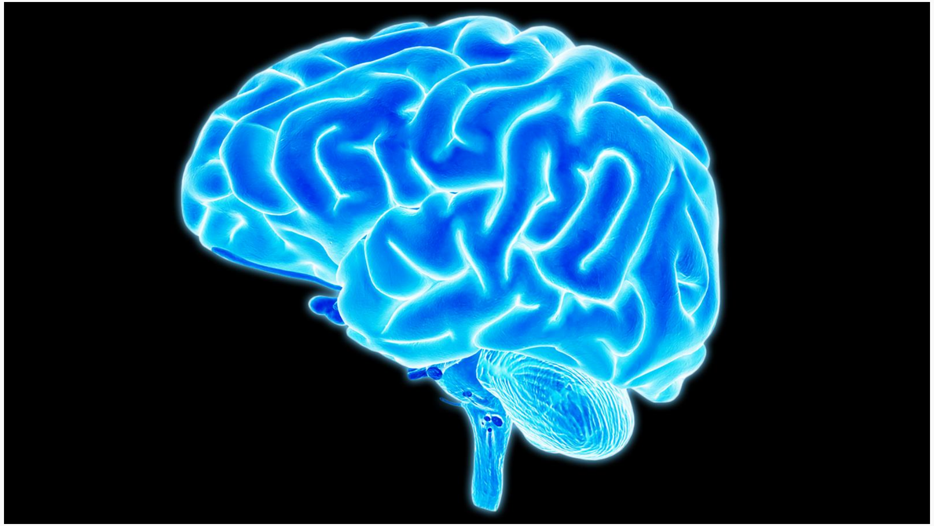
A Semantic Shared Response Model

Kiran Vodrahalli*, Po-Hsuan Chen*, Janice Chen*, Esther Yong †, Christopher Honey †,
Peter J. Ramadge*, Kenneth A. Norman*, Sanjeev Arora*
ICML MVRL 2016
June 23, 2016

* = Princeton, † = U. Toronto

fMRI: Sensing Brain Signal



100 billion neurons in the brain

fMRI measures hemodynamic response at ~10^5 different 3mm x 3mm x 3mm voxels

Each voxel represents an average of the activity of the ~10^6 neurons it contains

Goal: **detect semantic meaning in this signal.**

Prior Work on Decoding Semantic Content from fMRI

[Mitchell et al ‘08] predicts fMRI responses induced by **pictures of concrete nouns**.

[Naselaris et al ‘09] predicts fMRI responses induced by **images of scenes.**

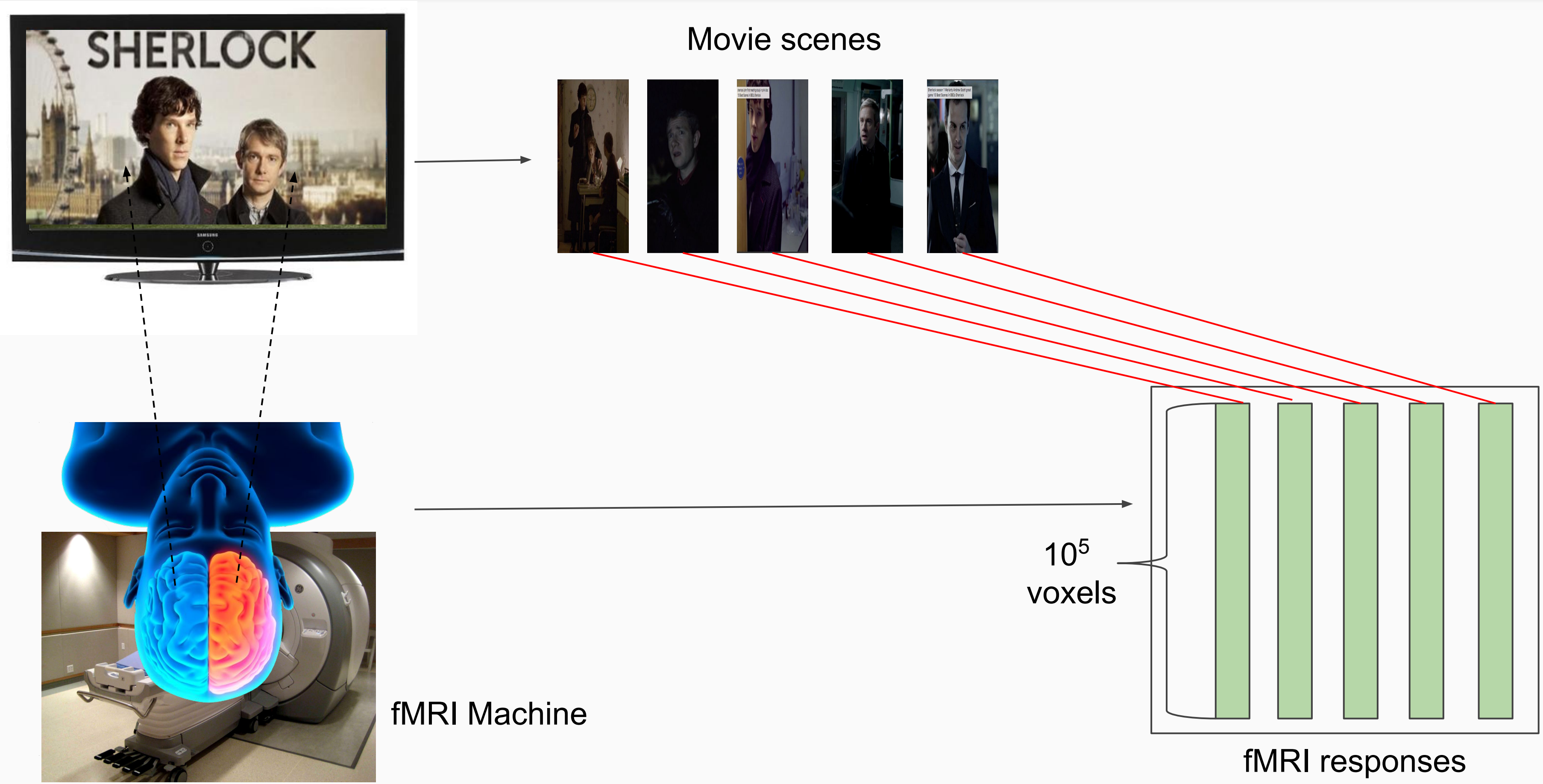
[Pereira et al ‘11] uses the same dataset as Mitchell ‘08, but focuses on **generating words** related to the concrete nouns.

[Naselaris et al ‘11] tries to **reconstruct movie images** from fMRI signals measured while subjects watched movies.

[Wehbe et al ‘14] has subjects **read a chapter of Harry Potter** and predicts fMRI responses for held-out time points.

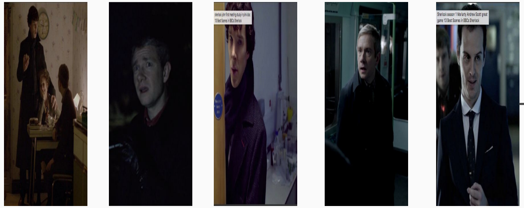
[Huth et al ‘16] reconstructs fMRI responses to **auditory stories**.

Goal 1: Decode fMRI Response Semantics



Goal 1: Match fMRI responses to annotations (Views: fMRI signal, text annotations)

Movie scenes



Annotations of movie scenes

Sherlock and John talk about the murder in an old room with Mrs. Hudson.

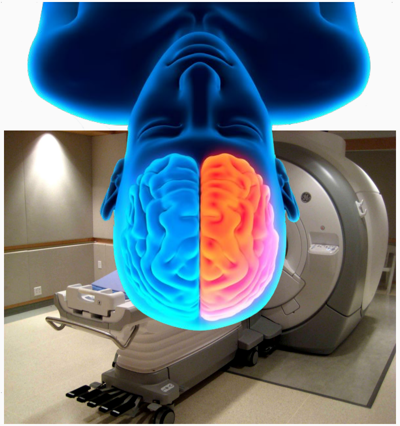
John is worried as Sherlock runs off.

Sherlock enters the door to the chemistry lab, saying "John, I was here the whole time."

Once they get on the subway, John exclaims, "No you weren't!"

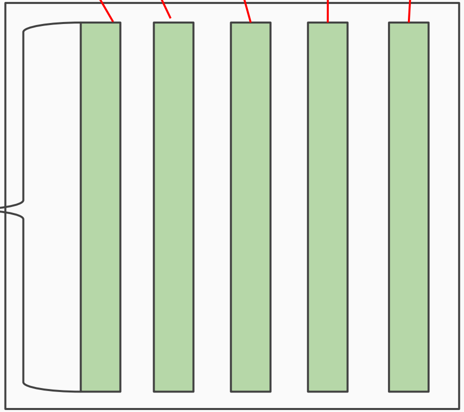
Moriarty arrives and says, "Hello Sherlock, John."

Each movie scene paired with text description from external party.



fMRI Machine

10^5 voxels



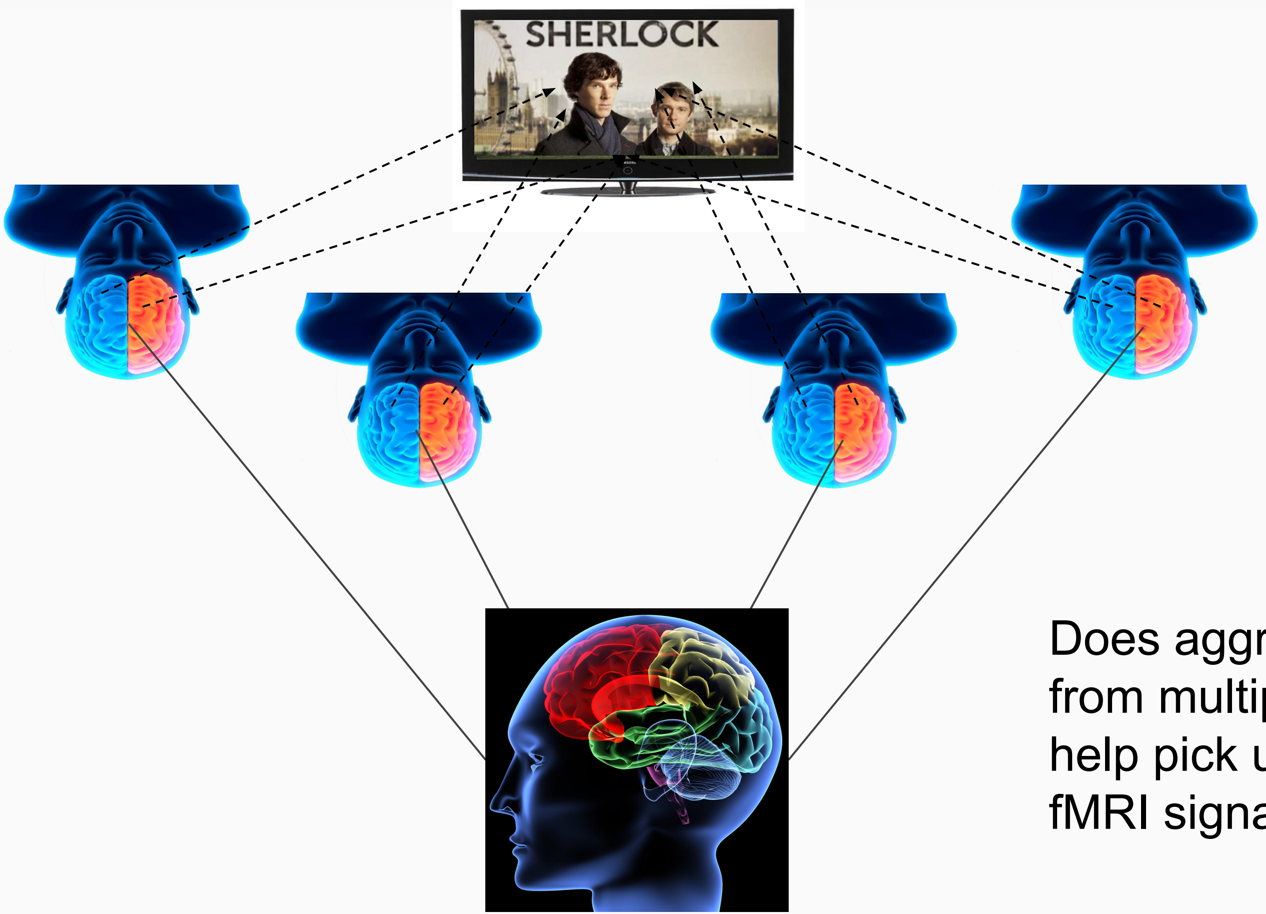
fMRI responses

Goal 2: Leverage Multiple Subject Views to Extract Better Semantics

Shared Movie Stimulus

Multiple Subject Responses

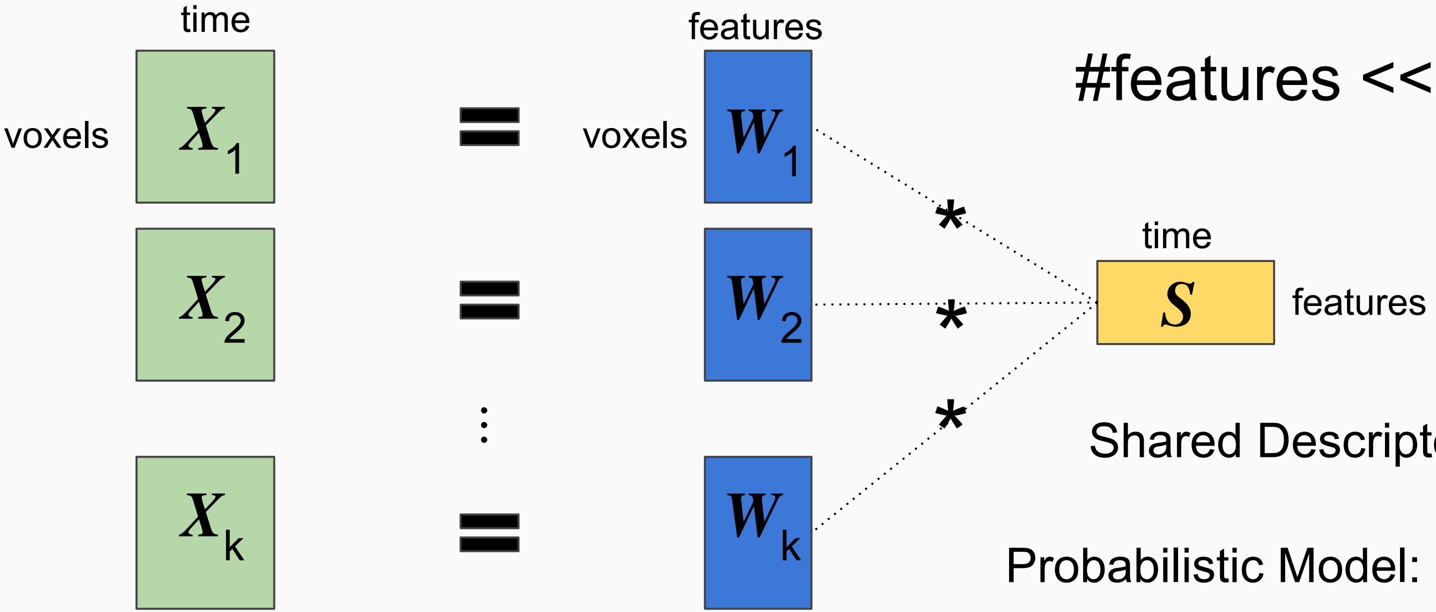
Shared fMRI Response



Does aggregating data from multiple individuals help pick up a stronger fMRI signal?

Shared Response Model (SRM, [Chen, Chen, Yeshurun, Hasson, Haxby, Ramadge ‘15])

#features << #voxels



Shared Descriptor of Semantic Signal

$$\underset{W^TW=I;S}{\operatorname{argmin}} \sum_{i=1}^{k} \|X_i - W_i S\|_F$$

Probabilistic Model:

$$s_t \sim \mathcal{N}(0, \Sigma_s)$$

$$x_{it} | s_t \sim \mathcal{N}(W_i s_t + \mu_i, \rho_i^2 I)$$

Key Differences From Prior Work

Natural, audio-visual dataset + text annotations

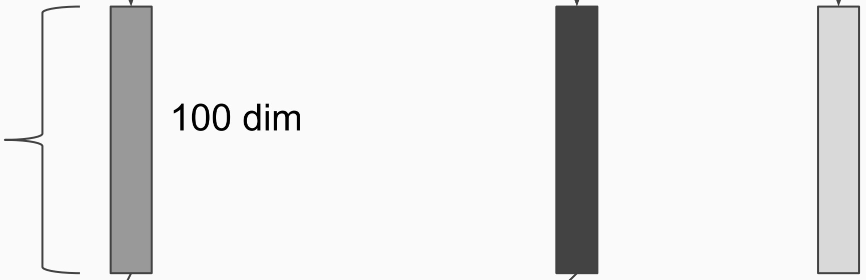
Aggregating multiple subjects improves performance.

We use **semantic word embeddings** and **atoms of discourse** to represent the text annotations.

Representing Annotations as “Semantic Vectors” [Arora, Li, Liang, Ma, Risteski ‘15, ‘16]

Annotation Text: … door at the murder scene …

Word Vectors from Wikipedia:



Then find a 3-sparse “basis” for the word vectors to get **atoms of meaning**.

Decomposition into Atoms: $\omega_1^1 + \omega_1^2 + \omega_1^3$

$\omega_2^1 + \omega_2^2 + \omega_2^3$

$\omega_3^1 + \omega_3^2 + \omega_3^3$

Sort the atoms by their aggregate weights and pick the top 4: $\omega_*^1 + \omega_*^2 + \omega_*^3 + \omega_*^3$ = Final Context Vector

Semantic Context Example

``Donovan looks up at the reporters and continues: `Preliminary investigations...' Lestrade looks distressed. Donovan continues: `... suggest that this was suicide. We can confirm that this...''

After creating the semantic vector for this annotation, the words nearby are:

1) *investigation* (corr. = 0.78)

2) *suicide* (corr. = 0.74)

3) *CNN* and *Reuters* (corr. = 0.71)

4) *police* (corr. = 0.70)

Shared fMRI ↔ Semantic Embeddings

Brain ROIs: We construct shared fMRI space for several ROIs, including the **Default Mode Network (DMN)** which prior work suggests encodes semantics.

Dimensionality: We learn maps between the low-dimensional shared space (k = 20, 50, 100 dims) and semantic space (100 dim). Empirically, k = 20 was best and is justified by the approx. low-rank of the fMRI data for the DMN region.

Learning Linear Maps: 1) Ridge regression regularizes via $\|\ \|_2$

2) Procrustes problem regularizes via orthogonality

Performs poorly ~~3) Apply SRM to (shared fMRI space, semantic space)~~

Classification Results for DMN Region

	fMRI → Text	Text → fMRI
Binary Classification Leave 2 scenes out and match (chance 50%)	70%	83%
Scene Classification Train first ½, test second ½ (Top-5 rank: chance 20%)	49%	50%

Shared Response Model Improves Voxel Reconstruction

Voxel Reconstruction measures the Pearson correlation between held-out fMRI response and predicted fMRI response from semantic embeddings via ridge regression.

	Corr. (true fMRI, pred. fMRI) (DMN Region)
Without SRM	0.04
With SRM	0.11

Conclusions

Decoding accuracy for fMRI -> Text (70%) comparable to similar settings ([Pereira et al ‘11, ‘16]).

Decoding accuracy for Text -> fMRI (82%) comparable to [Mitchell et al ‘08], [Wehbe et al ‘14] which use similar tasks.

SRM improves voxel reconstruction performance by factor of 3.

Results corroborate prior work suggesting the DMN plays a role in representing semantics.

Open Questions

We would like to output captions of fMRI stimulus as in the **image captioning** literature.

We would like to add **video** to the semantic representation.

Do **nonlinear** models work better than linear maps?

Explain the necessity of the **orthogonal constraint** for decoding text.

Temporal receptive windows: Learn map from surrounding variable-size window of fMRI time points to predict semantic vector.