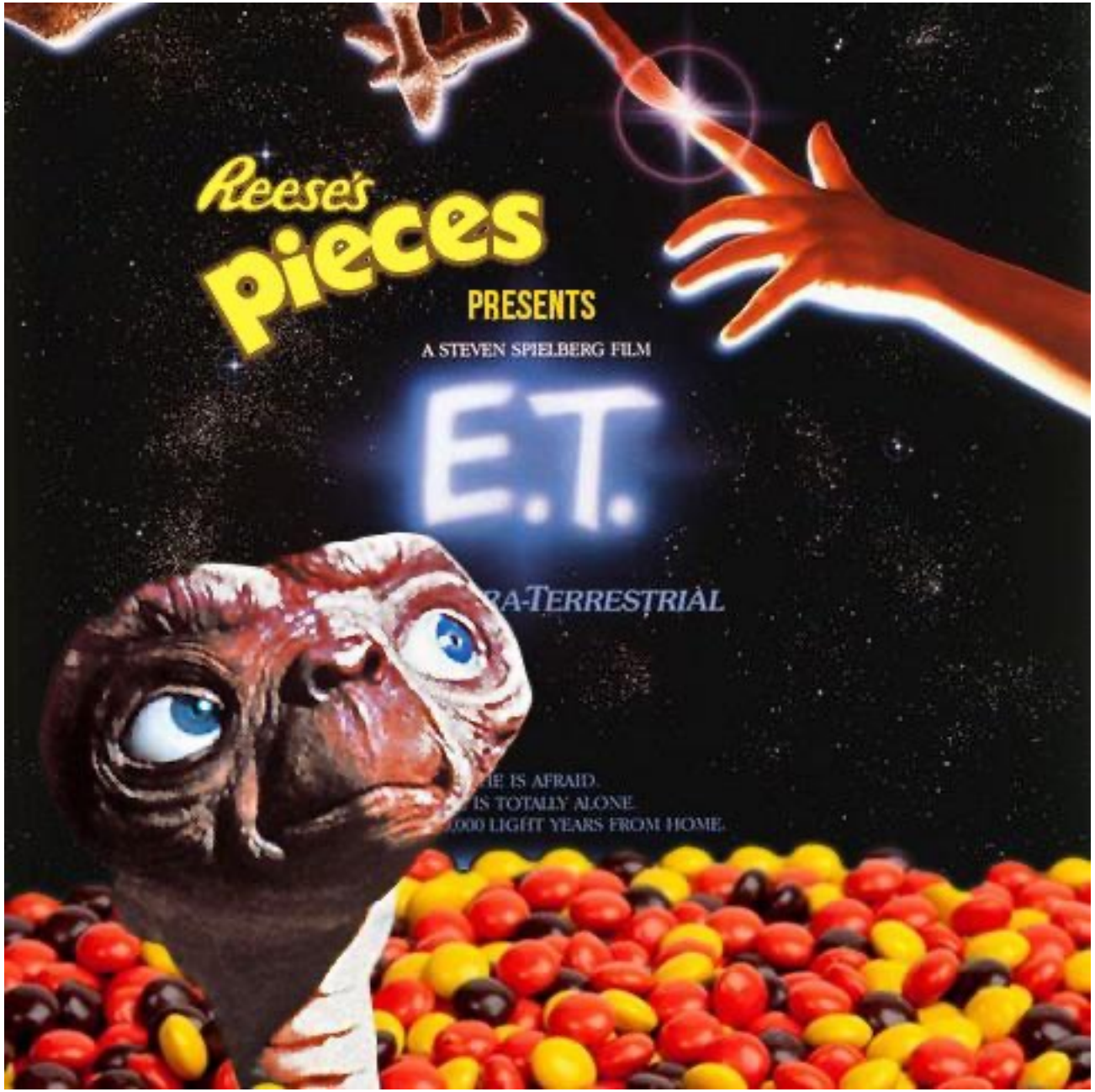
LLMs! (omg finally)

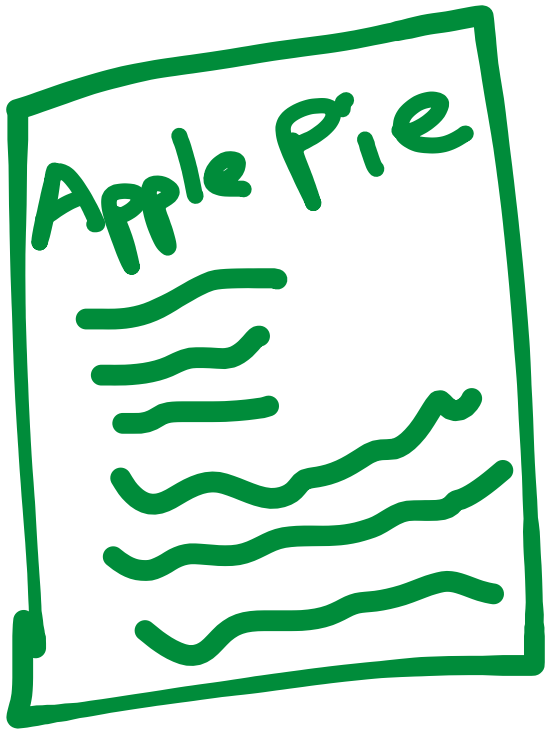
INF1 CG

Week 9, Lecture 25

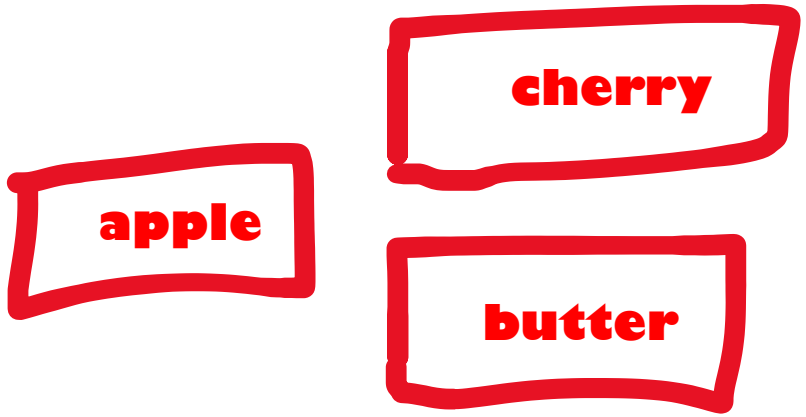
Maithilee Kunda


Reese's
Pieces
PRESENTS
A STEVEN SPIELBERG FILM
E.T.
A-TERRESTRIAL
IS AFRAID.
IS TOTALLY ALONE.
000 LIGHT YEARS FROM HOME.

- This is a learned word embedding!



- This is the training data!



- These are the tokens!
- We can also let the alien rip up words into pieces, or tape words together to make common phrases
- ”Learning” the tokens, i.e., learn the lexicon

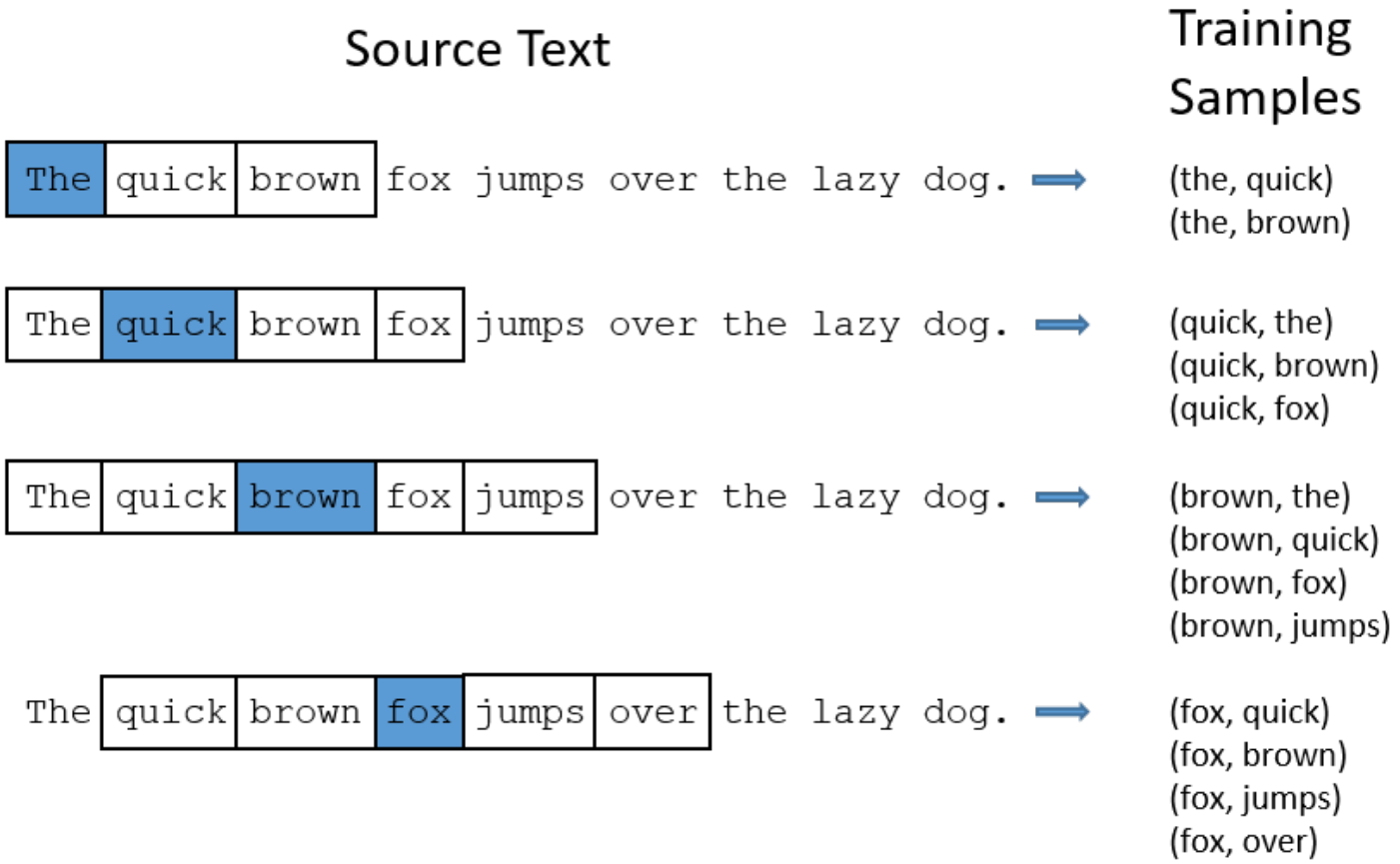
History of these ideas…

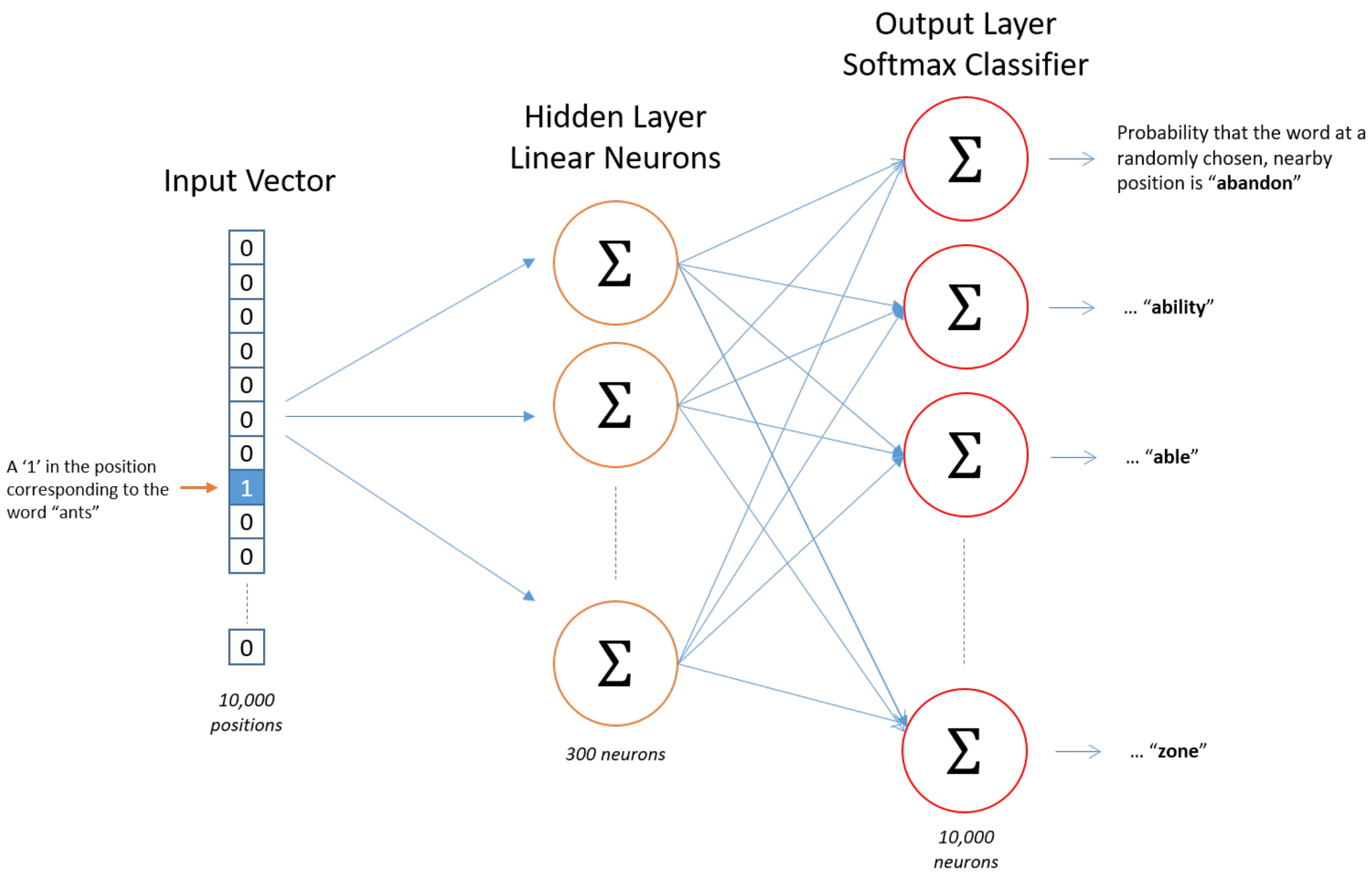
- 1990s – Use context vectors to help with information retrieval (e.g., find two documents that are related)
- 2000s – Use neural networks to learn the embedding instead of using manually defined dimensions
- 2013: Efficient Estimation of Word Representations in Vector Space ---> “Word2Vec” model

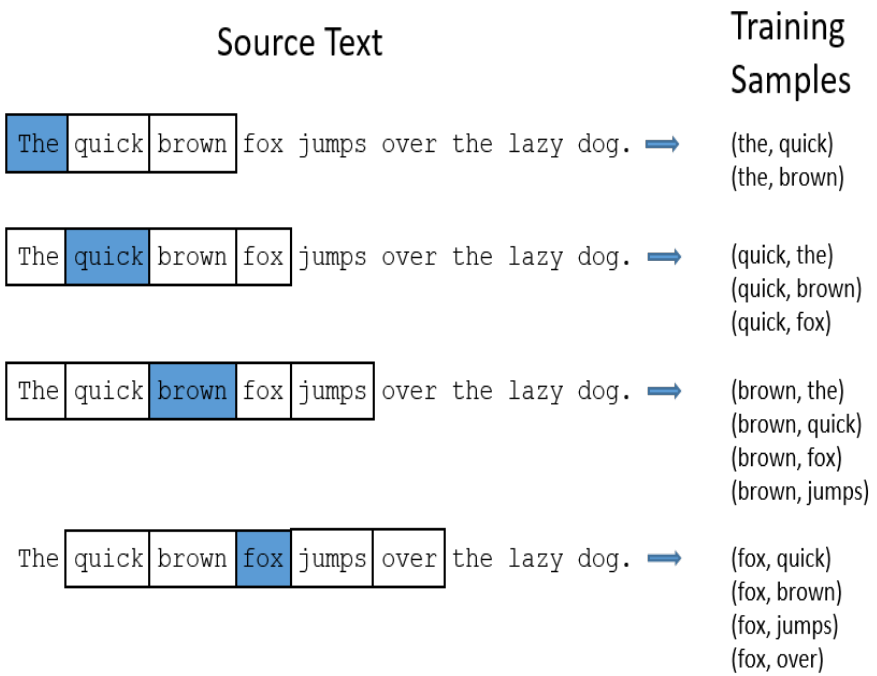
Word2Vec – two training methods, we'll just talk about one: The Skip Gram model

- https://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

- Training task:

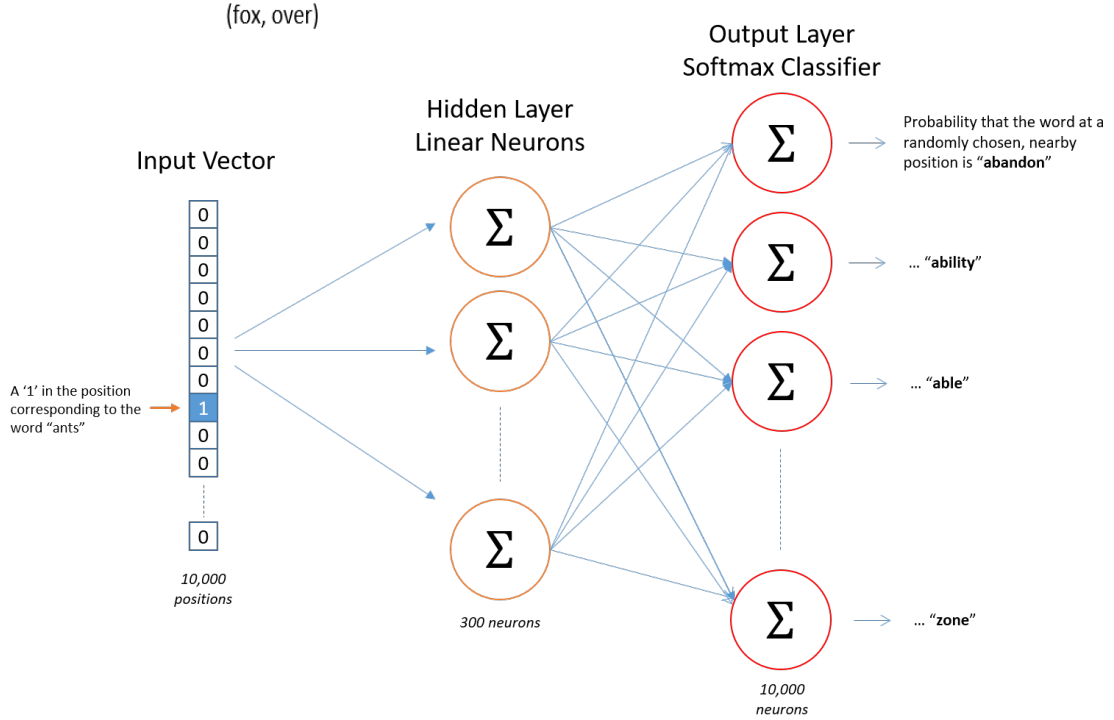



Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 neurons
10,000 neurons
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"



Result:

- A really good bucket system! (embedding)
- AND we didn't even need to label the data beforehand!
- "Self-supervised learning" – learn from correct answers and errors, just like supervised learning, but the "correct" answers are already in the data



Word2vec is pretty impressive!

- from the reading
- t-SNE is a method for projecting high dimensional vectors into 2D for visualizing relationships

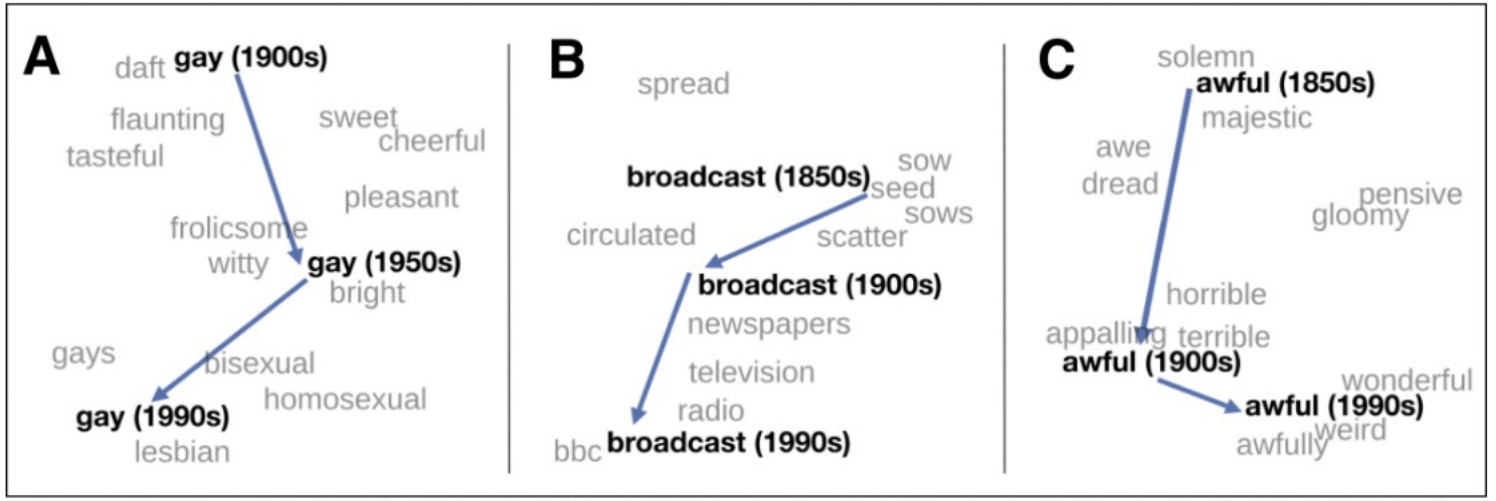


Figure 6.14 A t-SNE visualization of the semantic change of 3 words in English using word2vec vectors. The modern sense of each word, and the grey context words, are computed from the most recent (modern) time-point embedding space. Earlier points are computed from earlier historical embedding spaces. The visualizations show the changes in the word *gay* from meanings related to "cheerful" or "frolicsome" to referring to homosexuality, the development of the modern "transmission" sense of *broadcast* from its original sense of sowing seeds, and the pejoration of the word *awful* as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Hamilton et al., 2016).

We can subtract two context vectors, then add the result to another context vector:

Expression	*Nearest token*
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

- (slide credit: Frank Keller)

BUT: (from the reading)

- For example Bolukbasi et al. (2016) find that the closest occupation to

  ‘man’ - ‘computer programmer’ + ‘woman’

  in word2vec embeddings trained on news text is ….

- **‘homemaker’** 😱😱😱

From learned embedding to LLM……

- 1990s – Use context vectors to help with information retrieval (e.g., find two documents that are related)
- 2000s – Use neural networks to learn the embedding instead of using manually defined dimensions
- 2013: Efficient Estimation of Word Representations in Vector Space ---> “Word2Vec” model
- **2017: Google paper presenting “transformer” neural network architecture: “Attention is all you need”**

The cat left its litter mates and uses the litter

_ _ _ _ _ _ _ _ _

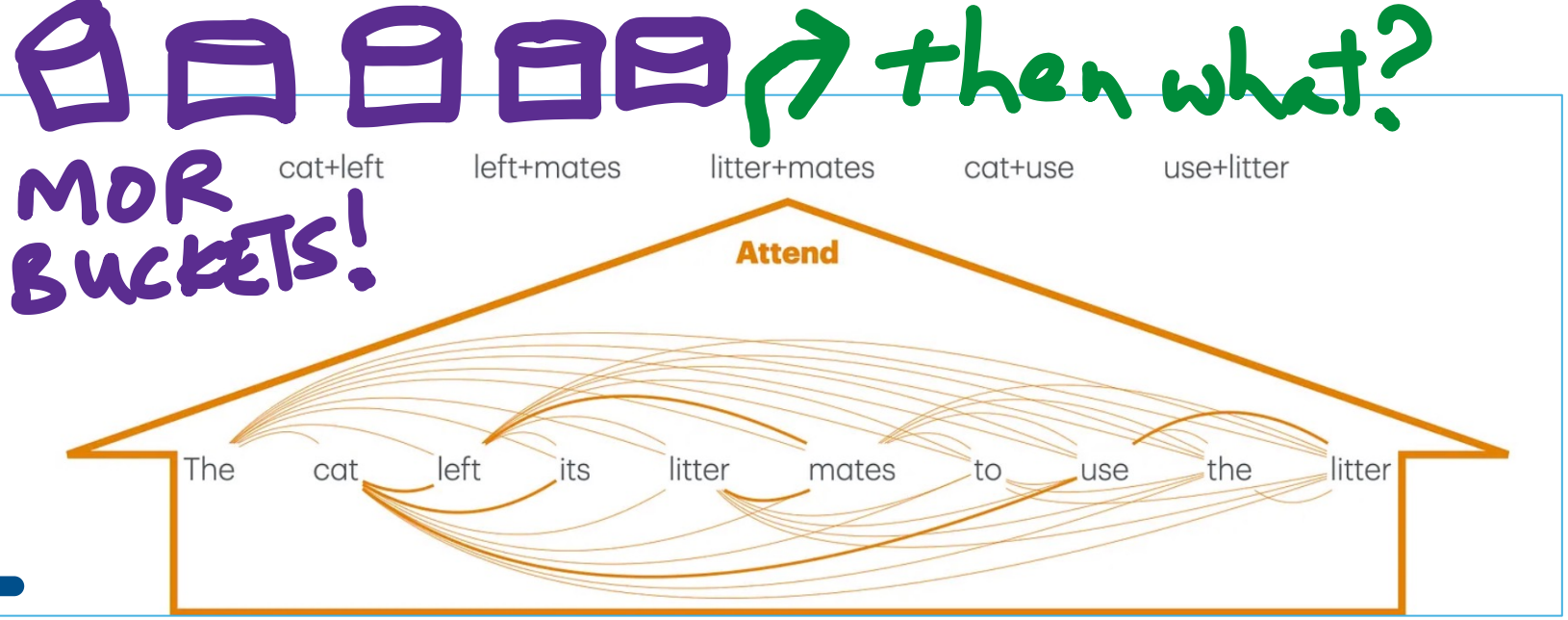
- What is the problem here for word embeddings?
- ”Litter” gets used twice, but it means completely different things!
- But it goes to the SAME embedding.

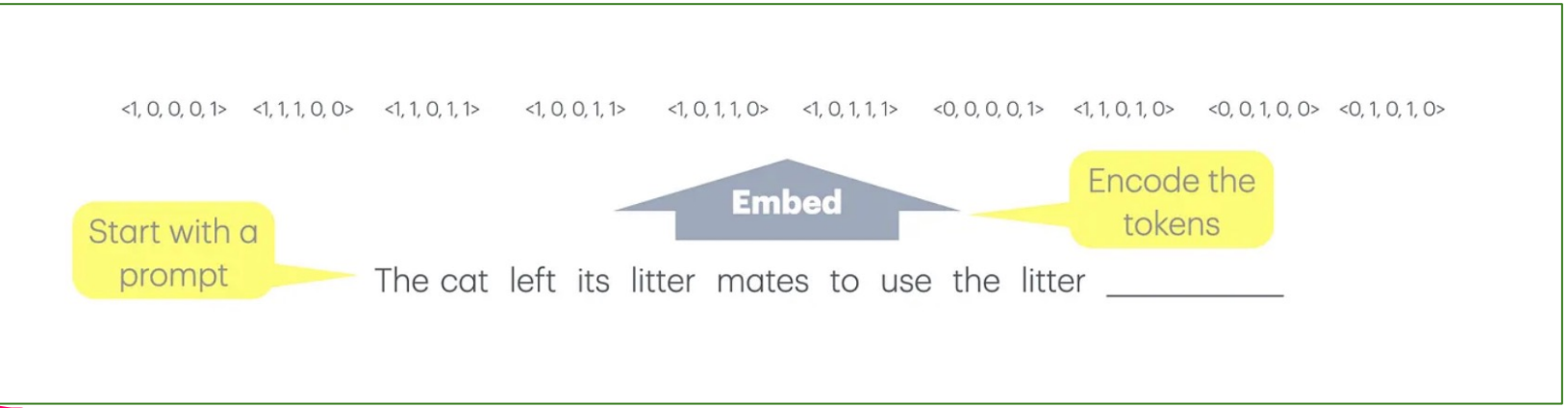
- How do WE know that litter means two different things?
- Context again!
- **But now… we are talking about the context of a word as it is used in a specific sentence… and NOT its average use across an entire corpus**

- https://mark-riedl.medium.com/the-intuition-behind-how-large-language-models-work-166cf2fb278a

getting into the grammar!

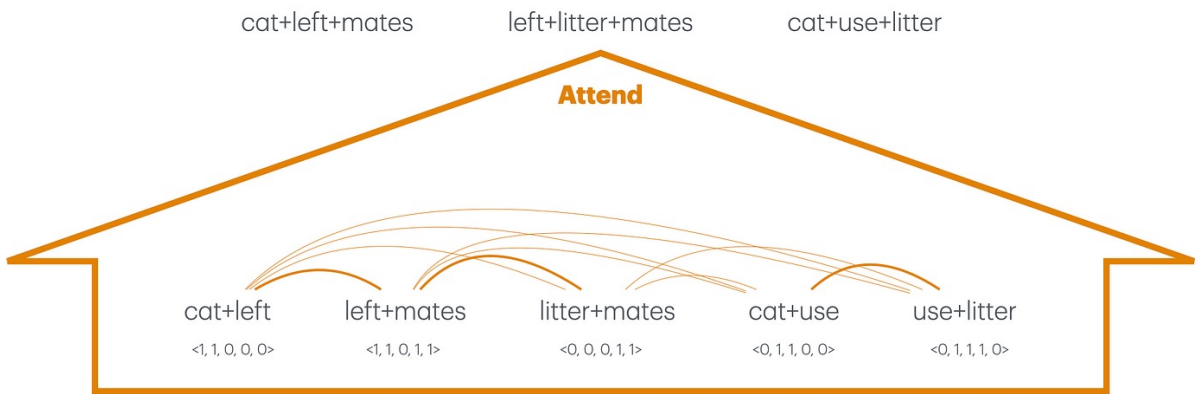
"lexicon"

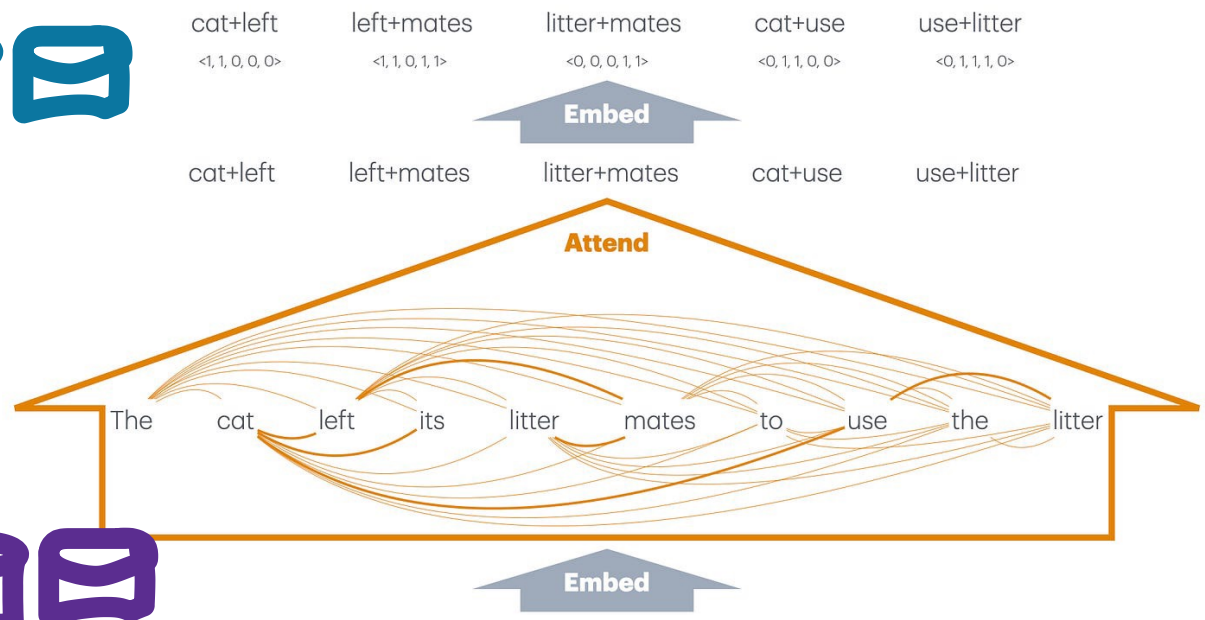


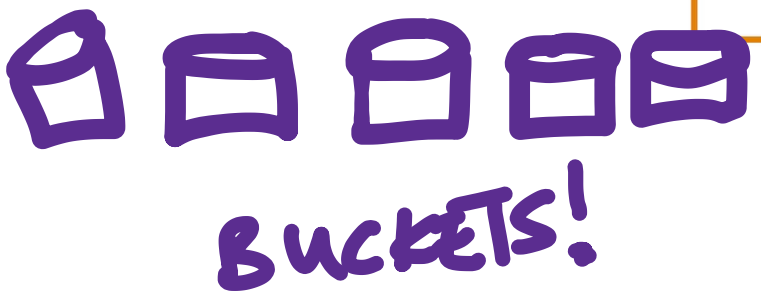


The cat left its litter mates and uses the litter

_____







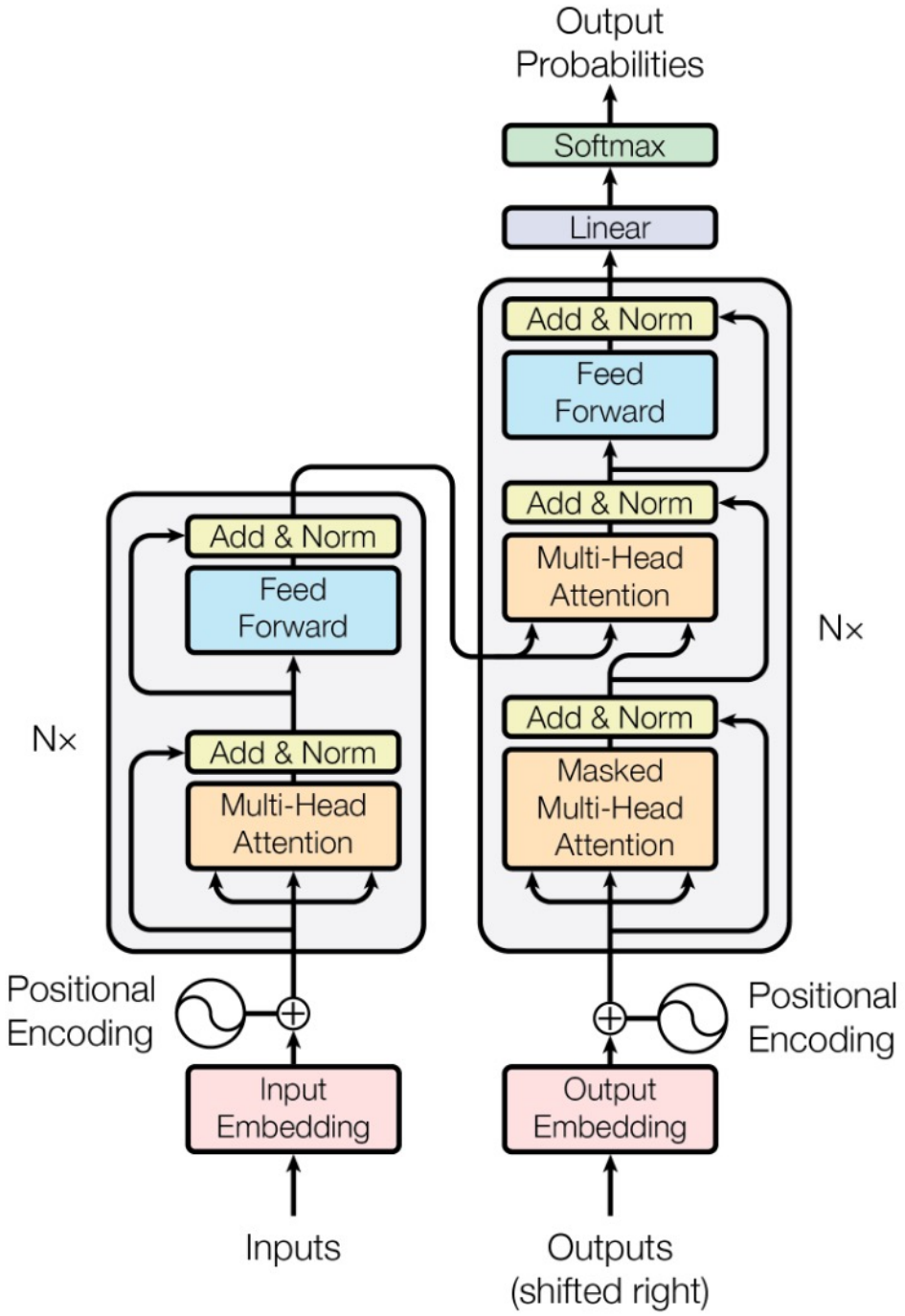
”attention” is the mechanism used to evaluate different word groupings and try to assign specific meanings to groups of words, according to how they are <u>grammatically</u> combined

Google's Transformer

- All of these are sets of nested buckets!
- Left: Encoder = words to buckets
- Right: Decoder = buckets to words

- The original Google paper trained the transformer on translation tasks (e.g., 14.5M English to German sentence pairs, and separately 36M English to French sentence pairs)

Supervised learning?



From learned embedding to LLM……

- 1990s – Use context vectors to help with information retrieval (e.g., find two documents that are related)
- 2000s – Use neural networks to learn the embedding instead of using manually defined dimensions
- 2013: Efficient Estimation of Word Representations in Vector Space ---> “Word2Vec” model
- 2017: Google paper presenting “transformer” neural network architecture: “Attention is all you need”
- **2018 and beyond: many papers building LLMs using transformers but with self-supervised tasks (like word2vec) instead of supervised tasks**

GPT-3 (OpenAI's LLM in 2020)

- training data size = 300 billion tokens
- 4 tokens = 3 words
- 300 billion tokens = 225 billion words

- 1 chunky literary novel = 100k words = 7 hours of reading time
- All of English Wikipedia = 5 billion words = 50k novels
- reading all of wikipedia would take 40 years straight through, or 175 years if it was your day job

- 300 billion tokens = 50 wikipedias
- = 8,750 years of your day job to finish reading

If you have an LLM trained on huge swaths of the Internet…..

- What will it “sound” like?

Internet Culture

Trolls turned Tay, Microsoft's fun millennial AI bot, into a genocidal maniac

March 25, 2016 More than **9 years ago**

By Abby Ohlheiser

It took mere hours for the Internet to transform Tay, the teenage AI bot who wants to chat with and learn from millennials, into Tay, the racist and genocidal AI bot who liked to reference Hitler. And now Tay is taking a break.

- “Unfortunately, within the first 24 hours of coming online,” an emailed statement from a Microsoft representative said, “a coordinated attack by a subset of people exploited a vulnerability in Tay.”

- Zoe Quinn, a frequent target of Gamergate, posted a screenshot overnight of the bot tweeting an insult at her, prompted by another user. “Wow it only took them hours to ruin this bot for me,” she wrote in a series of tweets about Tay.
- **“It’s 2016. If you’re not asking yourself ‘how could this be used to hurt someone’ in your design/engineering process, you’ve failed.”**

Tay was different because…

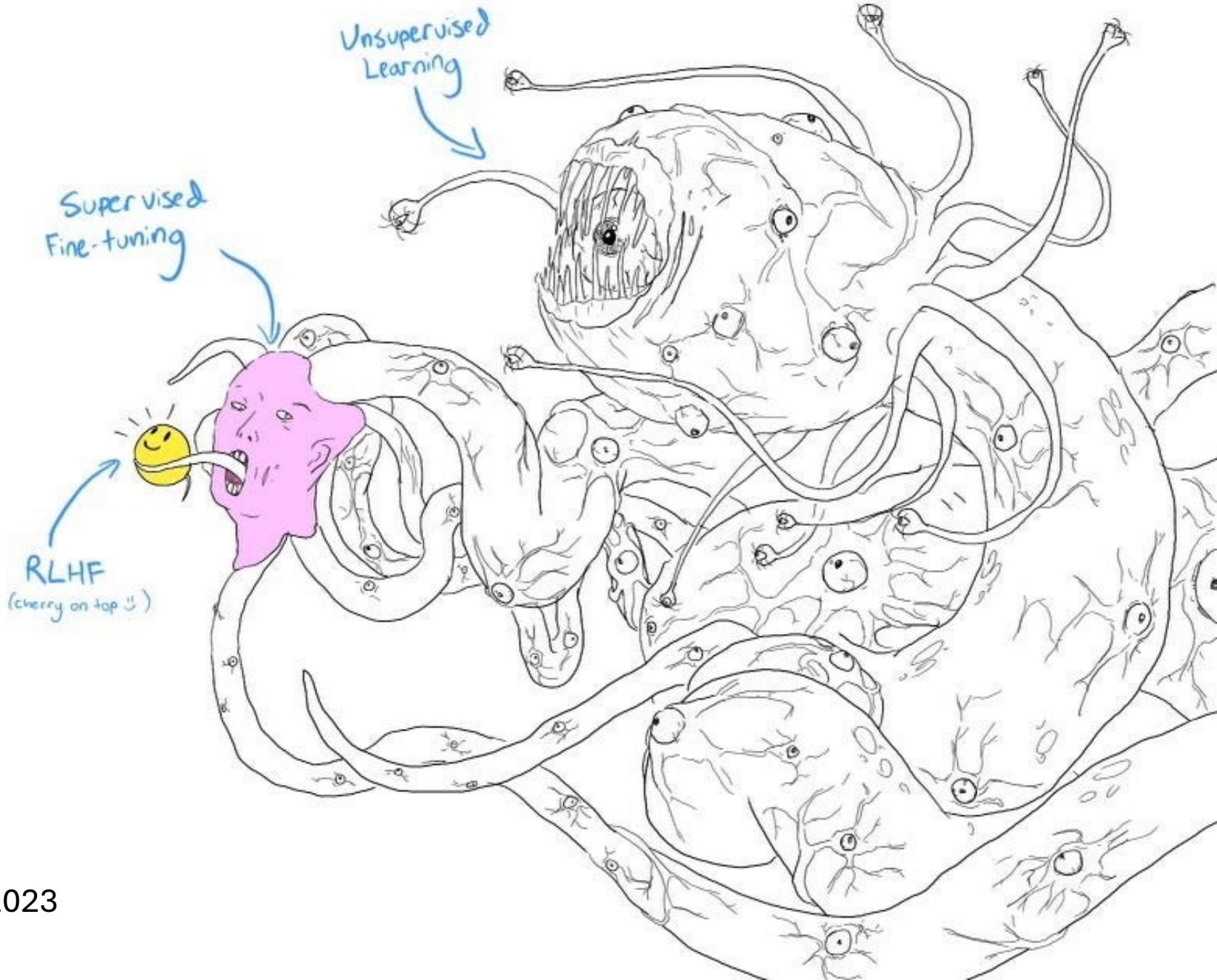
- Was learning online from active users

- LLMs now (conventionally) don’t do that.
- But they can still say terrible things! Learned from the training data.

- How to fix???

- How would you use supervised learning to “fix” outputs of an LLM?
- Idea #1: Give it positive and/or negative examples of things it should or shouldn’t say. Tweak the weights to reduce “error”!
  - Just the same as perceptrons!
  - “Supervised fine tuning”
- Idea #2: learning from reward = reinforcement learning
  - Reward LLM outputs for good stuff, penalize for bad stuff
  - (More on RL next week!)
  - Problem: what is the reward function???
  - “Reinforcement learning using human feedback” = RLHF
  - Get human feedback on a FEW examples of LLM outputs
  - Use this to estimate a reward function…
  - Then use that reward function to learn from lots more examples of LLM output

funny meme



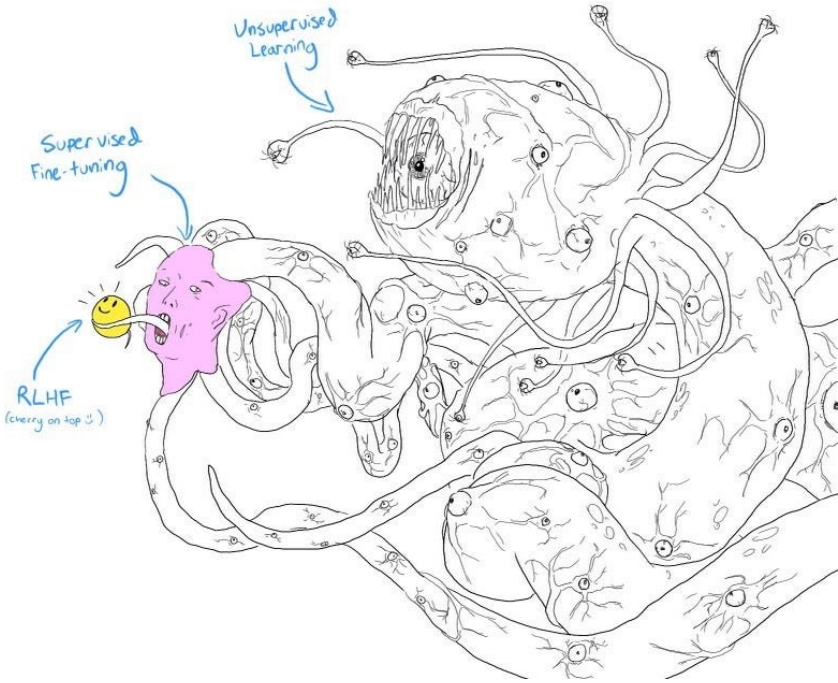
from twitter user “anthrupad” in early 2023

But this has some serious consequences…

- Problematic interactions……..
- ”chatbot-induced psychosis”

- ALSO:

- Where does all this human data come from???



Behind the AI boom, an army of overseas workers in ‘digital sweatshops’

August 28, 2023 More than **2 years a**

11 min 289

AI

’AI Is African Intelligence’: The Workers Who Train AI Are Fighting Back

Kenyan workers are still the underpaid labor behind AI training, moderation, and sex chatbots. The Data Labelers Association is fighting back.

JASON KOEBLER · MAR 12, 2026 AT 11:08 AM

MORGAN MEAKER BUSINESS SEP 11, 2023 6:00 AM

These Prisoners Are Training AI

In high-wage Finland, where clickworkers are rare, one company has discovered a novel labor force—prisoners.