Psych 252

Statistical Applications for Neuroimaging

Justin Gardner

P-values

P-values are overrated

Statistical analyses are really models

Models have assumptions

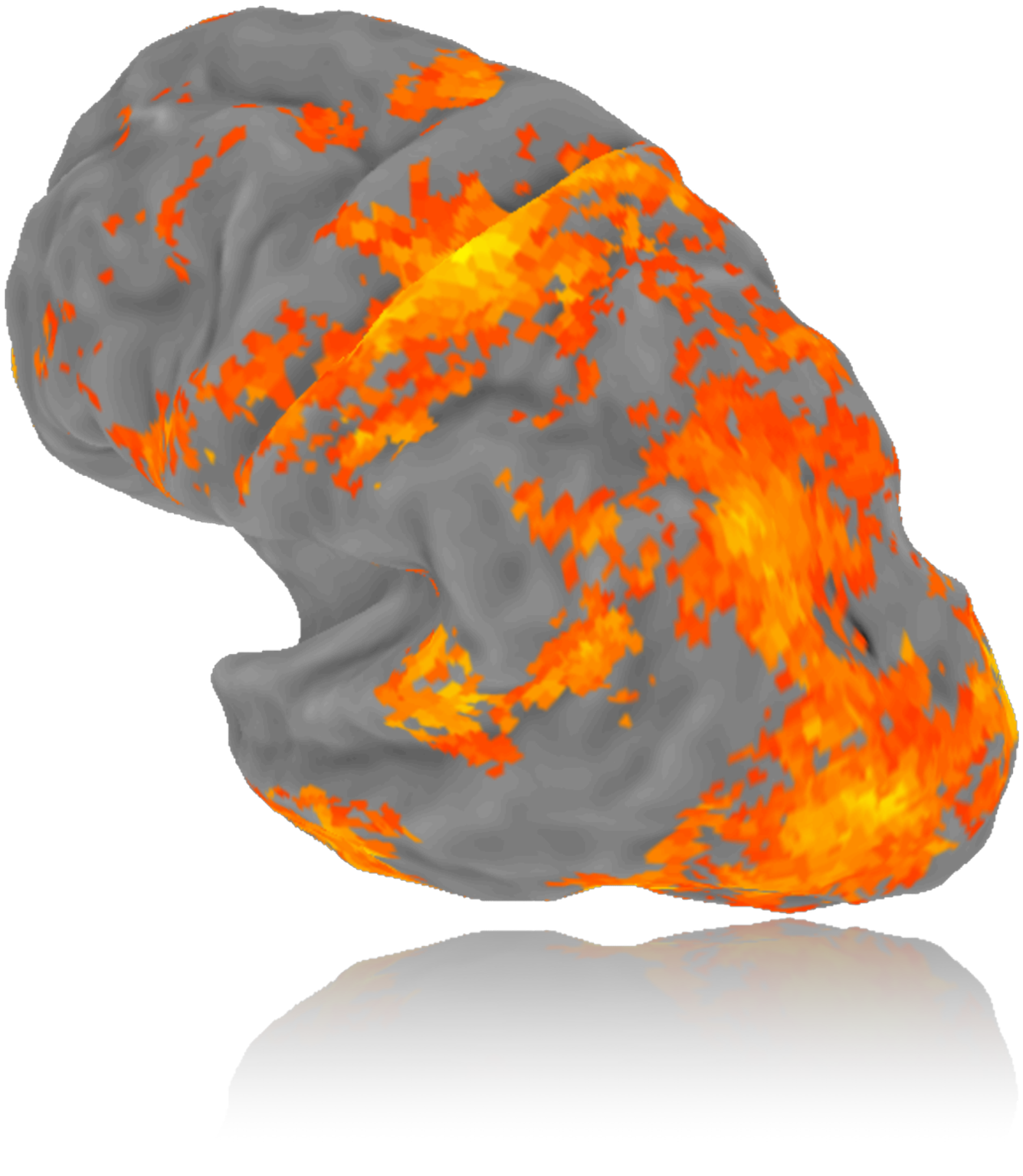
You need to test those assumptions

Evaluate how well your model fits the data

Then you can interpret parameters

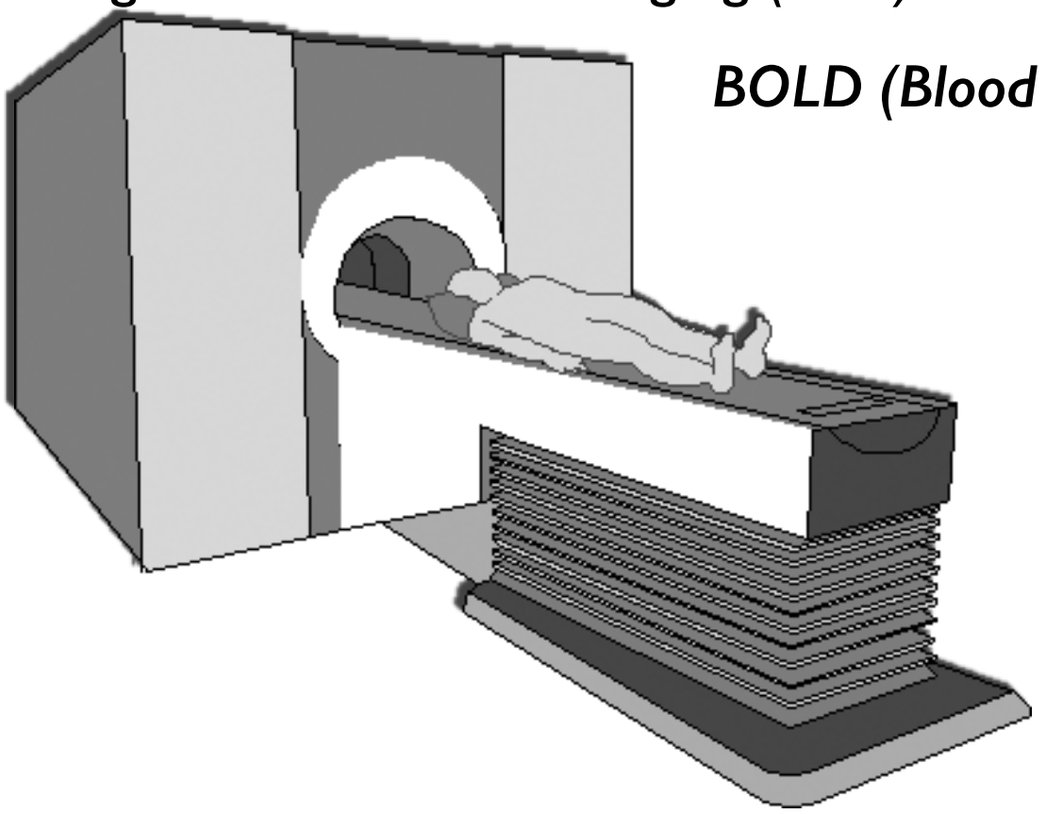
(well ok, here you might want to consider p-values)

P-values are overrated

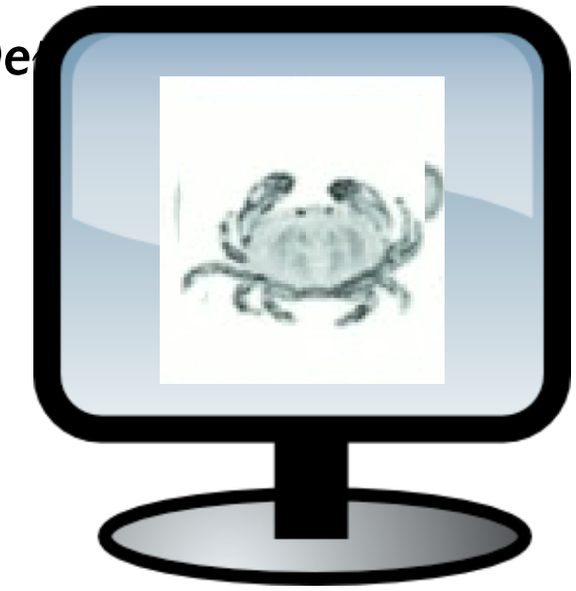


P-values are overrated

Magnetic Resonance Imaging (MRI)

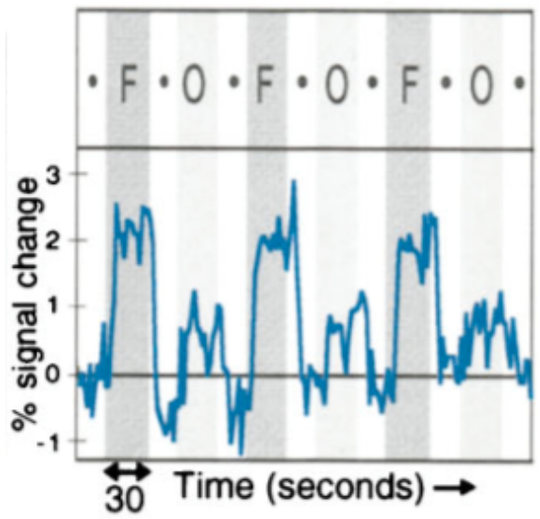


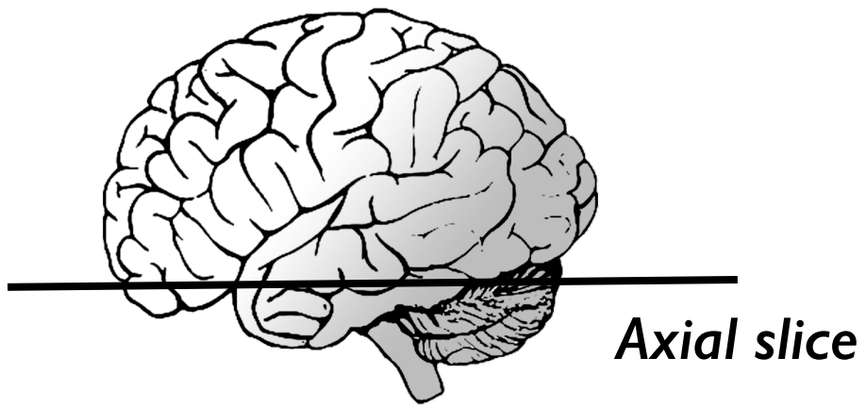
BOLD (Blood Oxygen-Level De



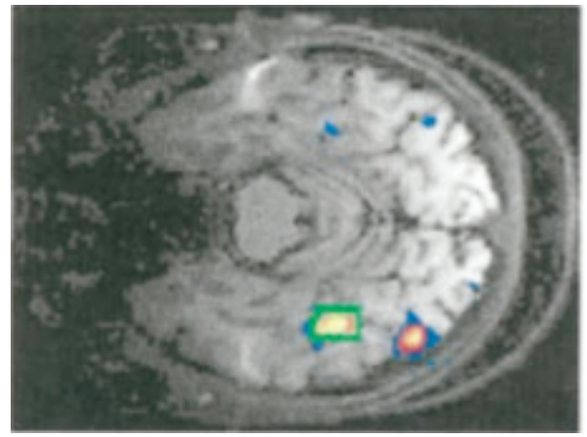
Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–43

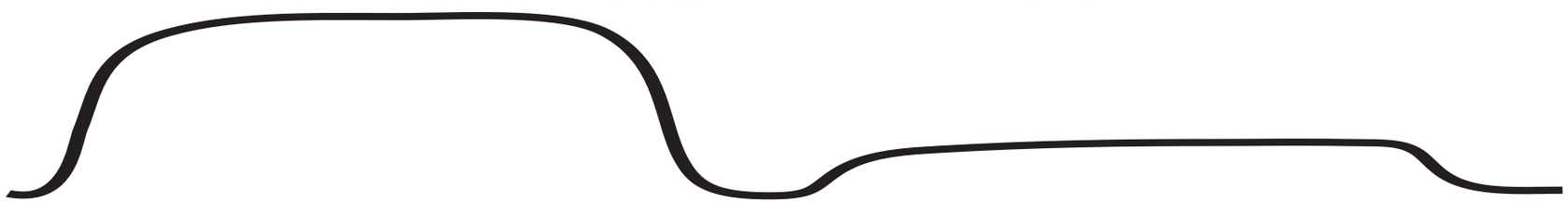
P-values are overrated





Fusiform face area





Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.

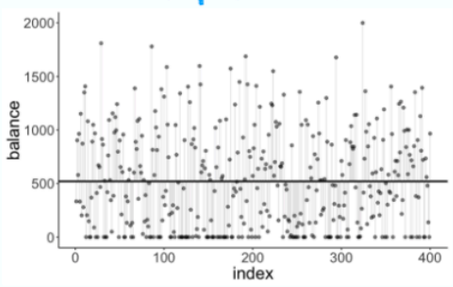
P-values are overrated

H_0: Students and non-students have the same balance.

Model C

$$Y_i = \beta_0 + \epsilon_i$$

Model prediction



Fitted model

$$Y_i = 520.02 + e_i$$

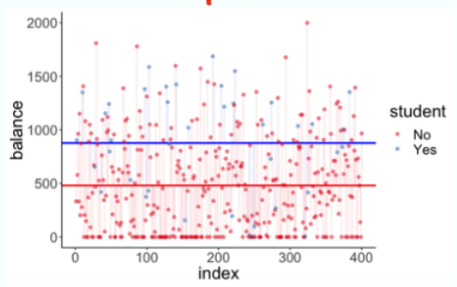
H_1: Students and non-students have different balances.

Model A

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

student

Model prediction



Fitted model

$$Y_i = 480.37 + 396.46 X_i + e_i$$

Worth it?

```
1 # fit the models
2 fit_c = lm(balance ~ 1, data = df.credit)
3 fit_a = lm(balance ~ student, data = df.credit)
4
5 # run the F test
6 anova(fit_c, fit_a)
```

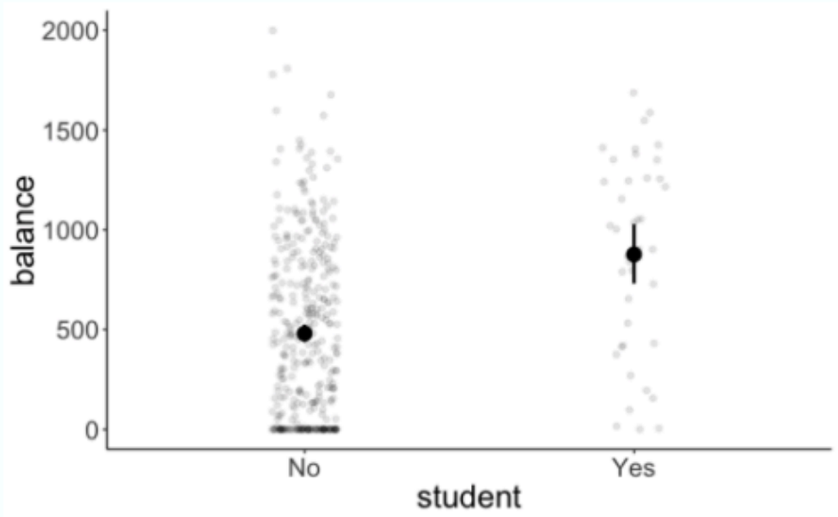
```
Analysis of Variance Table

Model 1: balance ~ 1
Model 2: balance ~ student
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    399 84339912
2    398 78681540  1   5658372 28.622 1.488e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Worth it!

Two sample t-test (with independent groups)

Reporting the results



Students have a significantly higher average credit card balance (*Mean* = 876.83, *SD* = 490.00) than non-students (*Mean* = 480.37, *SD* = 439.41), $F(1, 398) = 28.622$, $p < .001$.

P-values are overrated

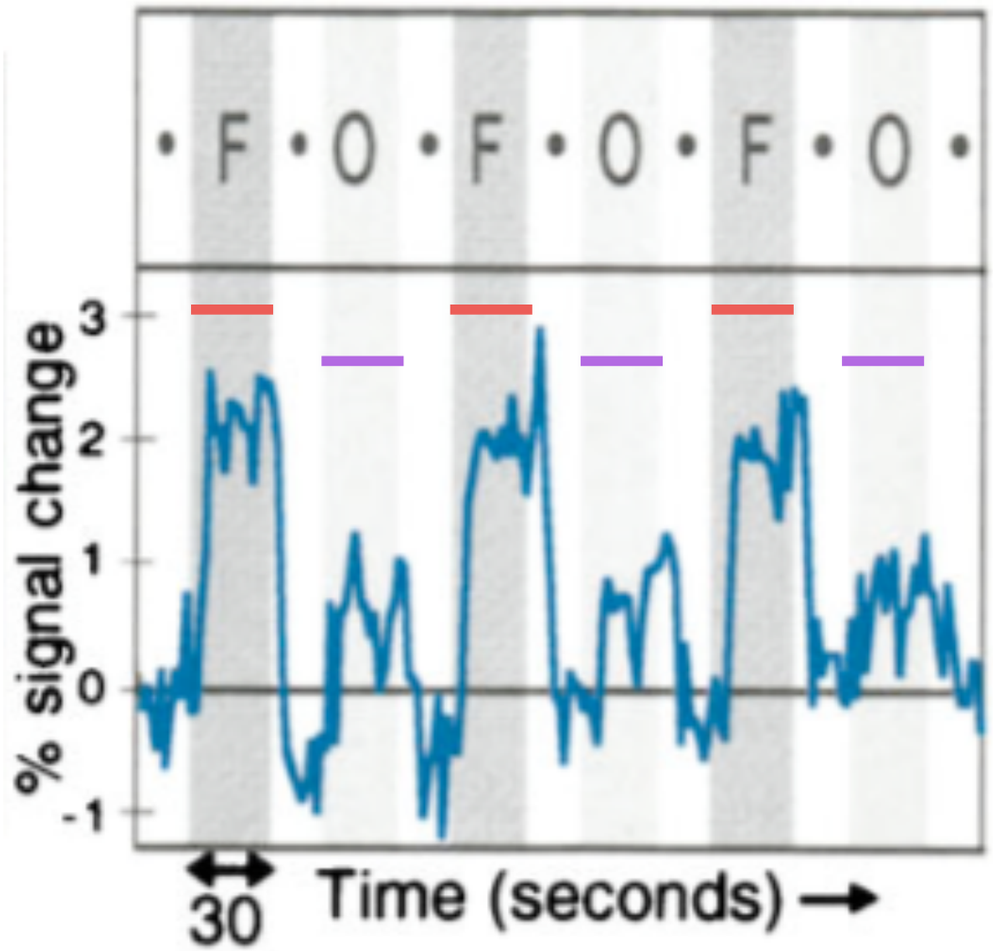
For every voxel, do the following procedure

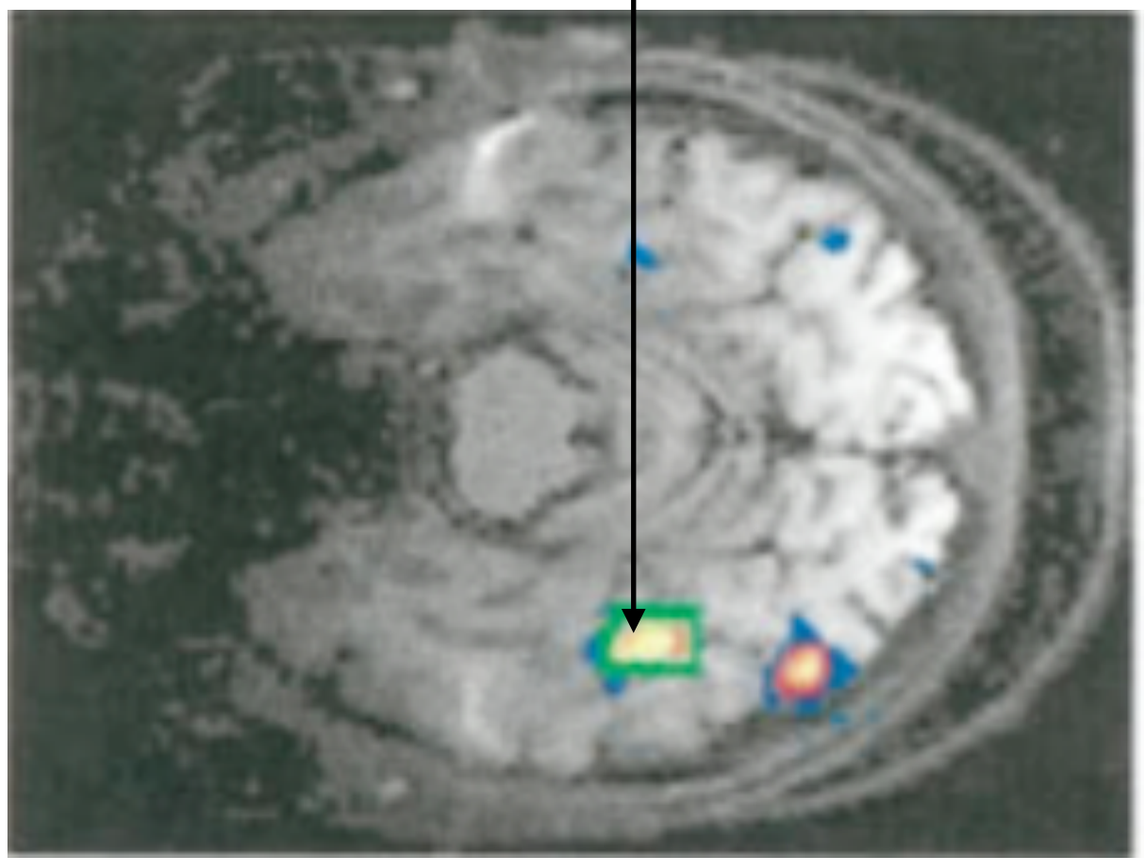
Average each red portion to get 3 "face" responses

Average each purple portion to get 3 "object" responses

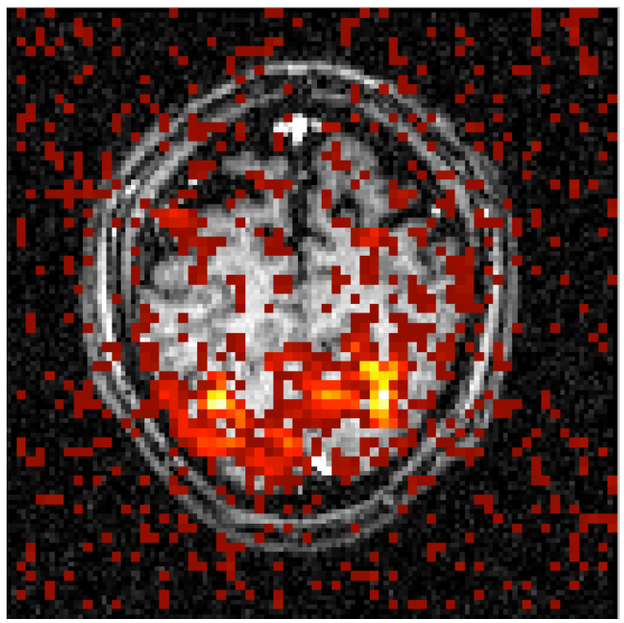
Use a t-test to compute p-value for comparison of face and object responses

Color each voxel according to how small its p-value is





P-values are overrated



p<0.01

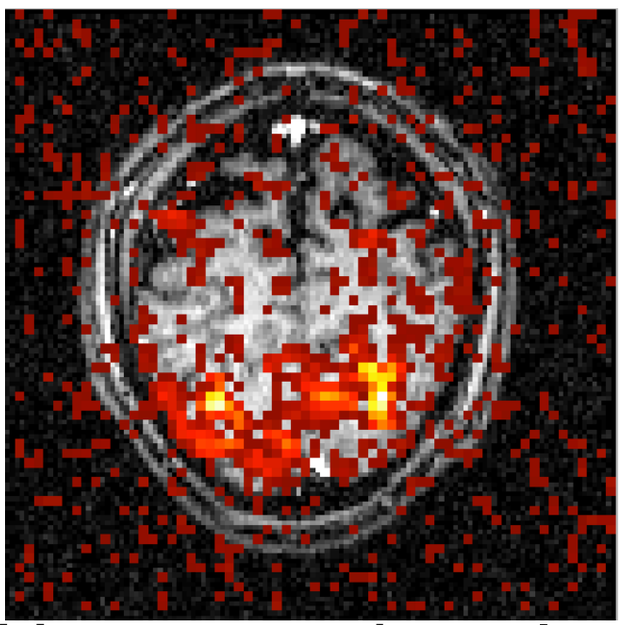
Every 1 in 100 times, the test will show a false-positive

For an imaging experiment with 25 64x64 slices,
you have more than 100,000 voxels and expect
a 1000 voxels to show up as false-positives.

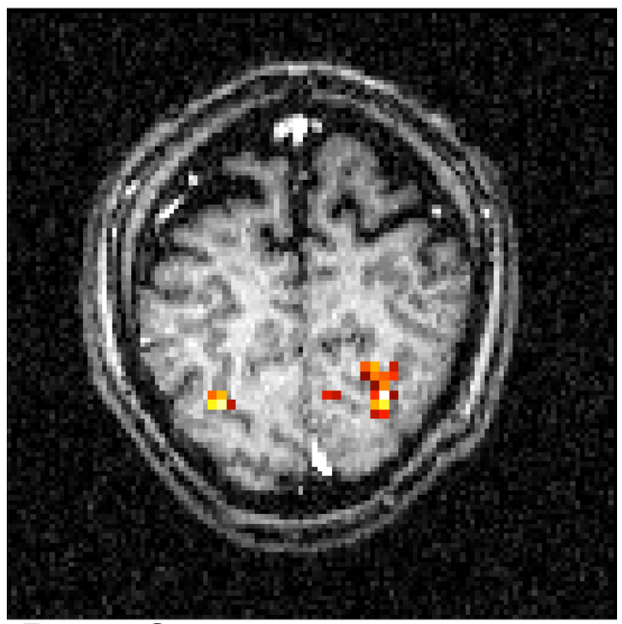
P-values are overrated

Bonferroni correction

- Divide p by the number of voxels n
- Assumes complete independence of voxels
- Very conservative (i.e. likely to cause many false-negative errors)



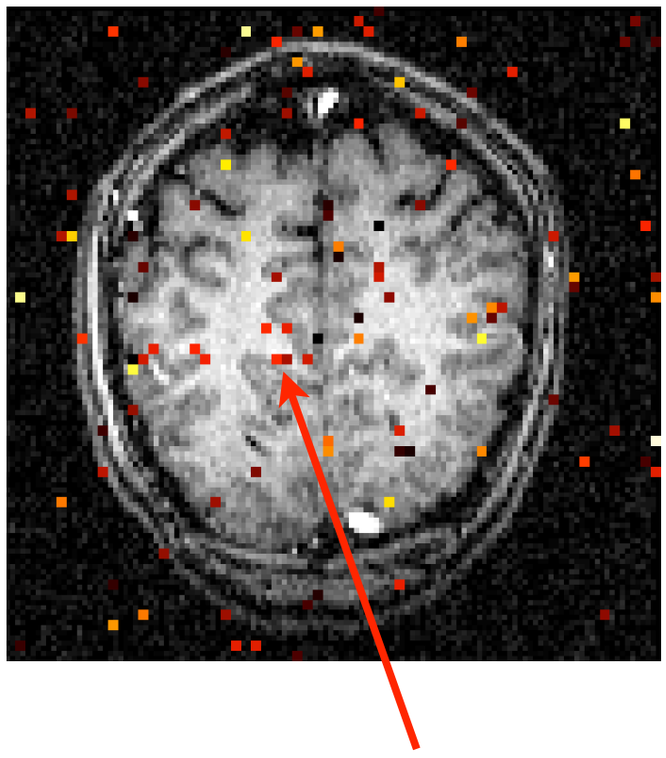
Uncorrected p-value



Bonferroni corrected p-value

P-values are overrated

Cluster threshold



Only “clusters” of 2 or more voxels are shown

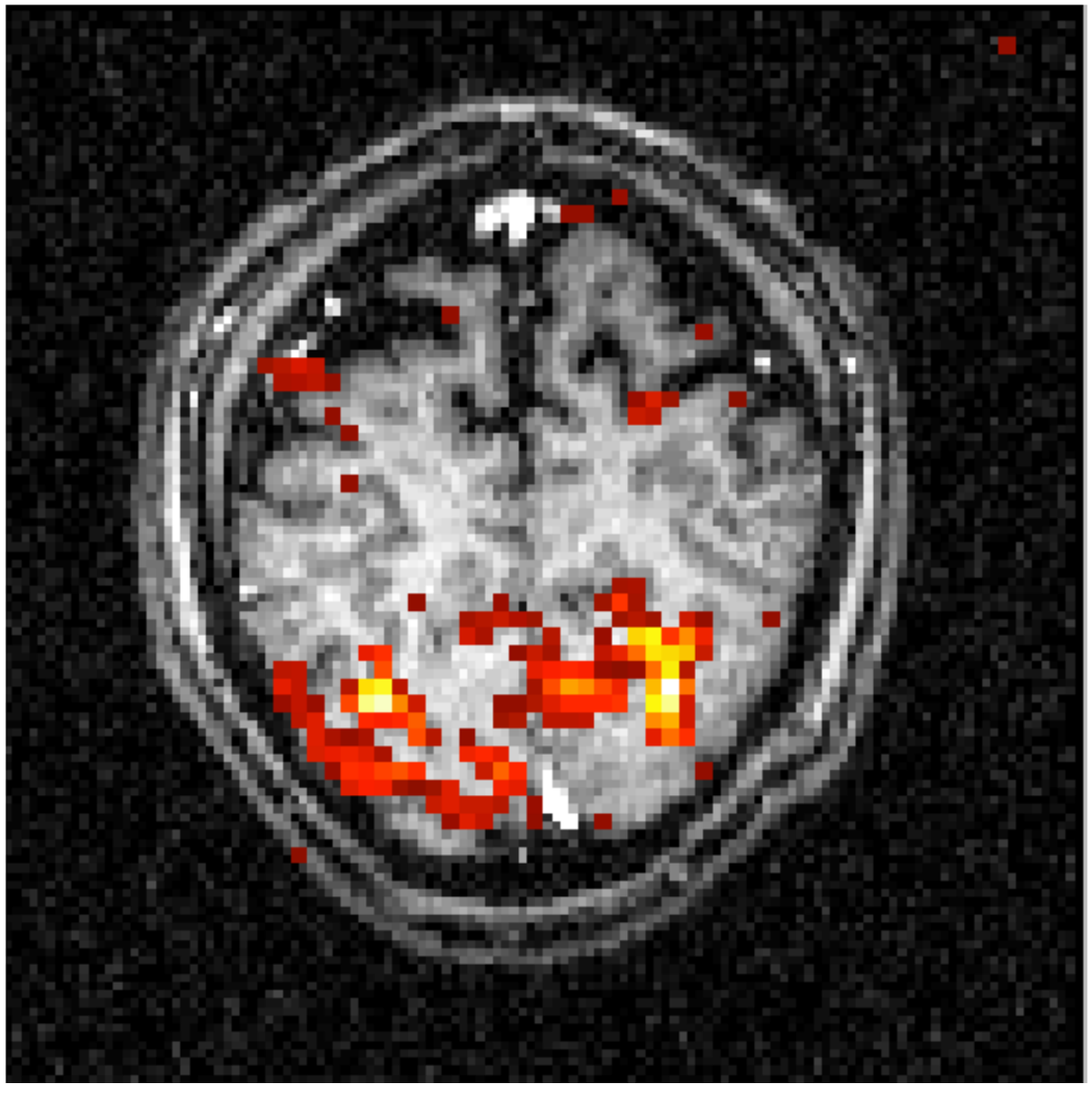
In a 64x64 images there are roughly 16,000 distinct clusters of 2. For a p-value of 0.001, then the probability of a cluster of two contiguous voxels being active is:

16,000 x 0.001 x 0.001 = 0.016

P-values are overrated

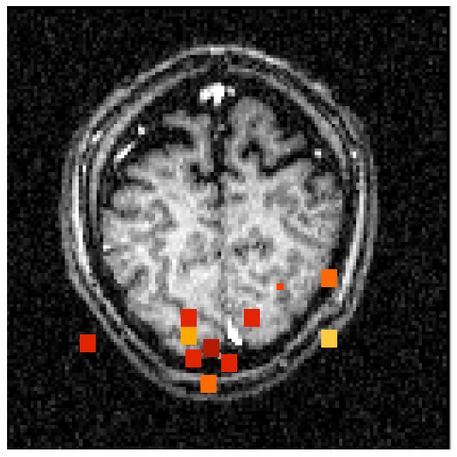
False discovery rate (FDR)

Of all the voxels "discovered" as active by the analysis, only FDR% of them are falsely labelled active.

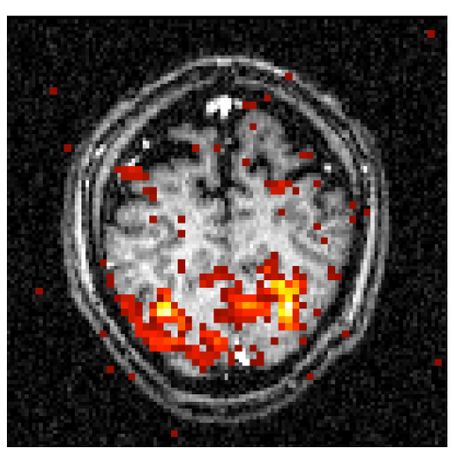


P-values are overrated

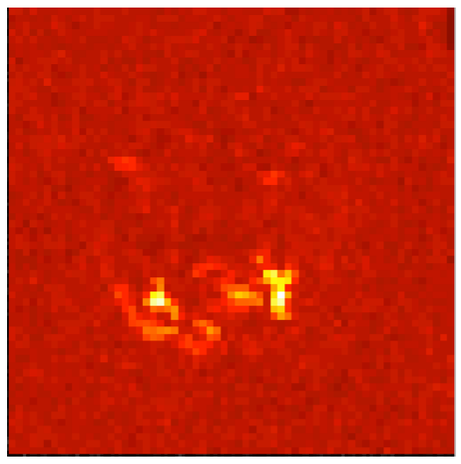
As opposed to "Familywise error rate", the number of possible false positives scales. e.g. if FDR=0.1...



If 10 voxel are active, then 1 of them may be falsely active.

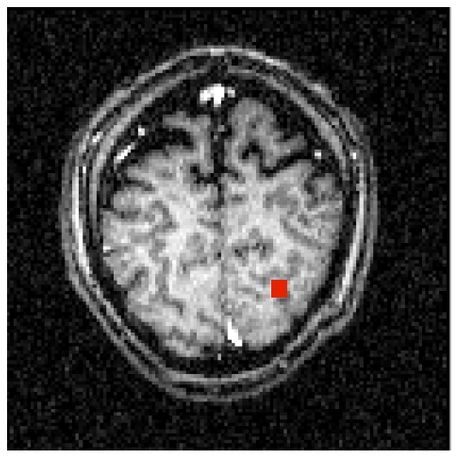


If 50 voxels are active, then 5 may be falsely active

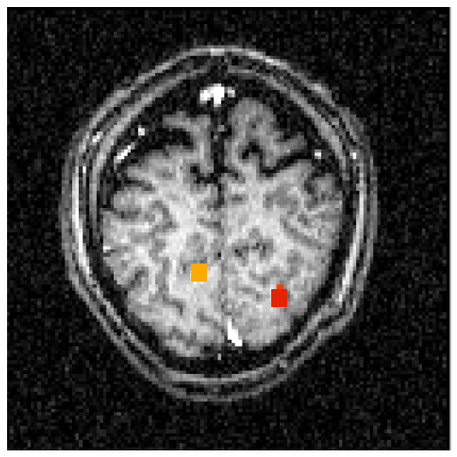


If all 4096 voxels are active, then 410 may be falsely active.

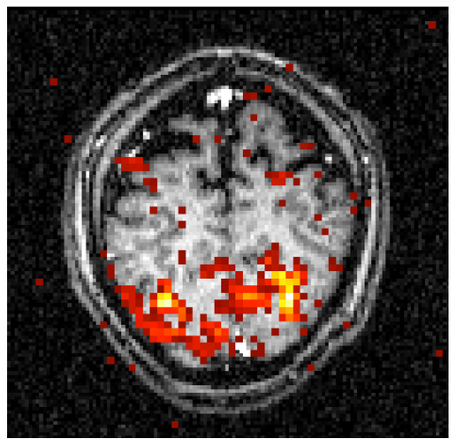
P-values are overrated



If 1 voxel has p=0.01, then the probability of a false activation is just FDR=p



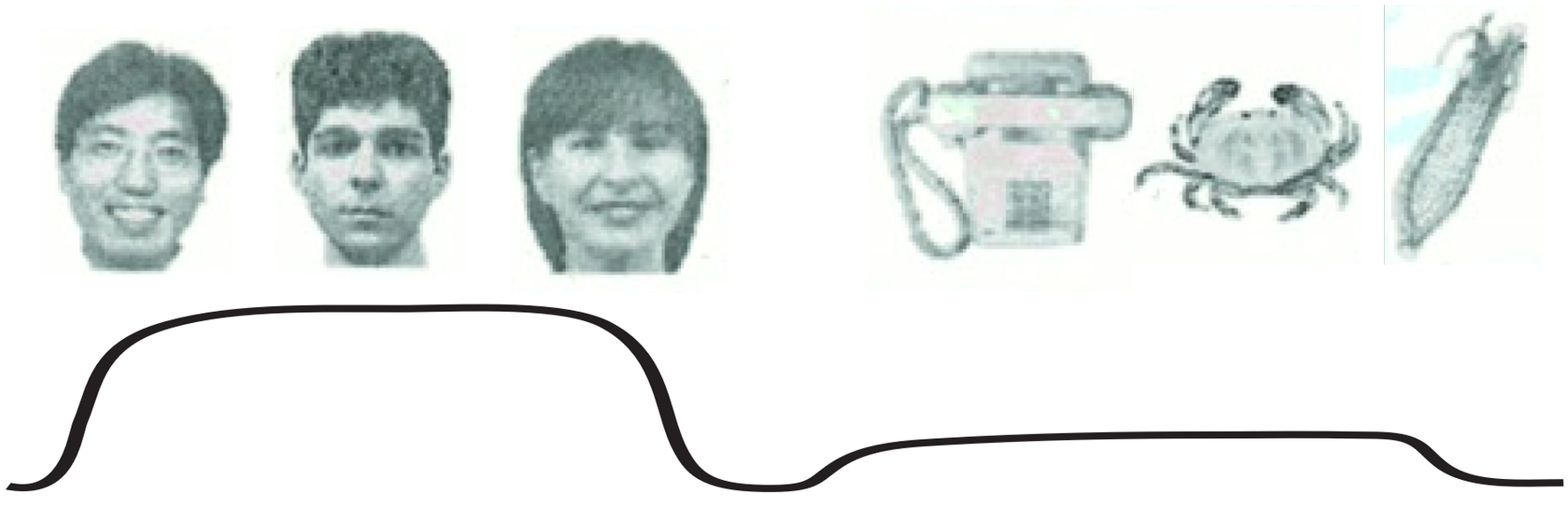
If 2 voxels have p=0.01, then the probability of a false activation, FDR, is 1-probability of no false activation: 1-(1-p)(1-p)



If k voxels have p=0.01, then the probability of a false activation, FDR=1-(1-p)^k

P-values are overrated

But, is the t-test really the right “model”

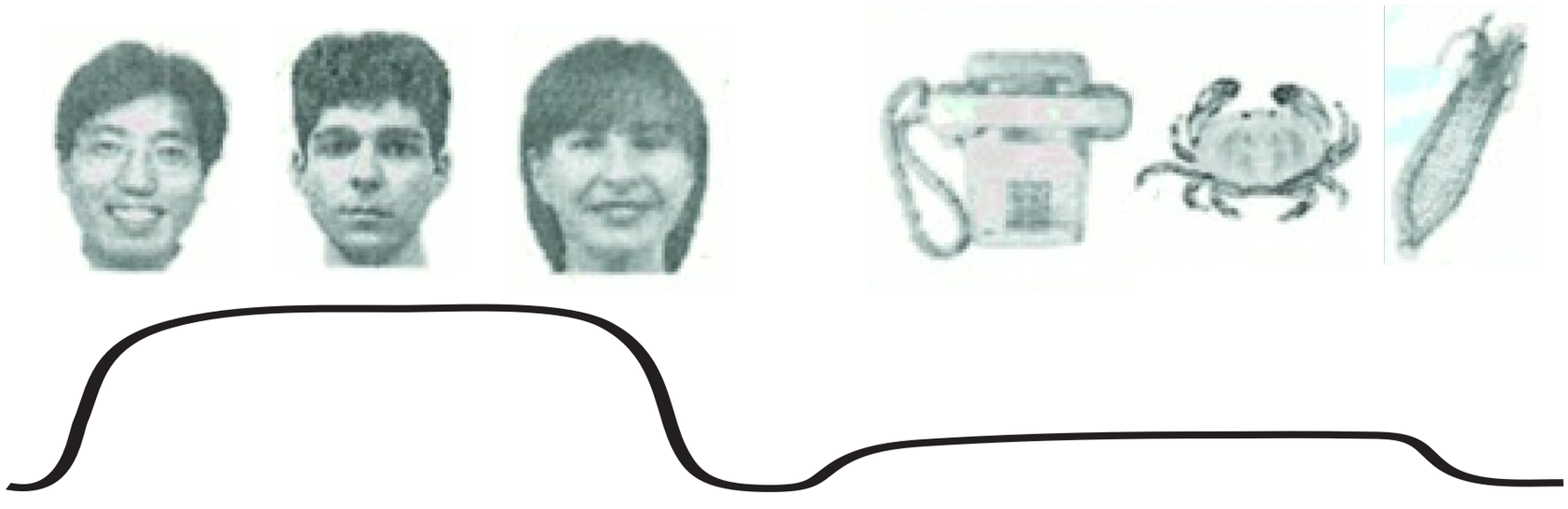


The way we did the analysis, assumes
a steady response during each period
(i.e. that’s the model of response)

But, if that is not true, then the t-test won’t work…

P-values are overrated

But, is the t-test really the right “model”



= 0 response

= 0 response

t-test might not find a difference
because it is the wrong model

P-values are overrated

Statistical analyses are really models

Models have assumptions

You need to test those assumptions

Evaluate how well your model fits the data

Then you can interpret parameters

(well ok, here you might want to consider p-values)

General Linear Model

Linear model

```
lm(formula = value ~ 1 + condition,
   data = df.original)
```

$$\text{value}_i = b_0 + b_1 \cdot \text{condition}_i + e_i$$

i = observation

$$e_i \sim \mathcal{N}(\text{mean} = 0,\ \text{sd} = s_{\text{error}})$$

3 parameters: b_0, b_1, s_{error}

Linear mixed effects model

```
lmer(formula = value ~ 1 + condition +
                (1 | participant),
     data = df.original)
```

$$\text{value}_{ij} = b_0 + b_1 \cdot \text{condition}_{ij} + U_i + e_{ij}$$

i = participant,
j = time point

$$e_{ij} \sim \mathcal{N}(\text{mean} = 0,\ \text{sd} = s_{\text{error}})$$

$$U_i \sim \mathcal{N}(\text{mean} = 0,\ \text{sd} = s_U)$$

b_0, b_1 = fixed effects
U_i = random effect

here: random intercept

4 parameters: b_0, b_1, s_{error}, s_U

(A)

Subjects' time courses

Combined time course

Combined statistical map

Fixed effects

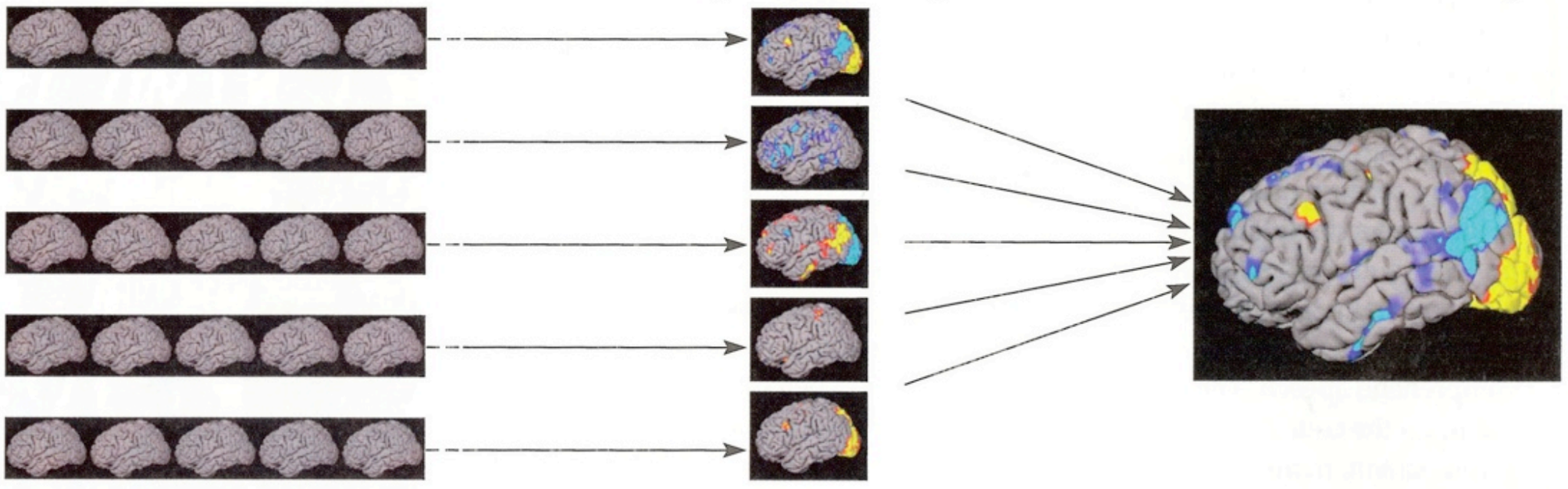
(B)

Subjects' time courses

Subjects' statistical maps

Combined statistical map

Random effects



Fixed-effects

Average first, then compute statistics.

More powerful (smaller false-negative rate), but conclusions are limited to the particular sample population chosen.

Random-effects

Compute statistics on each subject, then average.

Less powerful (requires many subjects - e.g. n>25), but conclusions can be generalized to the population from which the sample was chosen.

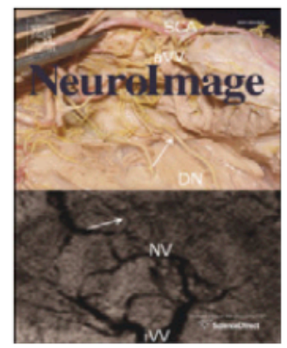
Maybe you should do mixed-effect modeling?!

NeuroImage 73 (2013) 176–190

Contents lists available at SciVerse ScienceDirect

NeuroImage

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



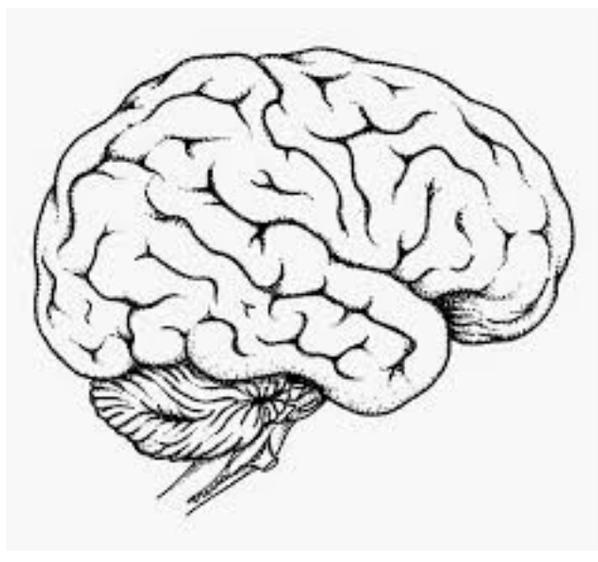
Linear mixed-effects modeling approach to FMRI group analysis

Gang Chen [a,*], Ziad S. Saad [a], Jennifer C. Britton [b], Daniel S. Pine [b], Robert W. Cox [a]

[a] *Scientific and Statistical Computing Core, NIMH/NIH/HHS, USA*
[b] *Section on Development and Affective Neuroscience, NIMH/NIH/HHS, USA*

Bigger problem for imaging data is how to combine different brains?

?
=



P-values are overrated

Statistical analyses are really models

Models have assumptions

You need to test those assumptions

Evaluate how well your model fits the data

Then you can interpret parameters

(well ok, here you might want to consider p-values)

Assumptions of the General Linear Model

1) Stationarity (same response through experiment)

2) Hemodynamic response has a particular shape

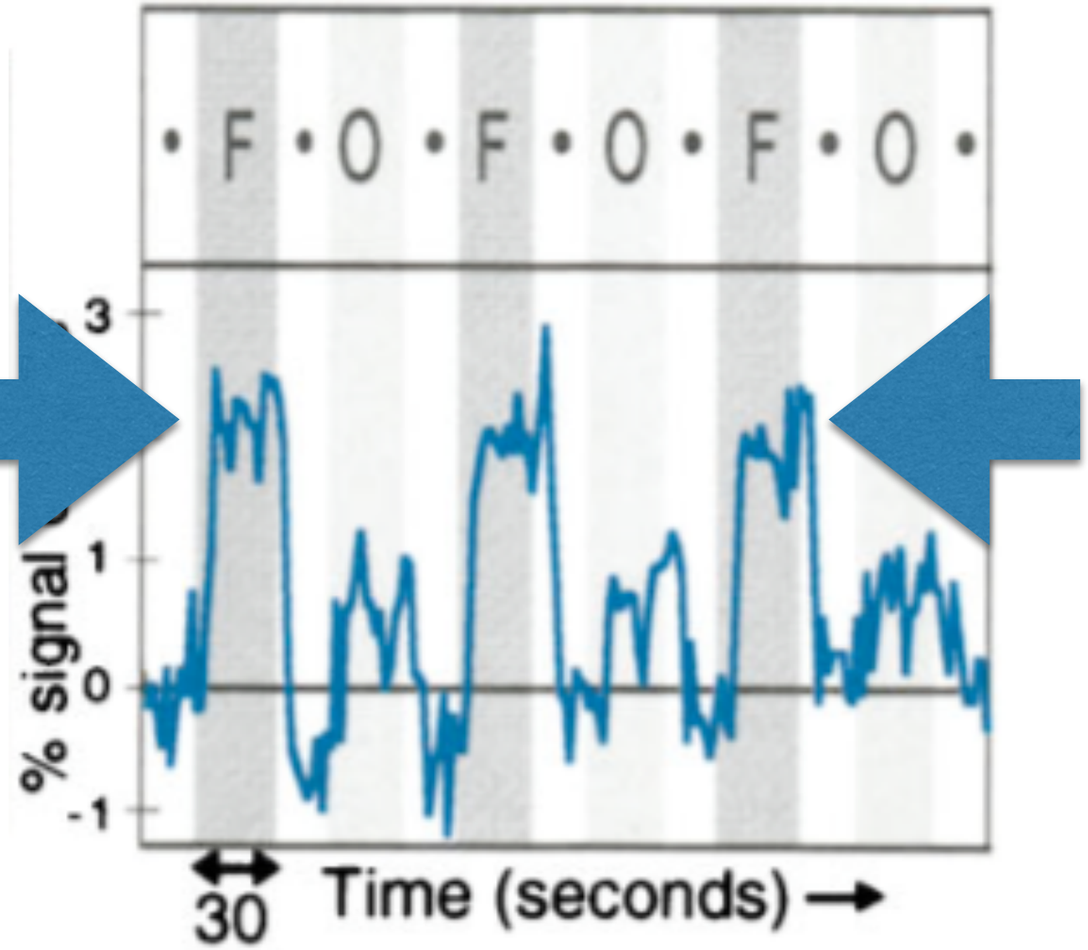
3) Temporal linearity (response overlaps sum)

4) Noise is “white”

1) Stationarity (same response through experiment)

For example…

Is the response the first time we present the stimulus…



…the same as the last time?

Assumptions of the General Linear Model

1) Stationarity (same response through experiment)

2) Hemodynamic response has a particular shape

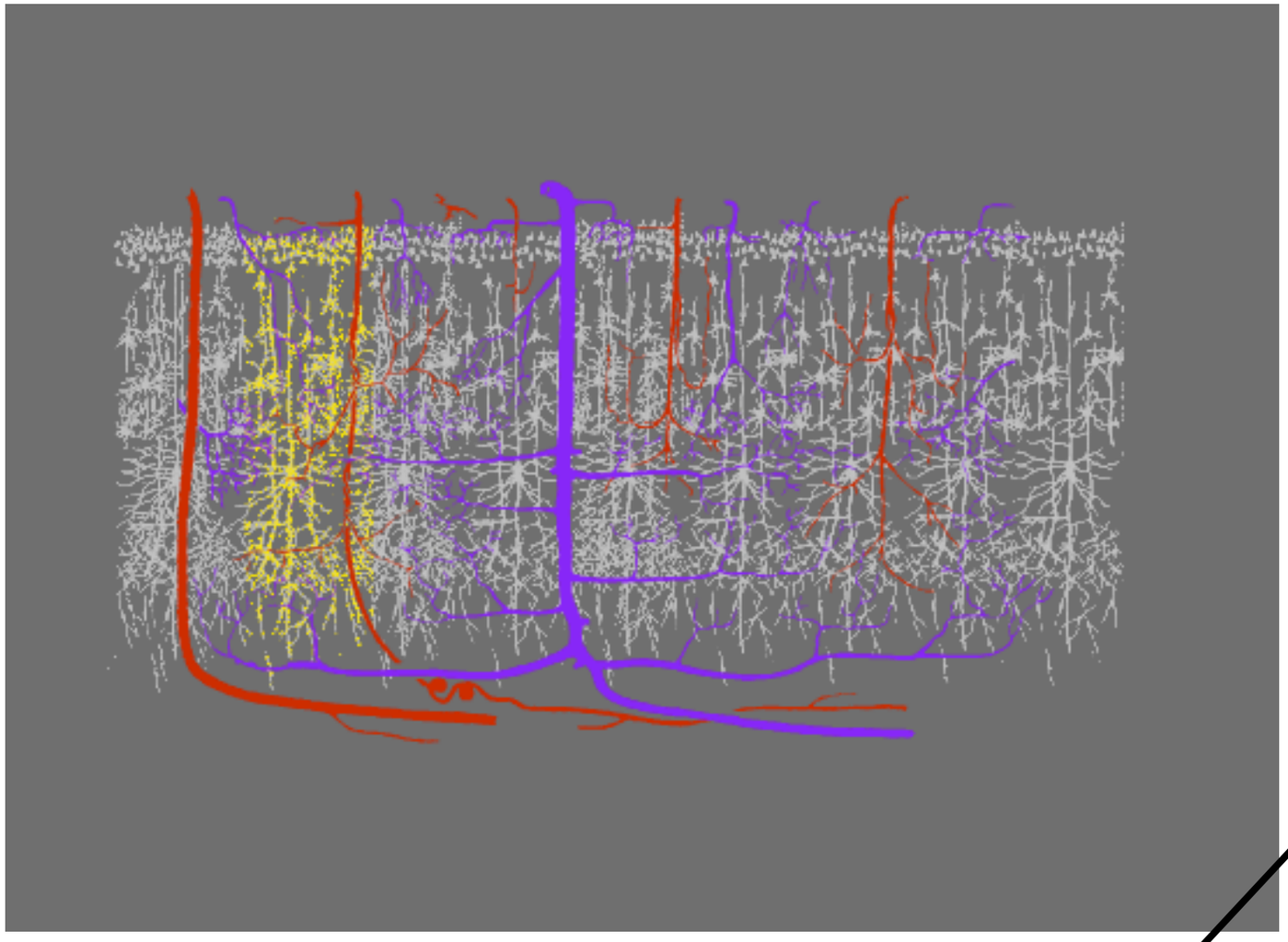
3) Temporal linearity (response overlaps sum)

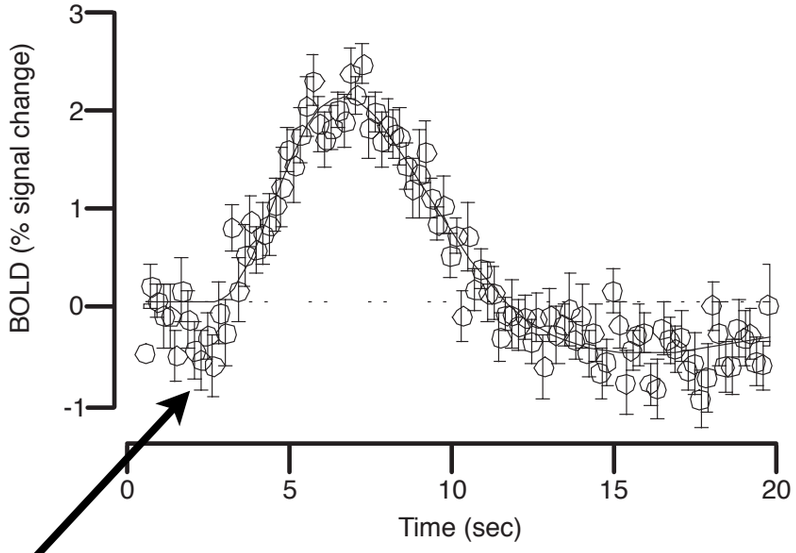
4) Noise is “white”

BOLD contrast

The higher the [deoxy hemoglobin]
the darker the image becomes.

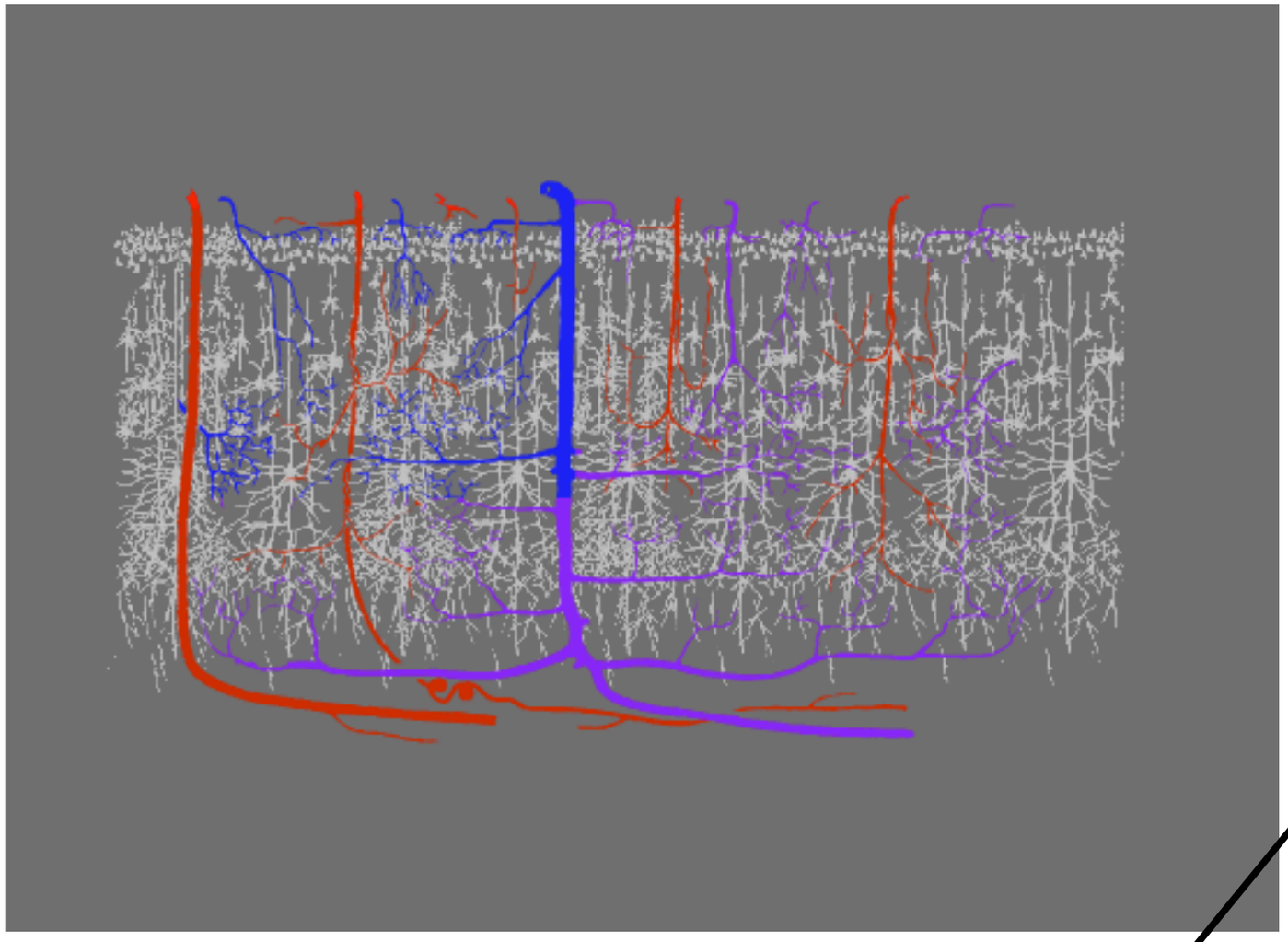
After neural activity, Cerebral Metabolic Rate of Oxygen consumption ($CMRO_2$) increases

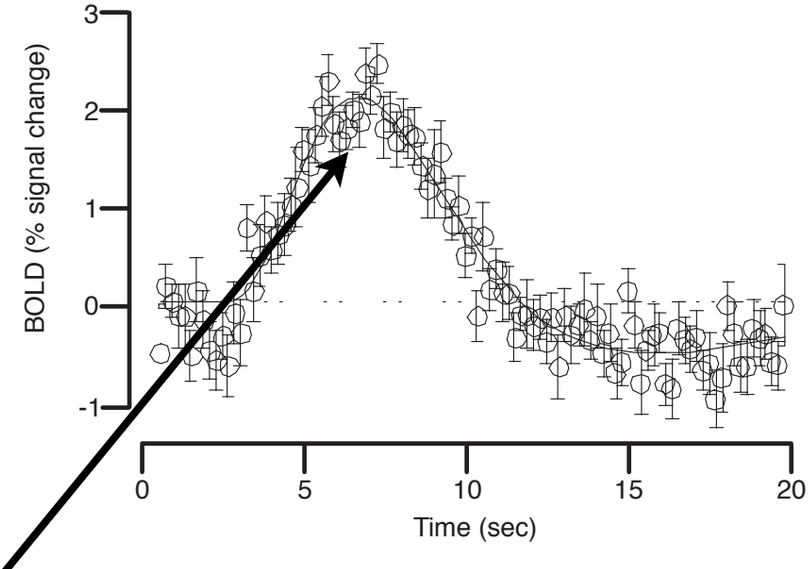




Increase in [deoxy hemoglobin] measured as a darkening of fMRI image

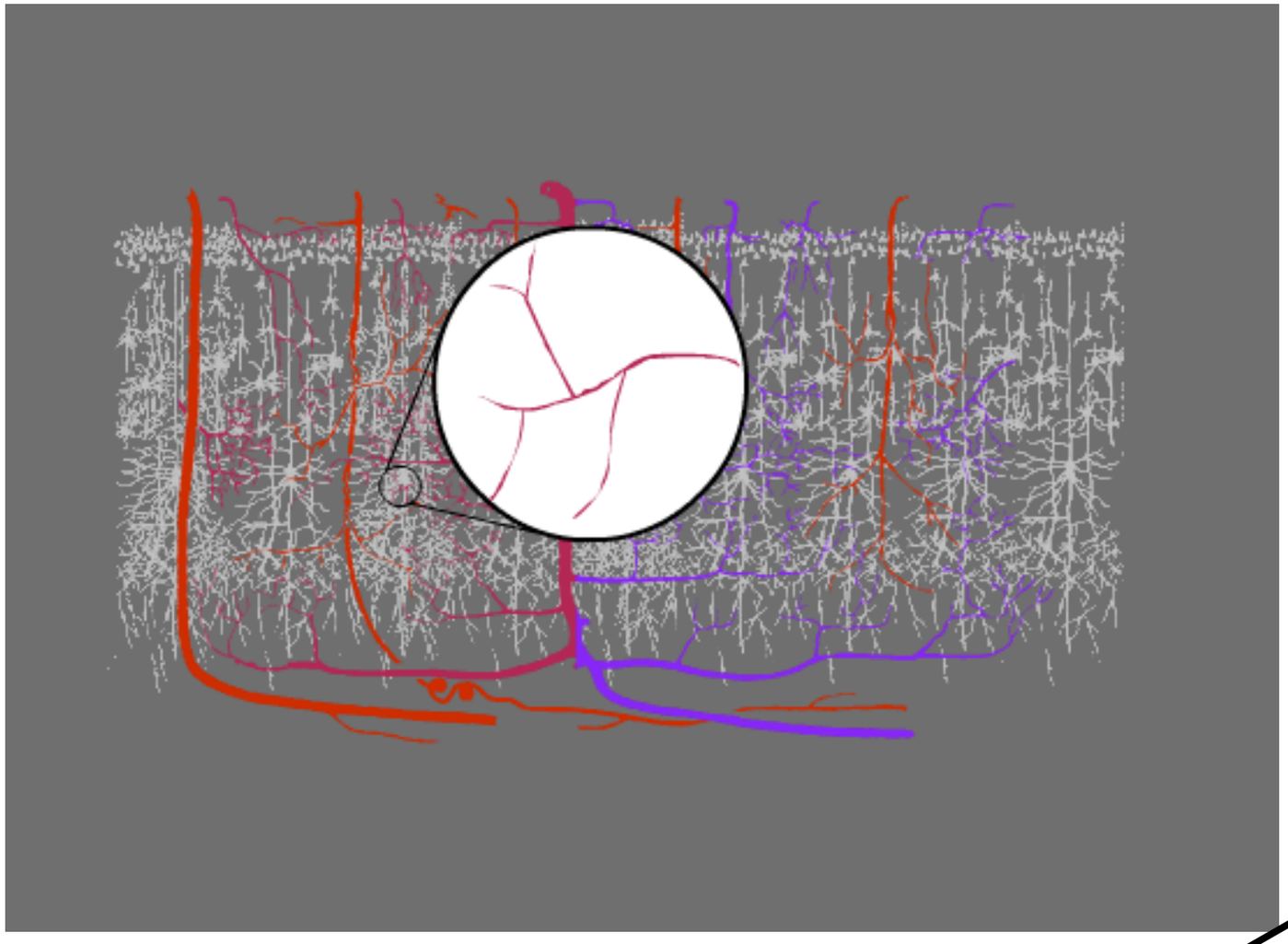
Cerebral Blood Flow (CBF) increases to bring in fresh oxygenated blood

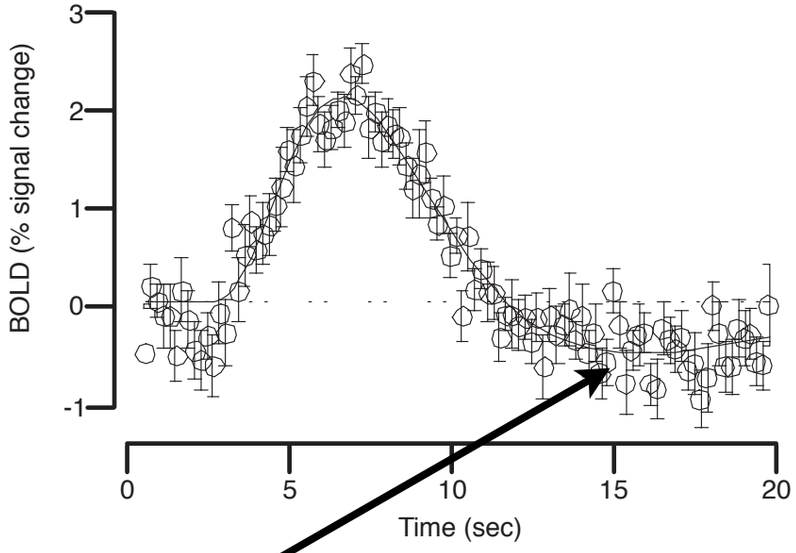




Decrease in [deoxy hemoglobin] measured as a brightening of fMRI image

Cerebral Blood Volume (CBV) increases resulting in more pooling of deoxygenated blood



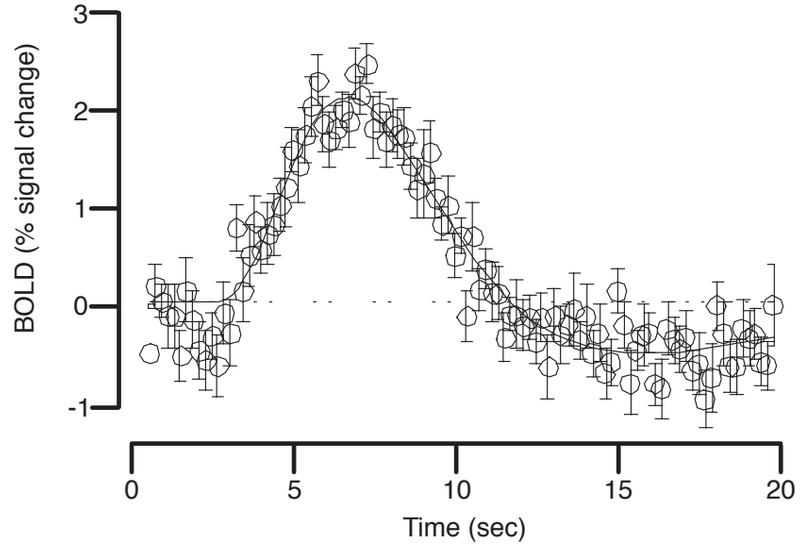


Increase in [deoxy hemoglobin] measured as a darkening of fMRI image

Time to peak
is usually
5-7s

Return to
baseline
can be
30s

Lag is
usually
1-2s



Rarely see initial dip

Can be fit with a difference of gamma functions

Assumptions of the General Linear Model

1) Stationarity (same response through experiment)

2) Hemodynamic response has a particular shape

3) Temporal linearity (response overlaps sum)

4) Noise is “white”

Imagine an experiment in which we present visual and audio stimuli

$\beta_1 \times$ 

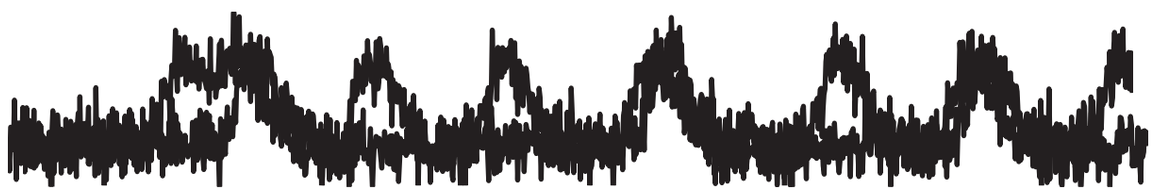
Expected visual response

$\beta_2 \times$ 

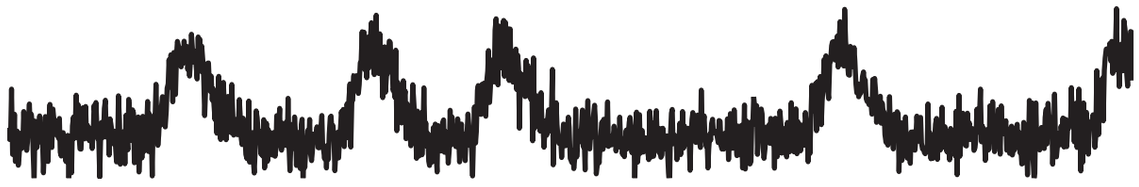
Expected audio response

\+ 

Noise

$\beta_1 = 0, \beta_2 = 1$ 

Auditory area

$\beta_1 = 1, \beta_2 = 0$ 

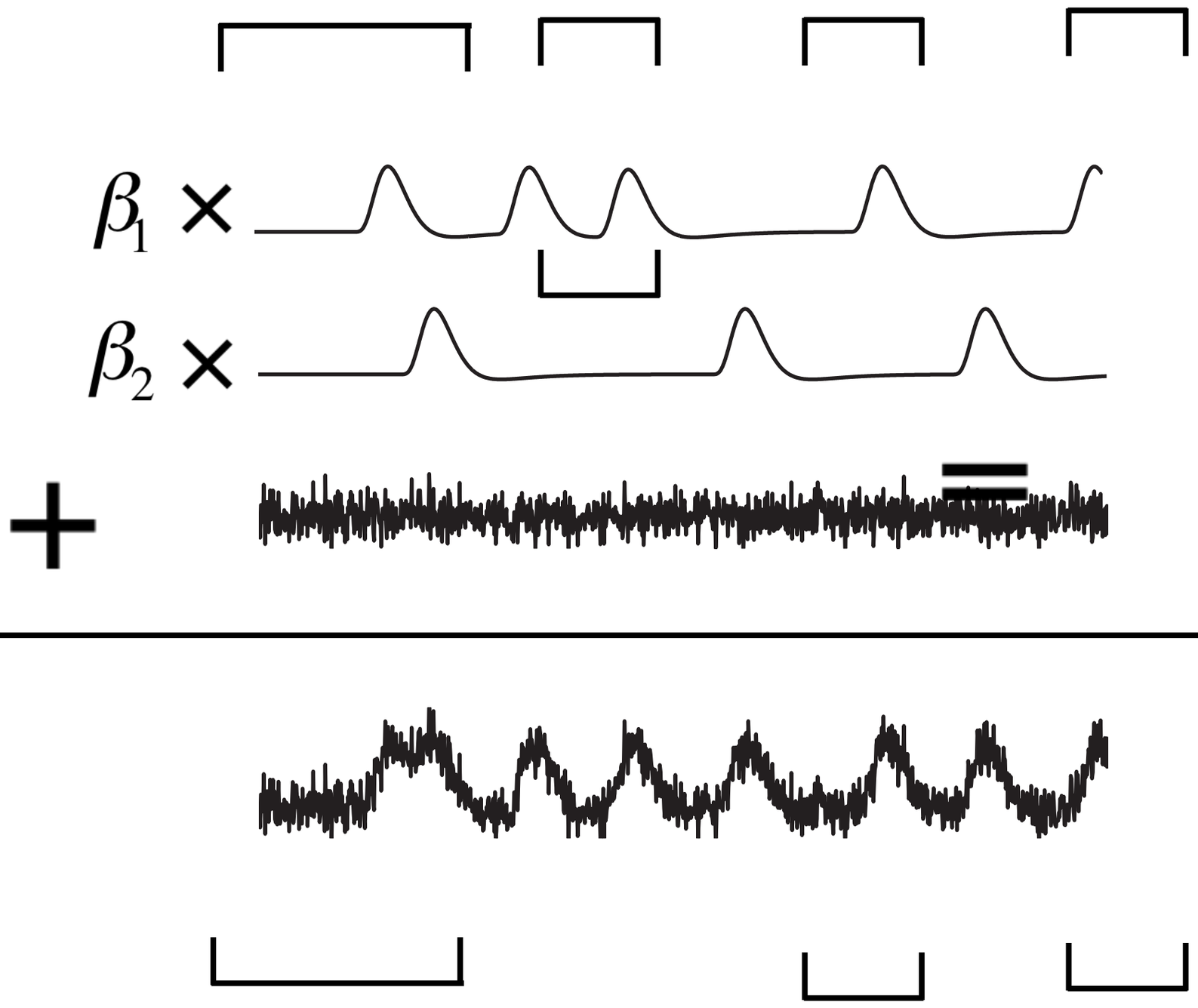
Visual area

$\beta_1 = 0.5, \beta_2 = 1$ 

Multimodal area

BOLD measurement

Now, it's just a linear algebra problem! (called a General Linear Model - GLM)



Assumptions of the General Linear Model

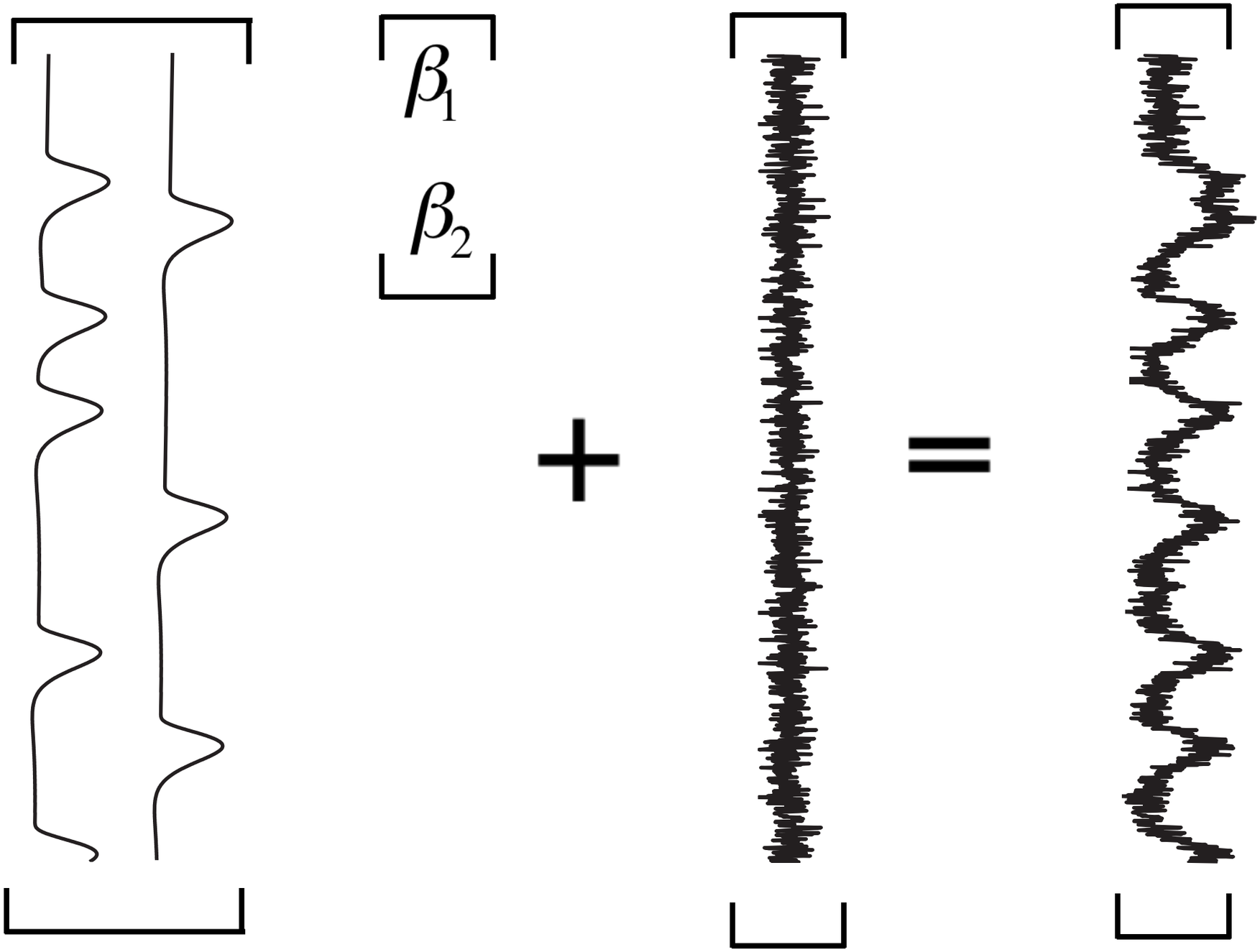
1) Stationarity (same response through experiment)

2) Hemodynamic response has a particular shape

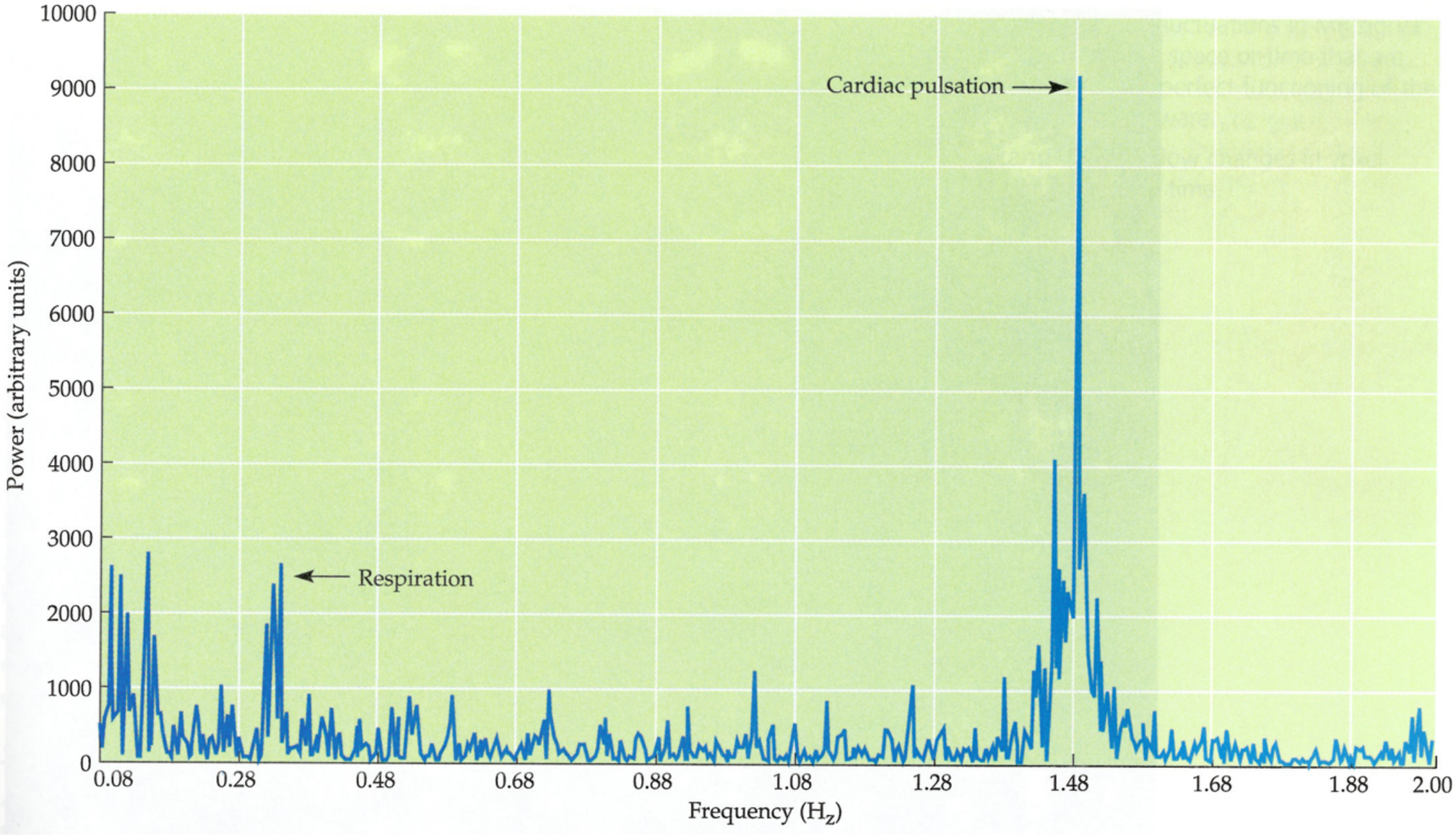
3) Temporal linearity (response overlaps sum)

4) Noise is “white”

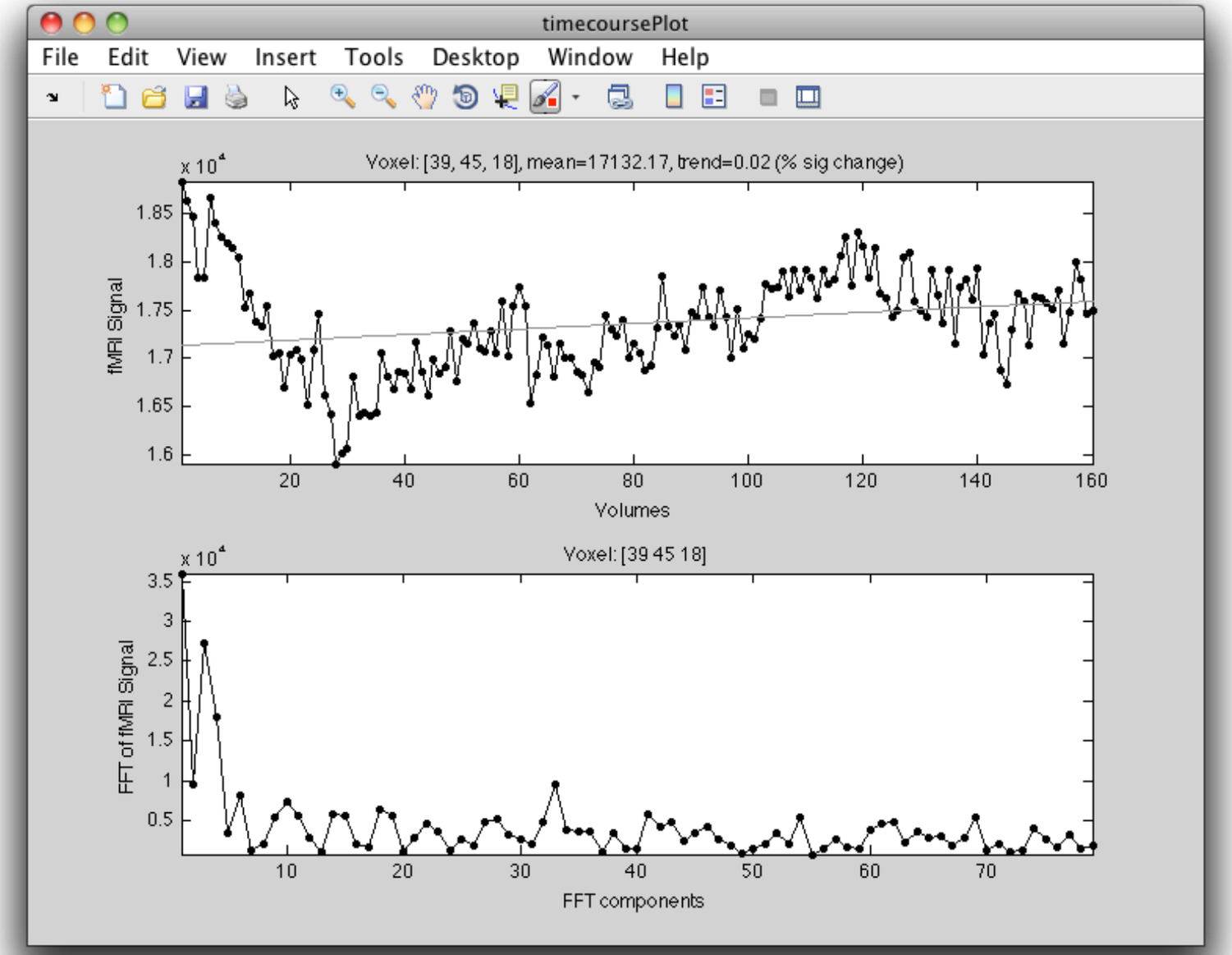
The General Linear Model



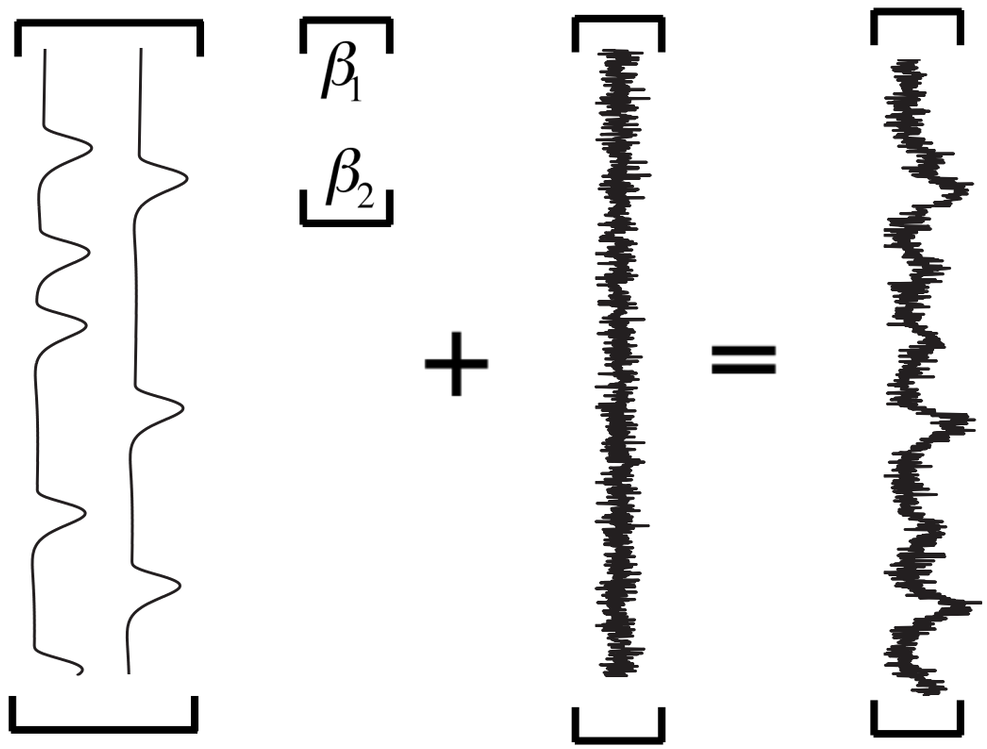
Respiration and cardiac noise



Noise is not always “white”



To remove noise that is not white, detrend and high-pass filter



If you detrend and high-pass filter the data, you MUST ALSO detrend and high-pass the columns of your GLM

P-values are overrated

Statistical analyses are really models

Models have assumptions

You need to test those assumptions

Evaluate how well your model fits the data

Then you can interpret parameters

(well ok, here you might want to consider p-values)

Assumptions of the General Linear Model

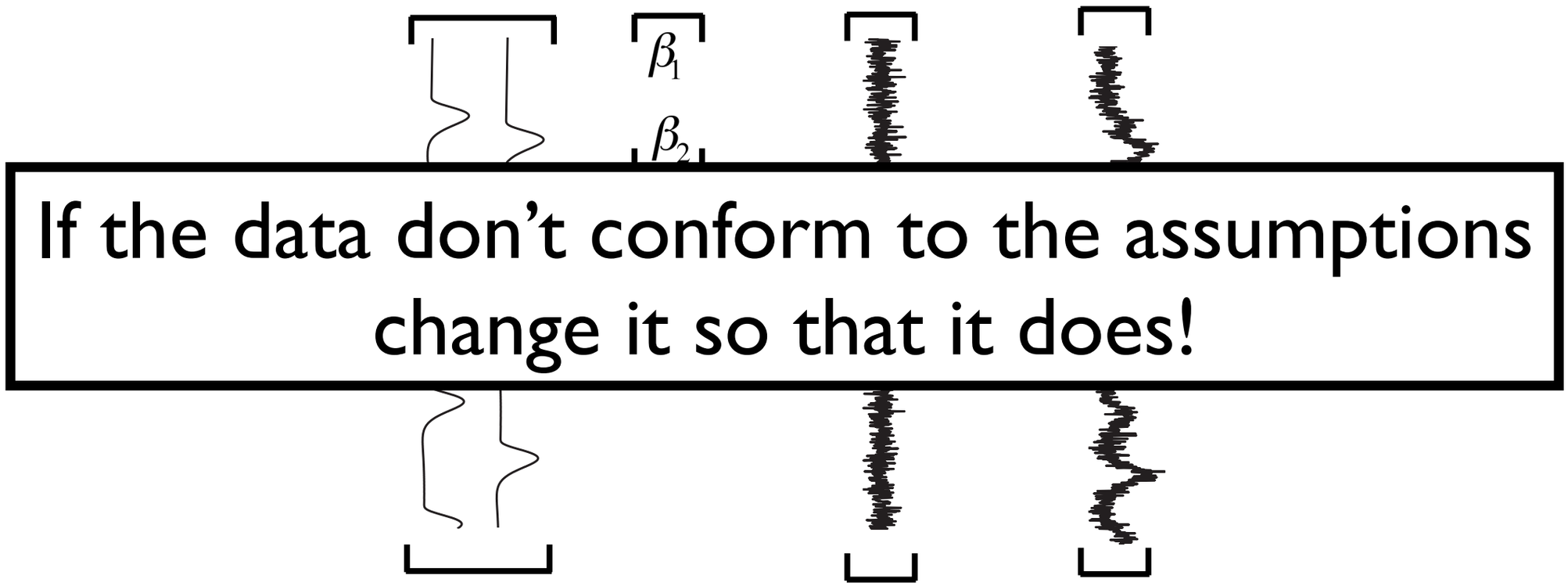
1) Stationarity (same response through experiment)

2) Hemodynamic response has a particular shape

3) Temporal linearity (response overlaps sum)

4) Noise is “white”

To remove noise that is not white,
detrend and high-pass filter



If the data don't conform to the assumptions
change it so that it does!

If you detrend and high-pass filter the data,
you MUST ALSO detrend and high-pass the
columns of your GLM

Assumptions of the General Linear Model

1) Stationarity (same response through experiment)

2) Hemodynamic response has a particular shape

3) Temporal linearity (response overlaps sum)

4) Noise is “white”

Linear functions obey the superposition principle

$$f(a+b) = f(a) + f(b)$$

The response to the sum of two stimuli is the sum of the responses to each stimulus presented alone.

Evidence for temporal linearity of BOLD

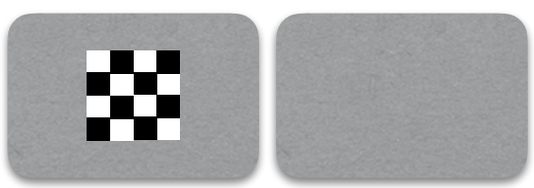
$$f(a+b) = f(a) + f(b)$$

Superposition principle:
Response to the sum of stimuli is the sum of responses to each stimulus presented individually

Stimulus

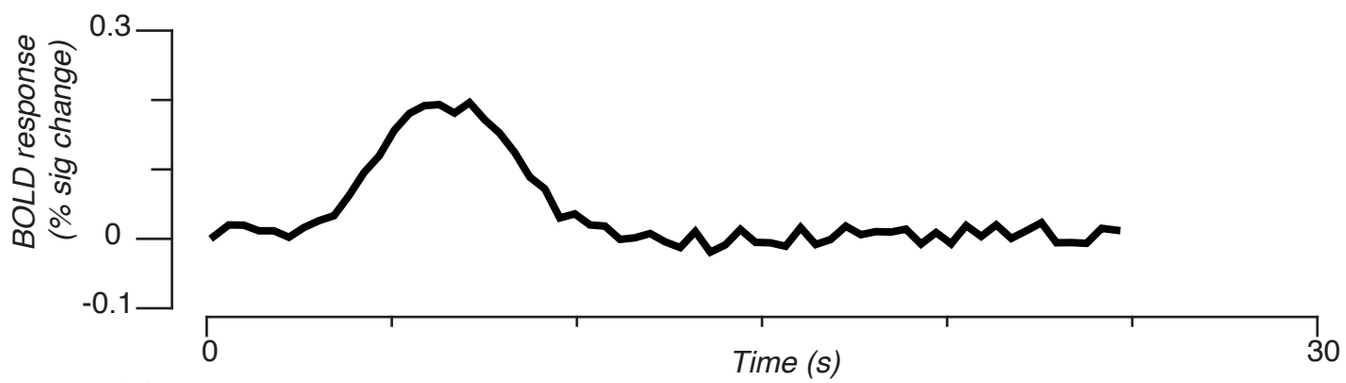
Response

a

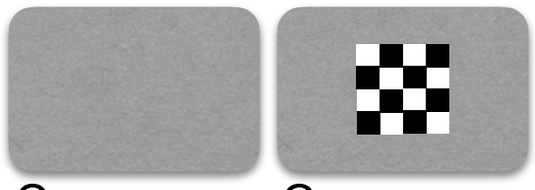


6s 6s

f(a)

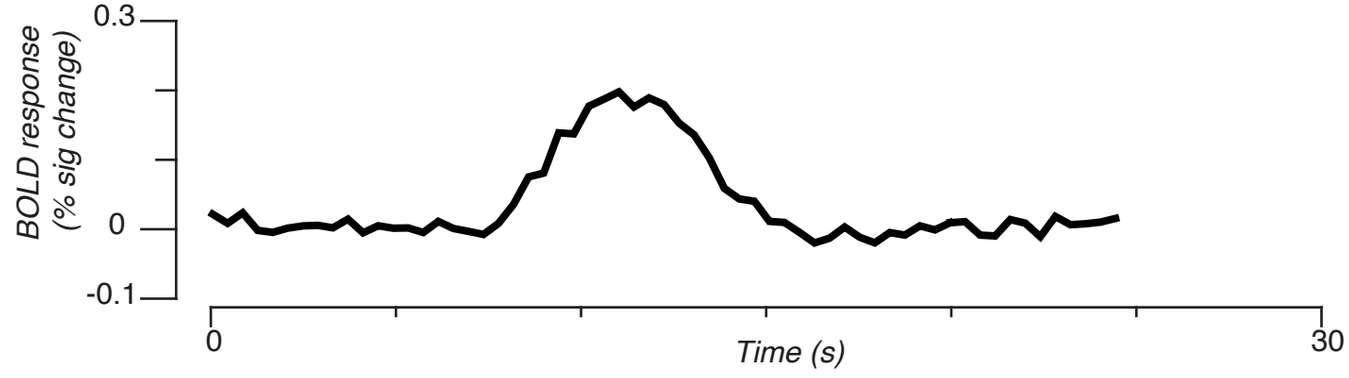


b

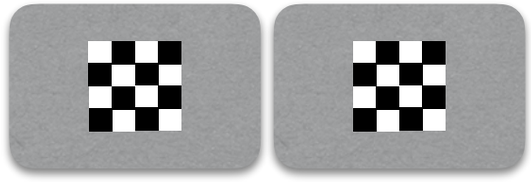


6s 6s

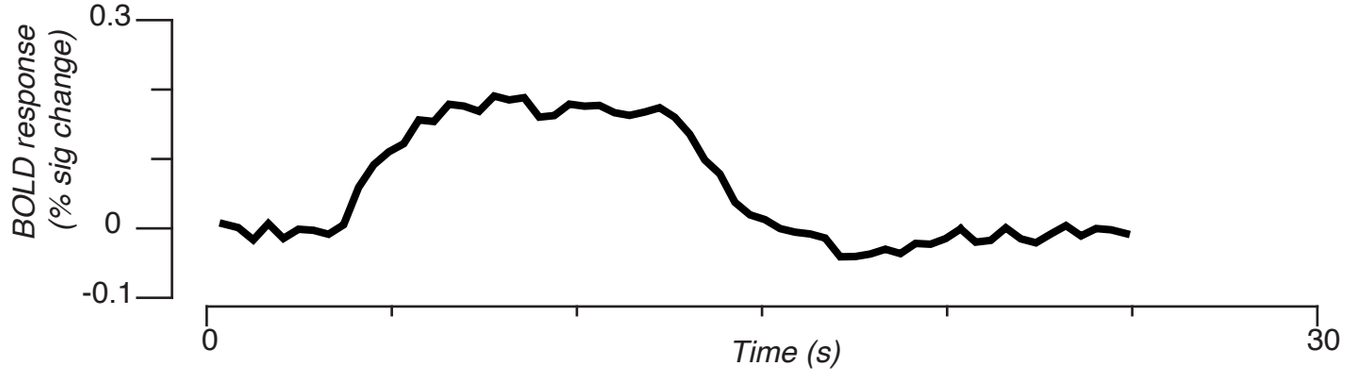
f(b)



a+b



f(a+b)



Evidence for temporal linearity of BOLD

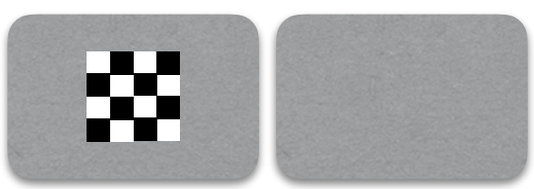
$$f(a+b) = f(a) + f(b)$$

Superposition principle:
Response to the sum of stimuli is the sum of responses to each stimulus presented individually

Stimulus

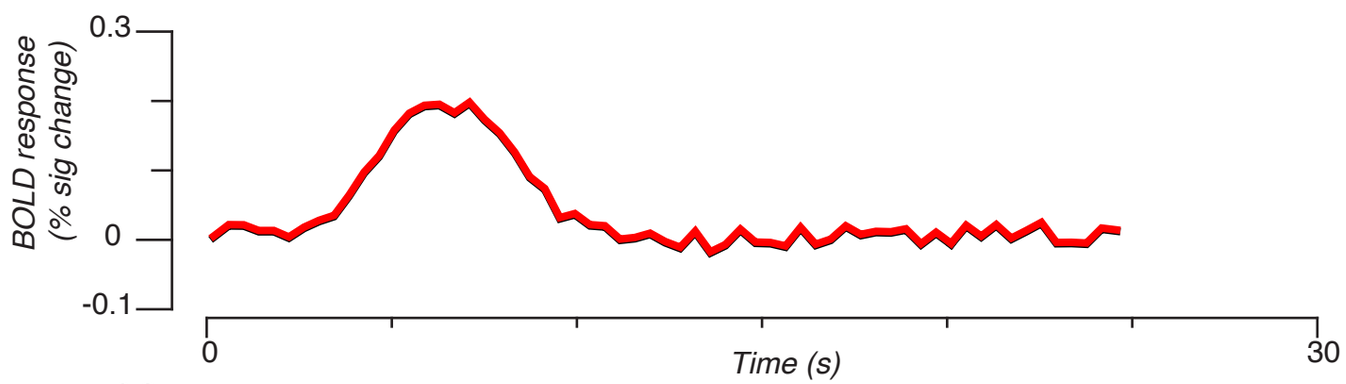
Response

a

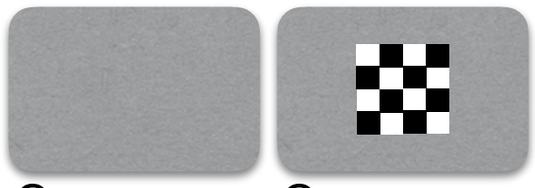


6s 6s

f(a)

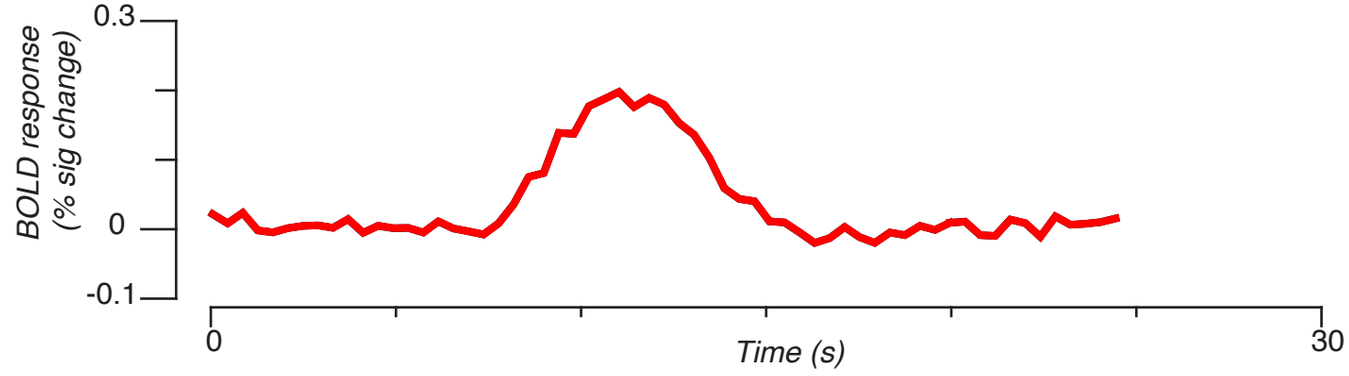


b

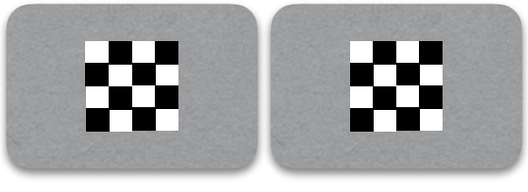


6s 6s

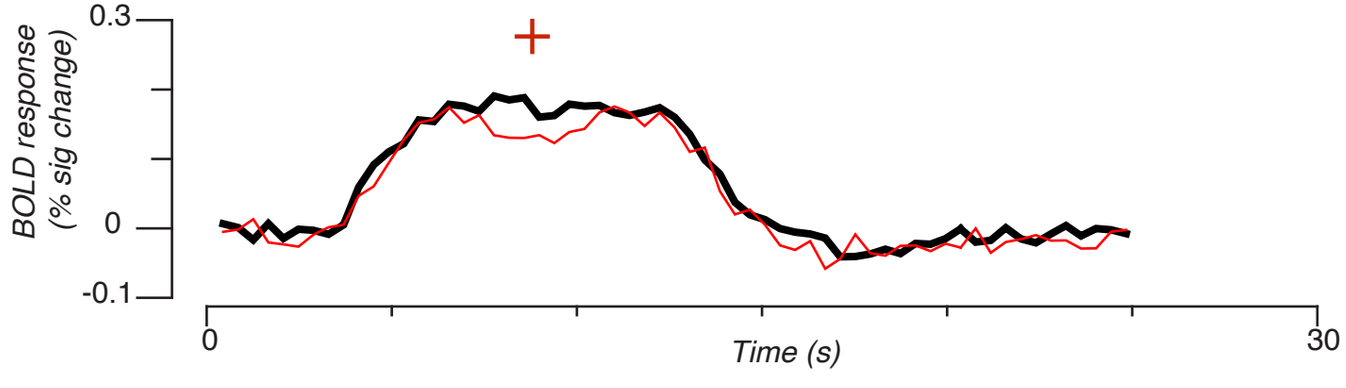
f(b)



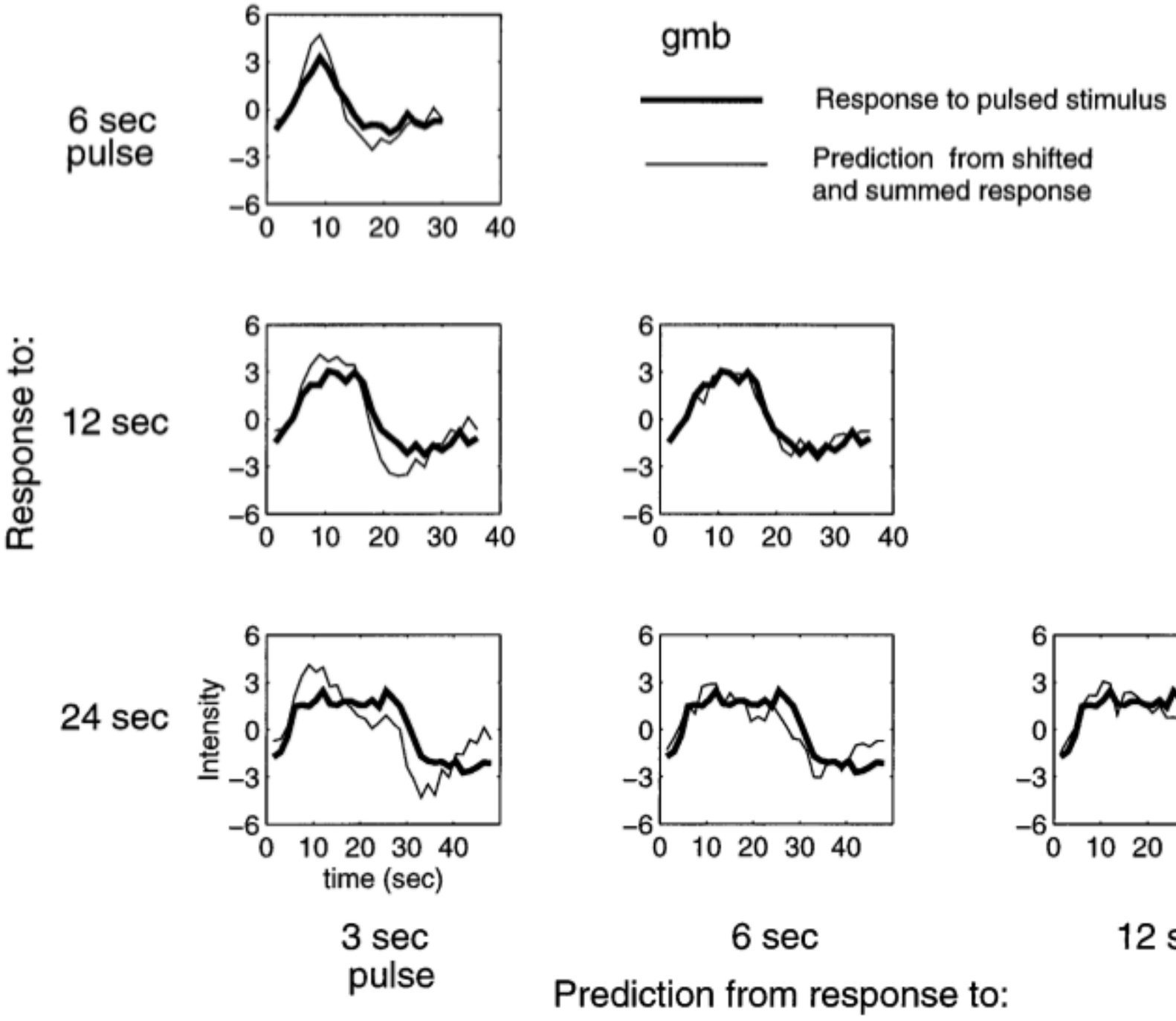
a+b



f(a+b)



Pretty good!



Bonyton GM, Engel SA, Glover GH & Heeger DJ (1996) Journal of Neuroscience 16:4207-21

P-values are overrated

Statistical analyses are really models

Models have assumptions

You need to test those assumptions

Evaluate how well your model fits the data

Then you can interpret parameters

(well ok, here you might want to consider p-values)

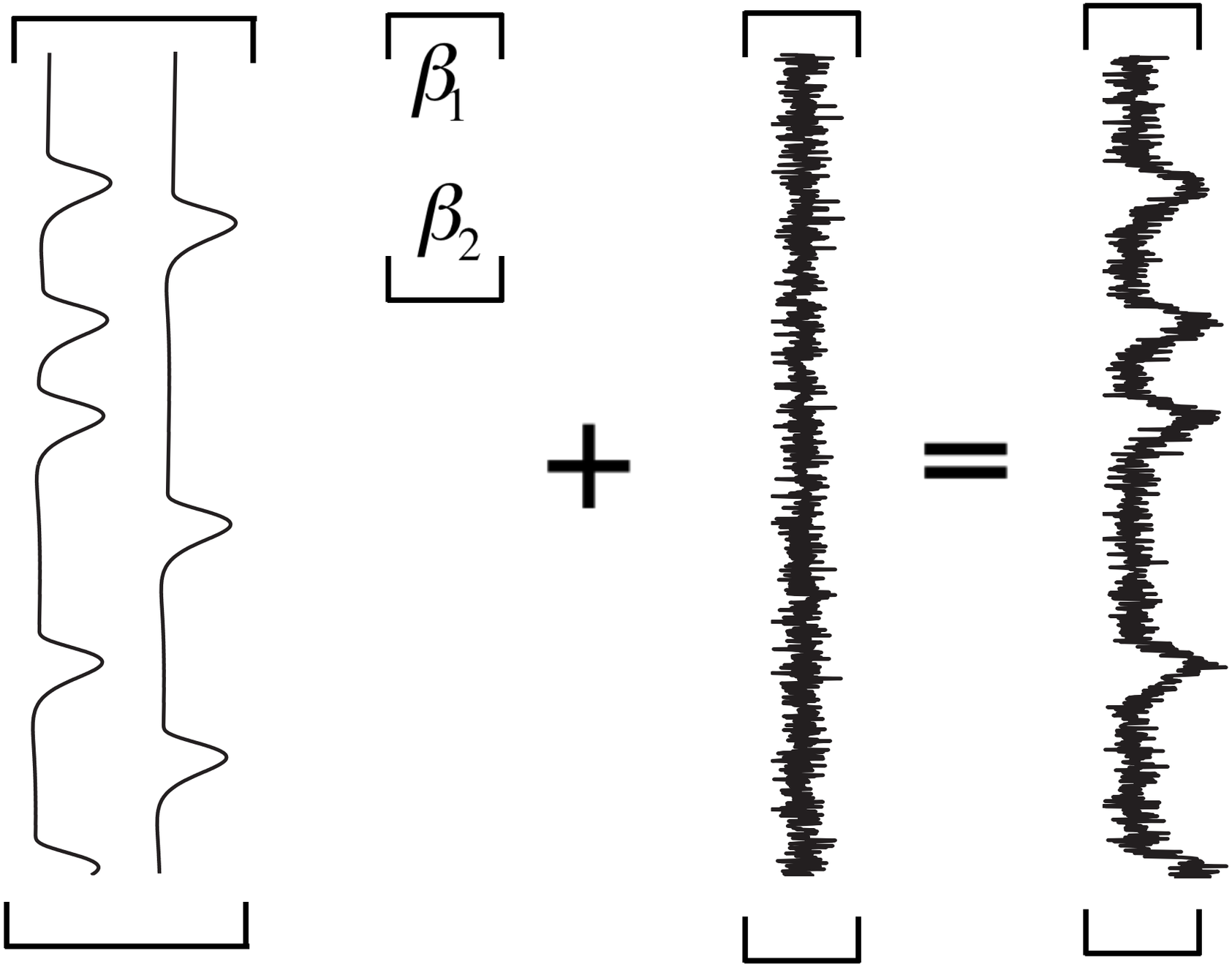
Evaluating the model: Model fit

```
fit_a %>%
  glance()
```

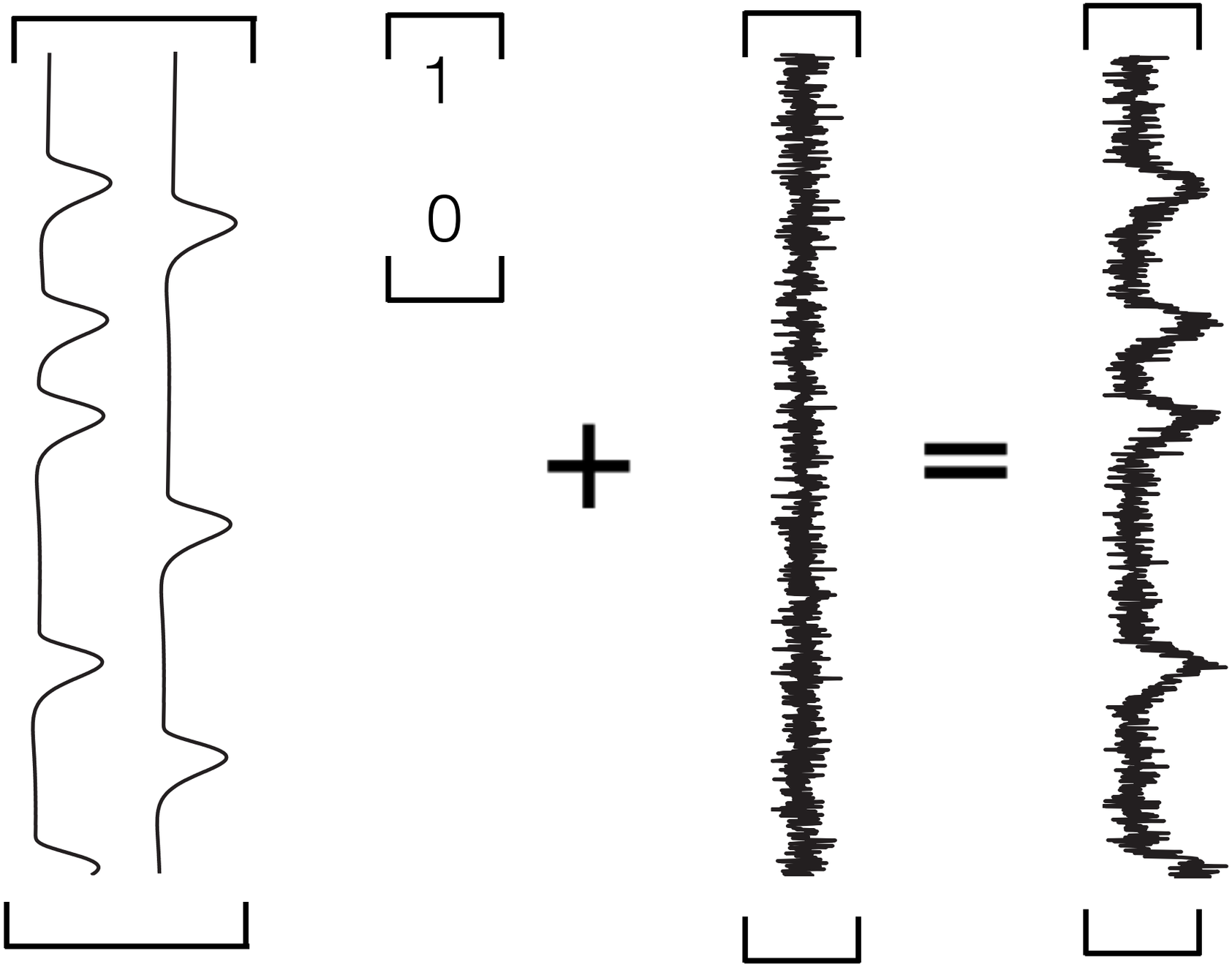
r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual
0.897	0.896	1.681	859.618	0	3	-386.197	780.394	793.587	556.914	197

`r.squared`	The percent of variance explained by the model
`adj.r.squared`	r.squared adjusted based on the degrees of freedom
`sigma`	The square root of the estimated residual variance
`statistic`	F-statistic
`p.value`	p-value from the F test, describing whether the full regression is significant
`df`	Degrees of freedom used by the coefficients
`logLik`	the data's log-likelihood under the model
`AIC`	the Akaike Information Criterion
`BIC`	the Bayesian Information Criterion
`deviance`	deviance
`df.residual`	residual degrees of freedom

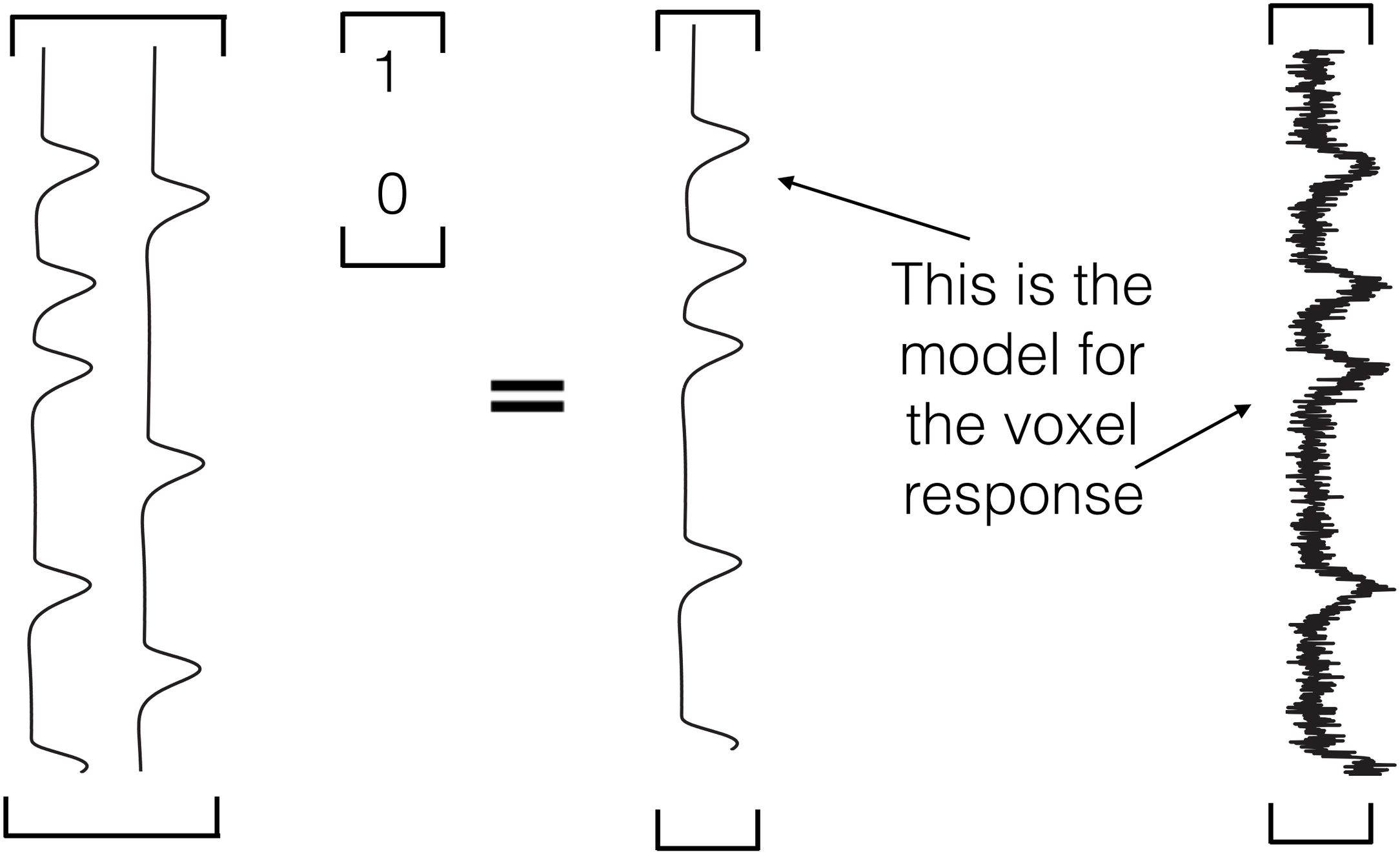
The General Linear Model



The General Linear Model

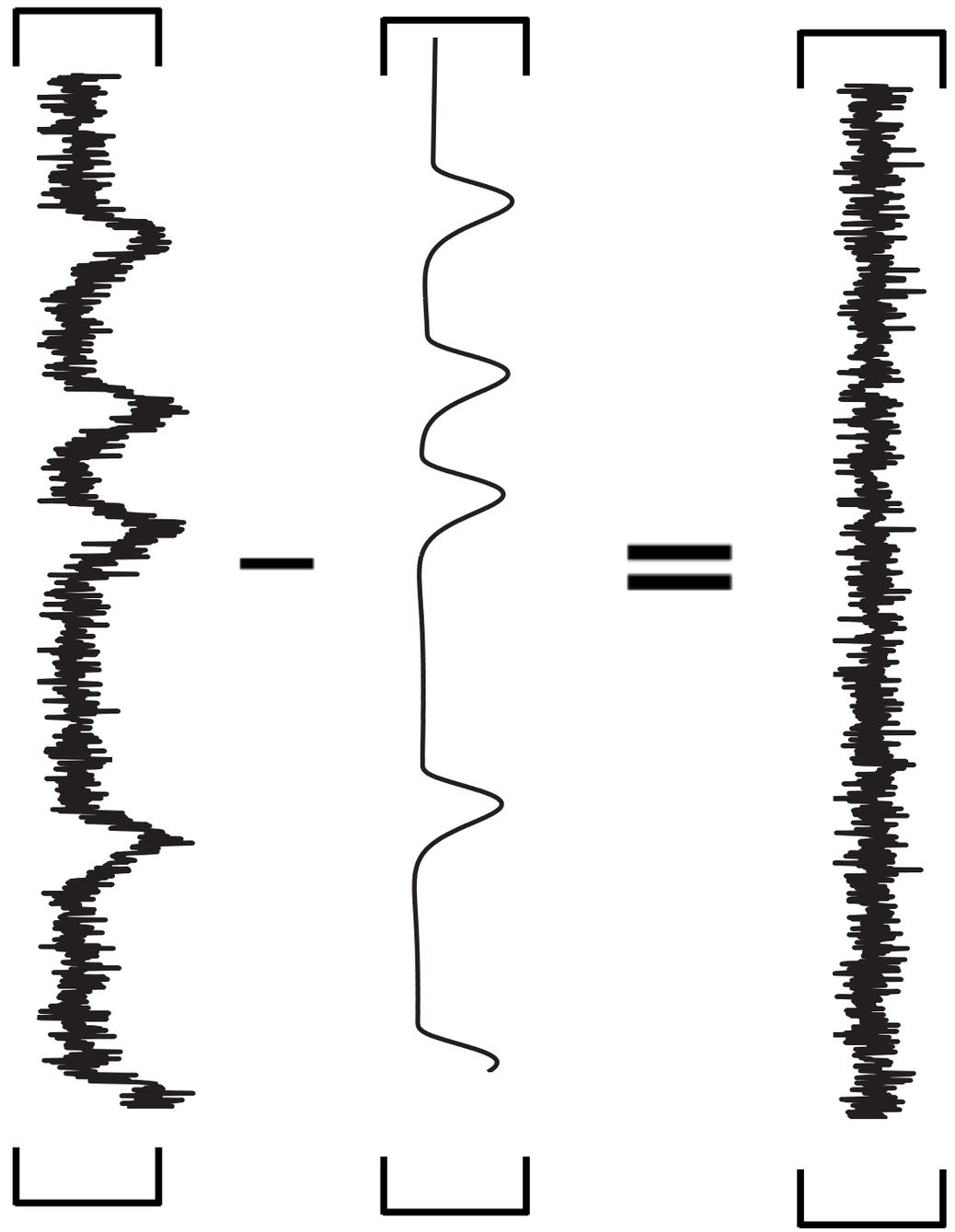


The General Linear Model



The General Linear Model

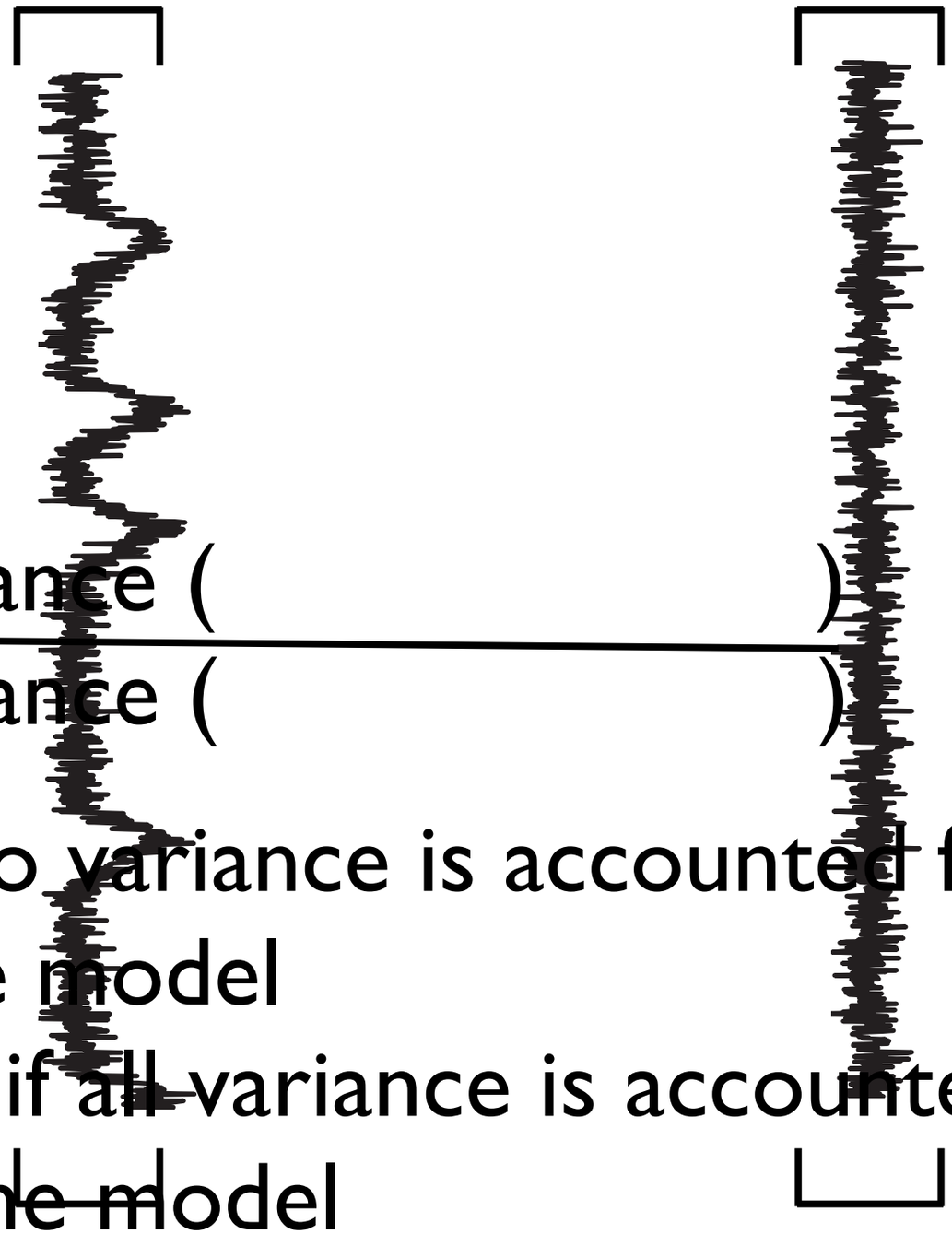
Subtract this model response from the actual measured response to get residual



Gardner, J.L., Sun, P., Waggoner, R.A., Ueno, K., Tanaka, K., and Cheng, K. (2005). Contrast adaptation and representation in human early visual cortex. Neuron *47*, 607–620.

The General Linear Model

Compute the ratio of the variance in the actual measured response to the residual variance



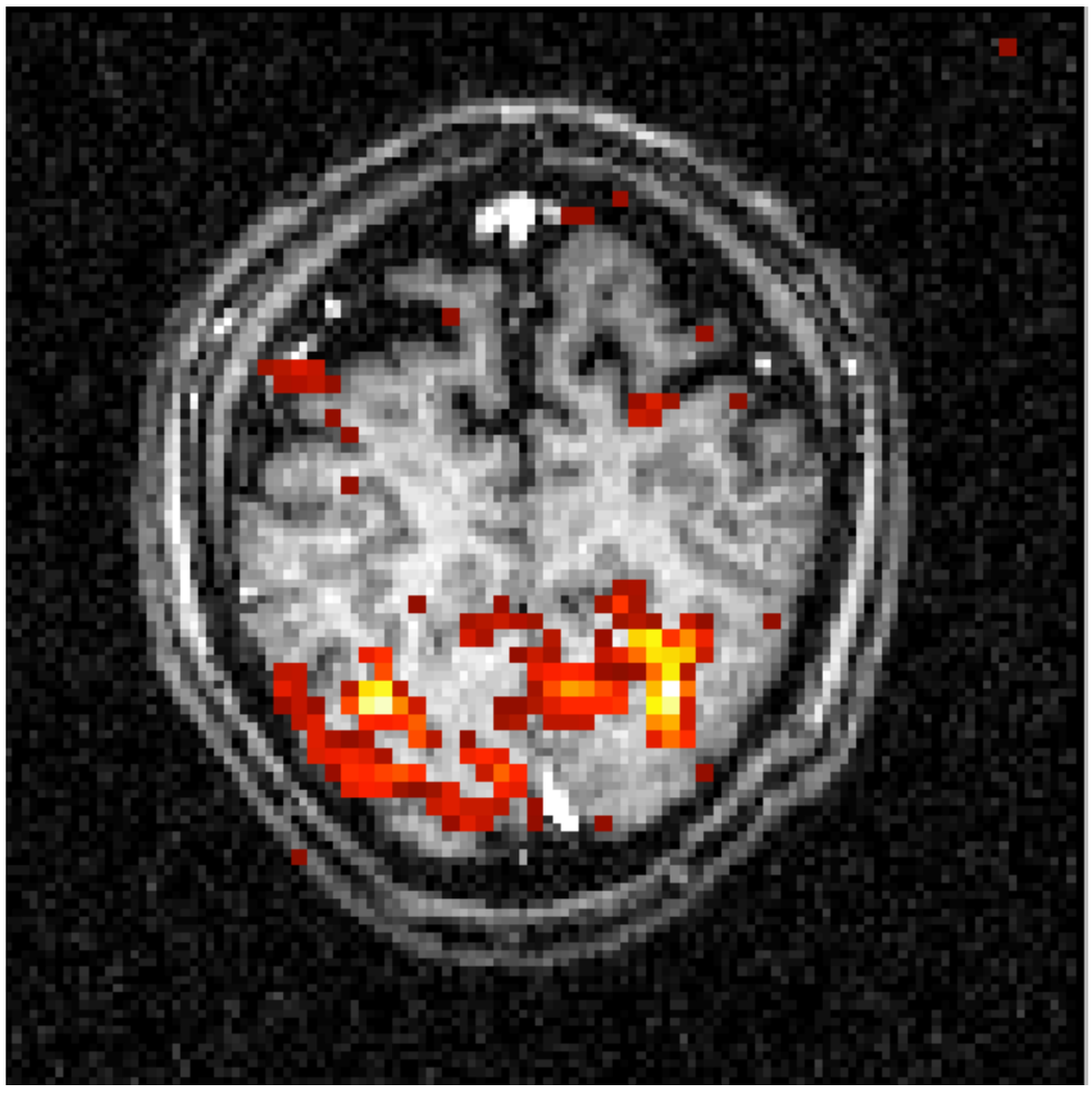
$$r^2 = 1 - \frac{\text{Variance (}\quad\text{)}}{\text{Variance (}\quad\text{)}}$$

r^2 is 0 (bad model fit) if no variance is accounted for by the model

r^2 is 1 (perfect model fit) if all variance is accounted for by the model

Gardner, J.L., Sun, P., Waggoner, R.A., Ueno, K., Tanaka, K., and Cheng, K. (2005). Contrast adaptation and representation in human early visual cortex. Neuron *47*, 607–620.

The General Linear Model



r^2 map

Gardner, J.L., Sun, P., Waggoner, R.A., Ueno, K., Tanaka, K., and Cheng, K. (2005). Contrast adaptation and representation in human early visual cortex. Neuron *47*, 607–620.

P-values are overrated

Statistical analyses are really models

Models have assumptions

You need to test those assumptions

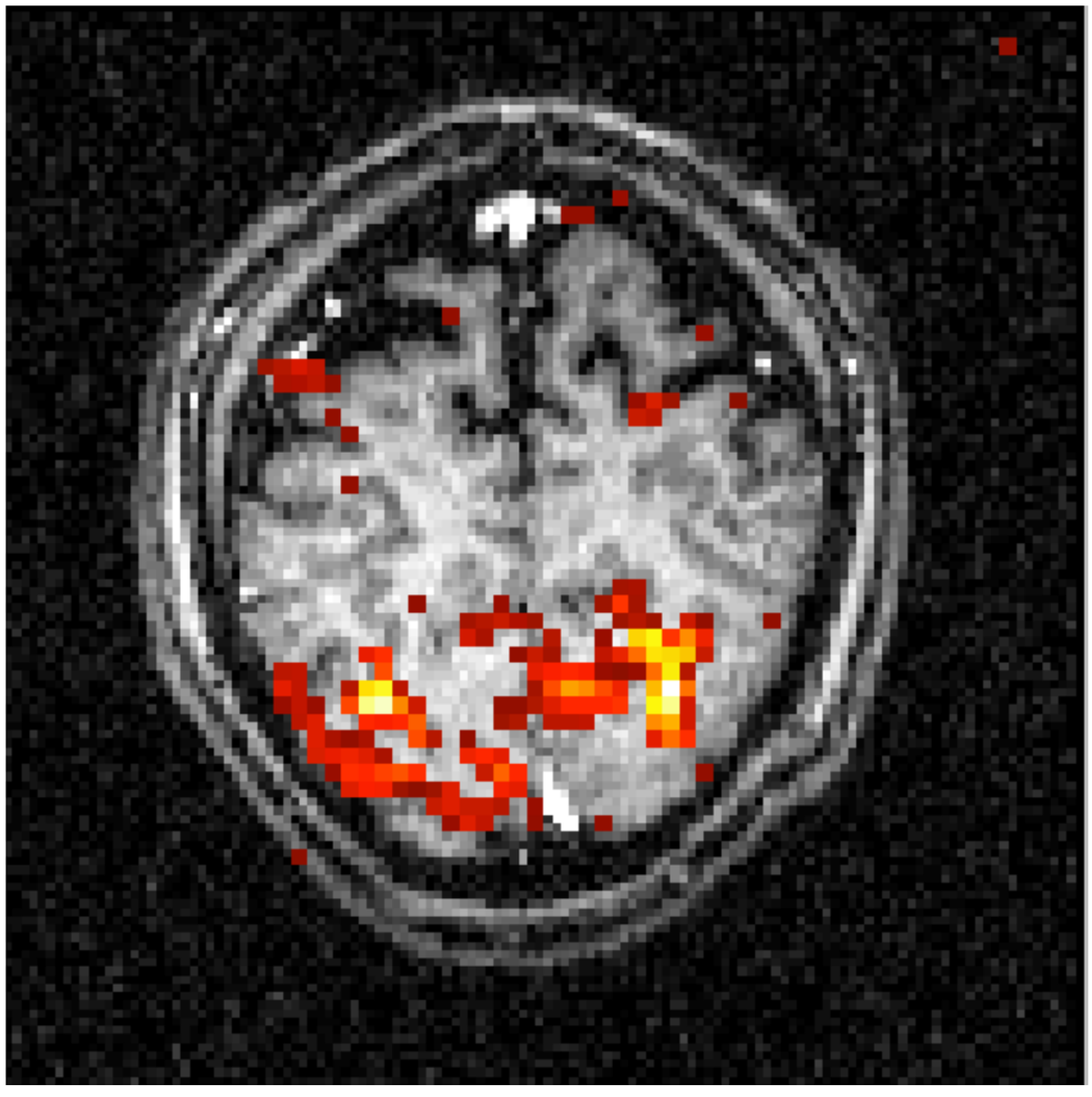
Evaluate how well your model fits the data

Then you can interpret parameters

(well ok, here you might want to consider p-values)

The General Linear Model

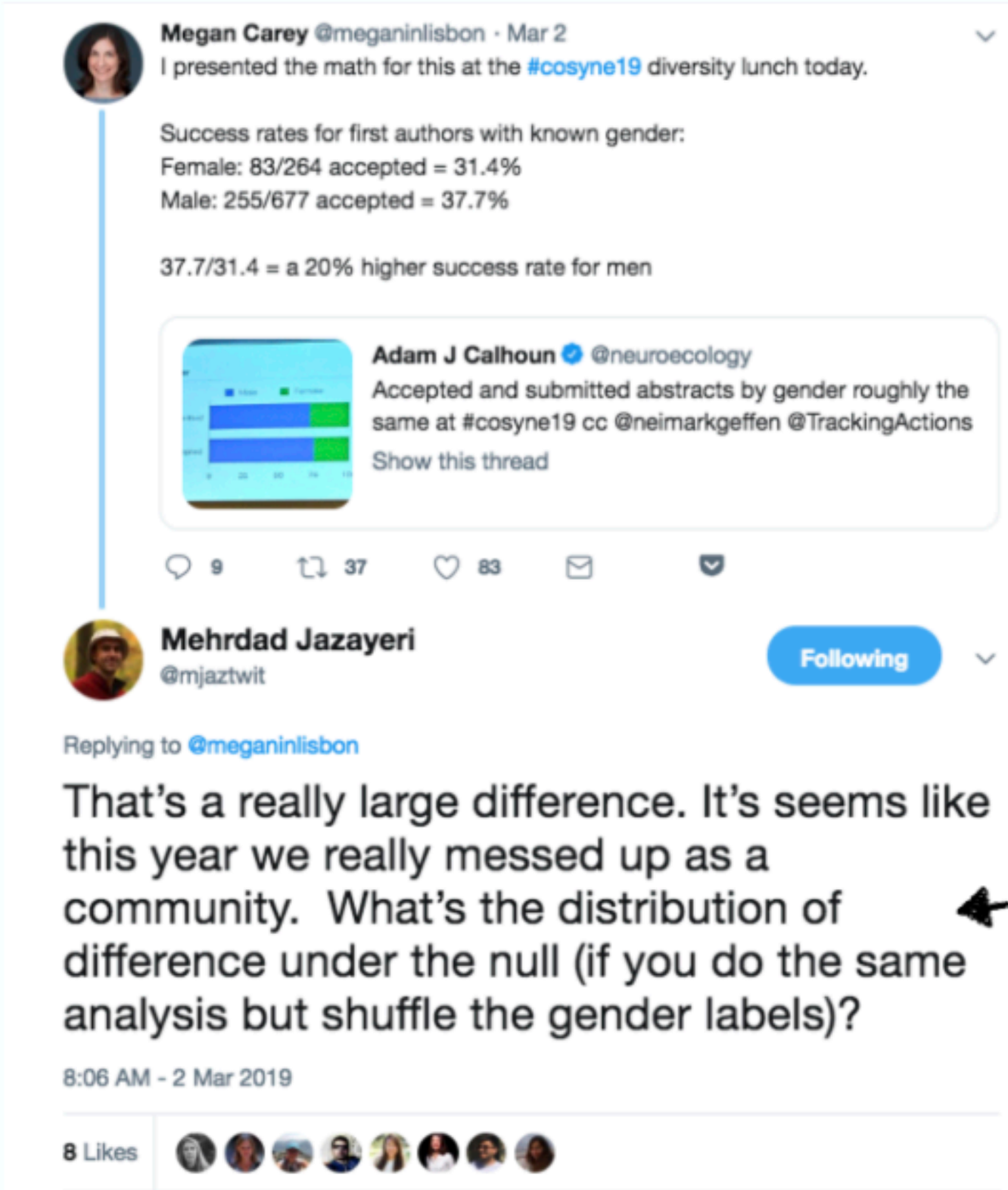
What is a significant r^2?



r^2 map

Gardner, J.L., Sun, P., Waggoner, R.A., Ueno, K., Tanaka, K., and Cheng, K. (2005). Contrast adaptation and representation in human early visual cortex. Neuron *47*, 607–620.

Bias in conference admission?



permutation test!

The General Linear Model

What is a significant r^2?

Permutation test.

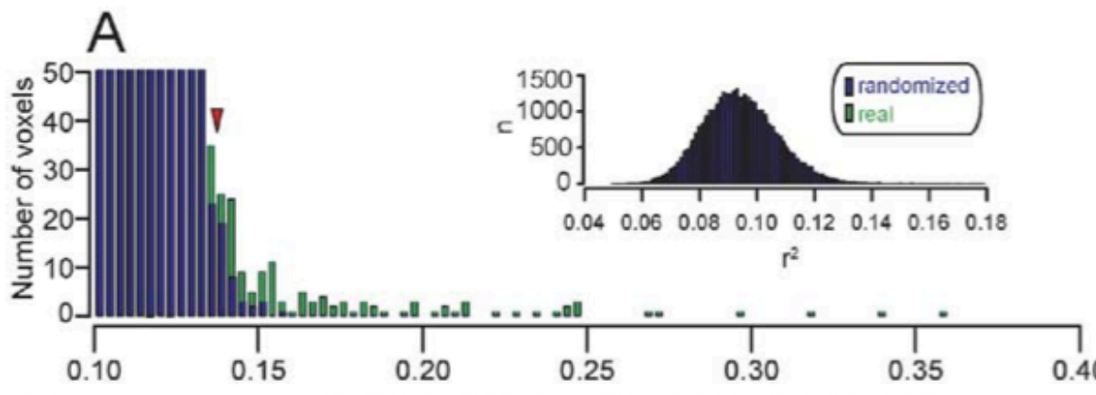
Randomize stimulus times

Gardner, J.L., Sun, P., Waggoner, R.A., Ueno, K., Tanaka, K., and Cheng, K. (2005). Contrast adaptation and representation in human early visual cortex. Neuron *47*, 607–620.

The General Linear Model

What is a significant r^2?

Permutation test.



Gardner, J.L., Sun, P., Waggoner, R.A., Ueno, K., Tanaka, K., and Cheng, K. (2005). Contrast adaptation and representation in human early visual cortex. Neuron *47*, 607–620.

P-values are overrated

Statistical analyses are really models

Models have assumptions

You need to test those assumptions

Evaluate how well your model fits the data

Then you can interpret parameters

(well ok, here you might want to consider p-values)