

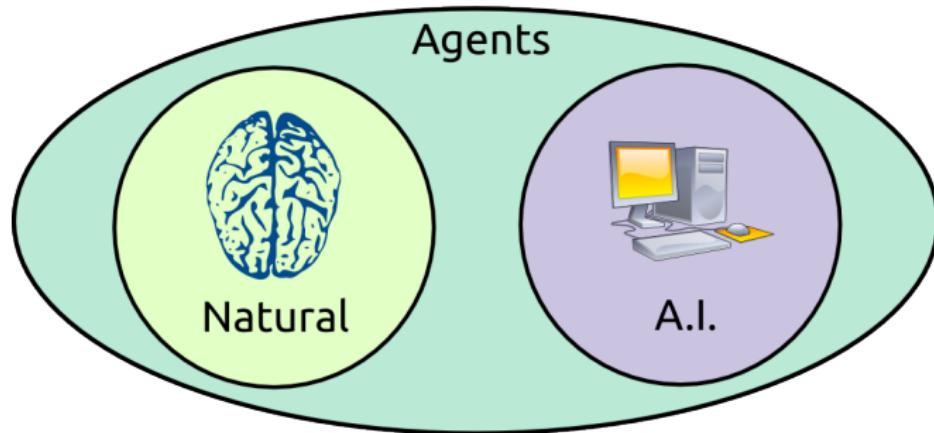
# Supervised Evaluation of Representations

Charles Zheng

Stanford University

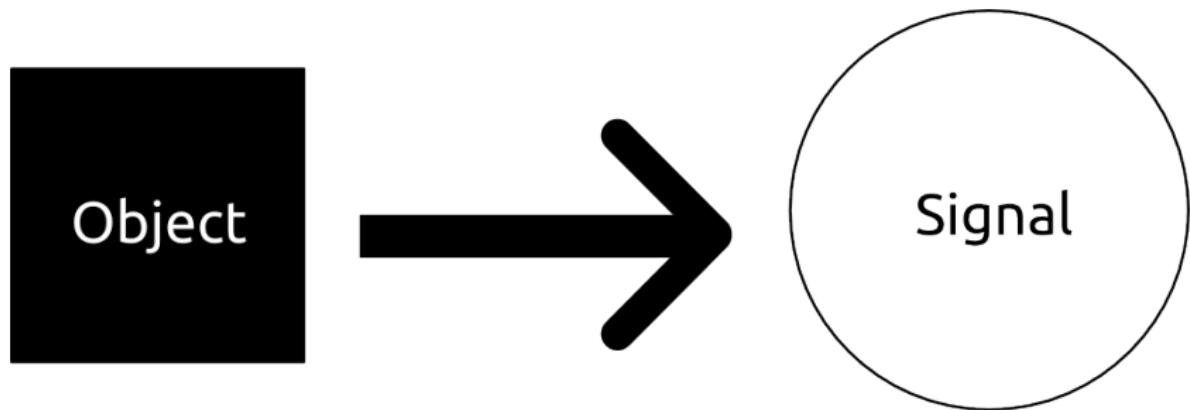
April 26, 2017

# Overview



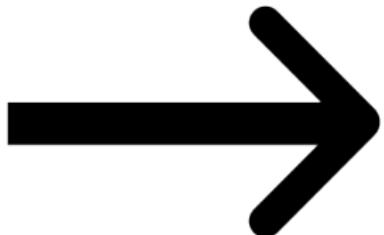
Human brains and machine learning algorithms tackle similar types of problems.

# Perception

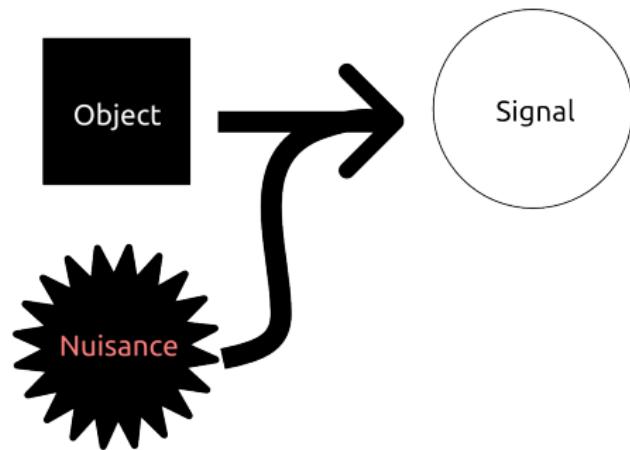


Perception: the problem of inferring *objects* in the environment given observed *signals*.

# Example: face recognition

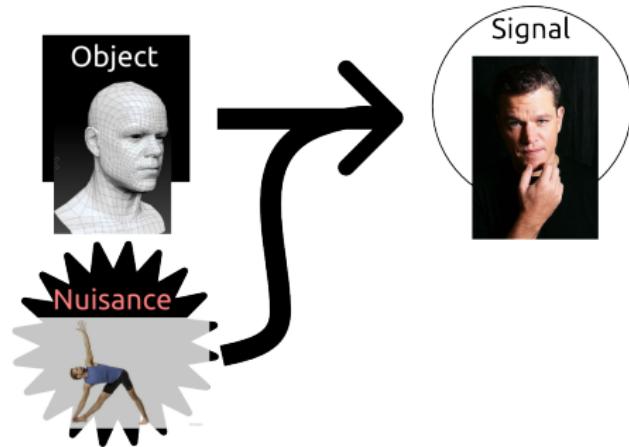


# Perception



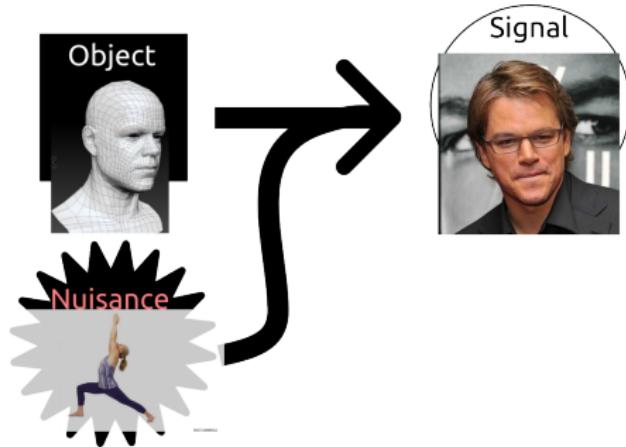
The problem is complicated because there exist some *nuisance parameters*, so the mapping from object to signal is not one-to-one.

# Example: face recognition



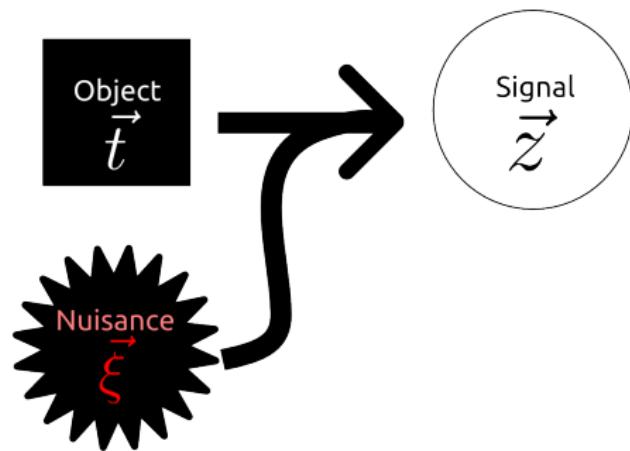
In face recognition, the *pose* (including hairstyle) and *lighting* are nuisance parameters.

# Example: face recognition



The same object can map to multiple signals.

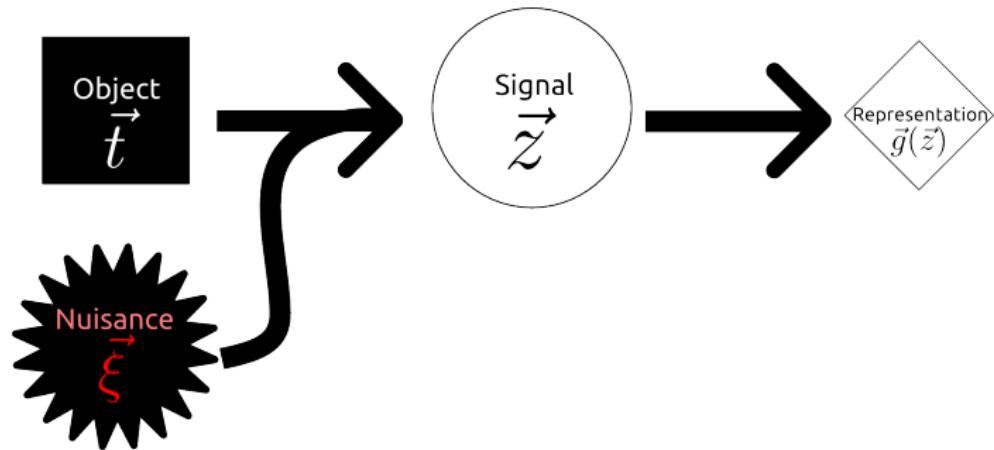
# Perception



Assume there exists a function  $\psi$  that maps objects and nuisance parameters to signals:

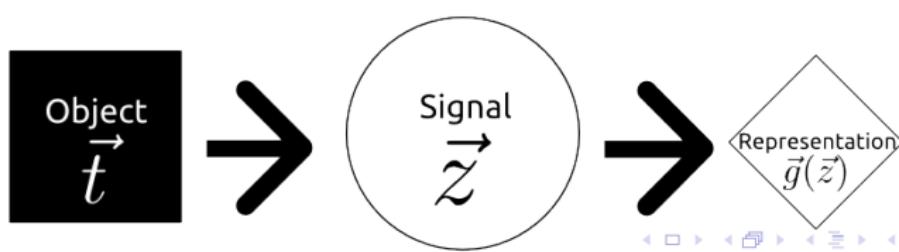
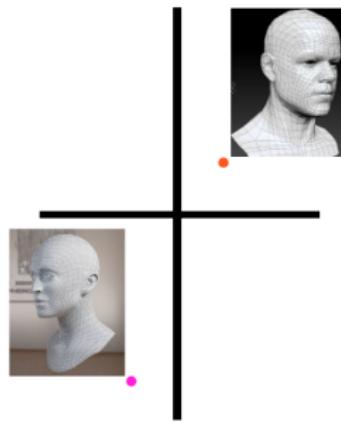
$$\vec{z} = \psi(\vec{t}, \vec{\xi}).$$

# What is a representation?

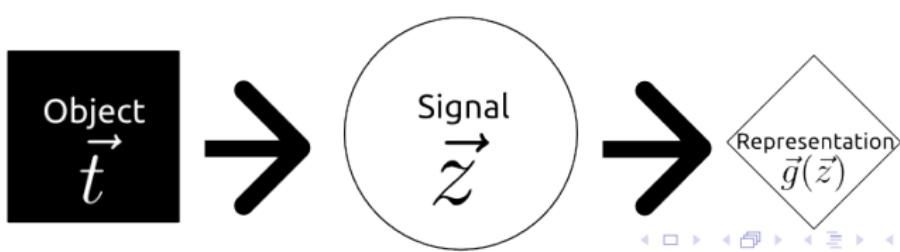
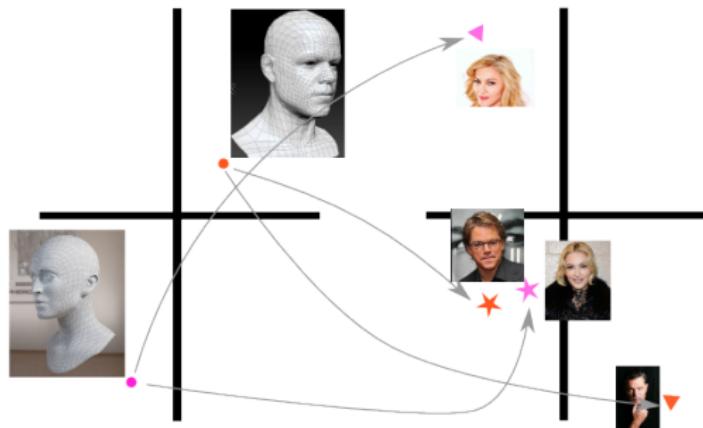


A dimensionality-reducing mapping  $\vec{g}$  of the signal.

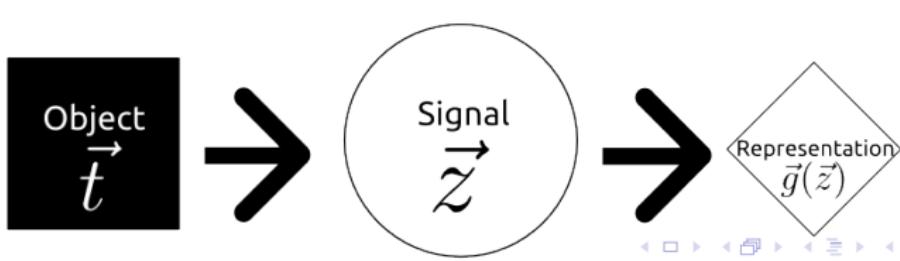
# A good representation...



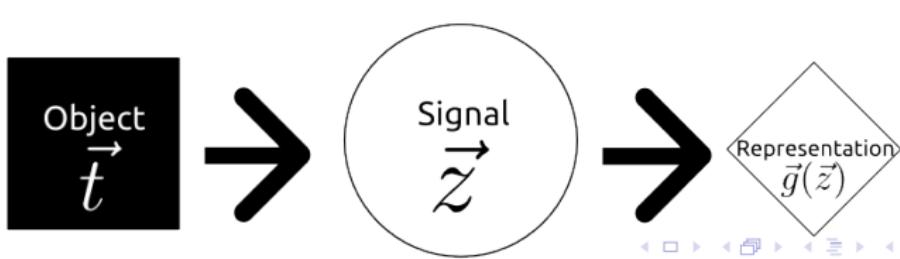
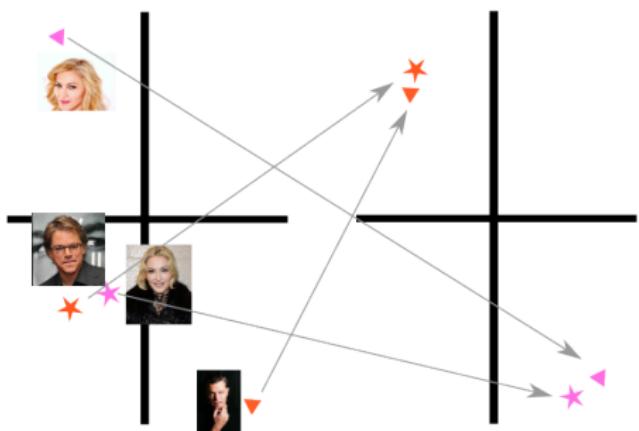
# A good representation...



# A good representation...



...captures the object space geometry



# Why do we care?

- 1) Neuroscience. The brain is hypothesized to use representations for cognitive purposes

# Why do we care?

- 1) Neuroscience. The brain is hypothesized to use representations for cognitive purposes
- 2) Machine learning. Representations turn out to be useful for many Machine Learning tasks!

# How can we tell if a representation is good?

- Method 1: *Ground truth*. If we happen to know the object parameters  $\vec{t}$  (e.g. we simulated the data).

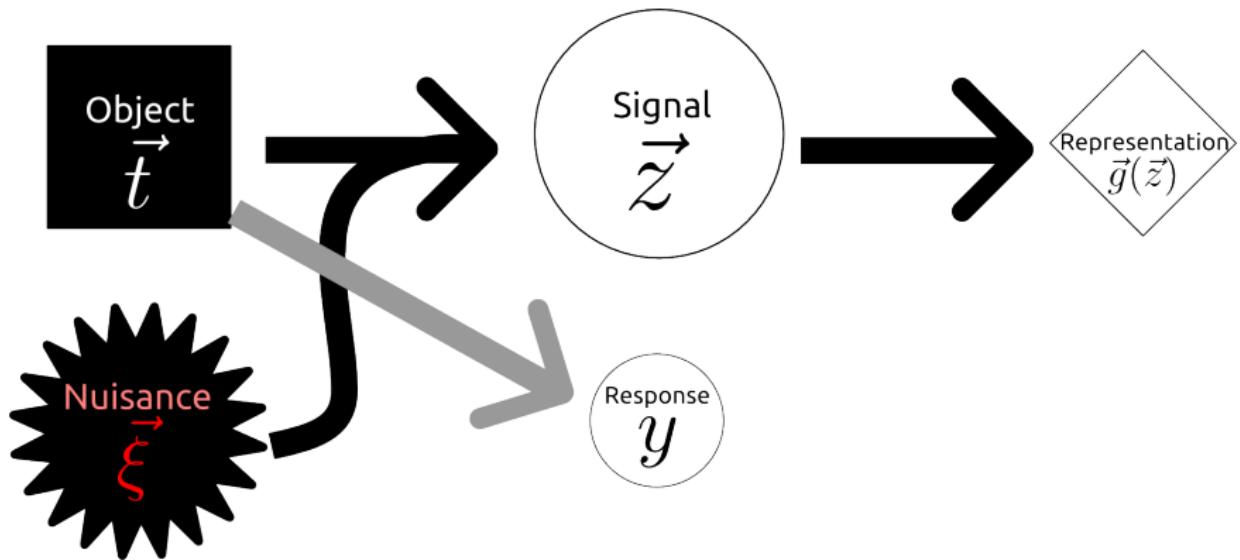
# How can we tell if a representation is good?

- Method 1: *Ground truth*. If we happen to know the object parameters  $\vec{t}$  (e.g. we simulated the data).
- Method 2: *End result*. By the performance of the representation on a machine learning task.

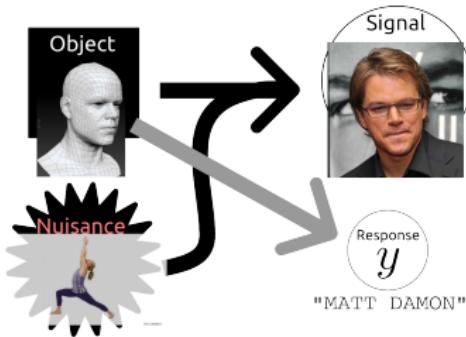
# How can we tell if a representation is good?

- Method 1: *Ground truth*. If we happen to know the object parameters  $\vec{t}$  (e.g. we simulated the data).
- Method 2: *End result*. By the performance of the representation on a machine learning task.
- Method 3: *Supervised*. If we have a *response variable*  $Y$  which can be used to infer distances in  $\vec{t}$ .

# Supervised evaluation of representations

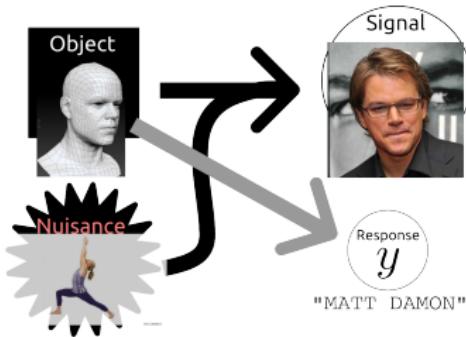


# Example: face recognition



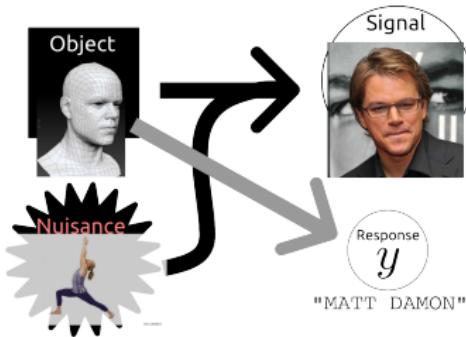
- The ID of the individual is an appropriate *response* variable...

# Example: face recognition



- The ID of the individual is an appropriate *response* variable...
- ...because two photos labeled with the same ID must belong to the same object  $\vec{t}$

## Example: face recognition



- The ID of the individual is an appropriate *response* variable...
- ...because two photos labeled with the same ID must belong to the same object  $\vec{t}$
- That is, for  $d(y, y')$  being the zero-one distance,

$$d(y, y') = 0 \Leftrightarrow d(\vec{t}, \vec{t}') = 0.$$

# Outline

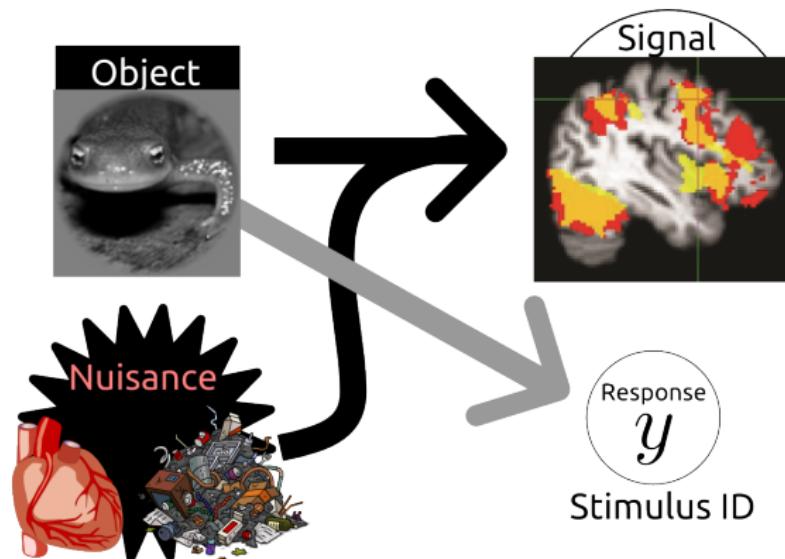
- *Previous work.* Identification accuracy, a method for evaluating representations
- *Contribution 1.* Extrapolation of identification accuracy.
- *Contribution 2.* Link between identification accuracy and mutual information.

## Section 2

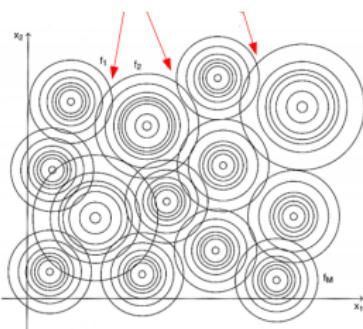
### Identification

# Identifying natural images from fMRI data

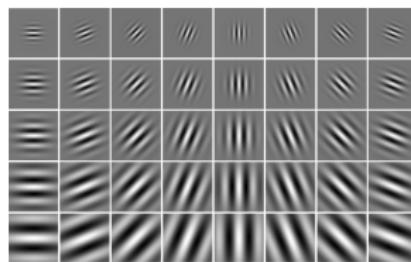
Kay, Naselaris, Prenger and Gallant (2008), *Nature*.



# Comparing two different natural image bases



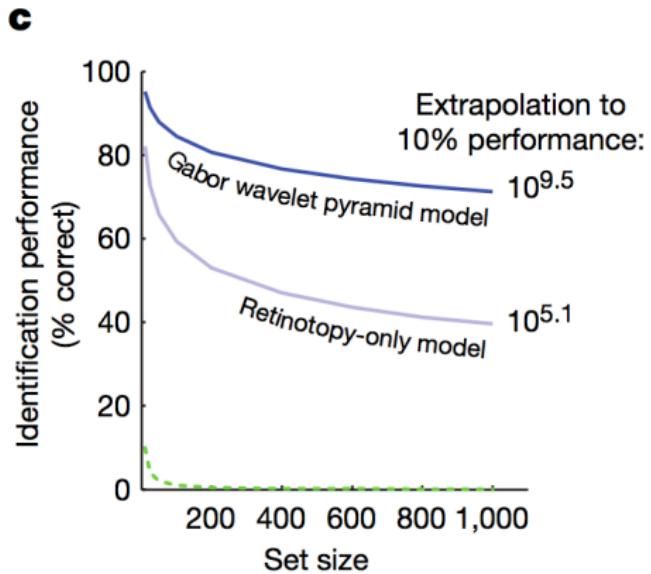
Retinotopic (Gaussian) basis



Gabor filter basis

# Identifying natural images from fMRI data

Kay et al. used *identification accuracy* as a metric for the quality of the representation



# Multiple-response regression

- Pairs  $(x_i, y_i)_{i=1}^n$ , where  $X$  is  $p$ -dimensional and  $Y$  is  $q$ -dimensional.
- Data matrices  $\mathbf{X}_{n \times p}$ ,  $\mathbf{Y}_{n \times q}$ .
- For each column of  $Y$ , fit sparse model  $Y^{(i)} \approx X^T \beta^{(i)} + \epsilon$ , e.g. by using elastic net (Zou 2008),

$$\hat{\beta}^{(i)} = \operatorname{argmin}_{\beta} \|\mathbf{X}^T \beta^{(i)} - Y^{(i)}\|^2 + \lambda_2 \|\beta^{(i)}\|_2^2 + \lambda_1 \|\beta^{(i)}\|_1$$

# Regression vs Identification accuracy

- Independent test set  $(x_i^*, y_i^*)_{i=1}^k$ .
- Use model to predict  $\hat{y}_i^* = (x_i^*)^T \hat{B}$  for  $i = 1, \dots, k$ .

Two ways to evaluate the predictive accuracy of the regression model:

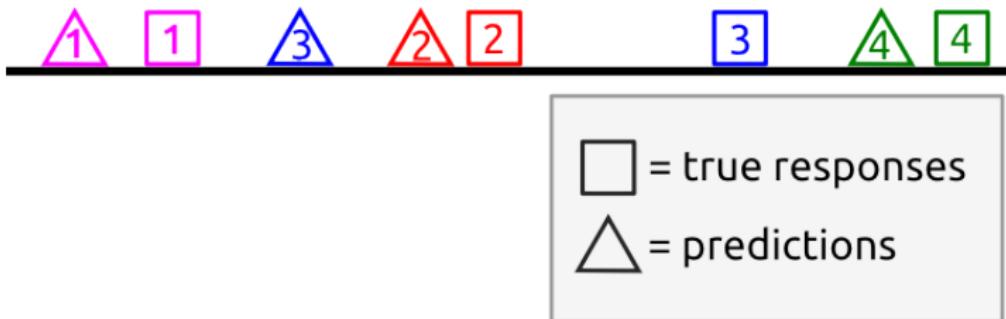
- Regression (mean squared-error) loss:

$$\text{MSE} = \frac{1}{k} \sum_{i=1}^k \|y_i^* - \hat{y}_i^*\|^2.$$

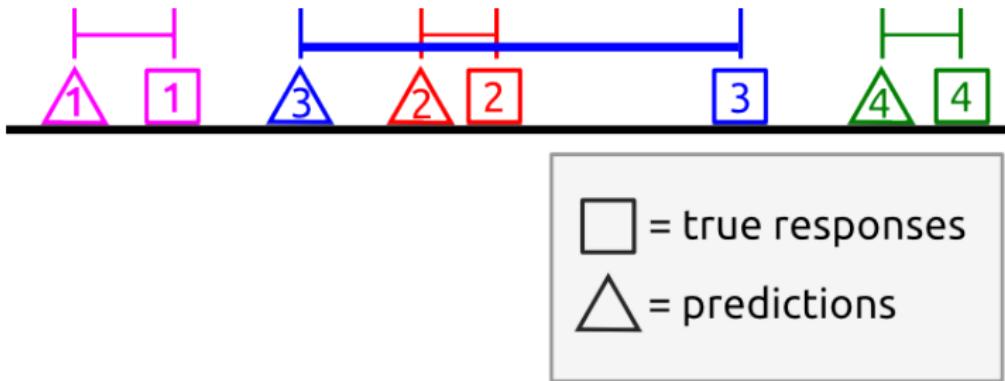
- Identification accuracy (Kay 2008):

$$\text{IdAcc}_k = \frac{1}{k} \sum_{i=1}^k I\{\hat{y}_i^* \text{ is nearest neighbor of } y_i^*\}.$$

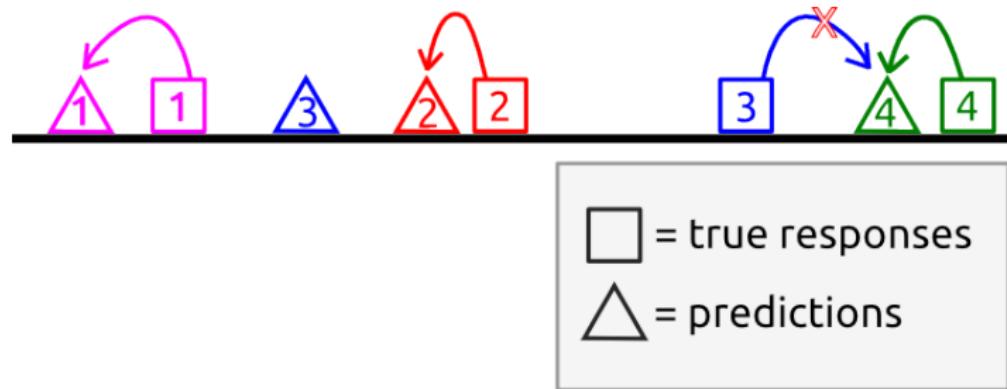
# Regression vs Identification accuracy



# Mean-squared error



# Identification accuracy



# Identification accuracy for comparing representations

- Suppose you have two different representation models:
  - $\vec{g}_1$ , the retinotopic model
  - $\vec{g}_2$ , Gabor filters

# Identification accuracy for comparing representations

- Suppose you have two different representation models:
  - $\vec{g}_1$ , the retinotopic model
  - $\vec{g}_2$ , Gabor filters
- For each model, train a linear model

$$\vec{Y} = B^T \vec{g}(\vec{Z}) + \epsilon$$

# Identification accuracy for comparing representations

- Suppose you have two different representation models:
  - $\vec{g}_1$ , the retinotopic model
  - $\vec{g}_2$ , Gabor filters
- For each model, train a linear model

$$\underbrace{\vec{Y}}_{\text{brain response}} = \underbrace{B^T}_{\text{coefficient matrix}} \underbrace{\vec{g}}_{\text{representation}} (\underbrace{\vec{Z}}_{\text{pixels}}) + \epsilon$$

# Identification accuracy for comparing representations

- Suppose you have two different representation models:
  - $\vec{g}_1$ , the retinotopic model
  - $\vec{g}_2$ , Gabor filters
- For each model, train a linear model

$$\underbrace{\vec{Y}}_{\text{brain response}} = \underbrace{B^T}_{\text{coefficient matrix}} \underbrace{\vec{g}}_{\text{representation}} (\underbrace{\vec{Z}}_{\text{pixels}}) + \epsilon$$

- Evaluate the identification accuracy on  $k$  test images.

# Identification accuracy for comparing representations

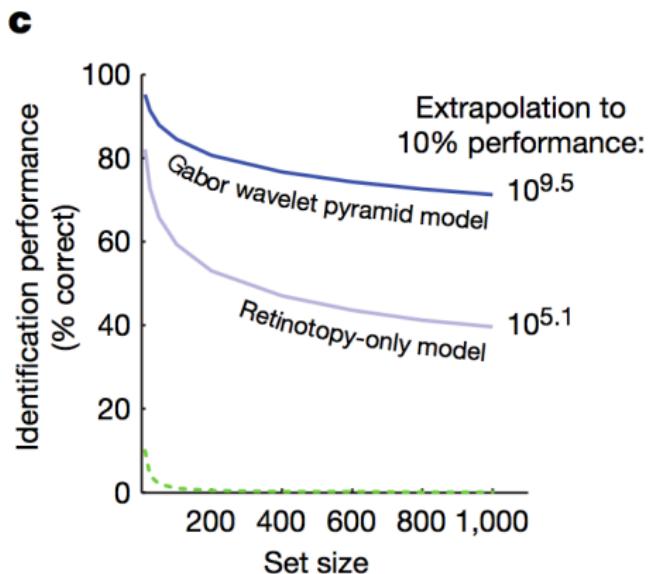
- Suppose you have two different representation models:
  - $\vec{g}_1$ , the retinotopic model
  - $\vec{g}_2$ , Gabor filters
- For each model, train a linear model

$$\underbrace{\vec{Y}}_{\text{brain response}} = \underbrace{B^T}_{\text{coefficient matrix}} \underbrace{\vec{g}}_{\text{representation}} (\underbrace{\vec{Z}}_{\text{pixels}}) + \epsilon$$

- Evaluate the identification accuracy on  $k$  test images.
- You can use a test set which is larger than  $k$ , and average the identification accuracy over  $k$ -sized subsamples

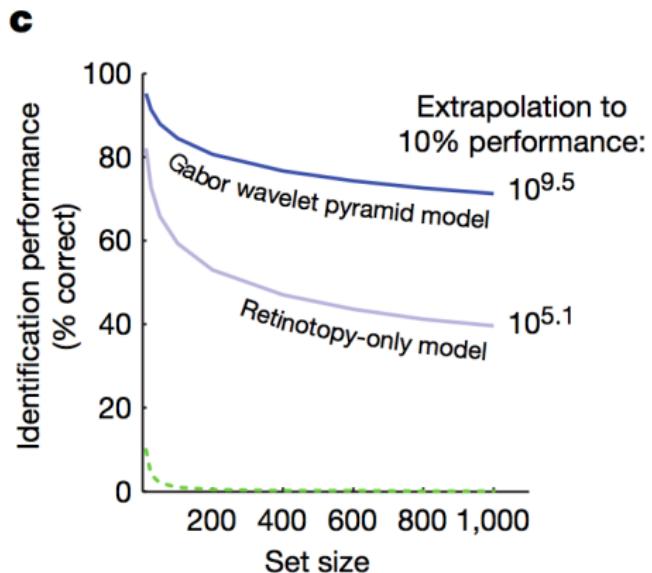
# Identifying natural images from fMRI data

Gabor filters yields consistently higher accuracy than retinotopic model



# Identifying natural images from fMRI data

Gabor filters yields consistently higher accuracy than retinotopic model



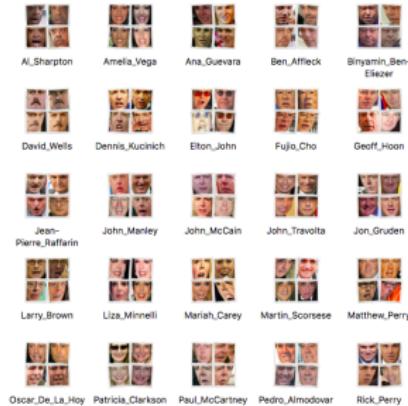
Q: Is this always the case? or could you also have intersecting curves?

## Section 3

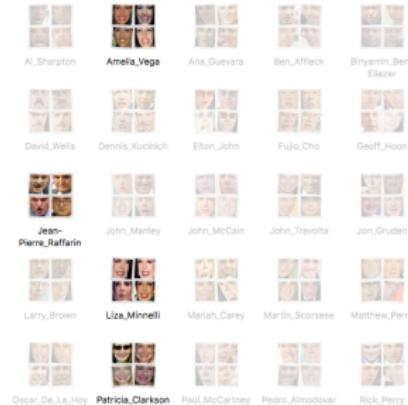
### Extrapolation

# Randomized multi-class classification

1. Population of categories  $\pi(y)$

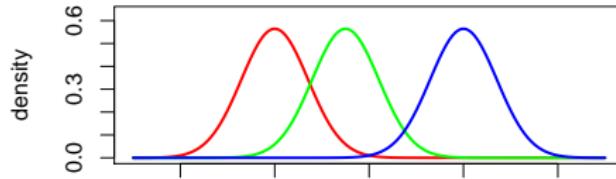


2. Subsample  $k$  labels,  $y_1, \dots, y_k$



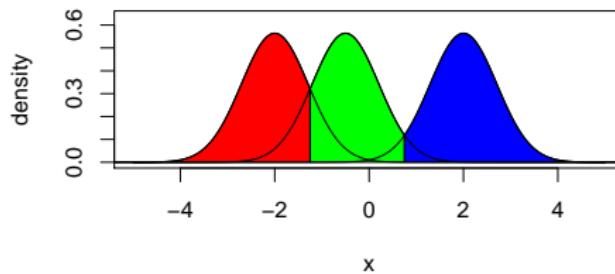
*Identification* is a special case of a *randomized classification* task.

# Toy example



- Suppose  $k = 3$ , and we draw  $Y_1, Y_2, Y_3$ .
- The *Bayes rule* is the optimal classifier and depends on knowing the true densities:
$$\hat{y}(x) = \operatorname{argmax}_{y_i} p(x|y_i)$$
- The *Bayes Risk*, which is the misclassification rate of the optimal classifier.

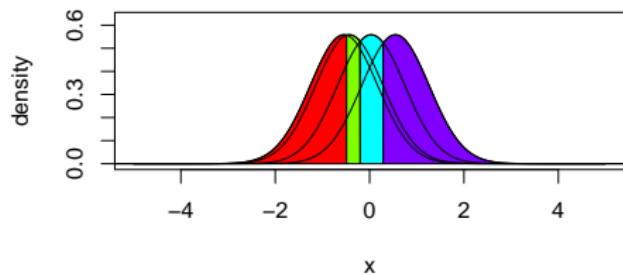
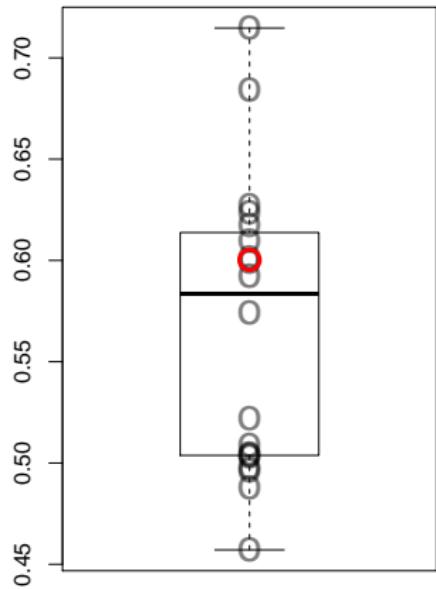
# Toy example



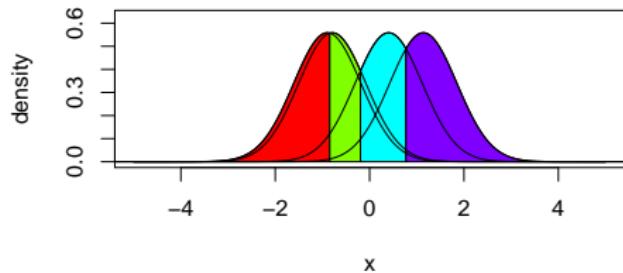
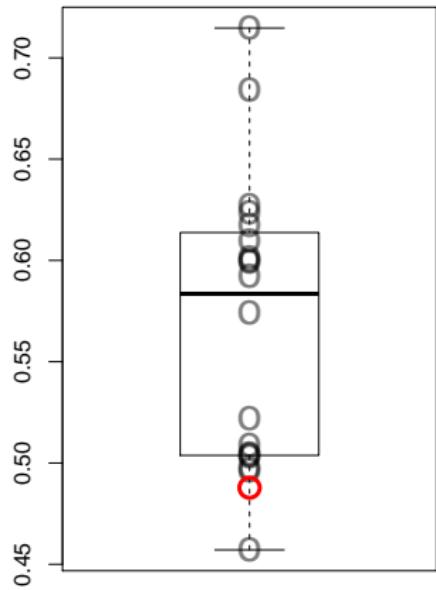
- The *Bayes Risk* is the expected test error of the Bayes rule,

$$\frac{1}{k} \sum_{i=1}^k \Pr[\hat{y}(x) \neq Y | Y = y_i]$$

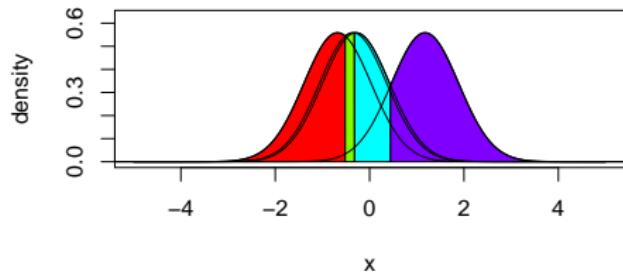
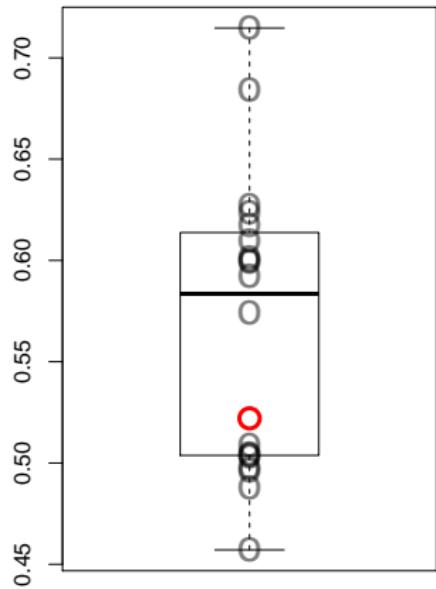
# Toy example



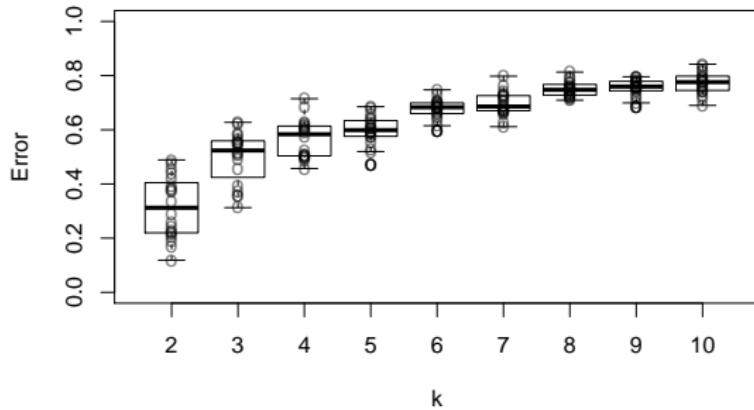
# Toy example



# Toy example

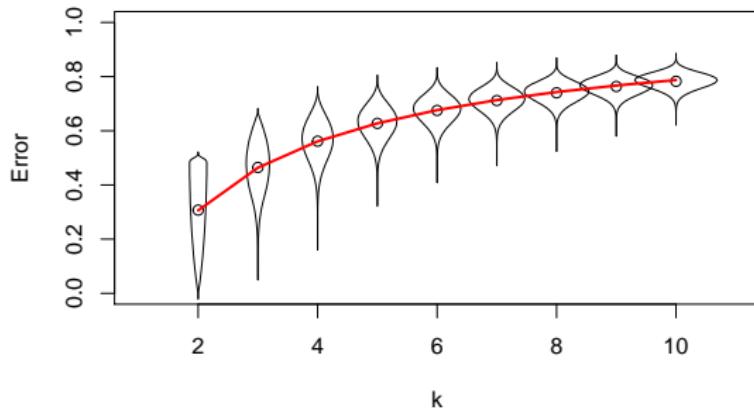


# Toy example



# Toy example

$\rho = 0.7$



# Theoretical Result

**Theorem.** (Z., Achanta, Benjamini.) Suppose  $\pi$ ,  $\{F_y\}_{y \in \mathcal{Y}}$  and marginal classifier  $\mathcal{F}$  satisfy (*some regularity condition*). Then, there exists some function  $\bar{D}(u)$  on  $[0, 1] \rightarrow [0, 1]$  such that the  $k$ -class average risk is given by

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$

# Theoretical Result

**Theorem.** (Z., Achanta, Benjamini.) Suppose  $\pi$ ,  $\{F_y\}_{y \in \mathcal{Y}}$  and marginal classifier  $\mathcal{F}$  satisfy (*some regularity condition*). Then, there exists some function  $\bar{D}(u)$  on  $[0, 1] \rightarrow [0, 1]$  such that the  $k$ -class average risk is given by

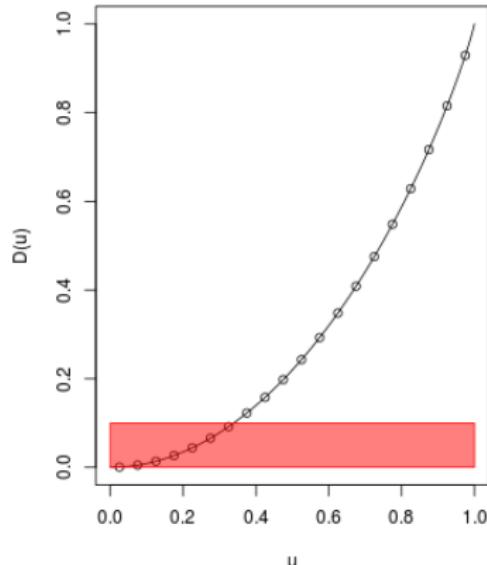
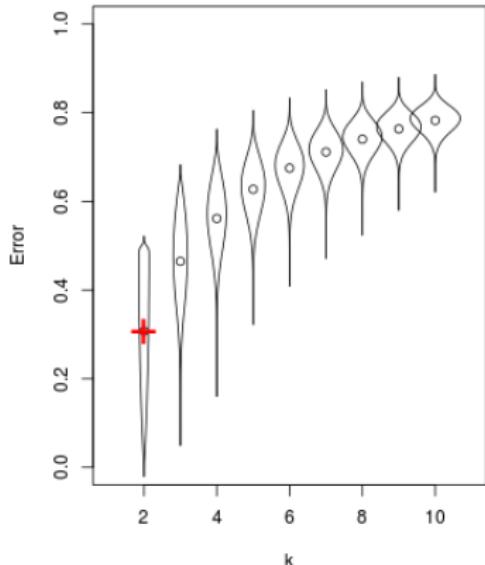
$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$

*Remark.* This result also applies to the *Bayes identification risk*.

# Computing average risk

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$

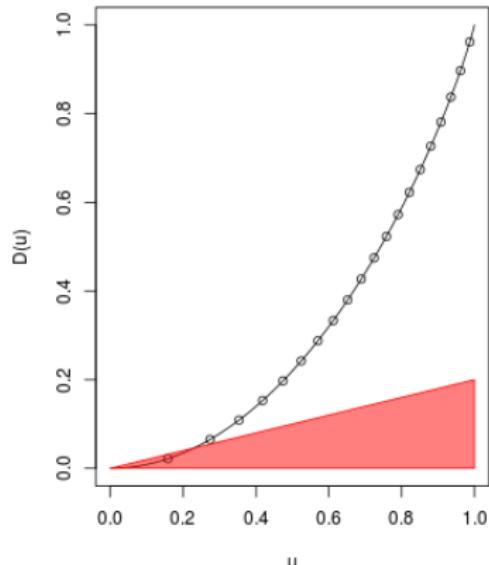
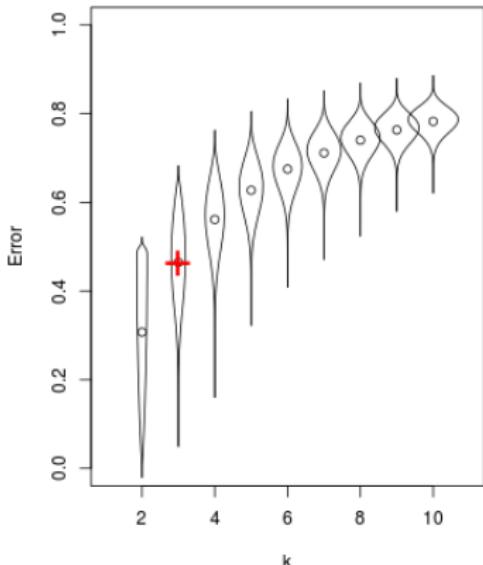
$(k = 2)$



# Computing average risk

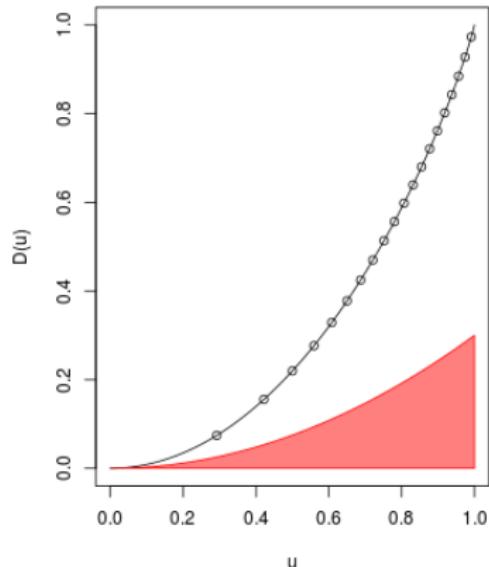
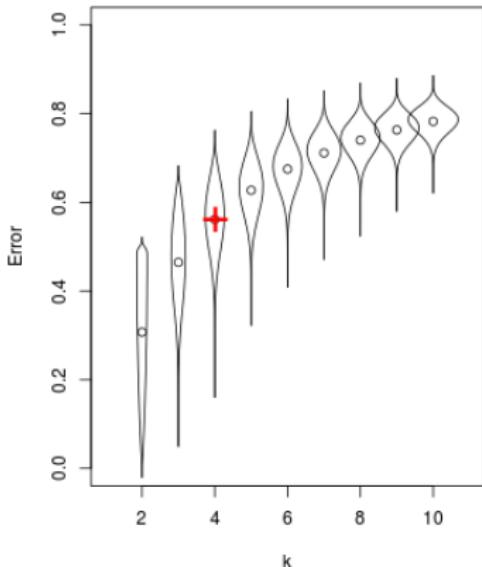
$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$

$(k = 3)$



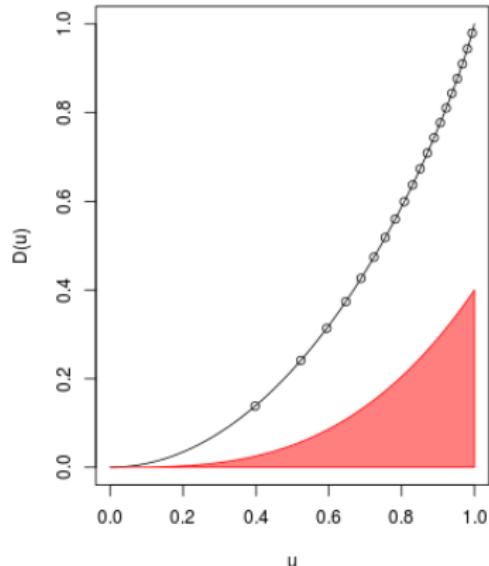
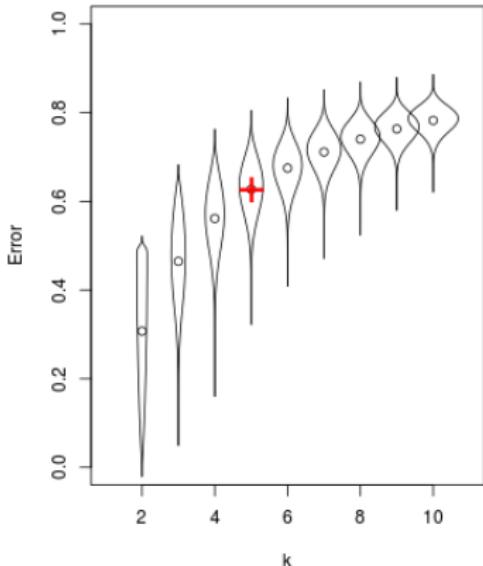
# Computing average risk

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$



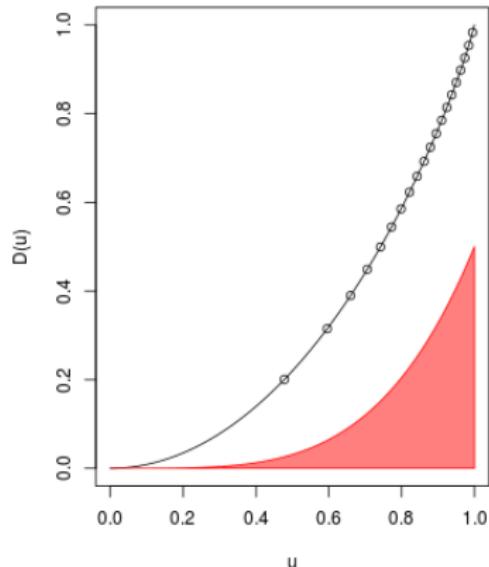
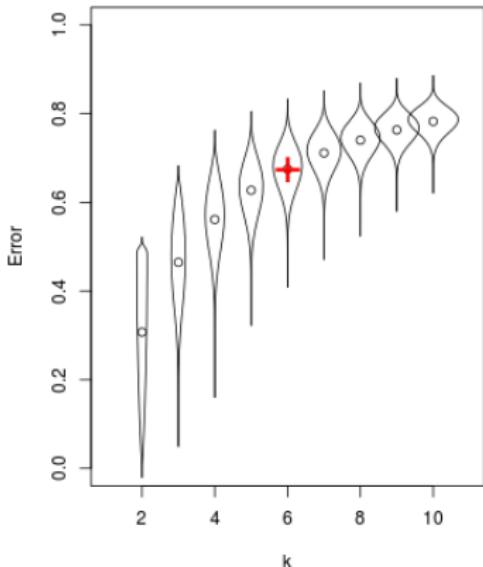
# Computing average risk

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$



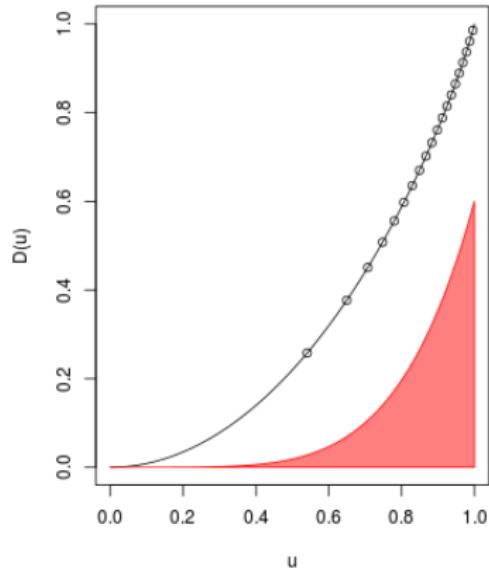
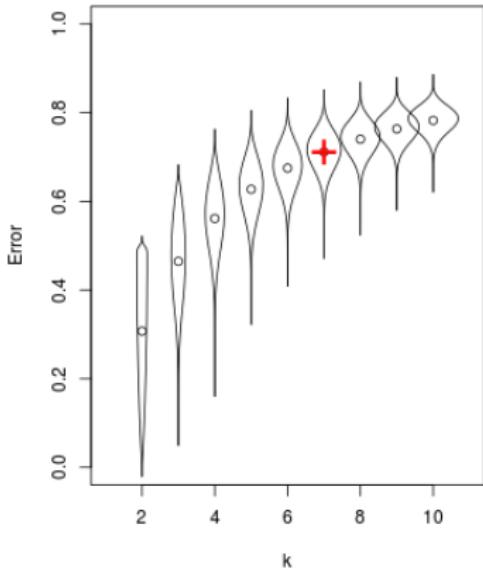
# Computing average risk

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$



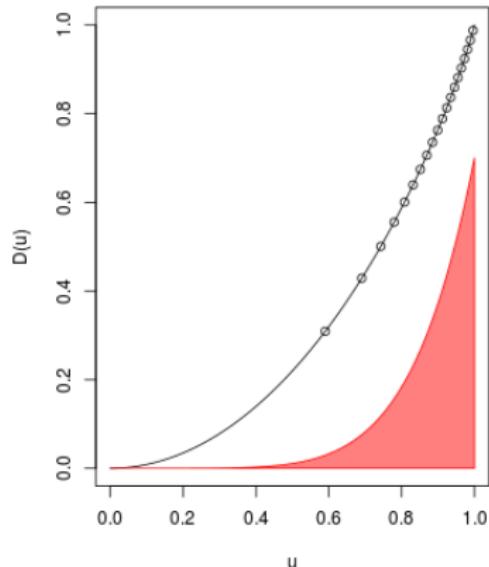
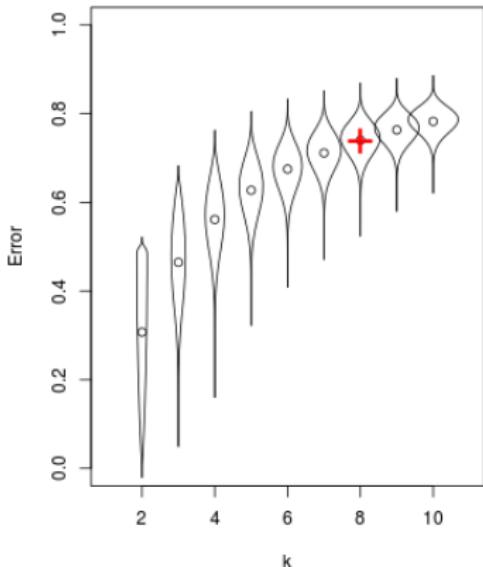
# Computing average risk

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$



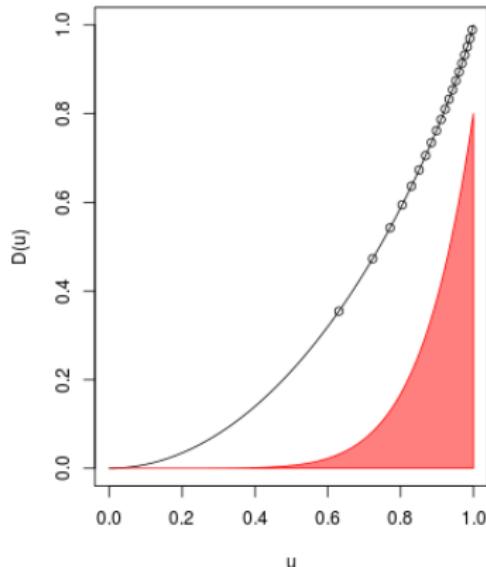
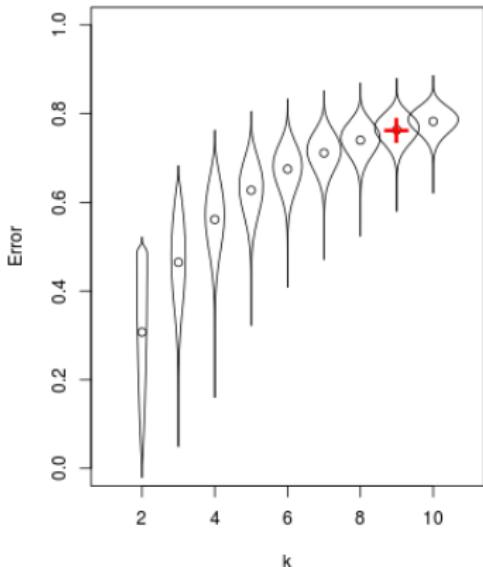
# Computing average risk

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$



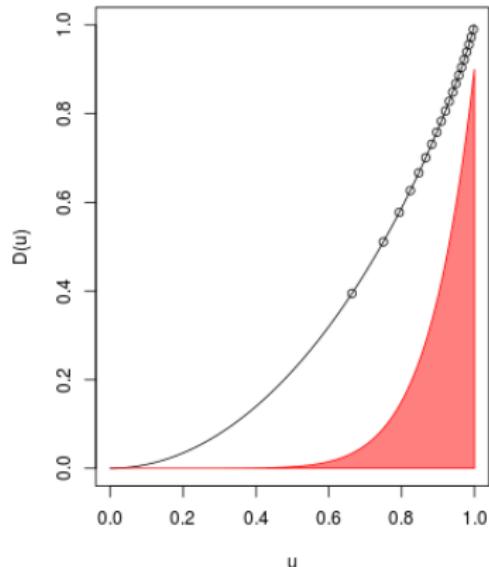
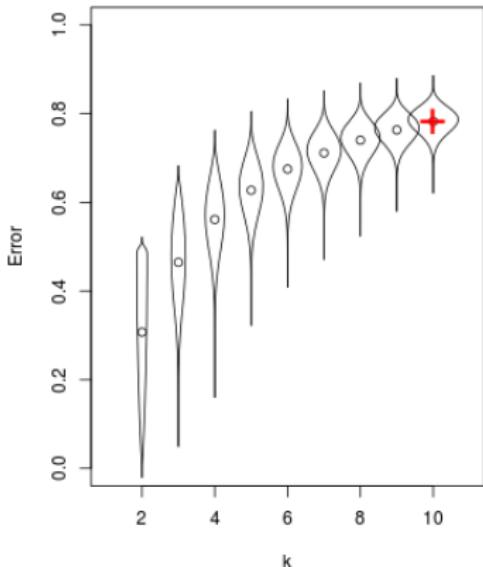
# Computing average risk

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$



# Computing average risk

$$\text{AvRisk}_k = (k - 1) \int \bar{D}(u) u^{k-2} du.$$



## Implication: estimate $\bar{D}(u)$ to predict risk

- Theoretical result links  $k$ -class average risk to  $\bar{D}(u)$  function
- In real data, we do not know  $\bar{D}(u)$  since it depends on the unknown joint distribution
- However, given a model, we can estimate  $\bar{D}(u)$

# So... can accuracy curves intersect?

- In general, the answer is yes.

# So... can accuracy curves intersect?

- In general, the answer is yes.
- However, we will see in the next section that under *high-dimension* assumptions, the Bayes accuracy curves do *not* intersect.

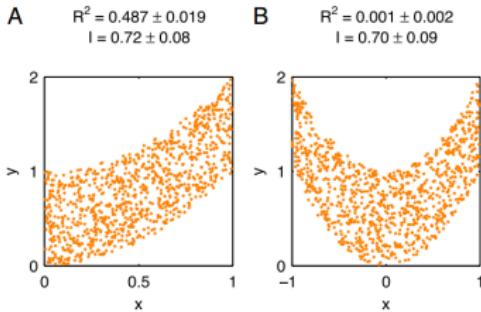
# So... can accuracy curves intersect?

- In general, the answer is yes.
- However, we will see in the next section that under *high-dimension* assumptions, the Bayes accuracy curves do *not* intersect.
- Therefore, a single parameter, the *mutual information*, suffices to summarize the entire curve.

## Section 4

### Mutual Information

# Mutual information $I(X; Y)$



Introduced in Shannon's 1948 paper, "A mathematical theory of communication"

$$I(X; Y) = \int \log \left( \frac{p(x, y)}{p(x)p(y)} \right) p(x, y) dx dy$$

Image credit Kinney et al. 2014.

# Result 1. Lower bound for mutual information

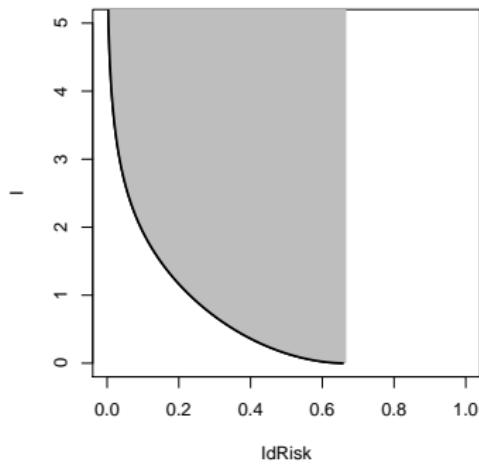
- Define the identification risk as the expected identification loss

$$\text{IdRisk}_k = \mathbf{E}[\text{IdLoss}_k]$$

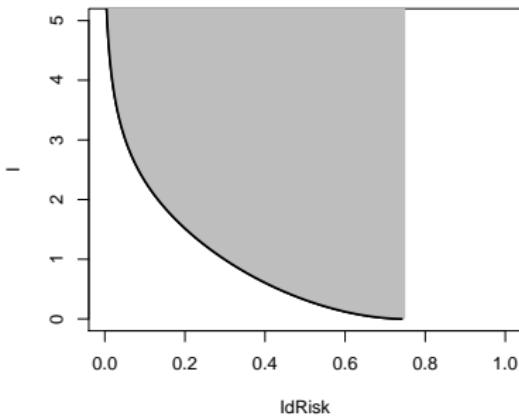
- Theorem.** (Z., Benjamini 2017) There exists a function  $h_k$  such that

$$I(\vec{g}(\vec{Z}); \vec{Y}) \geq h_k(\text{IdRisk}_k).$$

$h_3$



$h_4$



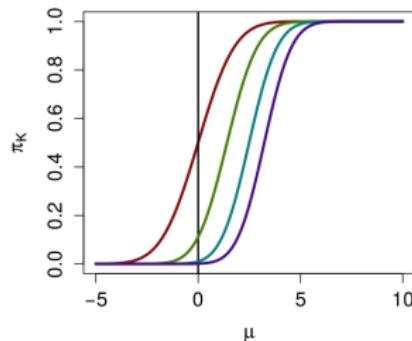
## Result 2. Limiting behavior of accuracy curves

Define  $\text{ABA}_k$  as the Bayes identification accuracy (or average Bayes classification accuracy). Then under a particular high-dimensional limit,

$$\text{ABA}_k \approx \pi_k(\sqrt{2I(X; Y)}) \quad (1)$$

The function  $\pi_k$  is given by

$$\pi_k(c) = \int_{\mathbb{R}} \phi(z - c) \Phi(z)^{k-1} dz.$$

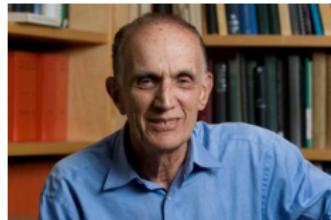


Legend:  $K = \{ 2, 9, 99, 999 \}$

## Section 5

### Acknowledgements







## Section 6

The end