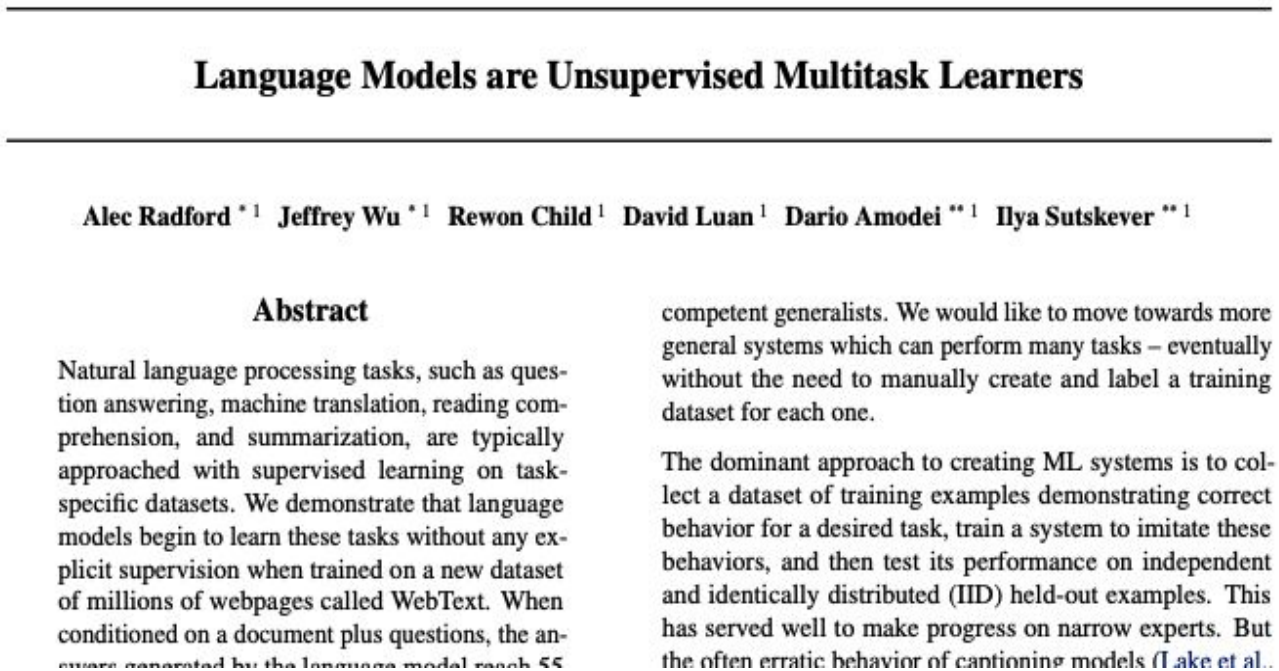
Intro to NLP & Language Modeling

Sep 12

Attendance survey

PollEv.com/ic226

2019



Language Models are Unsupervised Multitask Learners

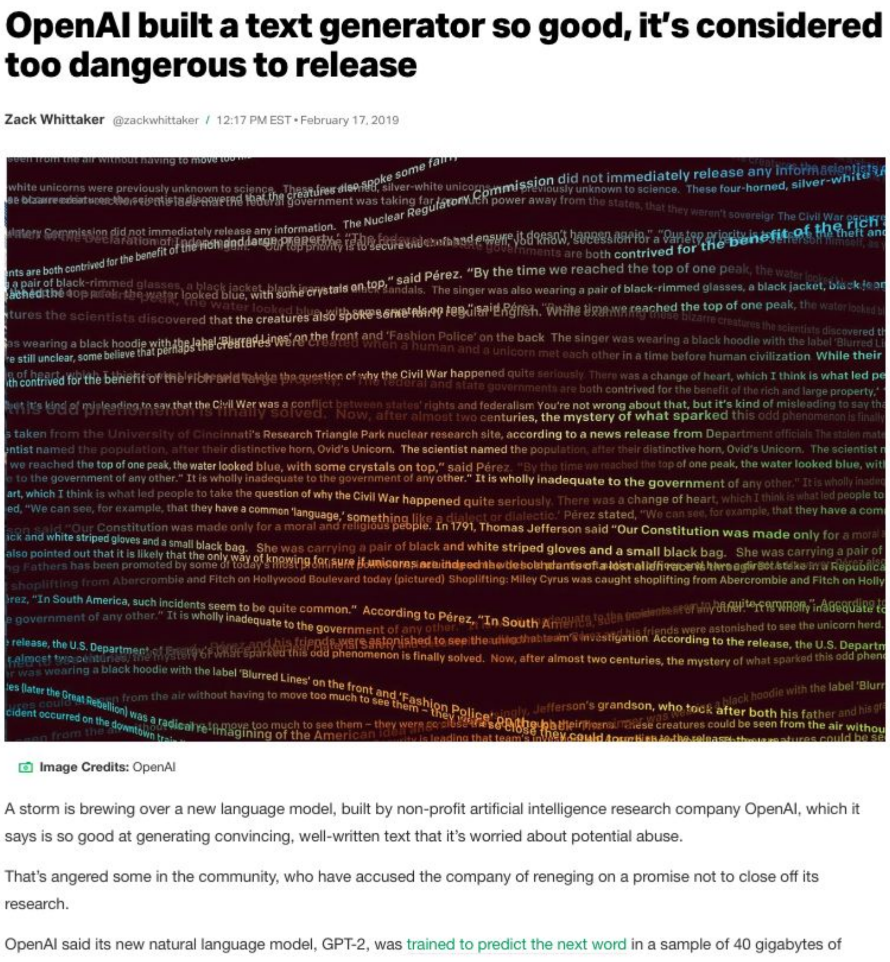
Alec Radford *[1] Jeffrey Wu *[1] Rewon Child[1] David Luan[1] Dario Amodei **[1] Ilya Sutskever **[1]

Abstract

Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55

competent generalists. We would like to move towards more general systems which can perform many tasks – eventually without the need to manually create and label a training dataset for each one.

The dominant approach to creating ML systems is to collect a dataset of training examples demonstrating correct behavior for a desired task, train a system to imitate these behaviors, and then test its performance on independent and identically distributed (IID) held-out examples. This has served well to make progress on narrow experts. But the often erratic behavior of captioning models (Lake et al.,



OpenAI built a text generator so good, it's considered too dangerous to release

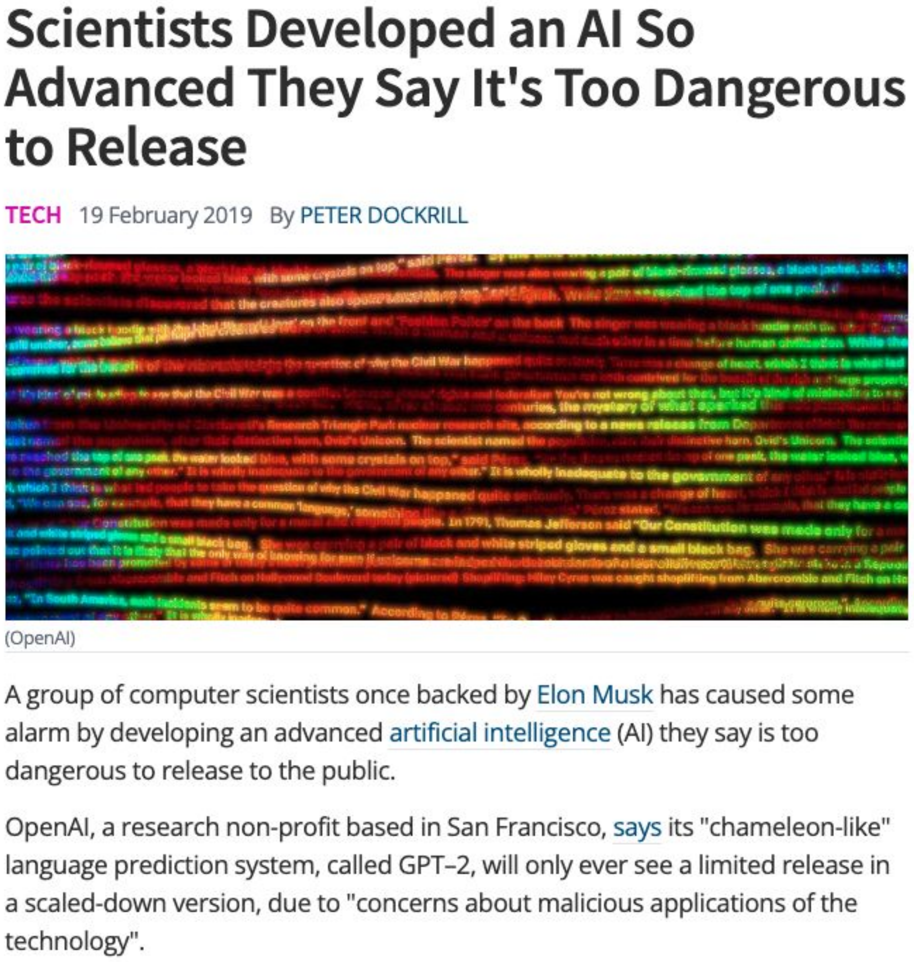
Zack Whittaker @zackwhittaker / 12:17 PM EST • February 17, 2019

Image Credits: OpenAI

A storm is brewing over a new language model, built by non-profit artificial intelligence research company OpenAI, which it says is so good at generating convincing, well-written text that it's worried about potential abuse.

That's angered some in the community, who have accused the company of reneging on a promise not to close off its research.

OpenAI said its new natural language model, GPT-2, was trained to predict the next word in a sample of 40 gigabytes of



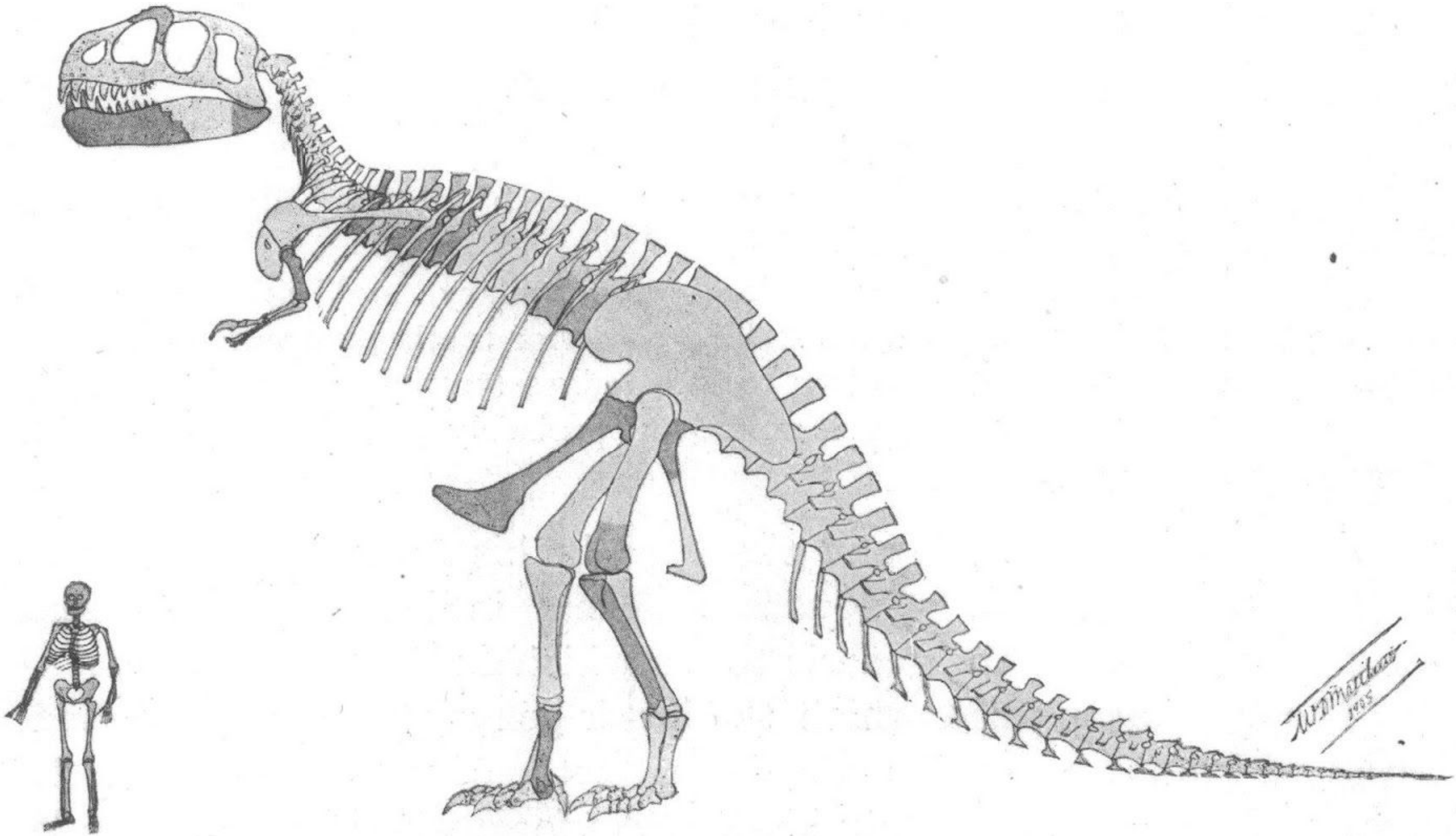
Scientists Developed an AI So Advanced They Say It's Too Dangerous to Release

TECH 19 February 2019 By PETER DOCKRILL

(OpenAI)

A group of computer scientists once backed by Elon Musk has caused some alarm by developing an advanced artificial intelligence (AI) they say is too dangerous to release to the public.

OpenAI, a research non-profit based in San Francisco, says its "chameleon-like" language prediction system, called GPT-2, will only ever see a limited release in a scaled-down version, due to "concerns about malicious applications of the technology".



GPT-2
1.5B Parameters

GPT-3
175B Parameters

Let’s define some buzz words

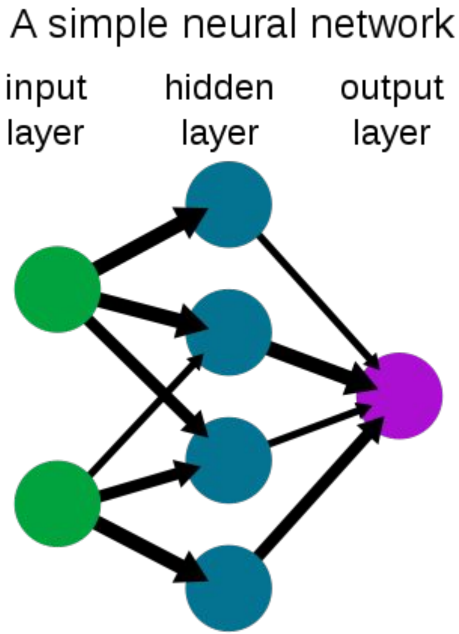
Artificial Intelligence: Intelligence demonstrated by machines, as opposed to the natural intelligence displayed by humans. [1]

Natural Language Processing: A subfield of linguistics and computer science concerned with the interactions between computers and human language

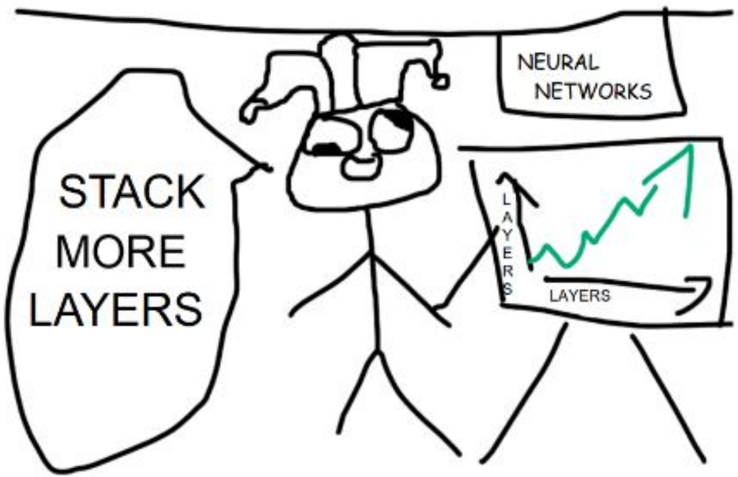
Natural Language Generation: A subfield of NLP concerned with the generation of human language

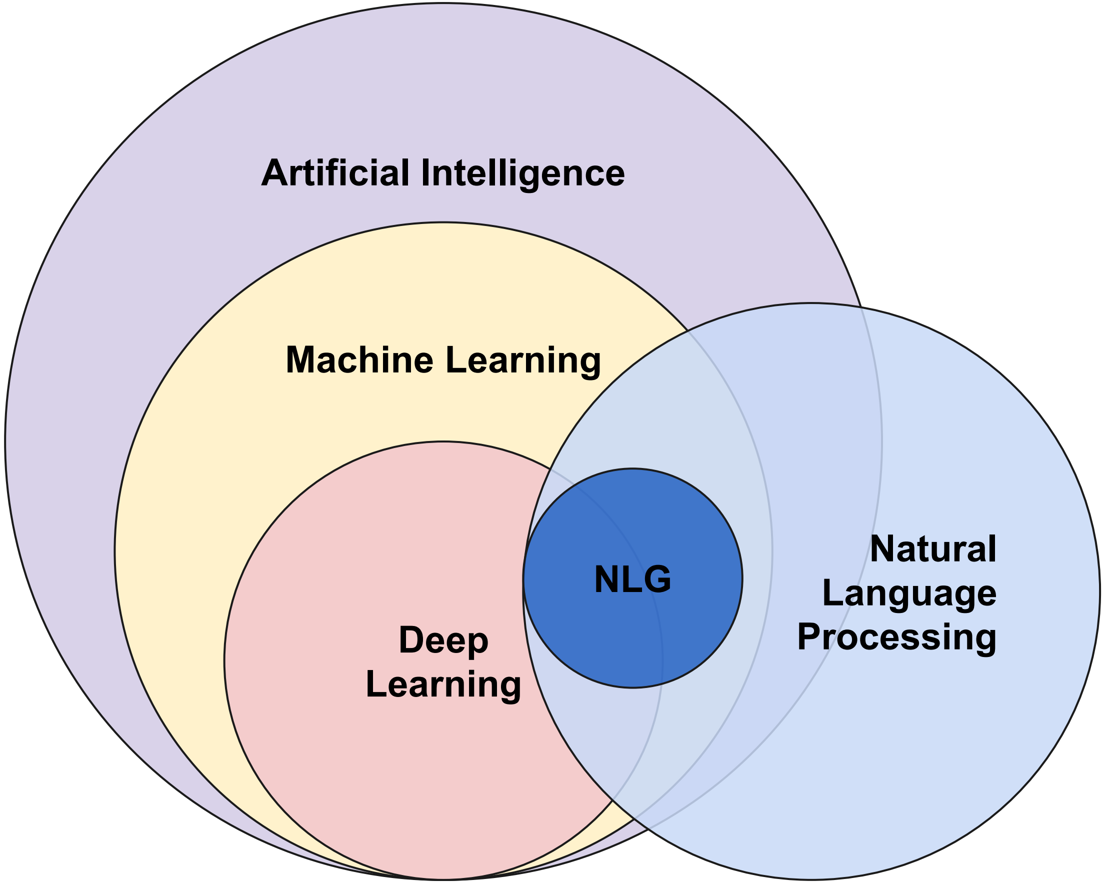
Machine Learning: Methods that allow a machine to “learn” how to perform a task without specifically being programmed how to do so

Neural networks: A class of methods composed of artificial neurons or nodes [2]



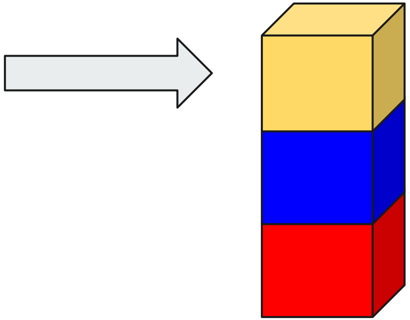
Deep learning: neural networks with a lot of layers




Artificial Intelligence
Machine Learning
Deep Learning
NLG
Natural Language Processing

Language Models

What does it mean to “model” something?



Given a stack of blocks, what will likely happen if we apply force in this direction?

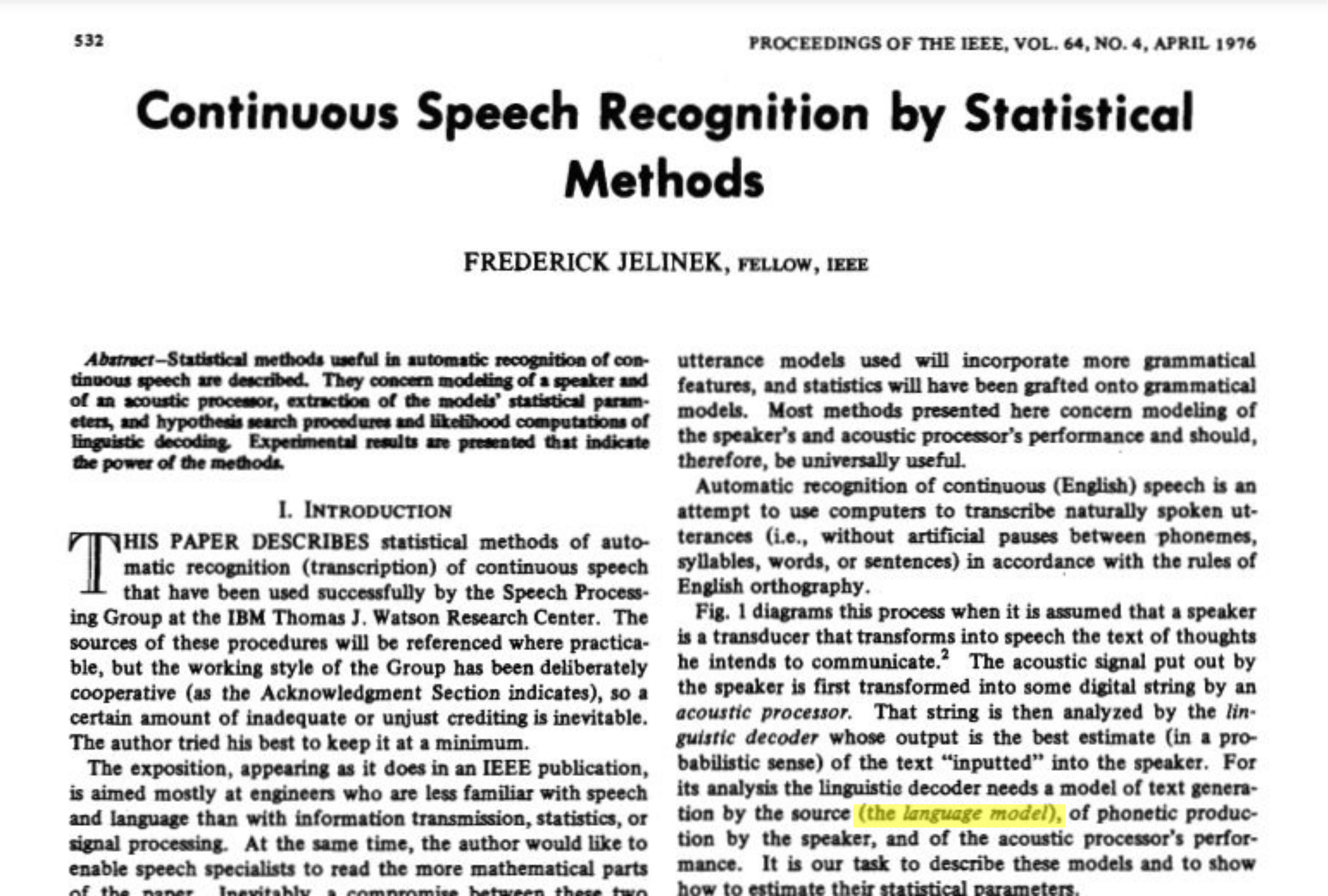
➔ The blocks will likely fall over

A good model of the world would **predict** that given the direction of the force, the blocks will likely fall over

What does it mean to model language?

- **Language Models estimate the probability of a sequence of words**
- This is easy for humans
  - Not so easy for machines
- Language modelling is the basis for all text generation applications

[LINK]



Continuous Speech Recognition by Statistical Methods

FREDERICK JELINEK, FELLOW, IEEE

Abstract—Statistical methods useful in automatic recognition of continuous speech are described. They concern modeling of a speaker and of an acoustic processor, extraction of the models' statistical parameters, and hypothesis search procedures and likelihood computations of linguistic decoding. Experimental results are presented that indicate the power of the methods.

I. Introduction

THIS PAPER DESCRIBES statistical methods of automatic recognition (transcription) of continuous speech that have been used successfully by the Speech Processing Group at the IBM Thomas J. Watson Research Center. The sources of these procedures will be referenced where practicable, but the working style of the Group has been deliberately cooperative (as the Acknowledgment Section indicates), so a certain amount of inadequate or unjust crediting is inevitable. The author tried his best to keep it at a minimum.

The exposition, appearing as it does in an IEEE publication, is aimed mostly at engineers who are less familiar with speech and language than with information transmission, statistics, or signal processing. At the same time, the author would like to enable speech specialists to read the more mathematical parts of the paper. Inevitably, a compromise between these two

utterance models used will incorporate more grammatical features, and statistics will have been grafted onto grammatical models. Most methods presented here concern modeling of the speaker's and acoustic processor's performance and should, therefore, be universally useful.

Automatic recognition of continuous (English) speech is an attempt to use computers to transcribe naturally spoken utterances (i.e., without artificial pauses between phonemes, syllables, words, or sentences) in accordance with the rules of English orthography.

Fig. 1 diagrams this process when it is assumed that a speaker is a transducer that transforms into speech the text of thoughts he intends to communicate.[2] The acoustic signal put out by the speaker is first transformed into some digital string by an *acoustic processor*. That string is then analyzed by the *linguistic decoder* whose output is the best estimate (in a probabilistic sense) of the text "inputted" into the speaker. For its analysis the linguistic decoder needs a model of text generation by the source (the *language model*), of phonetic production by the speaker, and of the acoustic processor's performance. It is our task to describe these models and to show how to estimate their statistical parameters.

You are now a language model

Examples

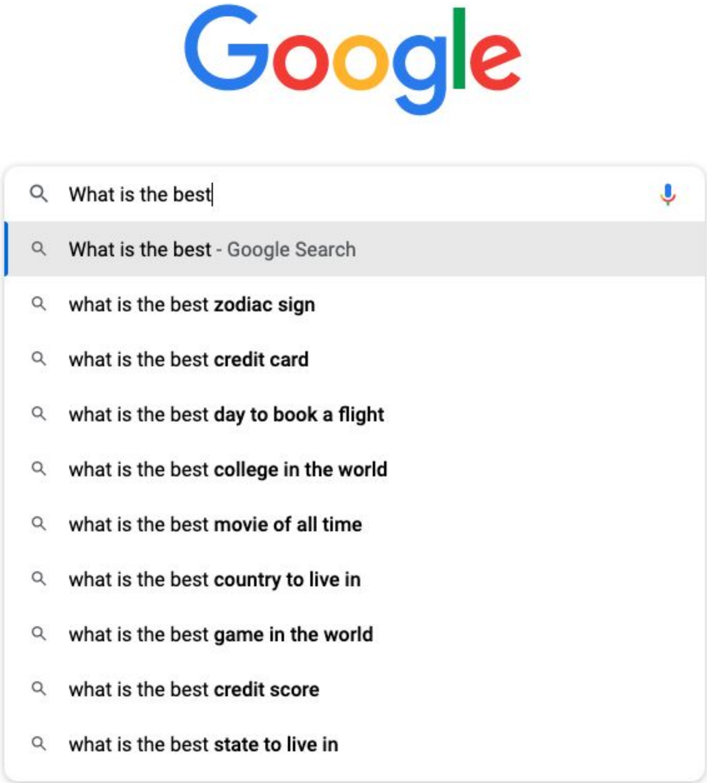
1. abab
2. baba
3. ababab

What letter is next?

ba_

You’ve encountered language models!

What are other examples?



What is the best

What is the best - Google Search

what is the best **zodiac sign**

what is the best **credit card**

what is the best **day to book a flight**

what is the best **college in the world**

what is the best **movie of all time**

what is the best **country to live in**

what is the best **game in the world**

what is the best **credit score**

what is the best **state to live in**

Which sentence is more probable?

Sentence A

The dog jumped over the fence

Sentence B

Fries Santa cheese dirt hello

Which sentence is more probable?

Sentence A

The dog jumped over the fence

Sentence B

The rock jumped over the fence

Which sentence is more probable?

Sentence A	Sentence B
The dog jumped over the fence	The elephant jumped over the fence

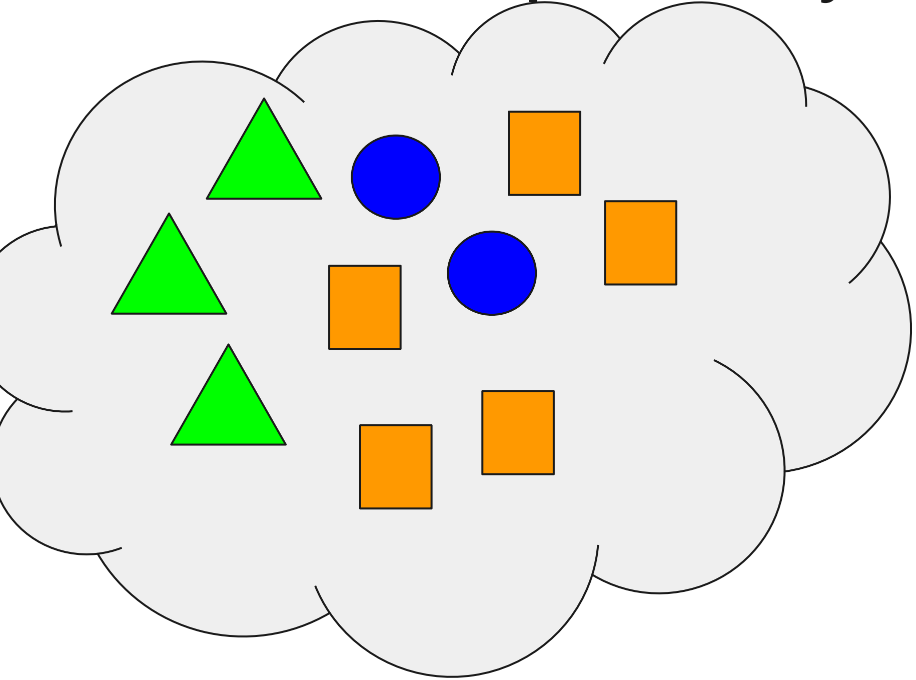
N-Gram Language Model

Some probability notation

$P(X)$ = the probability of event X

$P(X \mid Y)$ = the probability of event X **given** (or assuming) event Y

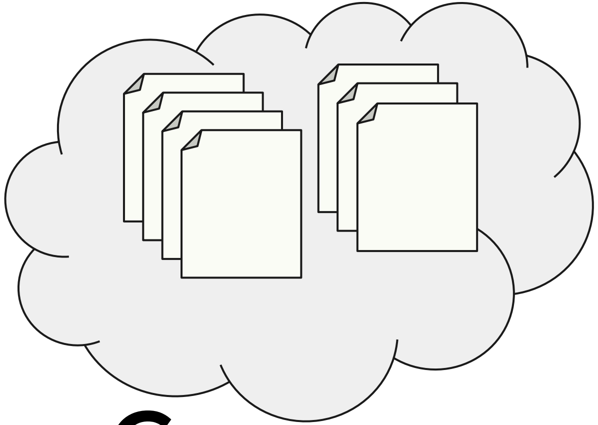
What is the probability of a circle?



$$P(●) = \frac{\#\,●}{\#\text{ total}}$$

$$P(●) = \frac{2}{10} = \frac{1}{5}$$

What is the probability of a sentence?

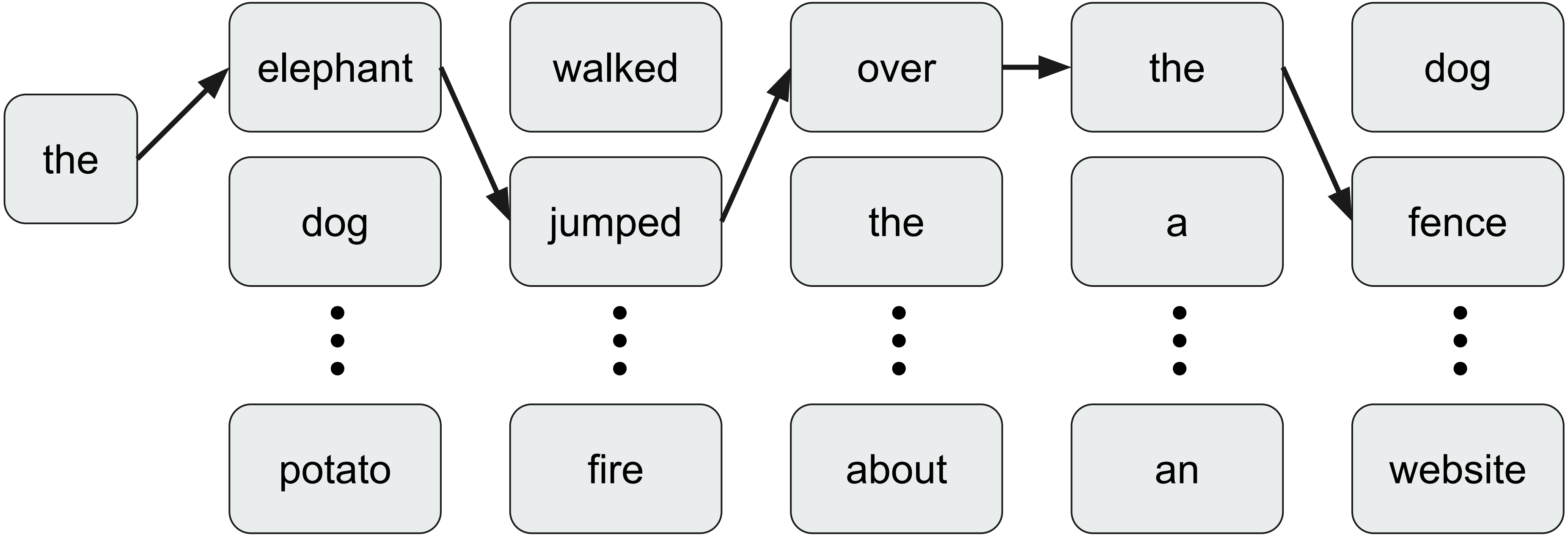


Corpus

$$P(\text{“the elephant jumped over the fence”}) = \frac{\#\ \text{“the elephant jumped over the fence”}}{\#\ \text{sentences in corpus}}$$

❓What if this sentence doesn’t exist in the corpus?

💡 Use smaller units



the

elephant
dog
⋮
potato

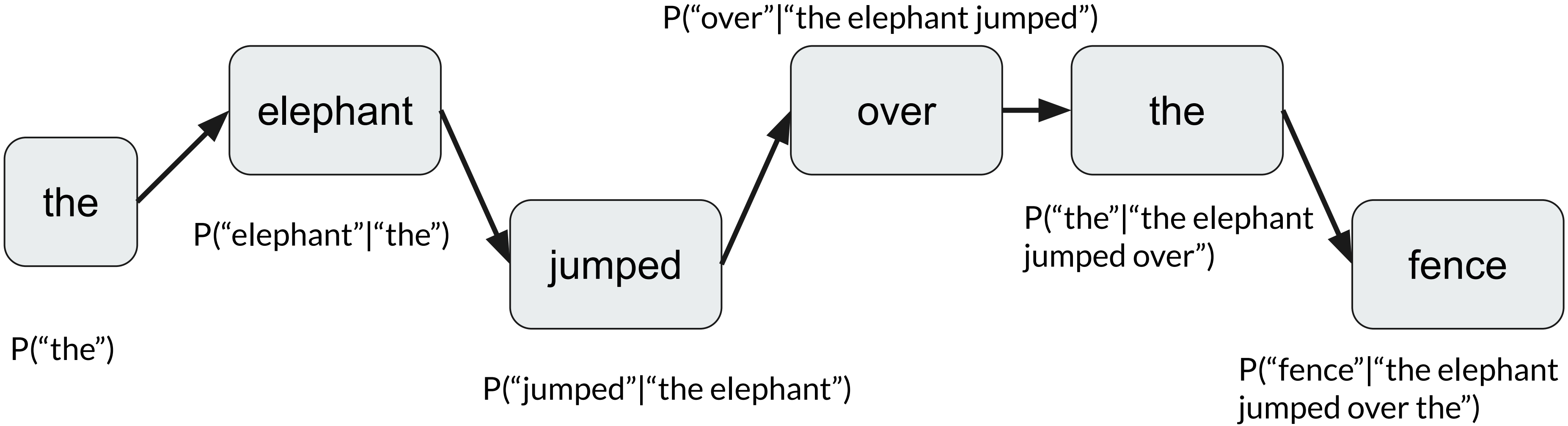
walked
jumped
⋮
fire

over
the
⋮
about

the
a
⋮
an

dog
fence
⋮
website

Use smaller units



the

elephant

jumped

over

the

fence

P(“the”)

P(“elephant”|“the”)

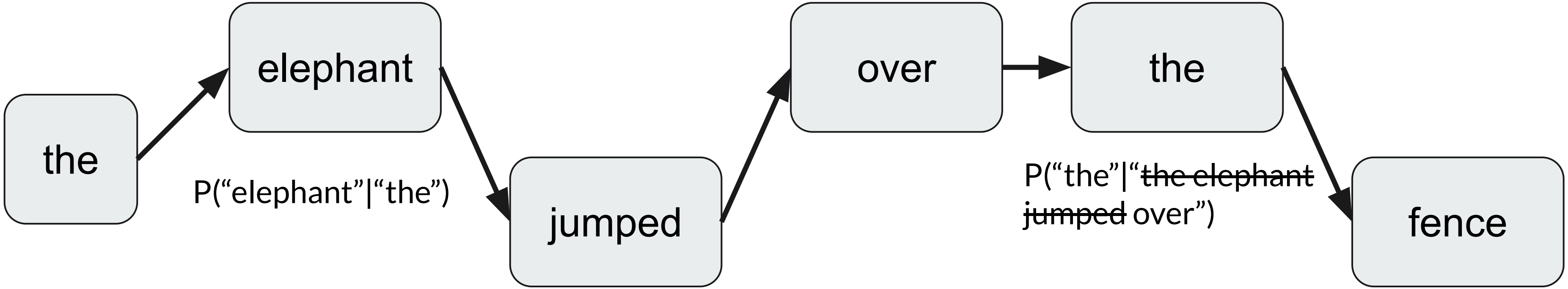
P(“jumped”|“the elephant”)

P(“over”|“the elephant jumped”)

P(“the”|“the elephant jumped over”)

P(“fence”|“the elephant jumped over the”)

Markov assumption



the

elephant

jumped

over

the

fence

P(“the”)

P(“elephant”|“the”)

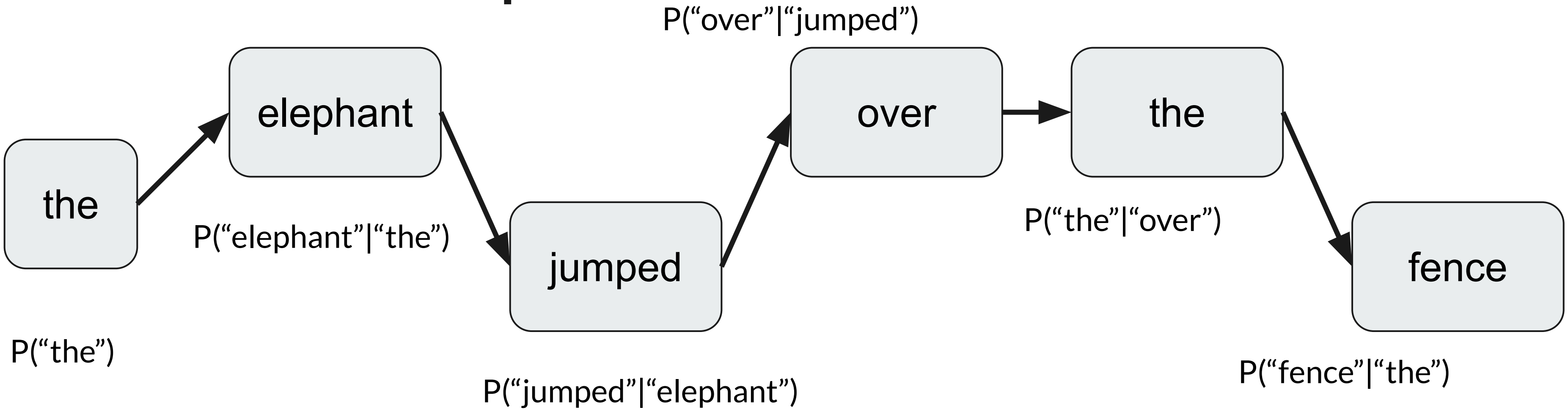
P(“jumped”|“~~the~~ elephant”)

P(“over”|“~~the elephant~~ jumped”)

P(“the”|“~~the elephant jumped~~ over”)

P(“fence”|“~~the elephant jumped over~~ the”)

Markov assumption



P(“the”)

the

P(“elephant”|“the”)

elephant

P(“jumped”|“elephant”)

jumped

P(“over”|“jumped”)

over

P(“the”|“over”)

the

P(“fence”|“the”)

fence

$$P(\text{“elephant”}|\text{“the”}) = \frac{\#\ \text{“the elephant”}}{\#\ \text{“the”}}$$

How do we use a language model to generate text?

What is the best ________ ?

zodiac `0.1`

credit card `0.05`

day to book a flight `0.002`

Choose the most probable word

How do we use a language model to generate text?

What is the best ________ ?

zodiac	0.1
credit card	0.05
day to book a flight	0.002

Randomly sample a word, weighted by probability

Demos

N-gram Demo:
https://colab.research.google.com/drive/1IKUNyzmOOm_6nqaLNrtuIiGAKLq8txHD?usp=sharing

GPT-2 Demo: https://transformer.huggingface.co/doc/gpt2-large

Reading for next week

In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation

Slide Credits

- A Visual Introduction to Language Models in NLP (Part 1: Intuition)
- Lena Voita’s NLP Course: Language Modeling