Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

November 27, 2017

Today:

Representation learning:
- PCA
- ICA
- CCA
- Matrix factorization
- Neural networks

Recommended reading:

- “A Tutorial on PCA,’ J. Schlens
- Wall et al., 2003

Representation Learning Overview

- Principle Components Analysis
  - Singular Value Decomposition
  - application to face image compression
- Matrix Factorization
  - relationship to linear neural networks
- Canonical Correlation Analysis
  - analysis of brain image data across human subjects
- Independent Components Analysis
- Latent Dirichlet Allocation
  - analysis of email interactions

Learning Lower Dimensional Representations

- Supervised learning of lower dimension representation
  - Hidden layers in Neural Networks
  - Fisher linear discriminant
- Unsupervised learning of lower dimension representation
  - Principle Components Analysis (PCA)
  - Matrix factorization
  - Independent components analysis (ICA)
  - Canonical correlation analysis (CCA)
  - Latent Dirichlet Allocation (LDA)

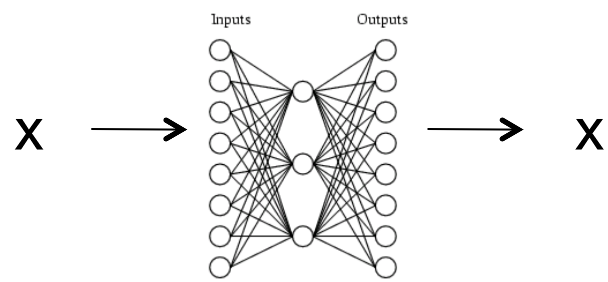
Principle Components Analysis

Principle Components Analysis

- Idea:
  - Given data points in d-dimensional space, project into lower dimensional space while preserving as much information as possible
    - E.g., find best planar approximation to 3D data
    - E.g., find best planar approximation to 10^4 D data
  - In particular, choose an orthogonal projection that <u>minimizes the squared error in reconstructing original data</u>

Principle Components Analysis

- Like auto-encoding neural networks, learn re-representation of input data that can best reconstruct it

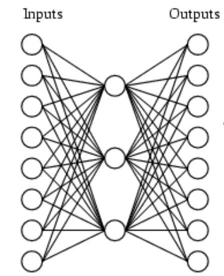


PCA:

- learned encoding is *linear* function of inputs
- No local minimum problems when training!
- Given d-dimensional data X, learns d-dimensional representation, where
  – the dimensions are orthogonal
  – top k dimensions are the k-dimensional linear re-representation that minimizes reconstruction error (sum of squared errors)

PCA Example

$$\text{face}_i = \Sigma_k \, c_{ik} \, \text{eigenface}_k$$

face →  → face

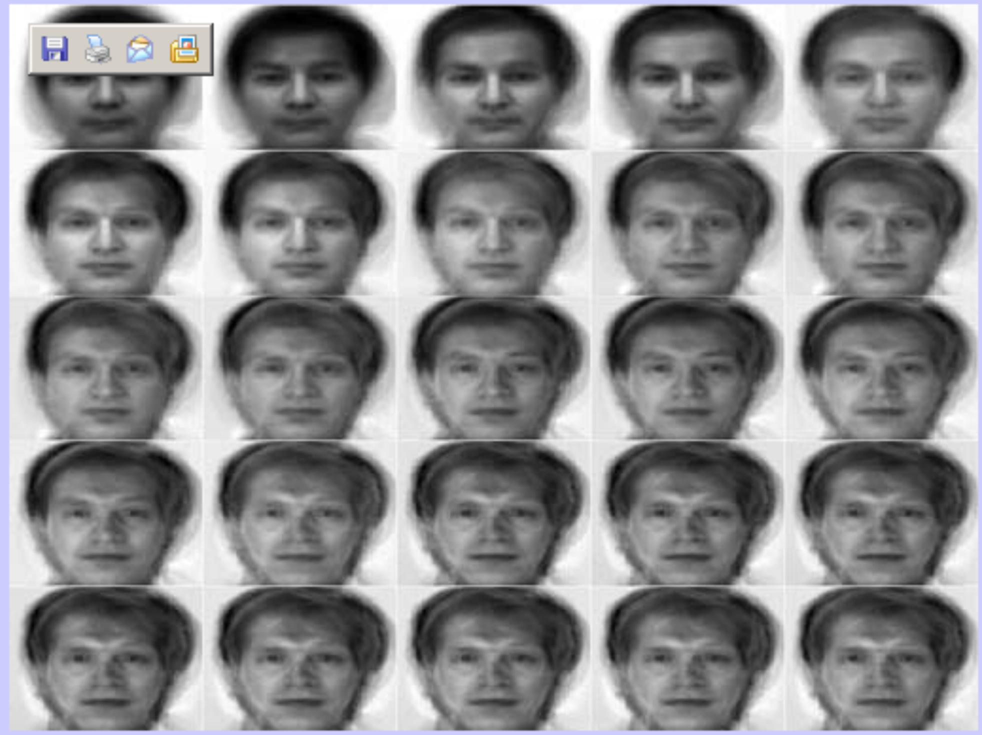
faces

eigenfaces

Thanks to Christopher DeCoro
see http://www.cs.princeton.edu/~cdecoro/eigenfaces/

Reconstructing a face from the first N components (eigenfaces)

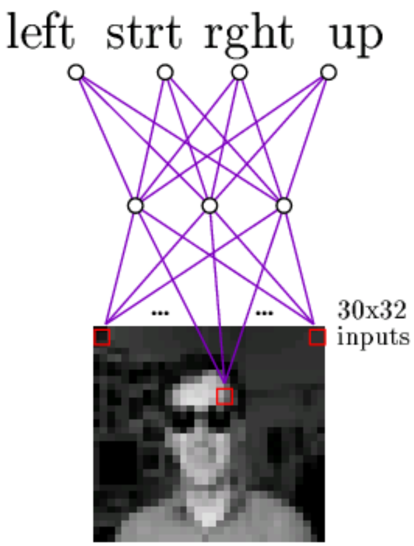
Adding 1 additional PCA component at each step

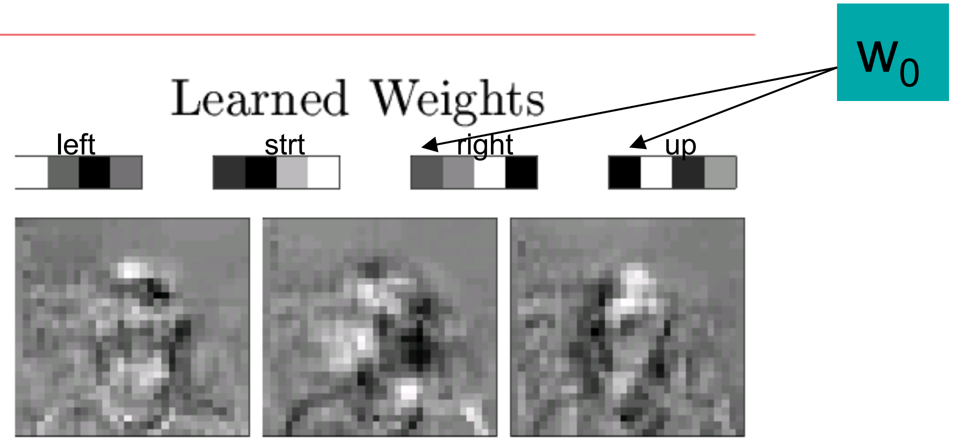


In this next image, we show a similar picture, but with each additional face representing an additional 8 principle components. You can see that it takes a rather large number of images before the picture looks totally correct.

Adding 8 additional PCA components at each step

Learned Hidden Unit Weights





Typical input images

http://www.cs.cmu.edu/~tom/faces.html

PCA: Find Projections to Minimize Reconstruction Error

Assume data is set of d-dimensional vectors, where nth vector is

$$\mathbf{x}^n = \langle x_1^n \ldots x_d^n \rangle$$

We can represent these in terms of any d orthogonal vectors $u_1 \ldots u_d$

$$\mathbf{x}^n = \sum_{i=1}^{d} z_i^n \mathbf{u}_i; \quad \mathbf{u}_i^T \mathbf{u}_j = \delta_{ij}$$

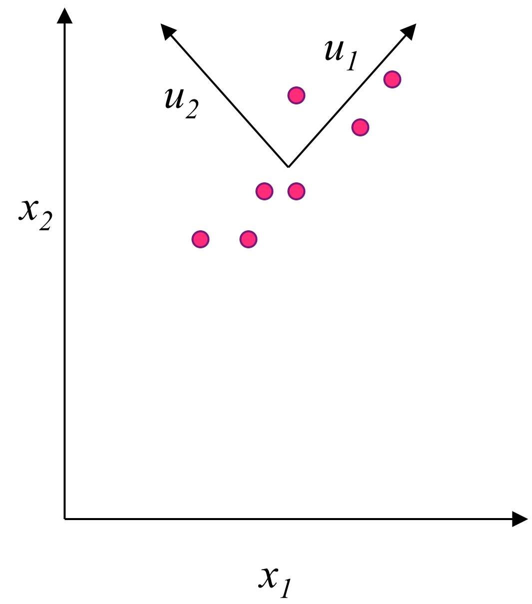
PCA: given M<d. Find $\langle \mathbf{u}_1 \ldots \mathbf{u}_M \rangle$

that minimizes $E_M \equiv \sum_{n=1}^{N} ||\mathbf{x}^n - \hat{\mathbf{x}}^n||^2$

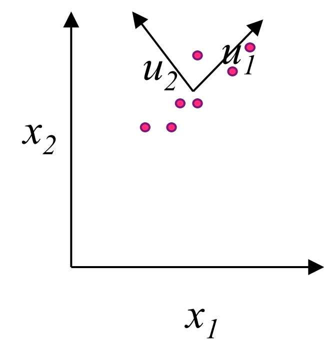
where $\hat{\mathbf{x}}^n = \bar{\mathbf{x}} + \sum_{i=1}^{M} z_i^n \mathbf{u}_i$

Mean

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^n$$



PCA



PCA: given M<d. Find $\langle \mathbf{u}_1 \ldots \mathbf{u}_M \rangle$

that minimizes $E_M \equiv \sum_{n=1}^{N} ||\mathbf{x}^n - \hat{\mathbf{x}}^n||^2$

where $\hat{\mathbf{x}}^n = \bar{\mathbf{x}} + \sum_{i=1}^{M} z_i^n \mathbf{u}_i$

Note we get zero error if M=d, so all error is due to missing components.

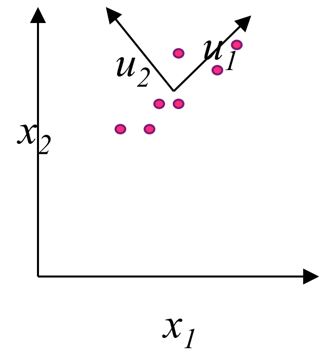
Therefore,
$$E_M = \sum_{i=M+1}^{d} \sum_{n=1}^{N} [\mathbf{u}_i^T (\mathbf{x}^n - \bar{\mathbf{x}})]^2$$
$$= \sum_{i=M+1}^{d} \mathbf{u}_i^T \Sigma \ \mathbf{u}_i$$

This minimized when $\boldsymbol{u}_i$ is eigenvector of Σ, the covariance matrix of X. i.e., minimized when: $\Sigma \mathbf{u}_i = \lambda_i \mathbf{u}_i$

Covariance matrix: $\Sigma = \sum (\mathbf{x}^n - \bar{\mathbf{x}})(\mathbf{x}^n - \bar{\mathbf{x}})^T$

$$\Sigma_{ij} = \sum_{n=1}^{N} (x_i^n - \bar{x}_i)(x_j^n - \bar{x}_j)$$

PCA



Minimize $E_M = \sum_{i=M+1}^{d} \mathbf{u}_i^T \mathbf{\Sigma}\ \mathbf{u}_i$

$$\rightarrow \quad \mathbf{\Sigma u}_i = \lambda_i \mathbf{u}_i$$

↖Eigenvector of Σ

↖ Eigenvalue (scalar)

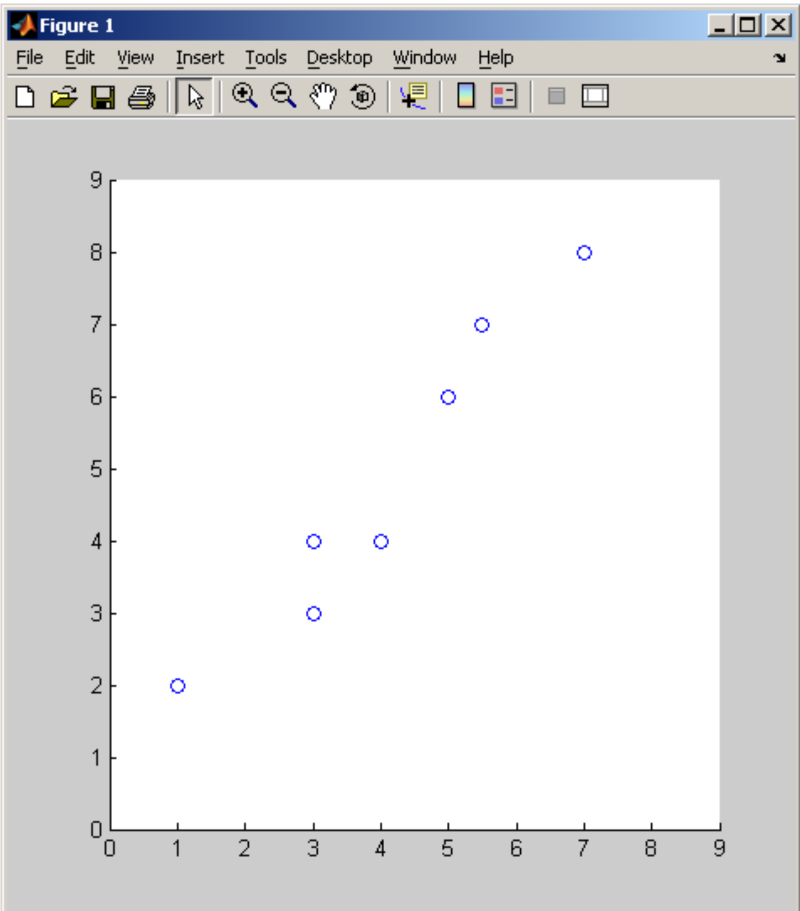
$$\rightarrow \quad E_M = \sum_{i=M+1}^{d} \lambda_i$$

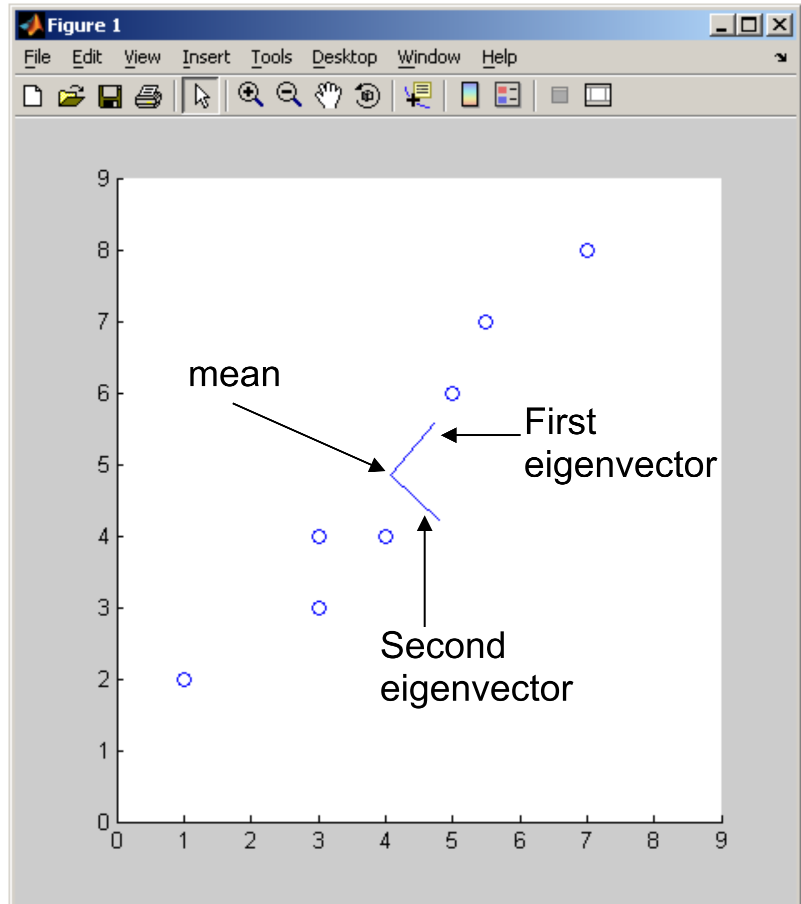
PCA algorithm 1:

1. X ← Create N x d data matrix, with one row vector x^n per data point
2. X ← subtract mean $\bar{x}$ from each row vector x^n in X
3. Σ ← covariance matrix of X
4. Find eigenvectors and eigenvalues of Σ
5. PC's ← the M eigenvectors with largest eigenvalues

PCA Example

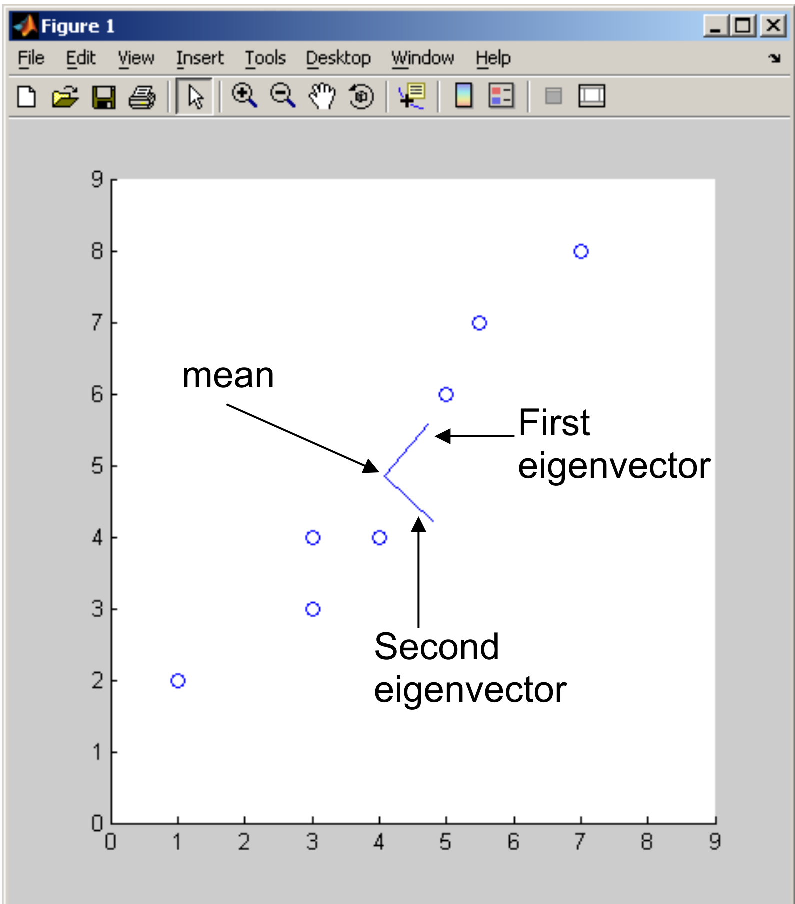
$$\widehat{\mathbf{x}}^n = \bar{\mathbf{x}} + \sum_{i=1}^{M} z_i^n \mathbf{u}_i$$



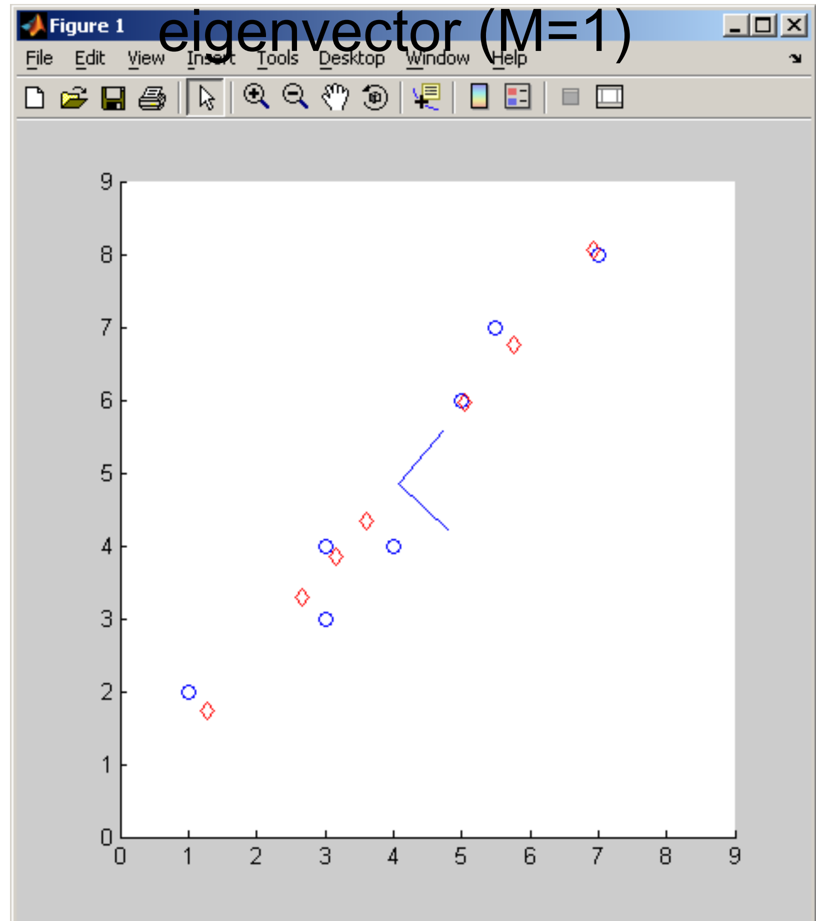


PCA Example

$$\hat{\mathbf{x}}^n = \bar{\mathbf{x}} + \sum_{i=1}^{M} z_i^n \mathbf{u}_i$$



Reconstructed data
using only first
eigenvector (M=1)



Very Nice When Initial Dimension Not Too Big

What if very large dimensional data?

- e.g., Images (d ¸ 10^4)

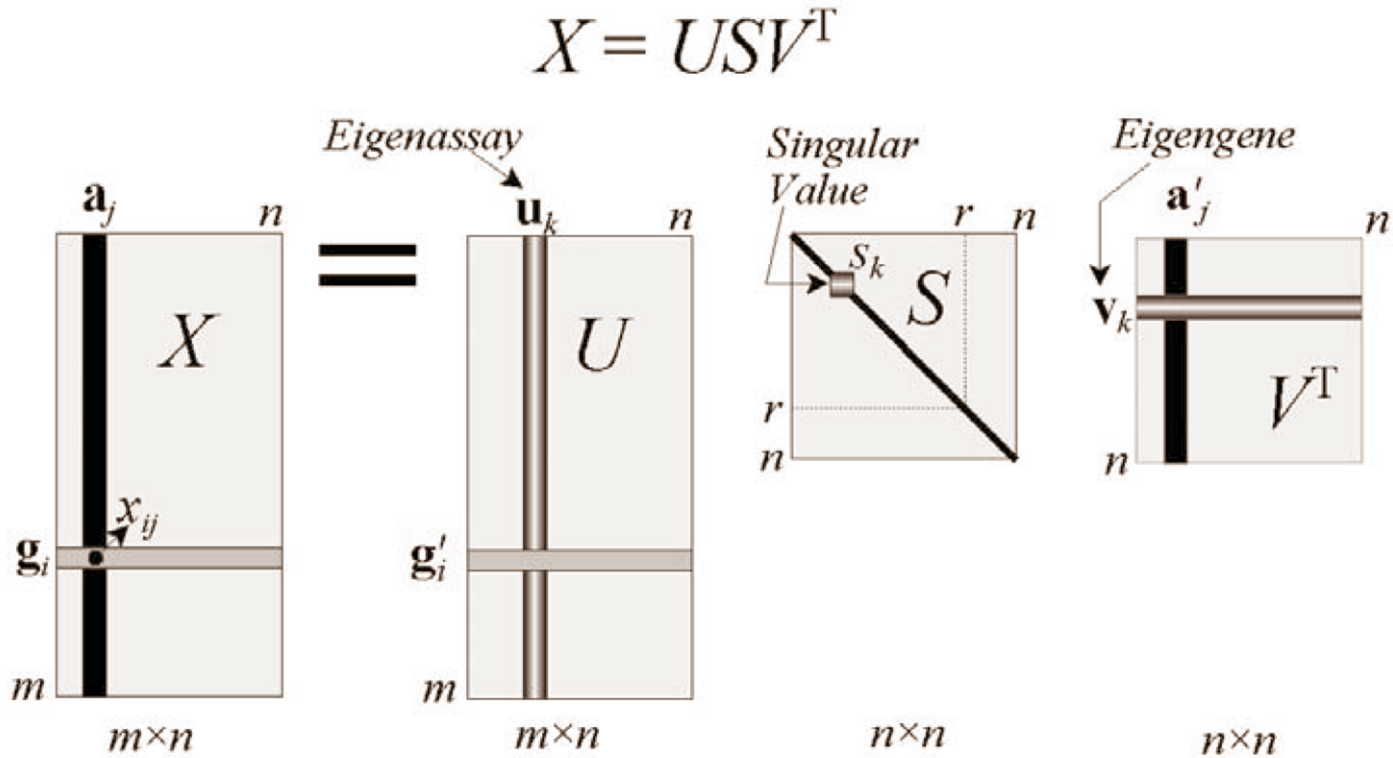
Problem:

- Covariance matrix Σ is size (d x d)
- $d=10^4 \rightarrow |\Sigma| = 10^8$

Singular Value Decomposition (SVD) to the rescue!

- pretty efficient algs available, including Matlab SVD
- some implementations find just top N eigenvectors

SVD



Data X, one row per data point

US gives coordinates of rows of X in the space of principle components

S is diagonal, $S_k > S_{k+1}$, S_k^2 is kth largest eigenvalue

Rows of V^T are unit length eigenvectors of X^TX

If cols of X have zero mean, then $X^TX = c\ \Sigma$ and eigenvects are the Principle Components

[from Wall et al., 2003]

Singular Value Decomposition

To generate principle components:

- Subtract mean $\bar{\mathrm{x}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}^n$ from each data point, to create zero-centered data
- Create matrix X with one row vector per (zero centered) data point
- Solve SVD: $X = USV^T$
- Output Principle components: columns of V (= rows of V^T)
  - Eigenvectors in V are sorted from largest to smallest eigenvalues
  - S is diagonal, with s_k^2 giving eigenvalue for kth eigenvector

Singular Value Decomposition

To project a point (column vector x) into PC coordinates:

$V^T x$

If x_i is ith row of data matrix X, then

- (ith row of US) $= V^T x_i^T$
- $(US)^T = V^T X^T$

To project a column vector x to M dim Principle Components subspace, take just the first M coordinates of $V^T x$

Independent Components Analysis

Independent Component Analysis (ICA)

Find a linear transformation

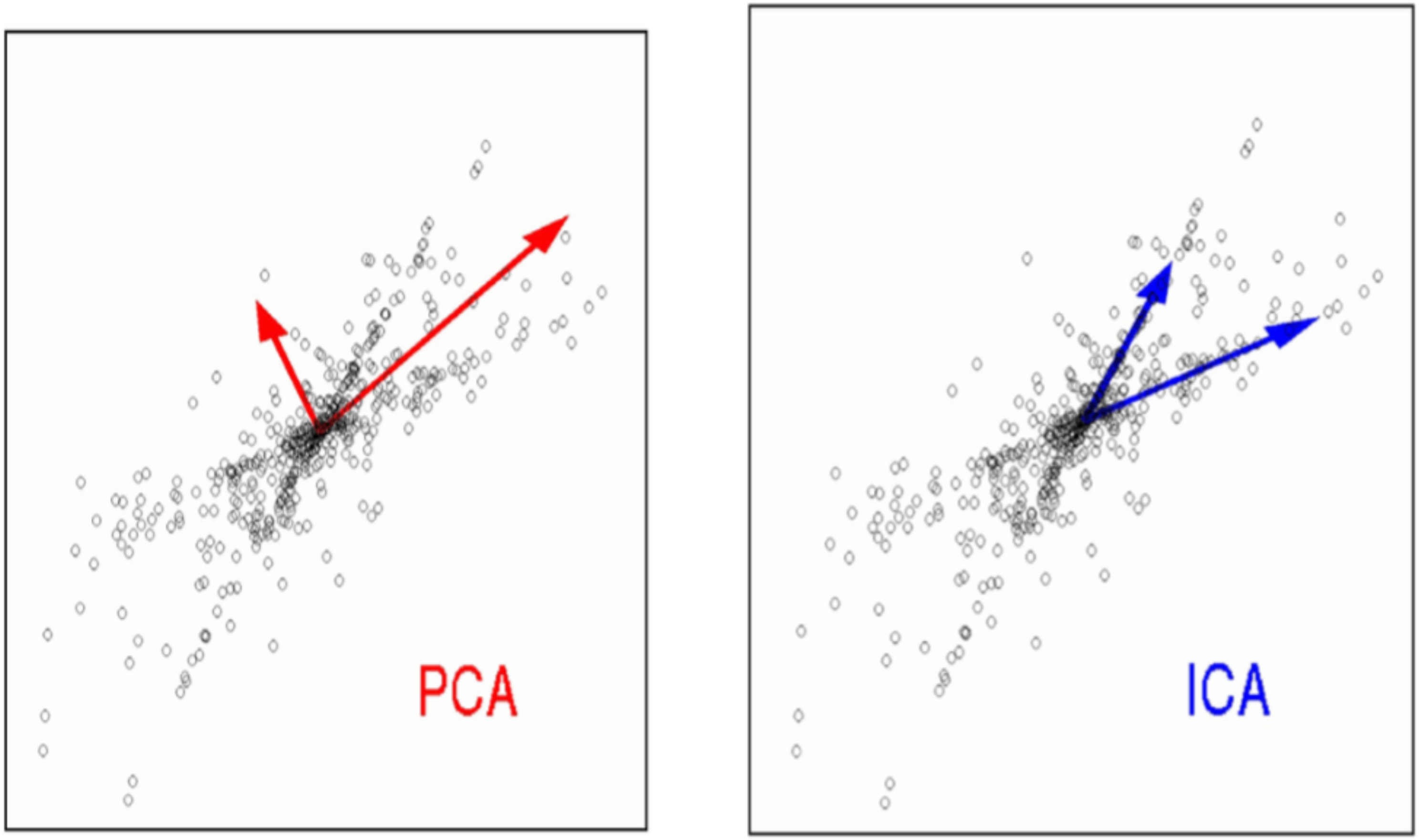
$$\boldsymbol{x} = V \cdot \boldsymbol{s}$$

for which coefficients $\boldsymbol{s} = (s_1, s_2, \dots, s_D)^T$ are **statistically independent**

$$p(s_1, s_2, \dots, s_D) = p_1(s_1)p_2(s_2) \dots p_n(s_D)$$

Algorithmically, we need to identify matrix V and coefficients s, s.t. under the condition $\boldsymbol{x} = V^T \cdot \boldsymbol{s}$ the **mutual information** between $s_1, s_2, \dots, s_D$ is minimized:

$$I(s_1, s_2, \dots, s_D) = \sum_{i=1}^{D} H(s_i) - H(s_1, s_2, \dots, s_D)$$



PCA finds directions of maximum variation,
ICA would find directions most "aligned" with data.

Canonical Correlation Analysis

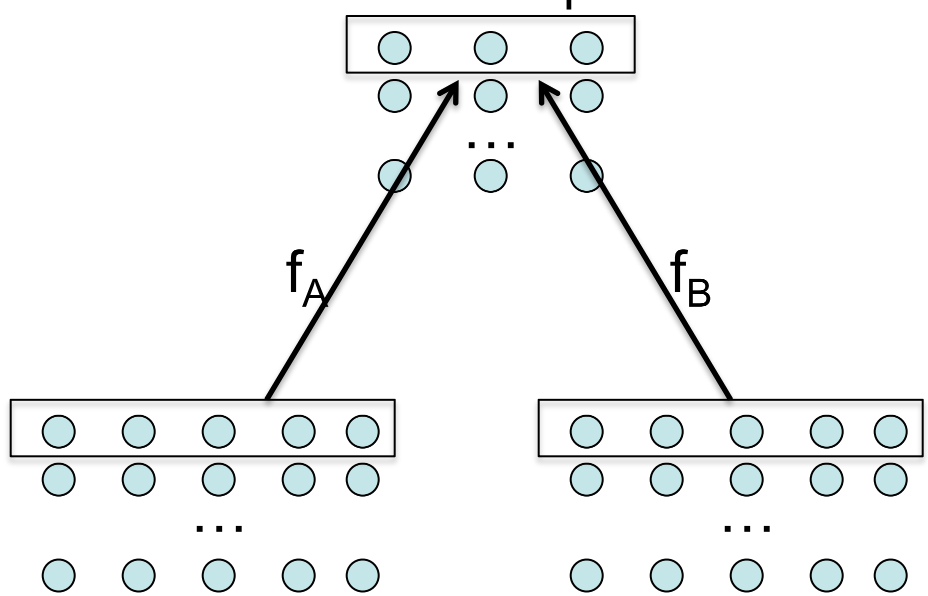
Learning Shared Representation across datasets

- Given data sets A and B, find linear projections of each into a <u>common</u> lower dimensional space
  - Canonical correlation analysis: $\max_{f_A,\, f_B} \Sigma_i \;\; \mathrm{corr}(f_A(A(i,:)), f_B(B(i,:)))$

$$Corr(A,B) = \frac{1}{N}\sum_{i=1}^{N} \frac{(A_i - \bar{A})}{\sigma_A} \; \frac{(B_i - \bar{B})}{\sigma_b}$$

where f_A, f_B are linear functions
(e.g., $f_A(A) = M\,A$)

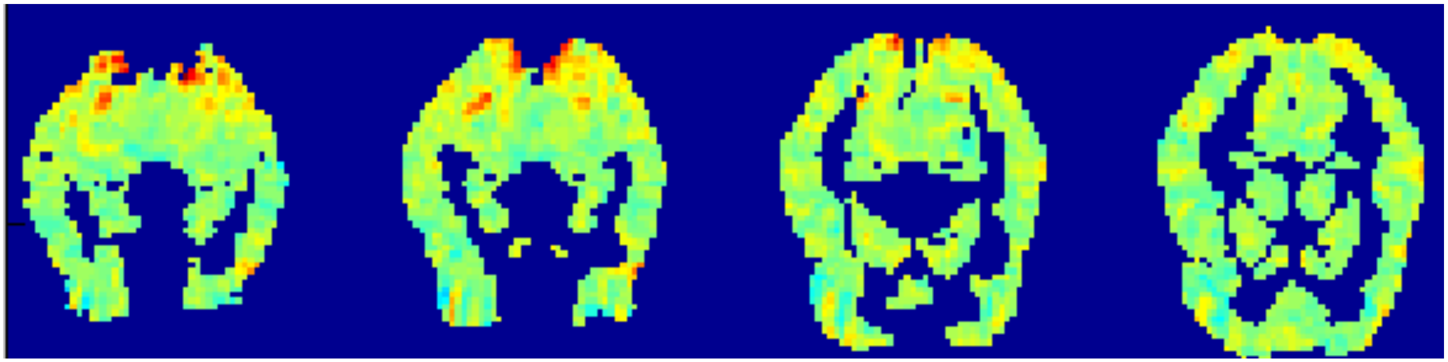
inferred shared representation

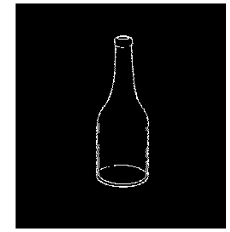


data set A

data set B

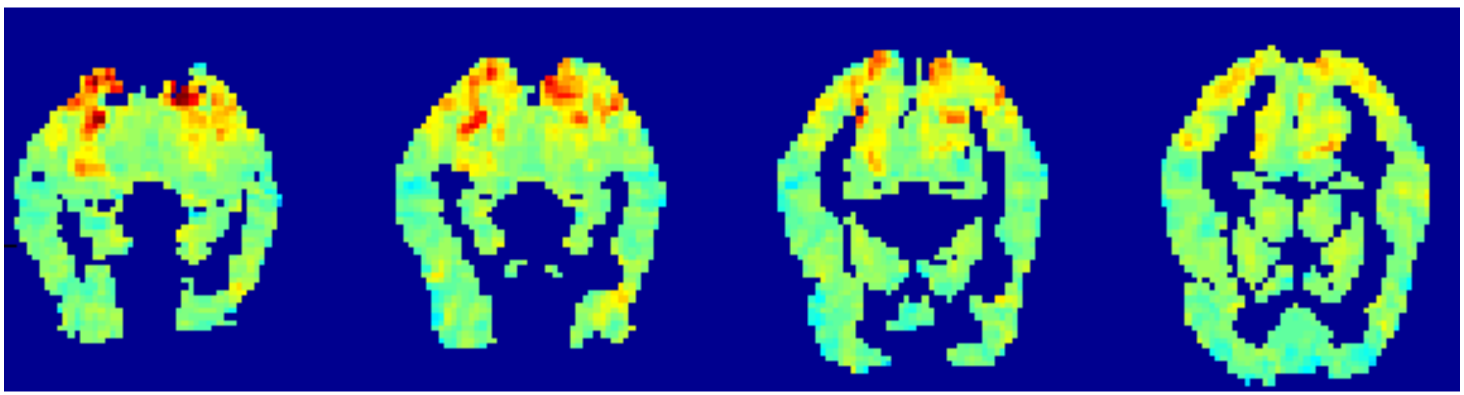
fMRI activation for "bottle":



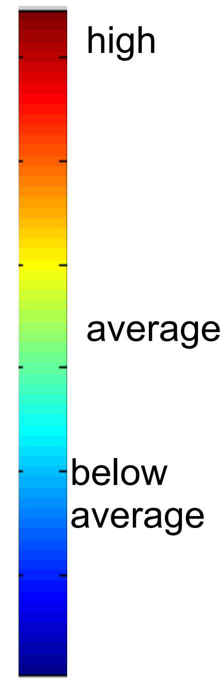


bottle

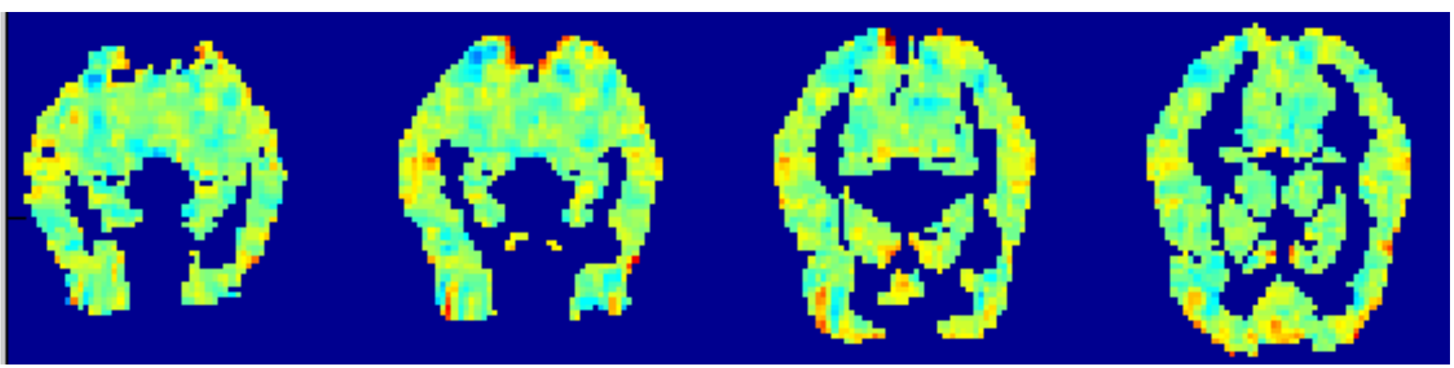
Mean activation averaged over 60 different stimuli:



fMRI activation



"bottle" minus mean activation:

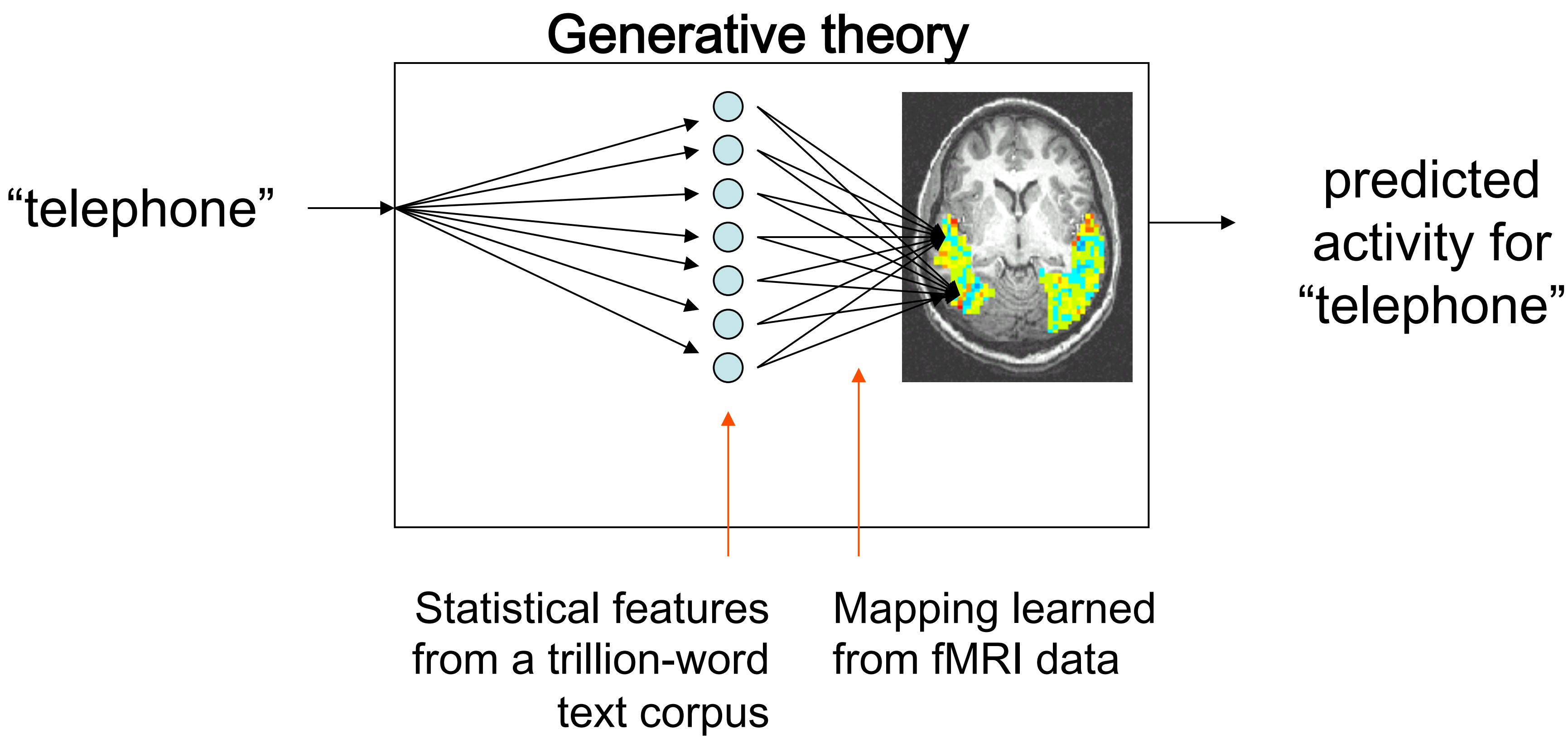


Can we discover underlying principles of neural encodings?

arbitrary word → **<u>Generative theory</u>** of word representation → predicted brain activity

Idea: Predict neural activity from corpus statistics of stimulus word

[Mitchell et al., *Science*, 2008]



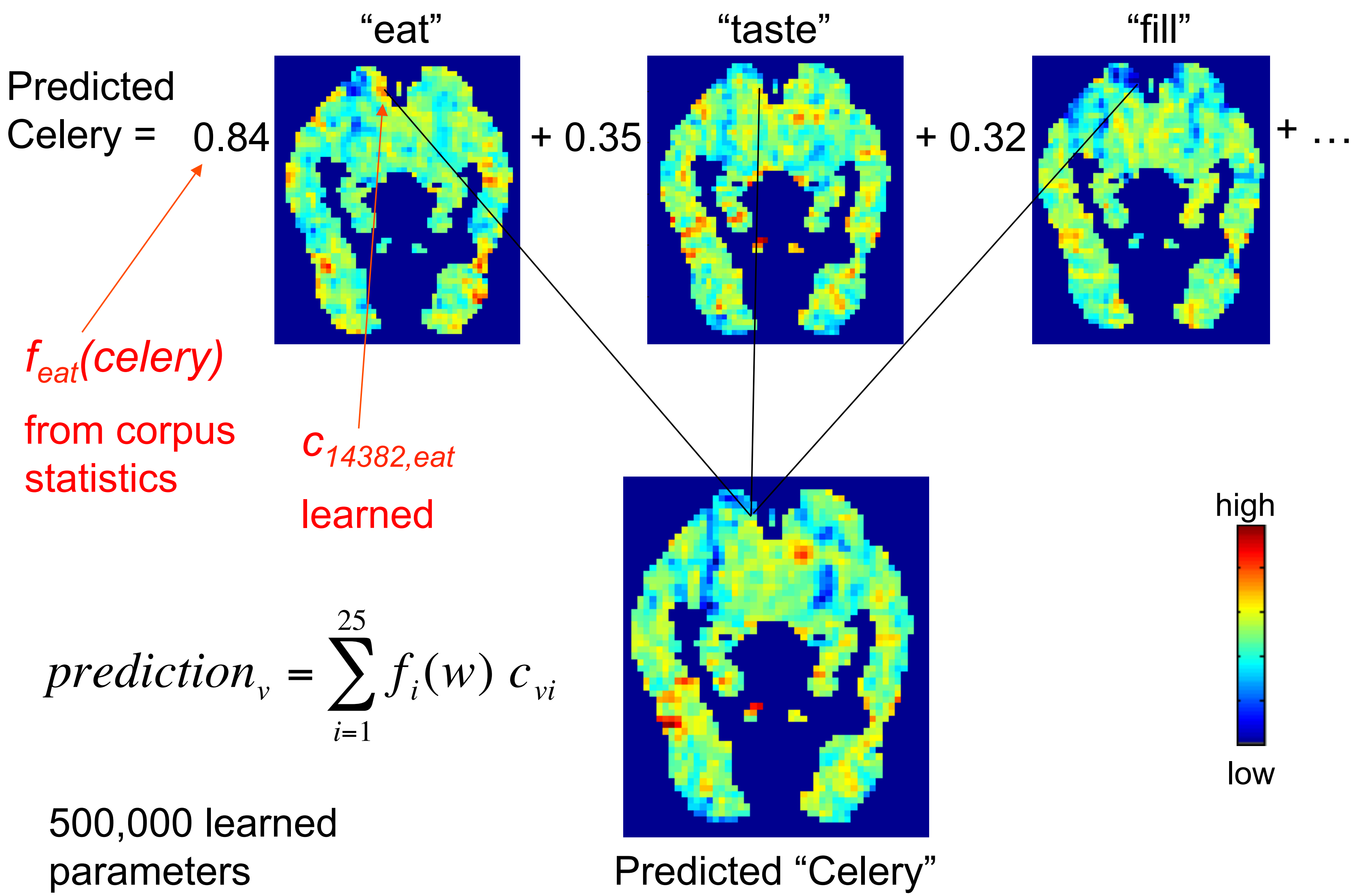
Semantic feature values:
"**celery**"

0.8368, eat
0.3461, taste
0.3153, fill
0.2430, see
0.1145, clean
0.0600, open
0.0586, smell
0.0286, touch
…
…
0.0000, drive
0.0000, wear
0.0000, lift
0.0000, break
0.0000, ride

Semantic feature values:
"**airplane**"

0.8673, ride
0.2891, see
0.2851, say
0.1689, near
0.1228, open
0.0883, hear
0.0771, run
0.0749, lift
…
…
0.0049, smell
0.0010, wear
0.0000, taste
0.0000, rub
0.0000, manipulate

Predicted Activation is Sum of Feature Contributions

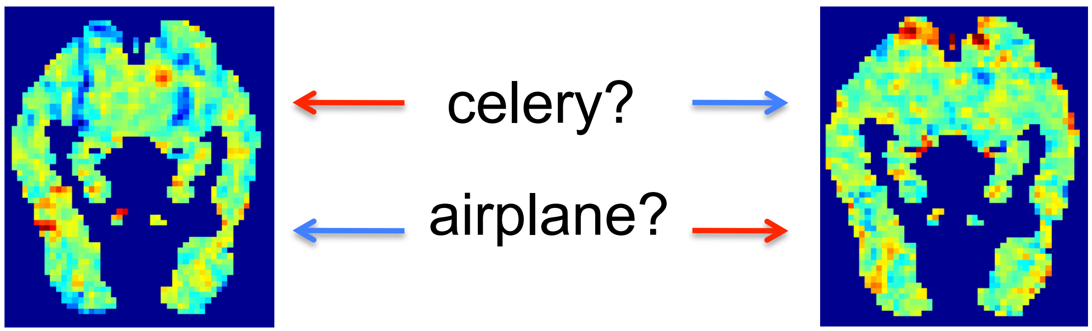


$$prediction_v = \sum_{i=1}^{25} f_i(w)\, c_{vi}$$

500,000 learned parameters

Evaluating the Computational Model

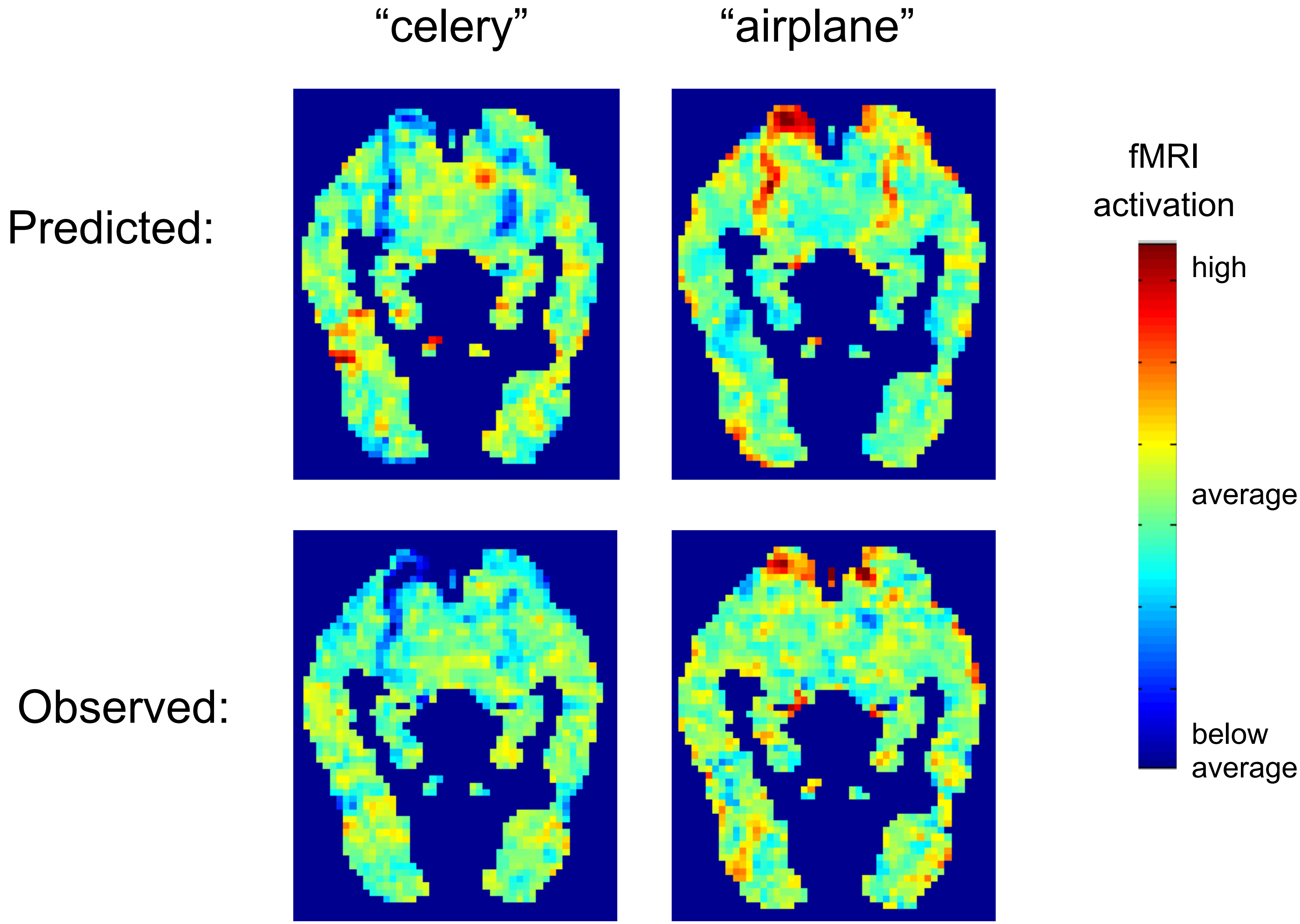
- Train it using 58 of the 60 word stimuli
- Apply it to predict fMRI images for other 2 words
- Test: show it the observed images for the 2 held-out, and make it predict which is which



1770 test pairs in leave-2-out:

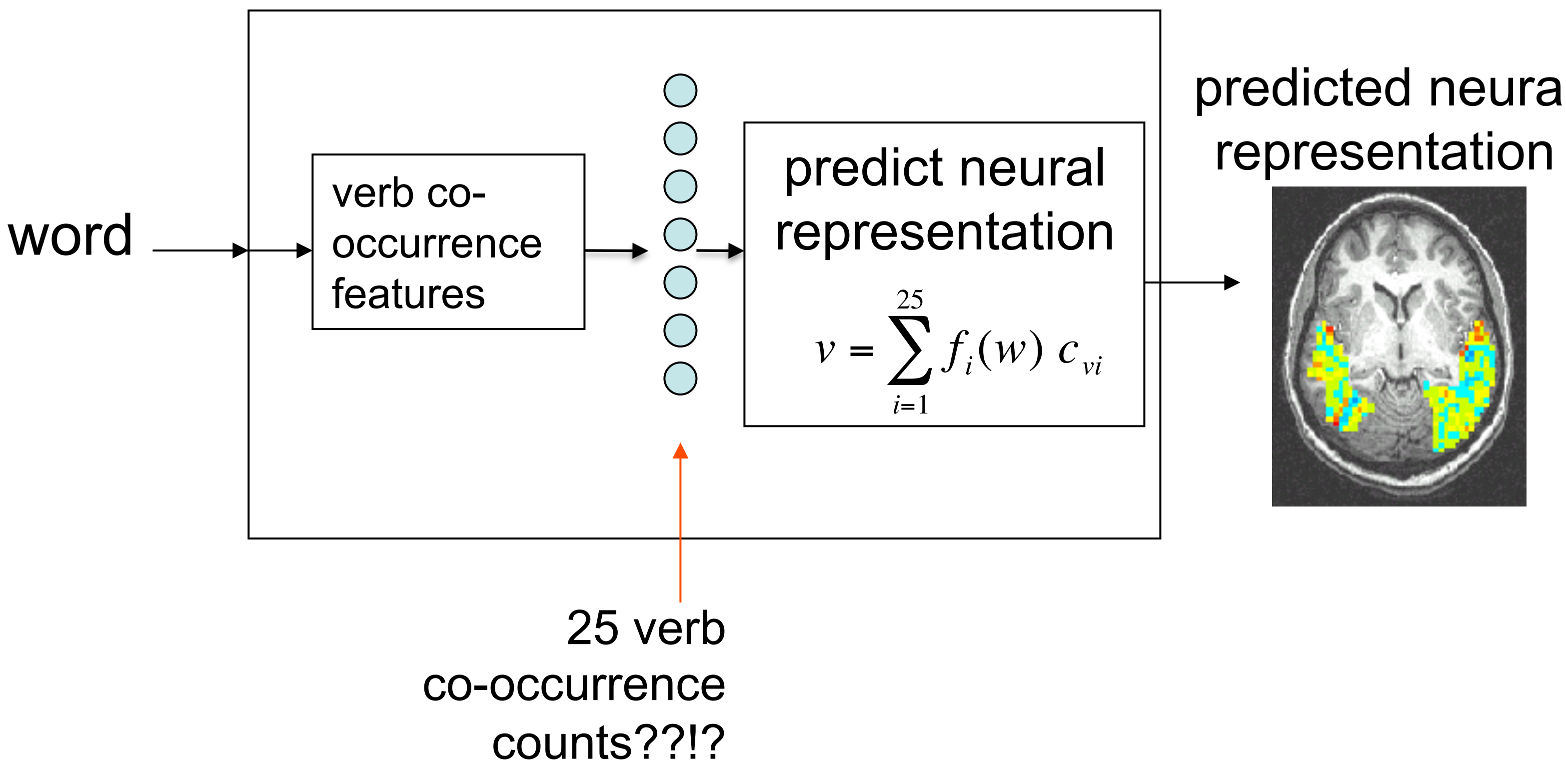
- Random guessing → 0.50 accuracy
- Accuracy above 0.61 is significant ($p<0.05$)

Mean accuracy over 9 subjects: 0.79



Predicted and observed fMRI images for “celery” and “airplane” after training on 58 other words.

Q: What are the actual semantic primitives from which neural encodings are composed?



Alternative semantic feature sets

PREDEFINED corpus features	**Mean Acc.**
25 verb co-occurrences	.79
486 verb co-occurrences	.79
50,000 word co-occurences	.76
300 Latent Semantic Analysis features	.73
50 corpus features from Collobert&Weston ICML08	.78
218 features collected using *Mechanical Turk**	**.83**
20 features discovered from the data**	**.86**

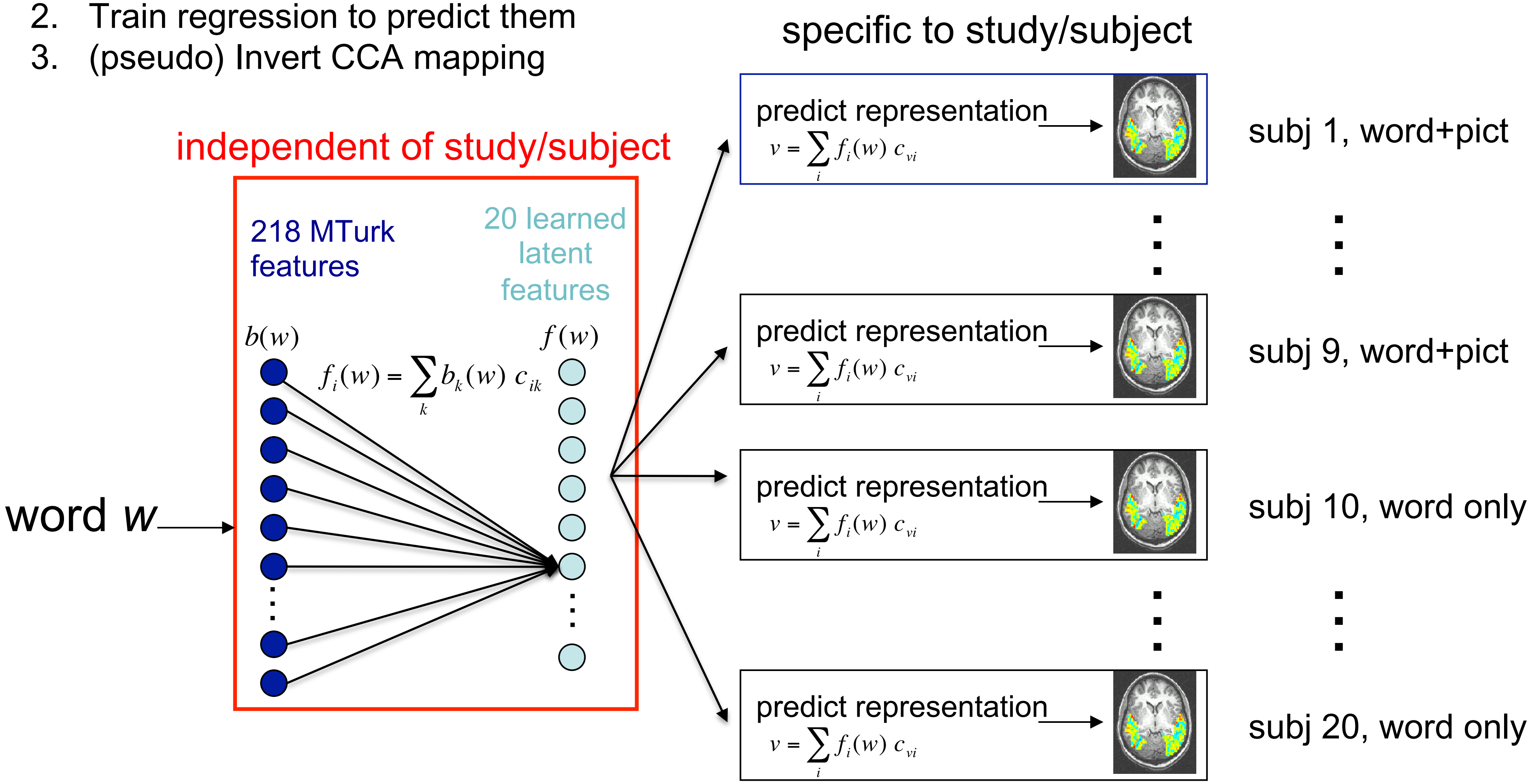
* developed by Dean Pommerleau

** developed by Indra Rustandi

Discovering shared semantic basis

[Rustandi et al., 2009]

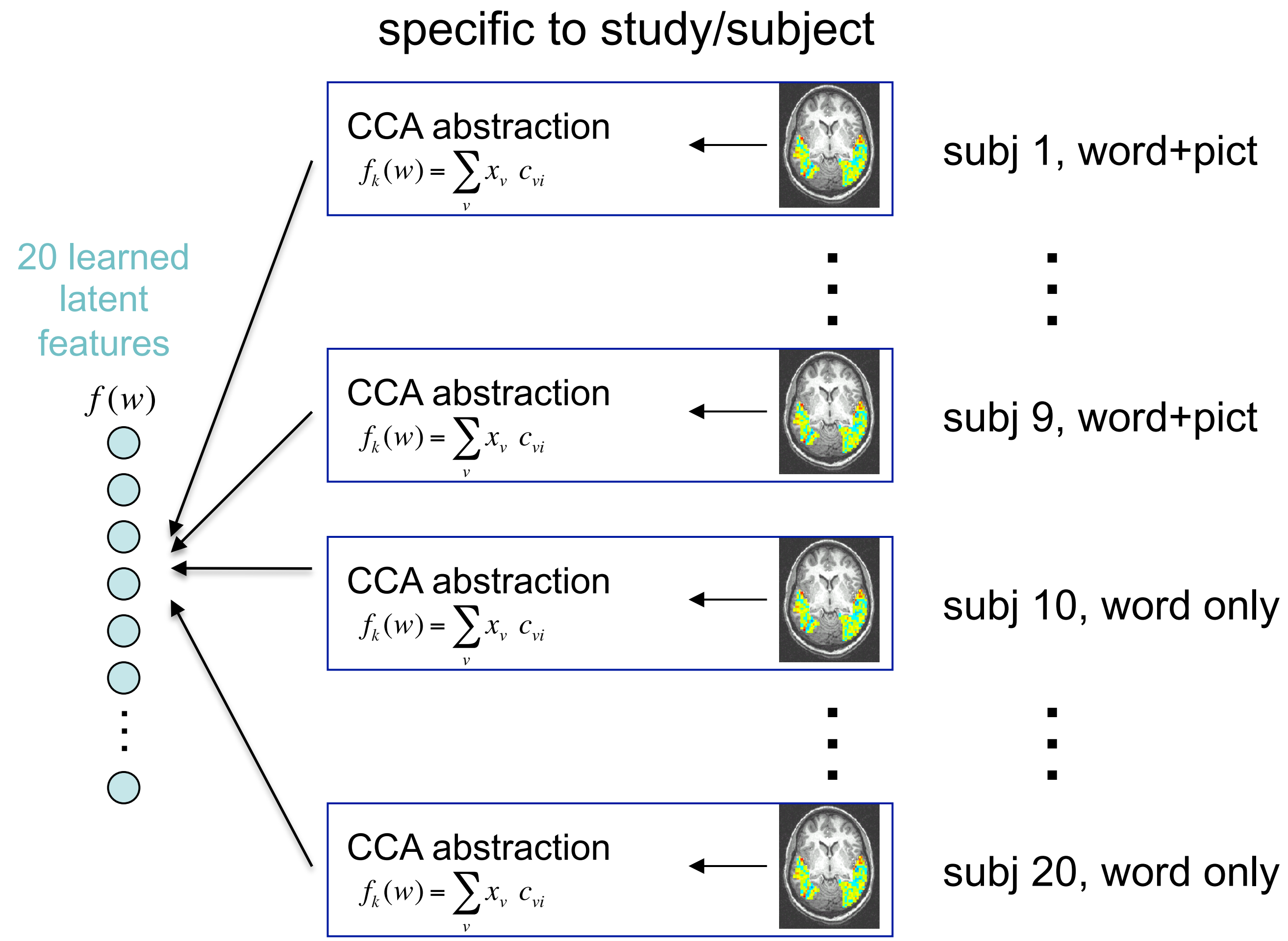
1. Use CCA to discover latent features
2. Train regression to predict them
3. (pseudo) Invert CCA mapping



Discovering shared semantic basis

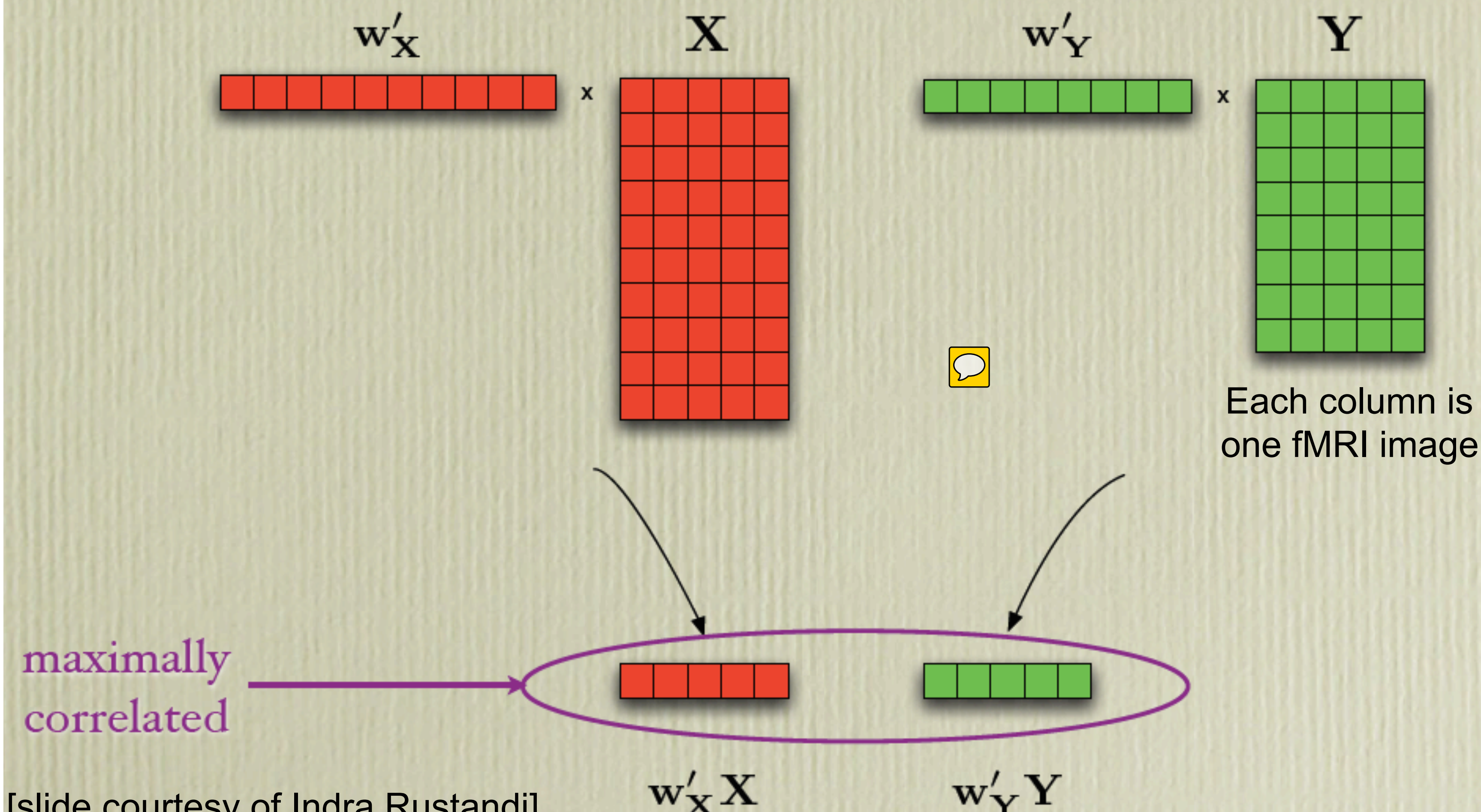
[Rustandi et al., 2009]

1. Use CCA to discover latent features across subjects



Canonical correlation analysis

$$Corr(A,B) = \frac{1}{N}\sum_{i=1}^{N}\frac{(A_i-\bar{A})}{\sigma_A}\ \frac{(B_i-\bar{B})}{\sigma_b}$$

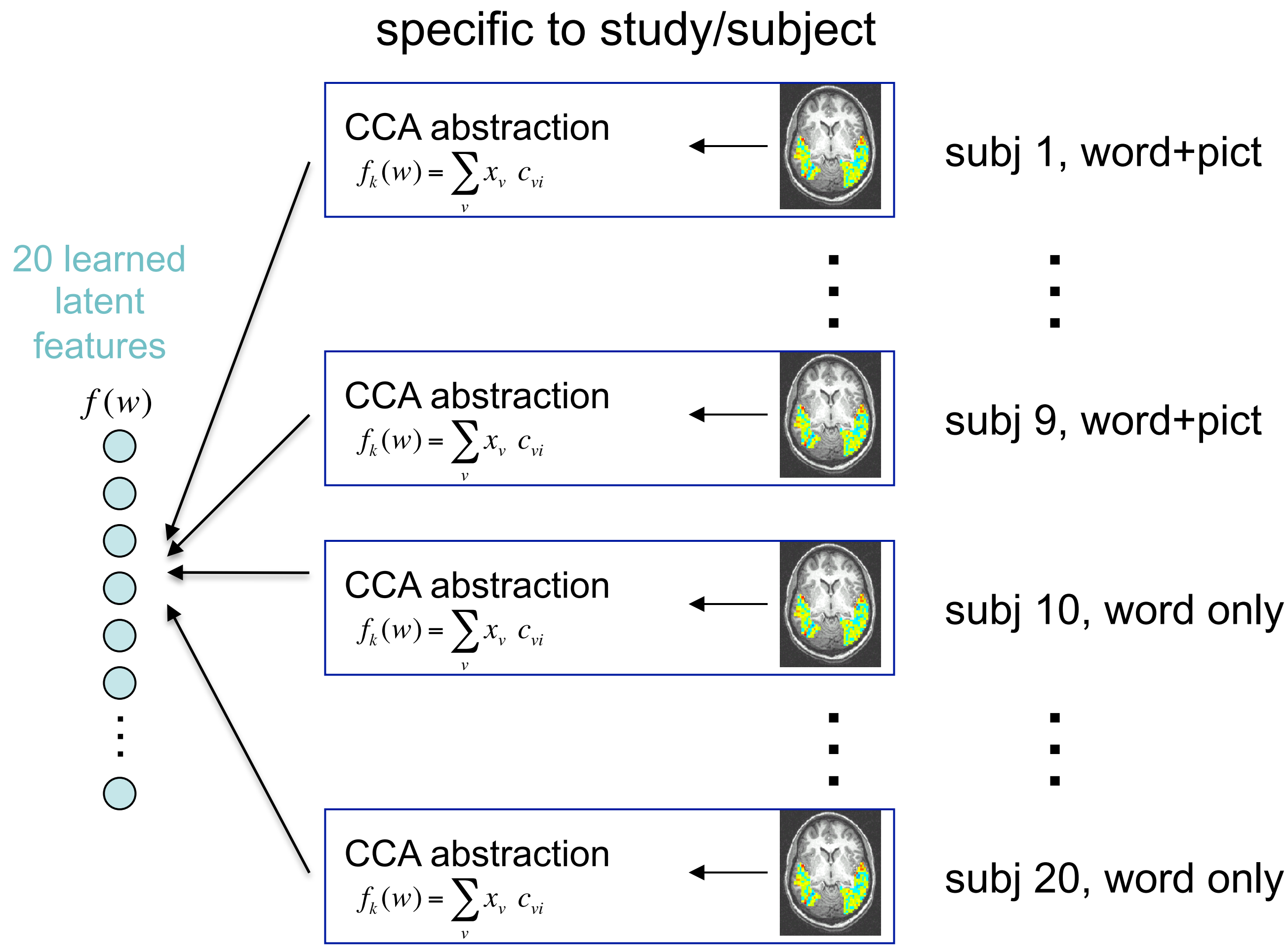


[slide courtesy of Indra Rustandi]

Discovering shared semantic basis

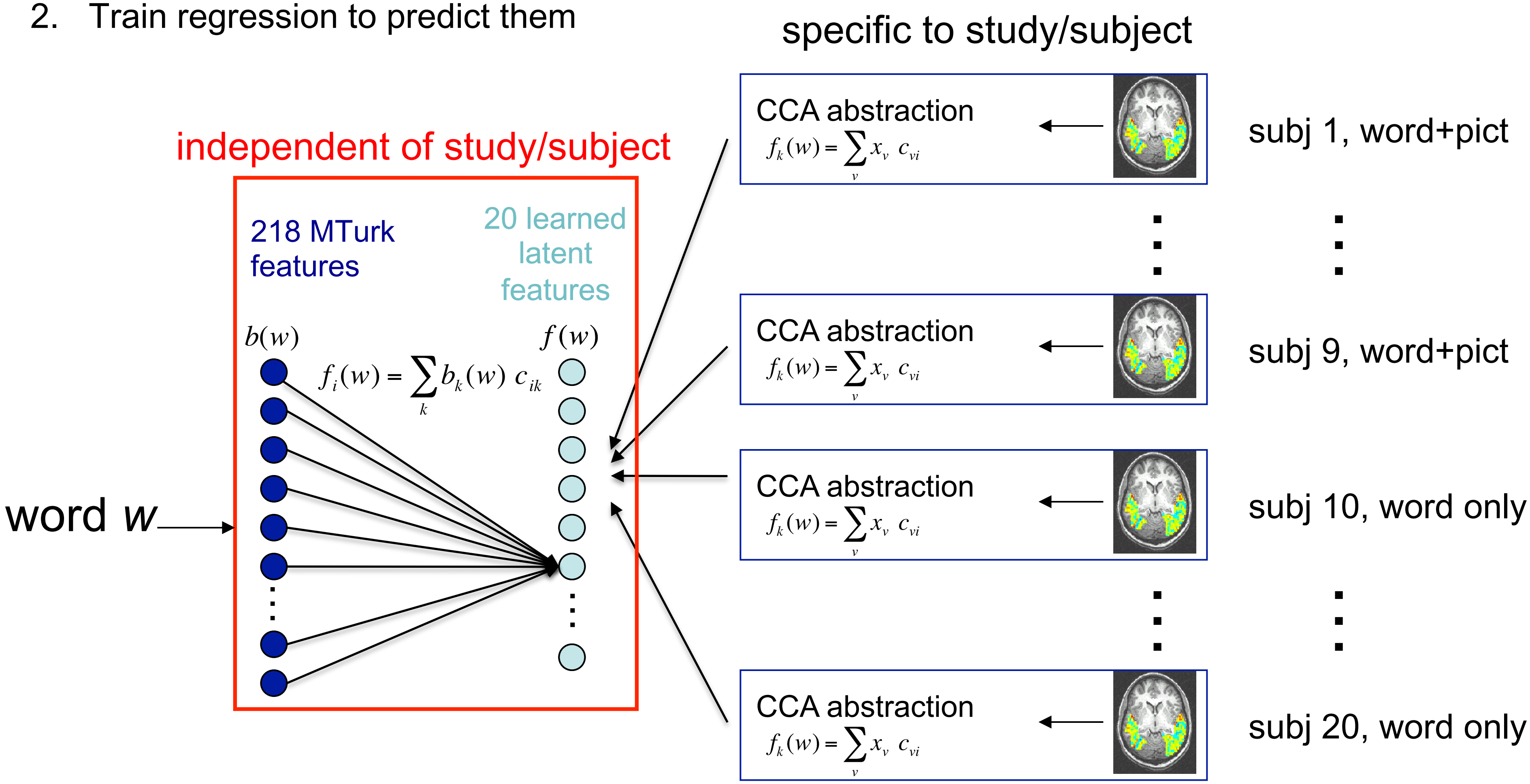
[Rustandi et al., 2009]

1. Use CCA to discover latent features across subjects



Discovering shared semantic basis [Rustandi et al., 2009]

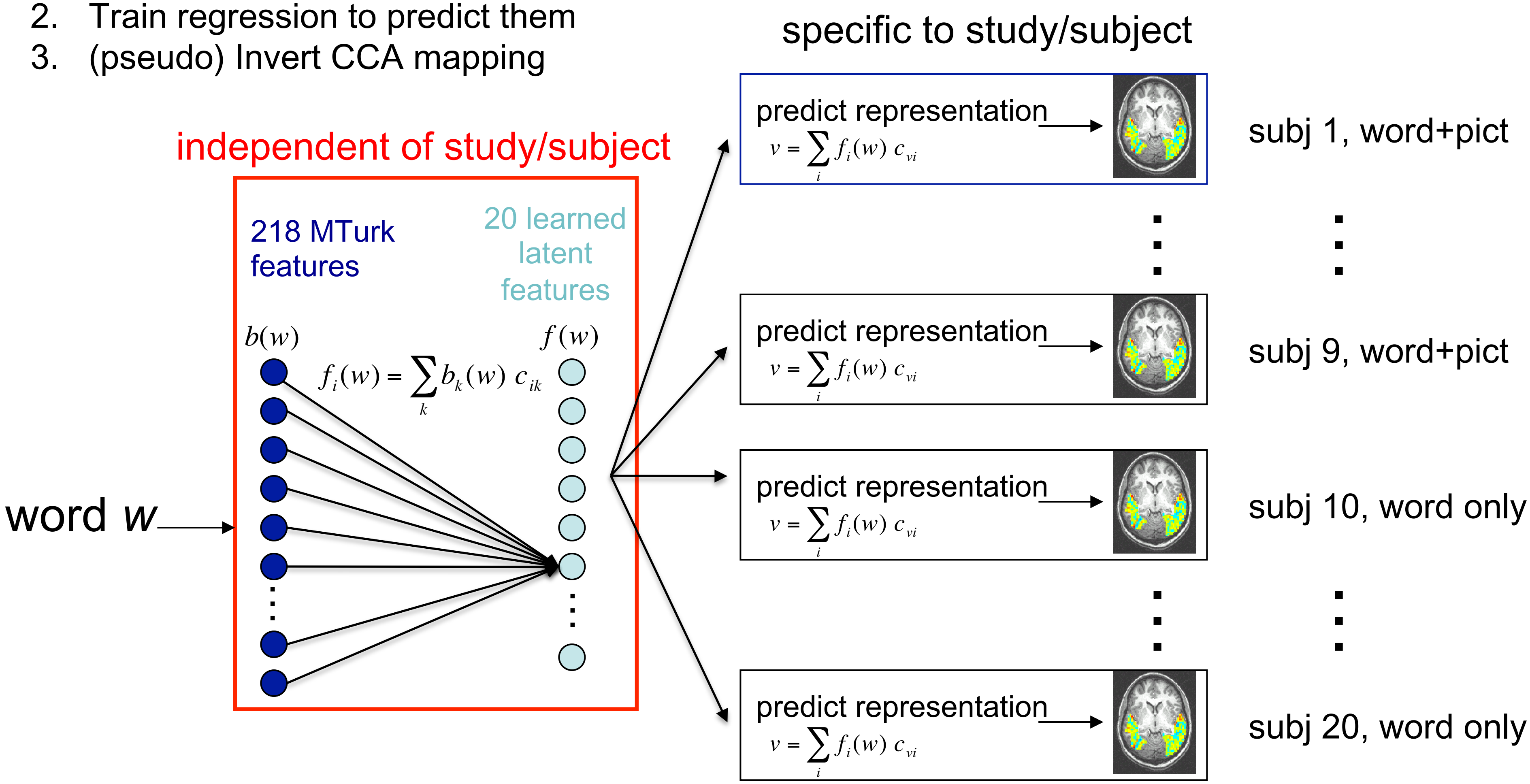
1. Use CCA to discover latent features
2. Train regression to predict them



Discovering shared semantic basis

[Rustandi et al., 2009]

1. Use CCA to discover latent features
2. Train regression to predict them
3. (pseudo) Invert CCA mapping

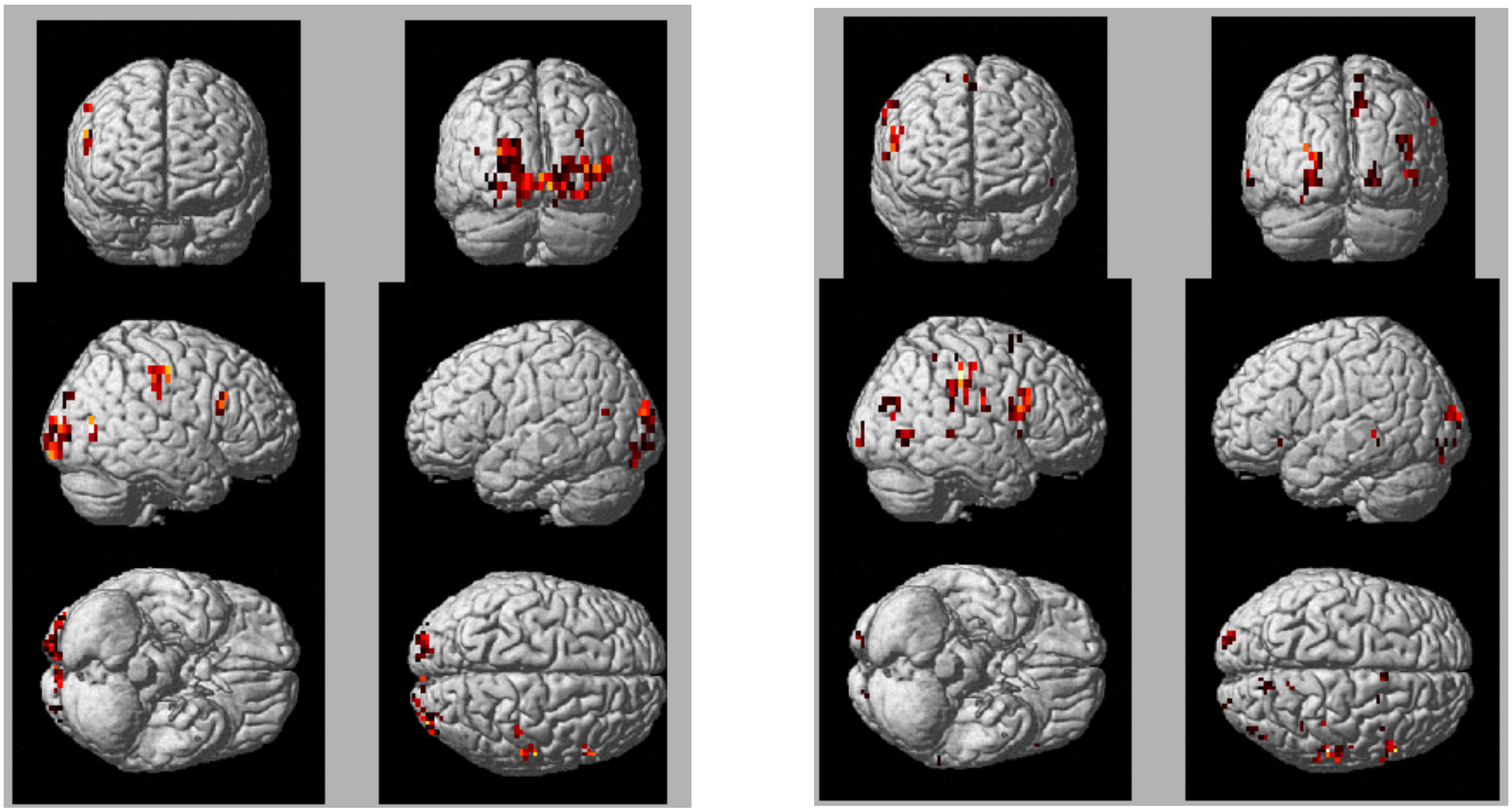


Multi-study (WP+WO) Multi-subject (9+11) CCA
Top Stimulus Words

	component 1	component 2	component 3	component 4
most active stimuli	apartment church closet house barn	screwdriver pliers refrigerator knife hammer	telephone butterfly bicycle beetle dog	pants dress glass coat chair
	shelter?	manipulation?		things that touch me?

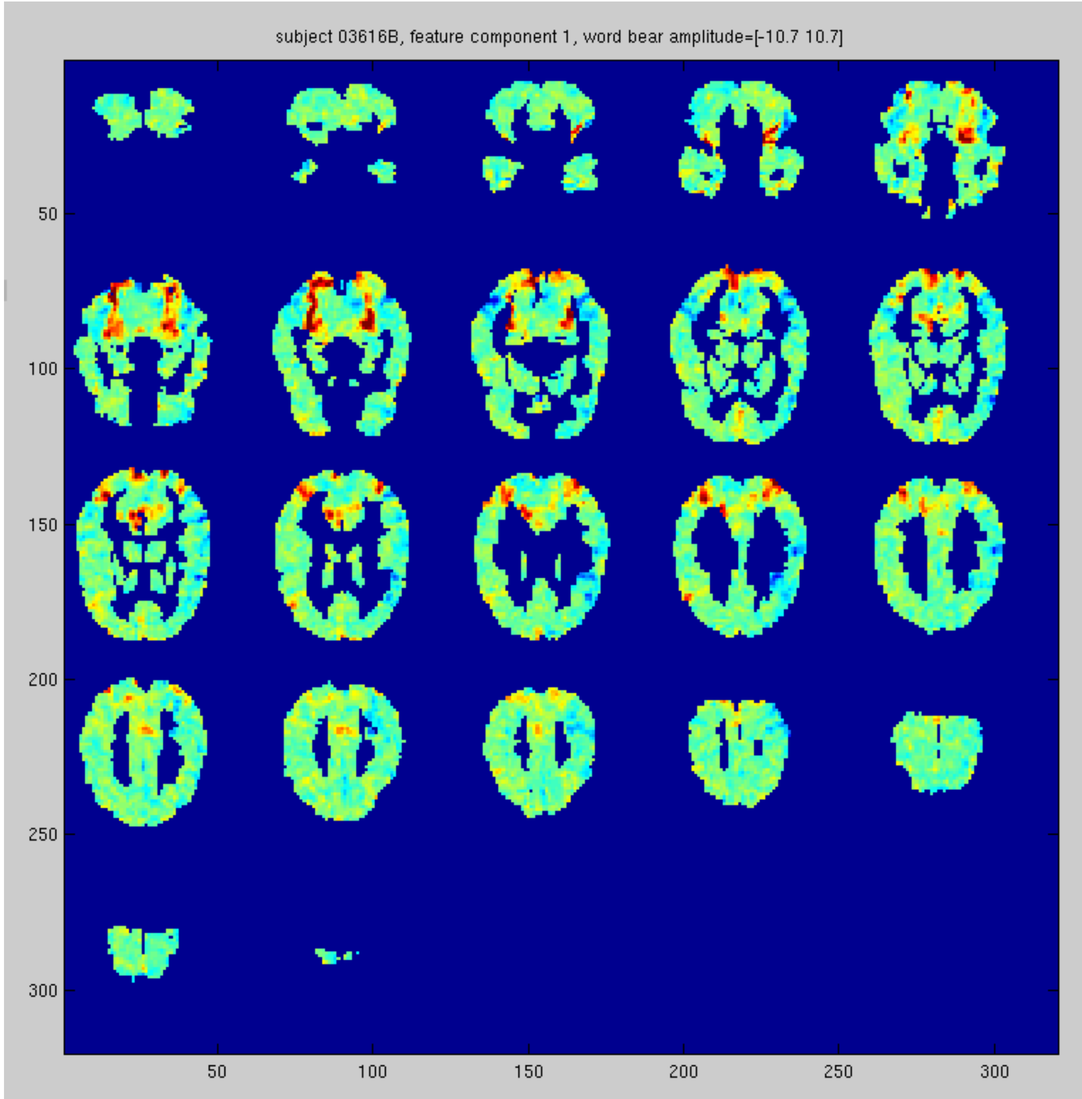
Projecting fMRI data from two brains into common space
CCA Component 1

[Rustandi, 2009]

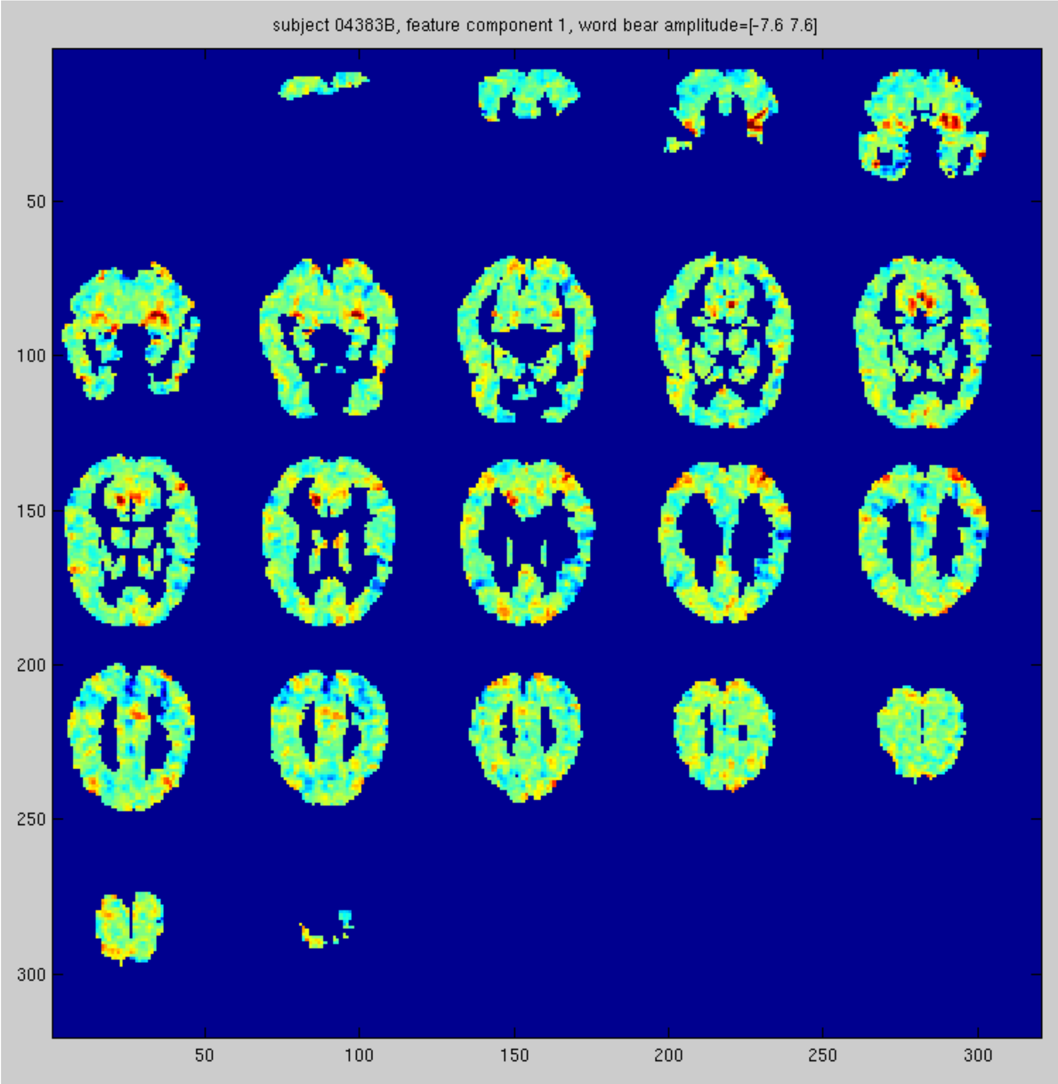


Text Top + verbs: bite chop cut tip ring grip hold bend dry remove
Top - verbs: build travel plan unite repair see sleep open walk live

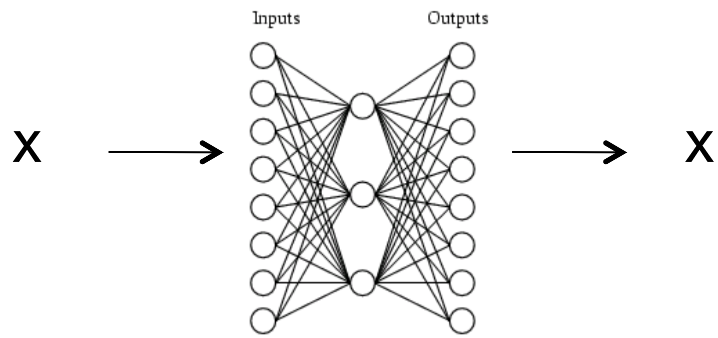
Subject 1 (Word-Picture stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 1



Subject 1 (Word-ONLY stimuli)
Multi-study (WP+WO) Multi-subject (9+11) CCA
Component 1



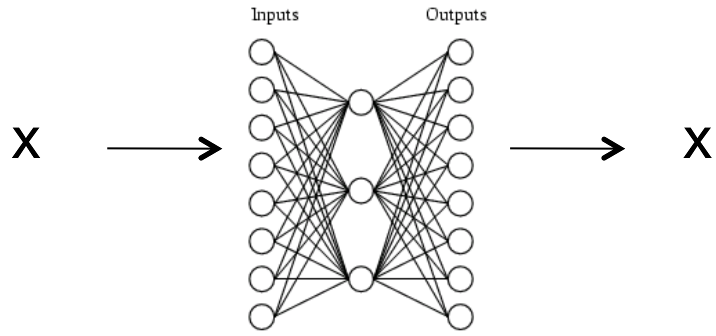
Matrix Factorization


X
Inputs
Outputs
X

Matrix Factorization

- $\text{minimize}_{H,W} \, || X - WH ||^2$
  such that ….

- Non-negative matrix factorization
- L1 regularized
- L2 regularized
- ...

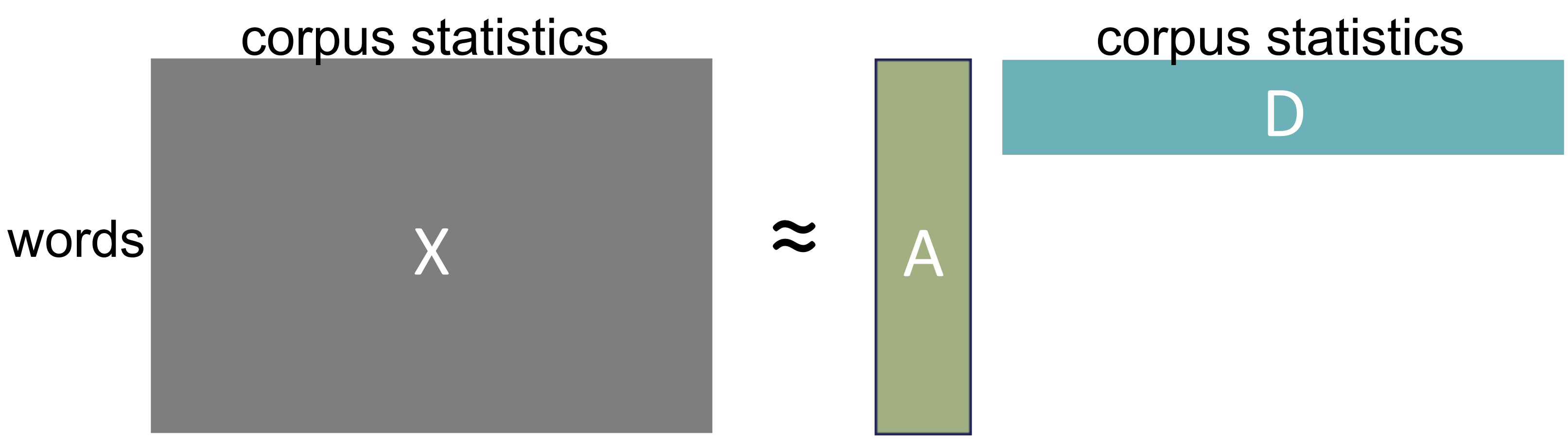


Non-Negative Sparse Embeddings

$$\underset{A,D}{\operatorname{argmin}} \sum_{i=1}^{w} \left(\left\| X_{i,:} - A_{i,:} \times D \right\|^2 + \lambda \left\| A_{i,:} \right\|_1 \right)$$

$$\text{subject to:} \quad A_{i,j} \geq 0, \forall 1 \leq i \leq w, \forall 1 \leq j \leq \ell$$

$$D_{i,:} D_{i,:}^T \leq 1, \forall\ 1 \leq i \leq \ell$$



Fyshe 2014

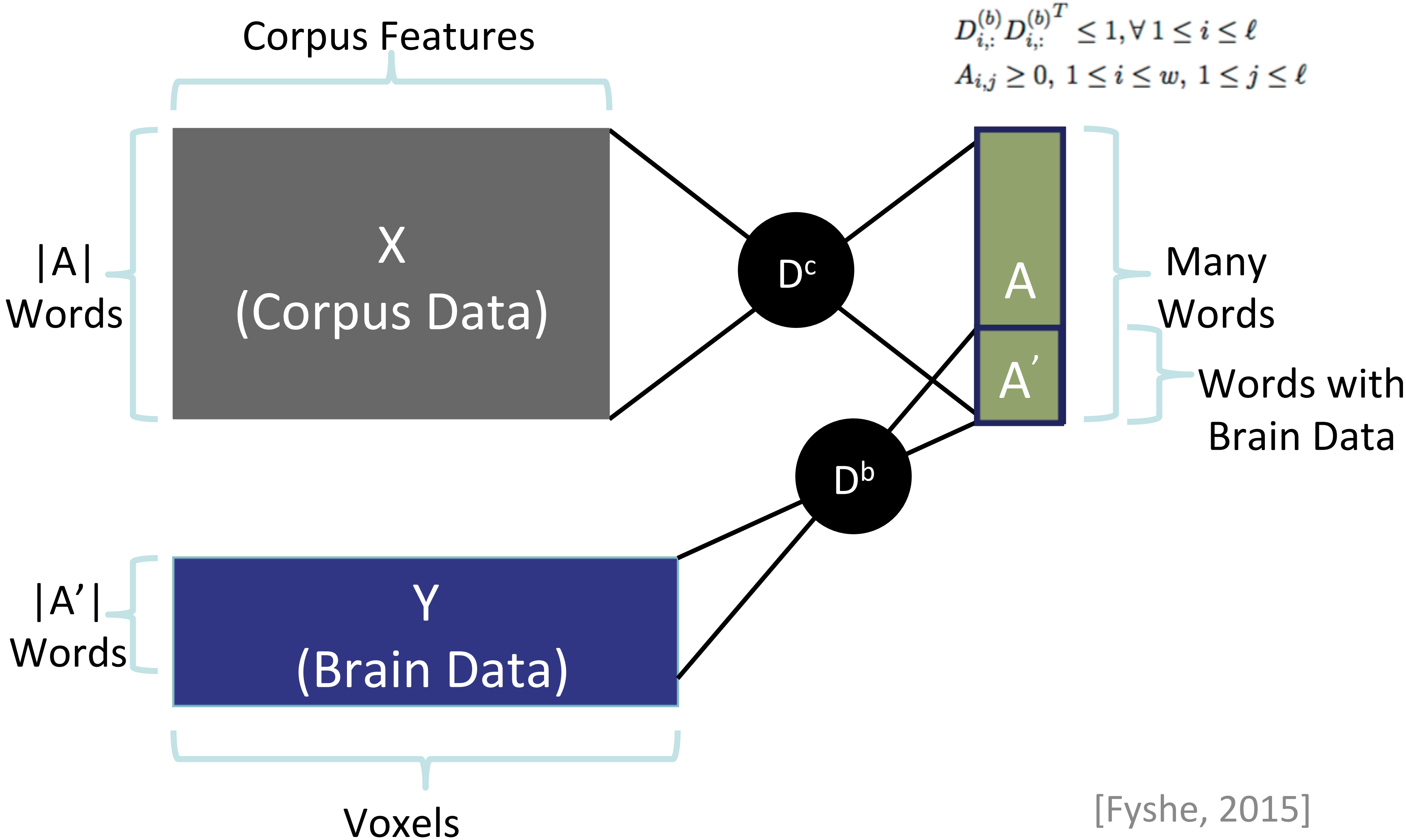
Interpretability

- PCA components
  - well, long, if, year, watch
  - plan, engine, e, rock, very
  - get, no, features, music, via
  - features, by, links, free, down
  - works, sound, video, building, section
- NNSE components
  - inhibitor, inhibitors, antagonists, receptors, inhibition
  - bristol, thames, southampton, brighton, poole
  - delhi, india, bombay, chennai, madras
  - pundits, forecasters, proponents, commentators, observers
  - nosy, averse, leery, unsympathetic, snotty

Text + Brain Matrix Factorization

$$\underset{A,D^c,D^b}{\operatorname{argmin}} \sum_{i=1}^{w} \left(\left\| X_{i,:} - A_{i,:} \times D^c \right\|^2 \right) + \sum_{i=1}^{w'} \left(\left\| Y_{i,:} - A_{i,:} \times D^b \right\|^2 \right) + \lambda |A|$$

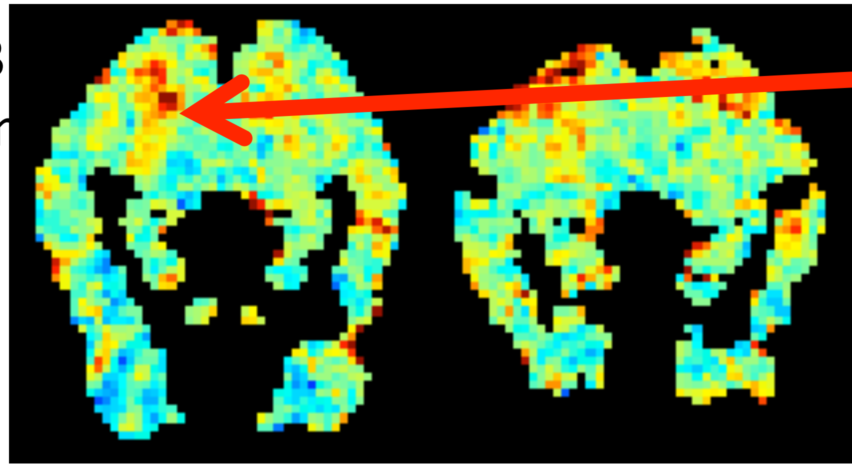
subject to:
$$D_{i,:}^{(c)} D_{i,:}^{(c)T} \leq 1, \forall 1 \leq i \leq \ell$$
$$D_{i,:}^{(b)} D_{i,:}^{(b)T} \leq 1, \forall 1 \leq i \leq \ell$$
$$A_{i,j} \geq 0,\ 1 \leq i \leq w,\ 1 \leq j \leq \ell$$



[Fyshe, 2015]

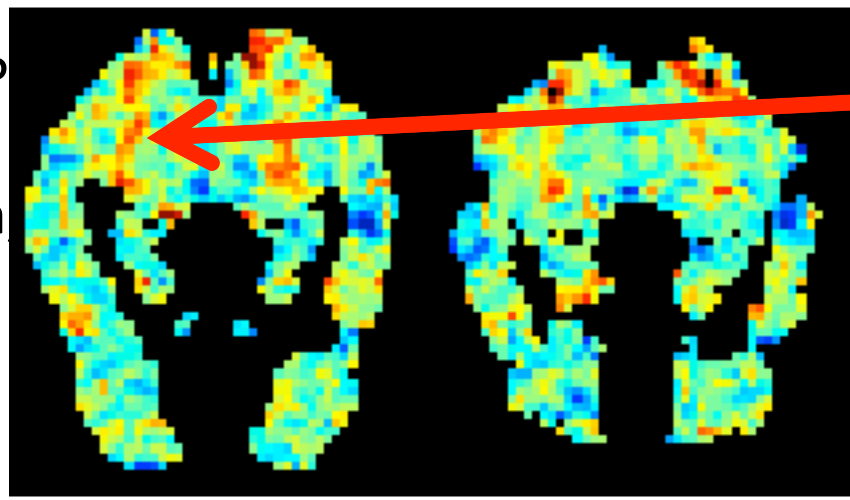
Mapping Semantics onto the Brain

Subject P3
“Bathroom
Balcony,
Kitchen”



Fusiform

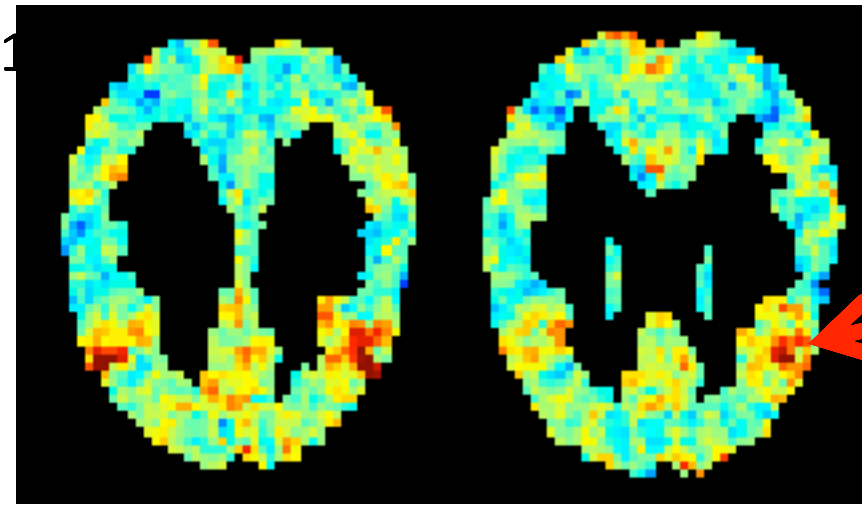
Subject P
“Parlor,
Cafeteria
Lounge”



Fusiform

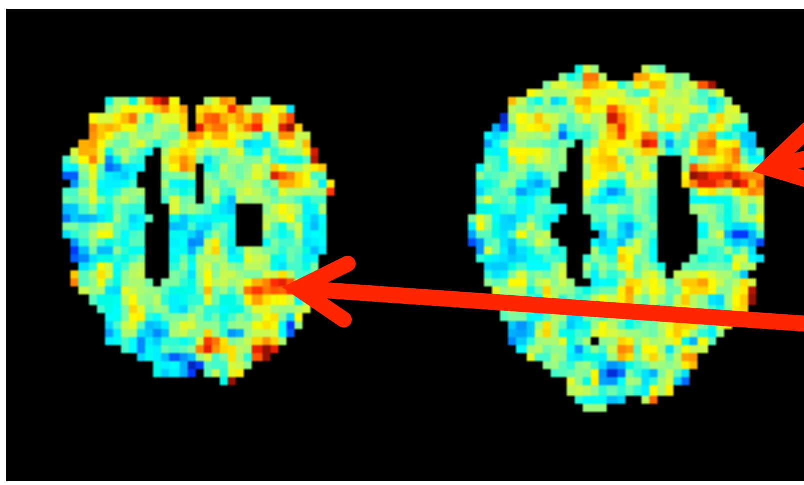
Mapping Semantics onto the Brain

Subject P1
“Buffet,
Brunch
Lunch”



Gustatory
Cortex

Subject P1
“Ankle,
Elbow,
Knee”



Postcentral

Precentral

What about Probabilistic Approaches?

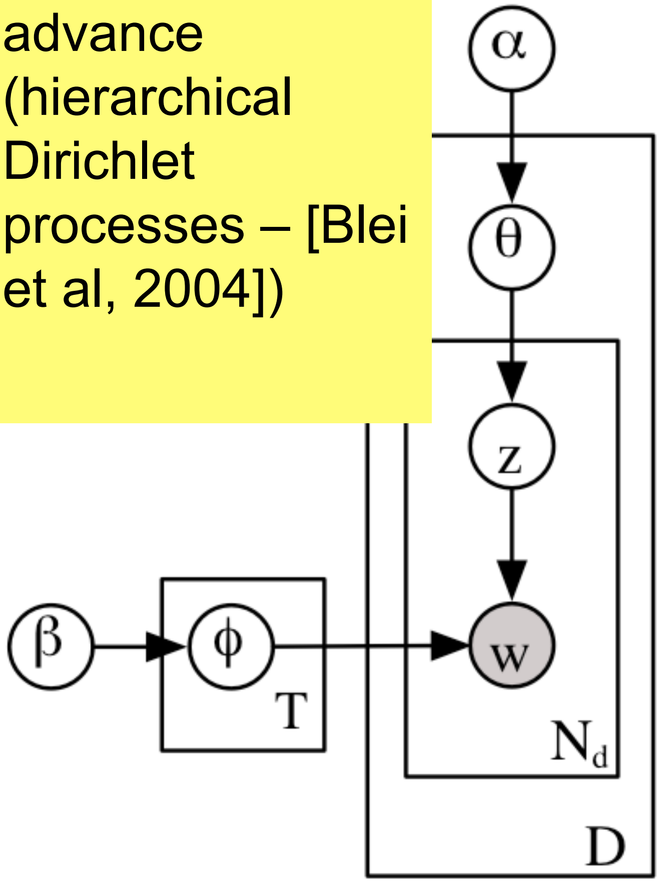
Latent Dirichlet Allocation

[David Blei]

Also extended to case where number of topics is not known in advance (hierarchical Dirichlet processes – [Blei et al, 2004])

ustering words into topics with ical Topic Models (unknown number of clusters)

[Blei, Ng, Jordan 2003]



Probabilistic model for generating document D:

1. Pick a $\theta \sim P(\theta|\alpha)$ to define $P(z|\theta)$
2. For each of the N_d words w
   - Pick topic z from $P(z \mid \theta)$
   - Pick word w from $P(w \mid z, \phi)$

Training this model defines topics (i.e., ϕ which defines $P(W|Z)$)

distributions are dirichlet, catergorical, dirichlet, categorical

Latent Dirichlet Allocation Model

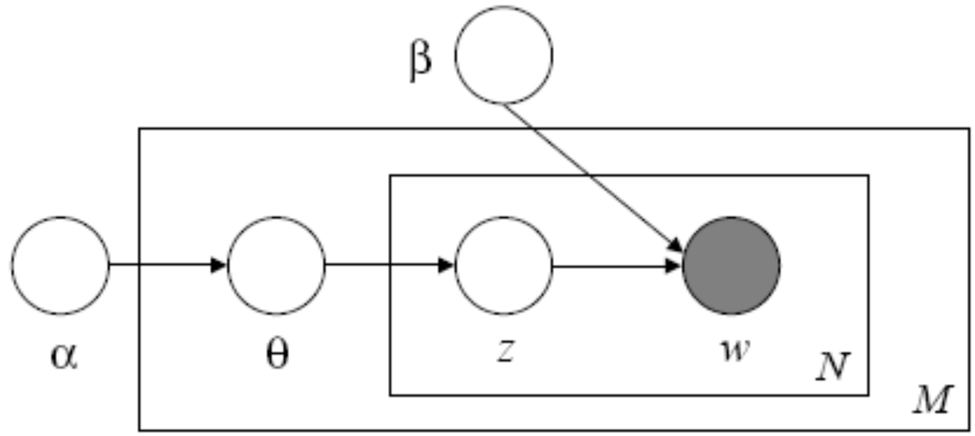


Figure 1: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

where $p(z_n \mid \theta)$ is simply θ_i for the unique i such that $z_n^i = 1$. Integrating over θ and summing over z, we obtain the marginal distribution of a document:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta. \tag{3}$$

Example topics induced from a large collection of text

DISEASE	WATER	MIND	STORY	FIELD	SCIENCE	BALL	JOB
BACTERIA	FISH	WORLD	STORIES	MAGNETIC	STUDY	GAME	WORK
DISEASES	SEA	DREAM	TELL	MAGNET	SCIENTISTS	TEAM	JOBS
GERMS	SWIM	DREAMS	CHARACTER	WIRE	SCIENTIFIC	FOOTBALL	CAREER
FEVER	SWIMMING	THOUGHT	CHARACTERS	NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CAUSE	POOL	IMAGINATION	AUTHOR	CURRENT	WORK	PLAYERS	EMPLOYMENT
CAUSED	LIKE	MOMENT	READ	COIL	RESEARCH	PLAY	OPPORTUNITIES
SPREAD	SHELL	THOUGHTS	TOLD	POLES	CHEMISTRY	FIELD	WORKING
VIRUSES	SHARK	OWN	SETTING	IRON	TECHNOLOGY	PLAYER	TRAINING
INFECTION	TANK	REAL	TALES	COMPASS	MANY	BASKETBALL	SKILLS
VIRUS	SHELLS	LIFE	PLOT	LINES	MATHEMATICS	COACH	CAREERS
MICROORGANISMS	SHARKS	IMAGINE	TELLING	CORE	BIOLOGY	PLAYED	POSITIONS
PERSON	DIVING	SENSE	SHORT	ELECTRIC	FIELD	PLAYING	FIND
INFECTIOUS	DOLPHINS	CONSCIOUSNESS	FICTION	DIRECTION	PHYSICS	HIT	POSITION
COMMON	SWAM	STRANGE	ACTION	FORCE	LABORATORY	TENNIS	FIELD
CAUSING	LONG	FEELING	TRUE	MAGNETS	STUDIES	TEAMS	OCCUPATIONS
SMALLPOX	SEAL	WHOLE	EVENTS	BE	WORLD	GAMES	REQUIRE
BODY	DIVE	BEING	TELLS	MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
INFECTIONS	DOLPHIN	MIGHT	TALE	POLE	STUDYING	BAT	EARN
CERTAIN	UNDERWATER	HOPE	NOVEL	INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

Example topics induced from a large collection of text

Significance:

• Learned topics reveal implicit semantic categories of words within the documents

• In many cases, we can represent documents with 10^2 topics instead of 10^5 words

• Especially important for short documents (e.g., emails). Topics overlap when words don't !

FIELD	SCIENCE	BALL	JOB
MAGNETIC	STUDY	GAME	WORK
MAGNET	SCIENTISTS	TEAM	JOBS
WIRE	SCIENTIFIC	FOOTBALL	CAREER
NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CURRENT	WORK	PLAYERS	EMPLOYMENT
COIL	RESEARCH	PLAY	OPPORTUNITIES
POLES	CHEMISTRY	**FIELD**	WORKING
IRON	TECHNOLOGY	PLAYER	TRAINING
COMPASS	MANY	BASKETBALL	SKILLS
LINES	MATHEMATICS	COACH	CAREERS
CORE	BIOLOGY	PLAYED	POSITIONS
ELECTRIC	**FIELD**	PLAYING	FIND
DIRECTION	PHYSICS	HIT	POSITION
FORCE	LABORATORY	TENNIS	**FIELD**
MAGNETS	STUDIES	TEAMS	OCCUPATIONS
BE	WORLD	GAMES	REQUIRE
MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
POLE	STUDYING	BAT	EARN
INDUCED	SCIENCES	TERRY	ABLE

[Tennenbaum et al]

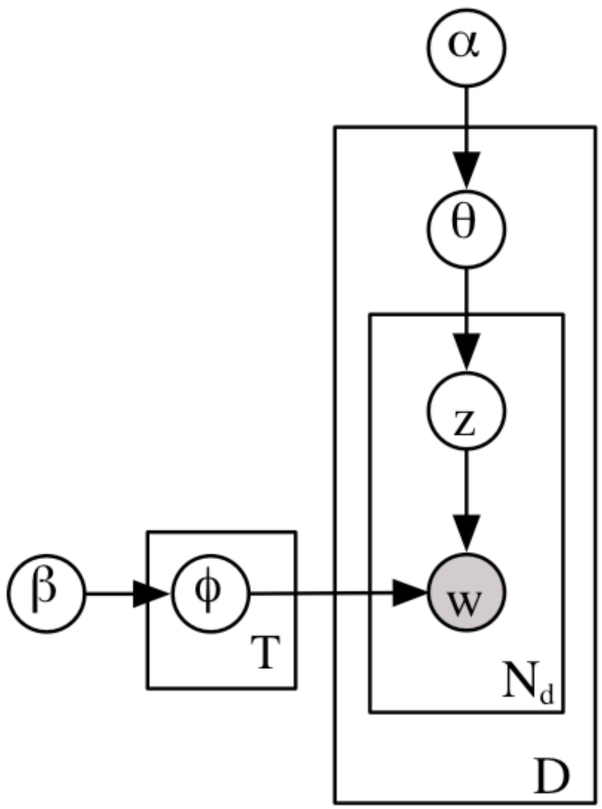
Analyzing topic distributions in email

Author-Recipient-Topic model for Email

Latent Dirichlet Allocation

(LDA)

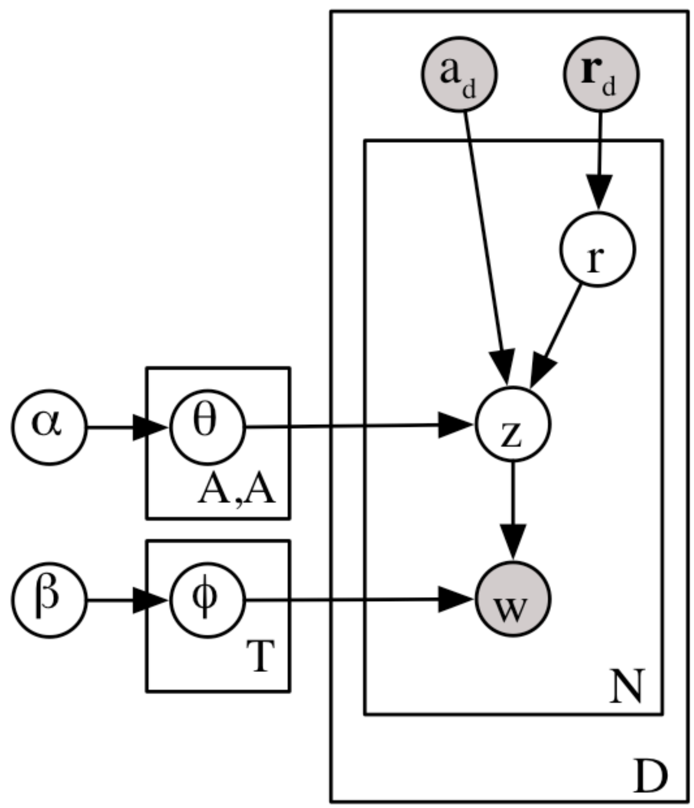
[Blei, Ng, Jordan, 2003]



Author-Recipient Topic

(ART)

[McCallum, Corrada, Wang, 2005]



Enron Email Corpus

- 250k email messages
- 23k people

```
Date: Wed, 11 Apr 2001 06:56:00 -0700 (PDT)
From: debra.perlingiere@enron.com
To: steve.hooser@enron.com
Subject: Enron/TransAltaContract dated Jan 1, 2001

Please see below. Katalin Kiss of TransAlta has requested an
electronic copy of our final draft?  Are you OK with this?  If
so, the only version I have is the original draft without
revisions.

DP

Debra Perlingiere
Enron North America Corp.
Legal Department
1400 Smith Street, EB 3885
Houston, Texas 77002
dperlin@enron.com
```

Topics, and prominent sender/receivers discovered by ART

[McCallum et al, 2005]

Top words within topic :

Top author-recipients exhibiting this topic

Topic 17 "Document Review"		Topic 27 "Time Scheduling"		Topic 45 "Sports Pool"	
attached	0.0742	day	0.0419	game	0.0170
agreement	0.0493	friday	0.0418	draft	0.0156
review	0.0340	morning	0.0369	week	0.0135
questions	0.0257	monday	0.0282	team	0.0135
draft	0.0245	office	0.0282	eric	0.0130
letter	0.0239	wednesday	0.0267	make	0.0125
comments	0.0207	tuesday	0.0261	free	0.0107
copy	0.0165	time	0.0218	year	0.0106
revised	0.0161	good	0.0214	pick	0.0097
document	0.0156	thursday	0.0191	phillip	0.0095
G.Nemec B.Tycholiz	0.0737	J.Dasovich R.Shapiro	0.0340	E.Bass M.Lenhart	0.3050
G.Nemec M.Whitt	0.0551	J.Dasovich J.Steffes	0.0289	E.Bass P.Love	0.0780
B.Tycholiz G.Nemec	0.0325	C.Clair M.Taylor	0.0175	M.Motley M.Grigsby	0.0522

Topics, and prominent sender/receivers discovered by ART

Topic 34 "Operations"		Topic 37 "Power Market"		Topic 41 "Government Relations"		Topic 42 "Wireless"	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	0.0239	website	0.0375
bob	0.0129	prices	0.0182	electricity	0.0203	report	0.0373
open	0.0126	high	0.0124	davis	0.0183	wireless	0.0364
meeting	0.0107	based	0.0120	utilities	0.0158	handheld	0.0362
gas	0.0107	buy	0.0117	commission	0.0136	stan	0.0282
business	0.0106	customers	0.0110	governor	0.0132	fyi	0.0271
houston	0.0099	costs	0.0106	prices	0.0089	named	0.0260
S.Beck L.Kitchen	0.2158	J.Dasovich J.Steffes	0.1231	J.Dasovich R.Shapiro	0.3338	R.Haylett T.Geaccone	0.1432
S.Beck J.Lavorato	0.0826	J.Dasovich R.Shapiro	0.1133	J.Dasovich J.Steffes	0.2440	T.Geaccone R.Haylett	0.0737
S.Beck S.White	0.0530	M.Taylor E.Sager	0.0218	J.Dasovich R.Sanders	0.1394	R.Haylett D.Fossum	0.0420

Beck = "Chief Operations Officer"

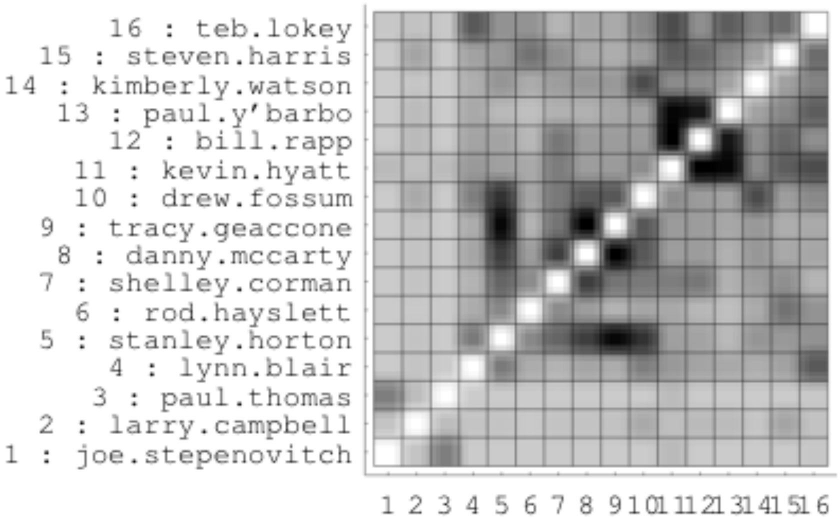
Dasovich = "Government Relations Executive"

Shapiro = "Vice Presidence of Regulatory Affairs'

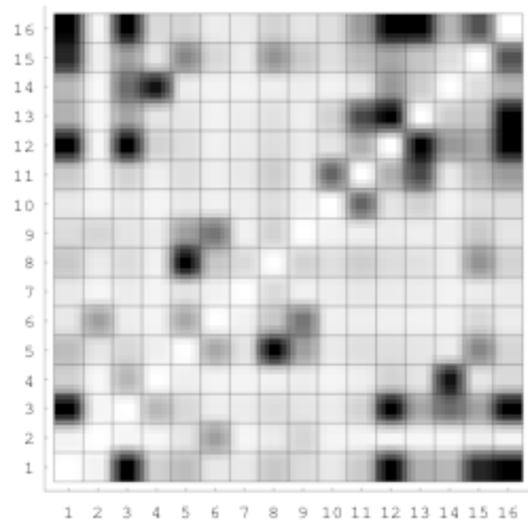
Steffes = "Vice President of Government Affairs"

Discovering Role Similarity

Traditional SNA



ART



connection strength (A,B) =

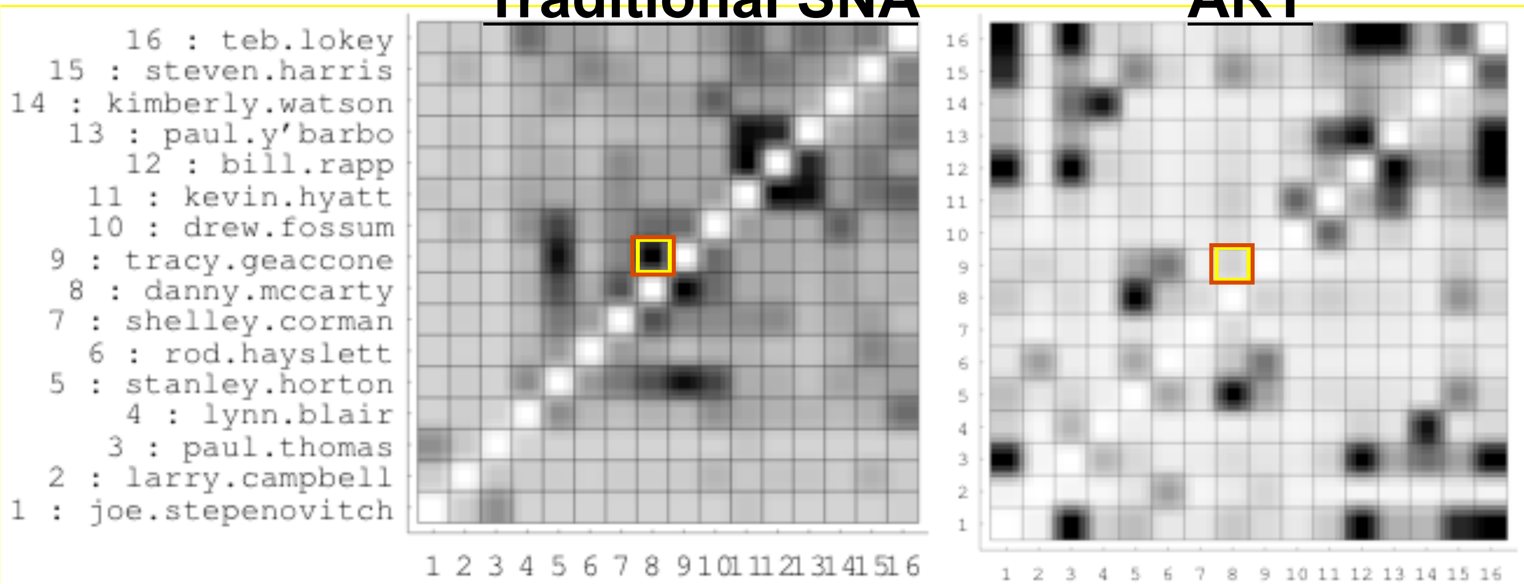
Similarity in recipients they sent email to

Similarity in authored topics, conditioned on recipient

Discovering Role Similarity

Tracy Geaconne ⇔ Dan McCarty

Traditional SNA **ART**



Similar
(send email to same individuals)

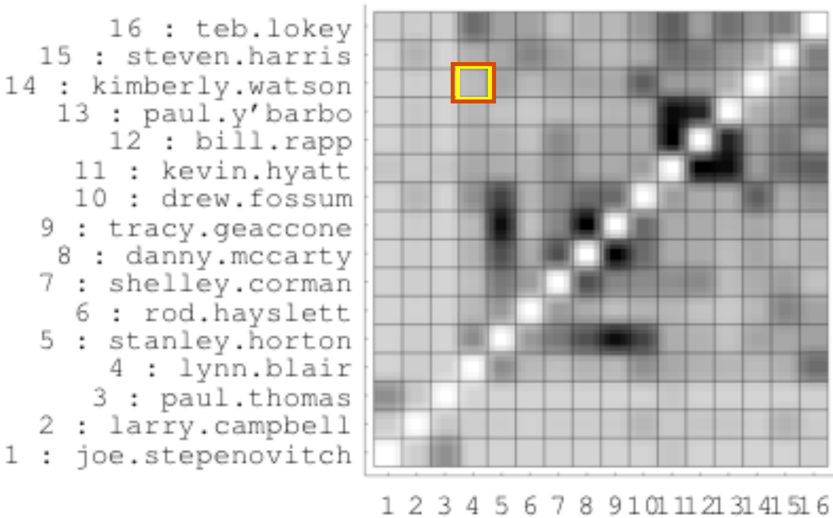
Different
(discuss different topics)

Geaconne = "Secretary"
McCarty = "Vice President"

Discovering Role Similarity

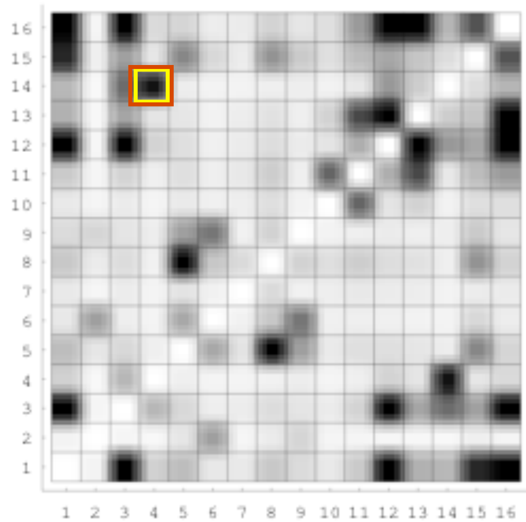
Lynn Blair ⇔ Kimberly Watson

Traditional SNA



Different
(send to different individuals)

ART



Similar
(discuss same topics)

Blair = "Gas pipeline logistics"
Watson = "Pipeline facilities planning"

What you should know

- Unsupervised dimension reduction using all features
  - Principle Components Analysis
    - Minimize reconstruction error
  - Singular Value Decomposition
    - Efficient PCA
  - Independent components analysis
  - Canonical correlation analysis
  - Probabilistic models with latent variables

- Supervised dimension reduction
  - Hidden layers of Neural Networks
    - Most flexible, local minima issues

- LOTS of ways of combining discovery of latent features with classification tasks