

# Reverse-engineering the brain

Charles Zheng

Stanford University

January 28, 2016

# HUMAN ENGINEERING

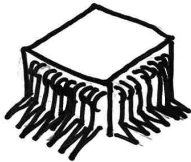
Coarse  
grained

$$\mathbb{E}[\cdot]$$

$$\text{Var}[\cdot]$$

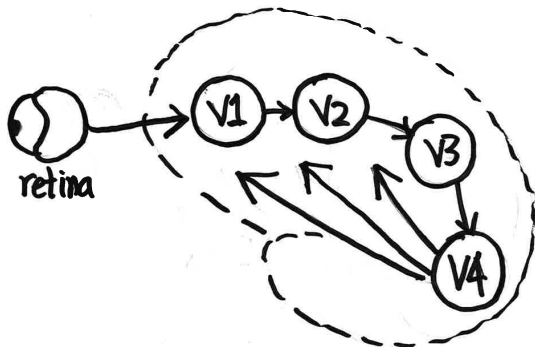
fine  
grained

$$\frac{dx}{dt} = f(x)$$

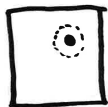


# NATURAL SYSTEM





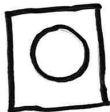
Retina



V1

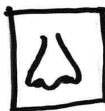


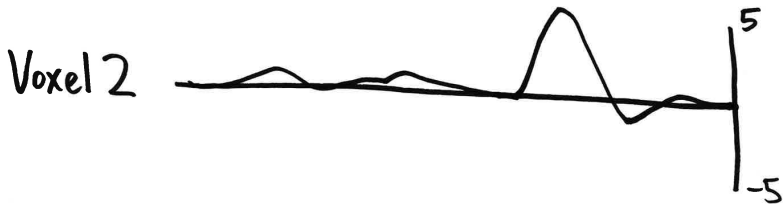
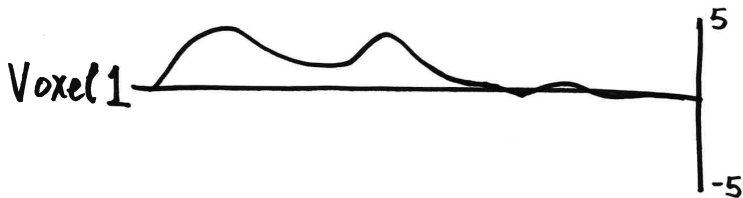
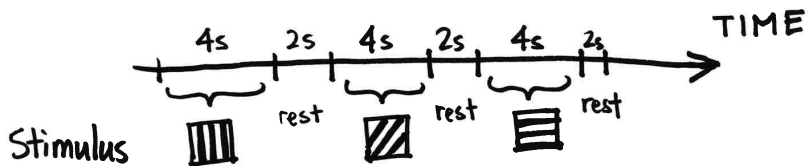
V2?

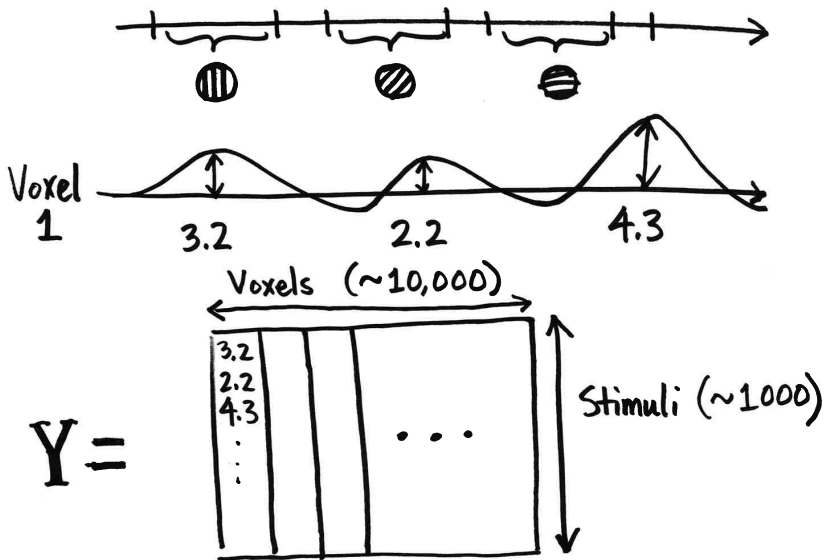


...

V4??

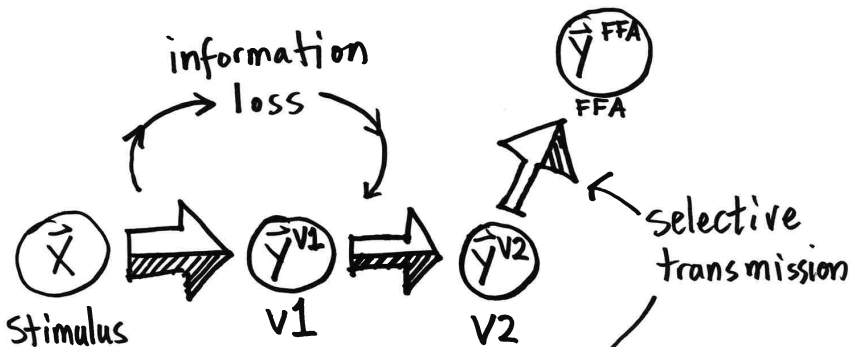






# Functional MRI

Stimuli $x$	Response $y$
$\begin{pmatrix} 1.0 \\ 0 \\ 3.0 \\ 0 \\ -1.2 \end{pmatrix}$	$\begin{pmatrix} 1.2 \\ 0 \\ -1.8 \\ -1.2 \end{pmatrix}$
$\begin{pmatrix} 0 \\ -2.2 \\ -3.1 \\ 4.5 \\ 0 \end{pmatrix}$	$\begin{pmatrix} -1.2 \\ -1.9 \\ 0.5 \\ 0.6 \end{pmatrix}$

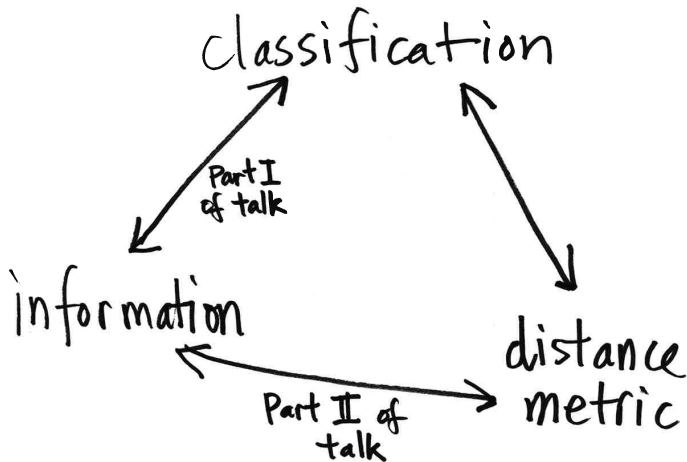


- explicit models

$$\vec{Y} = f(\vec{X}) + \vec{\epsilon}$$

- information

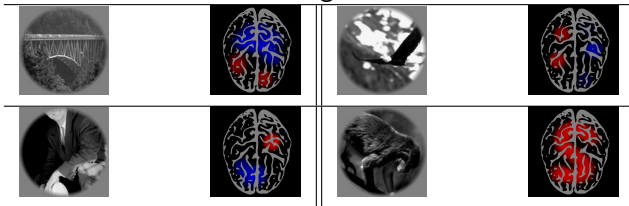
$$I(\vec{X}; \vec{Y})$$



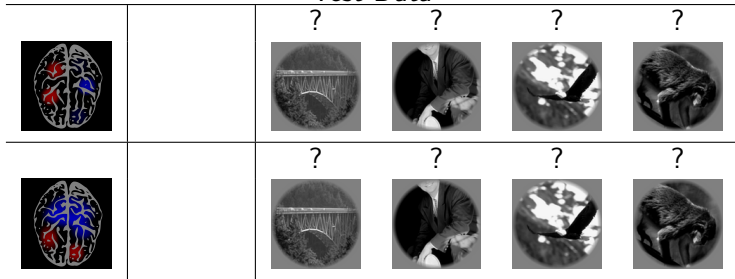


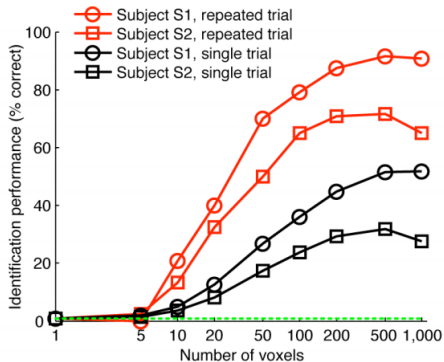
# A mind-reading game: Classification

Training Data

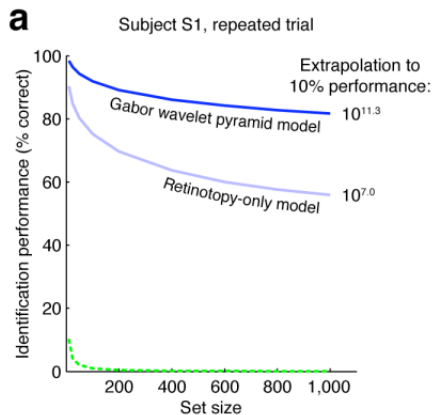


Test Data





Redundancy of neural coding



Can we extrapolate the classification curve?

# How many neurons does it take to classify a lightbulb?

Charles Zheng

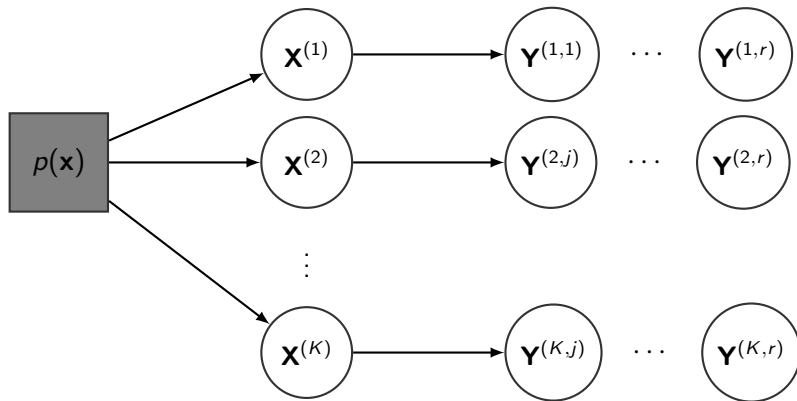
Stanford University

January 6, 2016

(Joint work with Yuval Benjamini.)

# Experimental design

Sample stimuli iid from  $p(\mathbf{x})$ . Repeated measures.



- Main question: How to estimate  $I(\mathbf{X}; \mathbf{Y})$  from such data?
- Side question (experimental design): How to choose  $K$ ?

# Entropy and mutual information

$X$  and  $Y$  have joint density  $p(x, y)$  with respect to  $\mu$ .

Quantity	Definition	Linear analogue
Entropy	$H(X) = - \int (\log p(x)) p(x) \mu_X(dx)$	$\text{Var}(X)$
Conditional entropy	$H(X Y) = \mathbf{E}[H(X Y)]$	$\mathbf{E}[\text{Var}(X Y)]$
Mutual information	$I(X; Y) = H(X) - H(X Y)$	$\text{Cor}^2(X, Y)$

The above definition includes both *differential* entropy and *discrete* entropy.  
Information theorists tend to use log base 2, we will use natural logs in this talk.

# Can we learn $I(\mathbf{X}; \mathbf{Y})$ from such data?

Answer: yes.

- We have  $I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X})$ .
- We can estimate  $H(\mathbf{Y})$  from the data
- We can estimate  $H(\mathbf{Y}|\mathbf{x}^{(i)})$  from the data, and define

$$\hat{H}(\mathbf{Y}|\mathbf{X}) = \frac{1}{K} \sum_{i=1}^K \hat{H}(\mathbf{Y}|\mathbf{x}^{(i)})$$

- As  $K$  and  $r$  both tend to infinity,

$$\hat{I}(\mathbf{X}; \mathbf{Y}) = \hat{H}(\mathbf{Y}) - \hat{H}(\mathbf{Y}|\mathbf{X})$$

is consistent for  $I(\mathbf{X}; \mathbf{Y})$ .

# Limitations with the 'naïve' approach

Naïve estimator:

$$\hat{l}(\mathbf{X}; \mathbf{Y}) = \hat{H}(\mathbf{Y}) - \frac{1}{K} \sum_{i=1}^K \hat{H}(\mathbf{Y} | \mathbf{x}^{(i)})$$

- If  $K$  is small, the naïve estimator may be quite biased, even for low-dimensional problems. Gastpar et al. (2010) introduced an *antropic correction* to deal with the small- $K$  bias.
- Difficult to estimate differential entropies  $H(\mathbf{Y})$ ,  $H(\mathbf{Y} | \mathbf{x}^{(i)})$  in high dimensions. Best rates are  $O(1/\sqrt{n})$  for  $d \leq 3$  dimensions. Convergence rates for  $d > 3$  unknown!



# Can we use machine learning to deal with dimensionality?

- Supervised learning becomes an extremely common approach for dealing with high-dimensional data, for numerous reasons!
- Perhaps we can use supervised learning to estimate  $I(\mathbf{X}; \mathbf{Y})$  as well.
- Existing approach uses confusion matrix (Treves et al, 1997)

# Why use supervised learning to estimate $I(\mathbf{X}; \mathbf{Y})$ ?

- Successful supervised learning exploits structure in the data, which *nonparametric methods ignore*.
- Using supervised learning to estimate mutual information can be viewed as *using prior information* to improve the estimate of  $I(\mathbf{X}; \mathbf{Y})$ .
- So while the general problem of information estimation is nearly impossible in high dimensions, the problem might become tractable if we can exploit known structure in the problem!

# Relationship between mutual information and classification

- Suppose  $X$  and  $Y$  are discrete random variables, and  $X$  is uniformly distributed over its support.
- Classify  $X$  given  $Y$ . The optimal rule is to guess

$$\hat{X} = \operatorname{argmax}_x p(Y|X = x).$$

- Bayes error:

$$p_e = \Pr[X \neq \hat{X}].$$

- Fano's inequality:

$$I(X; Y) \geq (1 - p_e) \ln K - \text{const.}$$

where  $K$  is the size of the support of  $X$ .

# Key Result

Define the  $K$ -class *average Bayes error* as

$$ABE_K = \mathbf{E}[\Pr[\hat{X}_{Bayes}(Y) \neq X]]$$

where  $\hat{X}_{Bayes}$  is the optimal classification rule, and the expectation averages over randomness in  $\mathbf{X}^{(1)}, \dots, \mathbf{X}^{(k)}$ .

Then, under a certain asymptotic regime, there is an *exact* relationship between  $ABE_K$  and  $I(\mathbf{X}; \mathbf{Y})$  (in contrast to the lower bound given by Fano's inequality or confusion matrix method.)

**Theorem.** Take a sequence of joint distributions  $(\mathbf{X}, \mathbf{Y})$  where the dimensionality of  $\mathbf{X}$  and  $\mathbf{Y}$  grow to infinity, but where

$$\lim I(\mathbf{X}; \mathbf{Y}) = \iota < \infty.$$

Under certain regularity conditions, the limiting  $K$ -class average Bayes error is given by

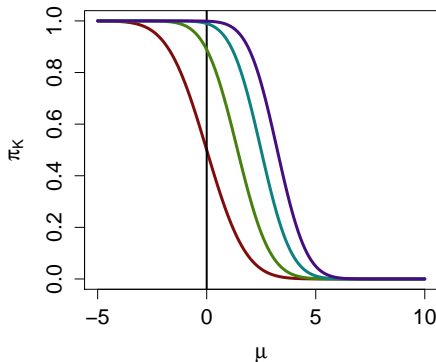
$$\lim_{d \rightarrow \infty} ABE_K = \pi_K(\sqrt{2\iota}).$$

where

$$\pi_K(\mu) = 1 - \int_{-\infty}^{\infty} \phi(z - \mu)(1 - \Phi(z))^{K-1} dz.$$

## Sidenote: interpretation of $\pi_K$

The function  $\pi_K(\mu)$  gives the probability that a  $N(\mu, 1)$  variable is smaller than the minimum of  $K - 1$  other  $N(0, 1)$  variables (all independent.)  
Hence  $\pi_K(0) = \frac{K-1}{K}$  due to symmetry. (This is also the misclassification rate from pure guessing.)



Legend:  $K = \{ \text{2, 9, 99, 999} \}$

# Regularity Conditions

- We require the conditional log-likelihoods

$$\log p(\mathbf{Y}^{(i)} | \mathbf{X}^{(j)})$$

to have a nondegenerate jointly multivariate normal limiting distribution (when scaled appropriately).

- Define

$$u(x, y) = \frac{p(x, y)}{p(x)p(y)} - 1.$$

Draw  $\mathbf{X} \sim p(x)$  and  $\mathbf{Y}^* \sim p(y)$  independently. We require  $U(\mathbf{X}, \mathbf{Y}^*)$  have a nondegenerate univariate normal limiting distribution (when scaled appropriately).

These conditions are satisfied quite generally, e.g. if  $(\mathbf{X}, \mathbf{Y})$  are multivariate normal and their marginal covariance matrix has a limiting spectrum.

# The low-SNR estimator of $I(\mathbf{X}; \mathbf{Y})$

Our proposed estimator for mutual information is

$$\hat{I}_{ls}(\mathbf{X}; \mathbf{Y}) = \frac{1}{2} \pi_K^{-1} (\widehat{\text{ABE}})^2$$

where  $\widehat{\text{ABE}}$  is the test error of the classifier. (The subscript  $ls$  stands for low-SNR.)

Note: there is still a lot of improvement for this estimator, mainly correcting for bias due to variability in  $\widehat{\text{ABE}}$ . This is the first version we tried.



## *Models.*

- Multiple-response logistic regression model

$$X \sim N(0, I_p)$$

$$Y \in \{0, 1\}^q$$

$$Y_i | X = x \sim \text{Bernoulli}(x^T B_i)$$

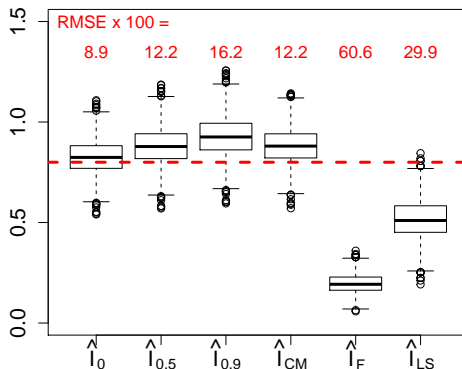
where  $B$  is a  $p \times q$  matrix.

## *Methods.*

- Nonparametric:  $\hat{l}_0$  naive estimator,  $\hat{l}_\alpha$  anthropic correction.
- ML-based:  $\hat{l}_{CM}$  confusion matrix,  $\hat{l}_F$  Fano,  $\hat{l}_{LS}$  low-SNR method.

# Fig 1. Low-dimensional results ( $q = 3$ )

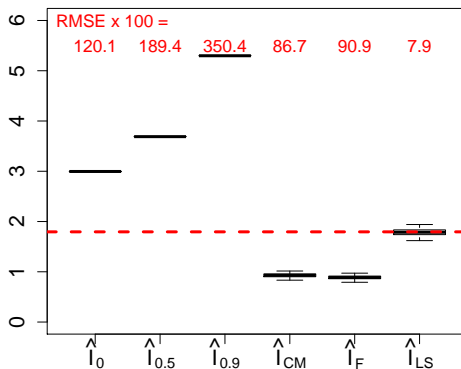
Sampling distribution of  $\hat{I}$  for  $\{p = 3, B = \frac{4}{\sqrt{3}}I_3, K = 20, r = 40\}$ .  
True parameter  $I(X; Y) = 0.800$  (dotted line.)



Naïve estimator performs best!  $\hat{I}_{LS}$  not effective.

## Fig 2. High-dimensional results ( $q = 50$ )

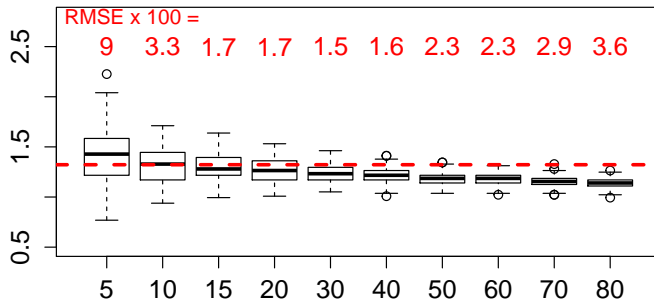
Sampling distribution of  $\hat{I}$  for  $\{p = 50, B = \frac{4}{\sqrt{50}} I_{50}, K = 20, r = 8000\}$ .  
True parameter  $I(X; Y) = 1.794$  (dashed line.)



Non-parametric methods extremely biased.

## Fig 5. Dependence on $K$ given fixed $N$ ( $q = 10$ )

Sampling distribution of  $\hat{I}_{LS}$  for  $\{p = 10, B = \frac{4}{\sqrt{10}}I_{10}, N = 80000\}$ ,  
and  $K = \{5, 10, 15, 20, \dots, 80\}$ ,  $r = N/k$ .  
True parameter  $I(X; Y) = 1.322$  (dashed line.)



Decreasing variance as  $K$  increases. Bias at large and small  $K$ .

# Conclusions

- We derive a relationship between average Bayes error (ABE) and mutual information (MI), motivating a novel estimator  $\hat{I}_{LS}$ .
- Theory based on high dimensional, low SNR limit, where

$$\text{ABE} \leftrightarrow \text{MI}.$$

- In ideal settings for supervised learning, ABE can be estimated effectively and  $\hat{I}_{LS}$  can recover MI at much lower sample sizes than nonparametric methods.
- In simulations,  $\hat{I}_{LS}$  works better than Fano's inequality or the confusion matrix approach.

# The geometry of human perception: RSA and multivariate models

Charles Zheng

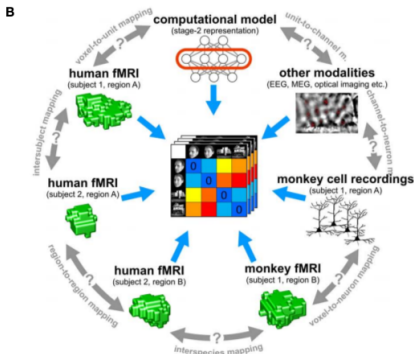
Stanford University

November 16, 2015

(Joint work with Yuval Benjamini and Oluwasanmi Koyejo)

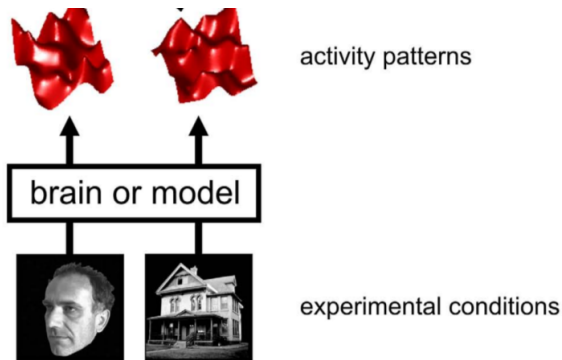
# Representation similarity analysis (RSA)

- Framework for studying how mental objects are represented in the brain, via brain activity (measured by fMRI, EEG) or behavior.
- Compare different brain regions or imaging modalities within a single subject, or compare multiple subjects.



# A typical RSA experiment

An experiment which demonstrates which regions of the brain differentiate between faces and objects.

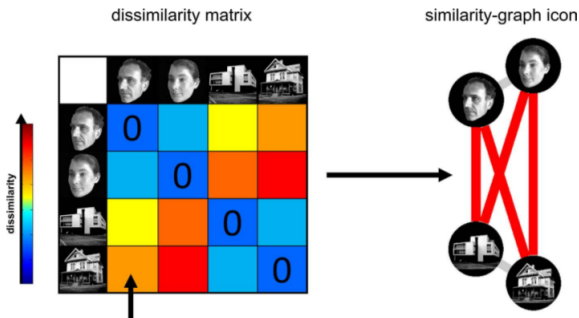


Step 1: Present the subject with visual stimuli, pictures of faces and houses. Record the subject's brain activity in the fMRI scanner.



# A typical RSA experiment

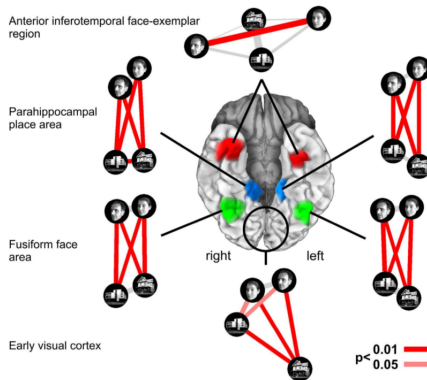
Step 2a: Process the data, and represent the brain activity of the subject for the  $i$ th stimulus as a real vector  $y_i$ . Form matrix of distances between  $y_i$  and  $y_j$ , the *representation distance matrix* (RDM).



Step 2b: Assess statistical significance of distances to form similarity graph.

# A typical RSA experiment

Step 3: Compare similarity graphs between different brain regions.



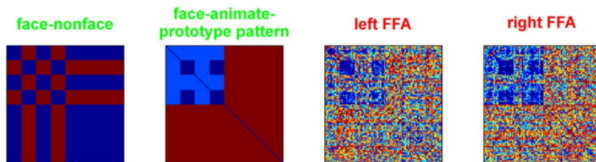
Step 4: Draw scientific conclusions. (Step 5: Profit!!...?)

- Core methodology presented by Kriegeskort et al (2008) and extended by others.
- Suppose each of the stimuli have  $r$  repeats, the responses for stimulus  $i$  are  $y_i^1, \dots, y_i^r$ , and the average over the repeats as  $\bar{y}_i$ .
- The representation distance matrix is computed as

$$D_{ij} = d(\bar{y}_i, \bar{y}_k)$$

where  $d$  may be Euclidean distance or correlation distance.

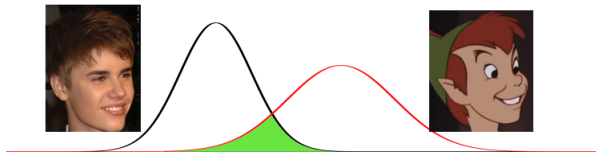
# Scientific interpretation of RSA results



A rejection of the independence null between  $D^A$  and  $D^B$  is taken to mean that outputs  $A$  and  $B$  are 'related.' For instance,  $D^A$  might be the response from a subject's brain region, and  $D^B$  is a 0-1 matrix reflecting *a priori* class membership. Rejection is taken to mean that the region  $A$  have differential activation depending on the classes represented  $D^B$ .

# Distribution-induced distance: motivation

Consider two stimuli  $x_1$  and  $x_2$  to be *distant* if the *response distributions* are statistically distant, or *close* if their response distributions overlap.



Note that the definition not only depends on the difference in *means* but also depends on the *noise distribution*.

# Distribution-induced distance: definition

- Let  $\mathcal{F}_x$  denote the distribution of the response  $y$  conditional on the stimulus  $x$ .
- Define the dissimilarity matrix

$$D_{ij} = \mathbb{D}(\mathcal{F}_{x_i}; \mathcal{F}_{x_j})$$

where  $\mathbb{D}$  is a measure of distance or divergence between probability measures.

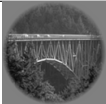



- Example: if  $y$  is conditionally multivariate normal, with covariance  $\Sigma$  not depending on  $x$ , then

$$D_{ij} = \frac{1}{2}(\mathbf{E}[y|x_i] - \mathbf{E}[y|x_j])^T \Sigma^{-1}(\mathbf{E}[y|x_i] - \mathbf{E}[y|x_j])$$

for either  $\mathbb{D} = \text{KL divergence}$  or  $\text{Hellinger distance}$ .

# Functional MRI allows us to infer the metric

- In fMRI, one records the subject's response  $y_i$  to a stimulus  $x_i$ , for  $i = 1, \dots, n$  (e.g.  $n = 3000$ )
- The recorded  $y_i$  is actually a 'filtered' version of the brain activity (some information is lost)
- By fitting a model to the data, one can estimate the conditional distribution of  $Y|x$  and hence estimate the perception-induced metric  $d_t$ .

Stimuli	Response
	
	

- Combine the parametric approach of multivariate regression with RSA.
- Model:

$$y \sim N(B^T x, \Omega^{-1}).$$

- The distribution-induced metric is therefore

$$D(x_i, x_j) = (x_i - x_j)^T B \Omega B^T (x_i - x_j).$$



$$y \sim N(B^T x, \Omega^{-1})$$

$$D(x_i, x_j) = (x_i - x_j)^T B \Omega B^T (x_i - x_j)$$

- Since all information about the distance is captured by the matrix  $\Sigma = B \Omega B^T$ , instead of testing

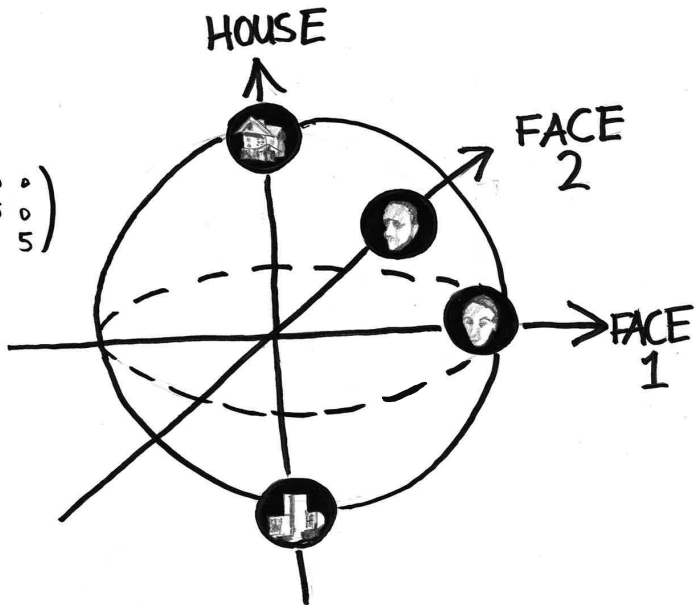
$$D^A = D^B$$

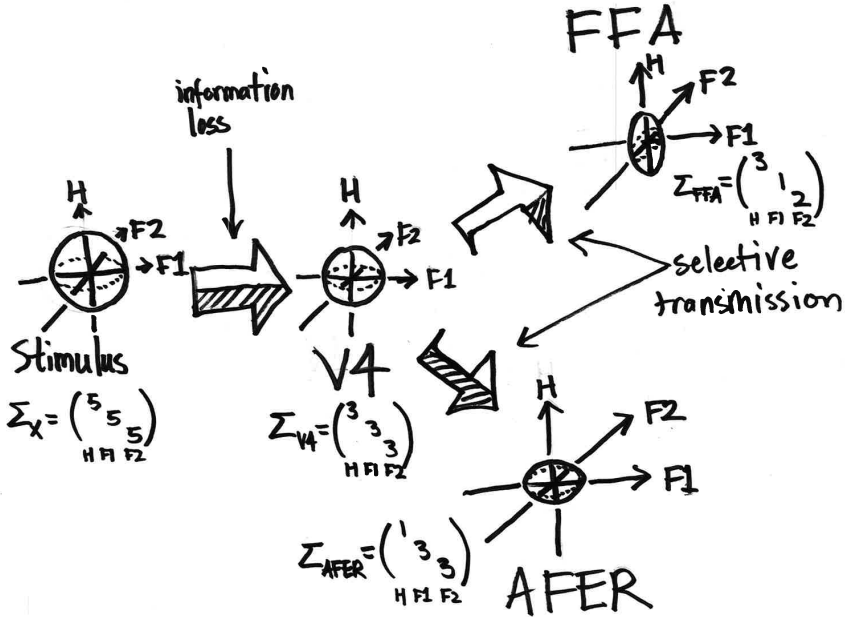
we can test

$$\Sigma^A = \Sigma^B.$$

- We can compare two datasets with non-overlapping stimuli
- The approach is scalable in the number of distinct stimuli, since the size of  $\Sigma^A$ ,  $\Sigma^B$  only depend on the number of *features* rather than the number of stimuli

$$\Sigma = \begin{matrix} F1 \\ F2 \\ H \end{matrix} \begin{pmatrix} 5 & 0 & 0 \\ 5 & 0 & 5 \end{pmatrix}$$





# Comparison with information approach

- Parametric RSA makes strong assumptions on the model (e.g. linearity.)
- We can track information loss:  $I(X; Y^{V^1}) > I(X; Y^{V^2})$
- But in cases of *selective transmission*, we can infer *which dimensions of information* are being lost (face features, house features, etc.)
- We get a finer-grained picture of information flow than the Shannon information approach.

## Section 2

To do

- Apply low-SNR information estimator and parametric RSA to real data.
- Develop bootstrap-based hypothesis tests for parametric RSA: e.g.  $H_0 : \Sigma^A = \Sigma^B$ .

# Possible extensions of information theory work

- Our proposed estimator performs well in small sample sizes, yet is *not consistent*... can it be fixed?
- Develop complimentary theory for estimating Bayes error. (This would allow us to state risk bounds for estimating  $I(X; Y)$ .)
- Even if you don't care about Shannon information, the theory allows you to potentially extrapolate classification curves:

$$ABE_{10} \rightarrow I(X; Y) \rightarrow ABE_{10000}$$

(But more work is needed.)

- Use our theory to address the question of optimal experimental design.

- Kay, KN., Naselaris, T., Prenger, R. J., and Gallant, J. L. “Identifying natural images from human brain activity”. *Nature* (2008)
- Naselaris, et al. “Bayesian reconstruction of natural images from human brain activity”. *Neuron* (2009)
- Vu, V. Q., Ravikumar, P., Naselaris, T., Kay, K. N., and Yu, B. “Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models”, *The Annals of Applied Statistics*. (2011)
- Chen, M., Han, J., Hu, X., Jiang, Xi., Guo, L. and Liu, T. “Survey of encoding and decoding of visual stimulus via fMRI: an image analysis perspective.” *Brain Imaging and Behavior*. (2014)