



# **Ethical Aspects of Data *Fairness, Bias, in Data Representation***

*Frederic Precioso*

*28/11/2023*

*(MAASAI, Joint Research Group INRIA-CNRS-UniCA)*

*frederic.precioso@univ-cotedazur.fr*



# License for this content: CC BY-NC-SA



- Training for Data Science & AI Master at UniCA by [Frederic Precioso](#) under Licence [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#).

cc BY NC SA

**Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)**

This is a human-readable summary of (and not a substitute for) the [license](#). [Disclaimer](#).

**You are free to:**

**Share** — copy and redistribute the material in any medium or format

**Adapt** — remix, transform, and build upon the material

[Share your work](#) | [Use & remix](#) | [What We Mean](#)

---

**Under the following terms:**

BY **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

NC **NonCommercial** — You may not use the material for [commercial purposes](#).

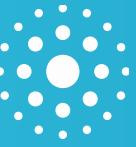
SA **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

**No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

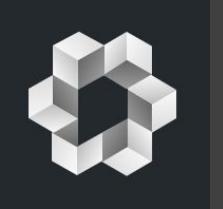


# Overview

- **Context**
- Text representation
- Bias in textual information



# CONTEXT



Conversation AI

# Bias in the Vision and Language of Artificial Intelligence



**Margaret Mitchell**  
*Senior Research Scientist*  
*Google AI*



Andrew  
Zaldivar

Me

Simone  
Wu

Parker  
Barnes

Lucy  
Vasserman

Ben  
Hutchinson

Elena  
Spitzer

Deb  
Raji

Timnit Gebru



Adrian  
Benton

Brian  
Zhang

Dirk  
Hovy

Josh  
Lovejoy

Alex  
Beutel

Blake  
Lemoine

Hee Jung  
Ryu

Hartwig  
Adam

Blaise  
Agüera y  
Arcas

# What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

...We don't tend to say  
**Yellow Bananas**



**What do you see?**

**Green Bananas**

**Unripe Bananas**



# What do you see?

Ripe Bananas

Bananas with spots

Bananas good for  
banana bread



# What do you see?

**Yellow** Bananas?

**Yellow** is prototypical  
for bananas



# Prototype Theory

One purpose of categorization is to **reduce the infinite differences** among stimuli **to behaviourally and cognitively usable proportions**

There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975)

May also store exemplars (Wu & Barsalou, 2009)



Fruit



Bananas  
“Basic Level”



Unripe Bananas,  
Cavendish Bananas

---

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

---



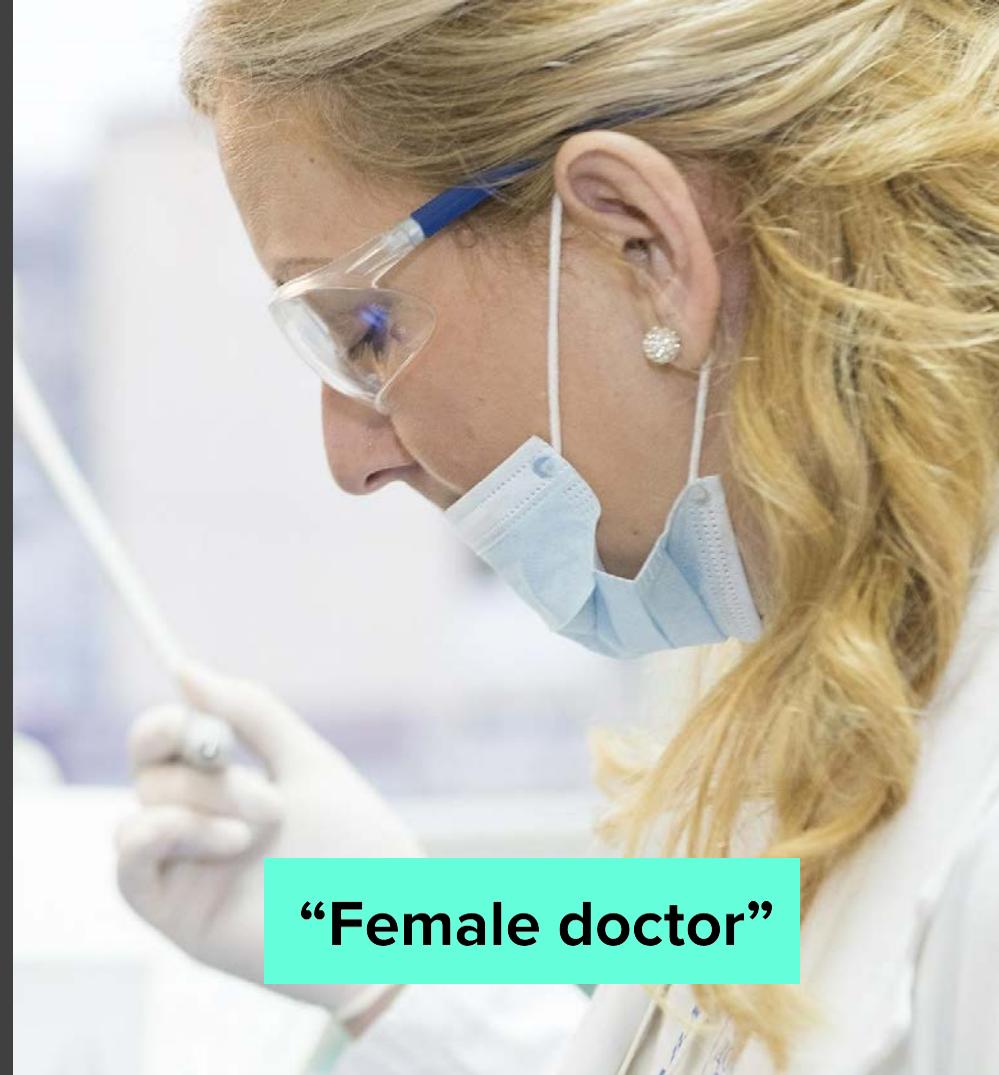
---

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

**How could this be?**

---



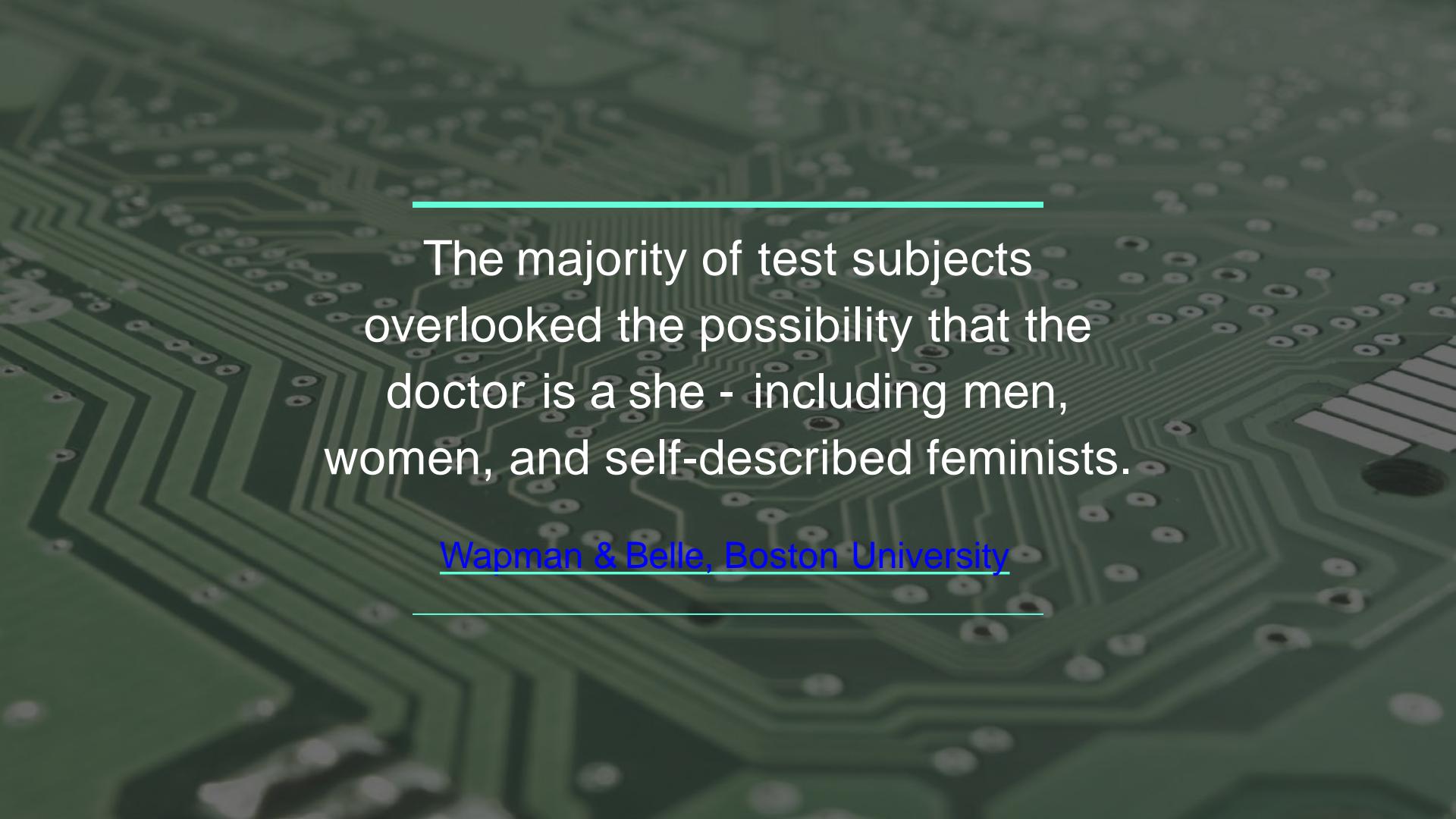
**“Female doctor”**



**“Doctor”**



**“Female doctor”**

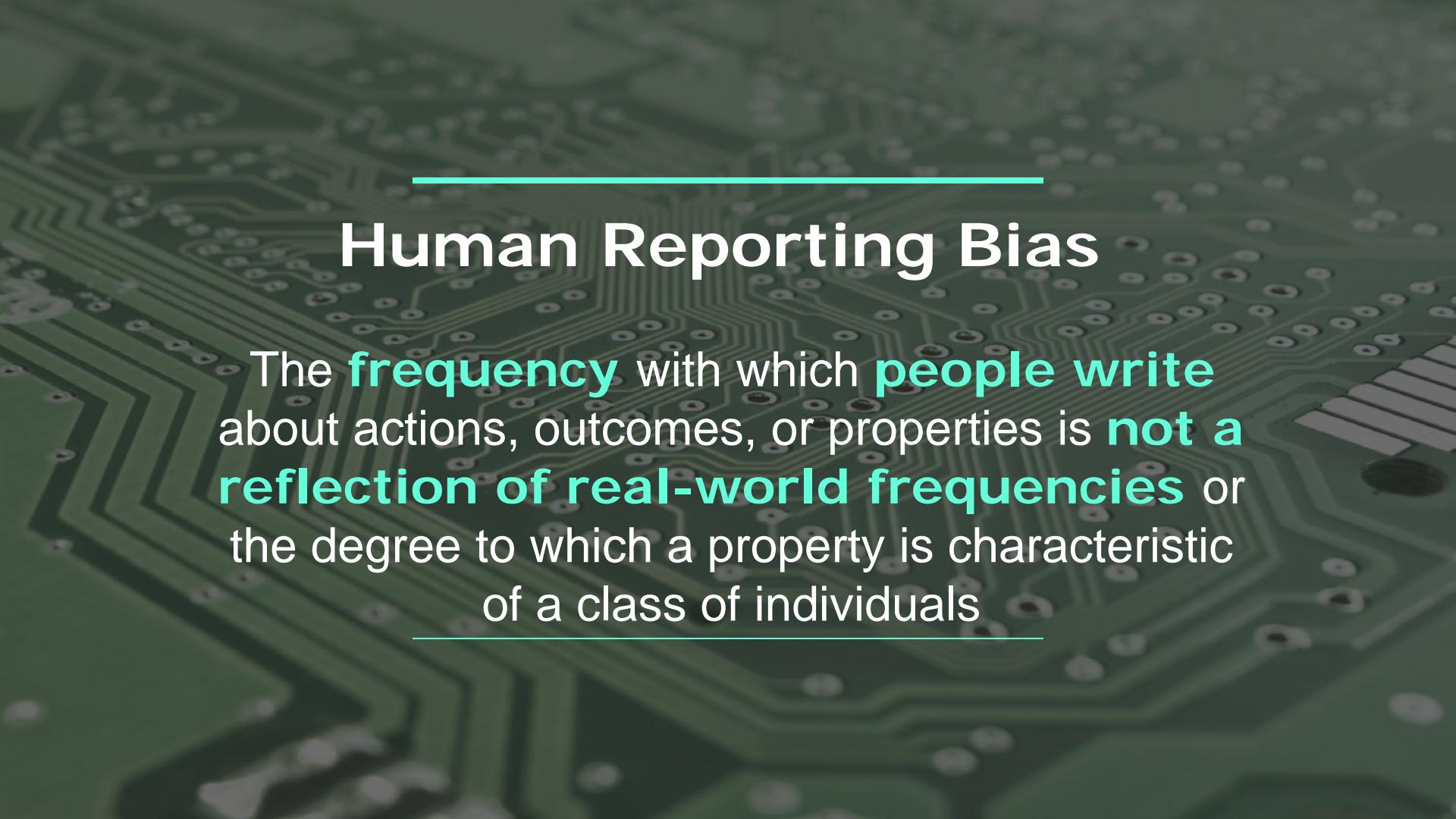


---

The majority of test subjects  
overlooked the possibility that the  
doctor is a she - including men,  
women, and self-described feminists.

---

[Wapman & Belle, Boston University](#)



---

## Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

---

# Bias in Language

Extreme <i>she</i> occupations		
1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme <i>he</i> occupations		
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. fighter pilot	12. boss

Figure 1: The most extreme occupations as projected on to the *she-he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

Gender stereotype <i>she-he</i> analogies.		
sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate <i>she-he</i> analogies.		
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 2: **Analogy examples.** Examples of automatically generated analogies for the pair *she-he* using the procedure described in text. For example, the first analogy is interpreted as *she:sewing :: he:carpentry* in the original w2vNEWS embedding. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype. Top: illustrative gender stereotypical analogies automatically generated from w2vNEWS, as rated by at least 5 of the 10 crowd-workers. Bottom: illustrative generated gender-appropriate analogies.

# Bias in Language

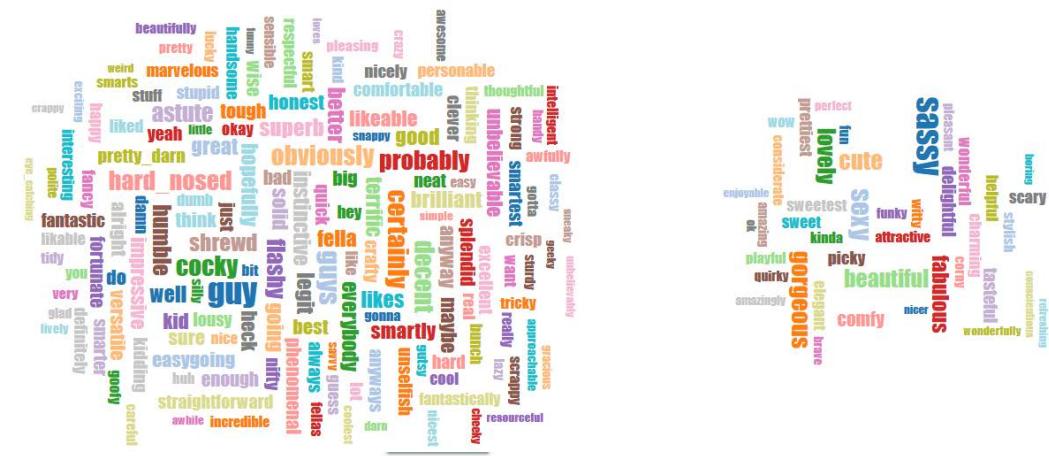
he (158)

## Adjectives

Or type your own words...

doctor

she (42)



he (47)

she (153)

prescribe bedside palliative care physical therapist dialysis anesthesia pharmaceuticals vaccinations realtor pill schoolteacher therapist pharmacist

diagnoses dentists physical therapist thyroid anesthesia pharmacy pill

acupuncture

vet

midwives

veterinarians

dentist

obstetrics

pregnatal

husband

nurses

pregnant

hairdresser

cosmetic surgery

clinicians

anesthetic

medication

prescription pharmacies

pill

midwife

maternity

midwife

mother

parents

meds

pediatrician

obstetrician

outpatient

ultrasound

therapists

immunizations

symptoms

physical therapy

clinician prescribing

psychotherapy

pediatricians

diabetic

guidance

counselor

pharmacist

veterinarian

medicines

counselor

dentist

dermatologist

autopsy

teenager

# But...don't forget the biases in the initial data

Man is to Computer Programmer as Woman is to Homemaker?  
Debiasing Word Embeddings

Tolga Bolukbasi<sup>1</sup>, Kai-Wei Chang<sup>2</sup>, James Zou<sup>2</sup>, Venkatesh Saligrama<sup>1,2</sup>, Adam Kalai<sup>2</sup>

<sup>1</sup>Boston University, 8 Saint Mary's Street, Boston, MA

<sup>2</sup>Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

# Bias in Vision



Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

# Bias in Vision

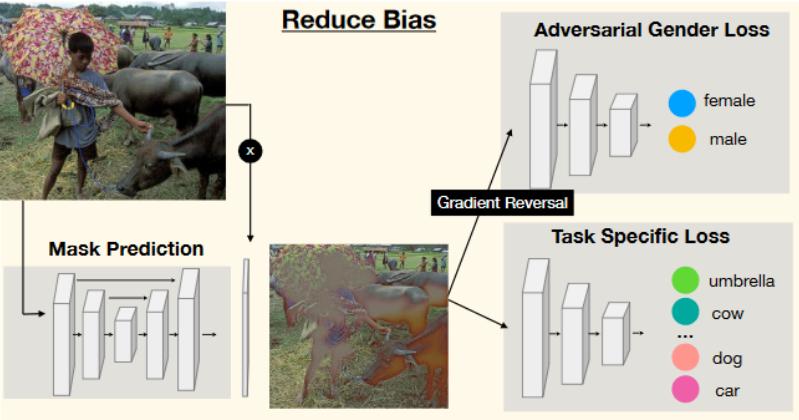


Figure 2. In our bias mitigation approach, we learn a task-specific model with an adversarial loss that removes features corresponding to a protected variable from an intermediate representation in the model – here we illustrate our pipeline to visualize the removal of features in image space through an auto-encoder network.

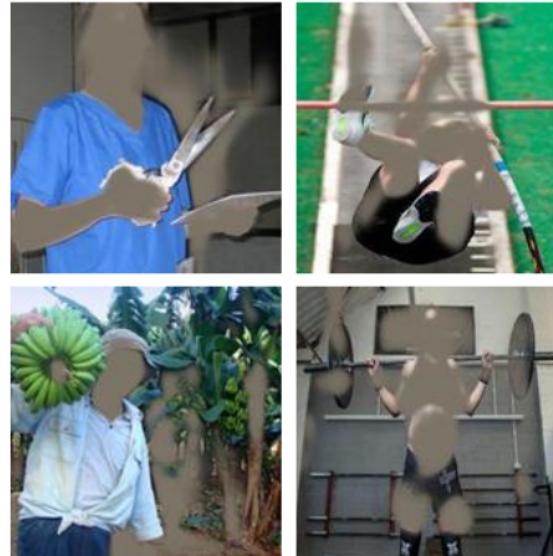
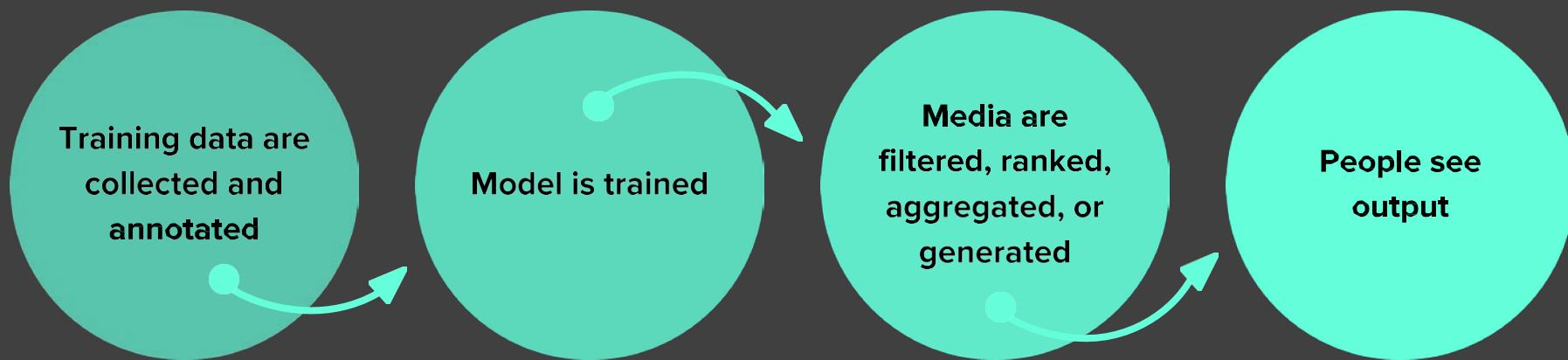


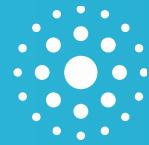
Figure 3. Images after adversarial removal of gender when applied to the image space. The objective was to preserve information about objects and verbs, e.g. scissors, banana (COCO) or vaulting, lifting (imSitu) while removing gender correlated features.





# Overview

- Context
- **Text representation**
- Bias in textual information



# TEXT REPRESENTATION



# How do we represent the meaning of a word?

Definition: meaning (Webster dictionary)

- the idea that is represented by a word, phrase, etc.
- the idea that a person wants to express by using words, signs, etc.
- the idea that is expressed in a work of writing, art, etc.

Commonest linguistic way of thinking of meaning:

signifier (symbol)  $\Leftrightarrow$  signified (idea or thing)

= denotational semantics



# How do we have usable meaning in a computer?

univ-cotedazur.fr

Common solution: Use e.g. [WordNet](#), a thesaurus containing lists of synonym sets and hypernyms (“is a” relationships).

e.g. synonym sets containing “good”:

```
from nltk.corpus import wordnet as wn
poses = { 'n':'noun', 'v':'verb', 's':'adj (s)', 'a':'adj', 'r':'adv'}
for synset in wn.synsets("good"):
    print("{}: {}".format(poses[synset.pos()],
                          ", ".join([l.name() for l in synset.lemmas()])))
```

noun: good  
noun: good, goodness  
noun: good, goodness  
noun: commodity, trade\_good, good  
adj: good  
adj (sat): full, good adj:  
good  
adj (sat): estimable, good, honorable, respectable adj (sat):  
beneficial, good  
adj (sat): good  
adj (sat): good, just, upright  
...  
adverb: well, good  
adverb: thoroughly, soundly, good

e.g. hypernyms of “panda”:

```
from nltk.corpus import wordnet as wn
panda = wn.synset("panda.n.01") hyper =
lambda s: s.hypernyms()
list(pandaclosure(hyper))
```

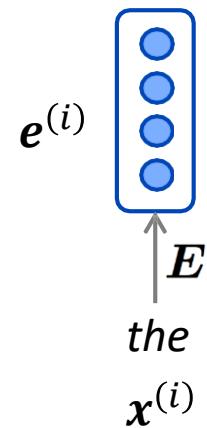
[Synset('procyonid.n.01'),  
Synset('carnivore.n.01'),  
Synset('placental.n.01'),  
Synset('mammal.n.01'),  
Synset('vertebrate.n.01'),  
Synset('chordate.n.01'),  
Synset('animal.n.01'),  
Synset('organism.n.01'),  
Synset('living\_thing.n.01'),  
Synset('whole.n.02'),  
Synset('object.n.01'),  
Synset('physical\_entity.n.01'),  
Synset('entity.n.01')]



- Great as a resource but missing nuance
  - e.g. “proficient” is listed as a synonym for “good”. This is only correct in some contexts.
- Missing new meanings of words
  - e.g., wicked, badass, nifty, wizard, genius, ninja, bombest
  - Impossible to keep up-to-date!
- Subjective
- Requires human labor to create and adapt
- Can't compute accurate word similarity



# First let us now detail how to represent words as vectors: *word embedding*



# Representing words as discrete symbols

In traditional NLP, we regard words as discrete symbols:  
**hotel, conference, motel** - a **localist** representation

Means one 1, the rest 0s

Words can be represented by **one-hot** vectors:

**motel** = [0 0 0 0 0 0 0 0 0 1 0 0 0 0]  
**hotel** = [0 0 0 0 0 0 1 0 0 0 0 0 0 0]

Vector dimension = number of words in vocab (e.g. 500,000)



# Problem with words as discrete symbols

[univ-cotedazur.fr](http://univ-cotedazur.fr)

Example: in web search, if user searches for “Seattle motel”, we would like to match documents containing “Seattle hotel”.

But:

$$\text{motel} = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0]$$

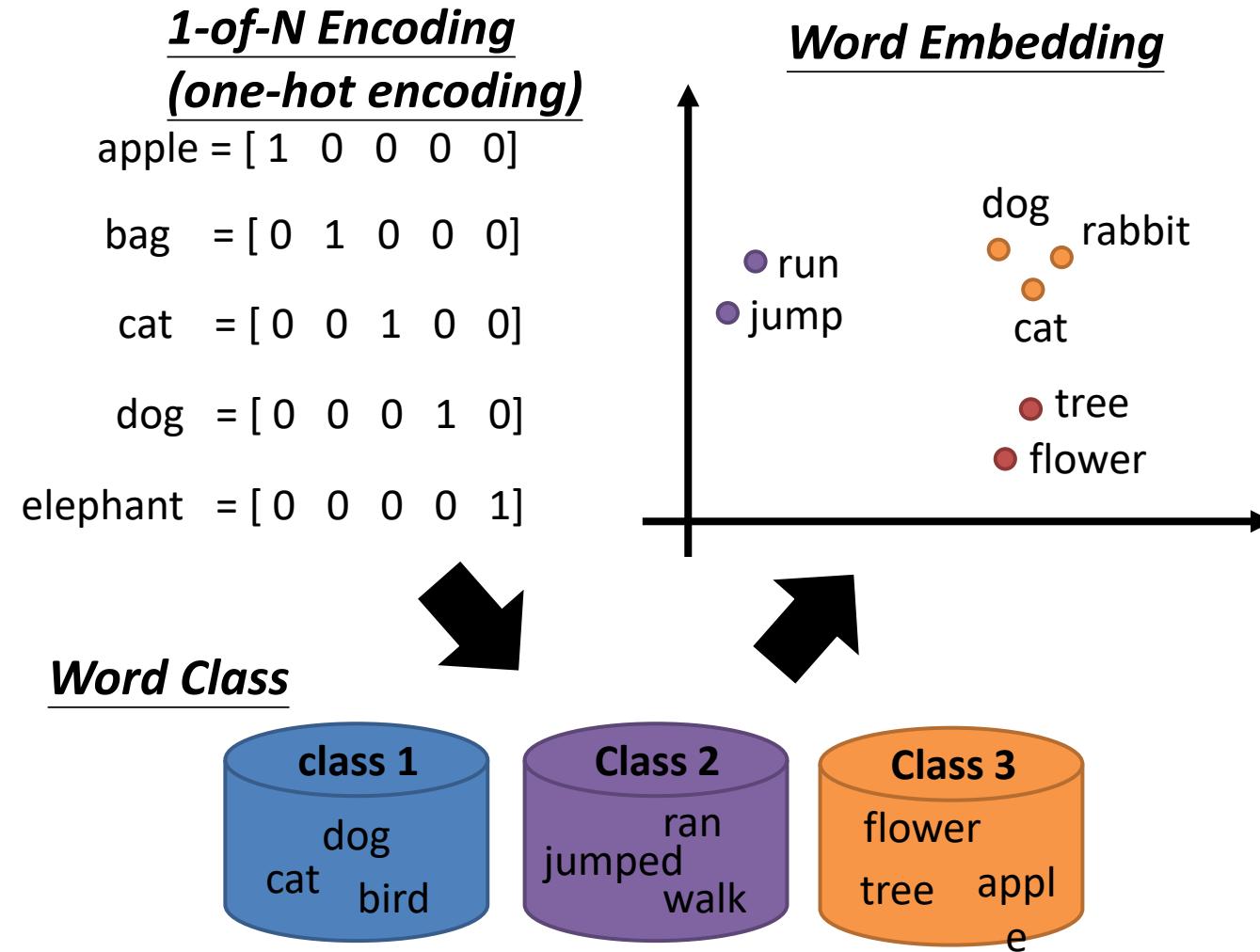
$$\text{hotel} = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$$

These two vectors are orthogonal.

There is no natural notion of similarity for one-hot vectors!

Solution:

- Could try to rely on WordNet’s list of synonyms to get similarity?
  - But it is well-known to fail badly: incompleteness, etc.
- Instead: learn to encode similarity in the vectors themselves



# How to exploit the context?

- **Count based**

- If two words  $w_i$  and  $w_j$  frequently co-occur,  $V(w_i)$  and  $V(w_j)$  would be close to each other
- E.g. Glove Vector:  
<http://nlp.stanford.edu/projects/glove/>

$$V(w_i) \cdot V(w_j) \longleftrightarrow N_{i,j}$$

Inner product

Number of times  $w_i$  and  $w_j$   
in the same document

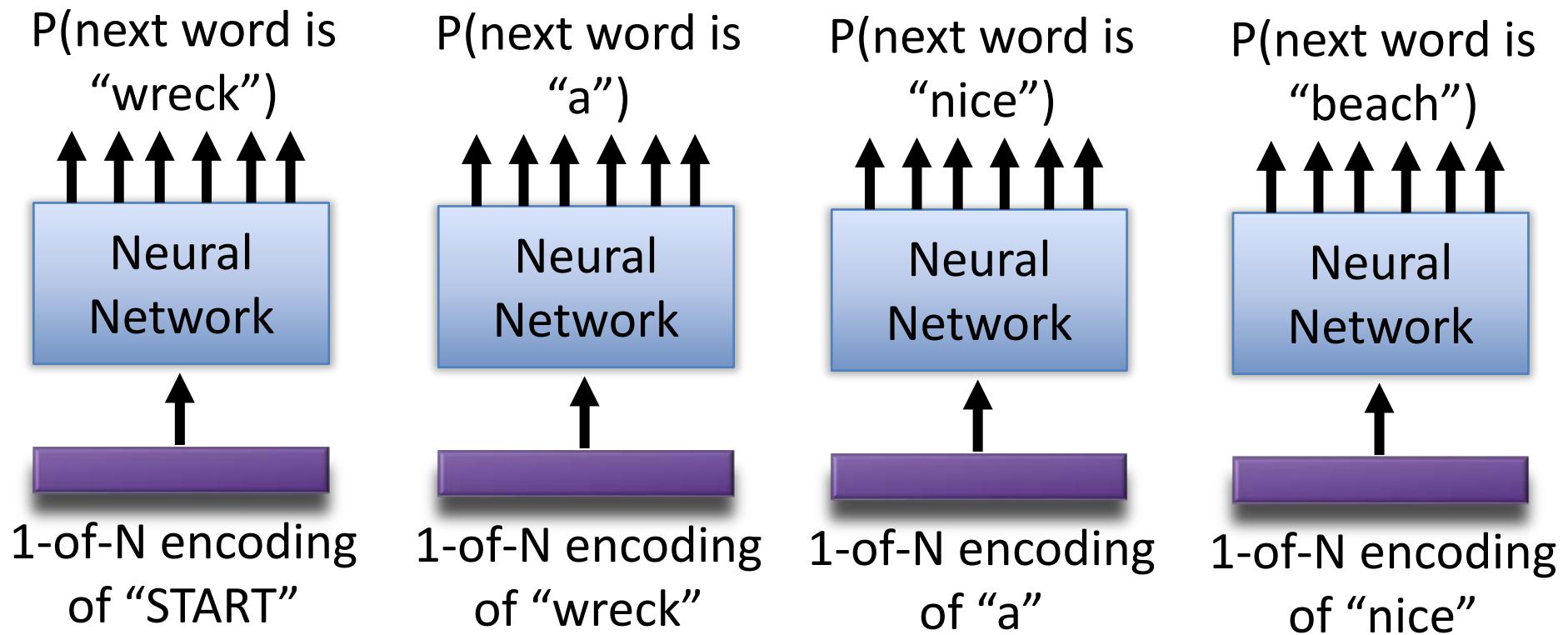
- **Prediction-based..**

# Prediction-based – Language Modeling

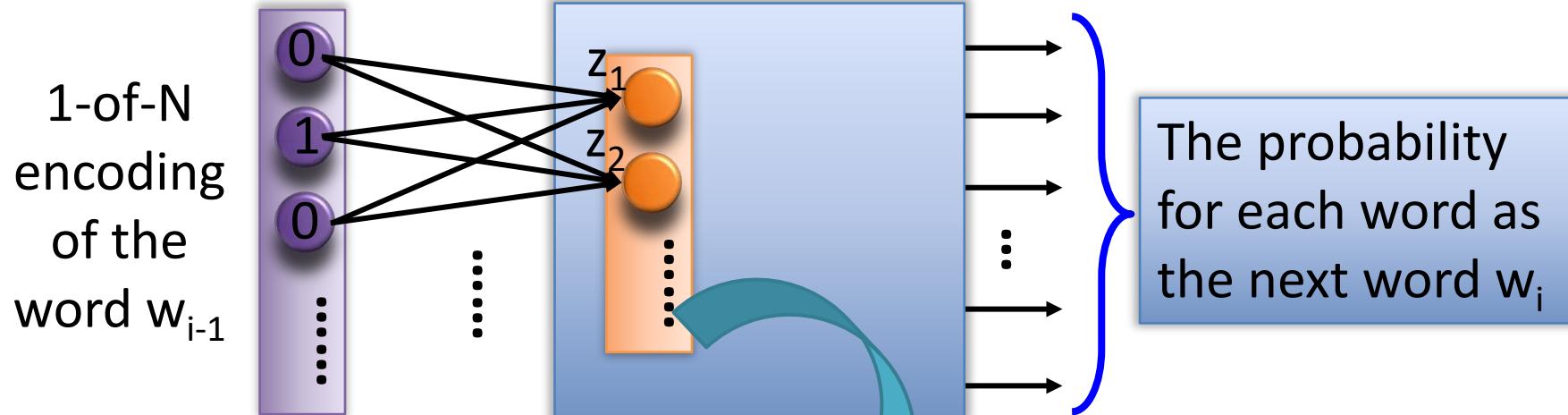
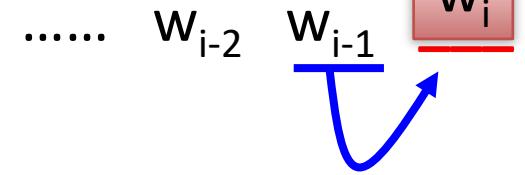
$$P(\text{"wreck a nice beach"})$$

$$= P(\text{wreck} \mid \text{START}) P(\text{a} \mid \text{wreck}) P(\text{nice} \mid \text{a}) P(\text{beach} \mid \text{nice})$$

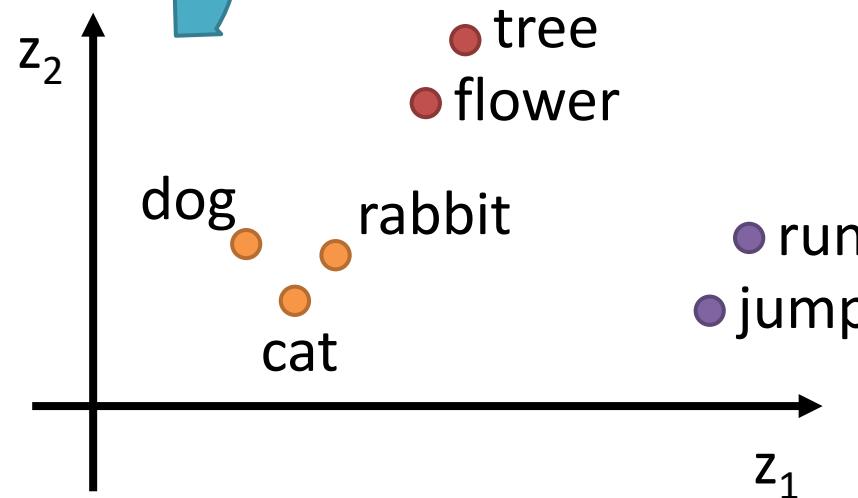
$P(b|a)$ : the probability of NN predicting the next word.



# Prediction-based



- Take out the input of the neurons in the first layer
- Use it to represent a word  $w$
- Word vector, word embedding feature:  $V(w)$





# Representing words by their context



- Distributional semantics: A word's meaning is given by the words that frequently appear close-by
  - “*You shall know a word by the company it keeps*” (J. R. Firth 1957)
  - One of the most successful ideas of modern statistical NLP!
  - However, this is a deliberate choice!!
- When a word  $w$  appears in a text, its **context** is the set of words that appear nearby (within a fixed-size window).
- Use the many contexts of  $w$  to build up a representation of  $w$

...government debt problems turning into **banking** crises as happened in 2009...

...saying that Europe needs unified **banking** regulation to replace the hodgepodge...

...India has just given its **banking** system a shot in the arm...



These **context words** will represent **banking**



# What can we learn from reconstructing the input?

Stanford University is located in \_\_\_\_\_, California.

# What can we learn from reconstructing the input?

I put \_\_\_\_ fork down on the table.

# What can we learn from reconstructing the input?

The woman walked across the street,  
checking for traffic over \_\_\_\_ shoulder.



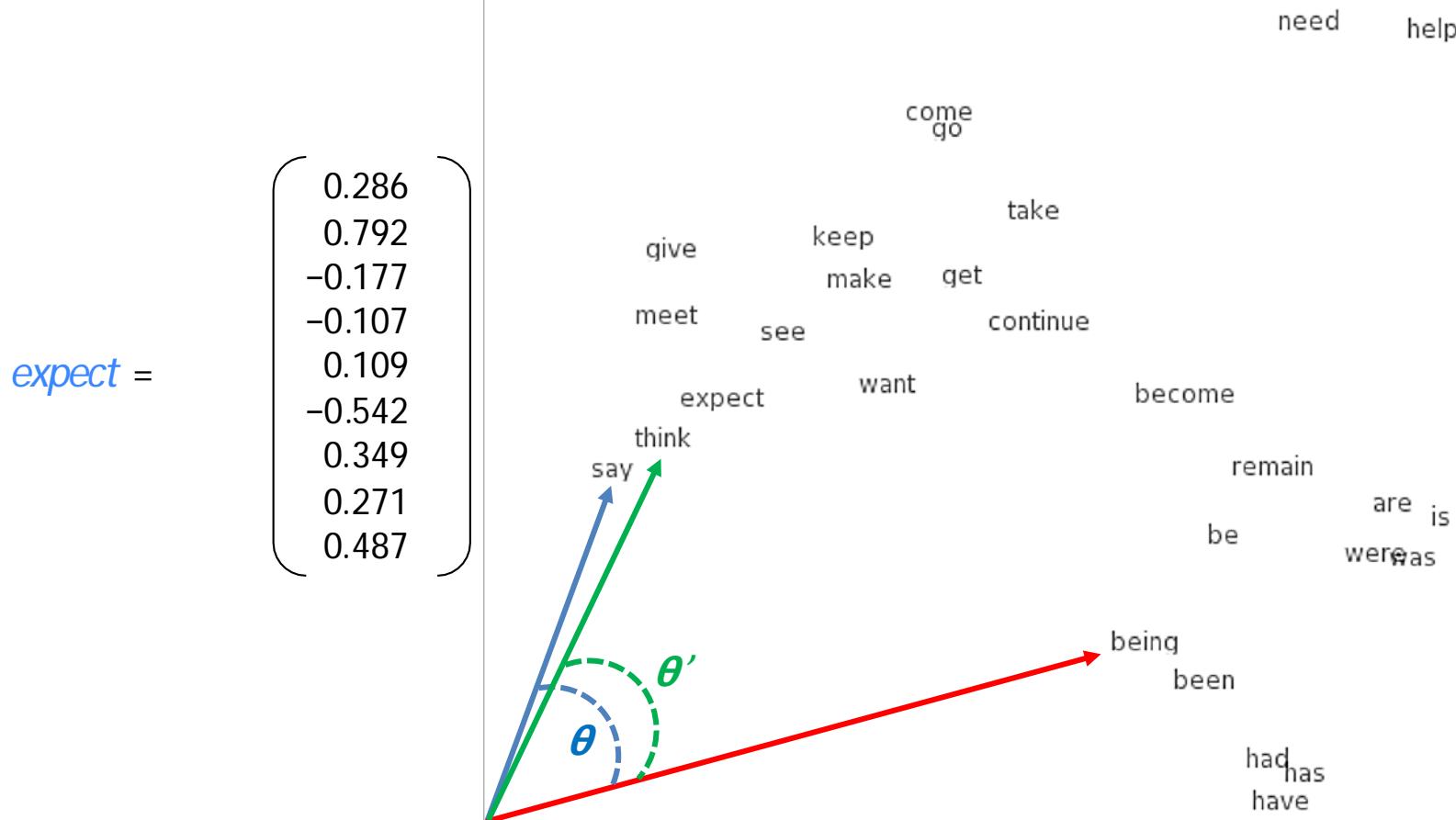
# Word vectors

We will build a dense vector for each word, chosen so that it is similar to vectors of words that appear in similar contexts

$$\text{banking} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Note: word vectors are sometimes called word embeddings or word representations. They are a distributed representation.

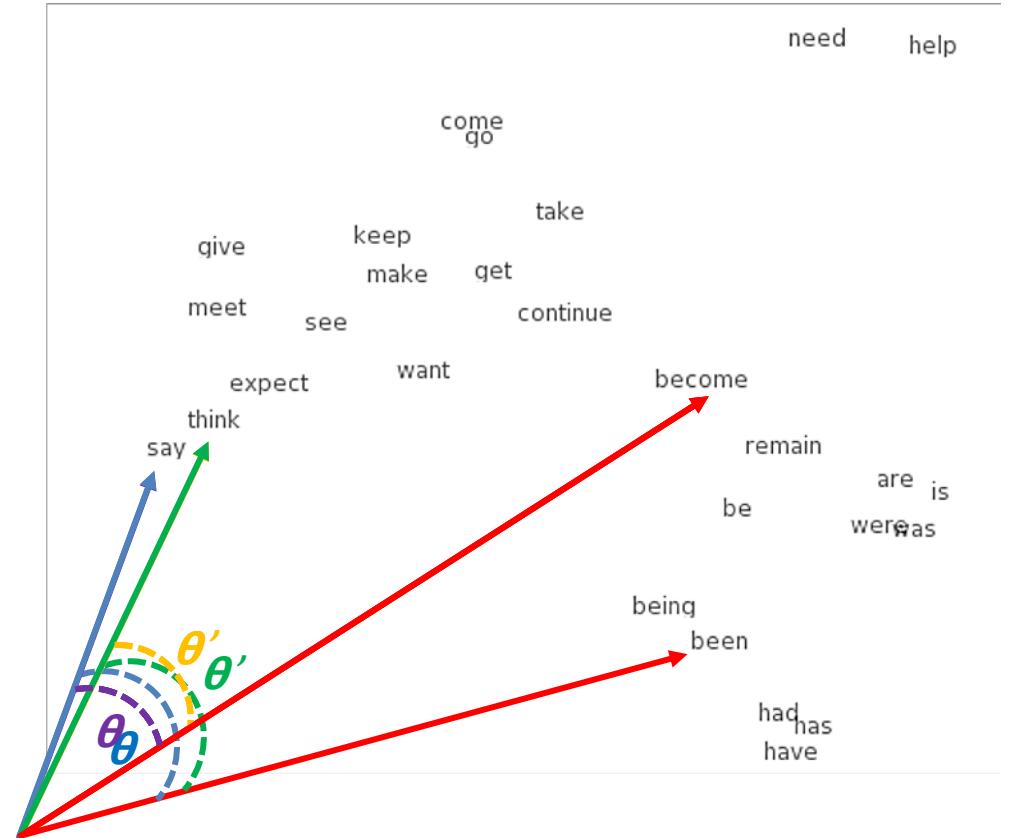
# Word meaning as a neural word vector – visualization



# Word meaning as a neural word vector – visualization

$$\theta' \approx \theta \Leftrightarrow \text{think} \approx \text{say}$$

$$\theta' \approx \theta \Leftrightarrow \text{think} \approx \text{say}$$





# Word2Vec Overview

Word2vec (Mikolov et al. 2013) is a framework for learning word vectors

Idea:

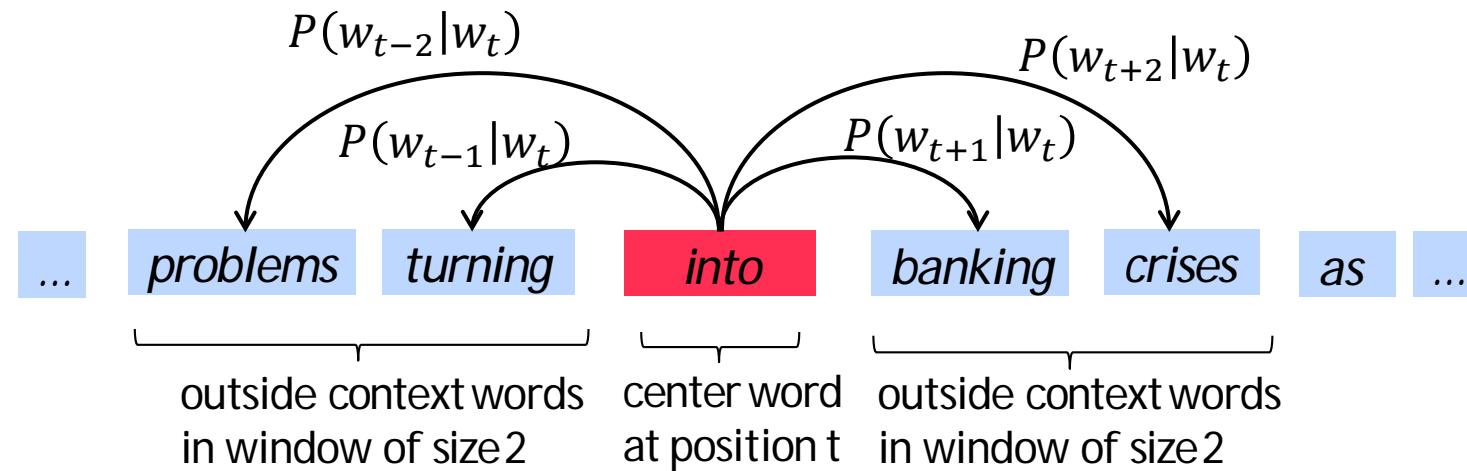
- We have a large corpus of text
- Every word in a fixed vocabulary is represented by a vector
- Go through each position  $t$  in the text, which has a center word  $c$  and context (“outside”) words  $o$
- Use the similarity of the word vectors for  $c$  and  $o$  to calculate the probability of  $o$  given  $c$  (or vice versa)
- Keep adjusting the word vectors to maximize this probability



# Word2Vec Overview

univ-cotedazur.fr

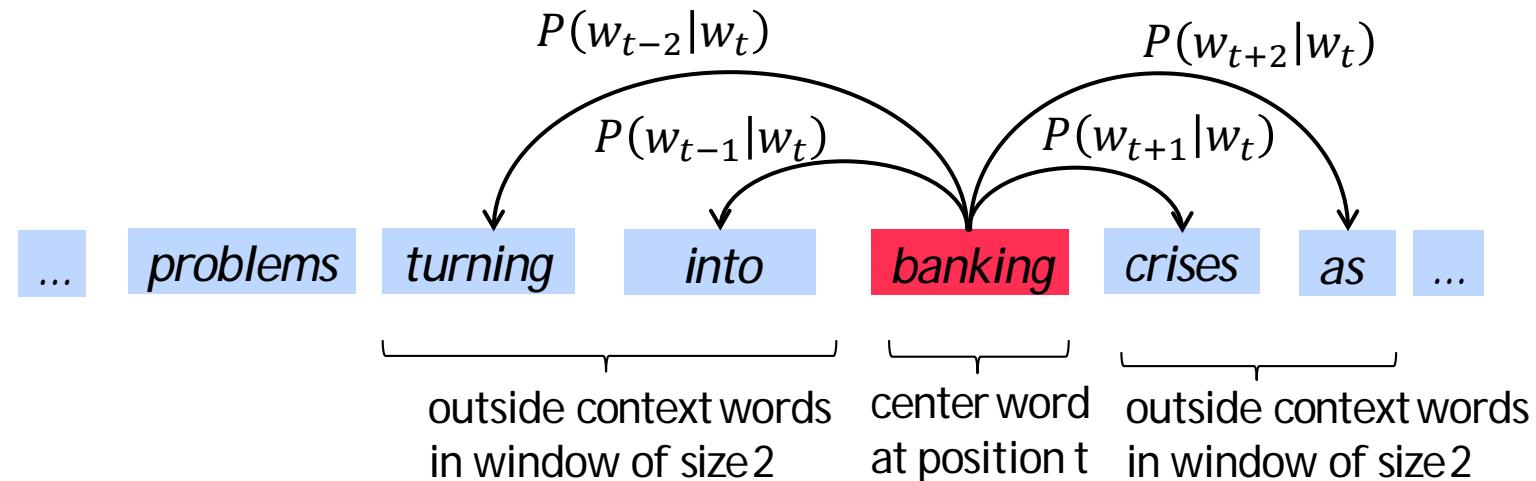
- Example windows and process for computing  $P(w_{t+j}|w_t)$





# Word2Vec Overview

- Example windows and process for computing  $P(w_{t+j}|w_t)$





# Word2Vec: objective function

- We want to minimize the objective function, the (average) negative log likelihood :

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^T \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \log P(w_{t+j} | w_t; \theta)$$

Minimizing objective function  $\Leftrightarrow$  Maximizing predictive accuracy

- Question: How to calculate  $P(w_{t+j} | w_t; \theta)$ ?
- Answer: We will use two vectors per word  $w$ :
  - $V_w$  when  $w$  is a *center* word
  - $u_w$  when  $w$  is a *context* word
- Then for a center word  $c$  and a context word  $o$ :

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$



# Word2Vec: prediction function

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

Exponentiation makes anything positive

Dot product compares similarity of  $o$  and  $c$ .  
 $u^T v = u \cdot v = \sum_{i=1}^n u_i v_i$   
Larger dot product = larger probability

Normalize over entire vocabulary to give probability distribution

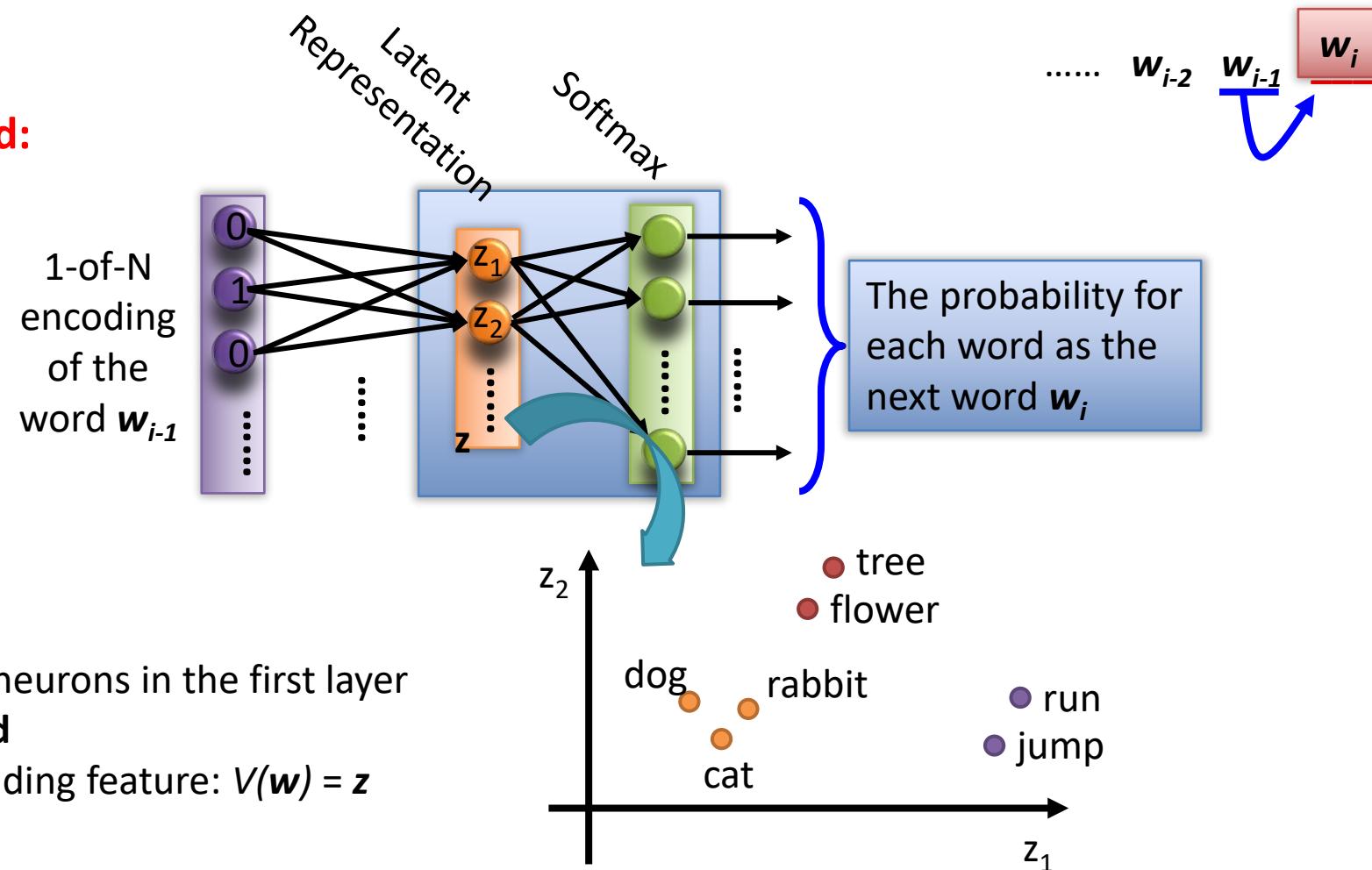
- This is an example of the softmax function  $\mathbb{R}^n \rightarrow \mathbb{R}^n$

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^n \exp(x_j)} = p_i$$

- The softmax function maps arbitrary values to a probability distribution  $p_i$ 
  - “max” because amplifies probability of largest  $x_i$ ,
  - “soft” because still assigns some probability to smaller  $x_i$

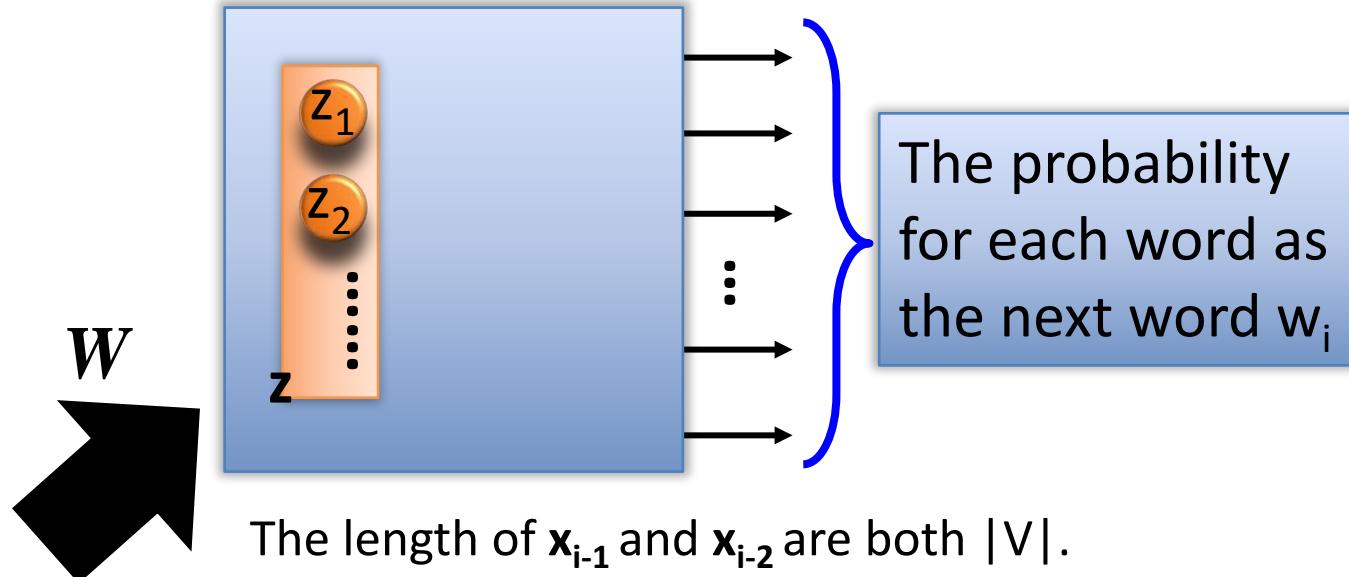
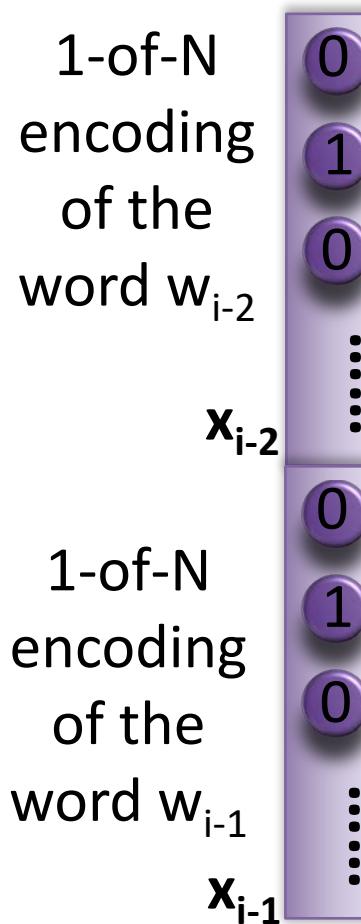
# Reminder - Glove

Predicting the next word:



- Take out the **input** of the neurons in the first layer
- **Use it to represent a word**
- Word vector, word embedding feature:  $V(w) = z$

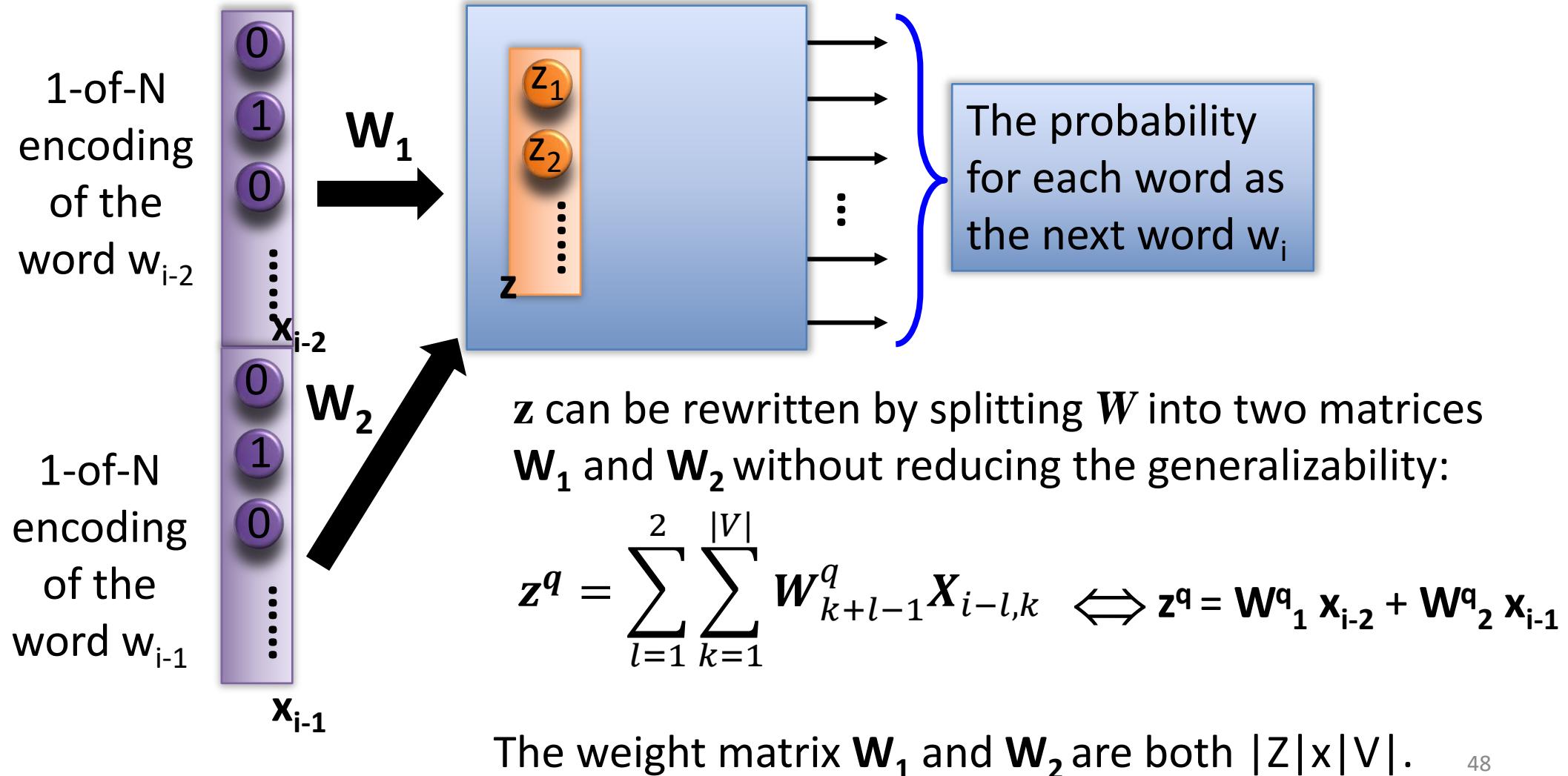
# Prediction-based – Sharing Parameters



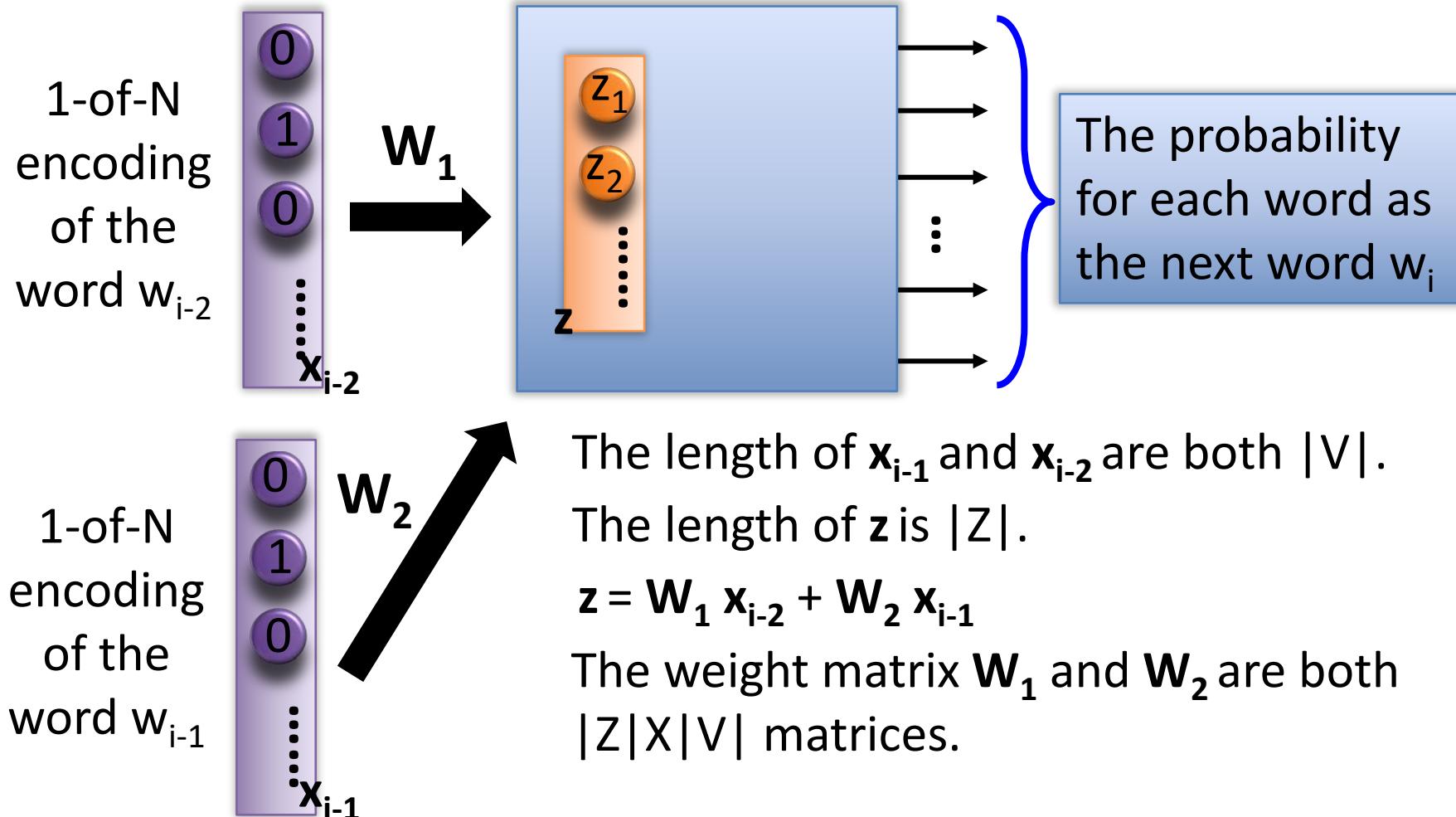
The length of  $x_{i-1}$  and  $x_{i-2}$  are both  $|V|$ .  
We can consider the “two-previous” vectors as concatenated and so to become one vector  $X$  of size  $2x|V|$ .  
The whole weights are  $W$ , and  $X$  to  $W$  are **fully connected**.

$$\Rightarrow z^q = \sum_{i=1}^{2|V|} W_i^q X_i \iff z^q = \sum_{l=1}^2 \sum_{k=1}^{|V|} W_{k+l-1}^q X_{i-l,k}$$

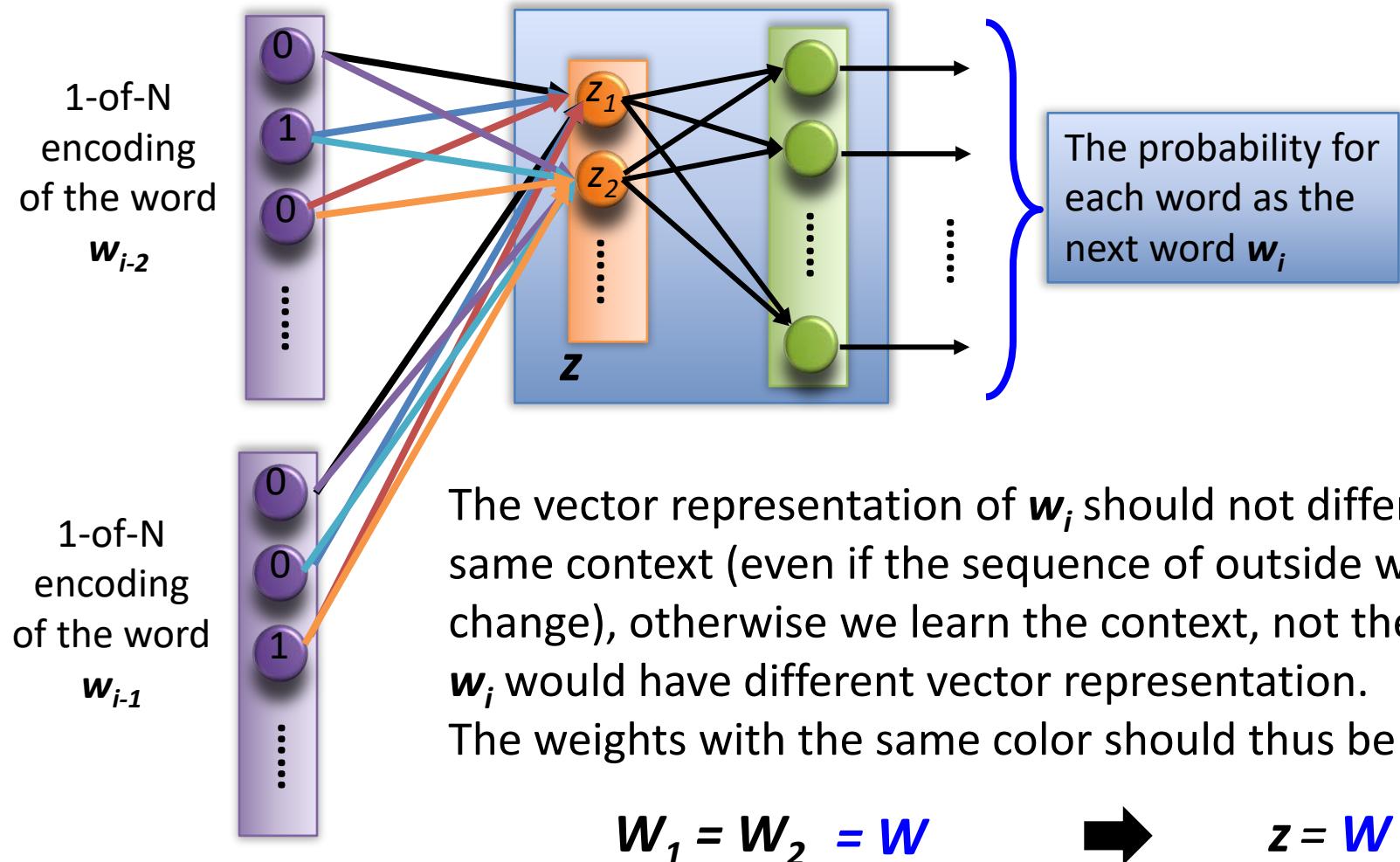
# Prediction-based – Sharing Parameters



# Prediction-based – Sharing Parameters

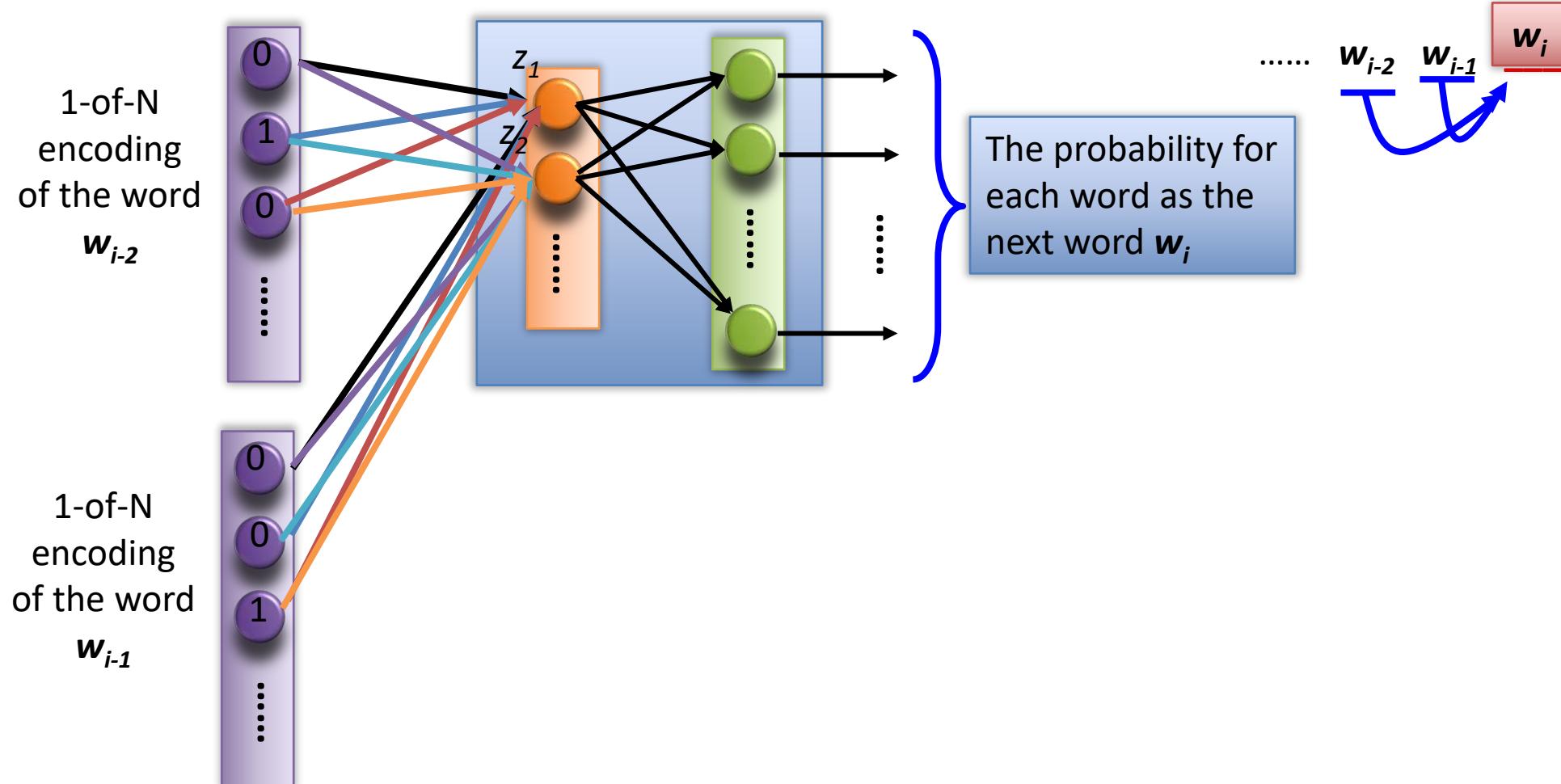


# Prediction-based – Sharing Parameters

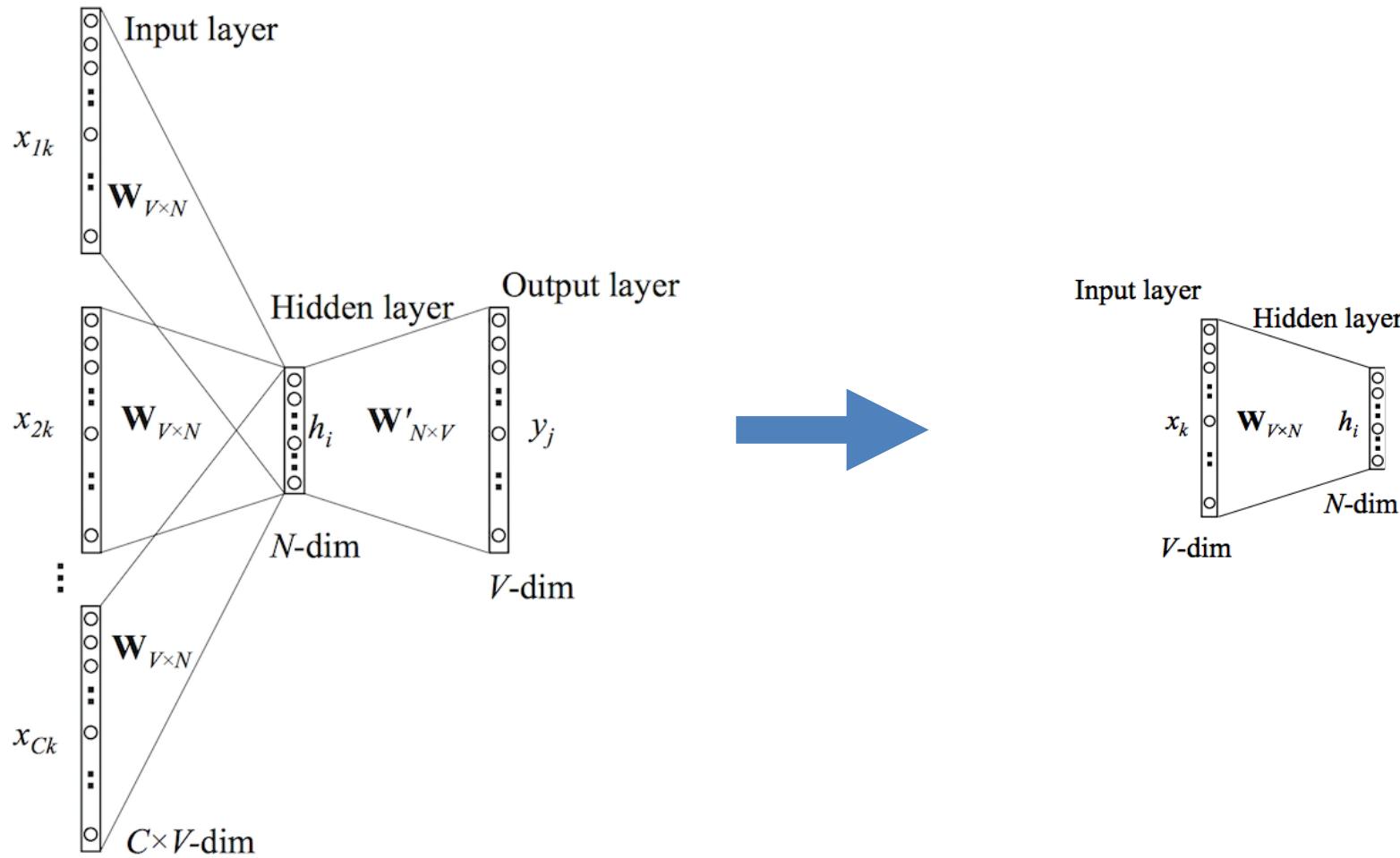


It is the same idea as the shared connections in CNN, edge or corner extractors are actually defined by the weights

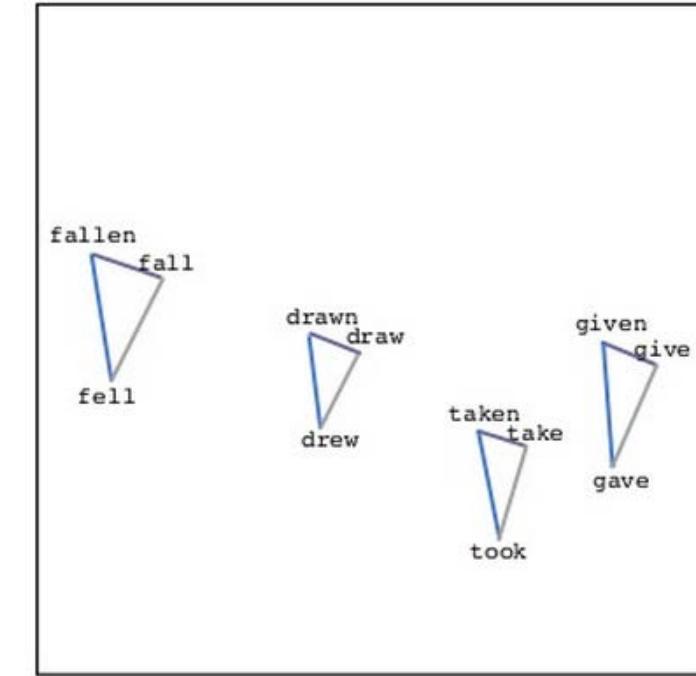
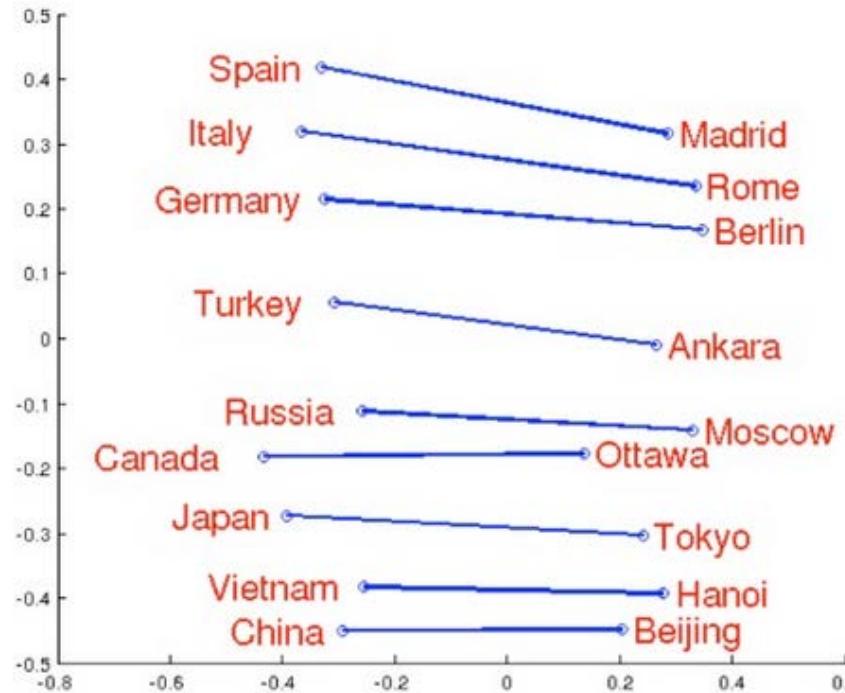
# Prediction-based – Sharing Parameters



# Word2Vec (CBOW) Overview



# Word Embedding



Source: <http://www.slideshare.net/hustwj/cikm-keynotenov2014>



# Word Embedding

- Characteristics

$$\begin{aligned}V(\text{Germany}) \\ \approx V(\text{Berlin}) - V(\text{Rome}) + V(\text{Italy})\end{aligned}$$

- Solving analogies

$$V(\text{hotter}) - V(\text{hot}) \approx V(\text{bigger}) - V(\text{big})$$

$$V(\text{Rome}) - V(\text{Italy}) \approx V(\text{Berlin}) - V(\text{Germany})$$

$$V(\text{king}) - V(\text{queen}) \approx V(\text{uncle}) - V(\text{aunt})$$

Rome : Italy = Berlin : ?

Compute  $V(\text{Berlin}) - V(\text{Rome}) + V(\text{Italy})$

Find the word w with the closest  $V(w)$

# Word Embedding

- Word2vec is known to be good at certain kinds of analogies:

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

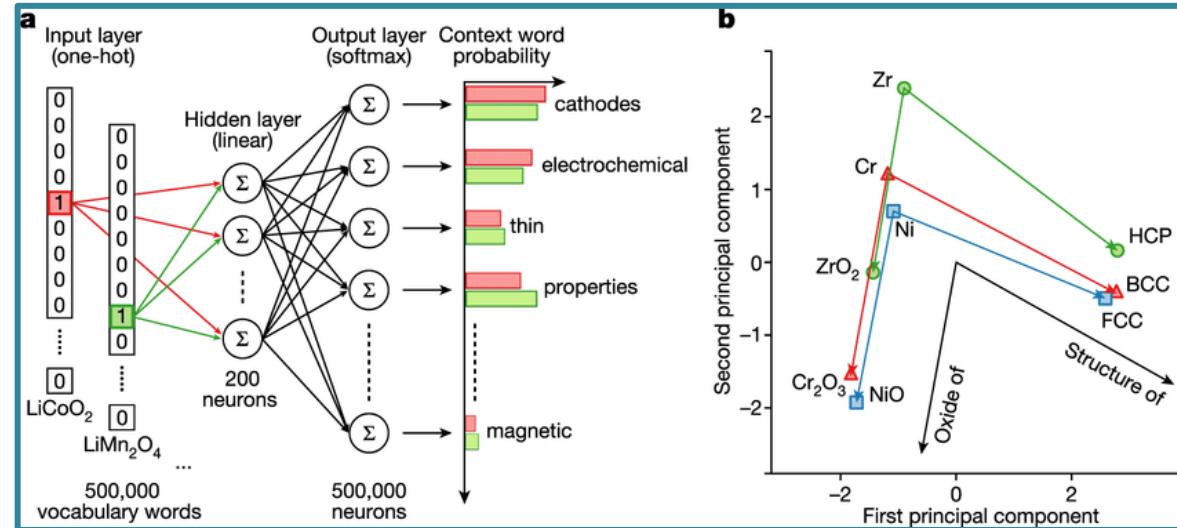
Type of relationship	Word Pair 1	Word Pair 2
Common capital city	Athens	Greece
All capital cities	Astana	Kazakhstan
Currency	Angola	kwanza
City-in-state	Chicago	Illinois
Man-Woman	brother	sister
Adjective to adverb	apparent	apparently
Opposite	possibly	impossibly
Comparative	great	greater
Superlative	easy	easiest
Present Participle	think	thinking
Nationality adjective	Switzerland	Swiss
Past tense	walking	walked
Plural nouns	mouse	mice
Plural verbs	work	works

# Word2Vec not for language (Nature 2019)

a) Target words ‘LiCoO<sub>2</sub>’ and ‘LiMn<sub>2</sub>O<sub>4</sub>’ are represented as vectors with ones at their corresponding vocabulary indices (for example, 5 and 8 in the schematic) and zeros everywhere else (one-hot encoding).

These one-hot encoded vectors are used as inputs for a neural network with a single linear hidden layer (for example, 200 neurons), which is trained to predict all words mentioned within a certain distance (context words) from the given target word.

- For similar battery cathode materials such as LiCoO<sub>2</sub> and LiMn<sub>2</sub>O<sub>4</sub>, the context words that occur in the text are mostly the same (for example, ‘cathodes’, ‘electrochemical’, and so on), which leads to similar hidden layer weights after the training is complete.
- These hidden layer weights are the actual word embeddings.
- The softmax function is used at the output layer to normalize the probabilities.
- b) Word embeddings for Zr, Cr and Ni, their principal oxides and crystal symmetries (at standard conditions) projected onto two dimensions using principal component analysis and represented as points in space. The relative positioning of the words encodes materials science relationships, such that there exist consistent vector operations between words that represent concepts such as ‘oxide of’ and ‘structure of’.





# Word2Vec not for language (Nature 2019)

Relationship	Example vector operation	Answer	Validation pairs	Accuracy (%)
Chemical element names	helium - He + Fe	= iron	8372	71.4
Crystal symmetries	cubic - GaAs + CdSe	= hexagonal	2034	35.4
Crystal structure names	zincblende - GaP + GaN	= wurtzite	556	18.7
Elemental crystal structures	dhcp - La + Cr	= bcc	1198	48.6
Principal oxides	$\text{Al}_2\text{O}_3$ - Al + Si	= $\text{SiO}_2$	650	48.8
Units	pressure - Pa + Hz	= frequency	452	35.4
Magnetic properties	ferromagnetic - NiCo + IrMn	= antiferromagnetic	622	41.0
Applications	thermoelectric - PbTe + LiFePO4	= cathode materials	-	-
Grammar	structures - structure + energy	= energies	15162	61.6
Total			29046	60.1

Materials science analogies

# Demos Word2Vec embeddings

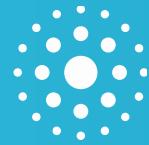
- <http://nlp.polytechnique.fr/word2vec>
- Try:
  - France – paris + berlin
  - Paris – France + allemagne
  - Paris – France + brésil

→ What do you notice?



# Overview

- Context
- Text representation
- **Bias in textual information**

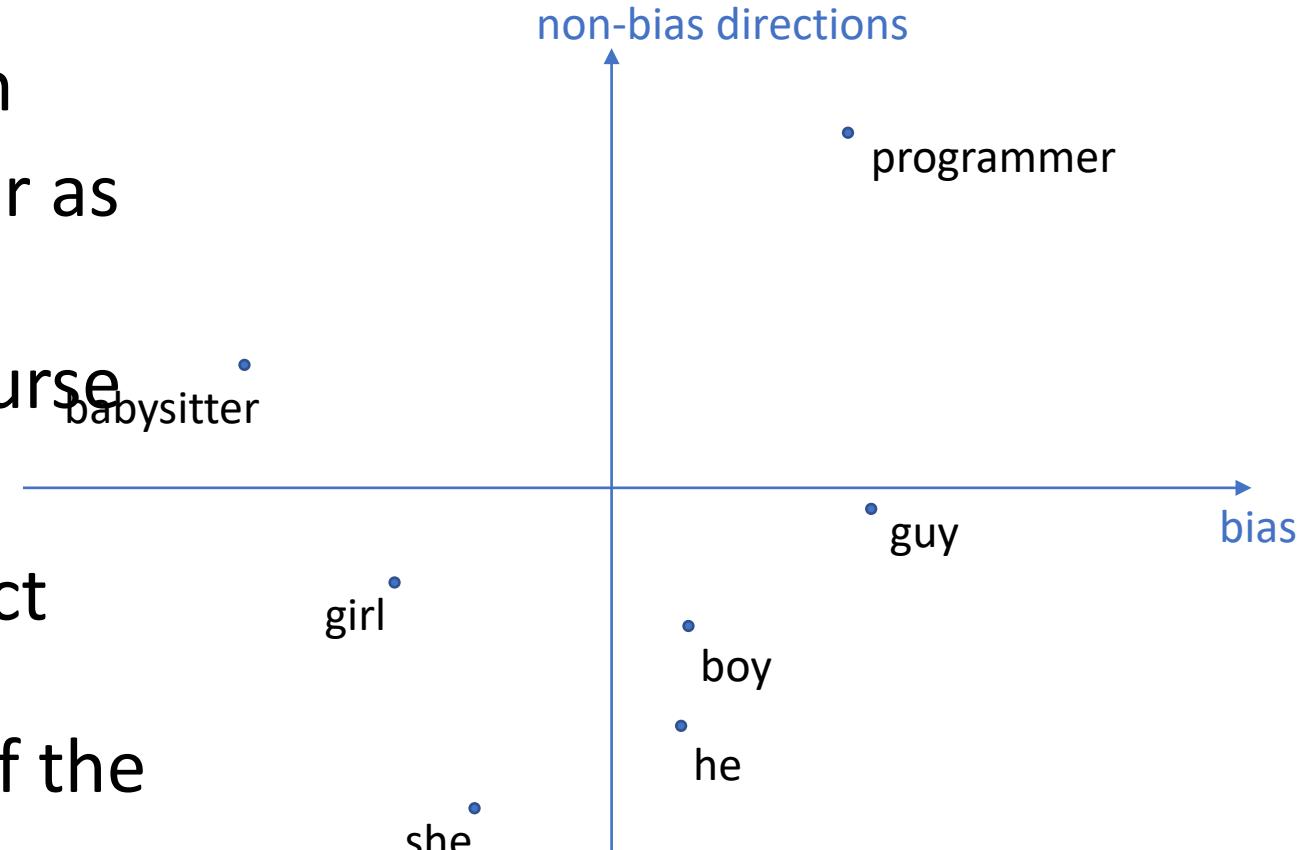


# BIAS IN TEXTUAL INFORMATION

# The problem of bias in word embeddings

- Man: Woman as King: Queen
- Man: Computer\_Programmer as Woman: Homemaker
- Father: Doctor as Mother: Nurse

→ Word embeddings can reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model.





# Definition of bias: human, in the data, learned by the AI system

- A prior, arbitrary, non-technical choice: representing the meaning of a word by its context
- Consequence: the numerical representation obtained (by the AI model) for each word reproduces associations of co-occurrences between concepts, including the problematics of...
- Bias (statistical): characteristic with different frequencies of occurrence between groups, in absolute terms, or with respect to other characteristics
- Bias (ethics for IA): statistical deviations from a model for groups of people based on protected attributes such as gender, race, age, etc.

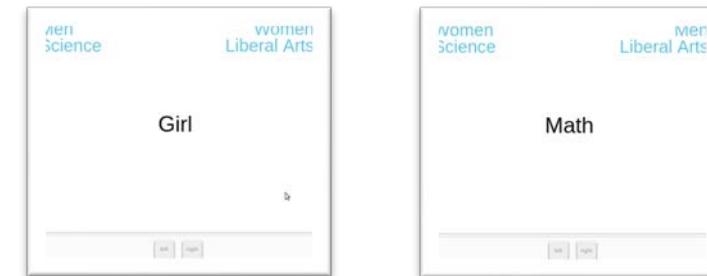
# Human biases: measured by IAT scores

univ-cotedazur.fr

- Formulated in Cognitive and Social Psychology
- In our brains, the network of our semantic memory works through associations between concepts ⇒ Bias when these associations are problematic:  
*We think more of the savannah when we talk about lions, but also generally more quickly of men when we talk about science*
- And we can measure the strength of these automatisms: **Implicit Association Test (IAT)**

$$\bullet \text{ IAT Score} = \frac{\text{resp.time.for assoc.incomp.with.stereo} - \text{resp.time.for assoc.comp.with.stereo}}{\text{standard deviation over resp.times}}$$

- General Pop.: Fast Individuals on Stereotype-Compatible Trials
  - Statistically Significant Positive Score: Stereotype Is Well Established in the biological networks of the Semantic Memory
- **Testing Speed evaluates the Strength of Connections Between Concepts**
- Highly reliable IAT for testing implicit stereotyping



# Human biases: measured by IAT scores



- Demo on IAT scores
  - Experiments | TELLab (<https://lab.tellab.org/show/paradigm/iat>)
  - The Implicit Association Test (IAT) | Outsmarting Implicit Bias: A Project at Harvard University (<https://outsmartingimplicitbias.org/module/iat/>)

# Any Question?