Computational Methods for Linguists

Ling 471

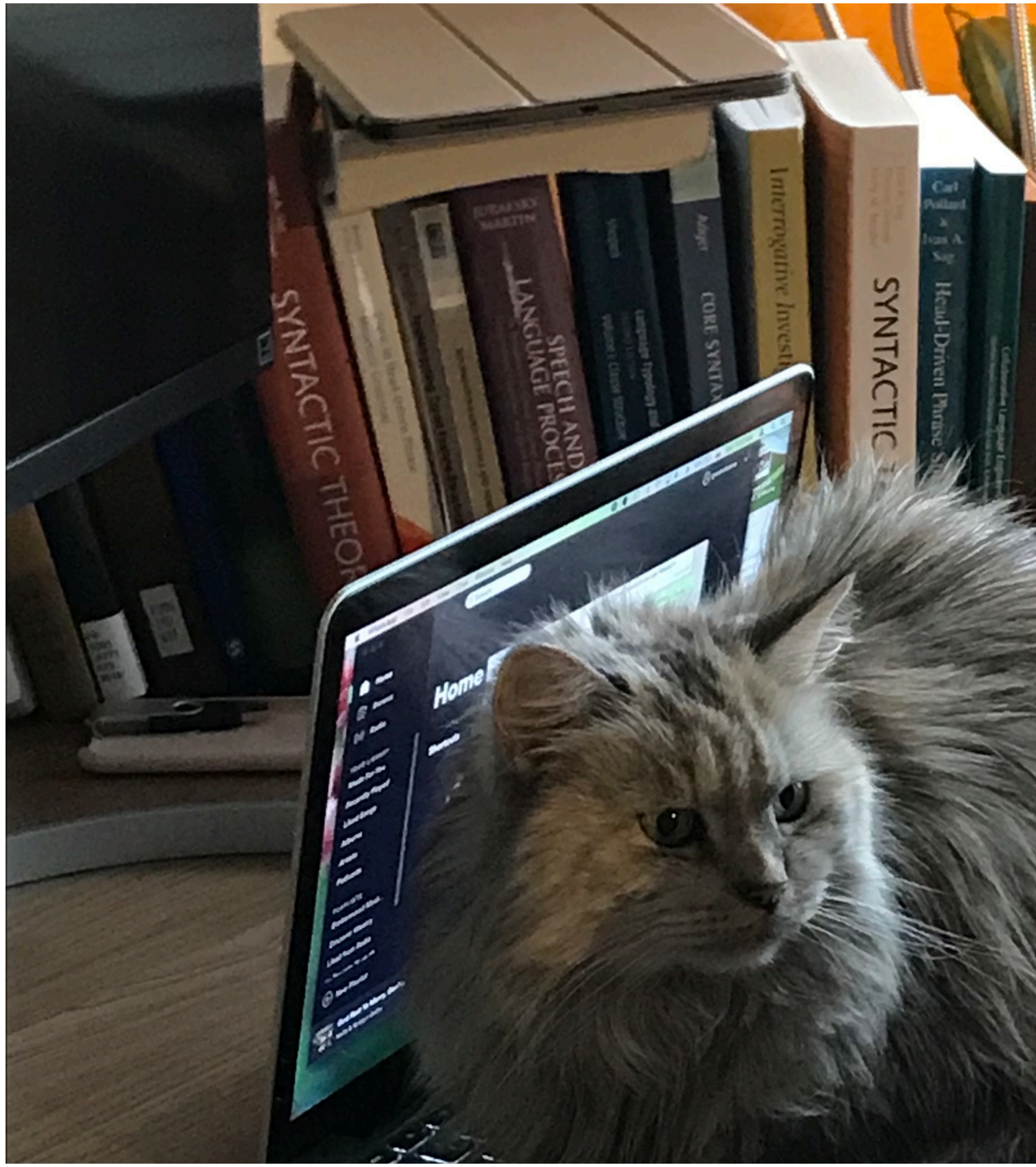
Olga Zamaraeva (Instructor)
Yuanhe Tian (TA)

05/20/21

Reminders

and announcements

- Presentation topic **suggestions** due next Tue via discussion board
  - they will have to **approved**
  - statistical approach to language data
  - peer-reviewed or established project
  - don't need to understand every detail but need to be able to clearly talk about:
    - what was done and why
    - why this is interesting
    - social impact (or lack thereof)



Plan for today

- Running programs in cmd
- Neural LMs
- Linguistic knowledge in NLP
  - philosophy and foundations of the debate
  - practicalities left to Ling472
  - (and to next week a bit)

How to easily run several programs?

Neural nets
and language models

XOR

A case for neural nets

$[x_1=1, x_2=0]$

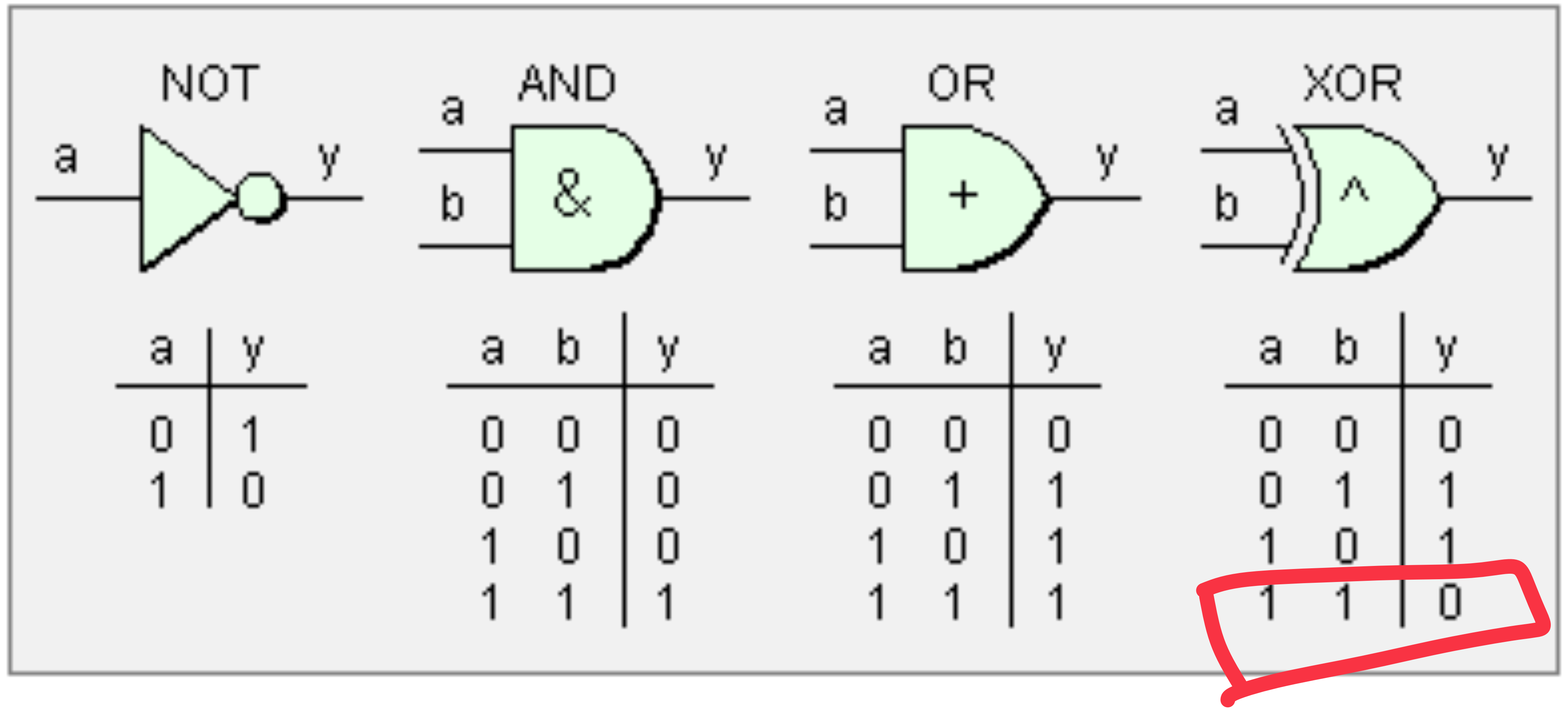
$h_1 = w_1x_1 + w_2x_2 + b =$

$= 1\cdot 1 + 0\cdot 1 + 0 = \boxed{1}$

$h_2 = 1\cdot 1 + 0\cdot 1 - 1 = \boxed{0}$

$y_1 = 1\cdot 1 + \boxed{0}(-2) + 0$

$= 0$



a	y
0	1
1	0

a	b	y
0	0	0
0	1	0
1	0	0
1	1	1

a	b	y
0	0	0
0	1	1
1	0	1
1	1	1

a	b	y
0	0	0
0	1	1
1	0	1
1	1	0

https://www.eetimes.com/how-to-invert-three-signals-with-only-two-not-gates-and-no-xor-gates-part-1/

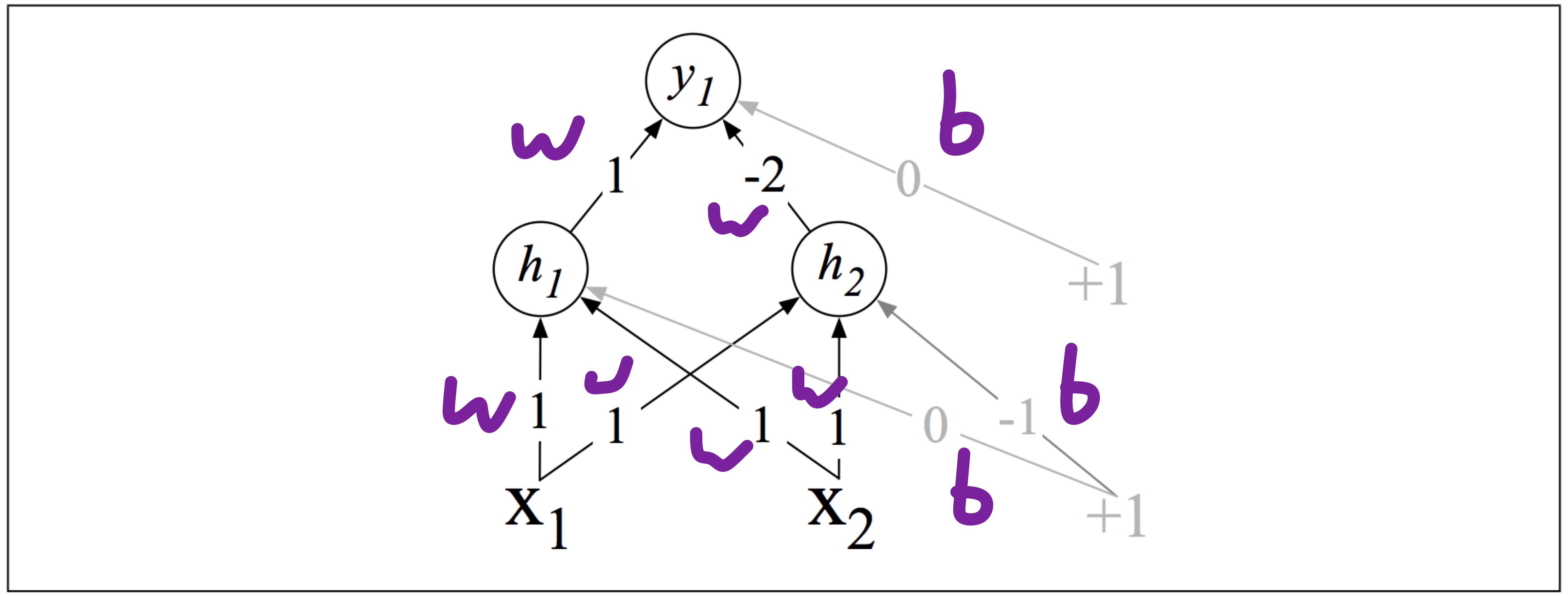


Figure 7.6 XOR solution after Goodfellow et al. (2016). There are three ReLU units, in two layers; we've called them h_1, h_2 (h for "hidden layer") and y_1. As before, the numbers on the arrows represent the weights w for each unit, and we represent the bias b as a weight on a unit clamped to +1, with the bias weights/units in gray.

Speech and Language Processing (Jurafsky and Martin 2004)

XOR

A case for neural nets

$$[x_1, x_2]$$

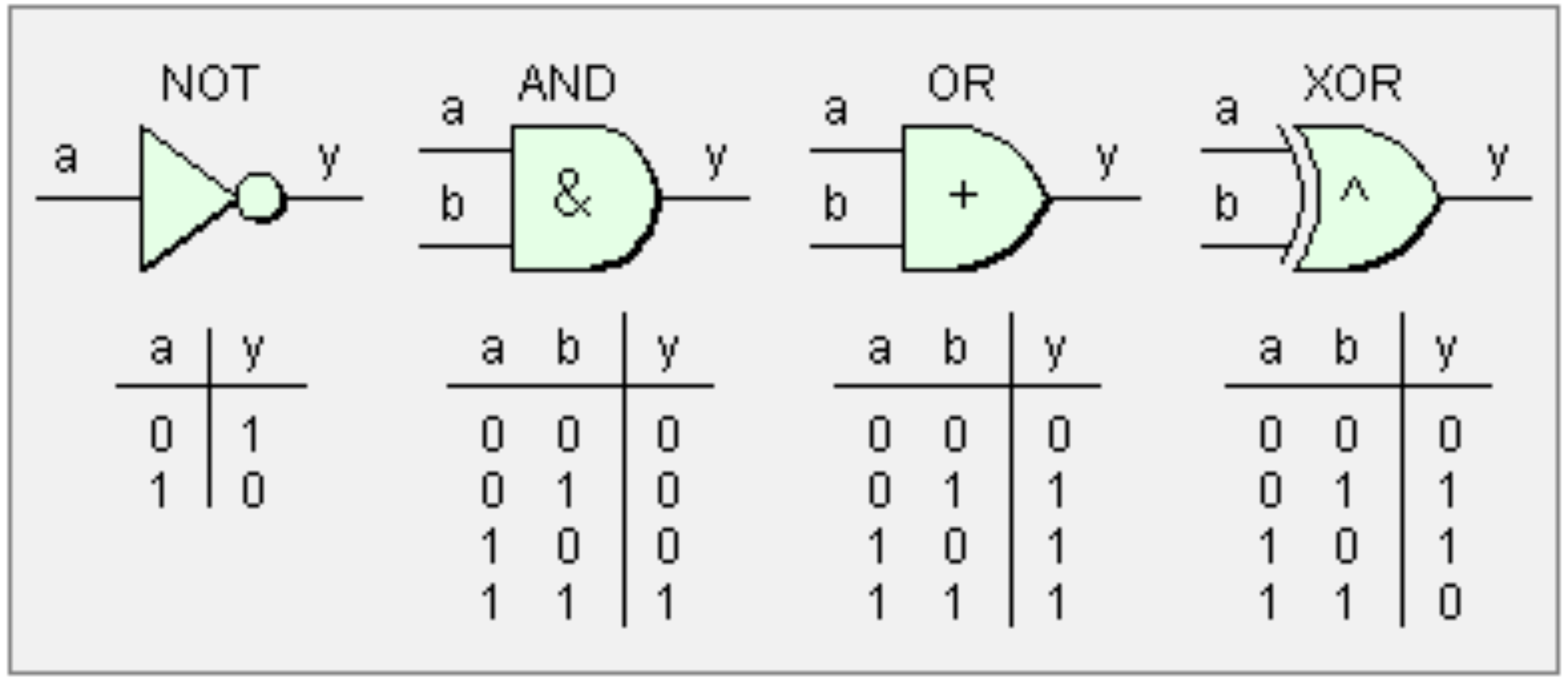
$$\begin{array}{ccc} & \downarrow & \downarrow \searrow \\ h_1 & w_1 & w_2 \\ h_2 & w_1 & w_2 \end{array}$$

$$\begin{array}{ccc} & h_1 & h_2 \\ y & w_1 & w_2 \end{array}$$

$$\downarrow$$

$$y$$

dimensions.



a	y
0	1
1	0

a	b	y
0	0	0
0	1	0
1	0	0
1	1	1

a	b	y
0	0	0
0	1	1
1	0	1
1	1	1

a	b	y
0	0	0
0	1	1
1	0	1
1	1	0

https://www.eetimes.com/how-to-invert-three-signals-with-only-two-not-gates-and-no-xor-gates-part-1/

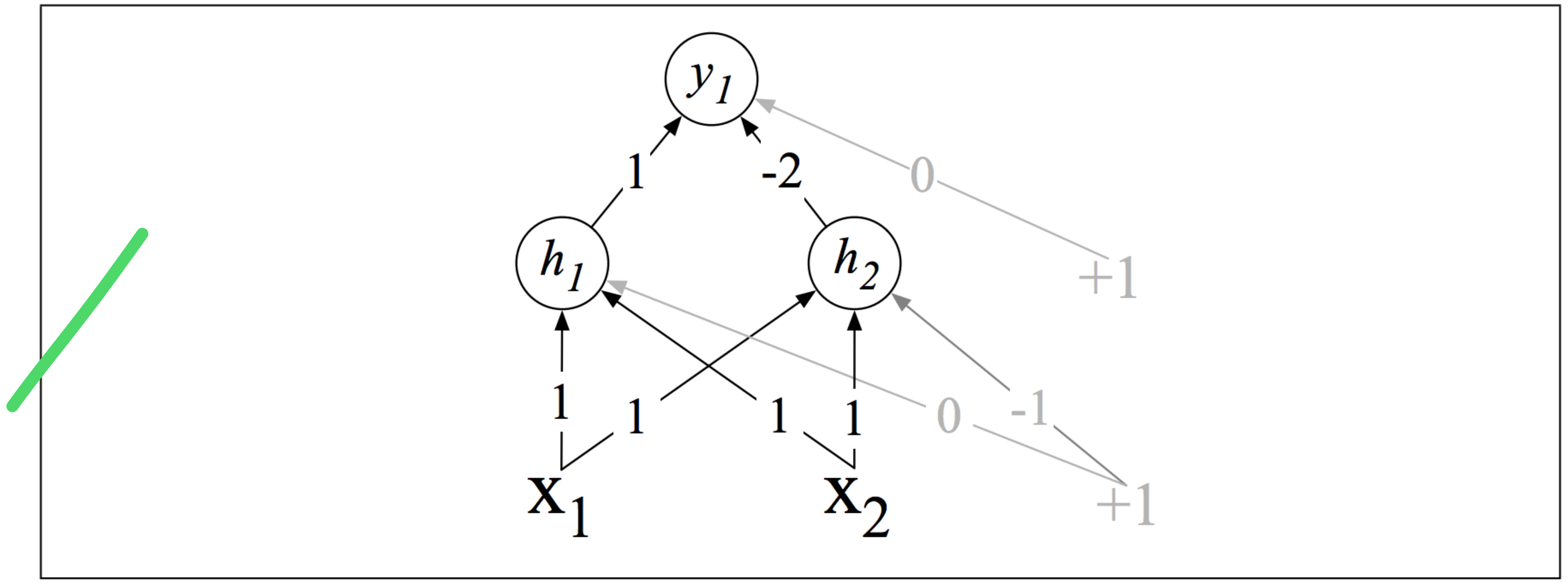
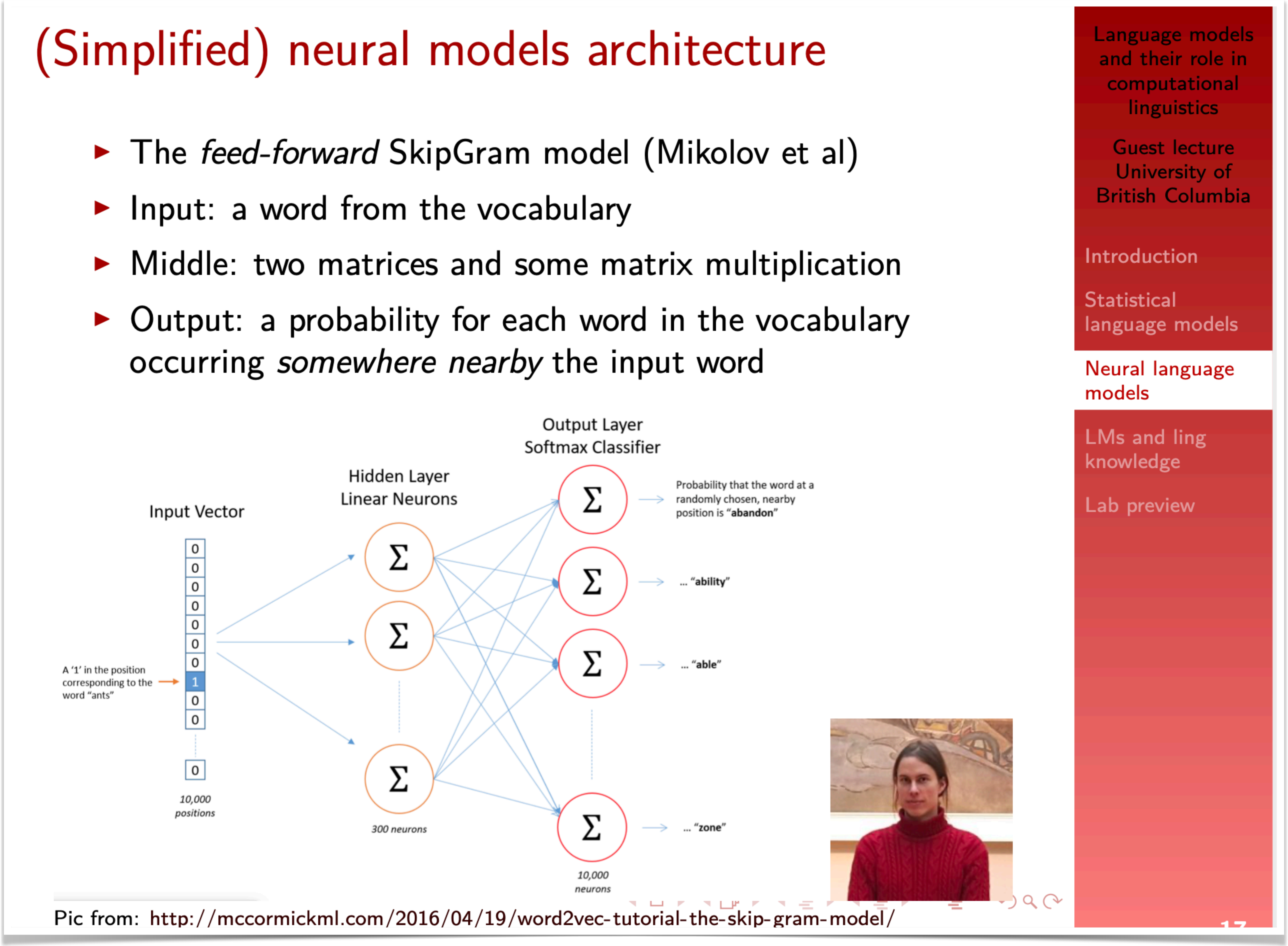


Figure 7.6 XOR solution after Goodfellow et al. (2016). There are three ReLU units, in two layers; we've called them h_1, h_2 (h for "hidden layer") and y_1. As before, the numbers on the arrows represent the weights w for each unit, and we represent the bias b as a weight on a unit clamped to +1, with the bias weights/units in gray.

Speech and Language Processing (Juratsky and Martin 2004)



(Simplified) neural models architecture

- The *feed-forward* SkipGram model (Mikolov et al)
- Input: a word from the vocabulary
- Middle: two matrices and some matrix multiplication
- Output: a probability for each word in the vocabulary occurring *somewhere nearby* the input word

Pic from: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Language models and their role in computational linguistics

Guest lecture University of British Columbia

Introduction

Statistical language models

Neural language models

LMs and ling knowledge

Lab preview

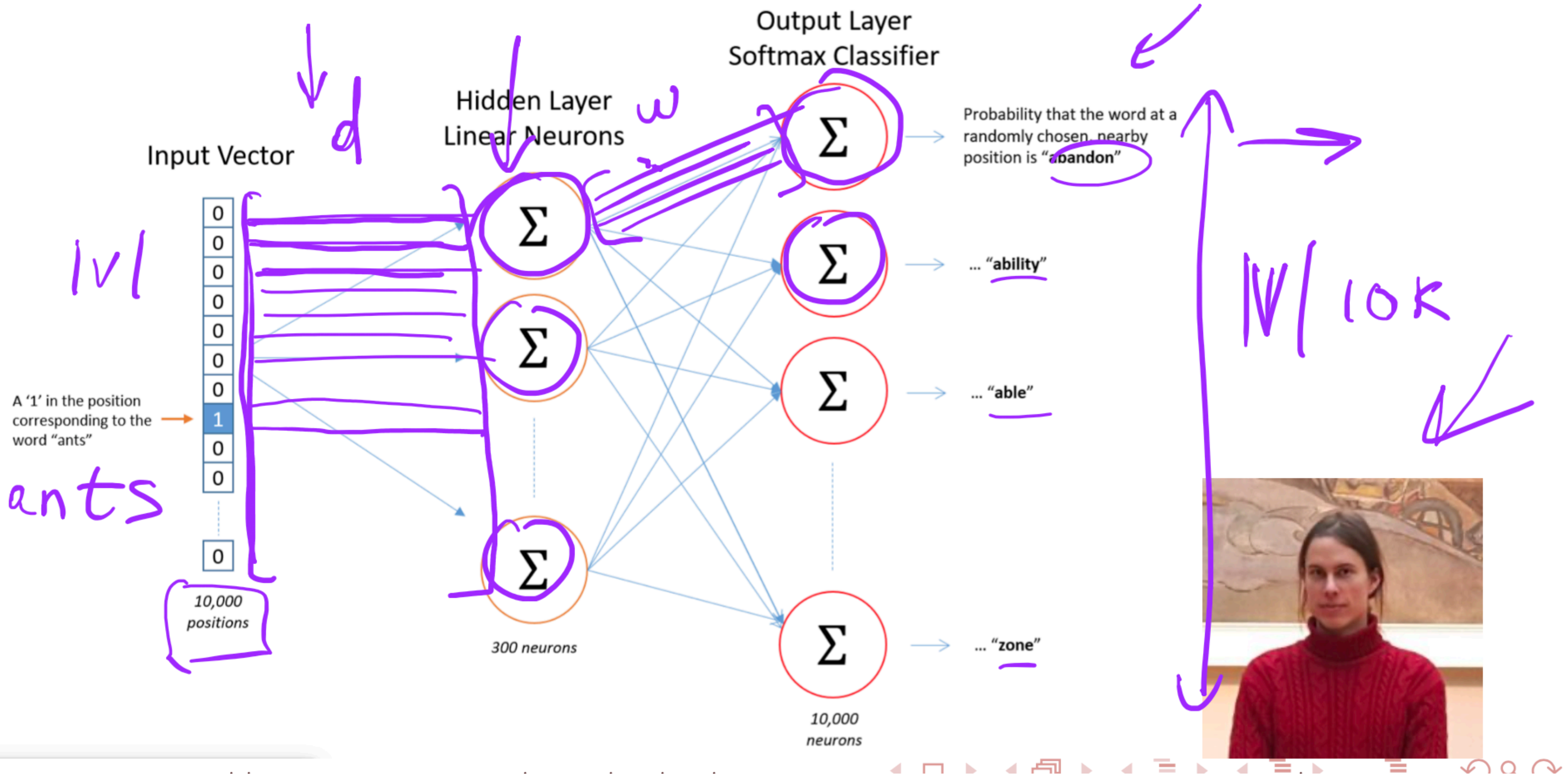
- What are the "two matrices"?!

(Simplified) neural models architecture

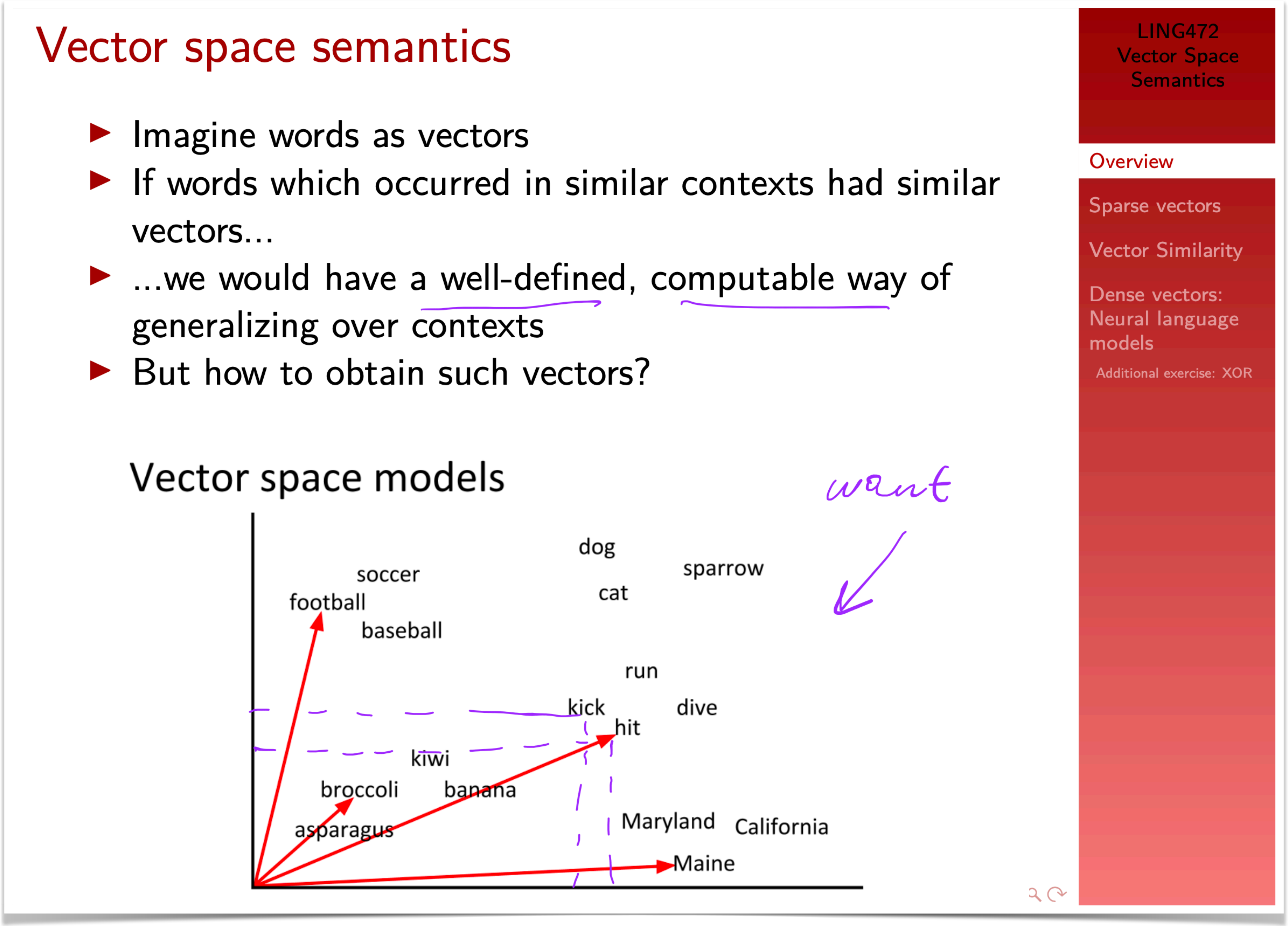
furry meows
[1 1]

- The *feed-forward* SkipGram model (Mikolov et al)
- Input: a word from the vocabulary
- Middle: two matrices and some matrix multiplication
- Output: a probability for each word in the vocabulary occurring *somewhere nearby* the input word

- Matrix2 is some coefficients/weights/parameters
- Matrix 1 is…
  - "dense word embeddings"
    - ?!?



Word embeddings



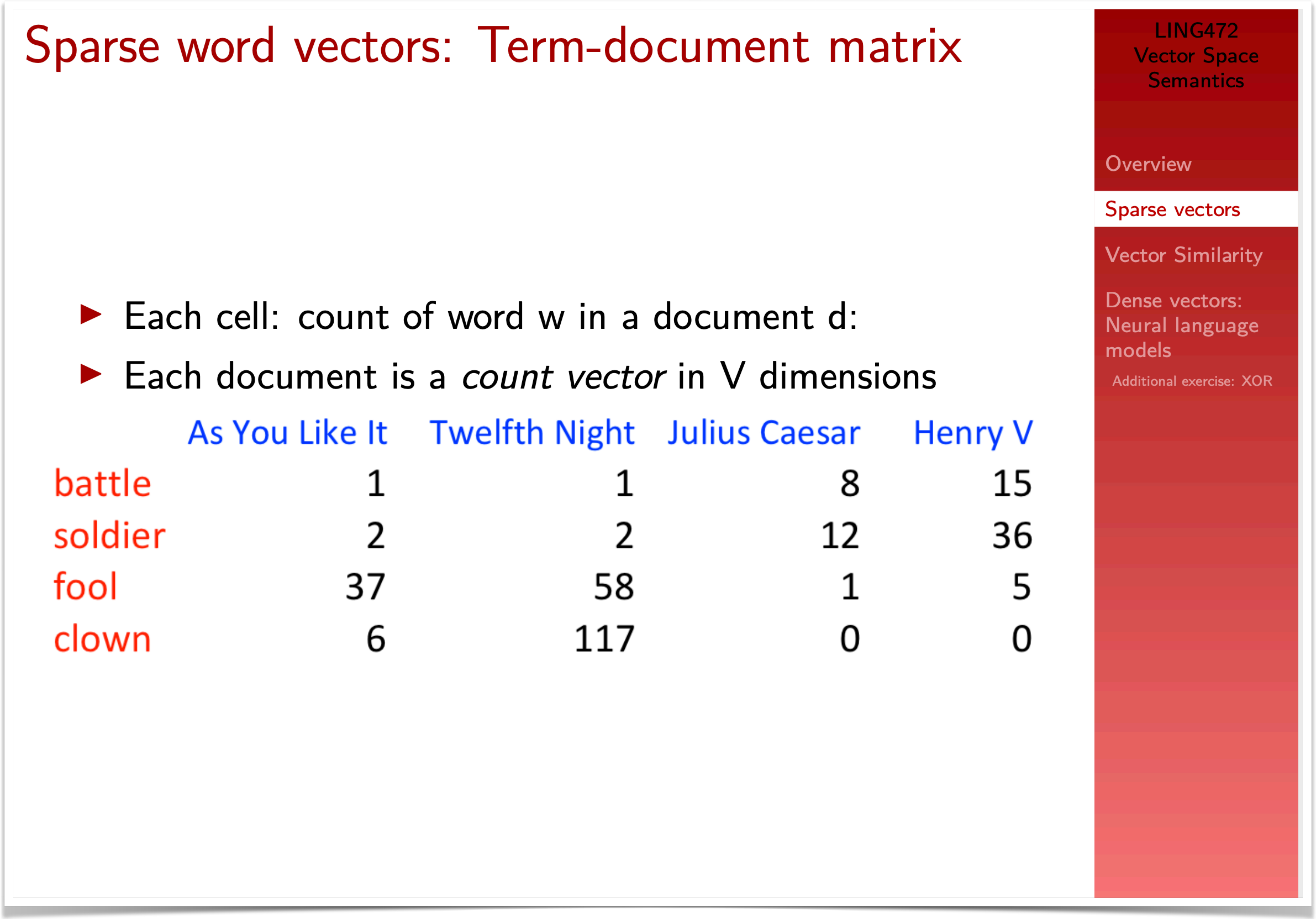
Vector space semantics

- Imagine words as vectors
- If words which occurred in similar contexts had similar vectors...
- ...we would have a well-defined, computable way of generalizing over contexts
- But how to obtain such vectors?

Vector space models

want

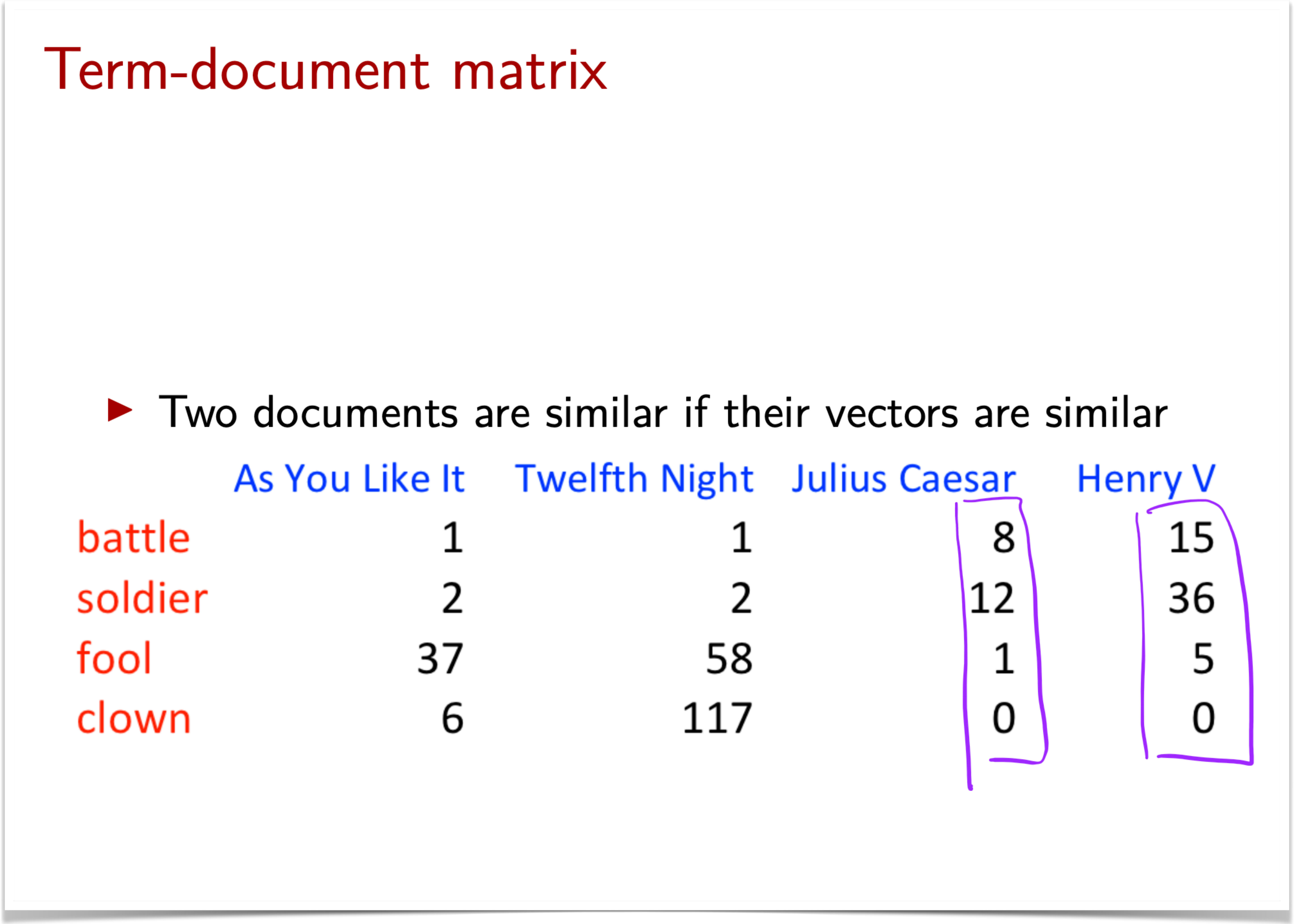
- "word embedding"
  - = "word vector"
    - it's a vector of e.g. counts



Sparse word vectors: Term-document matrix

- Each cell: count of word w in a document d:
- Each document is a *count vector* in V dimensions

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

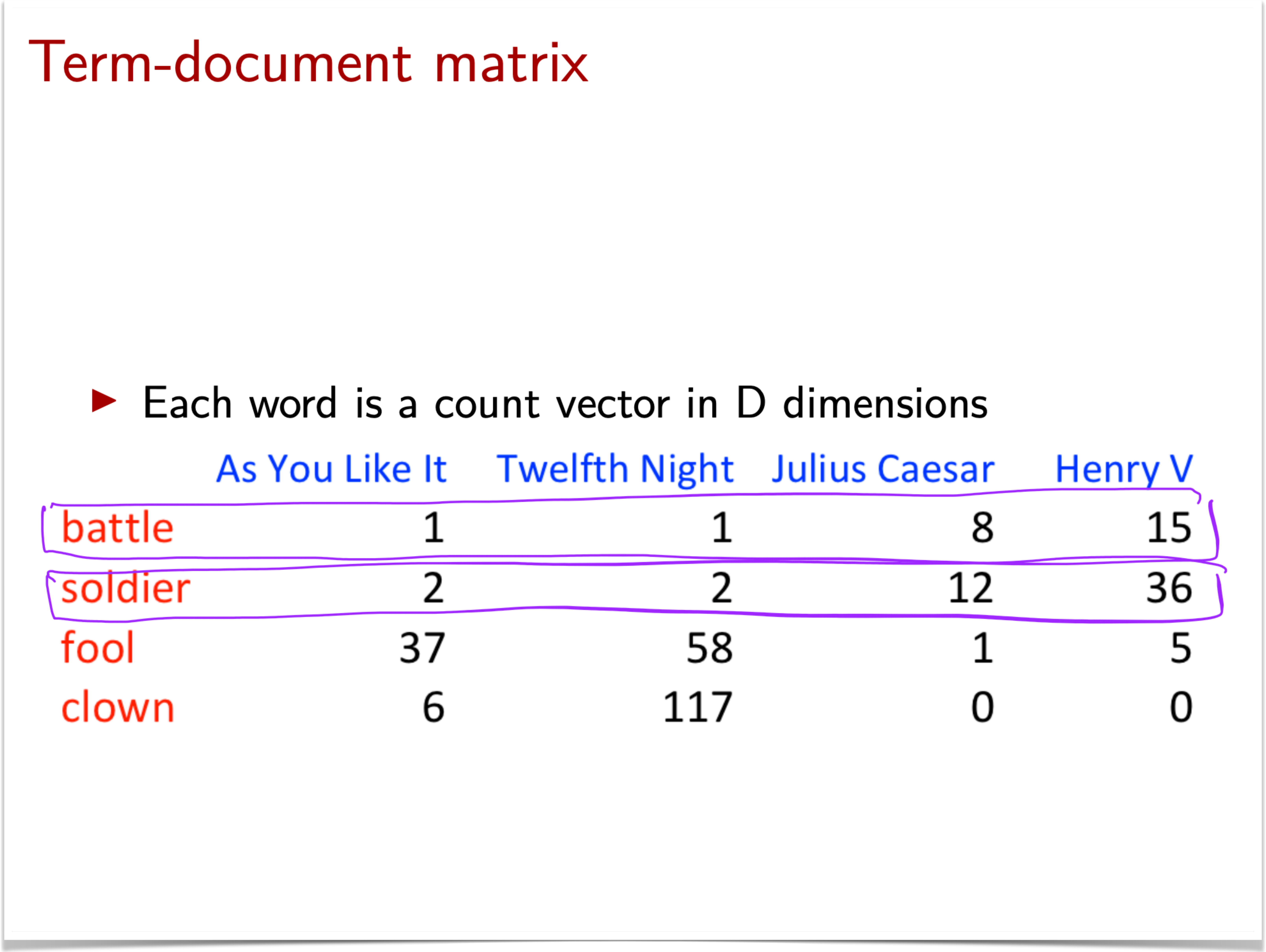


Term-document matrix

- Two documents are similar if their vectors are similar

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

- btw in HW4:
  - TF-IDF “vectorizer”
  - is building something like this…



Term-document matrix

- Each word is a count vector in D dimensions

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	6	117	0	0

- **Two words are similar if their vectors are similar!**

Sparse word vectors: Word-context matrix

- Instead of documents, use smaller *context windows* of e.g. 7 words
- Vector is defined in terms of how many times a word occurs near another word
- this is a VxV matrix
- V is very large (e.g. 50K)

	aardvark	computer	data	pinch	result	sugar	...
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	

Sparse vectors; Word-context matrix

Using a window of ±7 words:

7 7

sugar, a sliced lemon, a tablespoonful of	**apricot**	preserve or jam, a pinch each of,
their enjoyment. Cautiously she sampled her first	**pineapple**	and another fruit whose taste she likened
well suited to programming on the digital	**computer**.	In finding the optimal R-stage policy from
for the purpose of gathering data and	**information**	necessary for the study authorized in the

	aardvark	computer	data	pinch	result	sugar	6 ...
apricot	0	0	0	1	0	1	
pineapple	0	0	0	1	0	1	
digital	0	2	1	0	1	0	
information	0	1	6	0	4	0	

50K

4

50K

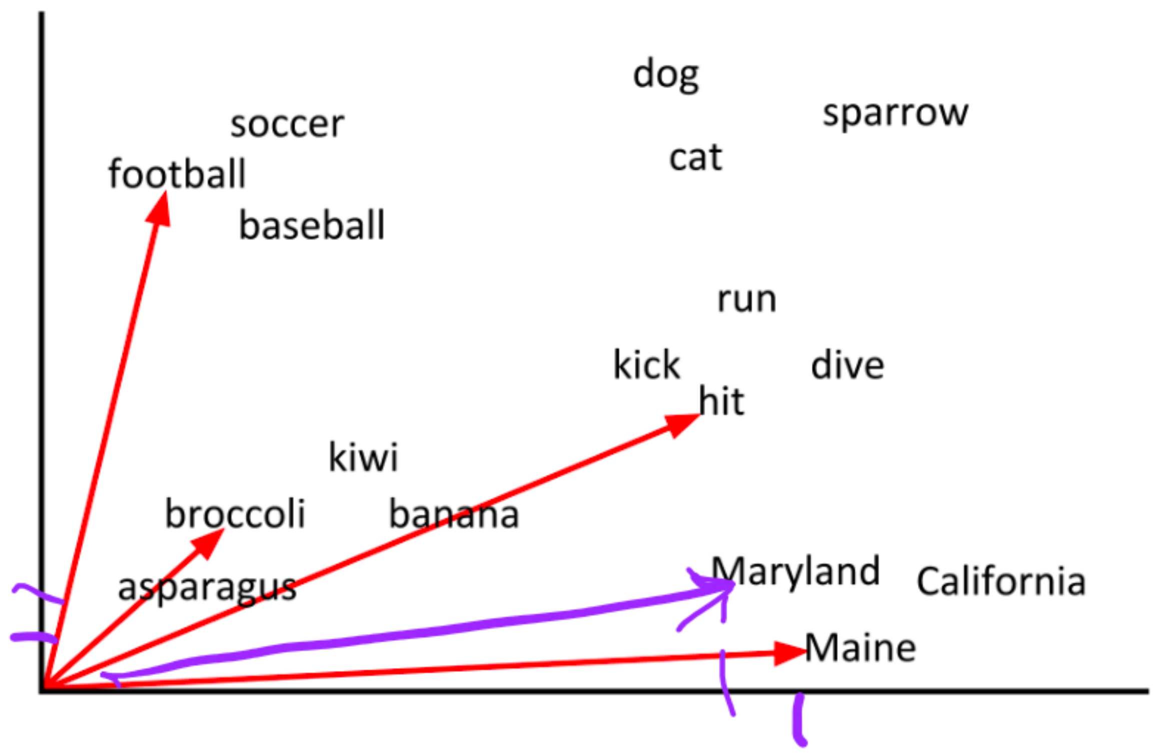
Vector similarity: Cosine

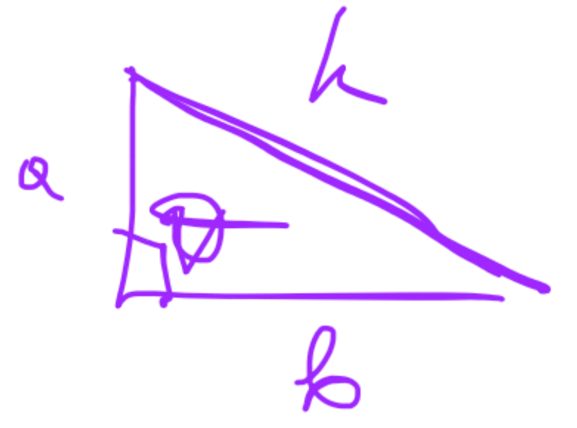
- Notion of vector similarity from linear algebra: the **dot product**

$$\underline{\text{dot-product}(\vec{v},\vec{w}) = \vec{v}\cdot\vec{w}} = \sum_{i=1}^{N} v_i w_i = \underline{v_1 w_1} + \underline{v_2 w_2} + \ldots + \underline{v_N w_N}$$

- The dot product is **high** when vectors have similar dimensions

Vector space models





$$\cos\theta = \frac{b}{h}$$

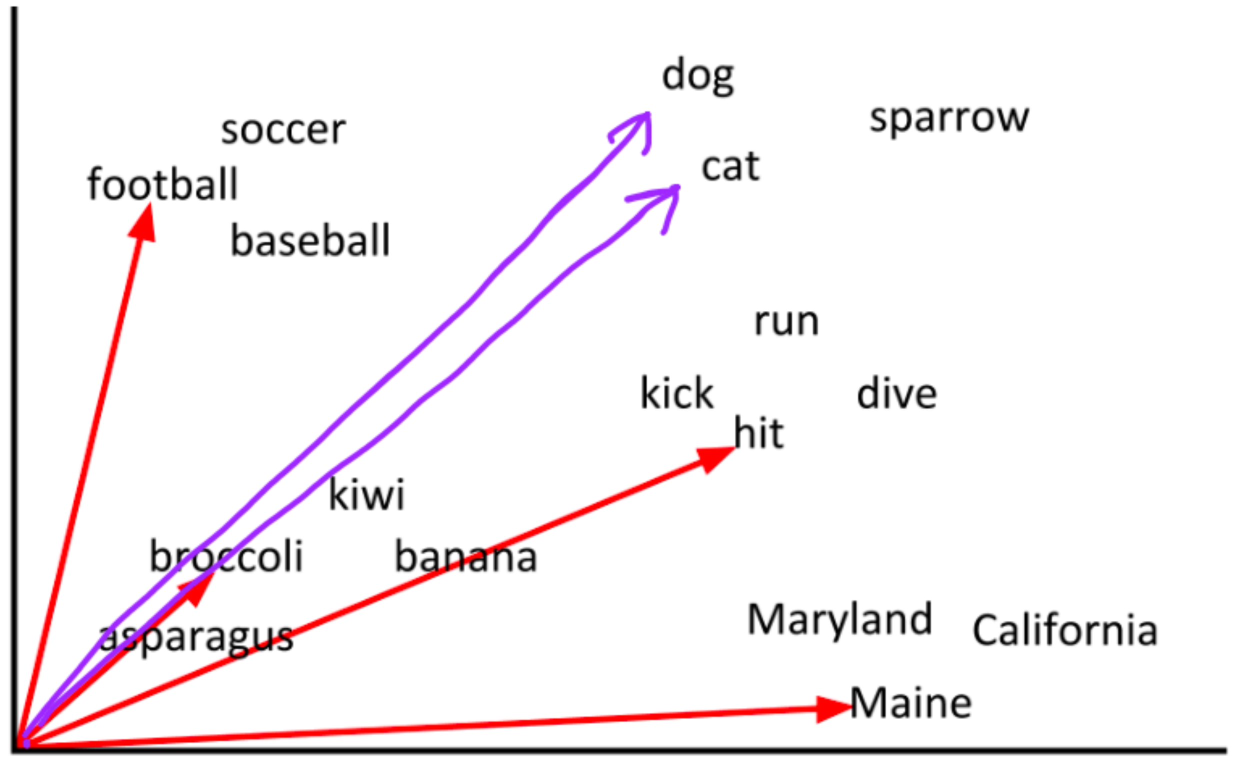
$$\vec{x} = [1,1] \qquad \vec{x}\cdot\vec{y} = 1\cdot 2 + 1\cdot 2$$
$$\vec{y} = [2,2]$$

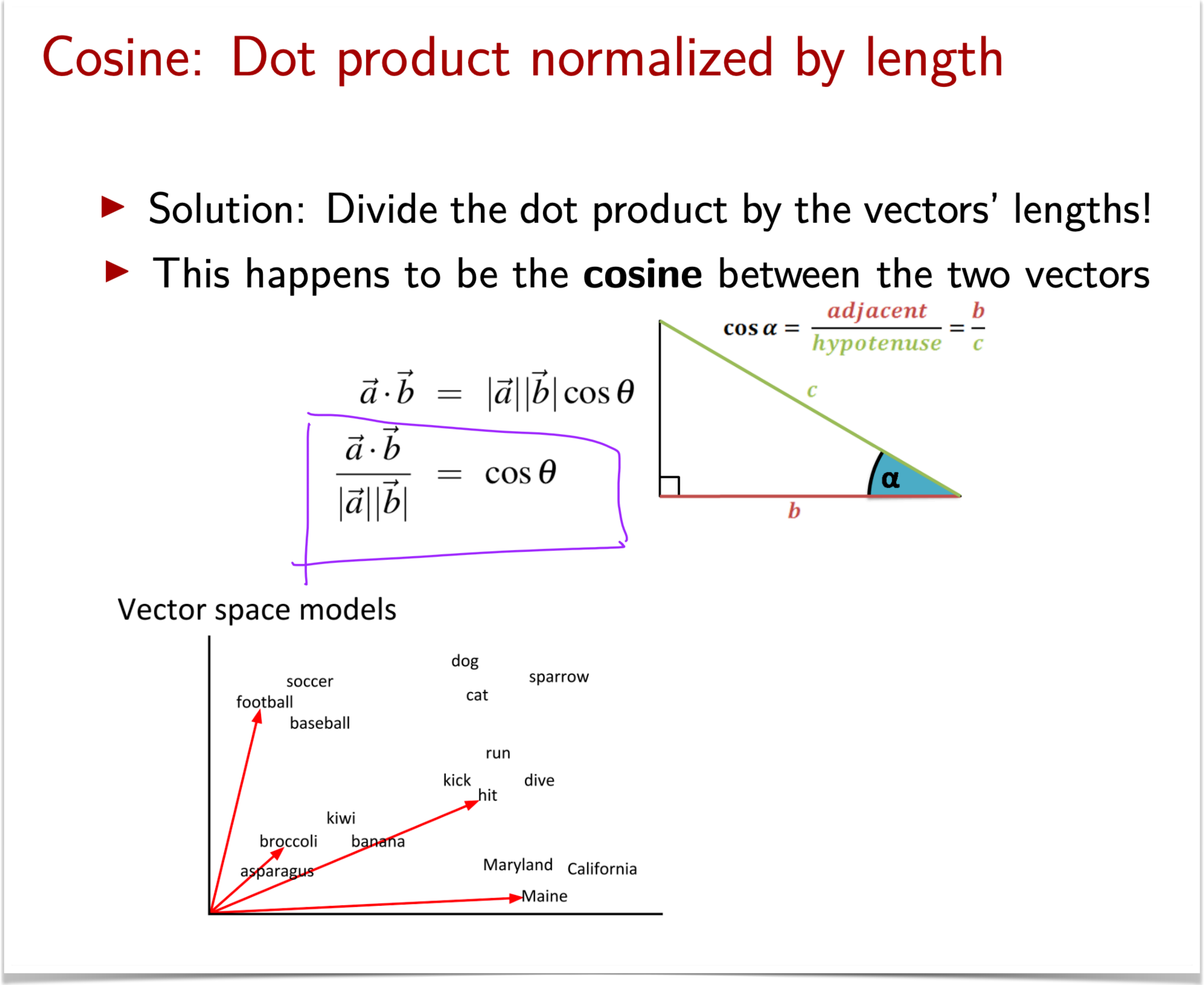
- dot product:
  - natural measure of similarity for vectors

Cosine: Dot product normalized by length

- A problem with dot product:
  - Dot product is higher is the vectors are longer
  - This means frequent words will have higher dot products
  - The model of *similarity* becomes sensitive to word *frequency*

Vector space models





Cosine: Dot product normalized by length

- ▶ Solution: Divide the dot product by the vectors’ lengths!
- ▶ This happens to be the **cosine** between the two vectors

$$\vec{a}\cdot\vec{b} = |\vec{a}||\vec{b}|\cos\theta$$

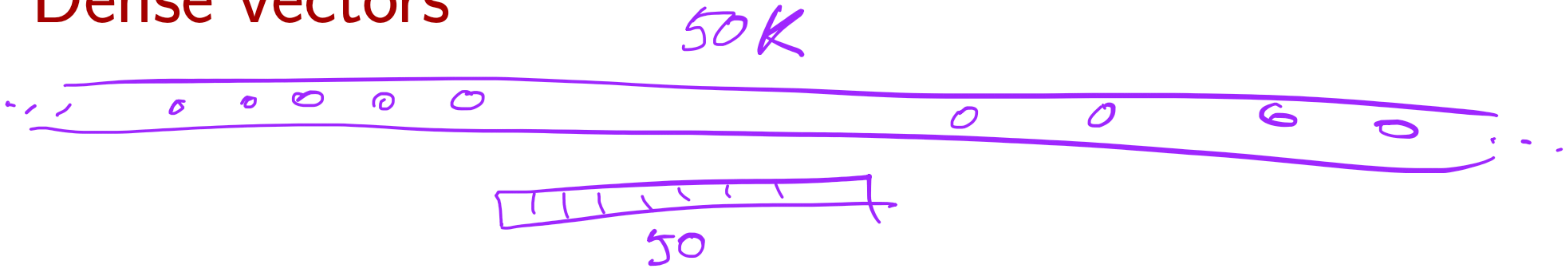
$$\boxed{\frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|} = \cos\theta}$$

$$\cos\alpha = \frac{adjacent}{hypotenuse} = \frac{b}{c}$$

Vector space models

- cosine:
  - happens to be useful measure of similarity
  - for **word** vectors

Dense vectors



- Easier to use as features in machine learning (less weights to tune)
- Generalize better than explicit counts
- May do better at capturing synonymy:
  - *car* and *automobile* are synonyms; but are represented as distinct dimensions; this captures that *car* and *automobile* are similar but fails to capture similarity between a word with *car* as a neighbor and a word with *automobile* as a neighbor

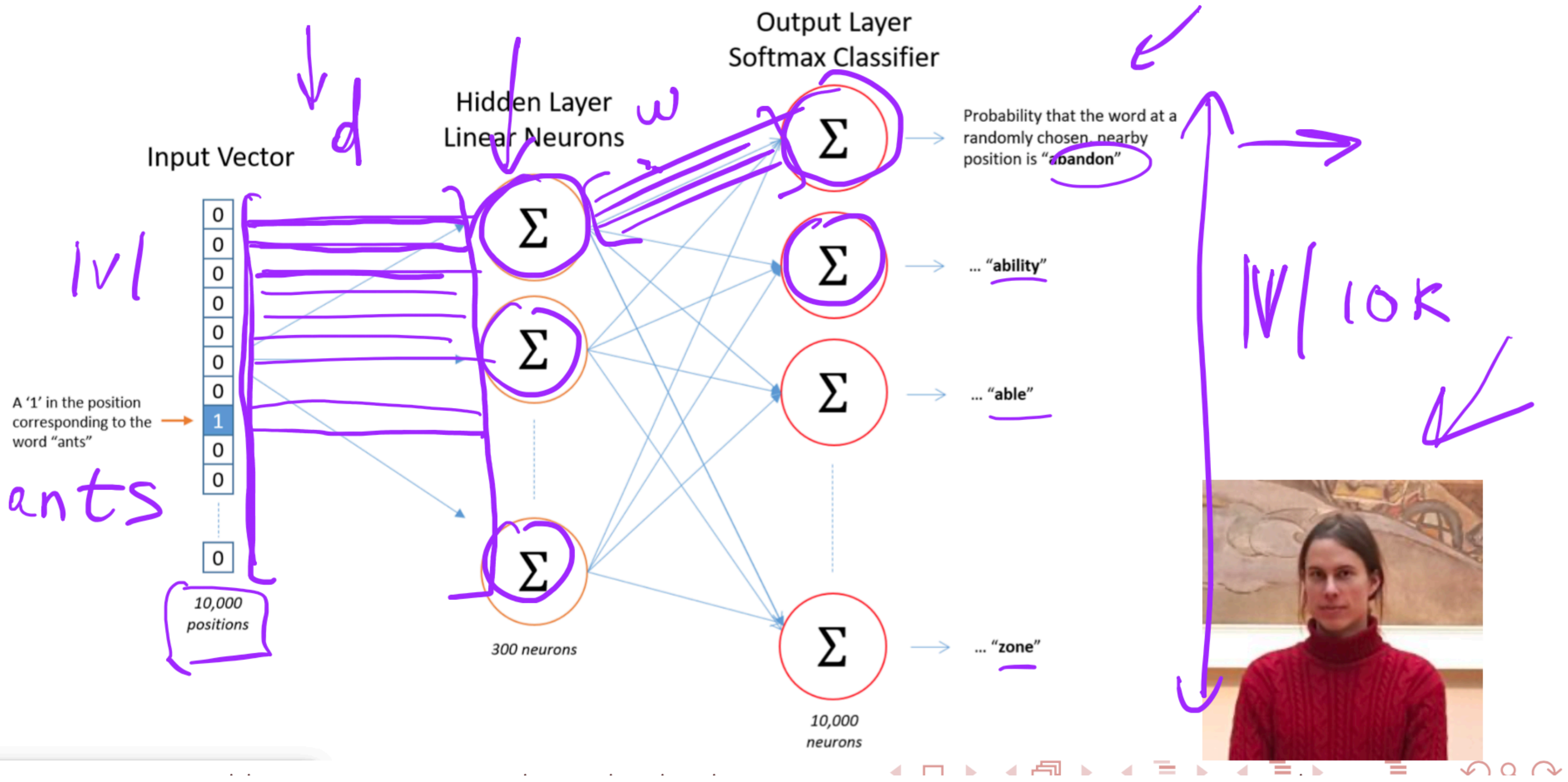
Dense vectors are a byproduct of neural LMs

(Simplified) neural models architecture

furry meows
[1 1]

- The *feed-forward* SkipGram model (Mikolov et al)
- Input: a word from the vocabulary
- Middle: two matrices and some matrix multiplication
- Output: a probability for each word in the vocabulary occurring *somewhere nearby* the input word

- Matrix2 is some coefficients/weights/ parameters
- Matrix 1 is…
  - "dense word embeddings"
  - !!!

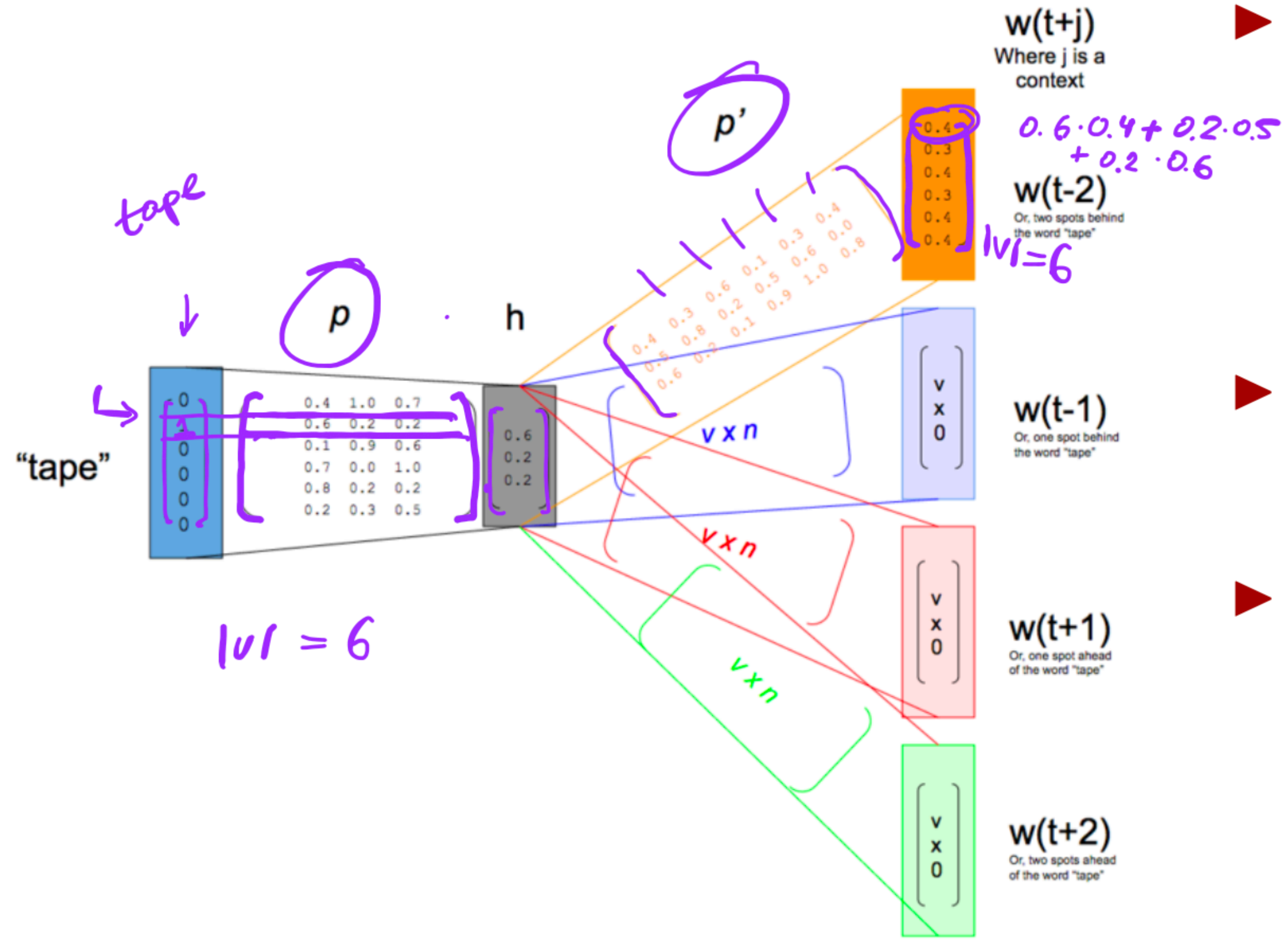


SkipGram training

- Input: a word
- Output: a probability distribution over the vocabulary
- **In the middle:** two *matrices*, “features” and “weights”
    - start with some random matrices
    - an input word is mapped to **some** vector (matrix row) at the start; call it the word vector
    - word vector is multiplied by weight matrix, the output is a vector of probabiltiies
    - iteratively find numbers for **both** the word vectors **and** the weights such that the output probabilities are “good enough”
    - unlike our cat example, the features are *learned* in the process along with feature weights

The SkipGram model in training

The e.g. orange output vector contains likelihood scores for each of the words in the vocabulary occurring 2 words before the word “tape”



w(t+j)
Where j is a context
p'
0.6·0.4+0.2·0.5 +0.2·0.6
w(t-2)
Or, two spots behind the word "tape"
|v|=6
tape
p
h
"tape"
v x n
w(t-1)
Or, one spot behind the word "tape"
v x n
w(t+1)
Or, one spot ahead of the word "tape"
v x n
|v| = 6
w(t+2)
Or, two spots ahead of the word "tape"

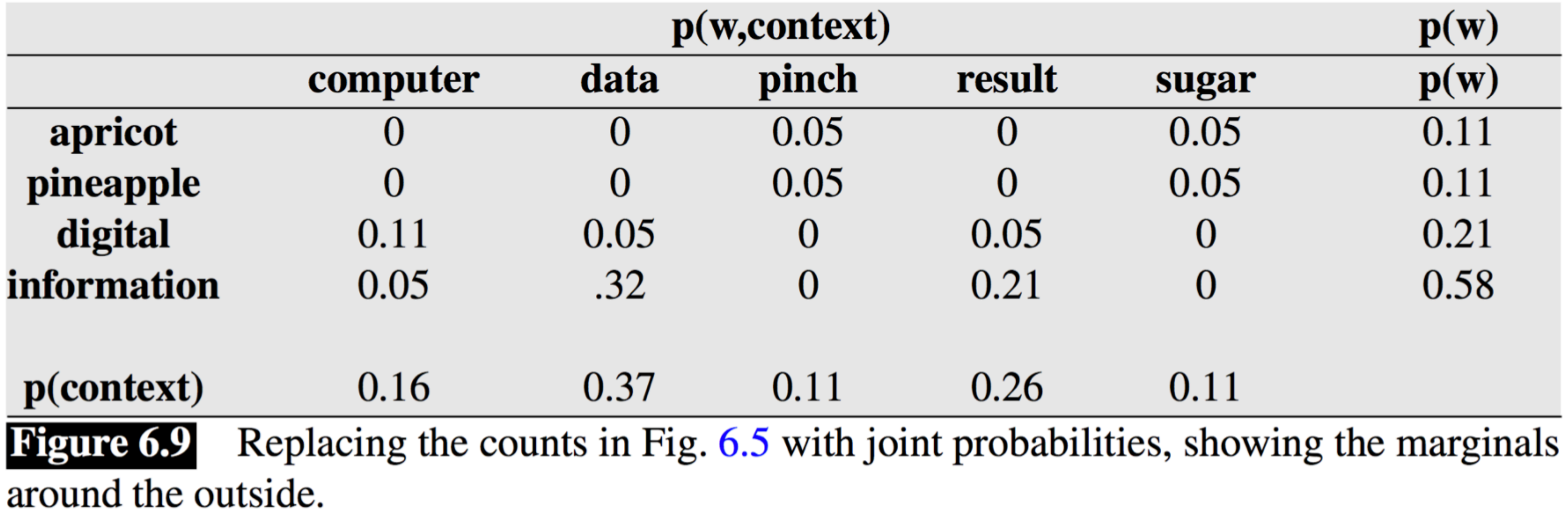
Sending w(t) = “tape” through the net once for context vector w(t-2) with randomized *p* and *p'* weight matrices

- computation is the *dot product*: $p \cdot h$, $h \cdot p'$
- (p’ has to be transposed)
- then need to map likelihood scores to actual probabilities (e.g. *softmax*)

Pic from: https://medium.com/district-data-labs/forward-propagation-building-a-skip-gram-net-from-the-ground-up-9578814b221

SkipGram training (simplified)

- Keep changing the p and p' matrices until the output probabilities are similar to the training corpus
- In the training corpus, count how many times a word occurs in the context of some other word and compute probabilities:



	p(w,context)					p(w)
	computer	**data**	**pinch**	**result**	**sugar**	**p(w)**
apricot	0	0	0.05	0	0.05	0.11
pineapple	0	0	0.05	0	0.05	0.11
digital	0.11	0.05	0	0.05	0	0.21
information	0.05	.32	0	0.21	0	0.58
p(context)	0.16	0.37	0.11	0.26	0.11	

Figure 6.9 Replacing the counts in Fig. 6.5 with joint probabilities, showing the marginals around the outside.

- Observe that we train to output accurate *probabilities* but it is the same as to train words that occur in similar contexts to have *similar vector representations*

Deep Learning
and neural nets

- SkipGram is a **basic** neural LM
- **Deep** networks:
  - have **many** node layers
  - leading to **huge** numbers of parameters
  - and **all sorts of things** happening **between** the layers
  - and **expensive** computation
- Deep Learning **architectures** change every year
- We are **not** experimenting with them in this class due to time it would take to train anything
  - but they **really are** kind of like the XOR :)

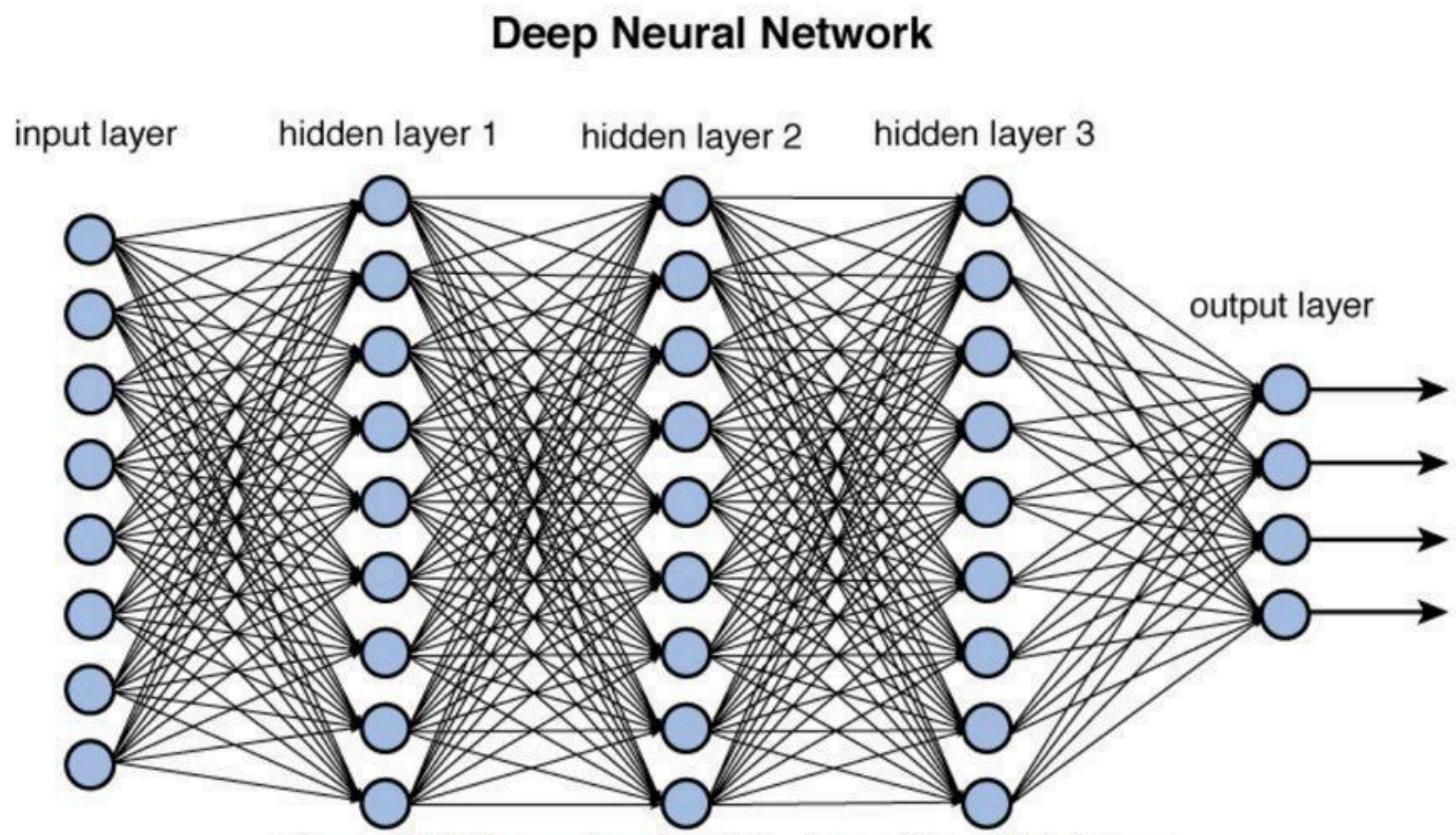
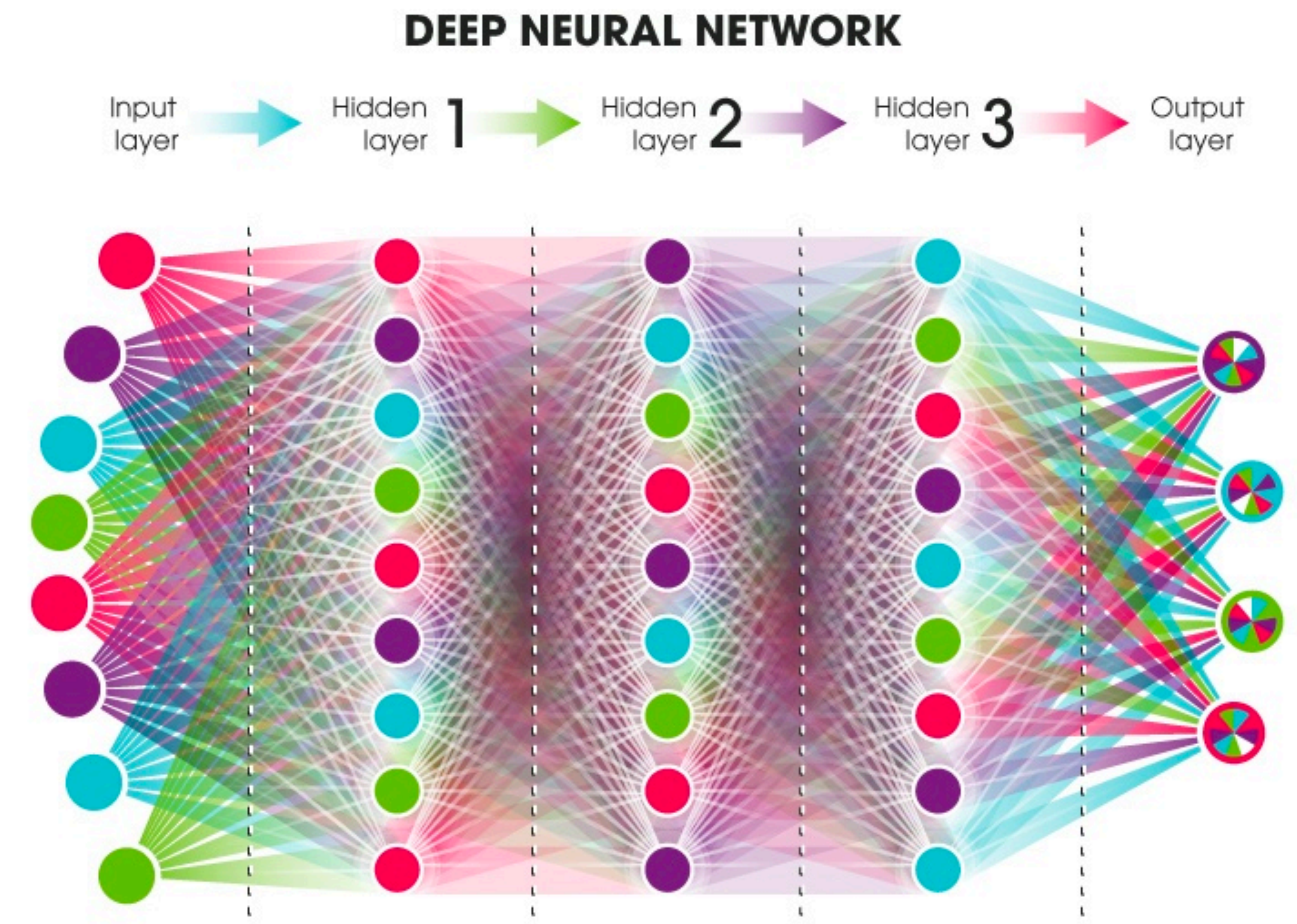


Figure 12.2 Deep network architecture with multiple layers.

https://towardsdatascience.com/training-deep-neural-networks-9fdb1964b964



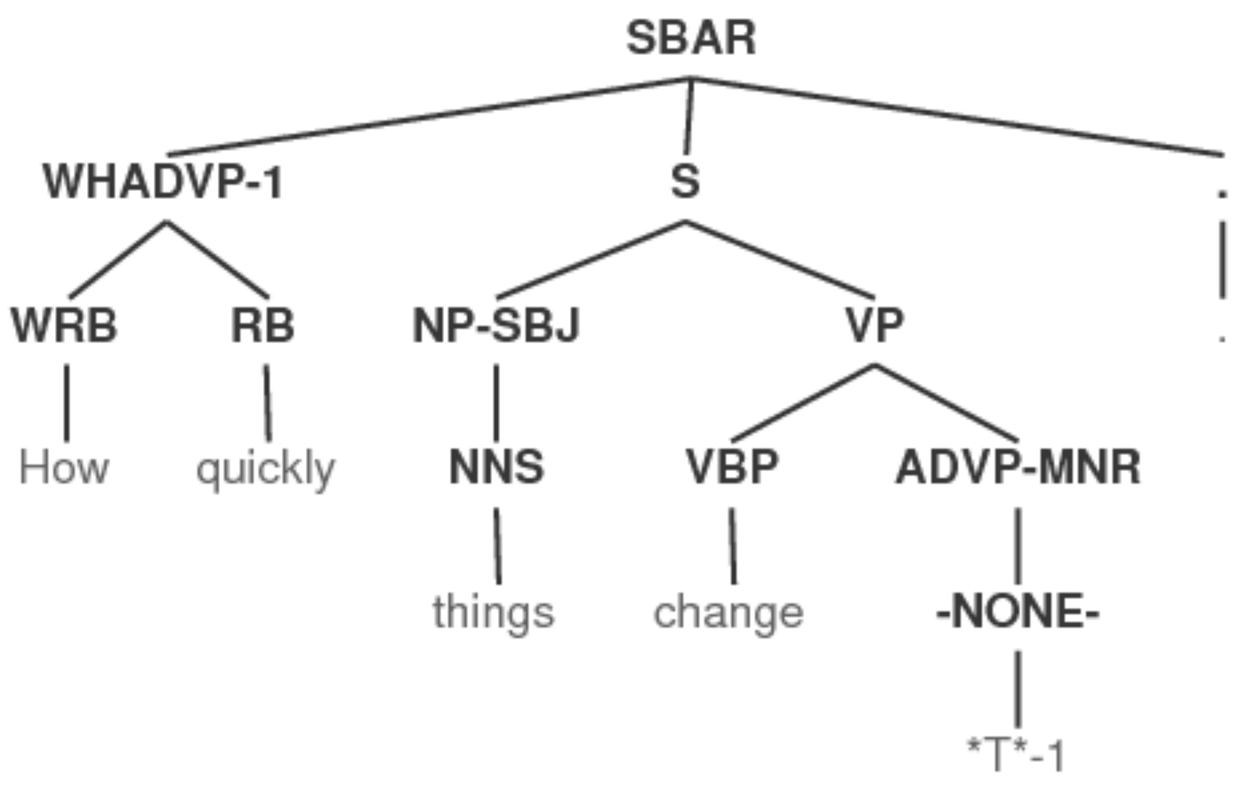
neuralnetworksanddeeplearning.com - Michael Nielsen, Yoshua Bengio, Ian Goodfellow, and Aaron Courville, 2016.

Conclusion: LMs and linguistic knowledge

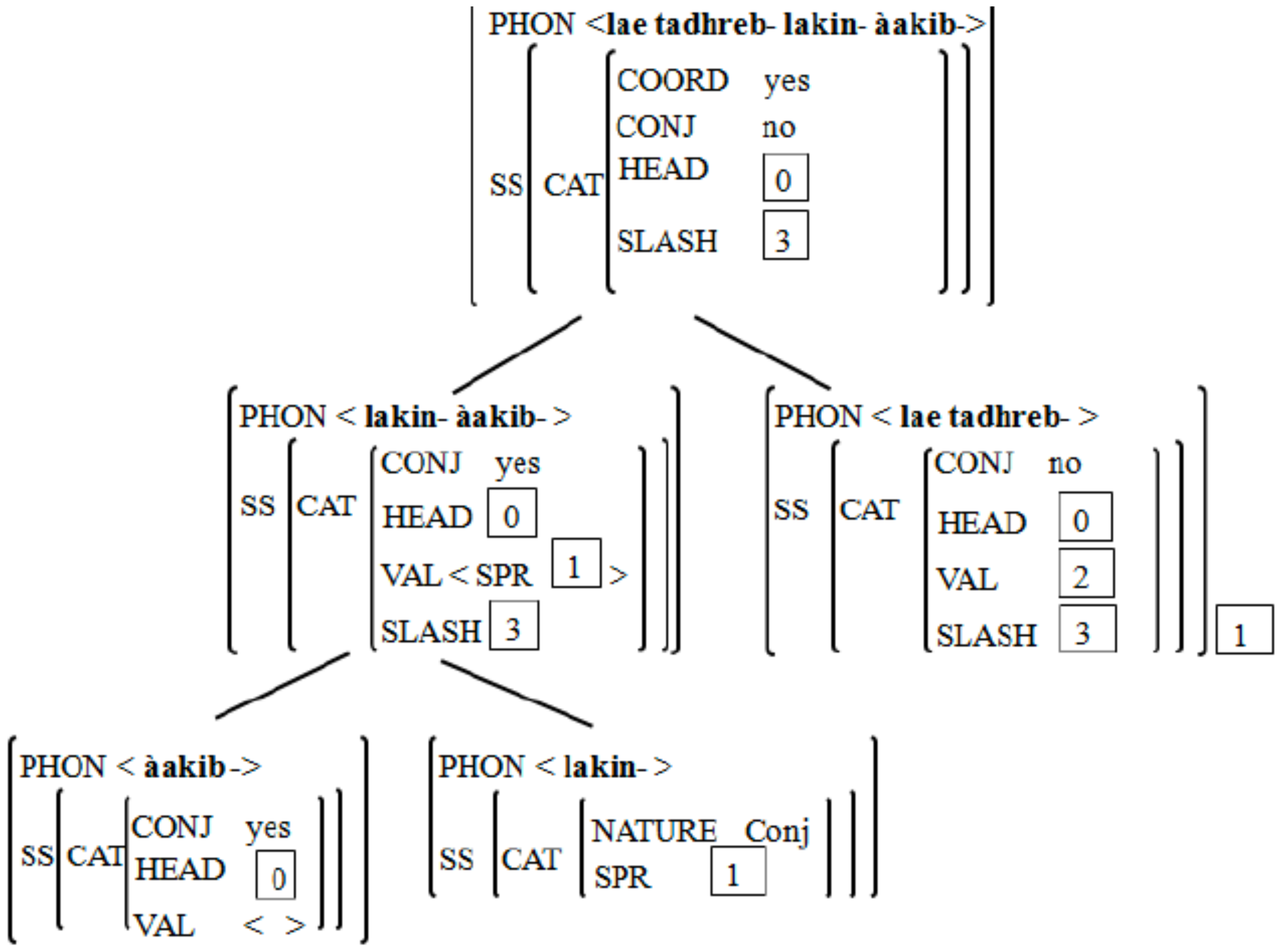
- Statistical and neural LMs are very successful in NLP
- They capture some **surface** information about the language (including the “world knowledge” that is on the surface)
- What about deeper structure, explanations, reasons of phenomena?

Linguistic knowledge in NLP
sample types for text

- Morphological analysis
- Syntax grammars
  - The Penn TreeBank was used to train numerous NLP systems
  - …which are in turn still often used today in pipelines
- Semantic representations



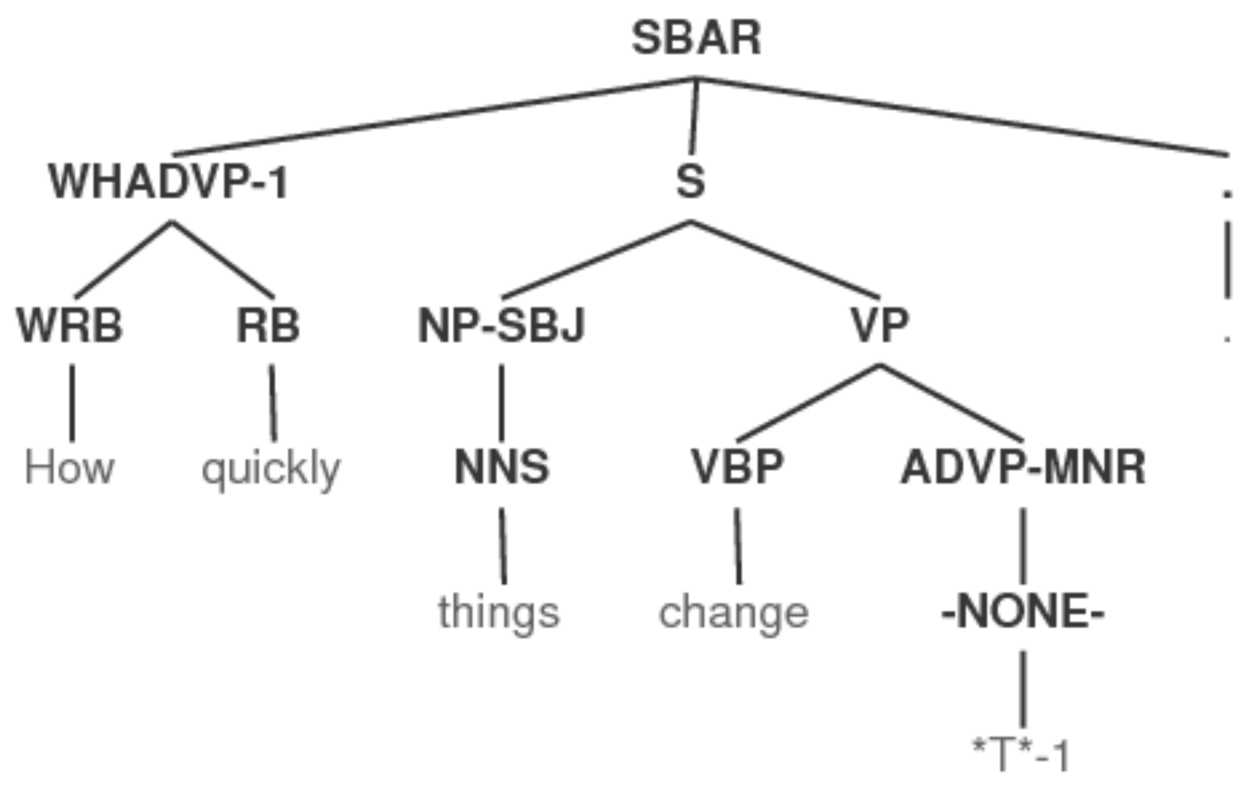
Penn Treebank (Marcus et al. 1993)



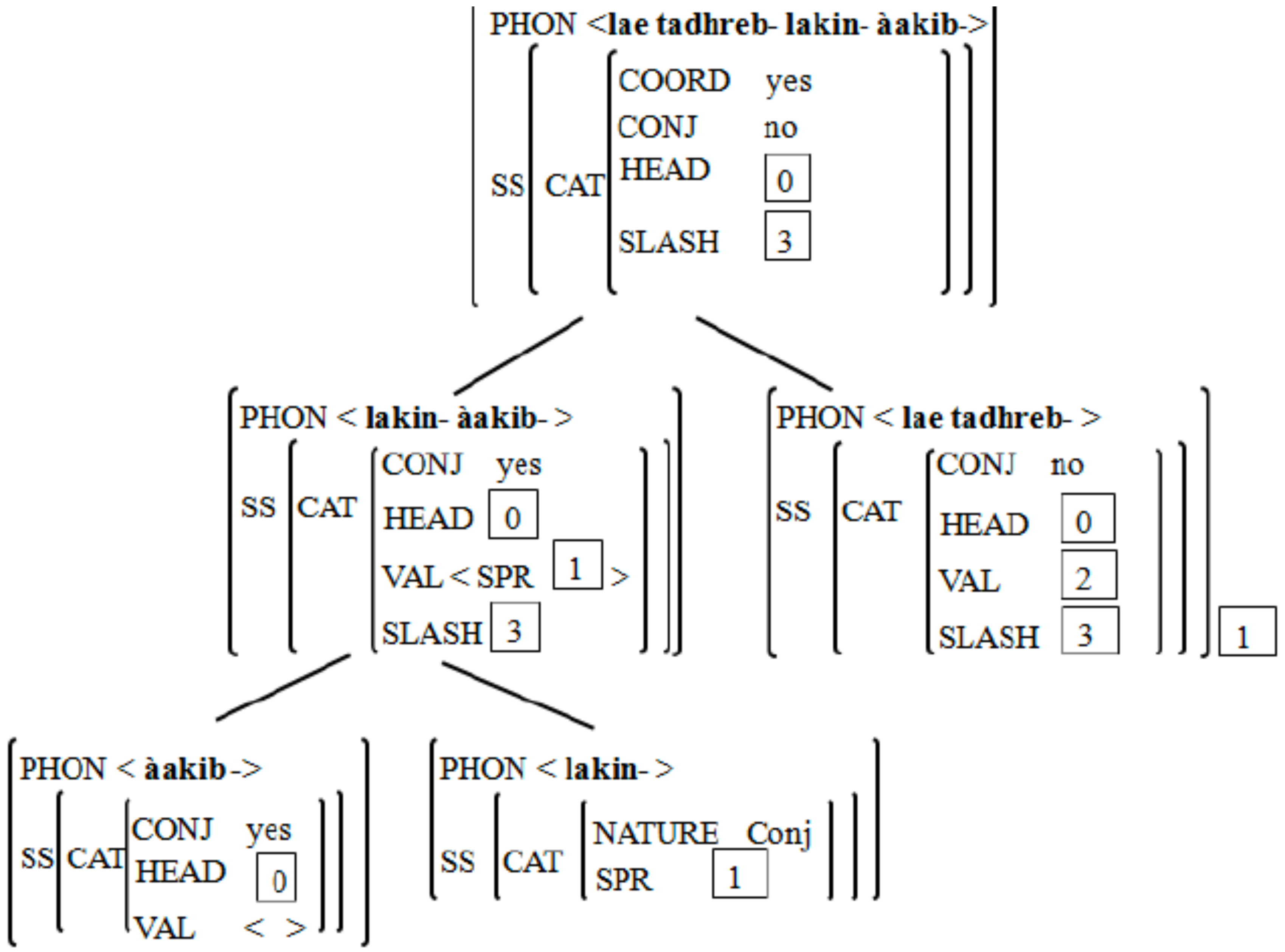
An HPSG tree (boukedi and Haddar 2014)

Linguistic knowledge in NLP
is PTB a theory?

- Yes.
- PTB may not stand up to HPSG or Minimalism in its elegance or power to generalize
- ...but it was conceptualized by people
- and as such represents linguistic knowledge
  - people sometimes forget that



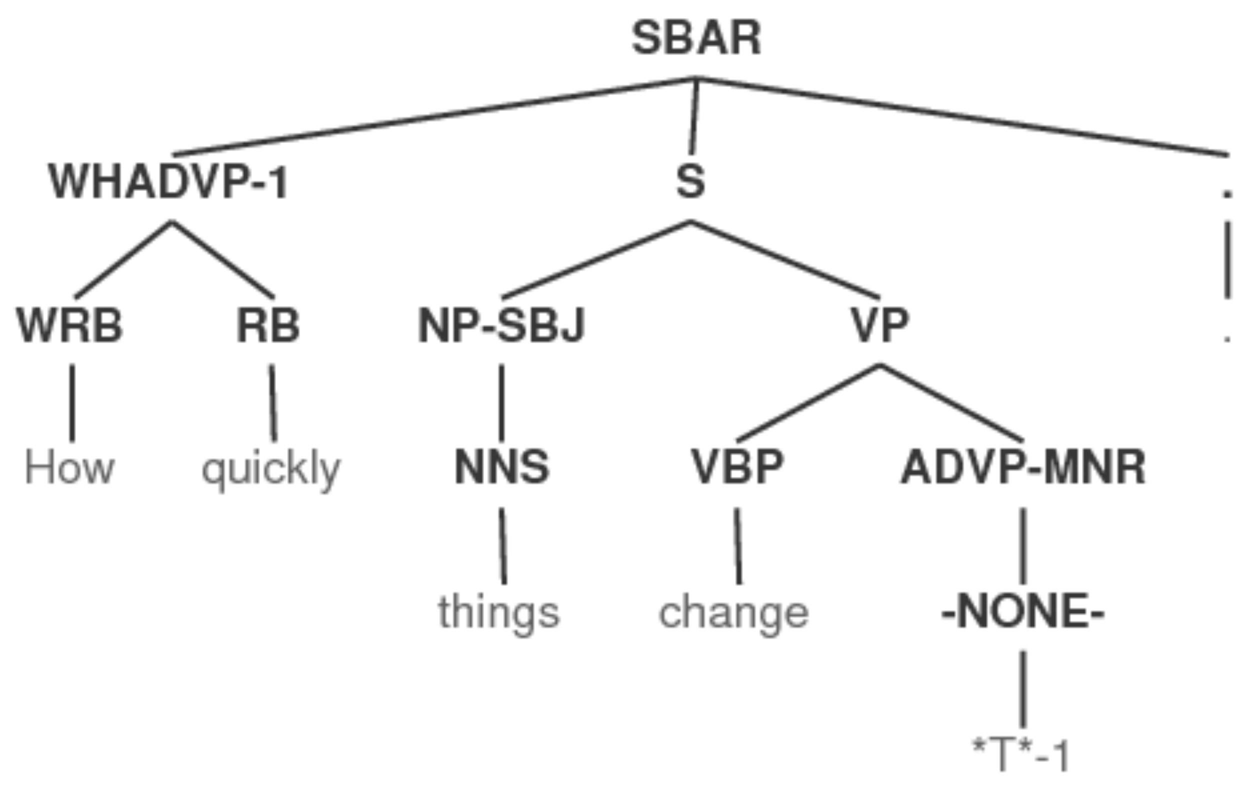
Penn Treebank (Marcus et al. 1993)



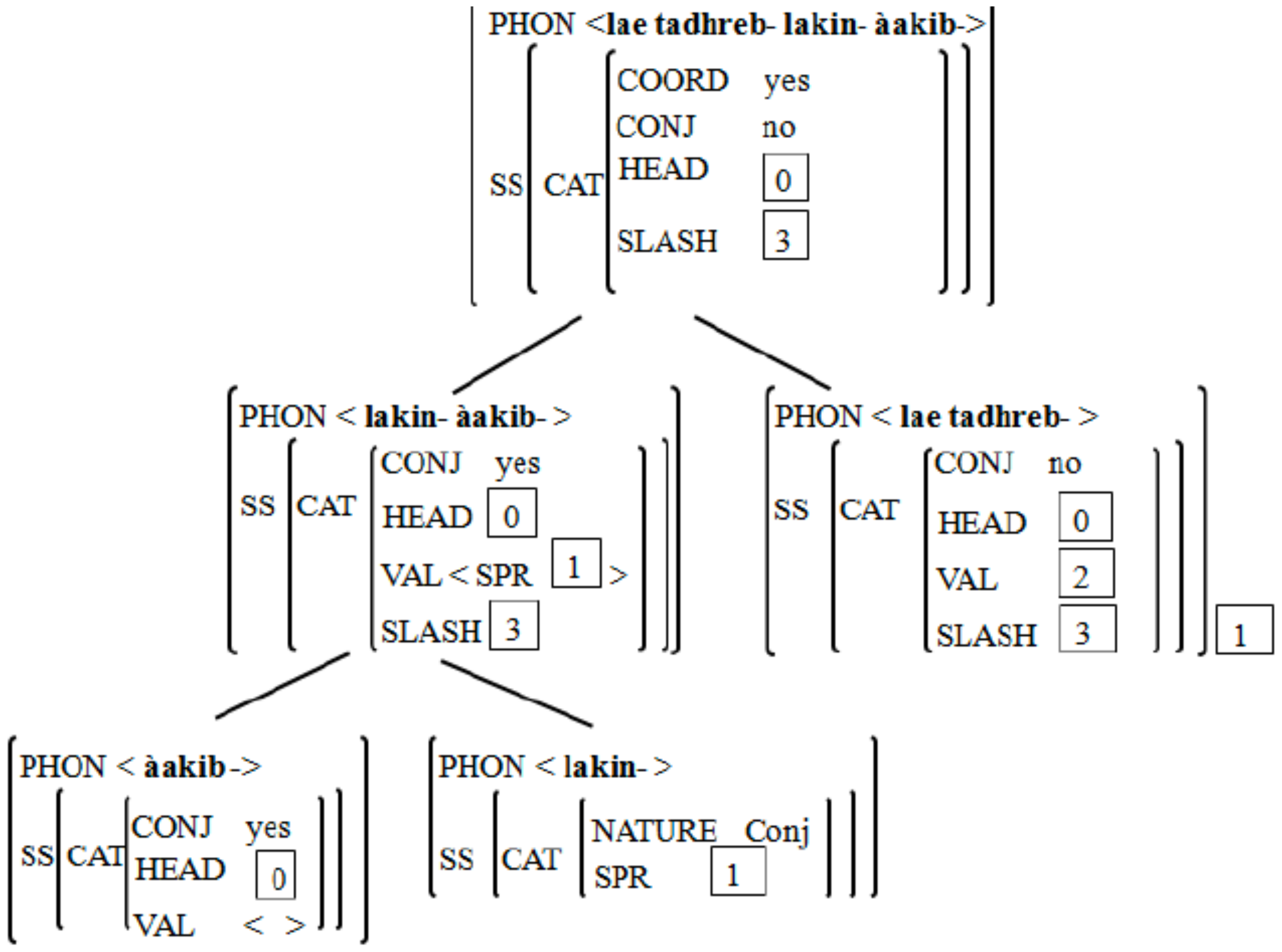
An HPSG tree (boukedi and Haddar 2014)

Linguistic knowledge in NLP

- How important is all of this?
  - in HW5: add different types of linguistic preprocessing to the data
  - observe no difference
  - Does it mean linguistics is not important?
  - No. It means linguistics is hard to use properly
  - ...which may make it less useful short term
  - but NLP relies on linguistic knowledge all the way and will continue to do so



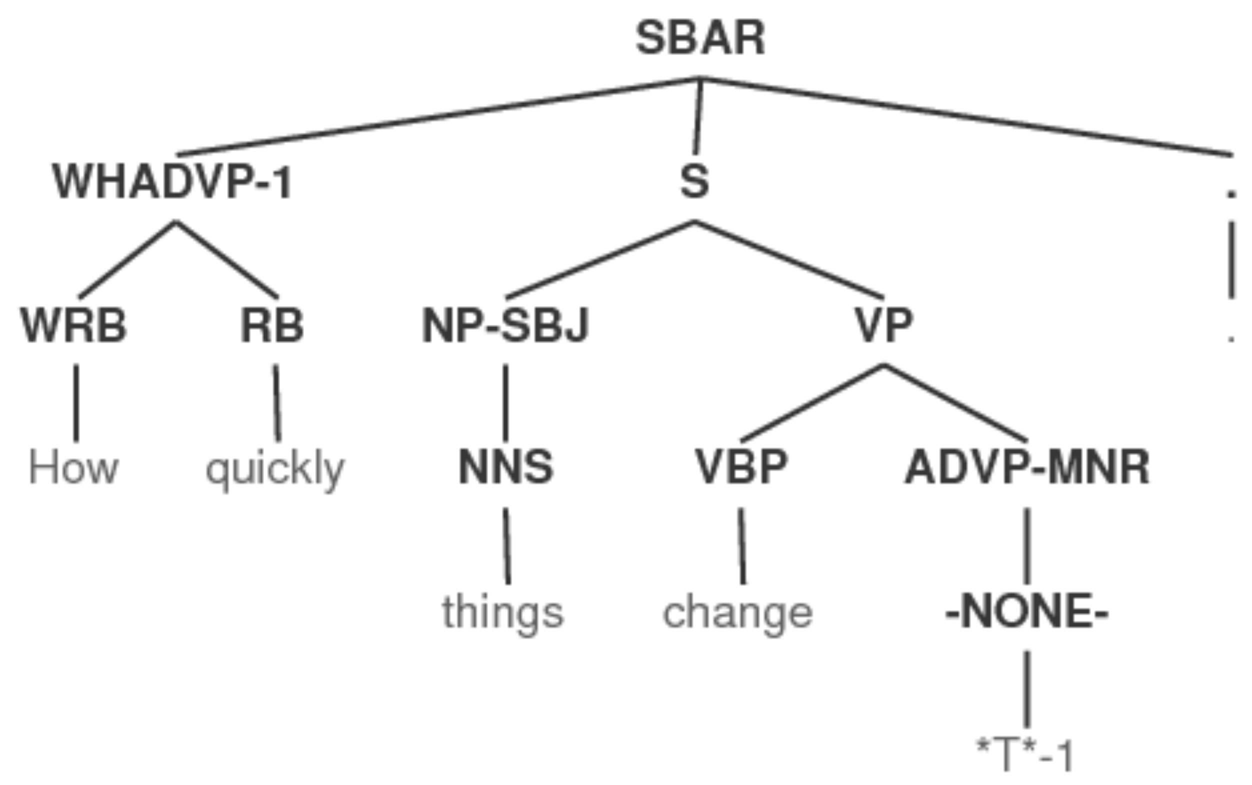
Penn Treebank (Marcus et al. 1993)



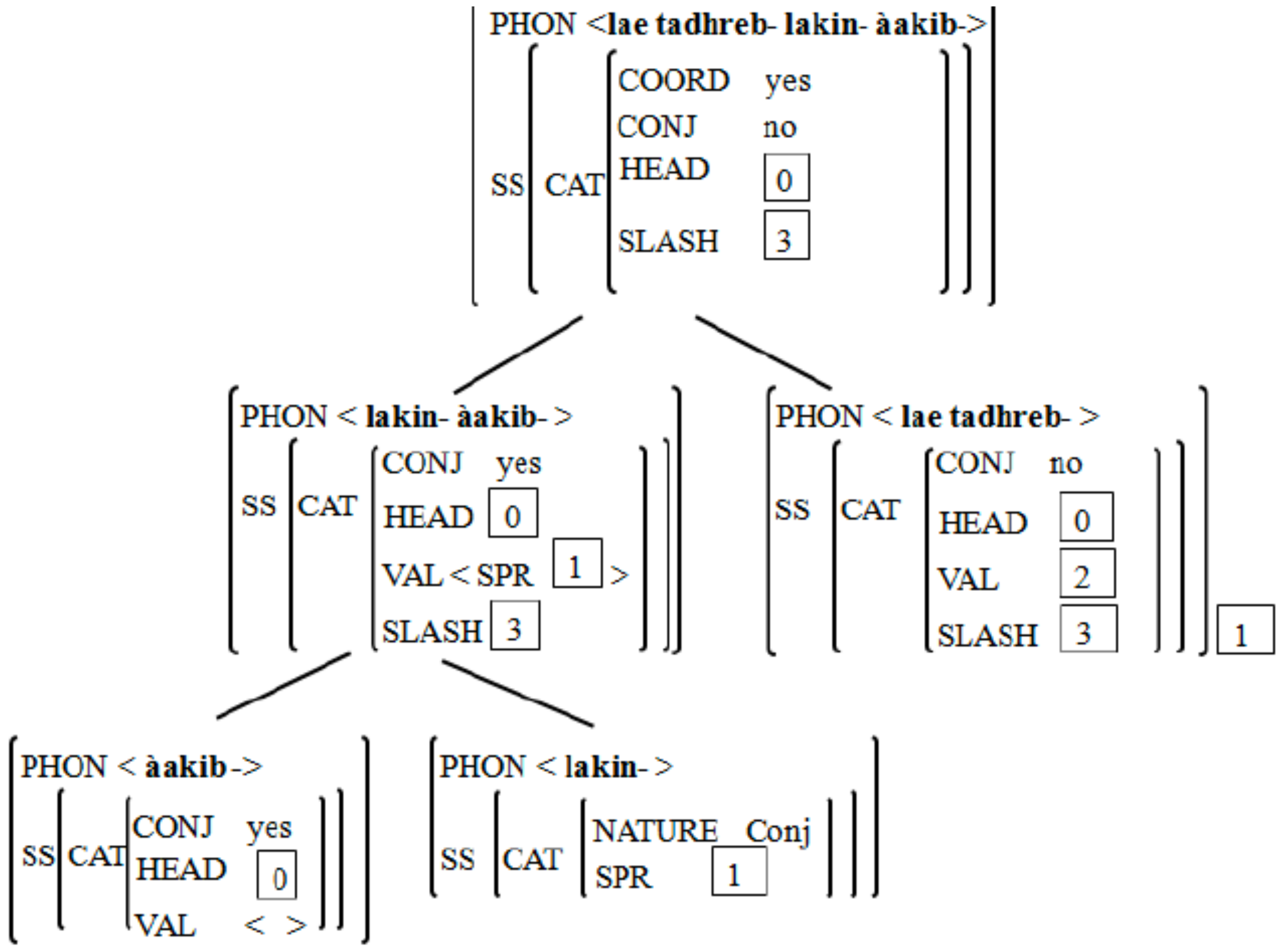
An HPSG tree (boukedi and Haddar 2014)

Linguistic knowledge in NLP

- Classical machine learning requires annotation
  - which requires knowledge
- What about Deep Learning?
  - DL learns features automatically
  - but still requires annotations for evaluation
  - ...also, the ever-improving metrics may be meaningless
  - and the question of whether we train linguistically competent systems remains open
- Whether or not we need to reason about systems is a philosophical question
  - going back to at least Plato and Democritus
  - Idealism and materialism
  - Rationalism and empiricism



Penn Treebank (Marcus et al. 1993)



An HPSG tree (boukedi and Haddar 2014)

Conclusion: LMs and linguistic knowledge

“What comes out of a 4-gram model of Shakespeare looks like Shakespeare because it is Shakespeare.”

D. Jurafsky

”In short, there is no free lunch – no way to generalize beyond the specific training examples, unless the learner commits to some additional assumptions.”

T. Mitchell

Some philosophy

Idealism and Materialism

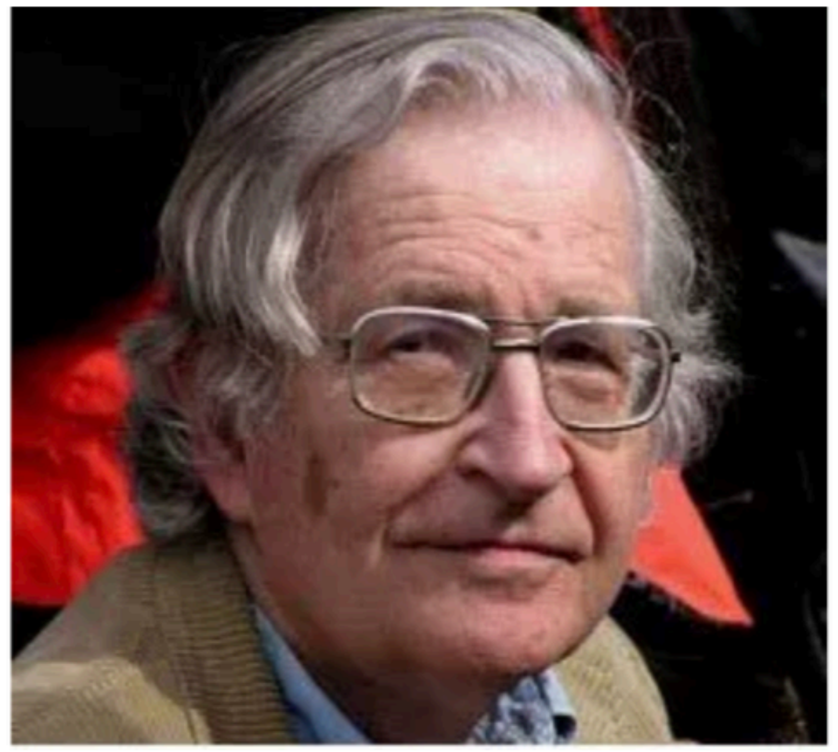
ancient schools of thought

- Idealism:
  - to reason about phenomena, need an idea first
- Materialism:
  - ideas/reasoning emerge from matter
  - …(or from data :))

School of Athens by Raphael

(Beyond LMs) The role of statistics

https://www.tor.com/2011/06/21/norvig-vs-chomsky-and-the-fight-for-the-future-of-ai/ (Kevin Gold's overview)



Chomsky: *To produce a statistically based simulation of ... a [bee] dance without attempting to understand why the bee behaved that way... is ...a notion of [scientific] success that's very novel. I don't know of anything like it in the history of science.*

(Beyond LMs) The role of statistics

https://www.tor.com/2011/06/21/norvig-vs-chomsky-and-the-fight-for-the-future-of-ai/ (Kevin Gold's overview)

Norvig: *Engineering success correlates with scientific success*

Linguistic knowledge in NLP
the great pendulum

- 1950s: Empiricism
- 1970s: Rationalism
- 1990s: Empiricism
- 2010s: Return to Rationalism?
    - (no)
- 2020s: Return to Rationalism?
- ?..

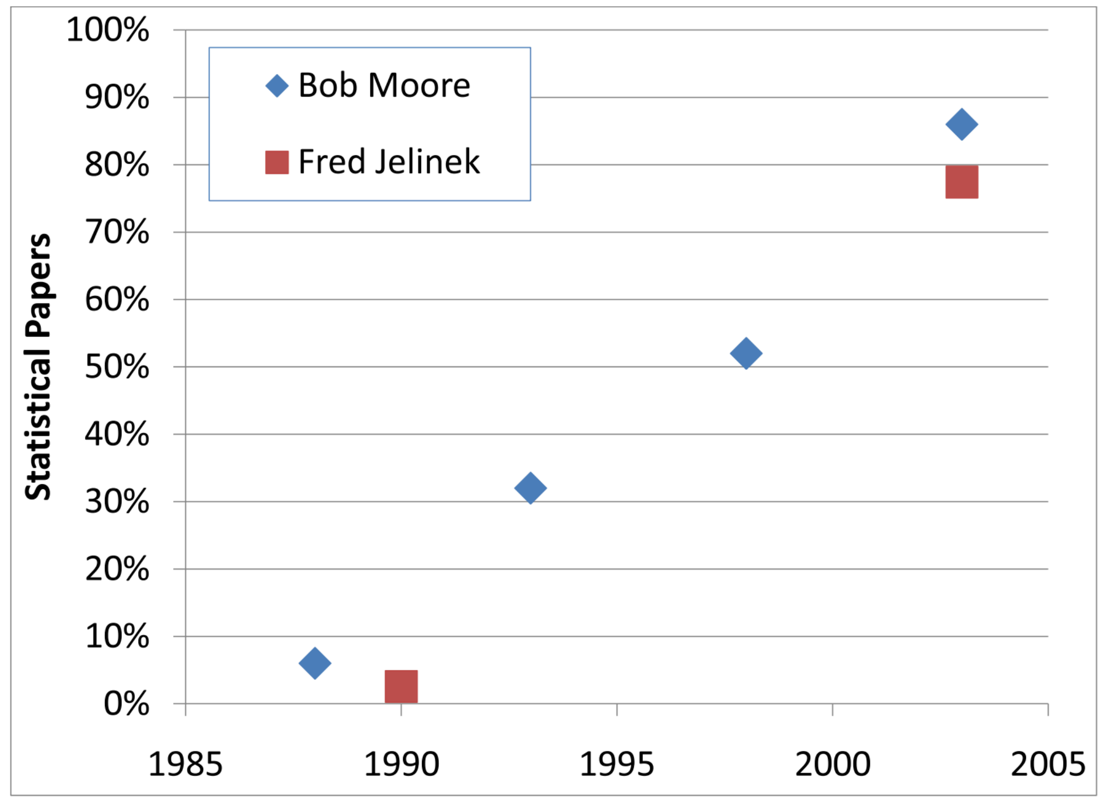
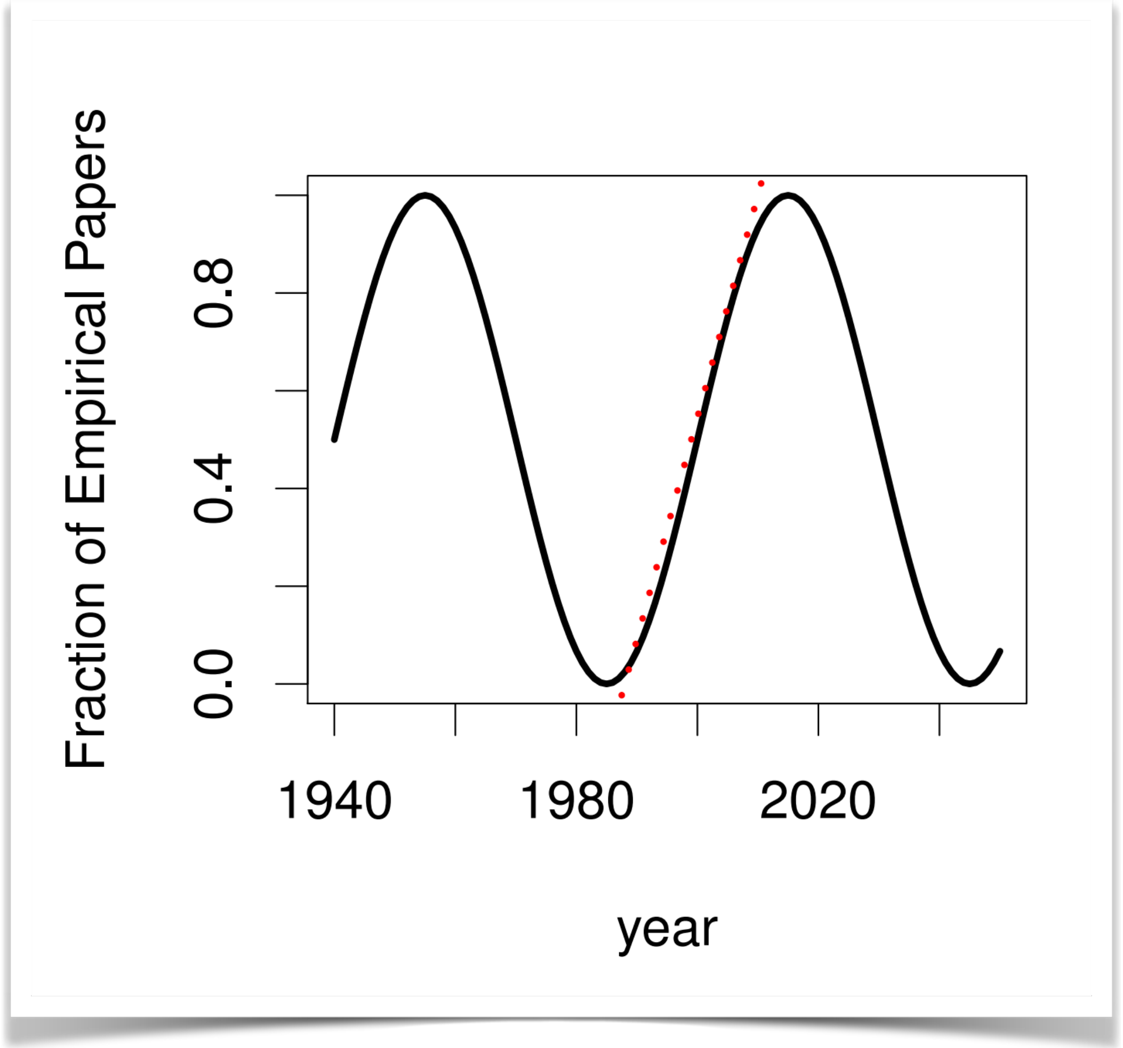


FIGURE 1 The shift from Rationalism to Empiricism is striking (and no longer controversial). This plot is based on two independent surveys of ACL meetings by Bob Moore and Fred Jelinek (personal communication).



Kenneth Church 2011. A Pendulum Swung Too Far

More philosophy,
by Julian Michael (UW CSE)

Syntactic Structures

Noam Chomsky, 1957

Chomsky: “Neither (a) 'colorless green ideas sleep furiously' nor (b) 'furiously sleep ideas green colorless', nor any of their parts, has ever occurred in the past linguistic experience of an English speaker. But (a) is grammatical, while (b) is not.”

Norvig: Actually, each bigram *has* occurred. And Pereira (2001) provided a model (based on word classes) which **does** distinguish correctly between the two.

What’s going on here is a **failure of imagination** on Chomsky’s part: just because you can’t *think* of a system which will generalize correctly **does not mean it cannot exist.**

slide from Julian Michael 2021

On Chomsky and the Two Cultures of Statistical Learning

Peter Norvig, 2011

Norvig: “There are usages which are rare in a language, but cannot be dismissed if one is concerned with actual data. For example, the verb quake is listed as intransitive in dictionaries, meaning that (1) below is grammatical, and (2) is not, according to a categorical theory of grammar.

1. The earth quaked.

2. ? It quaked her bowels.

But (2) actually appears as a sentence of English. This poses a dilemma for the categorical theory. When (2) is observed we must either arbitrarily dismiss it as an error that is outside the bounds of our model (without any theoretical grounds for doing so), or we must change the theory to allow (2), which often results in the acceptance of a flood of sentences that we would prefer to remain ungrammatical.”

My* take: this is a caricature of syntax as it is practiced. We have to go deeper.

*slide from Julian Michael 2021

On Chomsky and the Two Cultures of Statistical Learning

Peter Norvig, 2011

4. Norvig attributes their disagreement to the gap between Leo Breiman's "two cultures of statistical learning." Norvig (emphasis mine*):

- **The data modeling culture** (to which, Breiman estimates, 98% of statisticians subscribe) holds that nature can be described as a black box that has a relatively simple underlying model which maps from input variables to output variables (with perhaps some random noise thrown in). **It is the job of the statistician to wisely choose an underlying model that reflects the reality of nature**, and then use statistical data to estimate the parameters of the model.

- **The algorithmic modeling culture** (subscribed to by 2% of statisticians and many researchers in biology, artificial intelligence, and other fields that deal with complex phenomena), which holds that nature's black box cannot necessarily be described by a simple model. Complex algorithmic approaches (such as support vector machines or boosted decision trees or deep belief networks) are used to estimate the function that maps from input to output variables, but **we have no expectation that the form of the function that emerges from this complex algorithm reflects the true underlying nature**.

*slide from Julian Michael 2021

On the Role of Scientific Thought

Edsger W. Dijkstra, 1974

Dijkstra: “Let me try to explain to you, what to my taste is characteristic for all intelligent thinking. It is, that one is willing **to study in depth an aspect of one's subject matter in isolation for the sake of its own consistency, all the time knowing that one is occupying oneself only with one of the aspects**… It is what I have sometimes called **‘the separation of concerns’**, which, even if perfectly possible, is yet the only available technique for effective ordering of one’s thoughts, that I know of.”

“**A scientific discipline separates a fraction of human knowledge from the rest: we have to do so**, because, compared with what could be known, we have very, very small heads. It also separates a fraction of the human abilities from the rest; again, we have to do so, because the maintenance of our non-trivial abilities requires that they are exercised daily and a day — regretfully enough — has only 24 hours.”

slide from Julian Michael 2021

The ‘Quake’ Example

Due to Peter Norvig (2011)

1. The earth quaked.

2. ? It quaked her bowels.

When we say ‘quake’ is intransitive, we intend a **separation of concerns.**

- For example, we may understand this case as involving a *causative transformation* on the intransitive *quake*, and posit restrictions on the use of this transformation, a (perhaps discrete/categorical) markedness hierarchy for such uses, etc. — **here, it is Norvig who displays a lack of imagination w.r.t. modeling.**

Separation of concerns, furthermore, is the only way to **order one’s thoughts**, or a theory.

It is also **the only way we can express generalizations precisely**.

slide from Julian Michael 2021

On Chomsky and the Two Cultures of Statistical Learning

Peter Norvig, 2011

Chomsky seems to be objecting to the algorithmic modeling culture: no claim to represent nature means no insight into “why.” Norvig (and Breiman) disagree:

> **Basically, the conclusions made by data modeling are about the model, not about nature… The problem is, if the model does not emulate nature well, then the conclusions may be wrong.** For example, linear regression is one of the most powerful tools in the statistician's toolbox. Therefore, many analyses start out with "Assume the data are generated by a linear model..." and lack sufficient analysis of what happens if the data are not in fact generated that way… **Breiman is inviting us to give up on the idea that we can uniquely model the true underlying form of nature's function from inputs to outputs. Instead he asks us to be satisfied with a function that accounts for the observed data well, and generalizes to new, previously unseen data well, but may be expressed in a complex mathematical form that may bear no relation to the "true" function's form (if such a true function even exists).** Chomsky takes the opposite approach: he prefers to keep a simple, elegant model, and give up on the idea that the model will represent the data well. Instead, he declares that what he calls performance data—what people actually do—is off limits to linguistics; what really matters is competence—what he imagines that they should do.

slide from Julian Michael 2021

We have seen this before.

slide from Julian Michael 2021

Computing Machinery and Intelligence

Alan Turing, *Mind*, 1950

- Written to put to bed the question: *“Can machines think?”*
- Turing says the question is “too meaningless to deserve discussion.” Instead, he proposed we identify the testable behavior we associate with “thinking” (or intelligence), and test *that*.
    - His *Imitation Game* was an example of such a test.
- The rub: **we should put metaphysical questions (of the “true” nature of things) aside, in exclusive favor of behavioral tests.** Naturally: test generalization to novel situations, i.e., *success on unanalyzed data*.

slide from Julian Michael 2021

Trouble Brewing with Turing and Breiman

The Generalization Crisis

Recall: “Breiman … asks us to be satisfied with a function that accounts for the observed data well, **and generalizes to new, previously unseen data well.**”

How exactly do we measure generalization?

- The Turing Test? **DONE** — Eugene Goostman, 2014
- The Winograd Schema Challenge? **DONE** — 96.6%, T5 + Meena (Google)
- Many other existing benchmarks are saturated:
    - SQuAD (Rajpurkar et al., 2016)
    - GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019)
- Yet, models remain brittle and display striking weak spots
- **This was not a problem when Norvig wrote his piece: our systems were too transparent (interpretable) and not accurate enough to convince anyone of anything**

slide from Julian Michael 2021

Challenge Sets and Targeted Evaluation

A Game of Whack-a-Mole

- New tests of generalization? **TRAIN THEM OUT?**
    - Every time we invent a new “challenge set” a model fails, we can compensate by training on it
- The only consistent, robust gains we can get are from **LARGE**-scale pretraining on unlabeled data
    - But we have only vague ideas of what the resulting systems actually learn and how they generalize
    - And getting them to work with **high reliability** presents some of the same challenges

slide from Julian Michael 2021

Climbing Towards NLU (the Octopus Paper)

Emily Bender & Alexander Koller, *ACL*, 2020

- Argues that the current trend of regarding large-scale pretrained models as “language understanding systems” is misguided
- Crux of the argument: “meaning” is a relation between linguistic form and some external system (in particular, communicative intents, or the world in which these intents are grounded). Since this system is external to form, the connection to it cannot be learned from form alone.
- Rephrased: **form itself lacks the “why” of language, which is *to communicate something.***
- Can you characterize this “why” purely behaviorally, in terms of predicting linguistic form?

slide from Julian Michael 2021

To Dissect an Octopus: Making Sense of the Form/Meaning Debate

Julian Michael, 2020

- If understanding means **connecting explicitly to an external system**, then language models clearly cannot exhibit it — they lack the API. But then the argument is trivial.

- If understanding means **interpreting and manipulating form in a way compatible with human understanding** (i.e., that evinces an underlying model of reality, meanings, and intents), then our behavioral-testing hats apply — so how can the argument apply categorically? Isn't it an empirical question?

- **This conundrum vexed Chomsky as well**, though he may not have seen it that way. **How might one address it?**

slide from Julian Michael 2021

A Pragmatist Approach

Reconciling Rationalism and Empiricism

- We cannot rely on the wisdom of the theoretician to establish the right categories by decree…

- But the data alone can tell no story of how to generalize.

- **Pragmatism:** acknowledge the contingent nature of our theories, wield them **nimbly**, and always ground them in **'cash value'** — their ability to explain phenomena which themselves have practical & theoretical significance

- **In practice?** We should be constructing theories **automatically** on the basis of data which has **ecological significance in lingusitics** using algorithms which **instantiate general principles** of language structure, meaning, & use.

slide from Julian Michael 2021

Lecture survey in the chat!