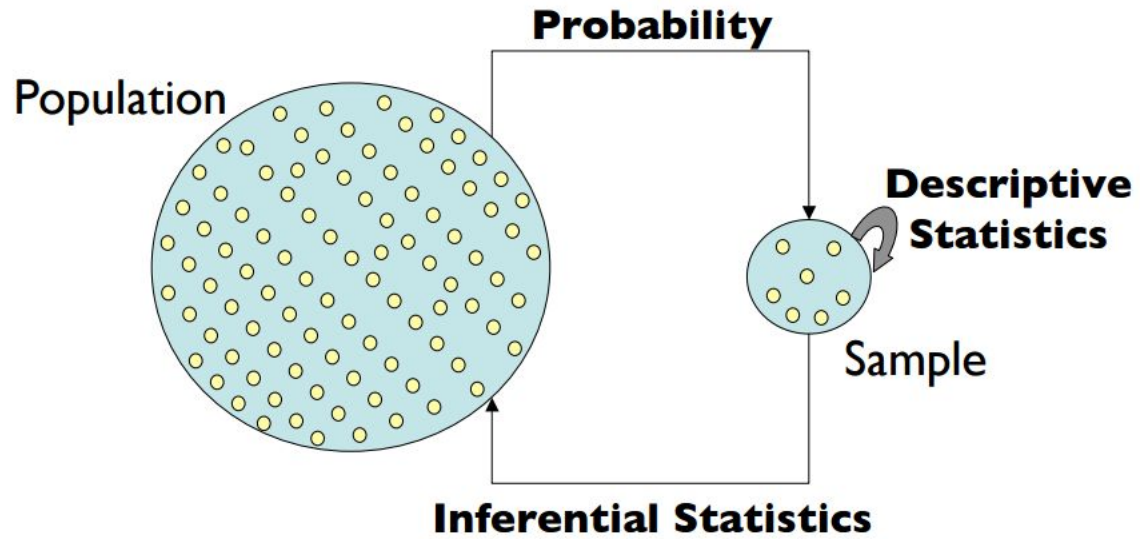
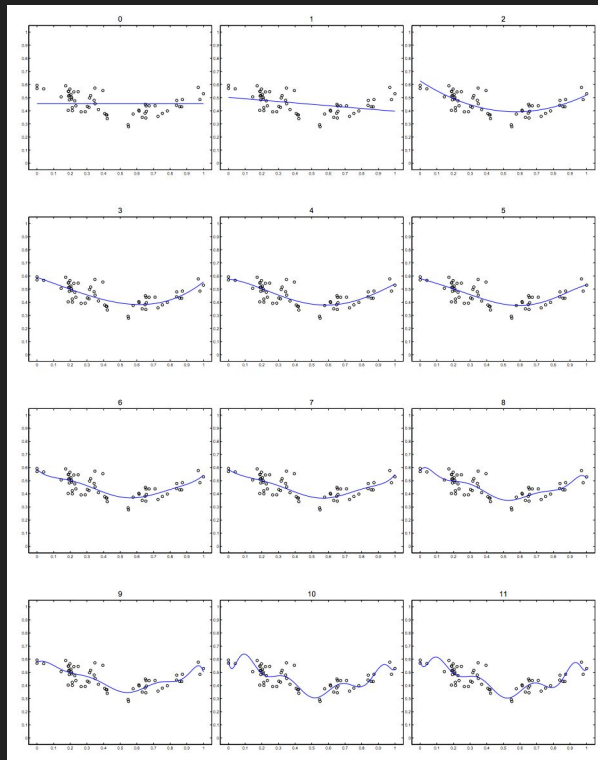


# Unbiased Decoding Through cross-validation permutation tests & the bootstrap

Patrick Beukema  
Ph.D. Candidate, CNUP & CNBC  
Tuesday, October 25



# Motivation: Example 1



Which is the best curve  
for the data?

# Additional applications of c-val

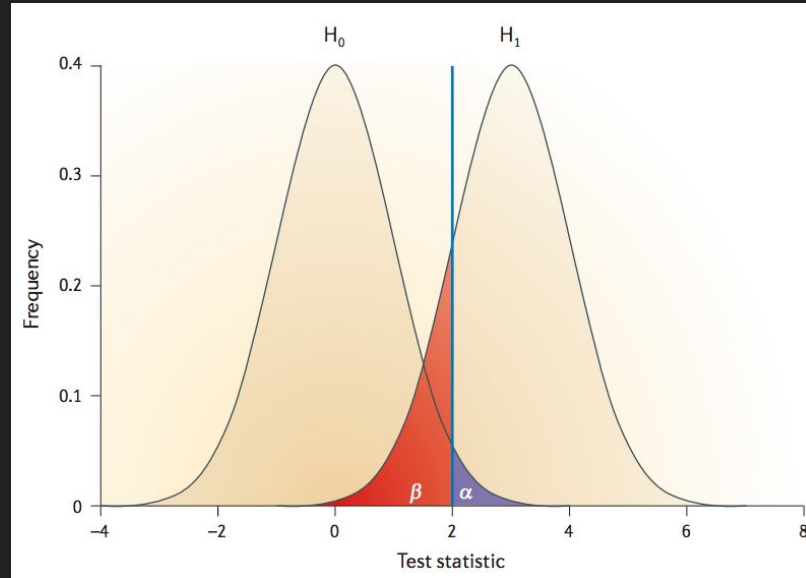
If I do clustering, how many clusters should I use?

If I fit a polynomial, what should its order be?

Should I use a gaussian, polynomial or linear kernel  
SVM?

# Example 2

Suppose that you are given a test statistic, how do you compute its significance?



Traditional way: Use Null distribution.

# But what if null is unknown?

Answer: Permutation Test

Permutation tests allow one to estimate the null distribution by constructing permuted samples for the data (more later)

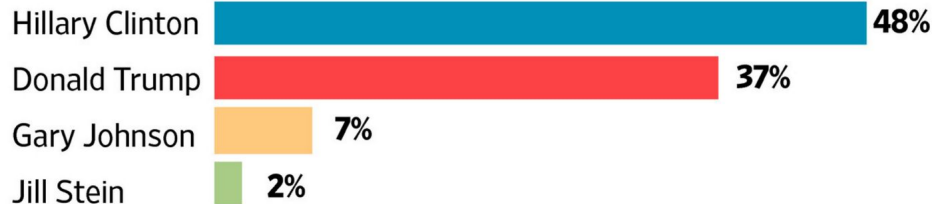
# How confident are we in this result?

Clinton Extends Lead to 11 Points

## A Clinton Advantage

Hillary Clinton leads Donald Trump by 11 percentage points in the latest WSJ/NBC News poll.

### Among likely voters:



Think about how you would answer this question, we will return to it later in the lecture.



# Motivation

- Less interested in fitting data to model than the truth
- Model fit statistics e.g  $R^2$   $\neq$  robustness of model
- Cross-validation remains gold standard (in many instances) throughout neuroscience and psychology
- Permutation tests enable hypothesis testing when population distribution is unknown or intractable
- Dim and size of data is  $\uparrow$  these issues are becoming more important

# Outline of Today's Lecture

1. Cross-validated distance measures are less biased than conventional estimates
2. Permutation tests enable estimating population distribution when data is not normal or unknown
3. Bootstrapping enables quantification of uncertainty

# Today's Goals

Better understanding of

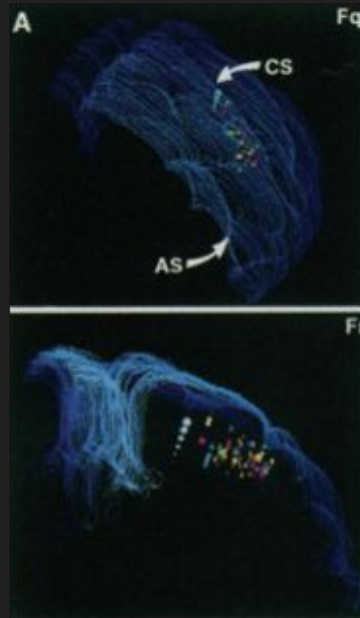
- 1) cross-validation,
- 2) permutation tests,
- 3) & bootstrap

through applications to Representational Similarity Analysis (RSA)

# Topic 1: Cross-validation: general idea

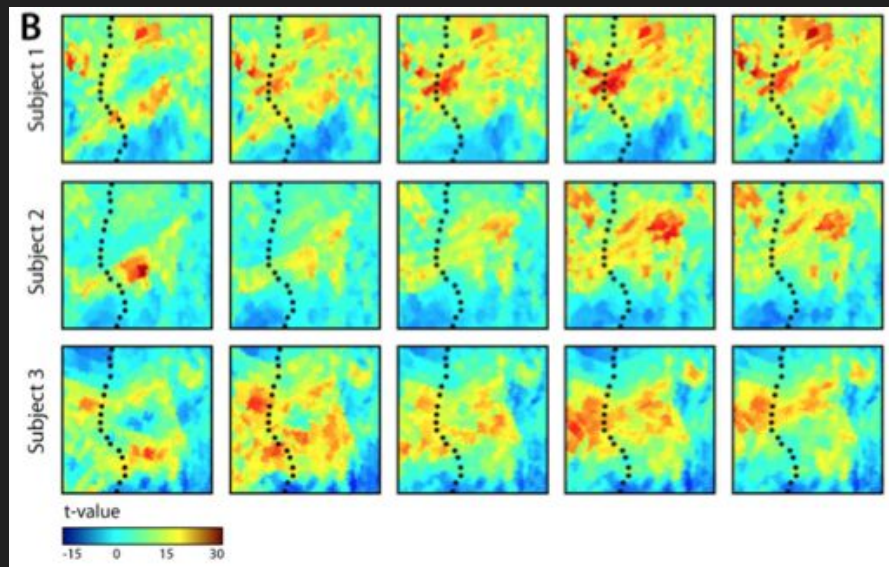
1. Split data into partitions (ideally iid)
  - Many sampling approaches (k-fold, MC, LVO)
2. Compute moments (1st, 2nd, 3rd,...) across partitions
3. Resulting statistic is less biased

# RSA: Application of Cross validation in MVPA



Problem: Lack of columnar organization and overlapping neuronal activity in M1 suggest that spatial analysis has limited utility.

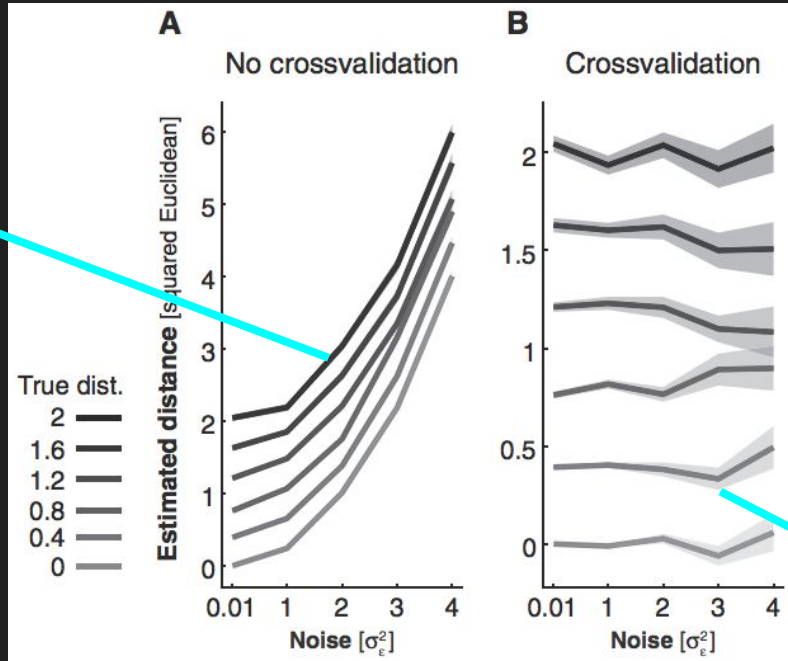
## Average activation patterns of finger movements across 3 example subjects



**While activity across subjects is different structure appears similar**

# Example of cross-validation to estimate distances

Without  
cross-validation,  
estimated distance  
>>  
true distance  
as noise increases



- Real noise points in random directions.
  - Cross-validation orthogonalizes those directions in higher dimensional (k-fold) space
- Estimated distance = true distance when using cross-validation

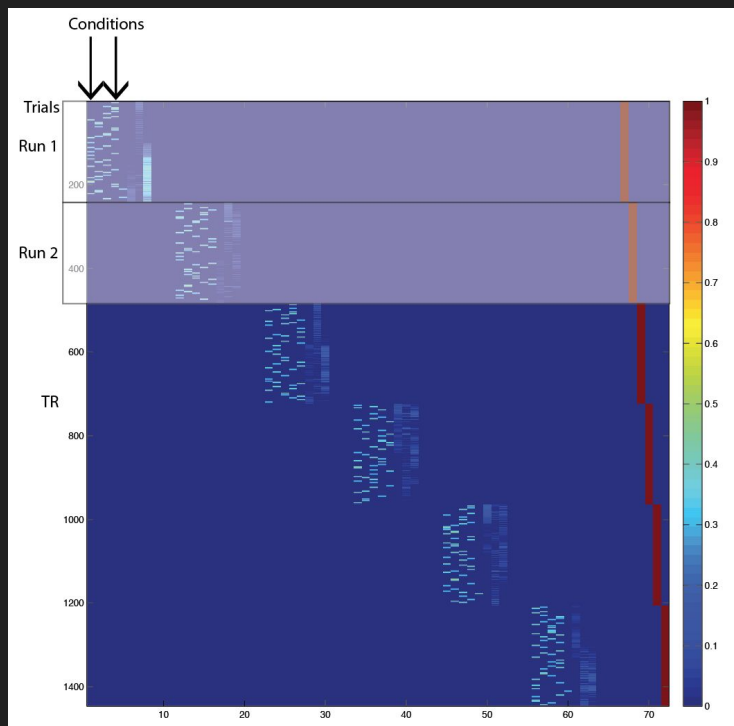


# Cross-validation across runs of an fMRI Experiment

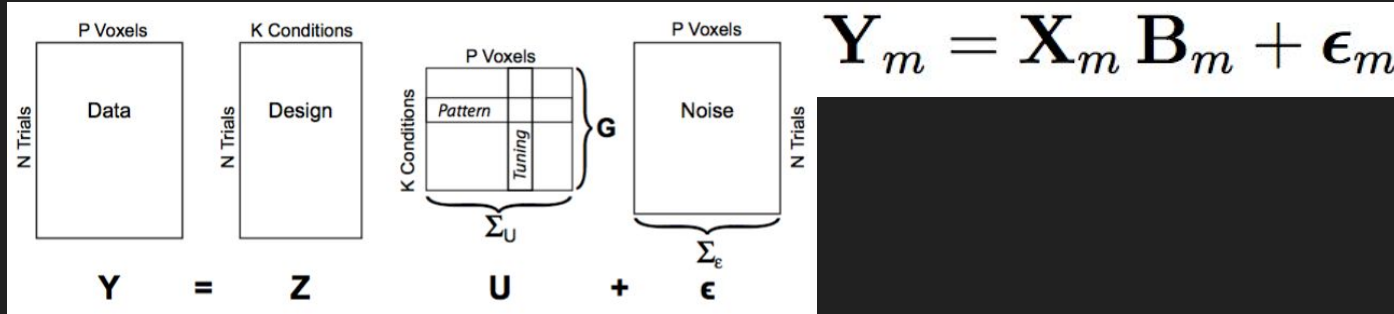
MRI data collected Pre and Post Training

- RSA Scan: 6 RER runs/session
- RSA Scans are independently collected.

# Design Matrix



# Cross validation across runs of an fMRI (or other neuroimaging) experiment



# Cross validation across runs of an fMRI (or other neuroimaging) experiment

The diagram illustrates the cross-validation setup for fMRI data. It shows the relationship between Data (Y), Design (Z), and Noise (epsilon) matrices, and the resulting model equation  $\mathbf{Y}_m = \mathbf{X}_m \mathbf{B}_m + \epsilon_m$ .

The Data matrix  $\mathbf{Y}$  (N Trials x P Voxels) is equal to the Design matrix  $\mathbf{Z}$  (N Trials x K Conditions) plus the Noise matrix  $\epsilon$  (N Trials x P Voxels).

The Design matrix  $\mathbf{Z}$  is partitioned into two components:  $\mathbf{U}$  (N Trials x P Voxels) and  $\mathbf{G}$  (N Trials x P Voxels). The covariance matrix  $\Sigma_U$  is associated with  $\mathbf{U}$ , and the covariance matrix  $\Sigma_\epsilon$  is associated with  $\epsilon$ .

The model equation is:

$$\mathbf{Y}_m = \mathbf{X}_m \mathbf{B}_m + \epsilon_m$$

The estimated parameter vector  $\hat{\mathbf{B}}_m$  is given by:

$$\hat{\mathbf{B}}_m = (\mathbf{X}_m^T \hat{\Sigma}_T^{-1} \mathbf{X}_m)^{-1} \mathbf{X}_m^T \hat{\Sigma}_T^{-1} \mathbf{Y}_m$$

# Cross validation across runs of an fMRI (or other neuroimaging) experiment

The diagram illustrates the cross-validation setup for fMRI data. It shows the relationship between Data (Y), Design (Z), and Noise (epsilon) matrices, and the resulting model parameters (B\_m) and cross-validation error (delta\_j,m).

**Data Matrix (Y):** Dimensions are N Trials (vertical) and P Voxels (horizontal).

**Design Matrix (Z):** Dimensions are N Trials (vertical) and K Conditions (horizontal).

**Model Matrix (U):** Dimensions are K Conditions (vertical) and P Voxels (horizontal). It is partitioned into a *Pattern* block and a *Tuning* block. A bracket labeled  $\Sigma_U$  spans the K Conditions dimension, and a bracket labeled  $\mathbf{G}$  spans the P Voxels dimension.

**Noise Matrix (epsilon):** Dimensions are N Trials (vertical) and P Voxels (horizontal). A bracket labeled  $\Sigma_\epsilon$  spans the P Voxels dimension.

The relationship is summarized by the equation:

$$\mathbf{Y} = \mathbf{Z} \mathbf{U} + \boldsymbol{\epsilon}$$

The model parameters are estimated as:

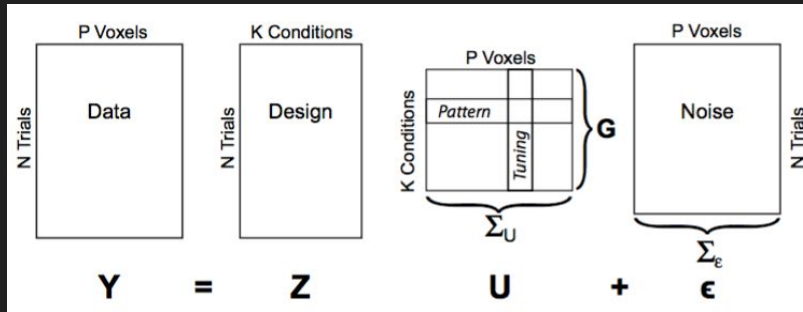
$$\mathbf{Y}_m = \mathbf{X}_m \mathbf{B}_m + \boldsymbol{\epsilon}_m$$

$$\hat{\mathbf{B}}_m = (\mathbf{X}_m^T \hat{\Sigma}_T^{-1} \mathbf{X}_m)^{-1} \mathbf{X}_m^T \hat{\Sigma}_T^{-1} \mathbf{Y}_m$$

The cross-validation error is defined as:

$$\hat{\delta}_{j,m} = \hat{u}_{i,m} - \hat{u}_{k,m}$$

# Cross validation across runs of an fMRI (or other neuroimaging) experiment



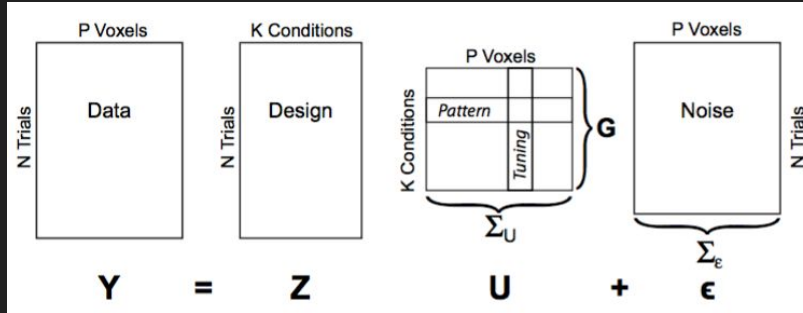
$$\mathbf{Y}_m = \mathbf{X}_m \mathbf{B}_m + \boldsymbol{\epsilon}_m$$

$$\hat{\mathbf{B}}_m = (\mathbf{X}_m^T \hat{\Sigma}_T^{-1} \mathbf{X}_m)^{-1} \mathbf{X}_m^T \hat{\Sigma}_T^{-1} \mathbf{Y}_m$$

$$\hat{\boldsymbol{\delta}}_{j,m} = \hat{\mathbf{u}}_{i,m} - \hat{\mathbf{u}}_{k,m}$$

$$\hat{d}_j = \frac{1}{MP} \sum_{i=1}^m \hat{\boldsymbol{\delta}}_{j,m} \hat{\boldsymbol{\delta}}_{j,\sim m}^T$$

# Cross validation across runs of an fMRI (or other neuroimaging) experiment



$$Y_m = X_m B_m + \epsilon_m$$

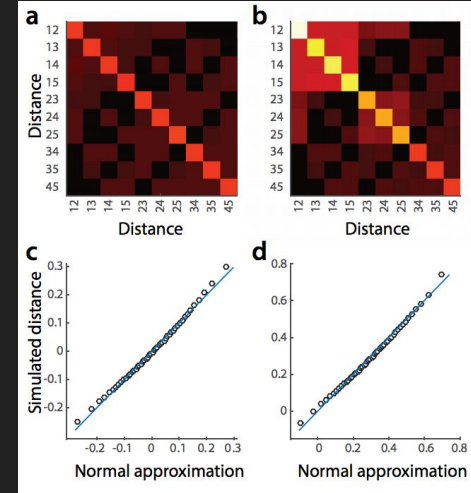
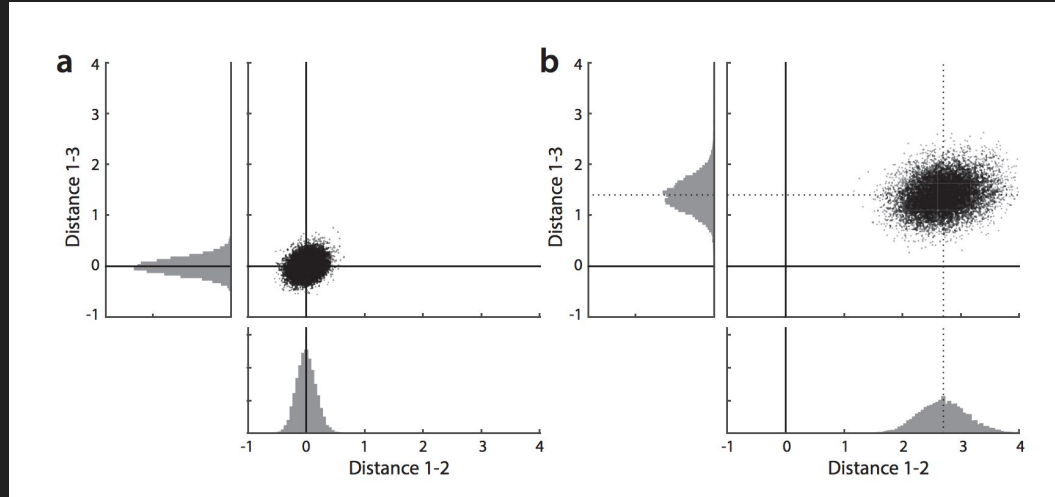
$$\hat{B}_m = (X_m^T \hat{\Sigma}_T^{-1} X_m)^{-1} X_m^T \hat{\Sigma}_T^{-1} Y_m$$

$$\hat{\delta}_{j,m} = \hat{u}_{i,m} - \hat{u}_{k,m}$$

$$\hat{d}_j = \frac{1}{MP} \sum_{i=1}^m \hat{\delta}_{j,m} \hat{\delta}_{j,\sim m}^T$$

$$H = \sum_{i \neq j}^K \hat{d}_{i,j}^2 / K(K-1)$$

# Joint sample distribution of distance estimates are multivariate normal\*



\*At least for simulated data

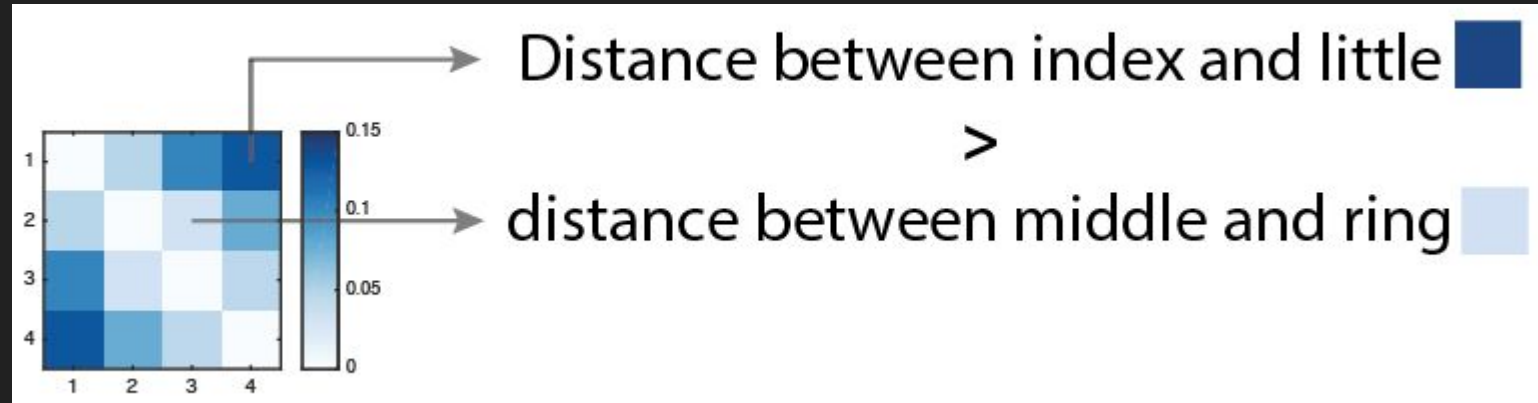


# Distances are unbiased

i.e.  $E(\hat{d}) = d$

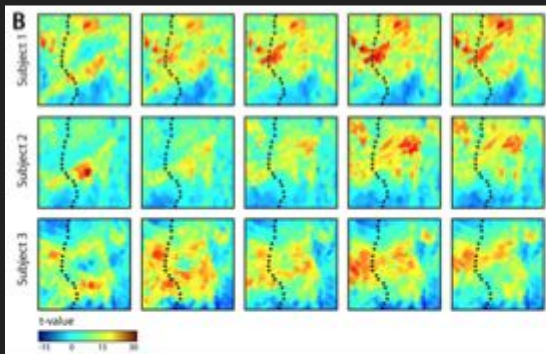
$$\begin{aligned} E(\hat{d}_j) &= E\left(\frac{1}{MP} \sum_{i=1}^m (\boldsymbol{\delta}_j + \boldsymbol{\eta}_{j,m})(\boldsymbol{\delta}_j + \boldsymbol{\eta}_{j,\sim m})^T\right) \\ &= E\left(\frac{1}{MP} \sum_{i=1}^m \boldsymbol{\delta}_j \boldsymbol{\delta}_j^T + \boldsymbol{\delta}_j \boldsymbol{\eta}_{j,m}^T + \boldsymbol{\delta}_j \boldsymbol{\eta}_{j,\sim m}^T + \boldsymbol{\eta}_{j,m} \boldsymbol{\eta}_{j,\sim m}^T\right) \\ &= \boldsymbol{\delta}_j \boldsymbol{\delta}_j^T / P = d_j. \end{aligned}$$

# Visualizing distances in a representational dissimilarity matrix



# Multivariate GLM resolves structure

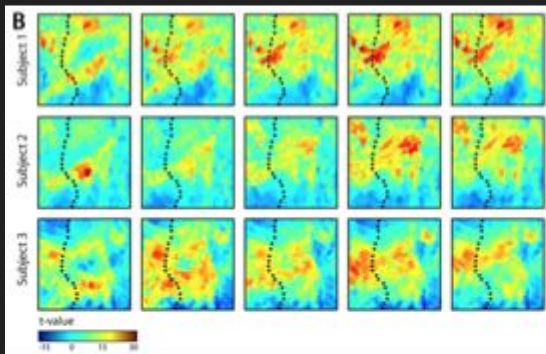
Univariate



No clear pattern

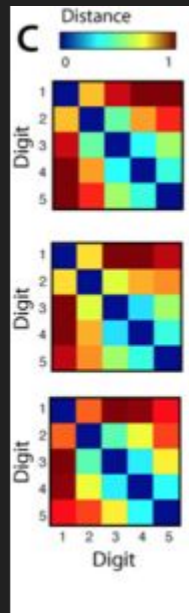
# Multivariate GLM resolves structure

# Univariate



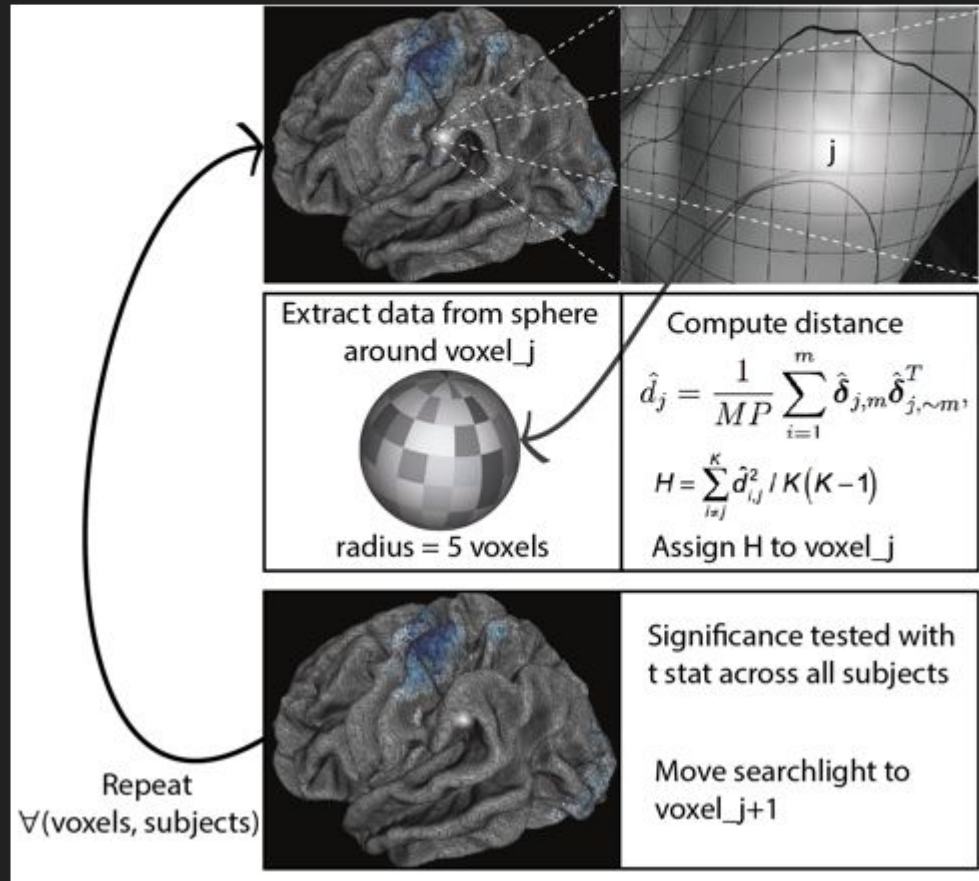
# No clear pattern

# Multivariate

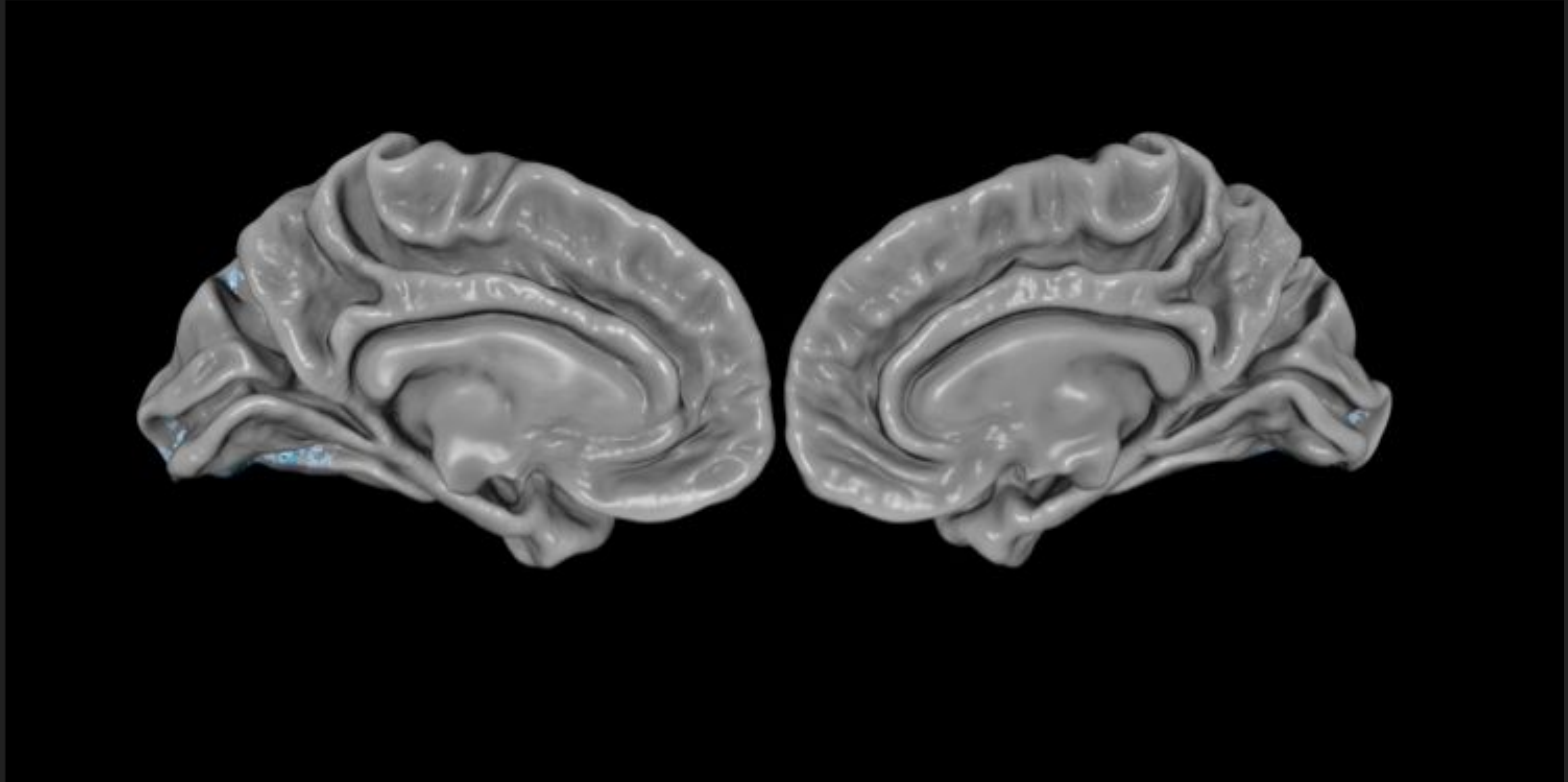


# Clear(er) pattern

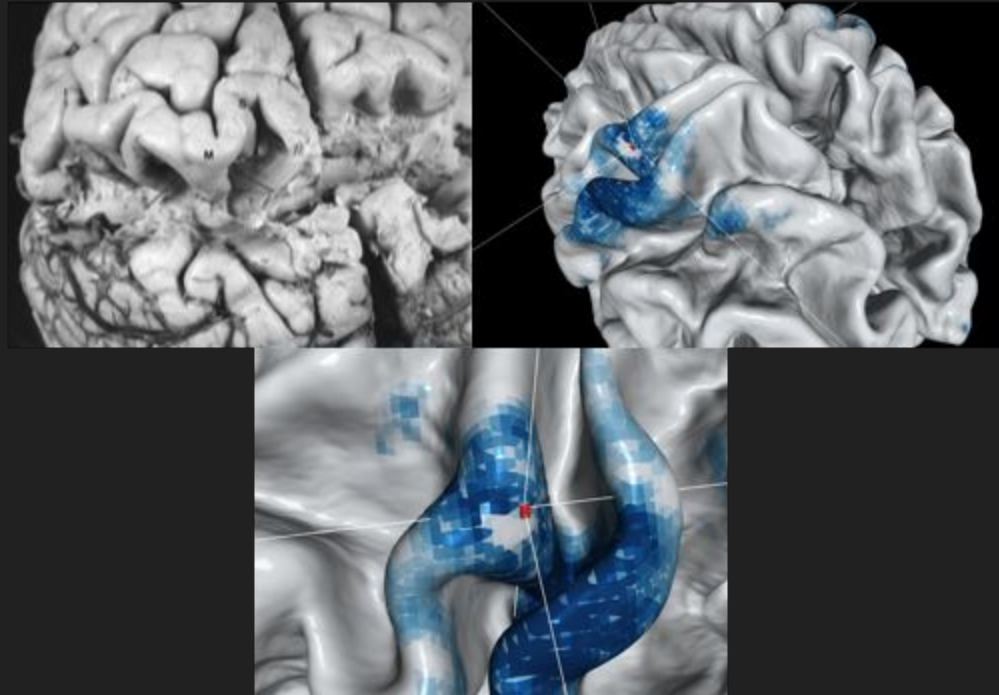
# Searchlight across entire brain

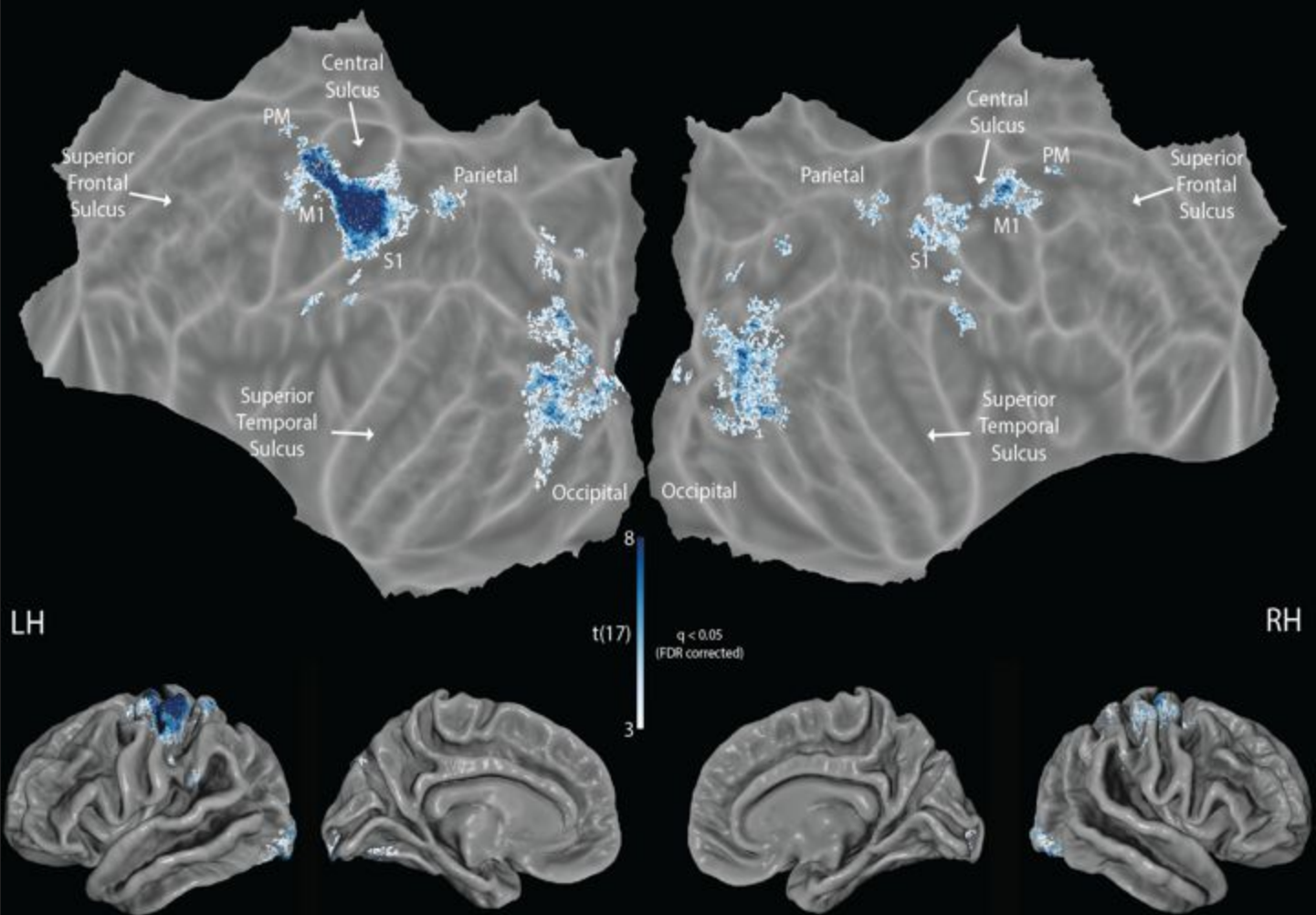


# Improve visualization by flattening cortex



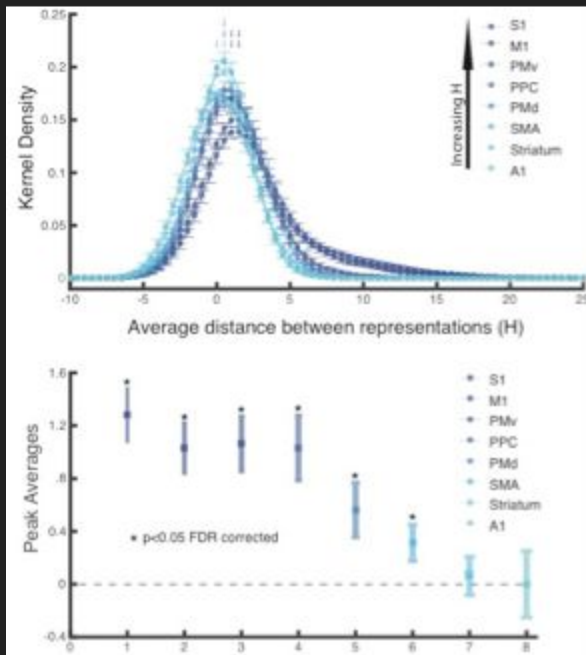
# High Anatomical Specificity





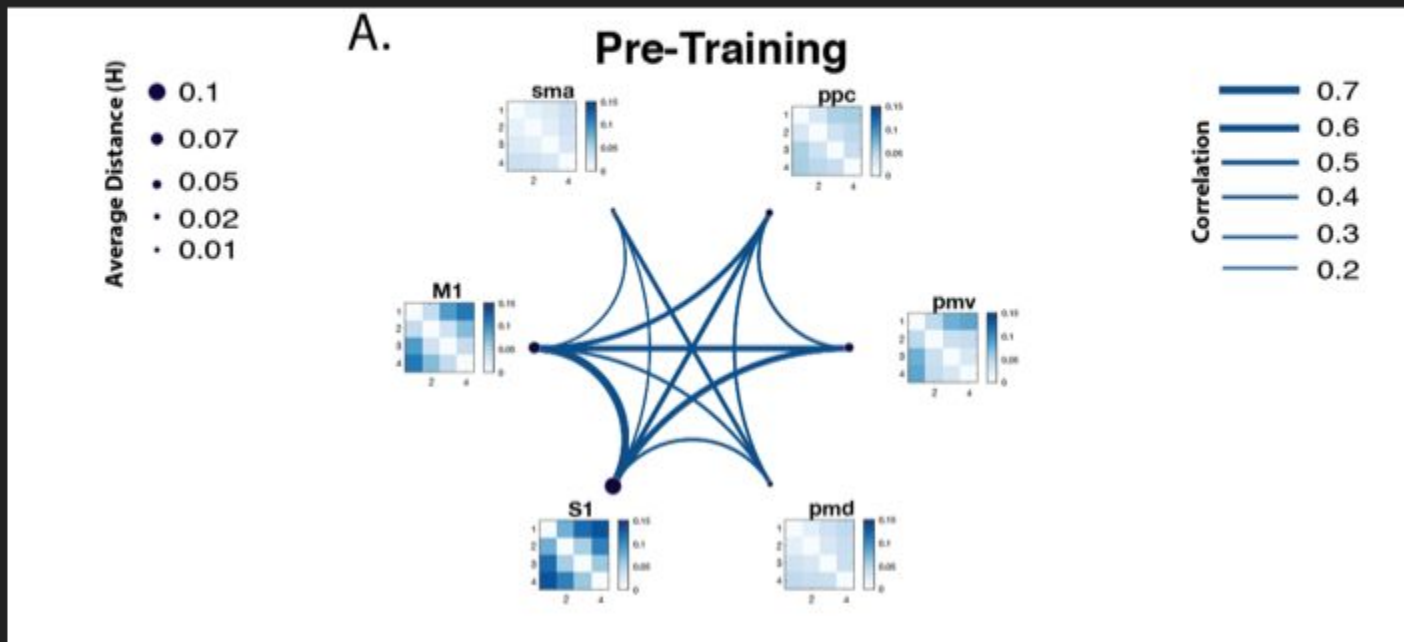


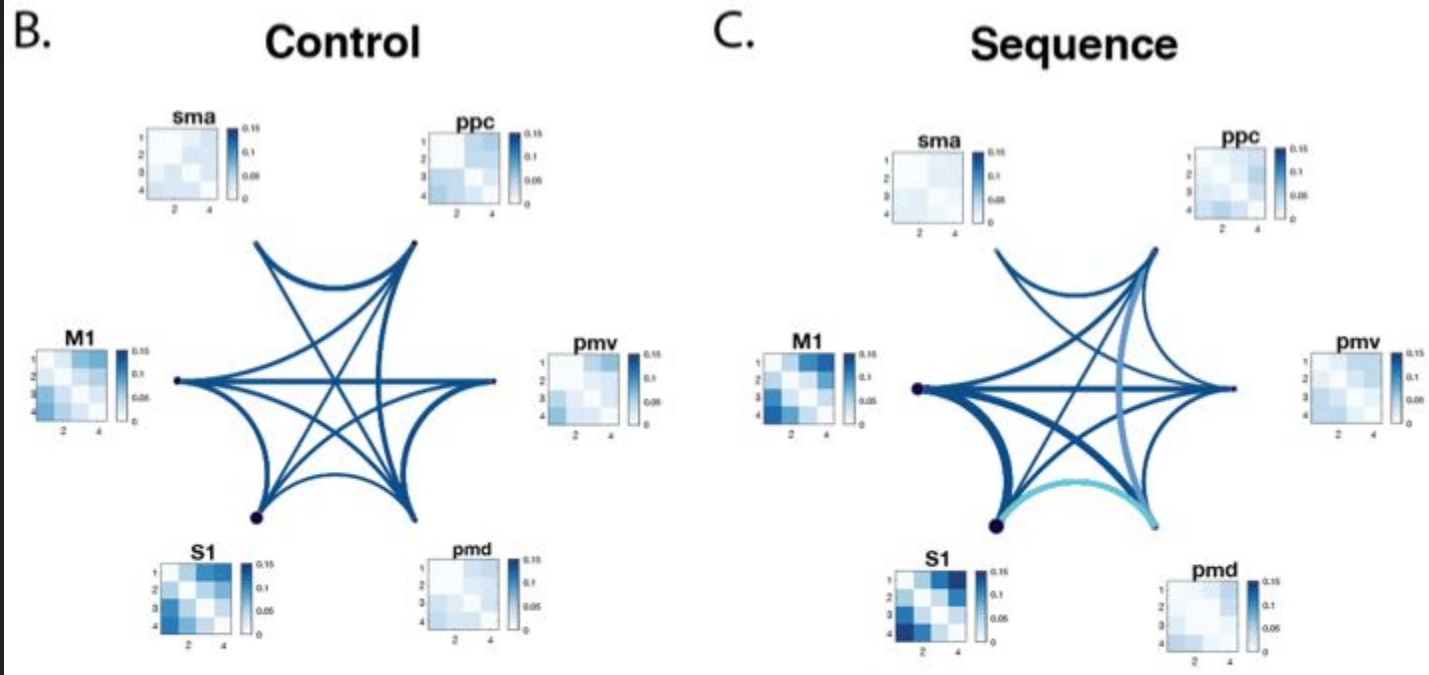
# ROI based decoding



Significant decoding in all regions of the motor network except striatum

# Structural network changes with learning





Significant group level differences in  
pattern structure correlation in premotor cortex

# Questions Regarding Cross-validation?

# RSA Based permutation test

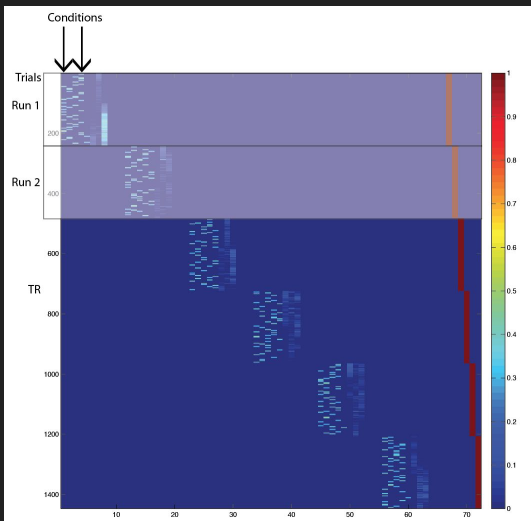
If we see a change in distances, suggestive of plasticity, how do we know that it is significant?

# Design of permutation test

1. Compute change in distance between finger representations Pre and Post
  - This is our observed test statistic
2. Compute permuted distance distribution
  - This is our 'chance' distribution
3. Check if observed is greater (or lesser than)

# Permute data in Pre and Post conditions

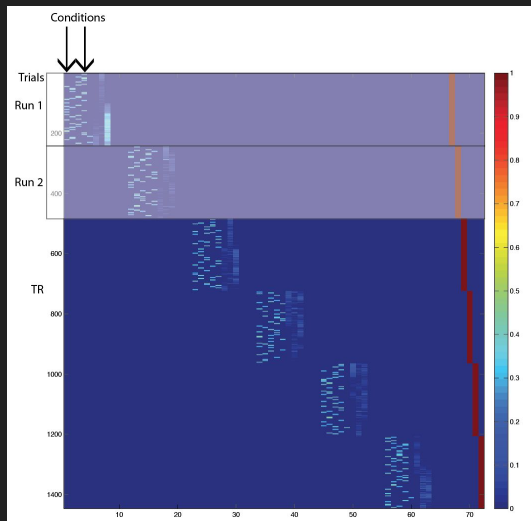
Pre Data



$$H = \sum_{i \neq j}^K \hat{d}_{i,j}^2 / K(K-1)$$

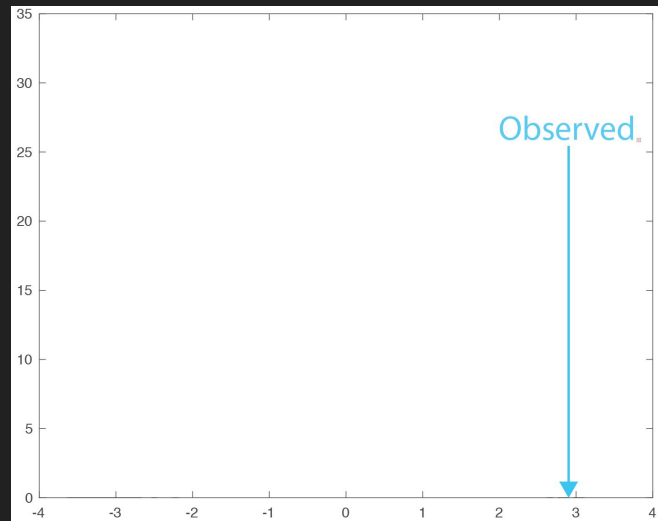
H Pre

Post Data



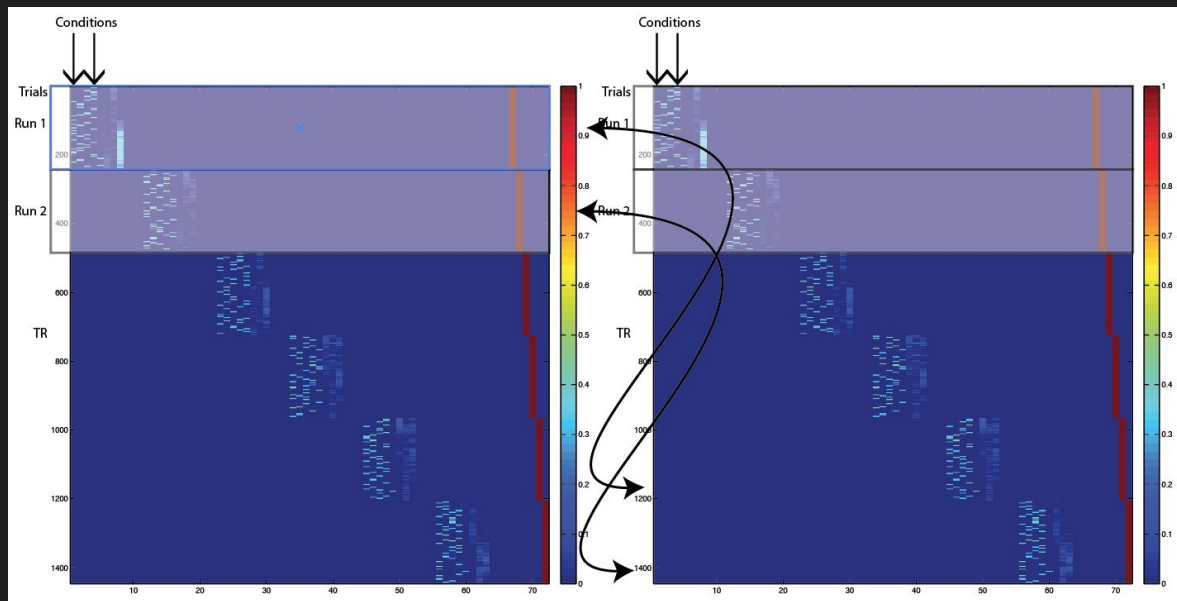
$$H = \sum_{i \neq j}^K \hat{d}_{i,j}^2 / K(K-1)$$

H post



Observed = H post - H Pre

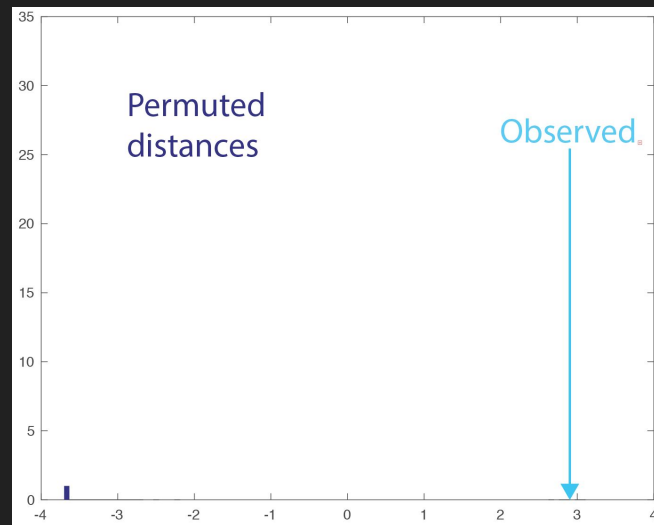
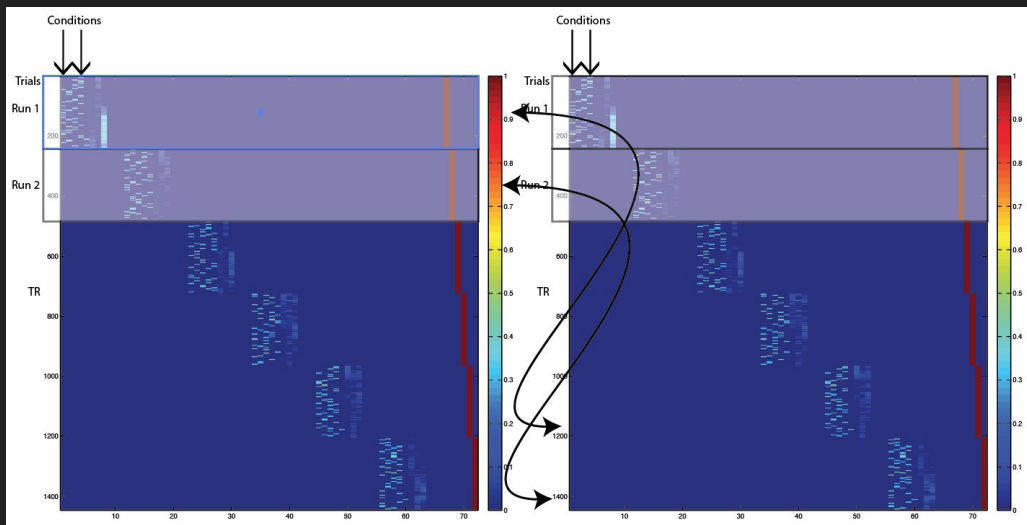
# Permute data in Pre and Post conditions



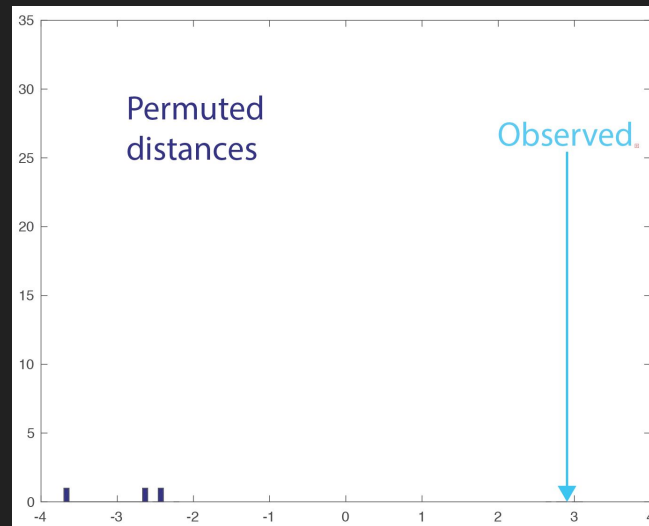
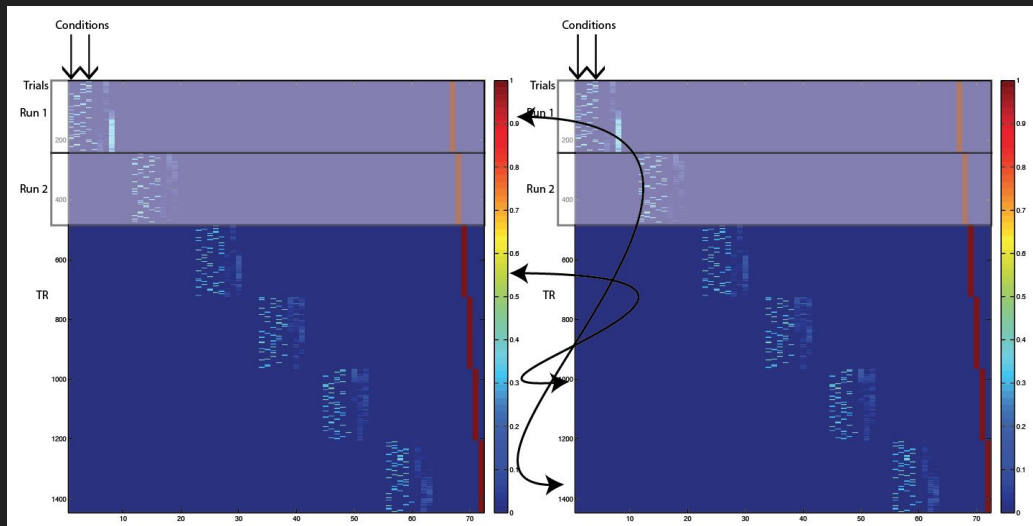
1. Recompute (now permuted) distance
2. Repeat for all possible permutations (12 choose 6)



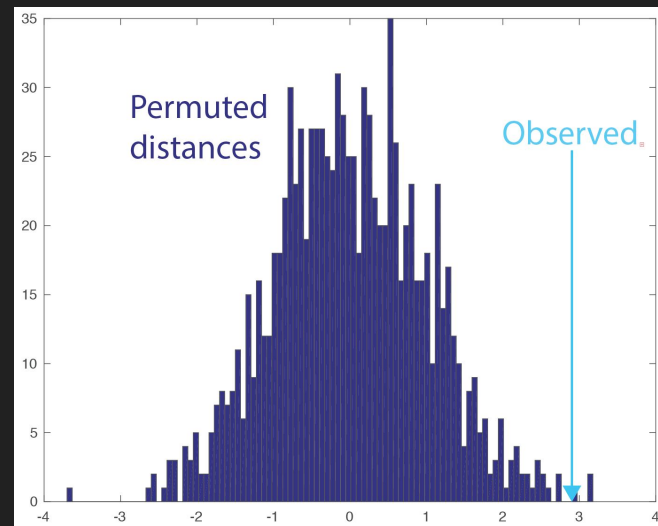
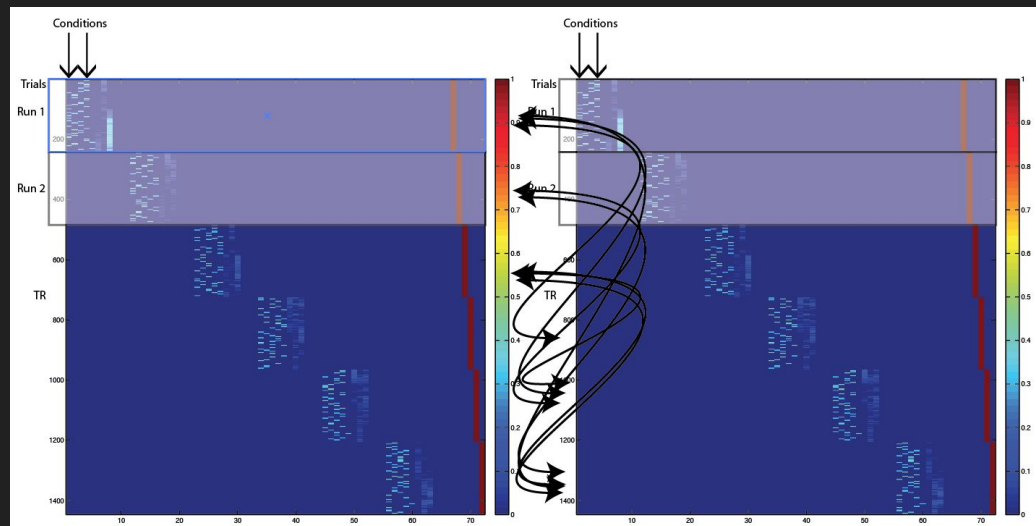
# Permute data, recompute H

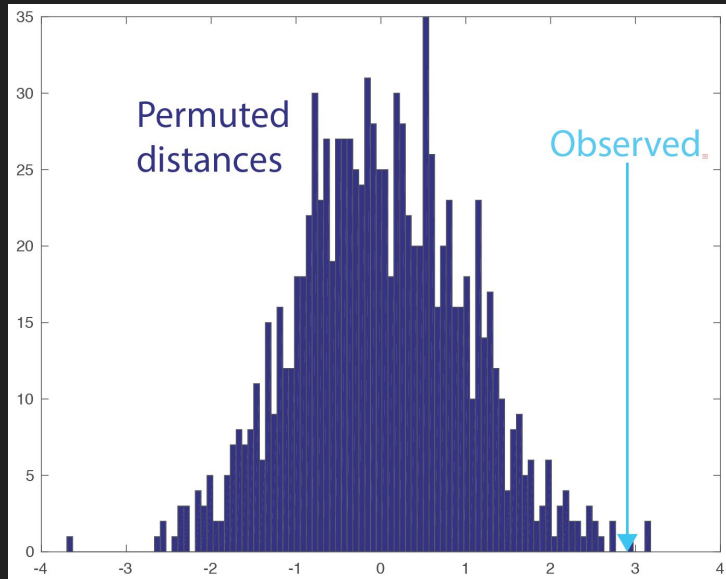


# Permute again, recompute H again



# Permute data again and again and again





Now we have a permutation distribution, and we can check if our observed statistic is significant with classical definition of p-value

$$\text{p-val} = \text{len}(\text{permuted distances} > \text{obs}) / n$$

where  $n = 12 \text{ choose } 6 = 924$

4 permuted distances are greater than observed

$$\text{p-val} = 4/924 = 0.004$$

Significant at 0.05 for this example.

# Limitations of Permutation Test

- Computationally expensive. (May not be possible to construct permutation distribution)
- e.g. Full permutation might entail 20000! Tests.
  - Not gonna happen
- Permutation test on searchlight might require cluster/AWS/GPU etc.

N.B. Always think carefully about what you are permuting

# Questions Regarding Permutation Tests?

# Topic 3: Pulling yourself up by the resampling of your data

Suppose you have computed a mean of your sample

Want to know about population, but all we have is a sample

You may not know if your population is Normal, or Bernoulli or you may not satisfy assumptions eg. CLT

Could repeat experiment 1000s of times, but not tractable, is there an alternative?

Let's take our collected data, and repeat the experiment as if the sample was the population.

# RSA and the Bootstrap

I have a sample size of 9 subjects, each with a distance change score.

Distance = [2.0424   2.1113   3.6934   1.8490  
0.8678   0.9481   3.2064   2.0855   1.4180]

Mean = 2.02. Let's construct the confidence interval.



# Bootstrap resamples.

First 5 samples:

2.2091	-2.3272	0.0459	0.9412	0.4037	0.9412	2.2091	1.1848	0.9412
0.4037	0.4037	1.0864	1.0864	1.4698	0.4037	1.0864	1.4698	1.0864
1.0864	0.4037	2.2091	0.4037	0.4037	1.0864	1.0864	1.4698	-2.3272
1.4698	1.1848	-2.3272	0.9412	1.1848	1.5377	0.4037	1.0864	1.4698
0.9412	2.2091	2.2091	0.0459	-2.3272	1.0864	0.0459	1.5377	0.9412

...

With means = [0.7277 0.9440 0.6469 0.7723 0.7433 ...]

Note samples are always of same length as original sample.

In other words, bootstrap samples are always WITH replacement

How many samples? Choice up to you, look for stability (min 1000)

# Construct CI from bootstrap

Compute standard deviation of means

Then, for 95% CI, find 2.5% and 97.5% percentiles in bootstrapped data.

Then observed mean  $\pm$  CI is your 95% bootstrapped confidence interval of your mean.

# Bootstrap Limitations

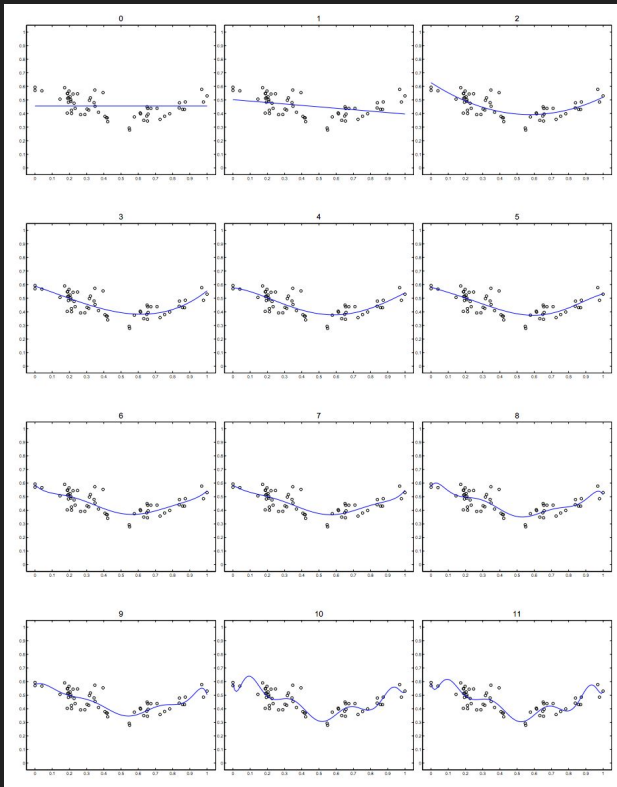
- Non-linearities can cause problems
- Boundaries can cause problems
- Like permutation test, can be computationally expensive depending on how many bootstrap resamples you require.

# Questions Regarding the Bootstrap

# Review

- 1) Cross-validated distances are unbiased
- 2) Permutation tests enable hypothesis testing regarding plasticity changes when null is unknown
- 3) Bootstrap enables constructing CIs of a single sample of data

# Problem 1



What's the best curve fit for this data?

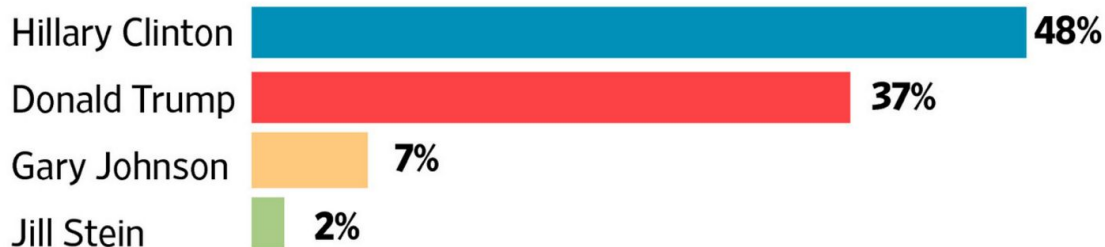
# Problem 2

## Clinton Extends Lead to 11 Points

### A Clinton Advantage

Hillary Clinton leads Donald Trump by 11 percentage points in the latest WSJ/NBC News poll.

#### Among likely voters:



# Problem 3

Suppose I wish to estimate the average IQ at CMU

I collect 10 data points and observe mean of 130.

I don't know the underlying distribution of IQ at CMU

What's my confidence in this estimate?



Questions about any topics  
covered today?