Eigenwords

Learning objectives
Distributional similarity
Word embeddings
SVD on words

Lyle Ungar
University of Pennsylvania

Represent each word by its context

I ate ham
You ate cheese
You ate

context

word	Word Before: ate	cheese	ham	I	You	Word After: ate	cheese	ham	I	You
ate	0	0	0	1	2	0	1	1	0	0
cheese	1	0	0	0	0	0	0	0	0	0
ham	1	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	1	0	0	0	0
You	0	0	0	0	0	2	0	0	0	0

Hypothesis: words with similar contexts have similar meanings

Eigenwords

- Project high dimensional context to low dimensional space (SVD/PCA)
- Similar words are close in this low dimensional space

I ate ham
You ate cheese
You ate

	Word Before					Word After				
	ate	cheese	ham	I	You	ate	cheese	ham	I	You
ate	0	0	0	1	2	0	1	1	0	0
cheese	1	0	0	0	0	0	0	0	0	0
ham	1	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	1	0	0	0	0
You	0	0	0	0	0	2	0	0	0	0

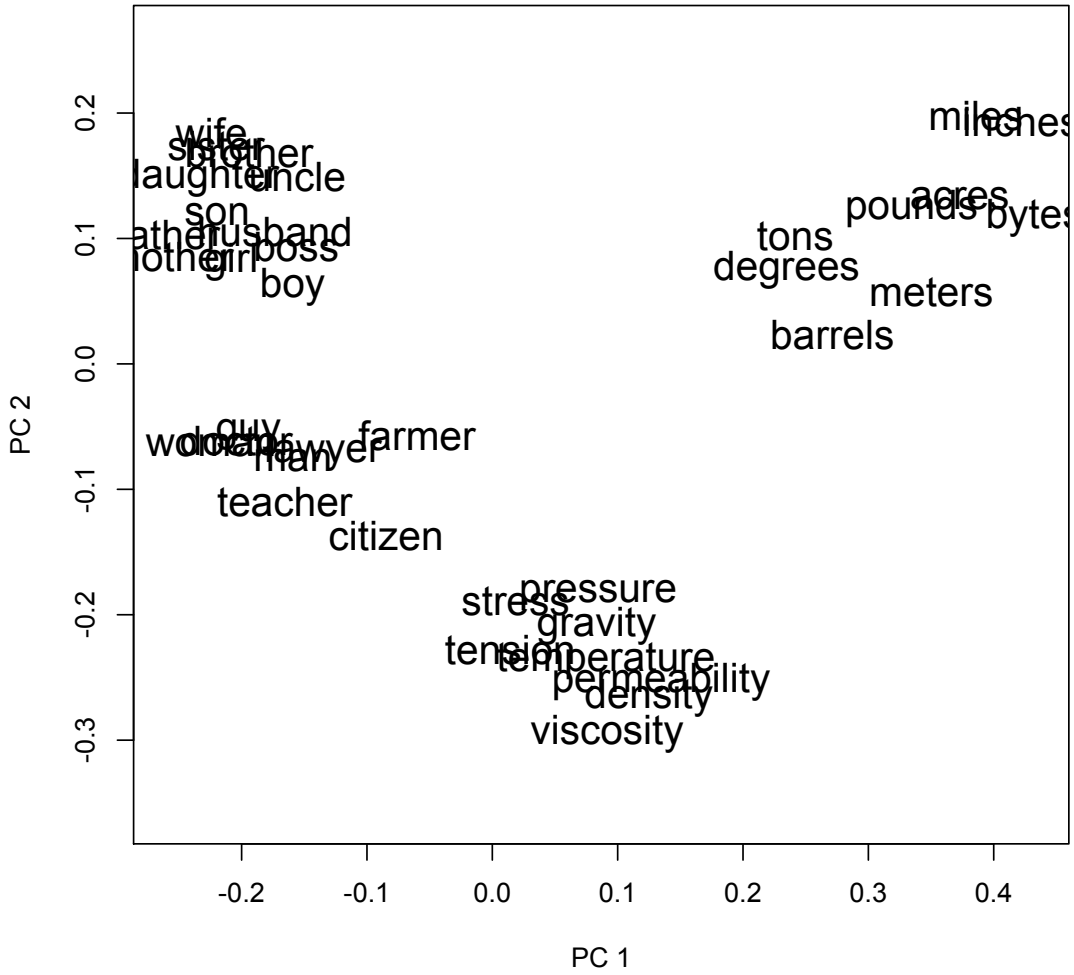
Eigenwords as SVD

- **Left singular vectors are *eigenwords***
  - a vector representing each word – ***"word embeddings"***
- **Right singular vectors times context give *eigentokens***
  - vectors mapping contexts to the latent space

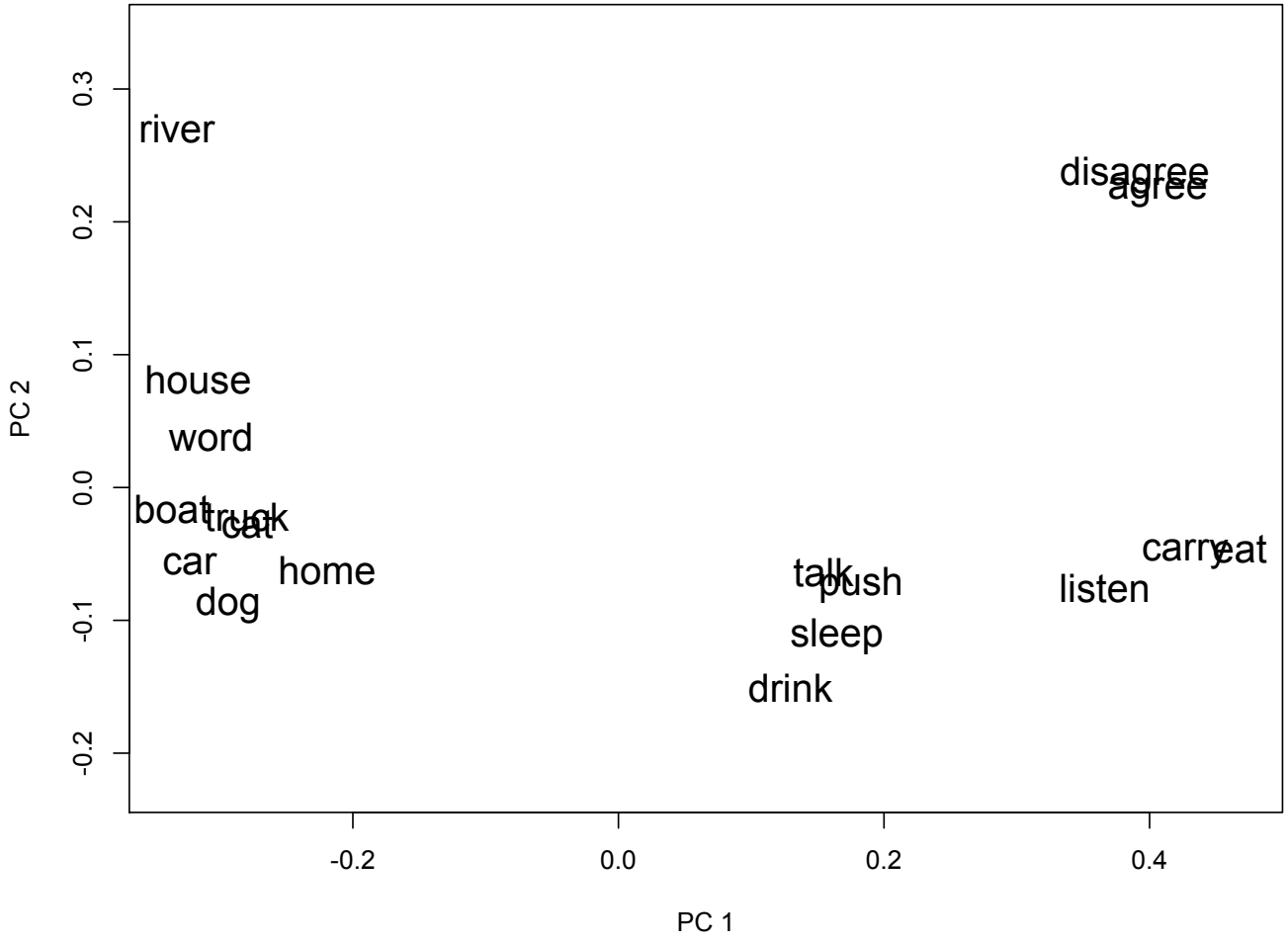
I ate ham
You ate cheese
You ate

	Word Before					Word After				
	ate	cheese	ham	I	You	ate	cheese	ham	I	You
ate	0	0	0	1	2	0	1	1	0	0
cheese	1	0	0	0	0	0	0	0	0	0
ham	1	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	1	0	0	0	0
You	0	0	0	0	0	2	0	0	0	0

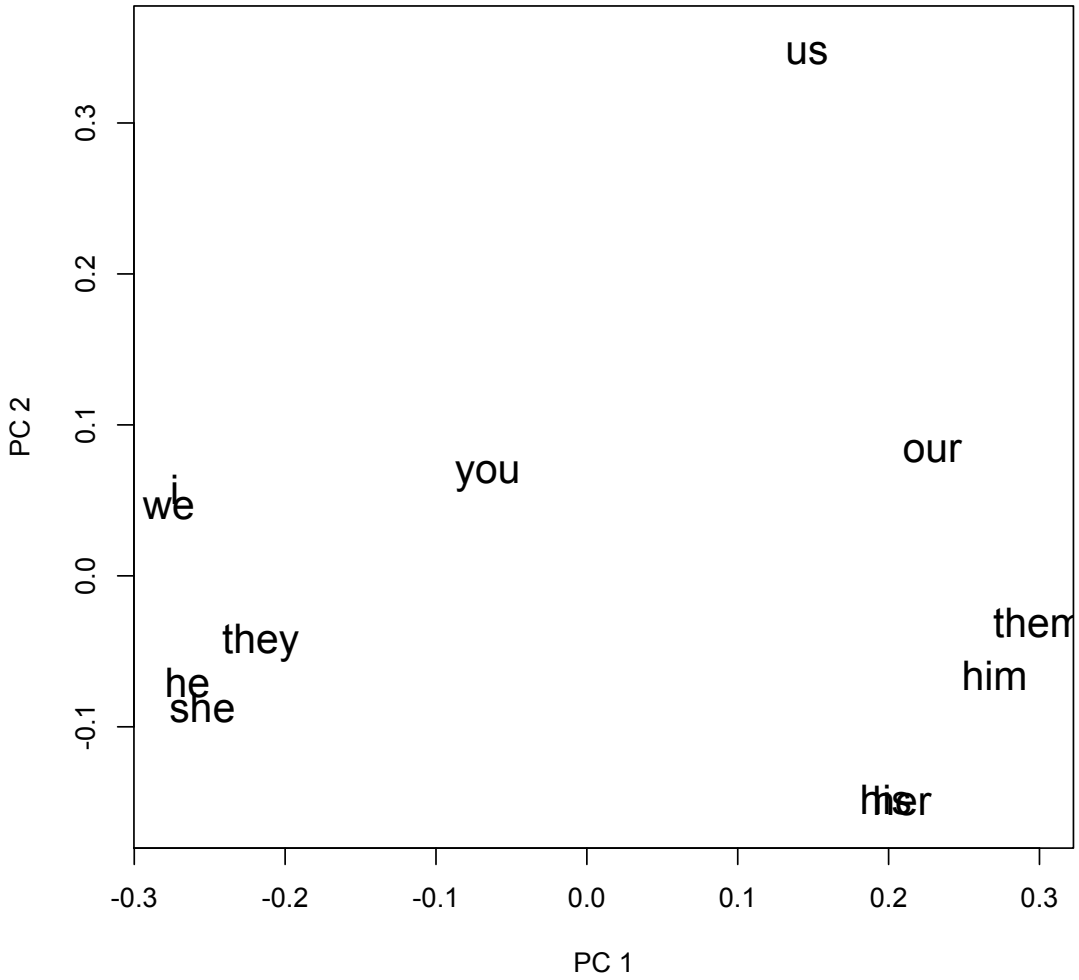
Similar words are close



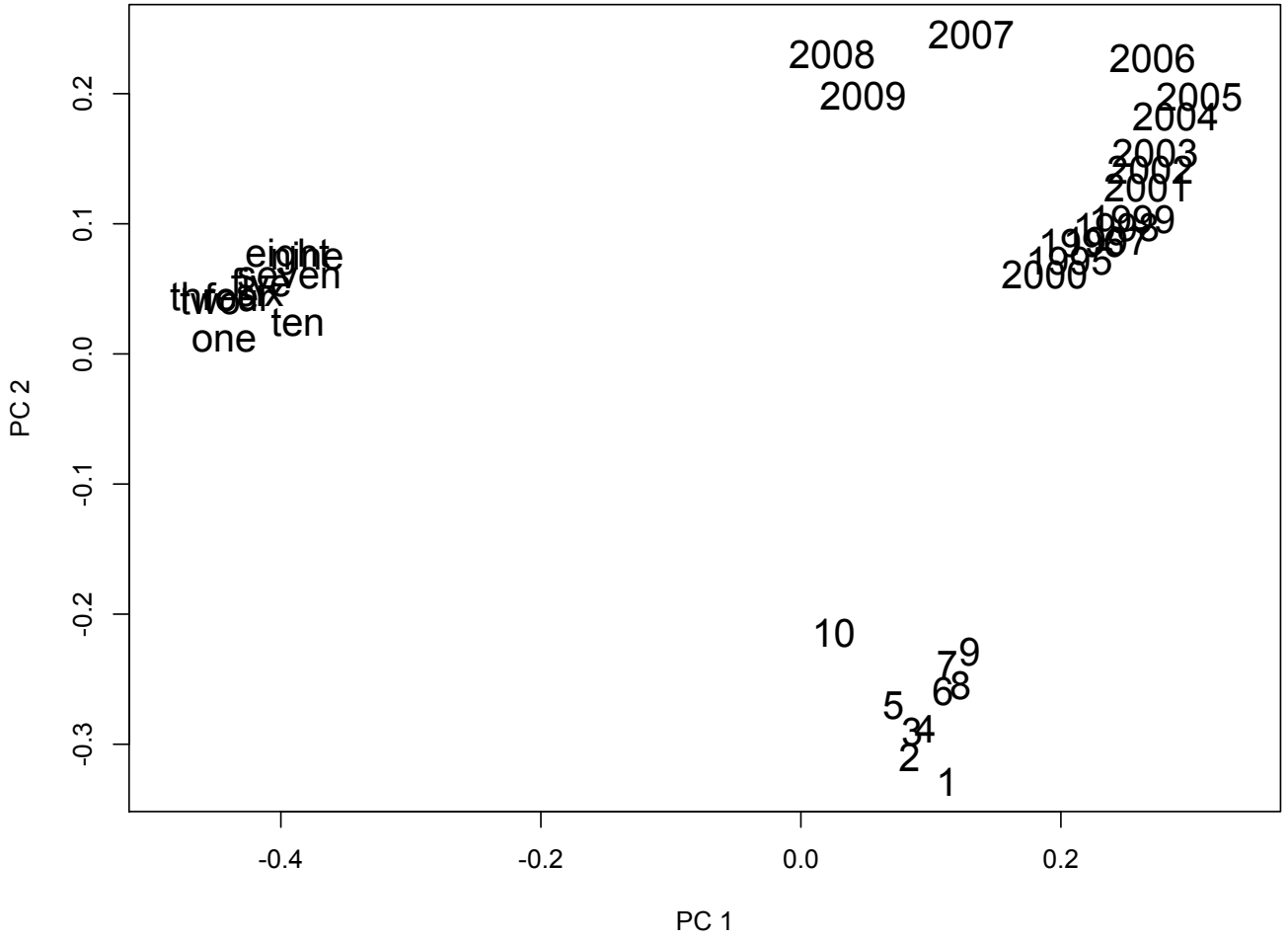
Nouns and verbs



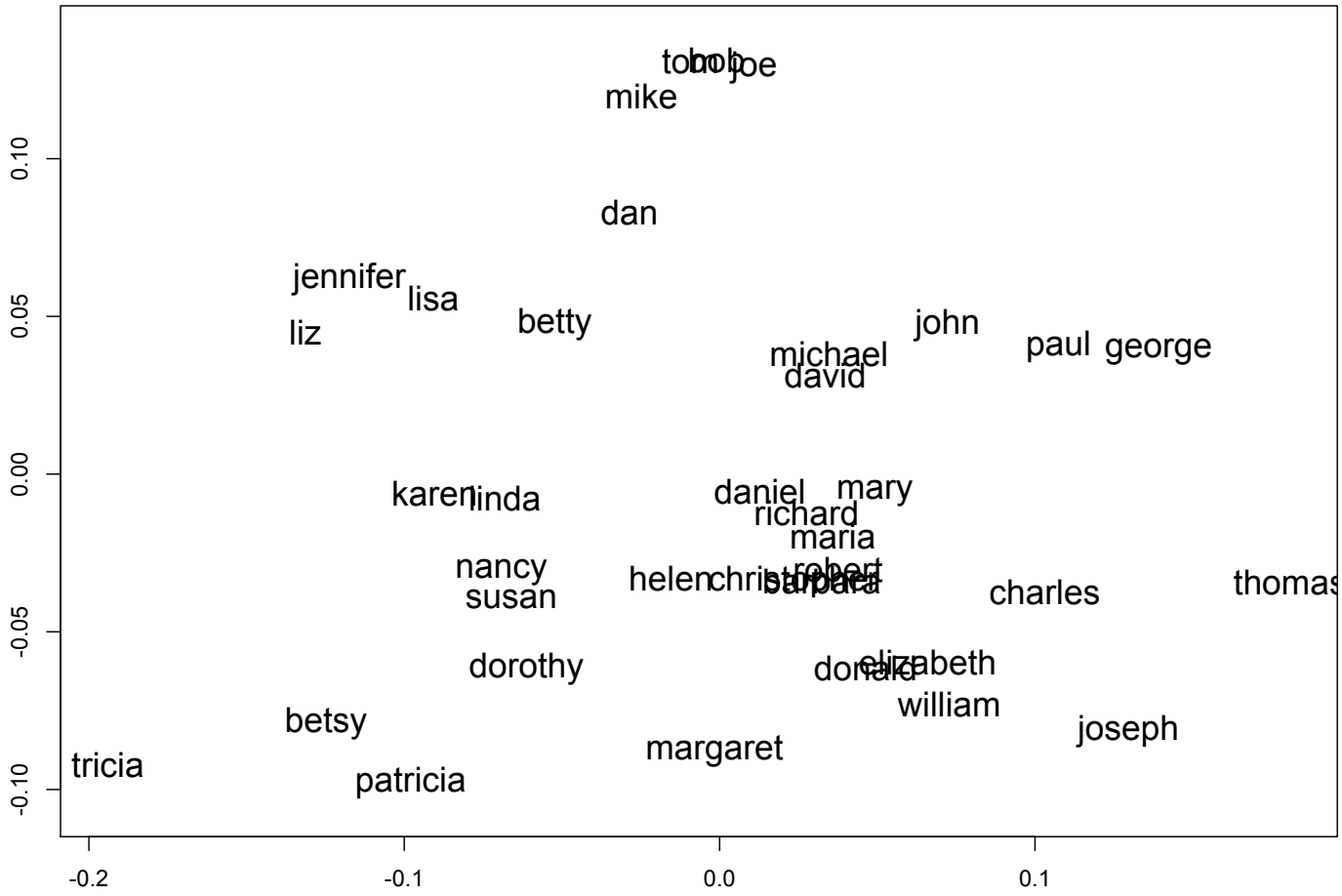
Pronouns


us
you
our
i
we
they
he
she
them
him
his
her
PC 2
PC 1
0.3
0.2
0.1
0.0
-0.1
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3

Numbers


2008
2007
2006
2009
2005
2004
2003
2001
2000
eight
seven
ten
one
10
9
7
8
6
5
4
3
2
1
PC 2
PC 1
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.2
0.0
0.2

Names


tom
bob
joe
mike
dan
jennifer
lisa
liz
betty
john
paul
george
michael
david
karen
linda
daniel
mary
richard
maria
nancy
susan
helen
barbara
charles
thomas
dorothy
donald
elizabeth
william
betsy
joseph
margaret
tricia
patricia
0.10
0.05
0.00
-0.05
-0.10
-0.2
-0.1
0.0
0.1

Word Sense Disambiguation

- **Estimate "state vector" for a word using right singular vectors**
- **Similar meanings will again be close.**
  - The ships dock in the port.
  - The port is loaded onto ships and sent to America

  - The meat is tender.
  - I have tender feelings for her.
  - The company will tender an offer.

Use eigenwords/eigentokens in supervised learning

- **'Similar' words have embeddings that are close**
- **Predict labels for tokens based on their estimated "state vector"**
  - Part of speech
  - Named entity type (person, place, thing…)
  - Word sense ("meaning") disambiguation
- **Or embed sentences**

Word2vec

- ***Word embeddings*, often found by deep learning, are very popular now**
  - Word2Vec has similar performance to the simpler eigenwords
- **Deep learning versions: BERT, ELMo work better**
  - To be covered later