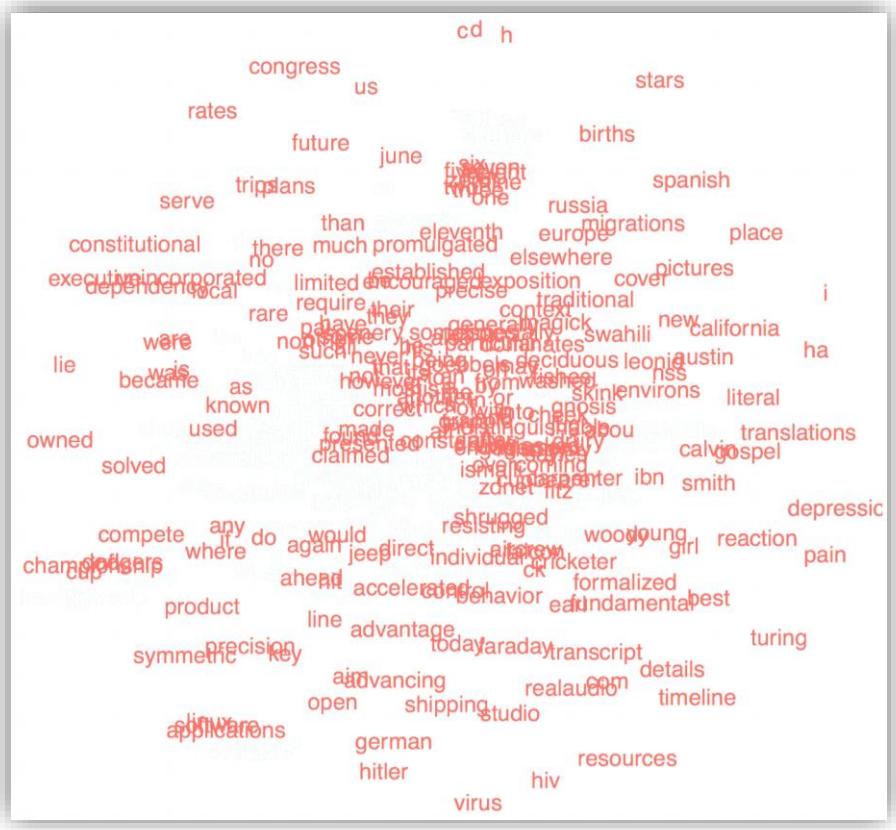
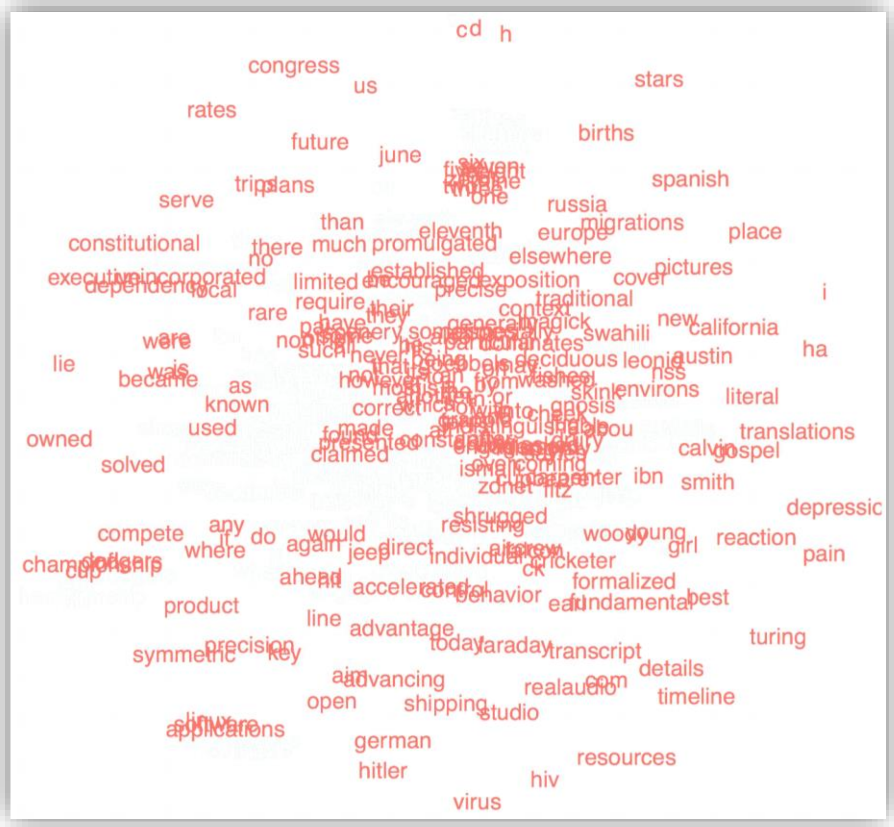
The Strange Geometry of Skip-Gram with Negative Sampling: A story of geometric observations



David Mimno and Laure Thompson
Cornell University

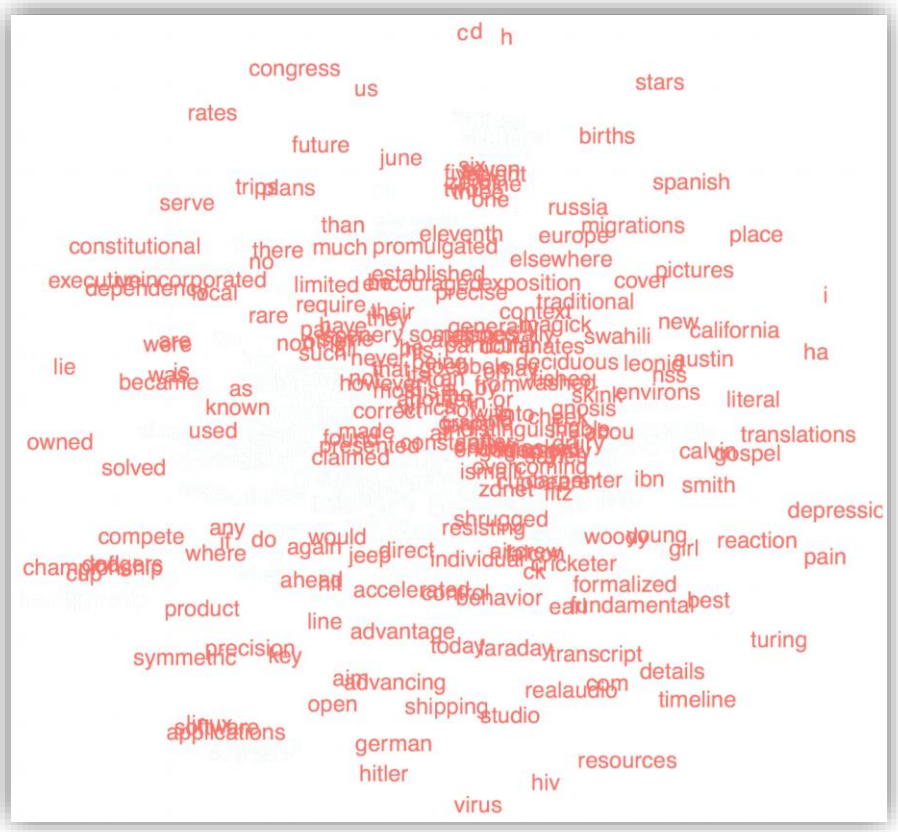
Ideally…



cd
h
congress
us
stars
rates
births
future
june
spanish
serve
russia
place
constitutional
migrations
than
much
promulgated
elsewhere
pictures
limited
exposition
traditional
california
swahili
rare
require
context
lie
ha
became
literal
known
translations
owned
used
solved
smith
ibn
depressic
compete
reaction
where
do
pain
formalized
product
fundamental
best
line
advantage
turing
transcript
details
open
shipping
timeline
studio
german
resources
hitler
hiv
virus

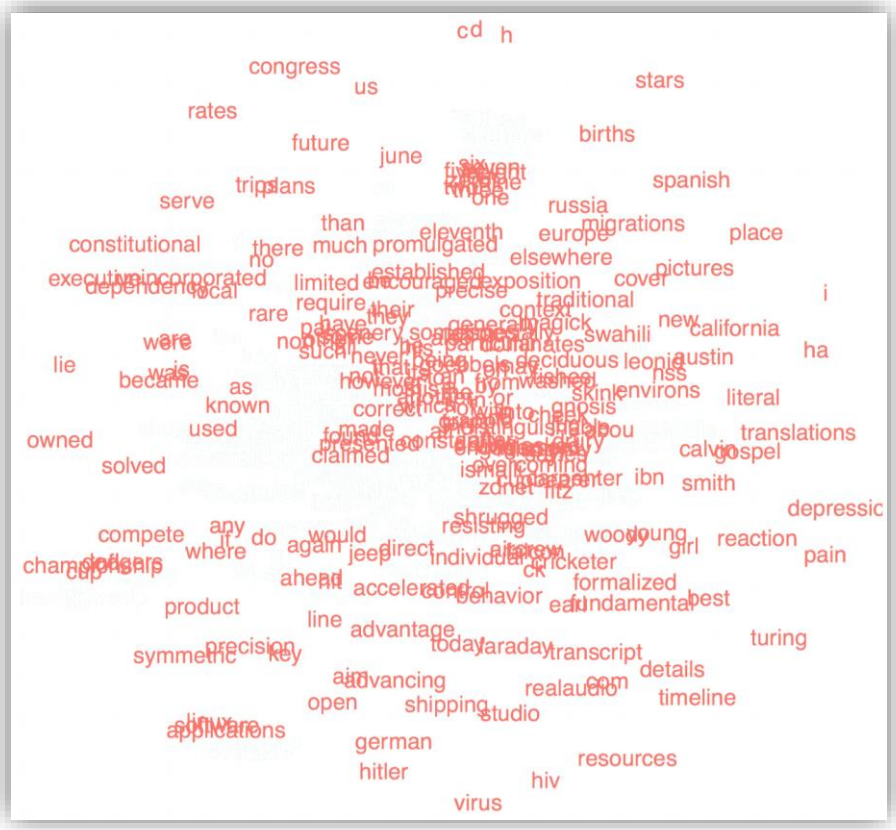
Ideally…

Words span the entire K-dimensional space



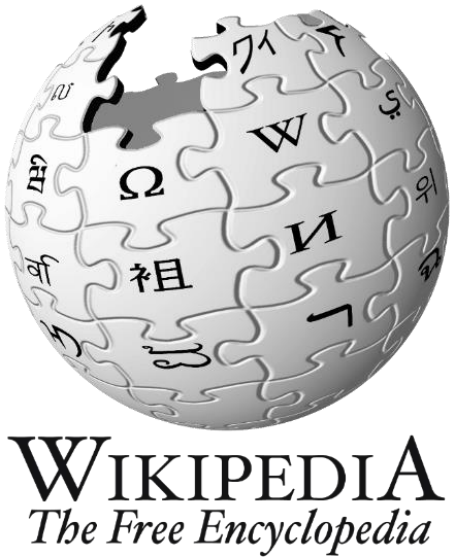
Ideally…

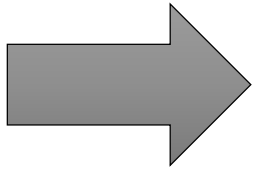
Words span the entire K-dimensional space



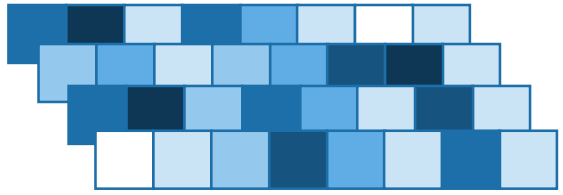
Distance reflects semantic relationships

Word Embeddings



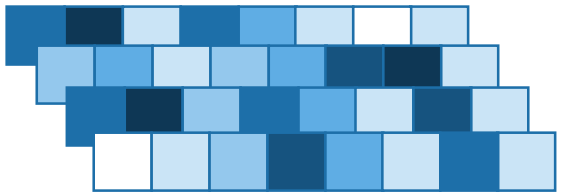


Dense Vectors

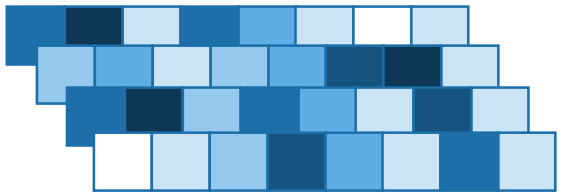


Skip-Gram with Negative Sampling (SGNS)

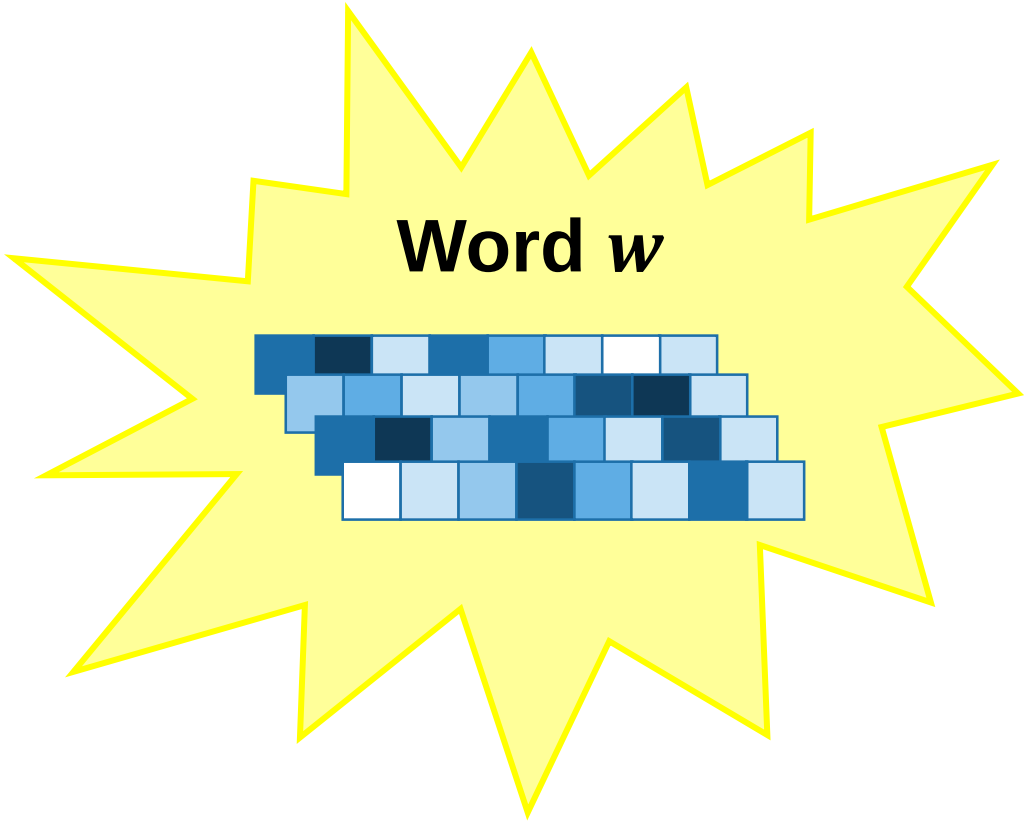
Word w

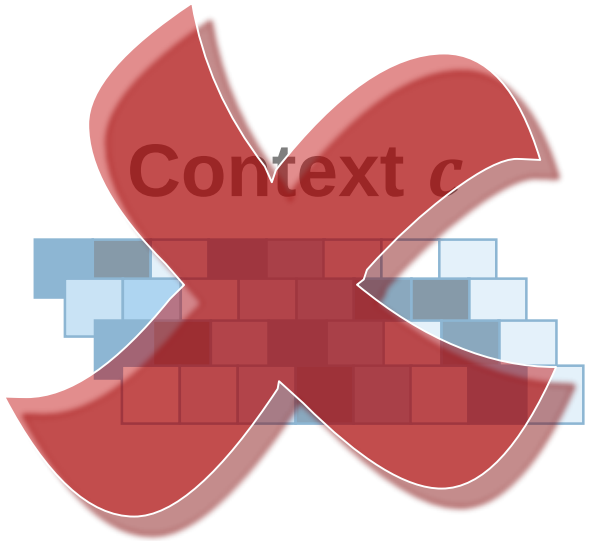


Context c



Skip-Gram with Negative Sampling (SGNS)





SGNS: Skip-Gram Model

The brown fox jumps over the lazy dog.

SGNS: Skip-Gram Model

The brown fox **jumps** over the lazy dog.

SGNS: Skip-Gram Model

The brown fox **jumps** over the lazy dog.

Context Window Size = 2

SGNS: Skip-Gram Model

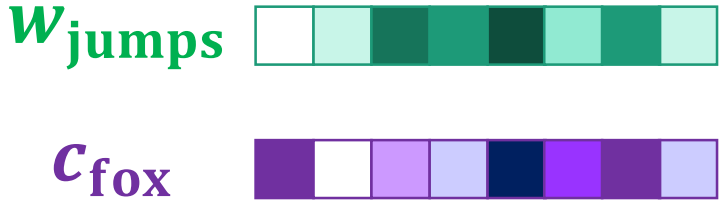
The brown fox **jumps** over the lazy dog.

Context Window Size = 2

jumps → { brown, fox, over, the }

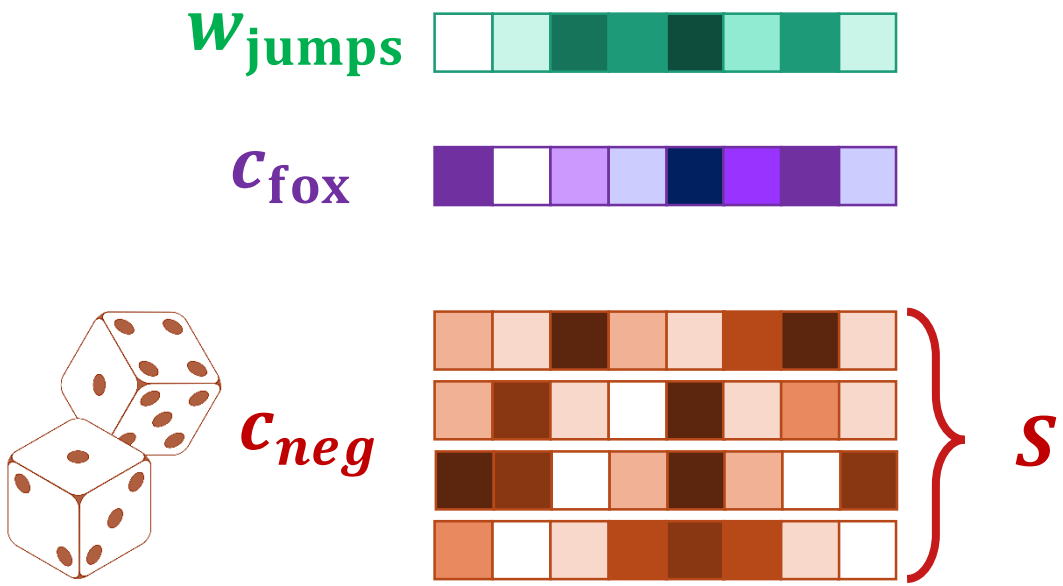
SG**NS**: Negative Sampling

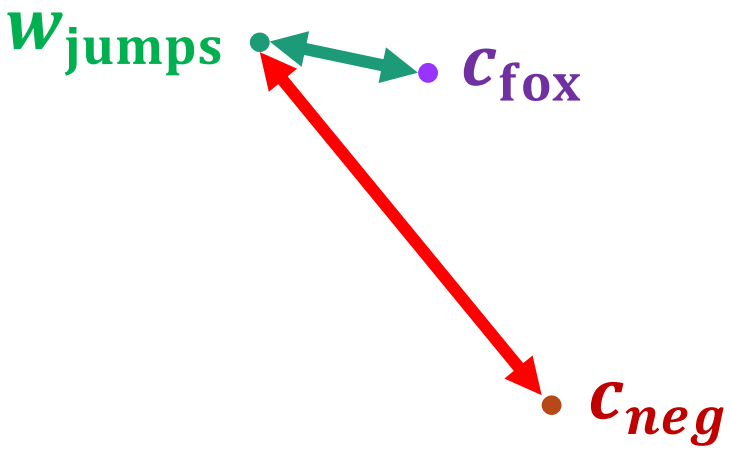
Co-occurrence **jumps** , **fox**:



SG**NS**: Negative Sampling

Co-occurrence **jumps** , **fox**:





Experimental Setup

Corpus

WIKIPEDIA
The Free Encyclopedia

Embeddings

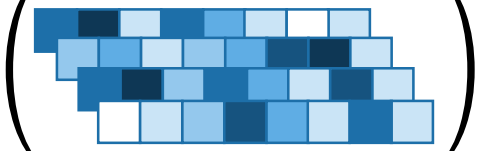
word2vec
GloVe

Parameters

Vector Size: 50
Window Size: 5

Mean word vector

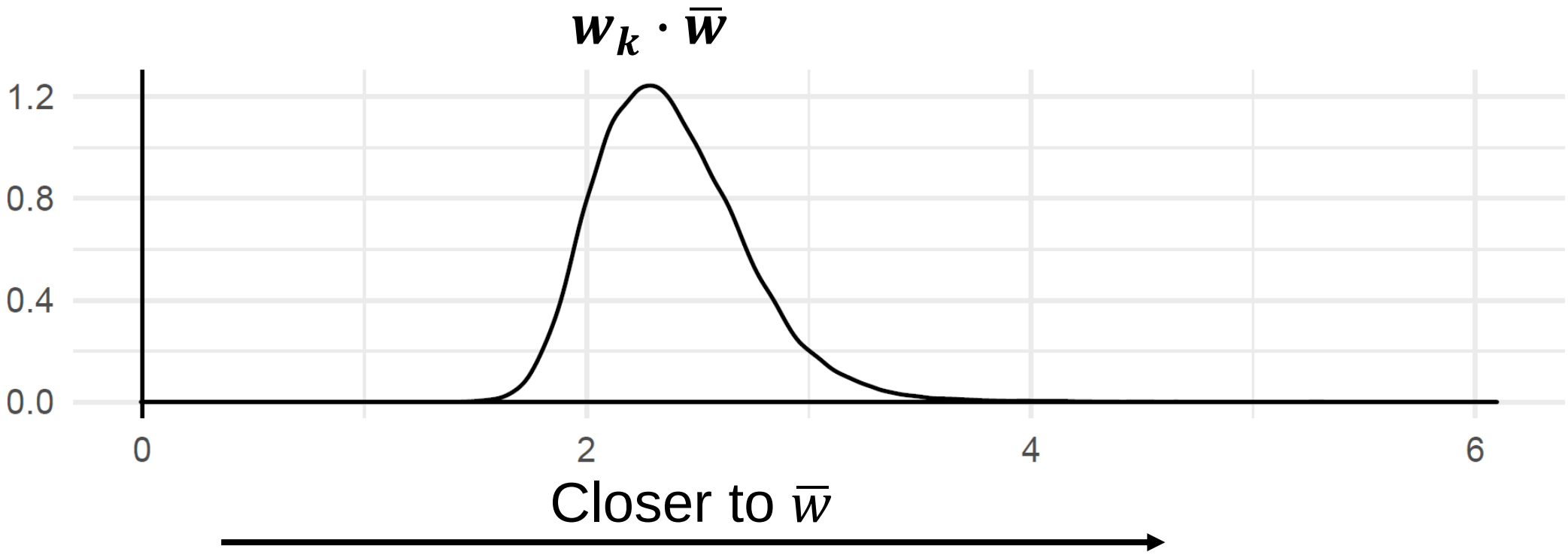
$$\bar{w} = \text{mean}\left(\right)$$



Observation #1:

SGNS vectors arrange along a primary axis

SGNS vectors point toward mean word vector



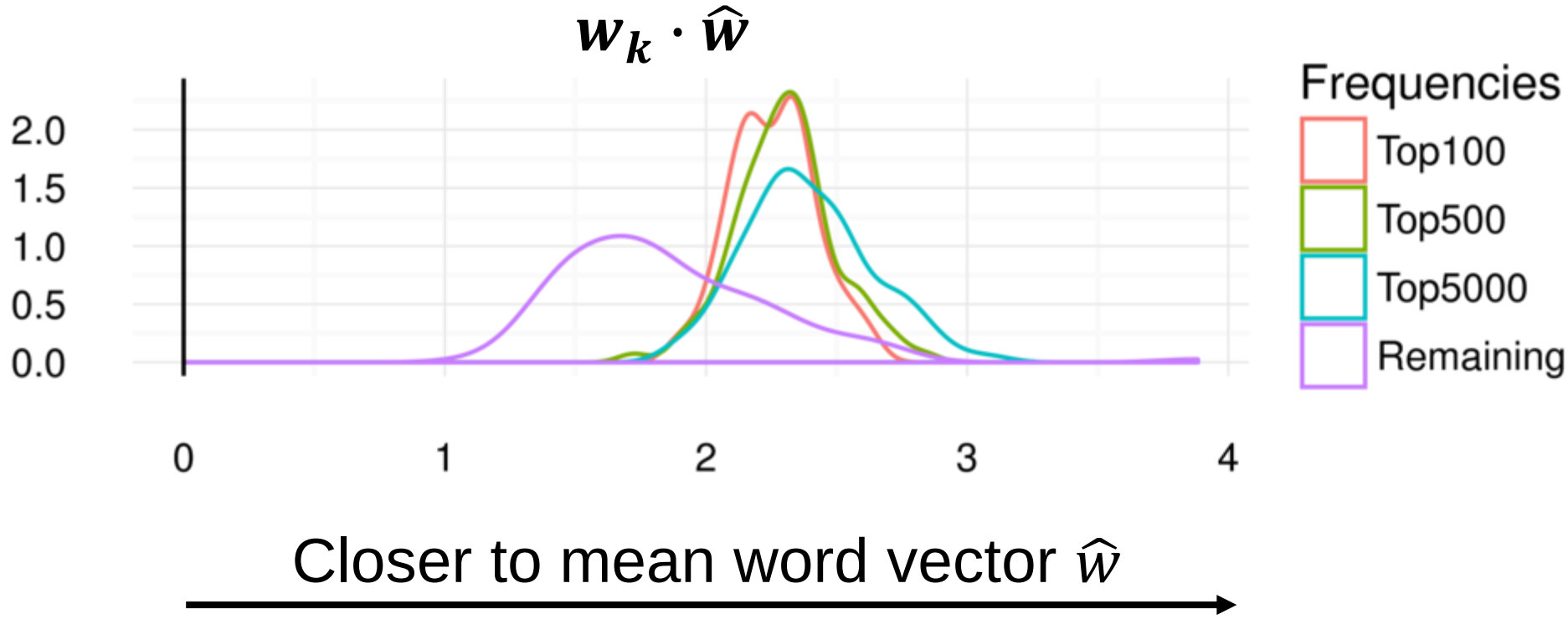
Artifact of word frequency?

4 Frequency Levels:

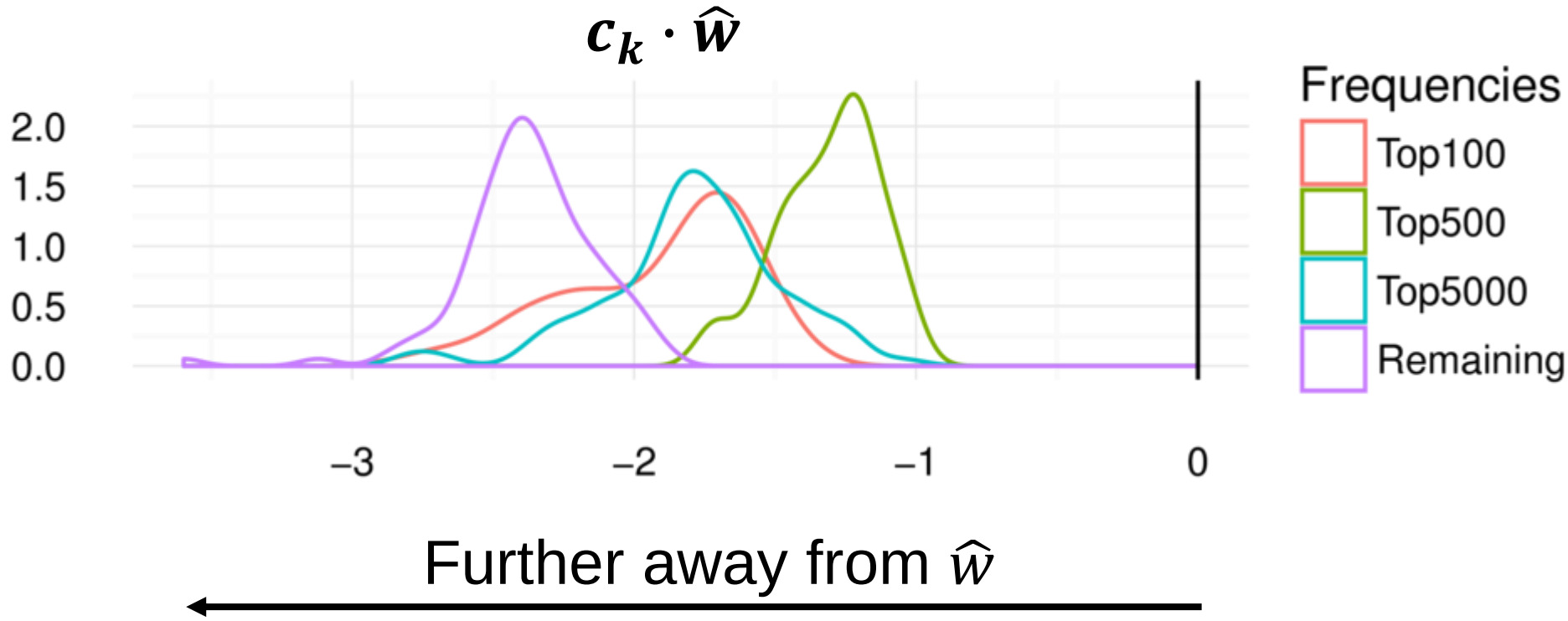
- Ultra-high (1–100)
- High (101–500)
- Moderate (501–5000)
- Low (5001+)

- Sample 100 from each
- Use sample mean vector $\hat{w}$ instead of global mean $\bar{w}$

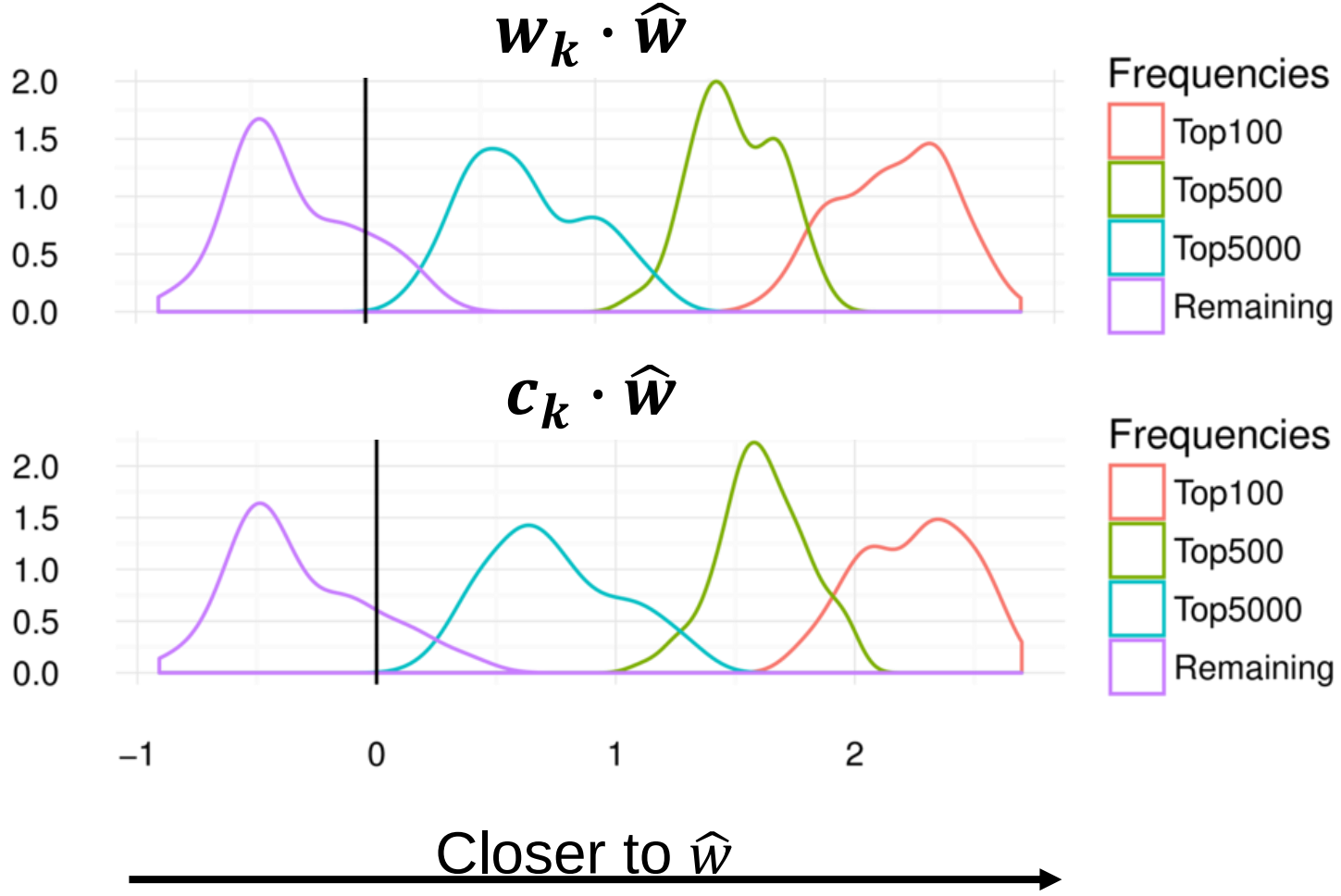
…true for all frequency classes



…and away from the context vectors

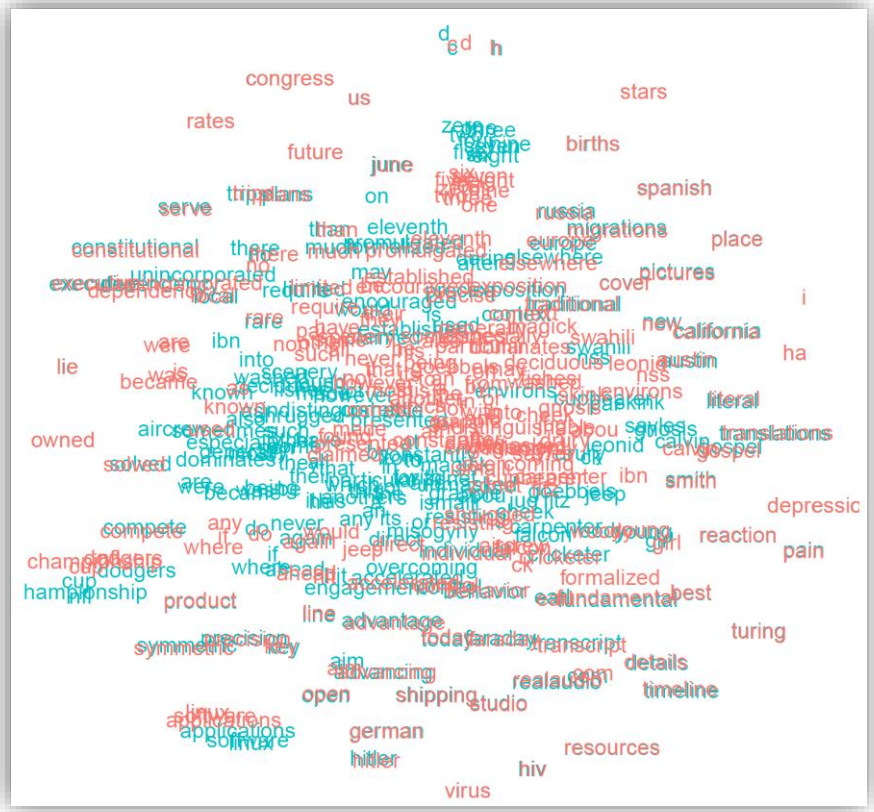


Not true for GloVe



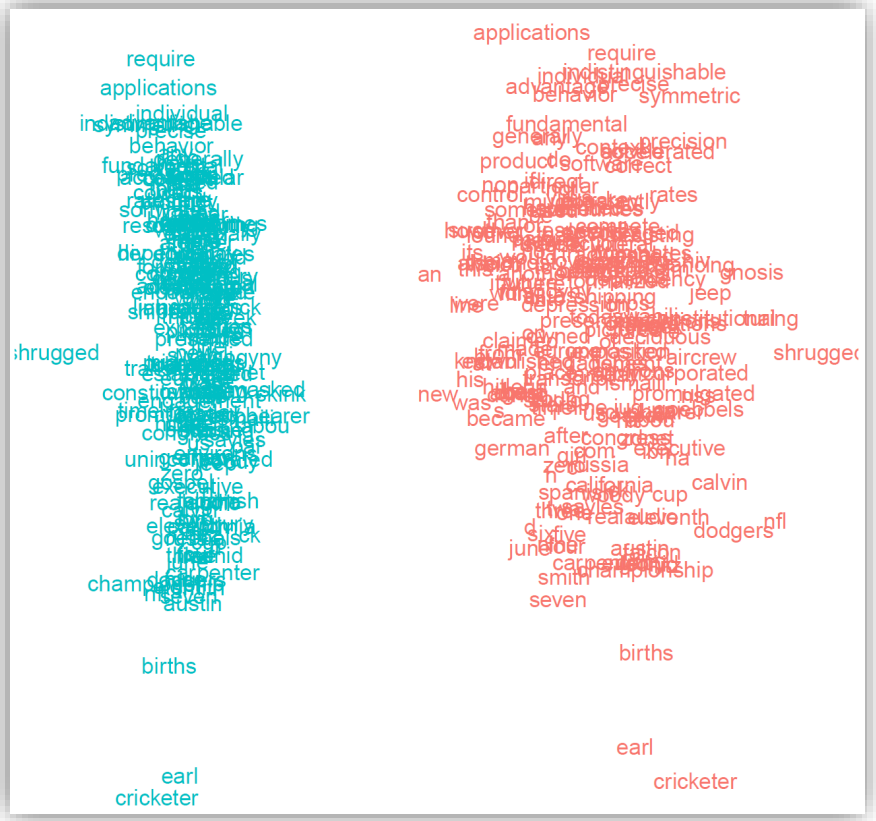
Visualization Disconnect

TSNE



VS

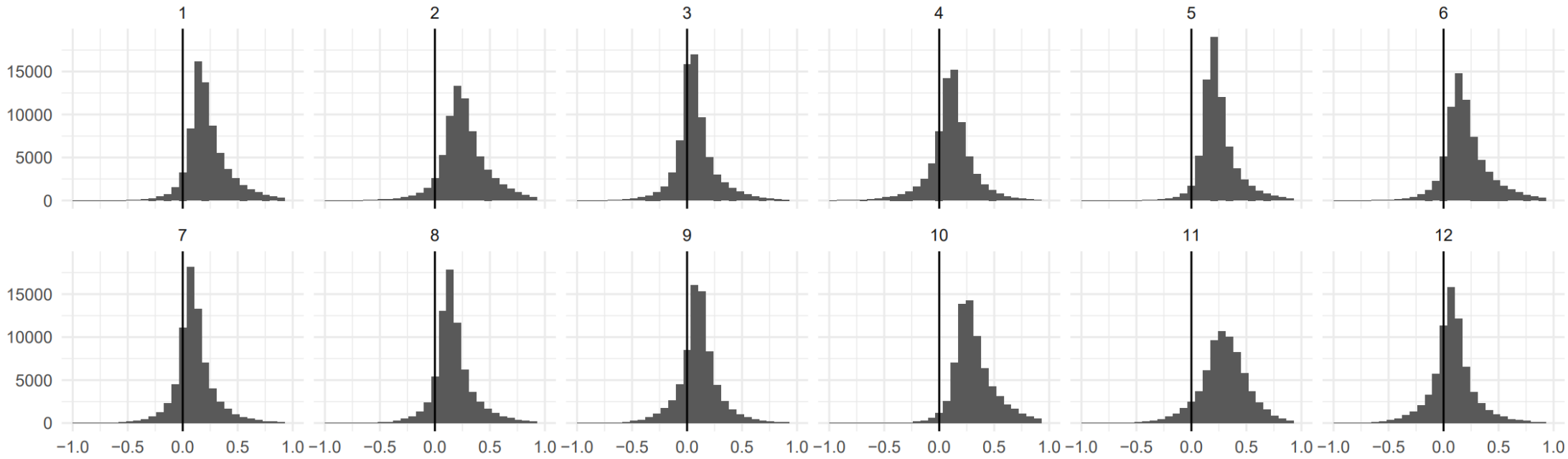
PCA



Observation #2:

SGNS vectors are mostly non-negative

Latent dimensions skew “positive”



…inefficient use of K-dimensional space?

Preserve Semantic Properties:

1. Dropping “negative” entries
   $w'_k = \max(0, w_k * sign(\overline{w}_k))$

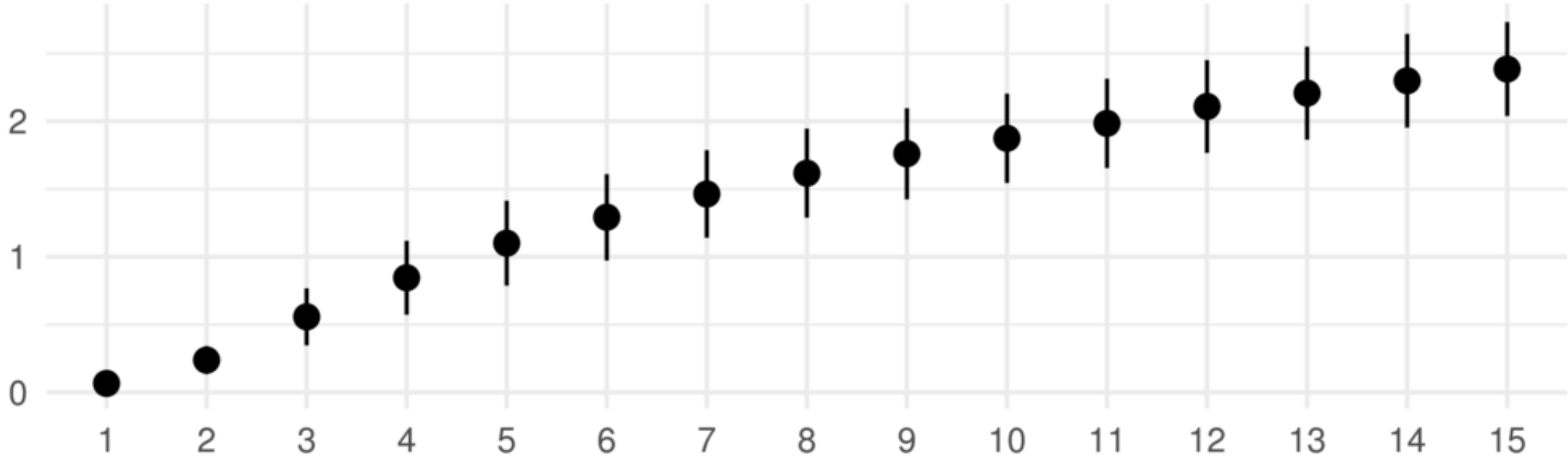
2. Subtracting mean vector
   $w' = w - \overline{w}$

Observation #3:

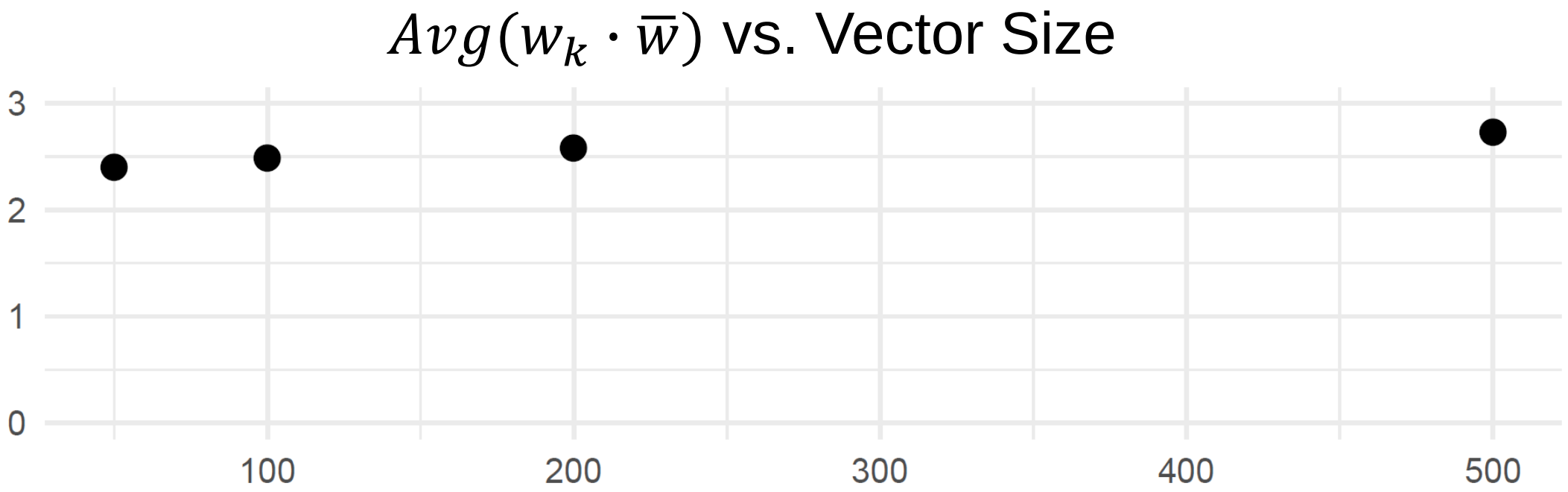
Negative sampling affects SGNS geometry

More negative samples, better alignment

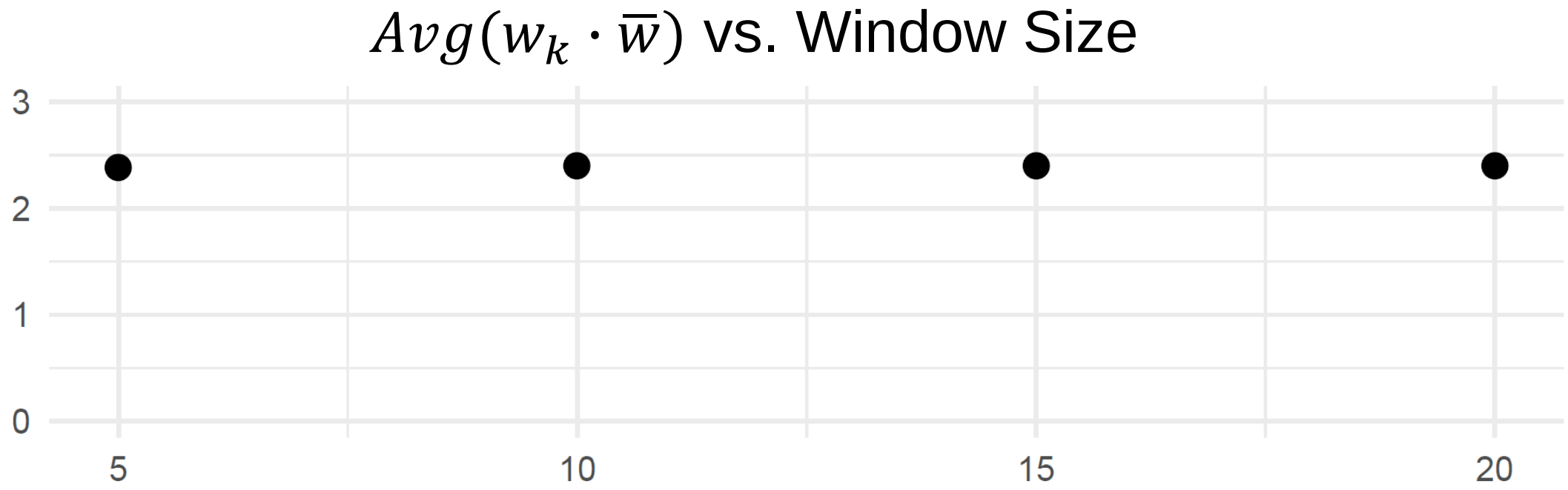
$Avg(w_k \cdot \bar{w})$ vs. # of Negative Samples



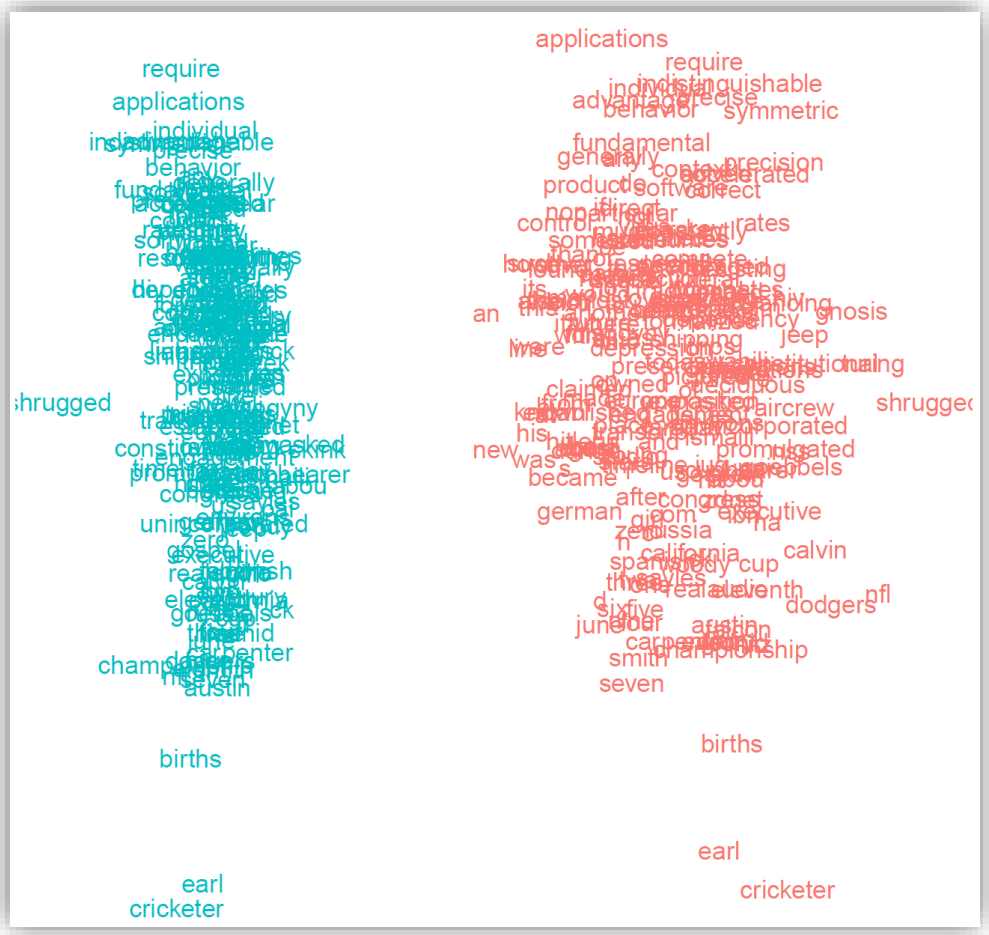
…this is not seen in other parameters



…this is not seen in other parameters



Thank You!


require
applications
individual
behavior
shrugged
constitutional
champion
austin
births
earl
cricketer
applications
require
indistinguishable
advantage
precise
behavior
symmetric
fundamental
precision
product
software
correct
rates
an
jeep
shrugged
aircrew
new
became
german
california
calvin
dodgers
nfl
championship
smith
seven
births
earl
cricketer