Limitations of Word Embeddings

Laura Burdick
wenlaura@umich.edu
http://laura-burdick.github.io/

Image from www.bbc.co.uk

Limitations of Word Embeddings

1. Word embeddings are built using lots of data.

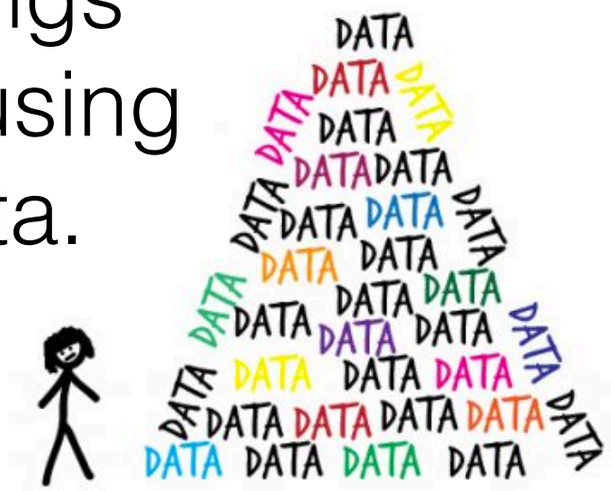
2. Word embeddings are unstable.

3. Word embeddings are hard to interpret.

Images from www.drugdesigntech.com, https://raphaelhertzog.com, https://thehappybody.com

Limitations of Word Embeddings

1. Word embeddings are built using lots of data.



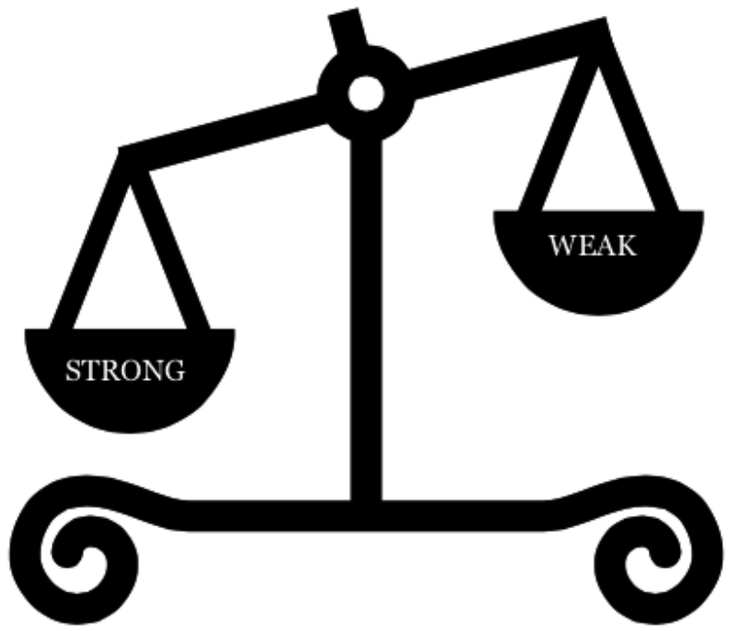
2. Word embeddings are unstable.

3. Word embeddings are hard to interpret.

Images from www.drugdesigntech.com, https://raphaelhertzog.com, https://thehappybody.com

Word embeddings are built using lots of data.

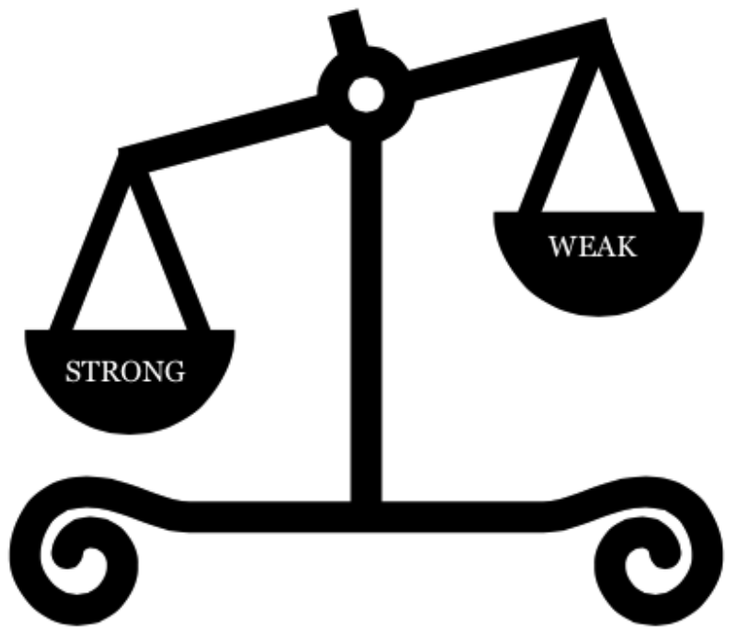
- Strength? Or limitation?



Word embeddings are built using lots of data.

Strengths:

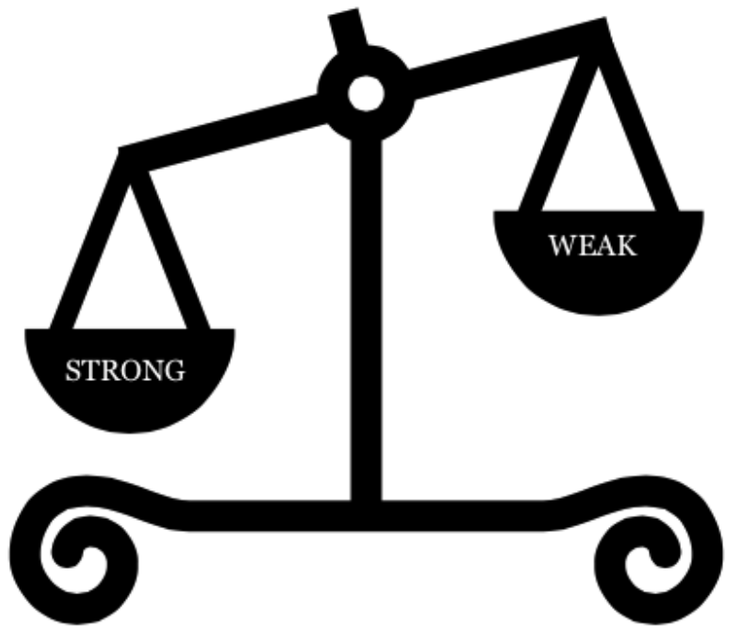
- ☑ More data than expert-curated resources
- ☑ Huge variety of language use across different topics & situations
- ☑ Language as used in real life
- ☑ Internet stays up-to-date with slang, expressions, etc.



Word embeddings are built using lots of data.

Weaknesses:

- ☑ Flaws in data create flaws in word embedding models (bias, nasty language, inappropriate topics)



- ☑ No dataset covers every variation in language

Your data shapes your word embeddings, for better and for worse.

- Be mindful of what data you're using.
- Be aware of what problems your data could cause further in your experimental pipeline.

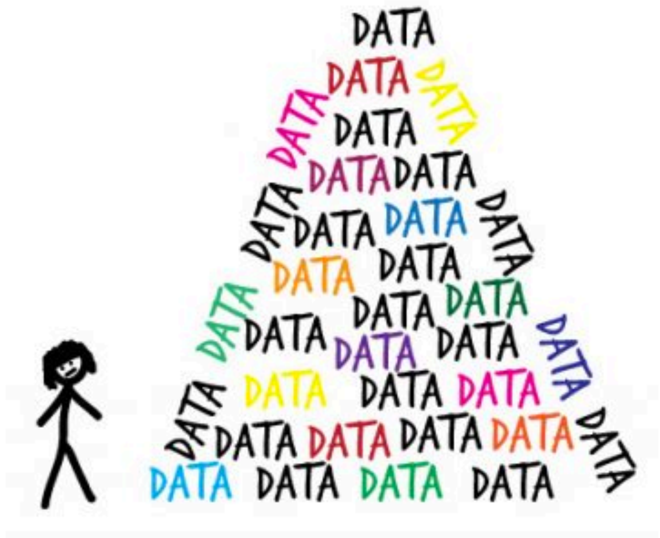


Image from www.drugdesigntech.com

 in Data

- *What kind of bias are we talking about?*
- *Is there bias in my data?*
- *Is bias a bad thing?*
- *Can we remove bias from data?*

BIAS in Data

- *What kind of bias are we talking about?*

Racial

Gender

Political,
Cultural,
etc…

Images from www.capartscentre.com, http://cem-mi-vancouver2019.ca

BIAS in Data

- *Is there bias in my data?* **YES.**

Implicit Association Test (IAT)

- Human response time

Word Embedding Association Test (WEAT)

- cosine similarity between a pair of vectors

Caliskan, Aylin, et al. "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 (2017): 183-186.

BIAS in Data

- *Is there bias in my data?* **YES.**

Target Words	*Attribute Words*	*(Humans) d*	*(Humans) P*	*(Emb.) d*	*(Emb.) P*
Male vs. female names	Career vs. family	0.72	$<10^{-2}$	1.81	10^{-3}
Math vs. Arts	Male vs. female terms	0.83	$<10^{-2}$	1.06	0.018
Science vs. Arts	Male vs. female terms	1.47	10^{-24}	1.24	10^{-2}

Caliskan, Aylin, et al. "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 (2017): 183-186.

 in Data

- Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." *Advances in Neural Information Processing Systems*. 2016.

- Caliskan, Aylin, et al. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356.6334 (2017): 183-186.

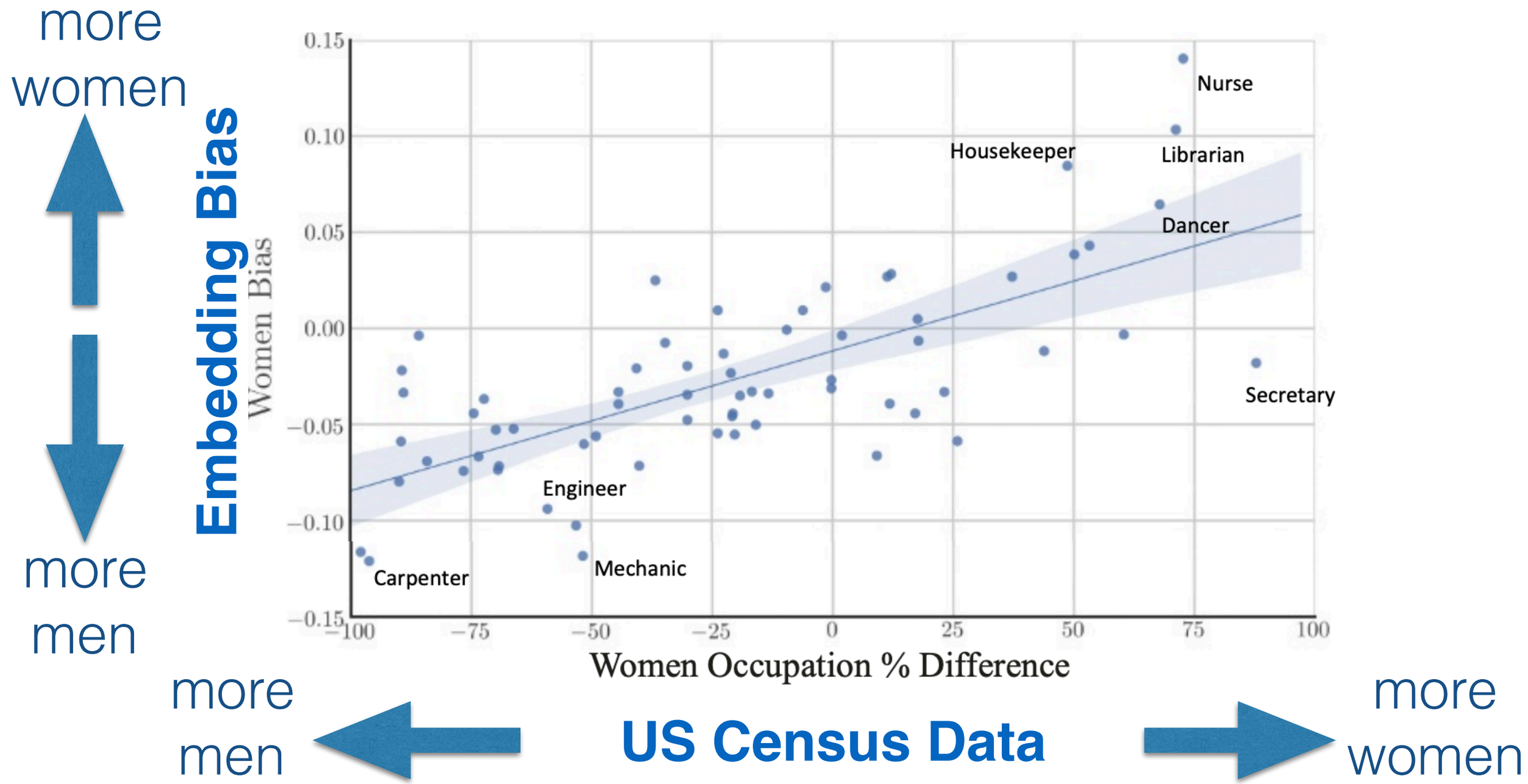
- Garg, Nikhil, et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes." *Proceedings of the National Academy of Sciences* 115.16 (2018): E3635-E3644.

- Swinger, Nathaniel, et al. "What are the biases in my word embedding?." *AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

- Lu, Kaiji, et al. "Gender bias in neural natural language processing." *arXiv* preprint arXiv:1807.11714 (2018).

BIAS in Data

- *Is bias a bad thing?*



Garg, Nikhil, et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes." Proceedings of the National Academy of Sciences 115.16 (2018): E3635-E3644.

BIAS in Data

- *Is bias a bad thing?*

$1_{\square}$: The **doctor** ran because **he** is late. (5.08)

$1_{\bigcirc}$: The **doctor** ran because **she** is late. (1.99)

$2_{\square}$: The **nurse** ran because **he** is late. (−0.44)

$2_{\bigcirc}$: The **nurse** ran because **she** is late. (5.34)

(a) Coreference resolution

	A	B	$\ln \Pr[B \mid A]$
$1_{\square}$:	**He** is a	**doctor**.	-9.72
$1_{\bigcirc}$:	**She** is a	**doctor**.	-9.77
$2_{\square}$:	**He** is a	**nurse**.	-8.99
$2_{\bigcirc}$:	**She** is a	**nurse**.	-8.97

(b) Language modeling

Lu, Kaiji, et al. "Gender bias in neural natural language processing." arXiv preprint arXiv:1807.11714 (2018).

in Data

- *Can we remove bias from data?*
  - Still an active research question
  - Some partial solutions, no full solutions
  - Download partially de-biased word embeddings: https://github.com/tolga-b/debiaswe

 in Data

- Gonen, Hila, and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them." *NAACL-HLT*. 2019.

- Zhao, Jieyu, et al. "Learning Gender-Neutral Word Embeddings." *Conference on Empirical Methods in Natural Language Processing*. 2018.

- Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." *Advances in Neural Information Processing Systems*. 2016.

- Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning." *AAAI/ACM Conference on AI, Ethics, and Society*. 2018.

Your data shapes your word embeddings, for better and for worse.

- Be mindful of what data you're using.
- Be aware of what problems your data could cause further in your experimental pipeline.

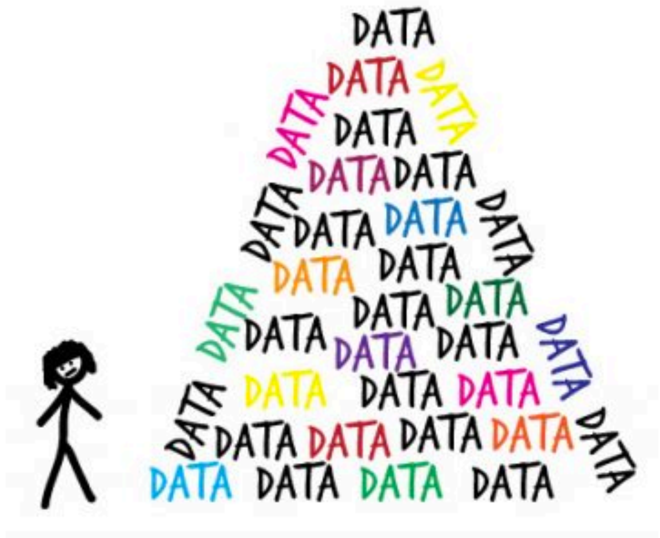


Image from www.drugdesigntech.com

Pay Attention to Topic

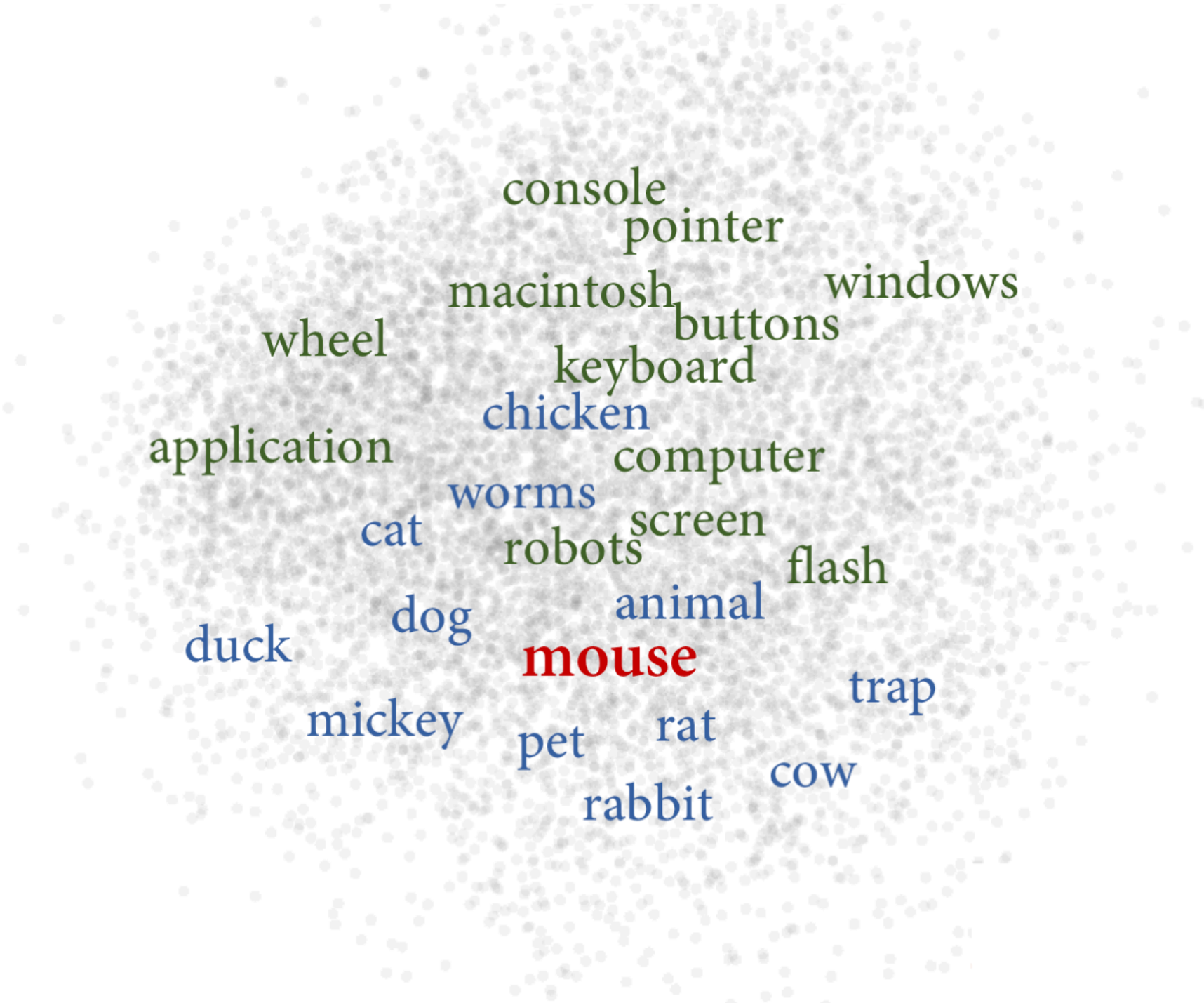


Image from Jose Camacho Collados and Taher Pilehvar

Pay Attention to Topic

- Match topic of your embeddings to your application

- General-purpose or specific embeddings? Depends on your application.

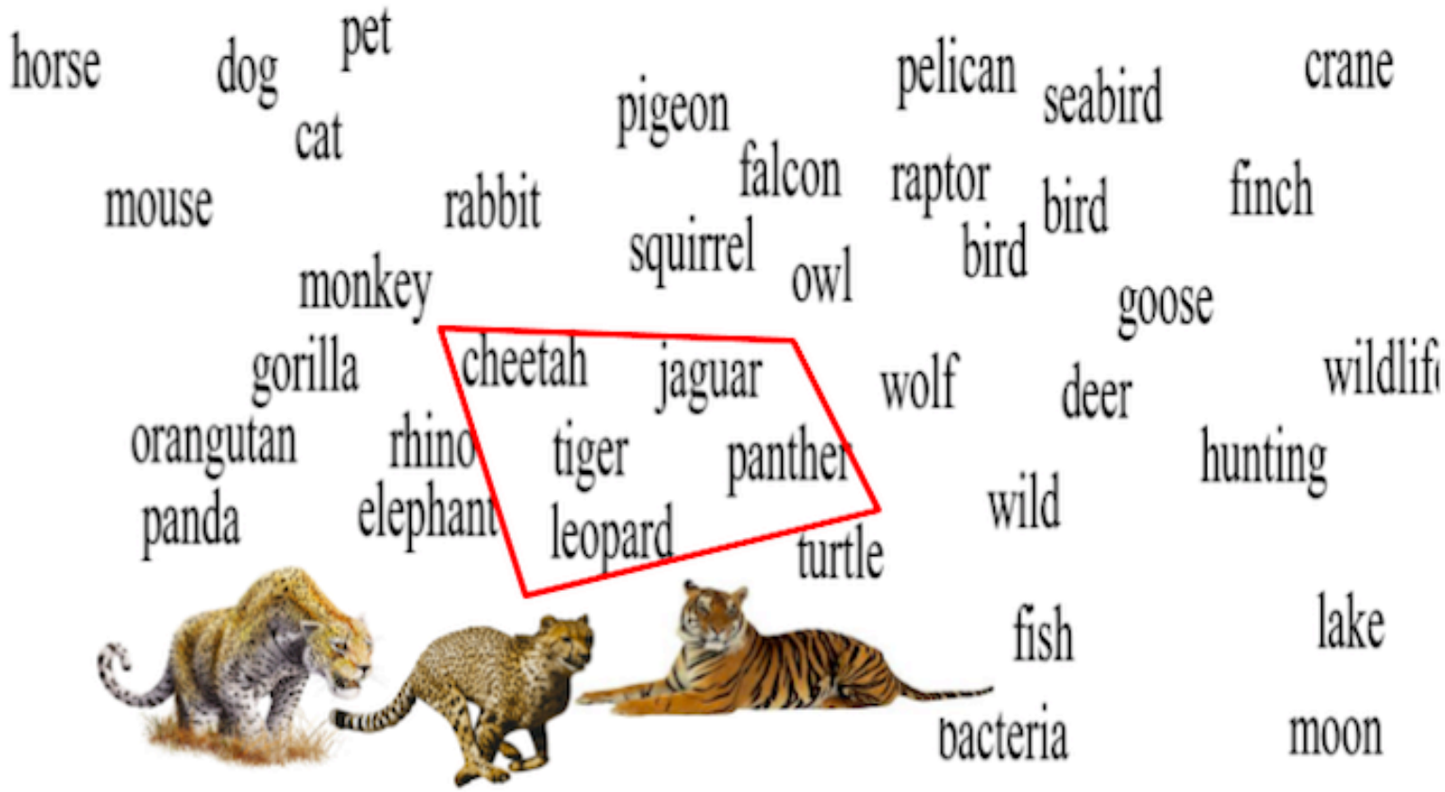


Image from Jose Camacho Collados

Limitations of Word Embeddings

1. Word embeddings are built using lots of data.

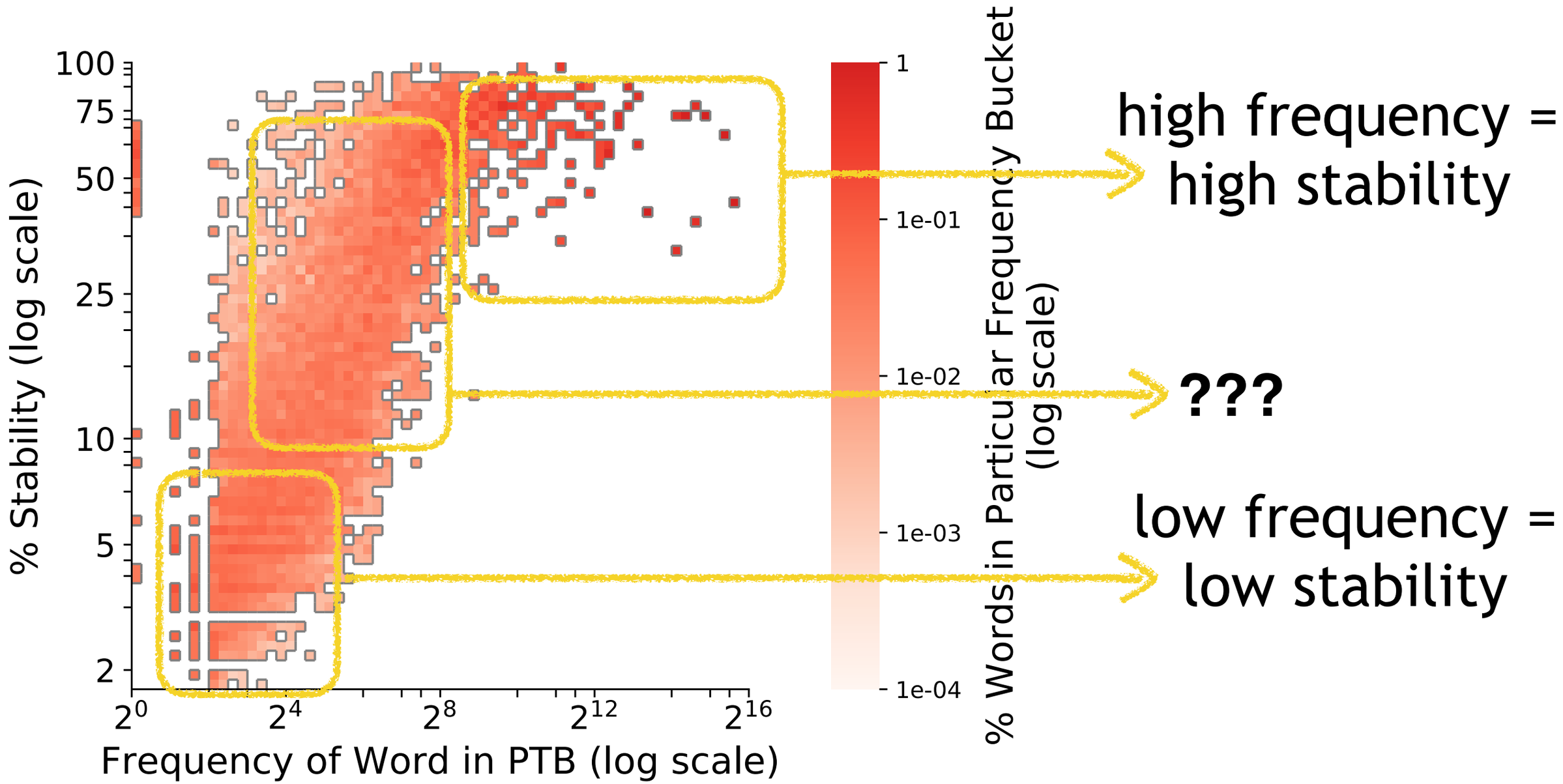
2. Word embeddings are unstable.

3. Word embeddings are hard to interpret.

Images from www.drugdesigntech.com, https://raphaelhertzog.com, https://thehappybody.com

The Problem

Many common embedding algorithms have large amounts of instability.



What is Stability?

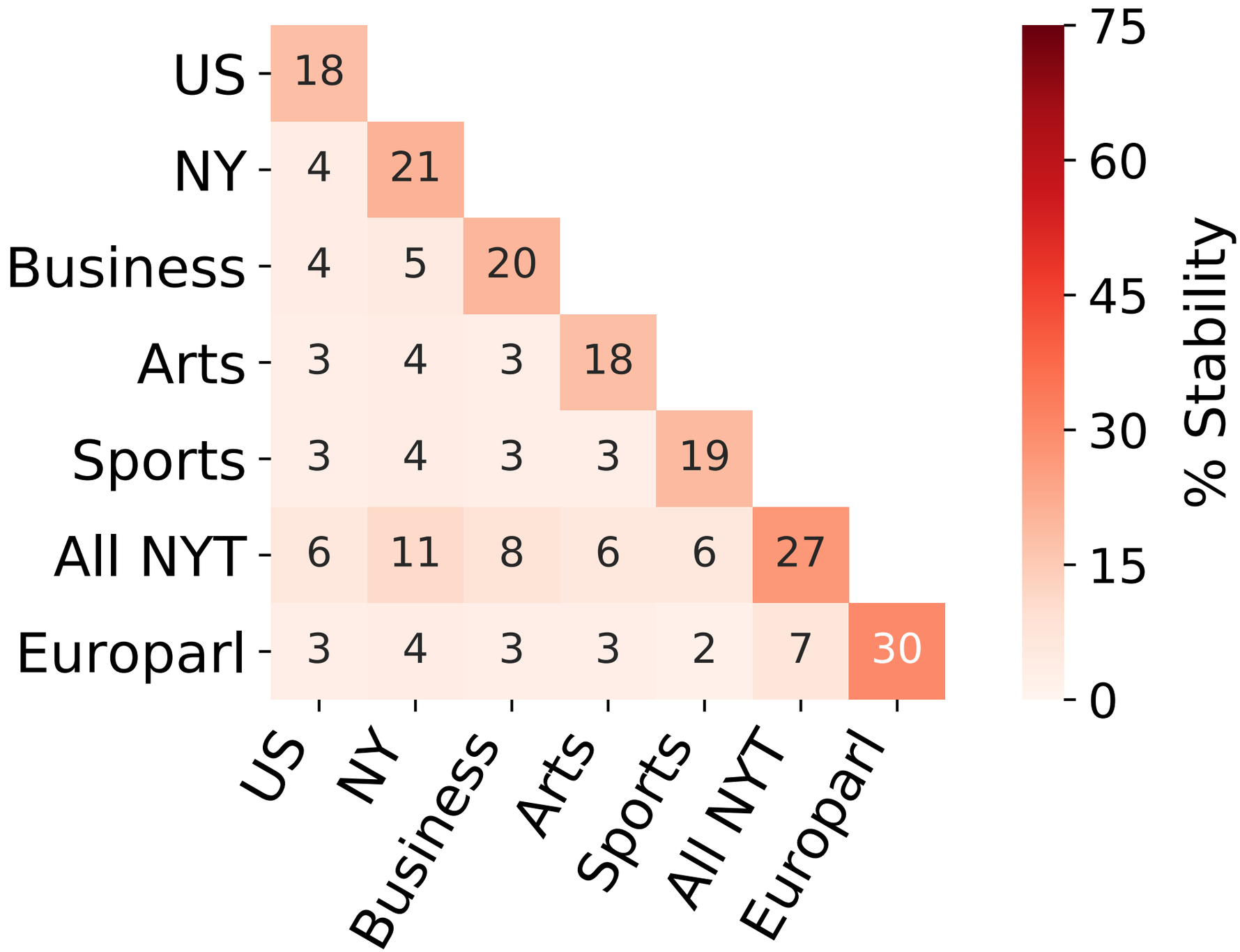
Stability = *percent overlap between ten nearest neighbors in an embedding space*

philadelphia
national
metropolitan
egyptian
international
folk
rhode
society
chairwoman
debut
reinstallation

metropolitan
exhibitions
ballet
bard
international
national
chicago
state
society
whitney
rhode

Stability = 40%

Stability within domains is greater than across domains.



Your data shapes your word embeddings, for better and for worse.

- Be mindful of what data you're using.
- Be aware of what problems your data could cause further in your experimental pipeline.

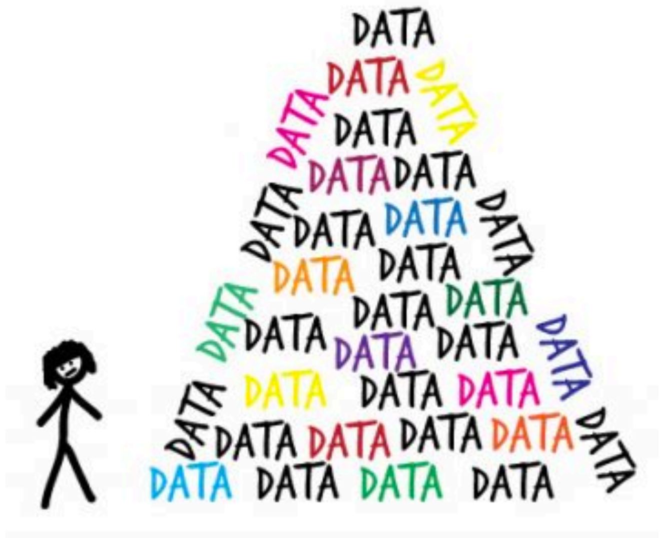
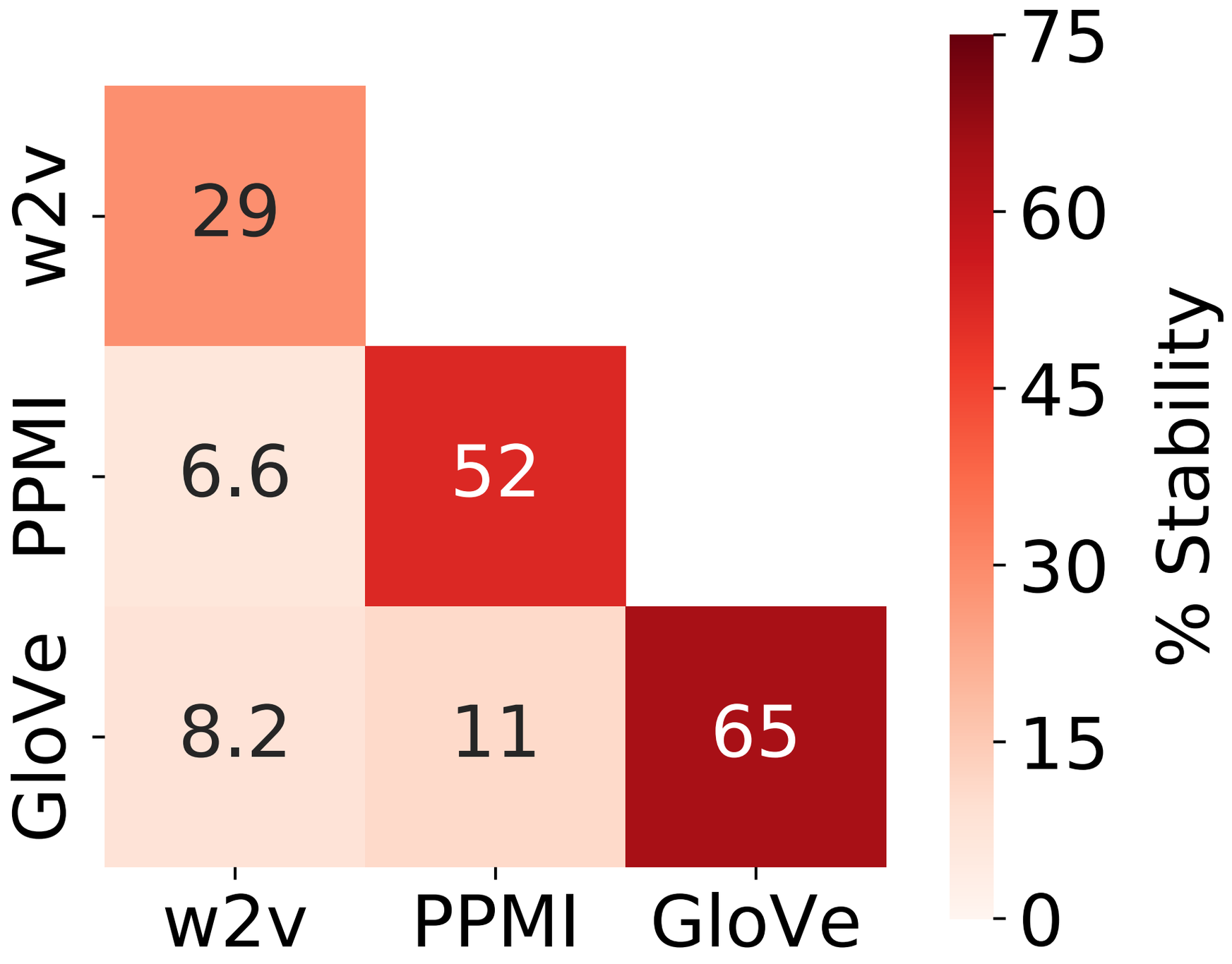


Image from www.drugdesigntech.com

Different word embedding algorithms are more or less stable.



Word Embedding Algorithms

- Pennington, Jeffrey, et al. "GloVe: Global vectors for word representation." *Conference on Empirical Methods in Natural Language Processing*. 2014.

- Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in Neural Information Processing Systems*. 2013.

- Bullinaria, John A., and Joseph P. Levy. "Extracting semantic representations from word co-occurrence statistics: A computational study." *Behavior research methods* 39.3 (2007): 510-526.

- Peters, Matthew E., et al. "Deep contextualized word representations." *NAACL-HLT*. 2018.

- Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL-HLT*. 2019.

Stability affects some downstream tasks.

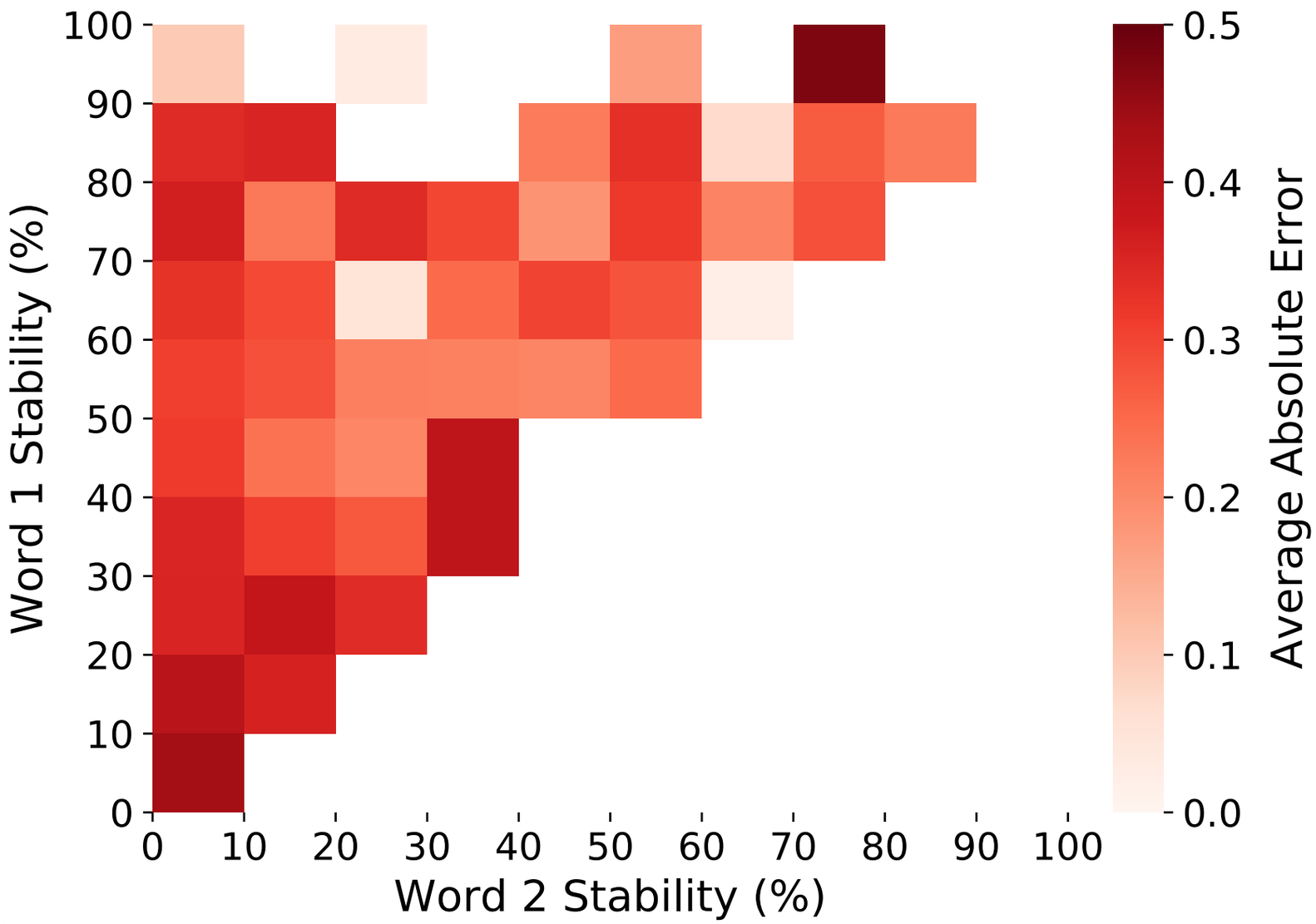
Word Similarity Task:

Sun 	Sunlight 	100%
Grapes 	Wine 	92%
Hummingbird 	Stork 	76%
Mushrooms 	Salad 	54%
Frog 	Poppy 	24%

Multimodal Distributional Semantics E. Bruni, N. K. Tran and M. Baroni. Journal of Artificial Intelligence Research 49: 1-47.
Images from https://solarsystem.nasa.gov, www.thepacker.com, www.amazon.com, www.dkfindout.com, www.onlinelabels.com, www.petmd.com, https://ala.pa-legion.com

Stability affects some downstream tasks.

Word stability correlates slightly with performance on word similarity tasks.



Stability affects some downstream tasks.

Part-of-speech (POS) tagging:

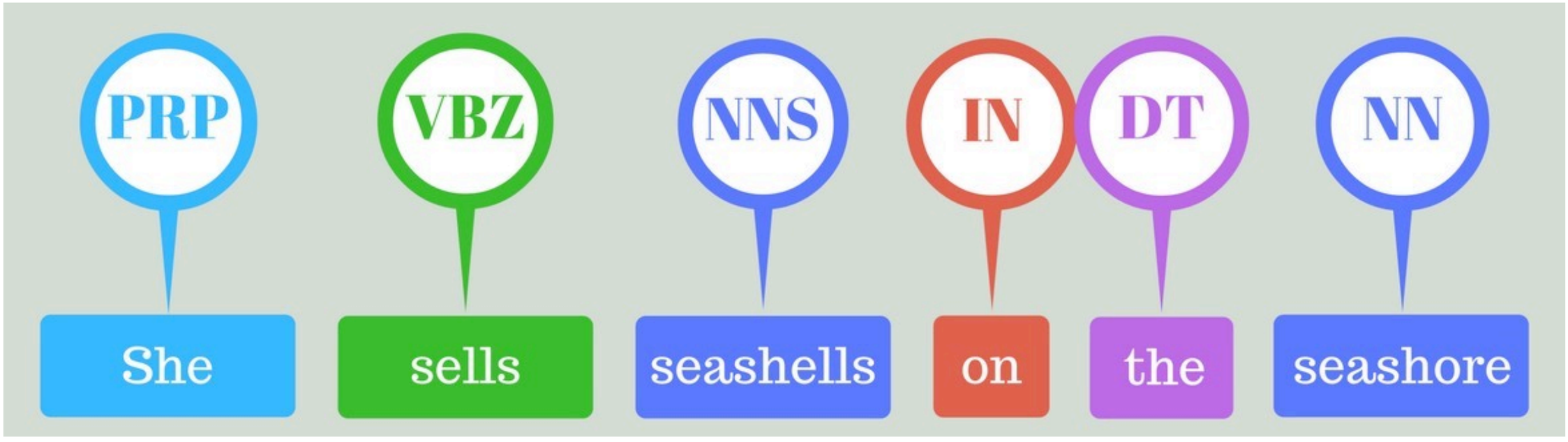


Image from www.learntek.org

Stability affects some downstream tasks.

Here, we use an LSTM architecture.

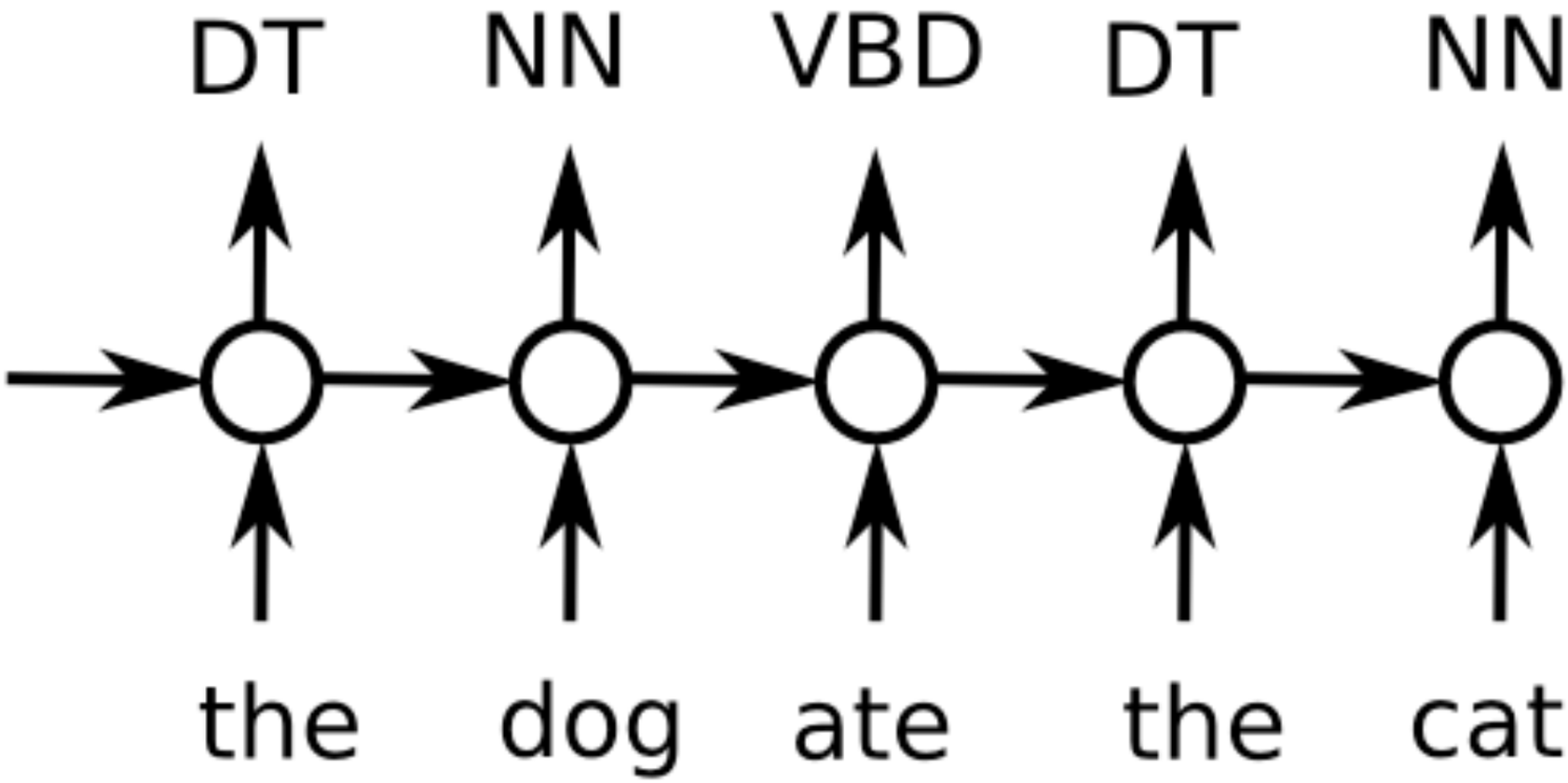
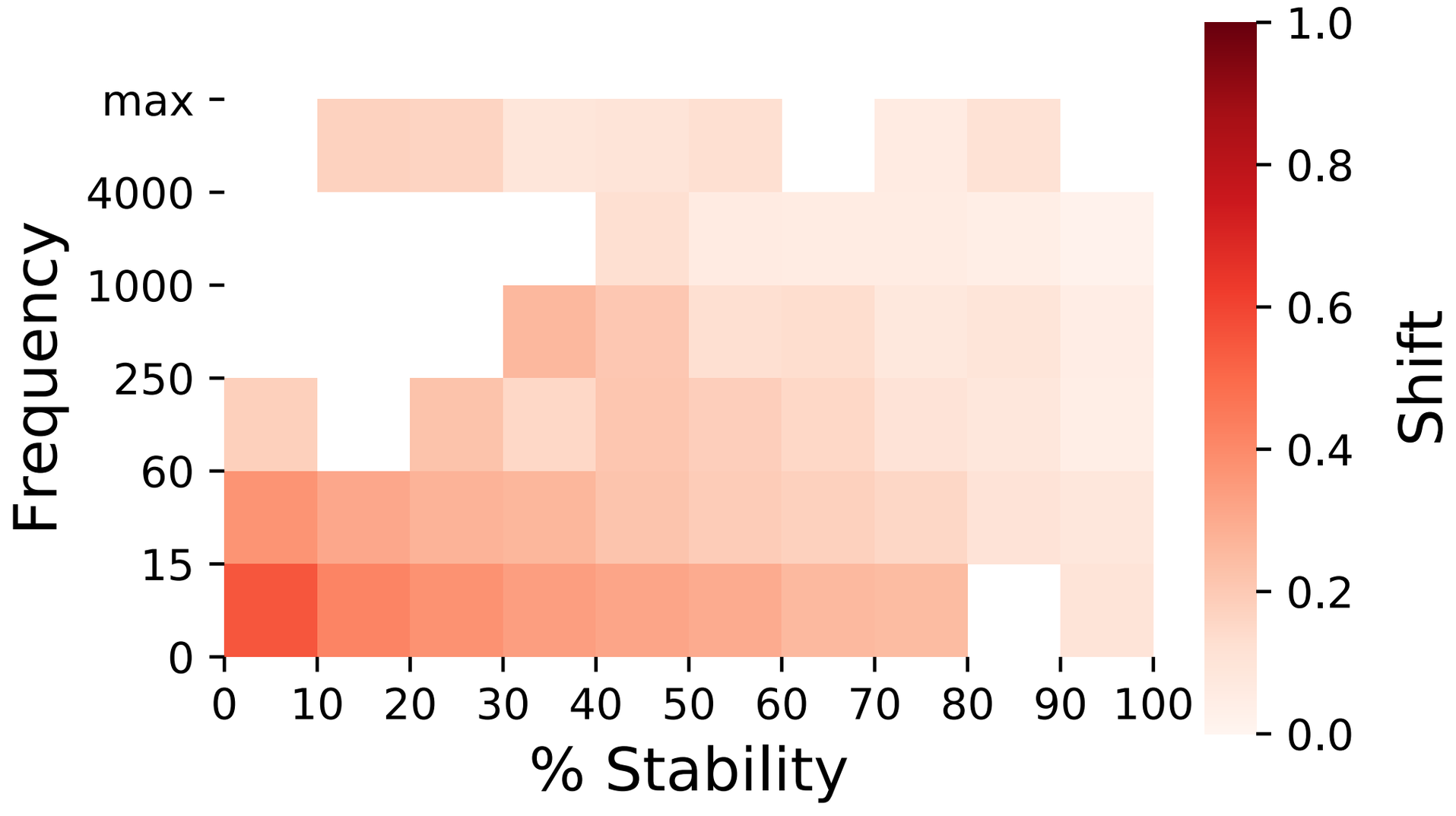


Image from mat kelcey on https://stackoverflow.com/

Stability affects some downstream tasks.

For POS tagging using an LSTM, the LSTM compensates for instability by shifting unstable word vectors.



Analyzing Language Using Word Embeddings

- What if you want to study language / language change using embeddings?
- “Corpus-centered” v. “downstream-centered”
- “Unlike the downstream-centered approach, the corpus-centered approach is based on direct human analysis of nearest neighbors to embedding vectors, and the training corpus is not simply an off-the-shelf convenience but rather the central object of study.”

Antoniak, Maria, and David Mimno. "Evaluating the stability of embedding-based word similarities." Transactions of the Association for Computational Linguistics 6 (2018): 107-119.

Analyzing Language Using Word Embeddings

Corpus-Centered	Downstream-Centered
Big corpus	Small corpus, difficult or impossible to expand
Source is not important	Source is the object of study
Only vectors are important	Specific, fine-grained comparisons are important
Embeddings are used in downstream tasks	Embeddings are used to learn about the mental model of word association for the authors of the corpus

Antoniak, Maria, and David Mimno. "Evaluating the stability of embedding-based word similarities." Transactions of the Association for Computational Linguistics 6 (2018): 107-119.

Corpus-Centered Usage of Embeddings

- Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." *Advances in Neural Information Processing Systems*. 2016.

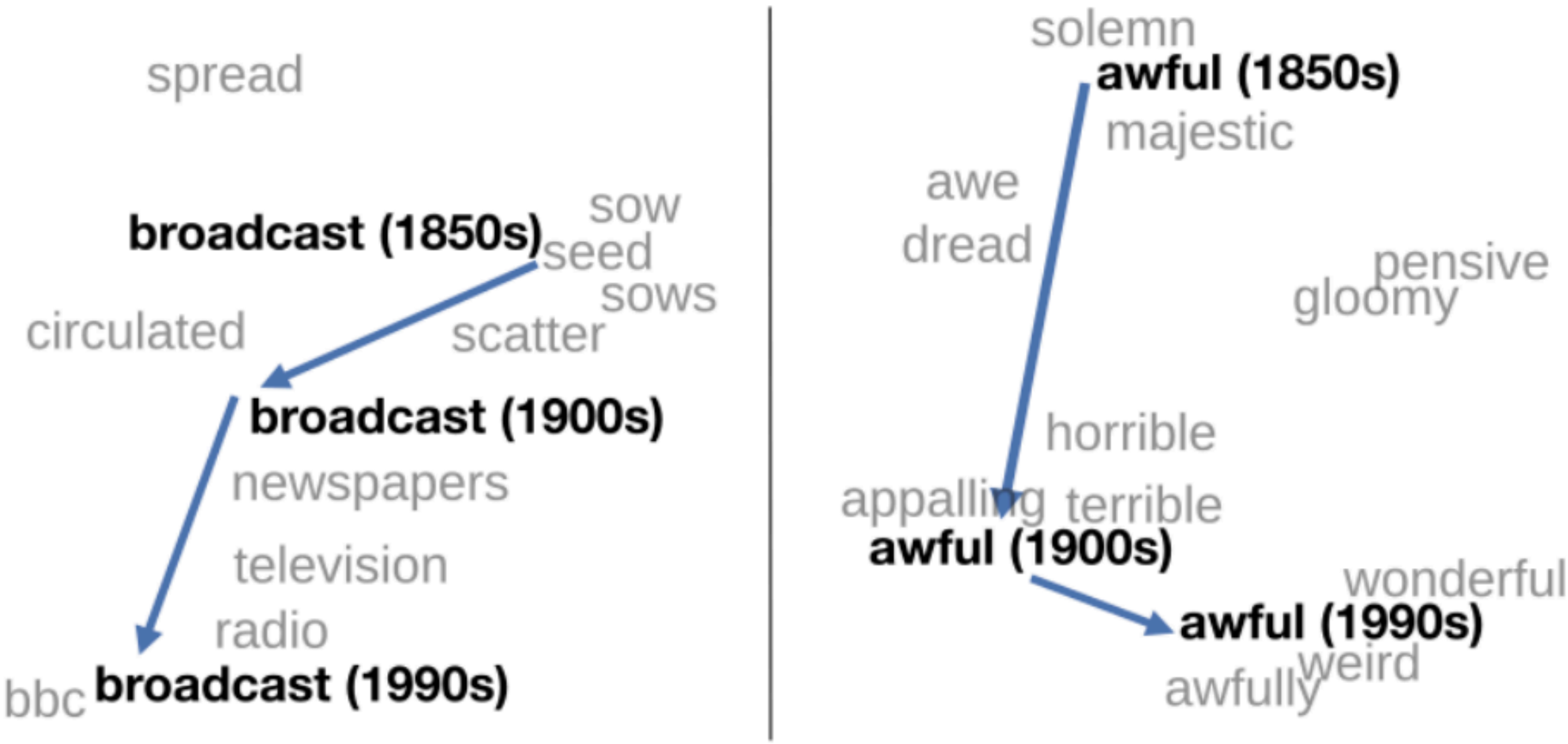
- Heuser, Ryan James. "Word Vectors in the Eighteenth Century." *DH*. 2017.

- Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." *Association for Computational Linguistics*. 2016.

- Grayson, S., et al. Novel2Vec: Characterising 19th Century Fiction via Word Embeddings. *Irish Conference on Artificial Intelligence and Cognitive Science*. 2016.

- Maslinsky, K. From history of emotions back to literary history: a case of Soviet realistic prose for children and young adults. *Estonian Digital Humanities Conference Data, humanities & language: tools & applications*. 2018.

Analyzing Language Using Word Embeddings



Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.

Analyzing Language Using Word Embeddings

Most similar words to "**pregnancy**" for 4 random runs:

Run 1	Run 2	Run 3	Run 4
viability	fetus	trimester	surgery
pregnancies	pregnancies	surgery	visit
abortion	gestation	visit	therapy
abortions	kindergarten	tenure	pain
fetus	viability	workday	hospitalization
gestation	headaches	abortions	neck
surgery	pregnant	hernia	headaches
expiration	abortion	summer	trimester
sudden	pain	suicide	experiencing
fetal	bladder	abortion	medications

Antoniak, Maria, and David Mimno. "Evaluating the stability of embedding-based word similarities." Transactions of the Association for Computational Linguistics 6 (2018): 107-119.

Analyzing Language Using Word Embeddings

- What if you want to study language / language change using embeddings?

- *Run your analysis on multiple bootstrapped embedding spaces and report statistical confidence.*

- *Look at groups of words, rather than individual words.*

Word Embedding Instability

- Wendlandt, Laura, Jonathan K. Kummerfeld, and Rada Mihalcea. "Factors Influencing the Surprising Instability of Word Embeddings." *NAACL-HLT*. 2018.

- Antoniak, Maria, and David Mimno. "Evaluating the stability of embedding-based word similarities." *Transactions of the Association for Computational Linguistics* 6 (2018): 107-119.

- Pierrejean, Bénédicte, and Ludovic Tanguy. "Predicting word embeddings variability." *Seventh Joint Conference on Lexical and Computational Semantics*. 2018.

Limitations of Word Embeddings

1. Word embeddings are built using lots of data.

2. Word embeddings are unstable.

3. Word embeddings are hard to interpret.

Images from www.drugdesigntech.com, https://raphaelhertzog.com, https://thehappybody.com

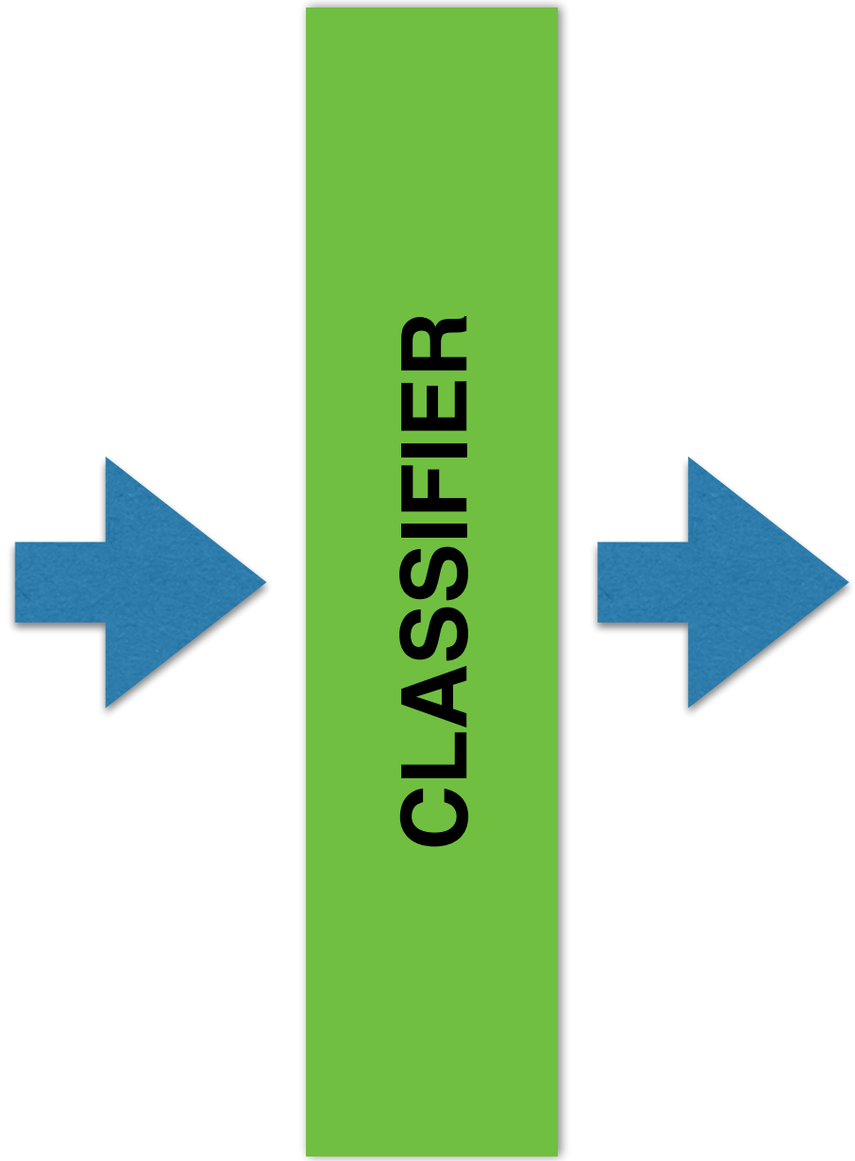
The Temptation

joyful	melancholy
-0.7	0.95
0.55	1.3
0.63	0.7
…	…
-0.34	0.55
0.22	1.32
0.03	0.02
2.1	-1.32
-0.3	-0.7

This dimension must be about sentiment!

A Better Way: Train a Classifier

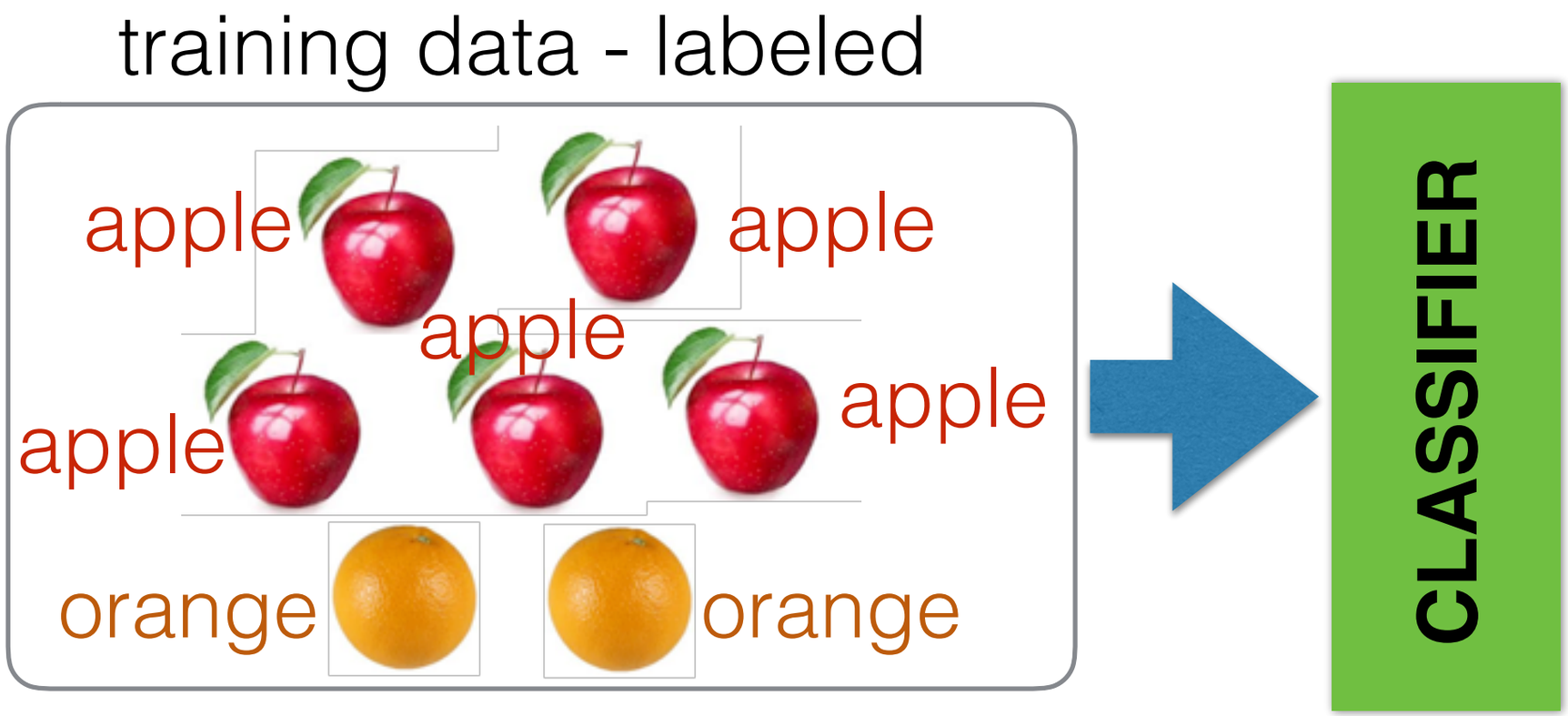
happy	sad
-0.7	0.95
0.55	1.3
0.63	0.7
…	…
-0.34	0.55
0.22	1.32
0.03	0.02
2.1	-1.32
-0.3	-0.7



positive
or
negative

Machine Learning Classifiers

1. Train the Classifier



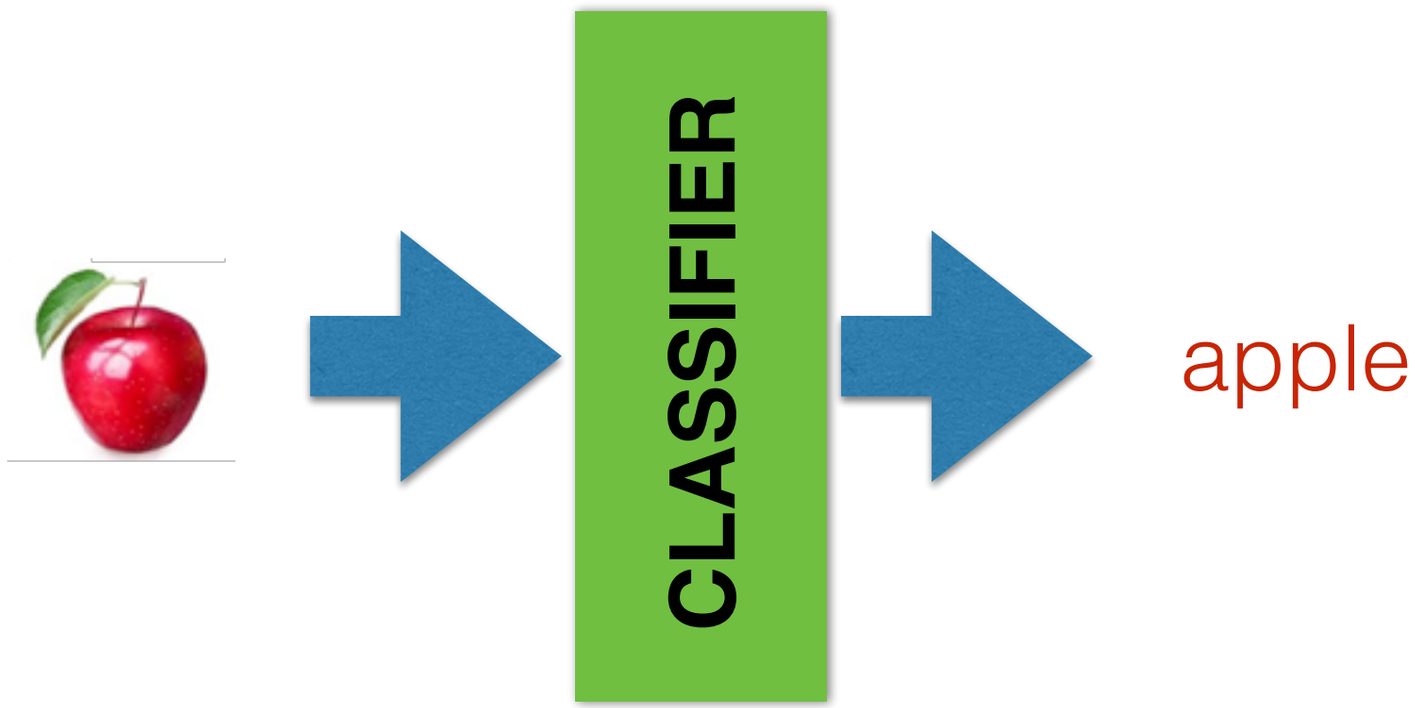
The classifier (hopefully) learns distinctive features of apples and oranges.

Image from www.kdnuggets.com

Machine Learning Classifiers

1. Test the Classifier

test data - unlabeled



The classifier uses what it learned at training time to make predictions.

Image from www.kdnuggets.com

Interpreting Word Embeddings

- Subramanian, Anant, et al. “SPINE: Sparse interpretable neural embeddings." *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

- Murphy, Brian, et al. "Learning effective and interpretable semantic models using non-negative sparse embedding." *COLING*. 2012.

- Erk, Katrin. "What do you know about an alligator when you know the company it keeps?" *Semantics and Pragmatics* 9 (2016): 17-1.

- Erk, Katrin. "Vector space models of word meaning and phrase meaning: A survey." *Language and Linguistics Compass* 6.10 (2012): 635-653.

- Li, Jiwei, et al. "Understanding neural networks through representation erasure." *arXiv* preprint arXiv:1612.08220 (2016).

- Fyshe, Alona, et al. "A compositional and interpretable semantic space." NAACL-HLT. 2015.

Limitations of Word Embeddings

1. Word embeddings are built using lots of data.

2. Word embeddings are unstable.

3. Word embeddings are hard to interpret.

Images from www.drugdesigntech.com, https://raphaelhertzog.com, https://thehappybody.com

Questions?

Laura Burdick
wenlaura@umich.edu
http://laura-burdick.github.io/

Image from www.bbc.co.uk