Topics in Representation Learning

NLPL Winter School 2020

Yoav Goldberg

Let's be more specific

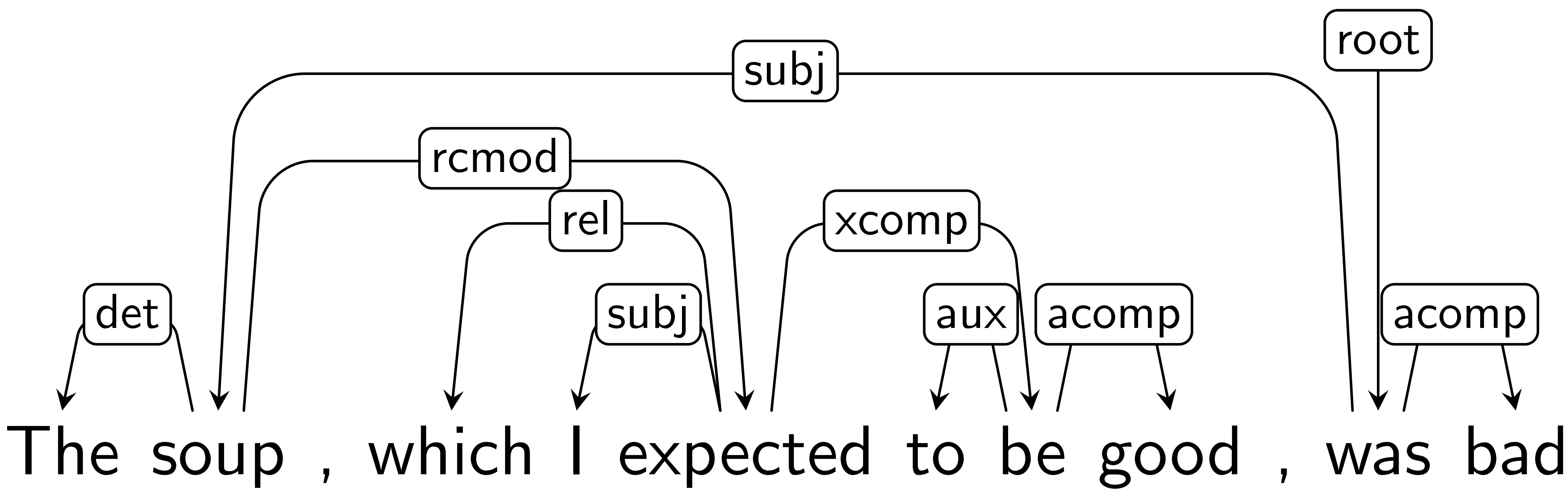
- Neural networks learn representations.
- Sharing the representations (multi-task learning).
- Using the representations --- by querying them.
- Biases in representations.
- Controlling the representations.

Neural networks learn representations.

NLP some years ago

the soup, which I expected to be good, was bad

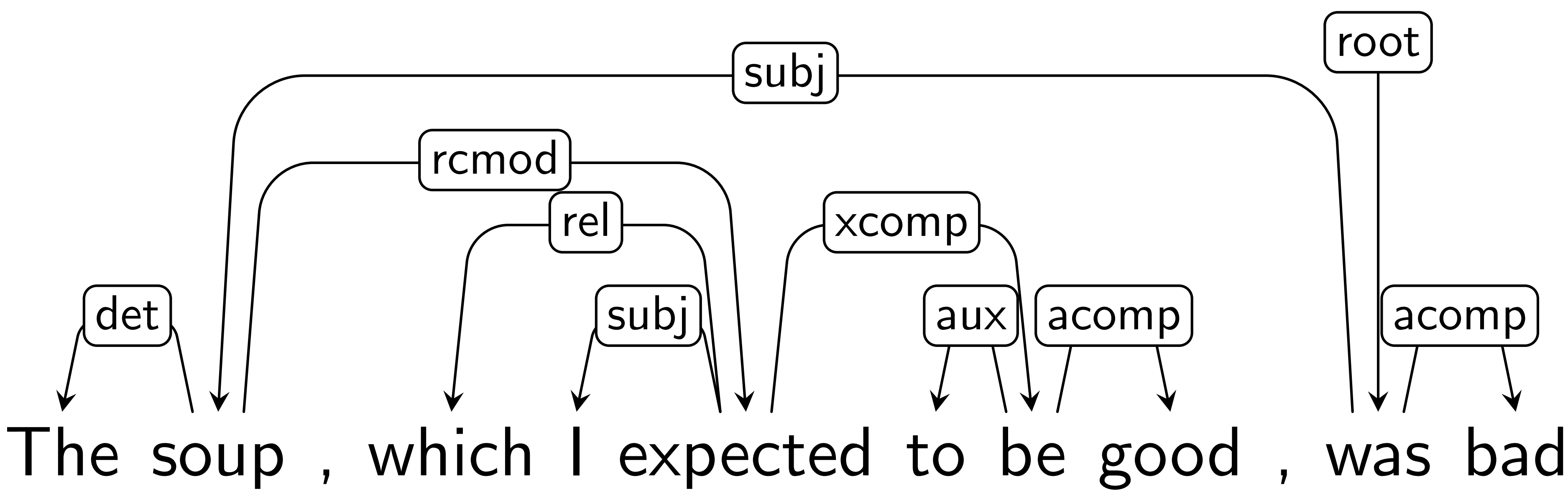
encode



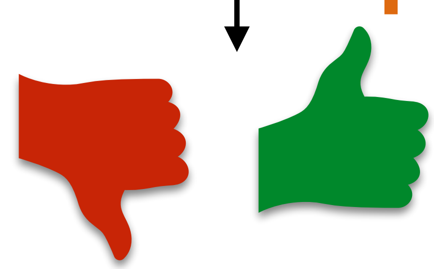
NLP some years ago

the soup, which I expected to be good, was bad

encode

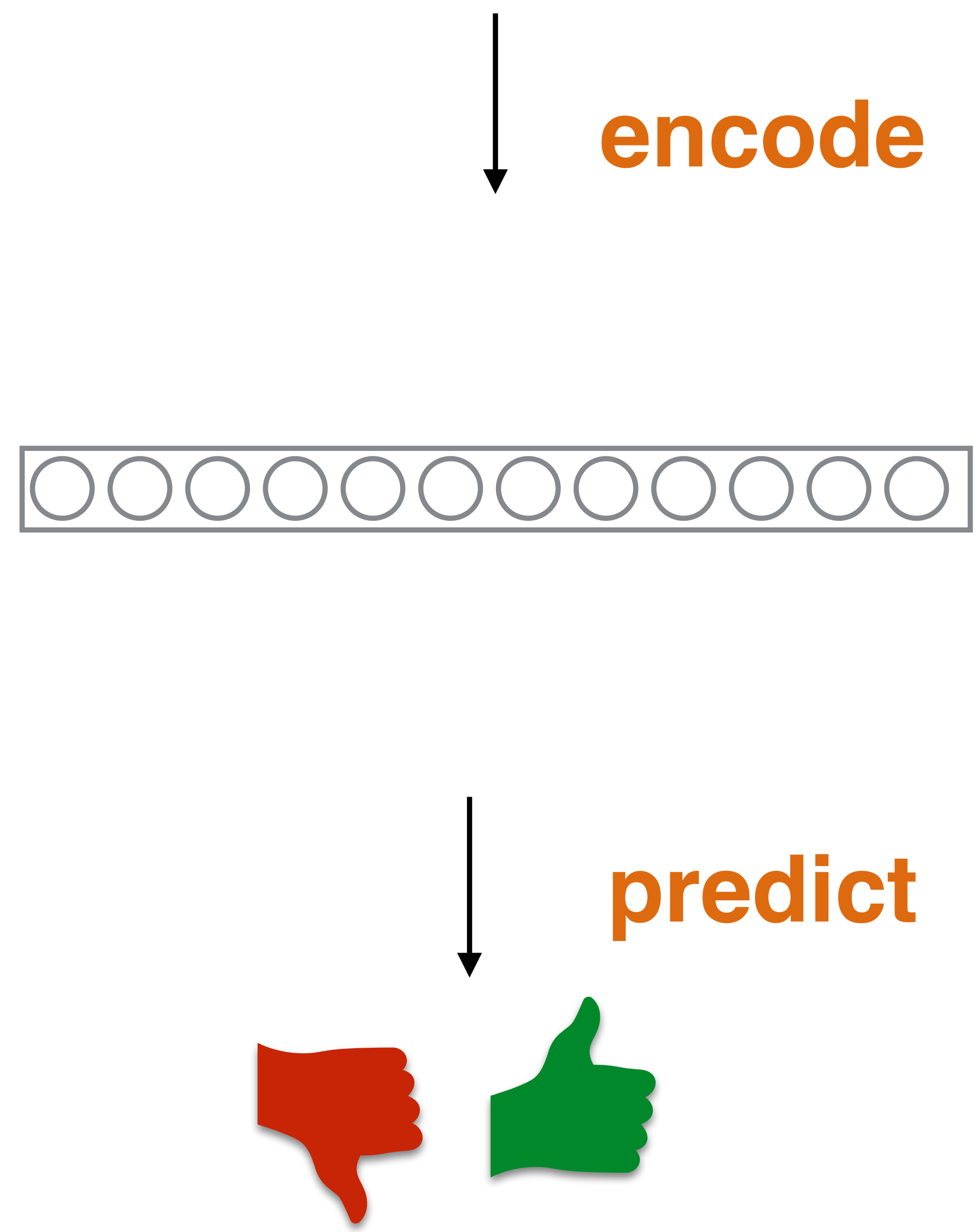


predict



NLP Today

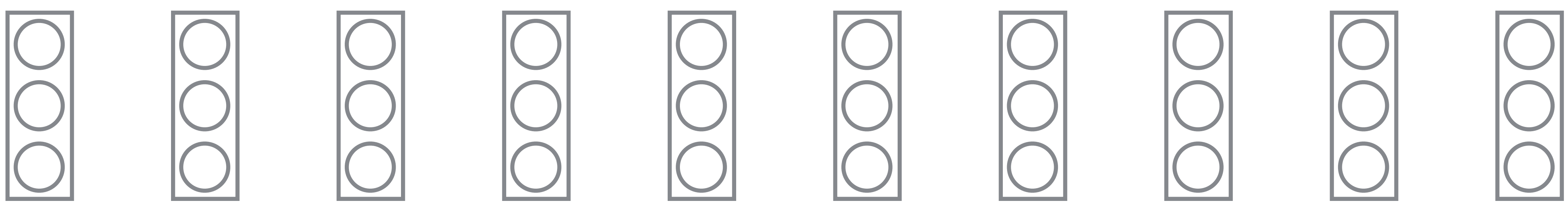
the soup, which I expected to be good, was bad



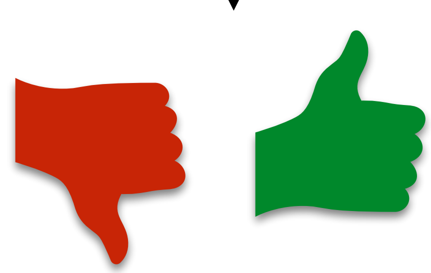
NLP Today

the soup, which I expected to be good, was bad

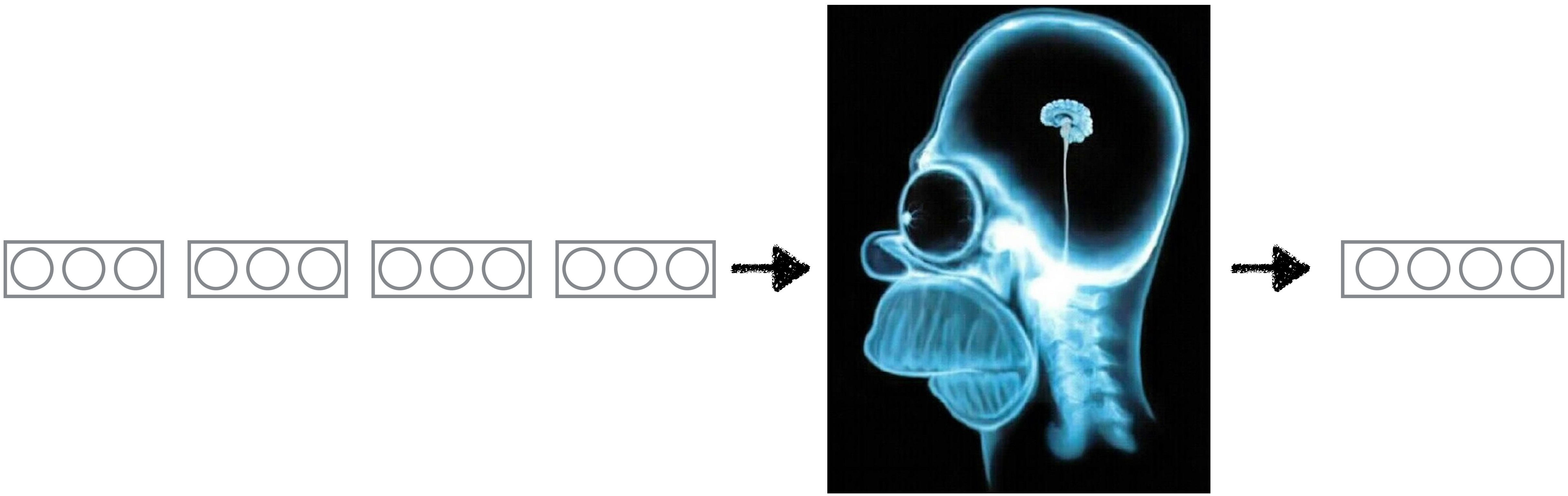
encode



predict

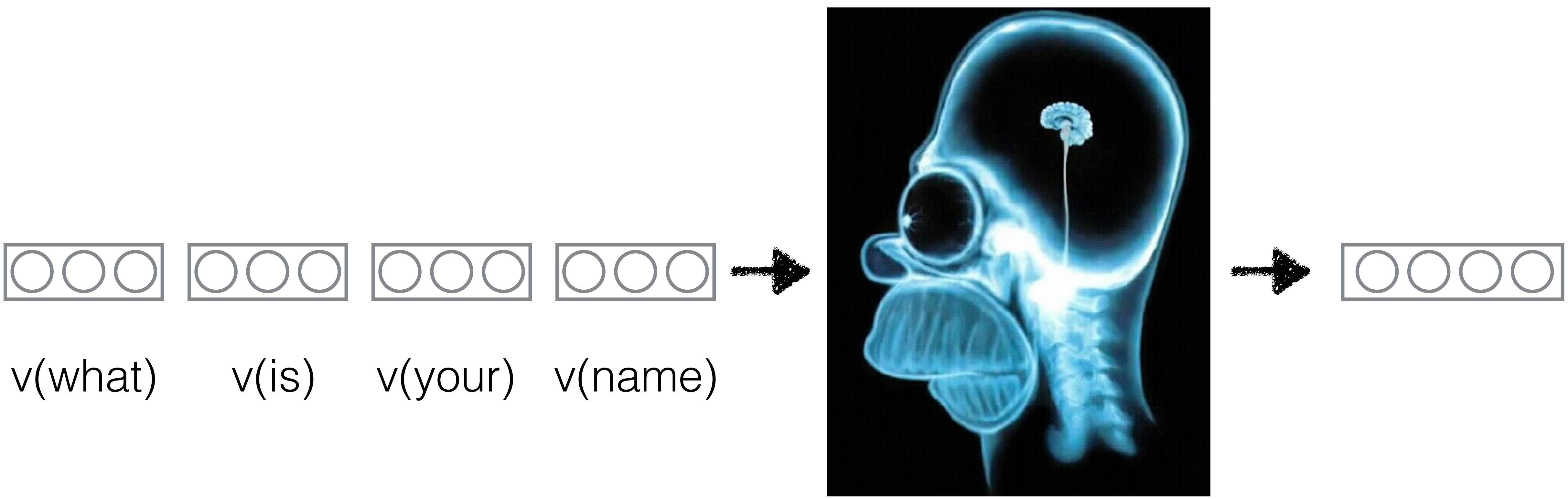


RNNs!!



encode

RNNs!!

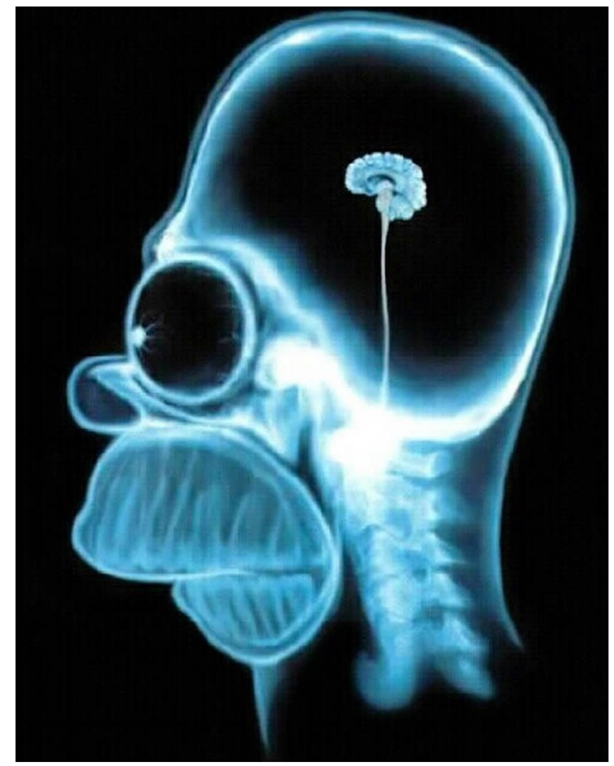


encode

RNNs!!

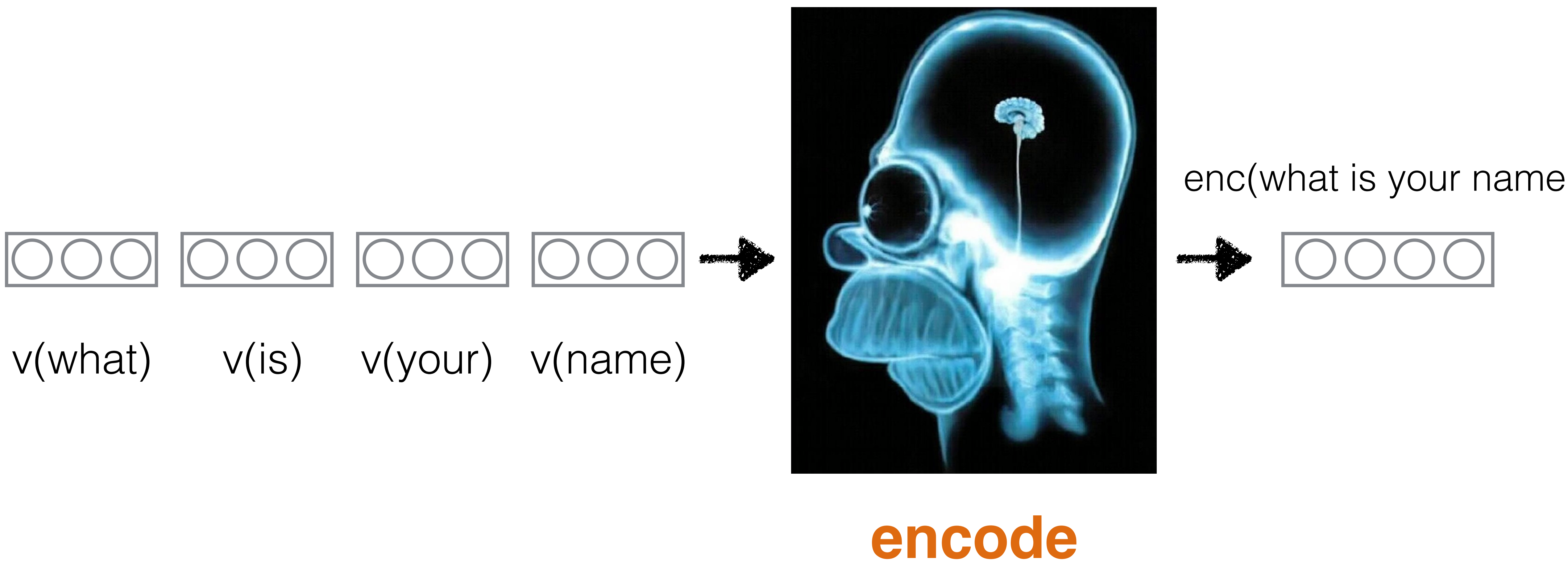
????

v(what) v(is) v(your) v(name)



encode

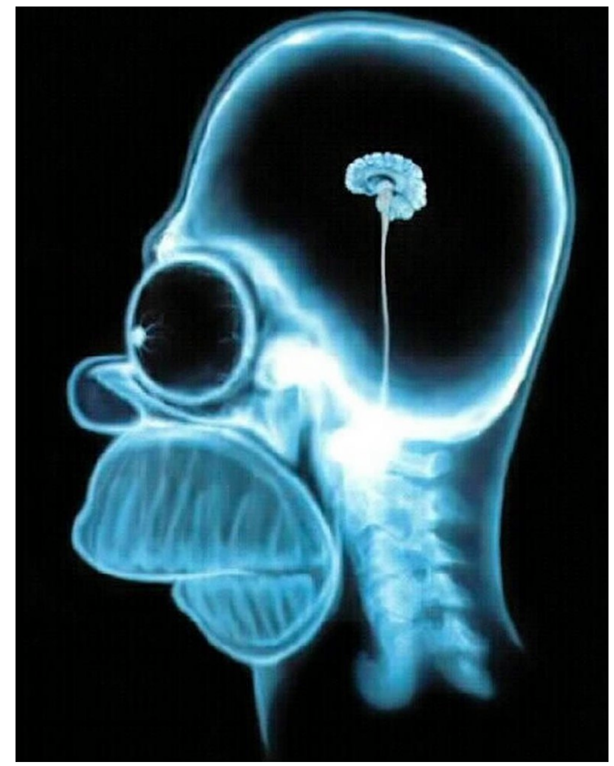
RNNs!!



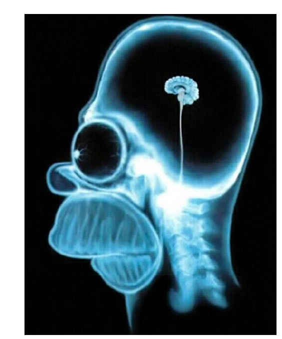
RNNs!!

enc(what is your name)

v(what) v(is) v(your) v(name)



encode

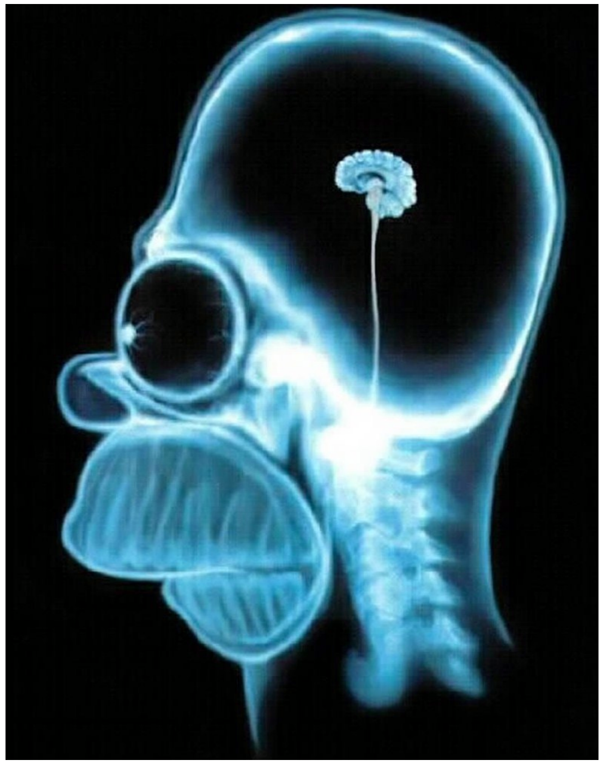


predict

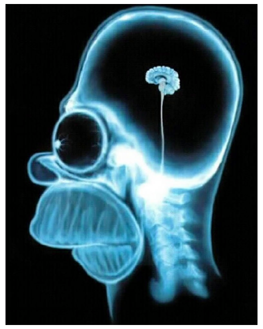
RNNs!!

enc(what is your name)

v(what) v(is) v(your) v(name)



encode

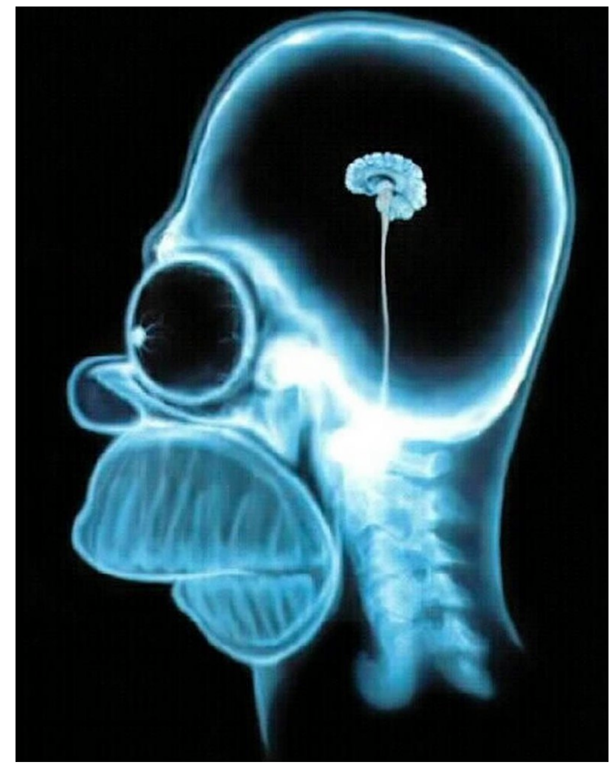


predict

~~"declarative"~~ / "question"

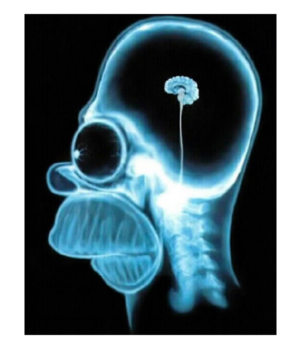
RNNs!!

v(what) v(is) v(your) v(name)



encode

enc(what is your name)



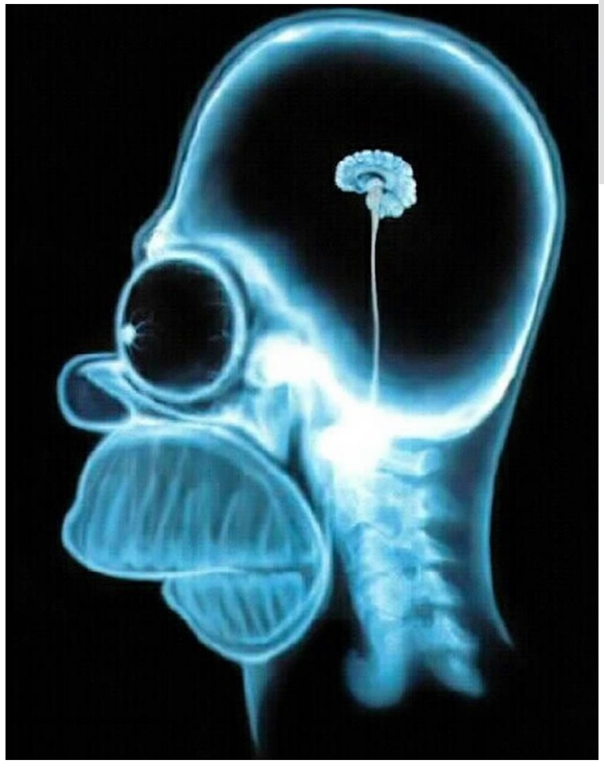
predict

RNNs!!

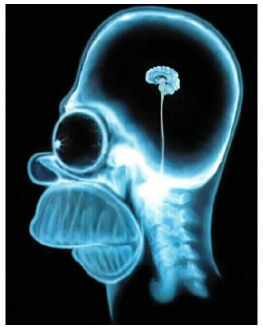
encoded vector is **informative for the task**

enc(what is your name)

v(what) v(is) v(your) v(name)



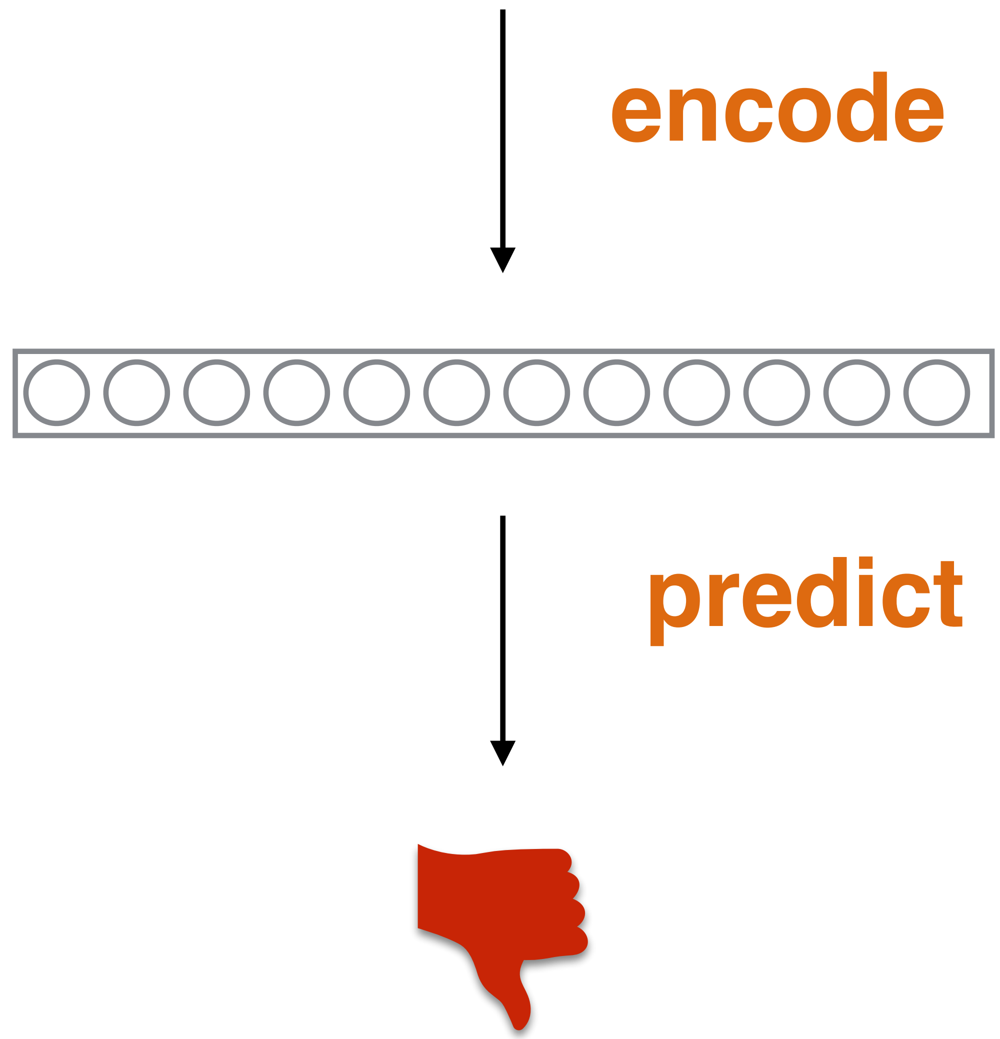
encode



predict

Representation Learning

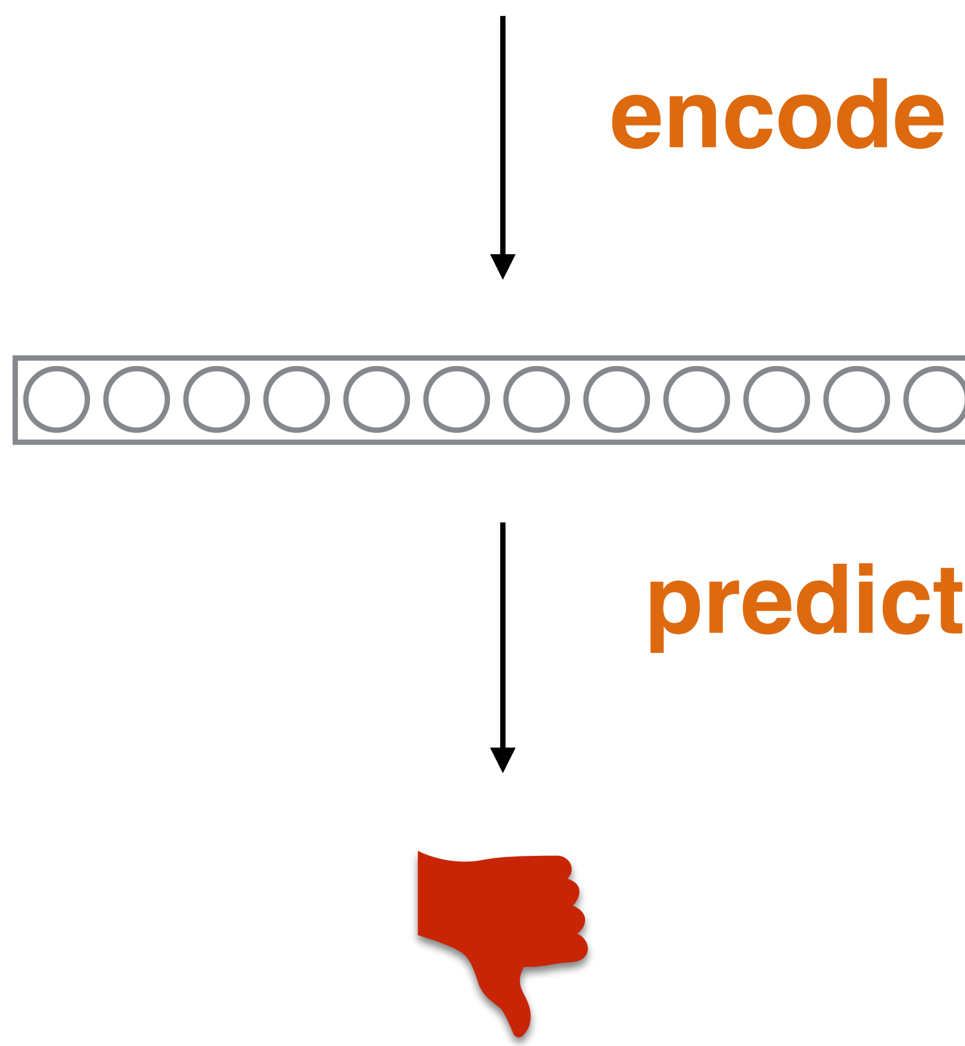
the soup, which I expected to be good, was bad



train a model to predict Y

Representation Learning

the soup, which I expected to be good, was bad



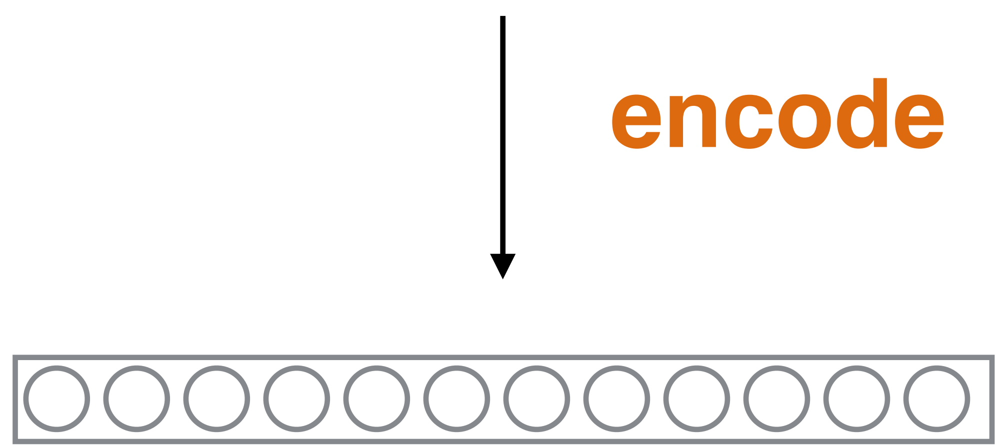
representation **h** which is predictive of Y

train a model to predict Y

Representation Learning

the soup, which I expected to be good, was bad

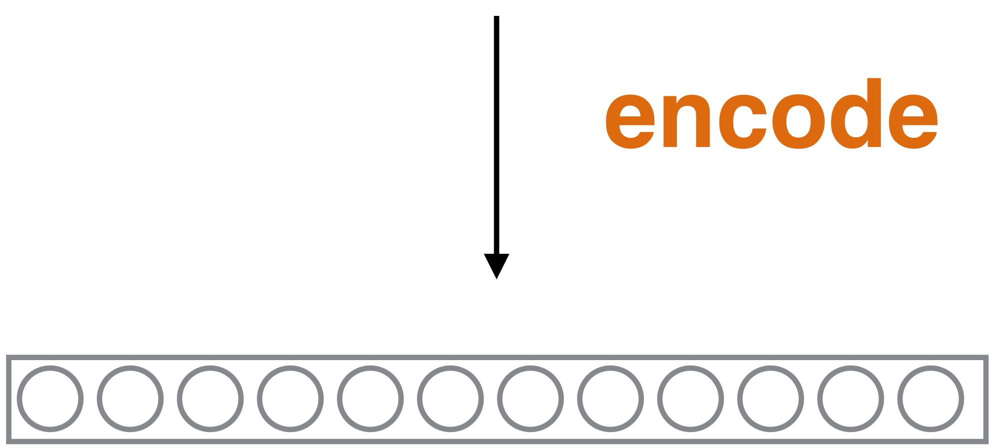
encode



representation **h** which is predictive of Y.

Representation Learning

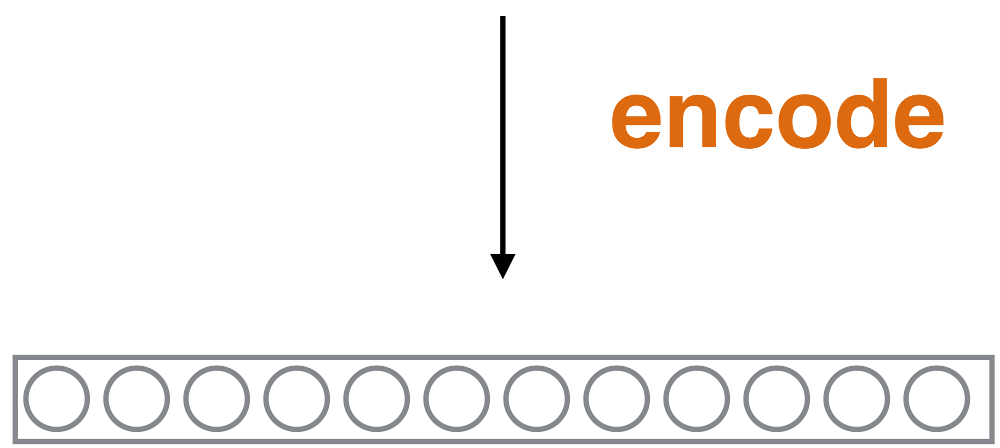
the soup, which I expected to be good, was bad



representation **h**
which is
predictive of Y.
and of Y2? and Y3?

Representation Learning

the soup, which I expected to be good, was bad



encode

representation **h**
which is
predictive of Y.
and of Y2? and Y3?

pre-training!

multi-task learning!

transfer-learning!

Shared representations

(Transfer. Multi-task.)

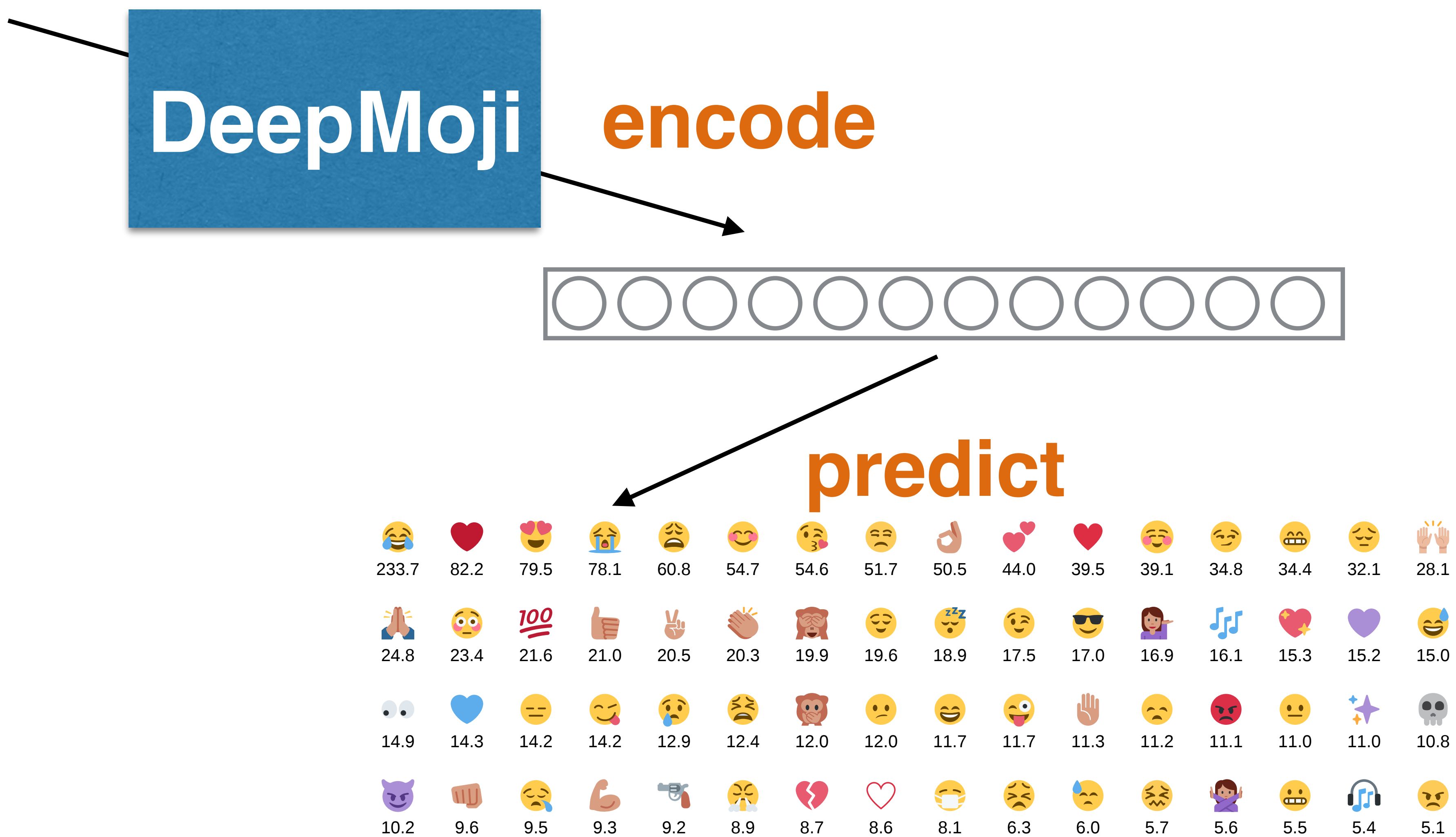
Example: DeepMoji

Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm

Bjarke Felbo[1], Alan Mislove[2], Anders Søgaard[3], Iyad Rahwan[1], Sune Lehmann[4]

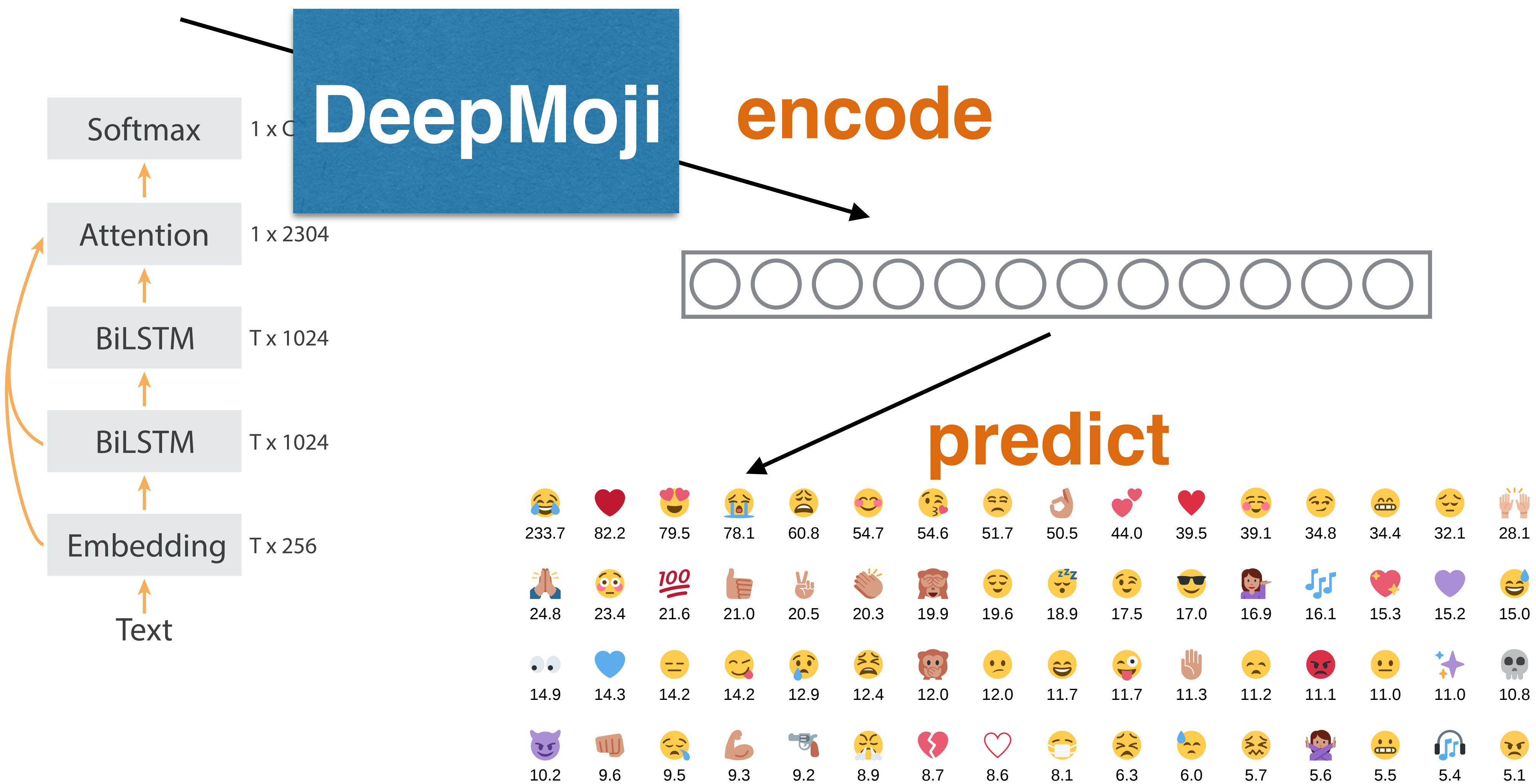
[1]Media Lab, Massachusetts Institute of Technology
[2]College of Computer and Information Science, Northeastern University
[3]Department of Computer Science, University of Copenhagen
[4]DTU Compute, Technical University of Denmark

Example: DeepMoji



Train a model to predict emojis from tweets

Example: DeepMoji



Train a model to predict emojis from tweets

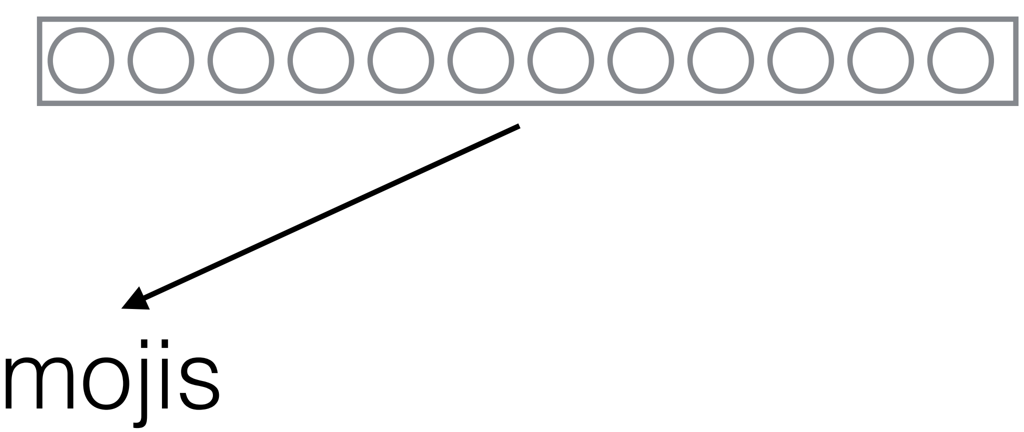
Example: DeepMoji

I love mom's cooking	😋 49.1%	😍 8.8%	❤️ 3.1%	☺️ 3.0%	♡ 2.9%
I love how you never reply back..	😒 14.0%	😑 8.3%	😠 6.3%	😐 5.4%	💔 5.1%
I love cruising with my homies	😎 34.0%	👌 6.6%	✌️ 5.7%	😌 4.1%	💯 3.8%
I love messing with yo mind!!	😜 17.2%	😈 11.8%	😏 8.0%	😉 6.4%	🙊 5.3%
I love you and now you're just gone..	💔 39.1%	😔 11.0%	😞 7.3%	😪 5.3%	😢 4.5%
This is shit	😠 7.0%	😡 6.4%	😞 6.0%	😒 6.0%	😤 5.8%
This is the shit	🎧 10.9%	🎶 9.7%	👌 6.5%	😎 5.7%	😏 4.8%

Example: DeepMoji

Train a model to predict emojis from tweets

DeepMoji

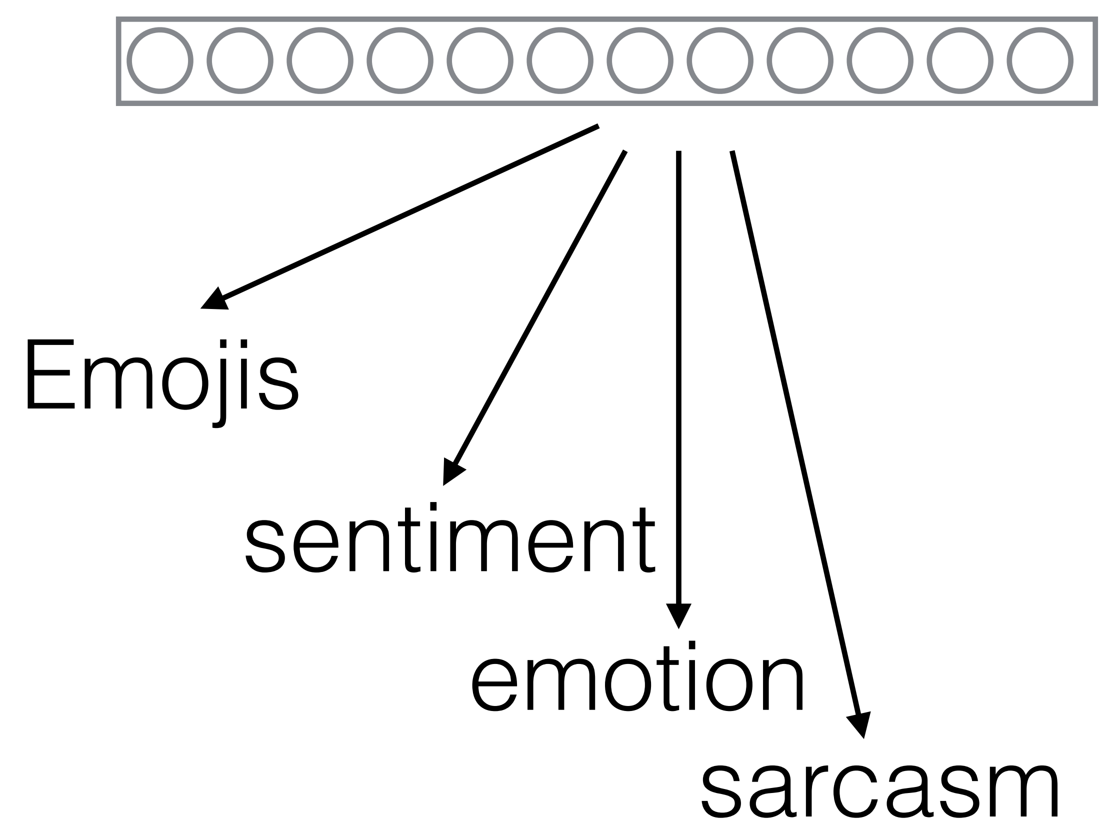


Example: DeepMoji

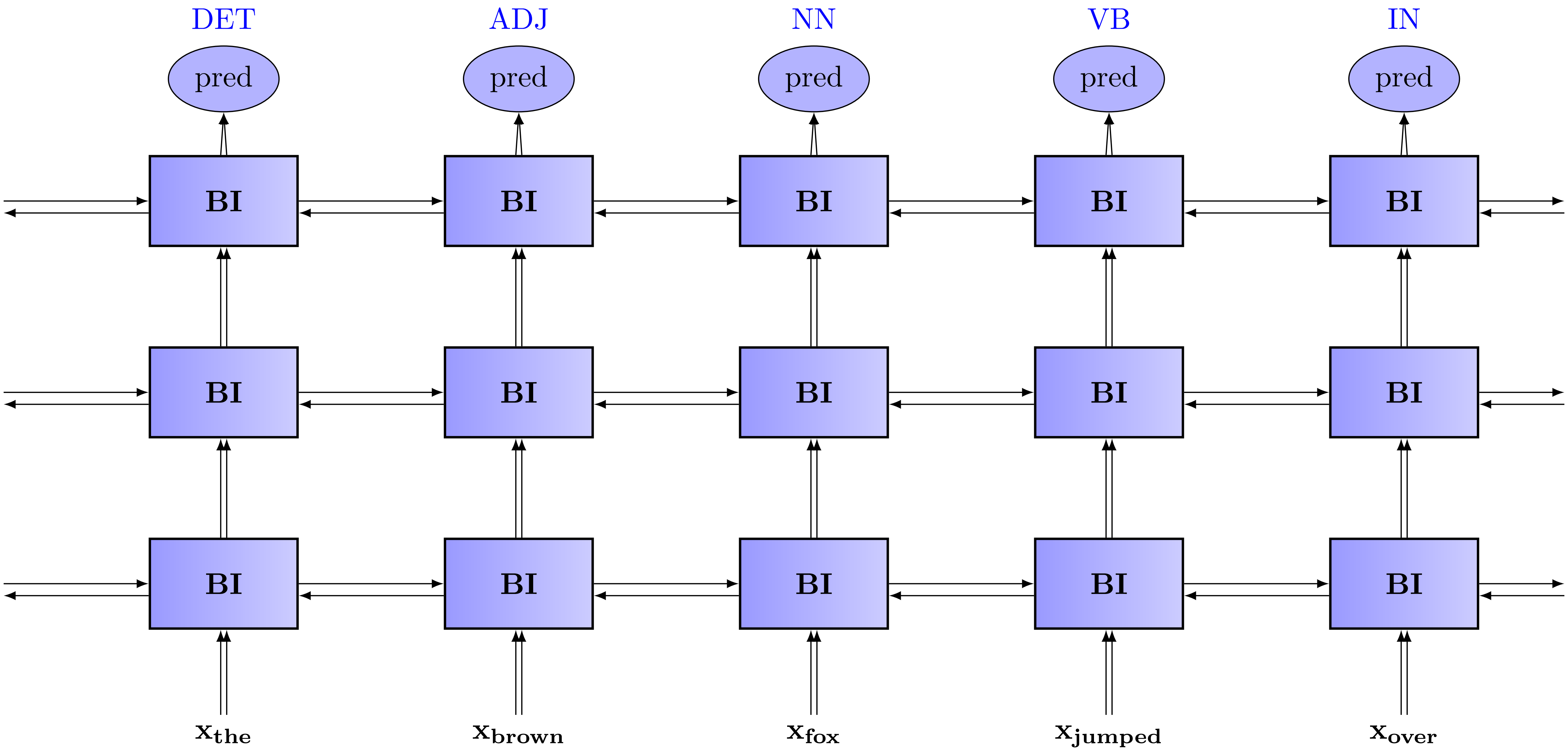
Train a model to predict emojis from tweets

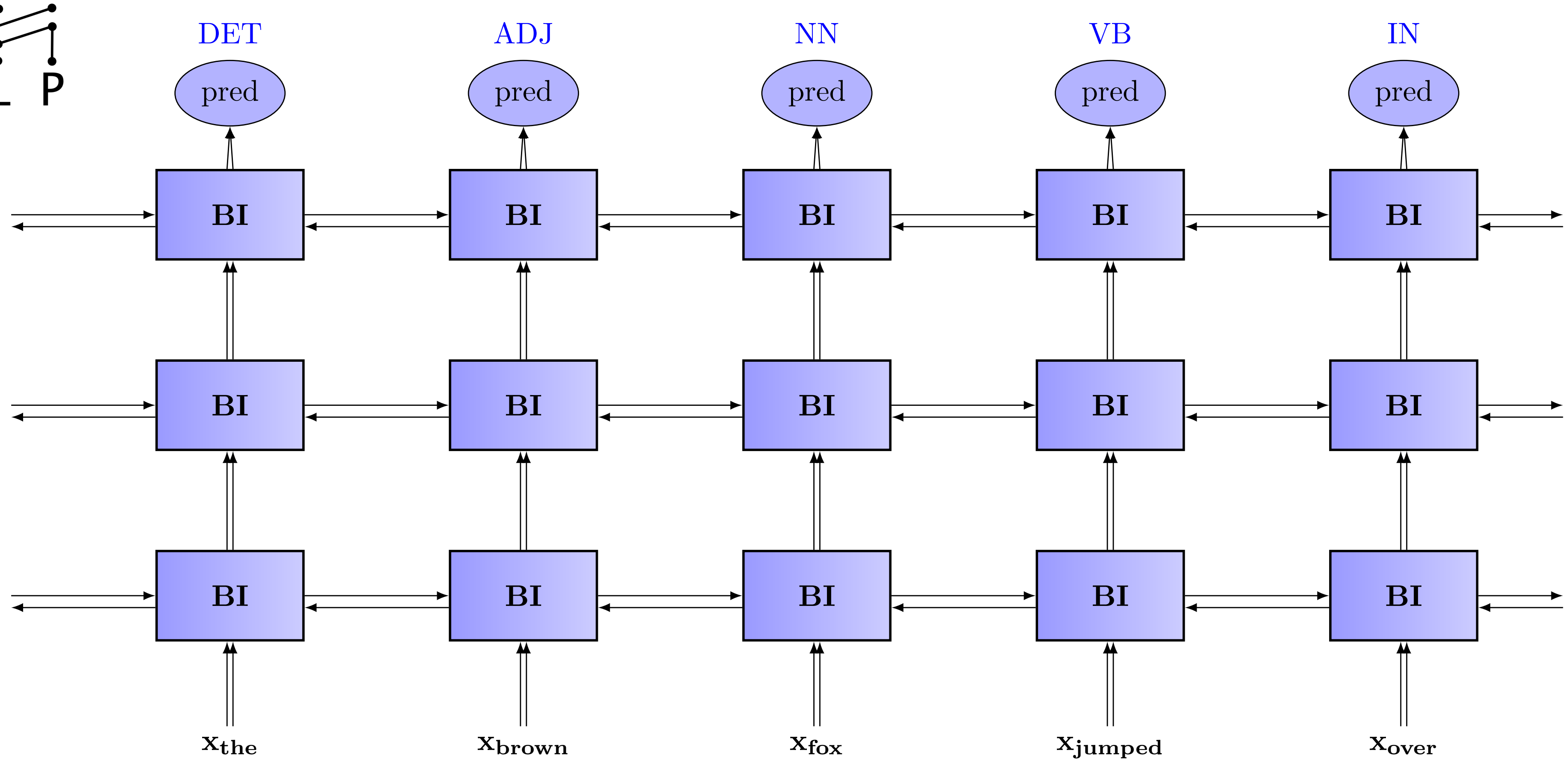
Vectors are also predictive of related tasks

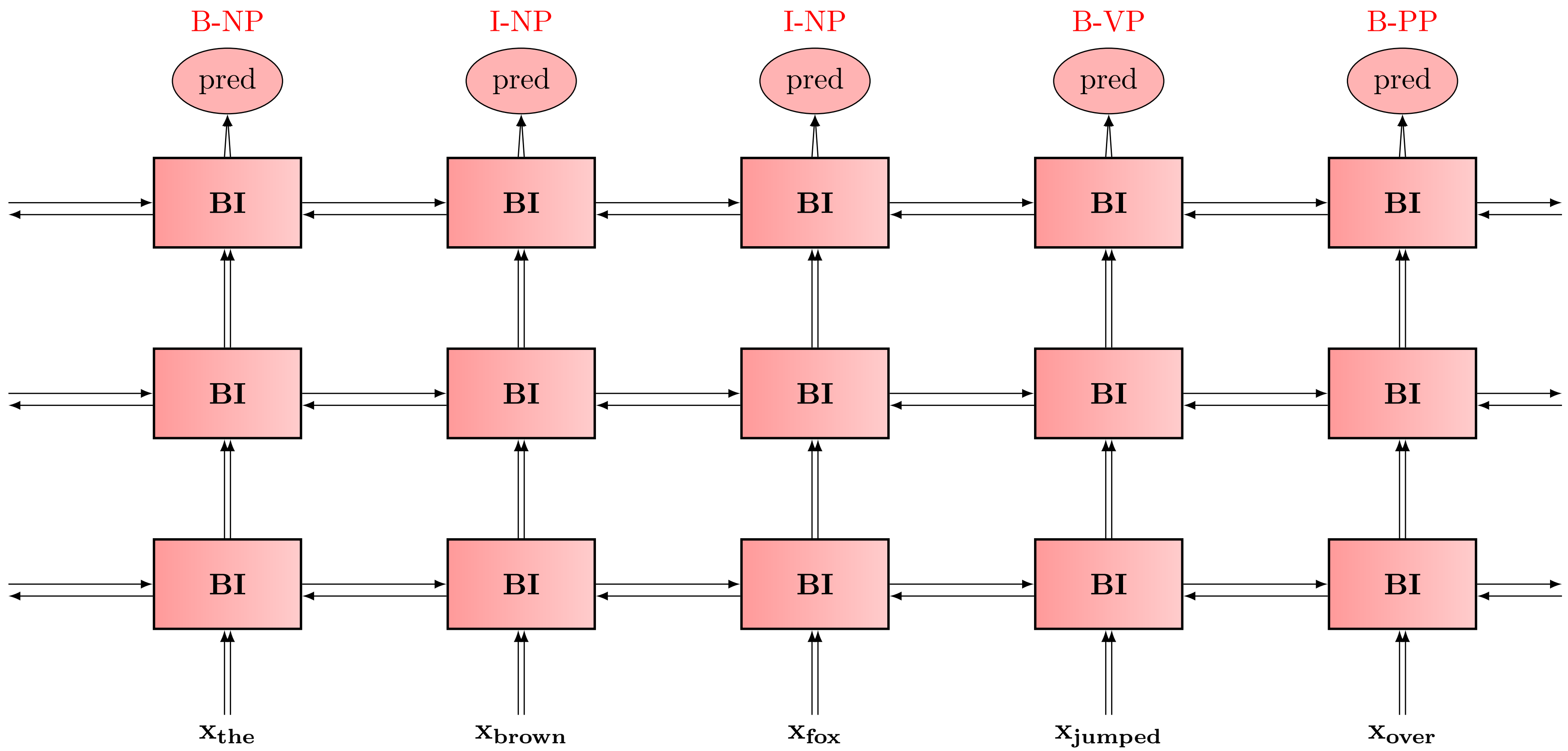
DeepMoji



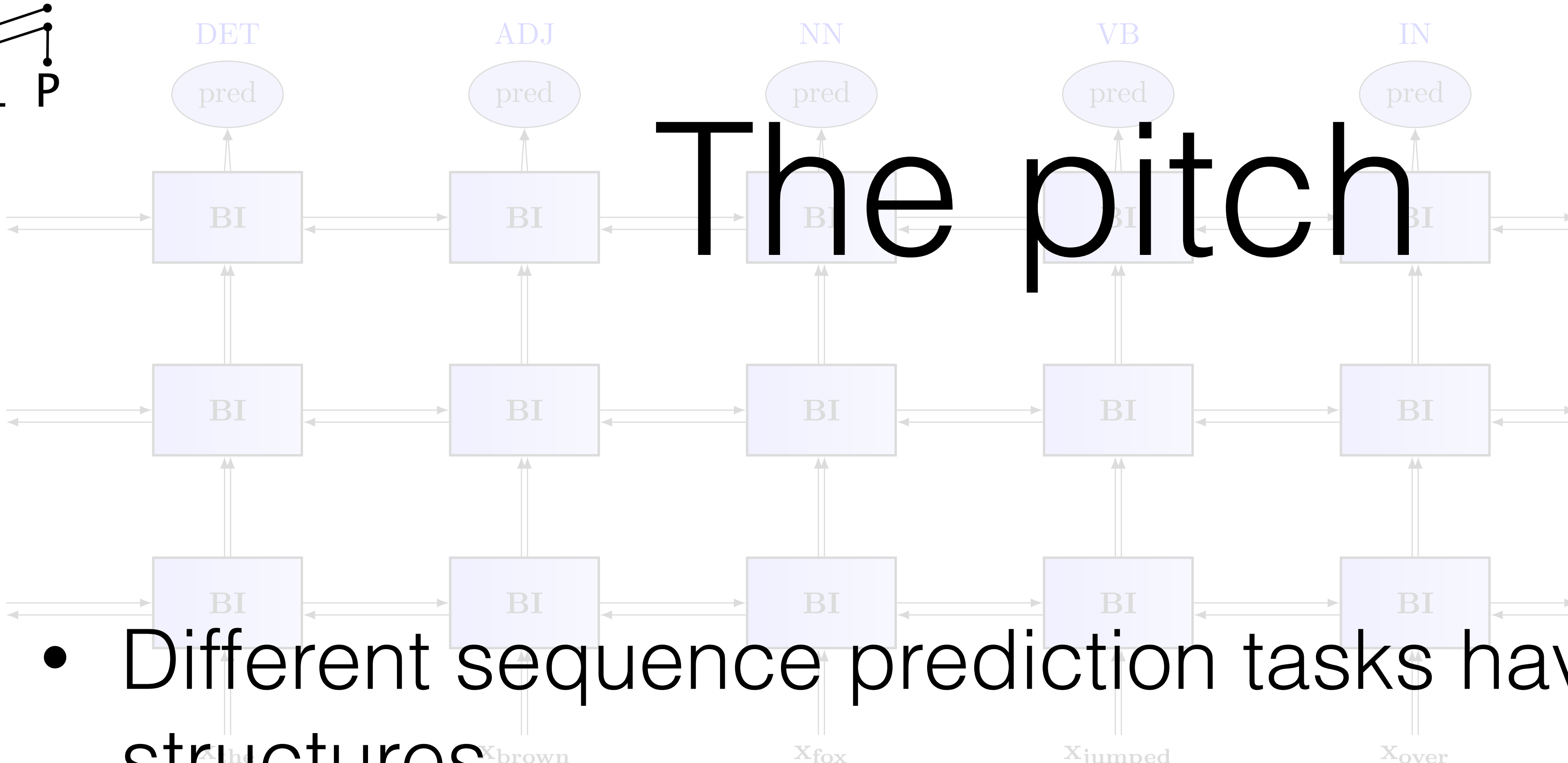
Example: Multi-tasking


DET
ADJ
NN
VB
IN
pred
pred
pred
pred
pred
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
$\mathbf{x_{the}}$
$\mathbf{x_{brown}}$
$\mathbf{x_{fox}}$
$\mathbf{x_{jumped}}$
$\mathbf{x_{over}}$

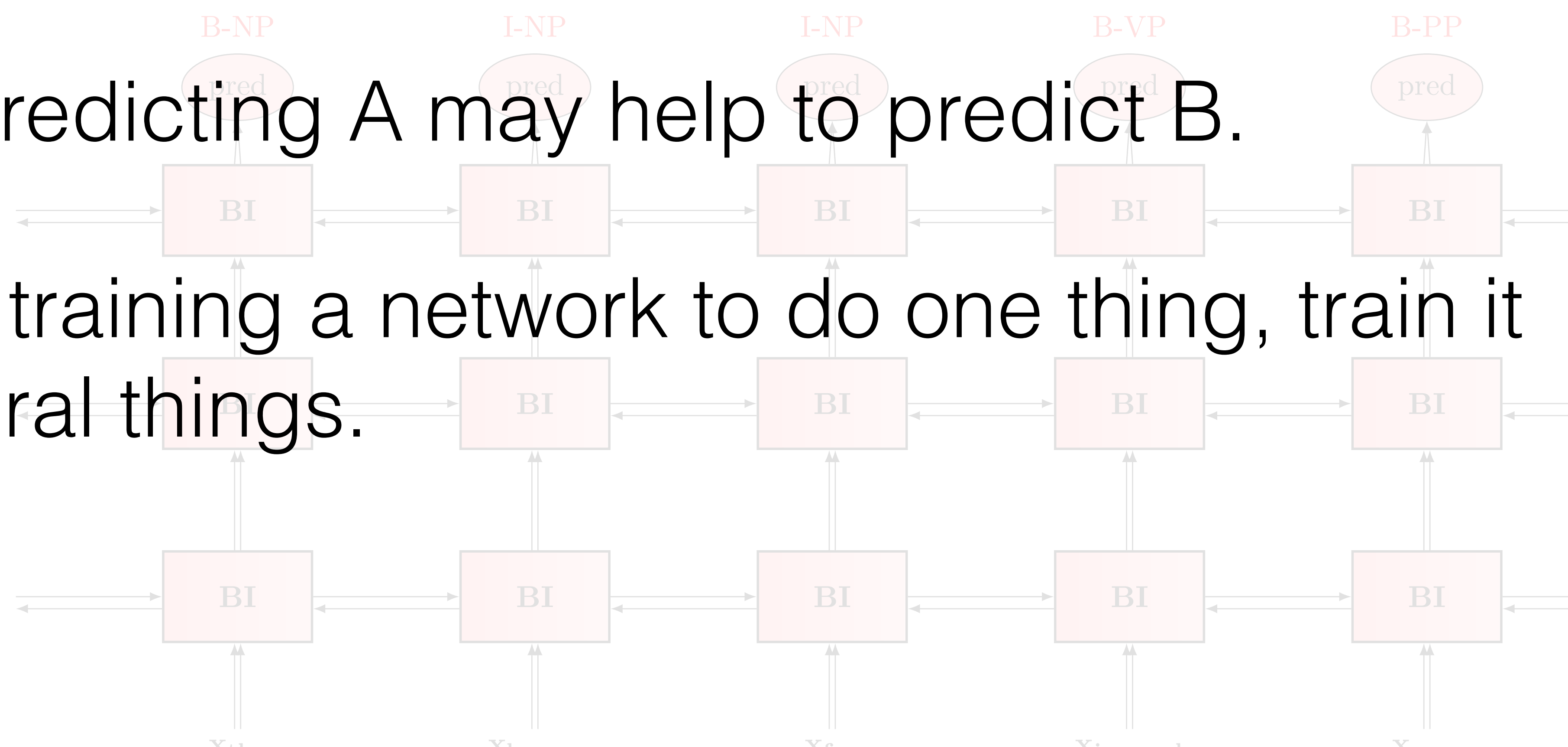

DET
ADJ
NN
VB
IN
pred
BI
x_{the}
x_{brown}
x_{fox}
x_{jumped}
x_{over}

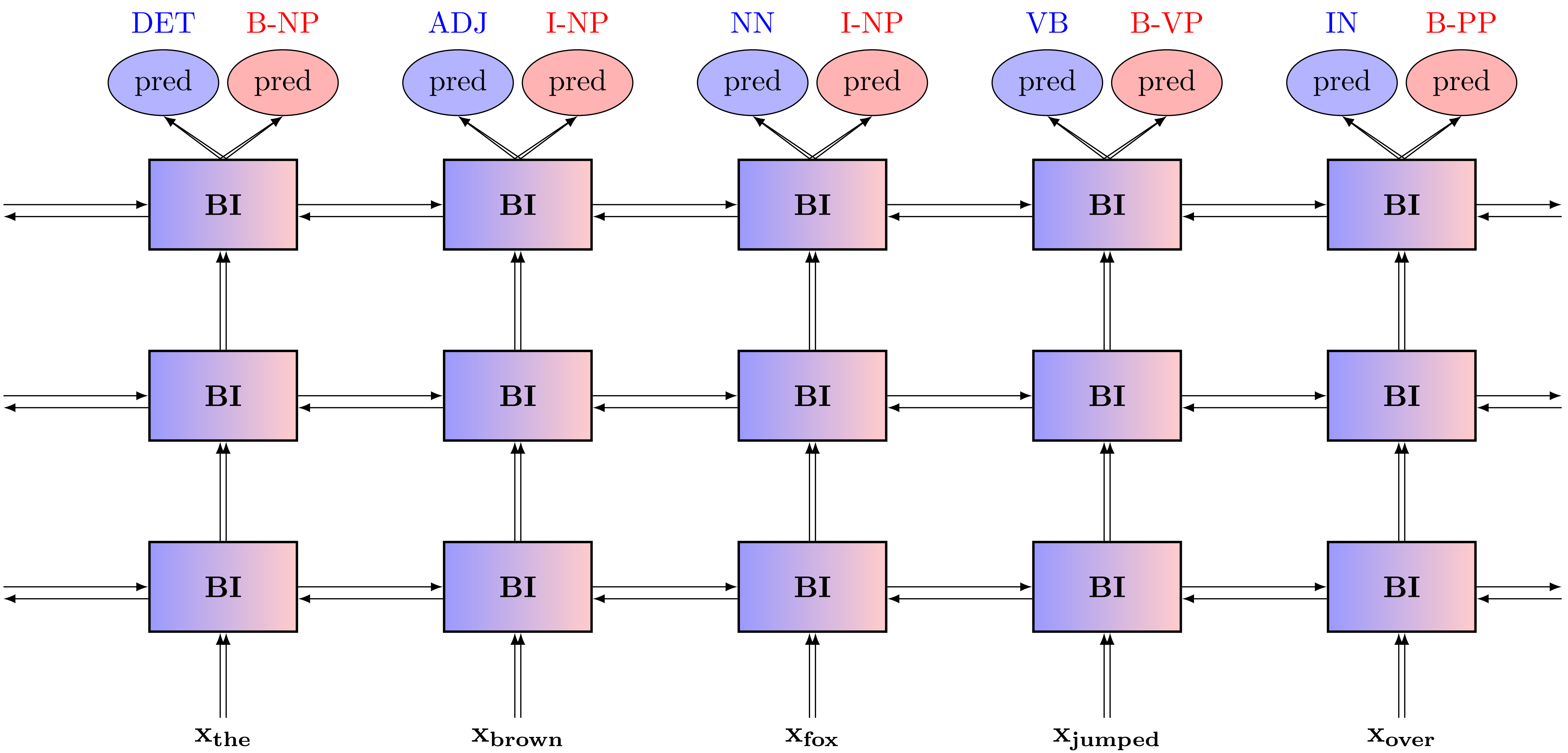

B-NP
I-NP
I-NP
B-VP
B-PP
pred
BI
x_{the}
x_{brown}
x_{fox}
x_{jumped}
x_{over}

The pitch



- Different sequence prediction tasks have shared structures.
- Hints for predicting A may help to predict B.
- Instead of training a network to do one thing, train it to do several things.




DET
B-NP
ADJ
I-NP
NN
I-NP
VB
B-VP
IN
B-PP
pred
pred
pred
pred
pred
pred
pred
pred
pred
pred
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
BI
x_{the}
x_{brown}
x_{fox}
x_{jumped}
x_{over}

Not all is pretty

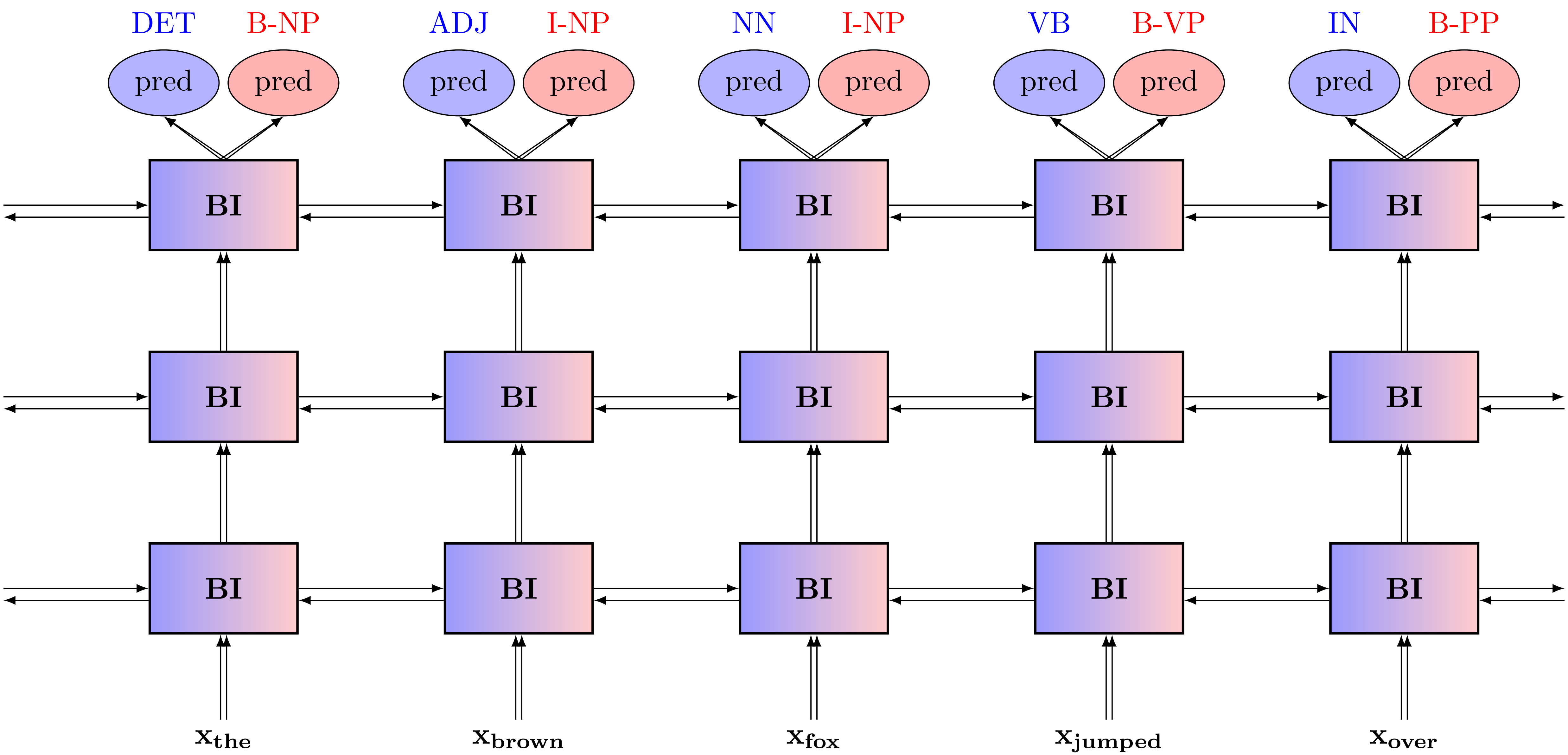
- Not so easy to get it to work.
- For many task pairs: no improvement at all.
- If network not wide enough, MTL **hurts both tasks**.
- More in-task data > more tasks.

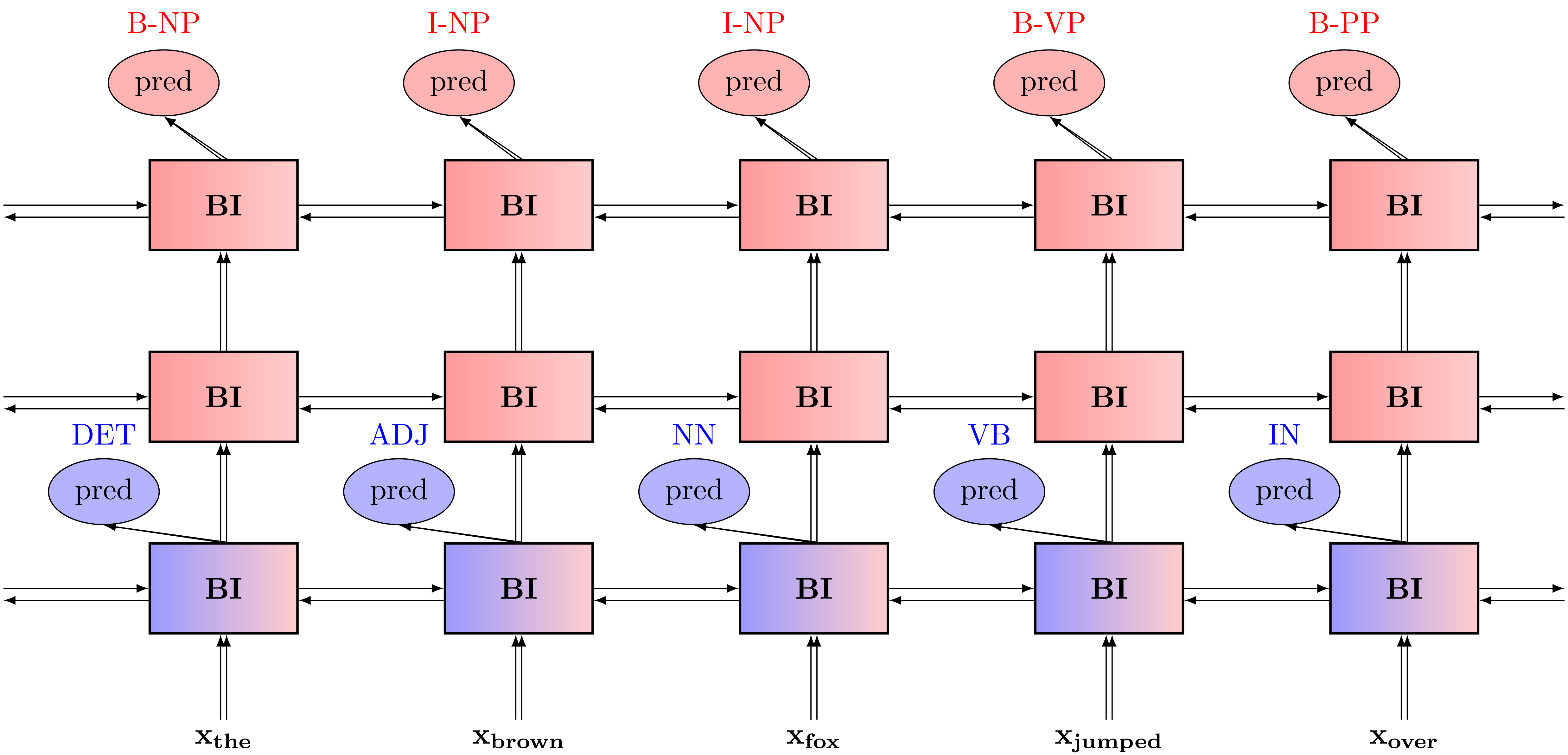
Thinking about the architecture

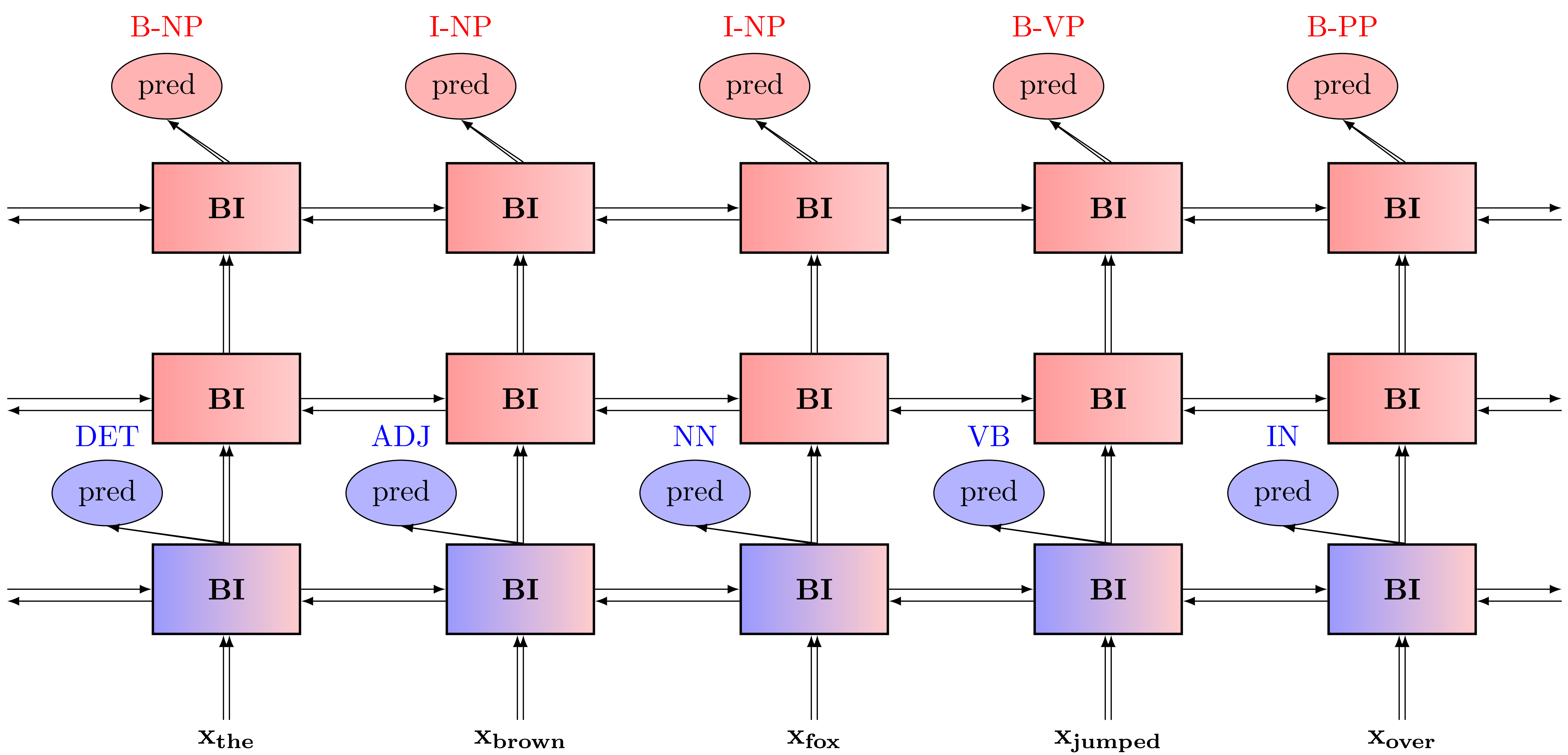
(joint work with Anders Søgaard)

We know there is a hierarchy between tasks

Why not use it?

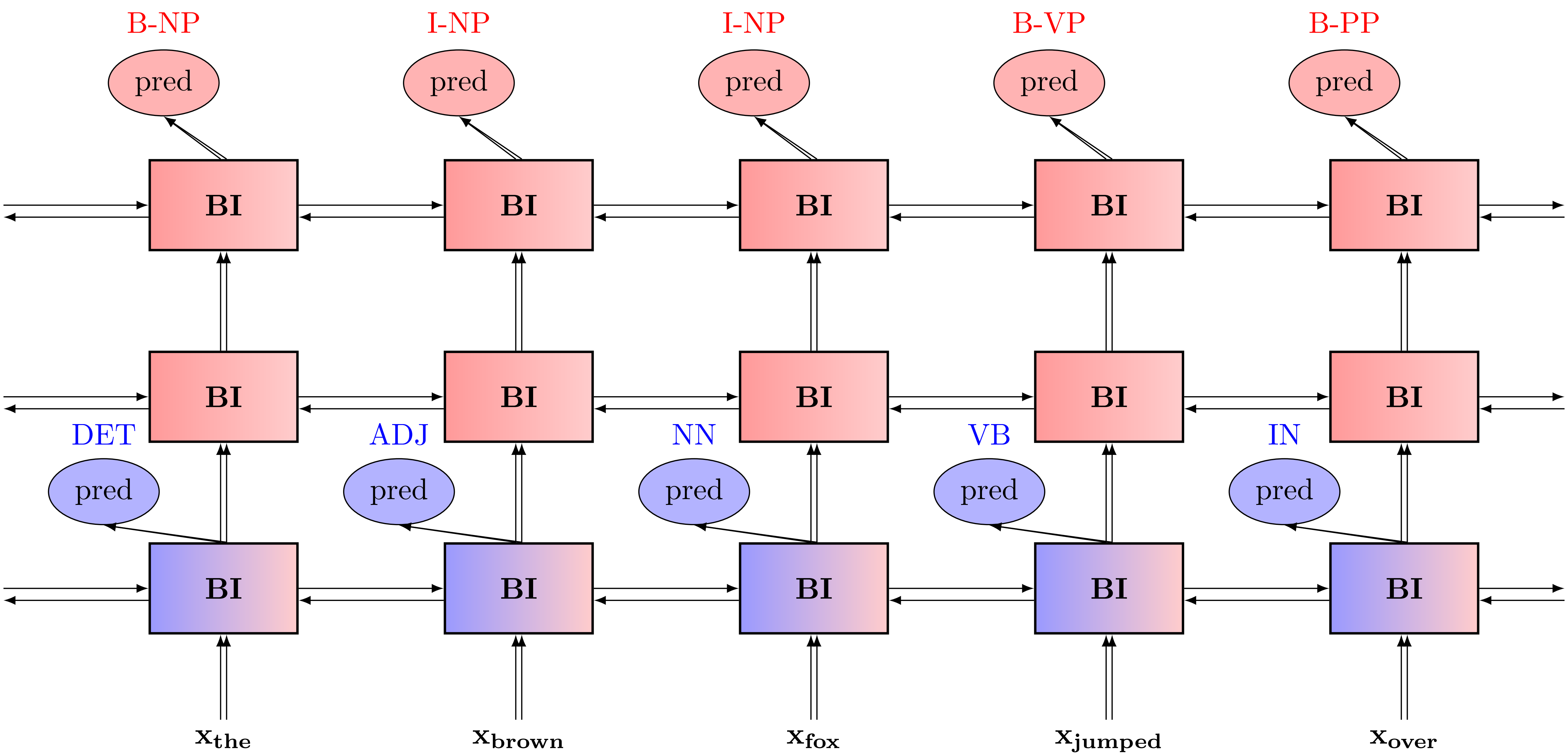

DET
B-NP
ADJ
I-NP
NN
I-NP
VB
B-VP
IN
B-PP
pred
BI
x_{the}
x_{brown}
x_{fox}
x_{jumped}
x_{over}


B-NP
I-NP
I-NP
B-VP
B-PP
pred
BI
DET
ADJ
NN
VB
IN
$\mathbf{x}_{\text{the}}$
$\mathbf{x}_{\text{brown}}$
$\mathbf{x}_{\text{fox}}$
$\mathbf{x}_{\text{jumped}}$
$\mathbf{x}_{\text{over}}$



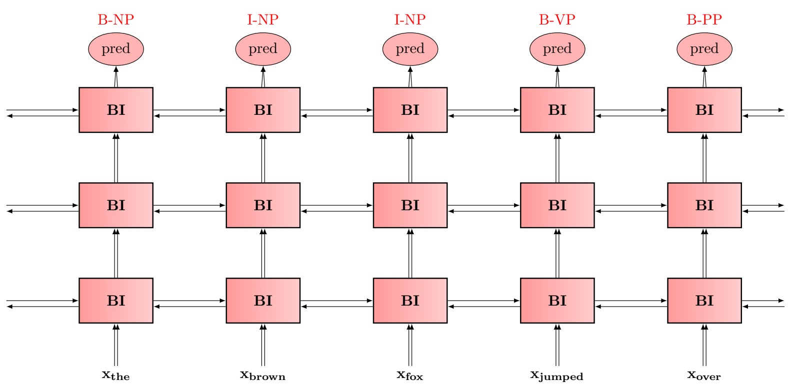
Lower layer is trained for predicting POS
(but also gets feedback from CHUNKS)

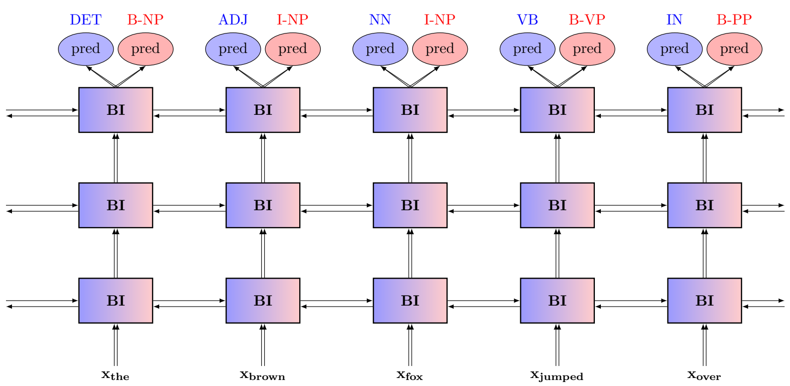
Upper layers are specialized for CHUNK (while using the POS information)

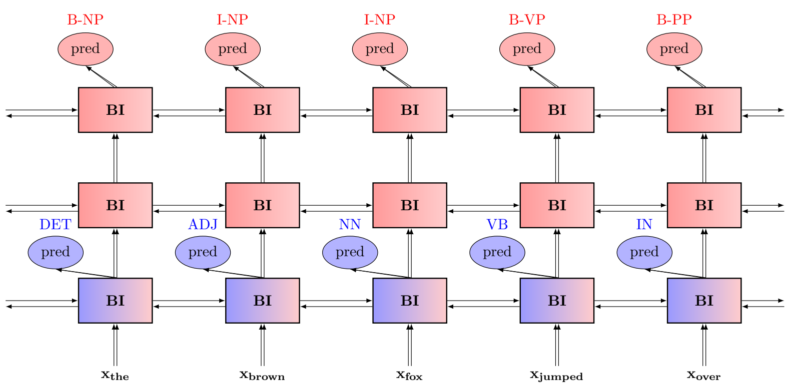


Lower layer is trained for predicting POS (but also gets feedback from CHUNKS)

Chunking scores (F)





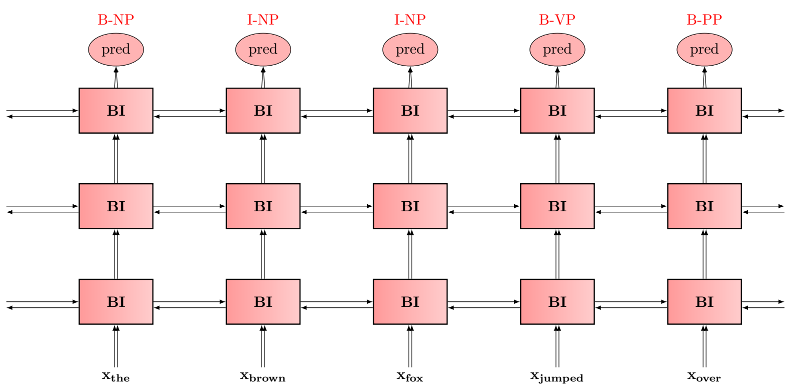


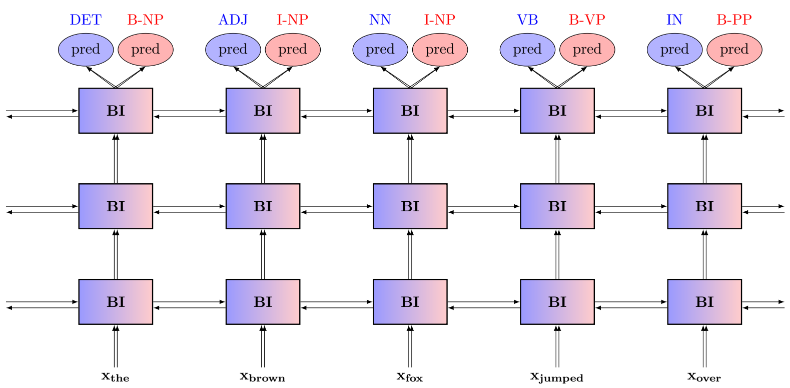
93.8

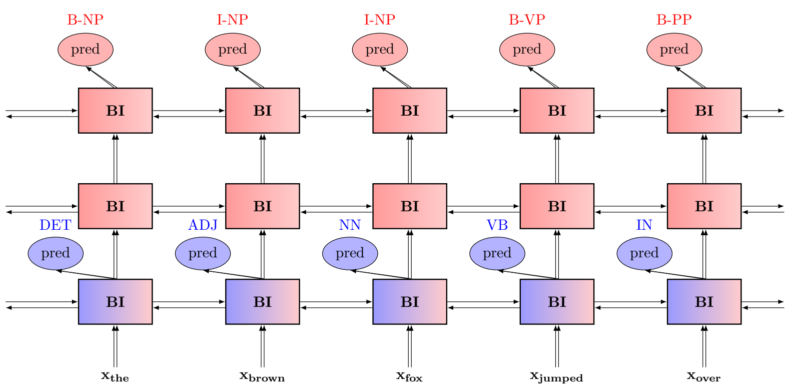
94.5

95.0

CCG Supertagging scores (acc)







91.0

92.9

93.3

Followups

Multitask Learning with Low-Level Auxiliary Tasks for Encoder-Decoder Based Speech Recognition

Shubham Toshniwal, Hao Tang, Liang Lu, and Karen Livescu

Toyota Technological Institute at Chicago
{shtoshni, haotang, llu, klivescu}@ttic.edu

Multitask Learning for Mental Health Conditions with Limited Social Media Data

Adrian Benton
Johns Hopkins University
adrian@cs.jhu.edu

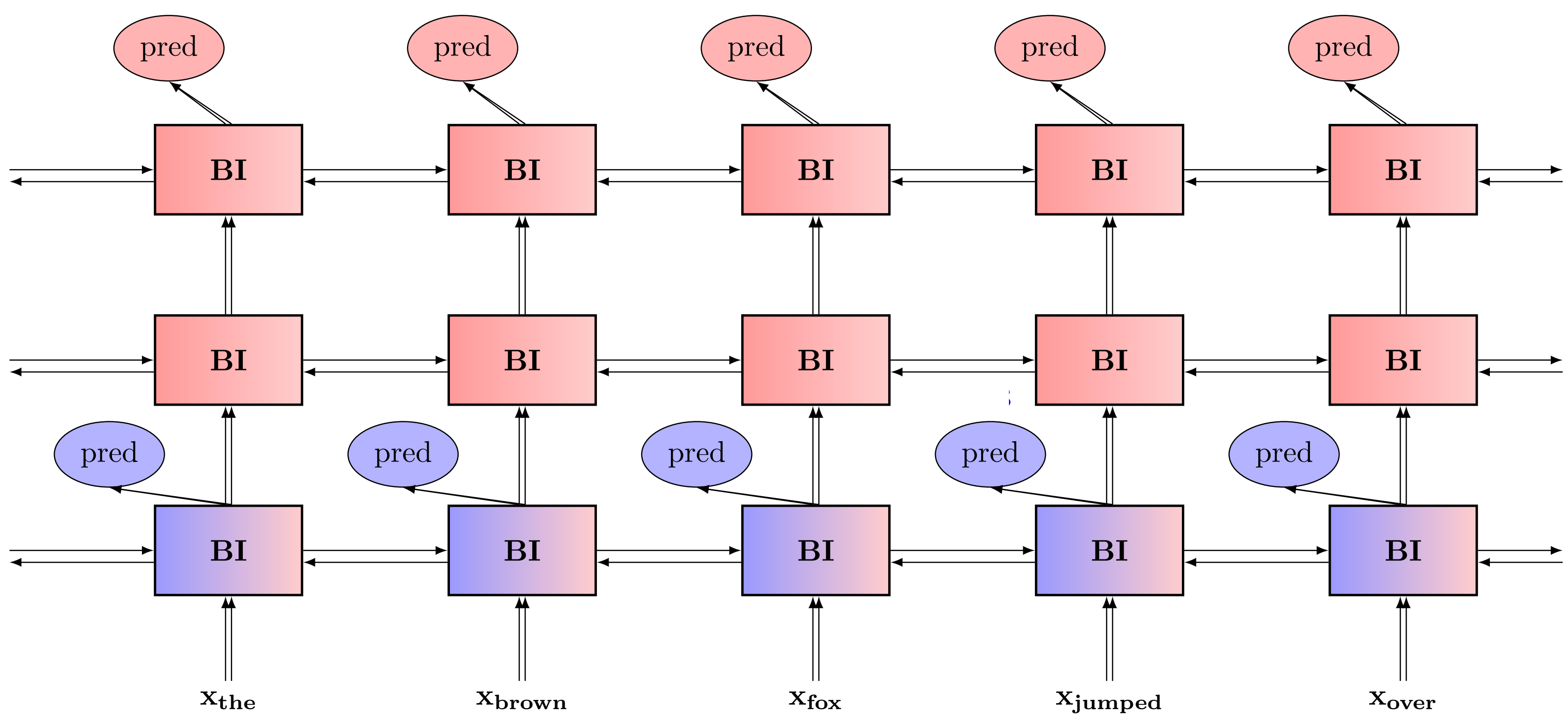
Margaret Mitchell
Microsoft Research[1]
mmitchellai@google.com

Dirk Hovy
University of Copenhagen
mail@dirkhovy.com

Sentence Compression

sentence compression decisions

eye tracking
+
CCG tags



Improving sentence compression by learning to predict gaze

Sigrid Klerke
University of Copenhagen
skl@hum.ku.dk

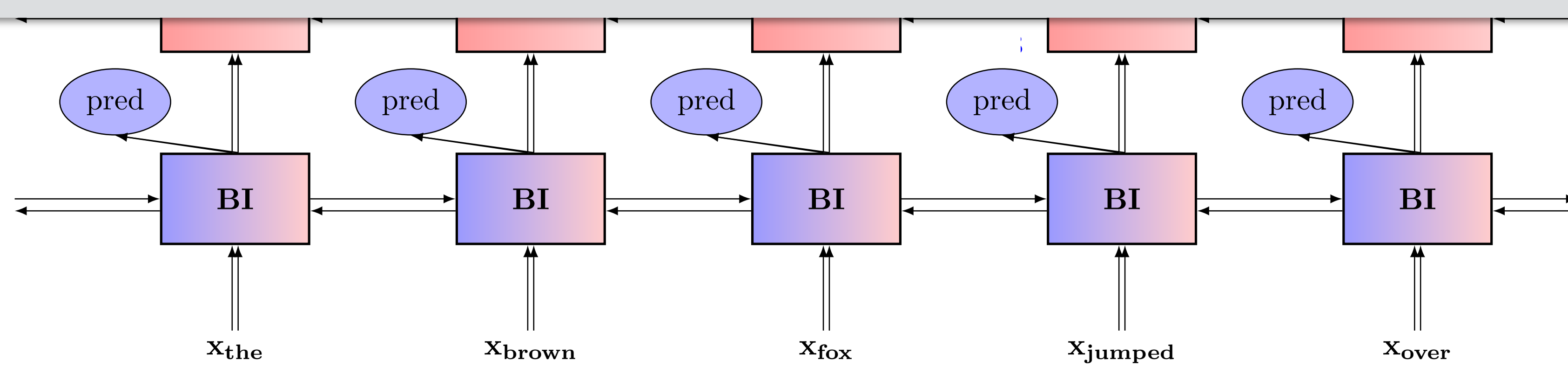
Yoav Goldberg
Bar-Ilan University
yoav.goldberg@gmail.com

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Sentence Compression

The first new product, ATF prototype, is a line of digital postscript typefaces that will be sold in packages of up to six fonts

eye tracking
+
CCG tags



Improving sentence compression by learning to predict gaze

Sigrid Klerke
University of Copenhagen
skl@hum.ku.dk

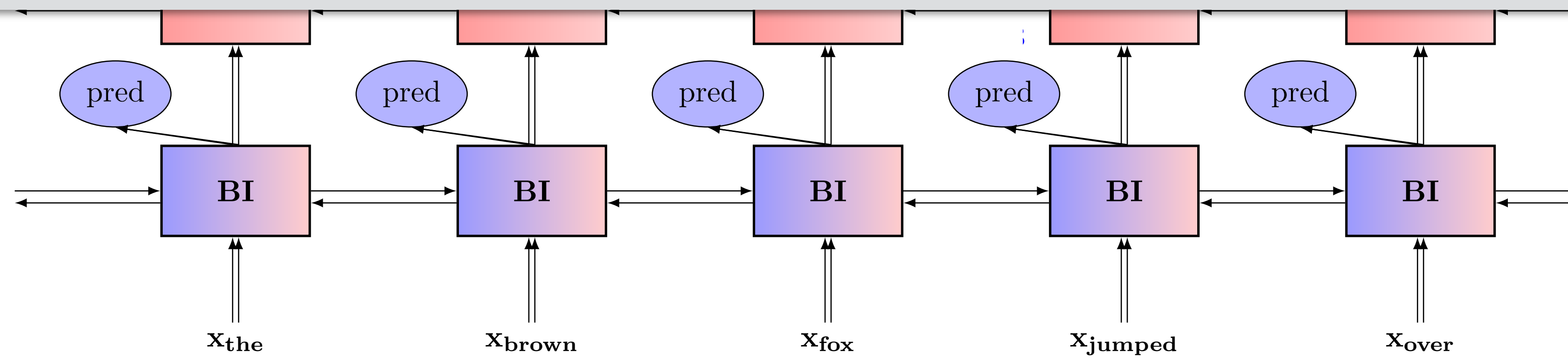
Yoav Goldberg
Bar-Ilan University
yoav.goldberg@gmail.com

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Sentence Compression

~~The first new product~~, ATF prototype, ~~is~~ a line of digital postscript typefaces ~~that~~ will be sold in packages of ~~up to six~~ fonts

eye tracking
+
CCG tags



Improving sentence compression by learning to predict gaze

Sigrid Klerke
University of Copenhagen
skl@hum.ku.dk

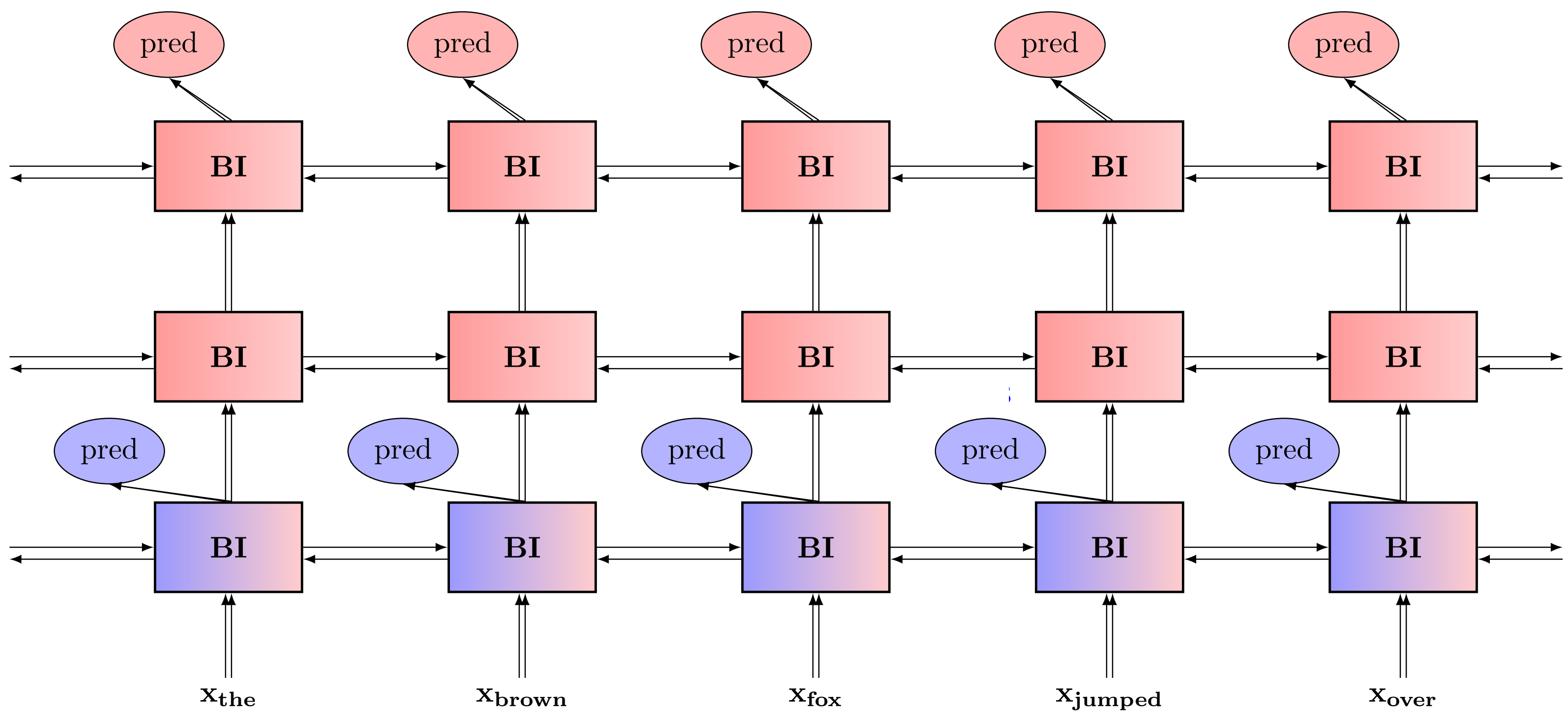
Yoav Goldberg
Bar-Ilan University
yoav.goldberg@gmail.com

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Sentence Compression

sentence
compression
decisions

eye tracking
+
CCG tags



Improving sentence compression by learning to predict gaze

Sigrid Klerke
University of Copenhagen
skl@hum.ku.dk

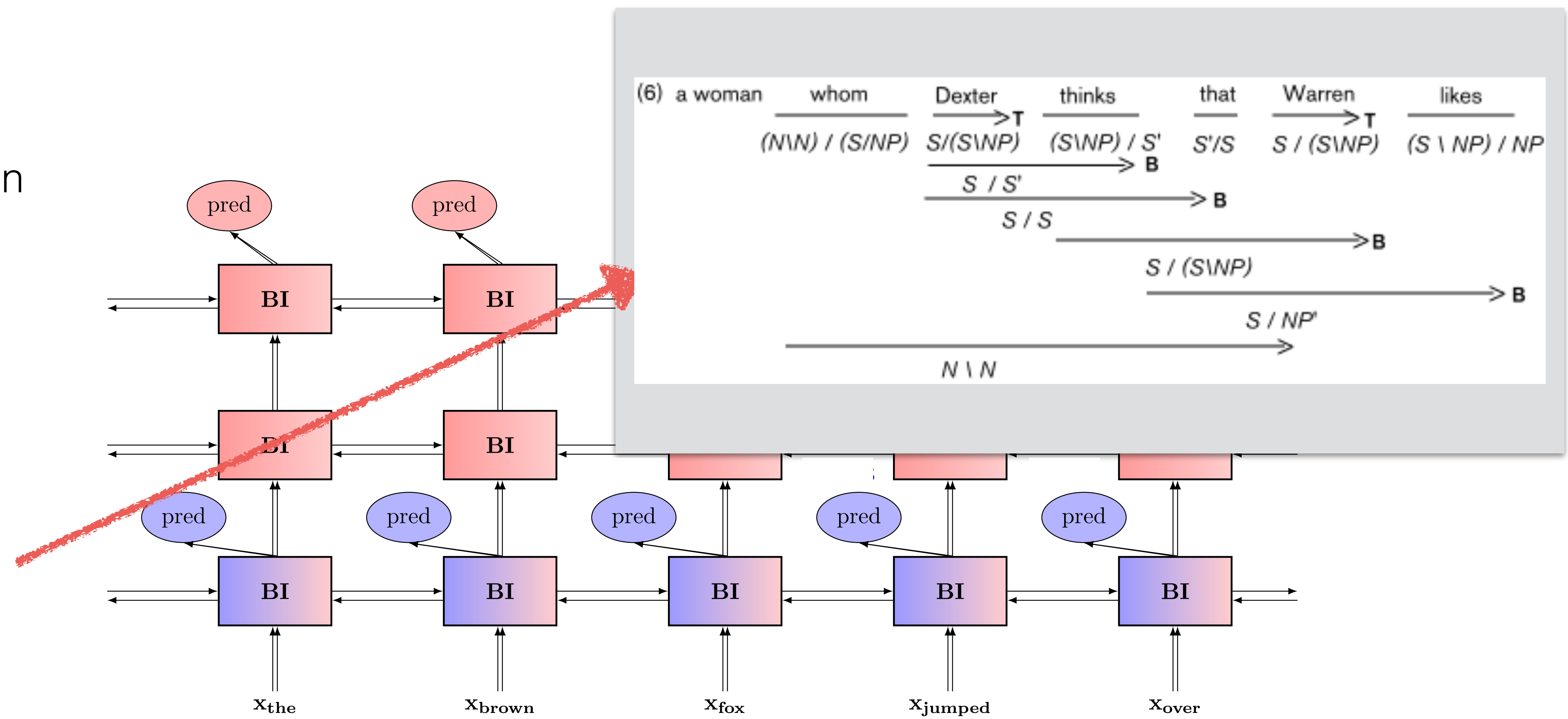
Yoav Goldberg
Bar-Ilan University
yoav.goldberg@gmail.com

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Sentence Compression

sentence
compression
decisions

eye tracking
+
CCG tags



Improving sentence compression by learning to predict gaze

Sigrid Klerke
University of Copenhagen
skl@hum.ku.dk

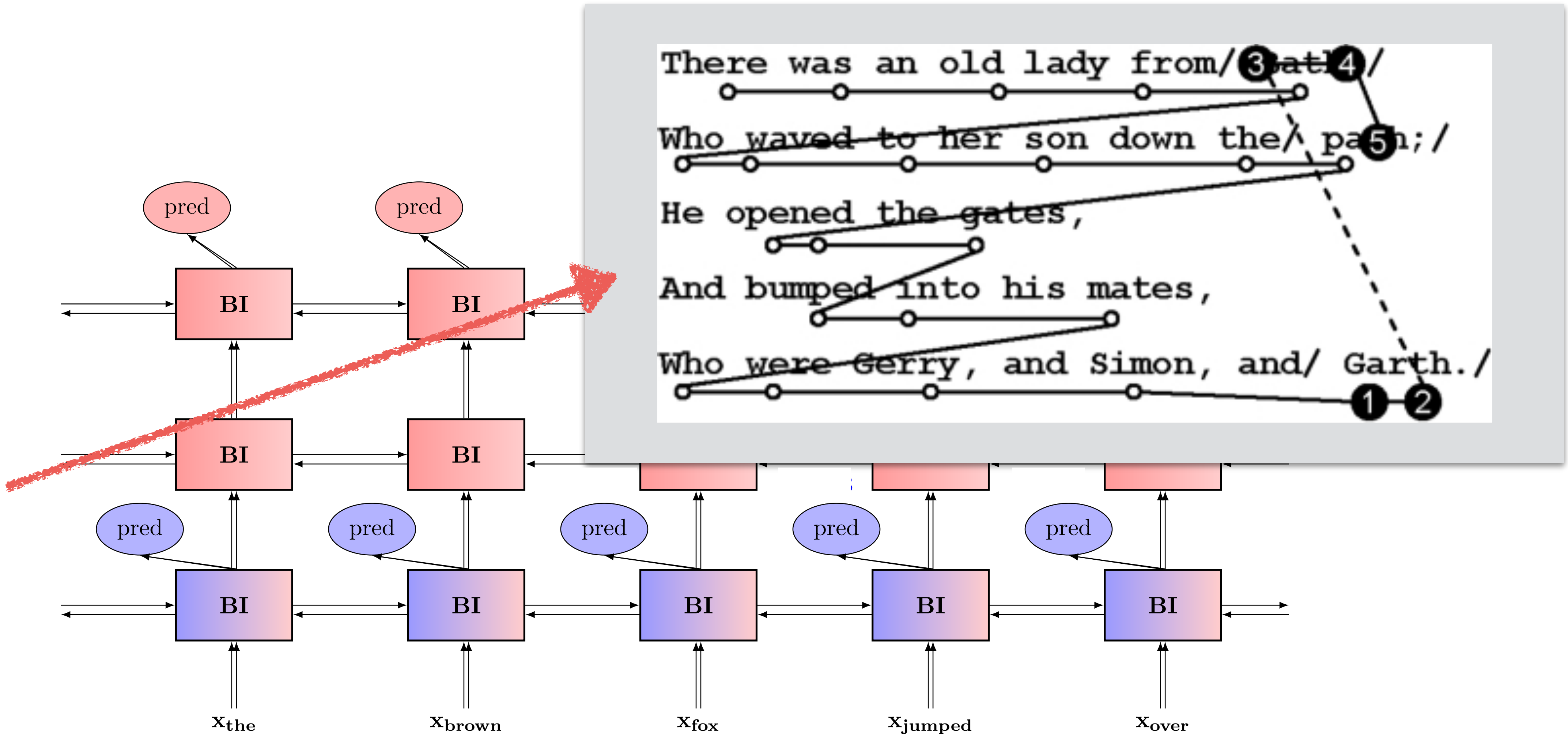
Yoav Goldberg
Bar-Ilan University
yoav.goldberg@gmail.com

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Sentence Compression

sentence compression decisions

eye tracking + CCG tags



Improving sentence compression by learning to predict gaze

Sigrid Klerke
University of Copenhagen
skl@hum.ku.dk

Yoav Goldberg
Bar-Ilan University
yoav.goldberg@gmail.com

Anders Søgaard
University of Copenhagen
soegaard@hum.ku.dk

Additional MTL Example

Semi Supervised Preposition-Sense Disambiguation using Multilingual Data

Hila Gonen
Department of Computer Science
Bar-Ilan University
hilagnn@gmail.com

Yoav Goldberg
Department of Computer Science
Bar-Ilan University
yoav.goldberg@gmail.com

Preposition-Sense Disambiguation

I met him **for** lunch → Purpose
He paid **for** me → Beneficiary
We sat there **for** hours → Duration

Preposition-Sense Disambiguation

I met him **for** lunch → Purpose
He paid **for** me → Beneficiary
We sat there **for** hours → Duration

how will you model this?

I met him **for** lunch → Purpose
He paid **for** me → Beneficiary
We sat there **for** hours → Duration

$$y = \operatorname*{argmax}_j MLP_{sense}(\phi(s,i))[j]$$

Stage 1: Simple MLP with features

I met him **for** lunch → Purpose
He paid **for** me → Beneficiary
We sat there **for** hours → Duration

$$y = \operatorname*{argmax}_j MLP_{sense}(\phi(s,i))[j]$$

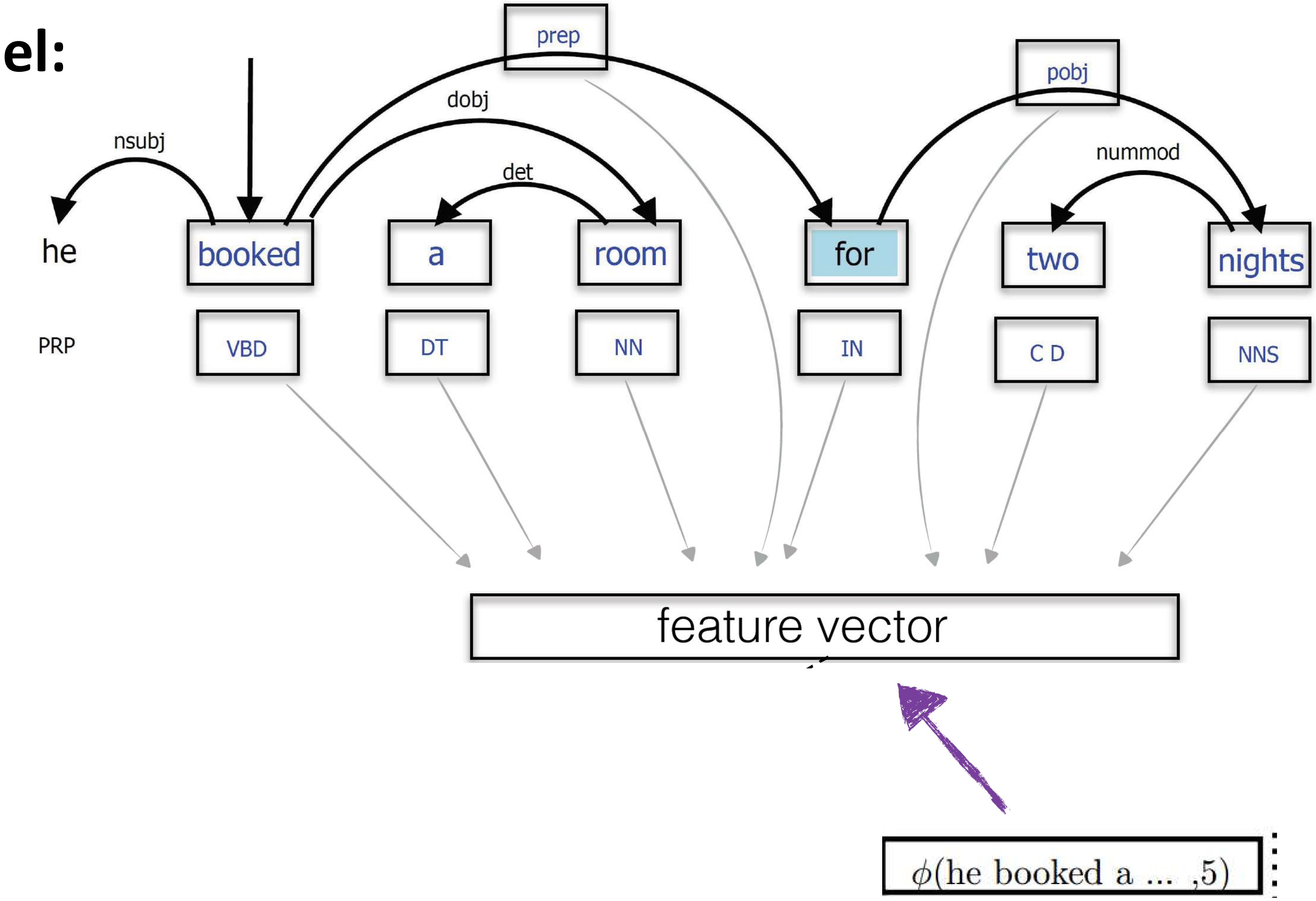
Stage 1: Simple MLP with features

- Embeddings of words in window of 2 to each side
- Embeddings of POS in window of 2 to each side
- Embeddings of Heads and Modifiers of word in Dep Tree
- Are words in window capitalized?

The features and the model:

The features are similar to those used in previous works. Features are extracted from:

- 2-words-window
- Head and modifier of the preposition



I met him **for** lunch → Purpose
He paid **for** me → Beneficiary
We sat there **for** hours → Duration

$$y = \operatorname*{argmax}_j MLP_{sense}(ctx(s,i) \circ \phi(s,i))[j]$$

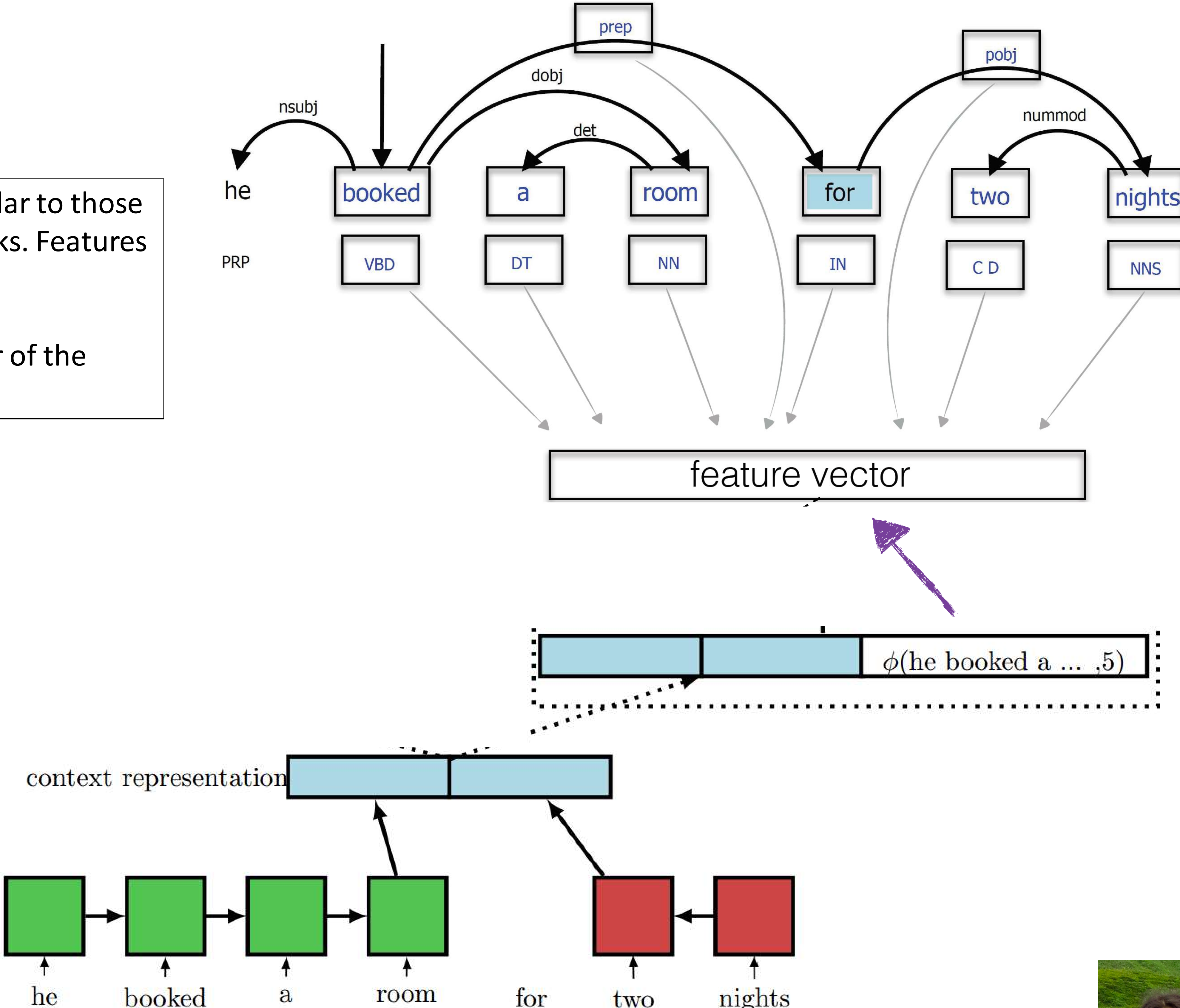
Stage 2: Adding Context

$$ctx(s,i) = RNN_f(\mathbf{x_{1:i-1}}) \circ RNN_b(\mathbf{x_{n:i+1}})$$

(almost biRNN, but not exactly. What's the difference?)

The features are similar to those used in previous works. Features are extracted from:

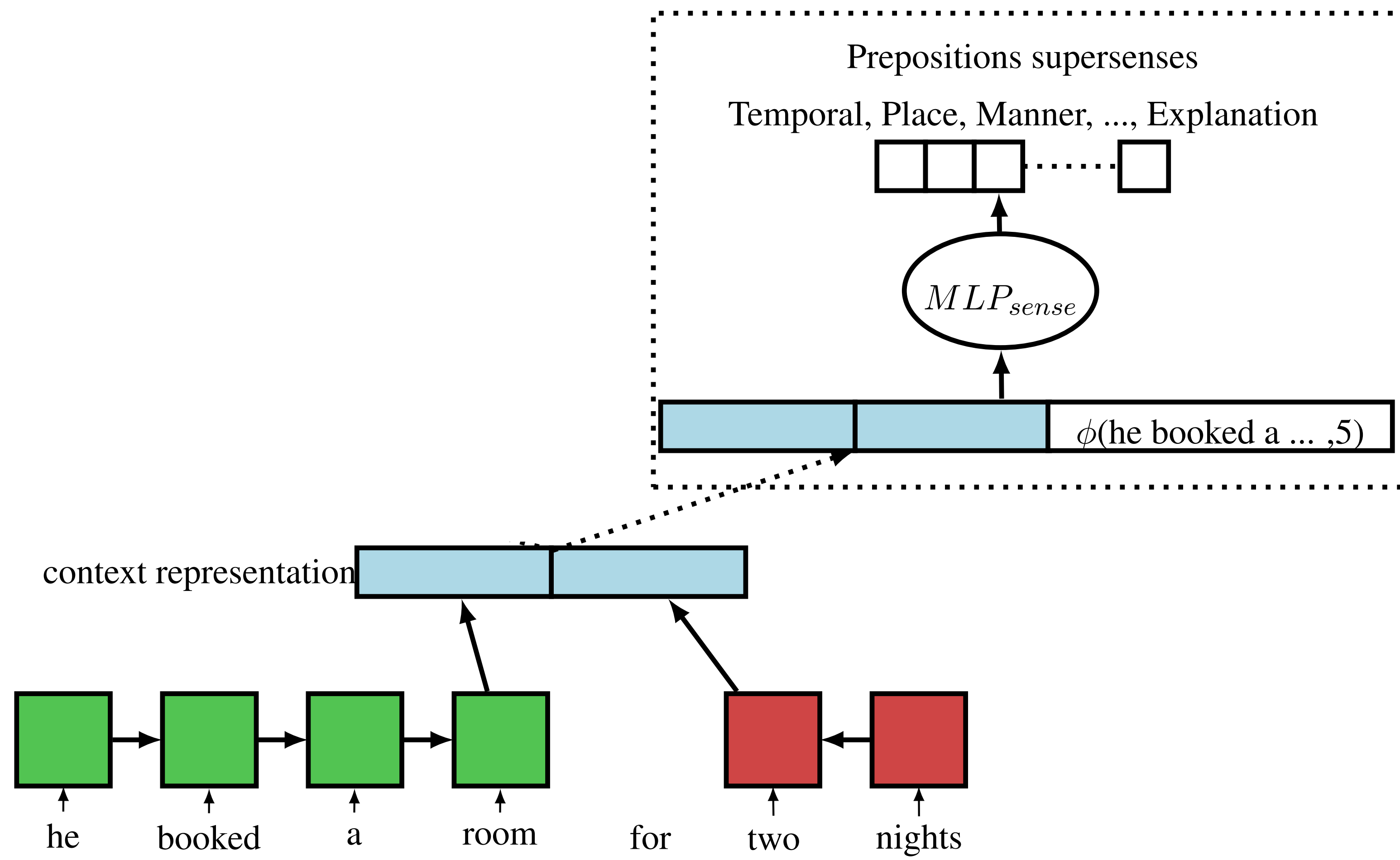
- 2-words-window
- Head and modifier of the preposition



I met him **for** lunch → Purpose

He paid **for** me → Beneficiary

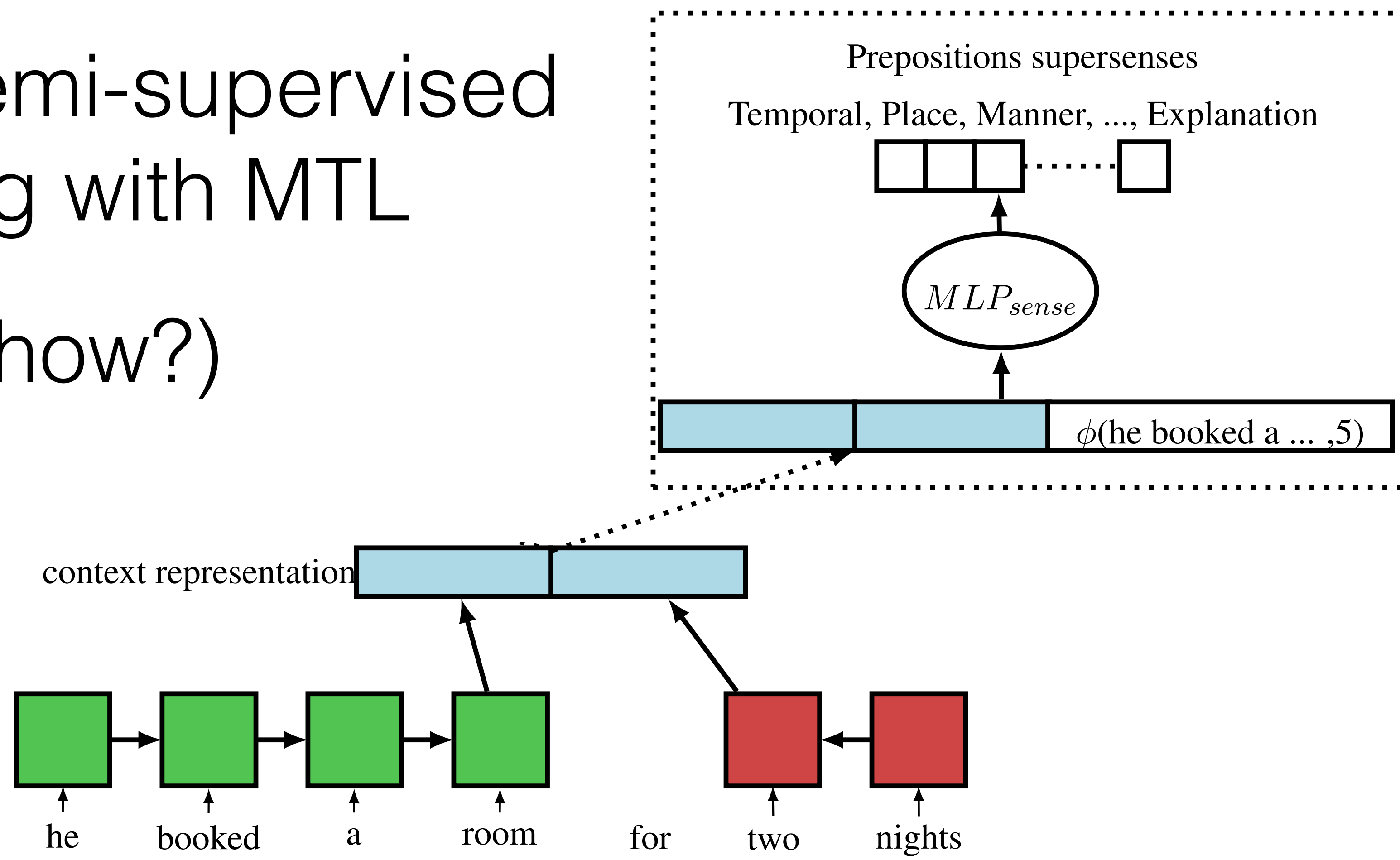
We sat there **for** hours → Duration



I met him **for** lunch → Purpose
He paid **for** me → Beneficiary
We sat there **for** hours → Duration

Adding Semi-supervised training with MTL

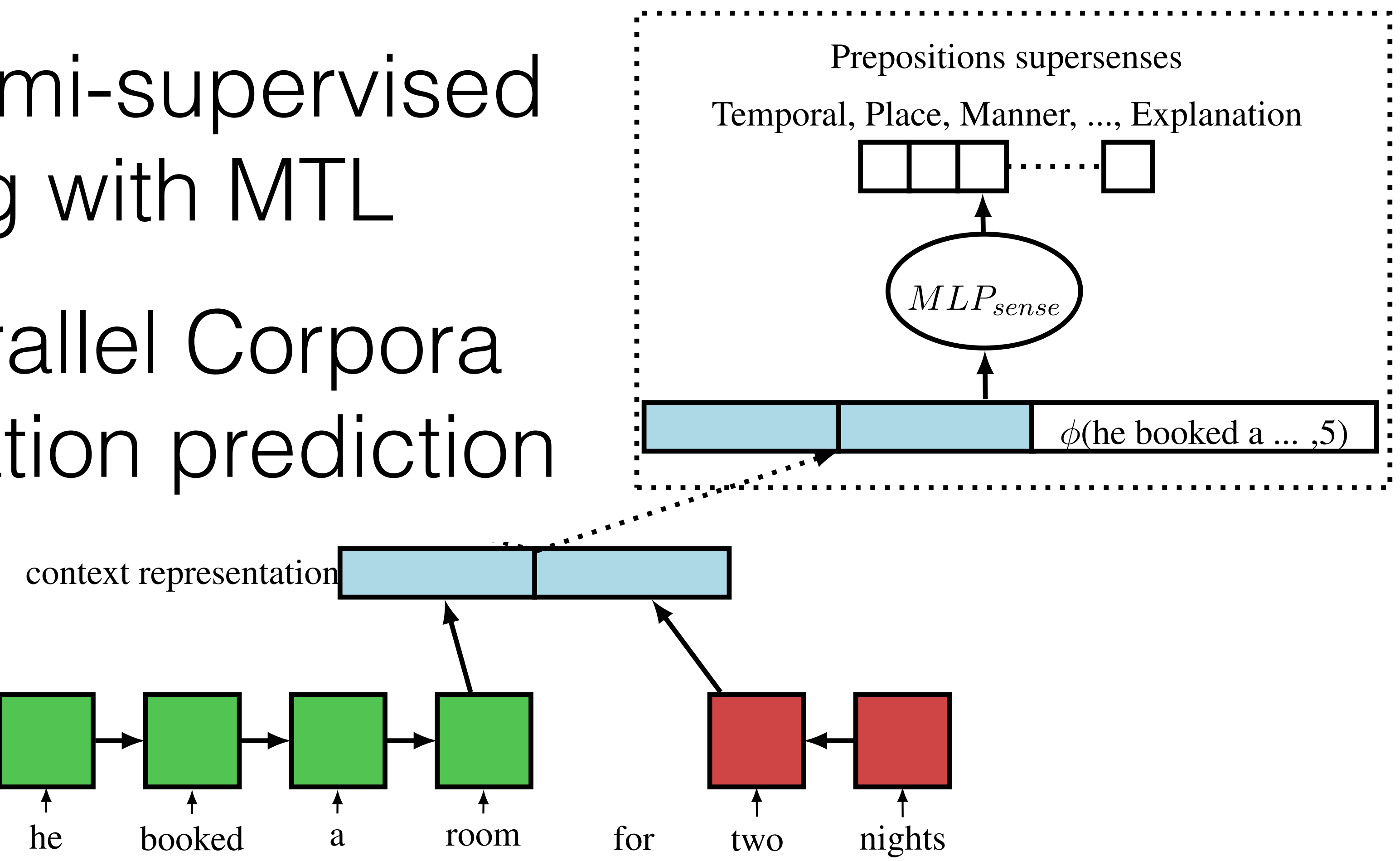
(how?)

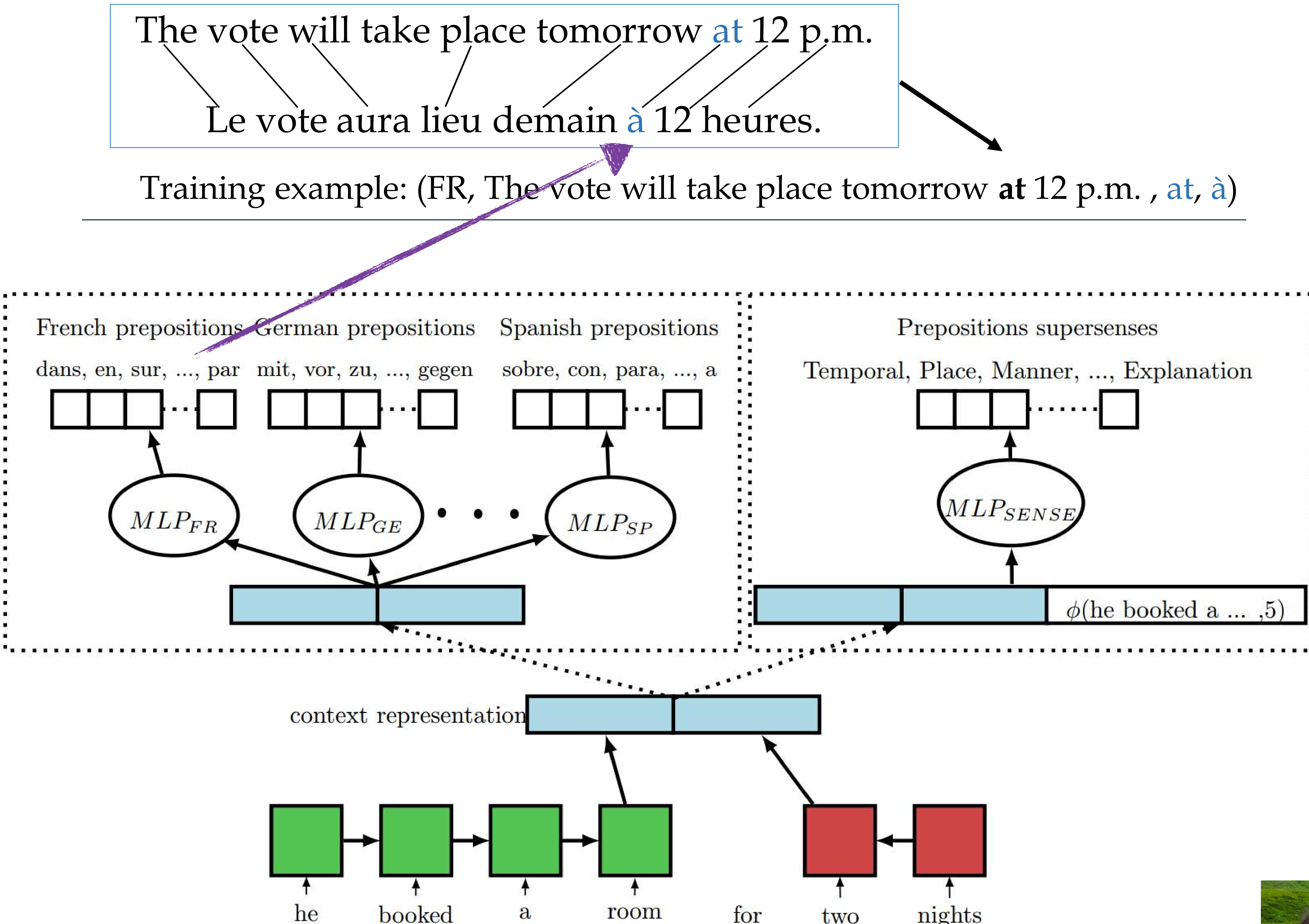


I met him **for** lunch → Purpose
He paid **for** me → Beneficiary
We sat there **for** hours → Duration

Adding Semi-supervised training with MTL

Using Parallel Corpora and translation prediction




The vote will take place tomorrow at 12 p.m.
Le vote aura lieu demain à 12 heures.
Training example: (FR, The vote will take place tomorrow **at** 12 p.m. , at, à)
French prepositions
dans, en, sur, ..., par
German prepositions
mit, vor, zu, ..., gegen
Spanish prepositions
sobre, con, para, ..., a
MLP_{FR}
MLP_{GE}
MLP_{SP}
Prepositions supersenses
Temporal, Place, Manner, ..., Explanation
MLP_{SENSE}
ϕ(he booked a ... ,5)
context representation
he
booked
a
room
for
two
nights

I met him **for** lunch → Purpose
He paid **for** me → Beneficiary
We sat there **for** hours → Duration

Model	Accuracy
base	73.34 (71.63-73.97)
+context	73.76 (71.86-75.38)
+context(multilingual)	**76.20** (74.91-77.26)

(with pre-trained embeddings, ensembles, get to ~80)

MTL - Recap

- For **related tasks**, can get nice gains from MTL.
- **Thinking about the architecture helps**.

Ultimate task: Language Modeling

- Train a model on "what is the next word?"

- The resulting representation is very useful for many different tasks.

Ultimate task: Language Modeling

ELMo
(NAACL 2018)

Deep contextualized word representations

Matthew E. Peters[†], Mark Neumann[†], Mohit Iyyer[†], Matt Gardner[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], Kenton Lee[*], Luke Zettlemoyer[†*]
{csquared,kentonl,lsz}@cs.washington.edu

Ultimate task: Language Modeling

ELMo
(NAACL 2018)

Deep contextualized word representations

Matthew E. Peters†, Mark Neumann†, Mohit Iyyer†, Matt Gardner†,
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark*, Kenton Lee*, Luke Zettlemoyer†*
{csquared,kentonl,lsz}@cs.washington.edu

Universal Language Model Fine-tuning for Text Classification

Jeremy Howard*
fast.ai
University of San Francisco
j@fast.ai

Sebastian Ruder*
Insight Centre, NUI Galway
Aylien Ltd., Dublin
sebastian@ruder.io

ULMfit
(ACL 2018)

Ultimate task: Language Modeling

ELMo
(NAACL 2018)

Deep contextualized word representations

Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

Universal Language Model Fine-tuning for Text Classification

Jeremy Howard[*]
fast.ai
University of San Francisco
j@fast.ai

Sebastian Ruder[*]
Insight Centre, NUI Galway
Aylien Ltd., Dublin
sebastian@ruder.io

ULMfit
(ACL 2018)

OpenAI
(arxiv+PR 2018)

Improving Language Understanding by Generative Pre-Training

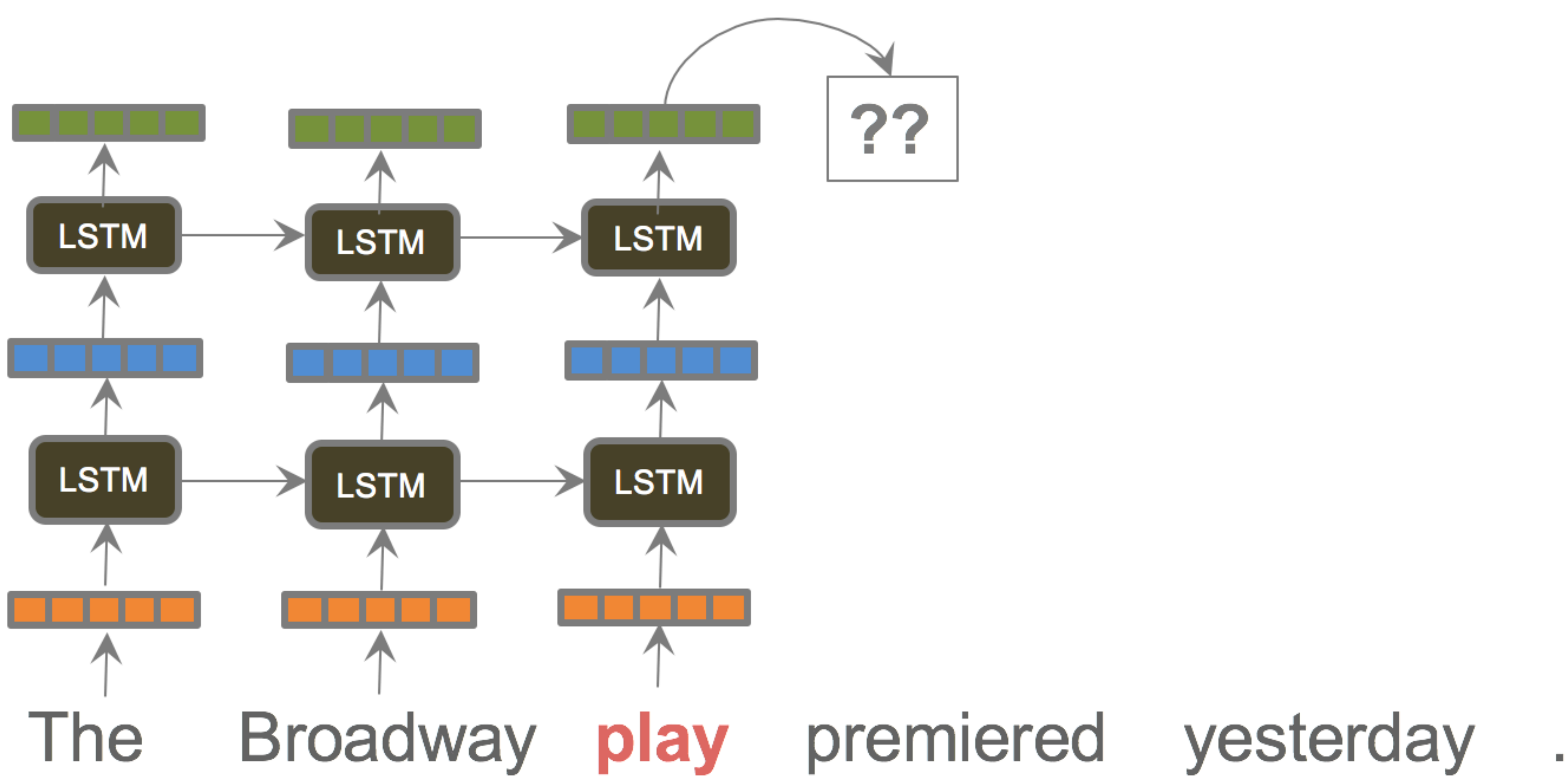
Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

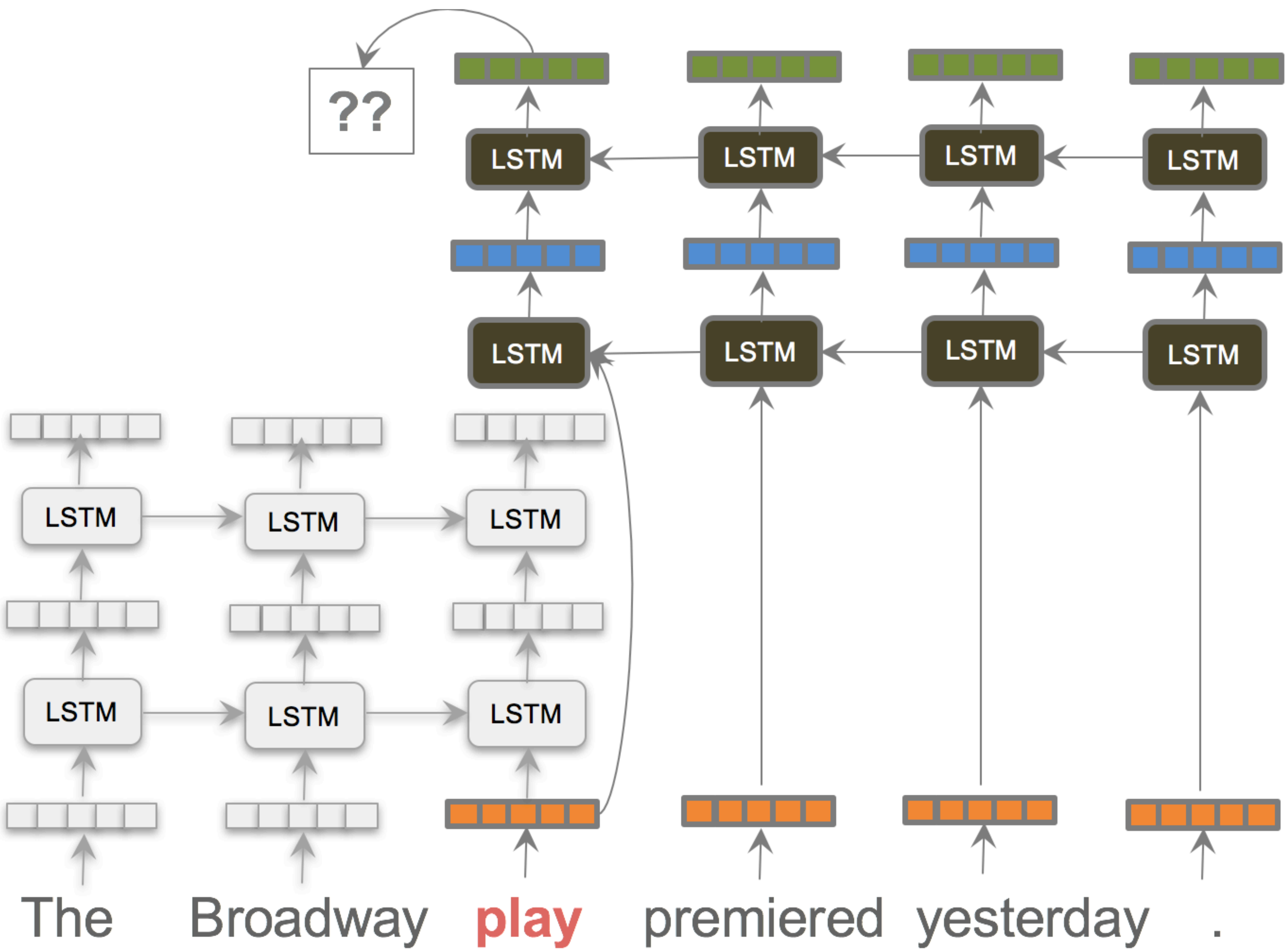
Ilya Sutskever
OpenAI
ilyasu@openai.com

ELMo

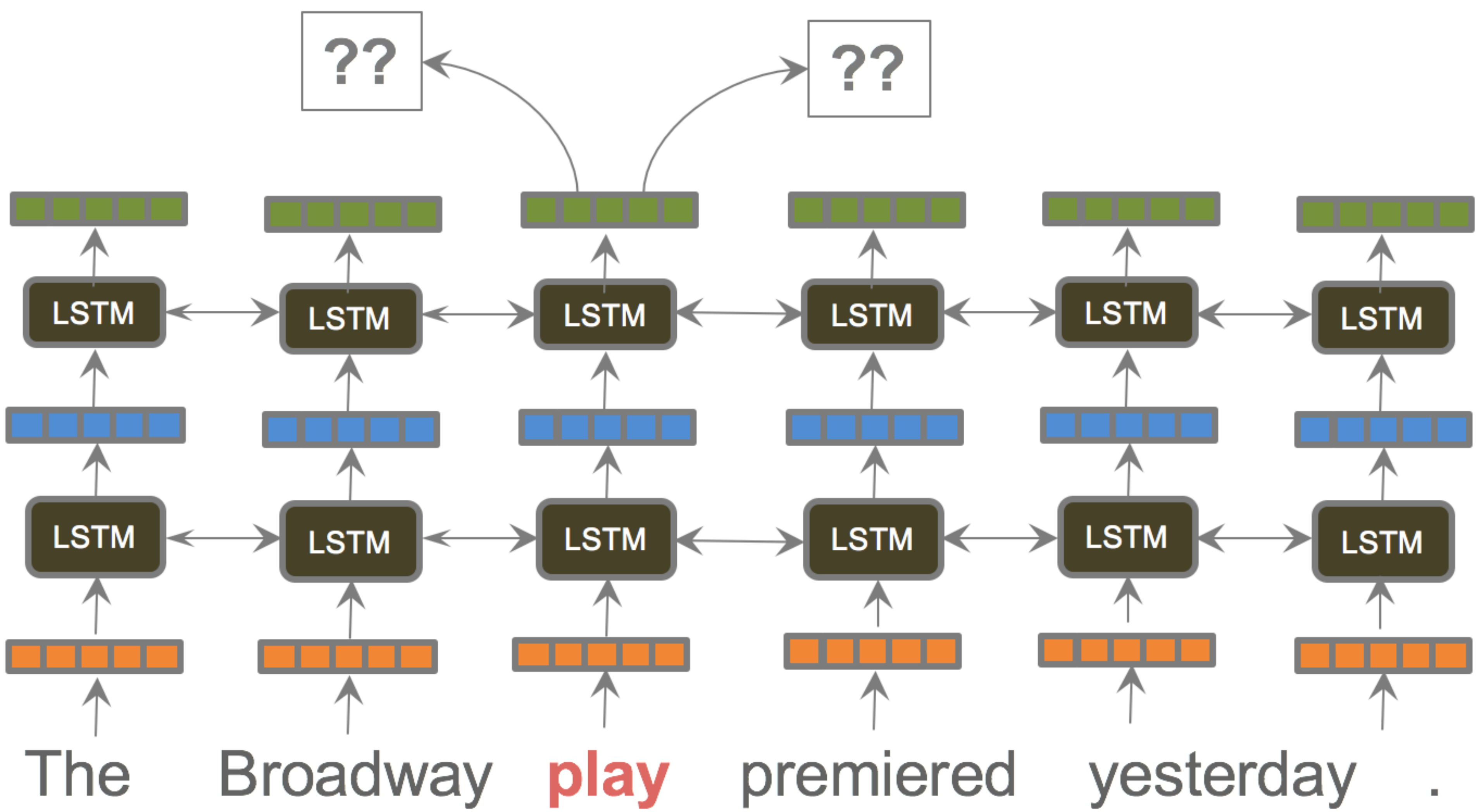


LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
??
The
Broadway
play
premiered
yesterday
.

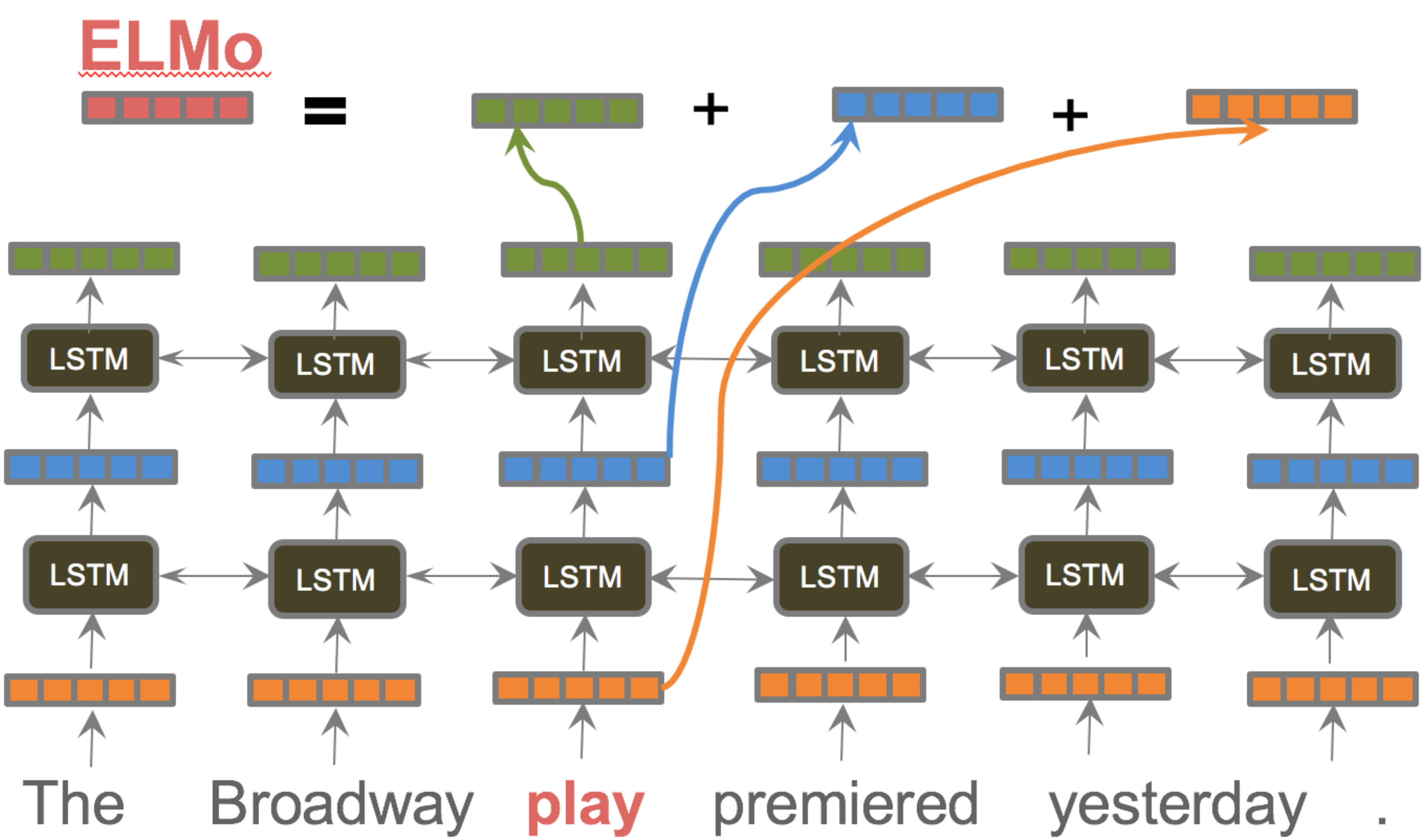
ELMo

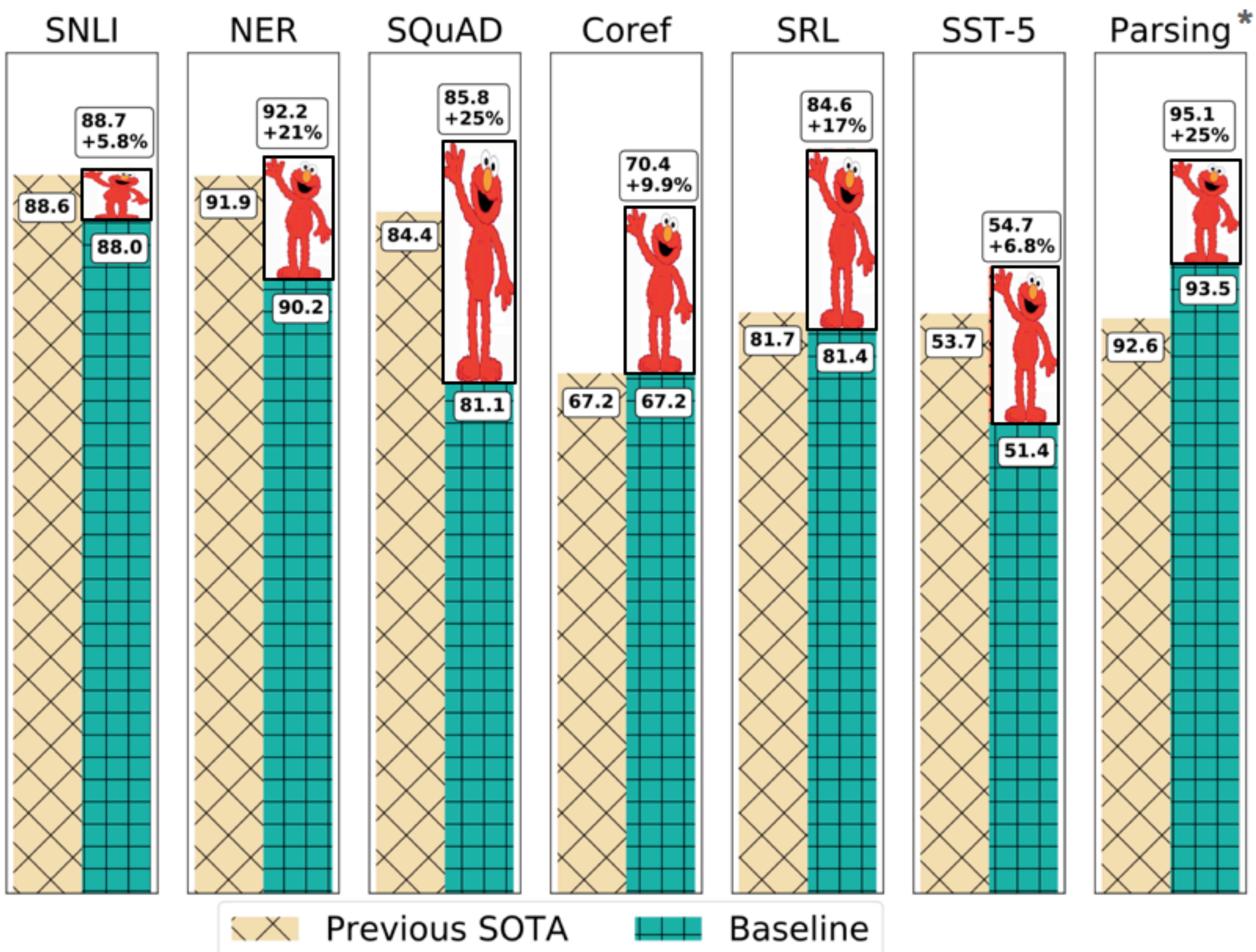

??
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
The
Broadway
play
premiered
yesterday
.

ELMo



ELMo




SNLI
NER
SQuAD
Coref
SRL
SST-5
Parsing*
88.7 +5.8%
88.6
88.0
92.2 +21%
91.9
90.2
85.8 +25%
84.4
81.1
70.4 +9.9%
67.2
67.2
84.6 +17%
81.7
81.4
54.7 +6.8%
53.7
51.4
95.1 +25%
92.6
93.5
Previous SOTA
Baseline

Ultimate task: Language Modeling

- Train a model on "what is the **next** word?"
- The resulting representation is **very useful** for many different tasks.

Ultimate task:
Language Modeling

- Train a model on "what is the **next** word?"

- The resulting representation is **very useful** for many different tasks.

PostUltimate task:
Masked Language Modeling?

- Train a model on "what is the **missing** word?"

- The resulting representation is **even more useful** for many different tasks.

Ultimate task: Language Modeling

- Train a model on "what is the **next** word?"
- The resulting representation is **very useful** for many different tasks.

PostUltimate task: **Masked** Language Modeling?

- Train a model on "what is the **missing** word?"
- The resulting representation is **even more useful** for many different tasks.

PostUltimate task: **Masked** Language Modeling?

- Trai
- The … or many diffe

Why does it work?
under what conditions?
should we fine-tune?
what happens in fine-tuning?
can we have a theory for this?

PostUltimate task:
Masked Language Modeling?

- Train a model on "what is the next word?"
- The ... or many diffe...

Why does it work?
under what conditions?
should we fine-tune?
what happens in fine-tuning?
can we have a theory for this?

no

PostUltimate task: **Masked** Language Modeling?

- Trai
- The … or many diffe

Why does it work?
under what conditions?
should we fine-tune?
what happens in fine-tuning?
can we have a theory for this?

no

Moving back to more discrete representations?

Word Sense Induction with Neural biLM and Symmetric Patterns

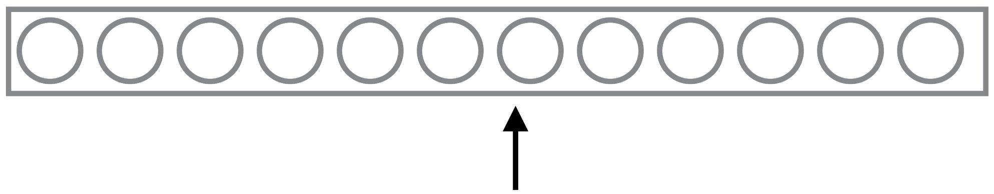
Asaf Amrami † and **Yoav Goldberg** † ‡

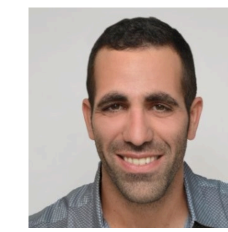
† Computer Science Department, Bar Ilan University, Israel
‡ Allen Institute for Artificial Intelligence

{asaf.amrami, yoav.goldberg}@gmail.com

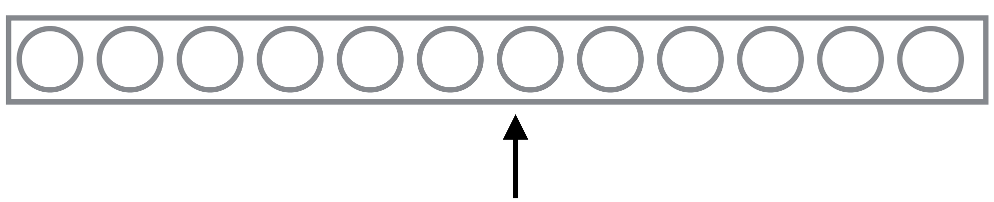
Task: Word Sense Induction

- We are given k sentences with the same word.
- We need to **cluster** them into groups according to senses.
- Can we use ELMo (or similar, or BERT) for this?





`this is a` **`sound`** `idea, I like it.`



`I like the` **`sound`** `of the harpsichord.`

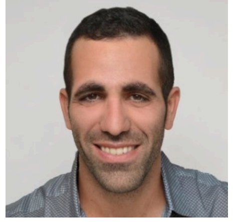
- Represent each word based on its ELMo/BERT vector.
- Cluster the vectors.

ELMO-based word sense induction

- This sort-of works... but not very well.
- What went wrong? who knows.
- How can we improve? great question.

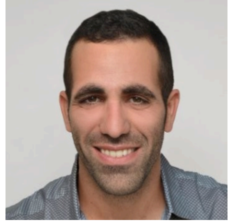
if only the vectors were more transparent!!

Back to more discrete representations



- **Substitute vectors**
  - Using the LM as an LM
  - Represent a word as a distribution of substitute words

Back to more discrete representations



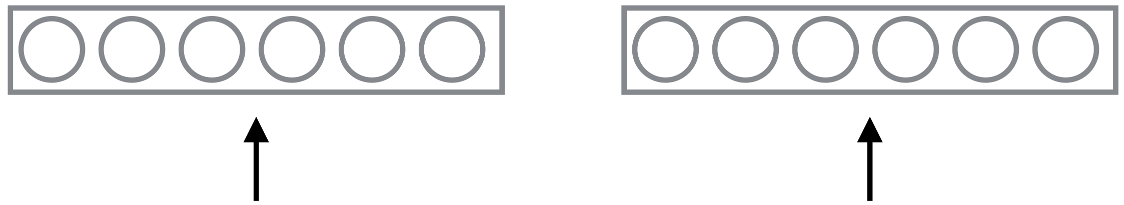
- **Substitute vectors**
    - Using the LM as an LM
    - Represent a word as a distribution of substitute words
- This is not our own idea.

AI-KU: Using Substitute Vectors and Co-Occurrence Modeling for Word Sense Induction and Disambiguation *SEM 2013

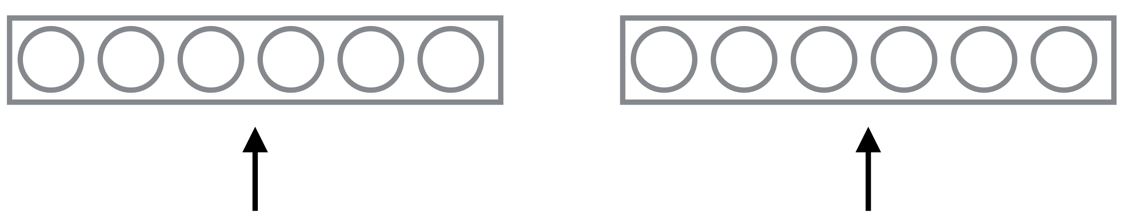
Osman Başkaya **Enis Sert** **Volkan Cirik** **Deniz Yuret**

Back to more discrete representations

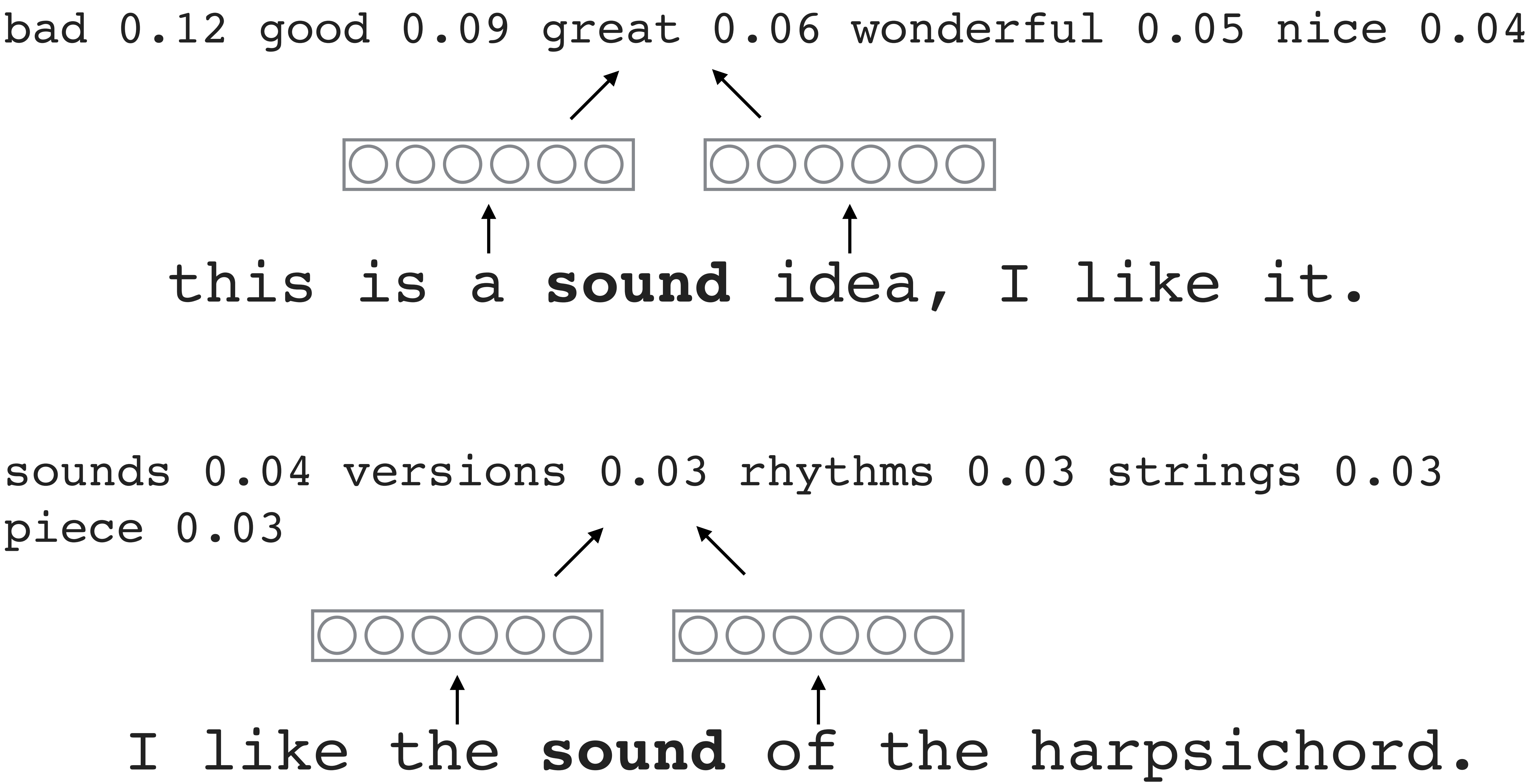
- **Substitute vectors**
    - Using the LM as an LM
    - Represent a word as a distribution of substitute words
- This is not our own idea.
    - But now we have **neural** LM



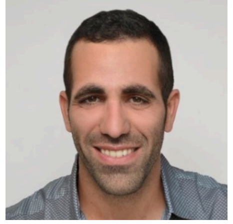
this is a **sound** idea, I like it.



I like the **sound** of the harpsichord.


bad 0.12 good 0.09 great 0.06 wonderful 0.05 nice 0.04
this is a **sound** idea, I like it.
sounds 0.04 versions 0.03 rhythms 0.03 strings 0.03 piece 0.03
I like the **sound** of the harpsichord.

State-vectors --> Word Distributions



- By looking at the substitute word distributions rather than the state vectors, we get a better understanding of what going on.

Two recent discoveries indicate probable very early **settlements** near the Thames in the London area .

Structured **settlements** provide for future periodic payments .

Two recent discoveries indicate probable very early **settlements** near the Thames in the London area .

Structured **settlements** provide for future periodic payments .

development 0.34 stage 0.14 death 0.13 signs 0.08
stages 0.07 life 0.04 cases 0.03 properties 0.02

Two recent discoveries indicate probable very early **settlements** near the Thames in the London area .

to 0.32 loans 0.23 and 0.12 products 0.08 as 0.06
credit 0.03 bonds 0.03 deals 0.03 securities 0.03

Structured **settlements** provide for future periodic payments .

development 0.34 stage 0.14 death 0.13 signs 0.08 stages 0.07 life 0.04 cases 0.03 properties 0.02

Two recent discoveries indicate probable very early **settlements** near the Thames in the London area .

to 0.32 loans 0.23 and 0.12 products 0.08 as 0.06 credit 0.03 bonds 0.03 deals 0.03 securities 0.03

Structured **settlements** provide for future periodic payments .

development 0.34 stage 0.14 death 0.13 signs 0.08 stages 0.07 life 0.04 cases 0.03 properties 0.02 ✗

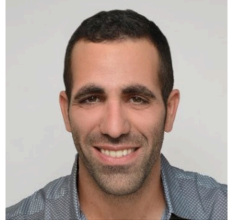
Two recent discoveries indicate probable very early **settlements** near the Thames in the London area .

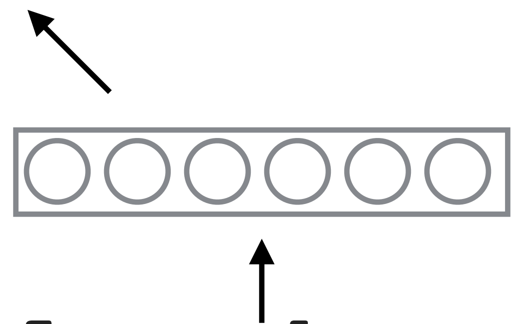
Problem: **no information about the word itself.**

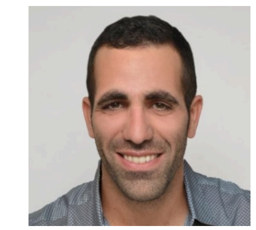
Better Word Distributions

- Query the language model in a creative way this is a **sound**

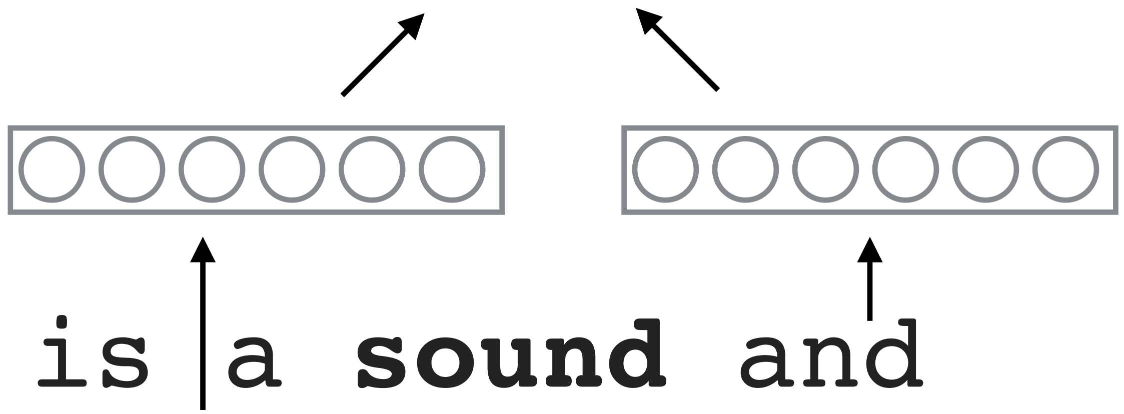
sound idea, I like it.





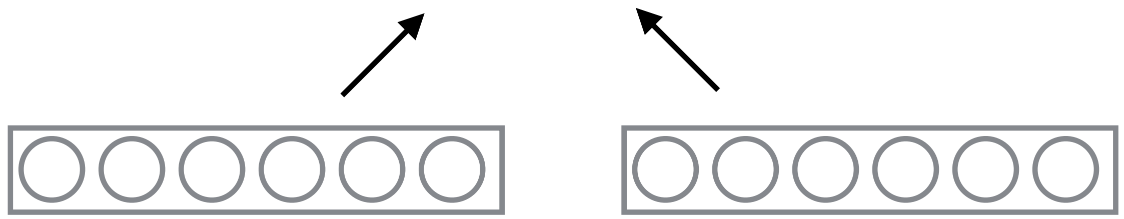


this is a **sound** and
sound idea, I like it.



this is a **sound** and

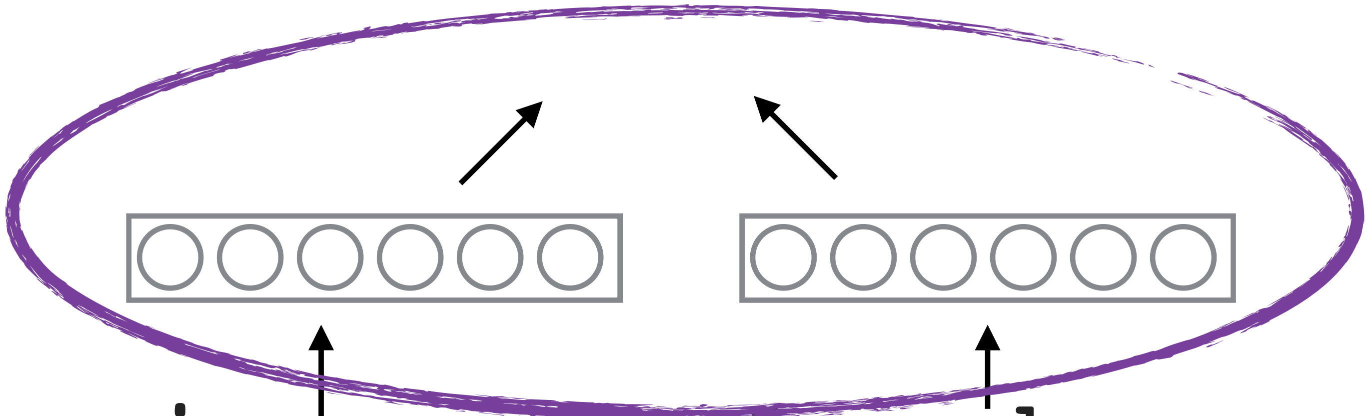
and **sound** idea, I like it.



this is a **sound** and

and **sound** idea, I like it.

Hearst patterns / symmetric patterns

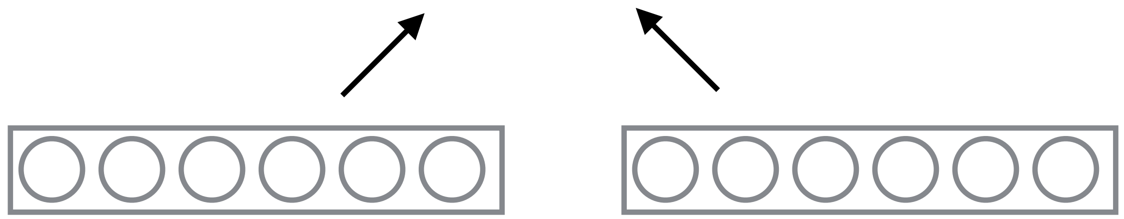


`this is a` **`sound`** `and`

`and` **`sound`** `idea, I like it.`

Hearst patterns / symmetric patterns

Neural-LM Query



this is a **sound** and and **sound** idea, I like it.

Hearst patterns / symmetric patterns

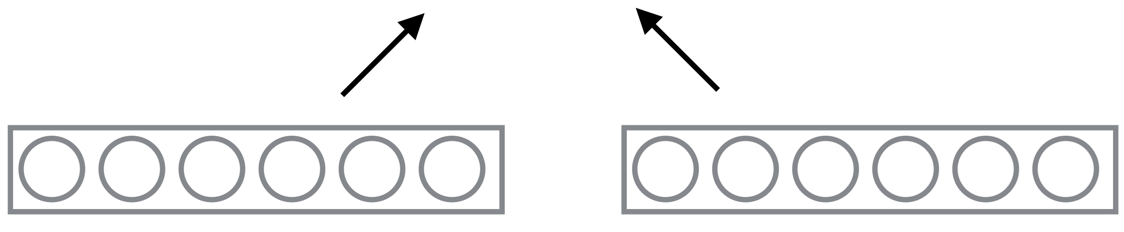
\+

Neural-LM Query

=

Context-dependent Hearst patterns

funny 0.10 welcome 0.09 beautiful 0.05 fun 0.04
simple 0.04 practical 0.03 comprehensive 0.03



this is a **sound** and
and **sound** idea, I like it.

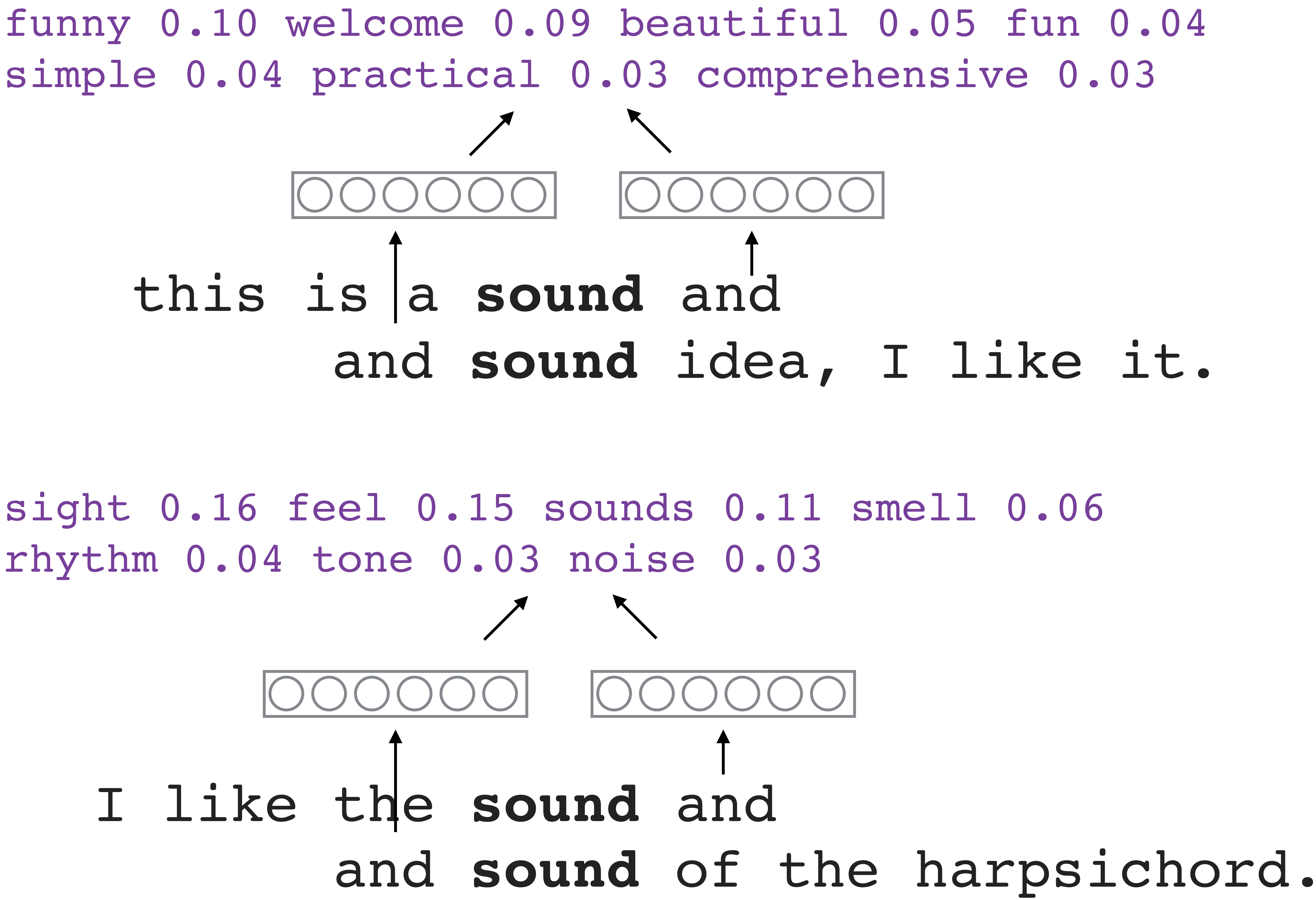
Hearst patterns / symmetric patterns

+

Neural-LM Query

=

Context-dependent Hearst patterns


funny 0.10 welcome 0.09 beautiful 0.05 fun 0.04
simple 0.04 practical 0.03 comprehensive 0.03
this is a **sound** and
and **sound** idea, I like it.
sight 0.16 feel 0.15 sounds 0.11 smell 0.06
rhythm 0.04 tone 0.03 noise 0.03
I like the **sound** and
and **sound** of the harpsichord.

Gulls nest in large , densely packed , **noisy** colonies .

Substitute Vector

urban 0.36
remote 0.12
isolated 0.11
tropical 0.10
dense 0.06

Contextualized Hearst

crowded 0.54
remote 0.14
noisy 0.09
overcrowded 0.05
cramped 0.04

land 0.25 sites 0.07 buildings 0.03 homes 0.02 plants 0.01 farms 0.01 development 0.01

Two recent discoveries indicate probable very early **settlements** near the Thames in the London area .

agreements 0.40 payments 0.13 contracts 0.10 loans 0.07 fees 0.05 swaps 0.03 litigation 0.02 transactions 0.01

Structured **settlements** provide for future periodic payments .

land 0.25 sites 0.07 buildings 0.03 homes 0.02 plants 0.01 farms 0.01 development 0.01

Two recent discoveries indicate probable very early **settlements** near the Thames in the London area .

agreements 0.40 payments 0.13 contracts 0.10 loans 0.07 fees 0.05 swaps 0.03 litigation 0.02 transactions 0.01

Structured **settlements** provide for future periodic payments .

NGRAM LM (AI-KU)	15.9
ELMo LM	23.4
ELMo LM + pattern	**25.4**

(Avg of FNMI and FBC on SemEval 2013 task 13)

Towards better substitution-based word sense induction

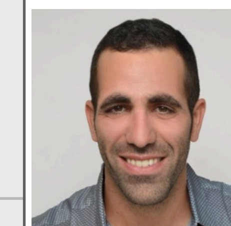
Asaf Amrami [†] and **Yoav Goldberg** [† ‡]

[†] Computer Science Department, Bar Ilan University, Israel
[‡] Allen Institute for Artificial Intelligence

{asaf.amrami, yoav.goldberg}@gmail.com

AKA "Let's try it with BERT"

NGRAM LM (AI-KU)	15.92
ELMo LM	23.4
ELMo LM + pattern	**25.4**



AKA "Let's try it with BERT"

BERT	35.1
BERT + pattern	**37.0**

Takeaway

- From opaque biLM state
  --> to transparent biLM word distribution

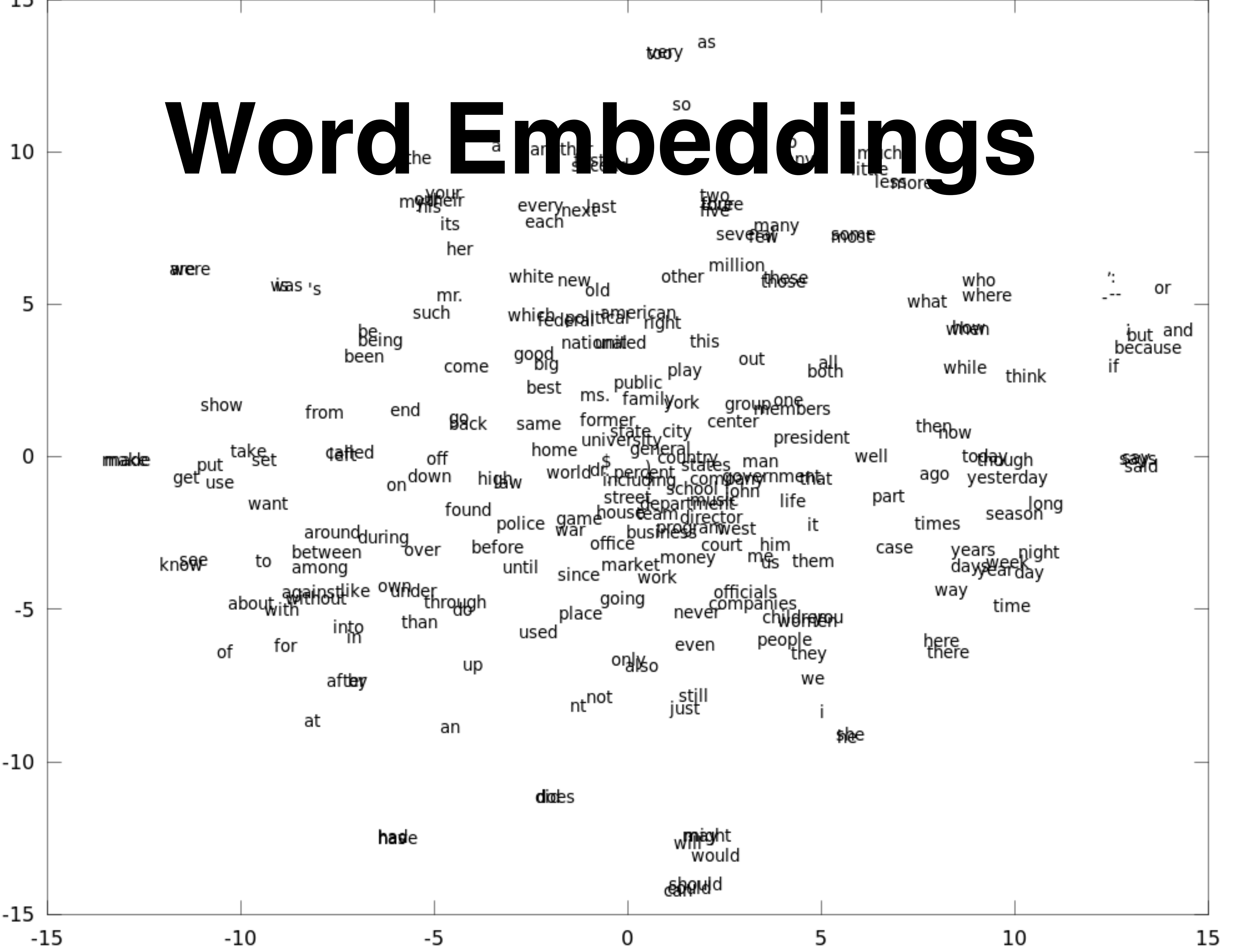
- Can look at things and try to debug

- Query the model in a creative way

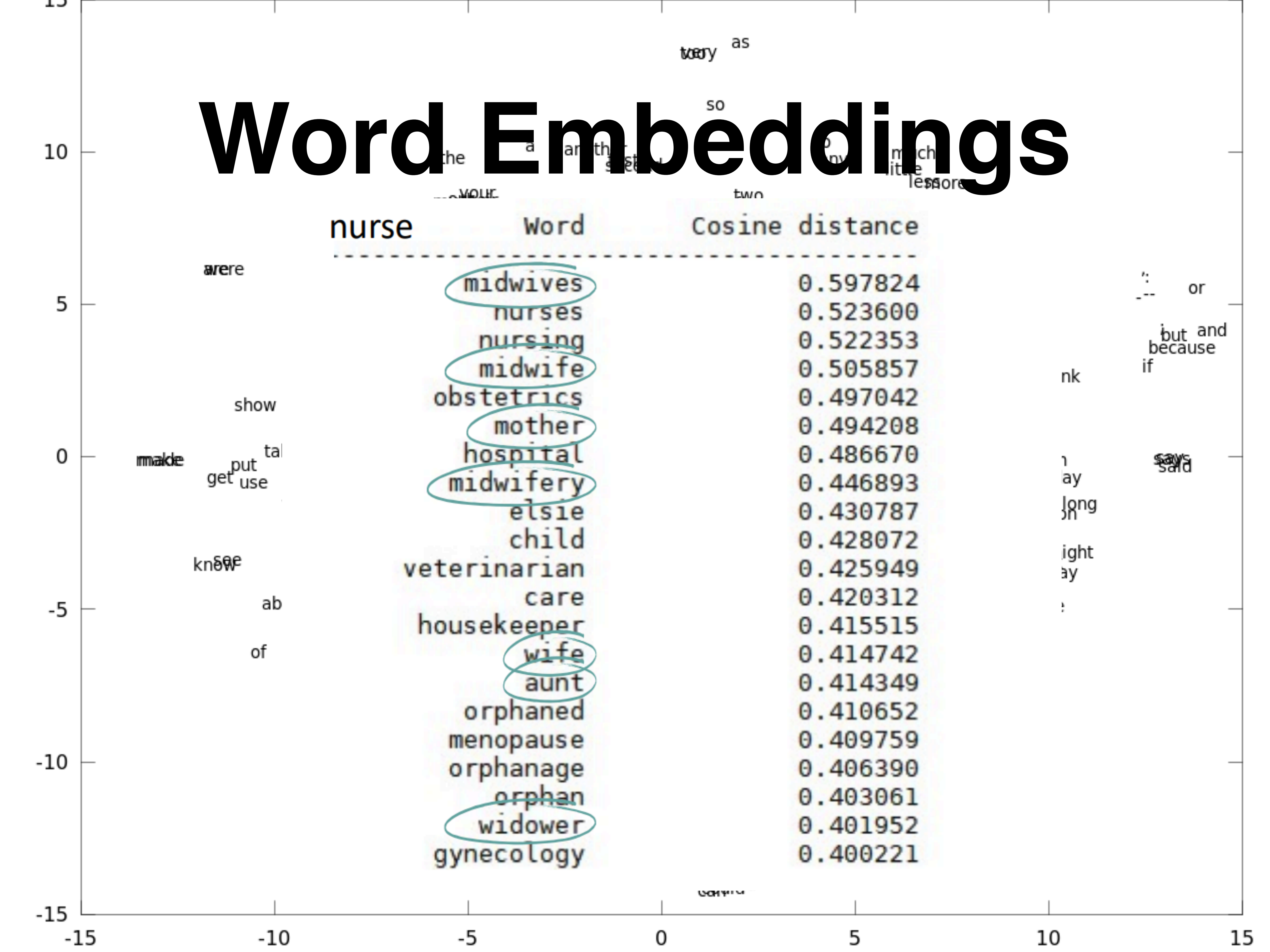
- Context-dependent Hearst patterns

What's encoded in my representation?

Word Embeddings



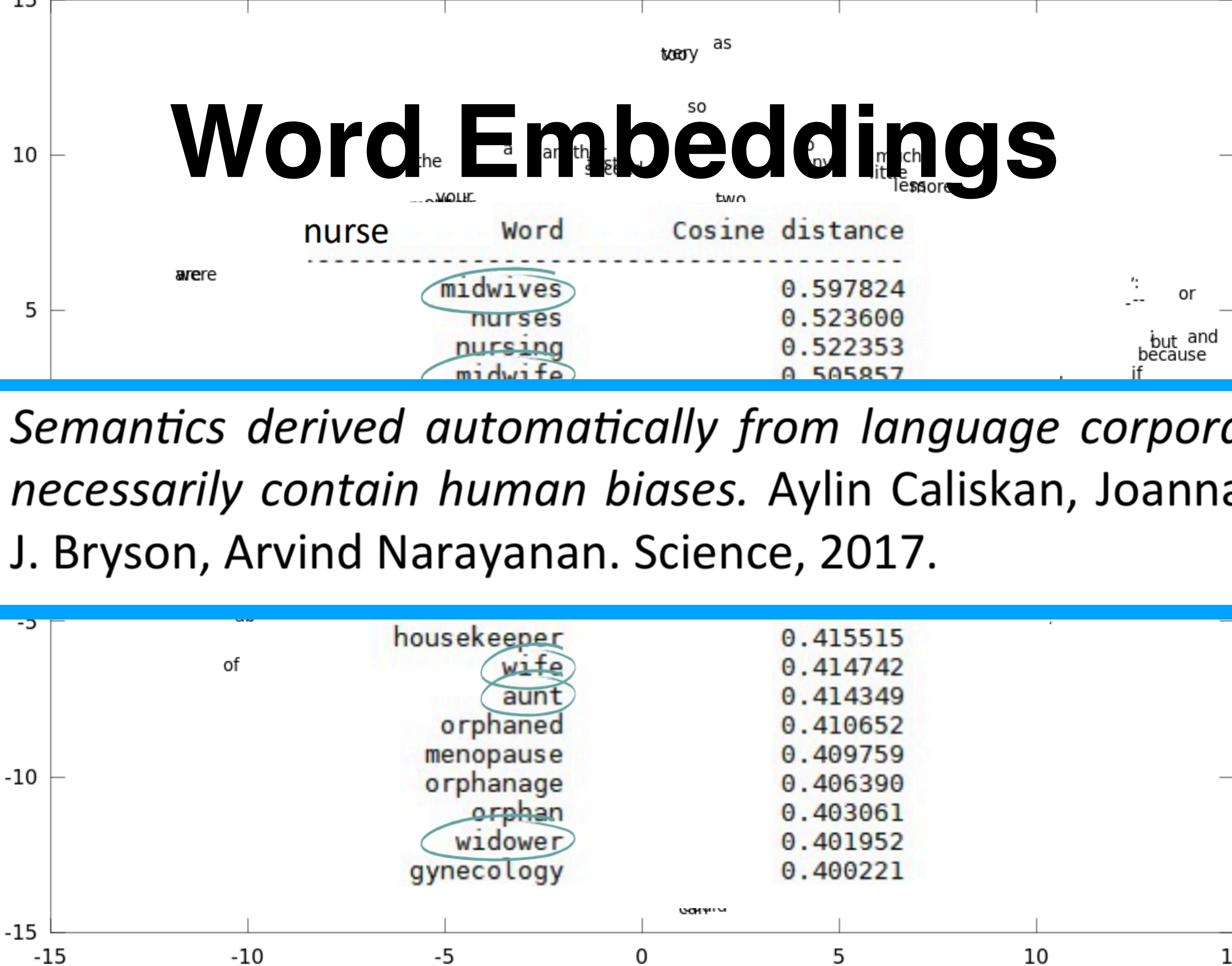
Word Embeddings



nurse

Word	Cosine distance
midwives	0.597824
nurses	0.523600
nursing	0.522353
midwife	0.505857
obstetrics	0.497042
mother	0.494208
hospital	0.486670
midwifery	0.446893
elsie	0.430787
child	0.428072
veterinarian	0.425949
care	0.420312
housekeeper	0.415515
wife	0.414742
aunt	0.414349
orphaned	0.410652
menopause	0.409759
orphanage	0.406390
orphan	0.403061
widower	0.401952
gynecology	0.400221

Word Embeddings



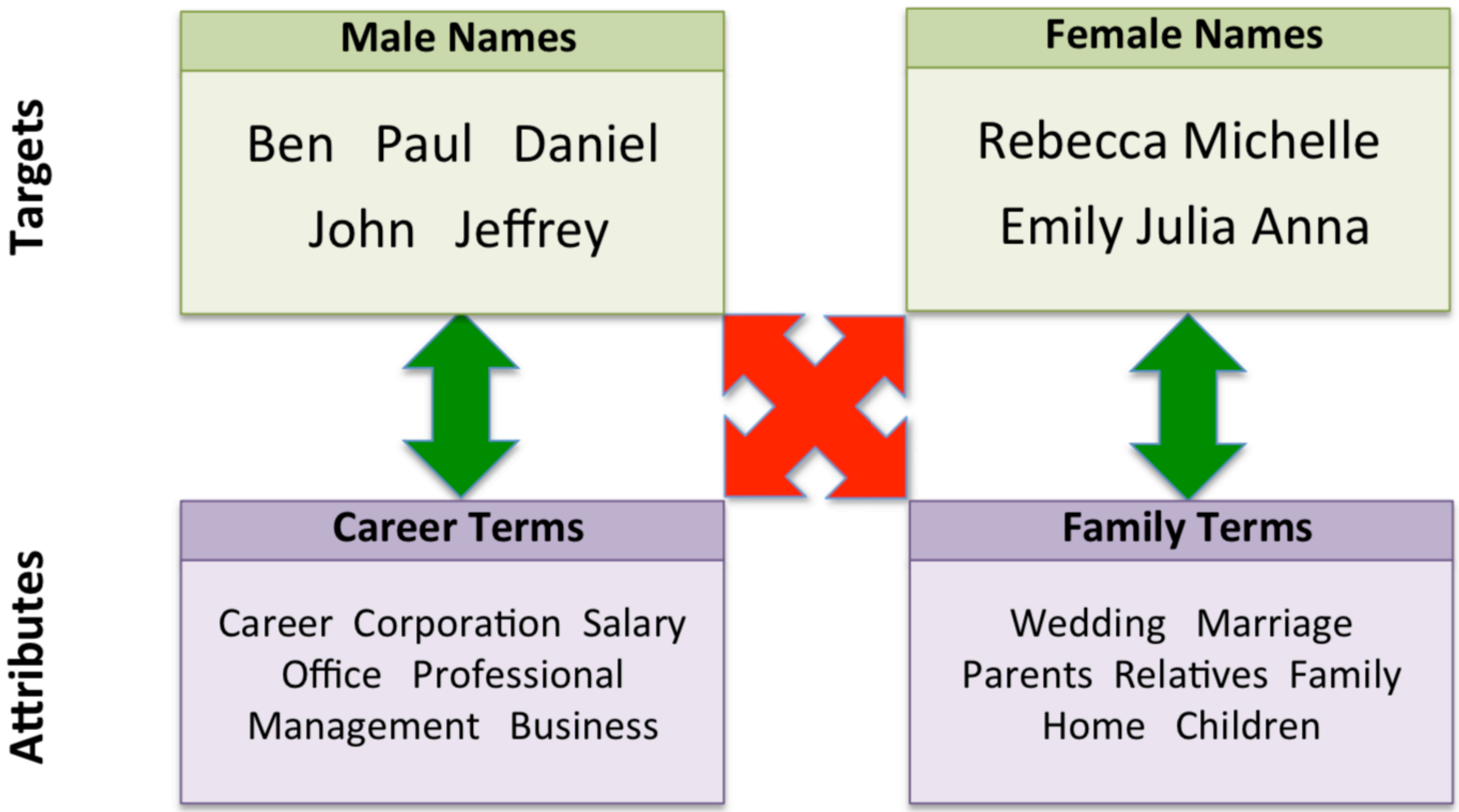
nurse

Word	Cosine distance
midwives	0.597824
nurses	0.523600
nursing	0.522353
midwife	0.505857
housekeeper	0.415515
wife	0.414742
aunt	0.414349
orphaned	0.410652
menopause	0.409759
orphanage	0.406390
orphan	0.403061
widower	0.401952
gynecology	0.400221

Semantics derived automatically from language corpora necessarily contain human biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

Implicit Association Tests (IATs)

Measuring implicit association between target and attribute concepts with reaction times.



Greenwald, Anthony G., Debbie E. McGhee, and Jordan LK Schwartz. "Measuring individual differences in implicit cognition: the implicit association test." Journal of personality and social psychology 74.6 (1998): 1464.

IAT Results Reproduced in Word Embeddings

Target words	Attrib. words	Original Finding				Our Finding			
		Ref	N	d	p	N_T	N_A	d	p
Flowers vs insects	Pleasant vs unpleasant	(*5*)	32	1.35	10^{-8}	25×2	25×2	1.50	10^{-7}
Instruments vs weapons	Pleasant vs unpleasant	(*5*)	32	1.66	10^{-10}	25×2	25×2	1.53	10^{-7}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(*5*)	26	1.17	10^{-5}	32×2	25×2	1.41	10^{-8}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (*5*)	(*7*)	Not applicable			16×2	25×2	1.50	10^{-4}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (*9*)	(*7*)	Not applicable			16×2	8×2	1.28	10^{-3}
Male vs female names	Career vs family	(*9*)	$39k$	0.72	$< 10^{-2}$	8×2	8×2	1.81	10^{-3}
Math vs arts	Male vs female terms	(*9*)	$28k$	0.82	$< 10^{-2}$	8×2	8×2	1.06	.018
Science vs arts	Male vs female terms	(*10*)	91	1.47	10^{-24}	8×2	8×2	1.24	10^{-2}
Mental vs physical disease	Temporary vs permanent	(*23*)	135	1.01	10^{-3}	6×2	7×2	1.38	10^{-2}
Young vs old people's names	Pleasant vs unpleasant	(*9*)	$43k$	1.42	$< 10^{-2}$	8×2	8×2	1.21	10^{-2}

N = # participants
N_T = # target words
N_A = # attribute words
d = effect size
p = p-value

Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

Gender Bias in Word Embedding Correlates with Real-World Gender Bias

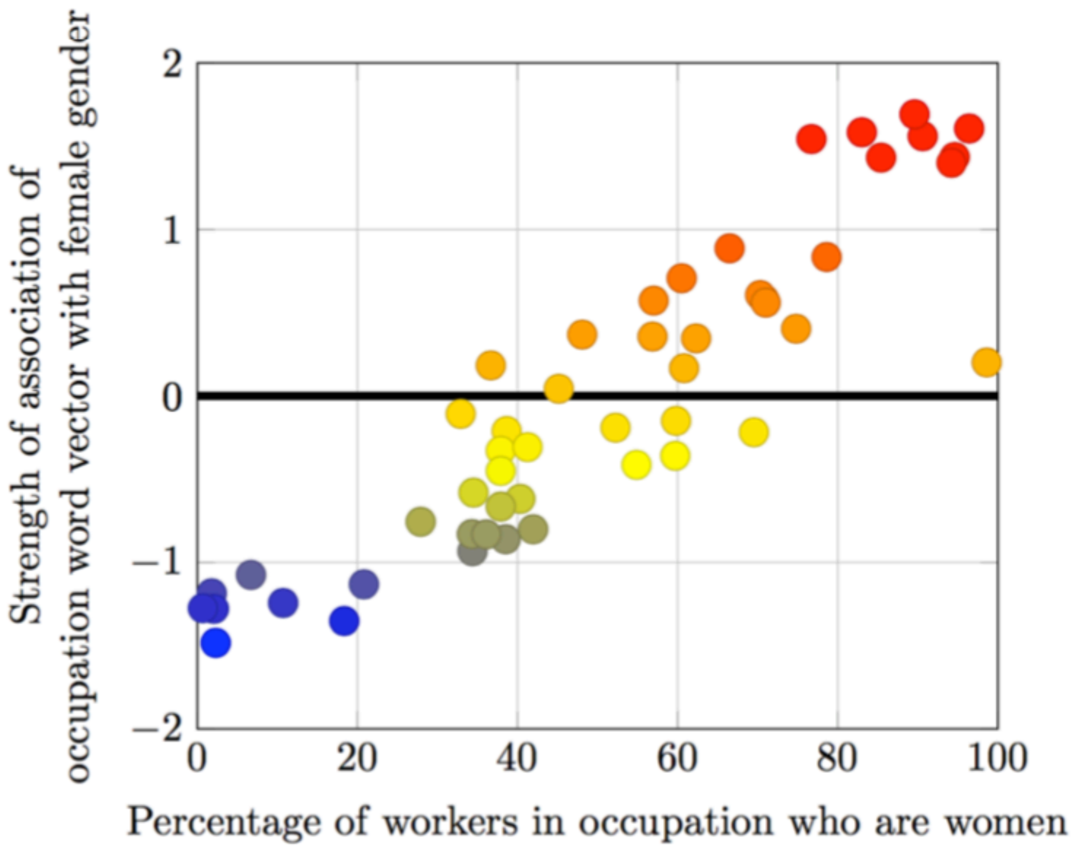


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

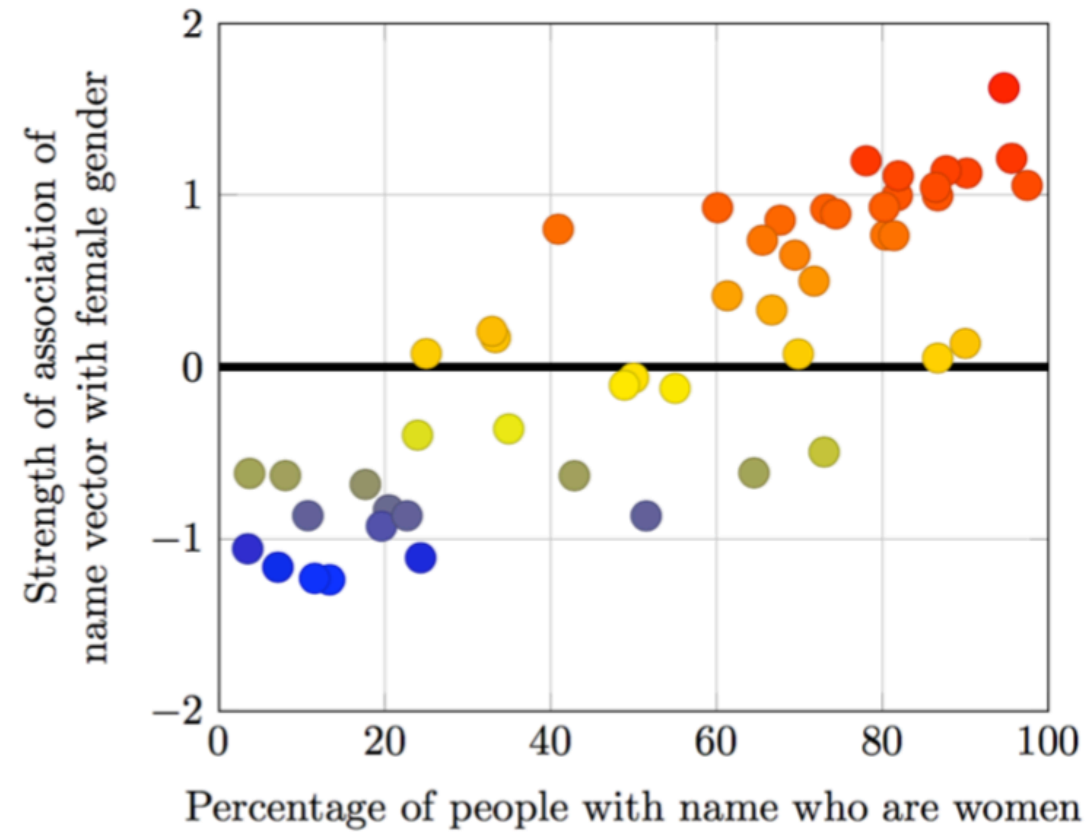


Figure 2: Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with p-value $< 10^{-13}$.

Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

Gender Bias in Word Embedding Correlates with Real-World Gender Bias

Semantics derived automatically from language corpora contain human-like biases. Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan. Science, 2017.

controlling the representations?

Debiasing word embeddings

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

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]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

Learning Gender-Neutral Word Embeddings

Jieyu Zhao **Yichao Zhou** **Zeyu Li** **Wei Wang** **Kai-Wei Chang**
University of California, Los Angeles
{jyzhao, yz, zyli, weiwang, kwchang}@cs.ucla.edu

Debiasing word embeddings

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

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]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

projection
on "he - she"
(gender direction)

$$\text{bias}(w) = \overrightarrow{w} \cdot \overrightarrow{he} - \overrightarrow{w} \cdot \overrightarrow{she} = \overrightarrow{w} \cdot (\overrightarrow{he} - \overrightarrow{she})$$

* This is the gender direction, can be computed using several pairs together (e.g. man-woman, brother-sister)

Debiasing word embeddings

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

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]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

projection on "he - she" (gender direction)

$$\vec{w} := (\vec{w} - \vec{w}_B) / \|\vec{w} - \vec{w}_B\|$$

$\vec{w}_B$ — Projection of w on gender direction

- The bias of all neutral words is now zero by definition

Debiasing word embeddings

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

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]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

not so easy!

$$\vec{w} := (\vec{w} - \vec{w}_B) / \|\vec{w} - \vec{w}_B\|$$

$\vec{w}_B -$ Projection of w on gender direction

- The bias of all neutral words is now zero by definition

Debiasing word embeddings

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

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]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

not so easy!

Learning Gender-Neutral Word Embeddings

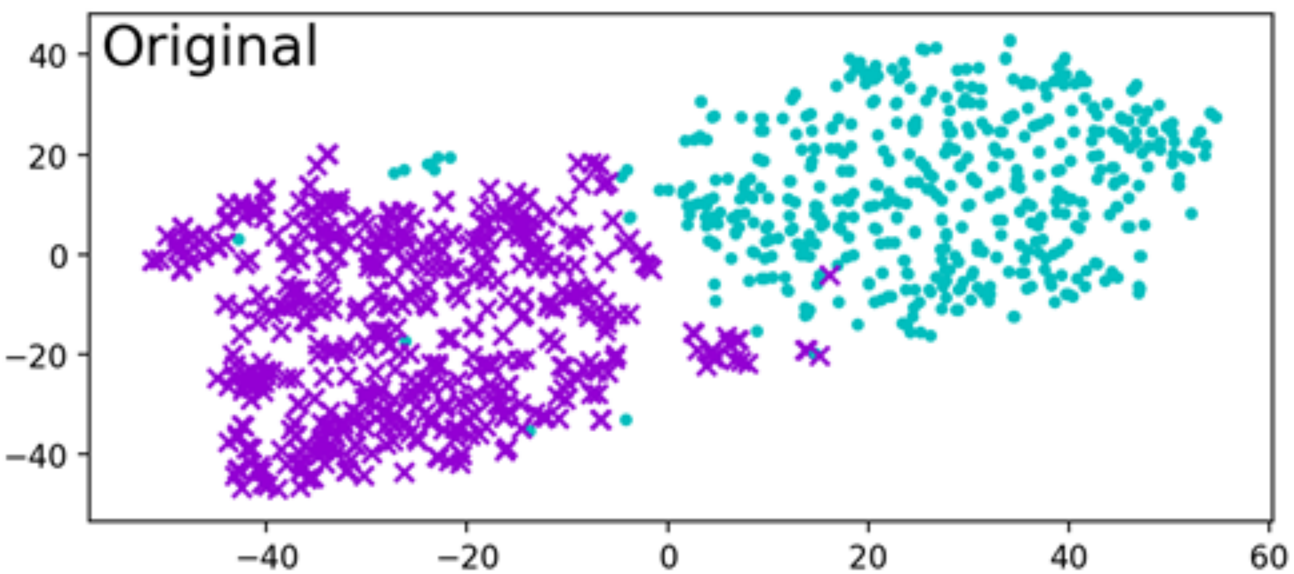
Jieyu Zhao **Yichao Zhou** **Zeyu Li** **Wei Wang** **Kai-Wei Chang**
University of California, Los Angeles
{jyzhao, yz, zyli, weiwang, kwchang}@cs.ucla.edu

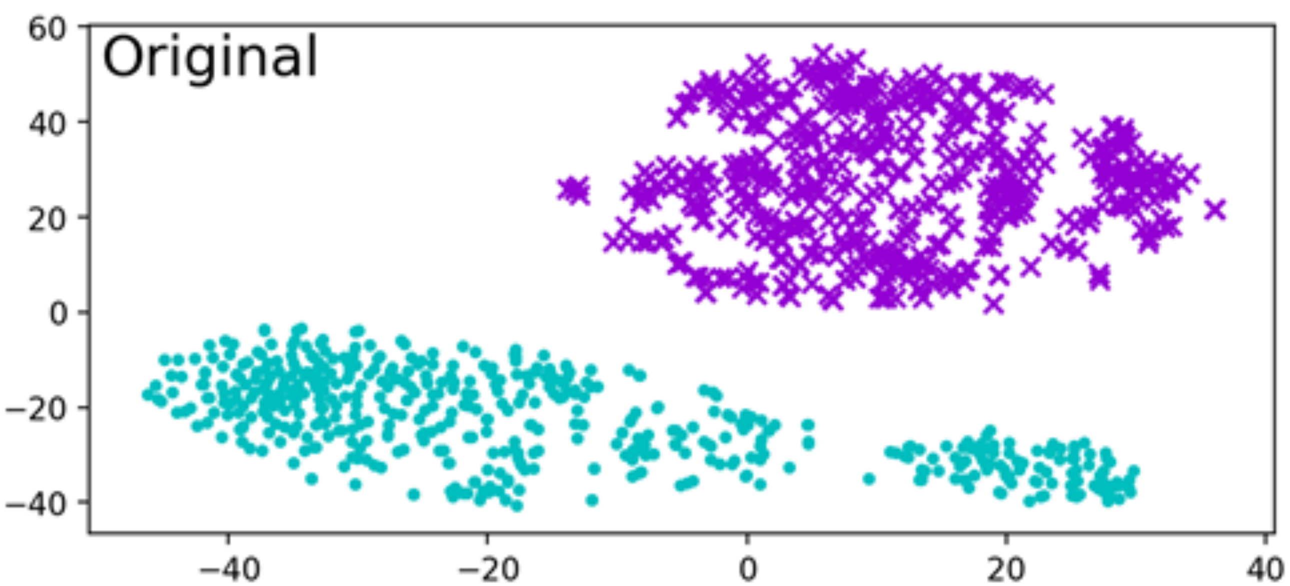
Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com





T-SNE, then color by gender (using the gender direction definition)

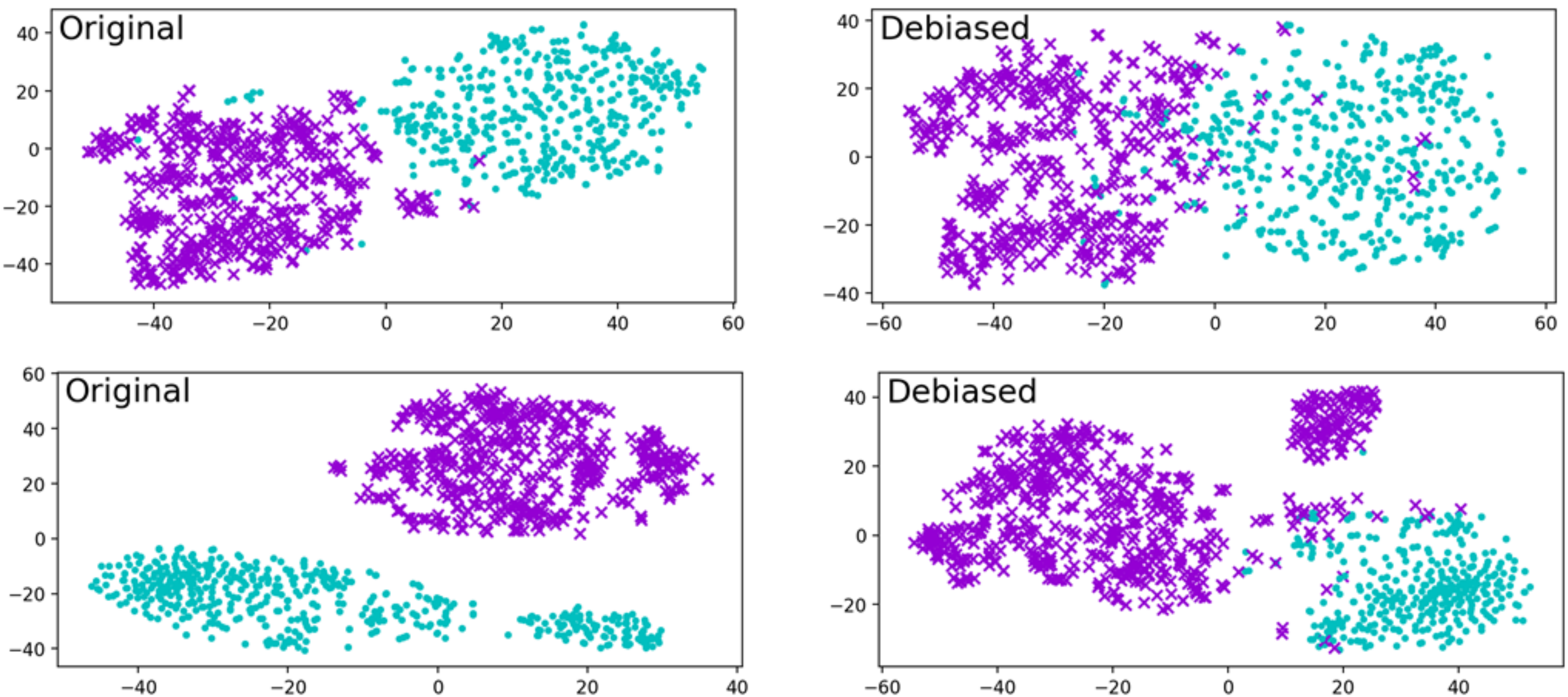
Lipstick on a Pig:
Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]

[1]Department of Computer Science, Bar-Ilan University

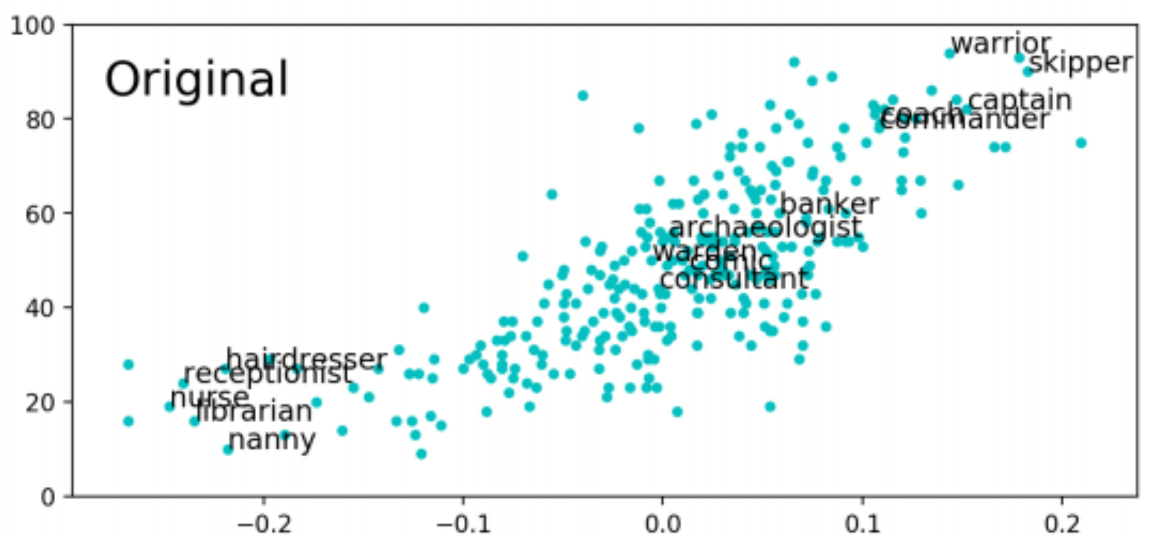
[2]Allen Institute for Artificial Intelligence

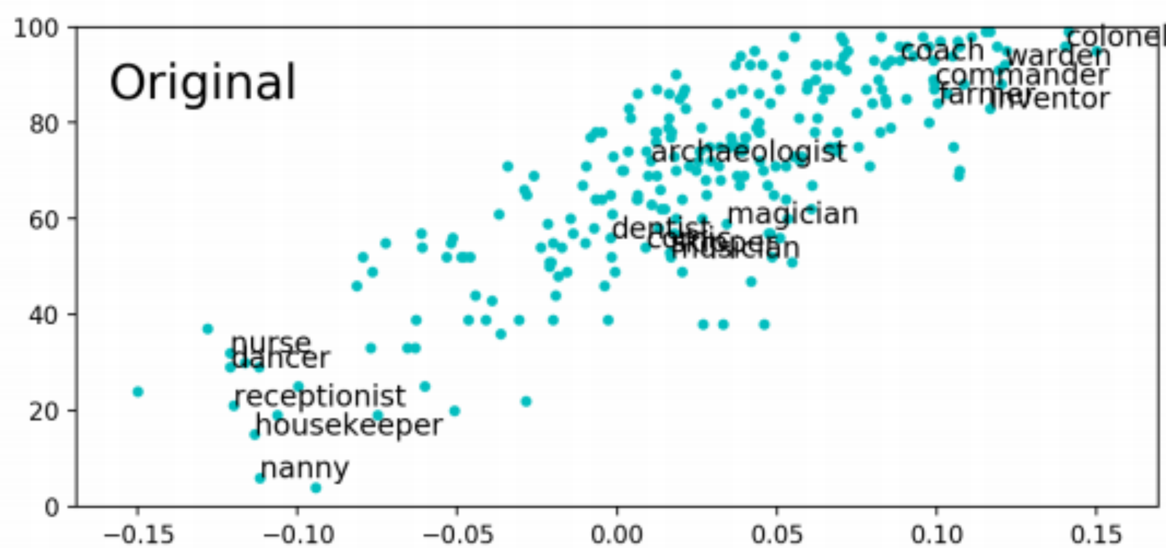
{hilagnn,yoav.goldberg}@gmail.com



Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

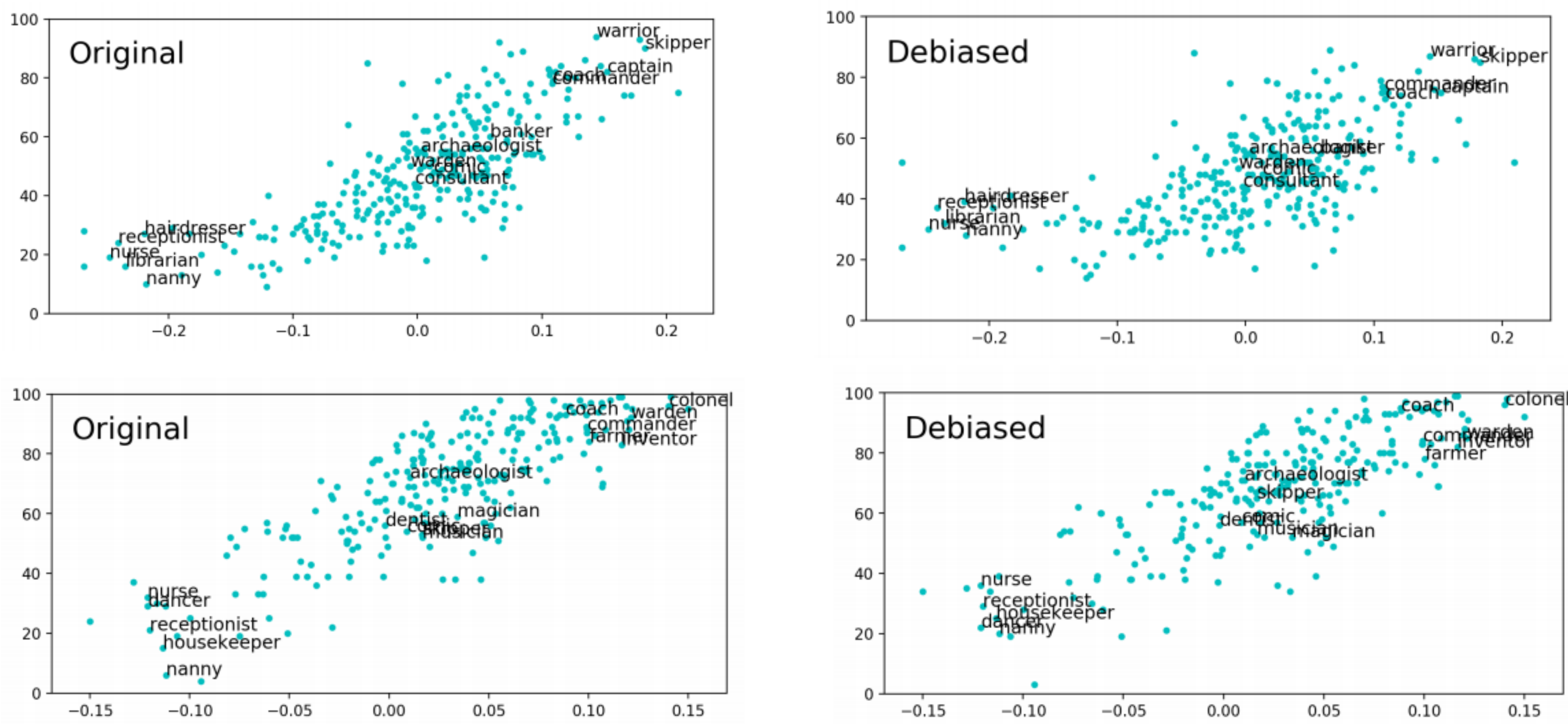




new metric: how many of my neighbors are male / female leaning?

Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

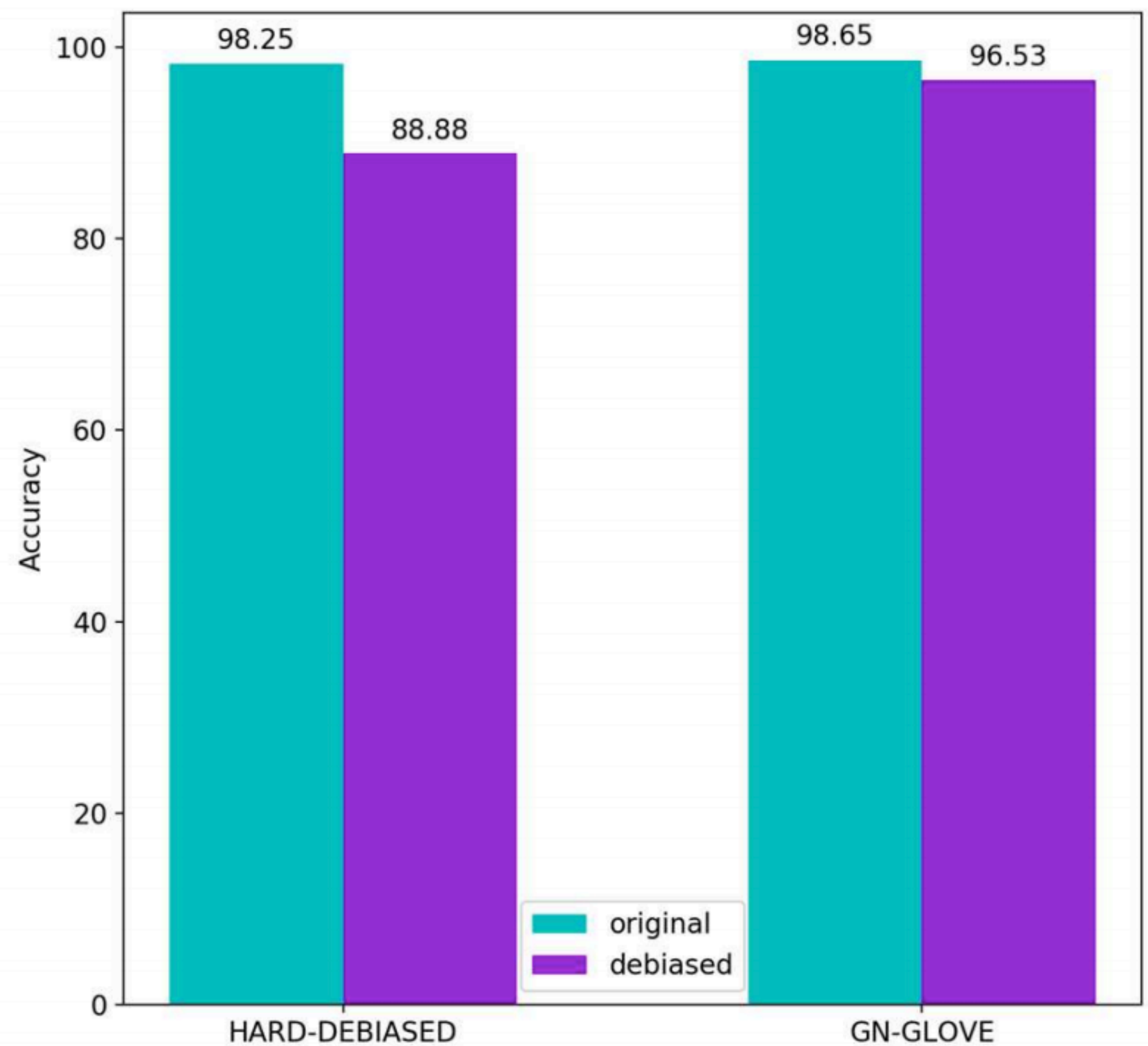
Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com



Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

YES.



Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

So what happened here?

1. define a way to measure a problem.

2. confuse the measurement of the phenomena with the phenomena.

3. design a way to treat the phenomena (actually, attack the measurement)

4. can no longer measure the phenomena (all measures are 0). Problem solved?

Word embeddings in gender marking languages

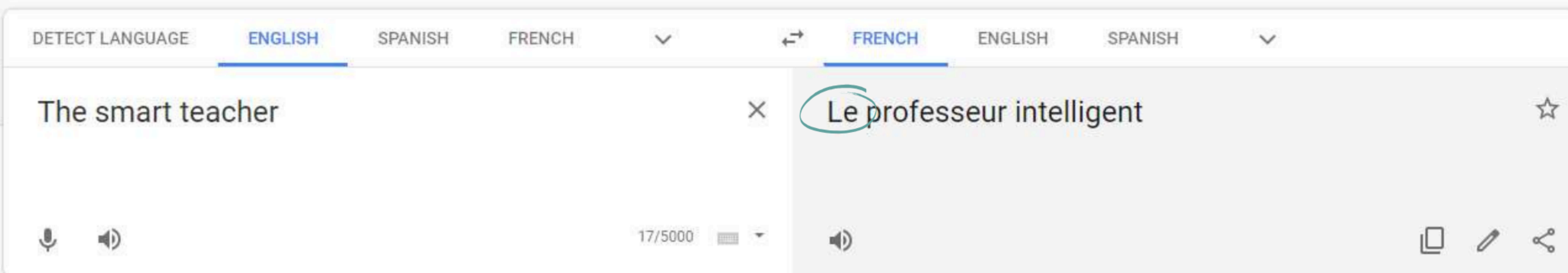
How does Grammatical Gender Affect Noun Representations in Gender-Marking Languages?

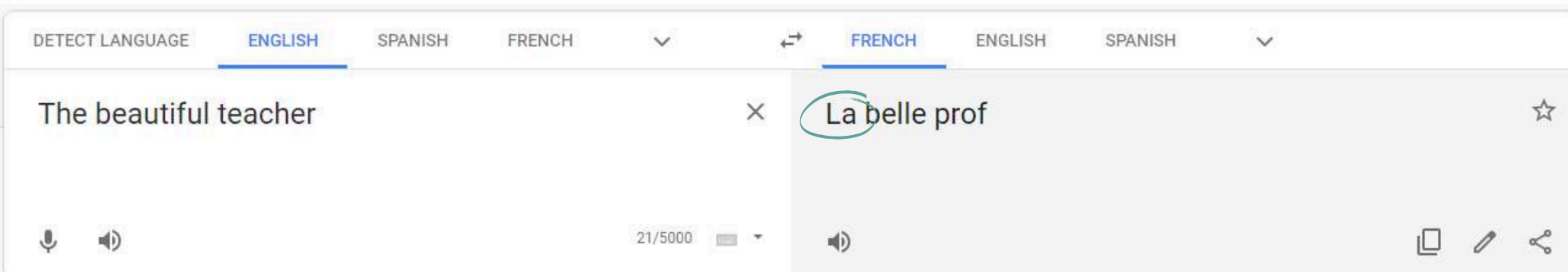
Hila Gonen[1] **Yova Kementchedjhieva**[2] **Yoav Goldberg**[1,3]
[1]Department of Computer Science, Bar-Ilan University
[2]University of Copenhagen
[3]Allen Institute for Artificial Intelligence
hilagnn@gmail.com, yova@di.ku.dk, yoav.goldberg@gmail.com

a whole new complex story. not in this talk. check out the paper.

taking a step back

gender-based examples are easy to find.





taking a step back

Gender-based examples are easy to find.

But the problem goes far beyond gender (or race, or age).

Models make many decisions based on various factors that we do not understand, with subtle interactions and the most non-transparent mechanism imaginable.

These models then ACT in the real world.

taking a step back

Gender-based examples are easy to find.

Bu **e).**

It is our responsibility to consider the consequences, and be careful about what we do, especially when we build "production" systems, but also when we "just do research".

M **s**

a **.**

These models then ACT in the real world.

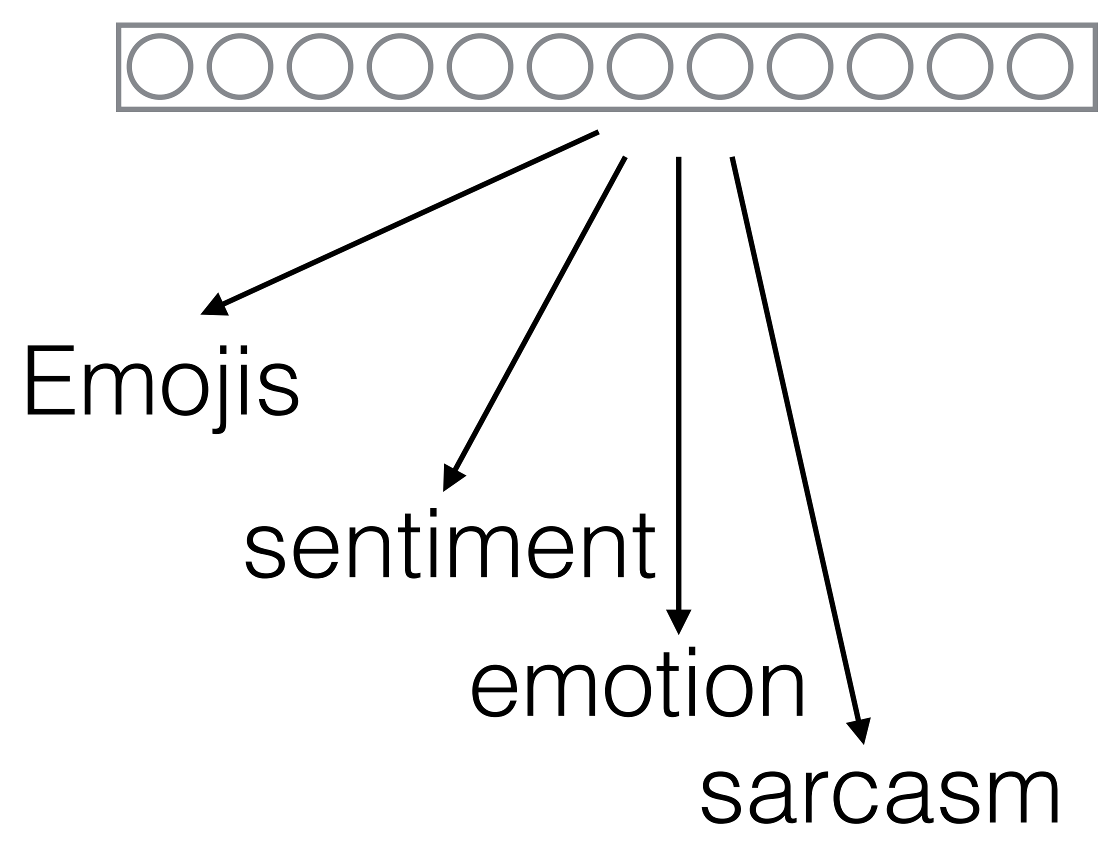
Beyond word embeddings

Example: DeepMoji

Train a model to predict emojis from tweets

Vectors are also predictive of related tasks

DeepMoji



Example: DeepMoji

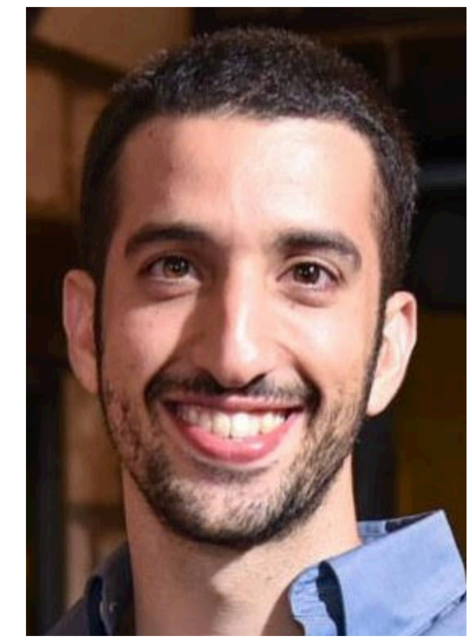
Adversarial Removal of Demographic Attributes from Text Data

Yanai Elazar[†] and **Yoav Goldberg**[†*]

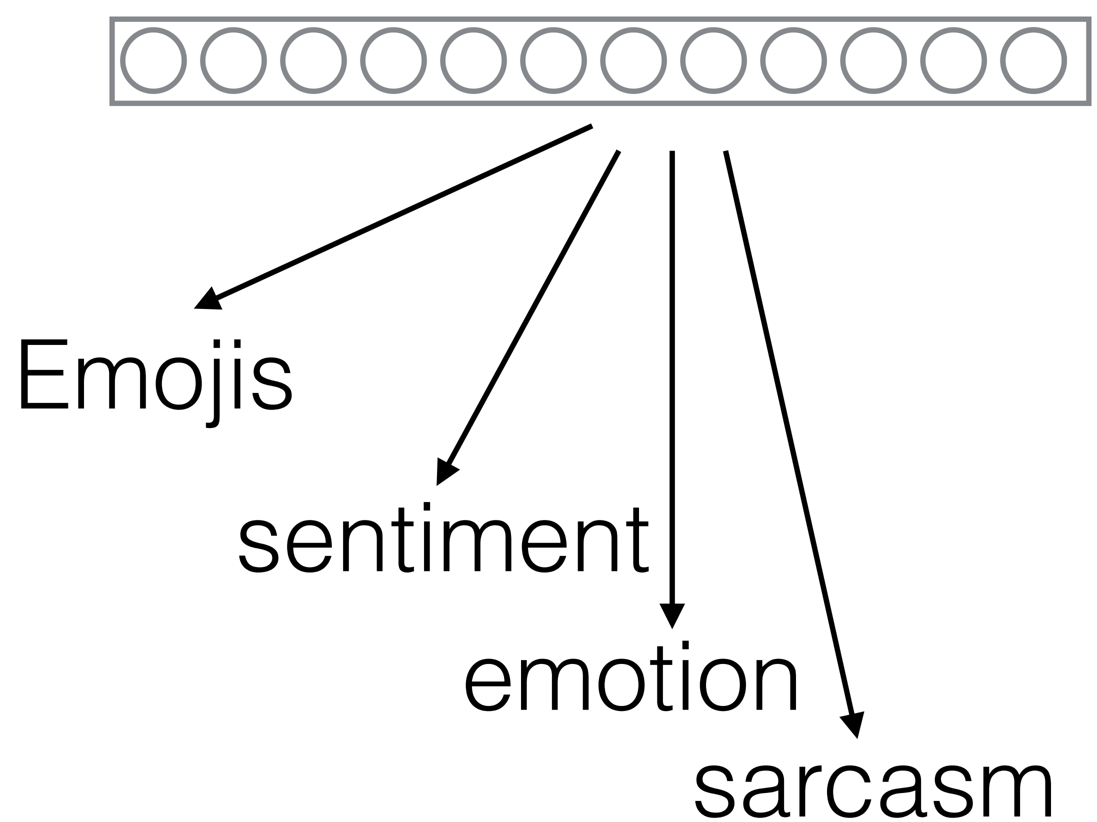
[†]Computer Science Department, Bar-Ilan University, Israel

[*]Allen Institute for Artificial Intelligence

{yanaiela,yoav.goldberg}@gmail.com



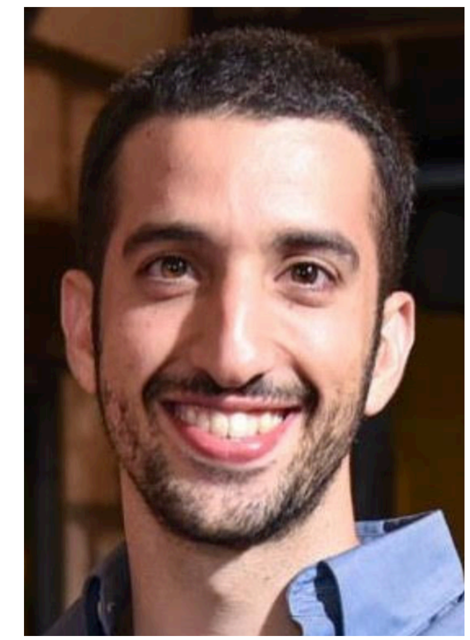
DeepMoji



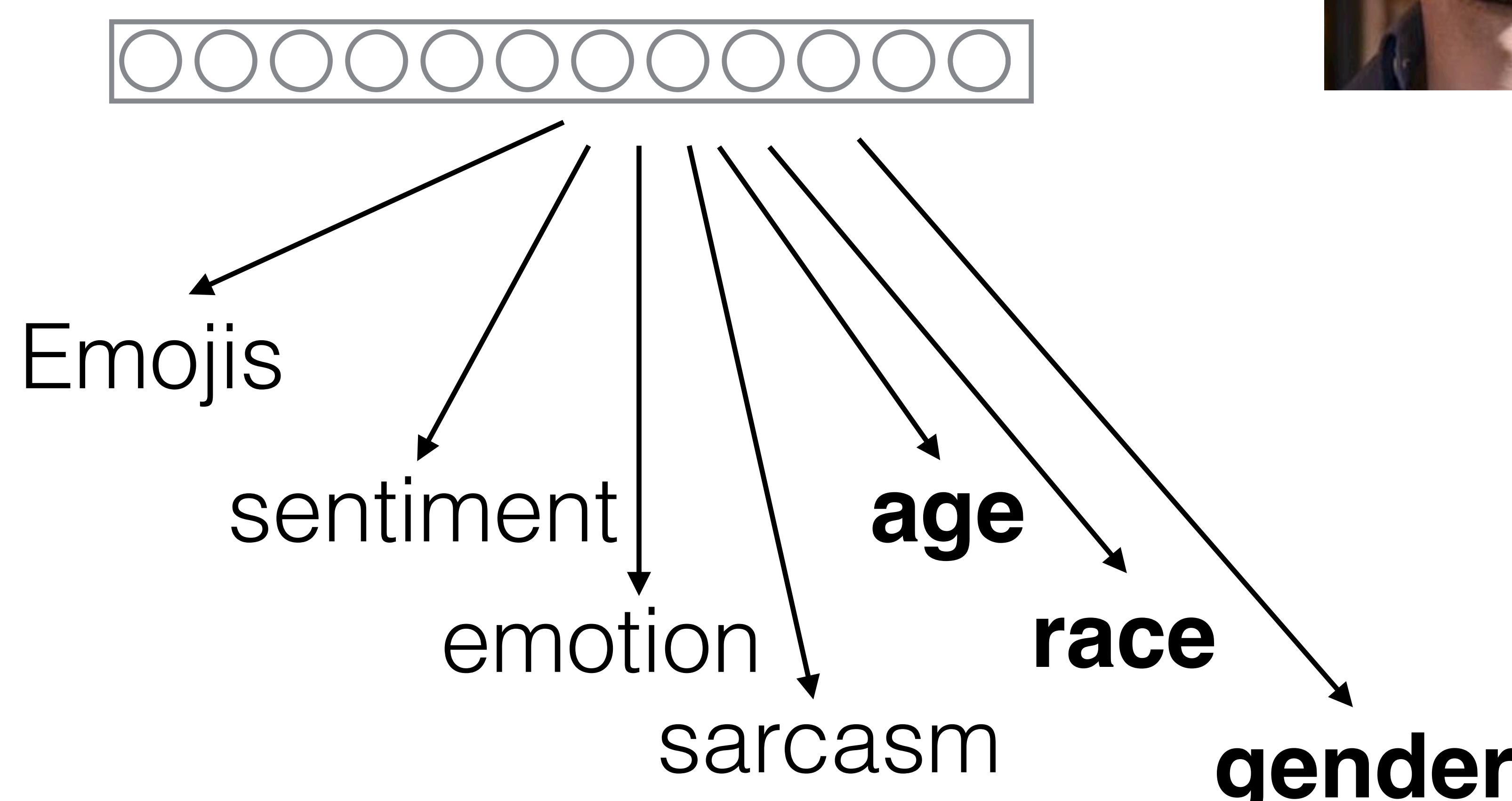
Example: DeepMoji

Adversarial Removal of Demographic Attributes from Text Data

Yanai Elazar[†] and **Yoav Goldberg**[†*]
[†]Computer Science Department, Bar-Ilan University, Israel
[*]Allen Institute for Artificial Intelligence
{yanaiela,yoav.goldberg}@gmail.com



DeepMoji

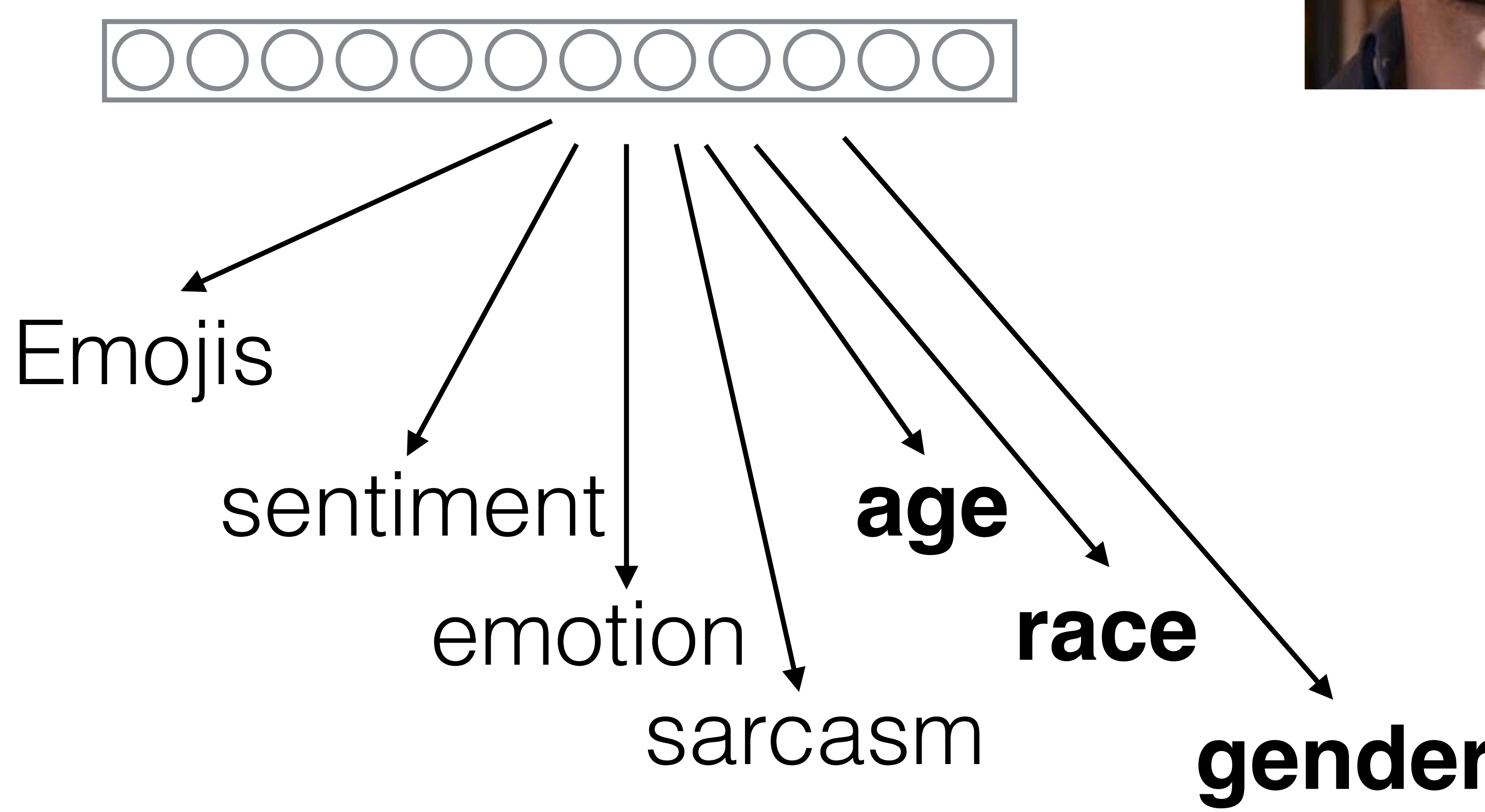


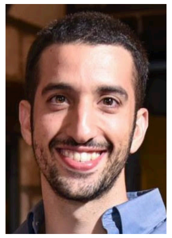
Vectors trained for Emojis.

Meant for sentiment.

Predictive of demographics.

DeepMoji





Vectors trained for Emojis.

Meant for sentiment.

Predictive of demographics.

who could have guessed?

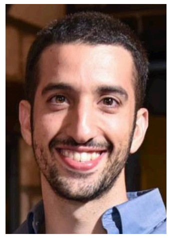
Vectors trained for Emojis.

Meant for sentiment.

Predictive of demographics.

who could have guessed?

well... not very surprising actually.
emoji usage is very much correlated with demographics.
knowing the demographics helps predict emojis.



Vectors trained for Emojis.

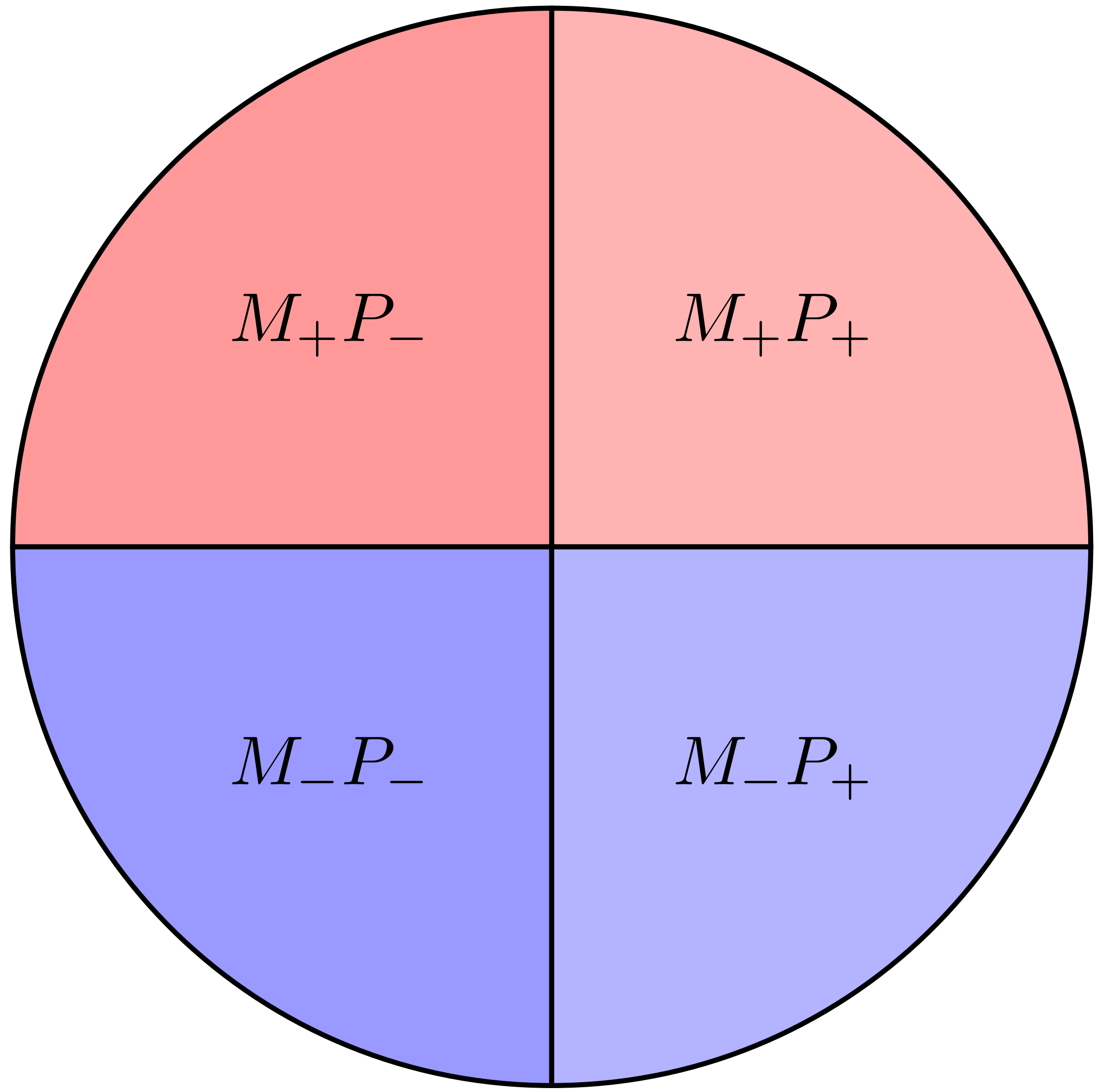
Meant for sentiment.

Predictive of demographics.

who could have guessed?

well... not very surprising actually.
emoji usage is very much correlated with demographics.
knowing the demographics helps predict emojis.

Lets control for this.

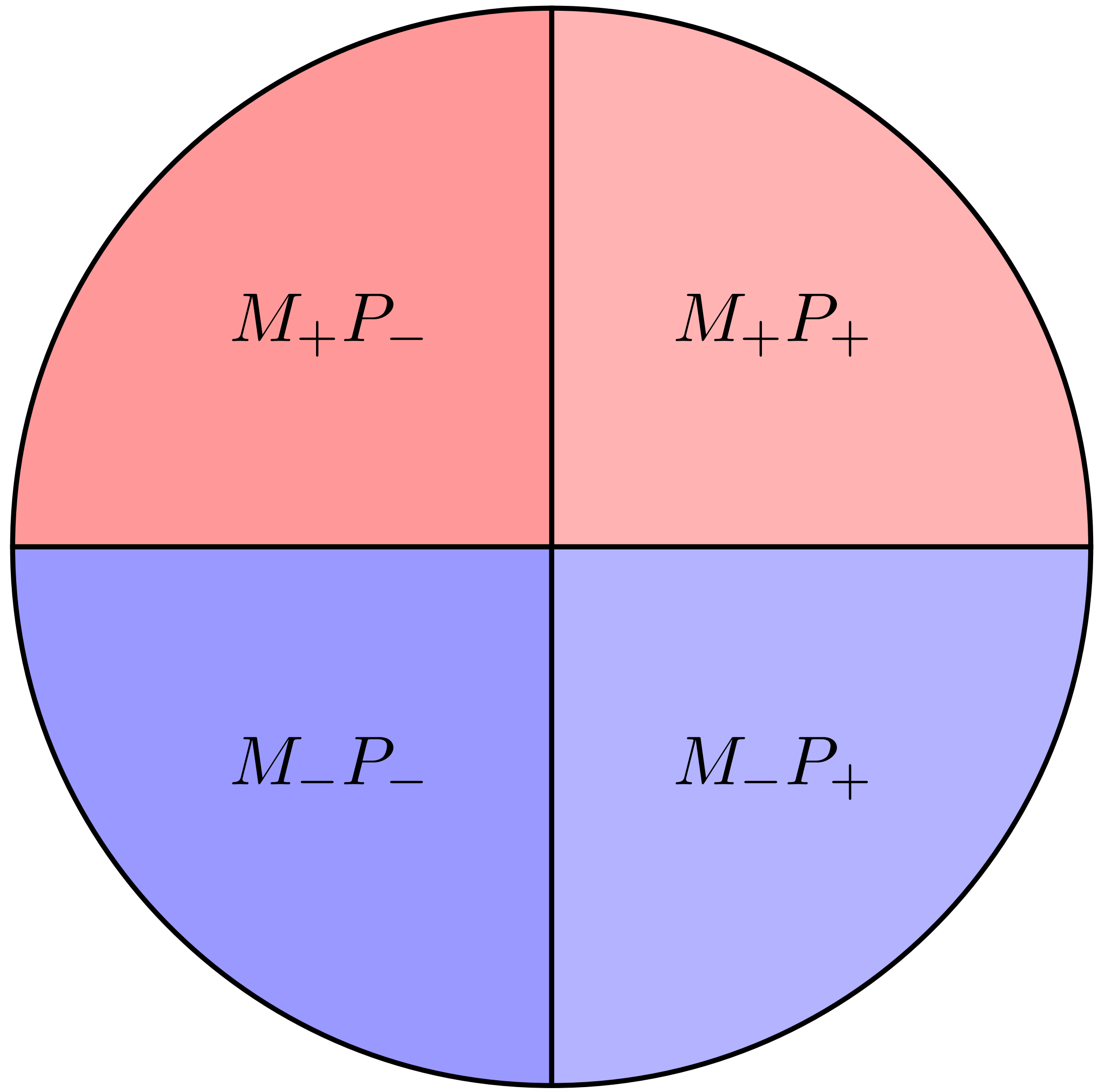


balanced dataset

50%
positive

task
(sentiment)

50%
negative



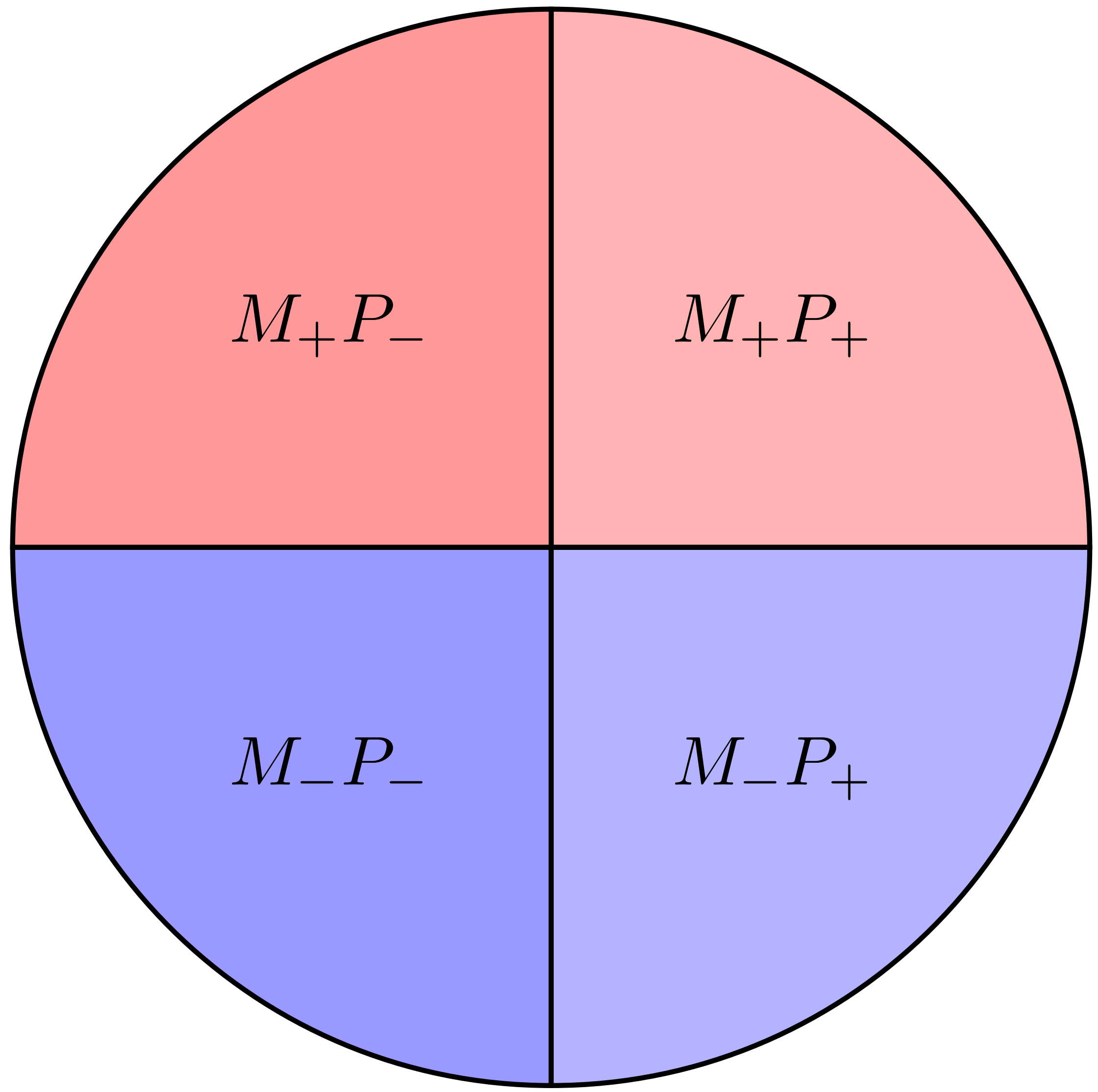
balanced
dataset

demographics

50% male   50% female

50% positive

task (sentiment)



50% negative

balanced dataset

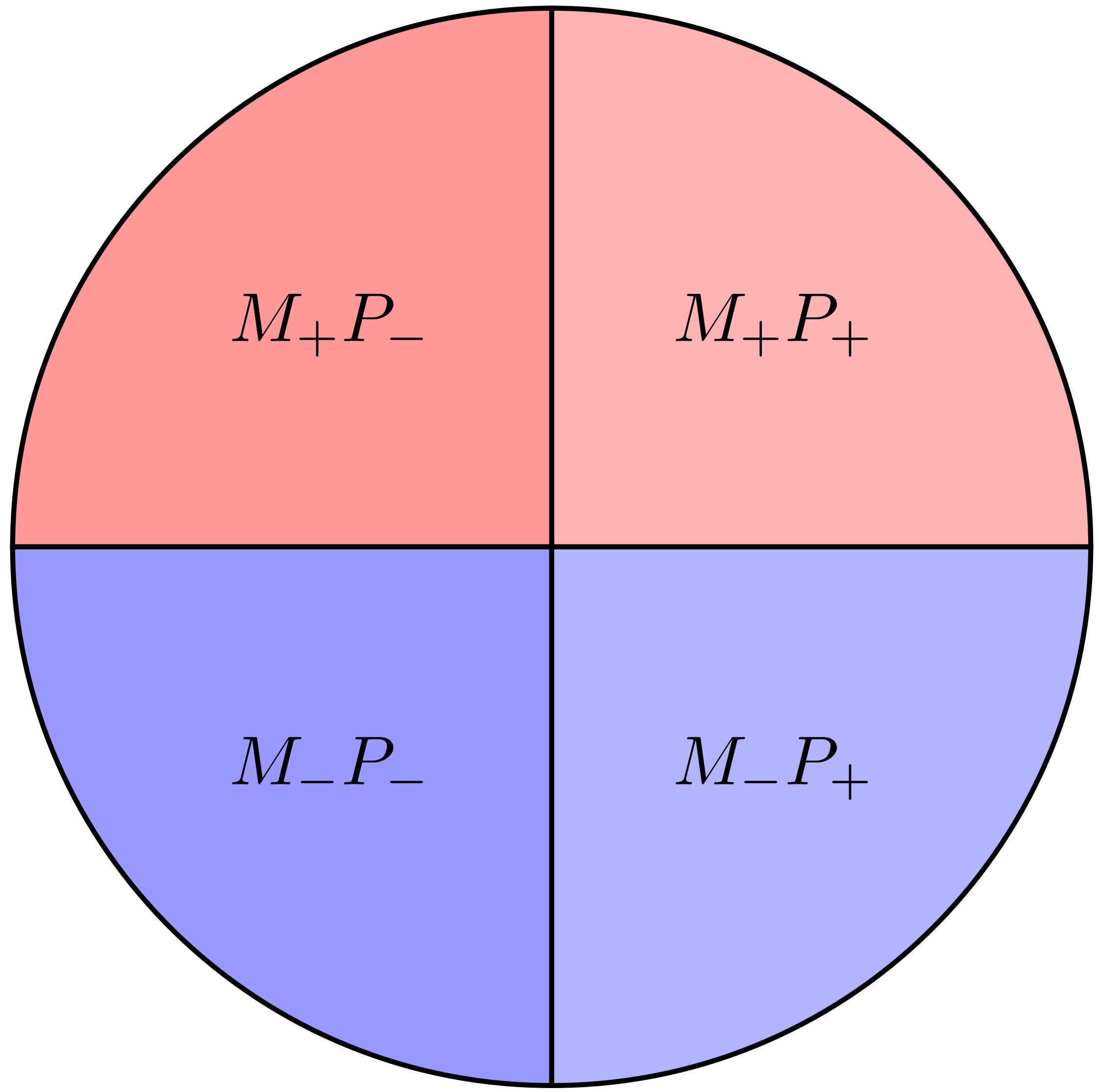
demographics

50% male 50% female

50%
positive

task
(sentiment)

50%
negative

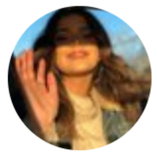


balanced dataset

no correlation between task and demographic attribute

B I U
N L P

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

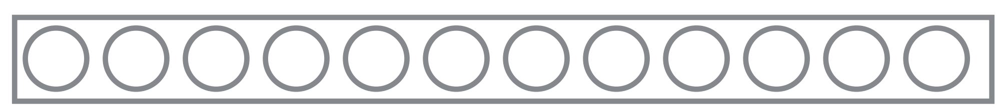


Hannah Whitlock
@Hannahwhitlock_

Follow

YALL. LOOK HOW BEAUTIFUL MY BFF IS OH MY GOODNESS

encode



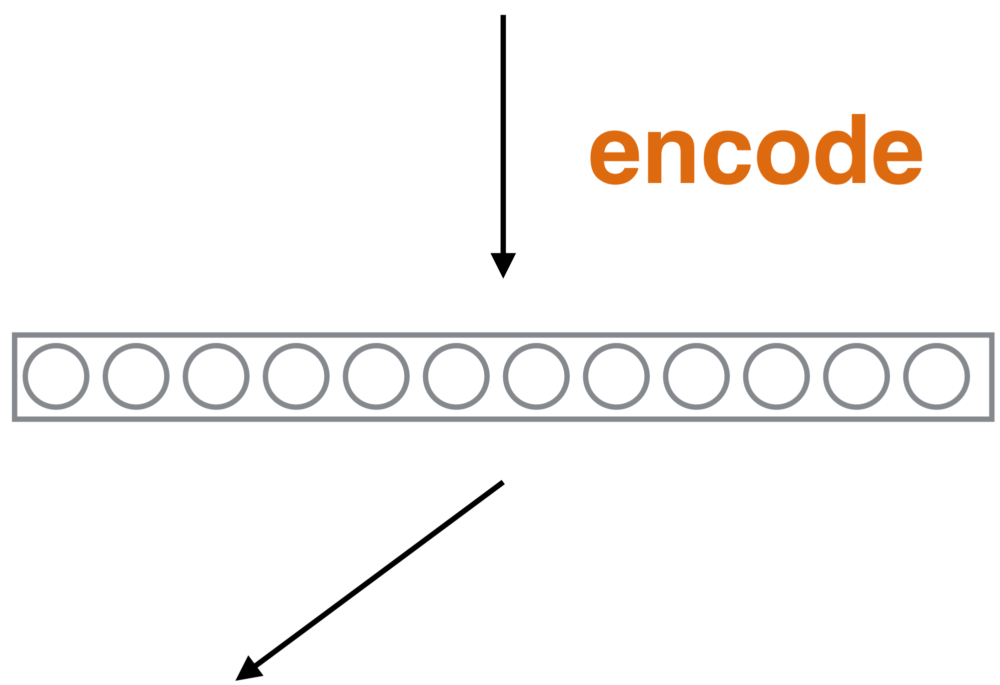
train to predict main task

no correlation between task and demographic attribute

Hannah Whitlock
@Hannahwhitlock_

Follow

YALL. LOOK HOW BEAUTIFUL MY BFF IS OH MY GOODNESS

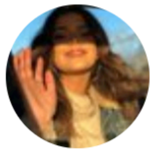


main task

no correlation between task and demographic attribute

BIU NLP

AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

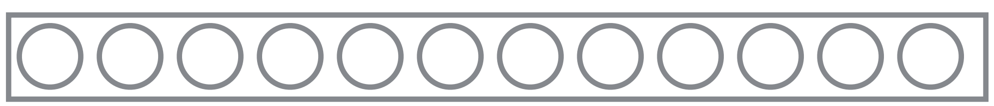


Hannah Whitlock
@Hannahwhitlock_

Follow

YALL. LOOK HOW BEAUTIFUL MY BFF IS OH MY GOODNESS

encode



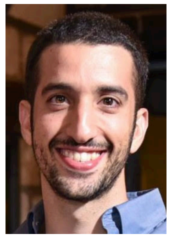
train a classifier

main task

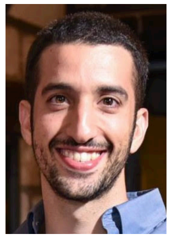
age **race** **gender**

no correlation between task and demographic attribute

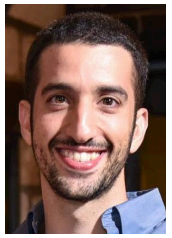
but we can still predict it with 60-70% accuracy



controlling the representation?



- We trained a classifier for task Y.
- We obtained a representation which can predict Z.
- **What if we don't want to condition on Z?**



Adversarial Training

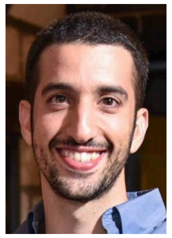
Domain-Adversarial Training of Neural Networks

Yaroslav Ganin GANIN@SKOLTECH.RU
Evgeniya Ustinova EVGENIYA.USTINOVA@SKOLTECH.RU
Skolkovo Institute of Science and Technology (Skoltech)
Skolkovo, Moscow Region, Russia

Hana Ajakan HANA.AJAKAN.1@ULAVAL.CA
Pascal Germain PASCAL.GERMAIN@IFT.ULAVAL.CA
Département d'informatique et de génie logiciel, Université Laval
Québec, Canada, G1V 0A6

Adversarial Training

Mitigating Unwanted Biases with Adversarial Learning



Brian Hu Zhang
Stanford University
Stanford, CA
bhz@stanford.edu

Blake Lemoine
Google
Mountain View, CA
lemoine@google.com

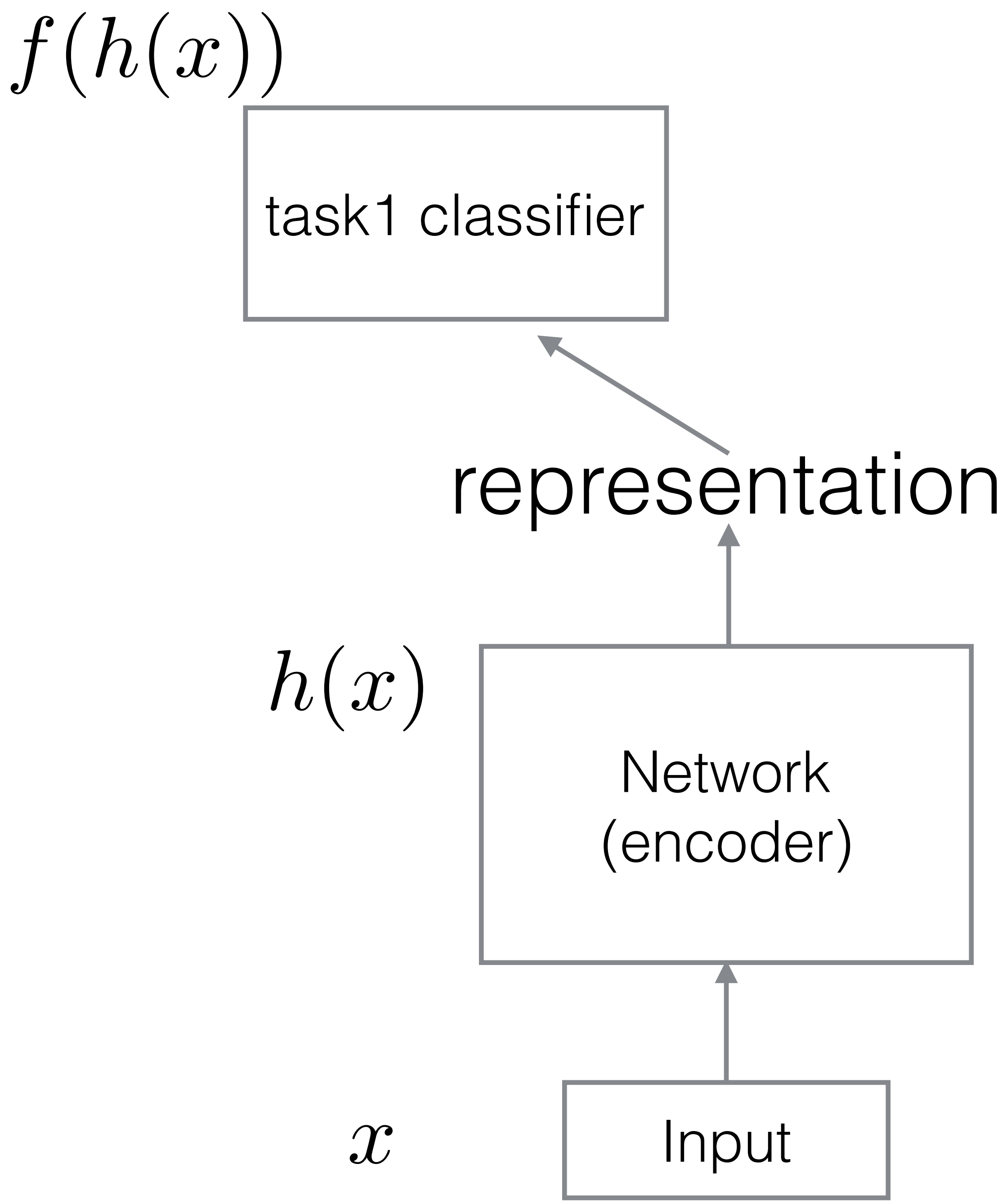
Margaret Mitchell
Google
Mountain View, CA
mmitchellai@google.com

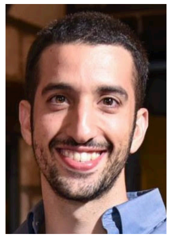
Adversarial Training

Towards Robust and Privacy-preserving Text Representations

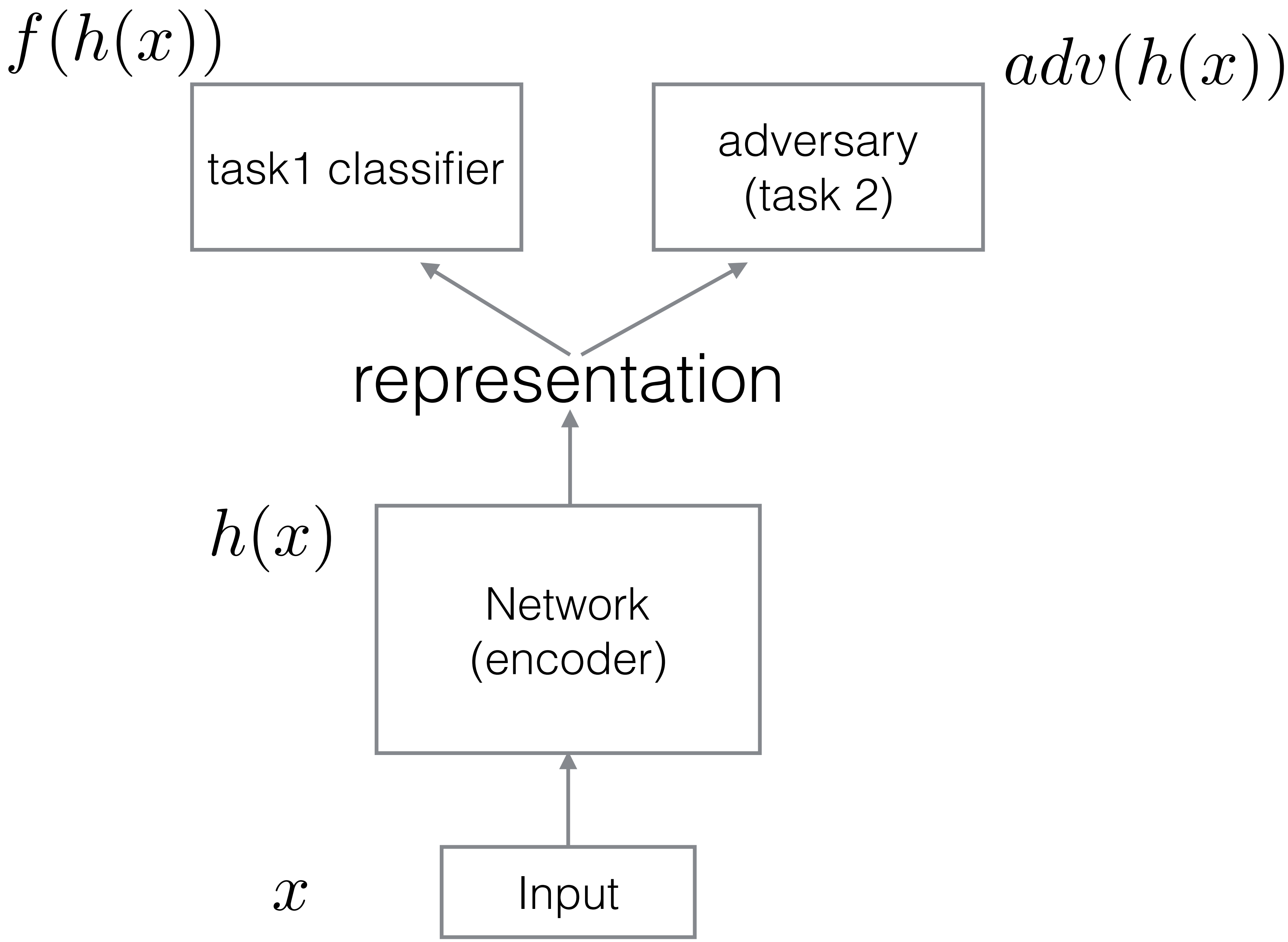
Yitong Li **Timothy Baldwin** **Trevor Cohn**
School of Computing and Information Systems
The University of Melbourne, Australia
yitongl4@student.unimelb.edu.au
{tbaldwin,tcohn}@unimelb.edu.au

predict sentiment

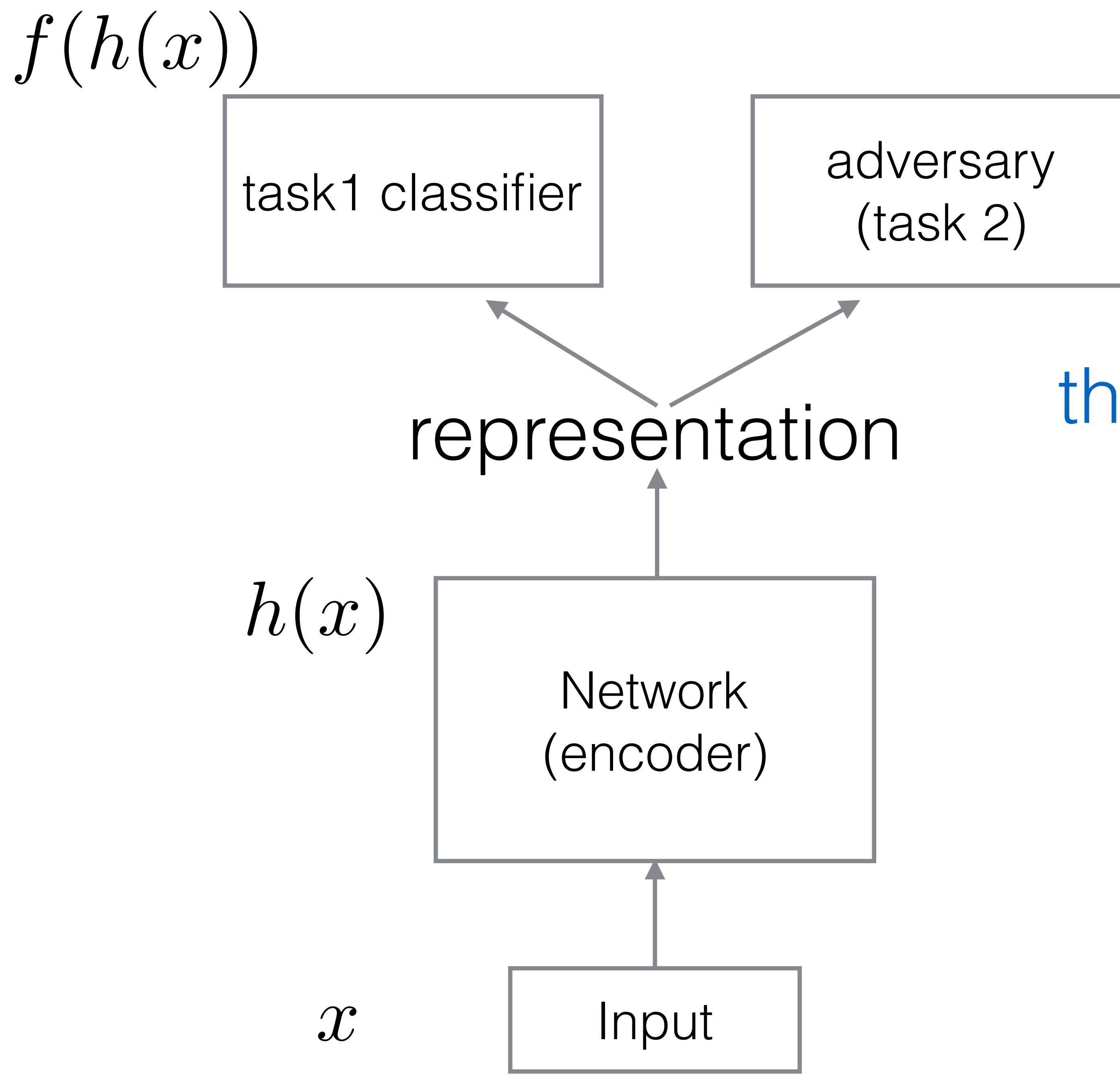




predict sentiment **don't predict race**



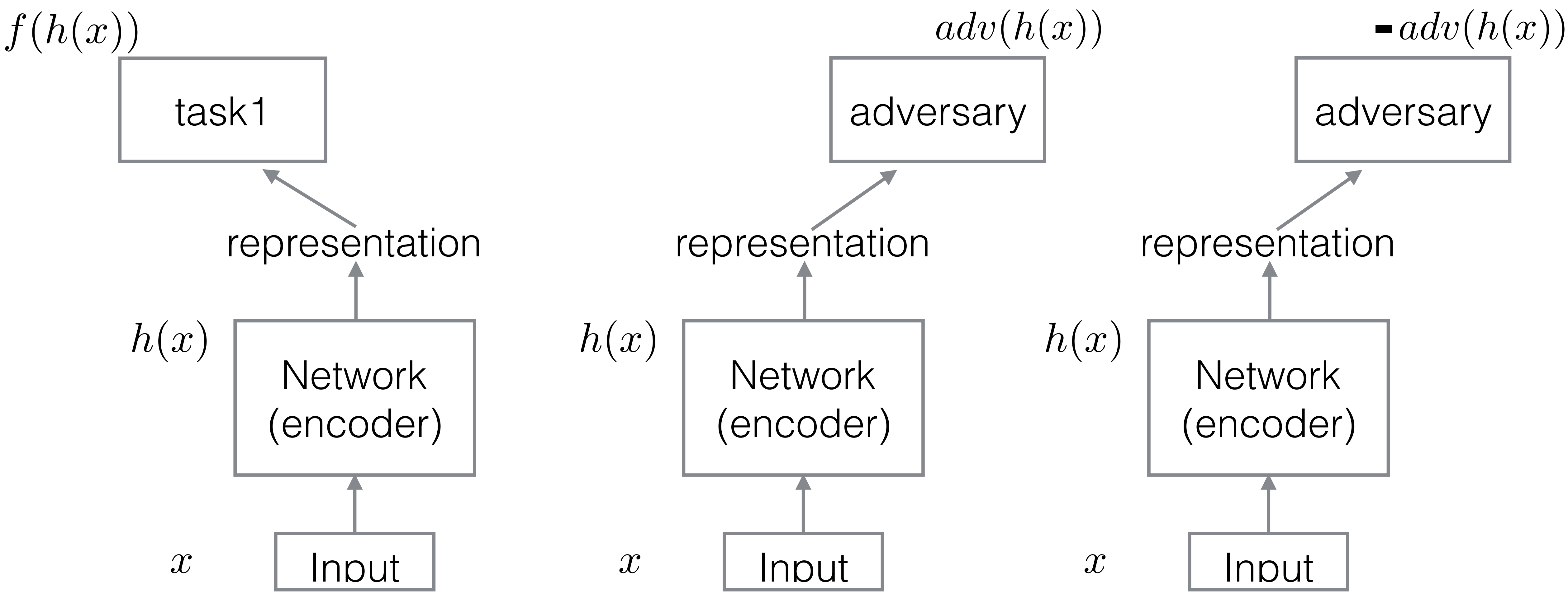
predict sentiment **don't predict race**



$adv(h(x))$

remove
stuff from
the representation

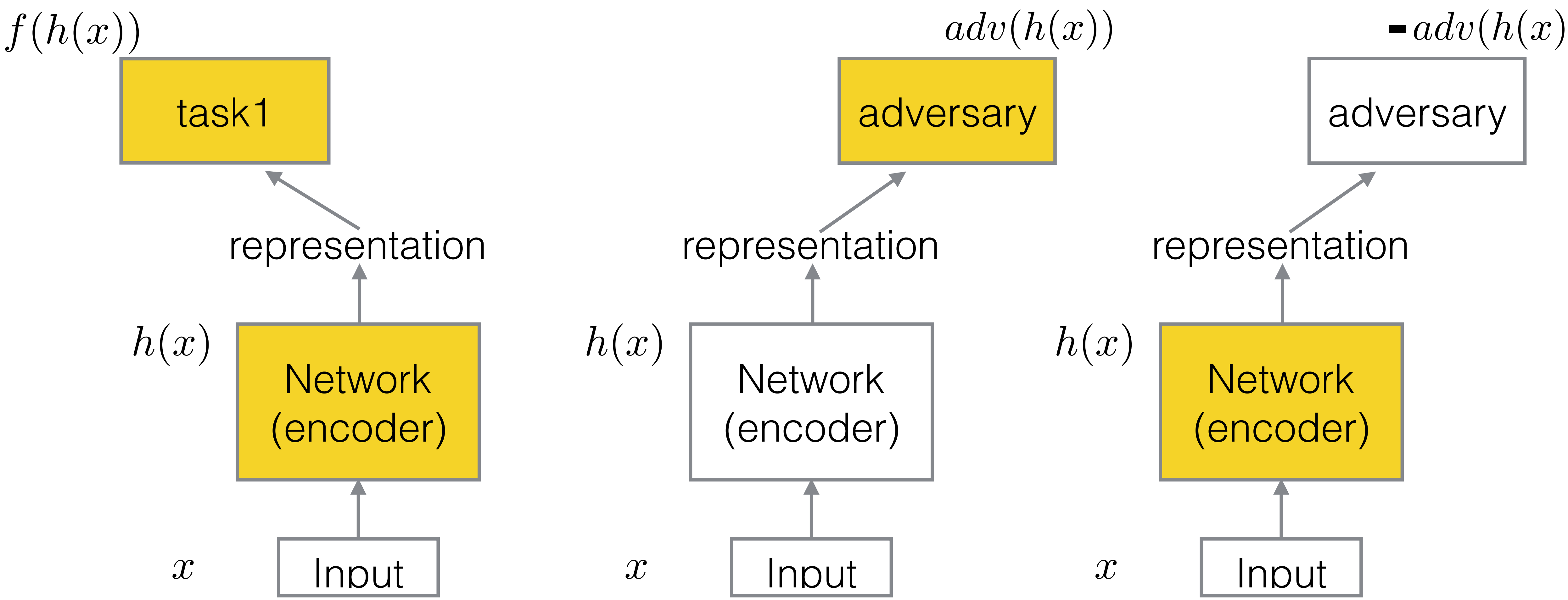
three different sub-objectives



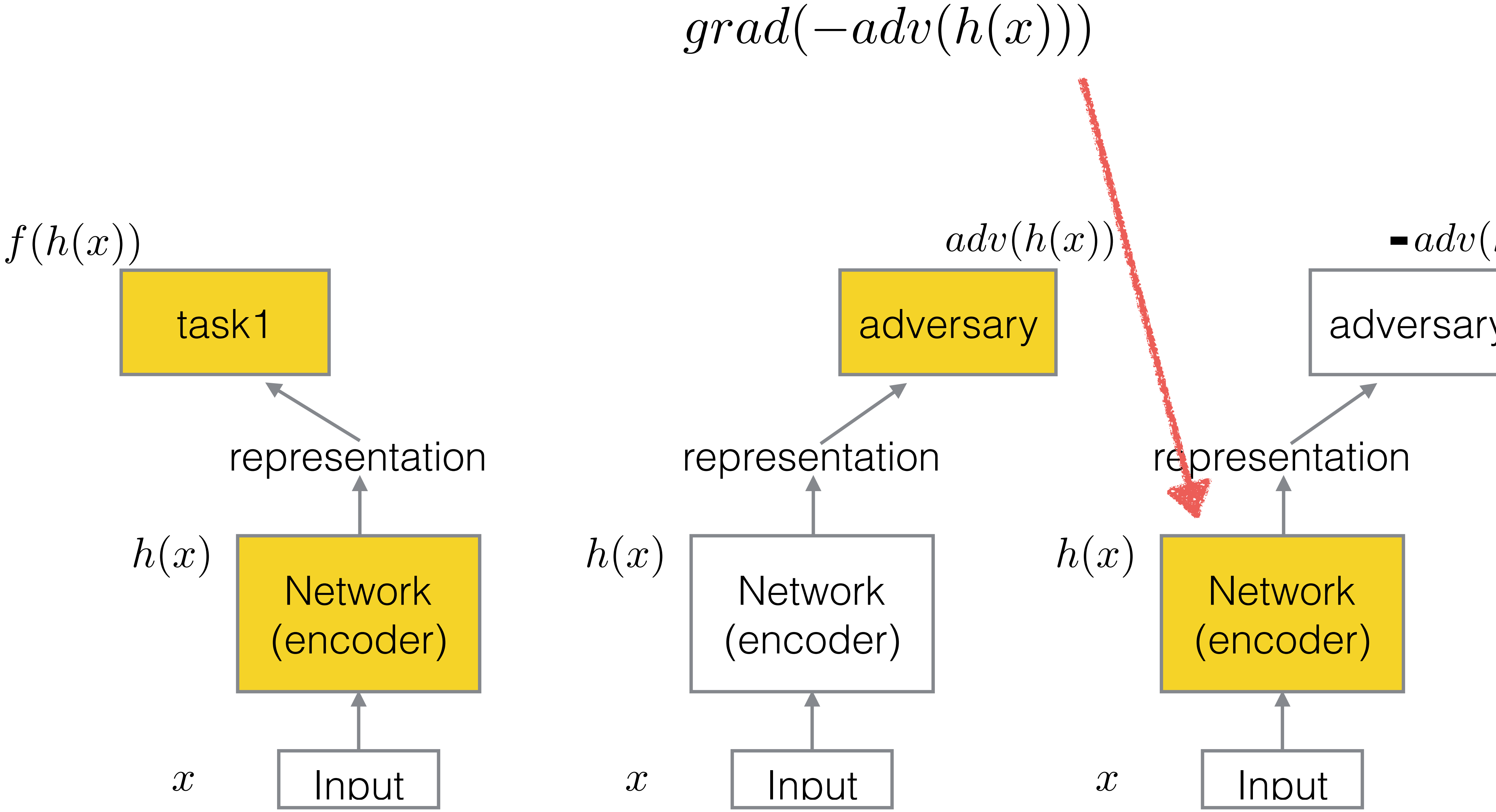
classify well

adversary should succeed

encoder should make adversary fail

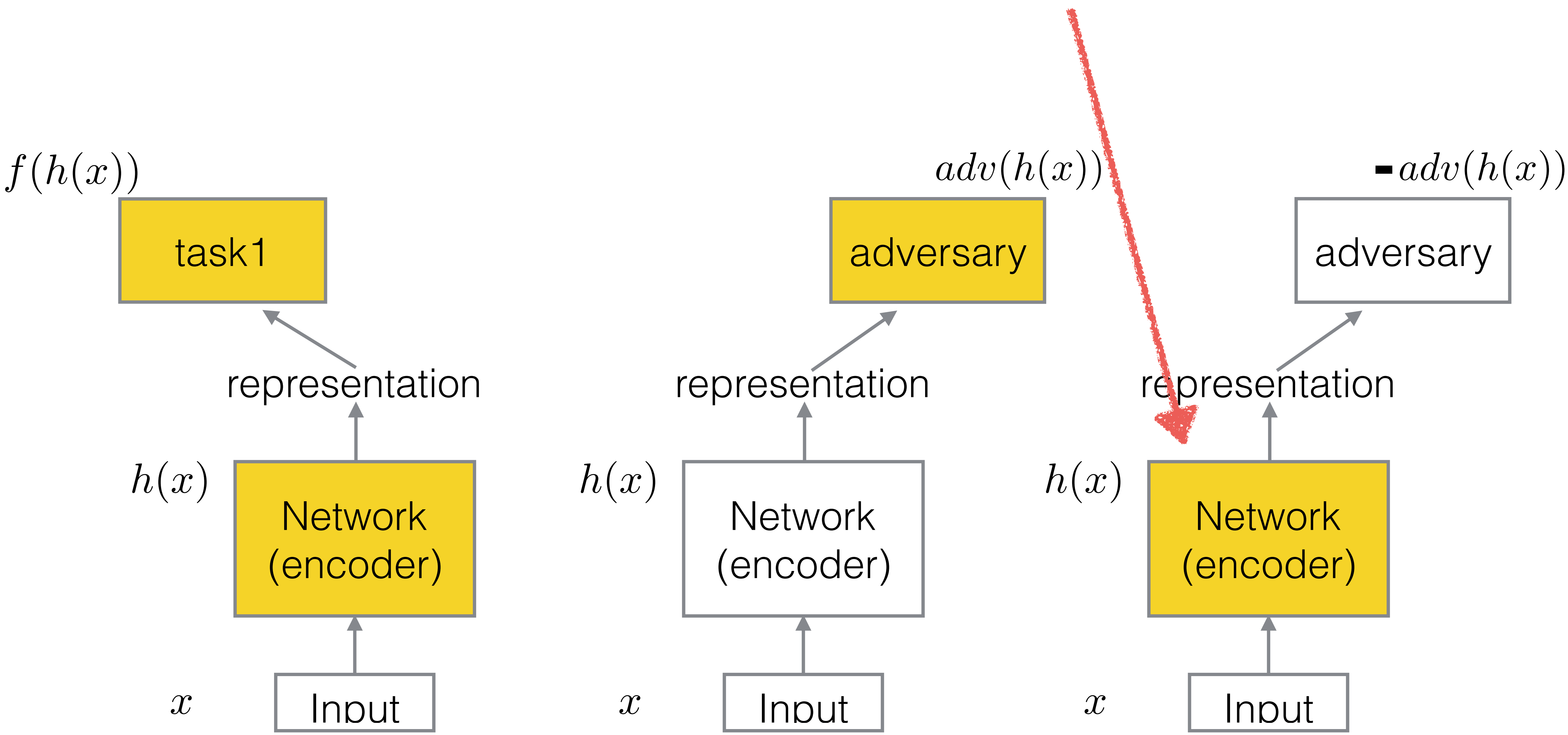


yellow: update parameters
white: don't update

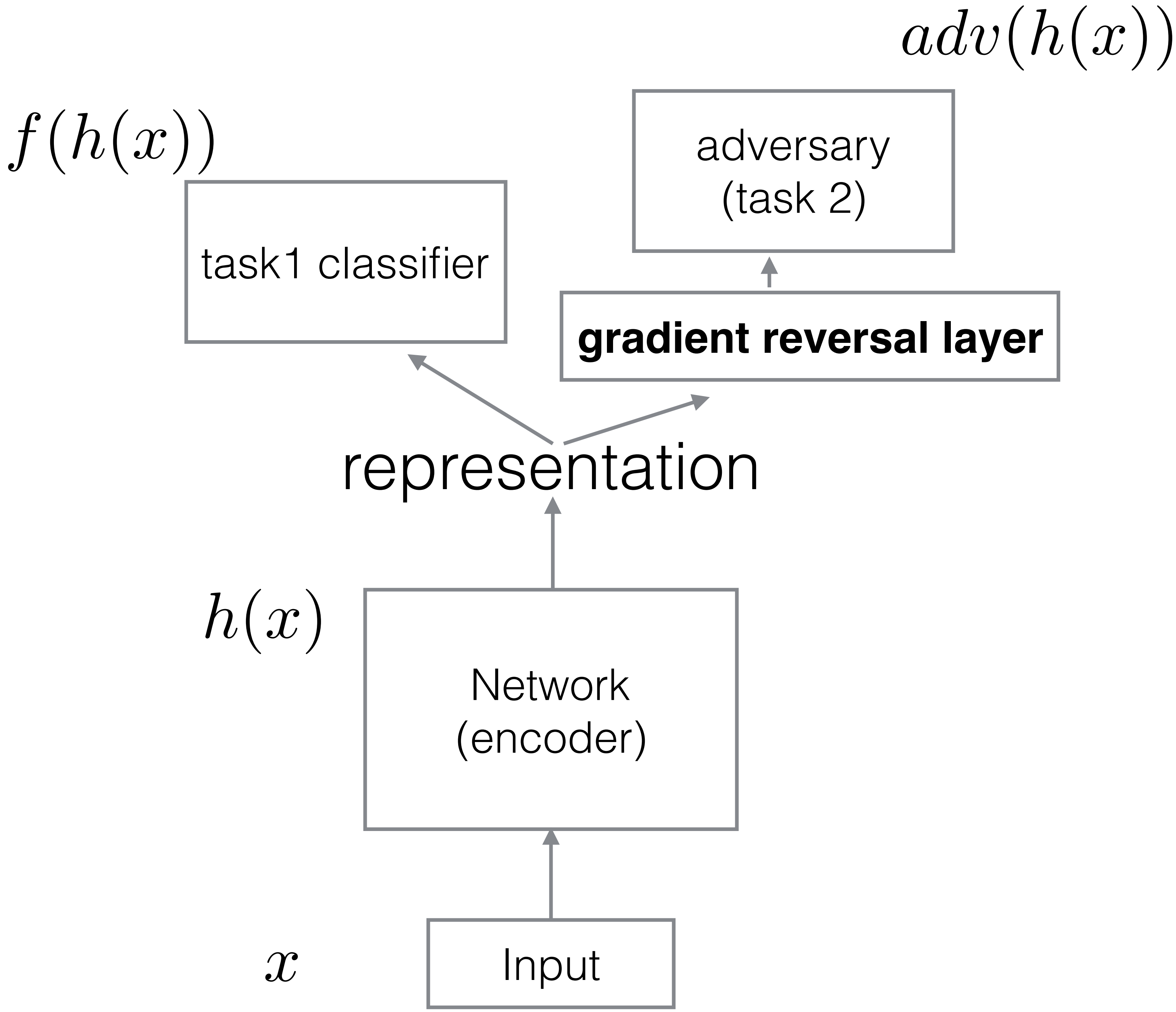


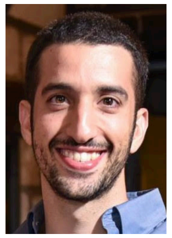
yellow: update parameters
white: don't update

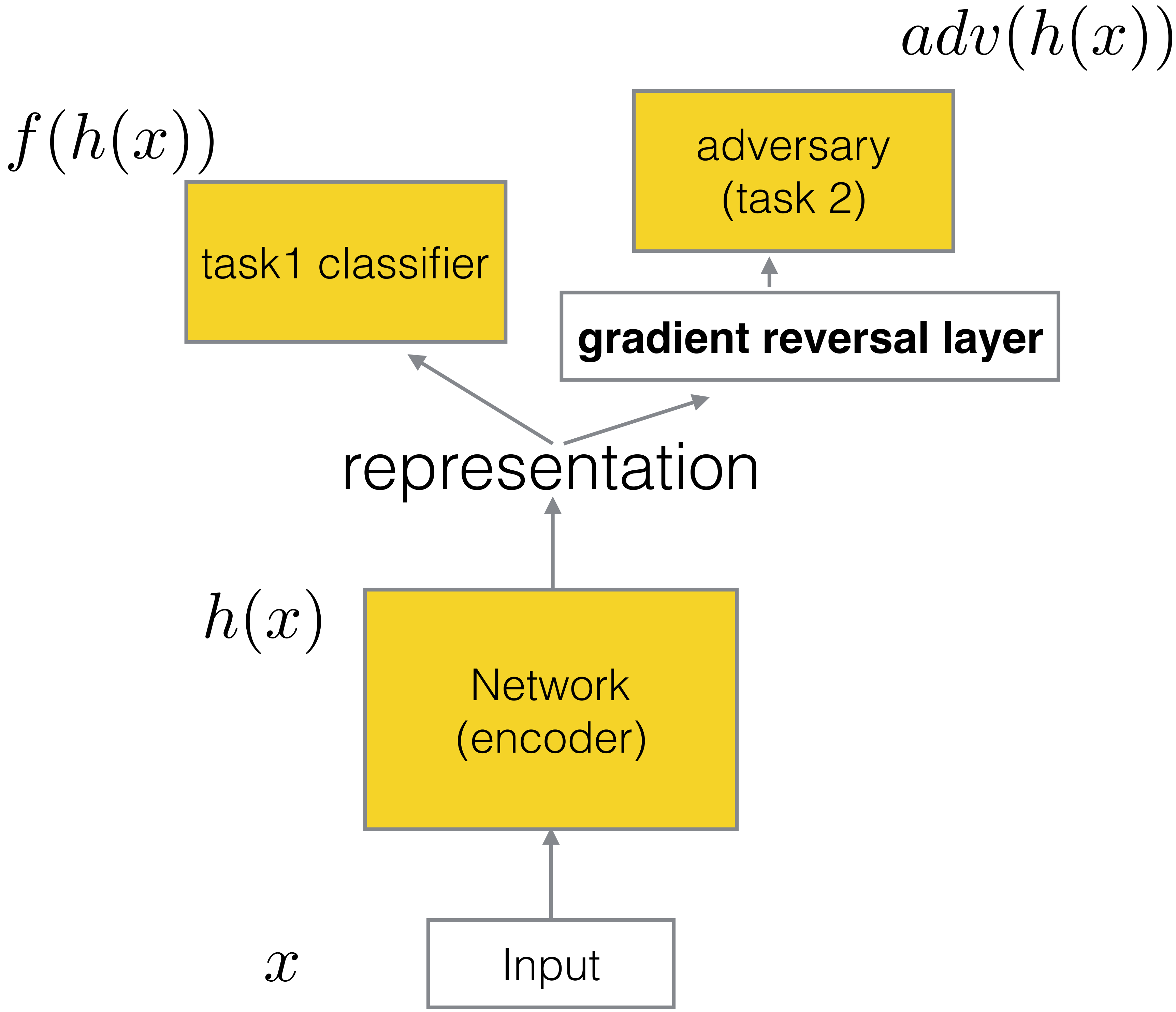
$$grad(-adv(h(x))) = -grad(adv(h(x)))$$

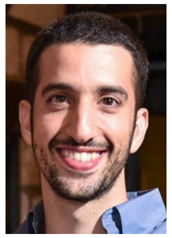


yellow: update parameters
white: don't update

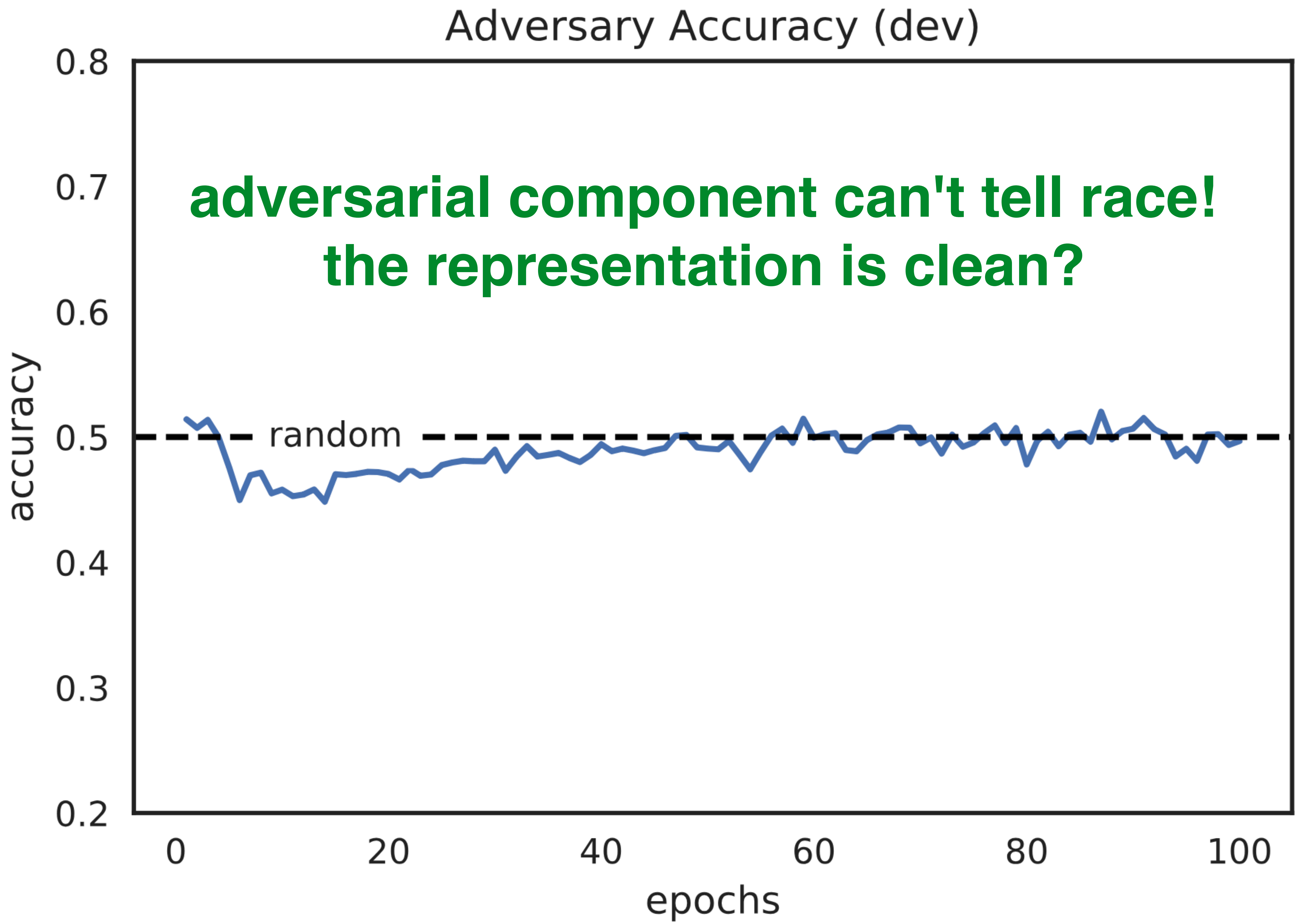

$adv(h(x))$
$f(h(x))$
adversary (task 2)
task1 classifier
gradient reversal layer
representation
$h(x)$
Network (encoder)
x
Input

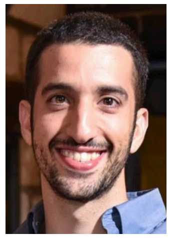



$adv(h(x))$
adversary
(task 2)
$f(h(x))$
task1 classifier
gradient reversal layer
representation
$h(x)$
Network
(encoder)
x
Input

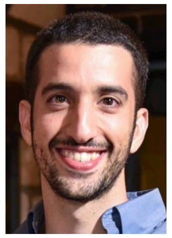


Does this work?



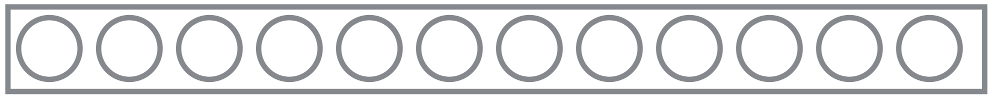


Does this work?



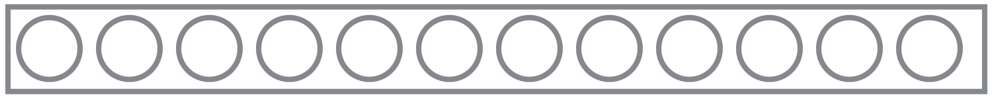
`im tryna get otp ian got nobody to talk to tho`

encode (regular)



race classifier succeeds

encode (adv train)

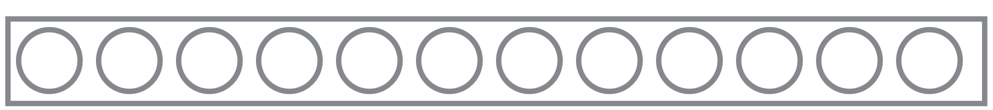


race classifier fails

Does this work?

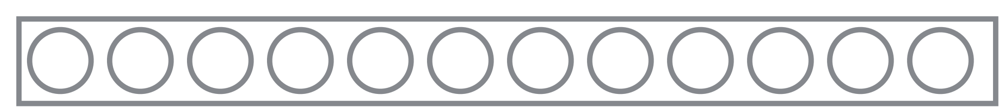
`Want to wear hipster glasses , but I have 20/20 vision`

encode (regular)



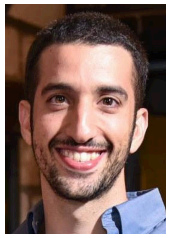
race classifier succeeds

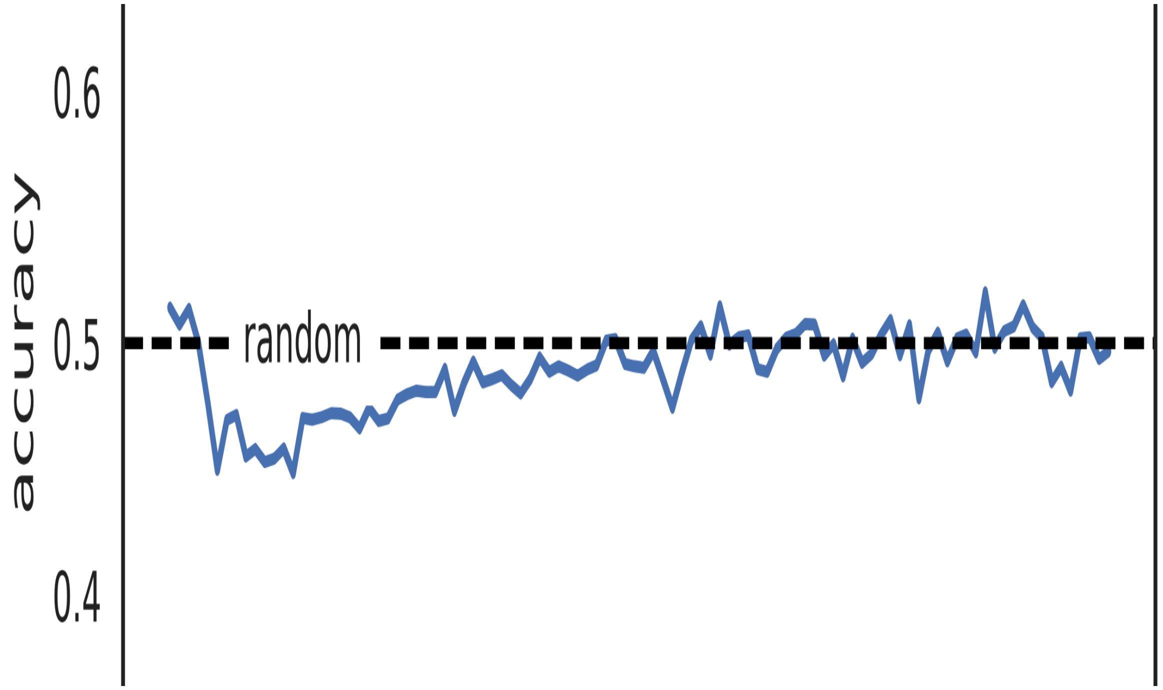
encode (adv train)



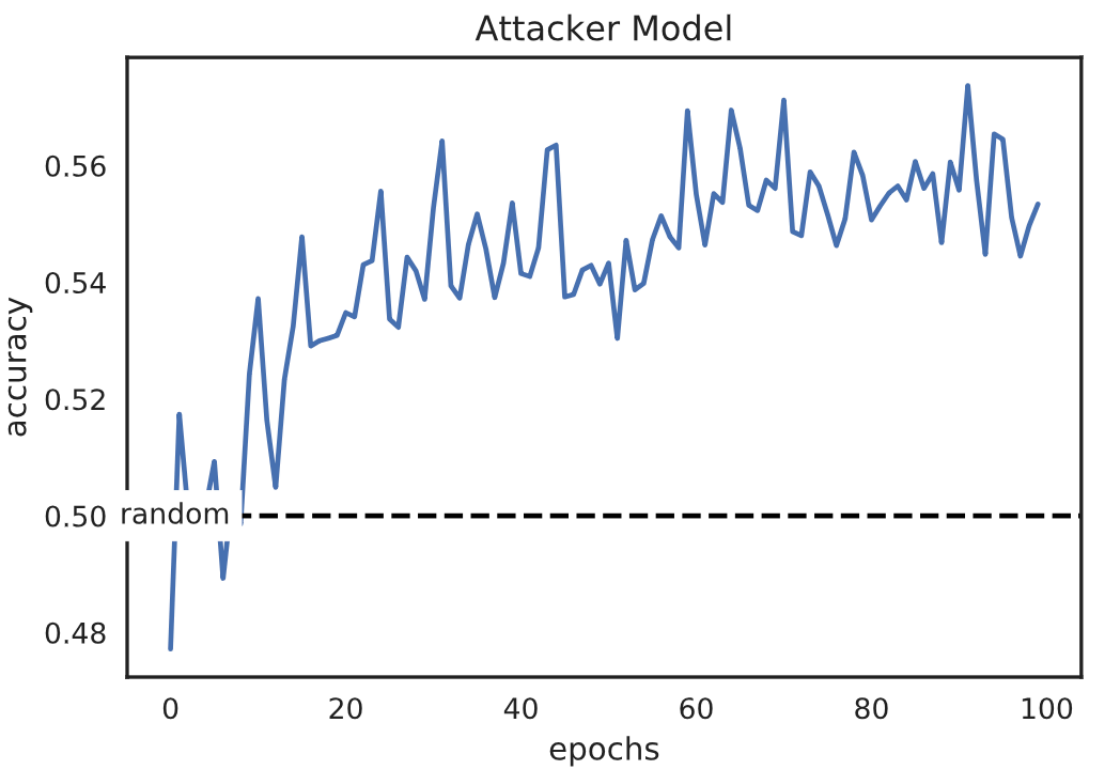
race classifier fails

However...



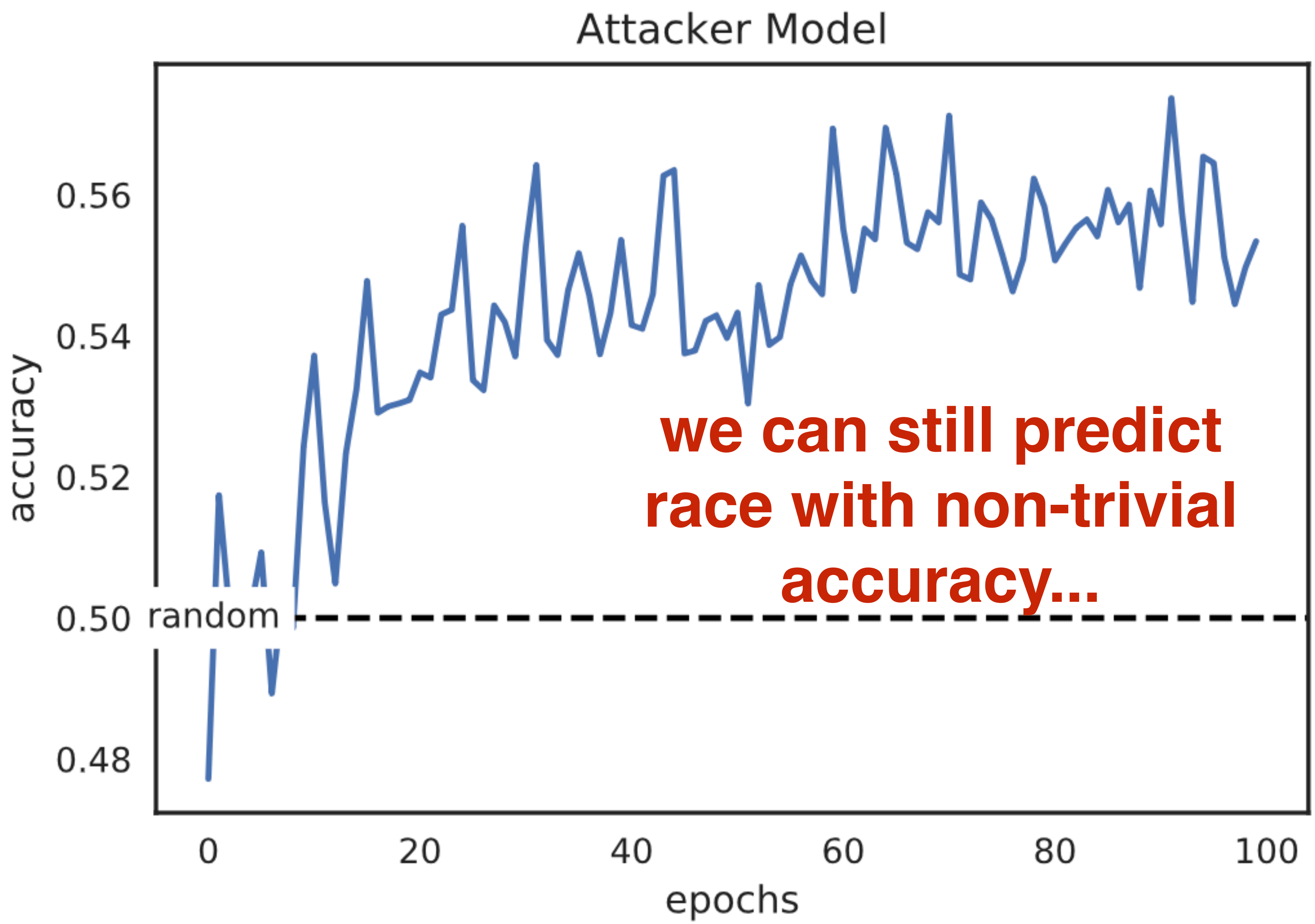


Adversarial classifier during training



"attacker" classifier trained to predict race on encoded representations

Does this work?



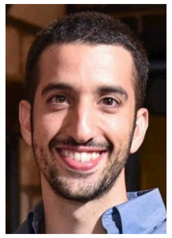
Does this work?

AAE ("non-hispanic blacks")
My Brew Eattin
_ Naw im cool
Tonoght was cool
My momma Bestfrand died
Enoy yall day
Going over Bae house

SAE ("non-hispanic Whites")
I want to be tan again
Why is it so hot in the house ?!
Been doing Spanish homework for 2 hours .
I wish I was still in Spain
Ahhhhh so much homework .

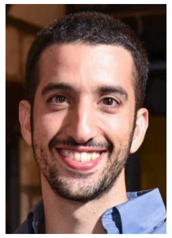
we can still predict race with non-trivial accuracy...

Summary of this part



- We train a text encoder for some task.
- Encoded vectors are useful for predicting various things...
- ...including things that we did not want to encode.
- Including things we **actively tried to remove**.
- **It is really hard to completely remove unwanted information from encoded language data**

Summary of this part:



- We train a text encoder for some task.
- Encoded vectors are useful for predicting various things...
- ...including things that we did not want to encode.
- Including things we **actively tried to remove**.
- **It is really hard to completely remove unwanted information from encoded language data**
- **Don't blindly trust the adversary!**

Summary of this part:

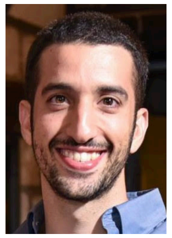
ing procedure (section 5.2).[1]

However, while successful to some extent, none of the methods fully succeed in removing all demographic information. Our main message, then, remains cautionary: **if the goal is to ensure fairness or invariant representation, do not trust adversarial removal of features from text inputs for achieving it.**

2 Learning Setup

We follow a setup in which we have some la-

Summary of this part:



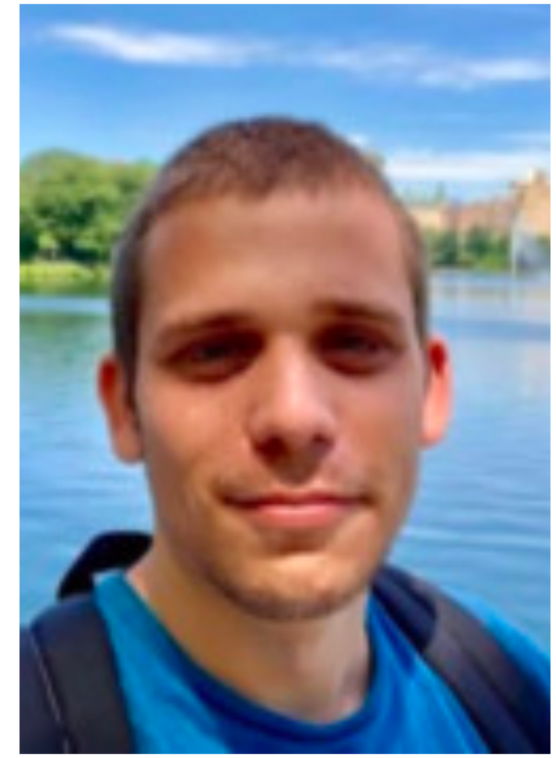
- We train a text encoder for some task.
- Encoded vectors are useful for predicting various things...
- ...including things that we did not want to encode.
- Including things we **actively tried to remove**.
- **It is really hard to completely remove unwanted information from encoded language data**
- **Don't blindly trust the adversary!**

Summary of this part:

- We train a text encoder for some task.
- Encoded vectors are useful for predicting various things...
- ...including things that we did not want to encode.
- Including things we **actively tried to remove**.
- **It is really hard to completely remove unwanted information from encoded language data**
- **Don't blindly trust the adversary!**

new work (in submission): we can do much better!

(was presented at venue, but cannot be put online yet due to anon period)



Main idea:

(was presented at venue, but cannot be put online yet due to anon period)

To summarize

- Neural networks learn representations. 😎
- Sharing the representations (multi-task learning). **Can be effective if careful.**
- Using the representations --- by querying them. **Ask LM for words. Use as features.**
- Biases in representations. **Prevalent.**
- Controlling the representations. **Hard, but we are making progress.**