Decoding fMRI to Text with Context

Kiran Vodrahalli*, Po-Hsuan Chen*, Chris Baldassano*, Janice Chen*, Esther Yong †,
Christopher Honey †, Peter J. Ramadge*, Kenneth A. Norman*, Sanjeev Arora*
Intel-PNI Meeting
July 22, 2016

* = Princeton, † = U. Toronto

Goals of this work

- Decode natural language descriptions of narrative stimuli from fMRI data (Sherlock dataset)
- Better methods for combining word vectors to create sentence/paragraph/etc. vectors (‘context’ vectors)
- Identify shifts in context in narrative; compare to psychological models

Interesting and Useful Discoveries

- The Shared Response Model (SRM, [Chen et al. 2015]) helps a lot for decoding text!
- Dictionary learning on word vectors → better semantic context vectors
- Orthogonal maps decode fMRI → text better than ridge regression

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.

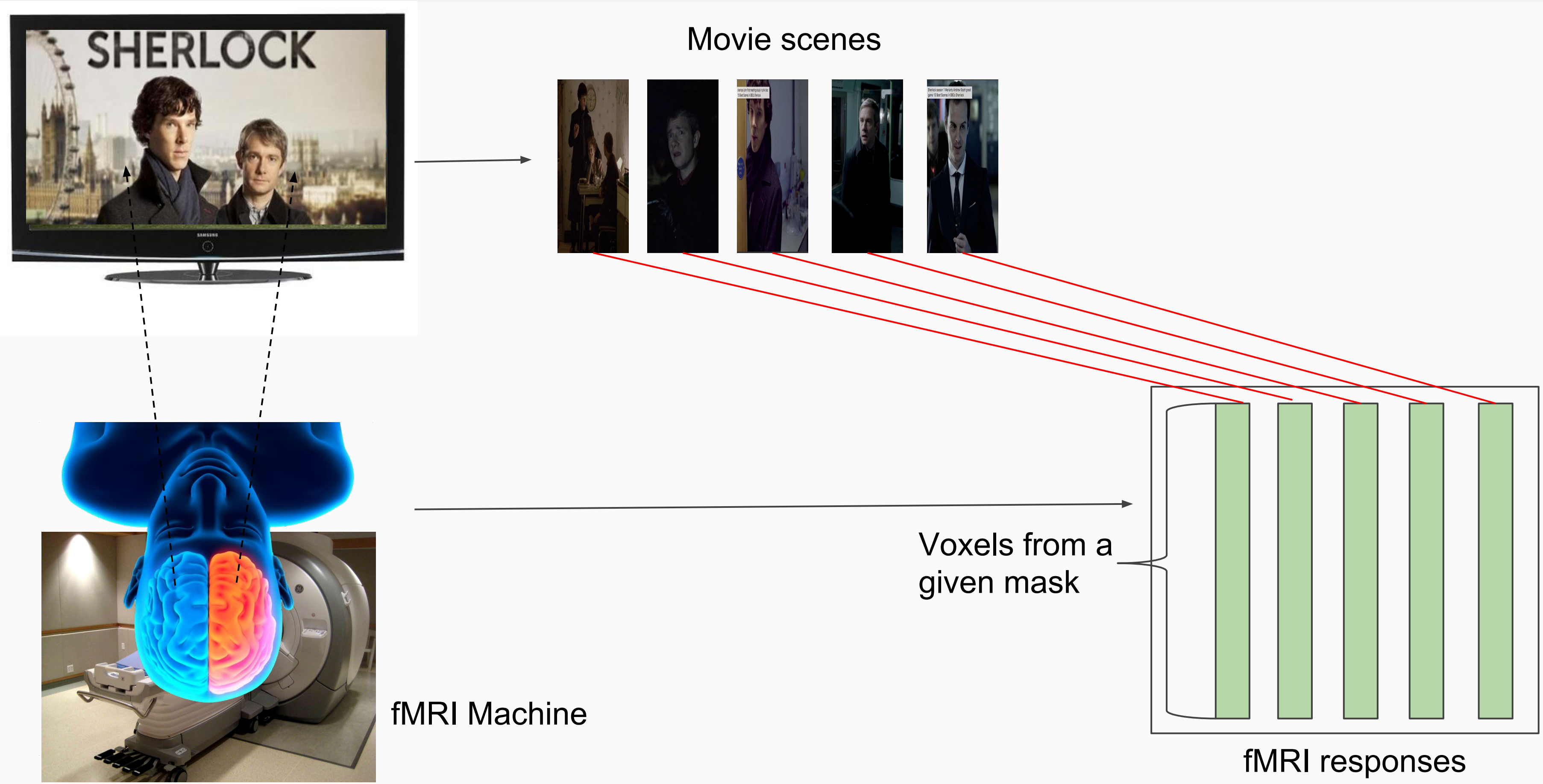
[Naselaris et al ‘11] tries to **reconstruct movie images** from fMRI signals measured while subjects watched movies.

[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**.

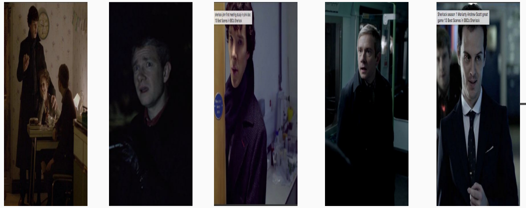
[Pereira et al ‘16] decodes fMRI responses to **word clouds and short sentences.**

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.

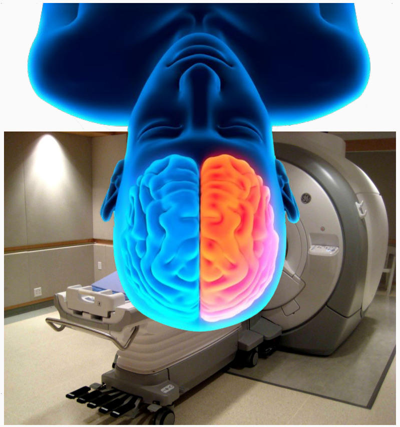
John is worried as Sherlock runs off.

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!"

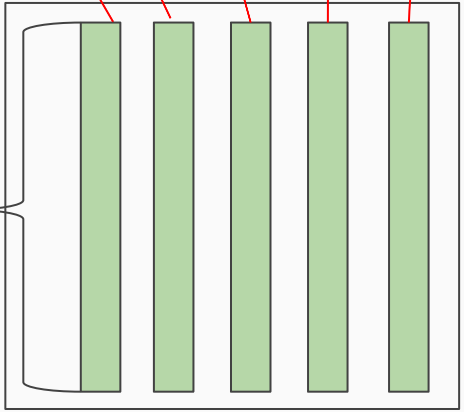
Moriarty arrives and says, "Hello Sherlock, John."

Each movie scene paired with text description from external party.



fMRI Machine

Voxels from a given mask



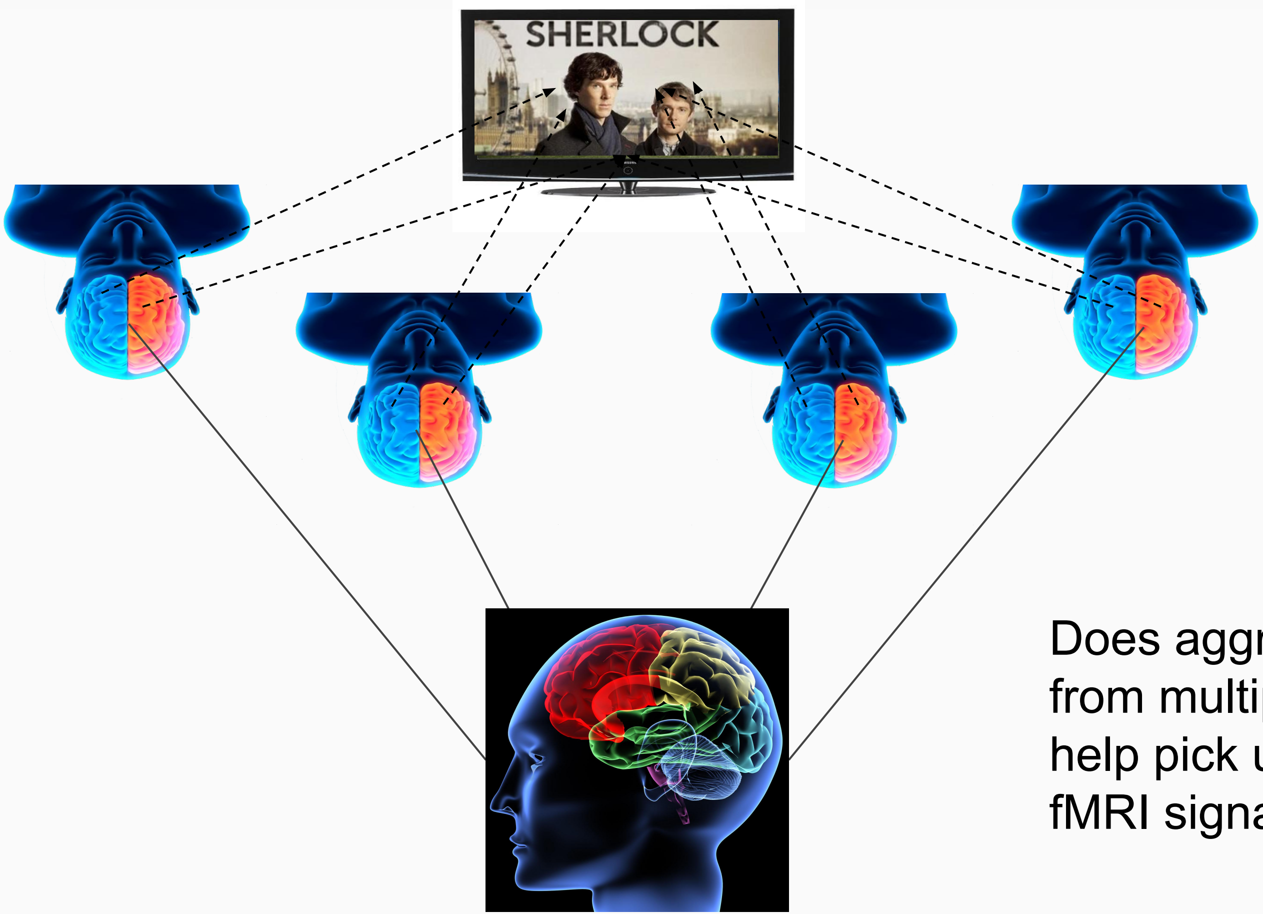
fMRI responses

Goal 2: Leverage Multiple Subject Views to Extract Better Semantics

Shared Movie Stimulus

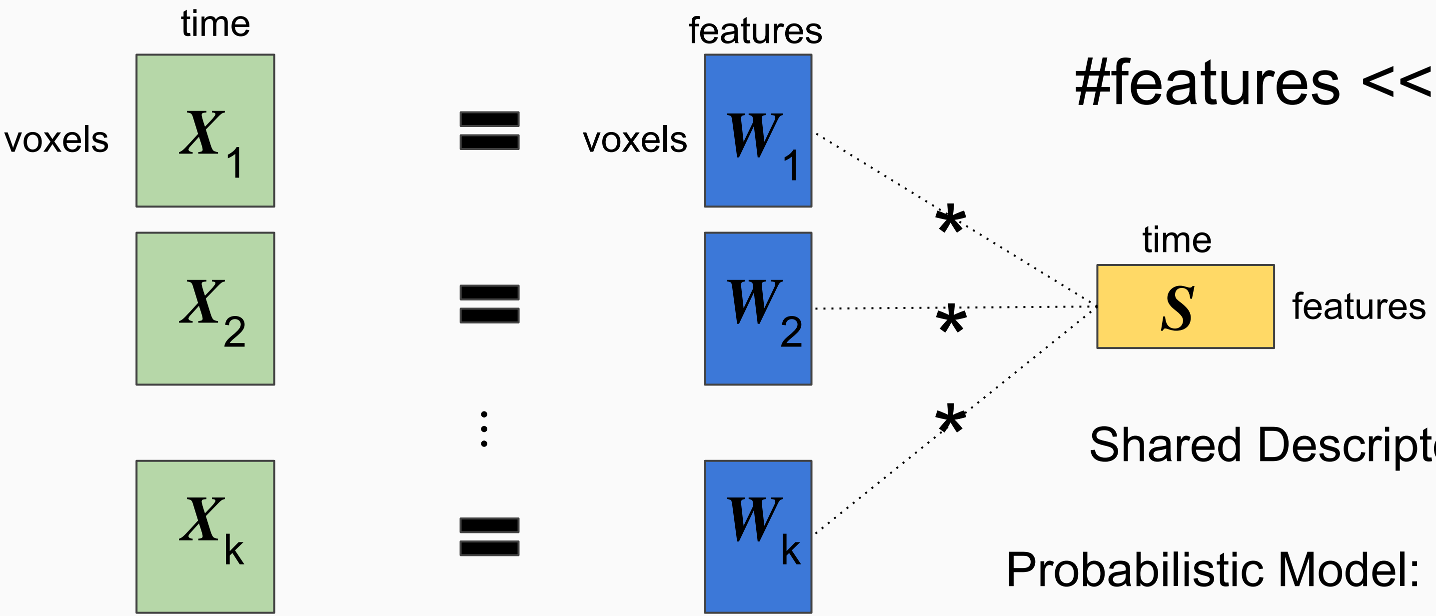
Multiple Subject Responses

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:

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

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

Semantic Vector and Dictionary Learning Background

- Large text corpus (Wikipedia) → map from words to vectors
  - Similar words are close by; linear algebraic relationships ([Mikolov et al '13], [Pennington et al '14], [Arora et al '15])
- Matrix Factorization approach [Arora et al '15]
- Dictionary learning (DL):
  - Given set of vectors, DL → set of building blocks (a basis)
  - Every vector ≅ linear combination of k building blocks
  - These building blocks are called **atoms**

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

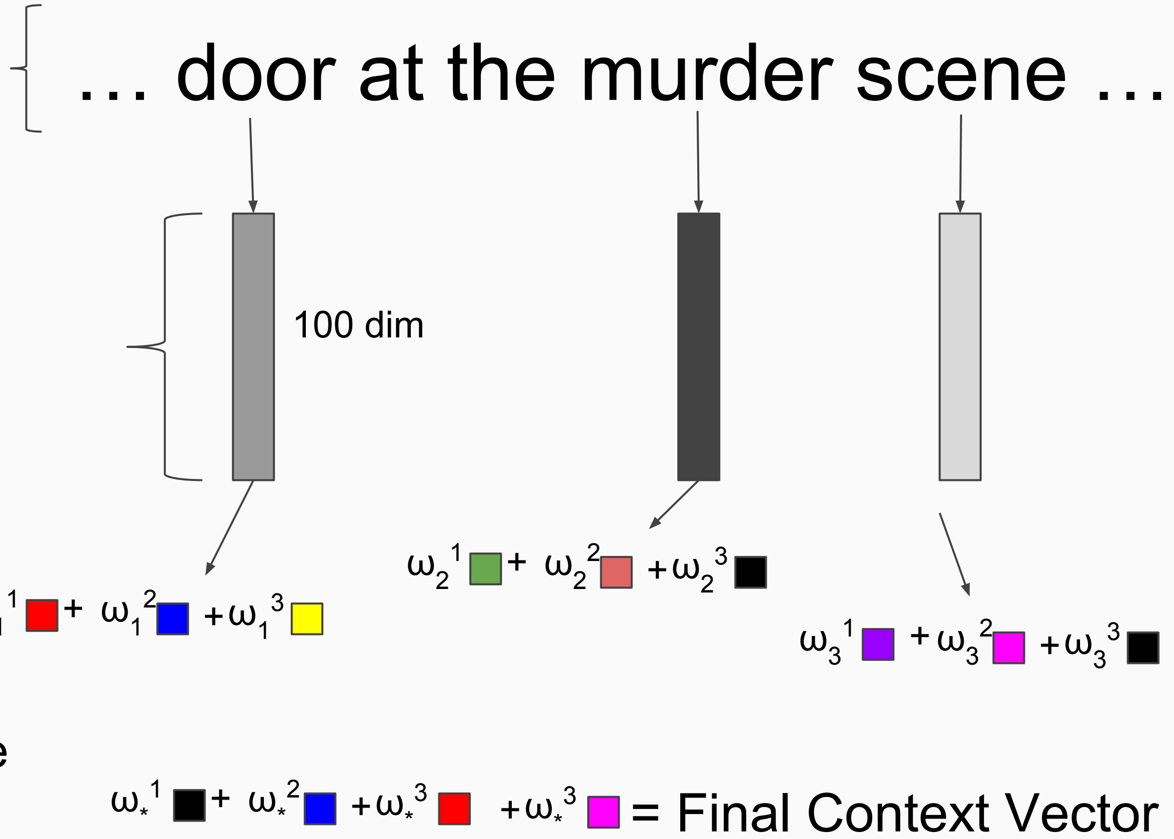
Annotation Text:

Word Vectors from Wikipedia:

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

Decomposition into Atoms:

Sort the atoms by their aggregate weights and pick the top 4:



Why use dictionaries?

- Think of atoms as topics in a topic model

Feet = $\boldsymbol{\alpha}$*{ankles, wrists, ...} + β*{inches, meters, …}

- The intuition is that we're essentially doing Word Sense Disambiguation

- [Arora et al '16] shows that word vectors are linear combinations of different senses - let's remove incorrect senses

Filtering Bad Atoms

Start off with 1550 atoms from Wikipedia corpus

End with 477 atoms by removing uninformative atoms

Currently, automating this process.

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.

Other ROIs: Auditory, Dorsal/Ventral Language areas, Occipital lobe, V1

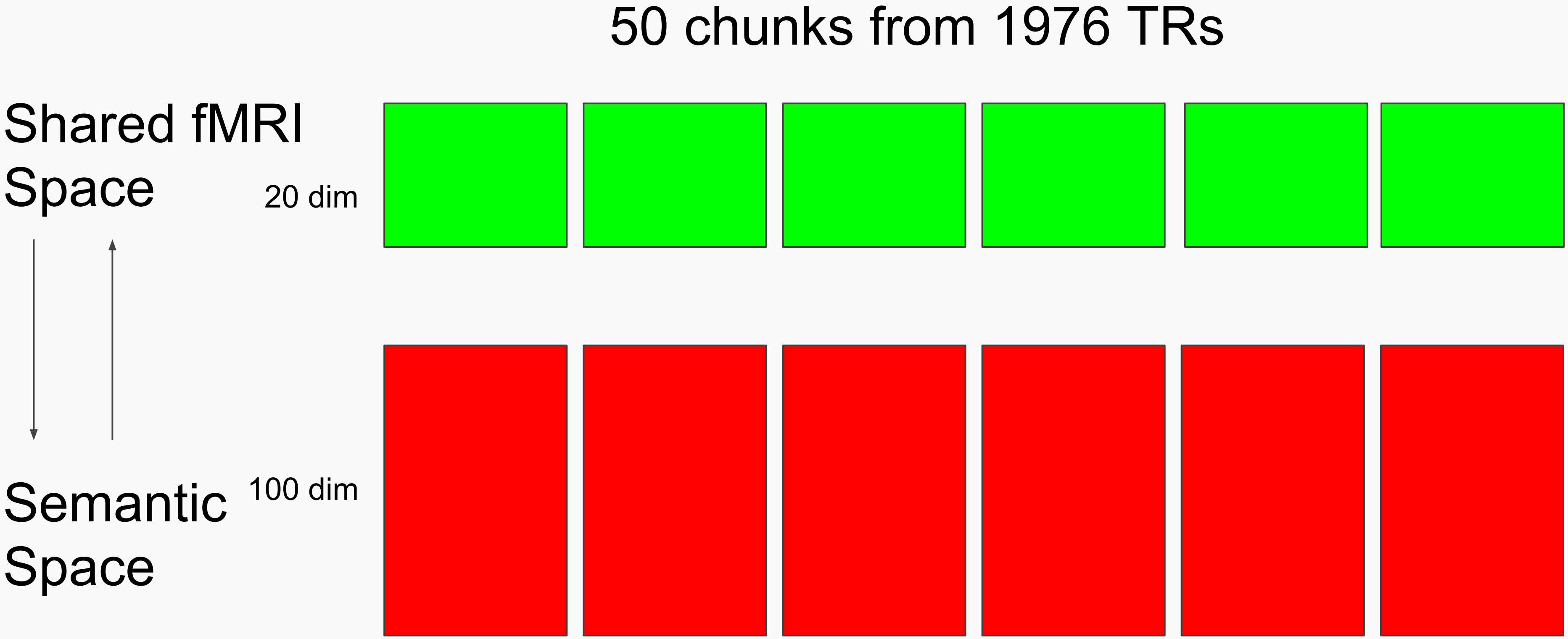
Dimensionality: We learn maps between the low-dimensional shared space (k = 20, 50, 100 dims) and semantic space (100 dim). Empirically, k = 20 was best and is justified by the approx. low-rank of the fMRI data for the DMN region.

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

2) Procrustes problem regularizes via orthogonality

Procrustes Problem: Minimize $\|X - WY\|$ such that W is a rotation matrix (X = fMRI, Y = text).

Scene Classification Experiment



Classification Results for DMN Region Using SRM

	S_fMRI → Text (Procrustes)	Text → S_fMRI (Ridge)
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%

Regularization Type Matters (Switch Ridge and Procrustes)

	S_fMRI → Text (Ridge)	Text → S_fMRI (Procrustes)
Binary Classification Leave 2 scenes out and match (chance 50%)	59% (< 70%)	71% (< 83%)
Scene Classification Train first ½, test second ½ (Top-5 rank: chance 20%)	34% (< 49%)	38% (< 50%)

Shared Response Model Improves Scene Classification

	A_fMRI → Text (Procrustes)	Text → A_fMRI (Ridge)
Scene Classification Train first ½, test second ½ (Top-5 rank: chance 20%)	28% (< 49%)	37% (< 50%)

Here, A_fMRI is the raw fMRI response averaged over all subjects.

Word Sense Disambiguation Improves Scene Classification

(Without performing word filtering)	S_fMRI → Text (Procrustes)	Text → S_fMRI (Ridge)
Scene Classification Train first ½, test second ½ (Top-5 rank: chance 20%)	24% (< 49%)	34% (< 50%)

Going Forward: Temporal Receptive Windows and Context

- Different regions of the brain operate on different time scales (DMN, Early Visual Cortex, etc.)

- Different stimuli (e.g. movie scenes) are relevant to current activity at different time scales

- If a particular brain area's state is informed by 10 TRs, we should **use all 10 TRs worth of matched text information** - not just a single TR's worth.