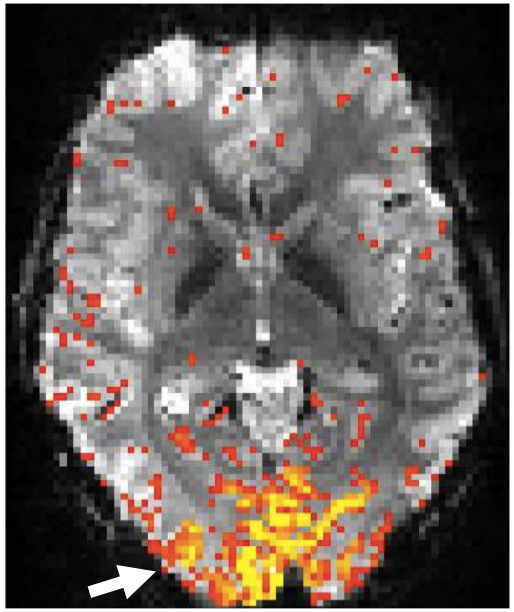
  - Assumptions of parametric approaches provide additional information that nonparametric approach must "discover"[1]

  - Become more powerful as the # of relabeling processes increases

- Higher computational load

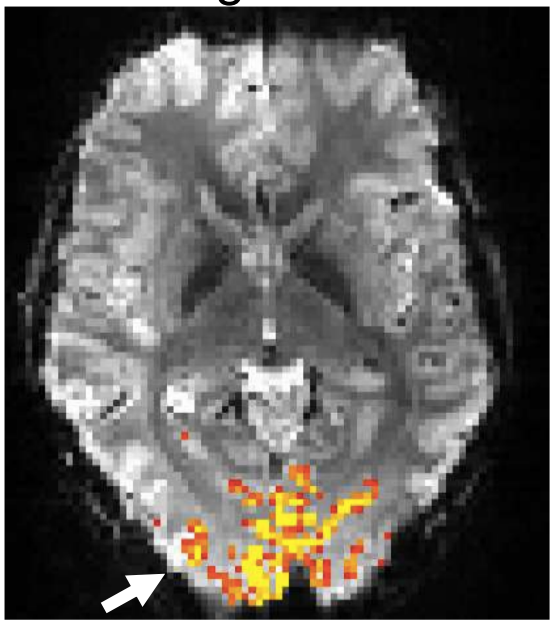
[1] Nichols et al., 2001

No correction



$p < 0.01$

Increasing the threshold



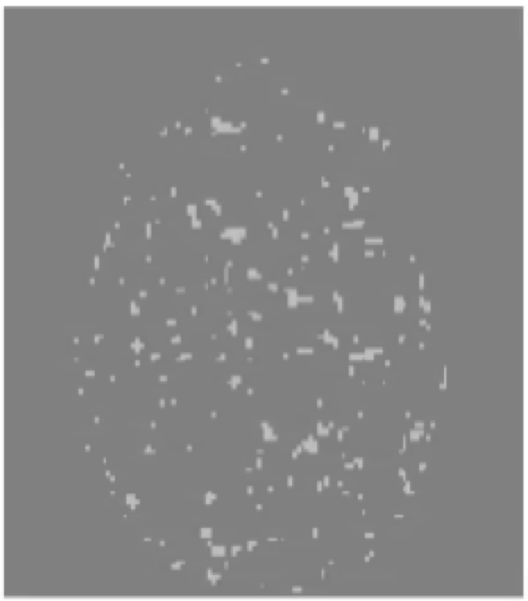
RFT, $\alpha_{FWE} = 0.01$

We can harness the spatial information to correct for multiple comparisons

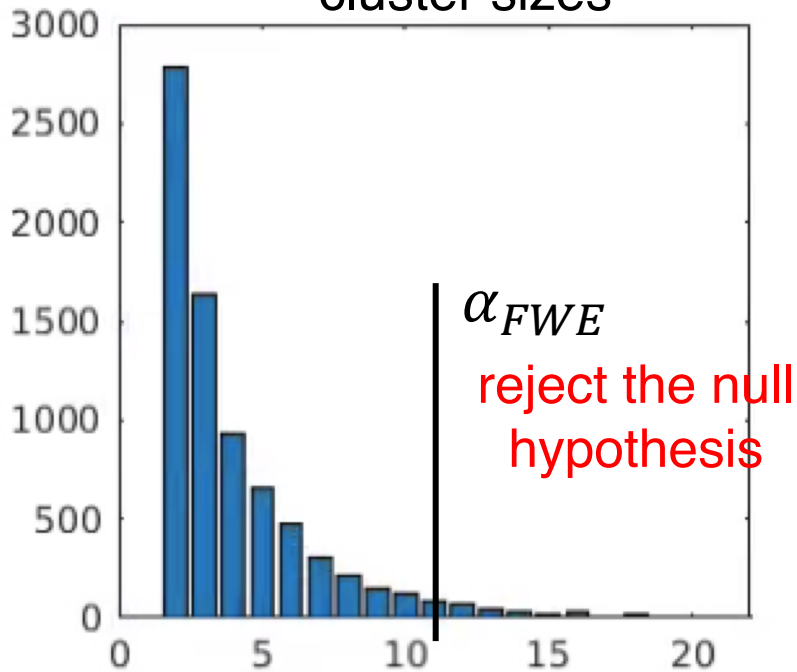
Monte Carlo Cluster-level Inference

- The probability of finding at least one above-threshold cluster under the null condition (no activation)

Thresholded, $p < 0.01$



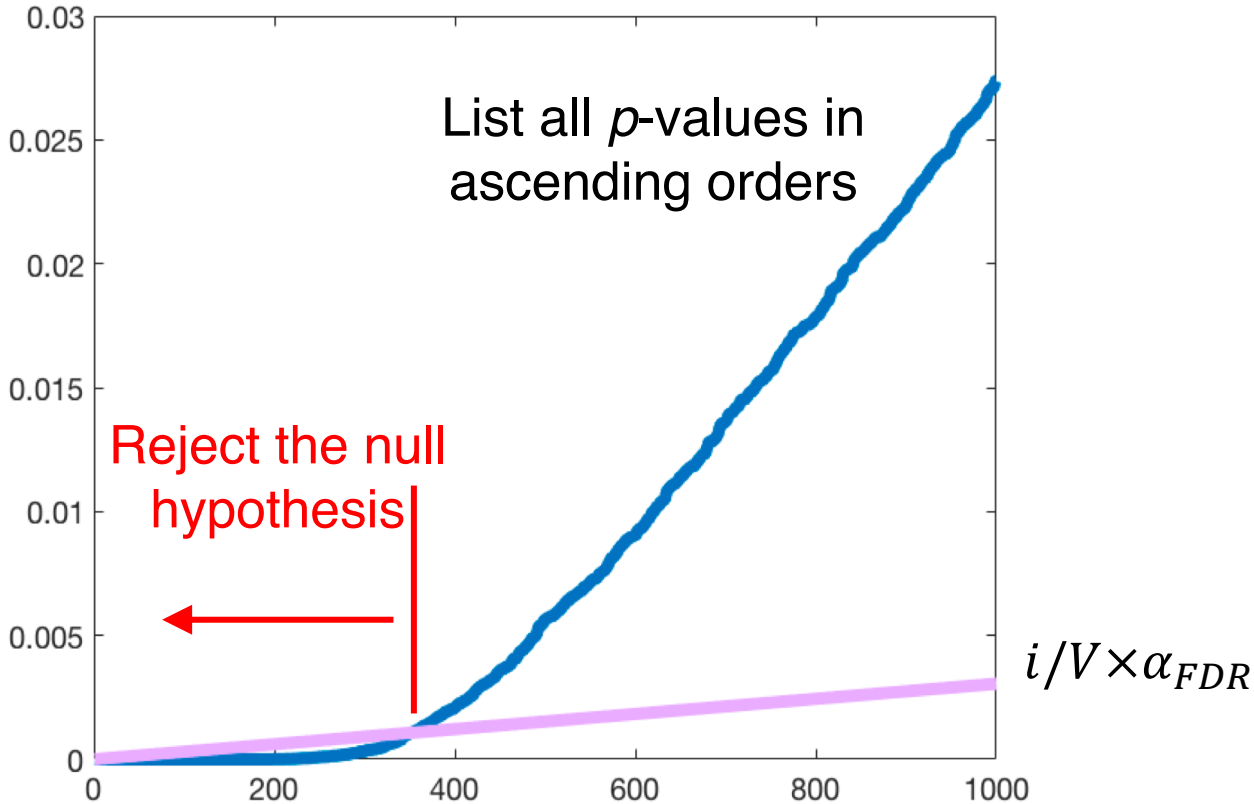
Null distribution of above-threshold cluster sizes



- Generally more powerful than single-threshold approach
- Better sensitivity, but poorer spatial specificity

Single-threshold method: False Discovery Rate

- Controlling the fraction of false positives in multiple comparisons (generally less conservative than FWE)
- V voxels, with α_{FDR}



Summary III: Correcting for Multiple Comparisons

- Given the large size of brain voxels, it is necessary to address the multiple comparison problem.

- "General" comments on different correction methods

  - Cluster-wise inference is more powerful than single threshold method, but looses spatial specificity

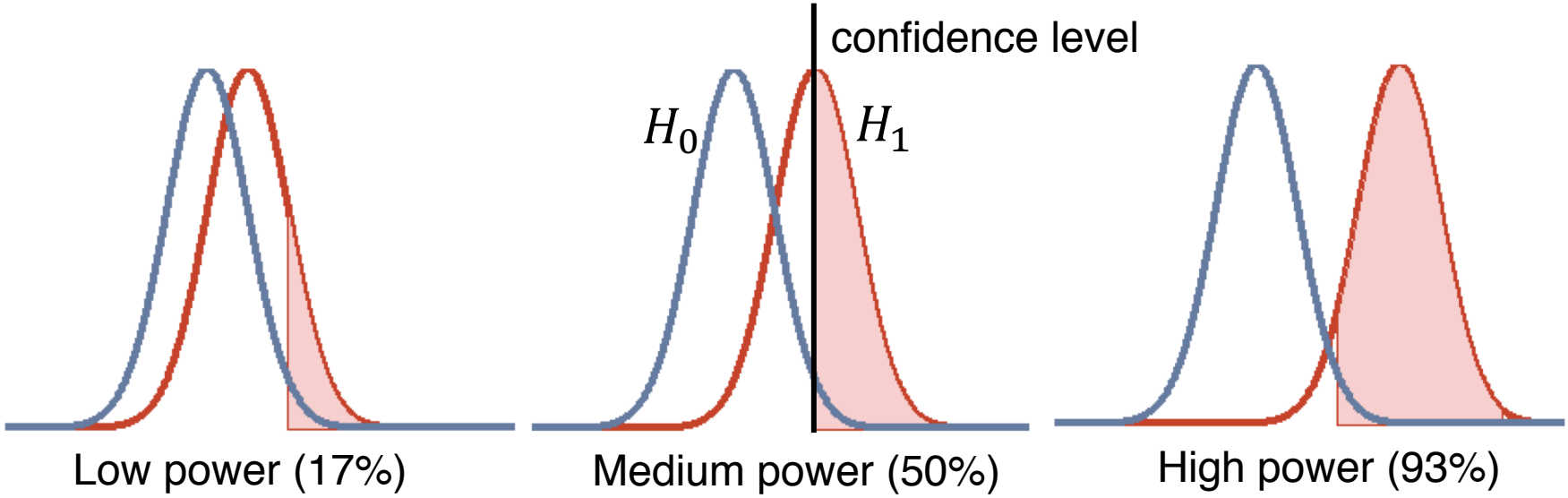
  - Nonparametric permutation tests make no prior assumptions on the data structure, but is less powerful compared to alternative parametric methods (if the assumption is correct)

  - FDR is less conservative compared to FWE

Power Analysis

Statistical power

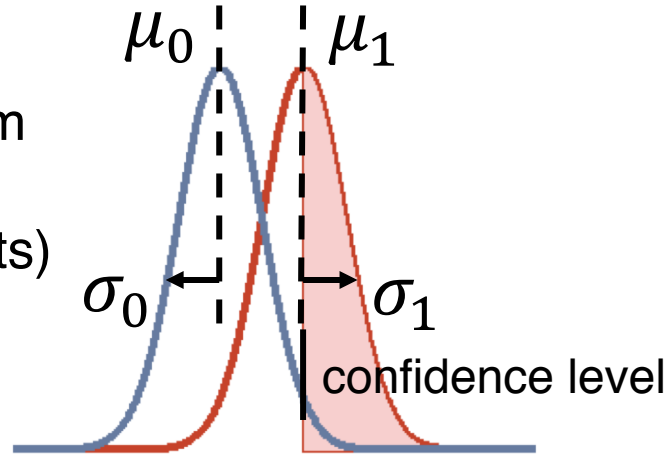
- Power = Pr (reject H_0 I H_1 is true)



- Power analysis is most commonly used for grant writing (how many subjects are needed for a particular study?)

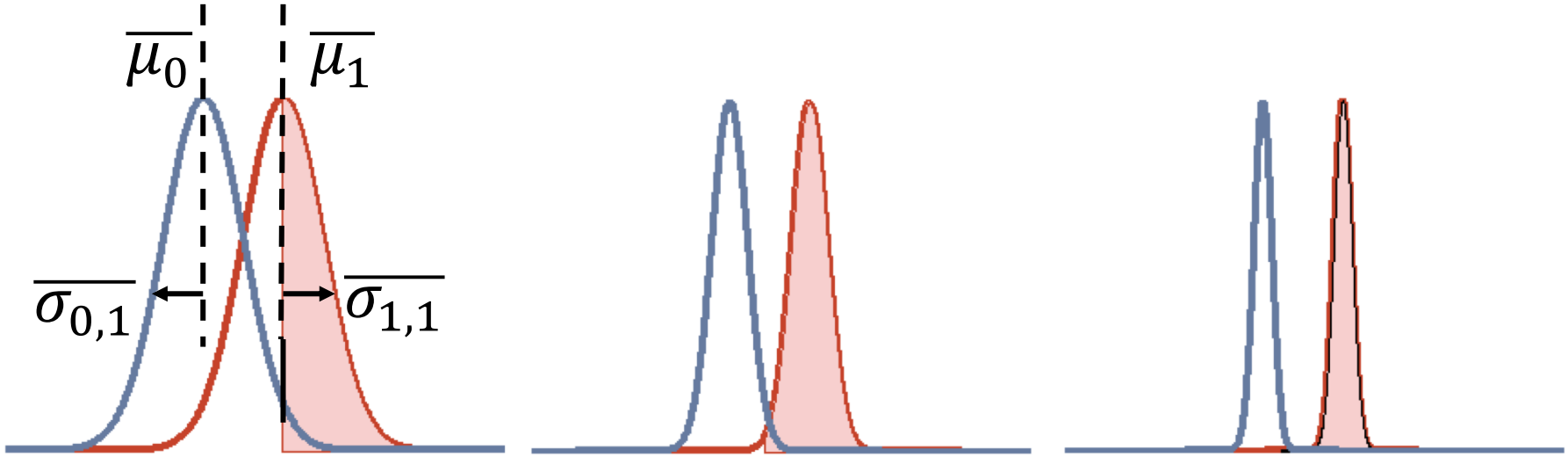
Power Analysis

estimated from
pilot data
(2nd-level results)



The more subjects,
the higher power

$$\overline{\sigma_{0/1,N}} = \frac{\sigma_{0/1}}{\sqrt{N}}$$



N = 1 (50%)

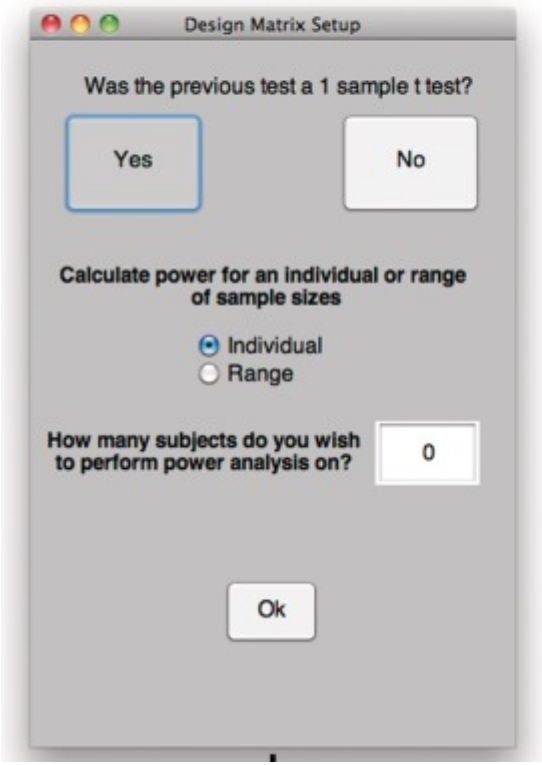
N = 5 (99.4%)

N = 20 (100%)

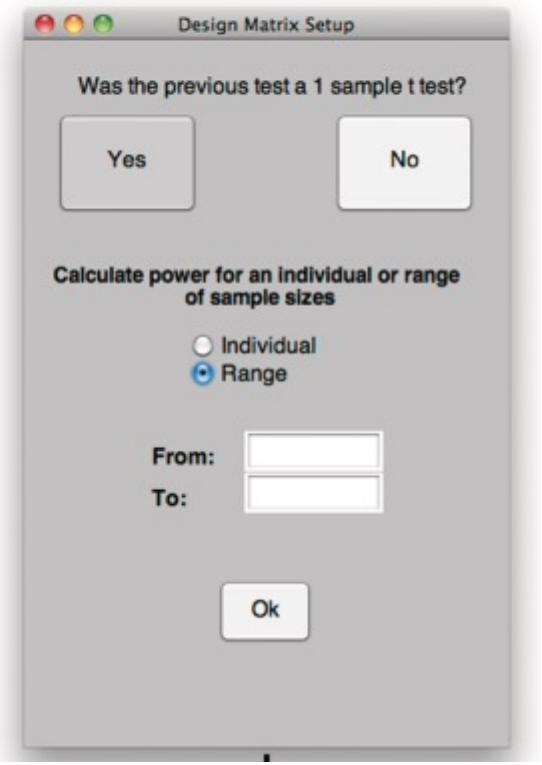
Power Analysis

- An automated toolbox http://fmripower.org[1]
- Takes the group-level design matrix from SPM (contrast '.nii' files) and FSL ('cope.feat').

Individual sample size option



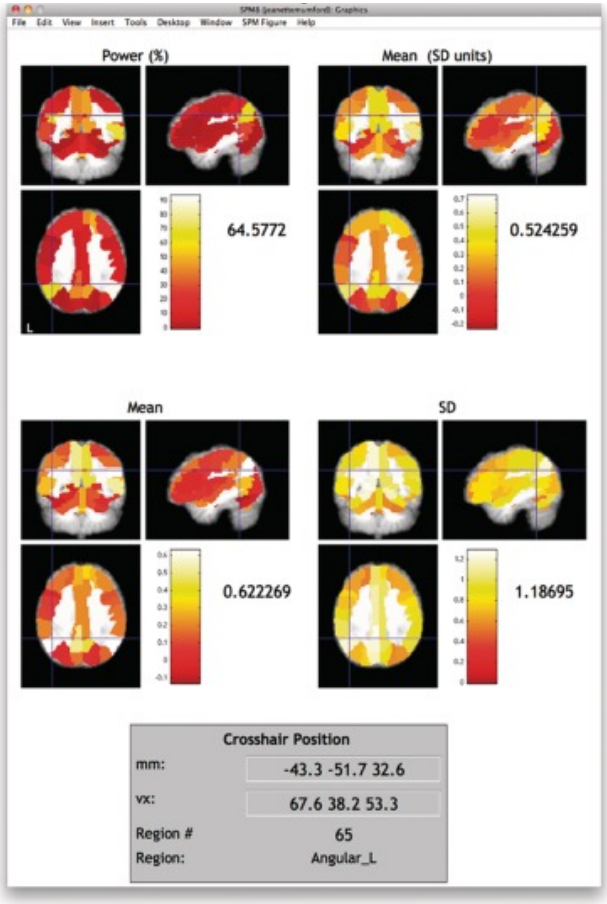
Range of sample sizes option

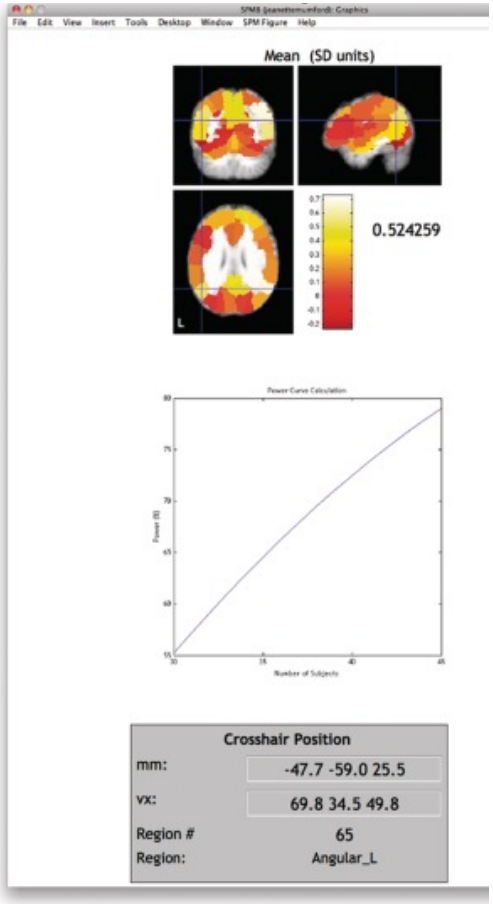


[1] Mumford et al., 2012

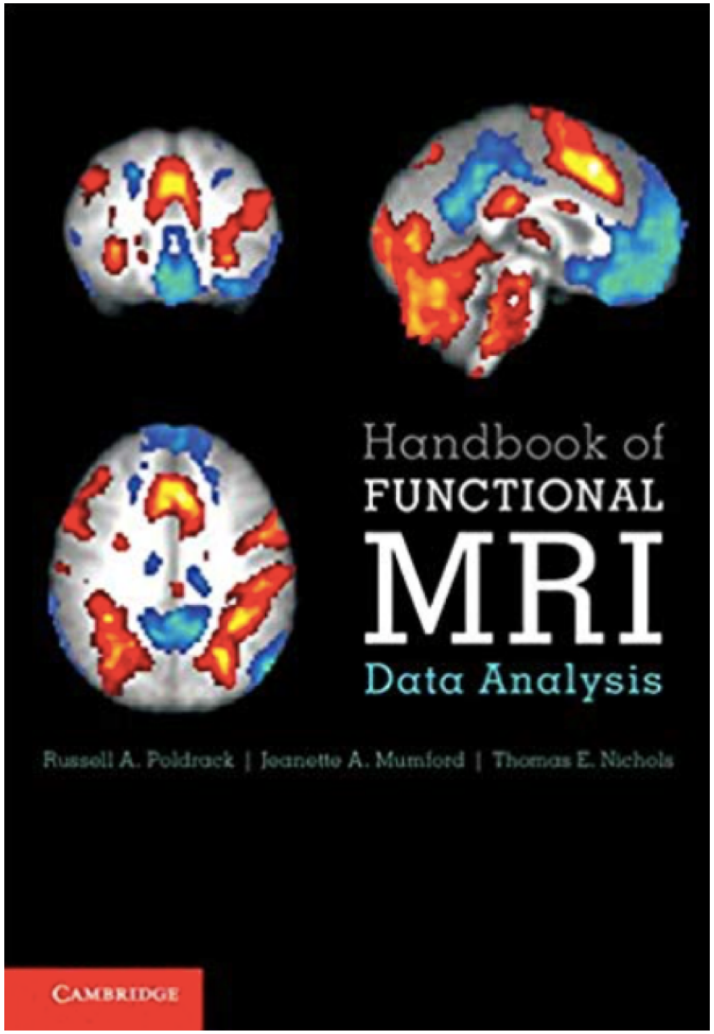
Power Analysis

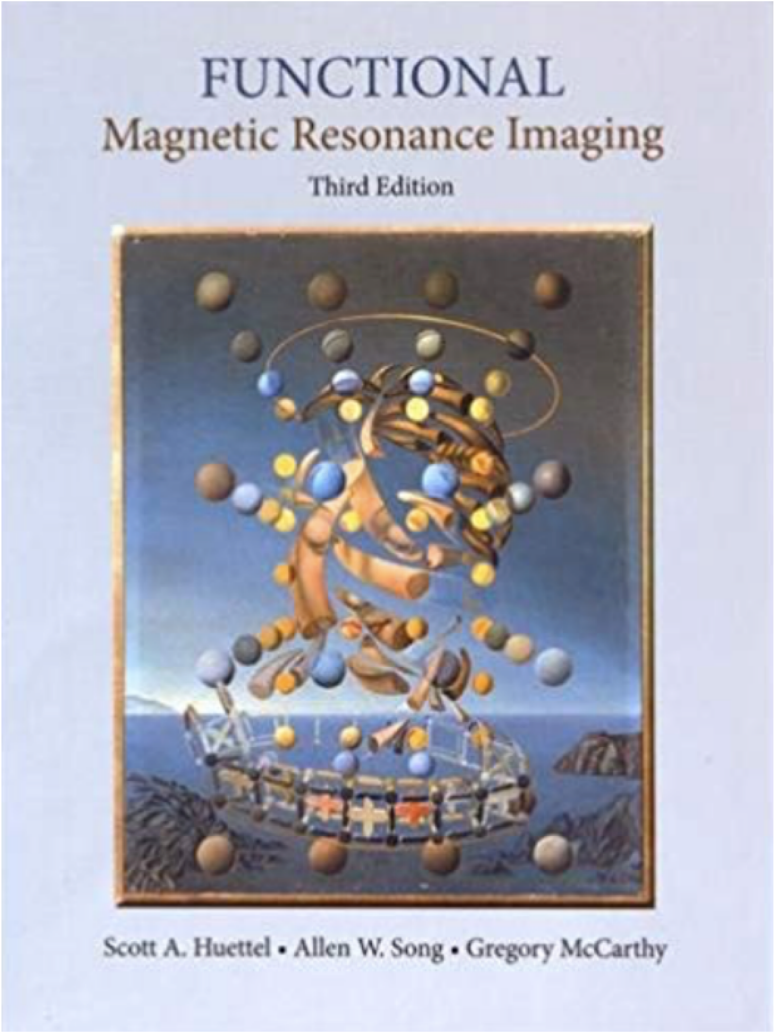
- An automated toolbox http://fmripower.org
- Specify test type, atlas or user-identified masks





References


Handbook of
FUNCTIONAL
MRI
Data Analysis
Russell A. Poldrack | Jeanette A. Mumford | Thomas E. Nichols
CAMBRIDGE


FUNCTIONAL
Magnetic Resonance Imaging
Third Edition
Scott A. Huettel • Allen W. Song • Gregory McCarthy

Thanks & questions