# Decoding topic vectors during memory encoding and retrieval

Jeremy R. Manning, David M. Blei, and Kenneth A. Norman

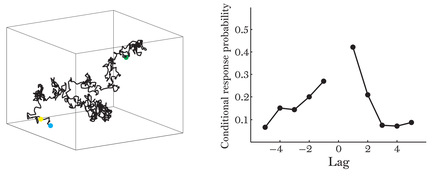
Princeton University

This research was supported by the NSF/NIH Collaborative Research in Computational Neuroscience Program, grant number NSF IIS-10009542

## Overview & experimental methods

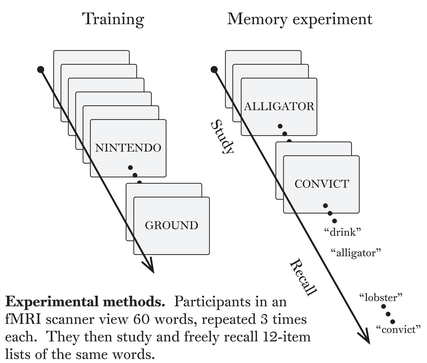
Context-based theories of memory posit that items on a studied list become associated with the mental contexts in which they are experienced.

We present a framework for tracking the neural correlates of individual items[6] and the contexts in which they are experienced,[5] during individual study and recall events.





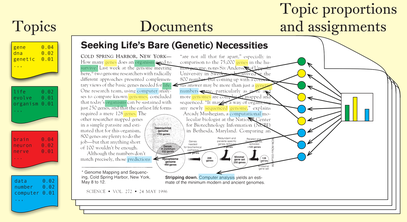

**Context-based theories of memory.** Context drifts gradually over time and becomes associated with each experienced event, giving rise to the contiguity effect in free recall.




**Experimental methods.** Participants in an fMRI scanner view 60 words, repeated 3 times each. They then study and freely recall 12-item lists of the same words.

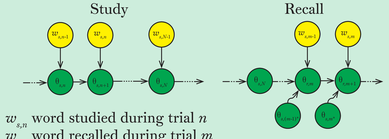

**Towards a unified model of corpora and cognition.** Our approach attempts to infer the evolving state of mental context using text, behavioral, and neural data.

## Model of word meanings




**Latent Dirichlet Allocation.** Topic vectors are derived by analyzing a collection of documents.[1]

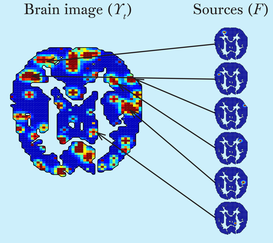
## Model of mental context




$w_{s,n}$ word studied during trial $n$
$w_{r,m}$ word recalled during trial $m$
$\theta_{s,n}$ context (topic) vector during study of word $n$
$\theta_{r,m}$ context (topic) vector during recall of word $m$
$\theta_{s,m'}$ context (topic) vector associated with the $m^{\text{th}}$ recalled word during study

**Temporal Context Model.** Mental context vectors are weighted averages of topic vectors for studied and recalled words.[3,8]

## Model of neural activity




$$\Upsilon_t = w_t F$$

**Topographic Latent Source Analysis.** Neural patterns are represented as linear combinations of spherical sources.[2,4]

## Decoding results

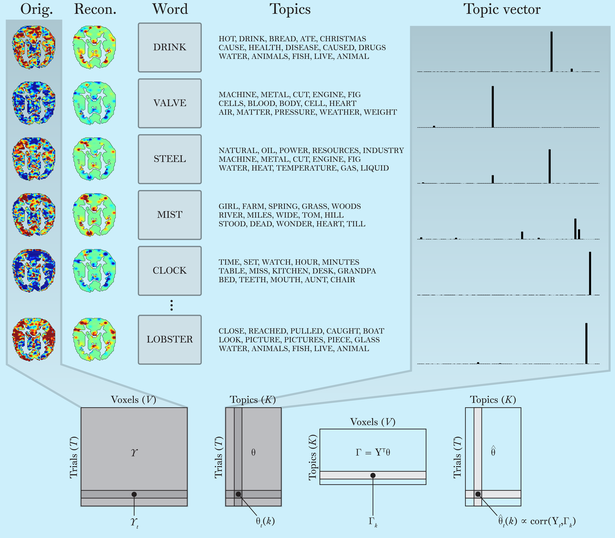
We collect brain images during each word presentation.

We use a topic model to obtain topic vectors for each presented word.

We infer the neural representation of each topic by computing a weighted average of the observed brain images. (Similar approach to 7.)

We interpret new neural patterns as topic vectors by computing the correlation between those patterns and the inferred representation of each topic.

Unfamiliar topics (biased towards less familiar words) are consistantly decoded more reliably than famliar topics.



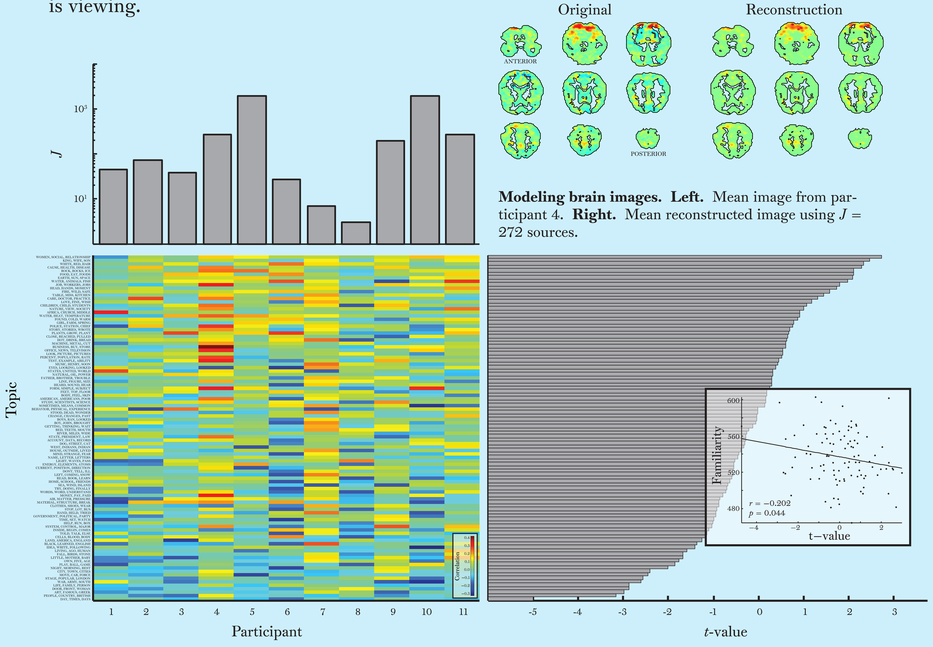

Brain images are modeled as linear combinations of spherical sources.

We use the observed brain images to infer the center locations and widths of each source.

We hold the center locations and widths fixed across images and fit per-image weights.

The vector of $J$ weights provides a low dimensional representation of the original image.

The model can learn the neural representations of words or topics by incorporating weights that depend on which word or topic the participant is viewing.




**Modeling brain images. Left.** Mean image from participant 4. **Right.** Mean reconstructed image using $J$ = 272 sources.

**Decoding topic vectors from brain images. Left.** Decoding methods. We train and test the decoder using 10-fold cross validation. **Right.** Decoding results. The heatmap displays, for each participant and topic, the correlations between the decoded topic activations across trials (inferred using the brain images) and "true" topic vectors (inferred using a topic model fit to the TASA corpus). The bar plot on the right shows that some topics tend to be decoded well across participants; the inset shows that decodability varies with topic familiarity (MRC Psychological Database). The upper bar plot shows the number of sources that yielded the highest overall decoding accuracy for each participant.

## Bibliography

[1] Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993 - 1022.
[2] Gershman SJ, Blei DM, Pereira F, Norman KA (2011) A topographic latent source model for fMRI data. *NeuroImage*, 57: 89 - 100.
[3] Howard MW, Kahana MJ (2002) A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46: 269 - 299.
[4] Manning JR, Norman KA, Blei DM (2012) Factor topographic latent source analysis: Factor analysis for brain images. *Submitted*.
[5] Manning JR, Polyn SM, Baltuch G, Litt B, Kahana MJ (2011) Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, USA*, 108: 12893 - 12897.
[6] Manning JR, Sperling MR, Sharan A, Rosenberg EA, Kahana MJ (2012) Spontaneously reactivated patterns in frontal and temporal lobe predict semantic clustering during memory search. *The Journal of Neuroscience*, 32: 8871 - 8878.
[7] Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA (2008) Predicting human brain activity associated with the meanings of nouns. *Science*, 320: 1191 - 1195.
[8] Socher R, Gershman SJ, Perotte AJ, Sederberg PB, Blei DM, Norman KA (2009) A Bayesian analysis of dynamics in free recall. *Advances in Neural Information Processing Systems*, 22: 1714 - 1722.

For reprints contact manning3@princeton.edu