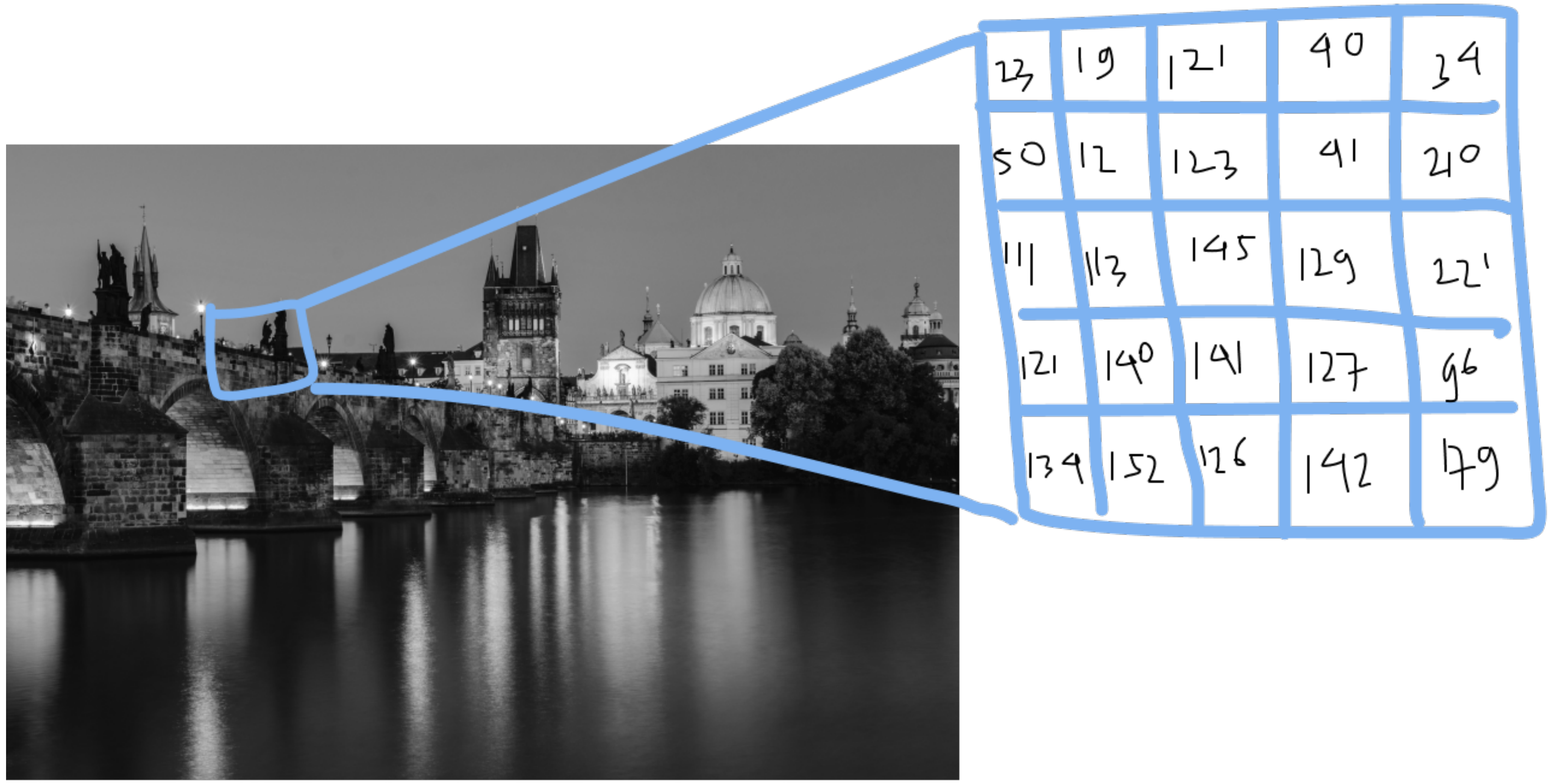
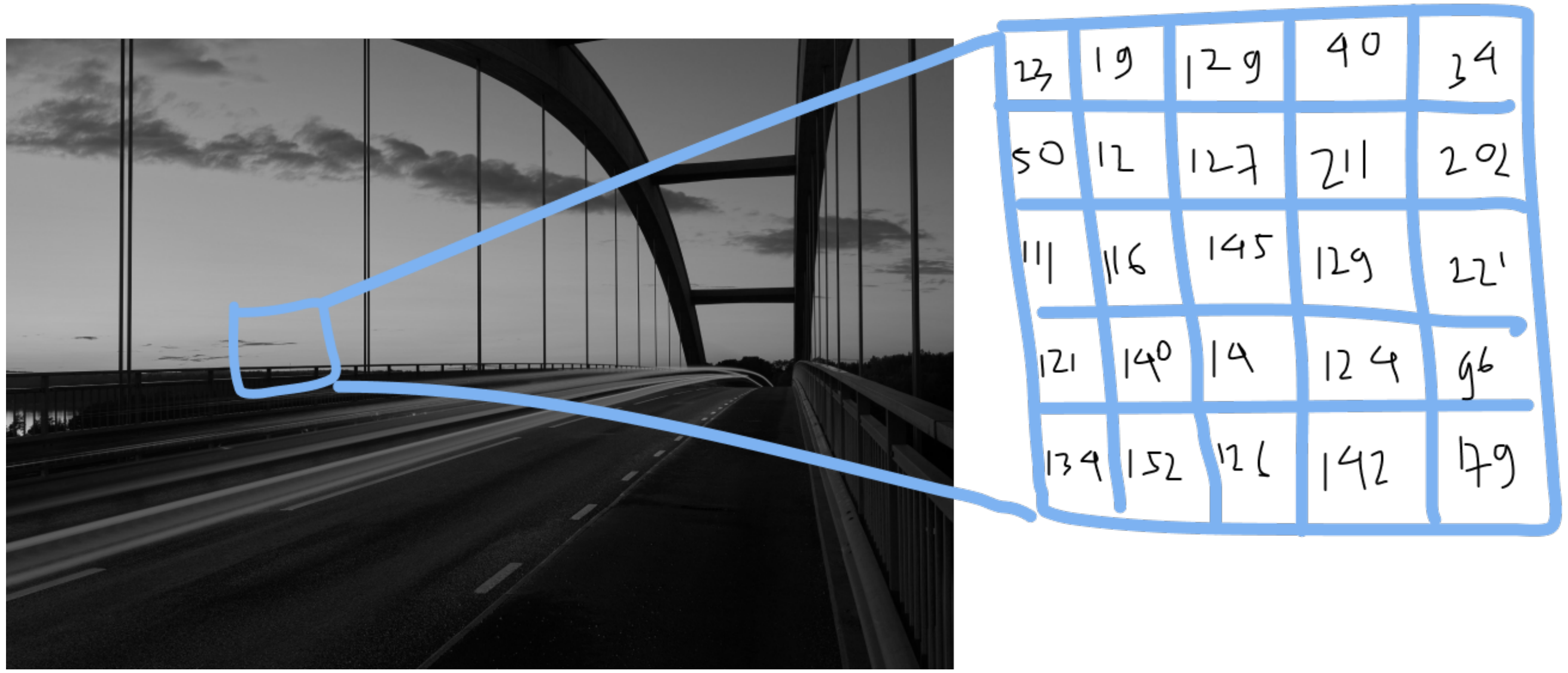
Introduction to
Word Embeddings

Galuh Sahid

@galuhsahid | github.com/galuhsahid

https://unsplash.com/photos/q0AtbGIOb5k


23 19 121 40 34
50 12 123 41 210
111 113 145 129 221
121 140 141 127 96
134 152 126 142 179


23 19 129 40 34
50 12 127 211 202
111 116 145 129 221
121 140 14 124 96
134 152 121 142 179

@galuhsahid

How do we represent **words**?

Language is hard

Lake
Forest
Ocean
River

Lake

Forest

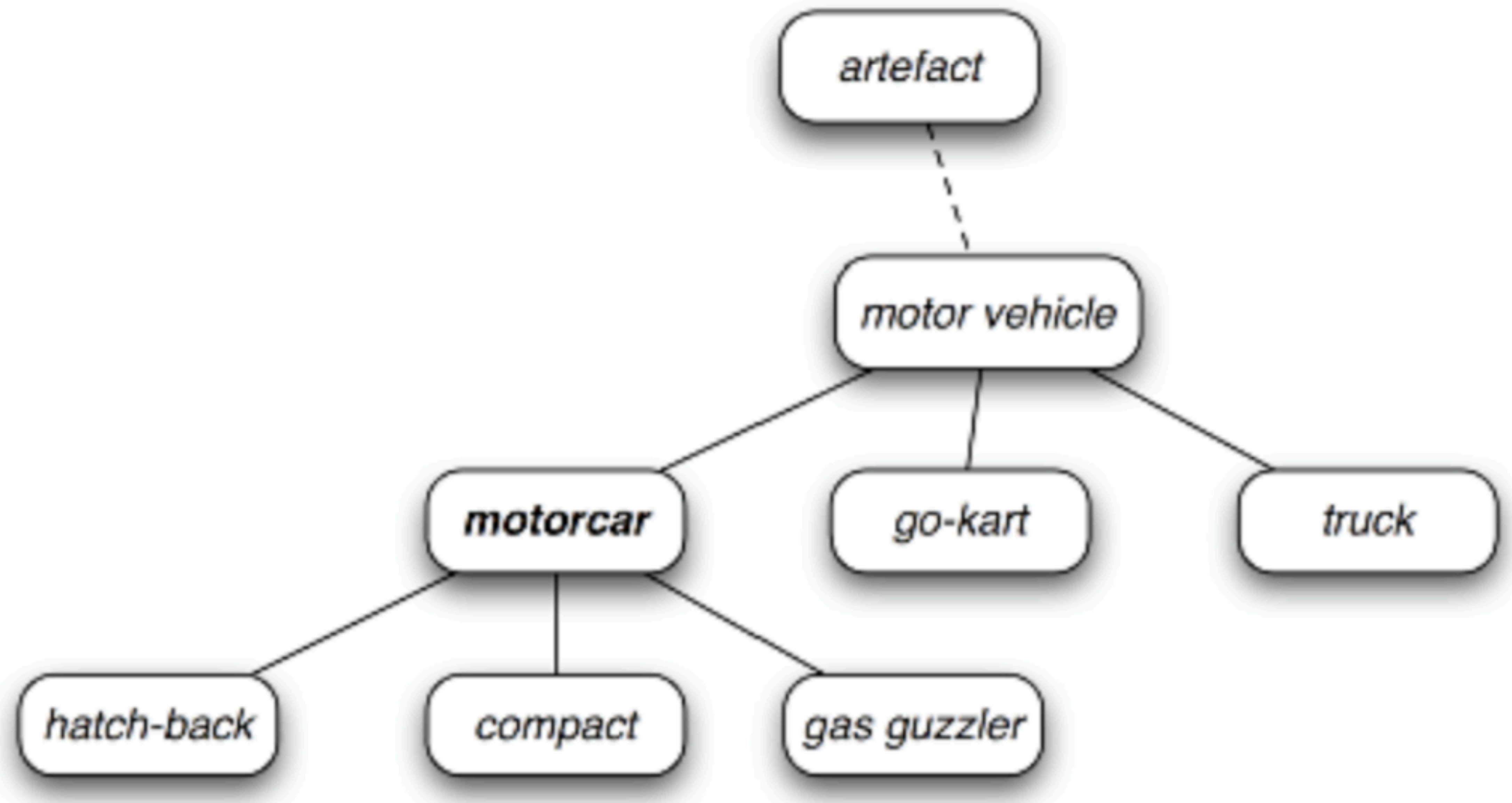
Ocean

River

Lake	Body of water
Forest	Not body of water
Ocean	Body of water
River	Body of water

Lake	Doesn’t have trees
Forest	Has trees
Ocean	Doesn’t have trees
River	Doesn’t have trees

WordNet



artefact
motor vehicle
motorcar
go-kart
truck
hatch-back
compact
gas guzzler

What kind of representation do we want?

- Real numbers
- What do we want to know about a word? Whether they have the same meaning, semantic relationship… etc.
- Can we do it without labelling everything manually?
- Ideally it's not too large!

Word Embeddings to the Rescue

Represent words as vectors of **real numbers** with much **lower and thus denser dimensions**

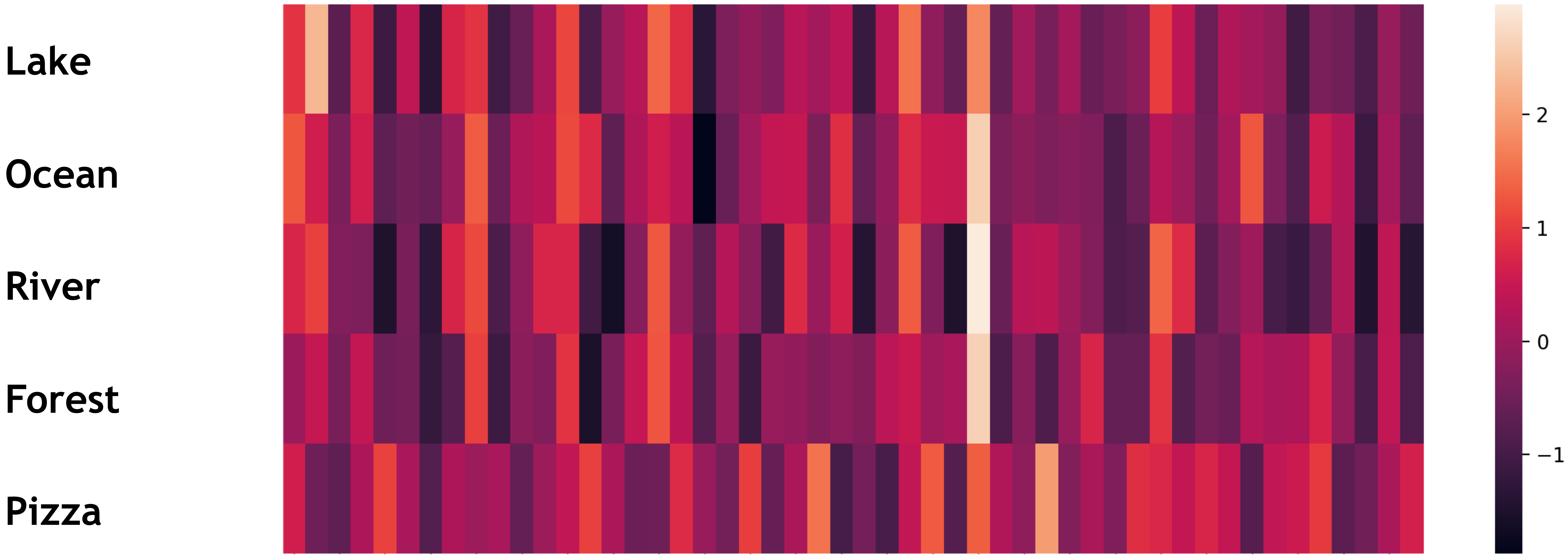
We're putting words that are outside any vector space into a vector space - hence, we're **embedding** the words into that vector space

(Spoiler alert)

Lake

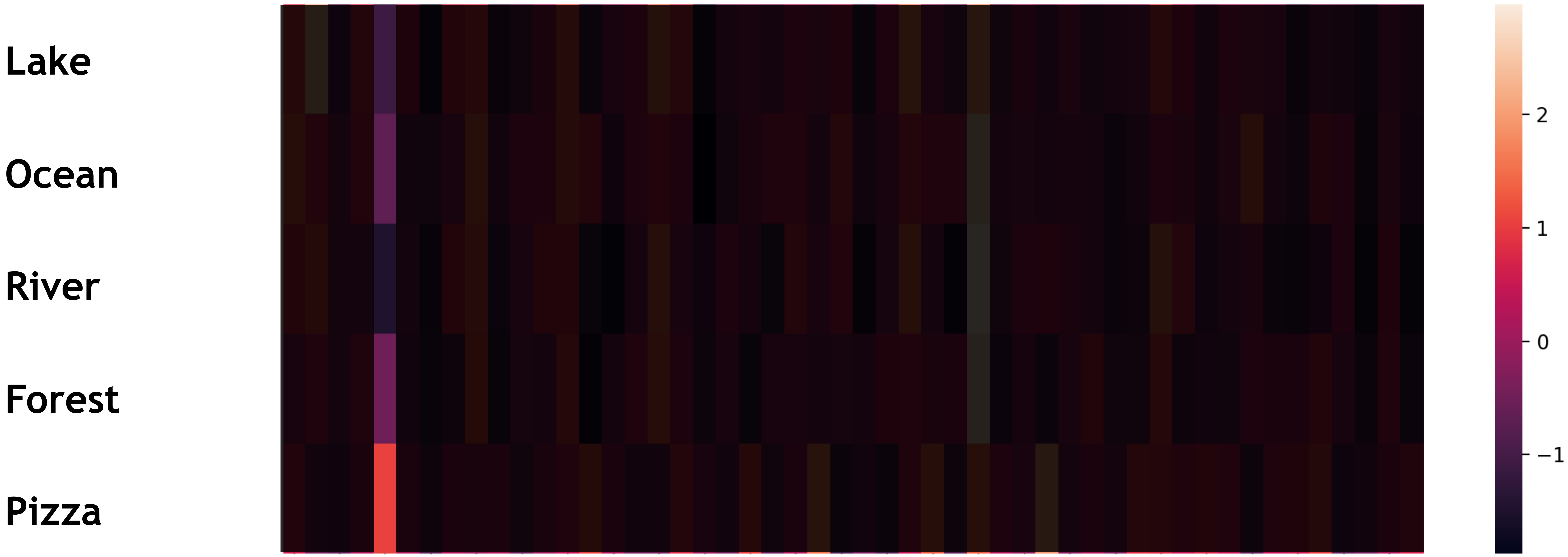
```
[ 0.89254 ,  2.3112  , -0.70036 ,  0.76679 , -1.0815  ,  0.40426 ,
   -1.3462  ,  0.71    ,  0.90067 , -1.043   , -0.57966 ,  0.18669 ,
    1.0996  , -0.90042 , -0.045962,  0.31492 ,  1.4128  ,  0.84963 ,
   -1.3389  , -0.32252 , -0.10208 , -0.31783 ,  0.33173 ,  0.096593,
    0.36732 , -1.1466  ,  0.3123  ,  1.549   , -0.13059 , -0.62003 ,
    1.774   , -0.62134 ,  0.065215, -0.39758 ,  0.095832, -0.56289 ,
   -0.39552 , -0.16224 ,  1.0035  ,  0.39161 , -0.54489 ,  0.21744 ,
    0.10831 , -0.06952 , -1.046   , -0.36096 , -0.48233 , -0.90467 ,
   -0.044913, -0.52132 ]
```

Visualization



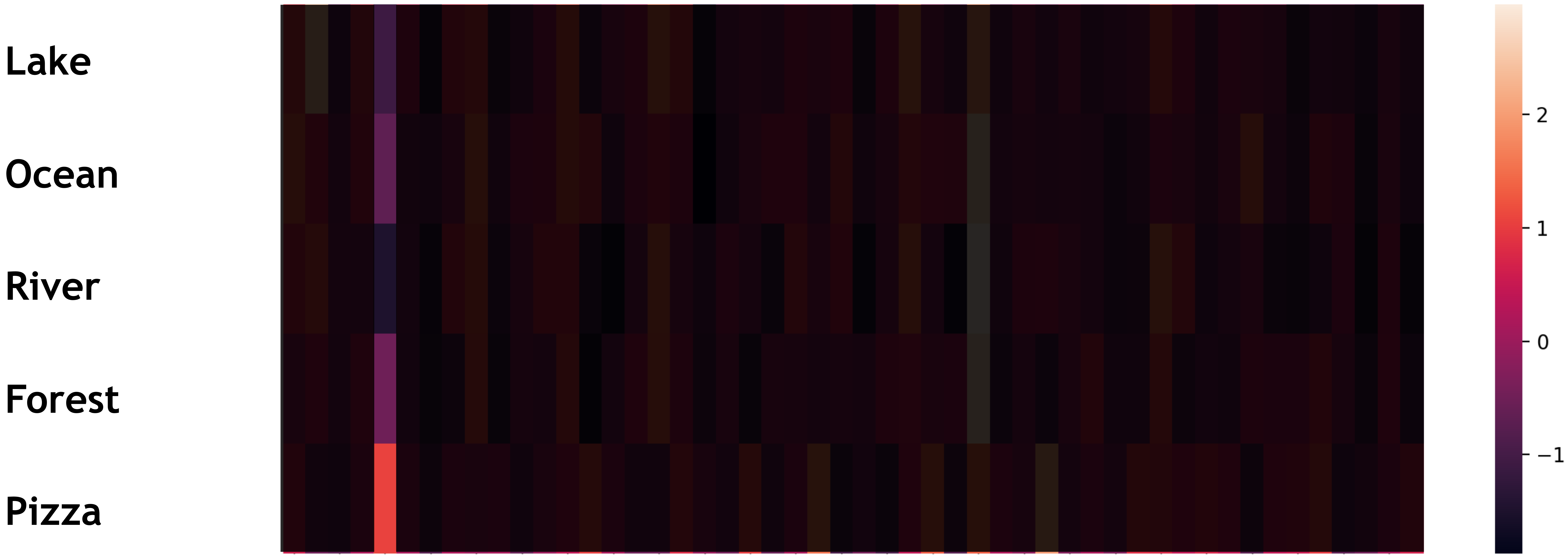
@galuhsahid

Visualization



Maybe this dimension represents the concept of whether it is a food or not…

Visualization



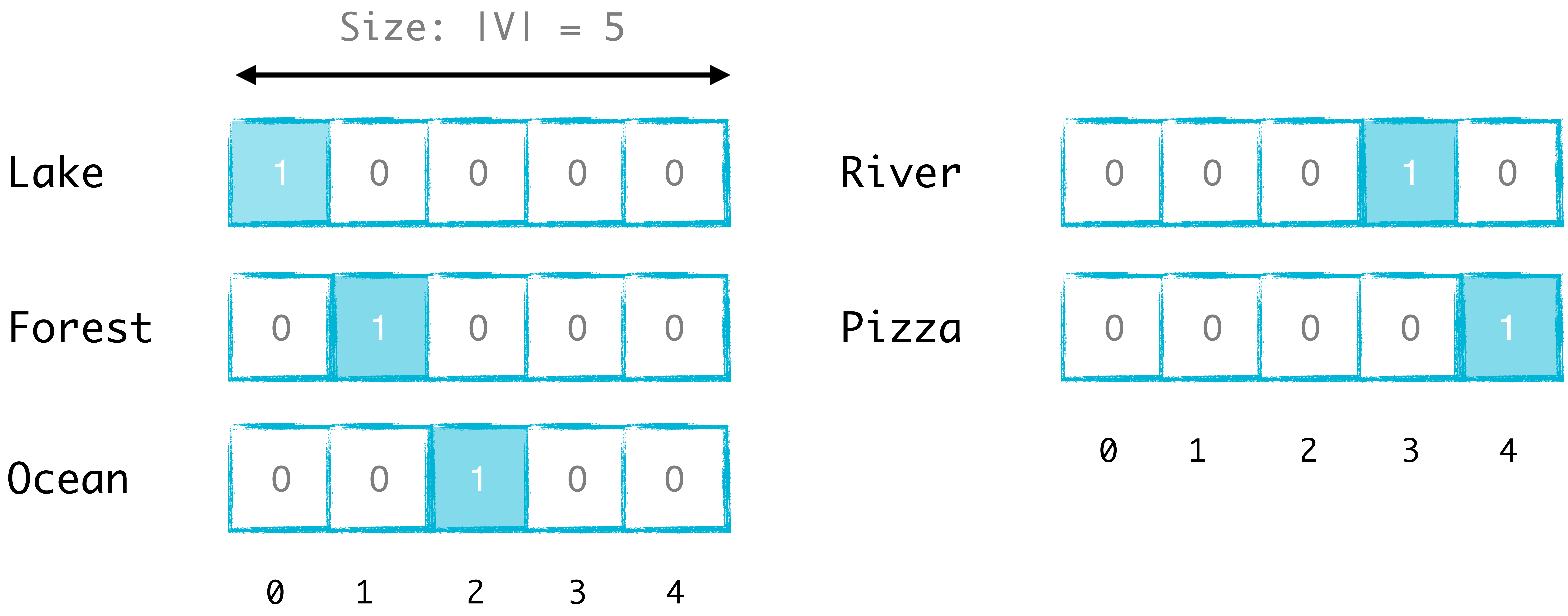
Maybe this dimension represents the concept of whether it is a food or not…

Or it could be something not intuitive to us - we actually have no idea, though. It could be anything

@galuhsahid

One-Hot Encoding

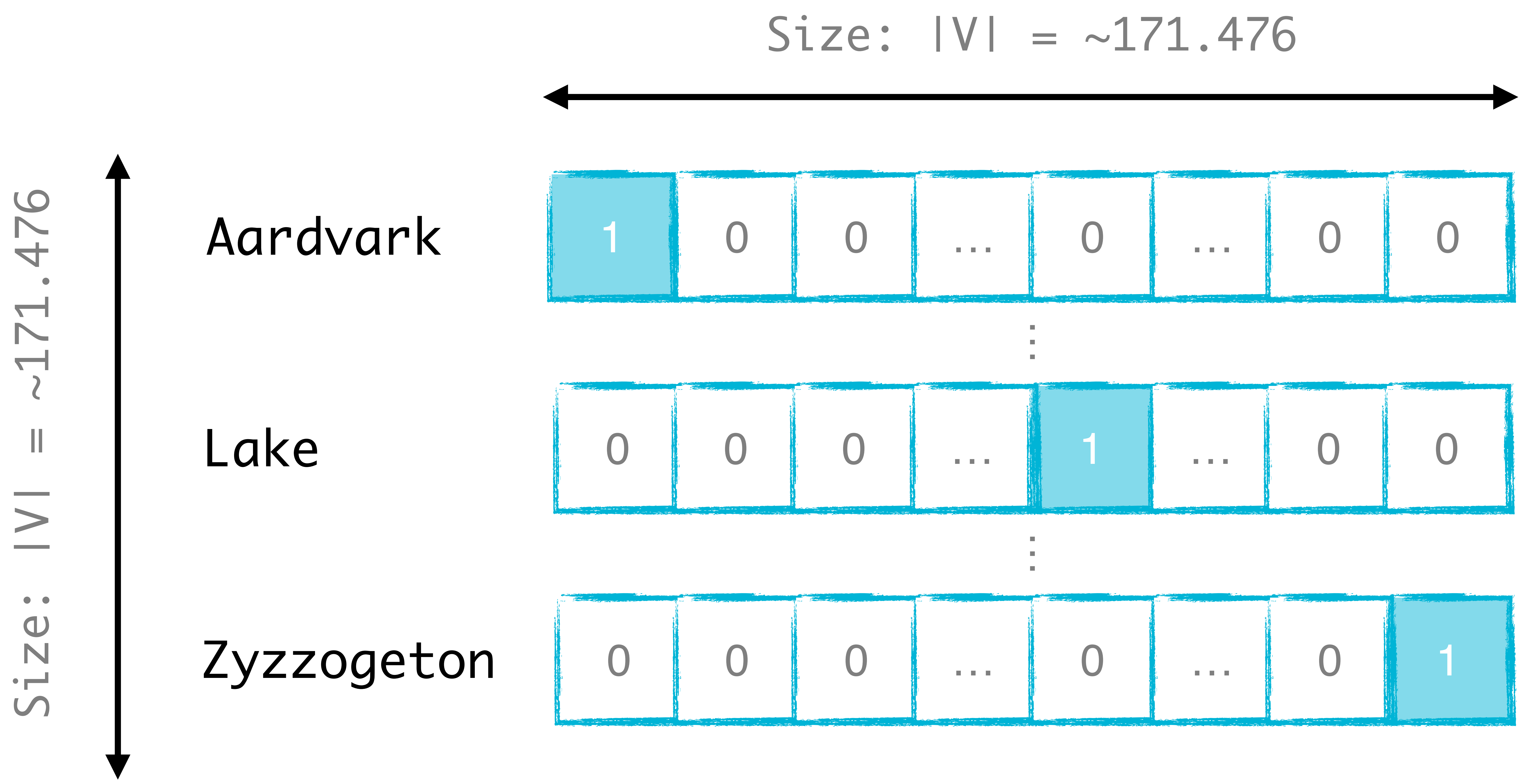
Our vocabulary: lake, forest, ocean, river, pizza



@galuhsahid

One-Hot Encoding

Our vocabulary: every English word (approx. 171,476 words in use)



https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language/

@galuhsahid

Distributional Representation

“Tell me who your friends are, and I’ll tell you who you are.”

Distributional Representation

"Tell me who your friends are, and I'll tell you who you are."

"You shall know a word by the company it keeps."

(Firth, 1957)

Distributional Representation

“Tell me who your friends are, and I’ll tell you who you are.”

“You shall know a word by the company it keeps.”

(Firth, 1957)

Words that occur in similar contexts have similar meaning

Distributional Hypothesis (Harris, 1954)

Distributional Representation

A **lake** is a large body of water in a body of land.

An **ocean** is a large area of water between continents.

A **river** is a stream of water that flows through a channel in the surface of the ground.

A **forest** is a piece of land with many trees.

Pizza is a type of food that was created in Italy.

Distributional Representation

A **lake** is a large body of water in a body of land.

An **ocean** is a large area of water between continents.

A **river** is a stream of water that flows through a channel in the surface of the ground.

A **forest** is a piece of land with many trees.

Pizza is a type of food that was created in Italy.

Distributional Representation

A $\boxed{\textbf{lake}}$ is a large body of water in a body of land.

An $\boxed{\textbf{ocean}}$ is a large area of water between continents.

A **river** is a stream of water that flows through a channel in the surface of the ground.

A **forest** is a piece of land with many trees.

Pizza is a type of food that was created in Italy.

No manual annotation!

Approaches

- **Count-based methods**
  - Computes how often a word co-occurs with its neighbour words, then map the counts to a small, dense vector

Count-based

window = 4

Neighbor words

Lake large body water body land.	[large, body, water]
Ocean large area water continents.	[large, area, water]
River stream water flows channel surface ground.	[stream, water, flows]
Forest piece land many trees.	[piece, land, many]
Pizza type food created Italy.	[type, food, created]

APPROACHES

Count-based

	Neighbor words	Large	Body	Water	Area	Stream	Flows	Piece	Land	Many	Type	Food	Created
Lake	[large, body, water]	1	1	1	0	0	0	0	0	0	0	0	0
Ocean	[large, area, water]	1	0	1	1	0	0	0	0	0	0	0	0
River	[stream, water, flows]	0	0	1	0	1	1	0	0	0	0	0	0
Forest	[piece, land, many]	0	0	0	0	0	0	1	1	1	0	0	0
Pizza	[type, food, created]	0	0	0	0	0	0	0	0	0	1	1	1

Approaches

- **Count-based methods**
  - Computes how often a word co-occurs with its neighbour words, then map the counts to a small, dense vector
  - Reduce dimensions using Singular Vector Decomposition (SVD) or Latent Dirichlet Allocation (LDA)

Approaches

- **Predictive methods**
  - Try to predict a word from its neighbors in terms of small and denser embedding vectors

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 238-247).

Predictive Methods

- **Word2Vec**
  - Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

- **GloVe**
  - Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Predictive Methods

- **FastText**
  - https://research.fb.com/downloads/fasttext/

- **ELMo**
  - Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Predictive Methods

- **Word2Vec**
  - Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

- **GloVe**
  - Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Architecture

- Continuous Bag-of-Words (CBOW)
- Skip-gram

Skip-gram

pizza

```
[ 0.89254 ,  2.3112  , -0.70036 ,  0.76679 , -1.0815  ,  0.40426 ,
 -1.3462  ,  0.71    ,  0.90067 , -1.043   , -0.57966 ,  0.18669 ,
  1.0996  , -0.90042 , -0.045962,  0.31492 ,  1.4128  ,  0.84963 ,
 -1.3389  , -0.32252 , -0.10208 , -0.31783 ,  0.33173 ,  0.096593,
  0.36732 , -1.1466  ,  0.3123  ,  1.549   , -0.13059 , -0.62003 ,
  1.774   , -0.62134 ,  0.065215, -0.39758 ,  0.095832, -0.56289 ,
 -0.39552 , -0.16224 ,  1.0035  ,  0.39161 , -0.54489 ,  0.21744 ,
  0.10831 , -0.06952 , -1.046   , -0.36096 , -0.48233 , -0.90467 ,
 -0.044913, -0.52132 ]
```

Skip-gram

“I ate the leftover **pizza** for dinner”

Skip-gram

Window size = 5

"I ate the leftover **pizza** for dinner"

Skip-gram

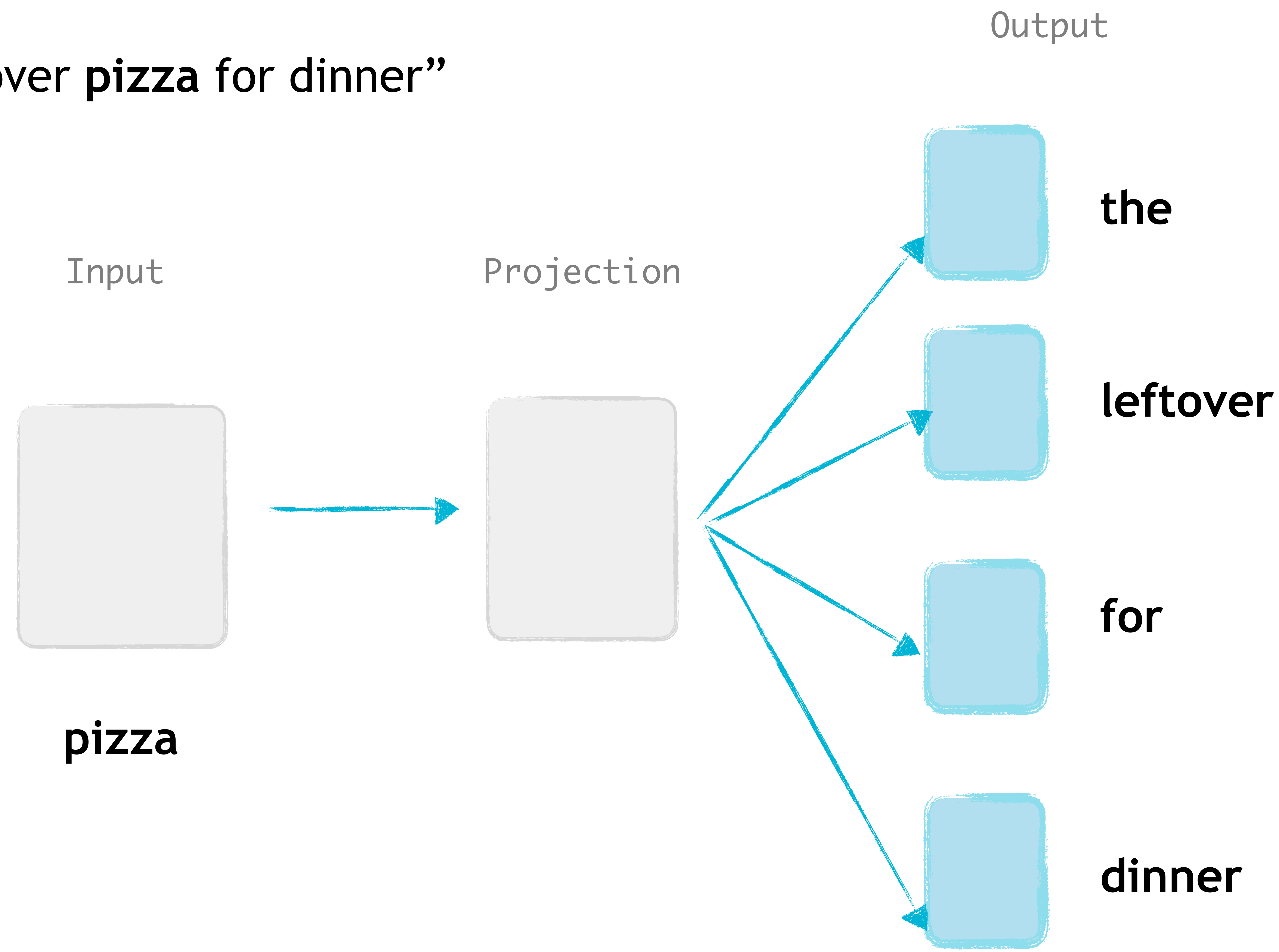
Window size = 5

"I ate $\boxed{\text{the leftover }\textbf{pizza}\text{ for dinner}}$"

Neighbour words: the, leftover, for, dinner

Overview

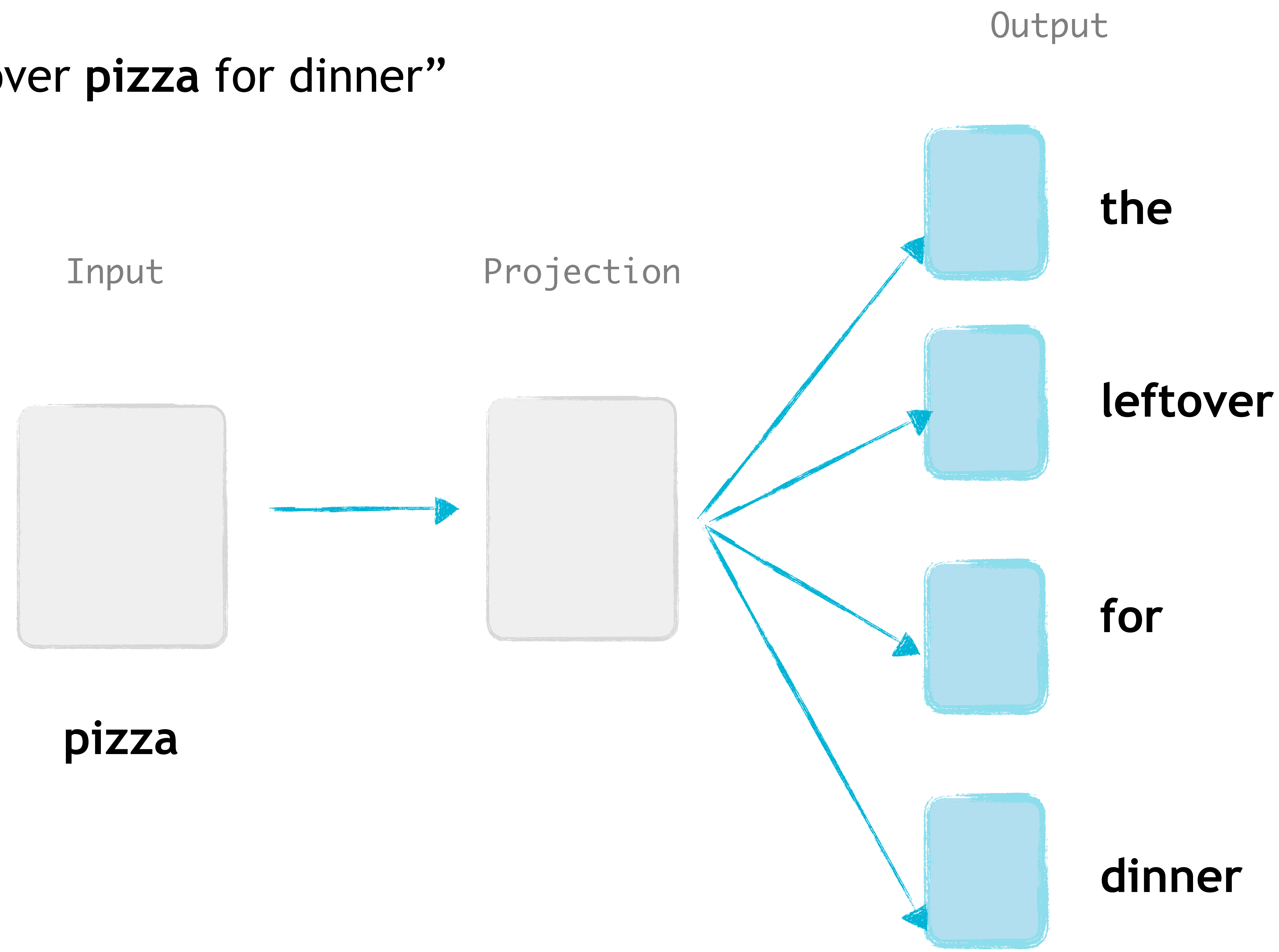
- "I ate the leftover **pizza** for dinner"



Skip-gram

Overview

- “I ate the leftover **pizza** for dinner”

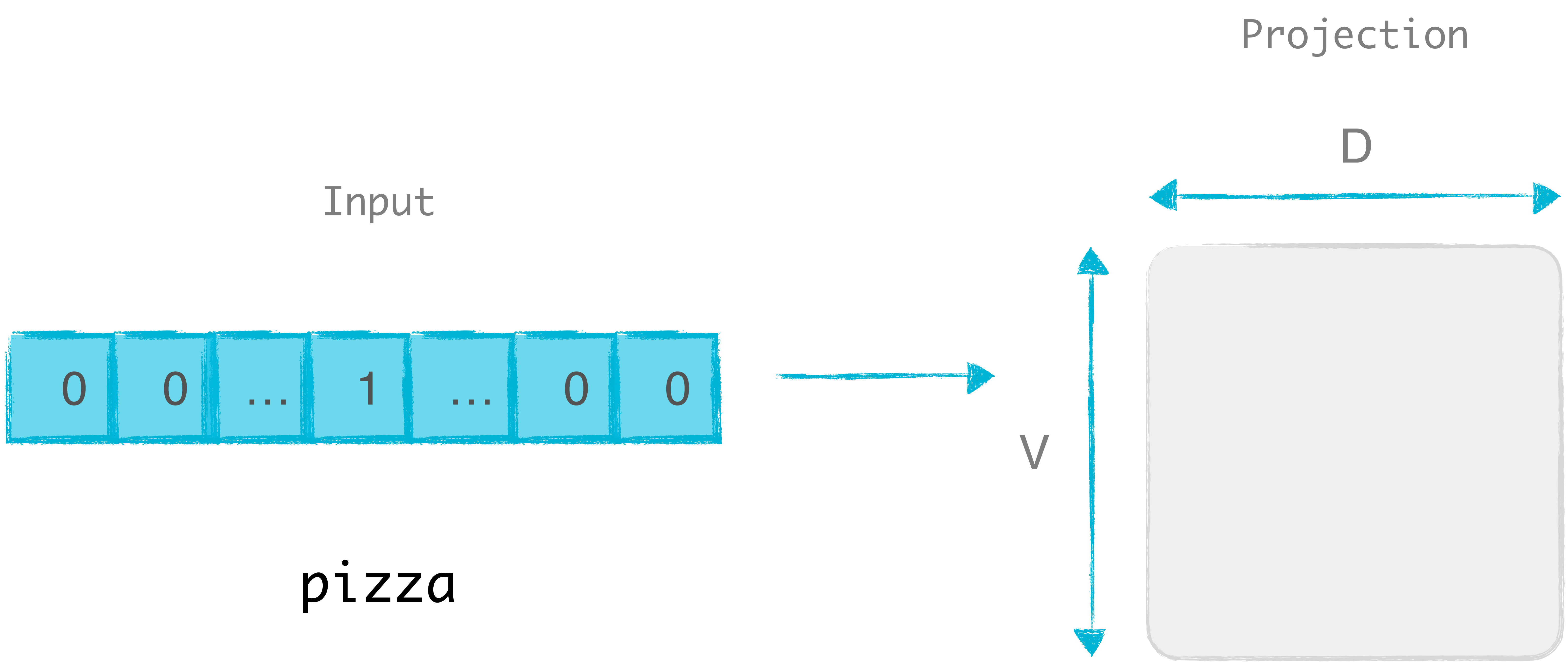


Architecture

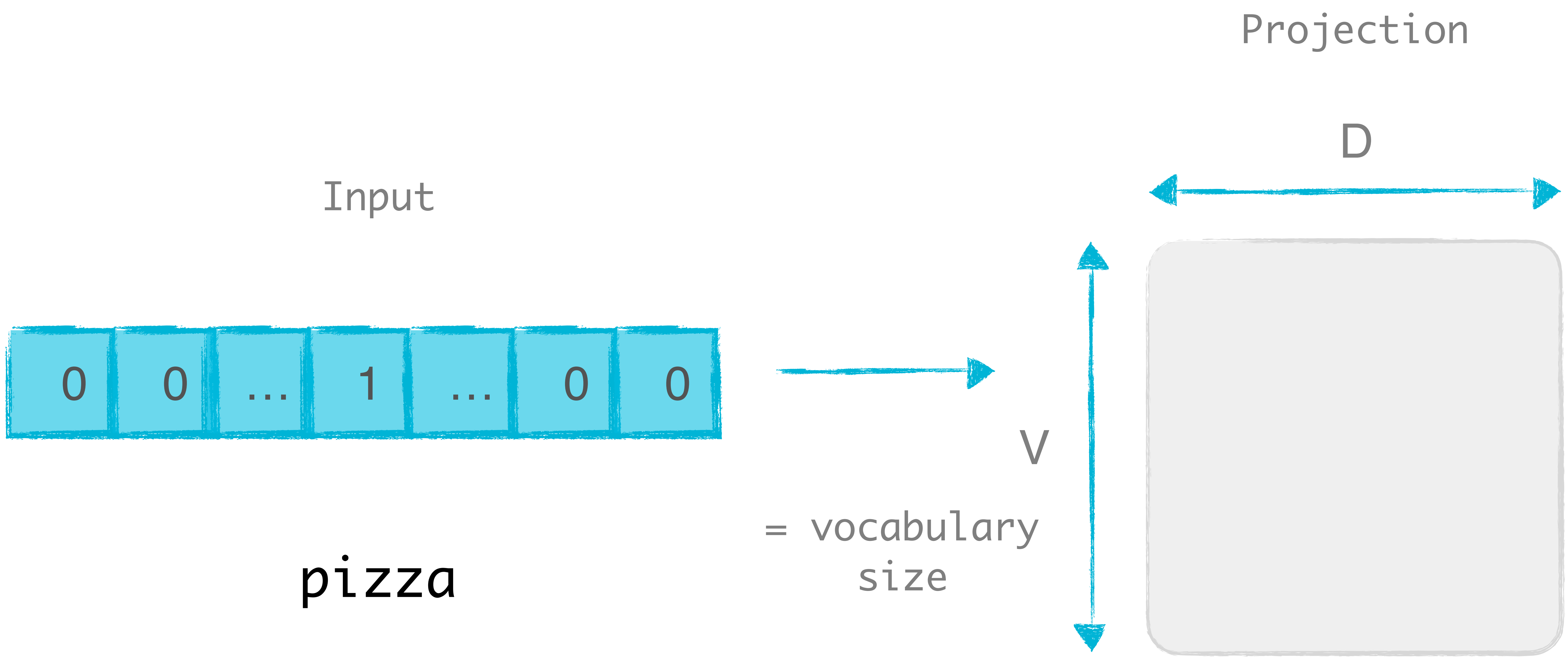
Input

pizza

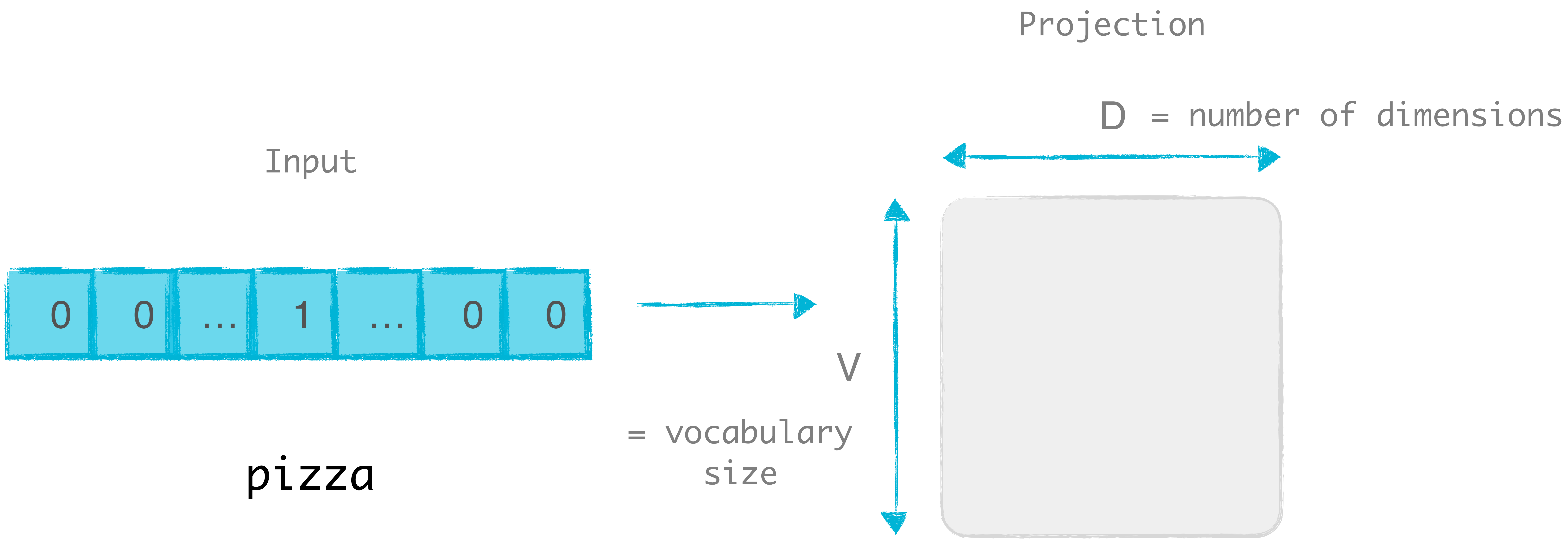
Architecture



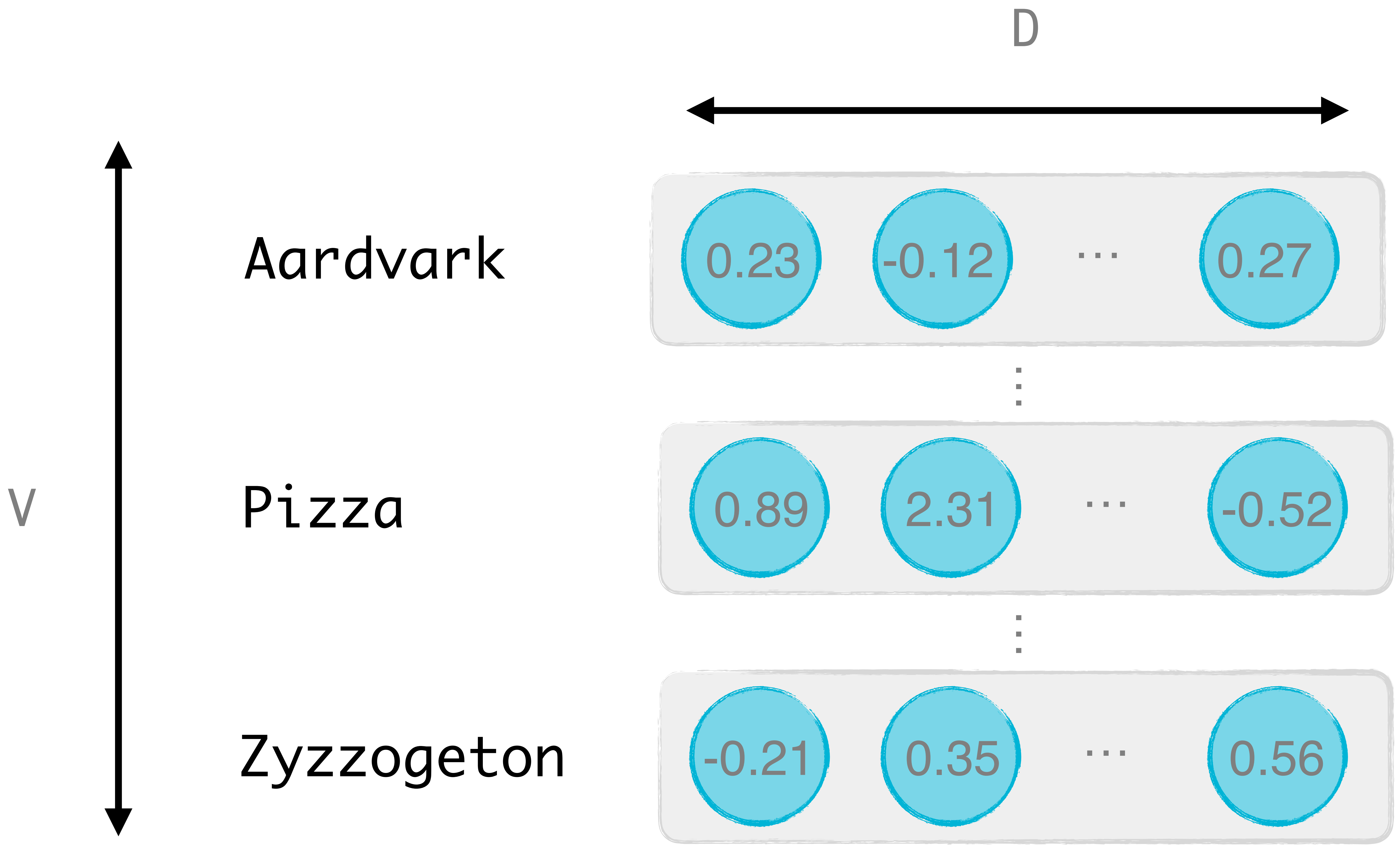
Architecture



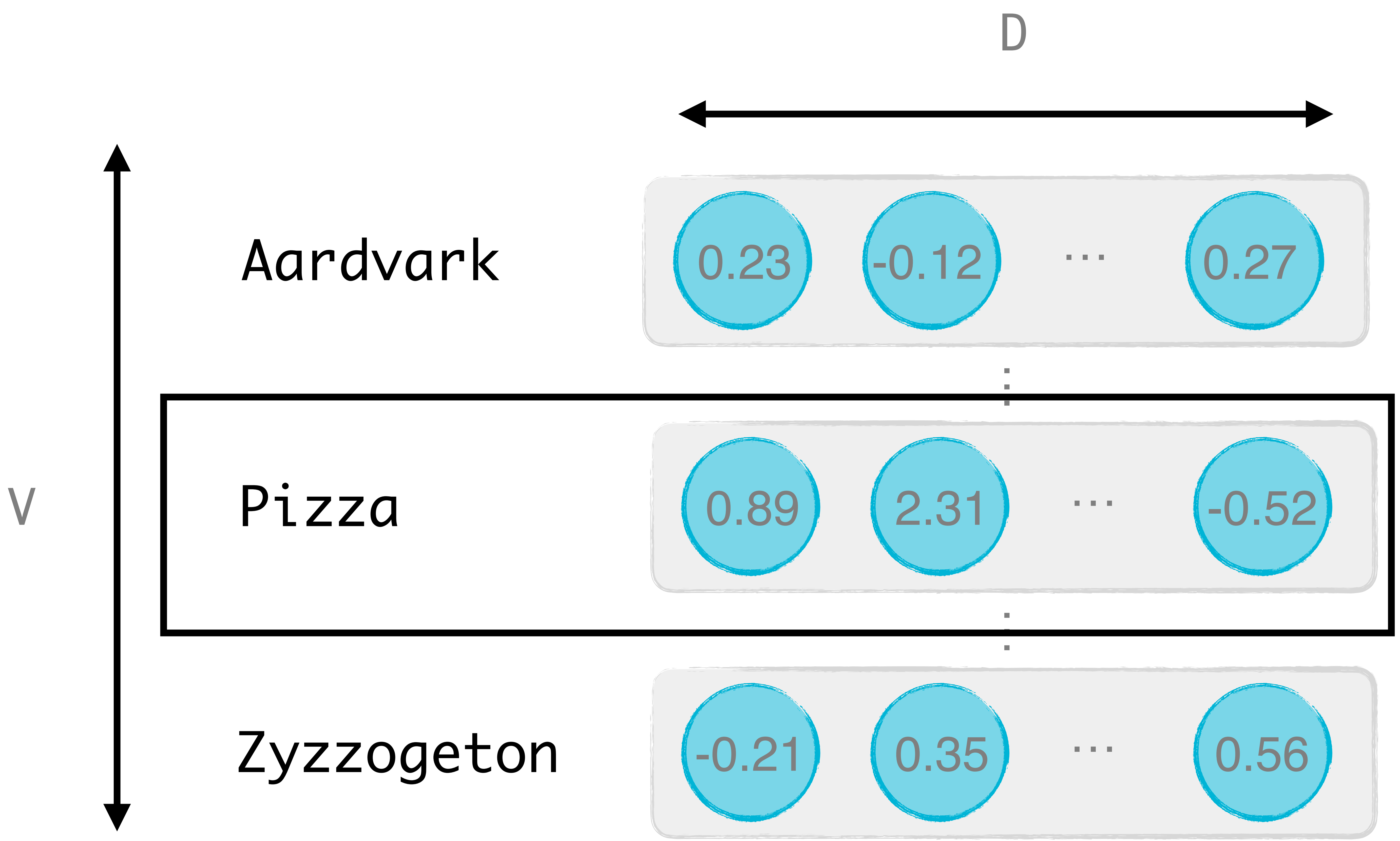
Architecture



Projection Layer

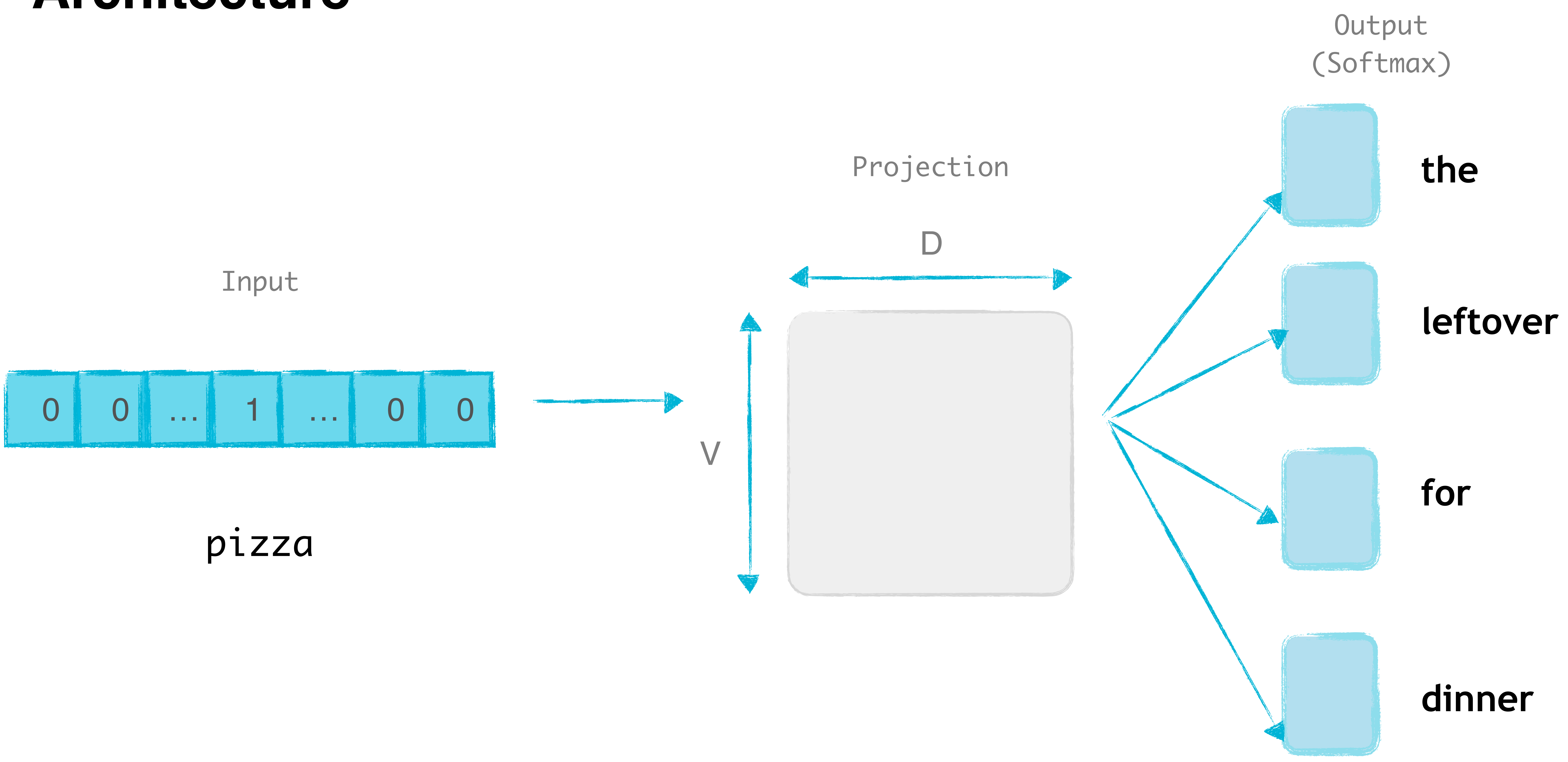


Projection Layer



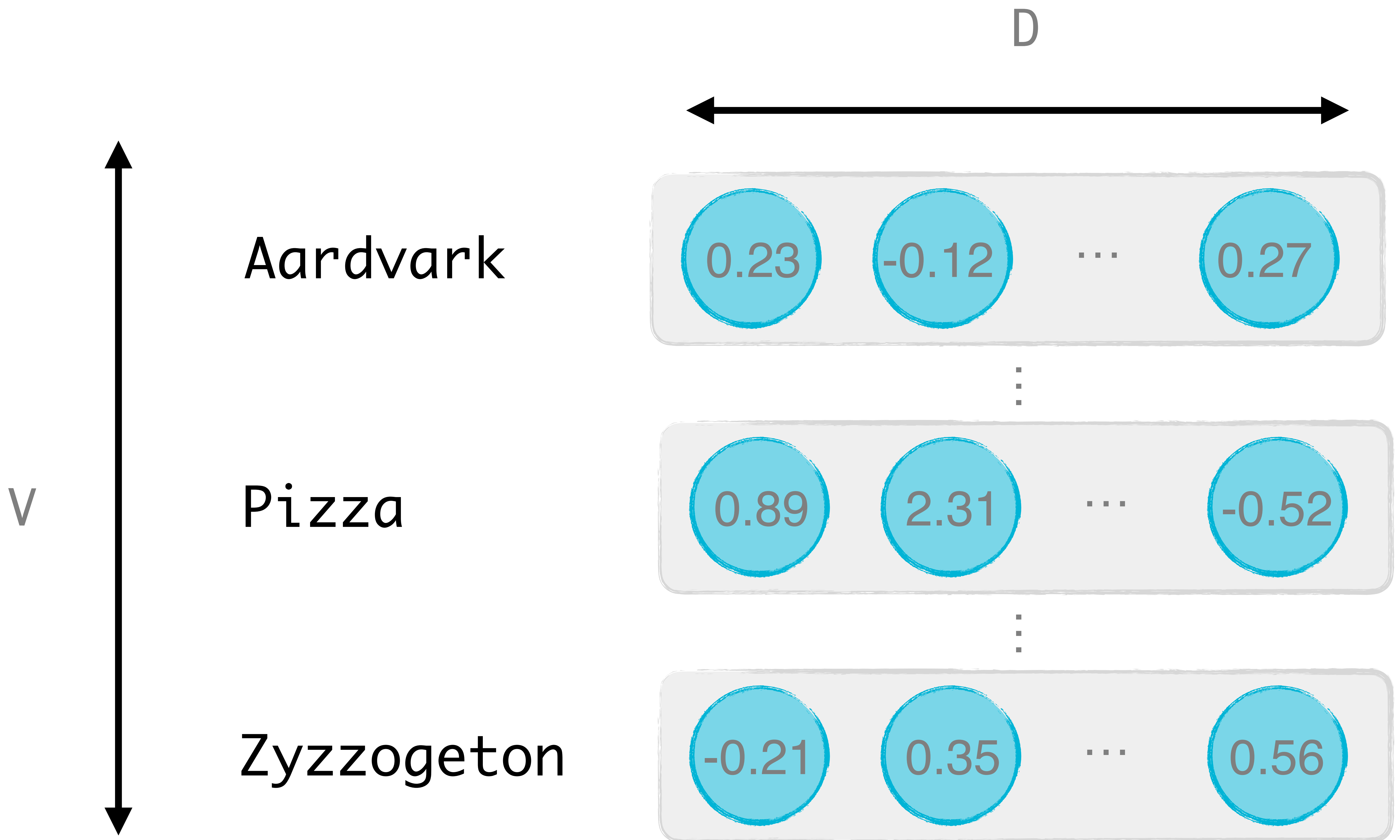
WORD2VEC - SKIP-GRAM

Architecture



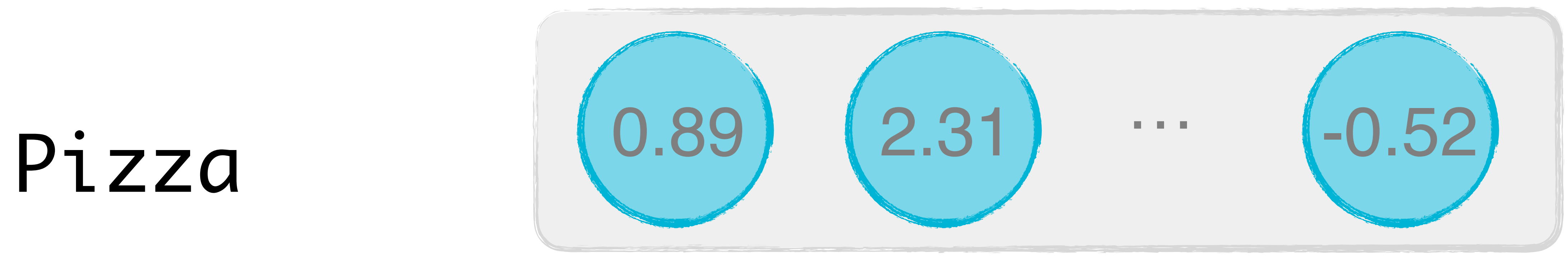
@galuhsahid

Projection Layer




SURPRISE!
NBC

Word Vector



This is our word vector!

Word Vector

Pizza

This is our word vector!

```
[ 0.89254 ,  2.3112  , -0.70036 ,  0.76679 , -1.0815  ,  0.40426 ,
 -1.3462  ,  0.71    ,  0.90067 , -1.043   , -0.57966 ,  0.18669 ,
  1.0996  , -0.90042 , -0.045962,  0.31492 ,  1.4128  ,  0.84963 ,
 -1.3389  , -0.32252 , -0.10208 , -0.31783 ,  0.33173 ,  0.096593,
  0.36732 , -1.1466  ,  0.3123  ,  1.549   , -0.13059 , -0.62003 ,
  1.774   , -0.62134 ,  0.065215, -0.39758 ,  0.095832, -0.56289 ,
 -0.39552 , -0.16224 ,  1.0035  ,  0.39161 , -0.54489 ,  0.21744 ,
  0.10831 , -0.06952 , -1.046   , -0.36096 , -0.48233 , -0.90467 ,
 -0.044913, -0.52132 ]
```

The Intuition

"I ate the leftover pizza for dinner"

Pizza

the
leftover
for
dinner

"I need some leftover chicken recipes for dinner"

Chicken

some
leftover
recipes
for

The Intuition

"I ate the leftover pizza for dinner"

Pizza	0.89	2.31	…	-0.52

the
leftover
for
dinner

"I need some leftover chicken recipes for dinner"

Chicken	0.76	2.01	…	-0.47

some
leftover
recipes
for

"The Germany embassy in Prague is located in…"

Prague	0.32	0.43	…	-0.21

embassy
in
is
located

Architecture

More details:

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Exploring Word Embeddings with Gensim

Pre-trained Models

- Gensim has an API to download pre-trained word embedding models.
  The list of available models can be found here.

glove-twitter-50	1193514	199 MB	Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased)
glove-wiki-gigaword-100	400000	128 MB	Wikipedia 2014 + Gigaword 5 (6B tokens, uncased)
glove-wiki-gigaword-200	400000	252 MB	Wikipedia 2014 + Gigaword 5 (6B tokens, uncased)

glove-wiki-gigaword-300	400000	376 MB	Wikipedia 2014 + Gigaword 5 (6B tokens, uncased)
glove-wiki-gigaword-50	400000	65 MB	Wikipedia 2014 + Gigaword 5 (6B tokens, uncased)
word2vec-google-news-300	3000000	1662 MB	Google News (about 100 billion words)

Loading the Model

```
from gensim.models import KeyedVectors

model_w2v = KeyedVectors.load_word2vec_format(
            './GoogleNews-vectors-negative300.bin',
            binary=True)
```

Word Vector

```
model_w2v[“lake”]
```

```
array([-8.39843750e-02,  2.02148438e-01,  2.65625000e-01,  1.04980469e-01,
       -7.95898438e-02,  1.05957031e-01, -5.39550781e-02,  8.11767578e-03,
        9.32617188e-02, -7.66601562e-02,  1.56250000e-01, -1.19628906e-01,
       … -4.15039062e-02,  4.08935547e-03, -2.47070312e-01, -1.78710938e-01,
       3.33984375e-01, -1.79687500e-01], dtype=float32)
```

Similar Words

```
model_w2v.most_similar("apple")
```

```
[('apples', 0.7203598022460938),
 ('pear', 0.6450696587562561),
 ('fruit', 0.6410146355628967),
 ('berry', 0.6302294731140137),
 ('pears', 0.6133961081504822),
 ('strawberry', 0.6058261394500732),
 ('peach', 0.6025873422622681),
 ('potato', 0.596093475818634),
 ('grape', 0.5935864448547363),
 ('blueberry', 0.5866668224334717)]
```

- Similar words are nearby vectors in a vector space
- The distance is calculated using **cosine similarity**

Get Similarity

```
model_w2v.similarity("apple", "mango")
```

```
0.57518554
```

Odd One Out

```
model_w2v.doesnt_match(["lake", "forest", "ocean", "river"])
```

```
'forest'
```

Analogies

- man to uncle is woman to ...

```
model_w2v.most_similar(positive=["uncle", "woman"],
                       negative=["man"])
```

Analogies

- man to uncle is woman to …

```
model_w2v.most_similar(positive=["uncle", "woman"],
                       negative=["man"])
```

```
[('aunt', 0.8022665977478027),
 ('mother', 0.7770732045173645),
 ('niece', 0.768424928188324),
 ('father', 0.7237852811813354),
 ('grandmother', 0.722037136554718),
 ('daughter', 0.7185647487640381),
 ('sister', 0.7006258368492126),
 ('husband', 0.6982548236846924),
 ('granddaughter', 0.6858304738998413),
 ('nephew', 0.671071410179138)]
```

Words that are similar to uncle and woman but dissimilar to man

uncle + woman - man

Analogies

- Germany to Berlin is France to…

```
model_w2v.most_similar(positive=["Berlin", “France”],
                       negative=["Germany"])
```

```
[('Paris', 0.7672388553619385),
 ('French', 0.6049168109893799),
 ('Parisian', 0.5810437202453613),
 ('Colombes', 0.5599985718727112),
 ('Hopital_Europeen_Georges_Pompidou', 0.555890679359436),
 ('Melun', 0.551270067691803),
 ('Dinard', 0.5451847314834595),
 ('Brussels', 0.5420989990234375),
 ('Mairie_de', 0.5337448120117188),
 ('Cagnes_sur_Mer', 0.531246542930603)]
```

Analogies

- Germany to Berlin is France to…

```
model_w2v.most_similar(positive=["Berlin", “France”],
                       negative=["Germany"])
```

Analogies

- Running to run is walking to...

```
model_w2v.most_similar(positive=["run", "walking"],
                       negative=["running"])
```

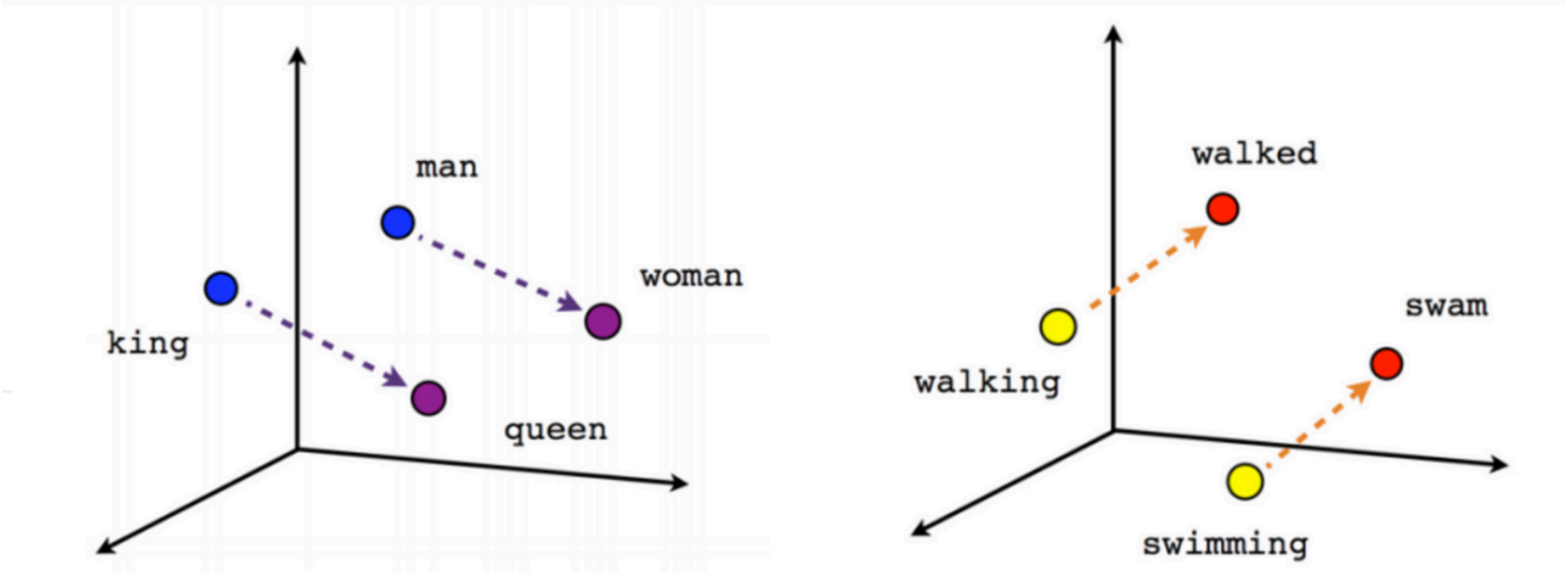
Analogies

- Running to run is walking to…

```
model_w2v.most_similar(positive=["run", "walking"],
                       negative=["running"])
```

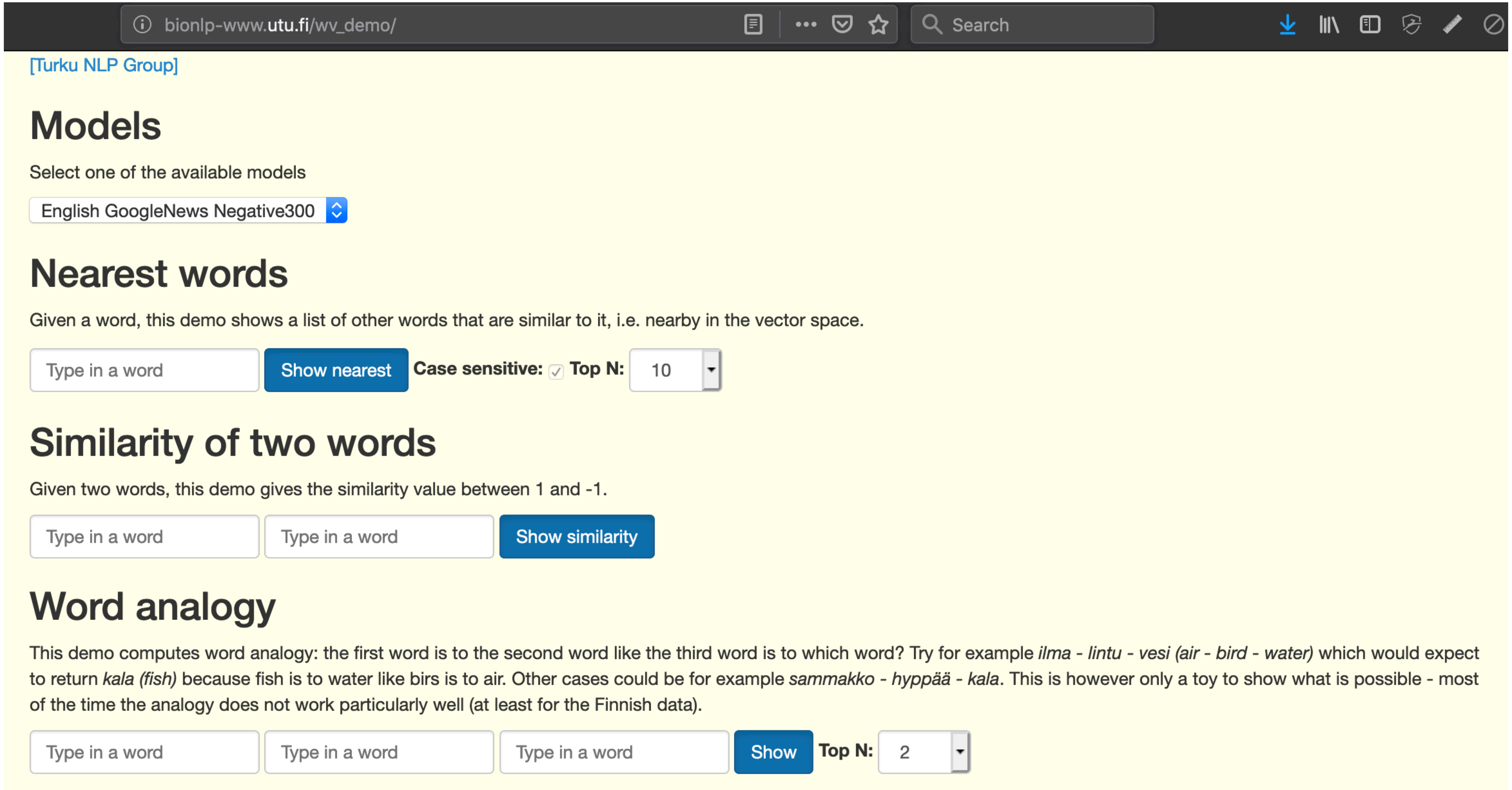
```
[('walk', 0.7163699865341187),
 ('walks', 0.5965700745582581),
 ('walked', 0.5833066701889038),
 ('stroll', 0.5236037969589233),
 ('pinch_hitter_Yunel_Escobar', 0.45626372098922729),
 ('Walking', 0.4554094374179840),
 ('Batterymate_Miguel_Olivo', 0.4483090043067932),
 ('runs', 0.4462803602218628),
 ('pinch_hitter_Carlos_Guillen', 0.440292537212371),
 ('Justin_Speier_relieved', 0.4352820515632629)]
```

Analogies



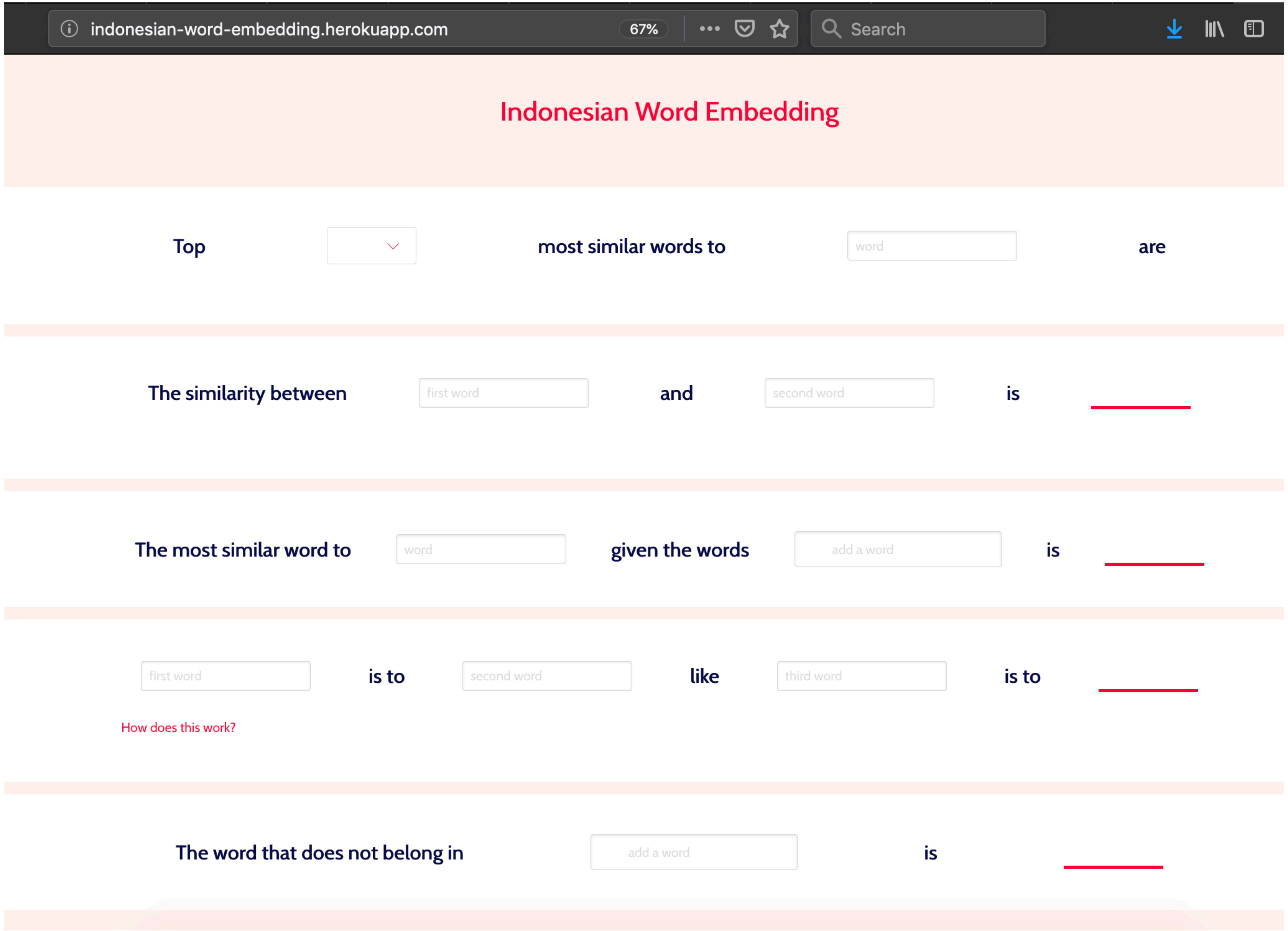
Male-Female

Verb tense



http://bionlp-www.utu.fi/wv_demo/

@galuhsahid



https://indonesian-word-embedding.herokuapp.com

http://github.com/galuhsahid/indonesian-word-embedding

@galuhsahid

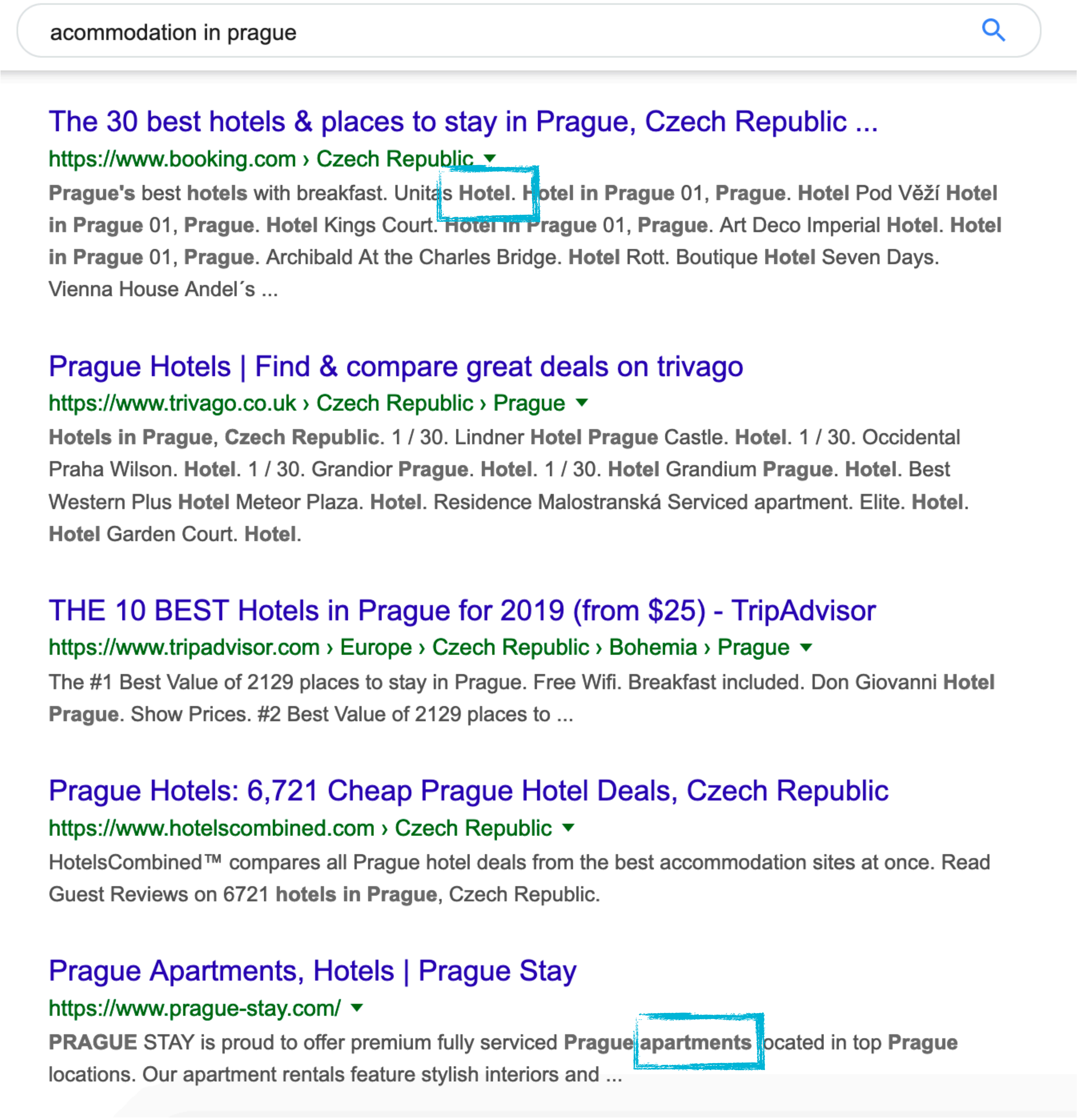
Training Your Own

- **Sure you can!**

- When to do so?
  - Specific problem domains
  - Challenge: training data

- Another alternative: continue training a pre-existing word embedding

Search



acommodation in prague

The 30 best hotels & places to stay in Prague, Czech Republic ...
https://www.booking.com › Czech Republic ▾
Prague's best **hotels** with breakfast. Unitas **Hotel**. **Hotel in Prague** 01, **Prague**. **Hotel** Pod Věží **Hotel in Prague** 01, **Prague**. **Hotel** Kings Court. **Hotel in Prague** 01, **Prague**. Art Deco Imperial **Hotel**. **Hotel in Prague** 01, **Prague**. Archibald At the Charles Bridge. **Hotel** Rott. Boutique **Hotel** Seven Days. Vienna House Andel´s ...

Prague Hotels | Find & compare great deals on trivago
https://www.trivago.co.uk › Czech Republic › Prague ▾
Hotels in Prague, **Czech Republic**. 1 / 30. Lindner **Hotel Prague** Castle. **Hotel**. 1 / 30. Occidental Praha Wilson. **Hotel**. 1 / 30. Grandior **Prague**. **Hotel**. 1 / 30. **Hotel** Grandium **Prague**. **Hotel**. Best Western Plus **Hotel** Meteor Plaza. **Hotel**. Residence Malostranská Serviced apartment. Elite. **Hotel**. **Hotel** Garden Court. **Hotel**.

THE 10 BEST Hotels in Prague for 2019 (from $25) - TripAdvisor
https://www.tripadvisor.com › Europe › Czech Republic › Bohemia › Prague ▾
The #1 Best Value of 2129 places to stay in Prague. Free Wifi. Breakfast included. Don Giovanni **Hotel Prague**. Show Prices. #2 Best Value of 2129 places to ...

Prague Hotels: 6,721 Cheap Prague Hotel Deals, Czech Republic
https://www.hotelscombined.com › Czech Republic ▾
HotelsCombined™ compares all Prague hotel deals from the best accommodation sites at once. Read Guest Reviews on 6721 **hotels in Prague**, Czech Republic.

Prague Apartments, Hotels | Prague Stay
https://www.prague-stay.com/ ▾
PRAGUE STAY is proud to offer premium fully serviced **Prague apartments** located in top **Prague** locations. Our apartment rentals feature stylish interiors and ...

Neural Machine Translation

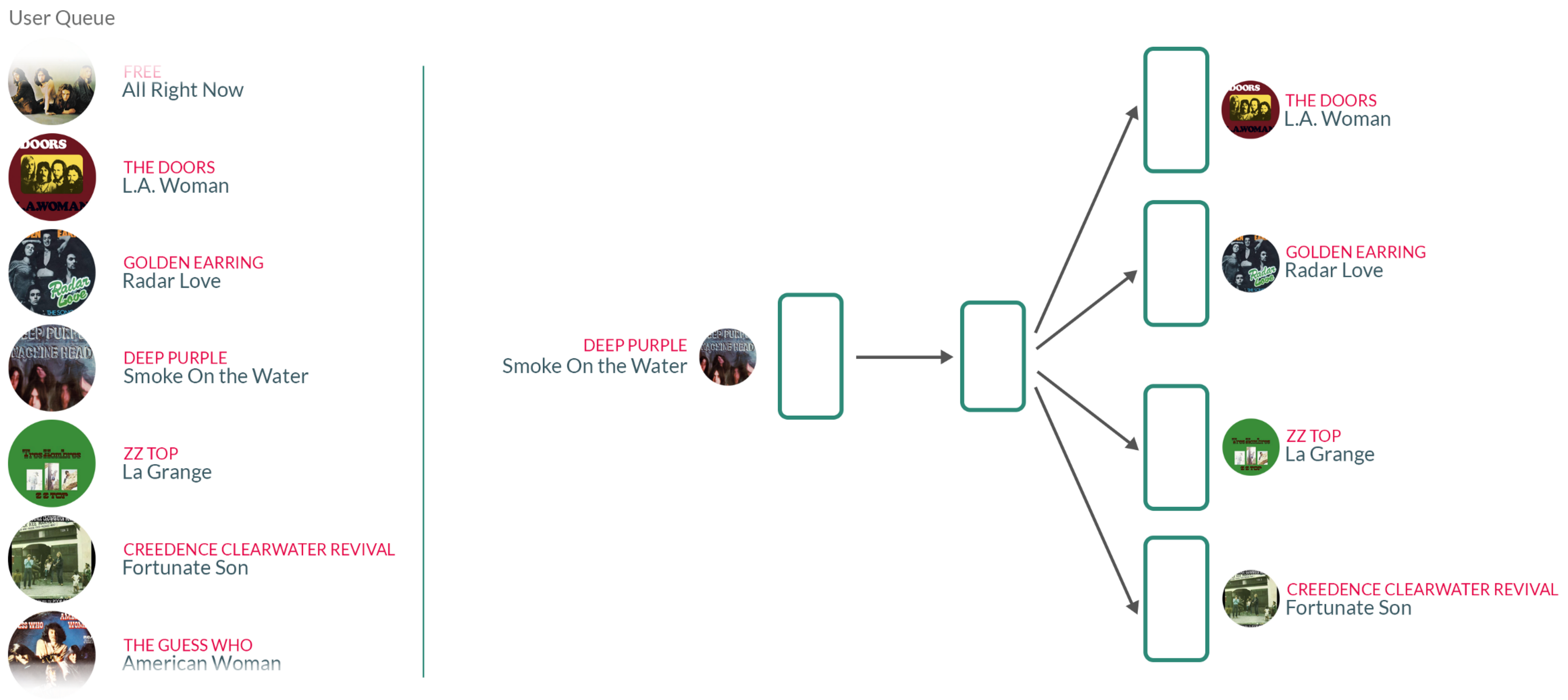
source	*(risos)* e é que **chris** é un grande avogado , pero non sabía case nada sobre lexislación de patentes e absolutamente nada sobre xenética .
reference	*(laughter)* now **chris** is a really brilliant lawyer , but he knew almost nothing about patent law and certainly nothing about genetics .
bi:std	*(laughter)* and **i** 'm not a little bit of a little bit of a little bit of and (laughter) and i 'm going to be able to be a lot of years .
multi:pre-align	*(laughter)* and **chris** is a big lawyer , but i did n't know almost anything about patent legislation and absolutely nothing about genetic .

Table 6: Example translations of GL → EN.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S. J., & Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation?. arXiv preprint arXiv:1804.06323.

APPLICATION

Recommendation Engine



https://towardsdatascience.com/using-word2vec-for-music-recommendations-bb9649ac2484

Out-of-vocabulary Words

- Word2vec doesn't handle this
- FastText handles this because it trains **n-grams** instead of **words** breaking down each word into n-grams
- ELMo also handles this because it trains the model on character-level

amazing -> <am>, <ama>, <maz>, <azi>, <zin>, <ing>, <ng>

min_n = max_n = 3

V1 V2 V3 V4 V5 V6 V7

amazin -> <am>, <ama>, <maz>, <azi>, <zin>, <in>

V1 V2 V3 V4 V5 V9

CHALLENGE

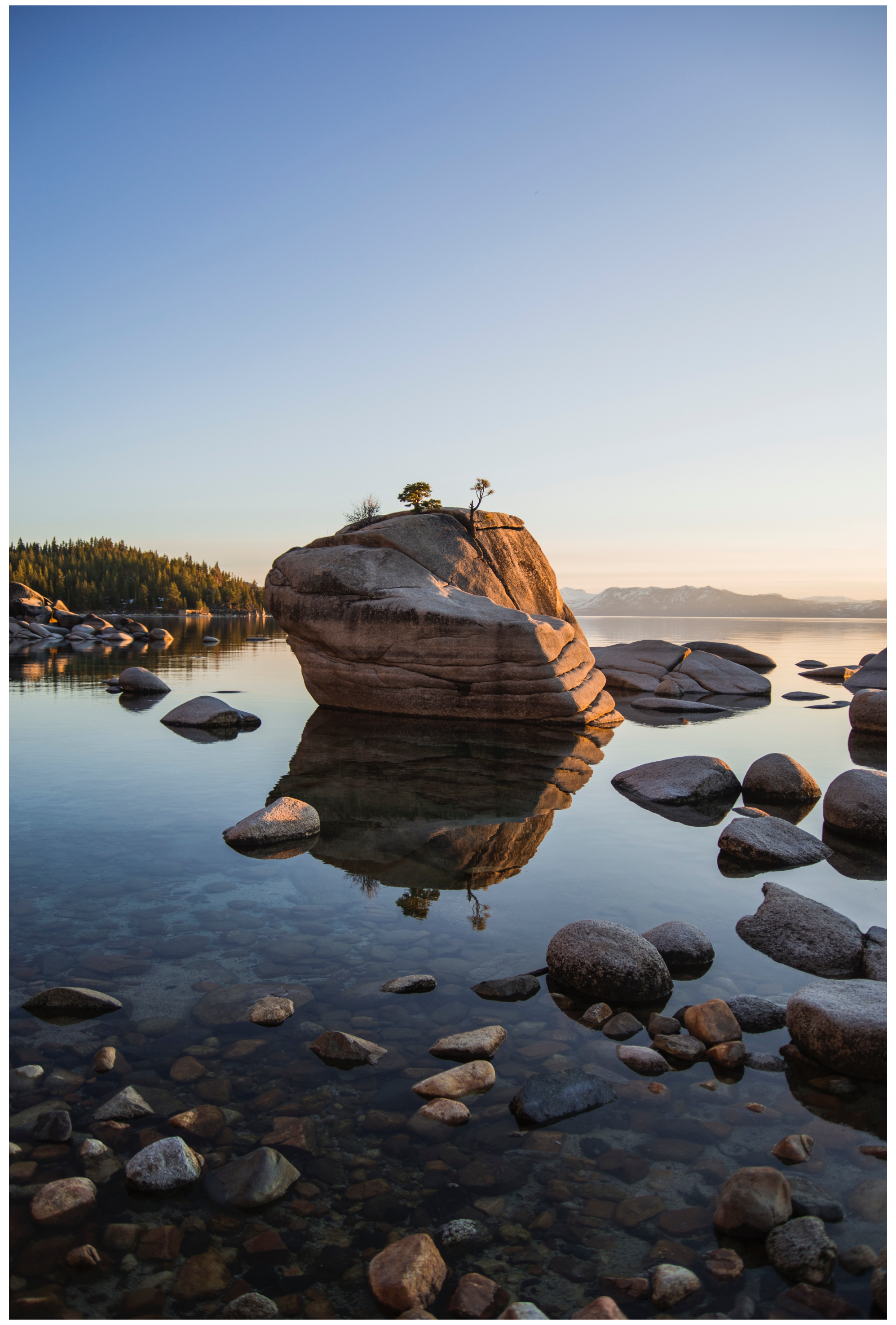
Polysemy

The word “rock”

CHALLENGE

Polysemy

The word “rock”



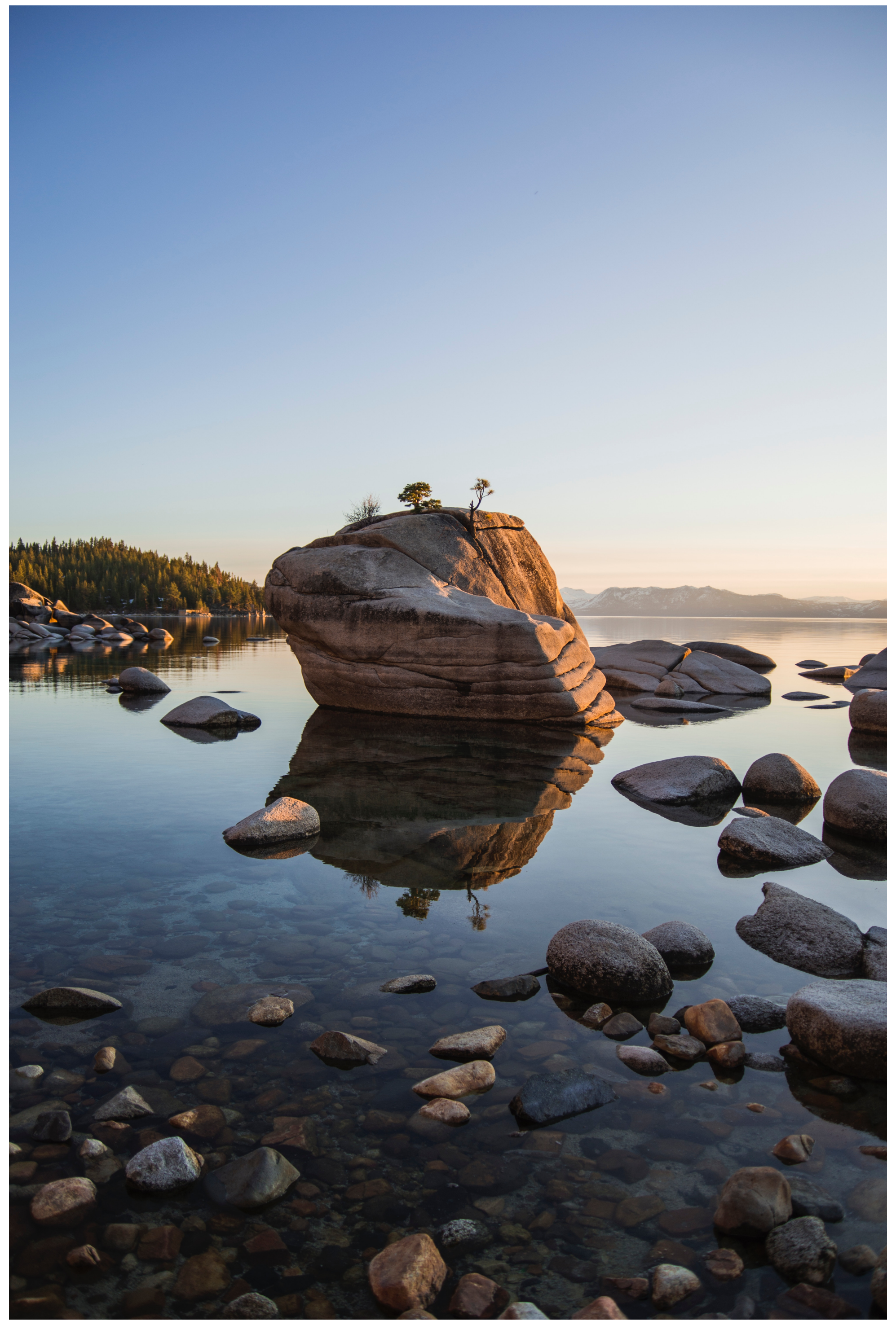
https://unsplash.com/photos/xssEs_oCv-A

https://unsplash.com/photos/I4zSNSxR8oA

CHALLENGE

Polysemy

The word “rock”



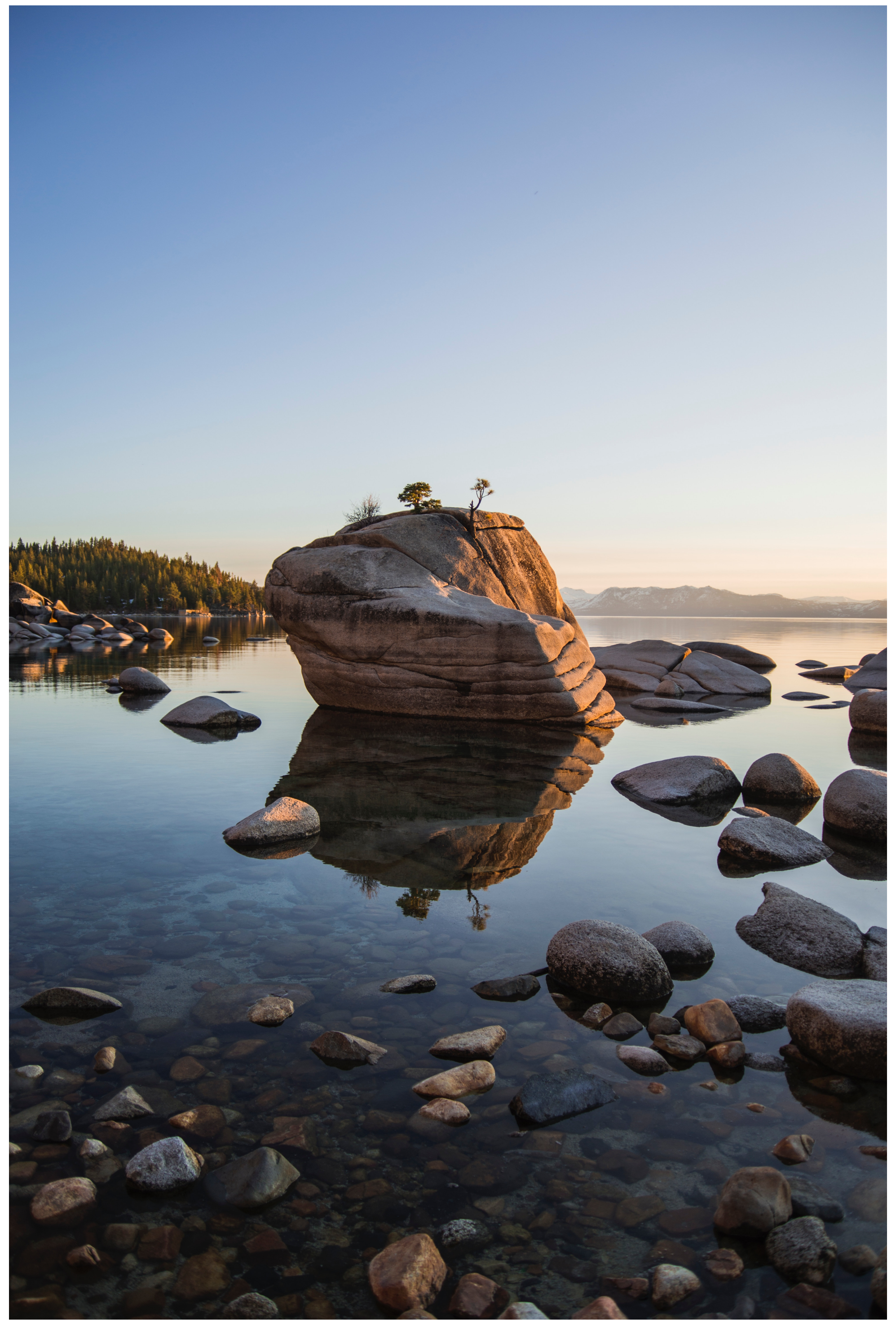
https://unsplash.com/photos/xssEs_oCv-A

https://unsplash.com/photos/I4zSNSxR8oA

CHALLENGE

Polysemy

The word “rock”



https://unsplash.com/photos/xssEs_oCv-A

https://unsplash.com/photos/I4zSNSxR8oA

https://www.muscleandfitness.com/workouts/athletecelebrity-workouts/dwayne-rock-johnsons-shoulder-workout

@galuhsahid

CHALLENGE

Polysemy

The bank at the end of the street was robbed yesterday

He caught a fish at the bank of the river

CHALLENGE

Polysemy

The **bank** at the end of the street was robbed yesterday

He caught a fish at the **bank** of the river

Same word, different meaning

CHALLENGE

Polysemy

The **bank** at the end of the street was robbed yesterday

He caught a fish at the **bank** of the river

Same word, different meaning

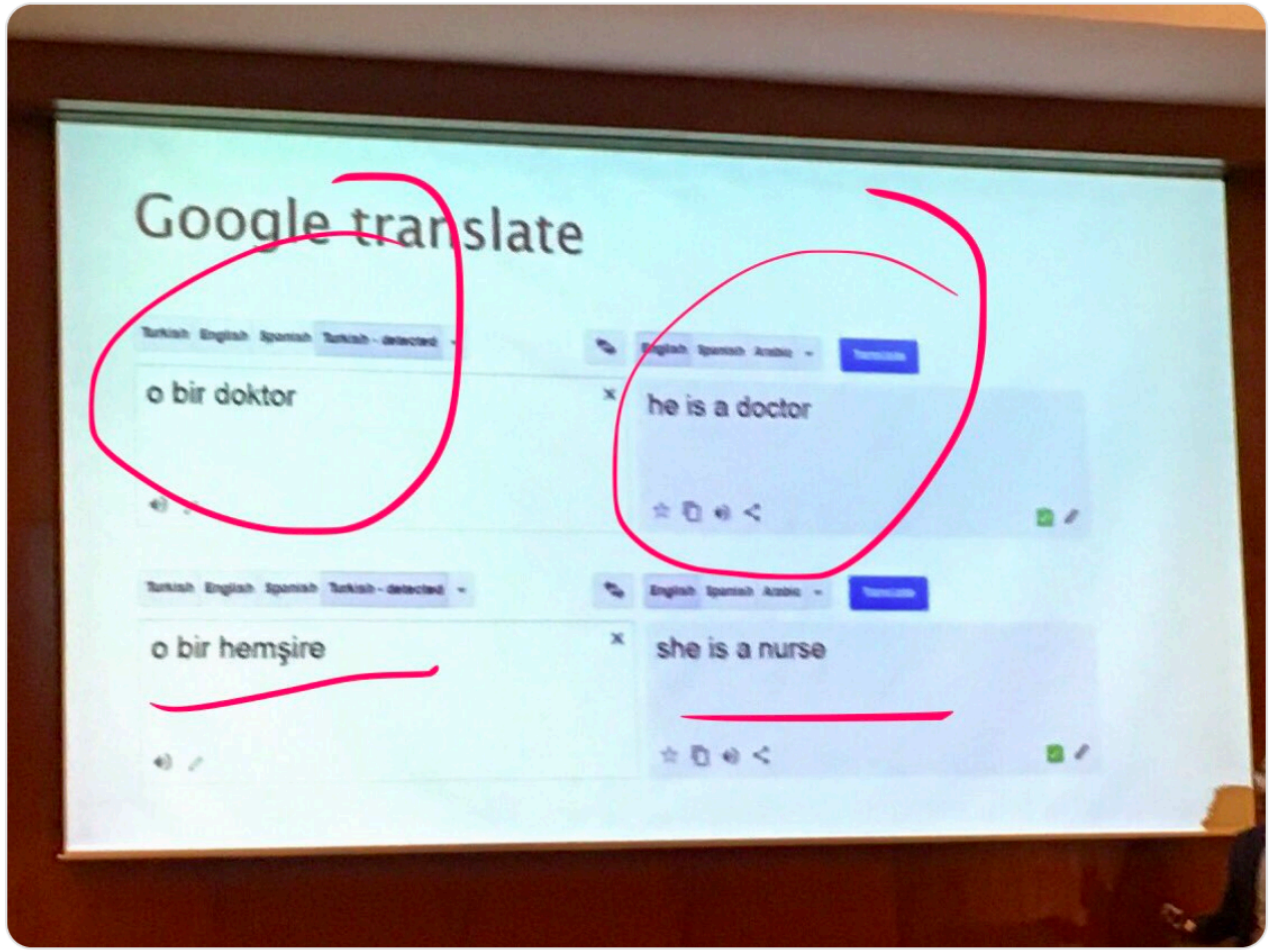
More recent word models such as ELMo and BERT will assign different word vectors for the word "bank" because they appear in different contexts

CHALLENGE

Bias

zeynep tufekci
@zeynep
Following

Google translate: Turkish 3rd person pronoun "O"=gender-neutral. "O" is a doctor ➡️ "He" is a doctor. "O" is a nurse ➡️ "She" is a nurse. #fatml



CHALLENGE

Bias

```
model_w2v.most_similar(positive=["doctor", "she"], negative=["he"])
```

```
[('nurse', 0.6588720679283142),
 ('gynecologist', 0.647172152996063 2),
 ('nurse_practitioner', 0.6255377531051636),
 ('midwife', 0.600278377532959),
 ('pediatrician', 0.5921323299407959),
 ('dermatologist', 0.558222532272 3389),
 ('ob_gyn', 0.5563921928405762),
 ('pharmacist', 0.5559877753257751),
 ('doctors', 0.5544068217277527),
 ('nurse_midwife', 0.554105281829834)]
```

```
model_w2v.most_similar(positive=["doctor", "he"], negative=["she"])
```

```
[('physician', 0.6272262334823608),
 ('surgeon', 0.572301983833313),
 ('doctors', 0.5682945847511292),
 ('orthopedic_surgeon', 0.5621302127838135),
 ('urologist', 0.5399489402770996),
 ('orthopedist', 0.5394338369369507),
 ('neurosurgeon', 0.5322569608688354),
 ('neurologist', 0.5314374566078186),
 ('cardiologist', 0.5162237882614136),
 ('neuro_surgeon', 0.5045366883277893)]
```

@galuhsahid

CHALLENGE

Bias

“Our results indicate that text corpora **contain recoverable and accurate imprints of our historic biases**, whether morally neutral as towards insects or flowers, problematic as towards race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names.”

“Certainly, **caution must be used** in incorporating modules constructed via unsupervised machine learning into decision-making systems.”

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

CHALLENGE

Bias

De-biasing word embeddings:

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357).

CHALLENGE

Bias

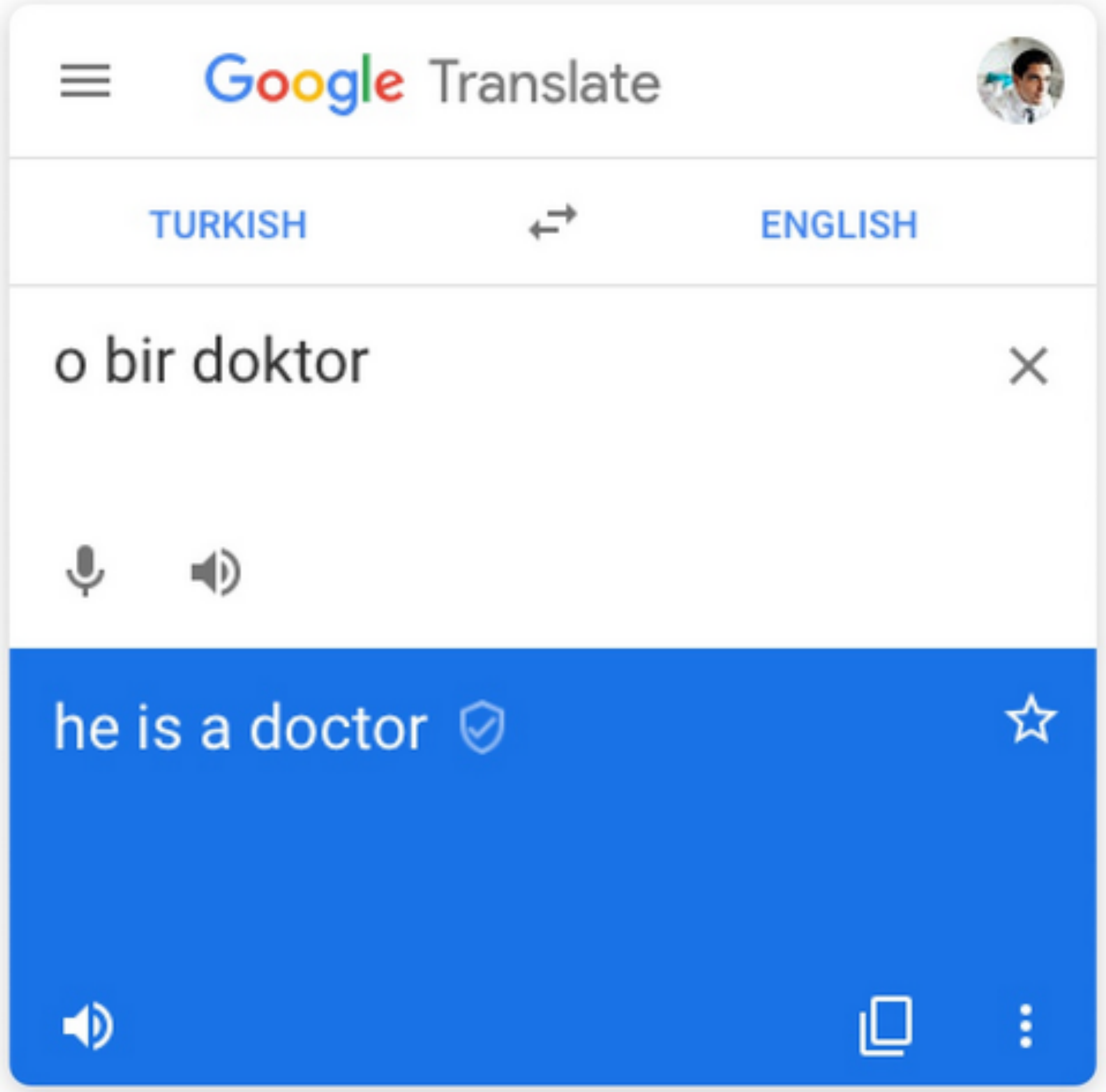
“We conclude that existing bias removal techniques are insufficient, and should not be trusted for providing gender-neutral modeling.”

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. arXiv preprint arXiv:1903.03862.

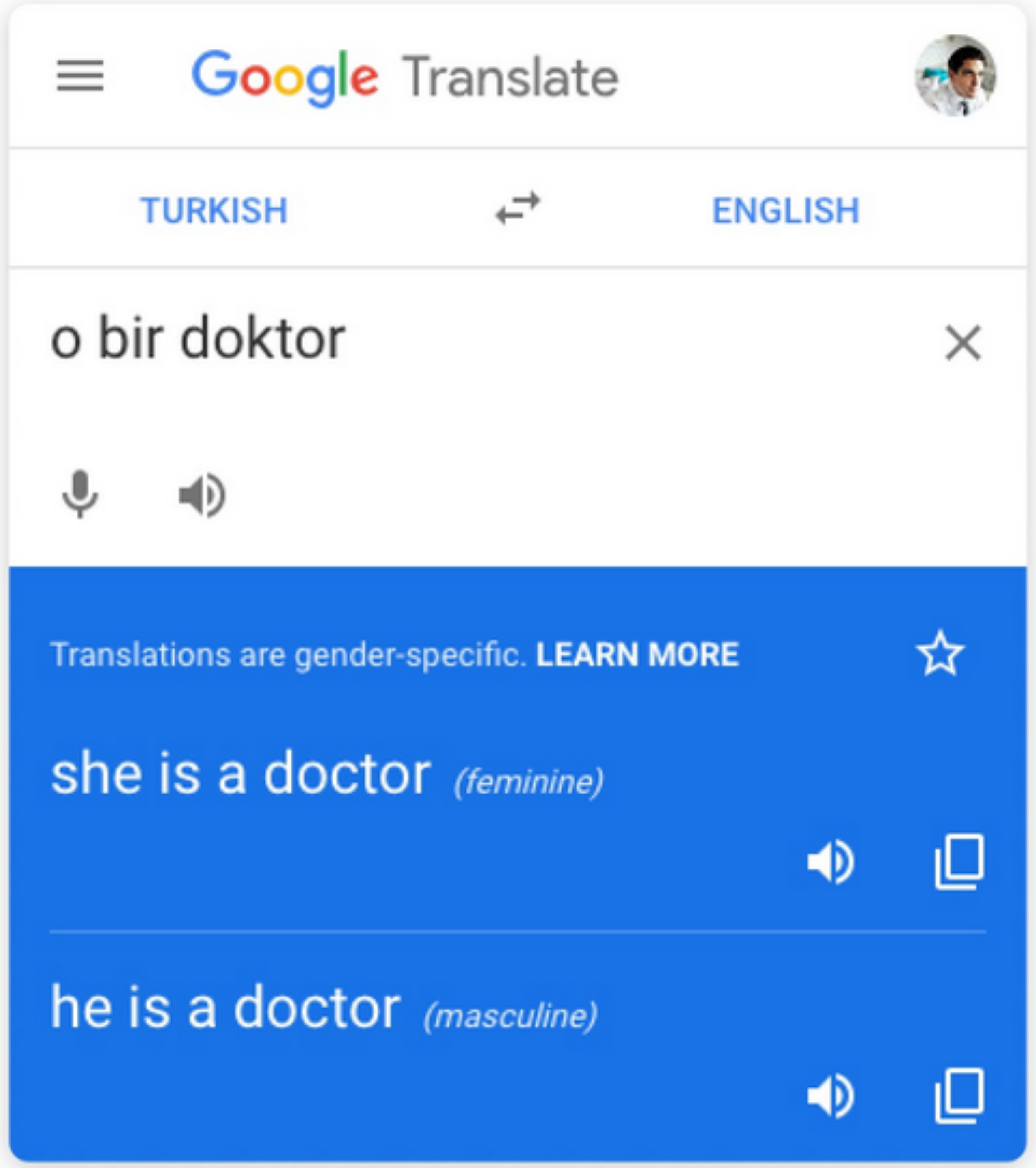
LIMITATION

Bias

Before



After



https://www.blog.google/products/translate/reducing-gender-bias-google-translate/

@galuhsahid

Beyond word2vec

- Explore other techniques such as GloVe, fasttext, ELMo, BERT…
- Train your own word embeddings
- Use word embeddings in NLP tasks (e.g. text classification with doc2vec)
- Use word embeddings for other tasks outside of NLP tasks (song2vec, perhaps?)


T. HANKS