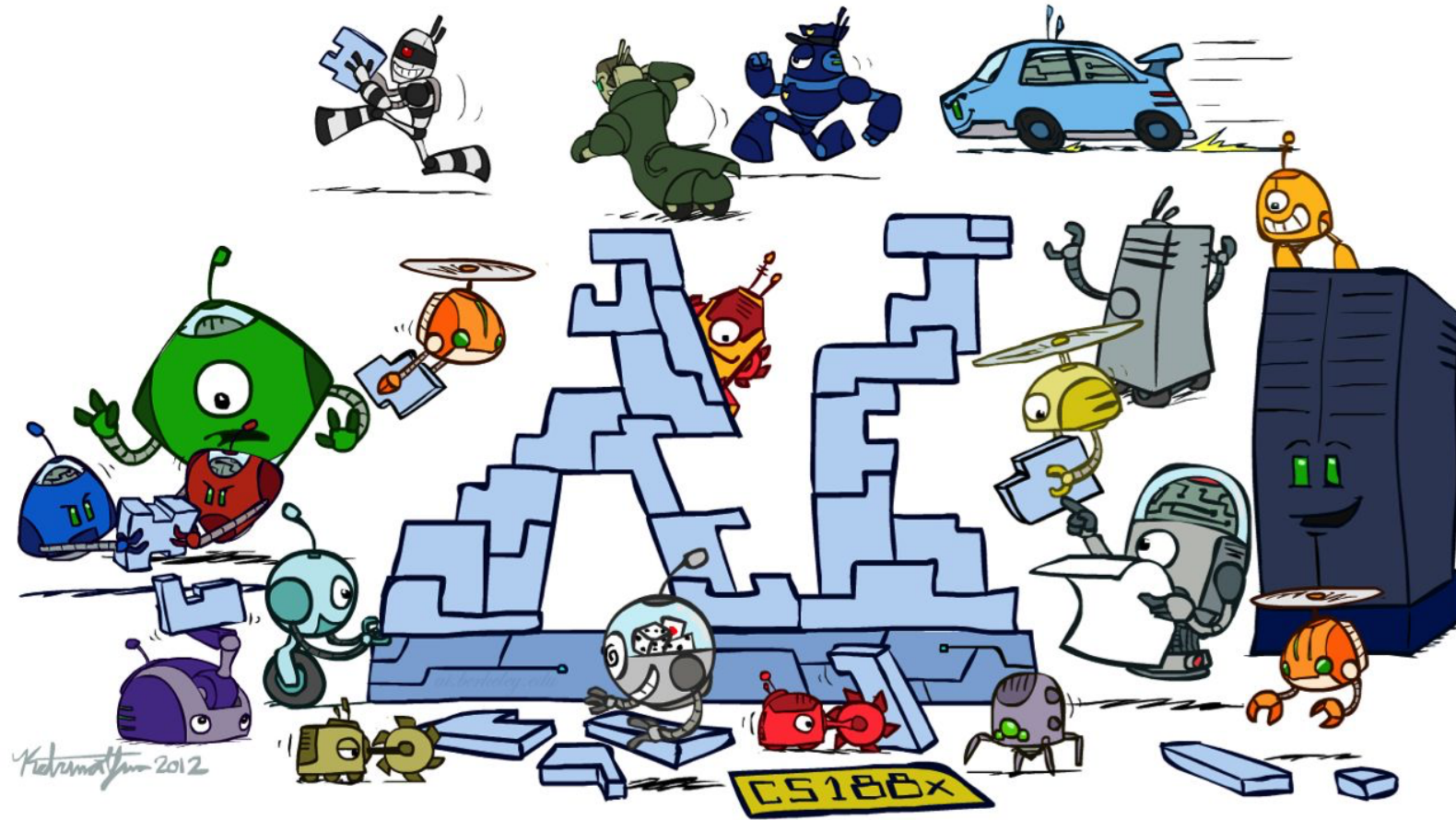


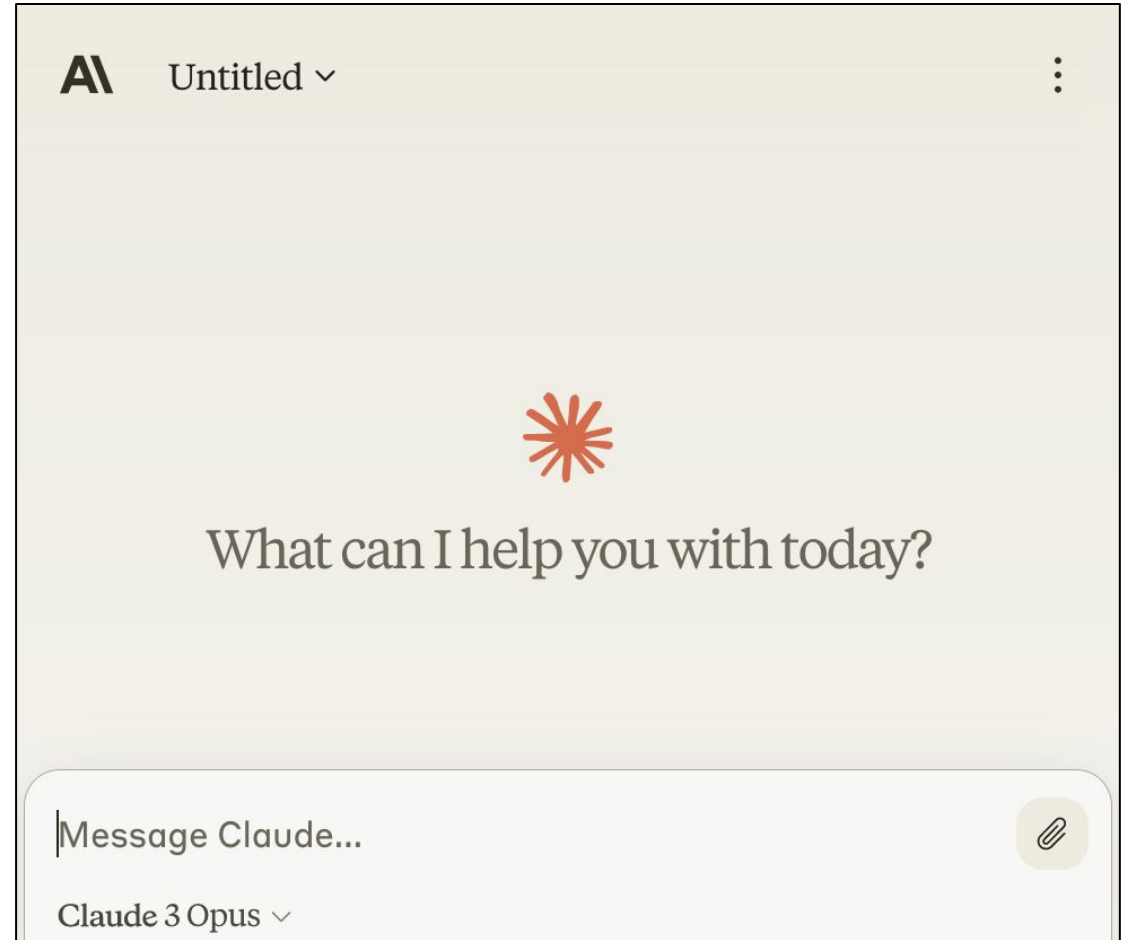
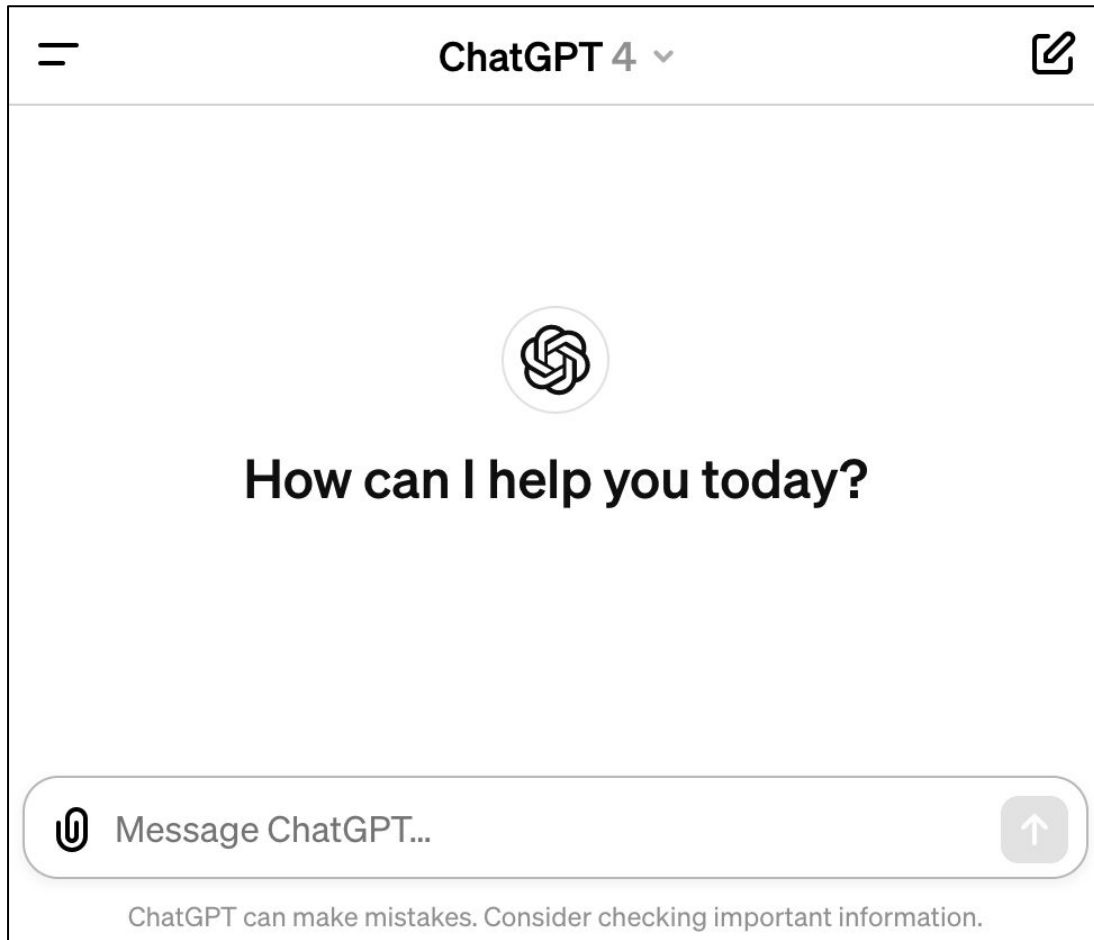
# Language Models I



[These slides were created by Cam Allen, Michael Cohen, Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.]

All CS188 materials are available at <http://ai.berkeley.edu>.]

# Today's AI



# Language Models

---



# Noisy Channel Models

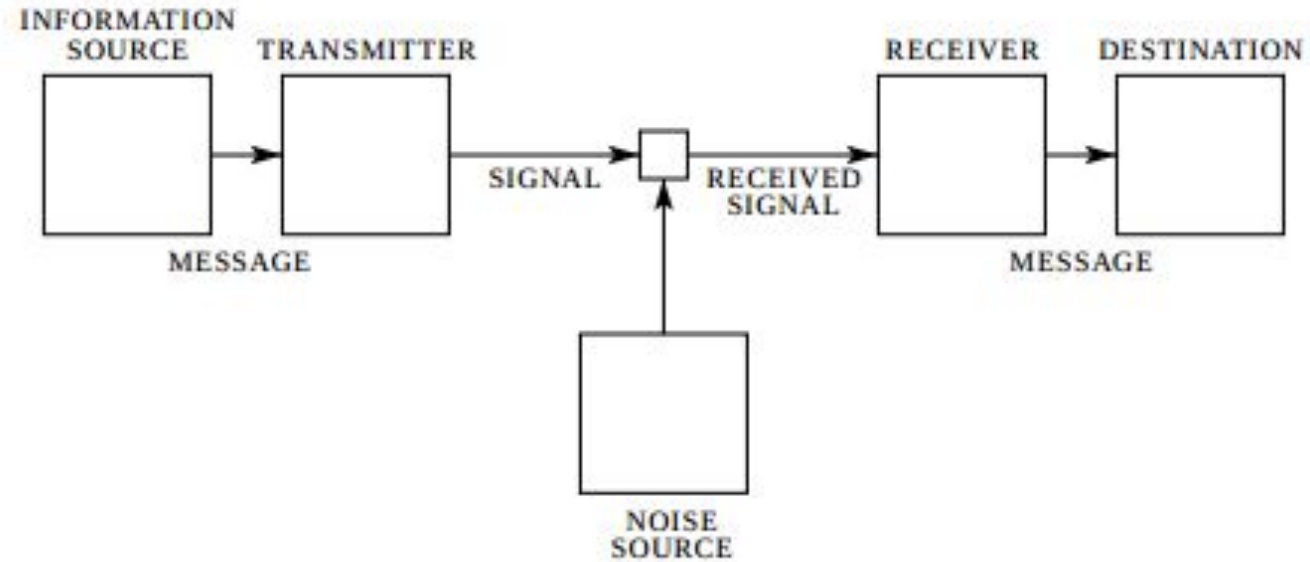
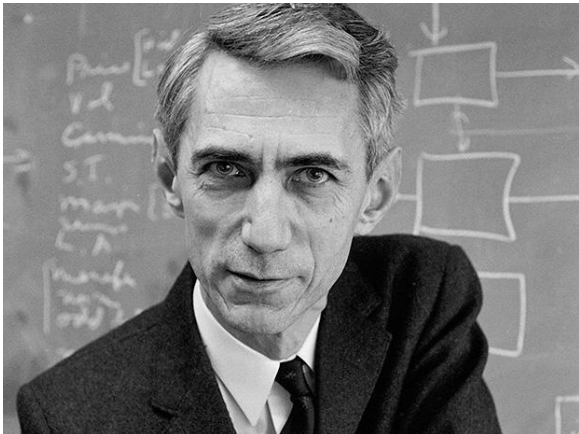


Fig. 1—Schematic diagram of a general communication system.





# Noisy Channel Model: ASR

We want to predict a sentence given acoustics:

$$w^* = \arg \max_w P(w|a)$$

The noisy channel approach:

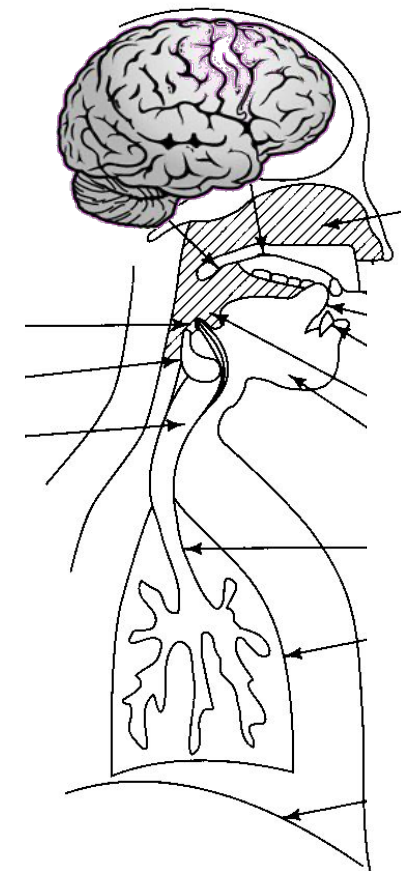
$$w^* = \arg \max_w P(w|a)$$

$$= \arg \max_w P(a|w)P(w)/P(a)$$

$$\propto \arg \max_w P(a|w)P(w)$$

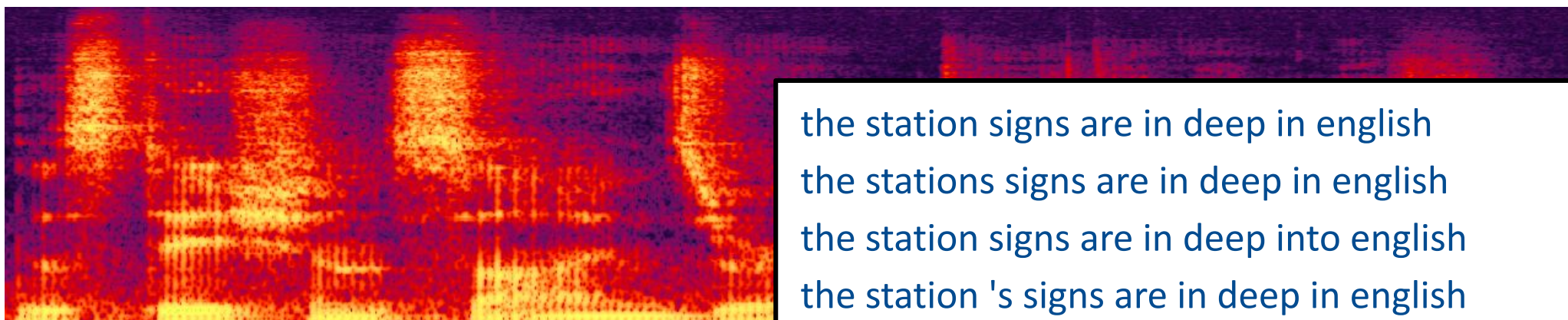
Acoustic model: score fit between  
sounds and words

Language model: score  
plausibility of word sequences





# Acoustic Confusions



the station signs are in deep in english	-14732
the stations signs are in deep in english	-14735
the station signs are in deep into english	-14739
the station 's signs are in deep in english	-14740
the station signs are in deep in the english	-14741
the station signs are indeed in english	-14757
the station 's signs are indeed in english	-14760
the station signs are indians in english	-14790



# Noisy Channel Model: Translation

---

“Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’ ”

Warren Weaver (1947)



# Model Quality: Likelihood

"when i eat pizza, i wipe off the"

Search results for "when i eat pizza, i wipe off the":

- <https://cal-cs288.github.io> › slides › PDF  
**SP20 CS288 -- Language Models (1) - GitHub Pages**  
 When I eat pizza, I wipe off the \_\_\_\_\_. Formally: test set log likelihood. Perplexity: "average per word branching factor" (not per-step) perp, = exp ...
- <https://people.eecs.berkeley.edu> › ~klein › slides › PDF  
**2PP - People @ EECS at UC Berkeley**  
 Unigrams are terrible at this game. (Why?) "Entropy": per-word test log likelihood (misnamed).  
 When I eat pizza, I wipe off the \_\_\_\_\_. Many children are allergic ...  
 2011
- <https://people.eecs.berkeley.edu> › ~klein › slides › PDF  
**Natural Language Processing - People @ EECS at UC Berkeley**  
 When I eat pizza, I wipe off the \_\_\_\_\_. Formally: define test set (log) likelihood. Perplexity: "average per word branching factor" (not per-step) perp X, exp.
- <https://courses.cs.washington.edu> › LanguageModels › PDF  
**CSEP 517 Natural Language Processing ... - Washington**  
 How good are we doing? Compute per word log likelihood (M words, m test sentences s i.):  
 When I eat pizza, I wipe off the \_\_\_\_\_. Many children are allergic to ...

actual word

er parameters  $\theta$

,  $\theta$ )

grease 0.5  
 sauce 0.4  
 dust 0.05  
 ....  
 mice 0.0001  
 ....  
 the 1e-100

3516 wipe off the excess  
 1034 wipe off the dust  
 547 wipe off the sweat  
 518 wipe off the mouthpiece  
 ...  
 120 wipe off the grease  
 0 wipe off the sauce  
 0 wipe off the mice  
 -----  
 28048 wipe off the \*

# N-Gram Models

---



# Generative Models

---

- Generative models describe a probability distribution over some structure, here a sequence of words.
- Commonly of the form: build sequence one by one, left to right

$$P(w_1 \dots w_n) = \prod_i P(w_i | w_1 \dots w_{i-1})$$

- You will also hear “autoregressive”: this term refers to example sequences being self-supervising examples for the function  $P(w | \text{context})$
- When trained to predict next words, models may capture many kinds of correlations



# N-Gram Models

---

- Use chain rule to generate words left-to-right

$$P(w_1 \dots w_n) = \prod_i P(w_i | w_1 \dots w_{i-1})$$

- Can't condition atomically on the entire left context

$P(??? \mid \text{The computer I had put into the machine room on the fifth floor just})$

- N-gram models make a Markov assumption

$$P(w_1 \dots w_n) = \prod_i P(w_i | w_{i-k} \dots w_{i-1})$$

$$P(\text{please close the door}) = P(\text{please} | \text{START}) P(\text{close} | \text{please}) \dots P(\text{STOP} | \text{door})$$



# Empirical N-Grams

- Use statistics from data (examples here from Google N-Grams)

Training Counts	198015222 the first
	194623024 the same
	168504105 the following
	158562063 the world
	...
	14112454 the door
	-----
23135851162 the *	

$$\hat{P}(\text{door}|\text{the}) = \frac{14112454}{23135851162} = 0.0006$$

- This is the maximum likelihood estimate, which needs modification
- N-gram models use such counts to compute probabilities on demand



# Increasing N-Gram Order

- Higher orders capture more correlations

## Bigram Model

198015222	the first
194623024	the same
168504105	the following
158562063	the world
...	
14112454	the door
-----	
23135851162	the *

$$P(\text{door} \mid \text{the}) = 0.0006$$

## Trigram Model

197302	close the window
191125	close the door
152500	close the gap
116451	close the thread
87298	close the deal
-----	
3785230	close the *

$$P(\text{door} \mid \text{close the}) = 0.05$$



# Increasing N-Gram Order

---

Unigram

- To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
- Every enter now severally so, let
- Hill he late speaks; or! a more to leg less first you enter
- Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like



# What's in an N-Gram?

---

- Just about every local correlation!
  - Word class restrictions: “it was very \_\_\_\_”
  - Morphology: “she \_\_\_\_”, “they \_\_\_\_”
  - Semantic class restrictions: “danced a \_\_\_\_”
  - Idioms: “add insult to \_\_\_\_”
  - World knowledge: “ice caps have \_\_\_\_”
  - Pop culture: “the empire strikes \_\_\_\_”
- But not the long-distance ones
  - “The **computer** which I had put into the machine room on the fifth floor just \_\_\_\_.”



# N-Grams on the Web

← → ↻ [https://corpora.linguistik.uni-erlangen.de/cgi-bin/demos/Web1T5/Web1T5\\_freq.perl?query=Berkeley+is+a+\\*&mode=Se...](https://corpora.linguistik.uni-erlangen.de/cgi-bin/demos/Web1T5/Web1T5_freq.perl?query=Berkeley+is+a+*&mode=Se...) ☆

Frequency list   Associations   Collocations

**The Google Web 1T 5-Gram Database – SQLite Index & Web Interface**

This is the Web interface of the [Web1T5-Easy package](#), using a [GOPHER](#) page design.  
(service provided by the [Corpus Linguistics group](#) at [FAU Erlangen-Nürnberg](#))

### Query Form

**Search pattern:**

• display first  N-grams with frequency  $\geq$

• variable elements are , constant elements are

Debug  Optim.

### Results

306	berkeley is a charming
242	berkeley is a five
165	berkeley is a city
155	berkeley is a great
134	berkeley is a very
115	berkeley is a public
88	berkeley is a good
85	berkeley is a sharp
66	berkeley is a member
63	berkeley is a place
58	berkeley is a small
56	berkeley is a wonderful
51	berkeley is a major
50	berkeley is a new
49	berkeley is a versatile

# N-Gram Models: Challenges

---



# Sparsity

---

*Please close the first door on the left.*

```
3380 please close the door
1601 please close the window
1164 please close the new
1159 please close the gate
...
0 please close the first
-----
13951 please close the *
```



# Back-off

*Please close the first door on the left.*

4-Gram

3380 please close the door  
1601 please close the window  
1164 please close the new  
1159 please close the gate  
...  
0 please close the first  
-----  
13951 please close the \*

0.0

3-Gram

197302 close the window  
191125 close the door  
152500 close the gap  
116451 close the thread  
...  
8662 close the first  
-----  
3785230 close the \*

0.002

2-Gram

198015222 the first  
194623024 the same  
168504105 the following  
158562063 the world  
...  
...  
-----  
23135851162 the \*

0.009

Specific but Sparse

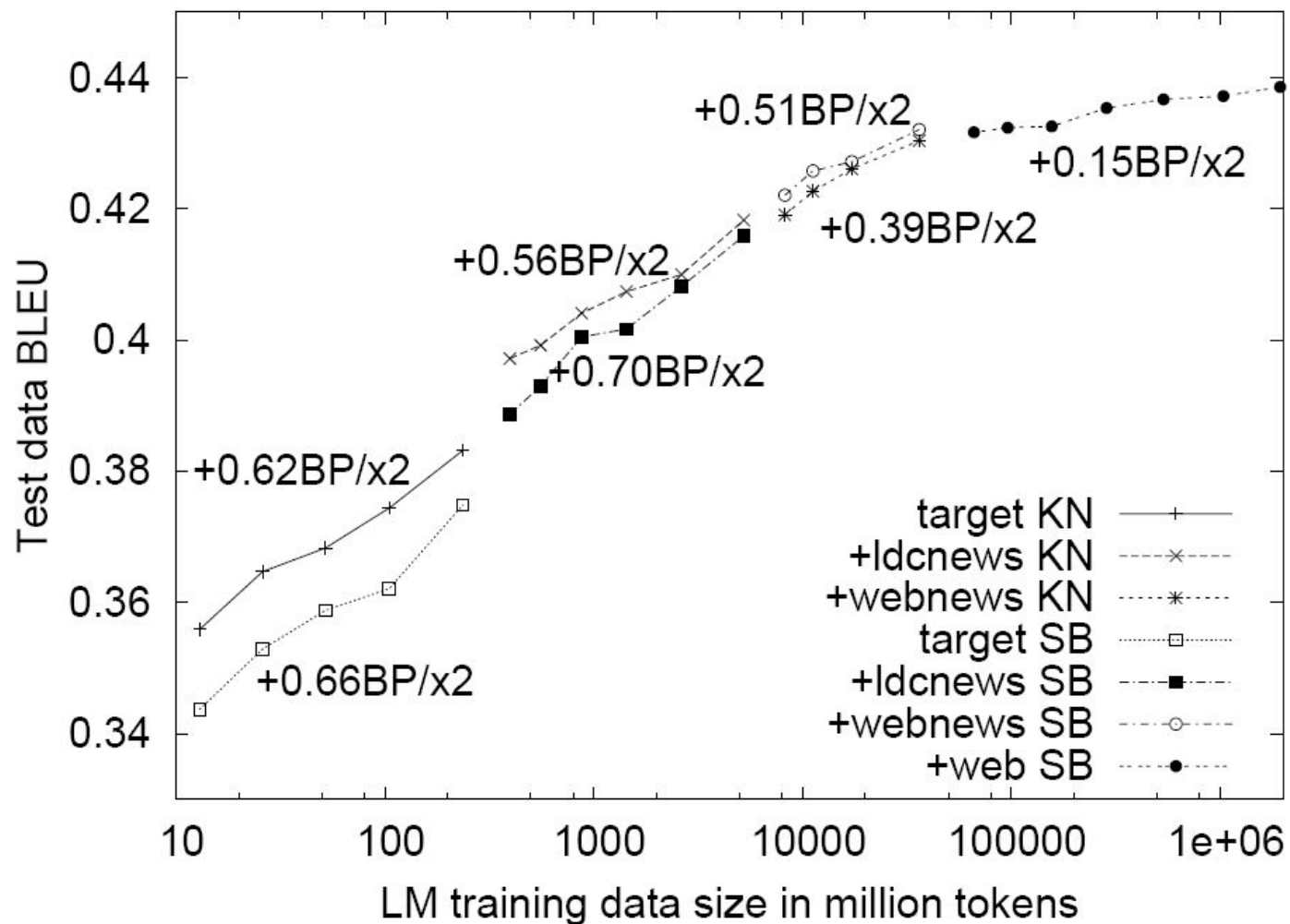


Dense but General

$$\lambda \hat{P}(w|w_{-1}, w_{-2}) + \lambda' \hat{P}(w|w_{-1}) + \lambda'' \hat{P}(w)$$



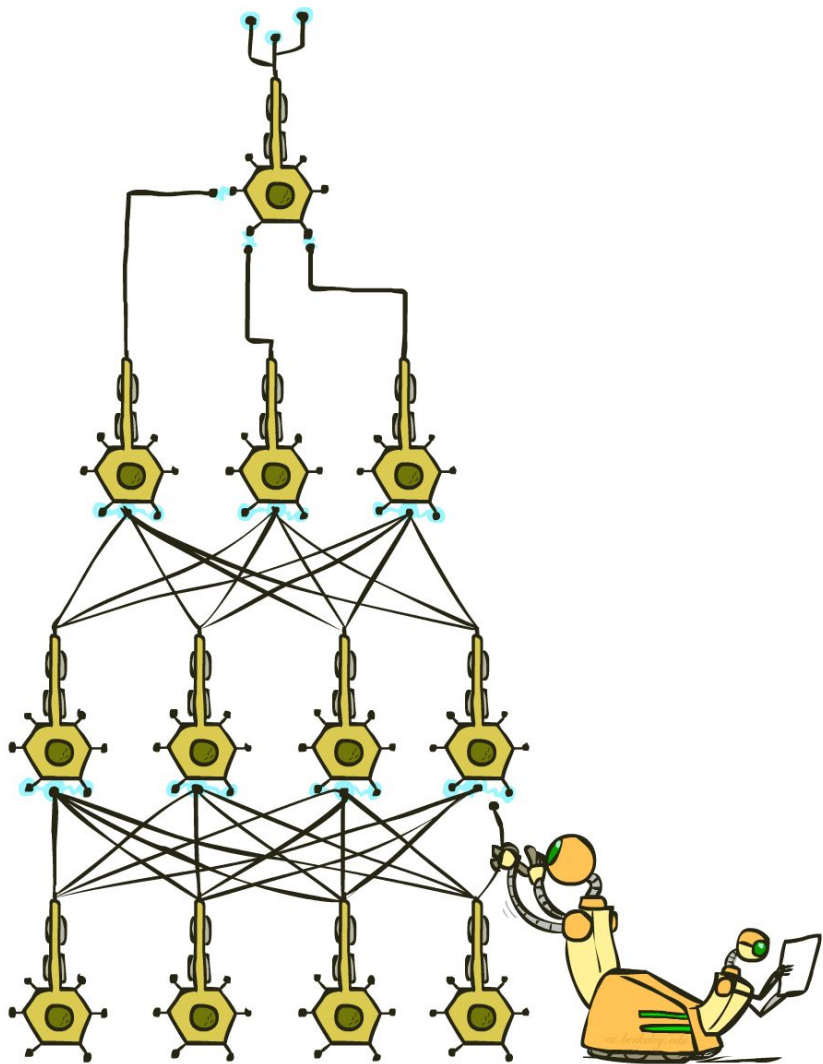
# More Data?



# Neural Language Models

---

# Deep Neural Networks?



- Input: some text

- “The dog chased the”

- Output: more text

- ... “ball”

- Implementation:

- Neural nets?
- Linear algebra?
- How??



# Neural LMs: Three Key Ideas

---

- **Word embeddings**
  - Different words are not entirely unrelated events
  - Words can be more and less similar, in complex ways
- **Partially factored representations**
  - Multiple semi-independent processes happen in parallel in language
  - It's too expensive to track language in an unfactored way, and too inaccurate to assume everything of interest is independent
- **Long distance dependencies**
  - Information can be relevant without being local
  - Different notions of locality are important at different times

# Words: Clusterings and Embeddings

---



# Clusterings

- Automatic (Finch and Chater 92, Shuetze 93, many others)

word	nearest neighbors
accompanied	submitted banned financed developed authorized headed canceled awarded barred
almost	virtually merely formally fully quite officially just nearly only less
causing	reflecting forcing providing creating producing becoming carrying particularly
classes	elections courses payments losses computers performances violations levels pictures
directors	professionals investigations materials competitors agreements papers transactions
goal	mood roof eye image tool song pool scene gap voice
japanese	chinese iraqi american western arab foreign european federal soviet indian
represent	reveal attend deliver reflect choose contain impose manage establish retain
think	believe wish know realize wonder assume feel say mean bet
york	angeles francisco sox rouge kong diego zone vegas inning layer
on	through in at over into with from for by across
must	might would could cannot will should can may does helps
they	we you i he she nobody who it everybody there

- Manual (e.g. thesauri, WordNet)



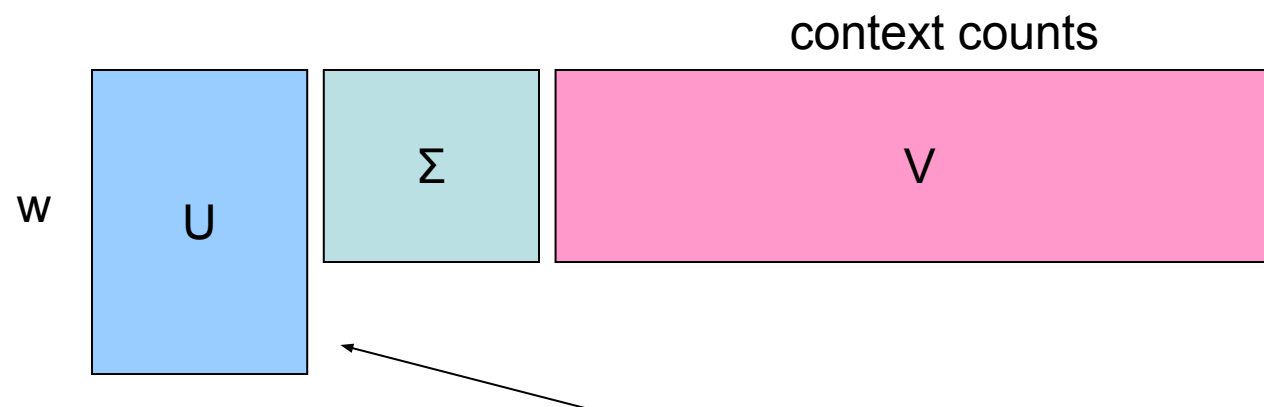
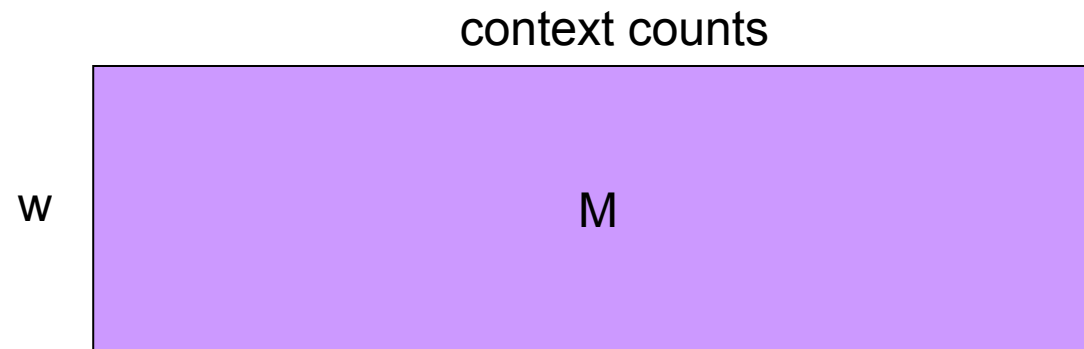
# Stuffing Words into Vector Spaces?





# Vector Space Methods

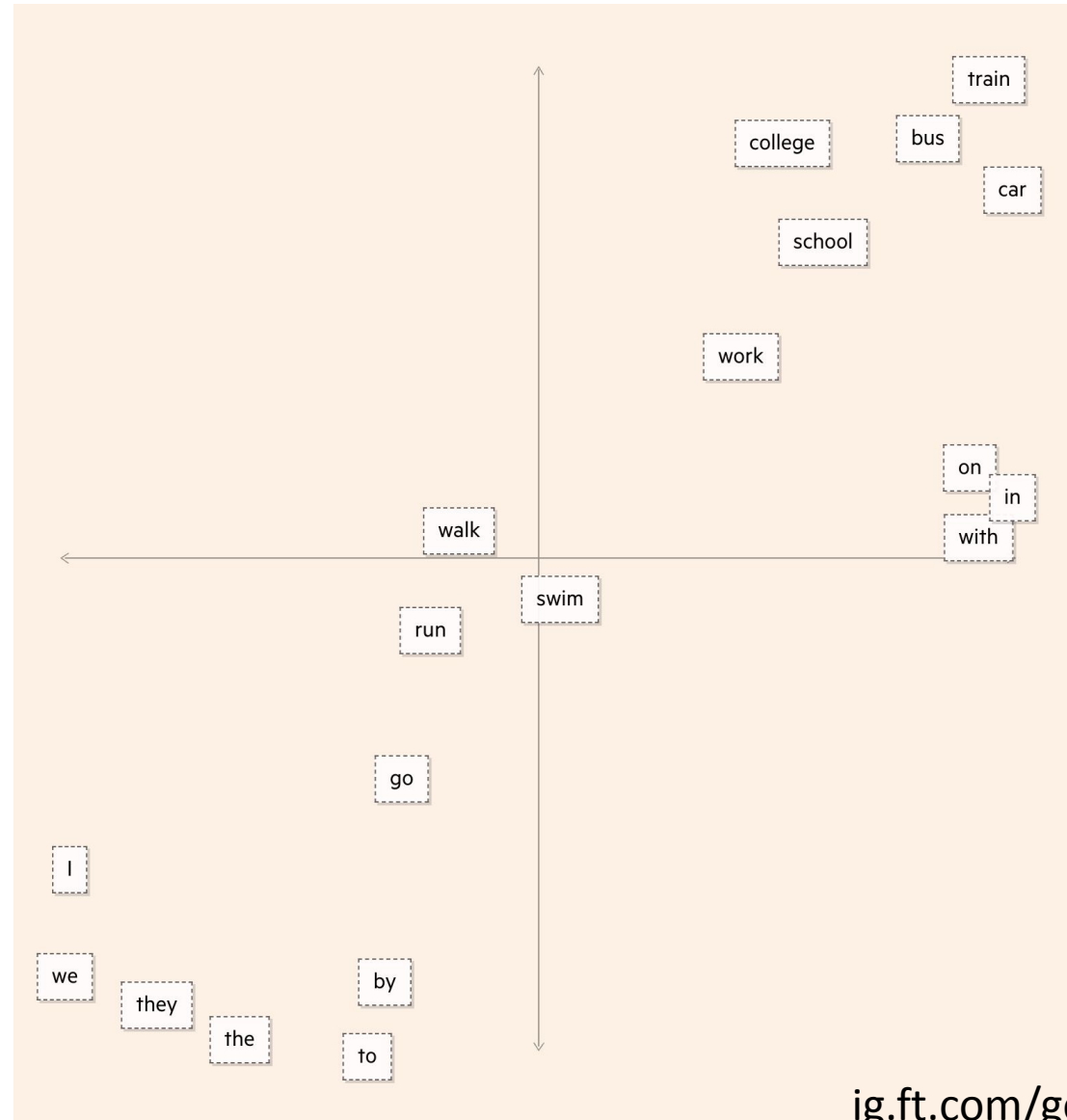
- Treat words as points in  $R^n$  (eg Shuetze, 93)
  - Form matrix of co-occurrence counts
  - SVD or similar to reduce rank
  - Cluster projections
  - People worried about things like: log of counts,  $U$  vs  $U\Sigma$
- Today we'd call this an embedding method (it's basically GLoVe)



Cluster these 50-200 dim vectors instead.

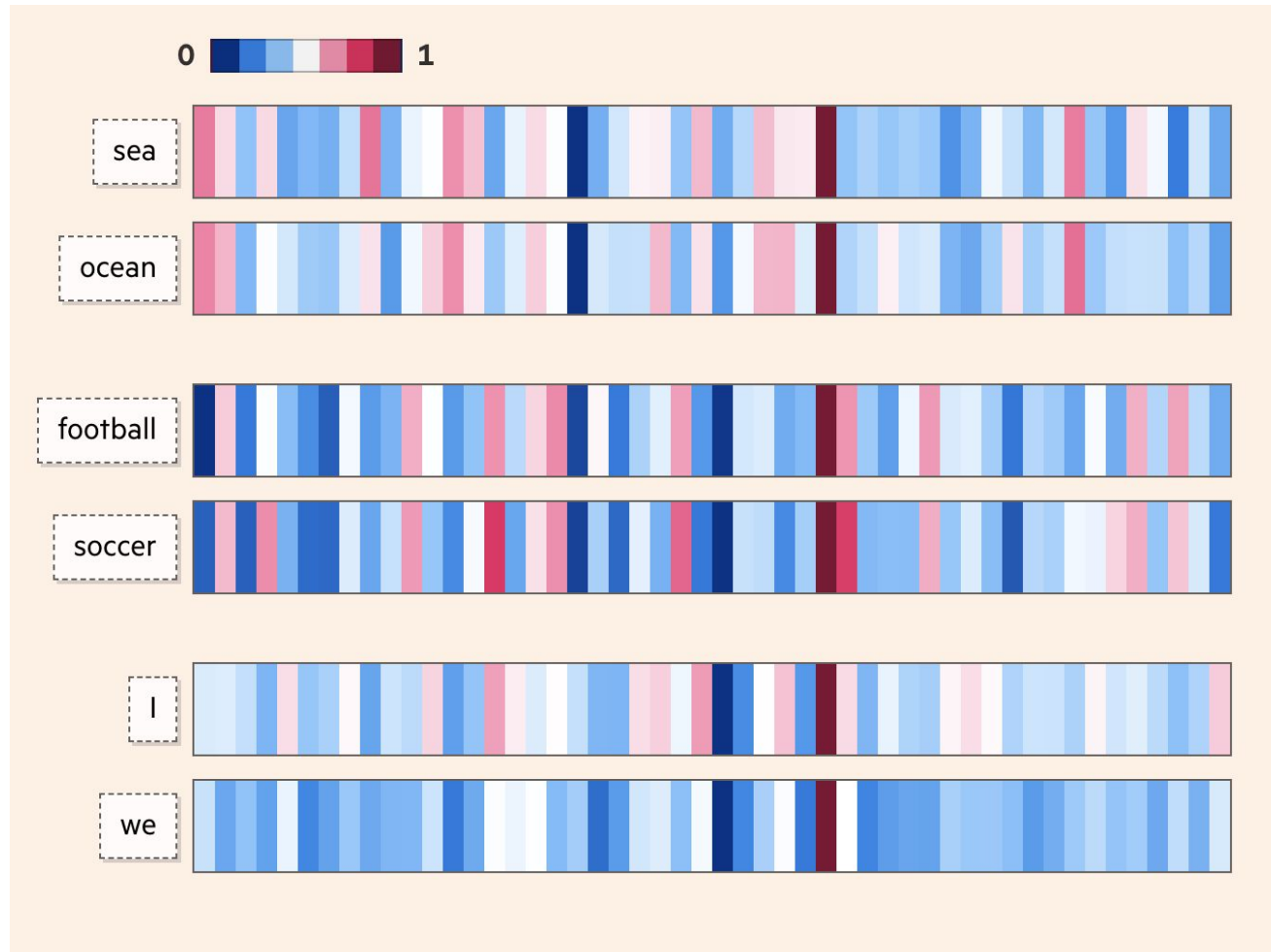
# What do word embeddings look like?

- Words cluster by similarity:



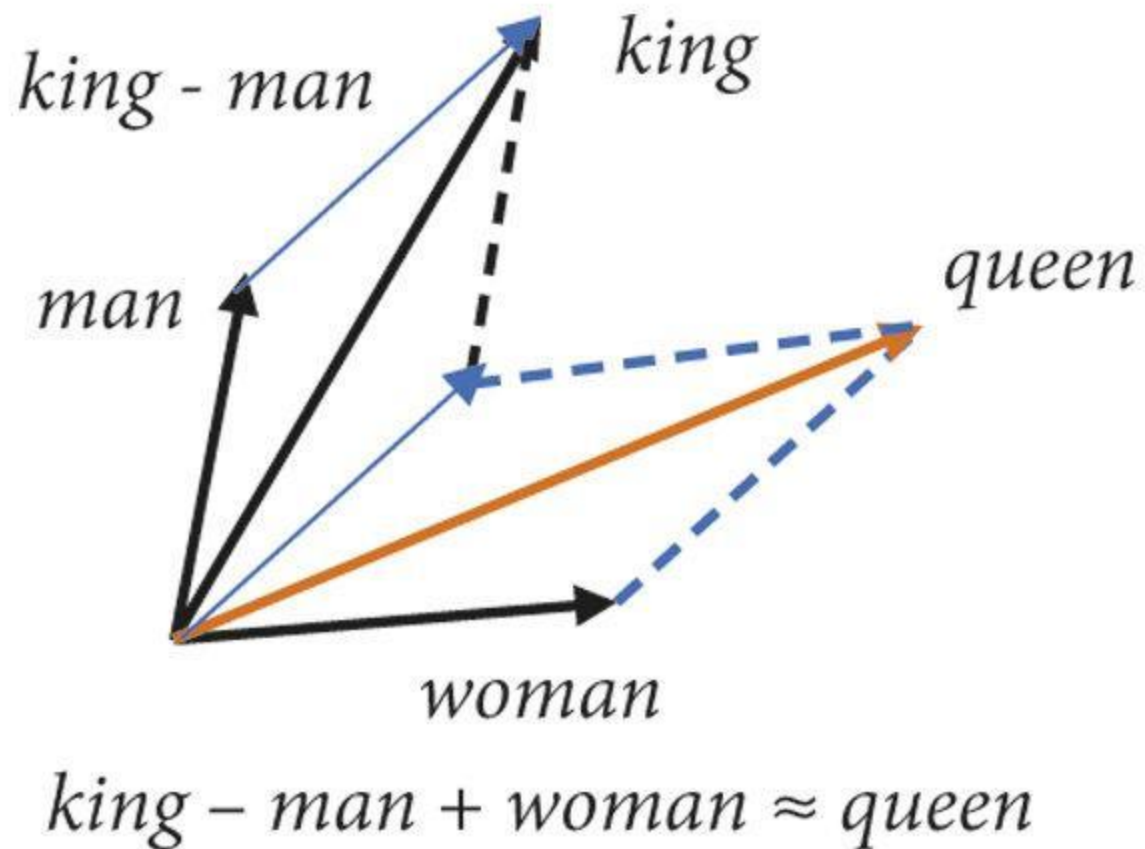
# What do word embeddings look like?

- Features learned in language models:



# What do word embeddings look like?

- Signs of semantic algebraic structure in embedding space?



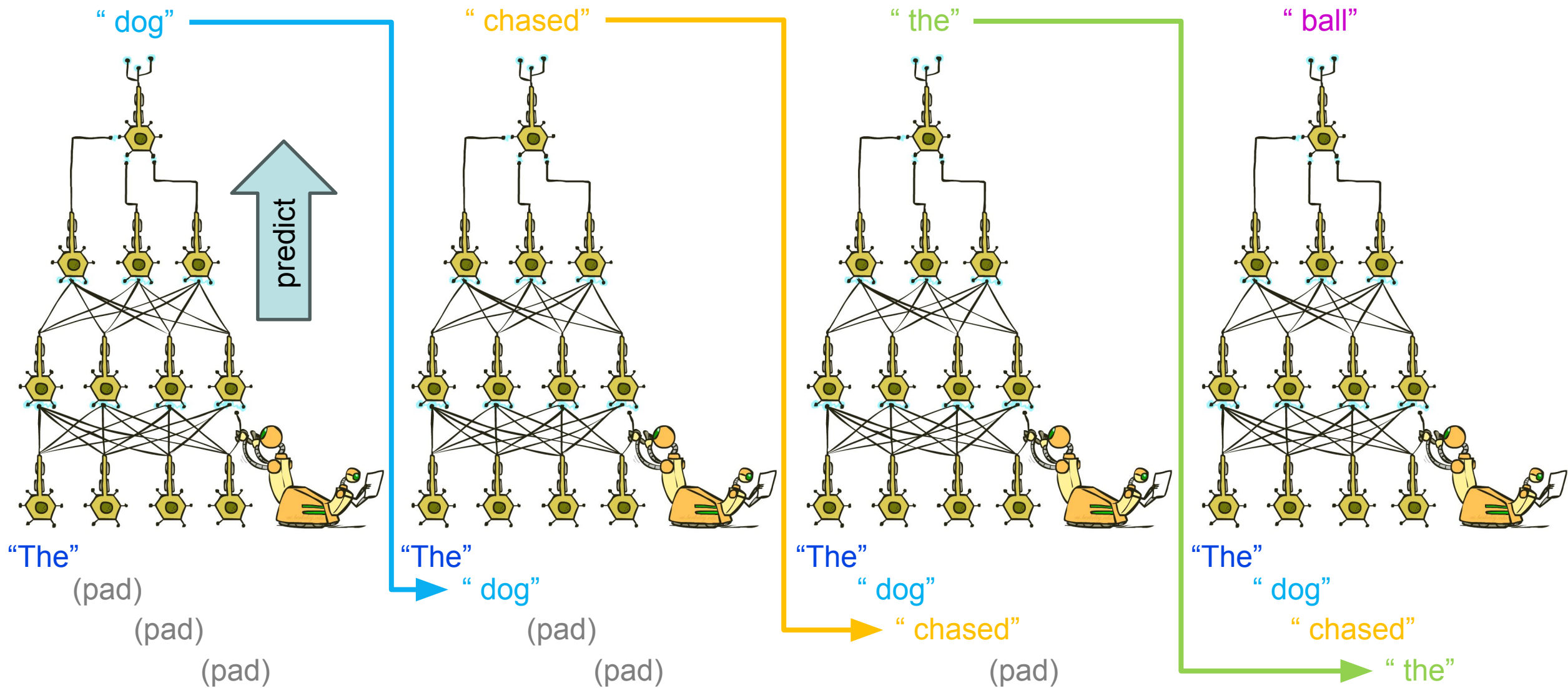


# Neural LMs: Three Key Ideas

---

- **Word embeddings**
  - Different words are not entirely unrelated events
  - Words can be more and less similar, in complex ways
- **Partially factored representations**
  - Multiple semi-independent processes happen in parallel in language
  - It's too expensive to track language in an unfactored way, and too inaccurate to assume everything of interest is independent
- **Long distance dependencies**
  - Information can be relevant without being local
  - Different notions of locality are important at different times

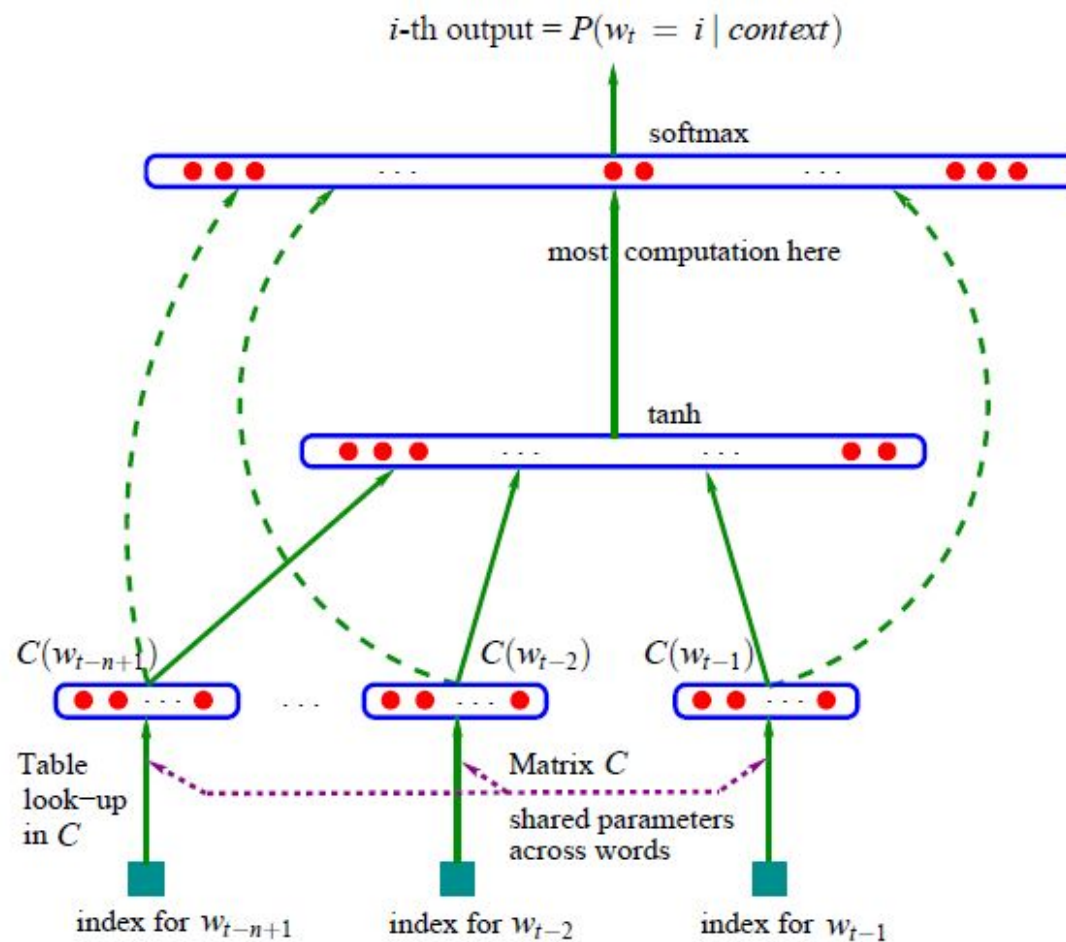
# Autoregressive Models





# Early Neural Language Models

- Fixed-order feed-forward neural LMs
  - Eg Bengio et al 03
  - Allow generalization across contexts in more nuanced ways than prefixing
  - Allow different kinds of pooling in different contexts
  - Much more expensive to train





# Neural LMs: Three Key Ideas

---

- **Word embeddings**
  - Different words are not entirely unrelated events
  - Words can be more and less similar, in complex ways
- **Partially factored representations**
  - Multiple semi-independent processes happen in parallel in language
  - It's too expensive to track language in an unfactored way, and too inaccurate to assume everything of interest is independent
- **Long distance dependencies**
  - Information can be relevant without being local
  - Different notions of locality are important at different times



# A GPT2 Continuation

---

- **The computer I had put into the machine room on the fifth floor just just** outside our landing was taken by a lot of people. It was going to be recovered from outside the machine room as soon as we could, but after the machine room was shut down, we had to open all of the windows and doors to save ourselves, and the computer would have been lost. Luckily we are prepared for this. We had five steel doors on each side of the landing to hold any rain and wind that might have hit, so we weren't affected by the storm.

The situation was really grim for days. A few days before, the Apache at Qush Tefah had been shot. This was the most damaged machine in the Armies Corps. The command was only meant to repair Qush



# A GPT2 Continuation

---

- **I shall go seek the traitor Gloucester**, when he hath declared his vengeance. Cunegund Shall I back her to kill my cousin in my stead? Witbane Far first shall the brother take his Queen's soul and that in the flames. Clotilda For to slay the King of Scotland with, the blood of my cousin, lie true; And she shall have the right to my feather." Sefton leapt into the rushes, and stole along in the sunlight: the small figure flashed like lightning in the west. In the likeness of a snare he had laid to catch the rushes and made of them a snares, a road to flee from his pursuers; but he now came to an oak where the branches were wreathed in an oak-

**Next Time: Transformers and LLMs**

---