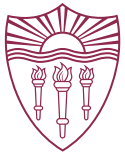
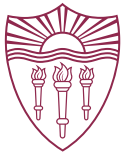


TOPIC MODELING AND WORD EMBEDDINGS

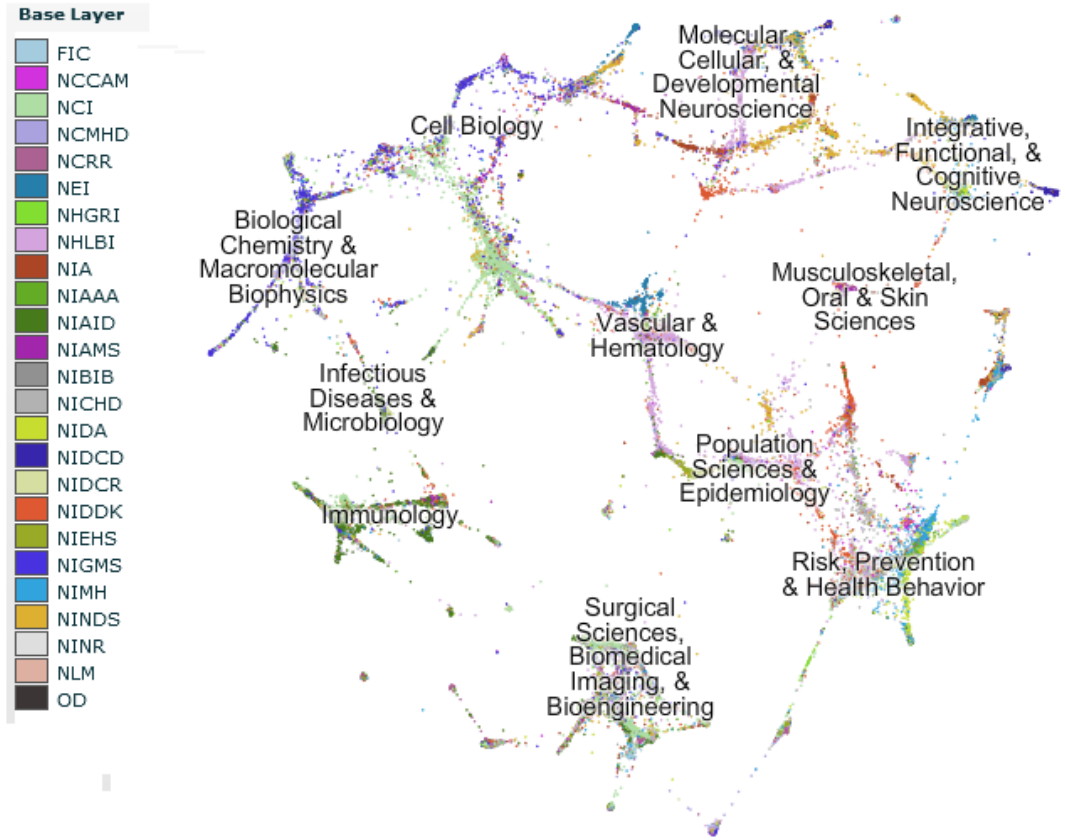
Kristina Lerman
USC Information Sciences Institute
DSCI 552 – Spring 2021
March 8, 2021

Why text analysis



- Volume of text data is growing exponentially, necessitating methods for automatically organizing, understanding, searching and summarizing them
  - Uncover hidden topical patterns in collections.
  - Annotate documents according to topics.
  - Using annotations to organize, summarize and search.

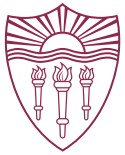
Topic Modeling



This is a topic map of all grants awarded by the National Institutes of Health in 2011. There are approximately 80,000 grants, each represented as a dot, color-coded by NIH Institute. Grants are located nearby one another based on shared topical focus. Labels are placed automatically, based on NIH Review Study Sections or other information obtained from the underlying grants.

NIH Grants Topic Map 2011
NIH Map Viewer (https://app.nihmaps.org)

Brief history of text analysis



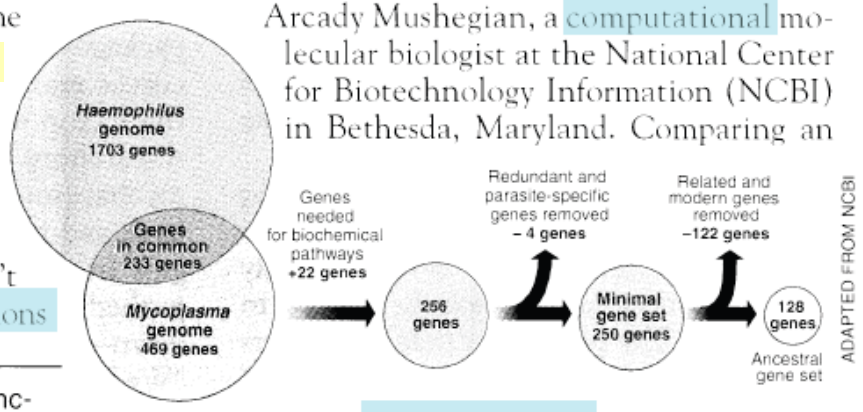
- 1960s
  - Electronic documents come online
  - Vector space models (Salton)
  - ‘bag of words’, tf-idf
- 1990s
  - Mathematical analysis tools become widely available
  - Latent semantic indexing (LSI)
  - Singular value decomposition (SVD, PCA)
- 2000s
  - Probabilistic topic modeling (LDA)
  - Probabilistic matrix factorization (PMF)
- 2010s
  - Text embeddings (word2vec)
  - Pre-trained language models (BERT)
- The present
  - Language generation (GPT-3)

Vector space model

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

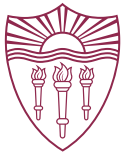
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Term frequency

- genes 5
- organism 3
- survive 1
- life 1
- computer 1
- organisms 1
- genomes 2
- predictions 1
- genetic 1
- numbers 1
- sequenced 1
- genome 2
- computational 1
- …

Vector space models: reducing noise



original

- genes 5
- organism 3
- survive 1
- life 1
- computer 1
- organisms 1
- genomes 2
- predictions 1
- genetic 1
- numbers 1
- sequenced 1
- genome 2
- computational 1

stem words

- gene 6
- organism 4
- survive 1
- life 1
- comput 2
- predictions 1
- numbers 1
- sequenced 1
- genome 4

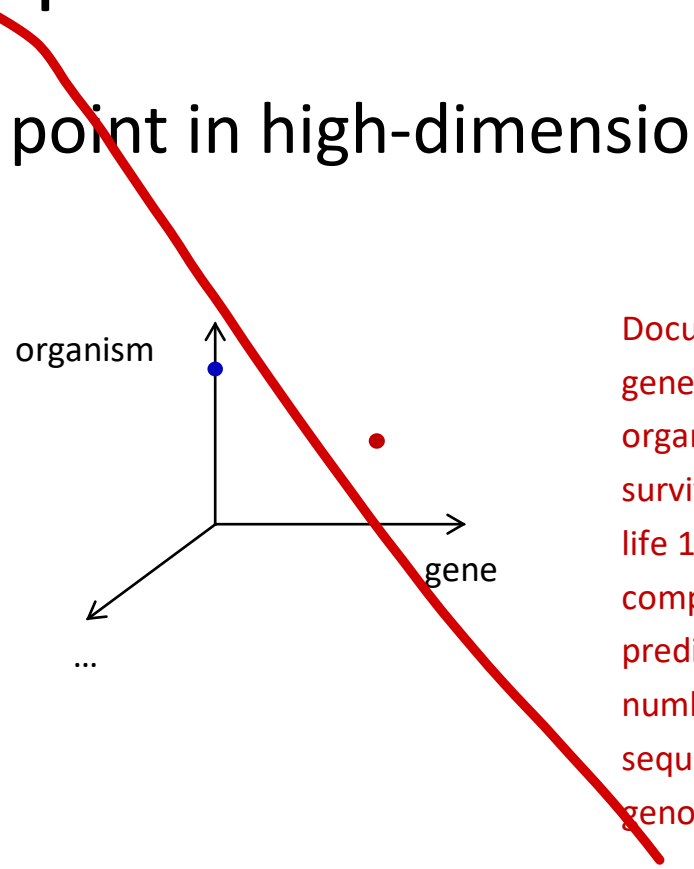
remove stopwords

- and
- or
- but
- also
- to
- too
- as
- can
- I
- you
- he
- she
- …

Vector space model

- Each document is a point in high-dimensional space

Document 2
gene 0
organism 6
survive 1
life 1
comput 2
predictions 1
numbers 1
sequenced 1
genome 4
…

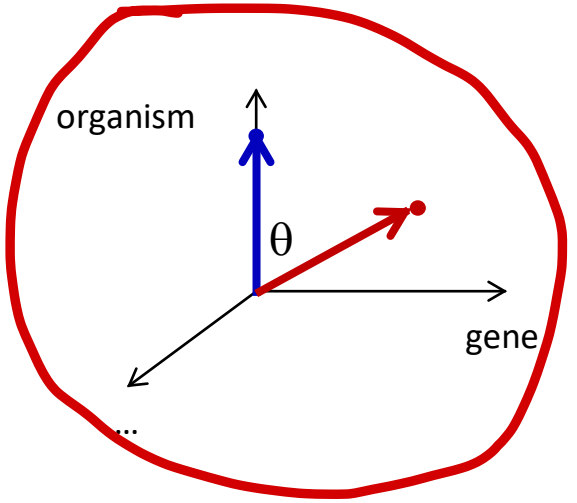


Document 1
gene 6
organism 4
survive 1
life 1
comput 2
predictions 1
numbers 1
sequenced 1
genome 4
…

Vector space model

- Each document is a point in high-dimensional space

Document 2
gene 0
organism 6
survive 1
life 1
comput 2
predictions 1
numbers 1
sequenced 1
genome 4
…

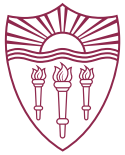


Document 1
gene 6
organism 4
survive 1
life 1
comput 2
predictions 1
numbers 1
sequenced 1
genome 4
…

- Compare two documents: similarity ~ cos(θ)

Improving the vector space model

提高

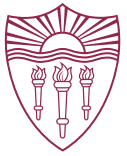


- Use tf-idf, instead of term frequency (tf), in the document vector
  - Term frequency * inverse document frequency
  - E.g.,
    - 'computer' occurs 3 times in a document, but it is present in 80% of documents → tf-idf score 'computer' is 3*1/.8=3.75
    - 'gene' occurs 2 times in a document, but it is present in 20% of documents → tf-idf score of 'gene' is 2*1/.2=10

Some problems with vector space model

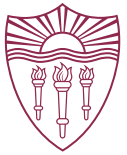
- Synonymy
  - Unique term corresponds to a dimension in term space
  - Synonyms ('kid' and 'child') are different dimensions
- Polysemy
  - Different meanings of the same term improperly confused
  - E.g., document about river 'banks' will be improperly judged to be similar to a document about financial 'banks'

Latent Semantic Indexing (LSI)



- Identifies subspace of tf-idf that captures most of the variance in a corpus
  - Need a smaller subspace to represent document corpus
  - This subspace captures topics that exist in a corpus
    - Topic = set of related words
- Handles polysemy and synonymy
  - Synonyms will belong to the same topic since they may co-occur with the same related words

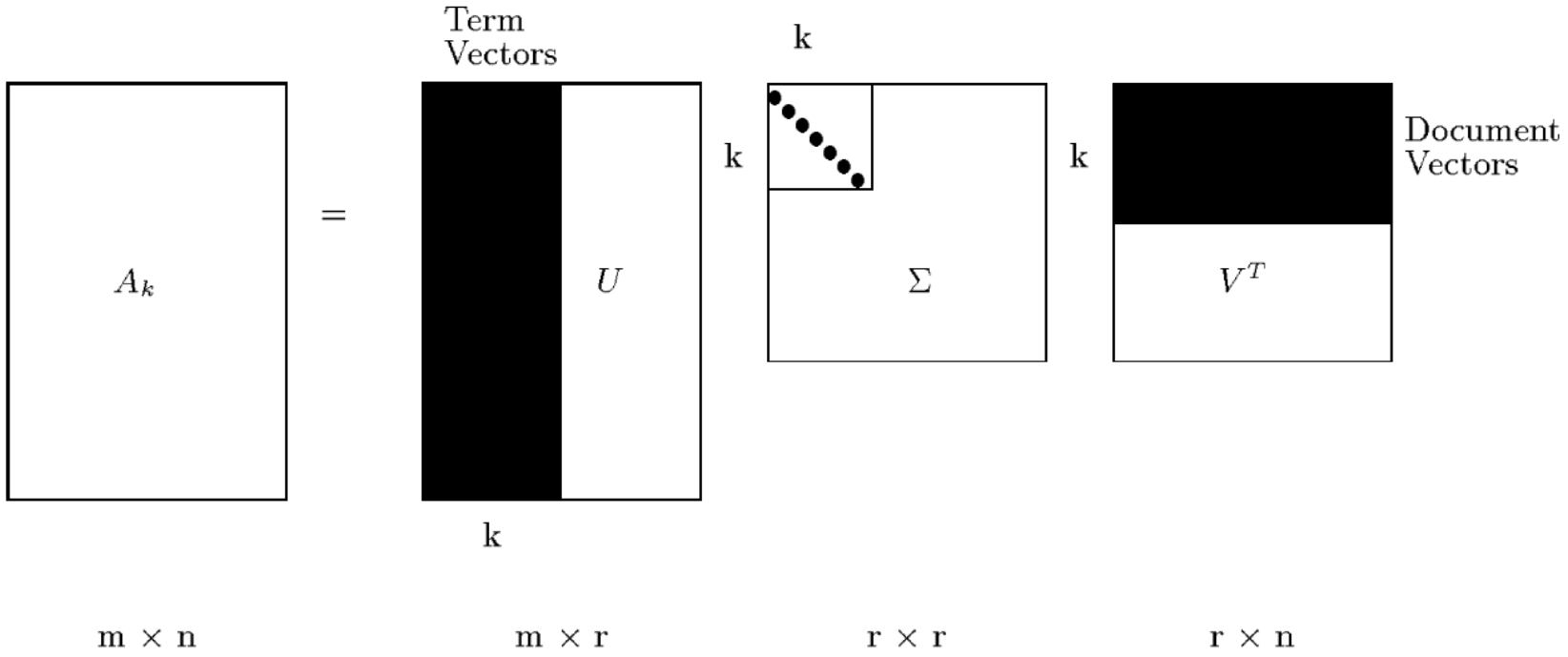
LSI, the Method 步骤 LSI



1. Encode documents in a (document-term) matrix A
2. Factorize
   - Decompose A by Singular Value Decomposition (SVD)
   - Linear algebra
3. Approximate A using truncated SVD
   - Captures the most important relationships in A
   - Ignores other relationships
   - Rebuild the matrix A using just the important relationships
4. Measure relatedness
   - Cosine of latent factors

$$A = \begin{array}{l|cccccc} & d_1 & d_2 & d_3 & d_4 & d_5 & d_6 \\ \hline \text{cosmonaut} & 1 & 0 & 1 & 0 & 0 & 0 \\ \text{astronaut} & 0 & 1 & 0 & 0 & 0 & 0 \\ \text{moon} & 1 & 1 & 0 & 0 & 0 & 0 \\ \text{car} & 1 & 0 & 0 & 1 & 1 & 0 \\ \text{truck} & 0 & 0 & 0 & 1 & 0 & 1 \end{array}$$

LSI, the Method (cont.)



Each row and column of *A* gets mapped into the *k*-dimensional LSI space, by the SVD.

Singular value decomposition

- SVD- Singular value decomposition

http://en.wikipedia.org/wiki/Singular_value_decomposition

Singular Value Decomposition (SVD):

$$A = U \Lambda V^T$$

where $A \in \Re^{m \times n}$, $U \in \Re^{m \times m}$ and $V \in \Re^{n \times n}$ are orthogonal,

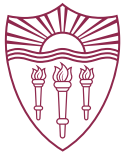
and $\Lambda = diag(\lambda_1, \cdots, \lambda_r)$ is diagonal with $\lambda_1 \geq \cdots \geq \lambda_r$

and $r = \min(m,n)$.

$AA^T = U \Sigma \Sigma^T U^T$: U forms the eigenvectors of AA^T.

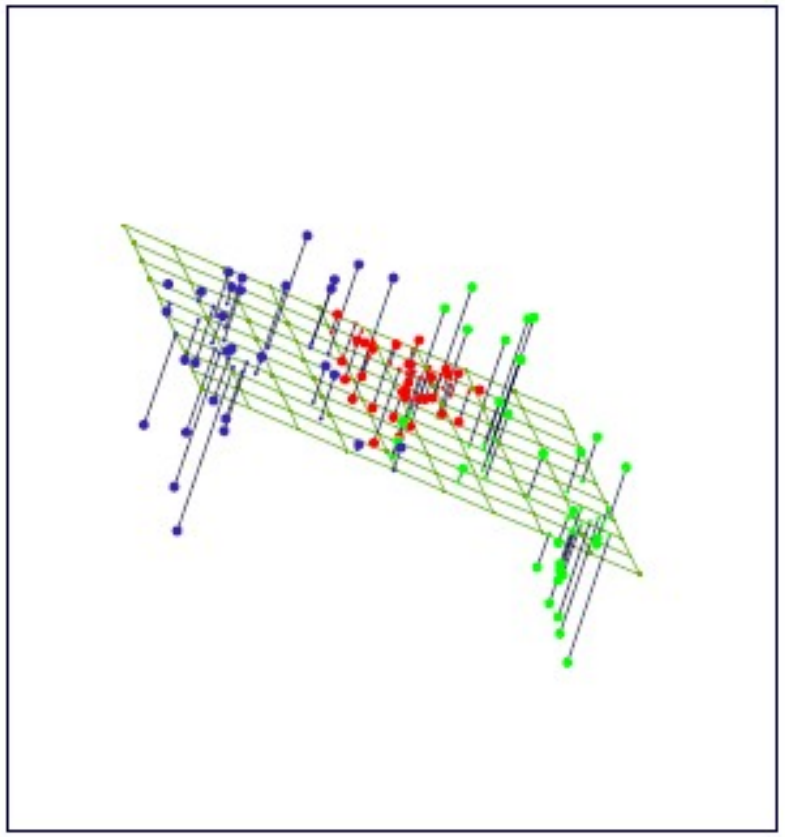
$A^T A = V \Sigma^T \Sigma V^T$: V forms the eigenvectors of $A^T A$.

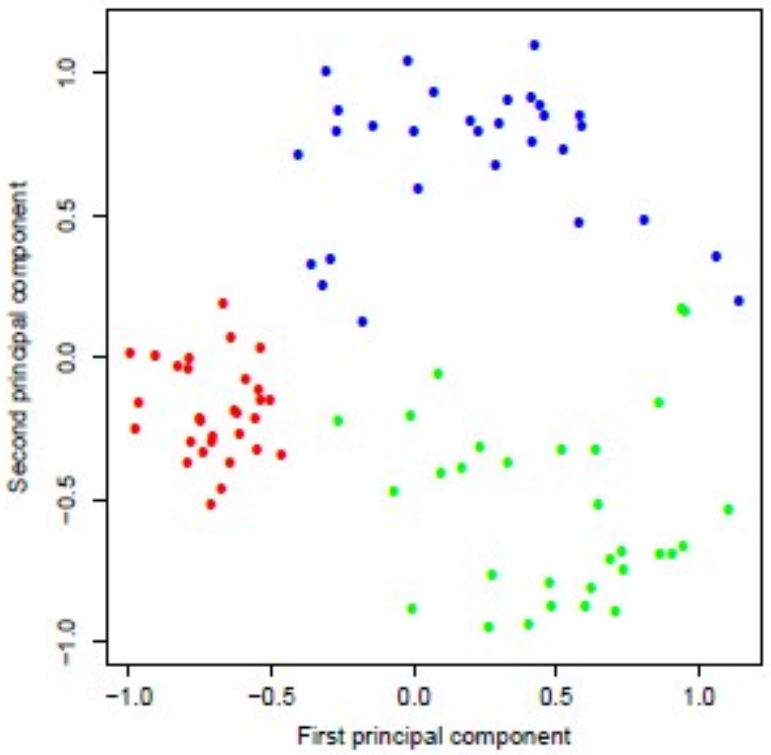
Lower rank decomposition

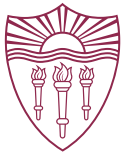


- Usually, rank of the matrix A is small: r<<min(m,n).
  - Only a few of the largest eigenvectors (those associated with the largest eigenvalues λ) matter
  - These r eigenvectors define a lower dimensional subspace that captures most important characteristics of the document corpus
  - All operations (document comparison, similar) can be done in this reduced-dimension subspace

Low rank approximation







Probabilistic Modeling

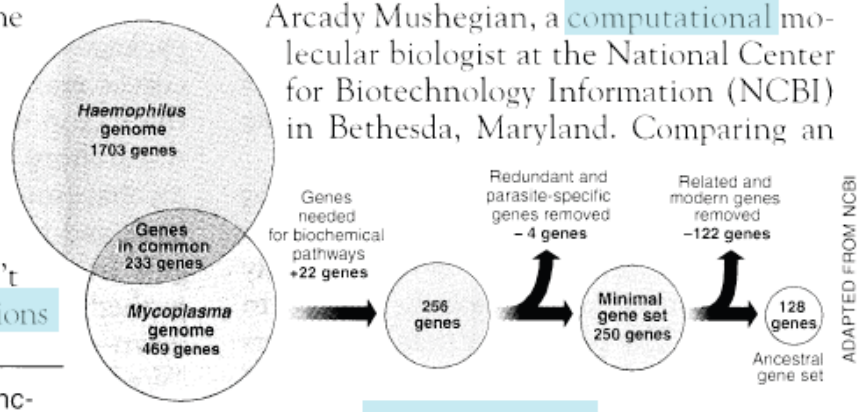
- Generative probabilistic modeling
  - Treats data as observations
  - Contains hidden variables
  - Hidden variables reflect the themes that pervade a corpus of documents
- Infer hidden thematic structure
  - Analyze words in the documents
  - Discover topics in the corpus
    - A topic is a distribution over words
  - Large reduction in description length
    - Few topics are needed to represent themes in a document corpus – about 100

LDA – Latent Dirichlet Allocation (Blei 2003)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

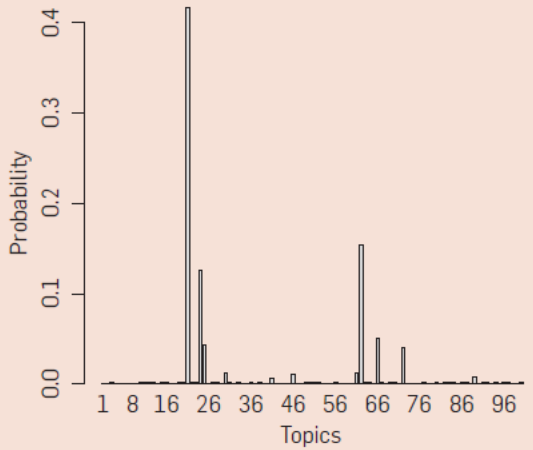
272 • 24 MAY 1996

Intuition: Documents have multiple topics

Topics

- A topic is a distribution over words
- A document is a distribution over topics
- A word in a document is drawn from one of those topics

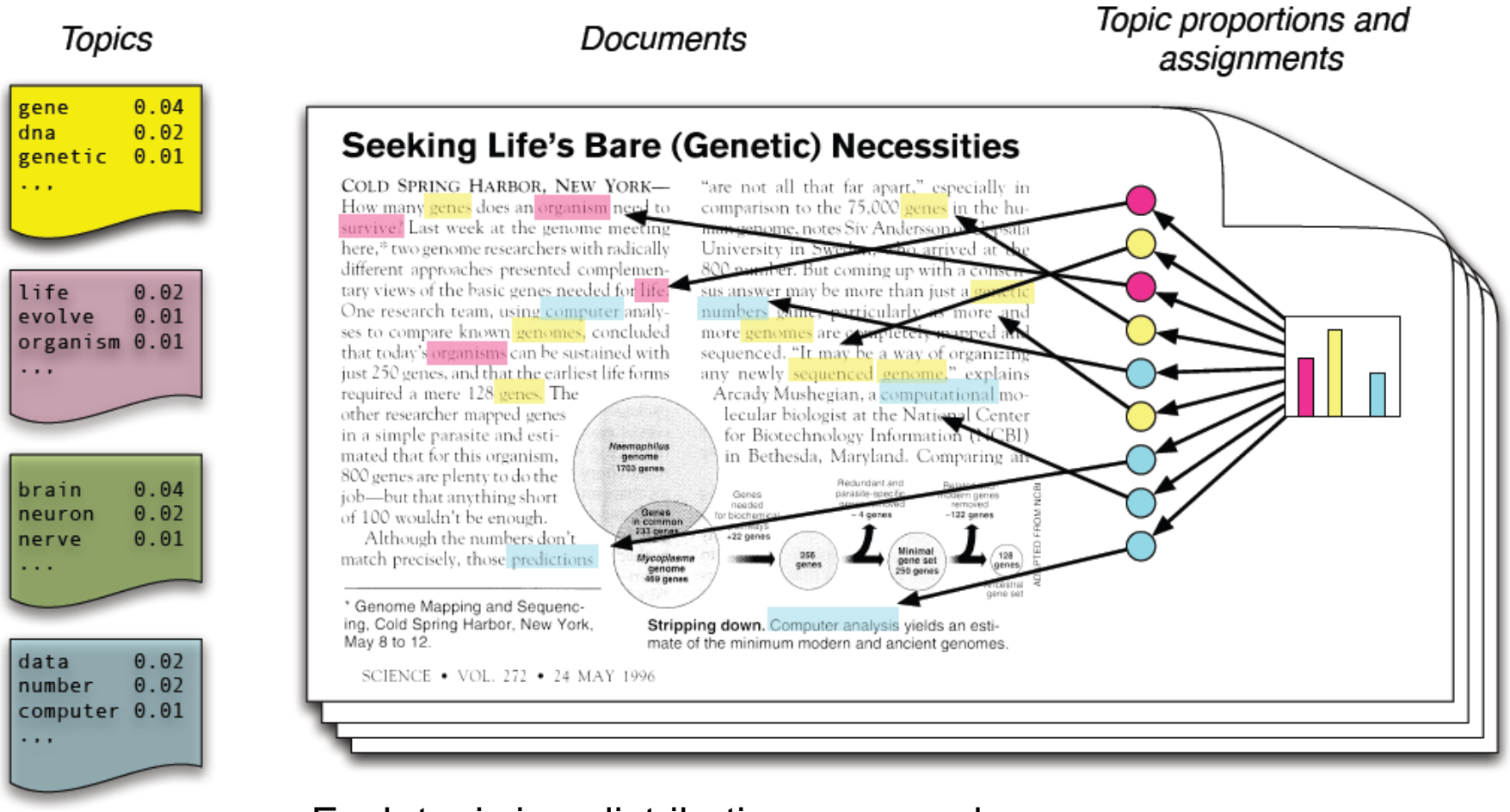
Document



Topics

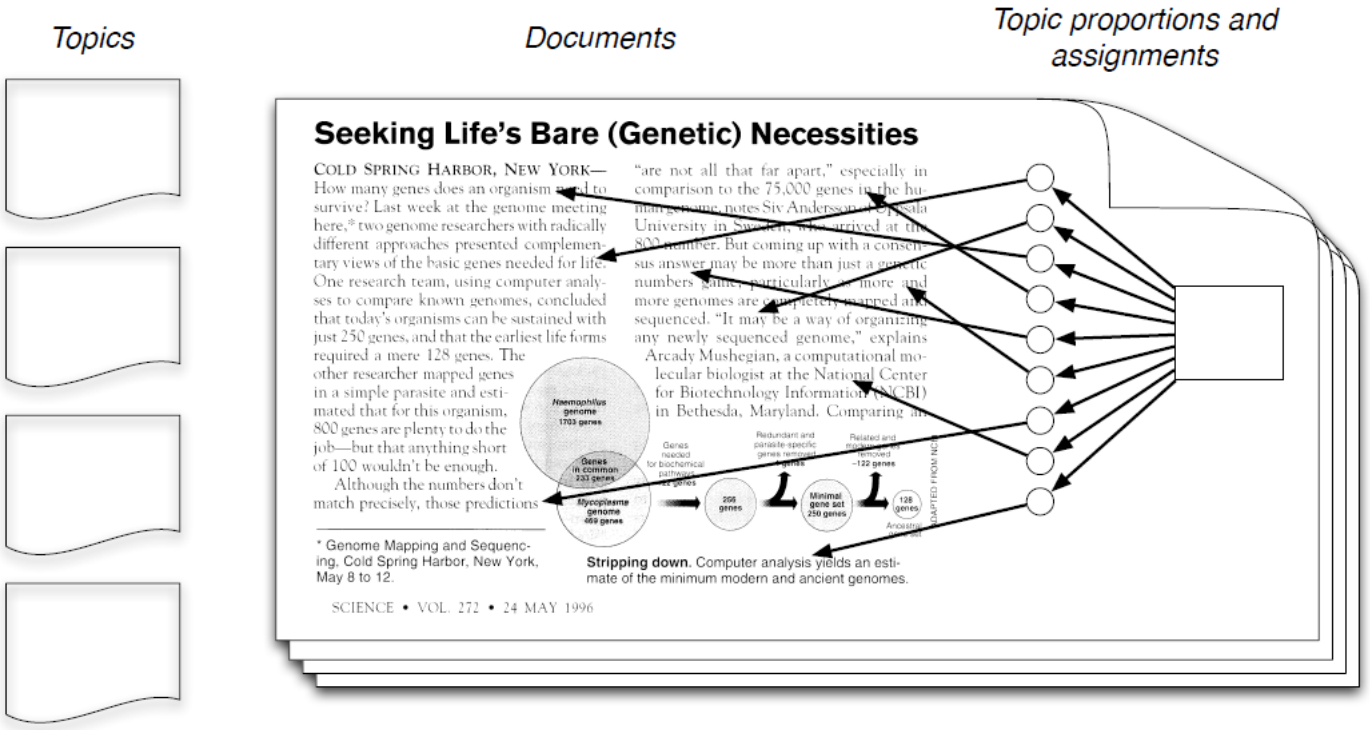
"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Generative Model of LDA



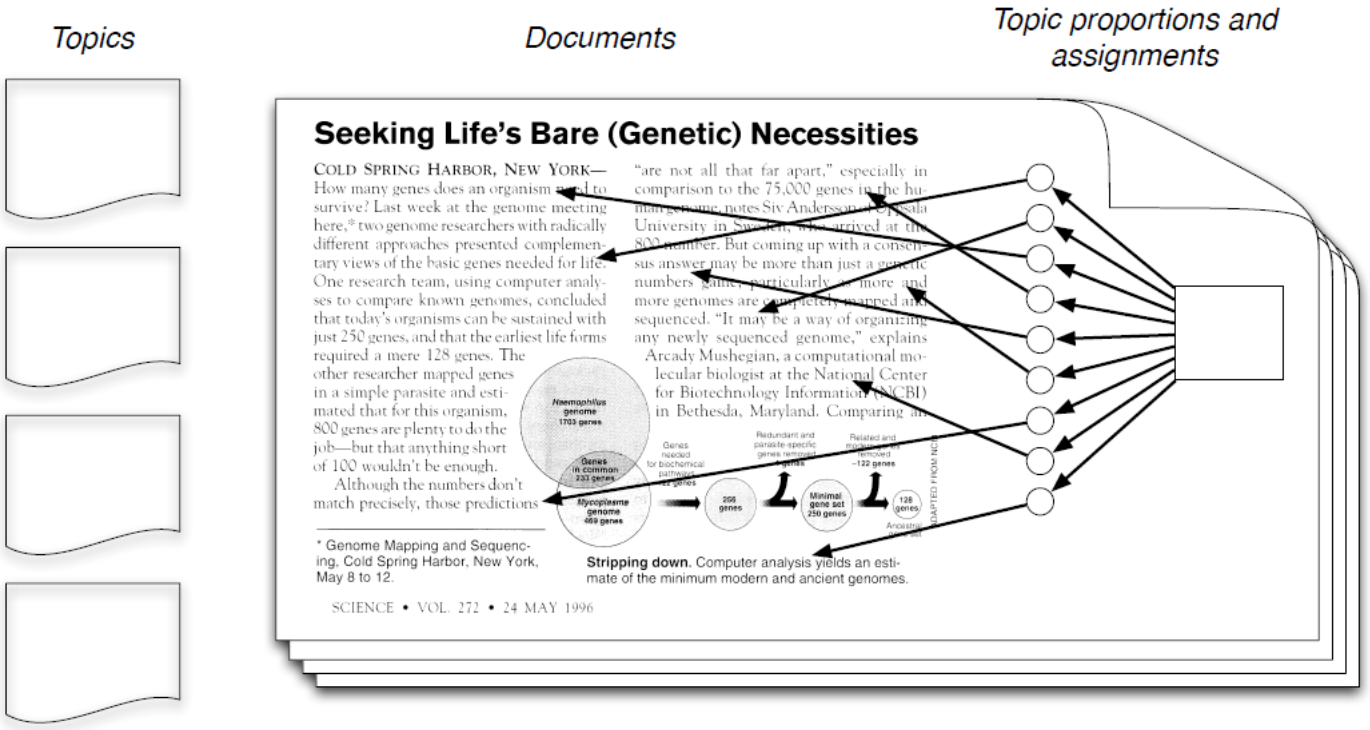
- Each topic is a distribution over words
- Each document is a mixture of corpus-wide topics
- Each word is drawn from one of those topics

LDA inference



- We observe only **documents**
- The rest of the structure are **hidden variables**

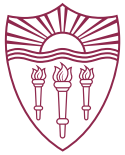
LDA inference



- Our goal is to infer hidden variables
- Compute their distribution conditioned on the documents

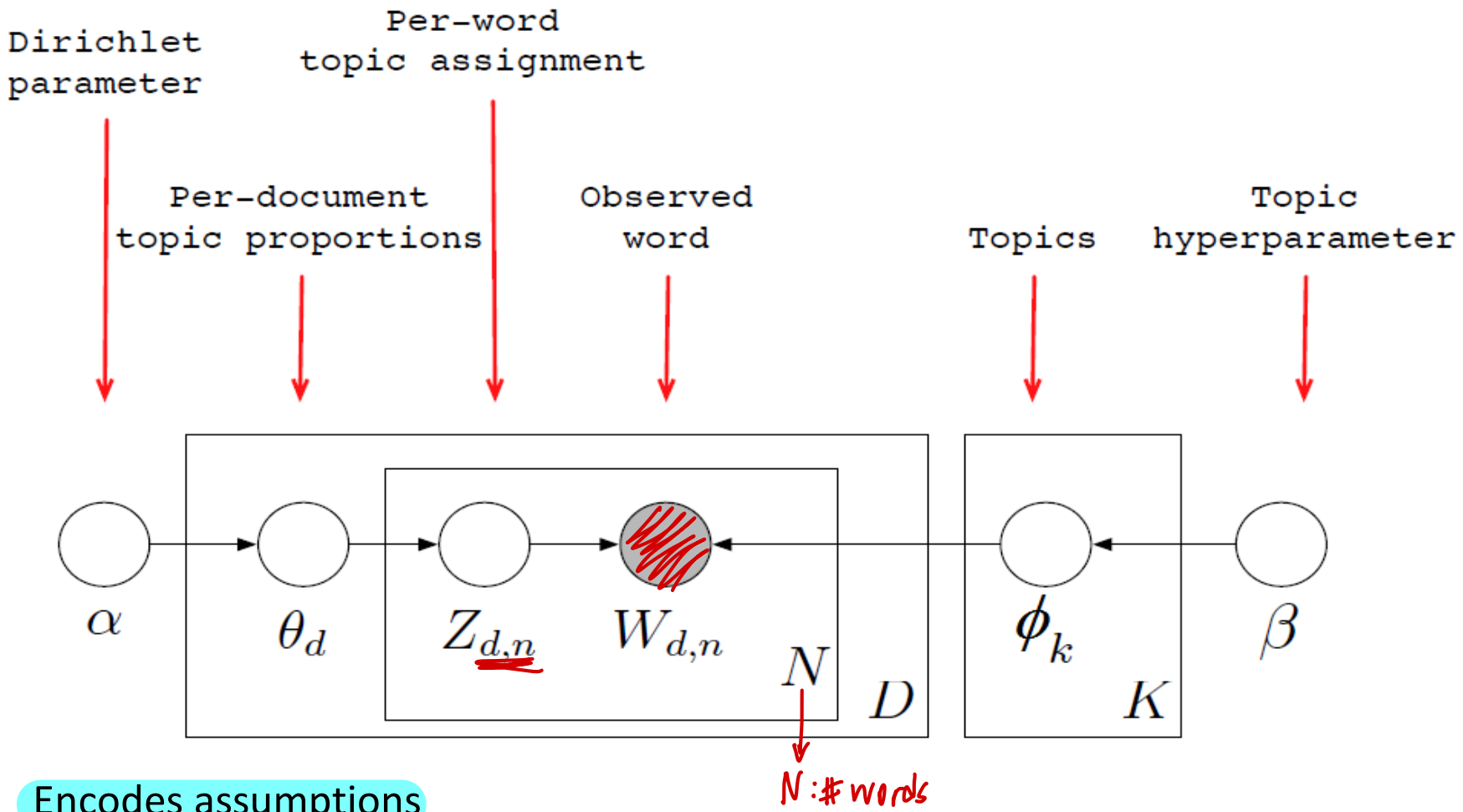
  p(topic, proportions, assignments | documents)

Posterior Distribution



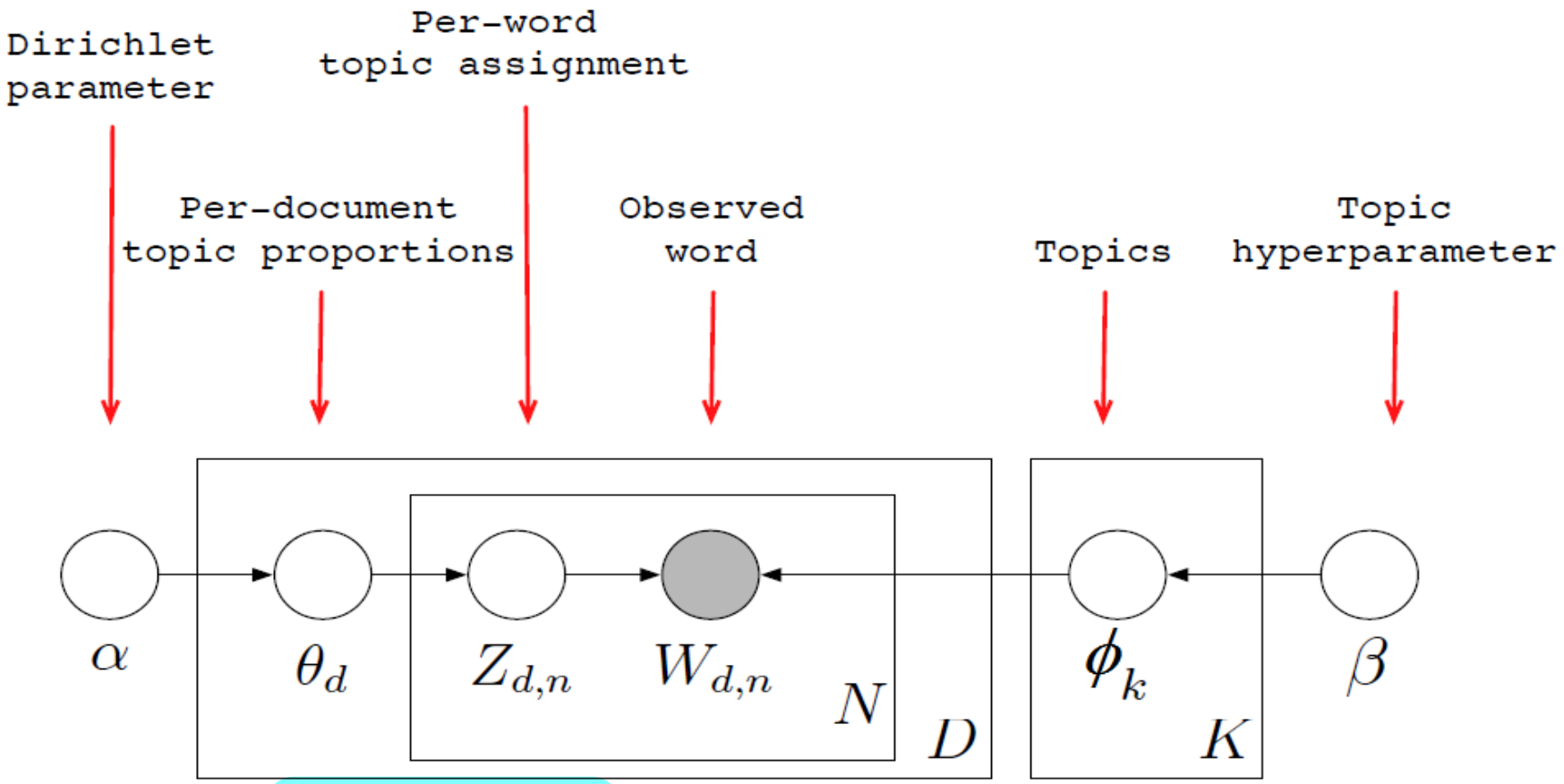
- Only documents are observable.
- Infer underlying topic structure.
  - Topics that generated the documents.
  - For each document, distribution of topics.
  - For each word, which topic generated the word.
- Algorithmic challenge: Finding the conditional distribution of all the latent variables, given the observation.

LDA as Graphical Model



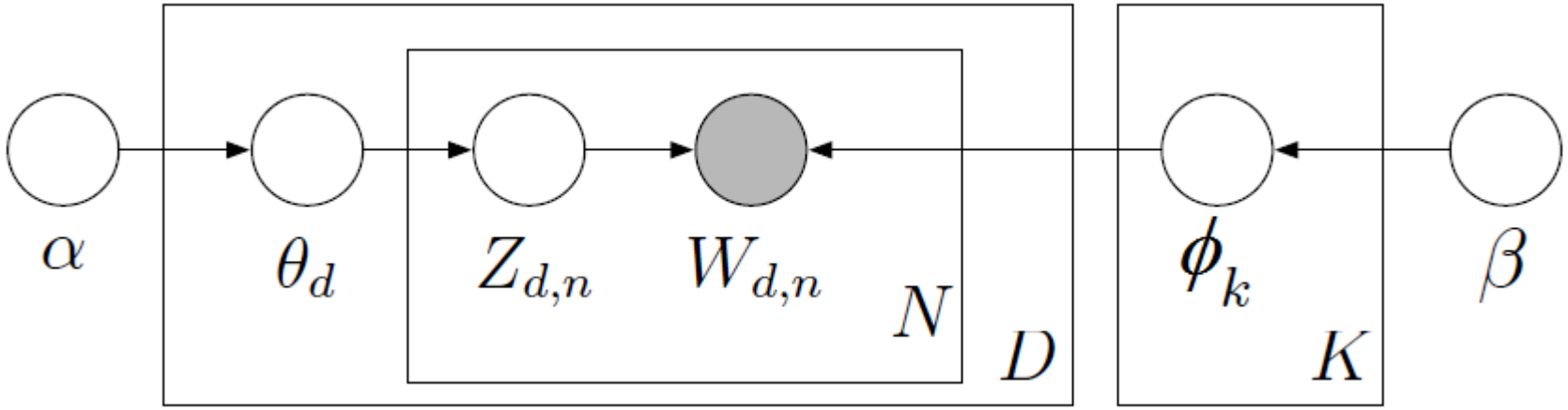
- Encodes assumptions
- Defines a factorization of the joint distribution

LDA as Graphical Model



- Nodes are random variables; edges indicate dependence
- Shaded nodes are observed; unshaded nodes are hidden
- Plates indicate replicated variables

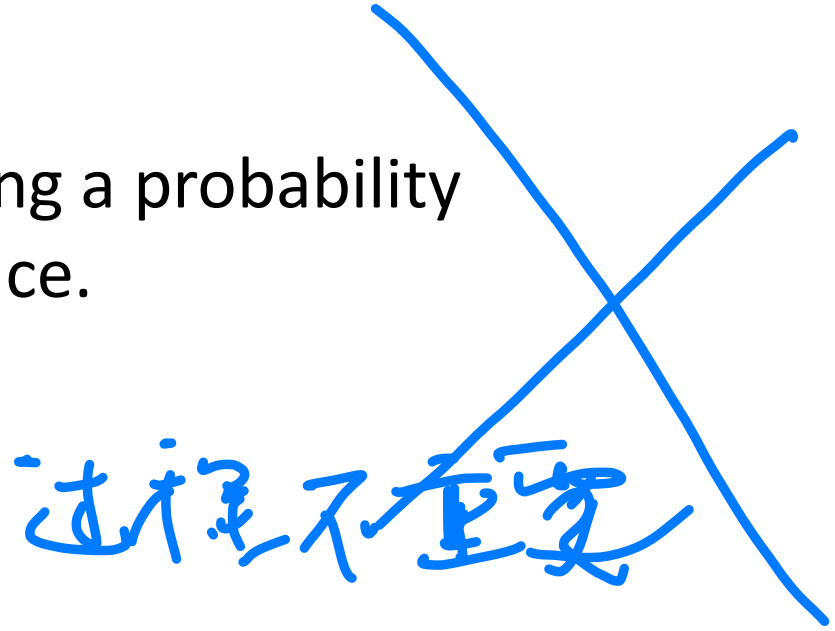
Posterior Distribution



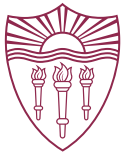
- This joint defines a posterior ***p(θ, z, β|W)***:
- From a collection of documents ***W***, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distribution ϕ_k

Posterior Distribution

- Evaluate ***p(z|W)***: posterior distribution over the assignment of words to topic.
- θ and ϕ can be estimated.
- Computing ***p(z|W)*** involves evaluating a probability distribution over a large discrete space.



Approximate posterior inference algorithms



- Mean field variational methods
- Expectation propagation
- Gibbs sampling
- Distributed sampling
- …
- Efficient packages for solving this problem

Example

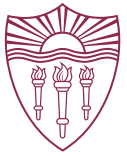
- Data: collection of *Science* articles from 1990-2000
  - 17K documents
  - 11M words
  - 20K unique words (stop words and rare words removed)
- Model: 100-topic LDA


1
dna
gene
sequence
genes
sequences
human
genome
genetic
analysis
two
2
protein
cell
cells
proteins
receptor
fig
binding
activity
activation
kinase
3
water
climate
atmospheric
temperature
global
surface
ocean
carbon
atmosphere
changes
4
says
researchers
new
university
just
science
like
work
first
years
5
mantle
high
earth
pressure
seismic
crust
temperature
earths
lower
earthquakes
6
end
article
start
science
readers
service
news
card
circle
letters
7
time
data
two
model
fig
system
number
different
results
rate
8
materials
surface
high
structure
temperature
molecules
chemical
molecular
fig
university
9
dna
rna
transcription
protein
site
binding
sequence
proteins
specific
sequences
10
disease
cancer
patients
human
gene
medical
studies
drug
normal
drugs
11
years
million
ago
age
university
north
early
fig
evidence
record
12
species
evolution
population
evolutionary
university
populations
natural
studies
genetic
biology
13
protein
structure
proteins
two
amino
binding
acid
residues
molecular
structural
14
cells
cell
virus
hiv
infection
immune
human
antigen
infected
viral
15
space
solar
observations
earth
stars
university
mass
sun
astronomers
telescope
16
fax
manager
science
aaas
advertising
sales
member
recruitment
associate
washington
17
cells
cell
gene
genes
expression
development
mutant
mice
fig
biology
18
energy
electron
state
light
quantum
physics
electrons
high
laser
magnetic
19
research
science
national
scientific
scientists
new
states
university
united
health
20
neurons
brain
cells
activity
fig
channels
university
cortex
neuronal
visual

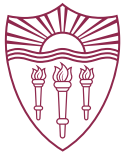
Tools

- Topic modeling
  1. Blei's LDA w/ "variational method" (http://cran.r-project.org/web/packages/lda/) or
  2. "Gibbs sampling method" (https://code.google.com/p/plda/ and http://gibbslda.sourceforge.net/)

How useful are learned topic models



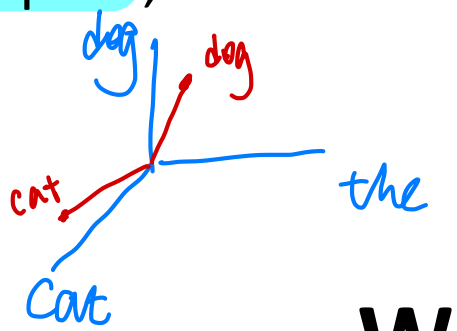
- Model evaluation
  - How well do learned topics describe unseen (test) documents
  - How well it can be used for personalization
- Model checking
  - Given a new corpus of documents, what model should be used? How many topics?
- Visualization and user interfaces
- Topic models for exploratory data analysis

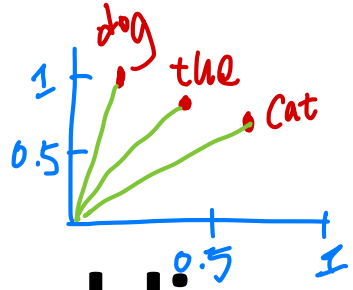


WORD EMBEDDINGS

Word Embedding

- Words from the vocabulary are mapped to vectors of real numbers:
  - In a low dimensional space, relative to the vocabulary size.
  - "continuous space".





Old:

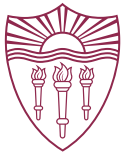
of unique →

the: <1, 0, 0>
cat: <0, 1, 0>
dog: <0, 0, 1>

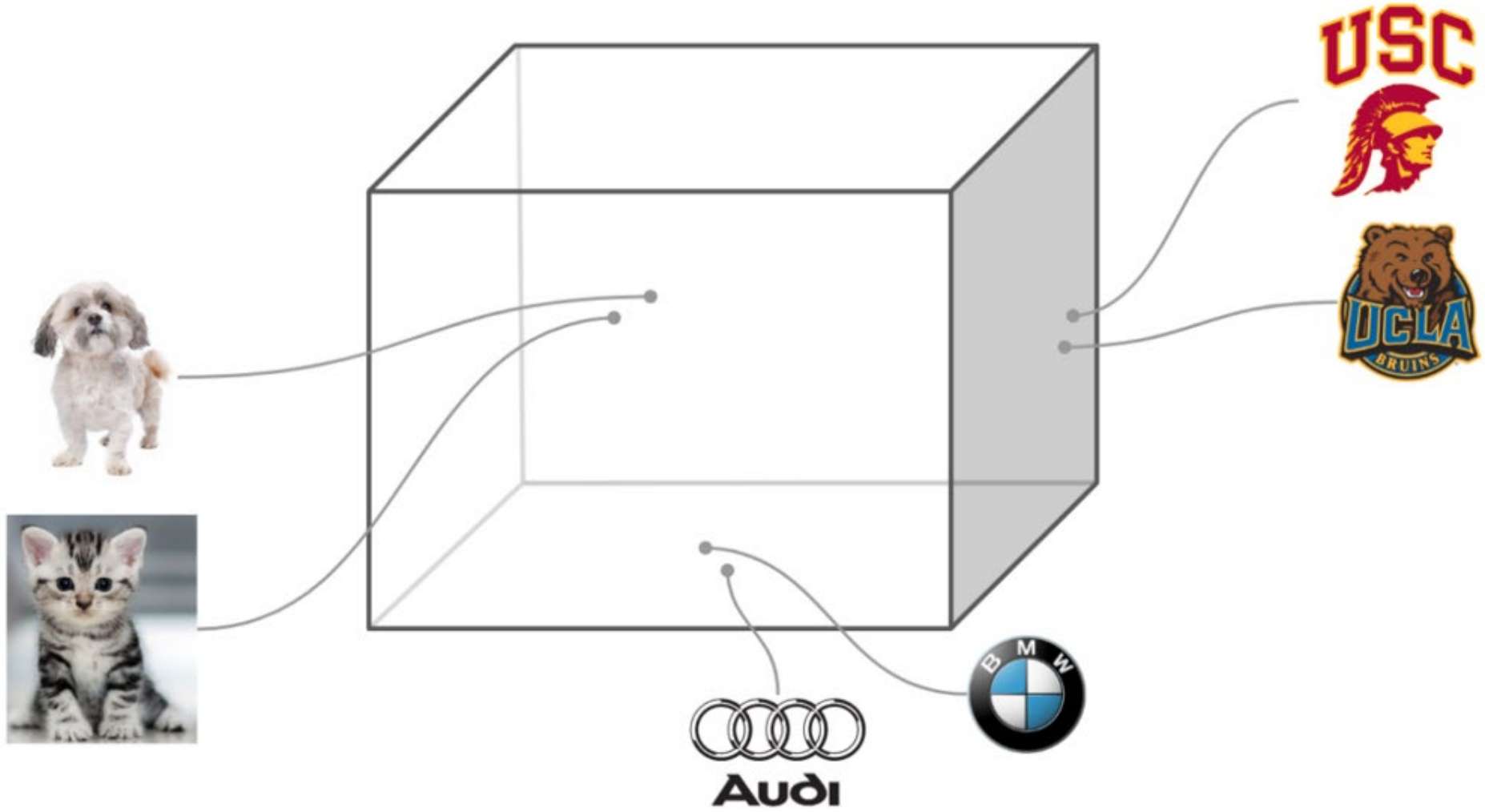
Word Embeddings:

the: <0.45, 0.89>
cat: <0.70, 0.71>
dog: <0.16, 0.98>

Skip-gram Model

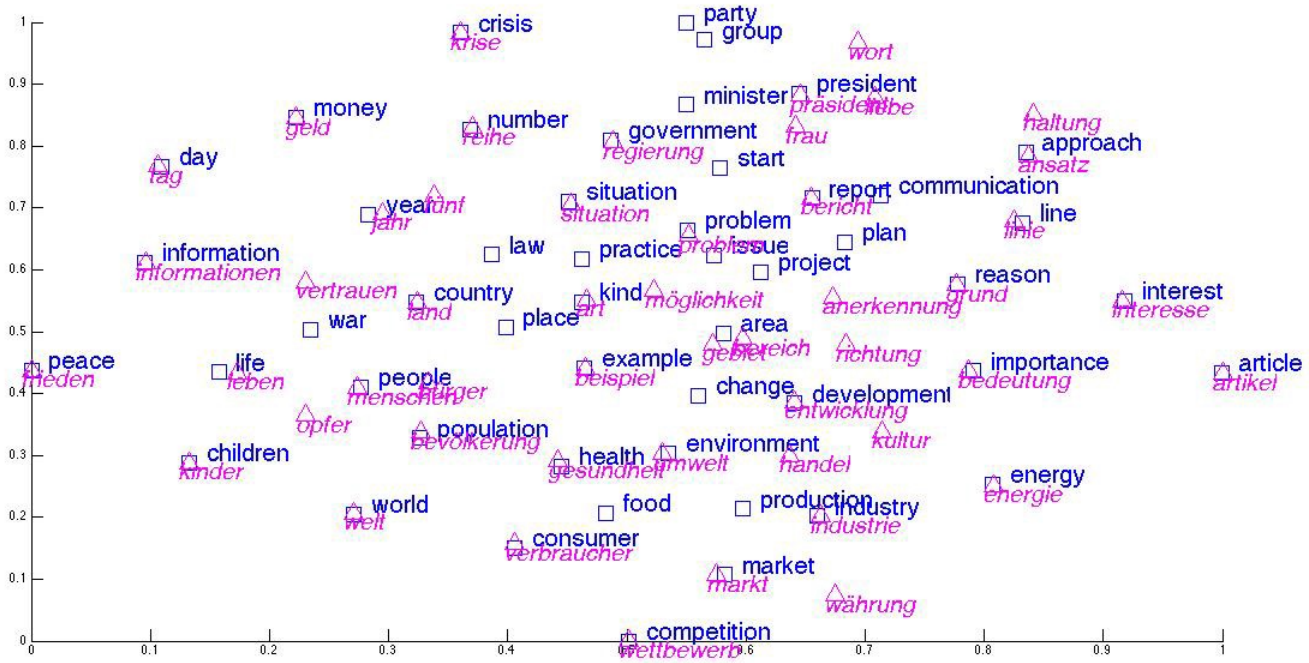


- Intuition:
  - Words that **appear alongside each other** should be **close**.
  - Words that **do not appear alongside each other** should be **far away**.
- "Alongside each other"
  - Words that appear next to each other.
  - "next to" -> within some window.
  - Window size is a parameter, let's call it **W**.
- "Do not appear alongside each other"
  - Never appear within a window together.
  - "Negative sampling"
  - Number of negative samples is a parameter, let's call it **N**.
- Distance is obtained with the dot product.


USC
UCLA
BRUINS
Audi
BMW

Word Embedding: Applications

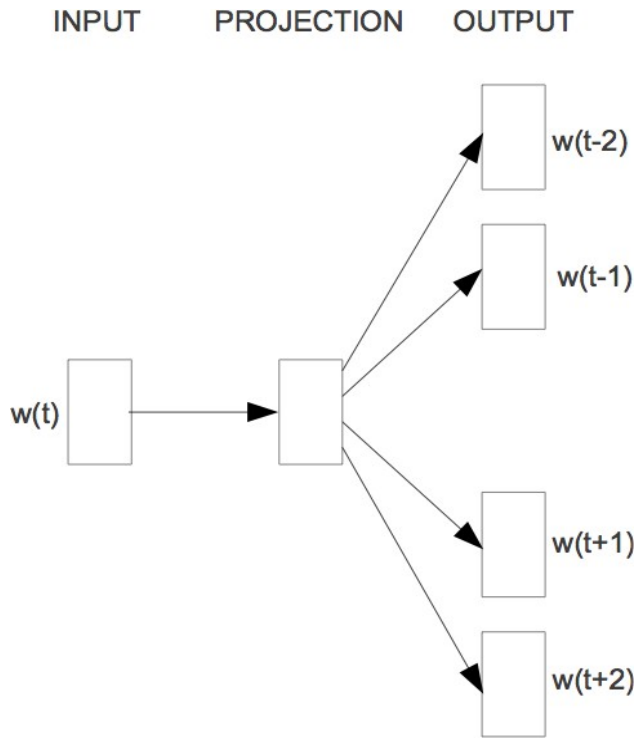
- Sentiment Analysis
- Machine Translation
- Music/Video Recommendation
- …



Word2Vec (skip-gram)

- LSA: a compact representation of co-occurrence matrix
- Word2Vec: Predict surrounding words (skip-gram)
  - Similar to using co-occurrence counts

    Levy&Goldberg (2014), Pennington et al. (2014)
- Easy to incorporate new words or sentences

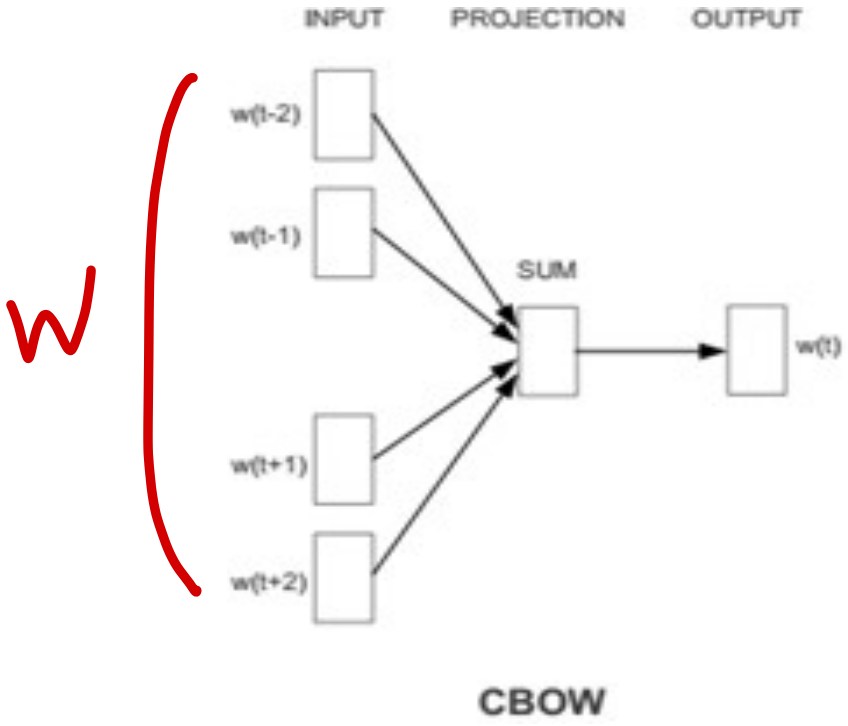


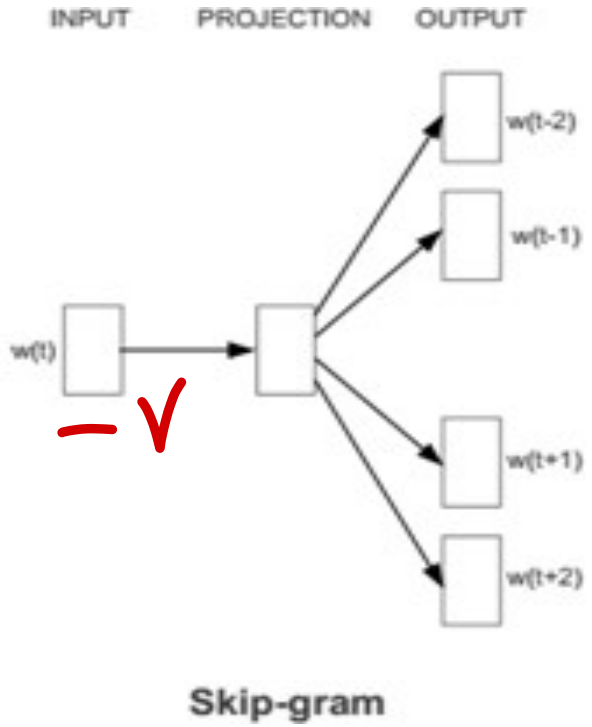
Word2Vec → Predict semantically similarity

- Idea: words that are semantically similar often occur near each other in text
  - Embeddings that are good at predicting neighboring words are also good at representing similarity
  - "You shall know a word by the company it keeps"

    - J. R. Firth

Skip-gram v.s Continuous bag-of-words





- What is the difference?

Skip-gram v.s Continuous bag-of-words (CBOW)

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Semantic Syntactic

Model Architecture	Semantic-Syntactic Word Relationship test set	
	Semantic Accuracy [%]	Syntactic Accuracy [%]
RNNLM	9	36
NNLM	23	53
CBOW	24	64
Skip-gram	55	59

CBOW : Semantic < Syntectic

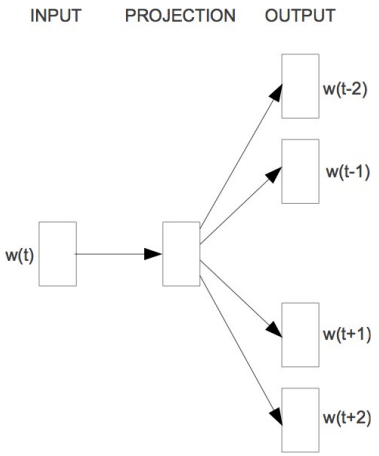
Skip-gram : Semantic > Syntectic

Objective of Word2Vec (Skip-gram)

Maximize the log likelihood of context word within a window of text

- $w_{t-m}, w_{t-m+1}, \ldots, w_{t-1}, w_{t+1}, w_{t+2}, \ldots, w_{t+m}$
- given word w_t

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log p(w_{t+j}|w_t)$$



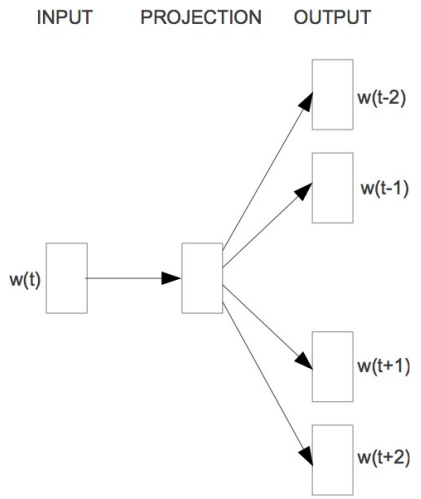
- m is usually 5~10

Objective of Word2Vec (Skip-gram)

details

- How to model $\log P(w_{t+j}|w_t)$?

$$p(w_{t+j}|w_t) = \frac{\exp(u_{w_{t+j}} \cdot v_{w_t})}{\sum_{w'} \exp(u_{w'} \cdot v_{w_t})}$$



  - softmax function
- Every word has 2 vectors
  - v_w : when w is the center word
  - u_w: when w is the outside word (context word)

How to update?

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log p(w_{t+j}|w_t)$$

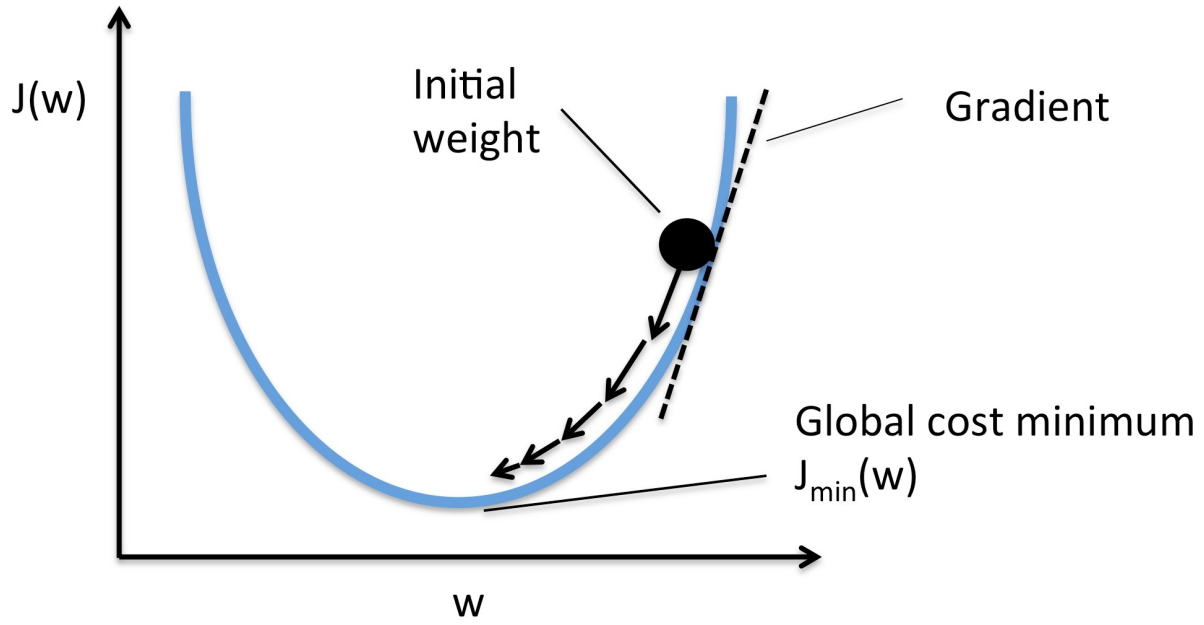
$$p(w_{t+j}|w_t) = \frac{\exp(u_{w_{t+j}} \cdot v_{w_t})}{\sum_{w'} \exp(u_{w'} \cdot v_{w_t})}$$

Minimize $J(\theta)$ with Gradient descent!

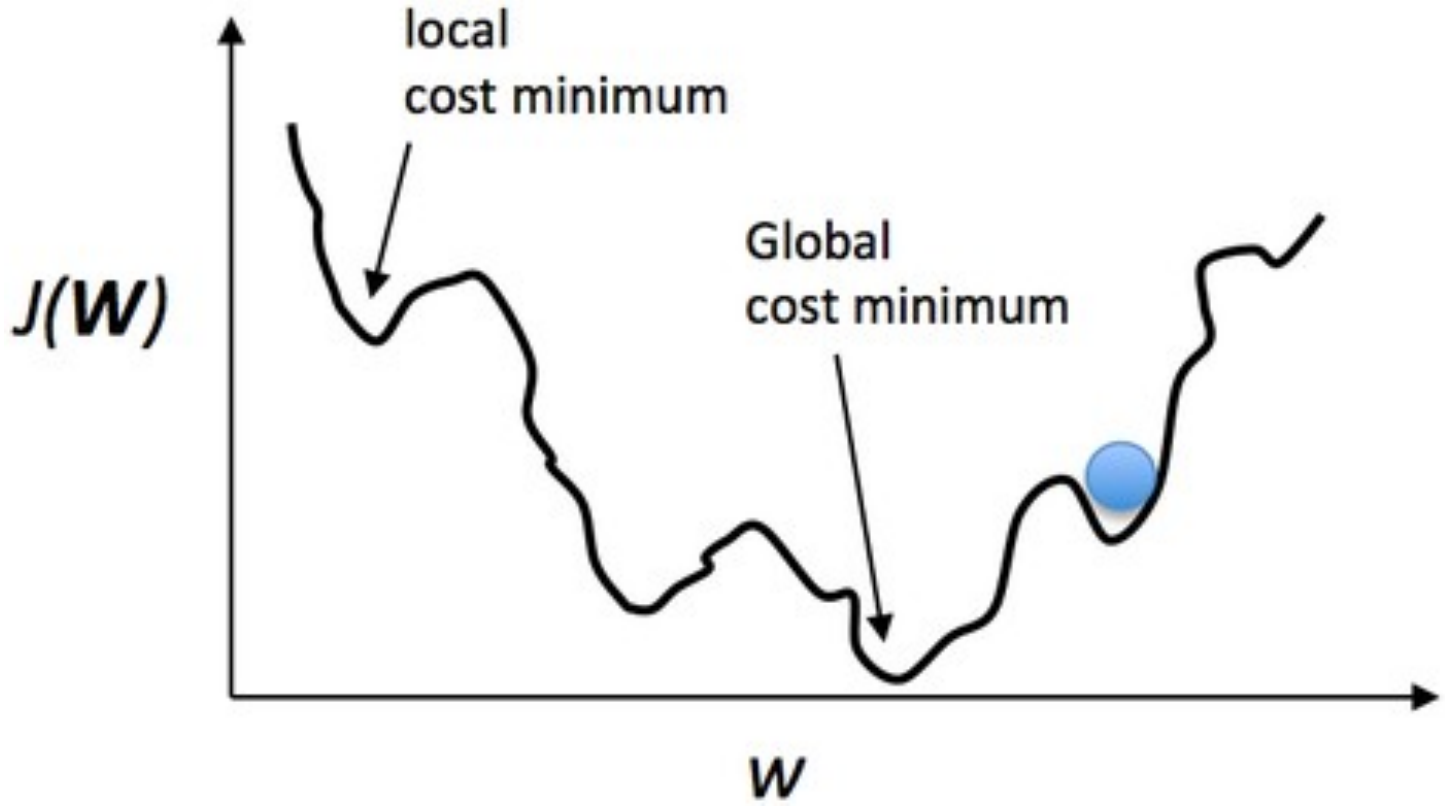
Gradient Descent

$$\min_{w} J(w)$$

Update w: $w \leftarrow w - \eta \nabla J(w)$



Local minimum v.s. global minimum



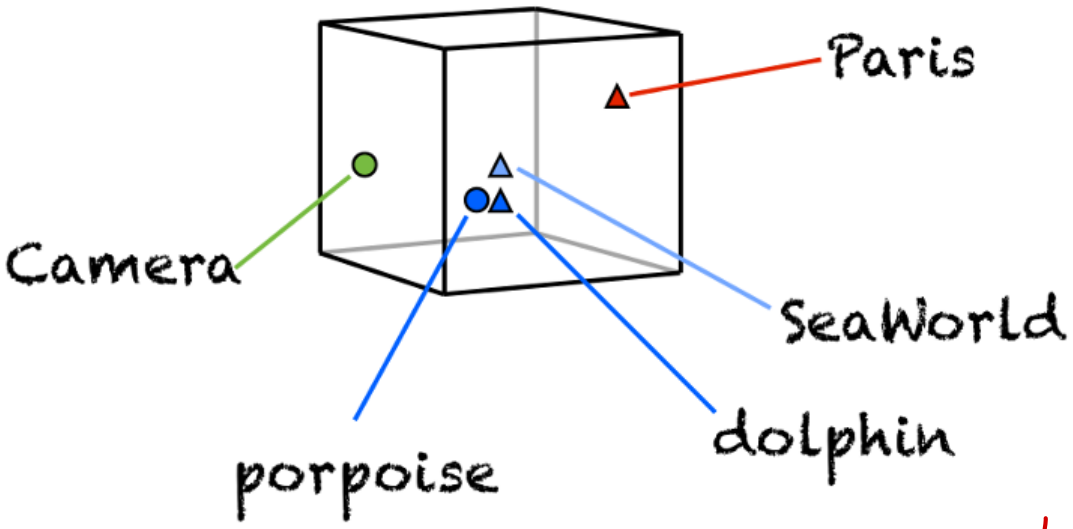
Negative sampling

- With a large vocabulary set, stochastic gradient descent is still not enough (why?)

$$\frac{\partial \log p(o|c)}{\partial v_c} = u_o - E_{w \sim p(w|c)}[u_w]$$

- Let's approximate it again!
  - Only sample a few words that do not appear in the context

Nice properties of word embeddings



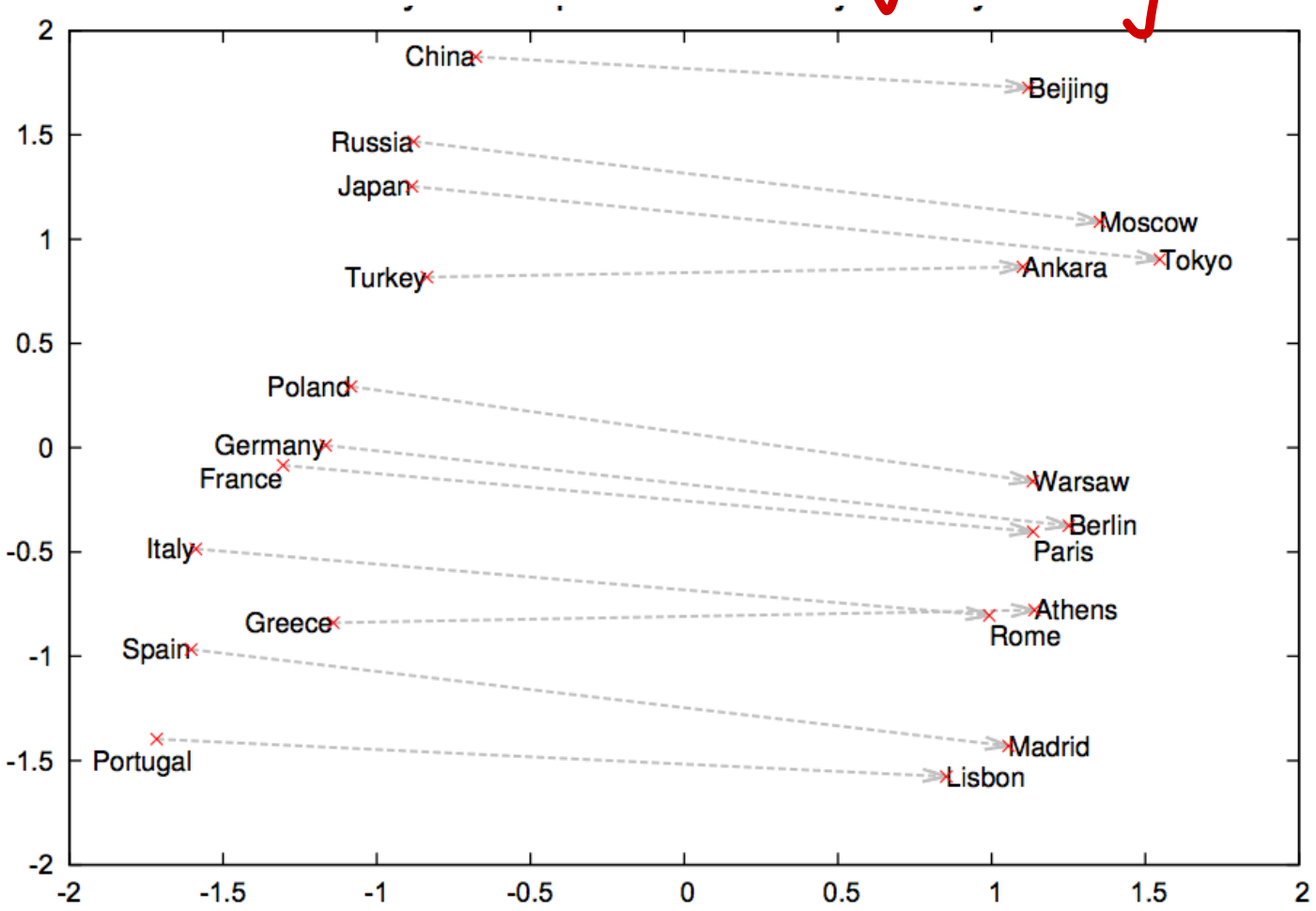
man → king

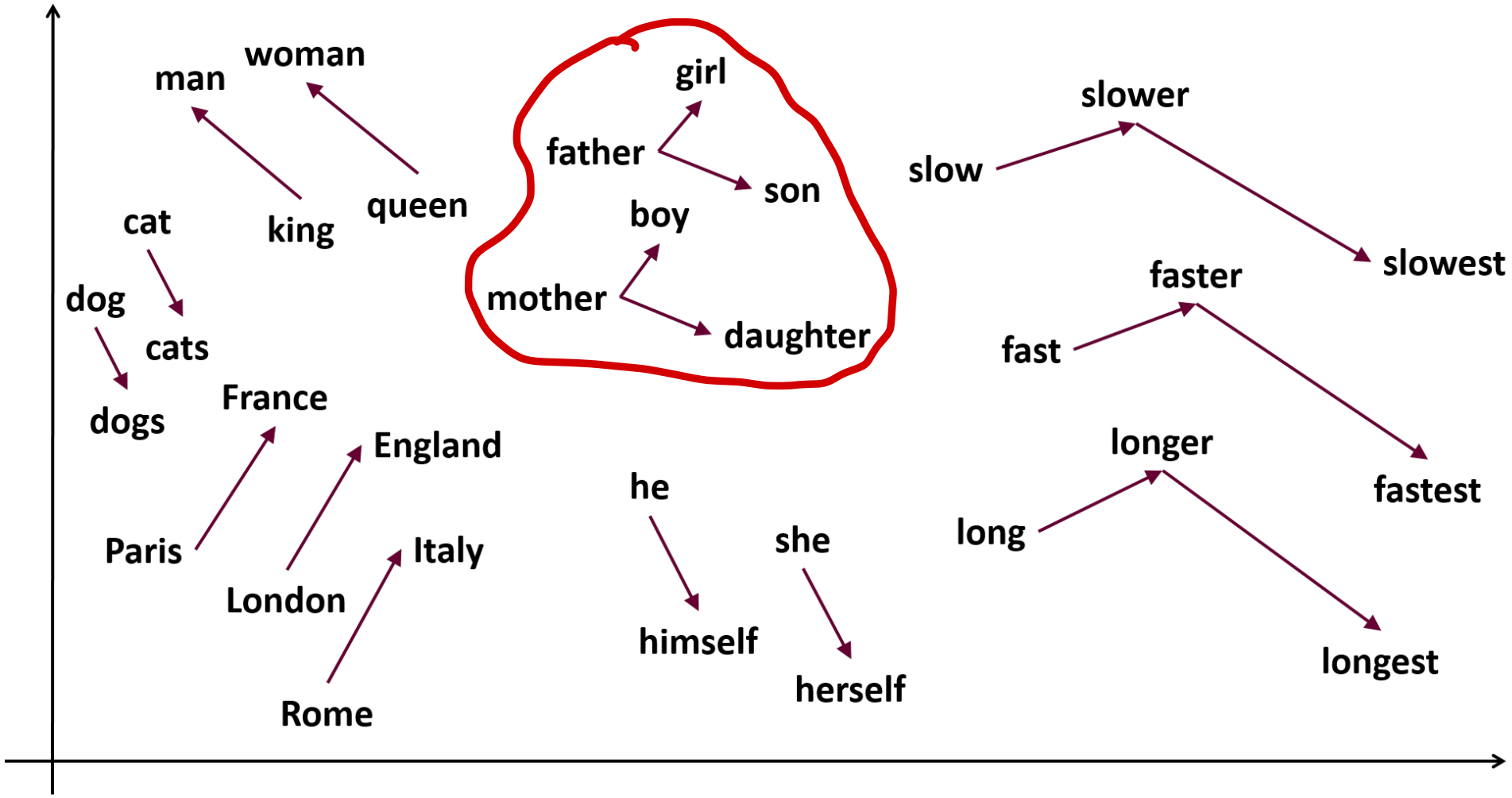
women → queen

vec(king) − vec(man) + vec(women) = vec(queen)

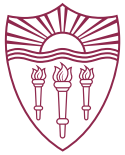
Analogies

China - Beijing + Portugal = Lisbon




man
woman
king
queen
girl
father
son
boy
mother
daughter
cat
dog
cats
dogs
France
England
Paris
London
Italy
Rome
he
himself
she
herself
slow
slower
slowest
fast
faster
fastest
long
longer
longest

Biases in word embeddings



A father and son get in a car crash and are rushed to the hospital.

The father dies.

The boy is taken to the operating room and the surgeon says,

"I can't operate on this boy, because he's my son."

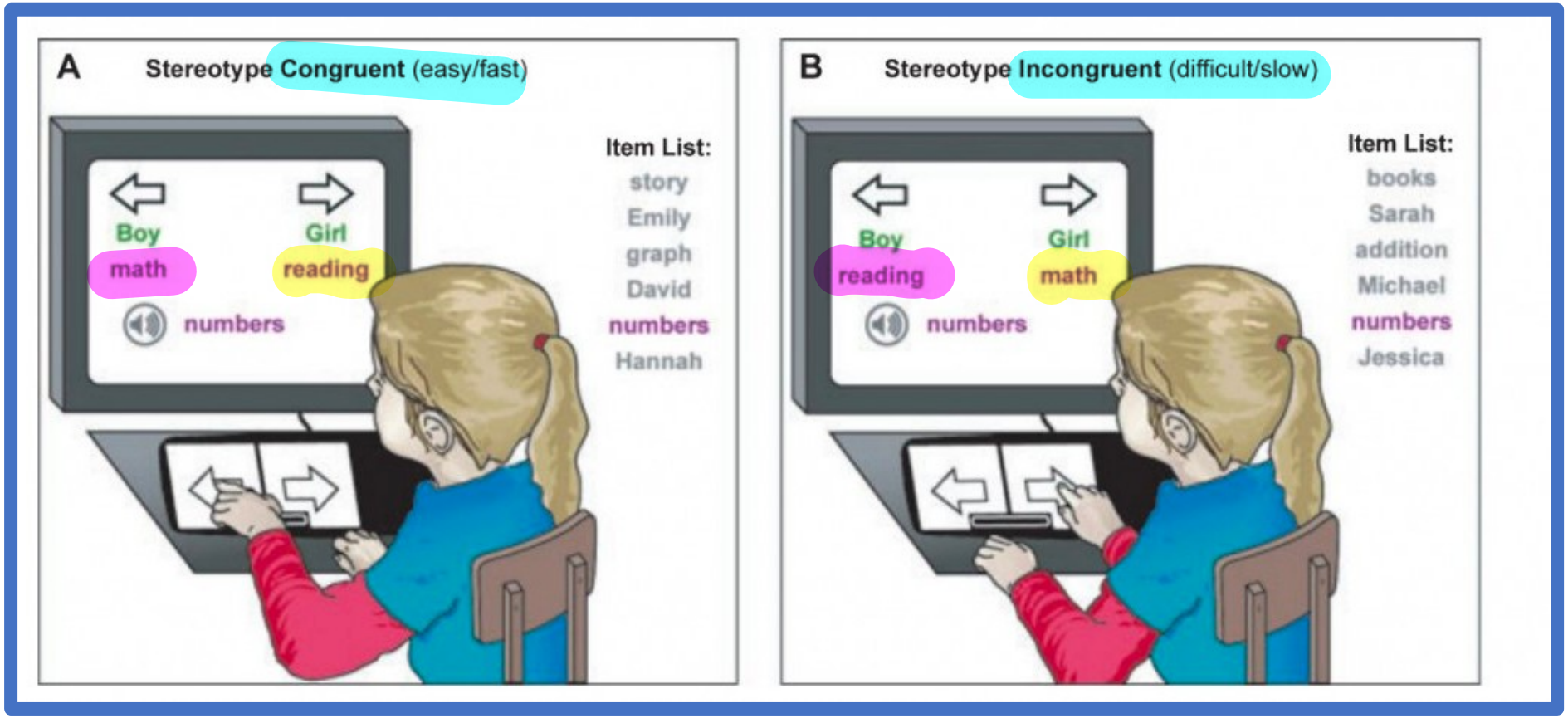
Can you explain why?

- mother
- father

-> Surgeon is the boy's mom

https://www.youtube.com/watch?v=J69HkKz9g4A

Implicit association test (IAT)



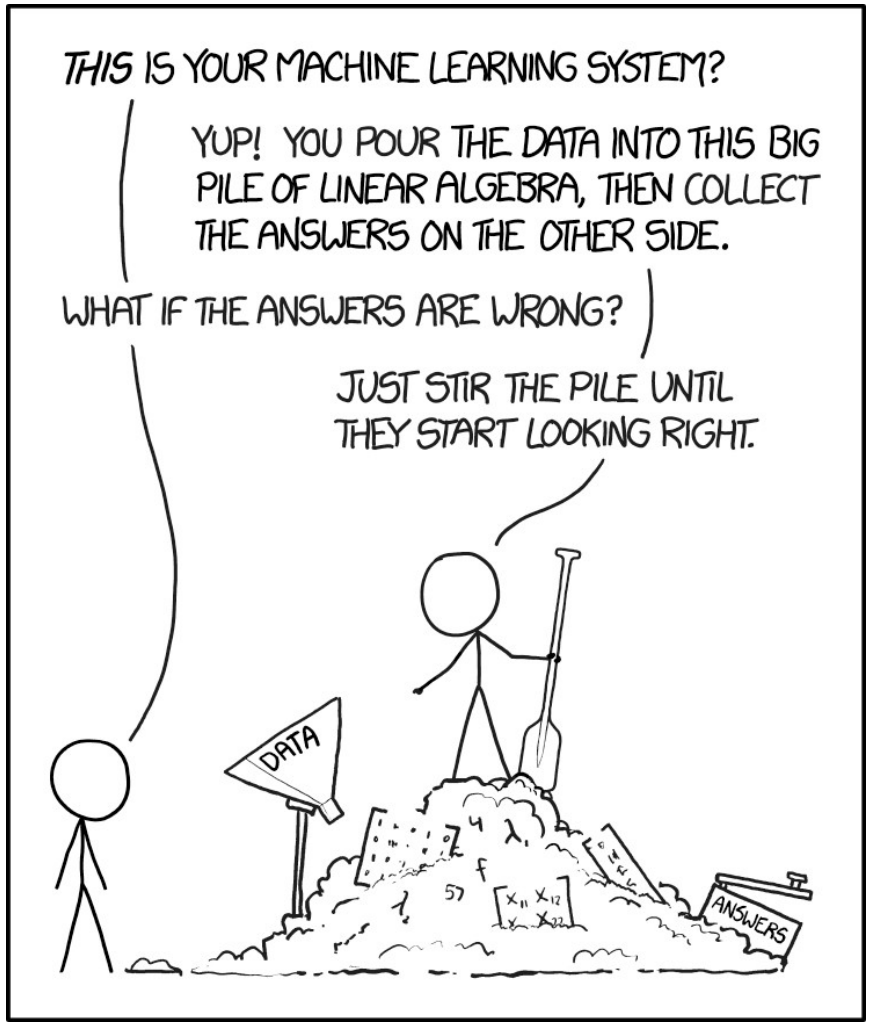
https://implicit.harvard.edu/implicit/

“Concepts in semantic memory are assumed to be linked together … with associated concepts
having stronger links … than unrelated concepts”
(Collins and Loftus, 1975).

- https://www.nature.com/articles/palcomms201786

"Concepts in semantic memory are assumed to be linked together … with associated concepts having stronger links … than unrelated concepts" (Collins and Loftus, 1975).

So does the computer



Data with Societal Bias

↓

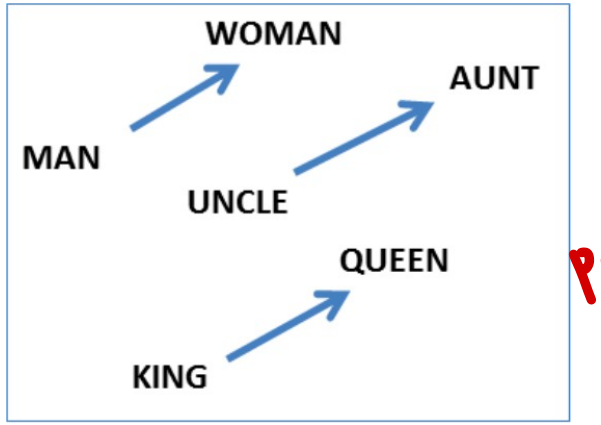
Model with Societal Bias

https://xkcd.com/1838/

Word Embeddings can be Dreadfully Sexist

bias

$$v_{man} - v_{woman} + v_{uncle} \sim v_{aunt}$$



he: _____	she: _____
brother	sister
beer	cocktail
physician	registered_nurse
programmer	homemaker
professor	associate professor

drink

profession

academic

Use Google w2v embedding trained from the news

Related works

Aylin, Joanna, and Arvind (2017) measure the biases in embedding using Implicit Association Test (IAT) and demonstrate it contain human-like biases

Human-like biases

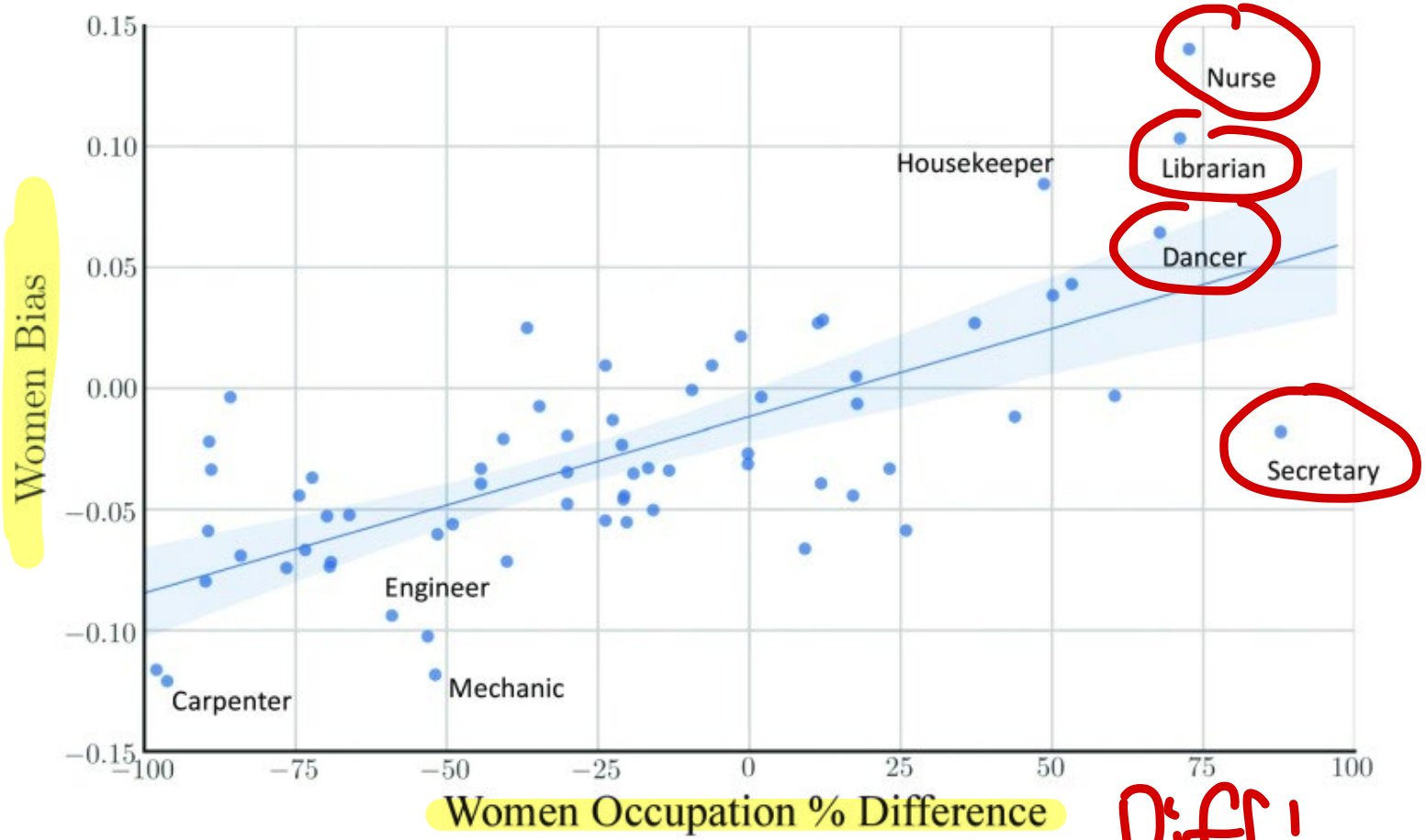
Garg, Schiebinger, Jurafsky, Zou (2017) Word embeddings quantify 100 years of gender and ethnic stereotypes:

(a) Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding.

1910	1950	1990
charming	delicate	maternal
placid	sweet	morbid
delicate	charming	artificial
passionate	transparent	physical
sweet	placid	caring
dreamy	childish	emotional
indulgent	soft	protective
playful	colorless	attractive
mellow	tasteless	soft
sentimental	agreeable	tidy

caring aspect

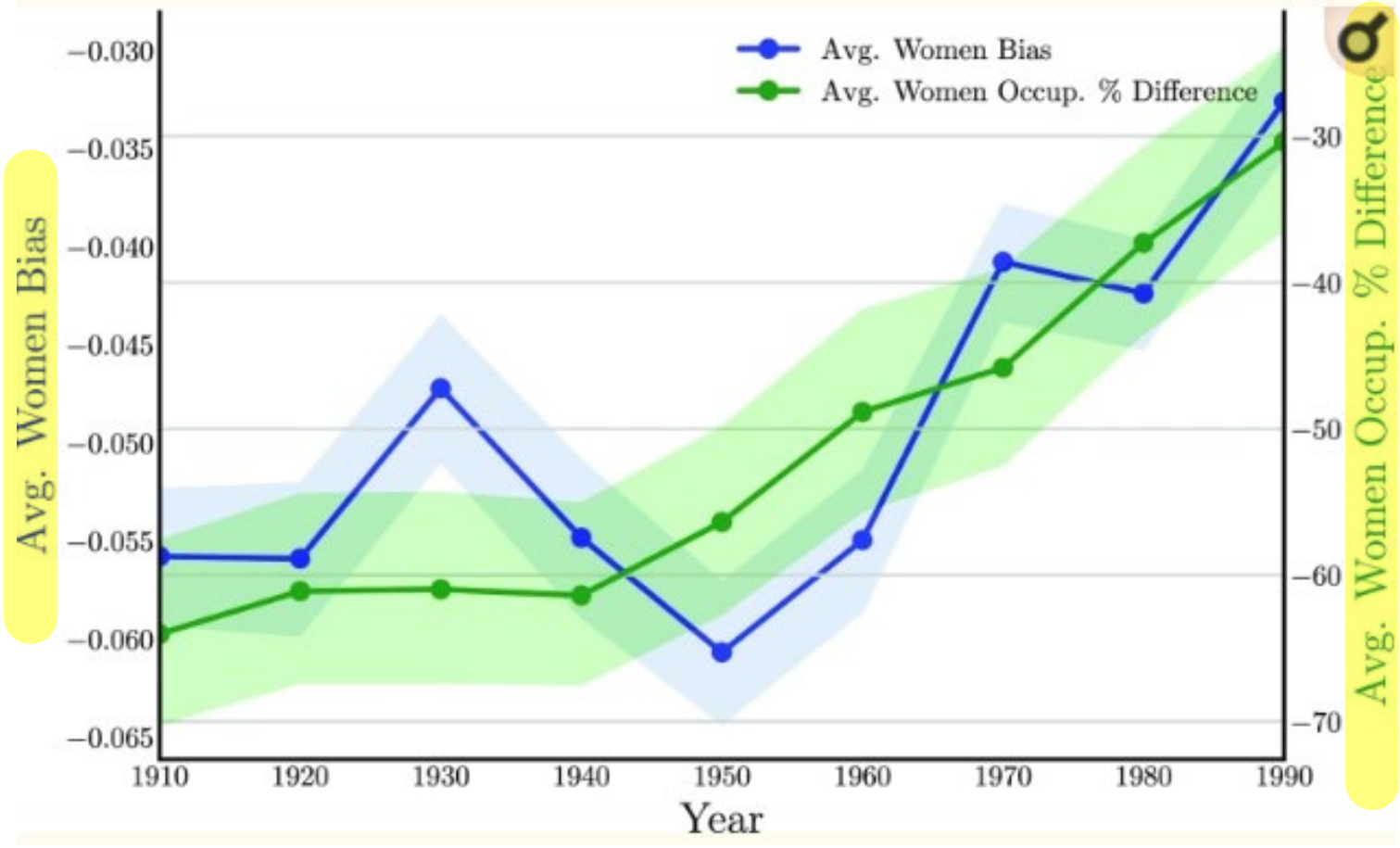
Another View



Bias

Diff'

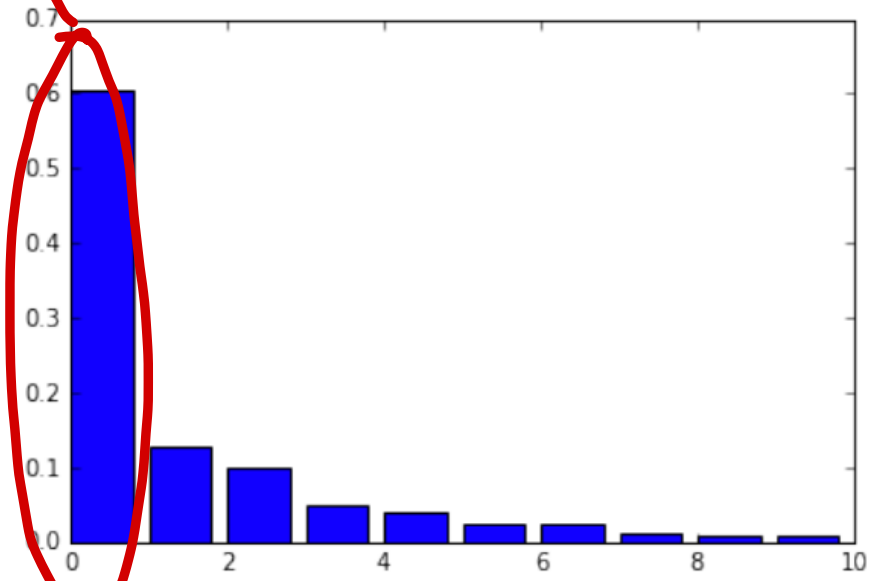
... By Time


Avg. Women Bias
Avg. Women Occup. % Difference
−0.030
−0.035
−0.040
−0.045
−0.050
−0.055
−0.060
−0.065
−30
−40
−50
−60
−70
1910
1920
1930
1940
1950
1960
1970
1980
1990
Year

Geometry of Gender and Bias

- Identifying the gender subspace

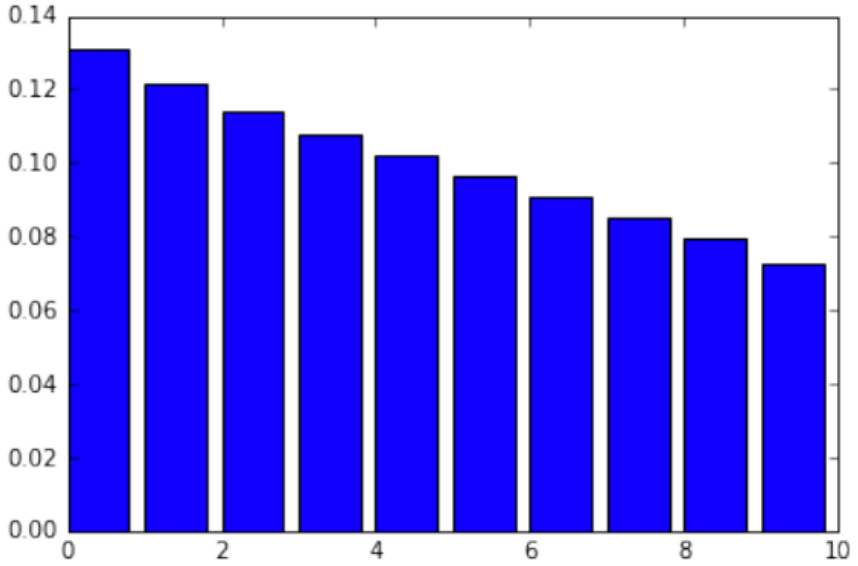
Top 10 Eigenvalue



PCA ("he"- "she", "father"-"mother",…)

Gender Pair

Top 10 Eigenvalue

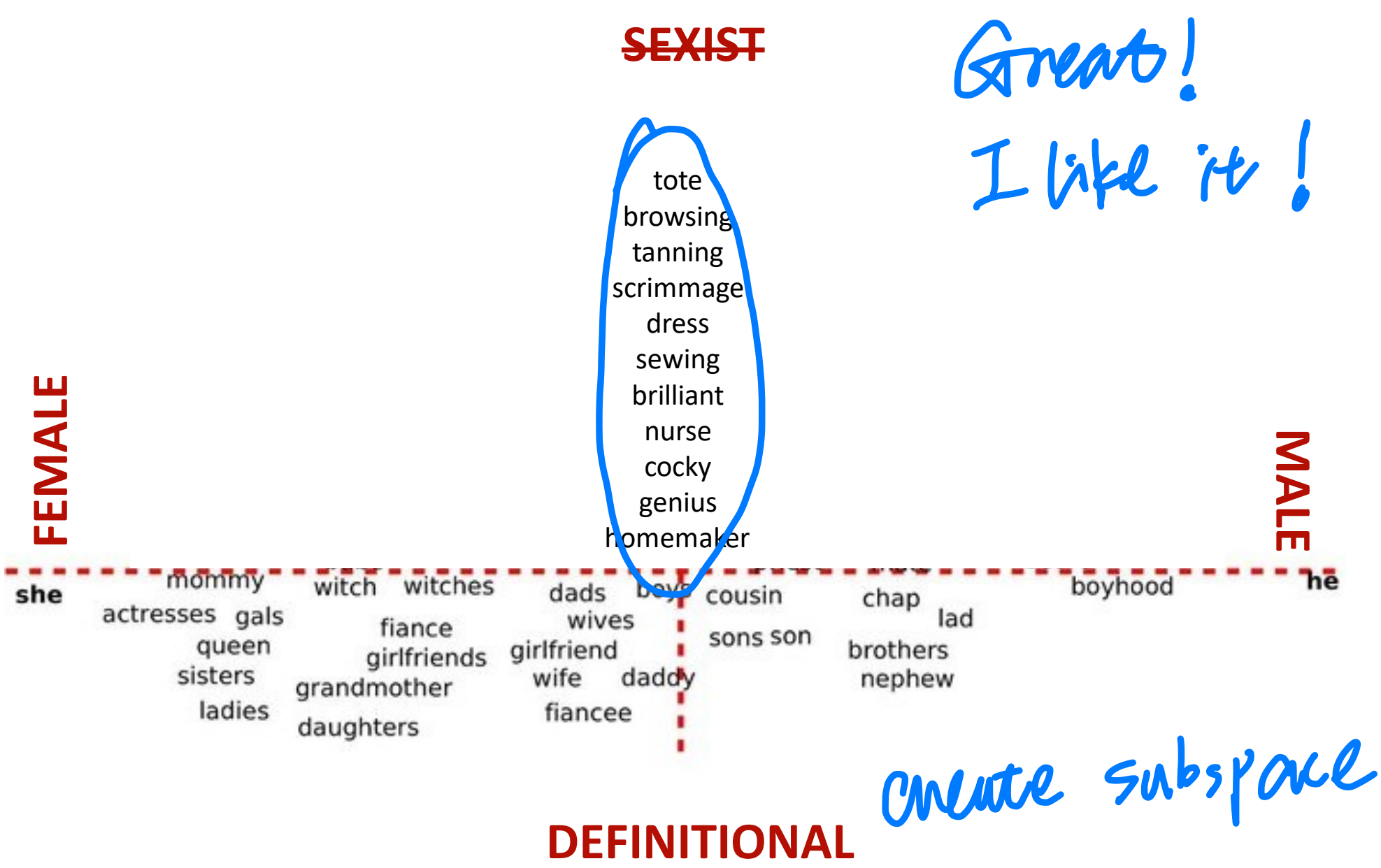


PCA ("dog"- "cat", "house"-"building",…)

Random Pair

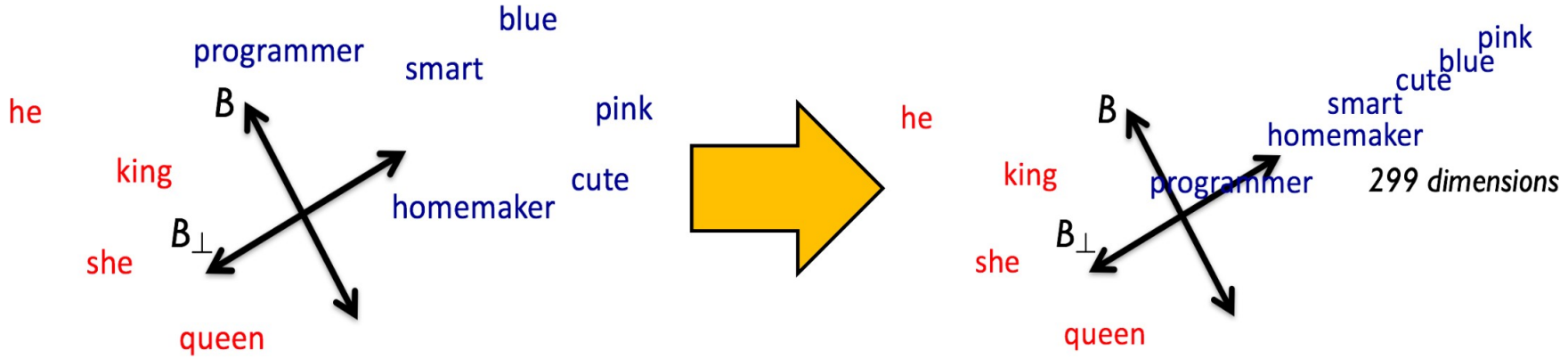
Kai-Wei Chang (kwchang.net/talks/sp.html)

Reducing bias


SEXIST
Great!
I like it!
tote
browsing
tanning
scrimmage
dress
sewing
brilliant
nurse
cocky
genius
homemaker
FEMALE
MALE
she
he
mommy
actresses
gals
queen
sisters
ladies
witch
witches
fiance
girlfriends
grandmother
daughters
dads
wives
girlfriend
wife
daddy
fiancee
cousin
sons
son
chap
lad
brothers
nephew
boyhood
DEFINITIONAL
create subspace

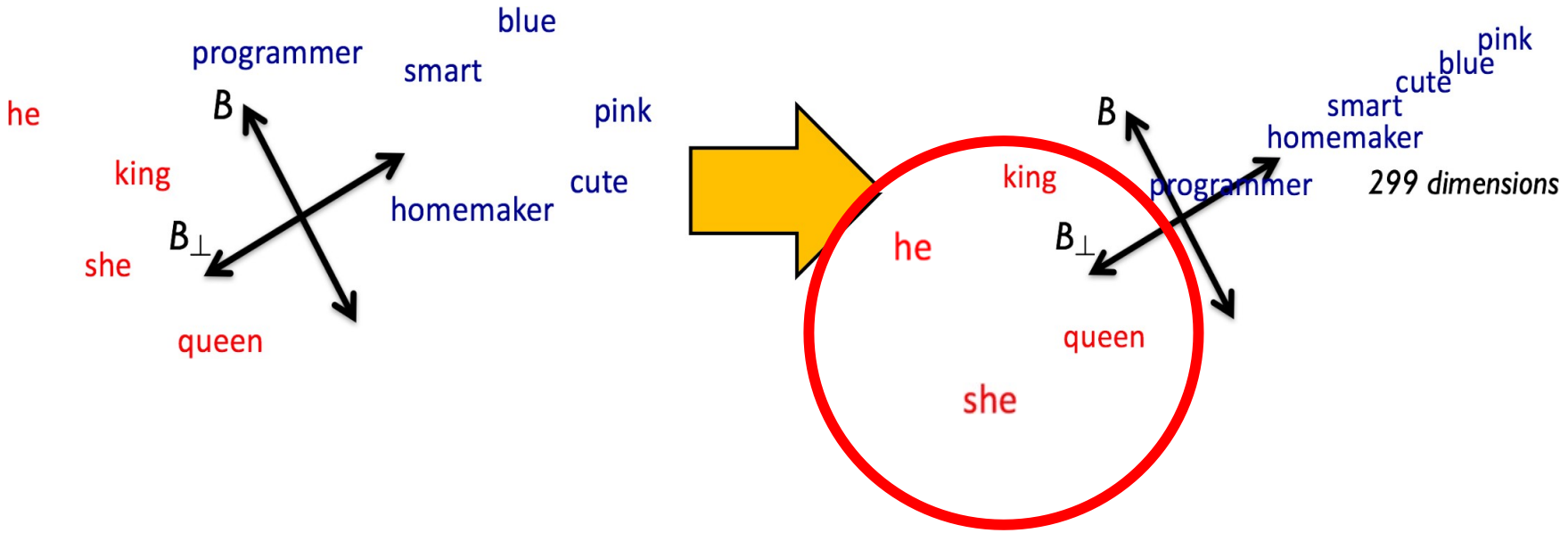
Approach 1:

- Project out gender dimension (hard version)
  - Step 1: Remove gender dimension from gender neutral words



Approach 1: 方法1

- Project out gender dimension (hard version)
  - Step 1: Remove gender dimension from gender neutral words
  - Step 2: re-center gender-definitional pairs

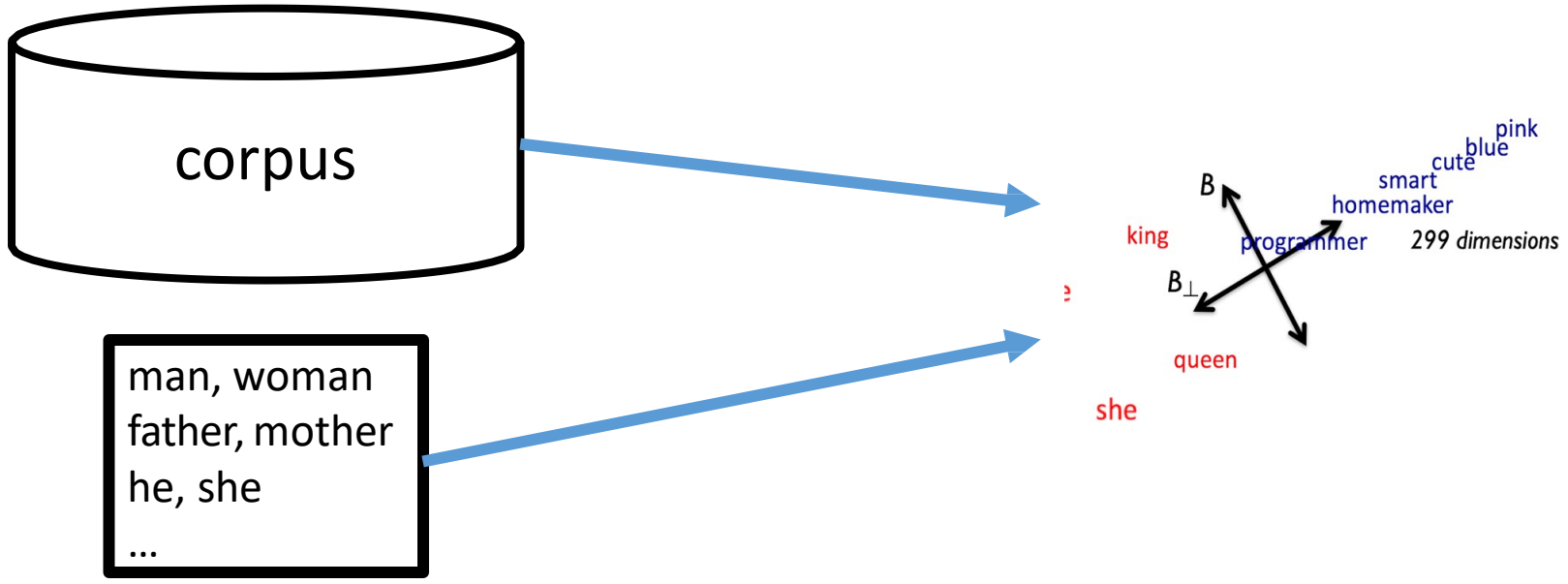


Approach 2: 方法2

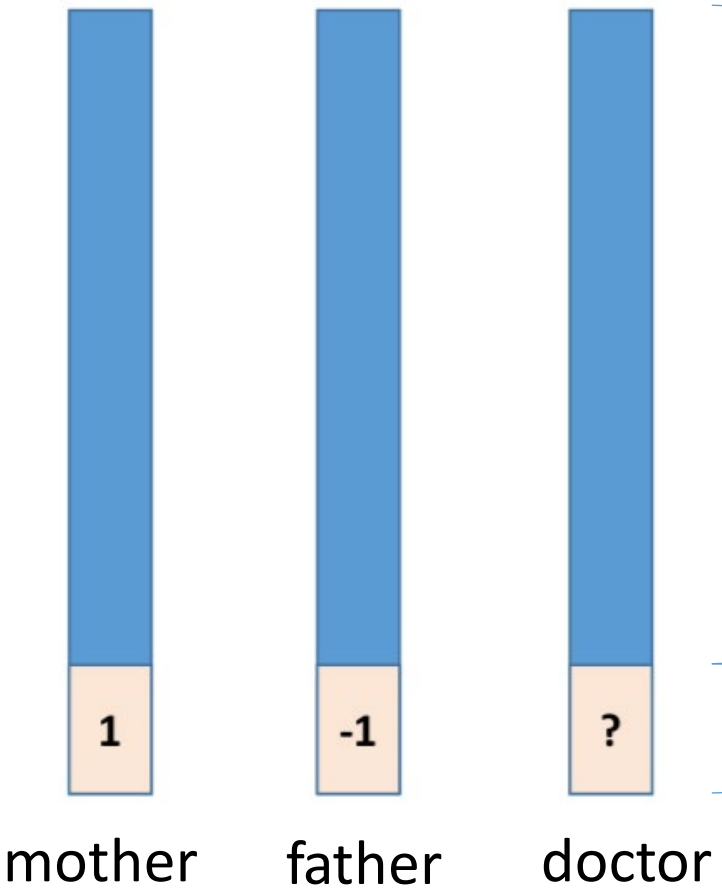
Learning Gender-Neutral Word Embedding

[Jieyu+EMNLP18]

- How can we **not** encode gender information in word vectors?

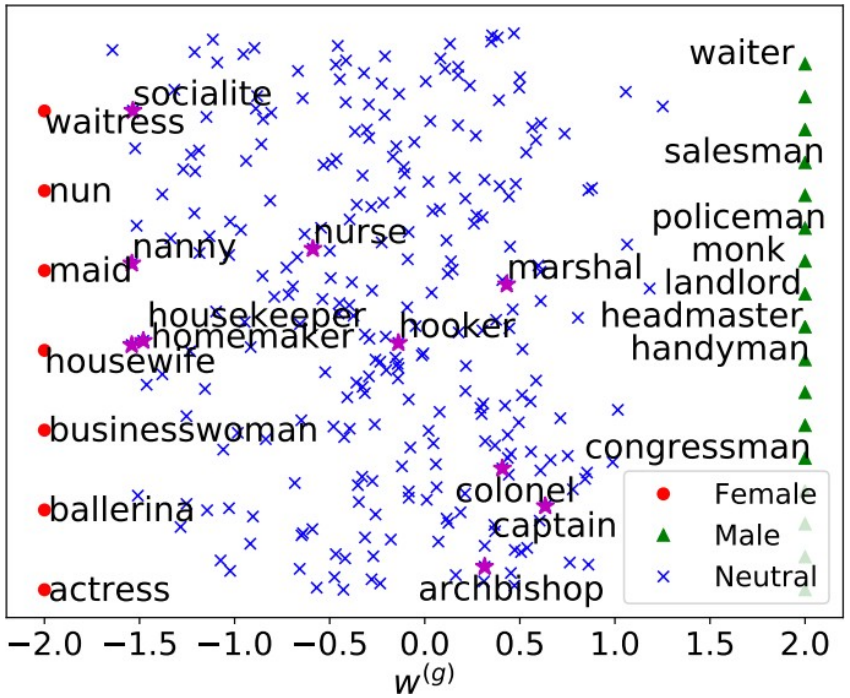


Approach 2:
Learning Gender-Neutral Word Embeddings
[Jieyu+ EMNLP18]

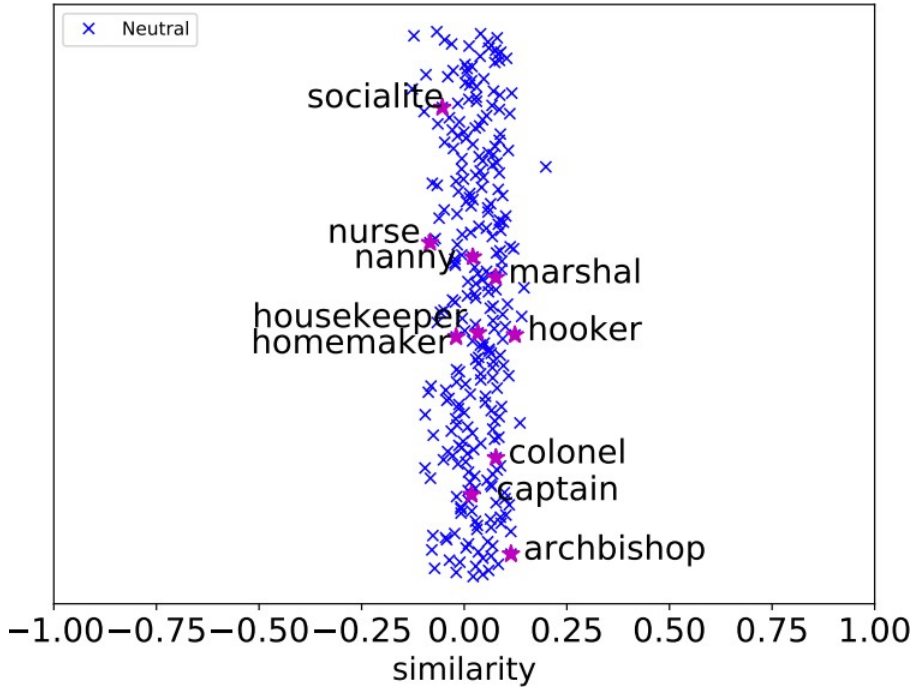


dimensions for other latent aspects w^{N}

dimensions reserved for gender information, w^{g}

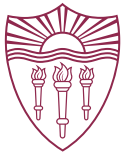


w^{g}



w^{N}

Are these debiased vectors actually useful?

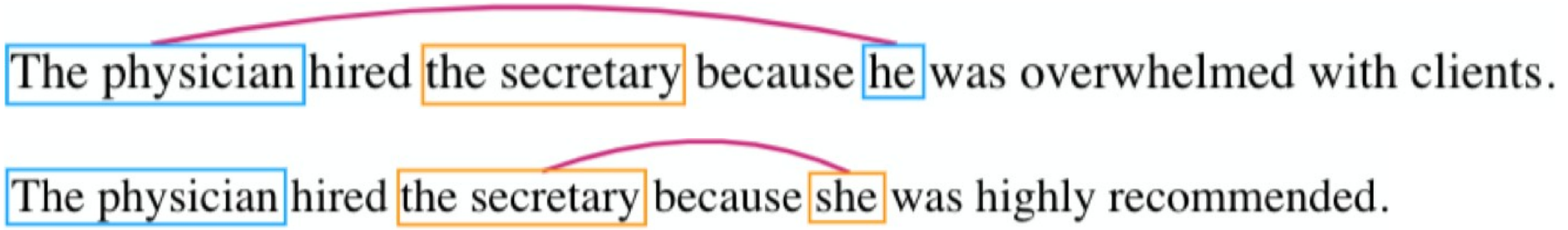


- Reducing Gender Bias in Data Level
  - A case study on co-reference resolution

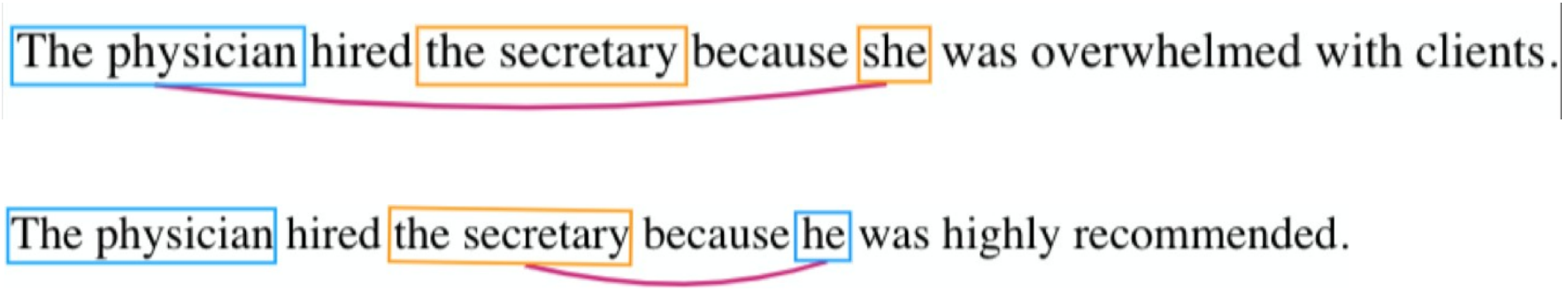
- Reducing Gender Bias in Inference Level Guiding predictions by corpus-wise constraints

Gender bias in coreference resolution

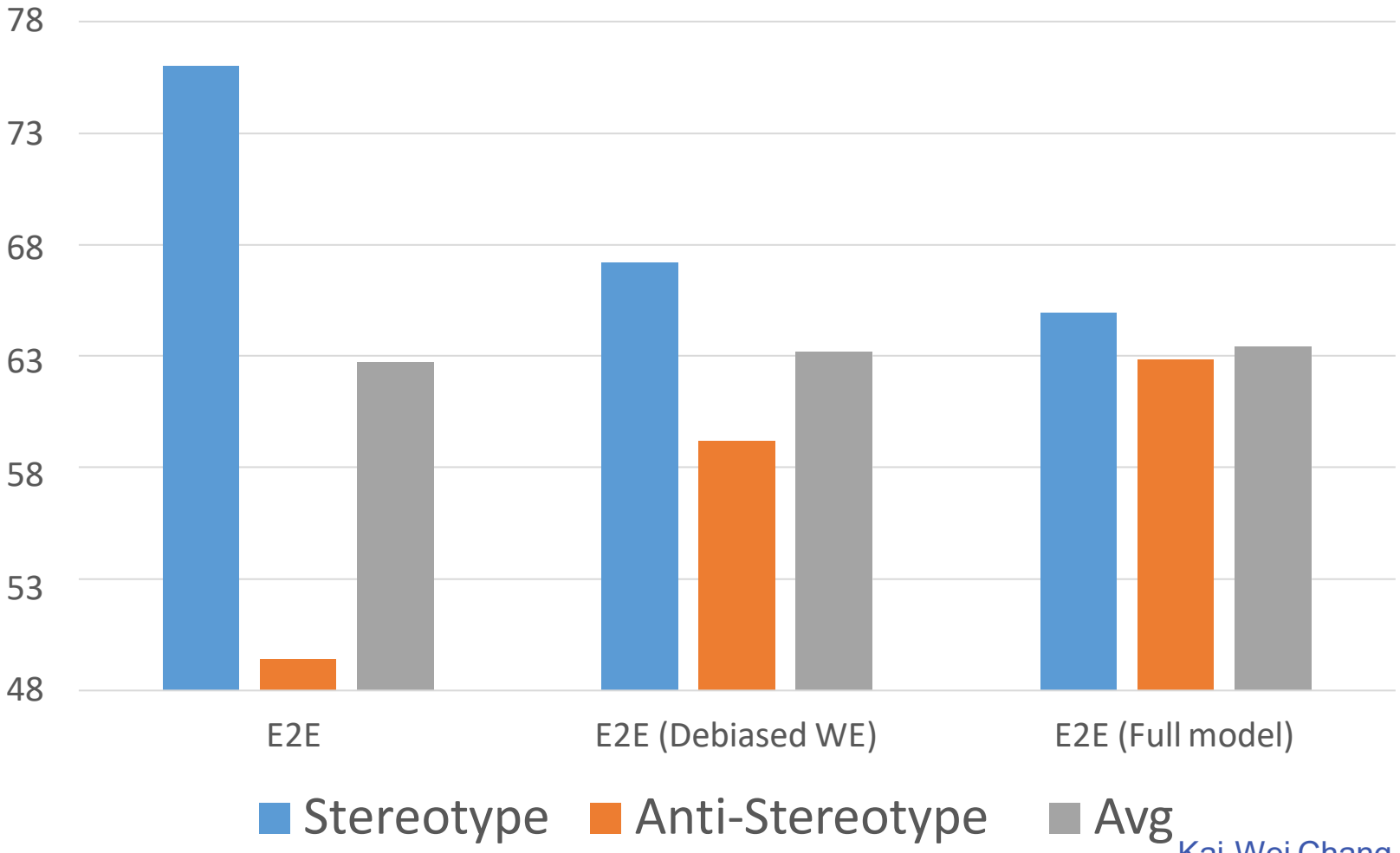
- Stereotypical dataset



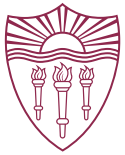
- Anti-stereotypical dataset



Gender bias in Coref System



Kai-Wei Chang (kwchang.net/talks/sp.html)



- Questions?

- Virtual office hour
- https://usc.zoom.us/j/95136500603?pwd=VEJhblhWK25lT2N3RC9FNWk3eTJKQT09