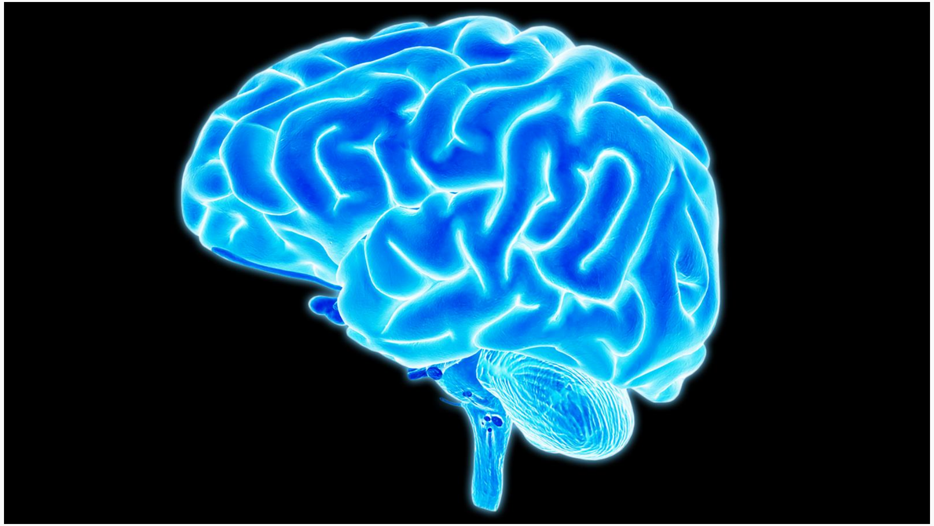
Temporally Dependent Mappings Between fMRI Responses and Natural Language Descriptions of Natural Stimuli

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

COS MSE Master's Thesis Presentation
May 10, 2017

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

fMRI: Sensing Brain Signal



100 billion neurons in the brain

fMRI measures hemodynamic response at $\sim 10^5$ different 3mm x 3mm x 3mm voxels

Each voxel represents an average of the activity of the $\sim 10^6$ neurons it contains

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

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.

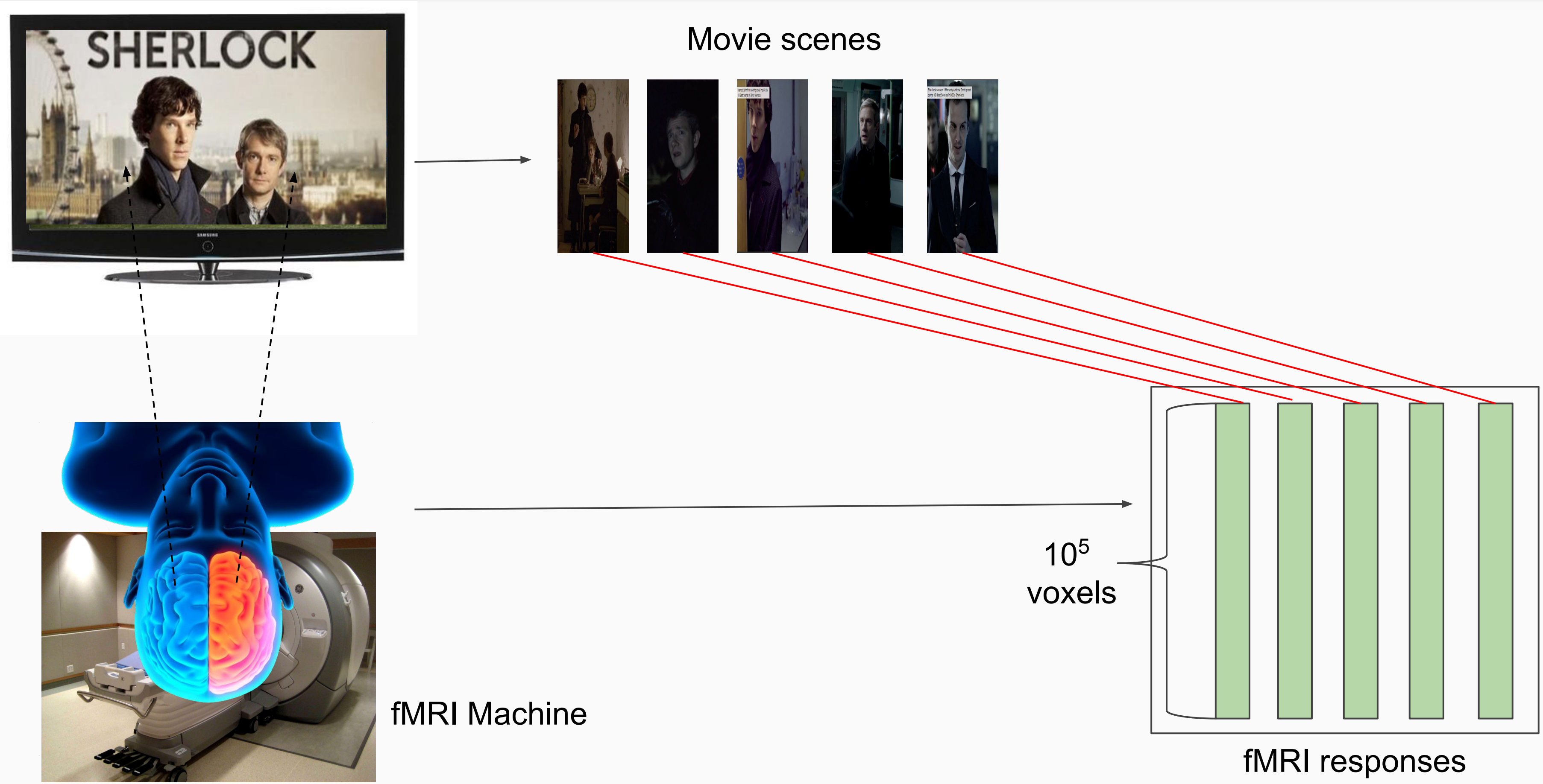
[Naselaris et al ‘11] tries to **reconstruct movie images** from fMRI signals measured while subjects watched movies.

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

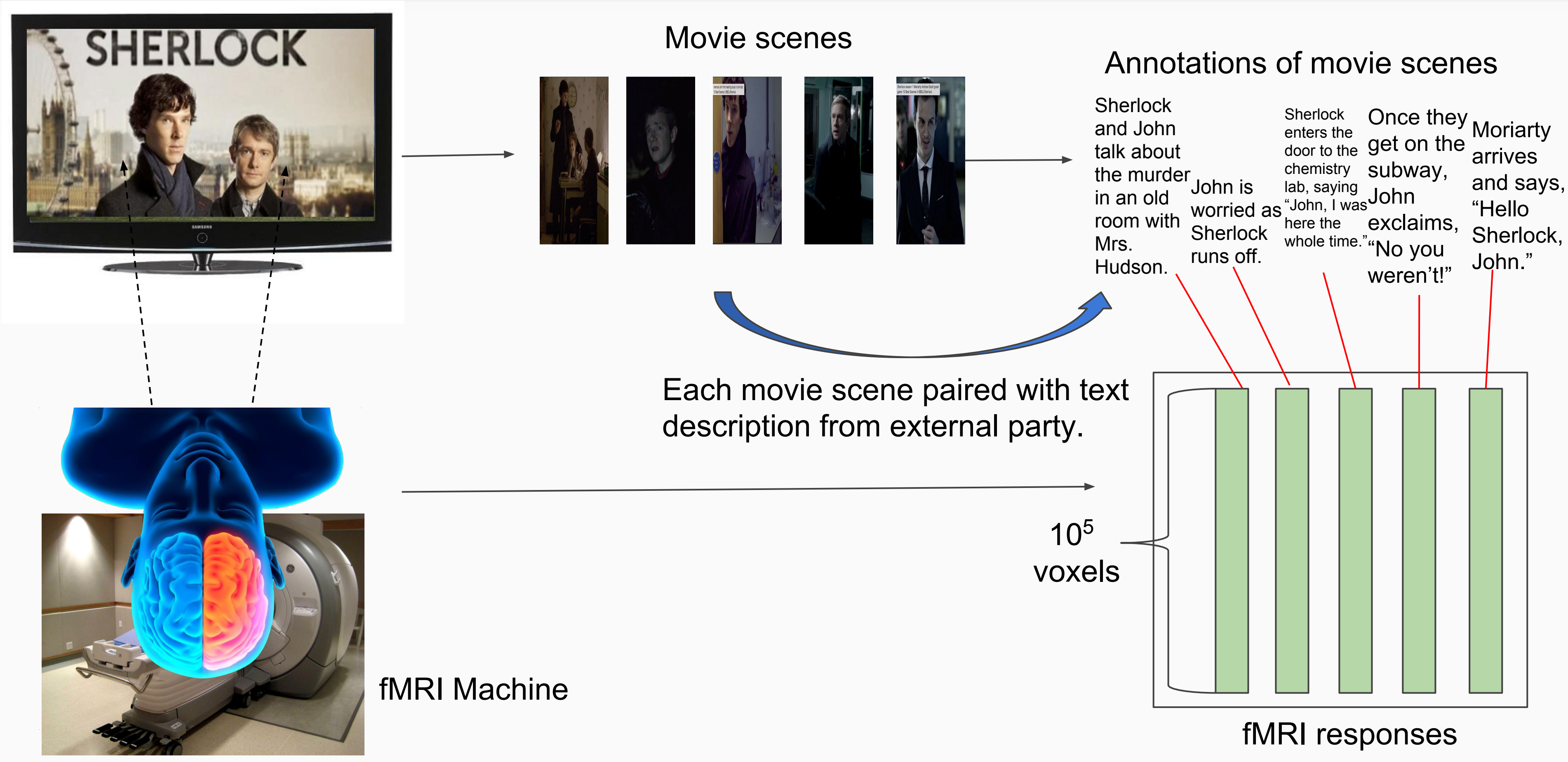
[Huth et al ‘16] reconstructs fMRI responses to **auditory stories**.

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

Main Goal: Decode fMRI Response Semantics



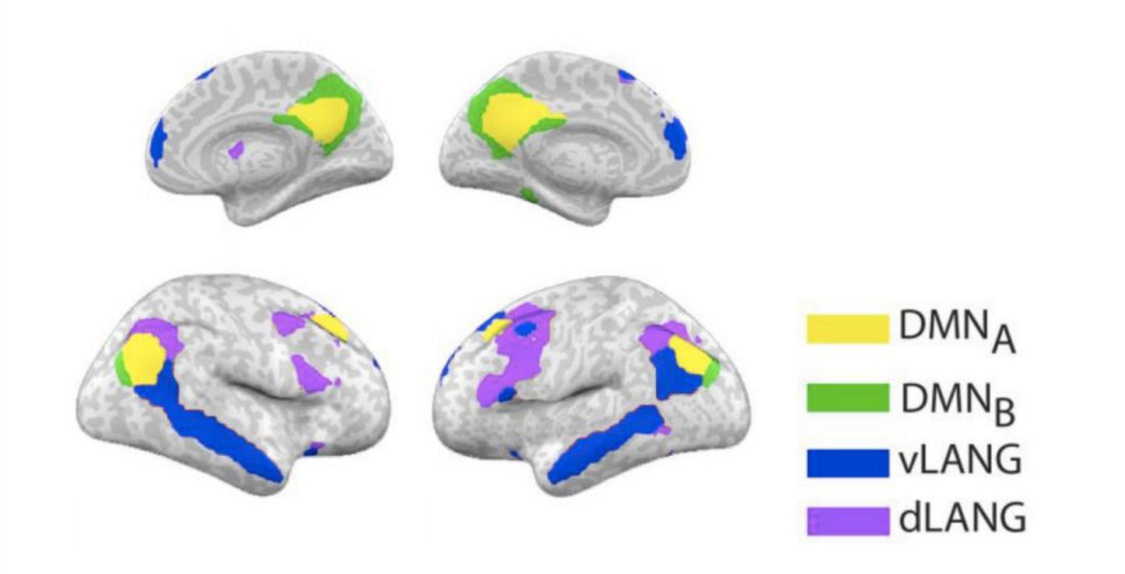
Matching fMRI responses to annotations (Views: fMRI signal, text annotations)



Interesting and Useful Discoveries

- The Shared Response Model (SRM, Chen et al. 2015) helps for decoding text!
- Weighted average word vectors $\rightarrow$ better semantic context vectors (ICLR 2017 paper, Arora et al)
- Using previous time points helps a lot for mapping fMRI $\rightarrow$ text, but hurts text $\rightarrow$ fMRI

Brain Regions (ROIs) Studied



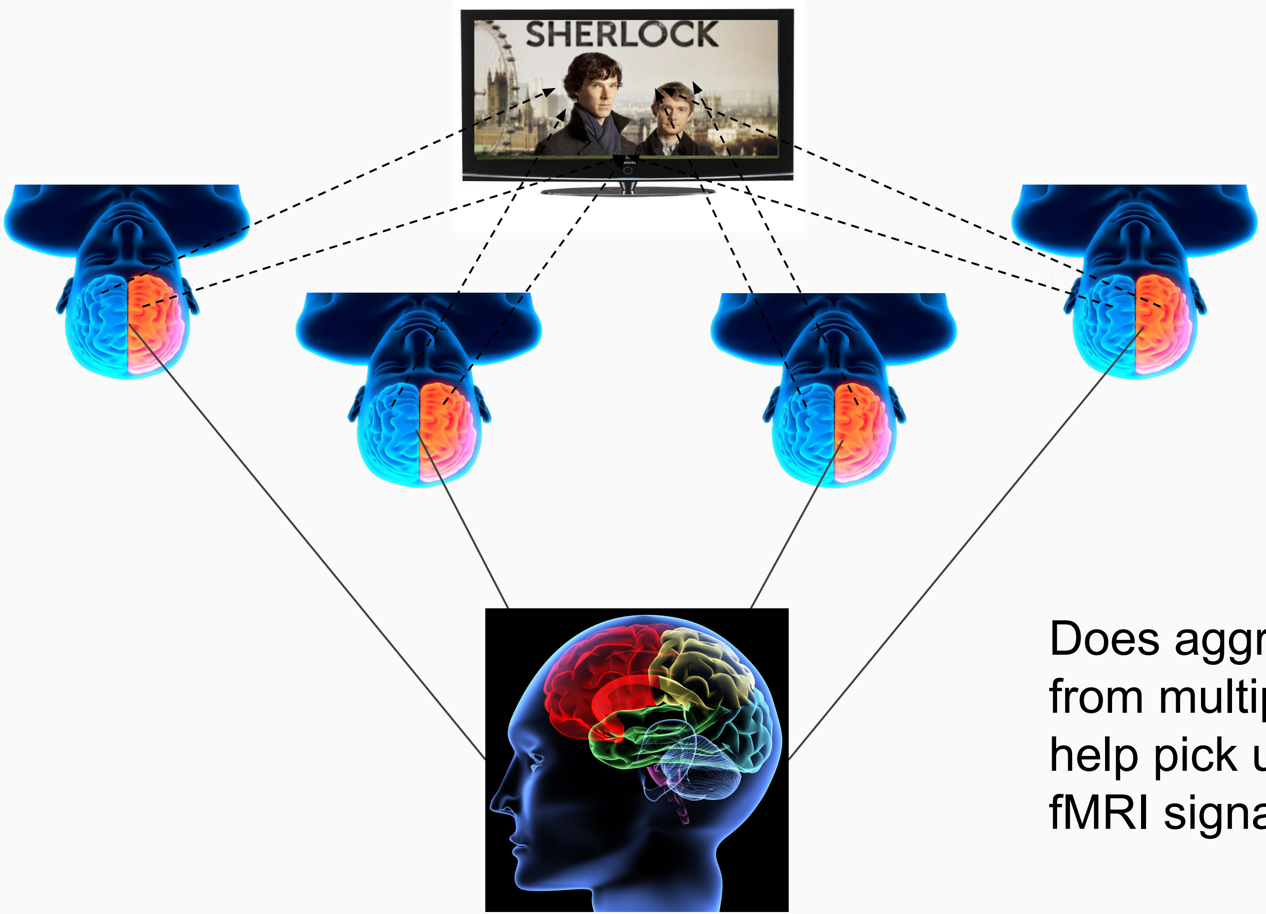
- Default Mode Network (DMN) standard area in literature
  - known to relate to narrative processing
  - DMN-A, -B (2000 voxels)
- Ventral/Dorsal Language (2000 voxels)
- Whole Brain (26000 voxels)
  - voxels with high inter-subject correlation
- Occipital Lobe (6000 voxels)

Leveraging Multiple Subject Views to Extract Better Semantics

Shared Movie Stimulus

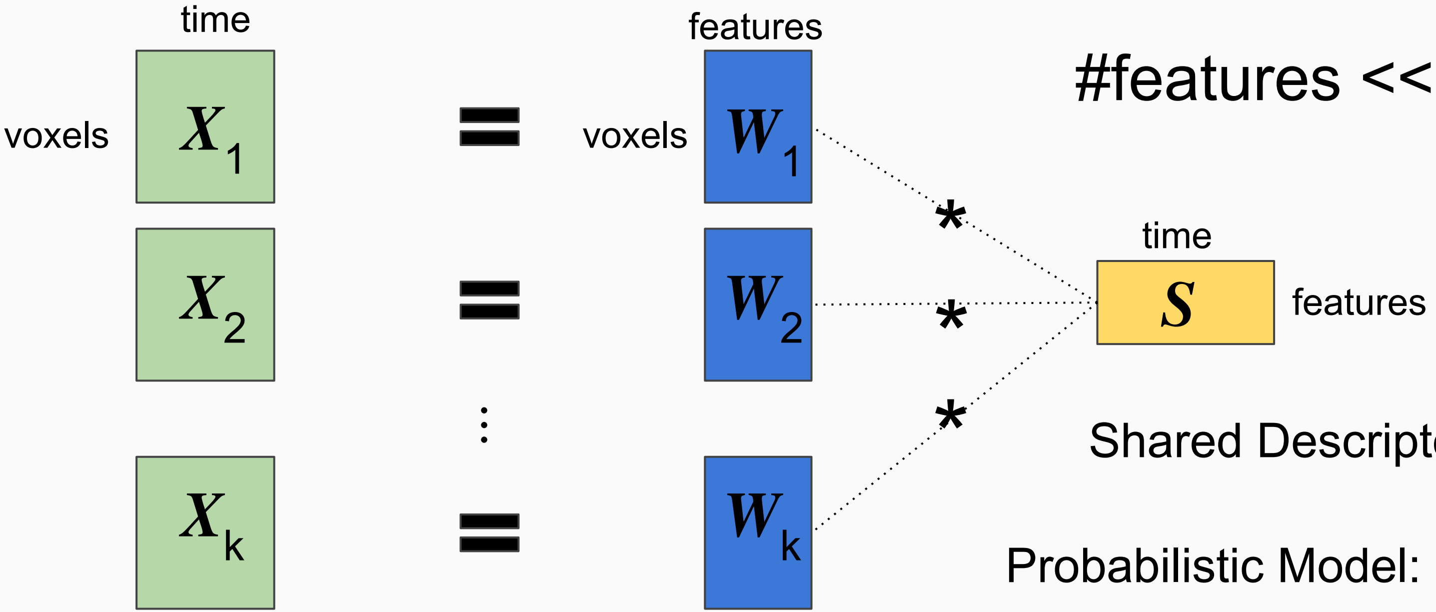
Multiple Subject Responses

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

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

Probabilistic Model:

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

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

Embedding Annotations with Weighted Sums of Word Vectors

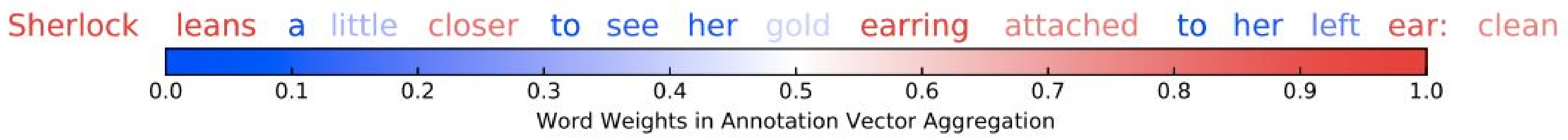


Fig. 3. Visualization of Semantic Annotation Vector Weightings: We display an example sentence from the Sherlock annotations, where we have colored important words red, and unimportant words blue. Brighter red means more important, and darker blue means less important.

Using Information (Temporal Dynamics) from the Past

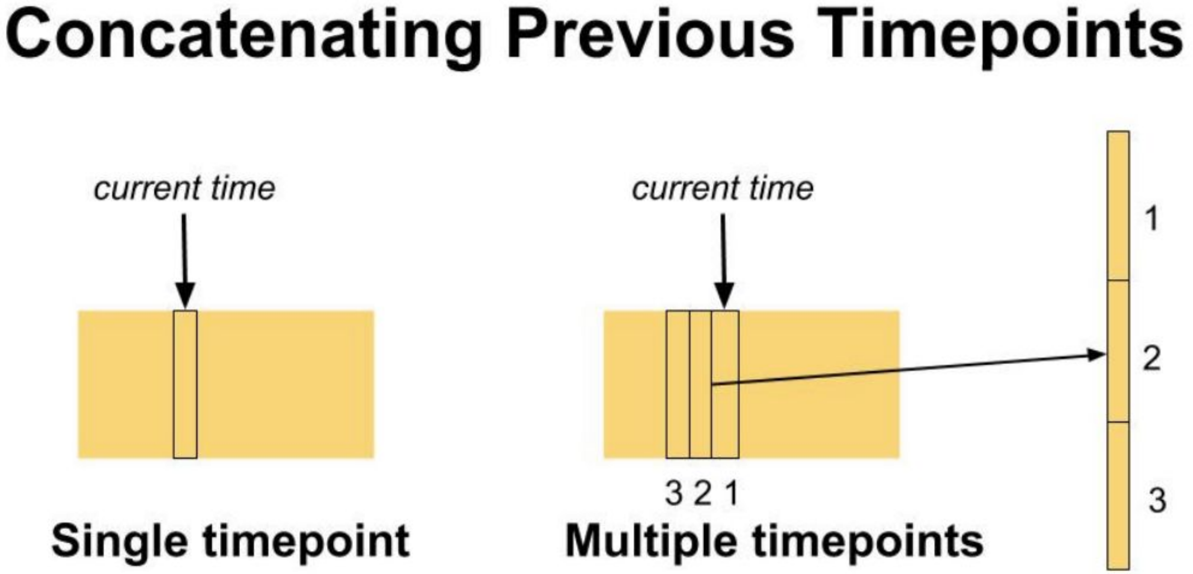


Fig. 4. Visualizing Concatenation: We visualize what the single timestep case looks like compared to a case where we use the previous two timesteps in our featurization as well. The latter case results in a more complicated model, since one of the dimensions of our linear map triples in size.

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)

Previous Time Step Model:

$$\hat{W}\hat{X} = Y \quad , \quad \hat{W} \in \mathbb{R}^{m \times n*(k+1)}$$

$$\hat{X} \in \mathbb{R}^{n*(k+1) \times T}$$

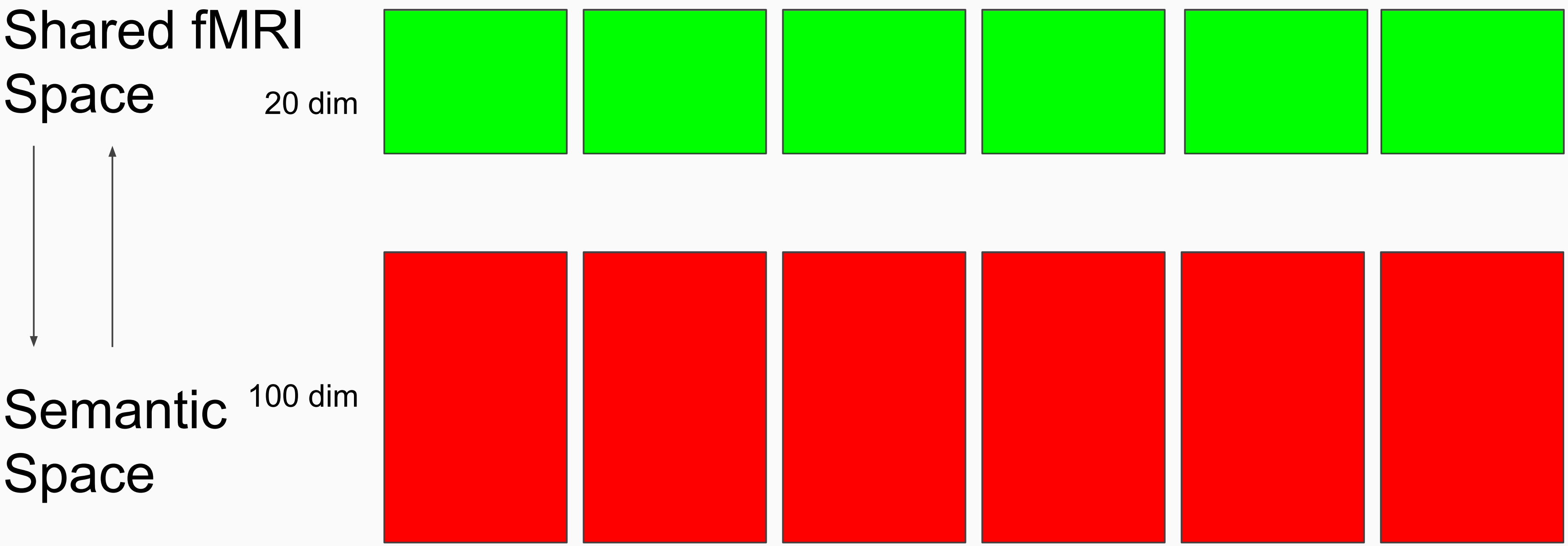
k is the number of previous timesteps used

Learning the Map:

- Procrustes ($W^T W = I$)
- Ridge Regression

Evaluation: Scene Classification/Ranking Experiments

25 test chunks from 1976 TRs



Results: Multiplicative Improvements with our Methods

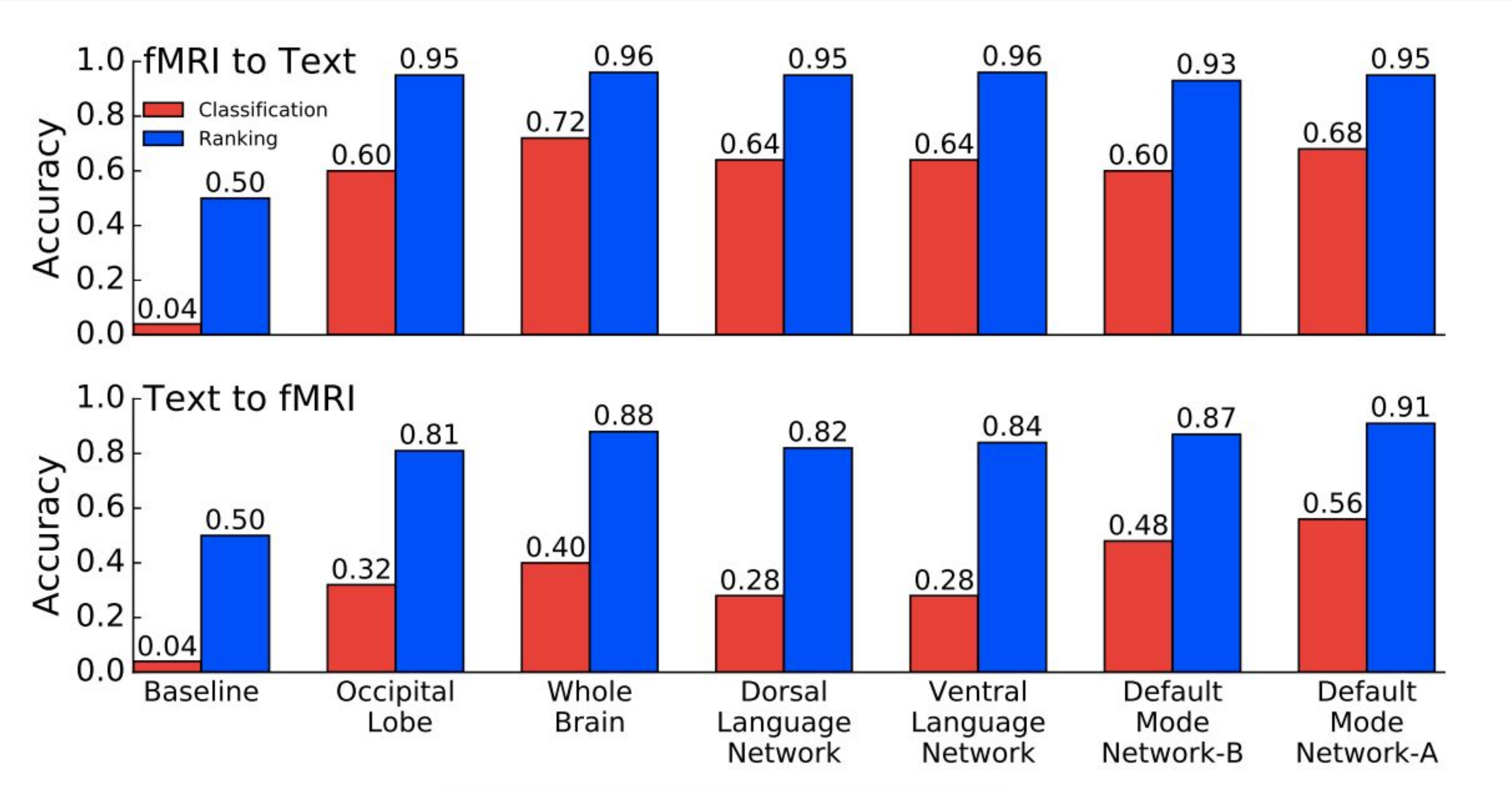
Mapping Between fMRI Responses and Semantic Representations

fMRI $\rightarrow$ Text	Maximum	Average
Previous Timesteps vs. None	5.3×	1.8×
Procrustes vs. Ridge	2.8×	1.3×
SRM/SRM-ICA vs. PCA	1.8×	1.3×
Weighted-SIF vs. Unweighted	1.6×	1.2×

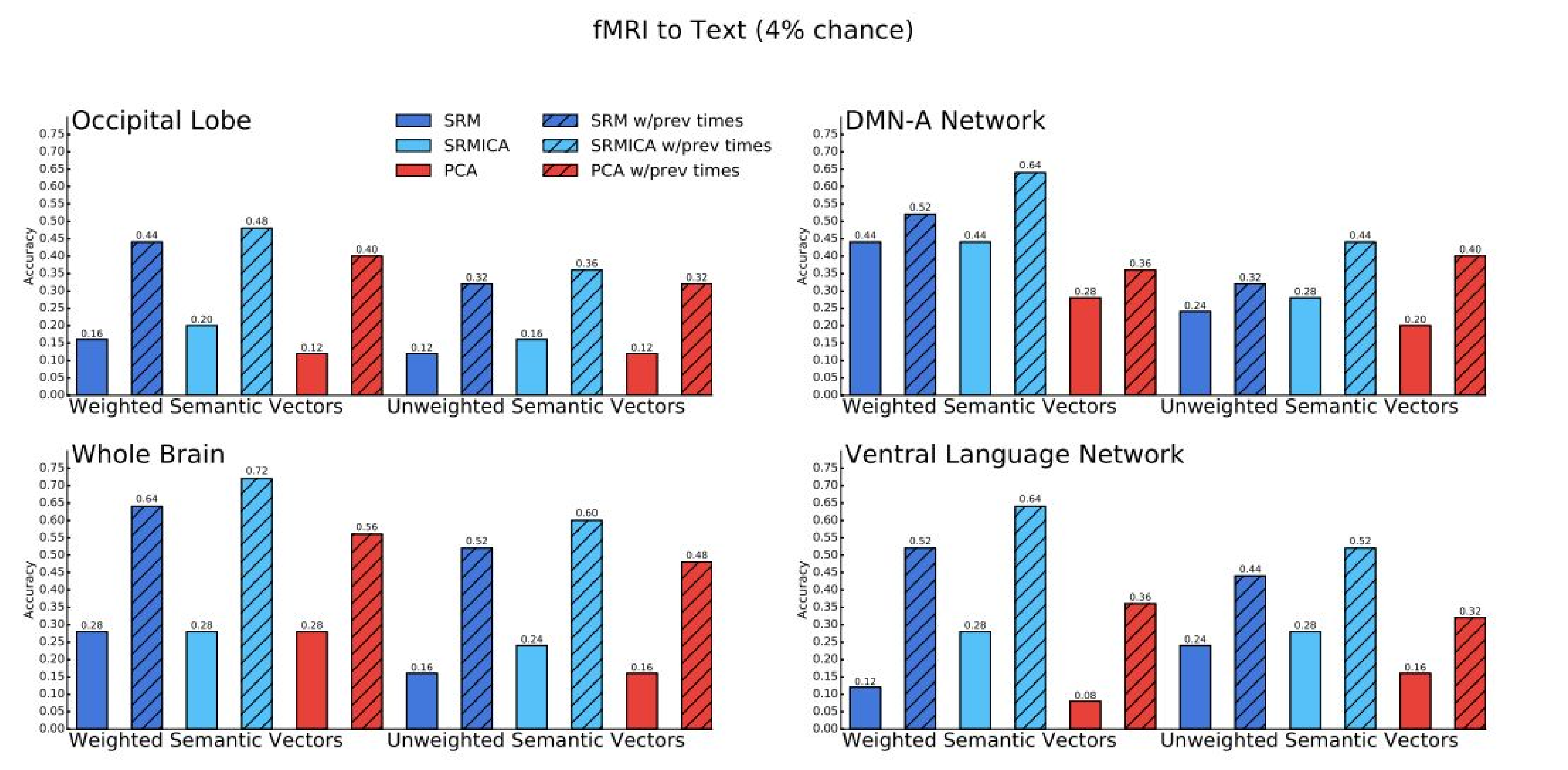
Text $\rightarrow$ fMRI	Maximum	Average
Previous Timesteps vs. None	2.5×	0.5×
Procrustes vs. Ridge	3.0×	0.8×
SRM/SRM-ICA vs. PCA	2.3×	1.2×
Weighted-SIF vs. Unweighted	1.8×	1.1×

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, including usage of previous time steps, SRM/SRM-ICA versus PCA, SIF-weighted annotation embeddings versus unweighted annotation embeddings, and Procrustes versus ridge regression for both fMRI $\rightarrow$ Text and Text $\rightarrow$ fMRI. When we use previous timesteps, we consider the results for using $5-8$ previous time steps. These numbers are all for the scene classification task. Note that the values from the maximum columns can be seen visually in Figures 6 and 7 respectively.

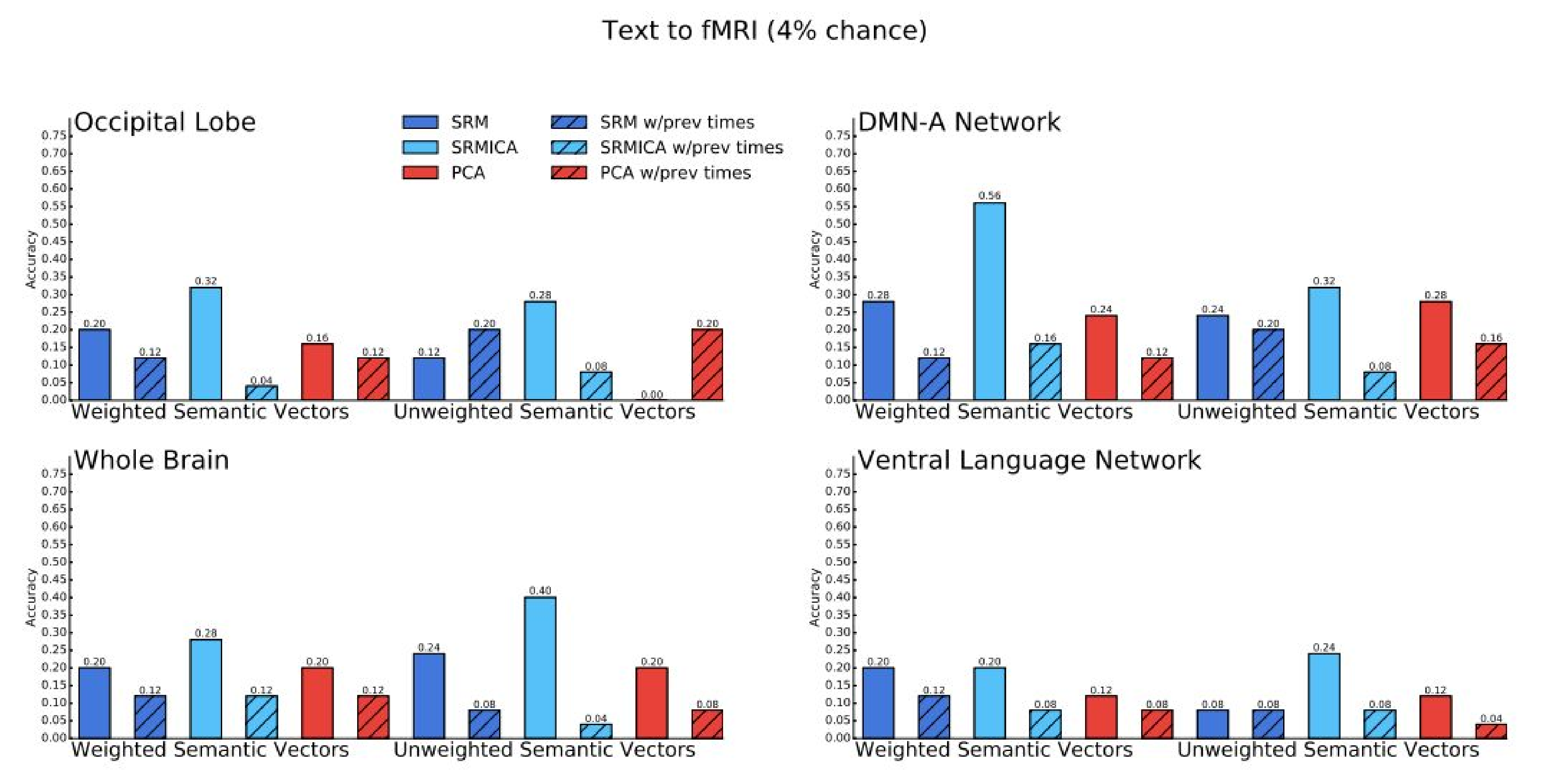
Results: Top-4% Classification and Average Rank



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)

Interpretable Methods for Using Previous Time Steps

Decay weights and Normalization: $\lambda = [\lambda_1, \cdots, \lambda_n]$, $Z_i = \sum_{j^*=t}^{t-k} e^{(t-j^*)\lambda_i}$

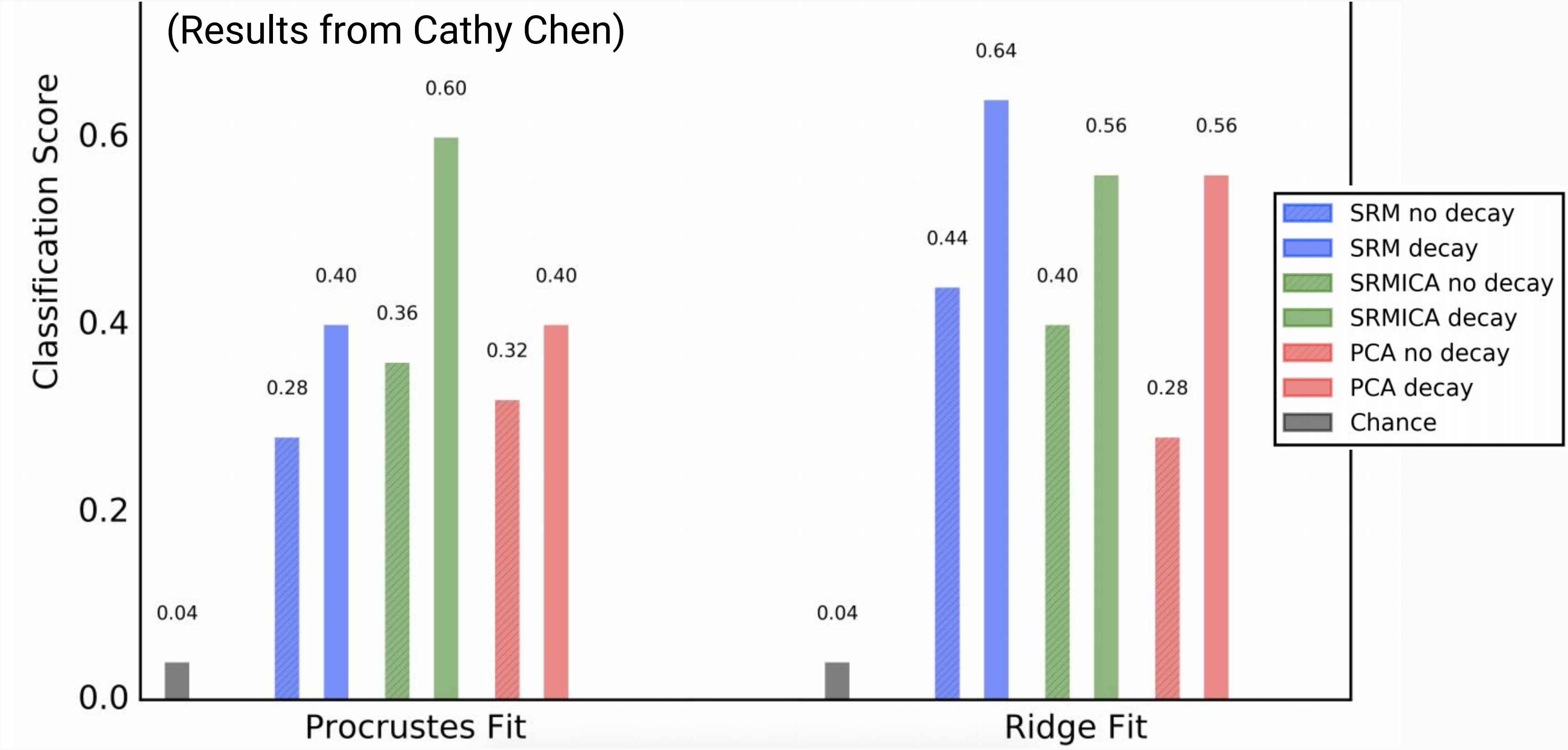
$$C_k = \begin{bmatrix} 1/Z_1 & 0 & \cdots & 0 & e^{\lambda_1}/Z_1 & 0 & \cdots & 0 & \cdots\cdots & e^{k\lambda_1}/Z_1 & 0 & \cdots & 0 \\ 0 & 1/Z_2 & & 0 & 0 & e^{\lambda_2}/Z_2 & & 0 & \cdots\cdots & 0 & e^{k\lambda_2}/Z_2 & & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & & \cdots\cdots & \vdots & & \ddots & \\ 0 & & & 1/Z_n & 0 & & & e^{\lambda_n}/Z_n & \cdots\cdots & 0 & & & e^{k\lambda_n}/Z_n \end{bmatrix}$$

Linear model: $W C_k \hat{X} = Y$

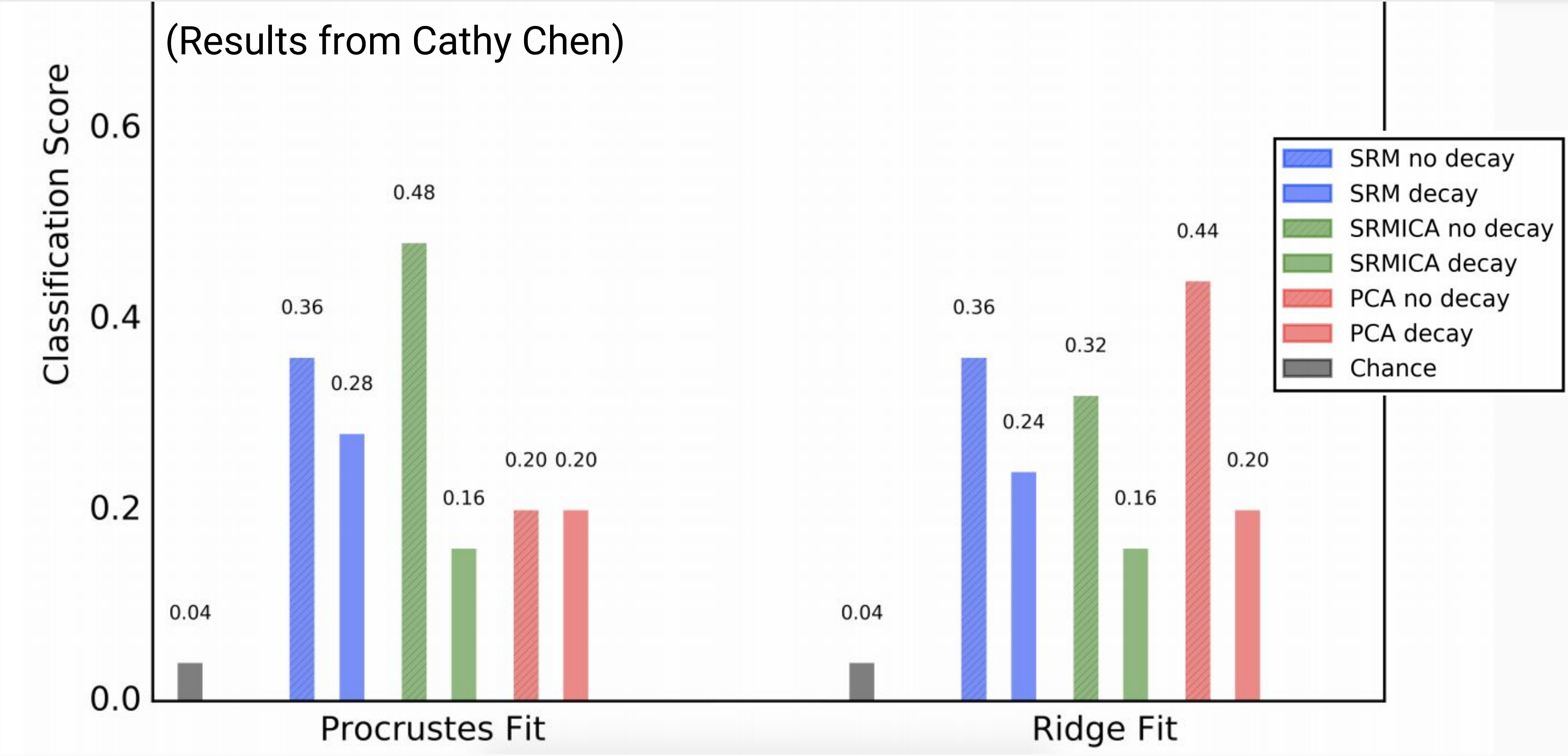
n = fMRI dimensions
m = text dimensions
k = prev. time steps

where $W \in \mathbb{R}^{m \times n}$, $C_k \in \mathbb{R}^{n \times n*(k+1)}$, $\hat{X} \in \mathbb{R}^{n*(k+1) \times T}$, and $Y \in \mathbb{R}^{m \times T}$

Comparison of Decay Weights, DMN-A Region fMRI → Text (4% Chance)



Comparison of Decay Weights, DMN-A Region Text → fMRI (4% Chance)



Ongoing and Future Work

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