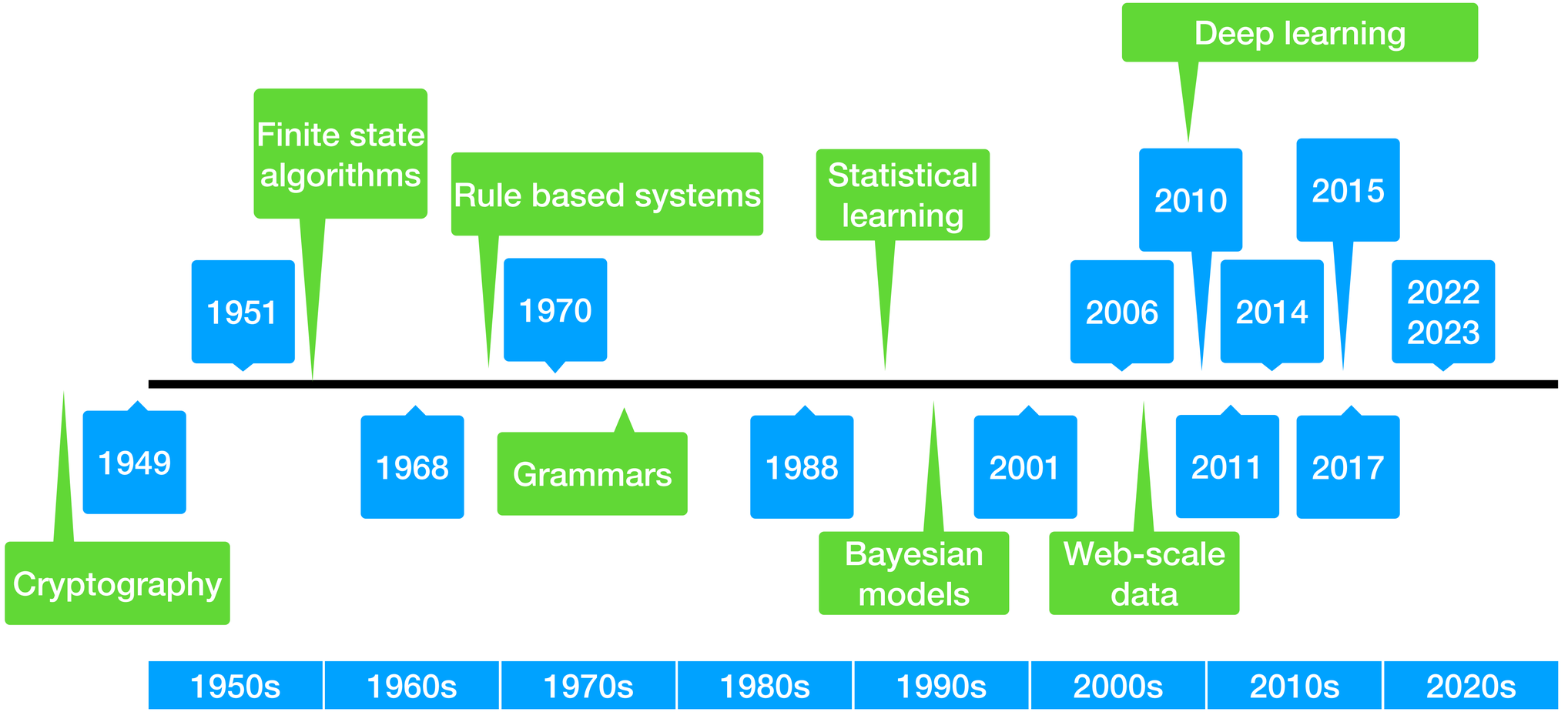
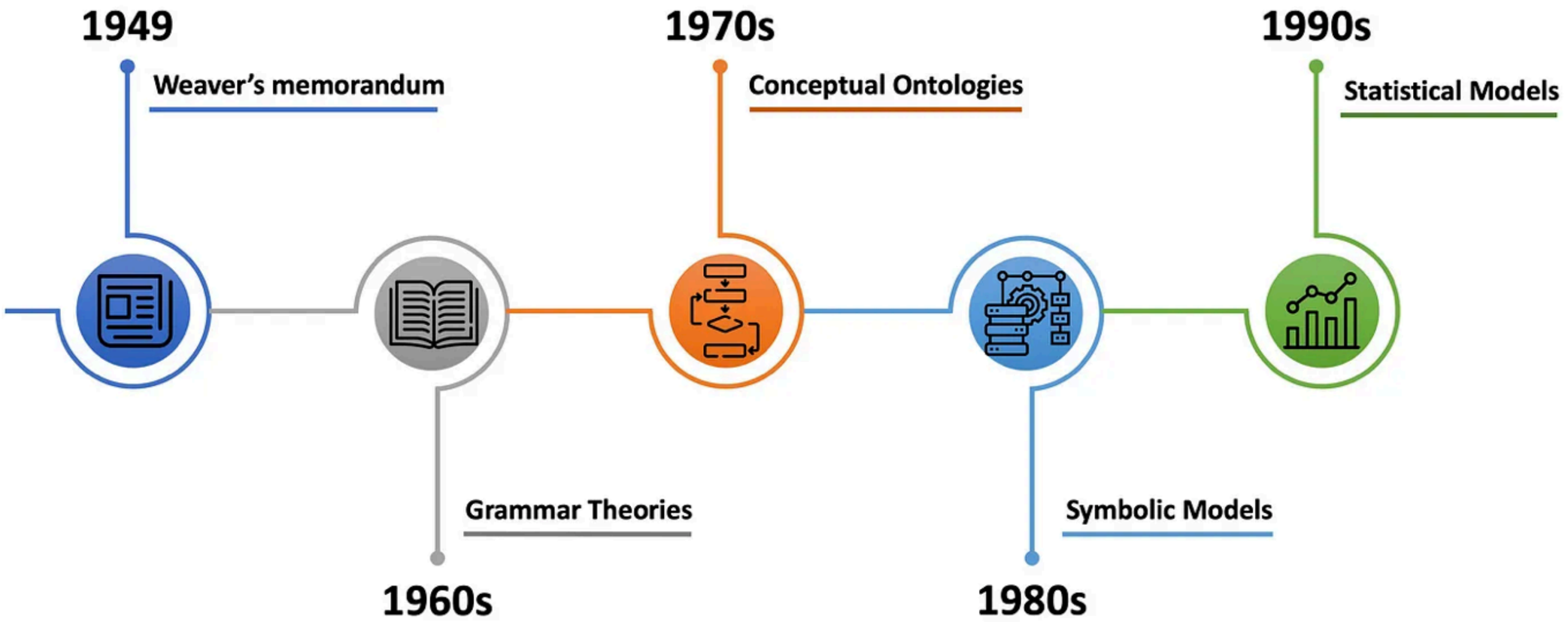
Brief History of NLP

Natural Language Processing

Fall 2023


Deep learning
Finite state algorithms
Rule based systems
Statistical learning
2010
2015
1951
1970
2006
2014
2022
2023
1949
1968
Grammars
1988
2001
2011
2017
Cryptography
Bayesian models
Web-scale data
1950s
1960s
1970s
1980s
1990s
2000s
2010s
2020s

NLP before deep learning



The big stages of NLP before the deep learning era.

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce

1949

Alan Turing

Instead we propose to try and see what can be done with a 'brain' which is more or less without a body, providing at most, organs of sight speech and hearing. We are then faced with the problem of finding suitable branches of thought for the machine to exercise its powers in. The following fields appear to me to have advantages:-

(i) Various games e.g. chess, noughts and crosses, bridge, poker.

(ii) The learning of languages.

(iii) Translation of languages.

(iv) Cryptography.

(v) Mathematics.

Of these (i), (iv), and to a lesser extent (iii) and (v) are good in that they require little contact with the outside world. For instance in order that the machine should be able to play chess its only organs need be 'eyes' capable of distinguishing the various positions on a specially made board, and means for announcing its own moves. Mathematics should preferably be restricted to branches where diagrams are not much used. Of the above possible fields the learning of languages would be the most impressive, since it is the most human of these activities. This field seems however to depend rather too much on sense organs and locomotion to be feasible.

The field of cryptography will perhaps be the most rewarding. There is a remarkably close parallel between the problems of the physicist and those of the cryptographer. The system on which a message is enciphered corresponds to the laws of the universe, the intercepted messages to the evidence available, the keys for a day or a message to important constants which have to be determined. The correspondence is very close, but the subject matter of cryptography is very easily dealt with by discrete machinery, physics not so easily.

Warren Weaver
Carlsbad, New Mexico
July 15, 1949

During the war a distinguished mathematician whom we will call P, an ex-German who had spent some time at the University of Istanbul and had learned Turkish there, told W.W. the following story.

Let's call him Peter

https://aclanthology.org/1952.earlymt-1.1.pdf

A mathematical colleague, let's call him Max, asks Peter to provide him a cipher.

Peter thinks, 'Max doesn't know that I can speak Turkish, so I'll encipher some Turkish text.' Peter reduces a sentence in Turkish into a column of five digit numbers.

Max comes back the next day and says he failed. All he could produce was some gibberish text "bu ne anlama geliyor"

Peter was amazed that Max had produced (with some minor errors) the original message in Turkish!

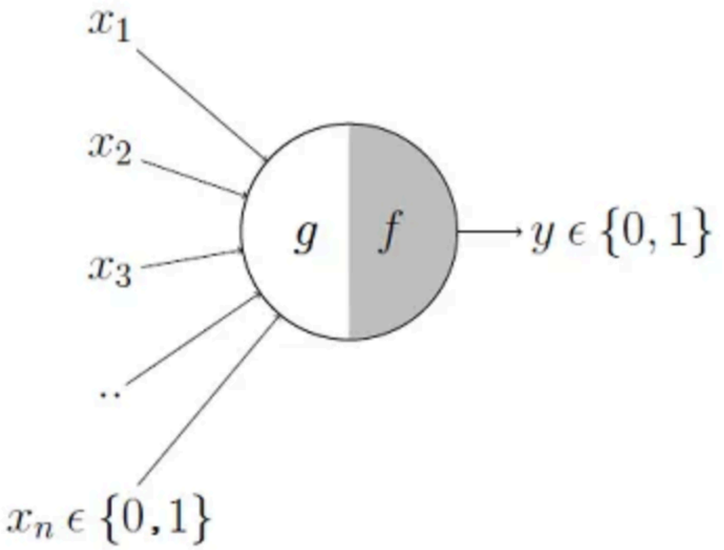
https://aclanthology.org/1952.earlymt-1.1.pdf

"Also knowing nothing official about, but having guessed and inferred considerable about, 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.

Every single NLP system that produces output uses what we now call a **decoder**

https://aclanthology.org/1952.earlymt-1.1.pdf

A more general basis for hoping that a computer could be designed which would cope with a useful part of the problem of translation is to be found in a theorem which was proved in 1943 by McCulloch and Pitts.



This is where it all began..

McCulloch-Pitts Neuron

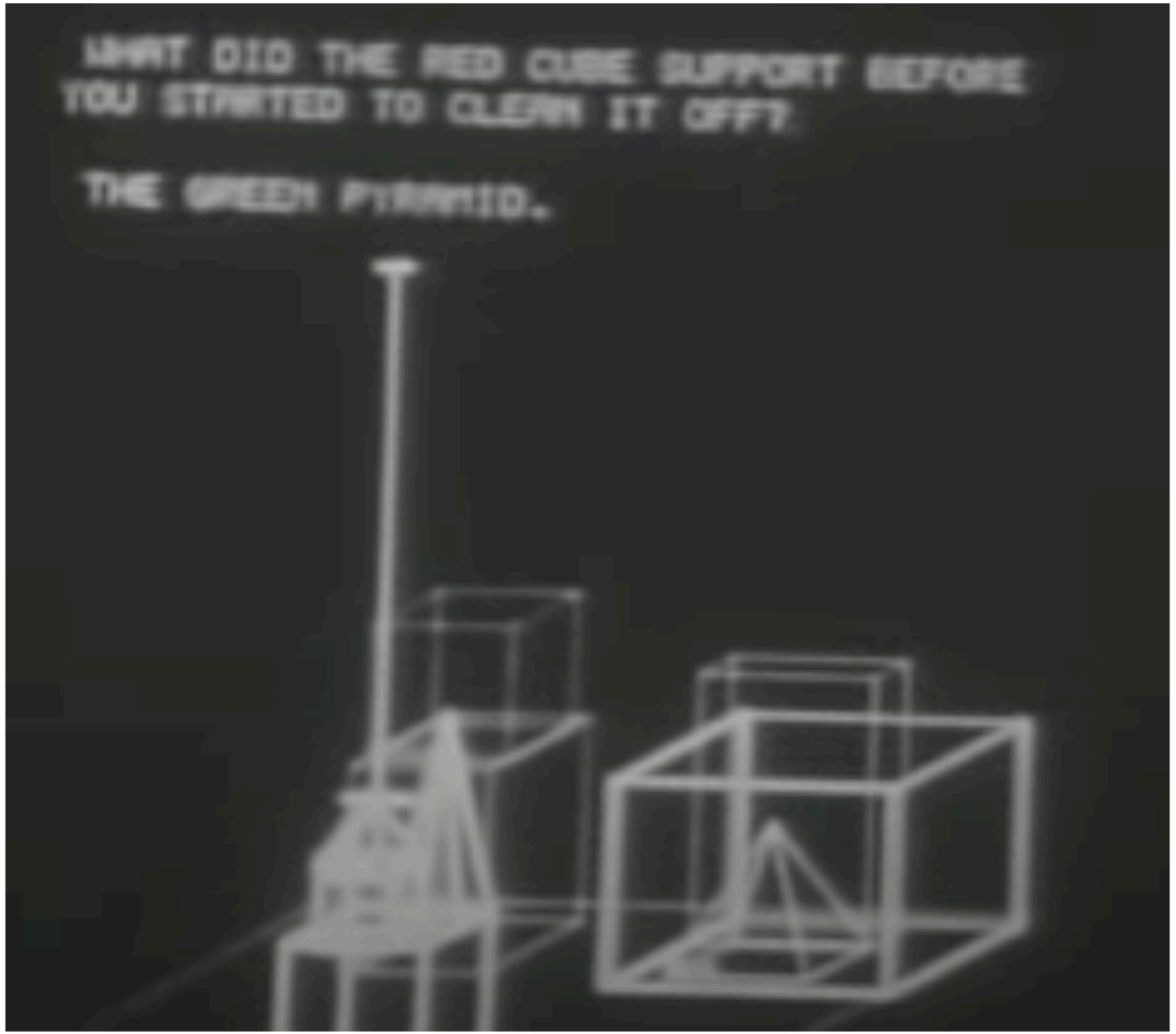
https://towardsdatascience.com/mcculloch-pitts-model-5fdf65ac5dd1

https://aclanthology.org/1952.earlymt-1.1.pdf

1968

SHRDLU
Terry Winograd

Rule-based NLP
(grounding)



https://www.youtube.com/watch?v=QAJz4YKUwqw

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK.
Person: What does the box contain?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: What is the pyramid supported by?
Computer: THE BOX.
Person: How many blocks are not in the box?
Computer: FOUR OF THEM.
Person: Is at least one of them narrower than the one which i told you to pick up?
Computer: YES, THE RED CUBE.

~1960-1970

~~TOP SECRET UMBRA~~

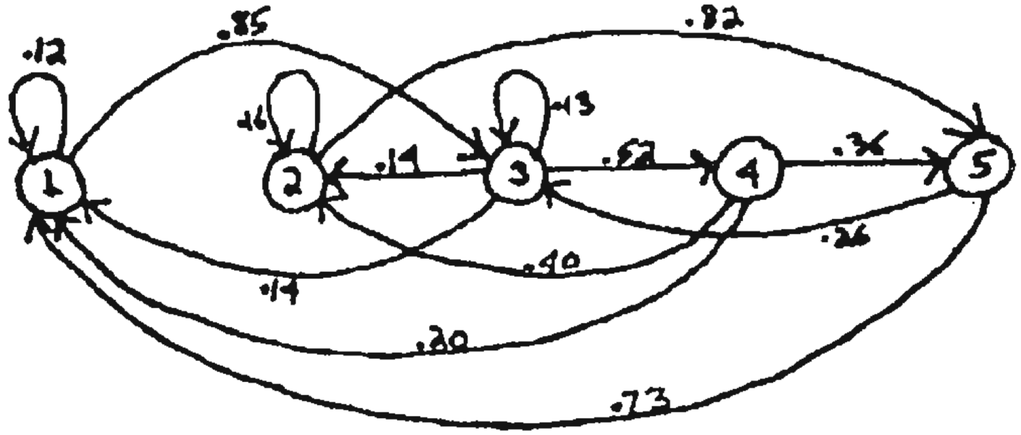
An Application of PTAH to the Voynich Manuscript (U)

BY MARY E. D'IMPERIO

~~*Top Secret Umbra*~~

(U) This article is the second in a series of studies applying some modern statistical techniques to the problems posed by the Voynich manuscript. This study attempts to discover and demonstrate regularities of patterning in the Voynich text subjectively noted by many earlier students of the manuscript. Three separate PTAH studies are described, attacking the Voynich text at three levels: single symbols, whole "words," and a carefully chosen set of substrings within "words." These analyses are applied to samples of text from the "Biological B" section of the manuscript, in Currier's transcription. A brief general characterization of PTAH is provided, with an explanation of how it is used in the present application.

Hidden Markov Models



UNCLASSIFIED

State Transition Diagram (U)

State	Associated English Letters
1	t b c j m k p v z w q
2	s y e d g
3	a o h i u
4	n r f l x
5	word space

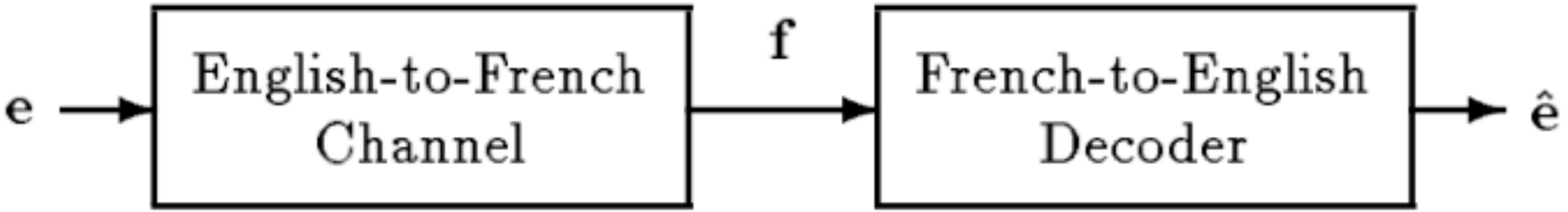
https://ia601301.us.archive.org/view_archive.php?archive=/19/items/NSA-FOIA-Vault/Tech Journals.7z&file=Tech Journals/Application_of_PTAH.pdf

1988

The Candide System for Machine Translation

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, Luboš Ureš

IBM Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598



$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{e} \mid \mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{f} \mid \mathbf{e}) \Pr(\mathbf{e})$$

Bayes Rule

2001

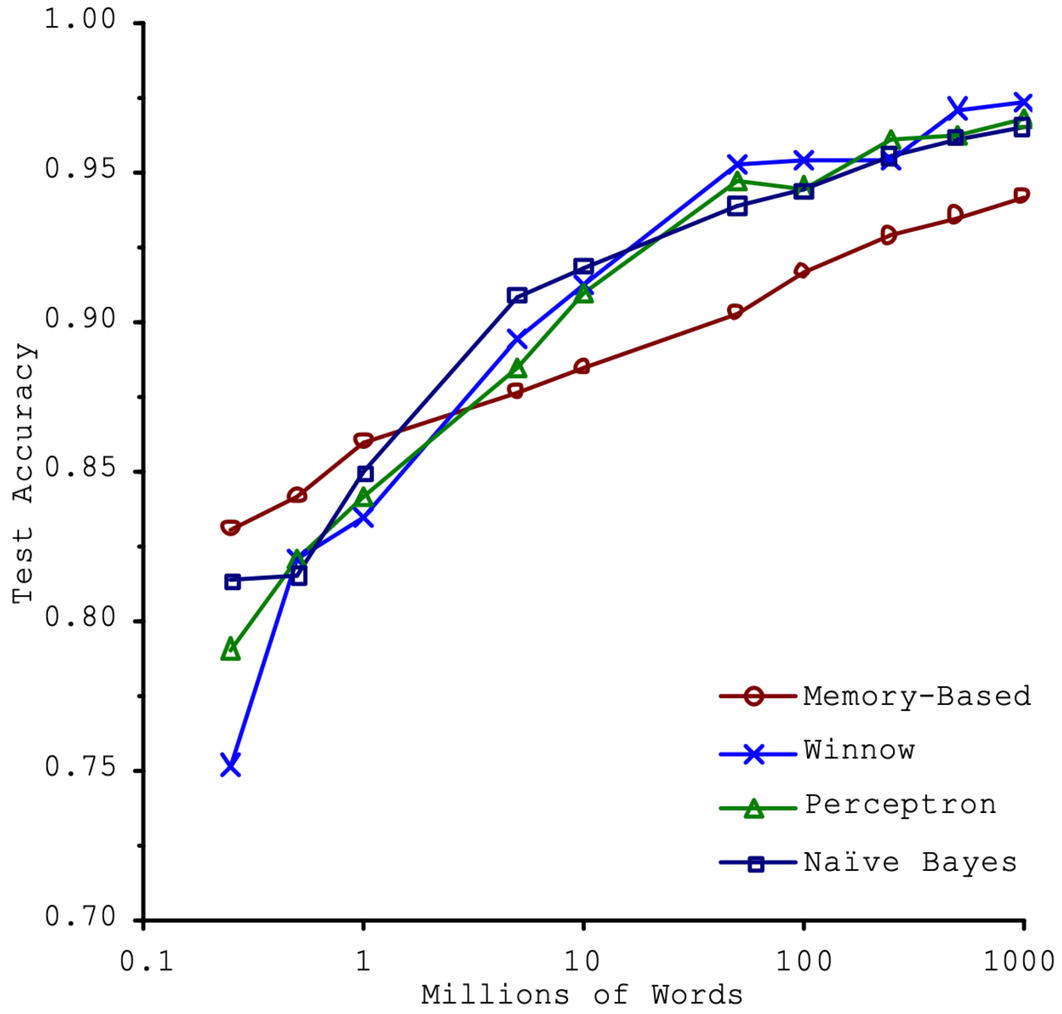


Figure 1. Learning Curves for Confusion Set Disambiguation

Scaling to very very large corpora for natural language disambiguation. Banko and Brill, 2001

2006

Franz Och, Apr 28 2006

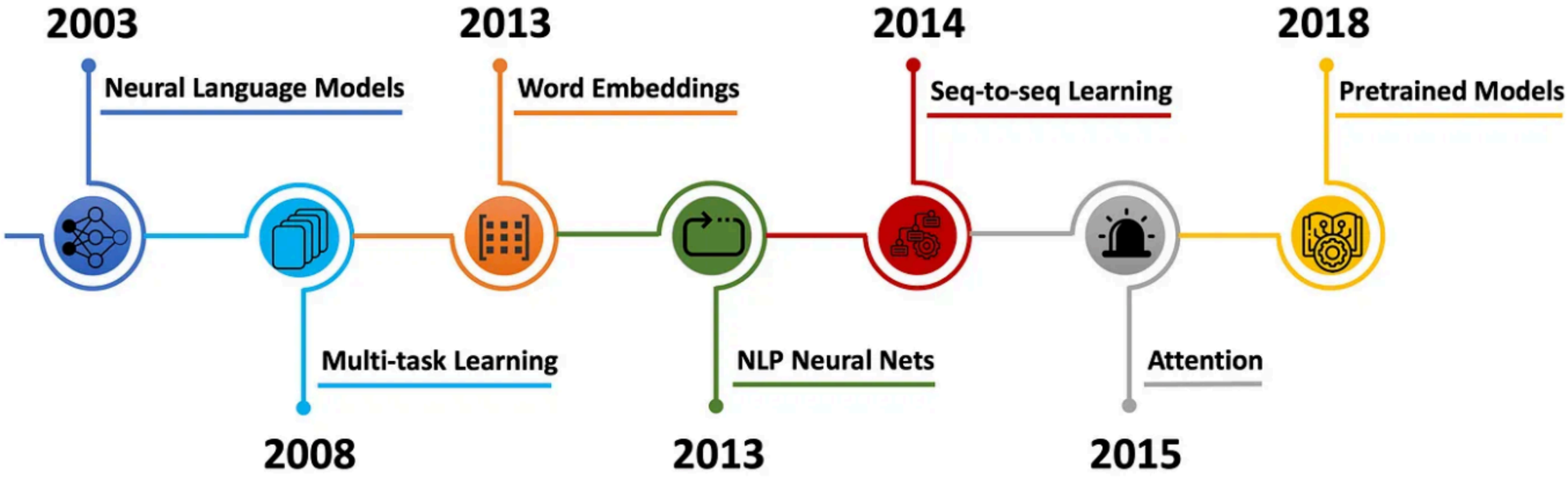
Most state-of-the-art commercial machine translation systems in use today have been developed using a rules-based approach ...

Several research systems, including ours, take a different approach: **we feed the computer with billions of words of text**, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages.

We then apply statistical learning techniques to build a translation model.

http://googleresearch.blogspot.ca/2006/04/statistical-machine-translation-live.html

NLP in the deep learning era

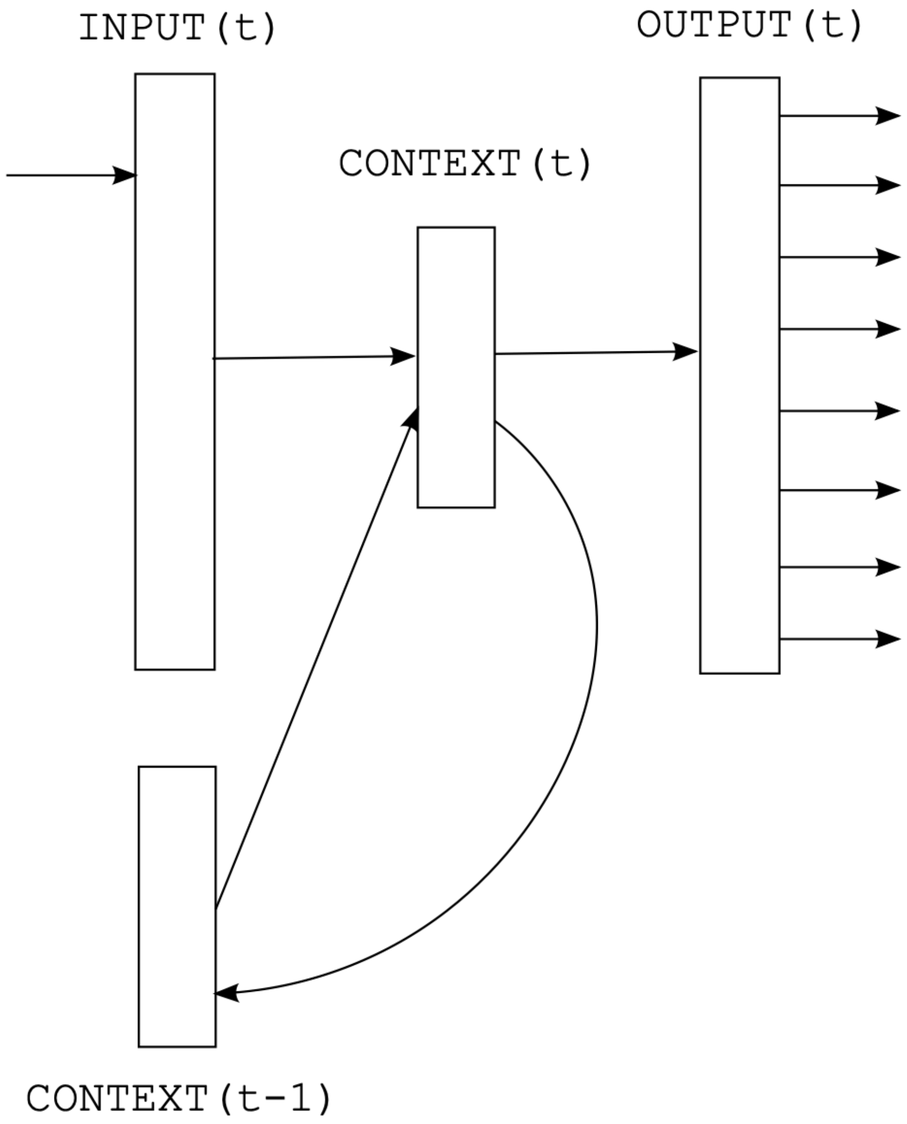


The big stages of NLP in the deep learning era.

https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37

2010

Mikolov, *Recurrent Networks for Language Models* (RNN-LM)



http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf

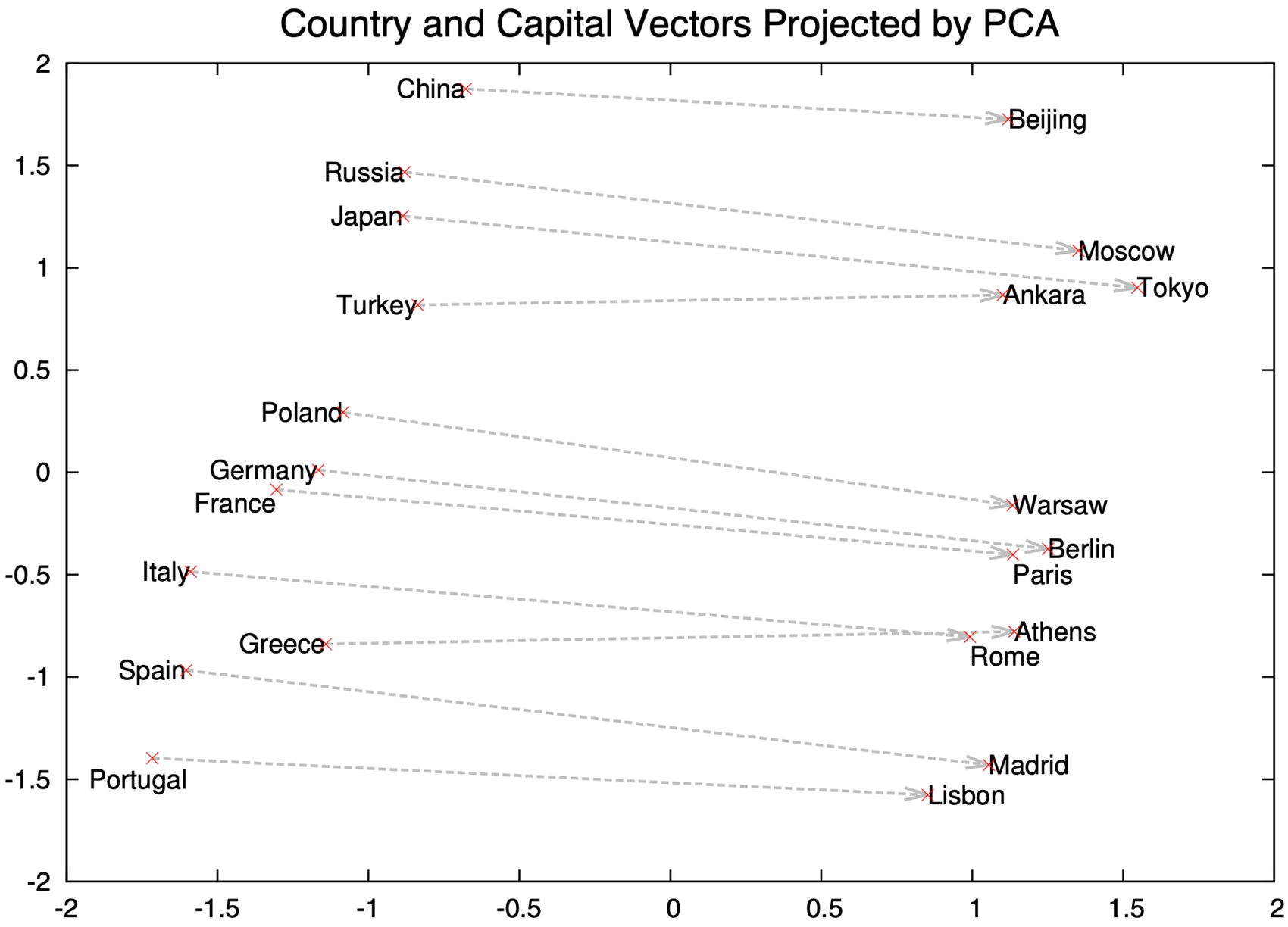
2011

IBM Watson plays Jeopardy!

and wins ...

2012

Mikolov, Continuous Bag of Words (word2vec)



https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

2014

“You can’t cram the meaning of a whole %&!$# sentence into a single $&!#* vector!”

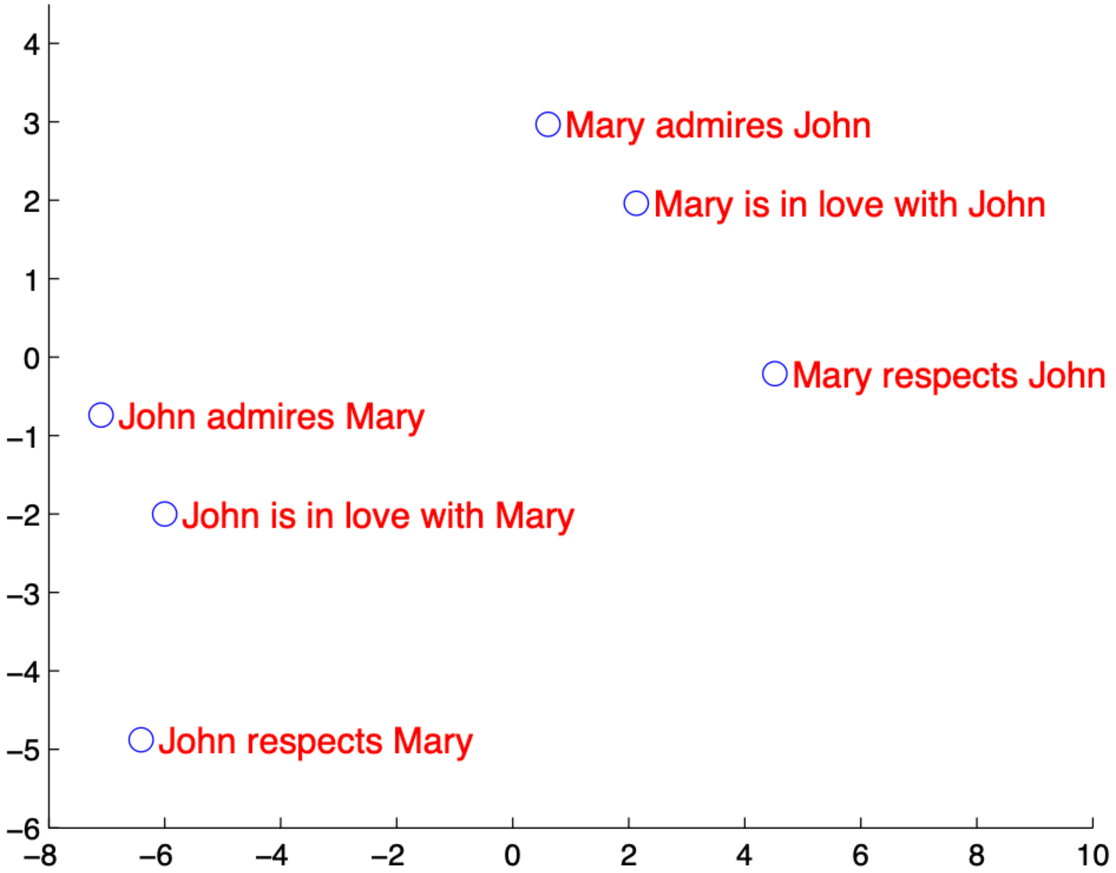
https://yoavartzi.com/sp14/slides/mooney.sp14.pdf

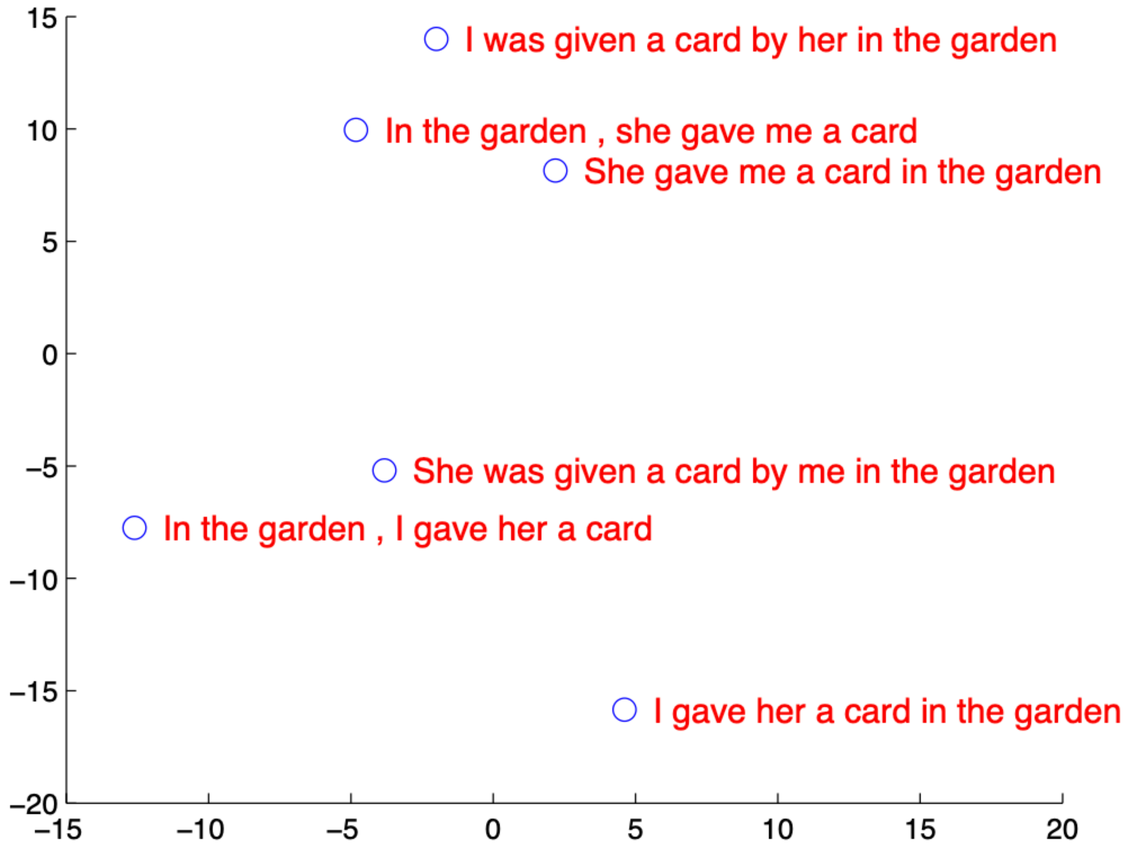
Ray Mooney

Invited talk at the ACL 2014 Workshop on Semantic Parsing

2014

Recurrent Neural Networks for Translation



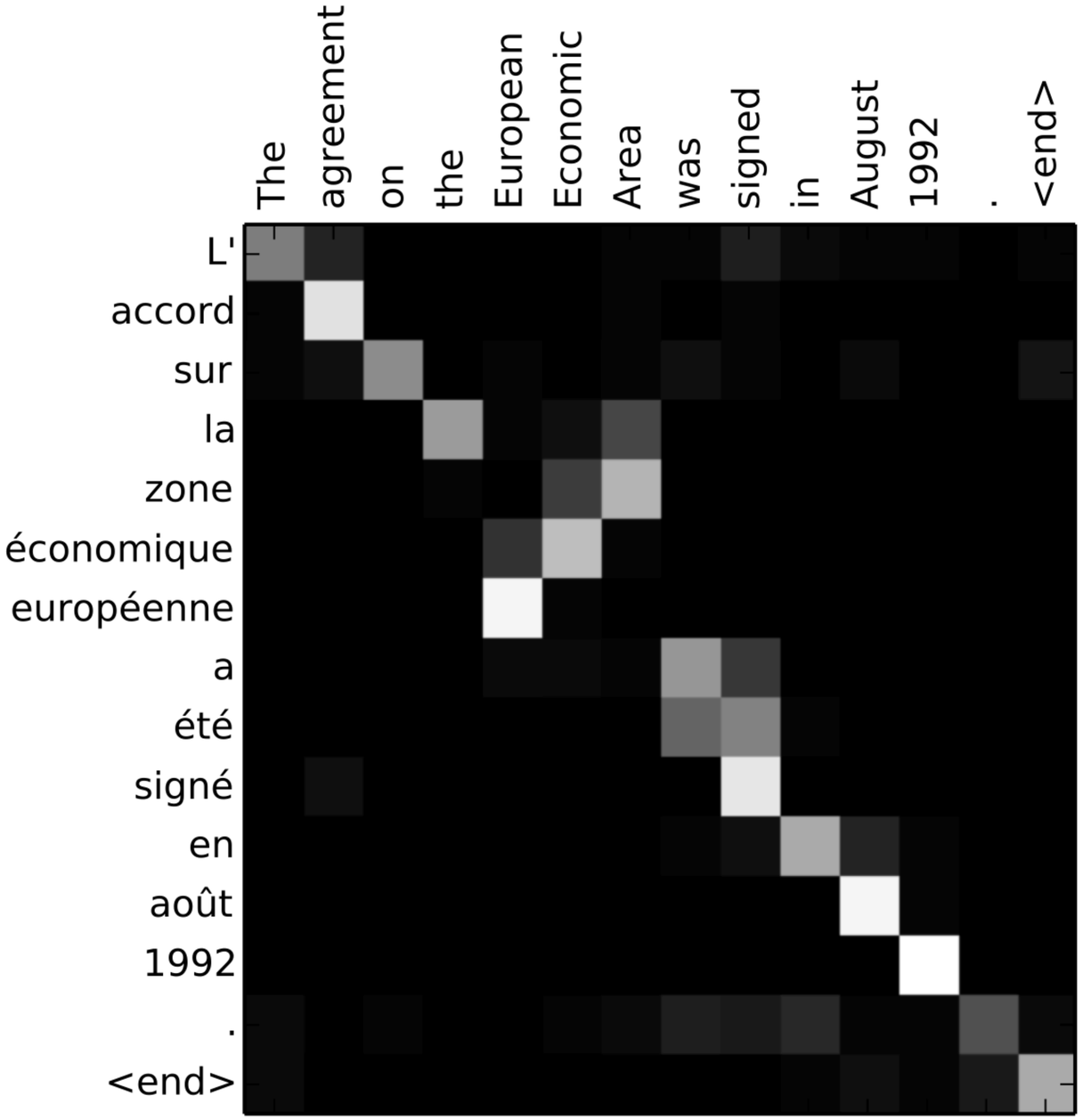


Sequence to Sequence Learning with Neural Networks https://arxiv.org/abs/1409.3215

2015

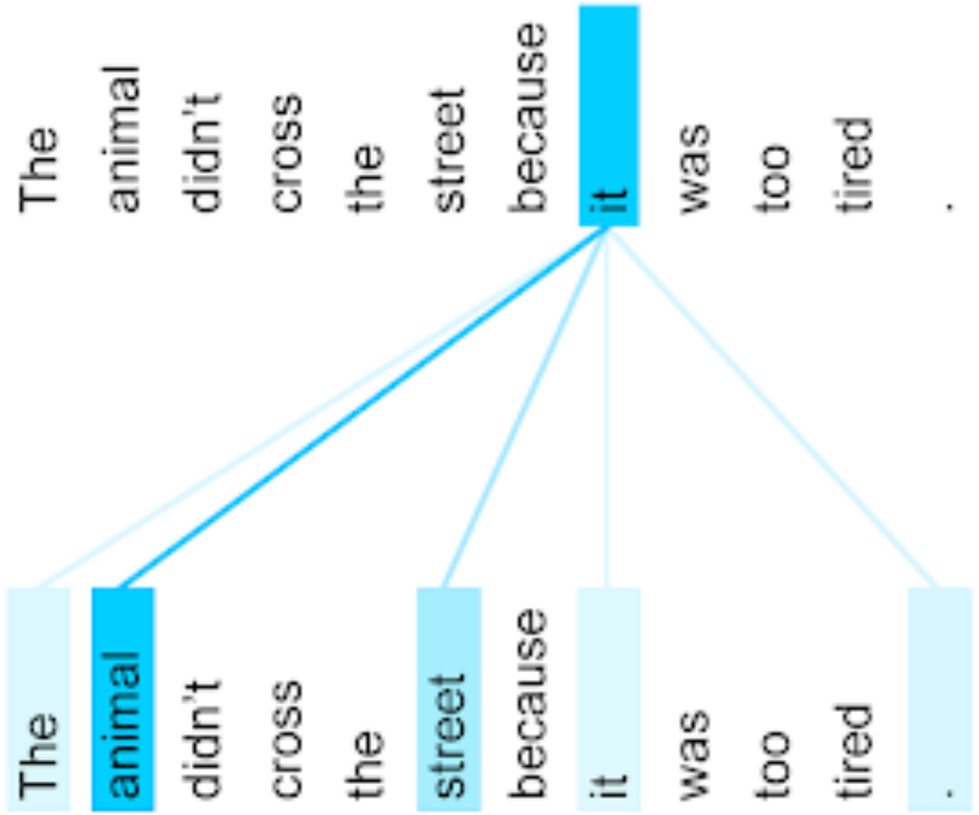
Skype Translator

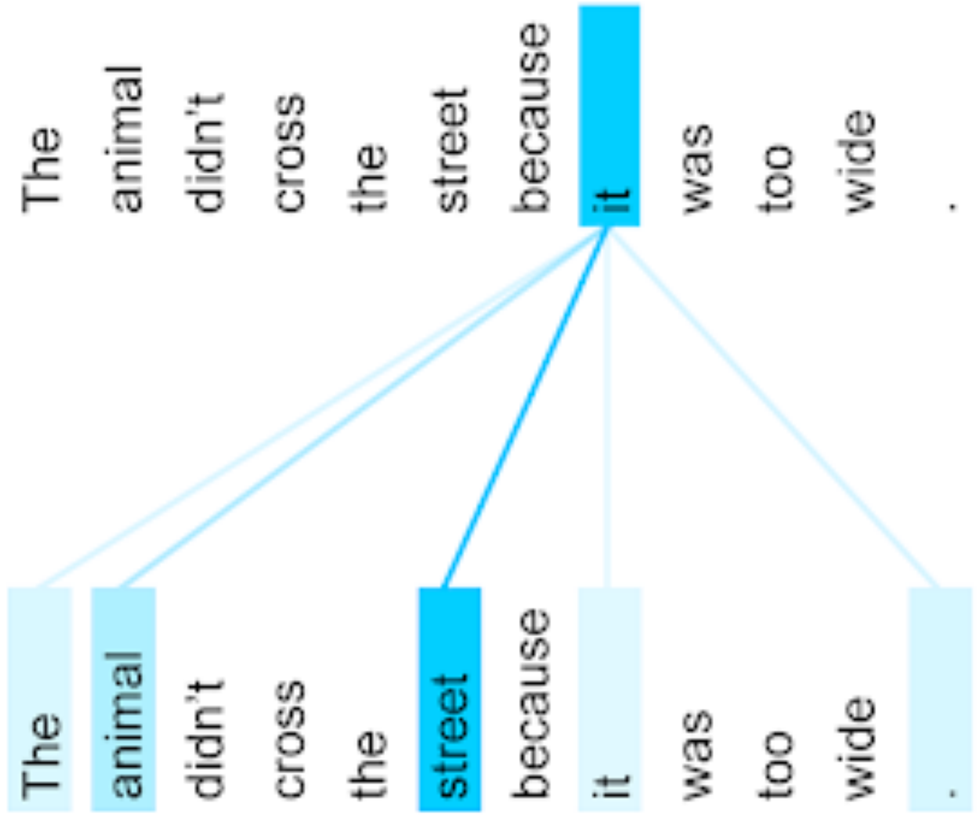
Attention


The
agreement
on
the
European
Economic
Area
was
signed
in
August
1992
.
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
1992
.
<end>

2017

Self-Attention

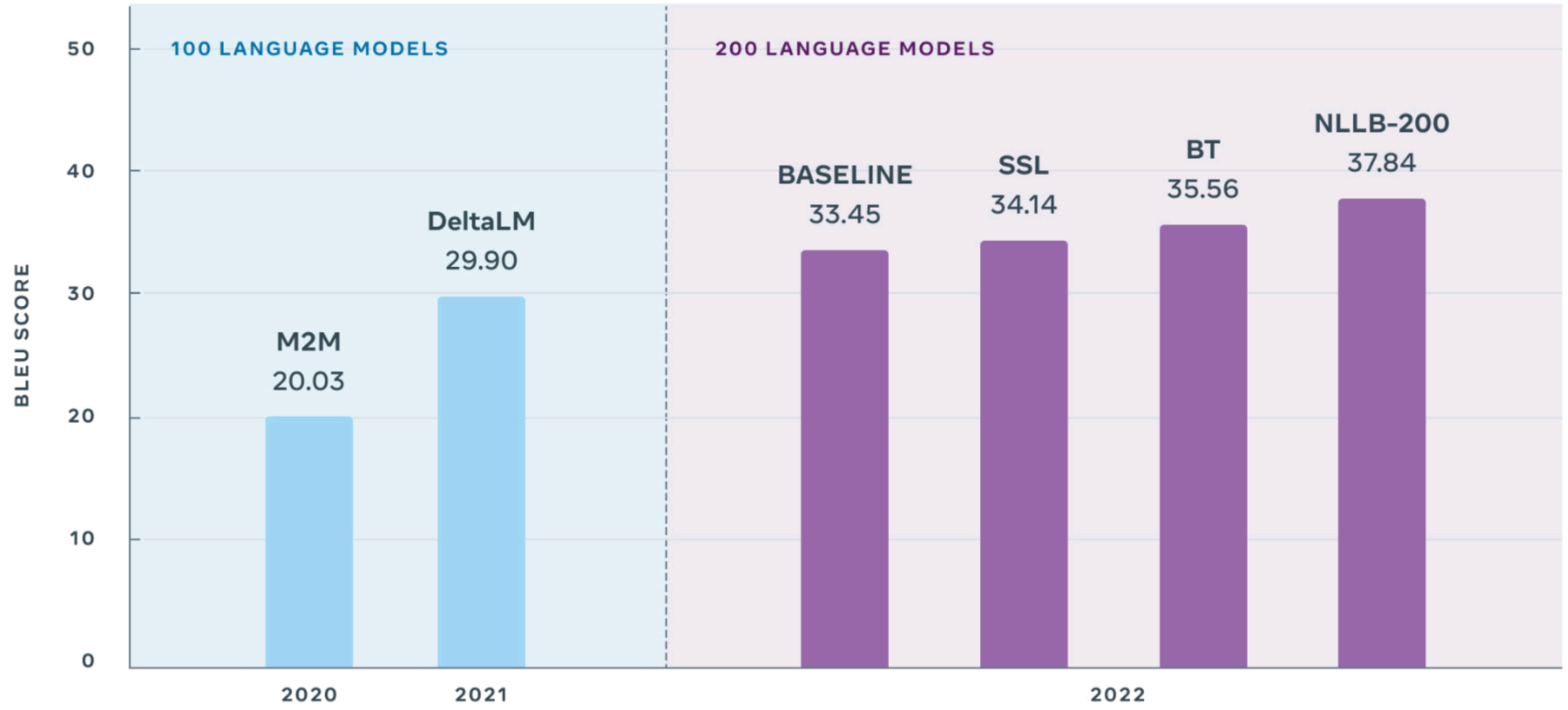




https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

2022

High-quality machine translation for 200 languages

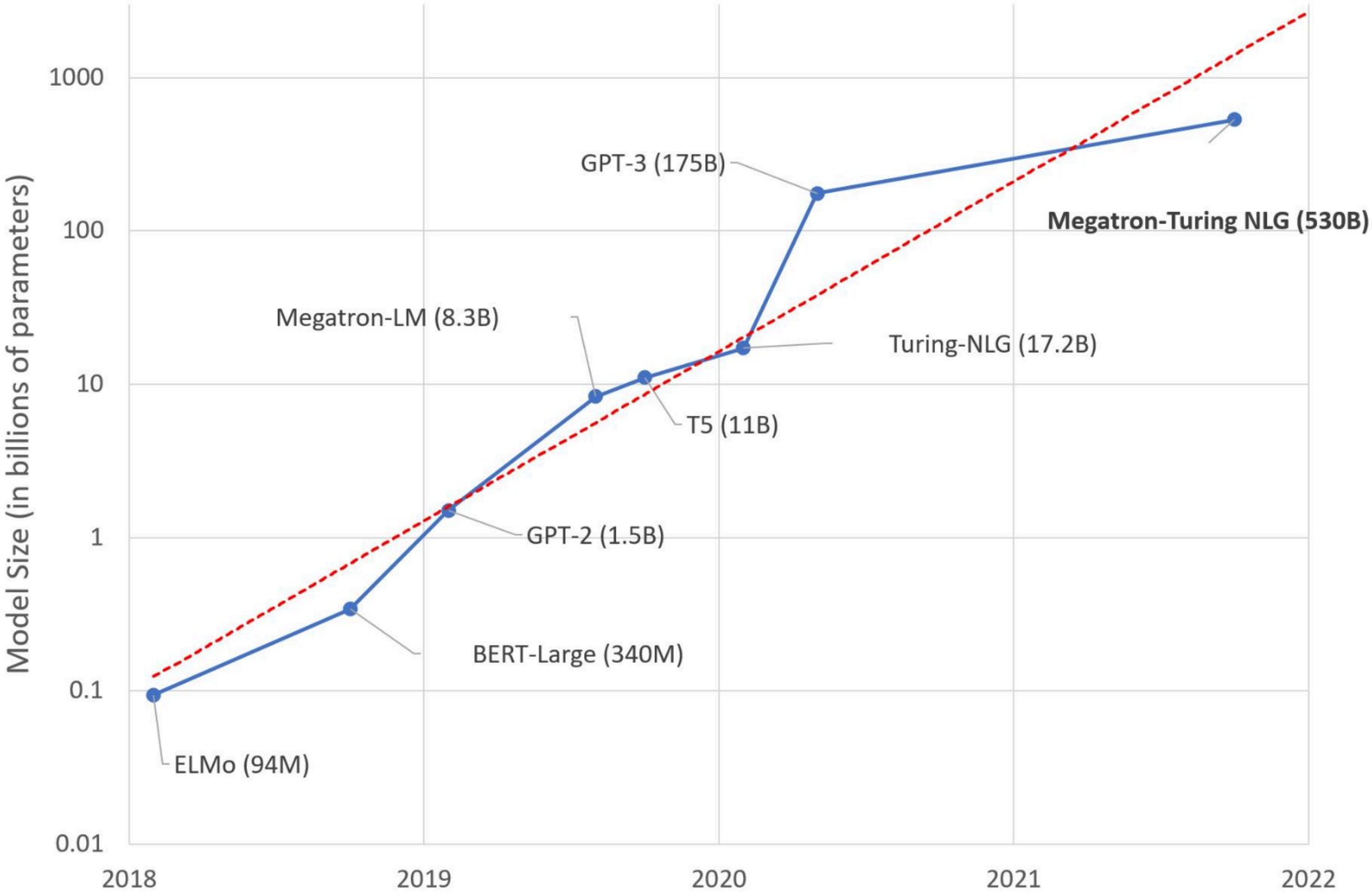


This graphic shows average BLEU score on FLORES-101 translations to and from English into 100 languages. On the left there are two published state-of-the-art models, M2M and Delta LM, that support 100 languages. Models on the right support 200 languages: A baseline Transformer model with 3.3B parameters, the baseline model with self-supervised learning (SSL), the baseline model with back translation (BT), and NLLB-200, a large mixture-of-experts based model that leverages both self-supervised learning and back translation.

https://ai.meta.com/blog/nllb-200-high-quality-machine-translation/

2023

Self-Supervised Pre-trained Language Models

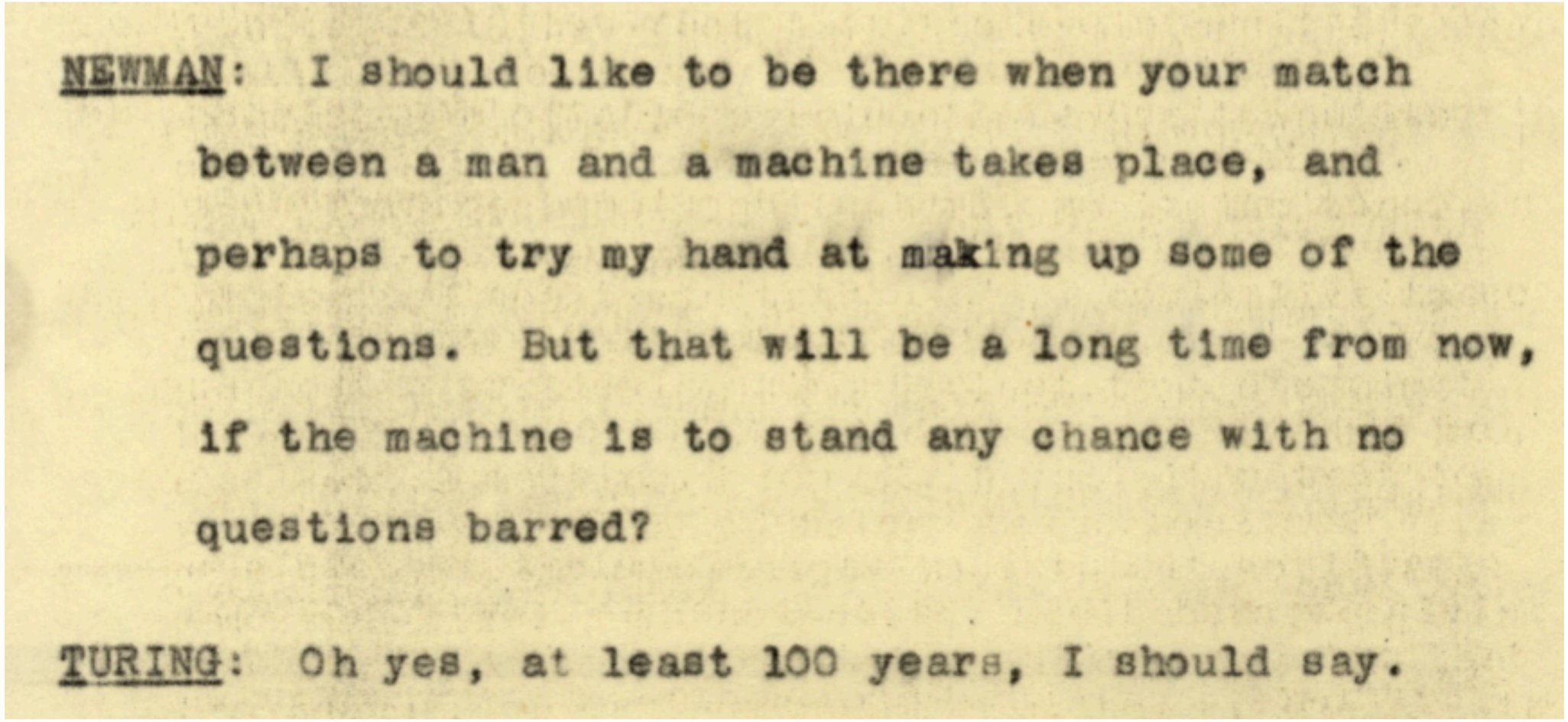


GPT-4 is estimated to have 500B-**1000B** parameters; trained using **2.2e25 FLOPs**

https://huggingface.co/blog/large-language-models

1951

The Turing Test

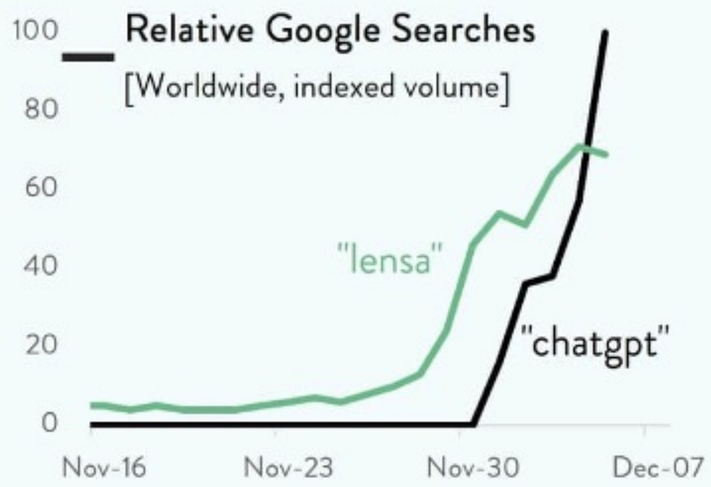


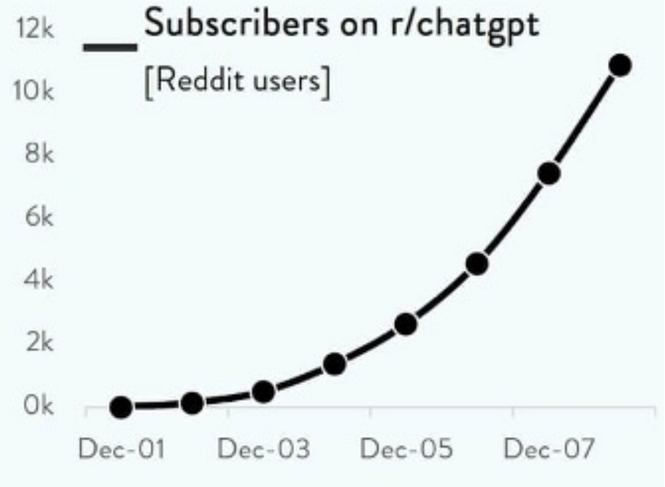
NEWMAN: I should like to be there when your match between a man and a machine takes place, and perhaps to try my hand at making up some of the questions. But that will be a long time from now, if the machine is to stand any chance with no questions barred?

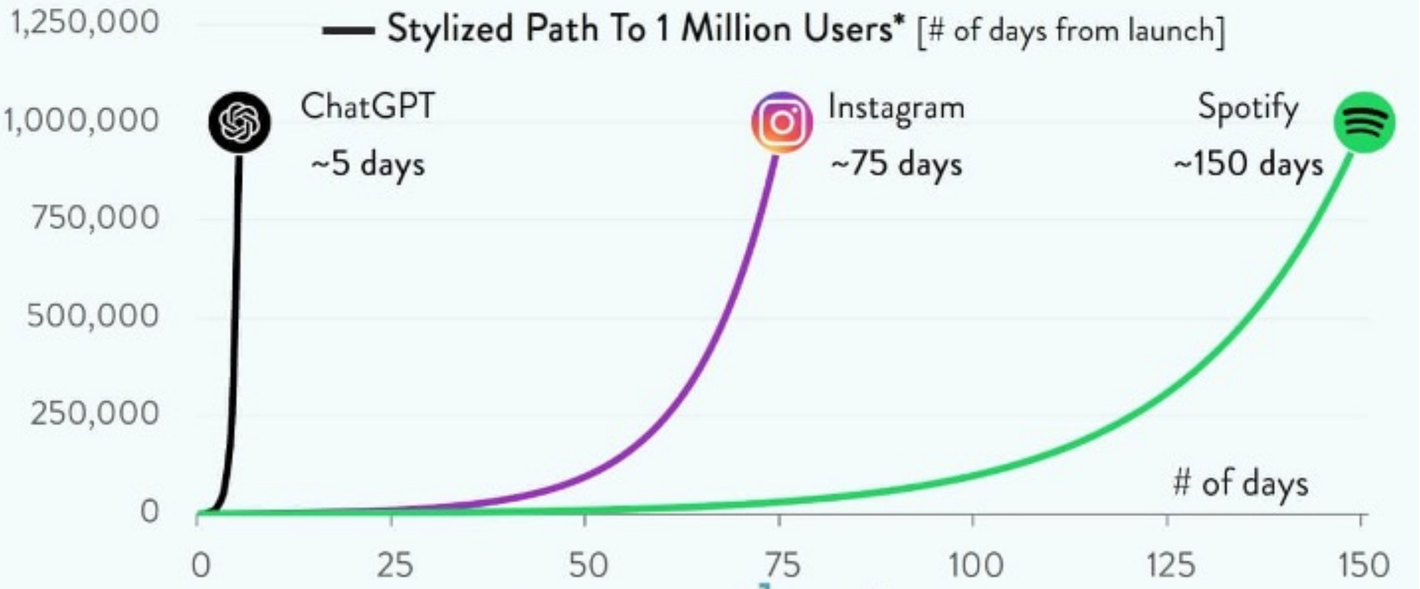
TURING: Oh yes, at least 100 years, I should say.

'Can digital computers think?'. Interview with Alan Turing.
BBC Third Programme, 15 May 1951.

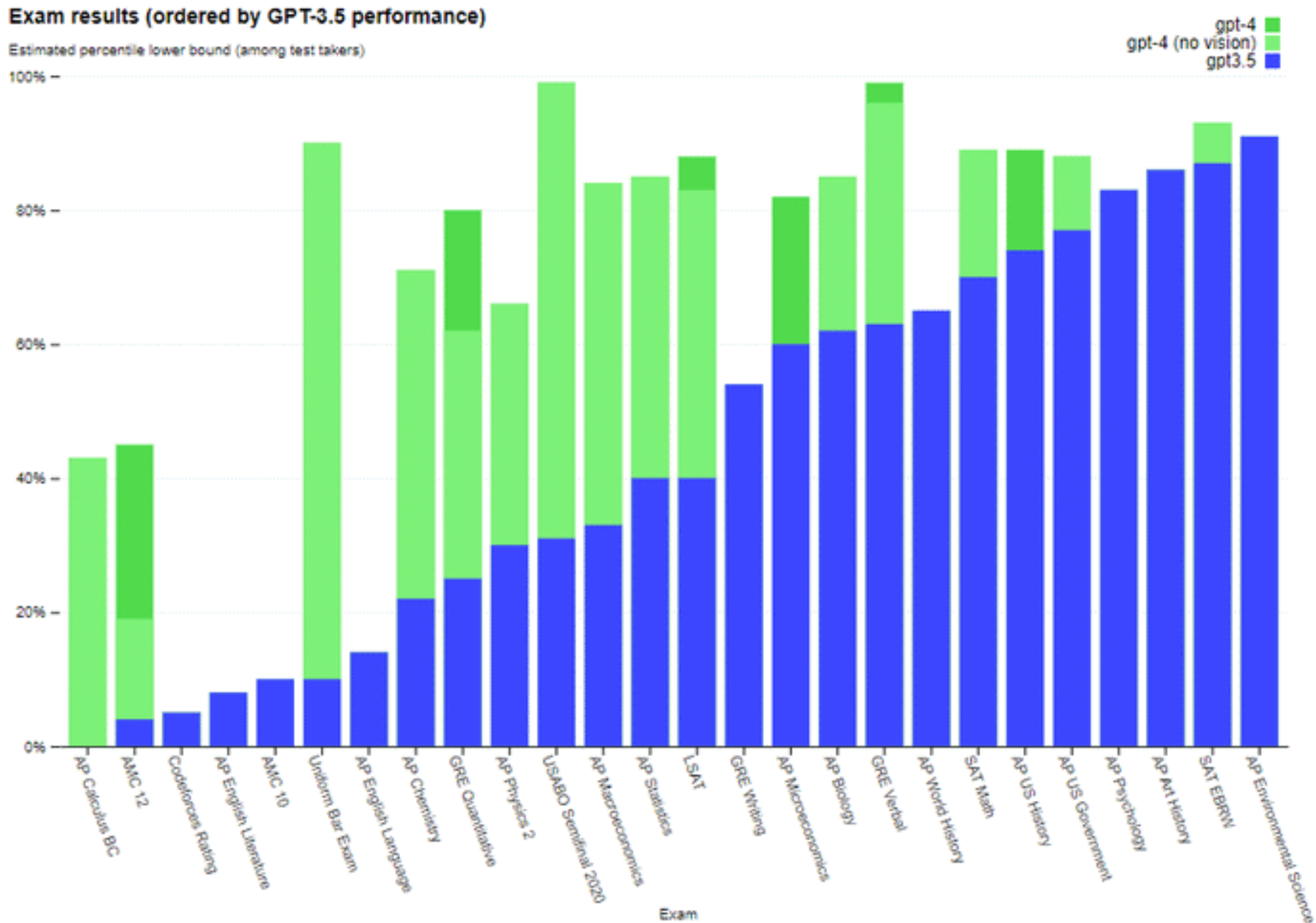
ChatGPT From OpenAI Is A Bot Taking The Tech World By Storm







Source: Chartr


Exam results (ordered by GPT-3.5 performance)
Estimated percentile lower bound (among test takers)
gpt-4
gpt-4 (no vision)
gpt3.5
100%
80%
60%
40%
20%
0%
AP Calculus BC
AMC 12
Codeforces Rating
AP English Literature
AMC 10
Uniform Bar Exam
AP English Language
AP Chemistry
GRE Quantitative
AP Physics 2
USABO Semifinal 2020
AP Macroeconomics
AP Statistics
LSAT
GRE Writing
AP Microeconomics
AP Biology
GRE Verbal
AP World History
SAT Math
AP US History
AP US Government
AP Psychology
AP Art History
SAT EBRW
AP Environmental Science
Exam

- Material for some of these slides comes from ideas borrowed from Kevin Knight, Angel Chang, Danqi Chen, Karthik Narasimhan and others on the Slack group for teaching NLP