Mappings Between fMRI Responses and Natural Language Descriptions of Natural Stimuli

Kiran Vodrahalli*, Po-Hsuan Chen*, Yingyu Liang*, Christopher Baldassano*, Janice Chen✦, Esther Yong †, Christopher Honey✦, Peter J. Ramadge*, Kenneth A. Norman*, Sanjeev Arora*

MURI Grant Kickoff Meeting
September 17, 2017

* = Princeton, ✦ = Johns Hopkins, † = U. Toronto

Prior Work on Connecting a Semantic Space to fMRI Data

[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.

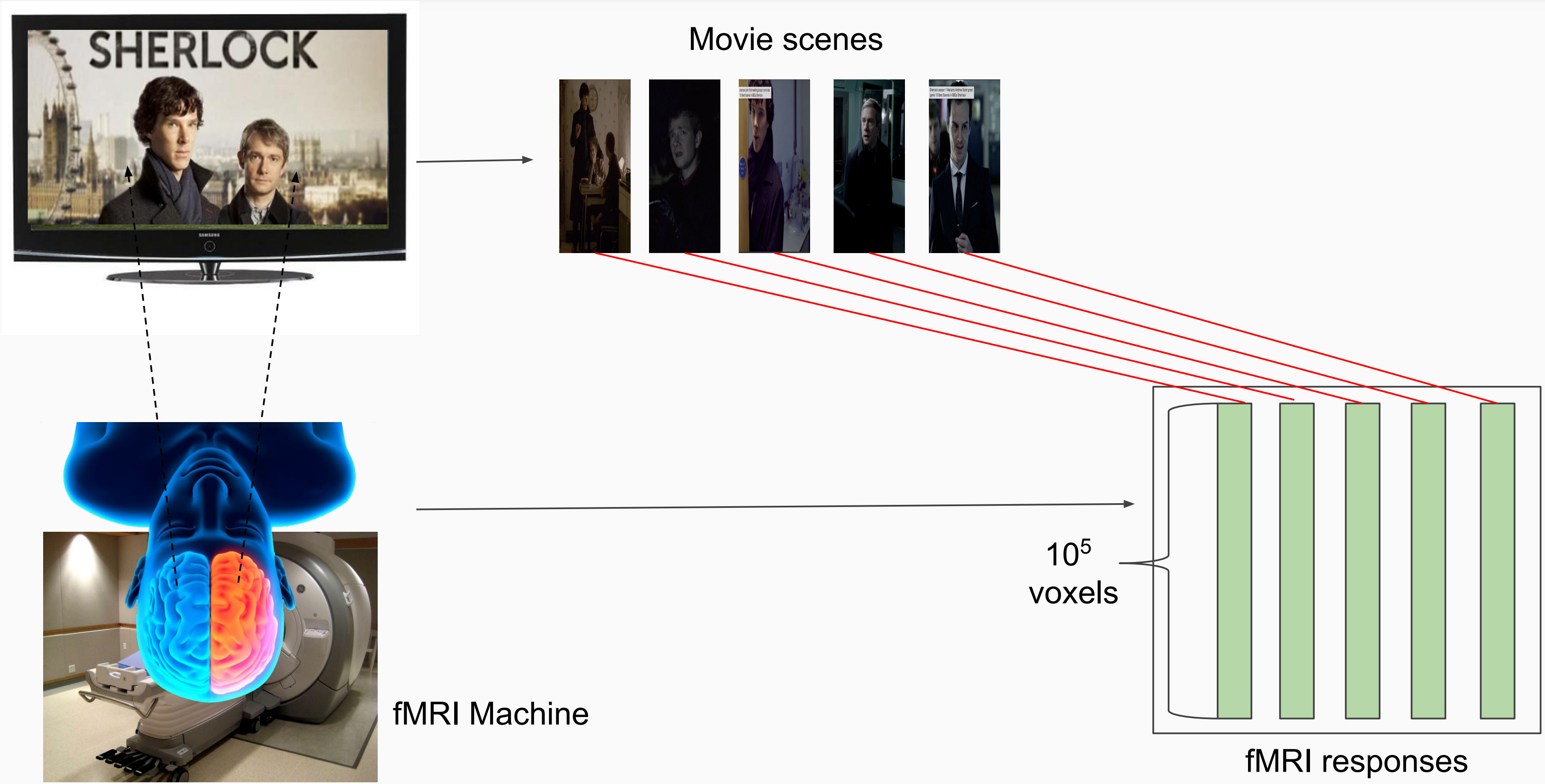
[Naselaris et al ‘11] tries to **reconstruct movie images** from fMRI signals measured while subjects watched movies.

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

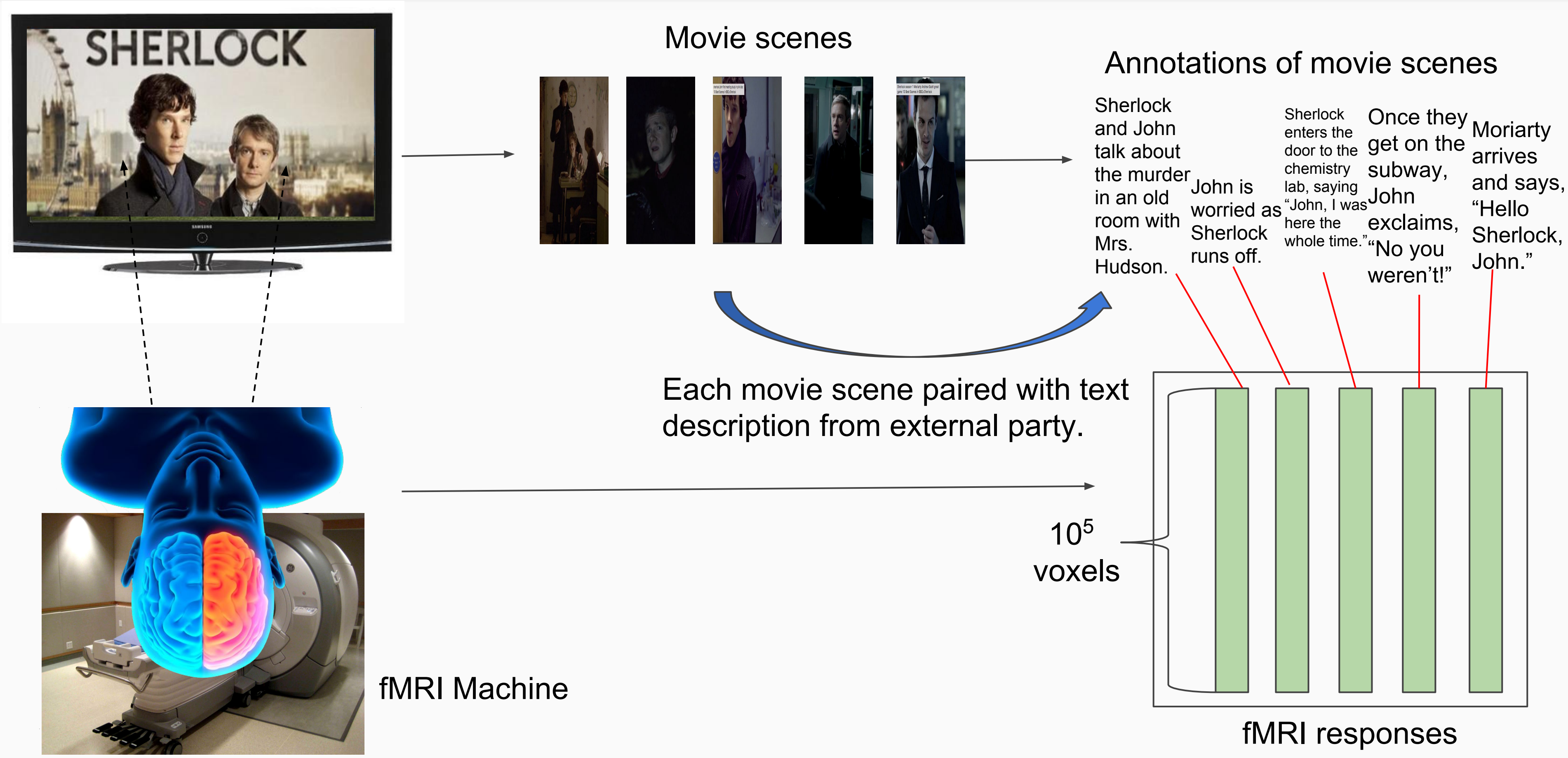
[Huth et al ‘16] reconstructs fMRI responses to **auditory stories**.

[Pereira et al ‘16] decodes fMRI responses to **word clouds and short sentences.**

Main Goal: Decode fMRI Response Semantics



Matching fMRI responses to annotations (Views: fMRI signal, text annotations)



Representing Annotations: Word Embeddings

- To map between fMRI vectors and words, we would like to represent words in vector space
- Goals of embeddings: Preserve some notions of similarity and distance that apply to the words
- Idea: Assign to every word a 100-dim vector
- How?

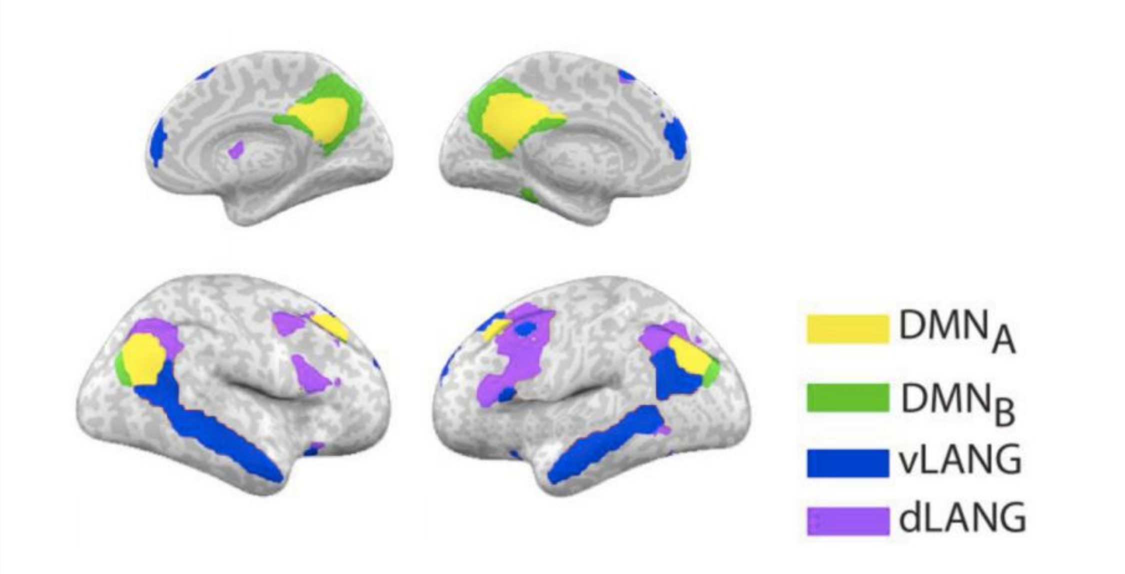
Training Word Embeddings

- Idea: Train a predictive model on some task which “captures the meaning of natural language”.
  - Can be an unsupervised task, i.e. “language modeling”.
- Parameters of the model include a subset which are assigned uniquely to each word
  - Initialized randomly
- Training the model on external training set → word vectors
- How to combine word vectors to get “annotation vectors”?

Interesting and Useful Discoveries

- The Shared Response Model (SRM, Chen et al. 2015) helps for decoding text!
- Weighted average word vectors → better semantic context vectors (ICLR 2017 paper, Arora et al)

Brain Regions (ROIs) Studied



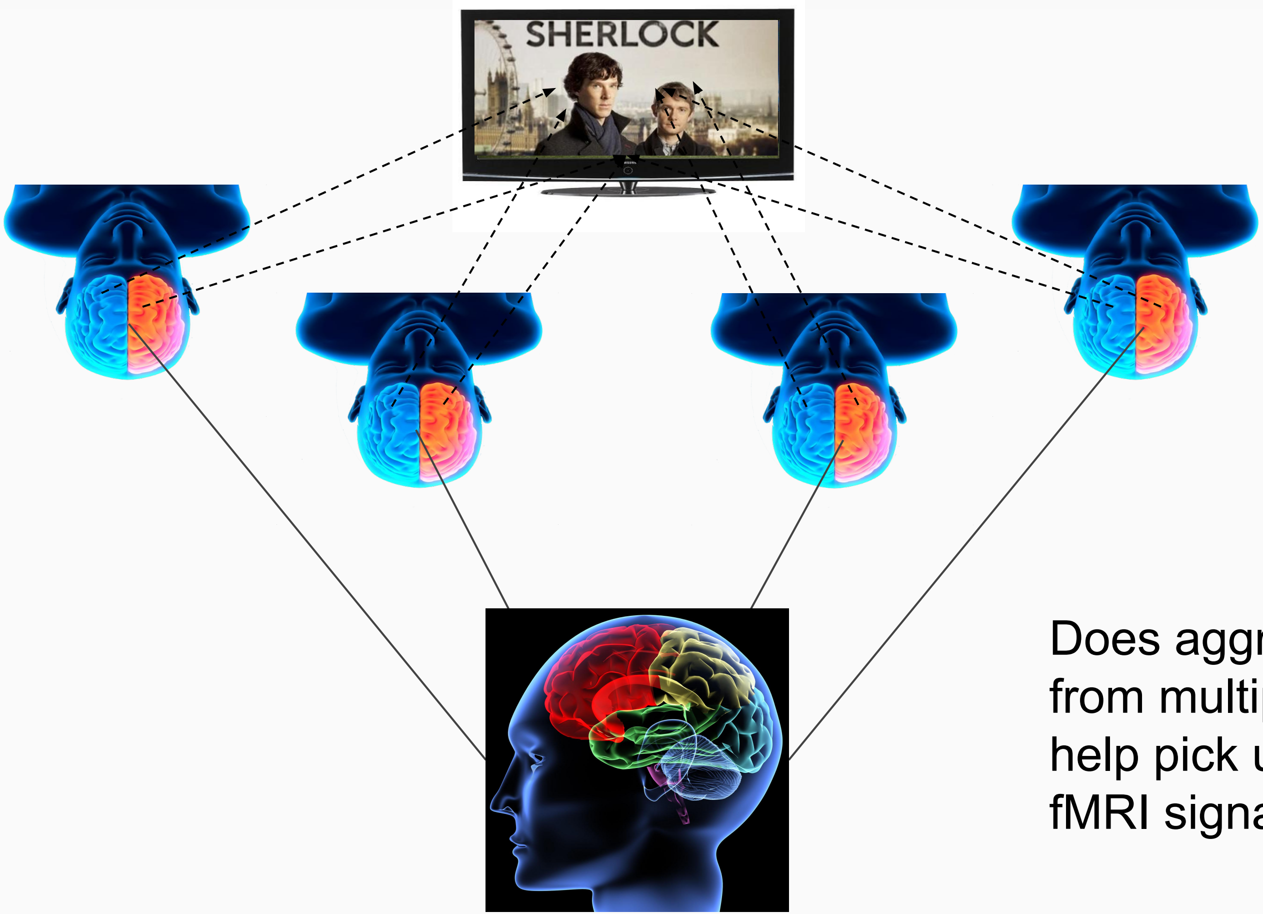
- Default Mode Network (DMN) standard area in literature
  - known to relate to narrative processing
  - DMN-A, -B (2000 voxels)
- Ventral/Dorsal Language (2000 voxels)
- Whole Brain (26000 voxels)
  - voxels with high inter-subject correlation
- Occipital Lobe (6000 voxels)

Leveraging Multiple Subject Views to Extract Better Semantics

Shared Movie Stimulus

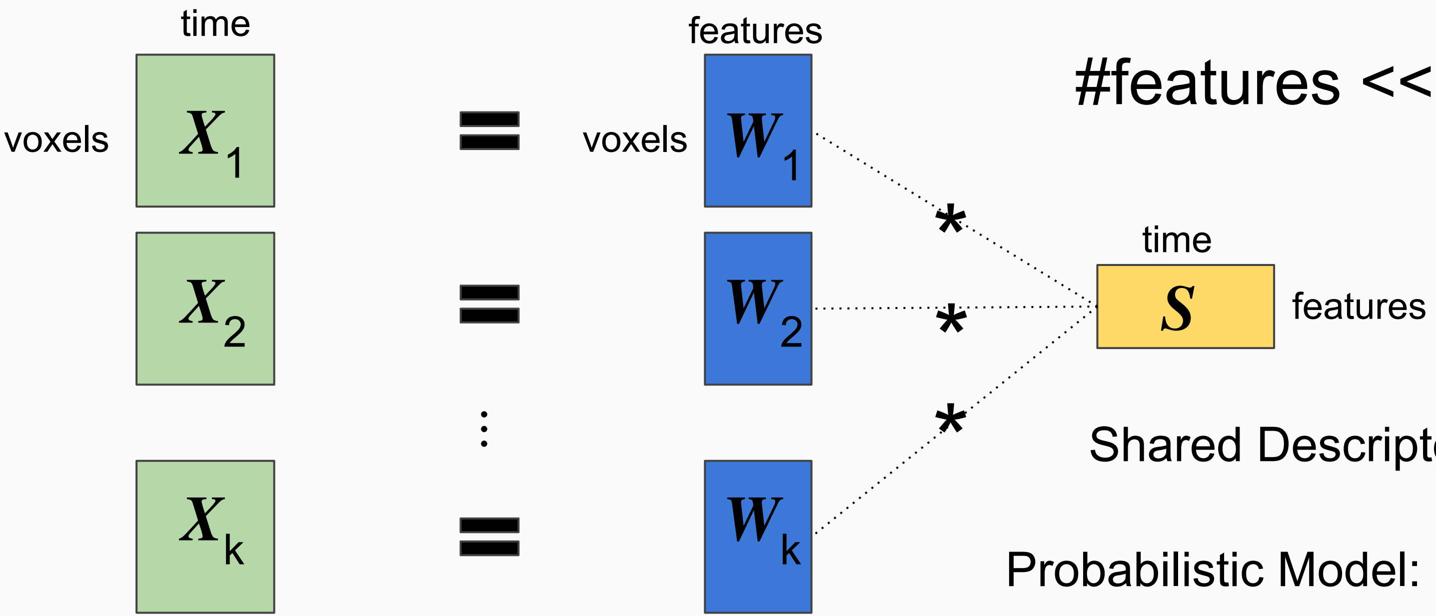
Multiple Subject Responses

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

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

Probabilistic Model:

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

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

Embedding Annotations with Weighted Sums of Word Vectors

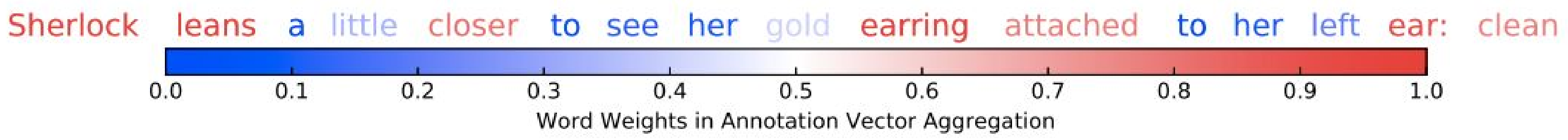


Fig. 3. Visualization of Semantic Annotation Vector Weightings: We display an example sentence from the Sherlock annotations, where we have colored important words red, and unimportant words blue. Brighter red means more important, and darker blue means less important.

Linear Maps Between fMRI and Text

Basic Model:

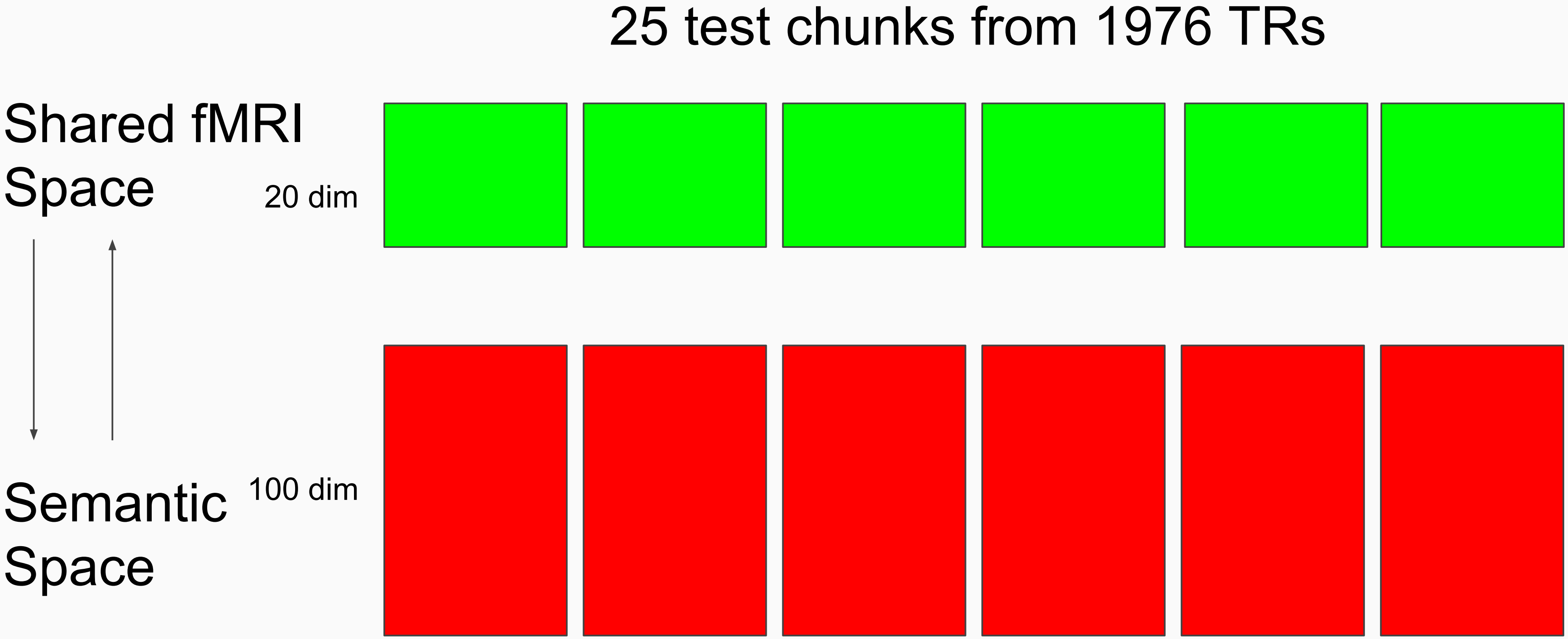
$$WX = Y \quad , \quad W \in \mathbb{R}^{m \times n}$$

X represents the fMRI data matrix (n x T)
Y represents the semantic annotation data matrix (m x T)

Learning the Map:

- Procrustes ($W^T W = I$)
  - Restrict map to "rotations" of the data.
  - Imposes strong constraint on map
- Ridge Regression (penalize l_2 norm $||w||_2$)
  - Classic linear regularization method
  - Restricts map weights to be uniformly small (not sparse)

Evaluation: Scene Classification/Ranking Experiments



Results: Multiplicative Improvements with our Methods

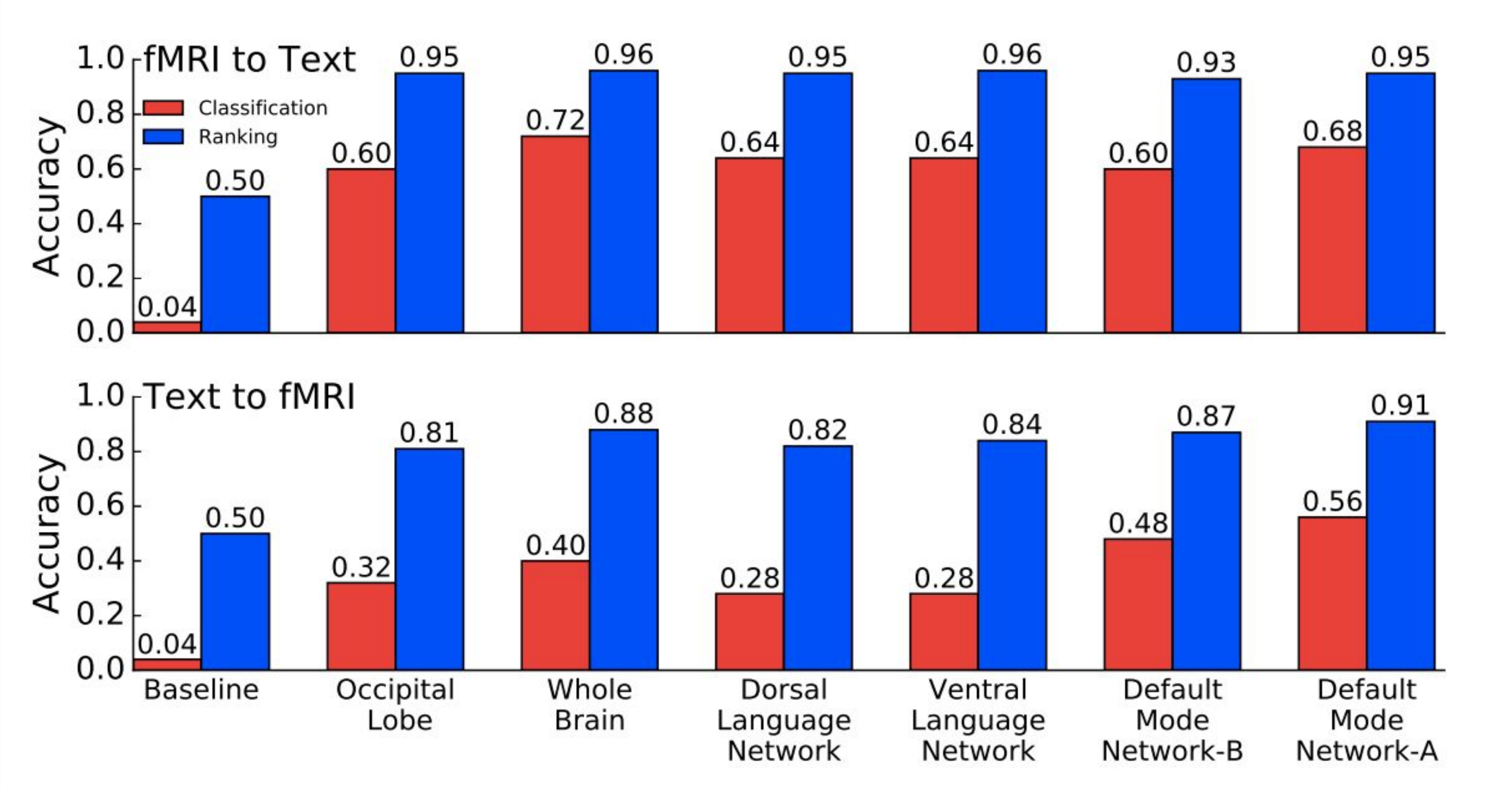
Mapping Between fMRI Responses and Semantic Representations

fMRI → Text	Maximum	Average
Procrustes vs. Ridge	2.8×	1.3×
SRM/SRM-ICA vs. PCA	1.8×	1.3×
Weighted-SIF vs. Unweighted	1.6×	1.2×

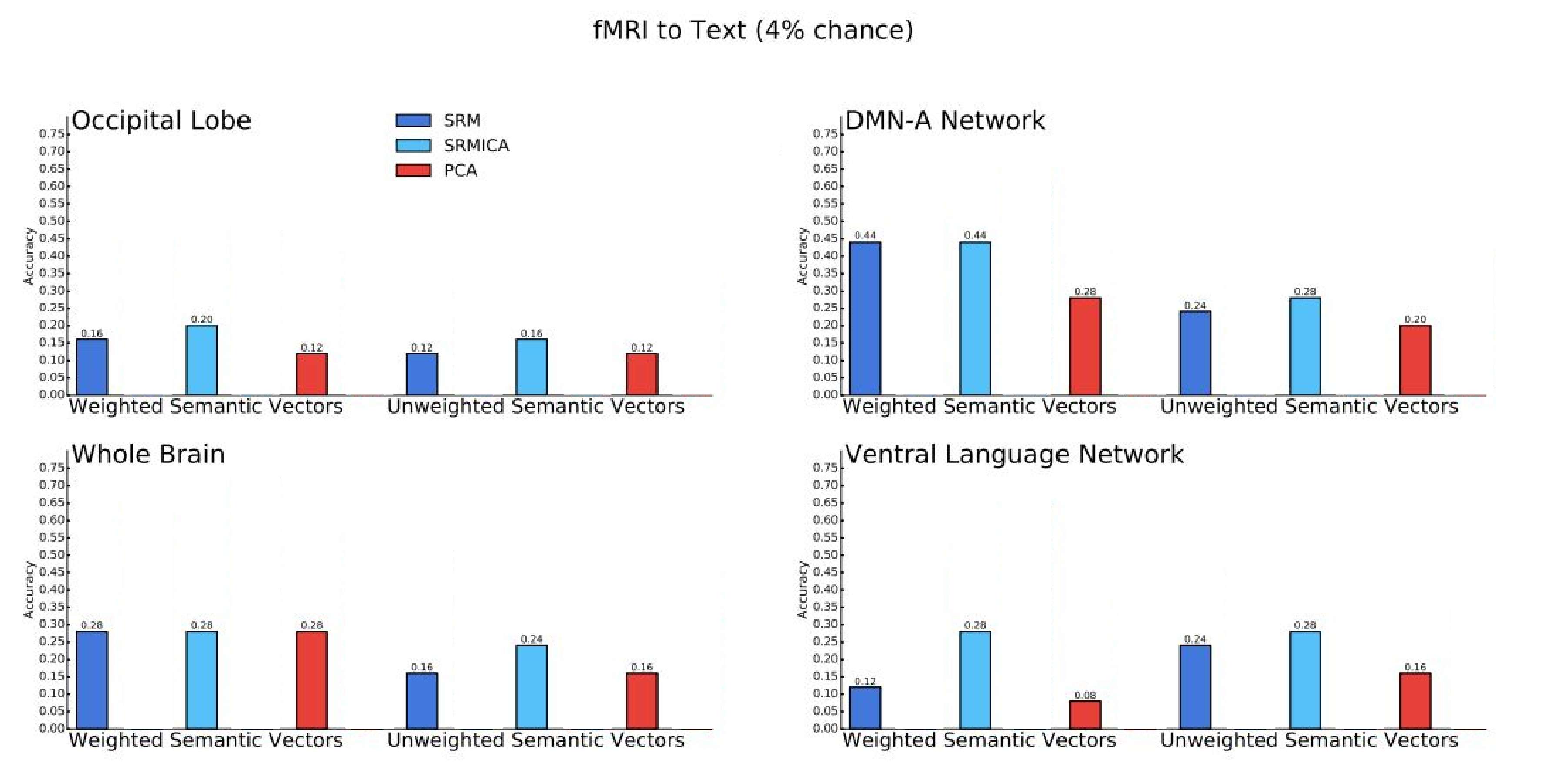
Text → fMRI	Maximum	Average
Procrustes vs. Ridge	3.0×	0.8×
SRM/SRM-ICA vs. PCA	2.3×	1.2×
Weighted-SIF vs. Unweighted	1.8×	1.1×

Table 1. Table of Improvement Ratios for Various Algorithmic Parameters: In this table we give the maximum and average improvement ratios for a specific algorithmic technique over another.

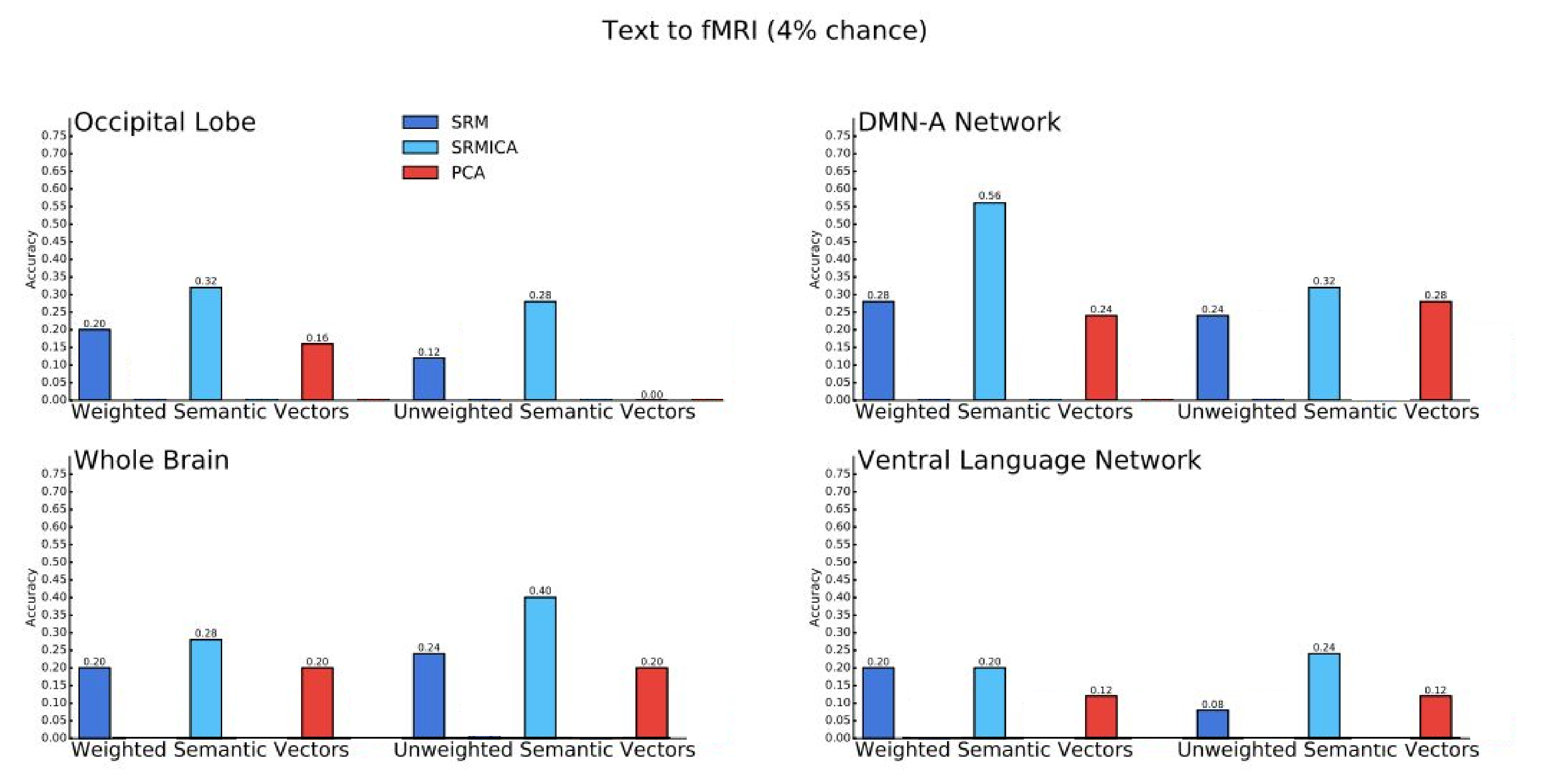
Results: Top-4% Classification and Average Rank



Results: Comparisons for fMRI → Text (4% Chance)



Results: Comparisons for Text → fMRI (4% Chance)



Performance on the Green Eyes Dataset (Yeshurun et al, 2017)

fMRI->Text Scene Classification	Maximum	Average
SRM vs. Average	4.547x	1.909348x
Weighted vs. Unweighted	2.182211x	1.17182x
Text->fMRI Scene Classification	Maximum	Average
SRM vs. Average	2.986x	1.431645x
Weighted vs. Unweighted	3.386167x	1.35073x

(Results from Viola Mocz)

Ongoing and Future Work

- Applying event segmentation to define scenes in classification and ranking tasks
- Using previous timesteps to get better classification accuracy
- Understanding gap between fMRI $\rightarrow$ Text and Text $\rightarrow$ fMRI
- Finer-grained annotation embeddings
- More datasets
- Genuine scene description decoding