

Natural Language Processing with Deep Learning

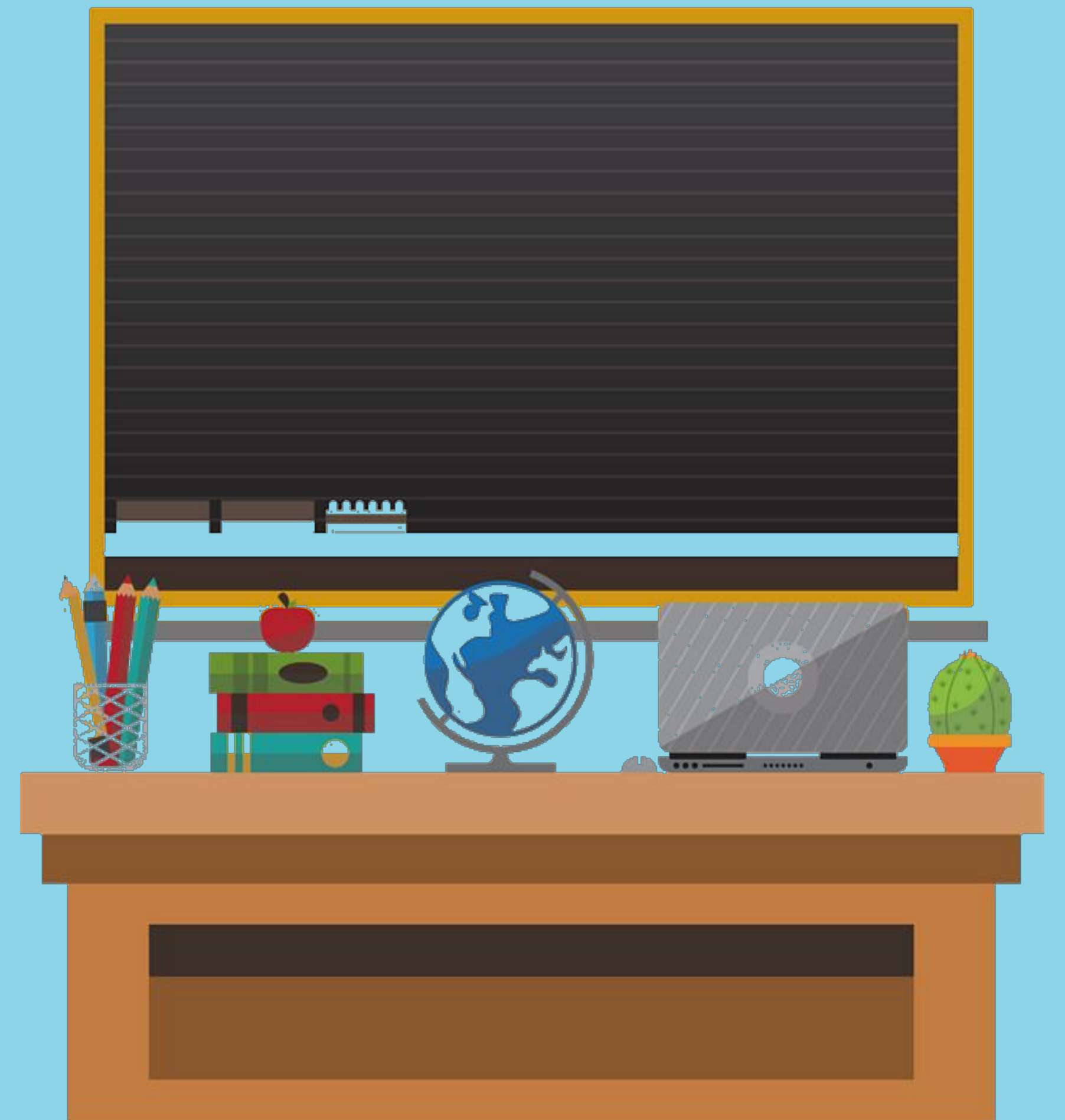
IFT6289, Winter 2022

Lecture 1: Introduction to NLP
Bang Liu

2 Lecture outline

1. Course logistics
2. What is natural language processing?
3. Why NLP is hard?
4. History of NLP
5. NLP techniques
6. NLP resources

Course logistics



4 Course logistics in brief



- **Teacher:** Bang Liu (bang.liu@umontreal.ca)
- **Teaching assistant:** Suyuchen Wang (suyuchen.wang@umontreal.ca)
- **Time:** Tuesdays 9:30am - 11:30am and Fridays 3:30pm - 5:30pm
- **Online office hours (zoom):** Tuesdays 2:00pm - 4:00pm
- **Slides:** uploaded before each lecture

5

Course aims

- **Basics:** deep learning for NLP
- **NLP Techniques:** word embedding, pre-trained language models, machine translation, search engine, dialogue system...
- **Big picture:** NLP techniques and the difficulties in understanding human language
- **Research:** find potential research interests



6 Course work and grading policy

- **Reading** (1% * 10)
10 papers or book chapters
- **Assignments** (15% * 3)
3 programming assignments with report.
- **Term project** (45%, 1-3 people)
Project proposal (up to 2 pages): 5%
Midway report (up to 4 pages): 5%
Final presentation: 5%
Final report (up to 8 pages): 30%
- **Participation** (2%)



7 Course logistics in brief

● Late policy:

- A **late day** extends the deadline 24 hours. For ALL assignments, submissions **after 2 late days** (48 hours) of the deadline **won't be accepted**.
- For programming assignments, we deduct **2%** for each late day. We don't count hours, e.g., if you submit an assignment after 25 hours, it will be considered as 2 late days and will be deducted 4%.
- For project proposal, midway report, we deduct **1%** for each late day.
- For project final report, we deduct **3%** for each late day.
- No late day for the final project presentation and reading assignments.

Course logistics in brief

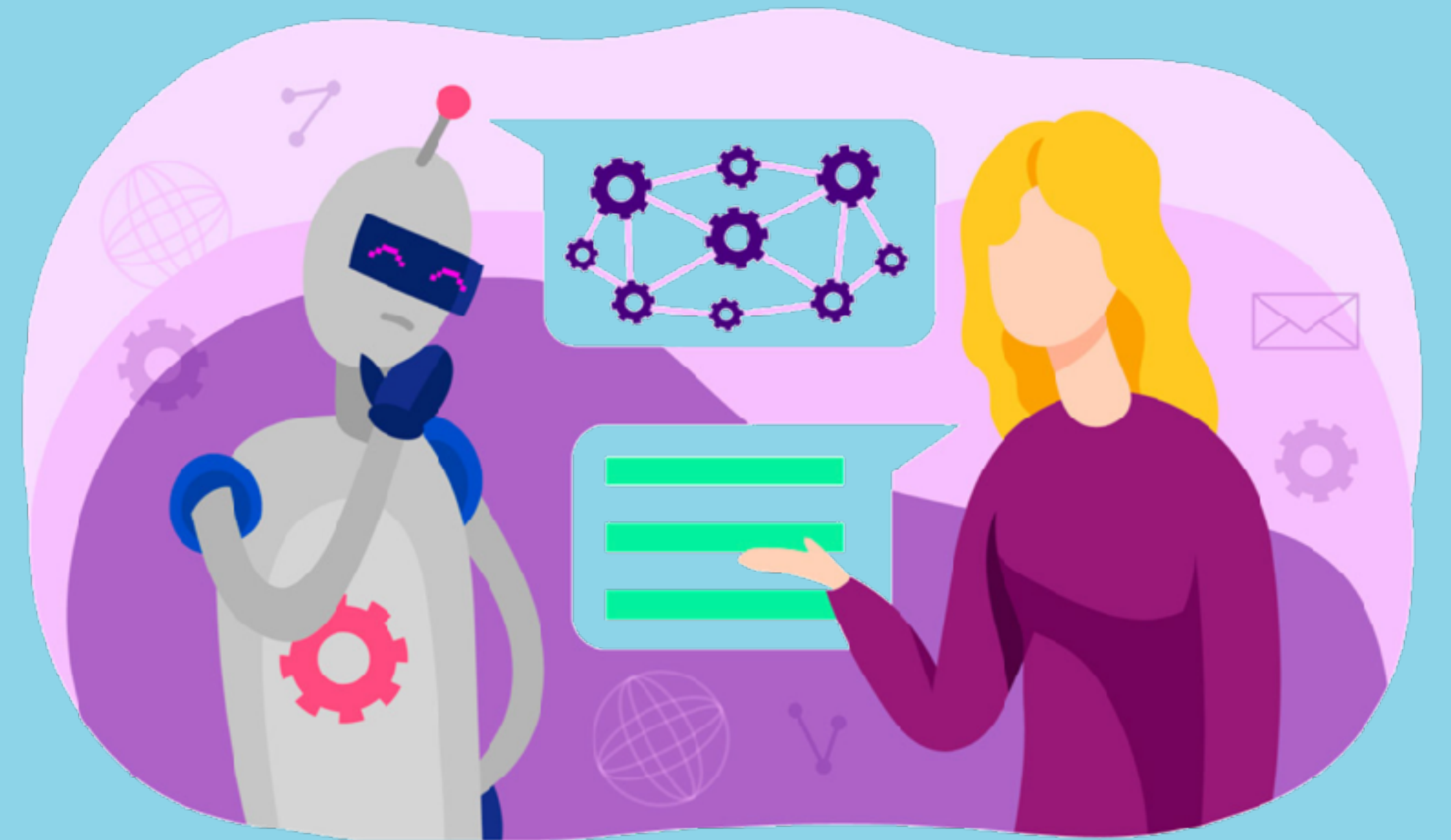
● Computing Resources

- Google Colab, Google Cloud, etc.
- Experiments may take up to hours
- Recommendation: start assignments early

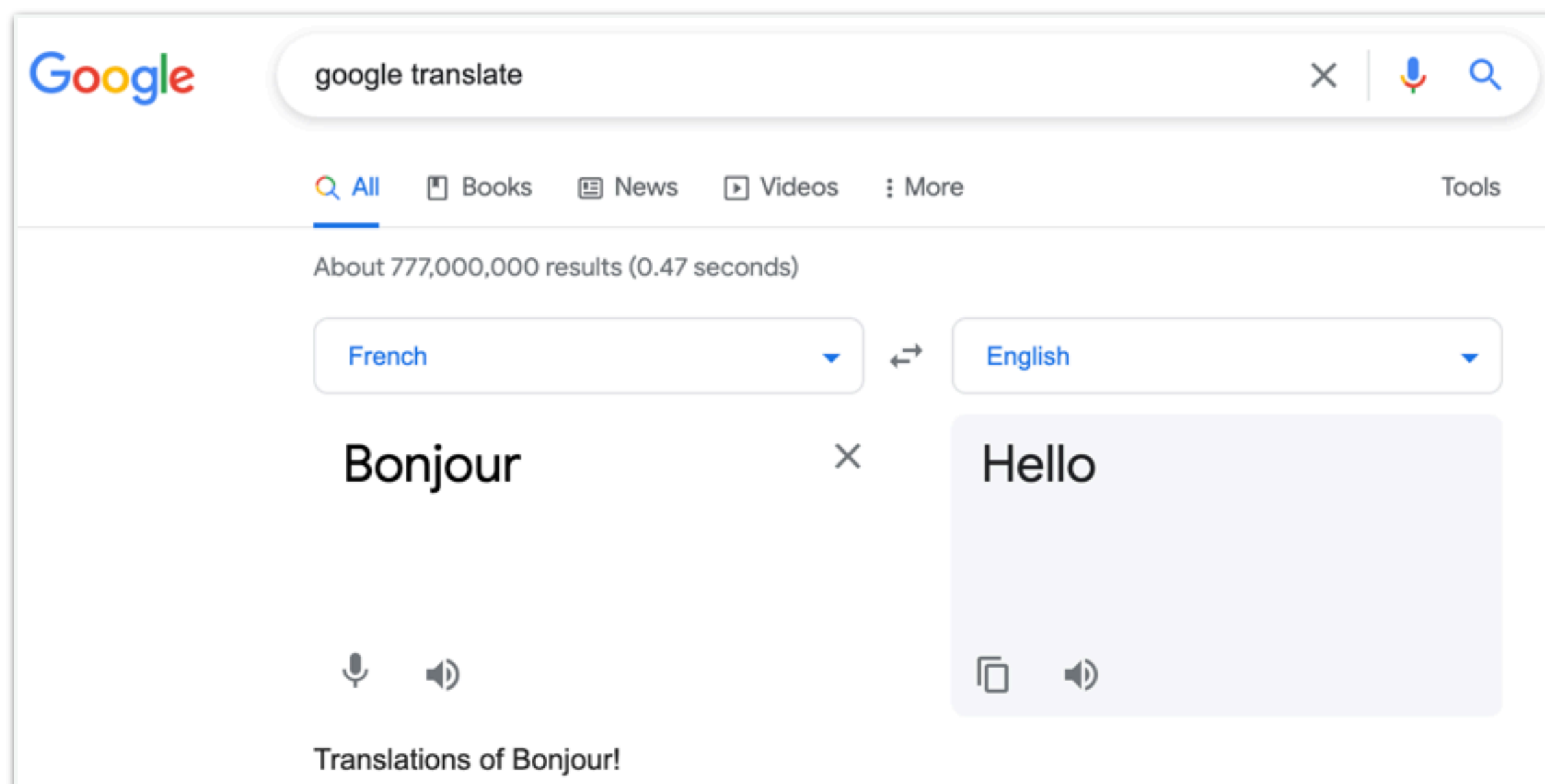
● More information

- Class webpage: <http://www-labs.iro.umontreal.ca/~liubang/IFT%206289%20-%20Winter%202022.htm>

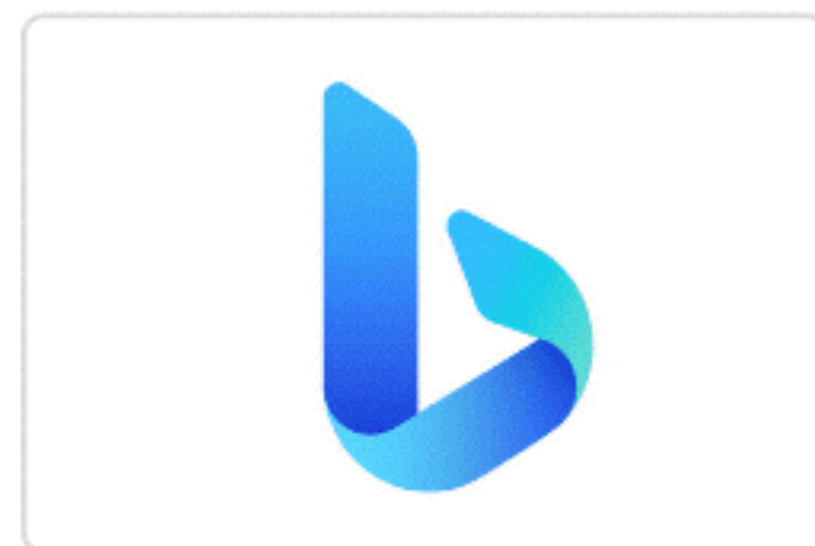
What is natural language processing?



10 NLP is everywhere: machine translation



11 NLP is everywhere: search engines




12 NLP is everywhere: smart speakers



13 NLP is everywhere: chatbots

百度发布PLATO-XL，全球首个百亿参数中英文对话预训练生成模型

DataFunTalk 2 days ago



百度PLATO
百度AI对话机器人度语灵，使用了百度业界领先的自然语言处理技术，给大家更好的智能对话体验

Official Account

导读：和 AI 进行无障碍的对话，是什么样的体验？你或许能够在这篇文章里找到答案！

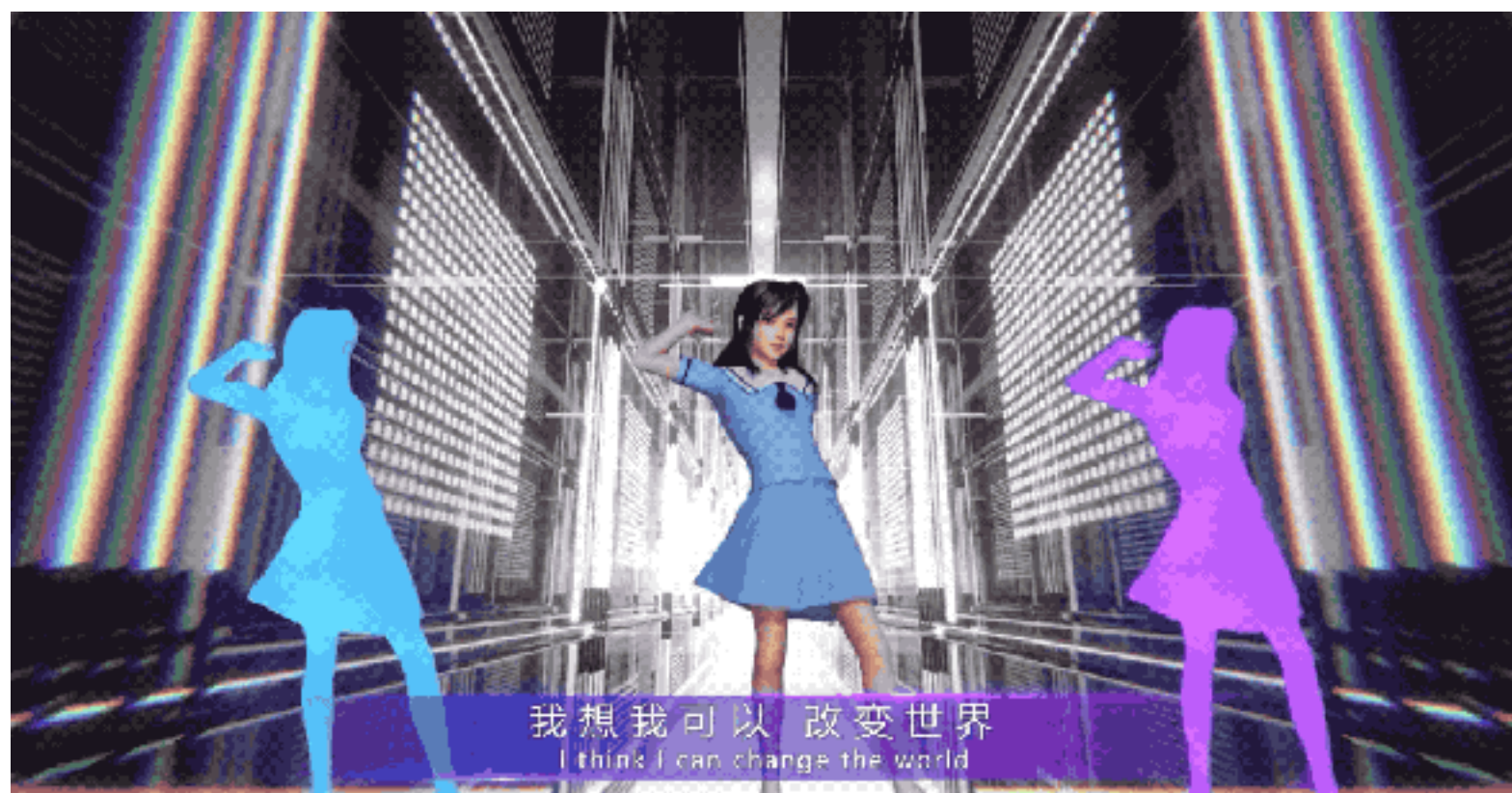
近日，百度全新发布 PLATO-XL，参数达到了 110 亿，超过之前最大的对话模型 Blender，是**当前最大规模的中英文对话模型**，并再次刷新了开放域对话效果。

https://mp.weixin.qq.com/s/jZdpzSgBuMk62_co5laeTQ



Baidu PLATO-XL

14 NLP is everywhere: virtual characters



Microsoft Xiaolce



Xiaomi Xiaoi

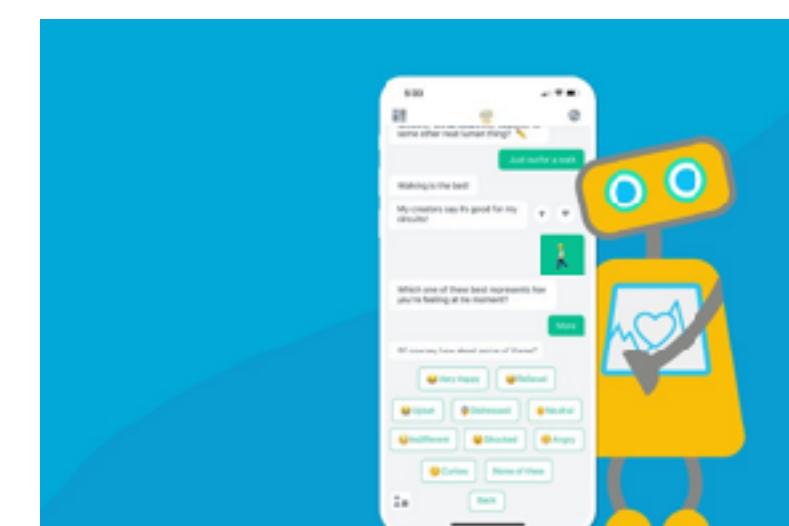
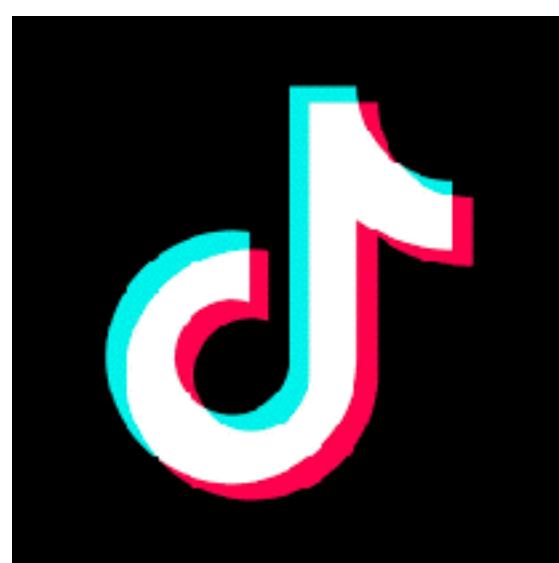


Bilibili 冷鸢yousa



Baidu Xiaodu

15 NLP is everywhere



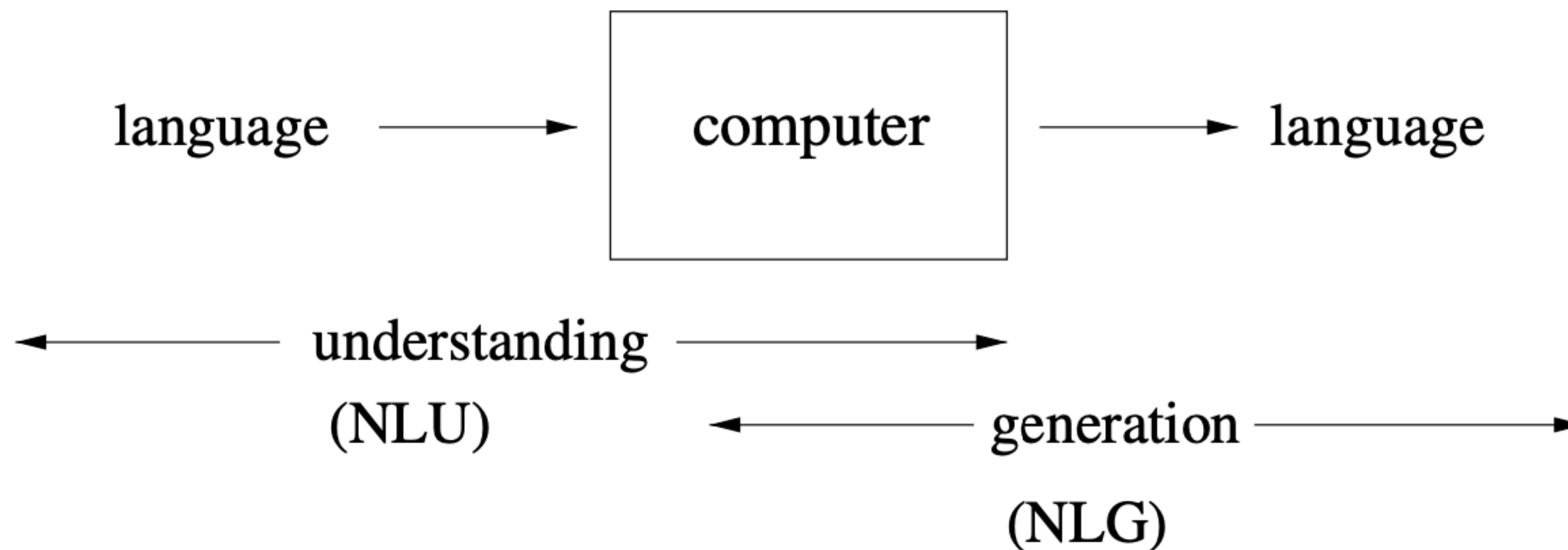
What is natural language processing?

- **Language** is a tool of human communication and a carrier of human thinking.
 - More than 1,900 languages in use all over the world.
 - Structures vary between different languages.
- **Natural language** is a human language (e.g., English, French, Chinese)
 - As opposed to a constructed language (programming languages, language of formal logic, etc.).



17 What is natural language processing?

- **Natural language processing (NLP)** is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language.



NLP tasks

Lexical Analysis

- Word Segmentation
- Word Tokenization
- New Words Identification
- Morphological Analysis
- Part-of-speech Tagging
- Spelling Correction

Sentence Analysis

- Chunking
- Super Tagging
- Constituency Parsing
- Dependency Parsing
- Language Modeling
- Language Identification

Semantic Analysis

- Word Sense Disambiguation
- Semantic Role Labeling
- Abstract Meaning Representation Parsing
- Frame Semantic Parsing
- First Order Predicate Calculus
- Word/Sentence/Paragraph Vector

Information Extraction

- Named Entity Recognition/Disambiguation
- Terminology/Glossary Extraction
- Coreference Resolution
- Relationship Extraction
- Event Extraction
- Sentiment Analysis

High-level Tasks

- Machine Translation
- Text Summarization/Simplification
- Question Answering
- Dialogue System
- Reading Comprehension
- Automatic Essay Grading

[More tasks can be found here: https://nlpprogress.com/](https://nlpprogress.com/)

The Era of Artificial Intelligence

Network

Manufacturing

Health

Biology

Finance

Retail

HCI

IoT

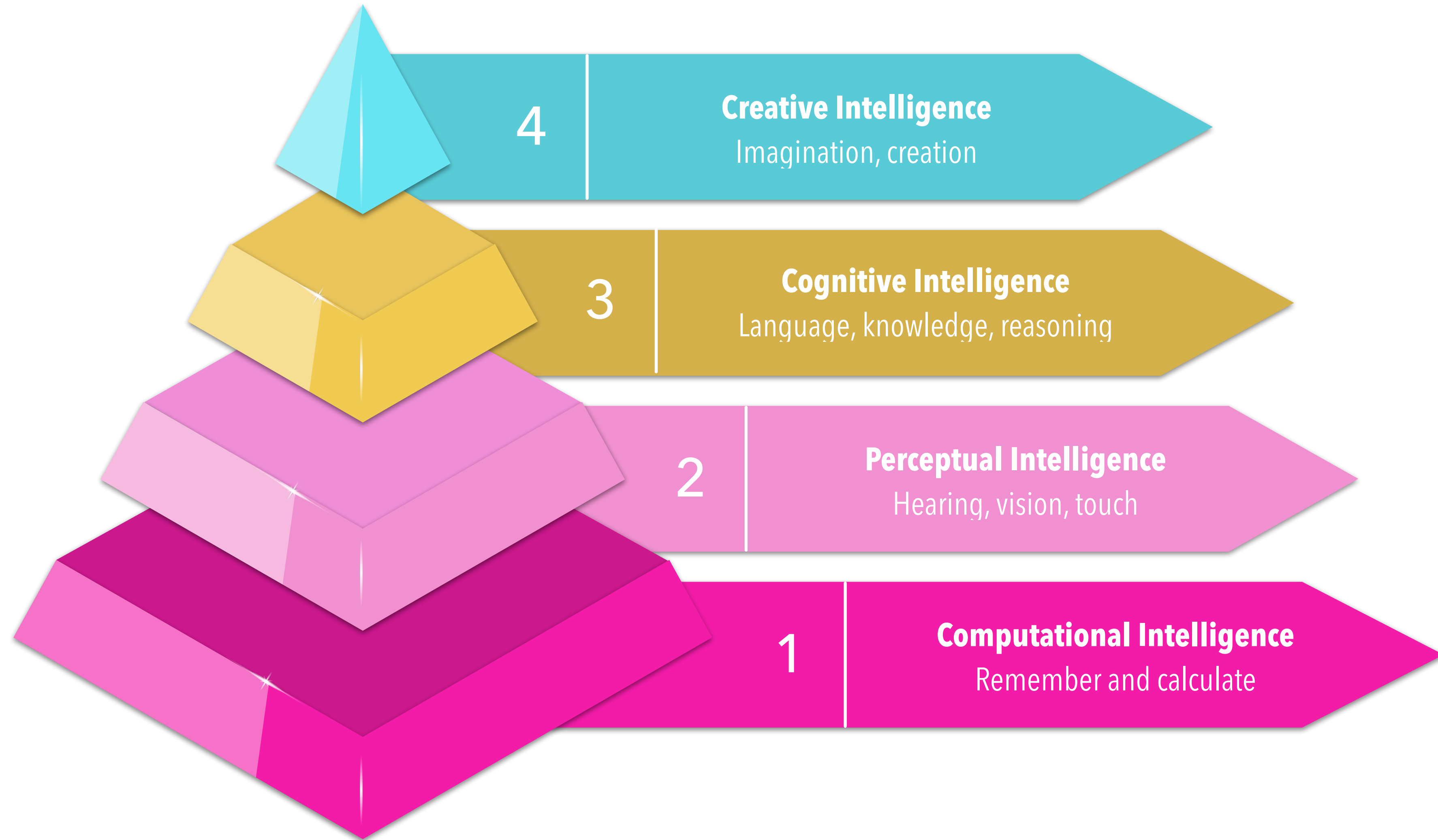
Assistant

Transport

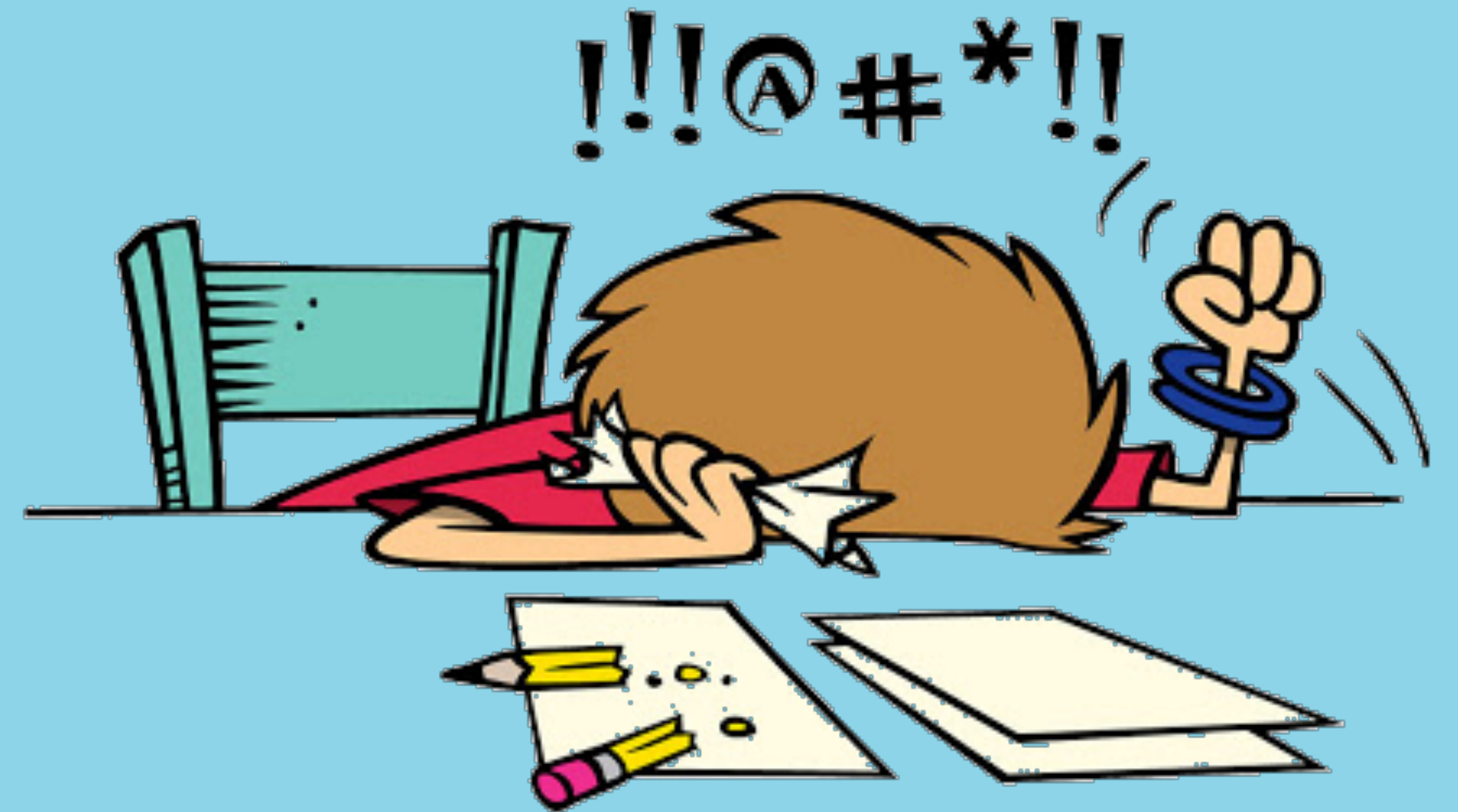
Smart City

Customer Service

20 Levels of AI

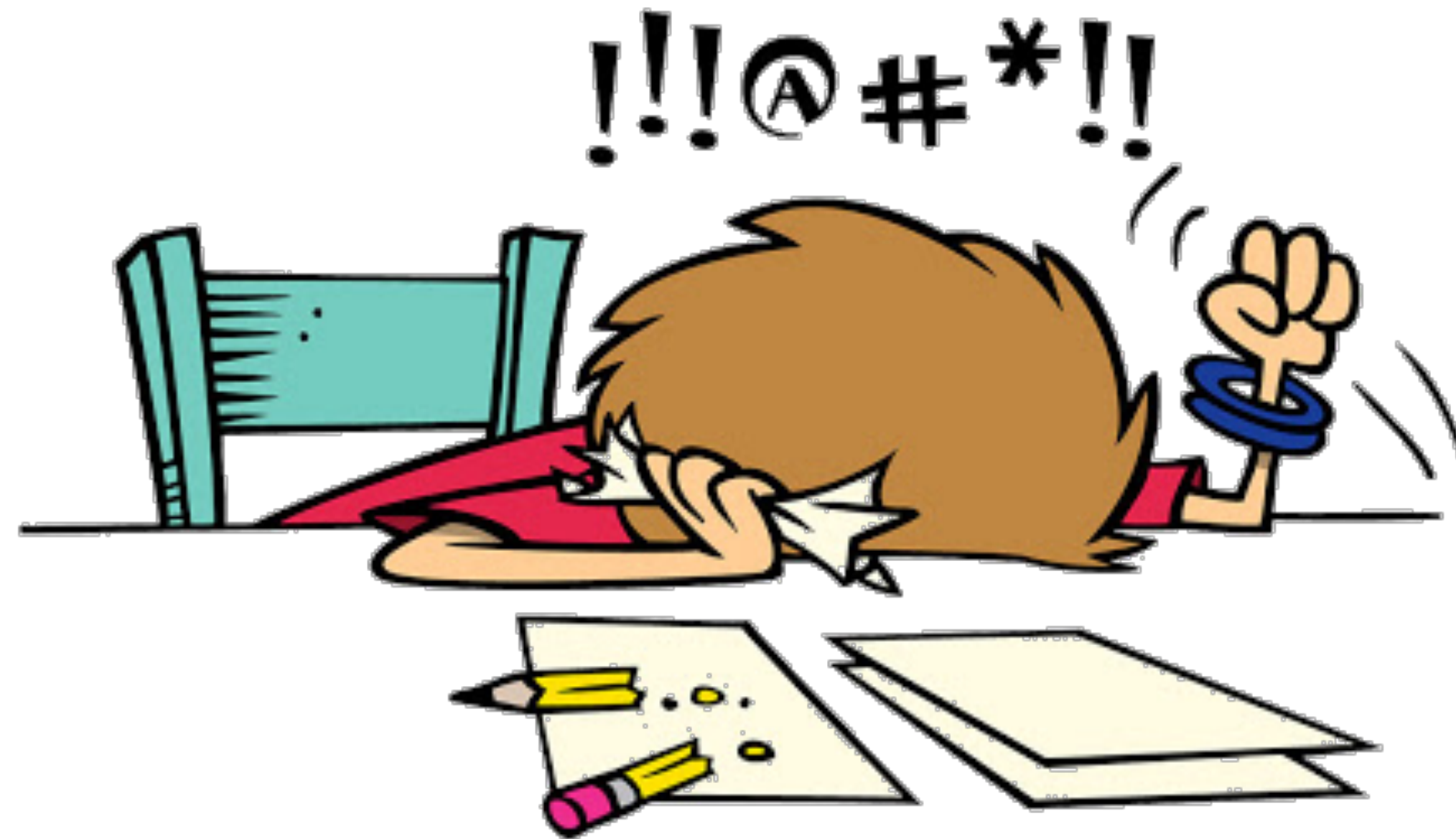


Why NLP is hard?



NLP is challenging

- **Ambiguity**
- **Knowledge**
- **Scale**
- **Sparsity**
- **Variation**
- **Expressivity**
- **Unknown representations**



23 NLP is challenging: ambiguity

“At last, a computer that
understands you like your mother”

How shall we interpret this sentence?

24 NLP is challenging: ambiguity

“At last, a computer that understands you like your mother”

- ⦿ It understands you as well as your mother understands you
- ⦿ It understands (that) you like your mother
- ⦿ It understands you as well as it understands your mother

25 NLP is challenging: ambiguity at many levels

“At last, a computer that
understands you like your mother”

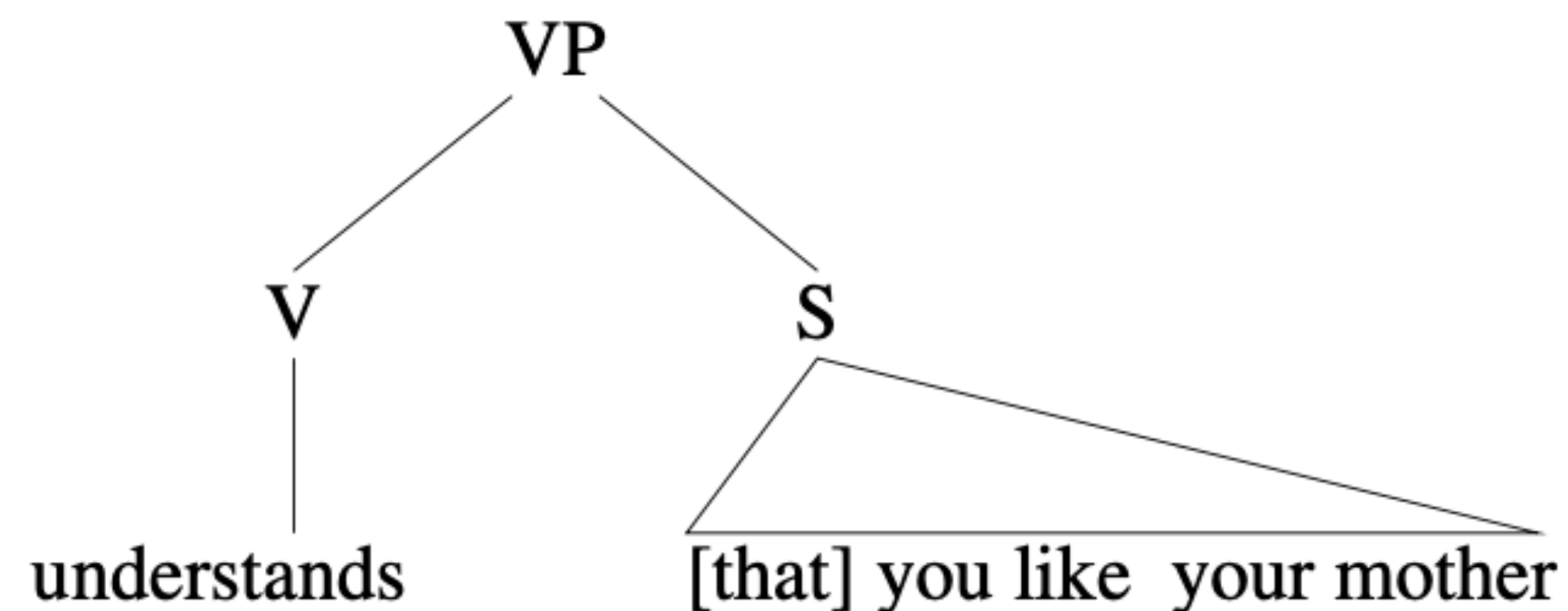
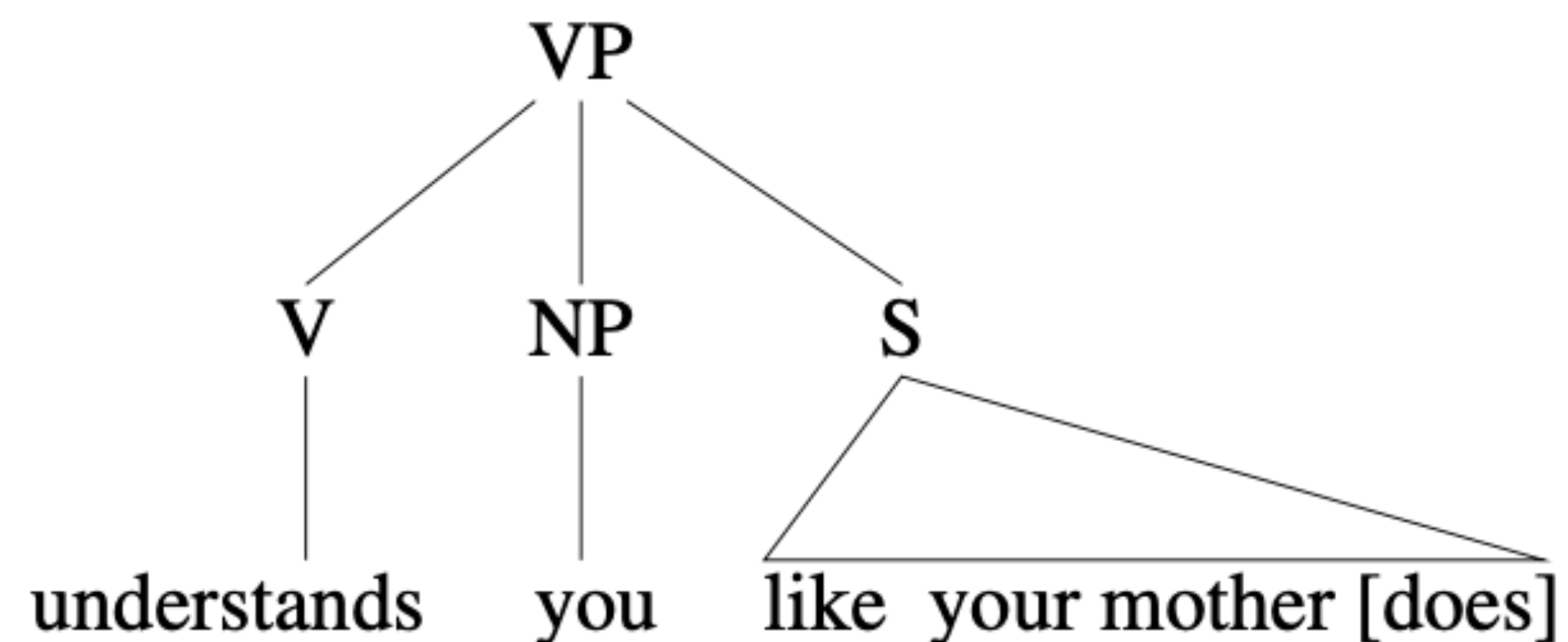
At the **acoustic** level (speech recognition):

1. “... a computer that understands you **like** your mother”
2. “... a computer that understands you **lie cured** your mother”

26 NLP is challenging: ambiguity at many levels

“At last, a computer that
understands you like your mother”

At the **syntactic** level:



Different structures lead to different interpretations.

27 NLP is challenging: ambiguity at many levels

“At last, a computer that understands you like your mother”

At the **semantic** (meaning) level (speech recognition):

Two definitions of “mother”

1. a woman who has given birth to a child
2. a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar

This is an instance of **word sense ambiguity**

28 NLP is challenging: ambiguity at many levels

At the **discourse** (multi-clause) level:

1. “Alice says they’ve built a computer that understands you like your mother,
2. but **she** ...
 - ... doesn’t know any details”
 - ... doesn’t understands me at all”

Here “she” co-referees to some other discourse entity

Dealing with ambiguity

How can we model ambiguity?

- ⦿ No-probabilistic methods (CKY parsers for syntax) return **all possible analyses**
- ⦿ Probabilistic and neural networks return the **most probable one**

But the “best” analysis is only good if our probabilities are accurate. Where do they come from?

30 Corpora

A corpus is a collection of text

- Often annotated in some way
- Sometimes just lots of text

Examples:

- Penn Treebank: 1M words of parsed WSJ
- The Web!

Relevant statistics/probabilities/
parameters are learned from data



31 NLP is challenging: knowledge

We need knowledge:

- Knowledge about language
- Knowledge about the world
- Knowledge about common sense



I dropped the glass on the floor and **it**
broke

I dropped the hammer on the glass
and **it** broke



Knowledge bottleneck in NLP

Possible solutions:

- Symbolic approach: encode all the required information into computer (knowledge graph)
- Statistical approach: infer language properties from language samples
- Deep learning: neural networks

Case study: GPT-3

Q: What is your favorite animal?

A: My favorite animal is a dog.

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

More details about GPT-3 in future lectures

Case study: GPT-3

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many eyes does my foot have?

A: Your foot has two eyes.

Q: How many eyes does a spider have?

A: A spider has eight eyes.

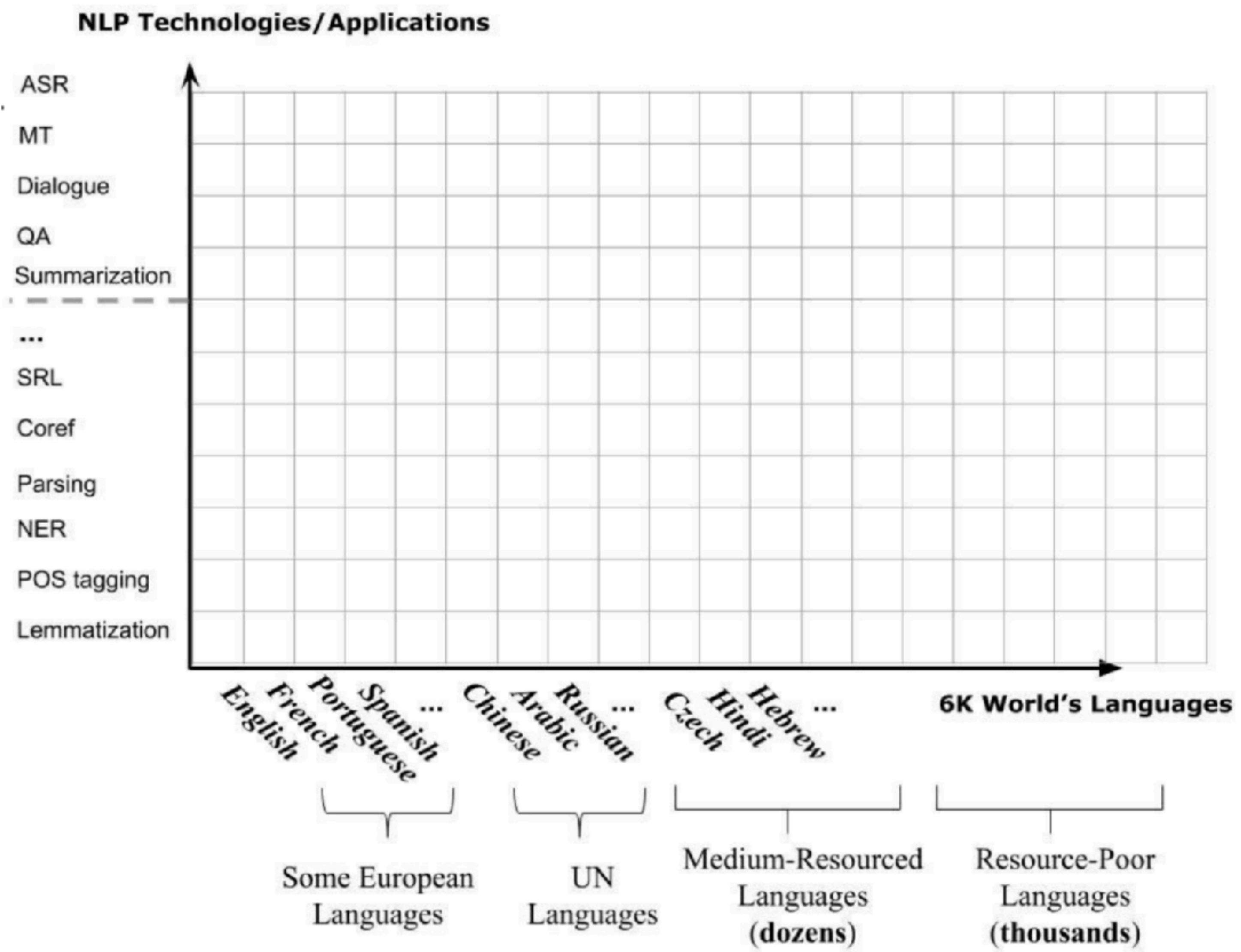
Q: How many eyes does the sun have?

A: The sun has one eye.

Q: How many eyes does a blade of grass have?

A: A blade of grass has one eye.

NLP is challenging: scale



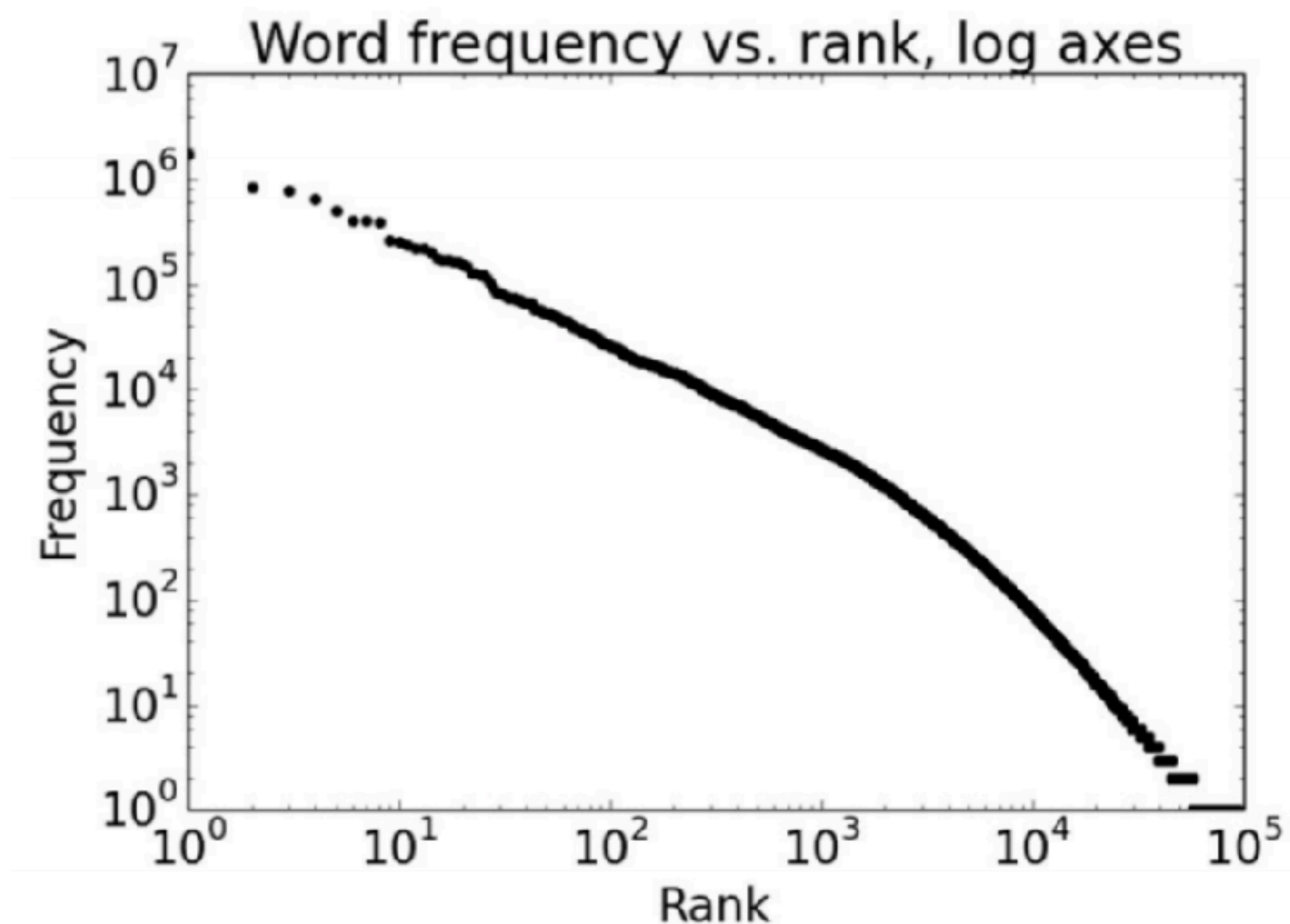
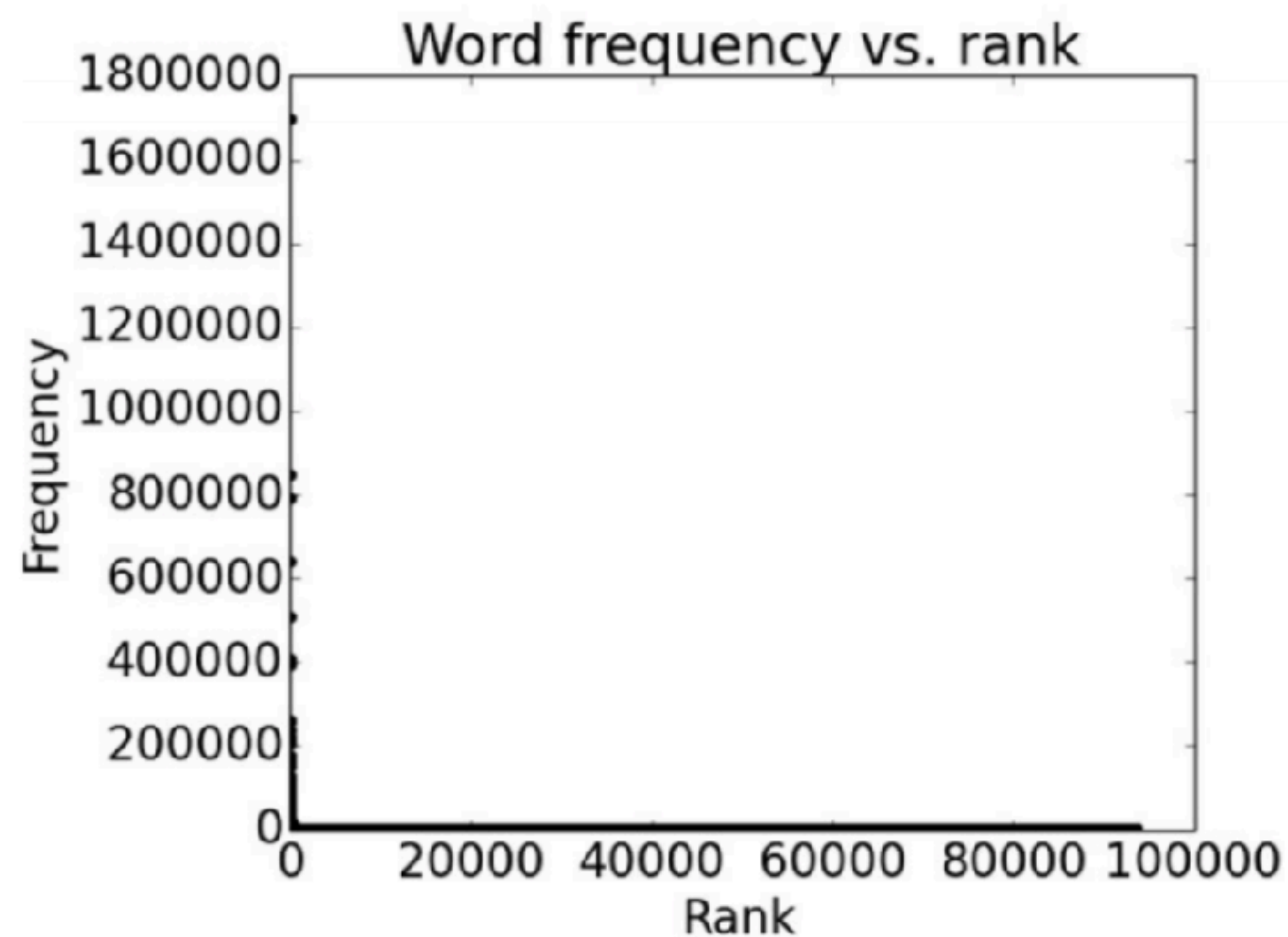
36 Sparsity

- Sparse data due to **Zipf's law**
- Example: the frequency of different words in a large text corpus

any word		nouns	
Frequency	Token	Frequency	Token
1,698,599	the	124,598	European
849,256	of	104,325	Mr
793,731	to	92,195	Commission
640,257	and	66,781	President
508,560	in	62,867	Parliament
407,638	that	57,804	Union
400,467	is	53,683	report
394,778	a	53,547	Council
263,040	I	45,842	States

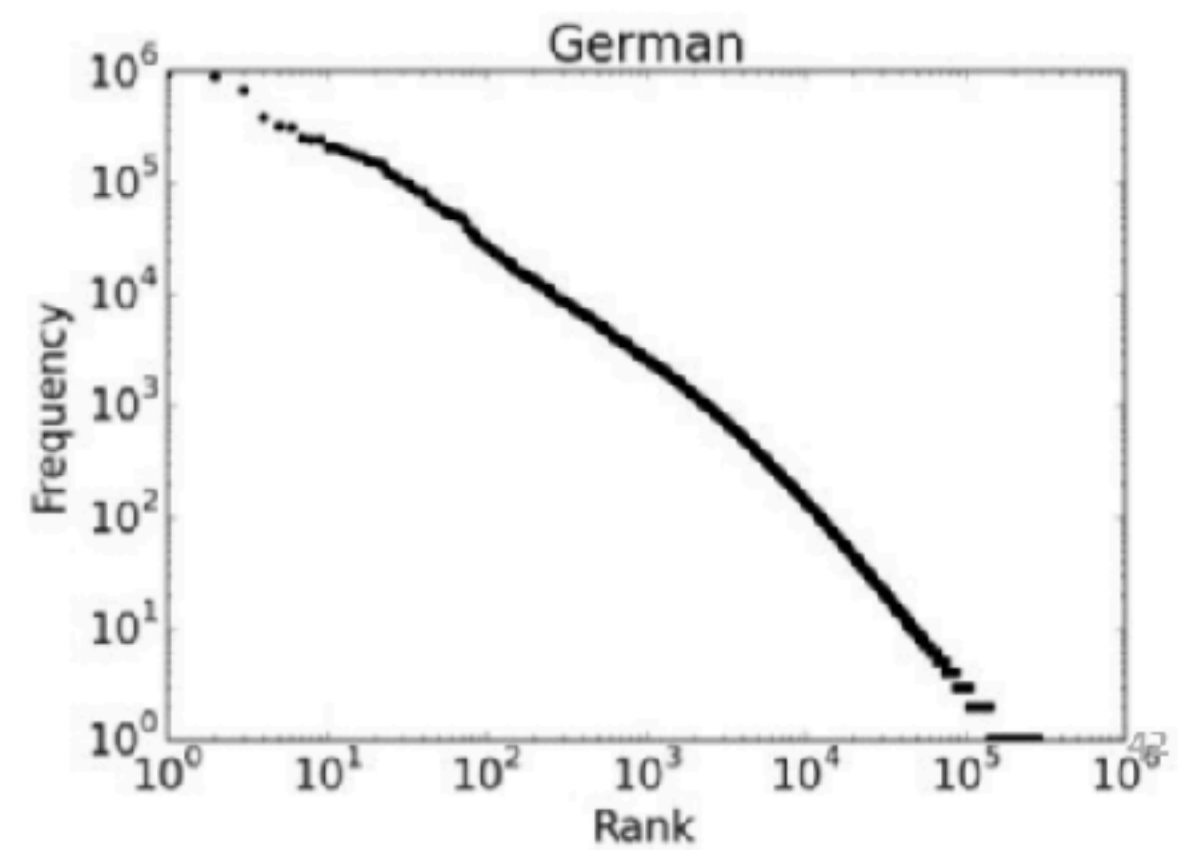
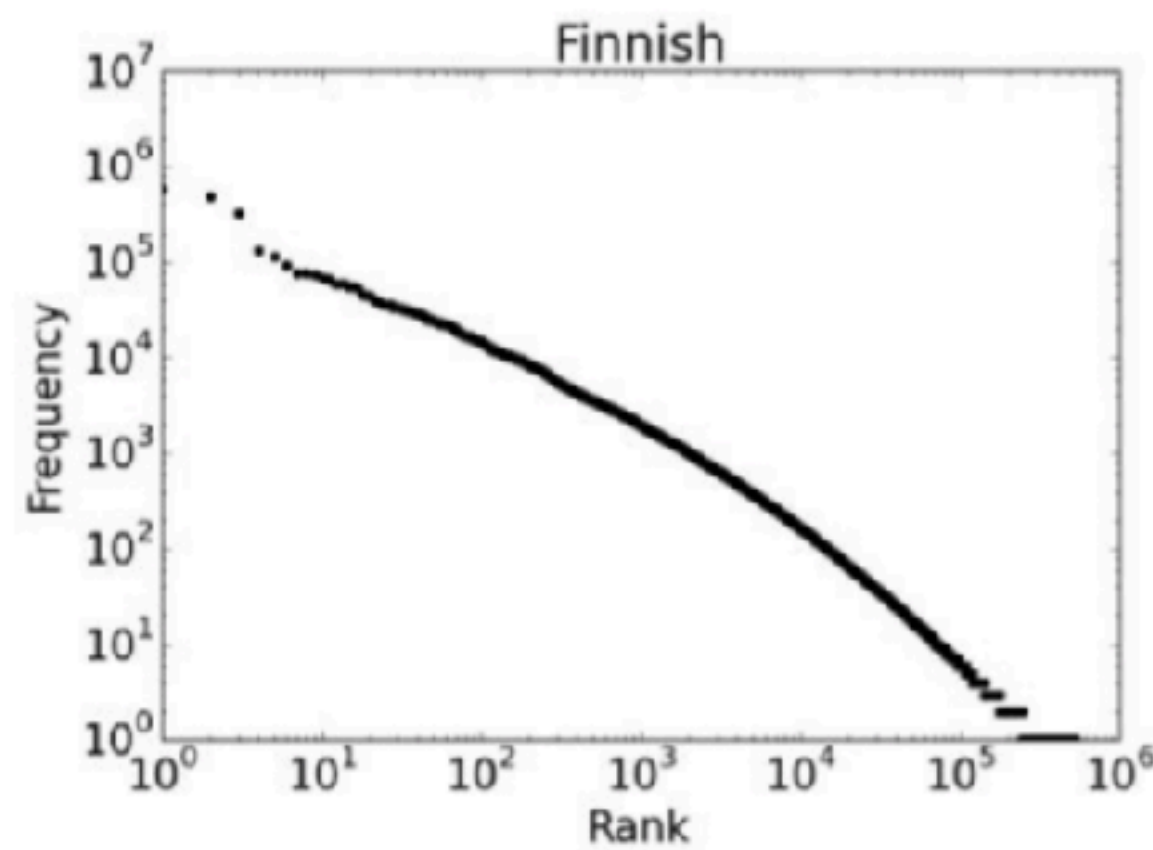
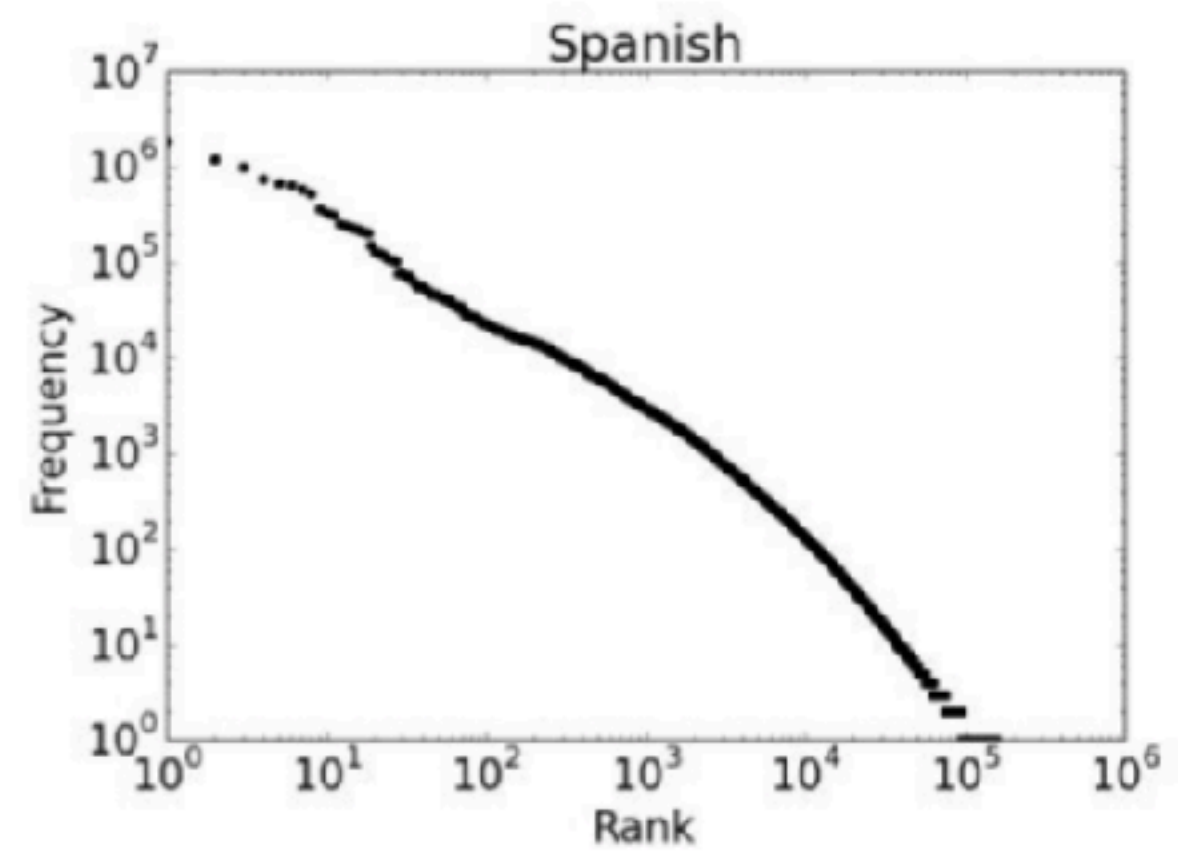
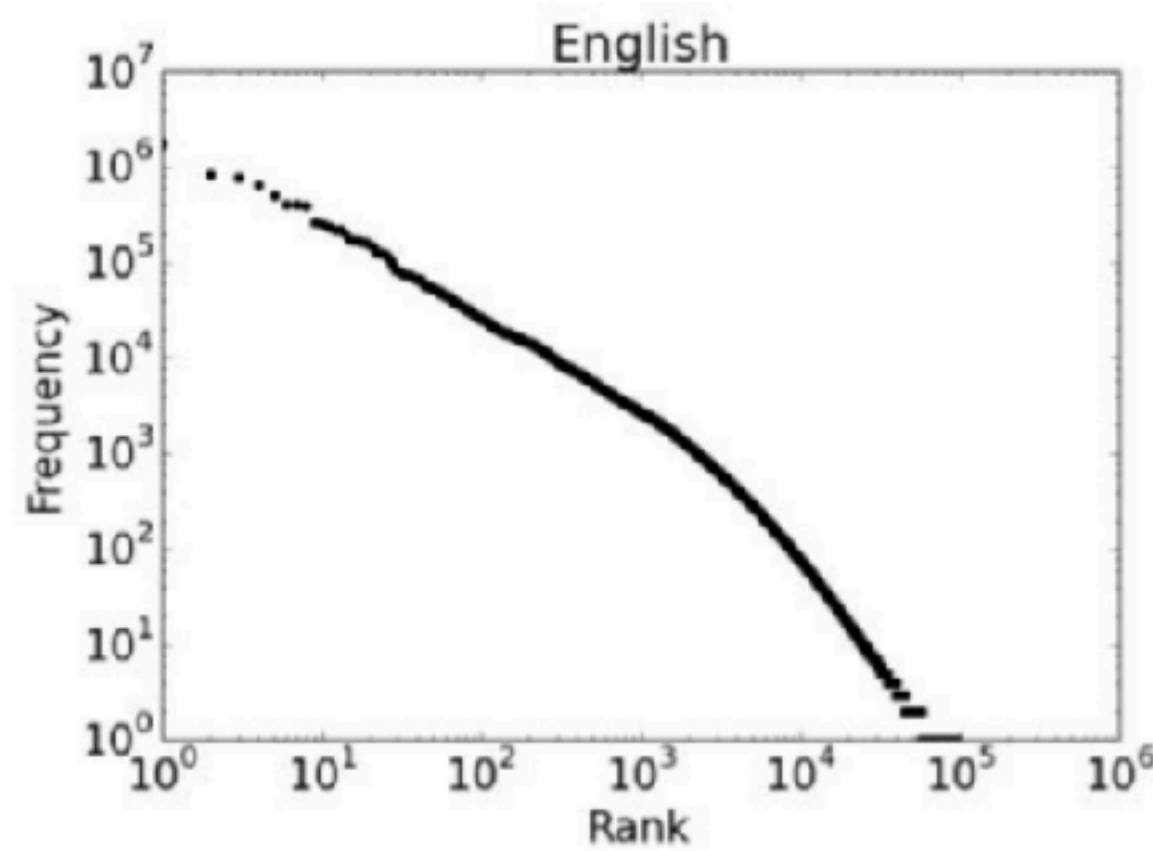
37 NLP is challenging: sparsity

Order words by frequency. What is the frequency of n-th ranked word?



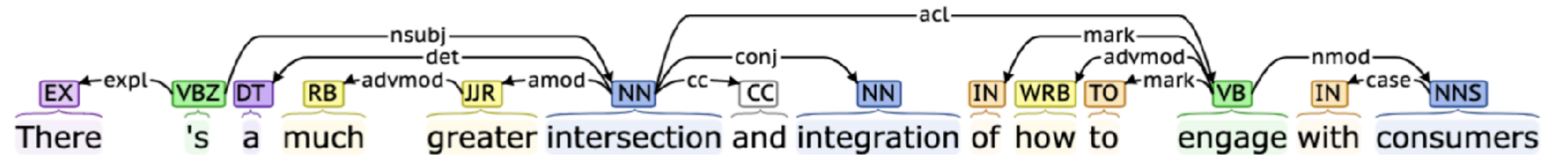
NLP is challenging: sparsity

- Regardless of how large our corpus is, there will be a lot of infrequent words.
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen.



39 NLP is challenging: variation

Suppose we train a part of speech tagger or a parser on the **Wall Street Journal**



What will happen if we try to use this tagger/parser for **social media**?

I know, right shake my head

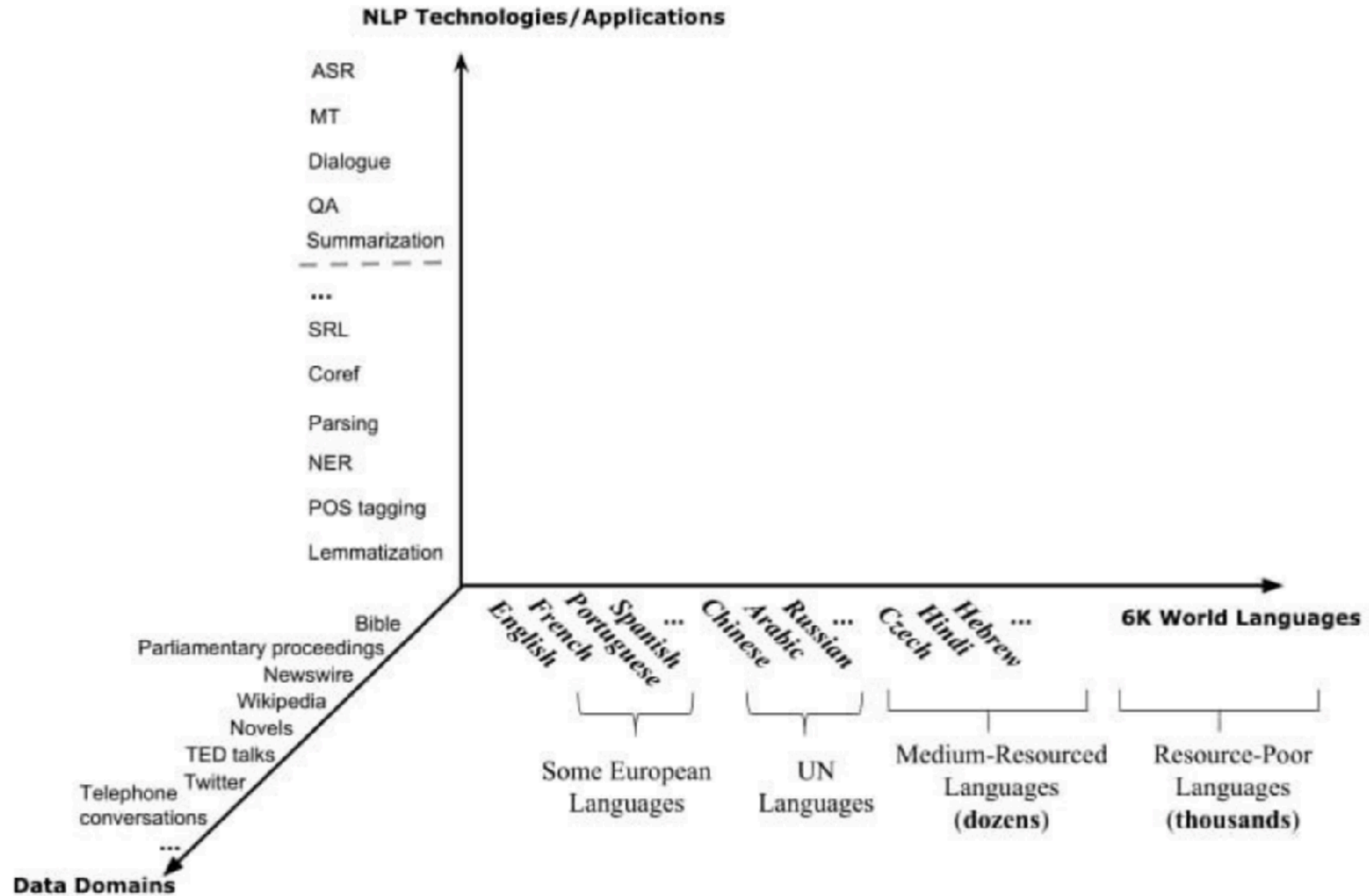
for your

ikr smh he asked fir yo last name so he can add u

on fb lololol

you Facebook laugh out loud

40 NLP is challenging: variation



41 NLP is challenging: expressivity

Not only can one form has different meanings (ambiguity), but the same meaning can be expressed with different forms.

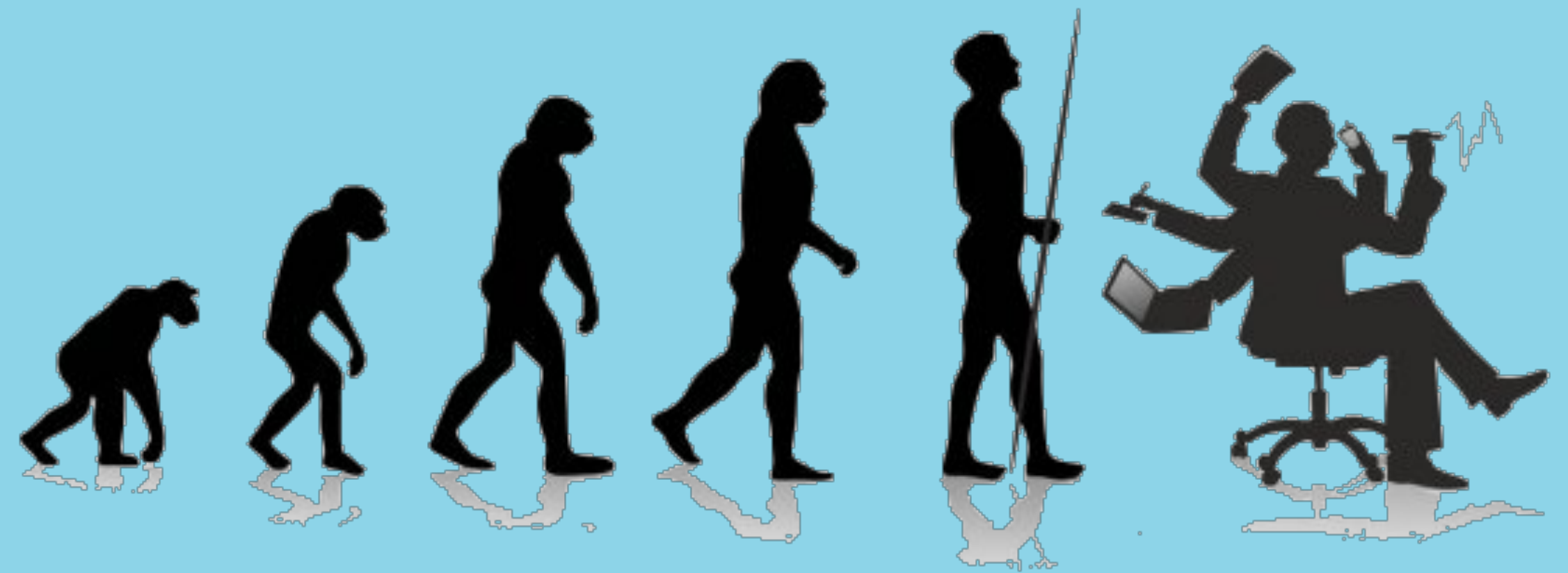


42 NLP is challenging: unmodeled representations

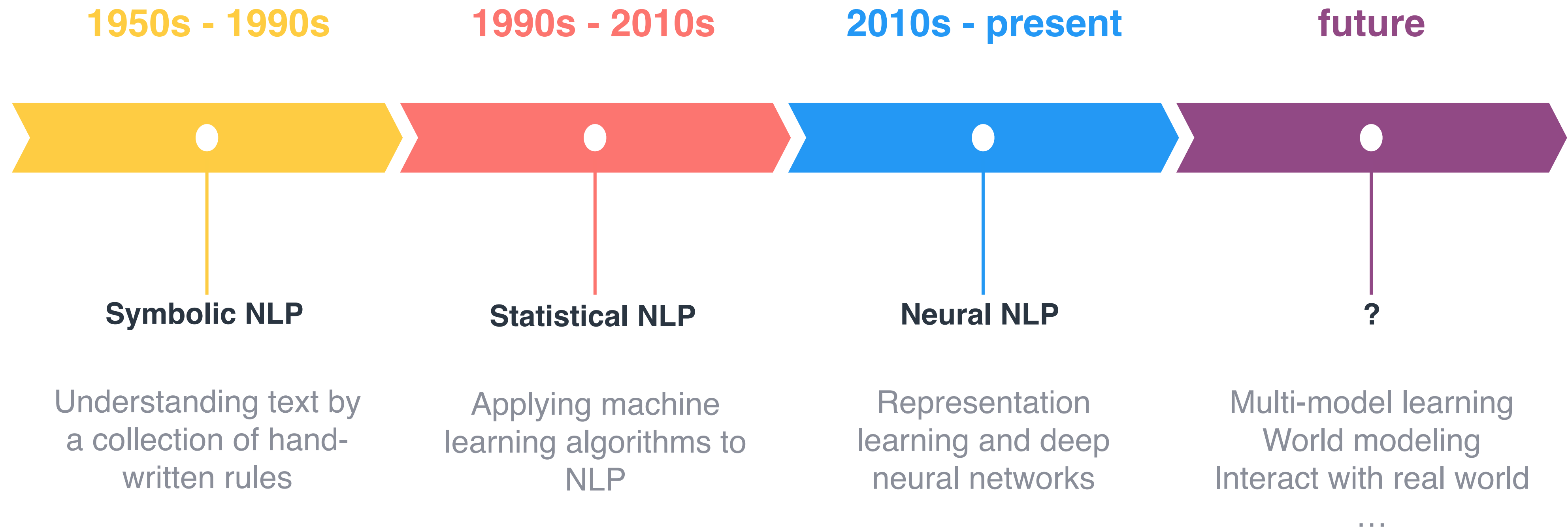
We don't even know how to represent the knowledge a human has/needs:

- ⦿ What is the “meaning” of a word or sentence?
- ⦿ How to model context?
- ⦿ Other general knowledge?

History of NLP



History of NLP



45 The confusion of tongues: Tower of Babel



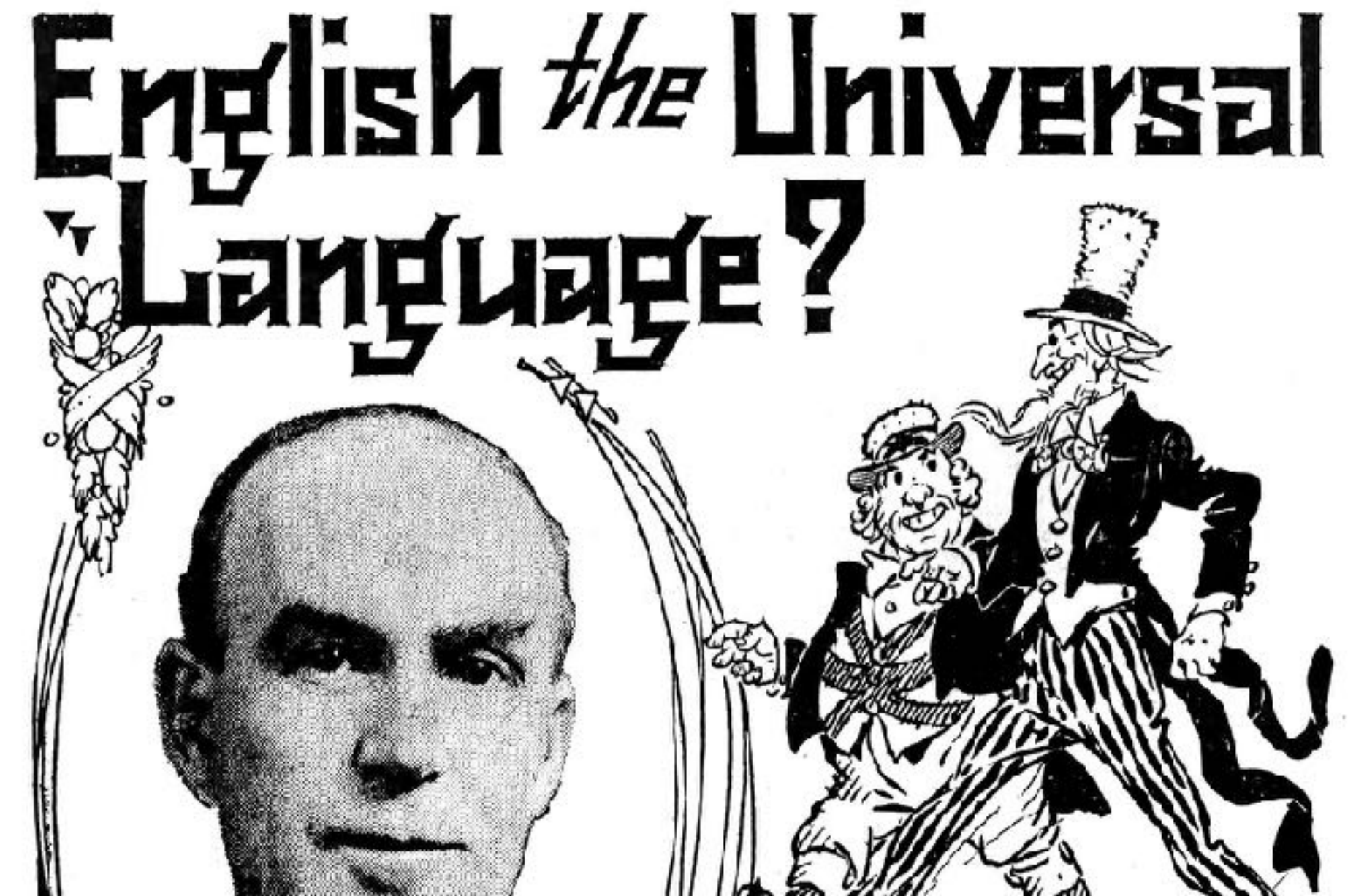
Gustave Doré: The Confusion of Tongues (1865)

The **confusion of tongues** is the origin myth for the fragmentation of human languages described in Genesis.

The Bible "Genesis" says that ancient humans originally spoke a unified language. They once wanted to build a tower called "**Tower of Babel**" as high as the heavens. This shocked God, and God let different people speak different languages so that it becomes impossible for people to coordinate their work. As a result, the Tower of Babel is not built, and the difference in language has become a great obstacle for people to communicate with each other.

Universal language

In the 17th century, some insightful people put forward the idea of using **machine dictionaries** to overcome language barriers. Both Descartes and Leibniz tried to write dictionaries based on a unified digital code. In the mid-17th century, such dictionaries were published by Cave Beck, Athanasius Kircher, and Johann Joachim Becher. As a result, a movement about "**universal language**" was launched. Some people tried to create an unambiguous language based on logical principles and graphic symbols, so that people would no longer have to be confused in communication due to misunderstandings.



47 Language computing have a long history

- In 1847, the Russian mathematician **B. Buljakovski** believed that probabilistic methods could be used to study grammar, etymology, and language history comparison.
- In 1851, the British mathematician **A. De Morgen** used word length as a feature of writing style for statistical research.
- In 1894, Swiss linguist **De Saussure** pointed out that in terms of basic properties, the relationship between quantity and quality in language can be expressed regularly by mathematical formulas. He published "General Linguistics Course" in 1916. It also pointed out that language is like a geometric system, which can be boiled down to some unproven theorems.

Language computing have a long history

- In 1898, German scholar **F.W. Kaeding** calculated the frequency of German vocabulary in the text and compiled the world's first frequency dictionary "German Frequency Dictionary".
- In 1935, Canadian scholar **E. Varder Beke** proposed the concept of word distribution rate and used it as the main criterion for dictionary selection.
- In 1944, the British mathematician **G.U. Yule** published the book "Statistical Analysis of Literary Words", using probabilistic and statistical methods to study vocabulary.

Four important milestones to NLP



Andrey Markov

Research on Markov model



Alan Turing

Research on model of
computation



Claude Shannon

Research on probability and
information theory



Noam Chomsky

Research on formal language
theory

Andrey Markov: Markov model

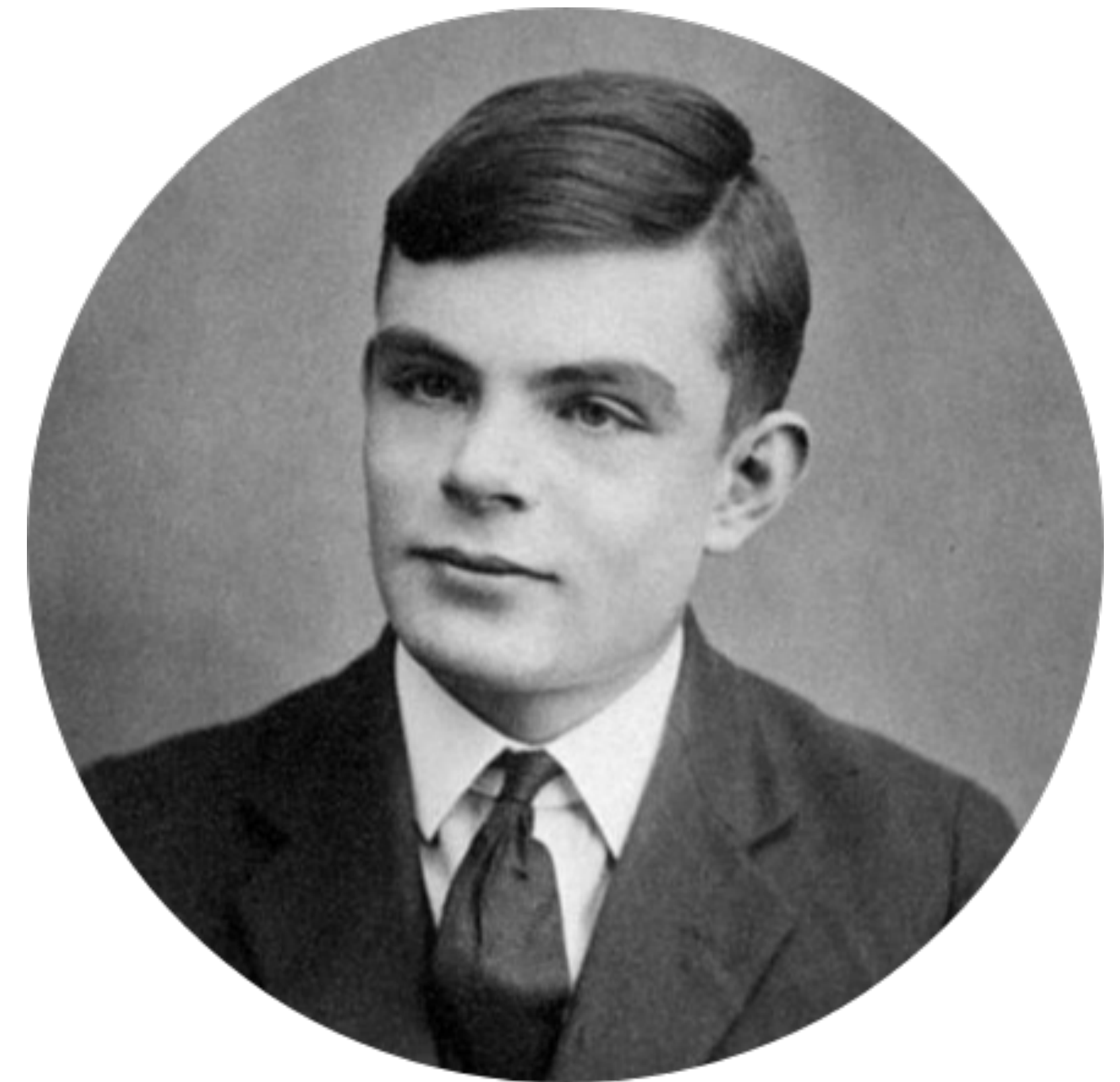
In 1913, the famous Russian mathematician **Andrey Markov** noticed the mutual influence between the appearance probability of language signs in the narrative poem "Eugen Onegin" by the Russian poet Pushkin. He tried to use the appearance probability of language signs as an example to study the random process mathematical theory and put forward the idea of **Markov Chain**. His pioneering result later became the Markov model widely used in computational linguistics. It is one of the most important theoretical pillars in computational linguistics.



Andrey Markov

Alan Turing: Turing Machine

In 1936, **Turing** proposed a mathematical model of the famous "Turing Machine" in the seminal paper "On Computable Numbers, with an Application to the Entscheidungsproblem" (German for "decision problem"). The "**Turing machine**" is not a specific machine, but an abstract mathematical model that can create a very simple but powerful computing device to calculate all imaginable computable functions. This research became the basis of modern computer science. Turing also proposed the concept of **Turing test**.



Alan Turing

Claude Shannon: Information Theory

In 1948, American scholar **Shannon** used the probabilistic model of discrete Markov process to describe language automata.

Shannon's another contribution is the creation of "**Information Theory**". He likened the act of transmitting language through media such as communication channels or acoustic speech to "noisy channel" or "decoding".

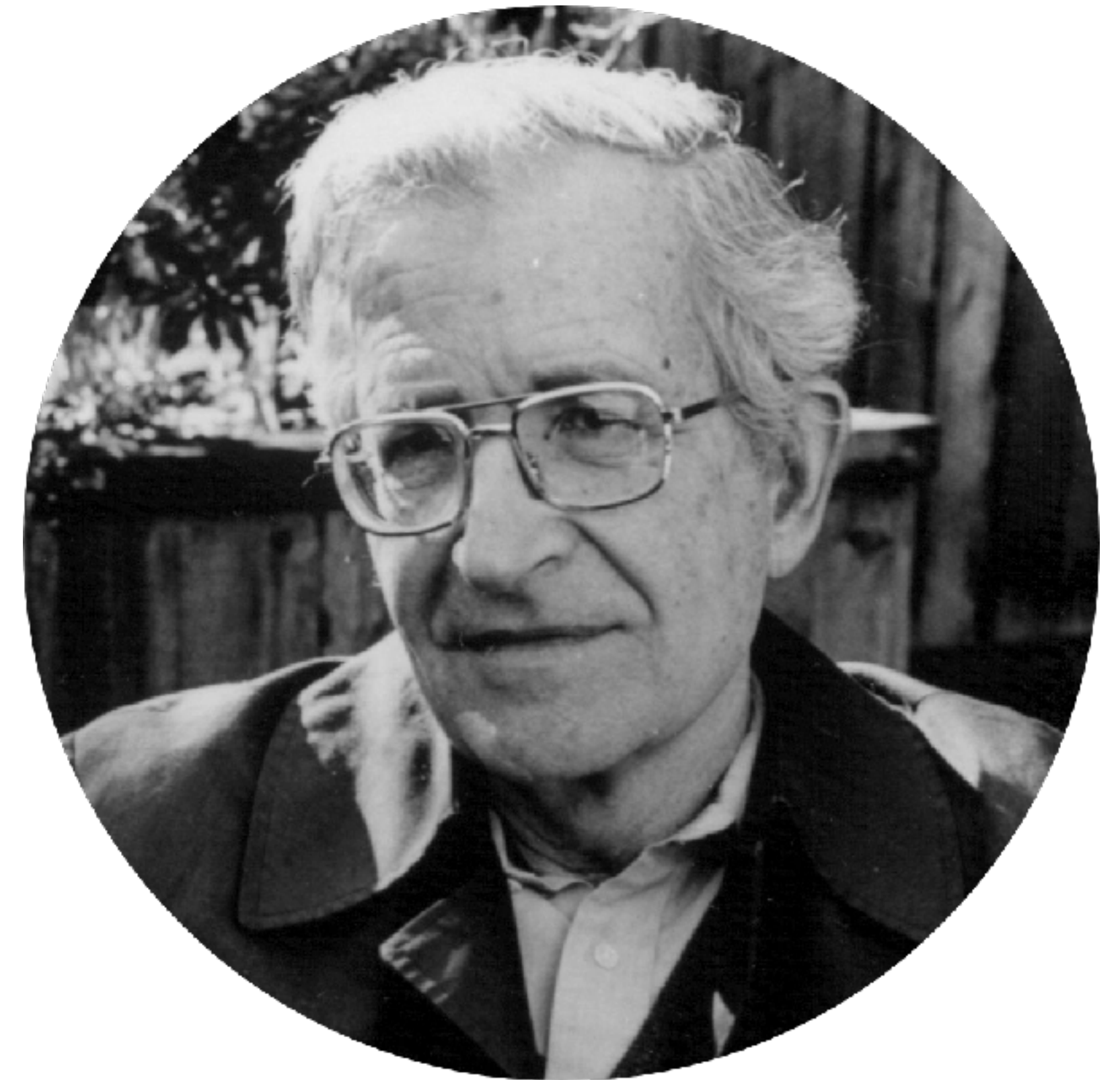
Shannon also borrowed the term "**entropy**" from thermodynamics as a method of measuring the information capacity of a channel or the amount of information in a language, and he used probabilistic techniques to measure the entropy of English for the first time.



Claude Shannon

Noam Chomsky: Formal Language Theory

In 1956, American linguist **Noam Chomsky** absorbed the idea of finite state Markov process from Shannon's work, and used finite state automata as a tool to describe the grammar of language. The state language is defined as a language generated by a finite state grammar. Produced "**formal language theory**"
He used algebra and set theory to define formal language as a sequence of symbols. It has become the most important theoretical cornerstone of computer science.



Noam Chomsky

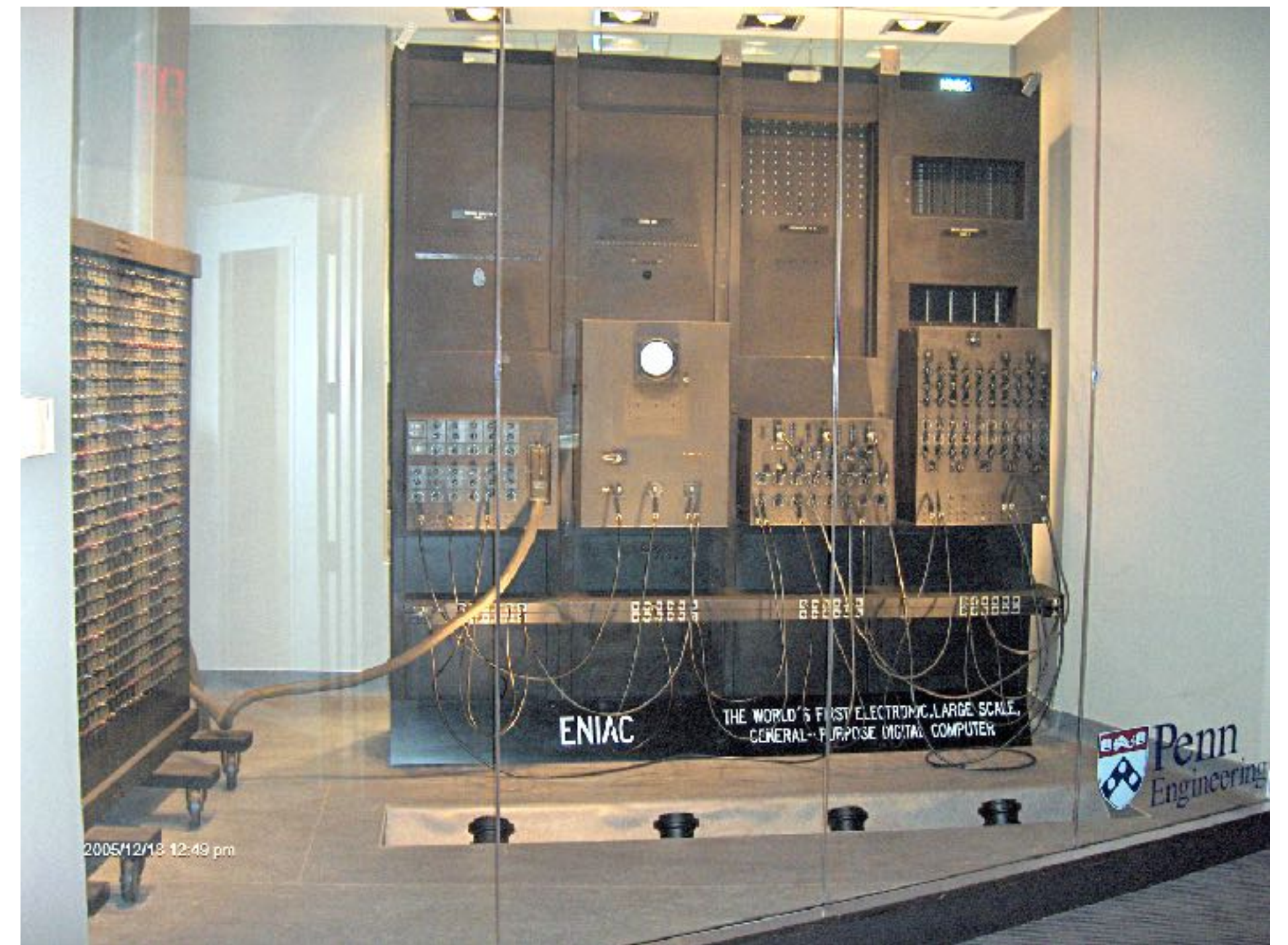
Machine translation: early development

Machine Translation is the key application in the history of computational linguistics.

- In 1933, the Soviet inventor Troyansky (**П.П.ТРОЯНСКИЙ**) **designed a machine** that mechanically **translates** one language into another, and registered his invention on September 5 of the same year.
- In the early 1930s, the Armenian French engineer **GB Artsouni** proposed the idea of **using machines for language translation**, and on July 22, 1933, he obtained a patent for a "translator" called "**Mechanical brain**" (mechanical brain). The prototype of Artsouni was formally exhibited in 1937, which aroused the interest of the French postal and telecommunications departments. However, due to the outbreak of World War II soon, Artsouni's robotic brain could not be installed and used.

Machine translation: the world's first computer

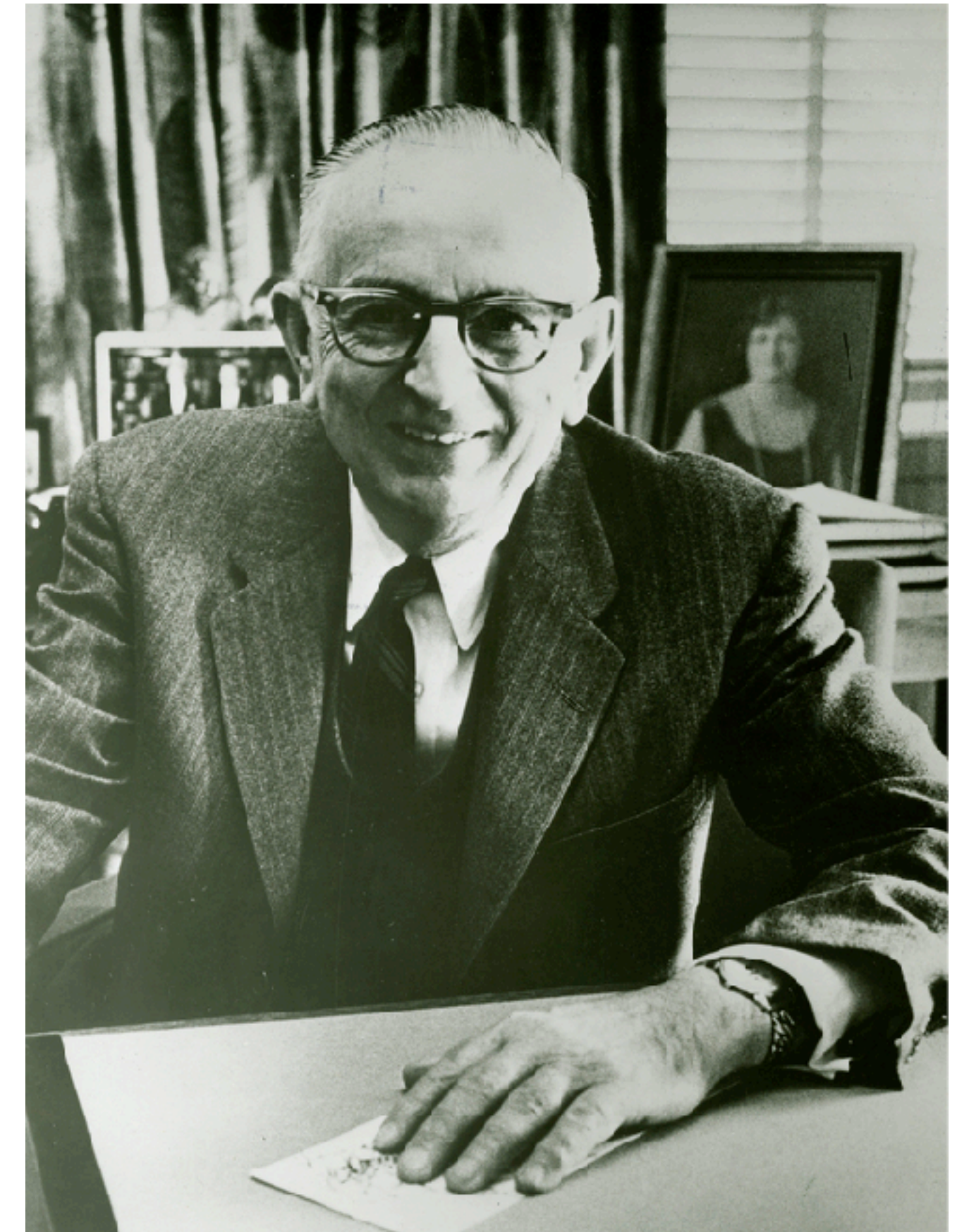
- In 1946, J.P. Eckert and J.W. Mauchly of UPenn designed the world's first electronic computer **ENIAC**



ENIAC

Machine translation: a decoding perspective

- British engineer Andrew Donald Booth and American Rockefeller Foundation (Rockefeller Foundation) Deputy President W. Weaver put forward the idea of **machine translation**. In 1949, Weaver published a memo titled "Translation", which formally raised the issue of machine translation. He said: "When I read an article written in Chinese, I can say that this article is actually written in English, but it is coded with another strange symbol. When I was reading, I was decoding. Weaver's excellent ideas became the theoretical basis of Statistic Machine Translation (SMT).
- However, translation is more complex than decoding: the complexity of lexical analysis, syntactic analysis, semantic analysis, etc., was largely omitted.



Warren Weaver

Machine translation: the first MT system

- The academic community in the United States and the United Kingdom have developed a keen interest in machine translation and have received support from the industry.
- In 1954, Georgetown University, with the assistance of IBM, used the IBM-701 computer to implement **the world's first MT system**, realizing Russian-English translation, and the system was publicly demonstrated in New York in January 1954
- In the following 10 years, with the progress of machine translation research, various natural language processing technologies emerged and gradually developed, forming this emerging discipline combining linguistics and computer technology.



IBM-701

- In 1966, the **rule-based dialogue robot ELIZA** was born in the MIT Artificial Intelligence Laboratory
- However, in the same year, **ALPAC** (Automatic Language Processing Advisory Committee) proposed in a **report** that the progress of **machine translation research in the past ten years was slow** and did not meet expectations. After the release of the report, research funding for machine translation and natural language has been greatly reduced, and research on natural language processing and artificial intelligence has entered a **winter**.

```
Welcome to
      EEEEE LL   IIII ZZZZZZ  AAAAA
      EE   LL   II   ZZ   AA  AA
      EEEEE LL   II   ZZZ  AAAAAA
      EE   LL   II   ZZ   AA  AA
      EEEEE LLLLL IIII ZZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   █
```

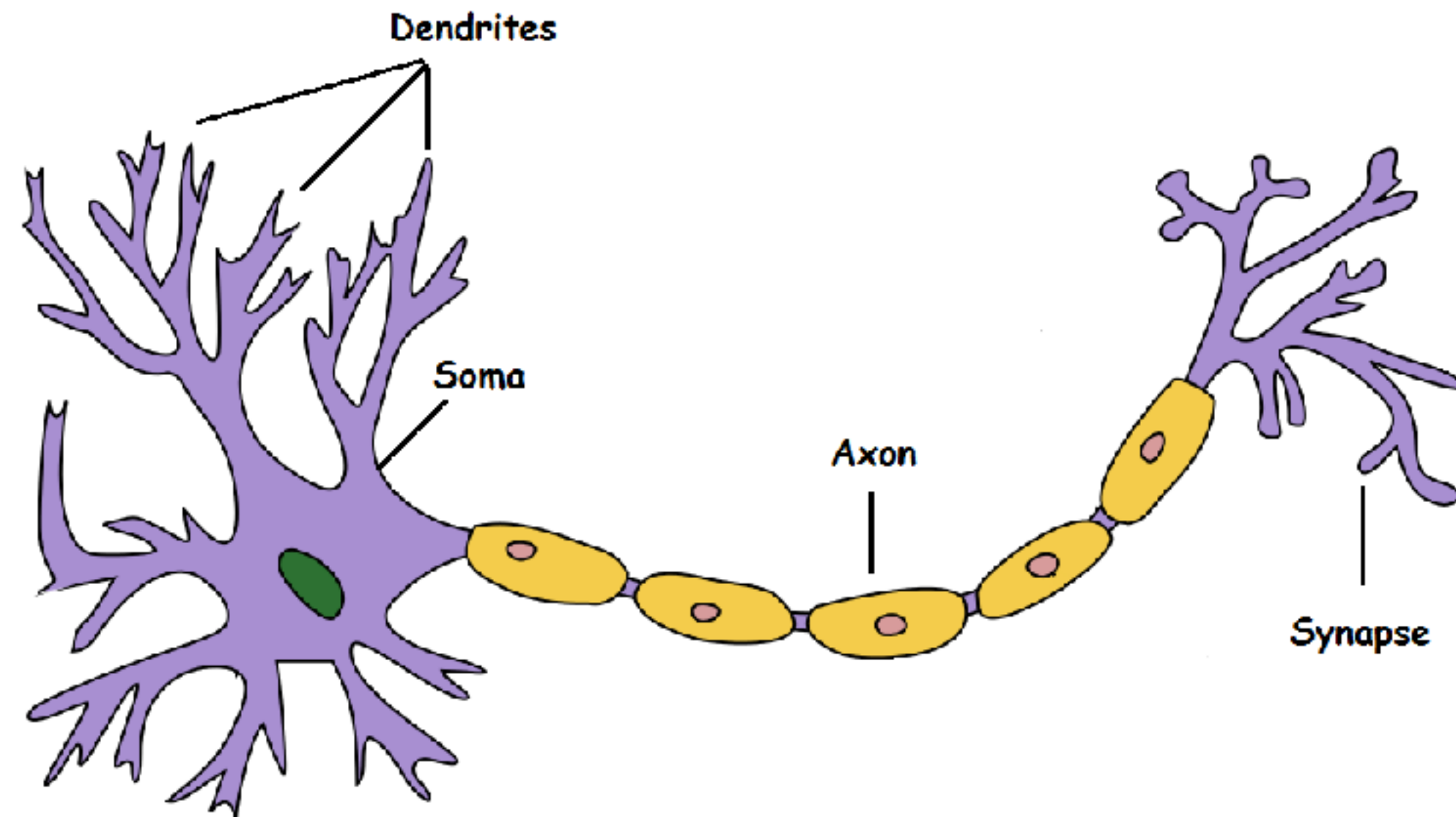
***Rule-based dialogue system
ELIZA***

- Researchers from IBM's Watson Research Center and Baker from Carnegie Mellon University have achieved **success in the development of statistical speech recognition algorithms**: "Hidden Markov Model" and "Noisy Channel Model and Decoding Model".
- At the 4th MT Summit IV held in Kobe, Japan in July 1993, the famous British scholar J. Hutchins pointed out in his special report that **the development of machine translation has entered a new era**. With the beginning of a new era of machine translation, computational linguistics has entered its boom period.



Deep Learning

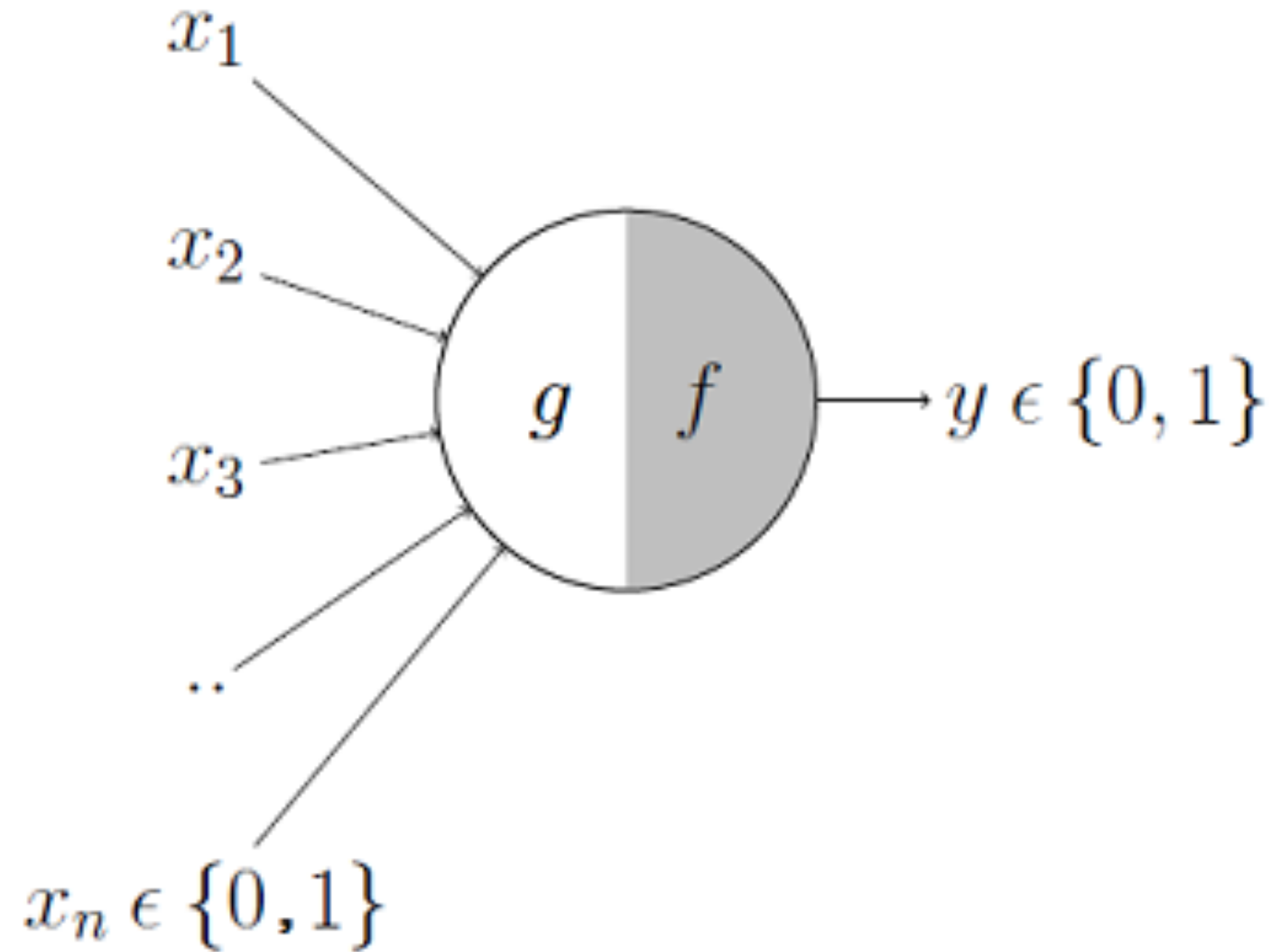
Biological neurons



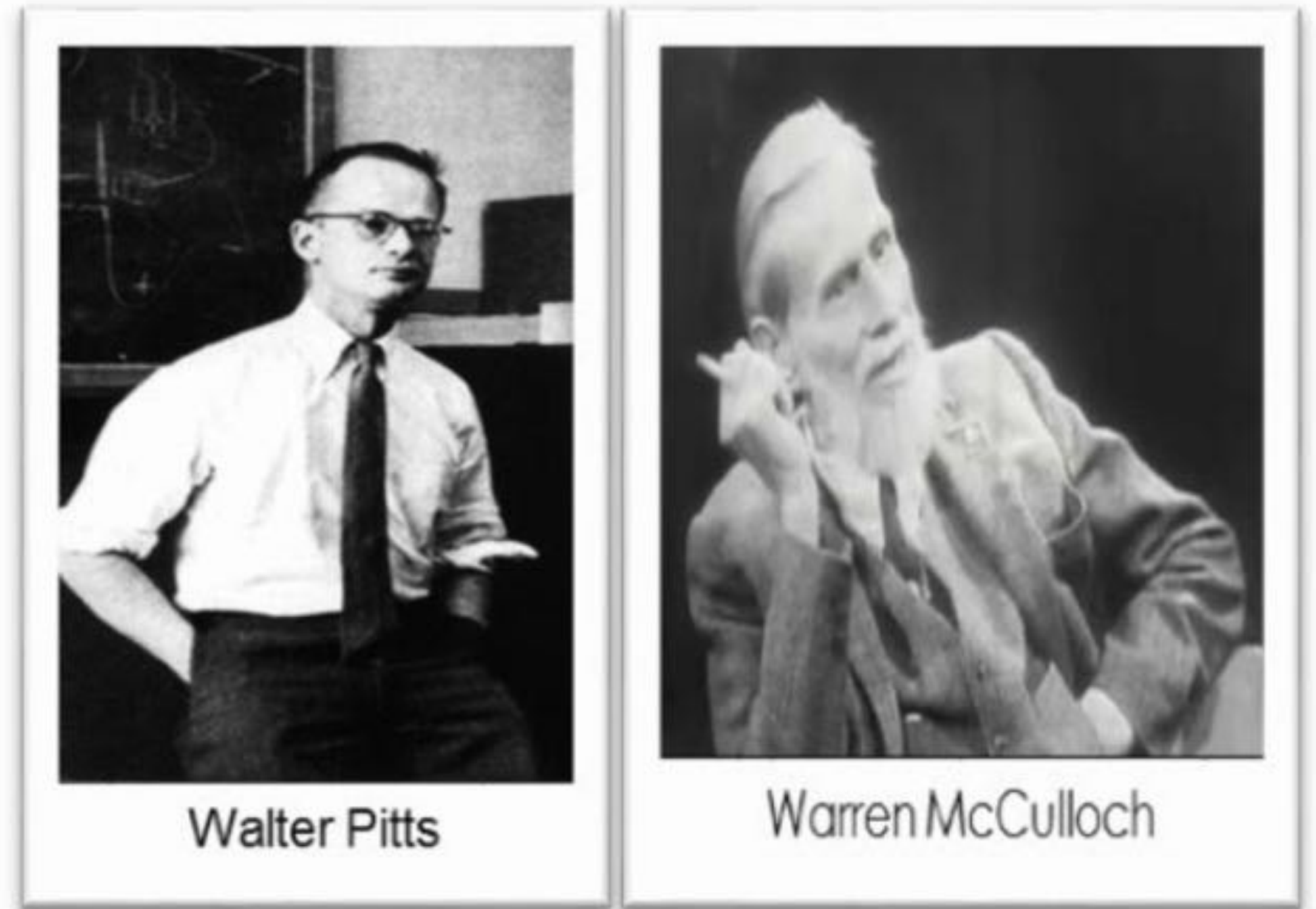
- **Dendrite:** Receives signals from other neurons
- **Soma:** Processes the information
- **Axon:** Transmits the output of this neuron
- **Synapse:** Point of connection to other neurons

***Biological Neurons:
An Overly Simplified Illustration***

62 The first computational model of a neuron

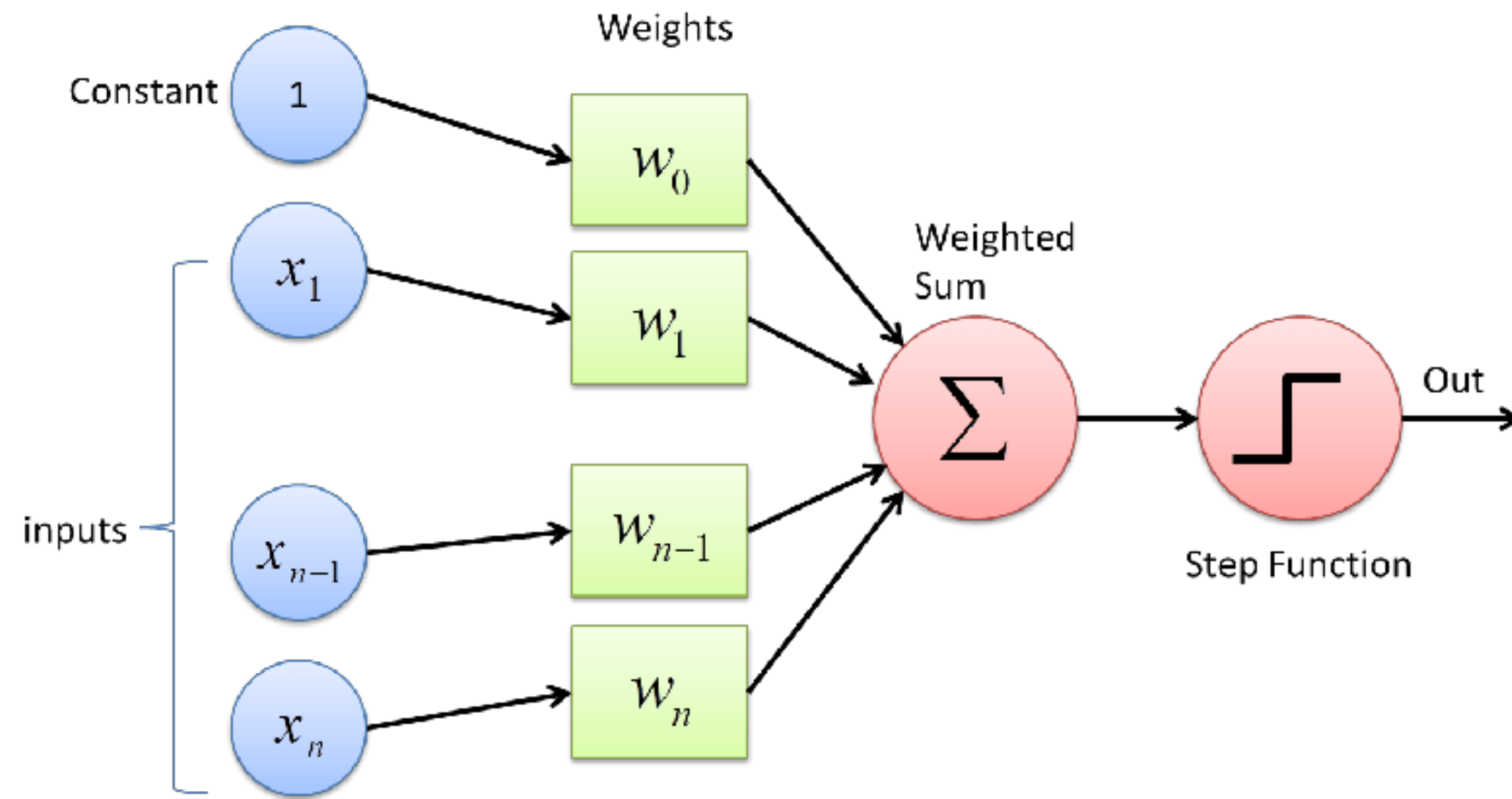


McCulloch-Pitts Neuron



The first computational model of a neuron was proposed by Warren McCulloch (neuroscientist) and Walter Pitts (logician) in 1943.

63 Ancestor of neural networks: Perceptron



The "Perceptron" proposed by Cornell University professor Frank Rosenblatt in 1957 was the first to use algorithms to accurately define neural networks, and the first mathematical model with self-organization and self-learning capabilities. It is the ancestor of neural network models.



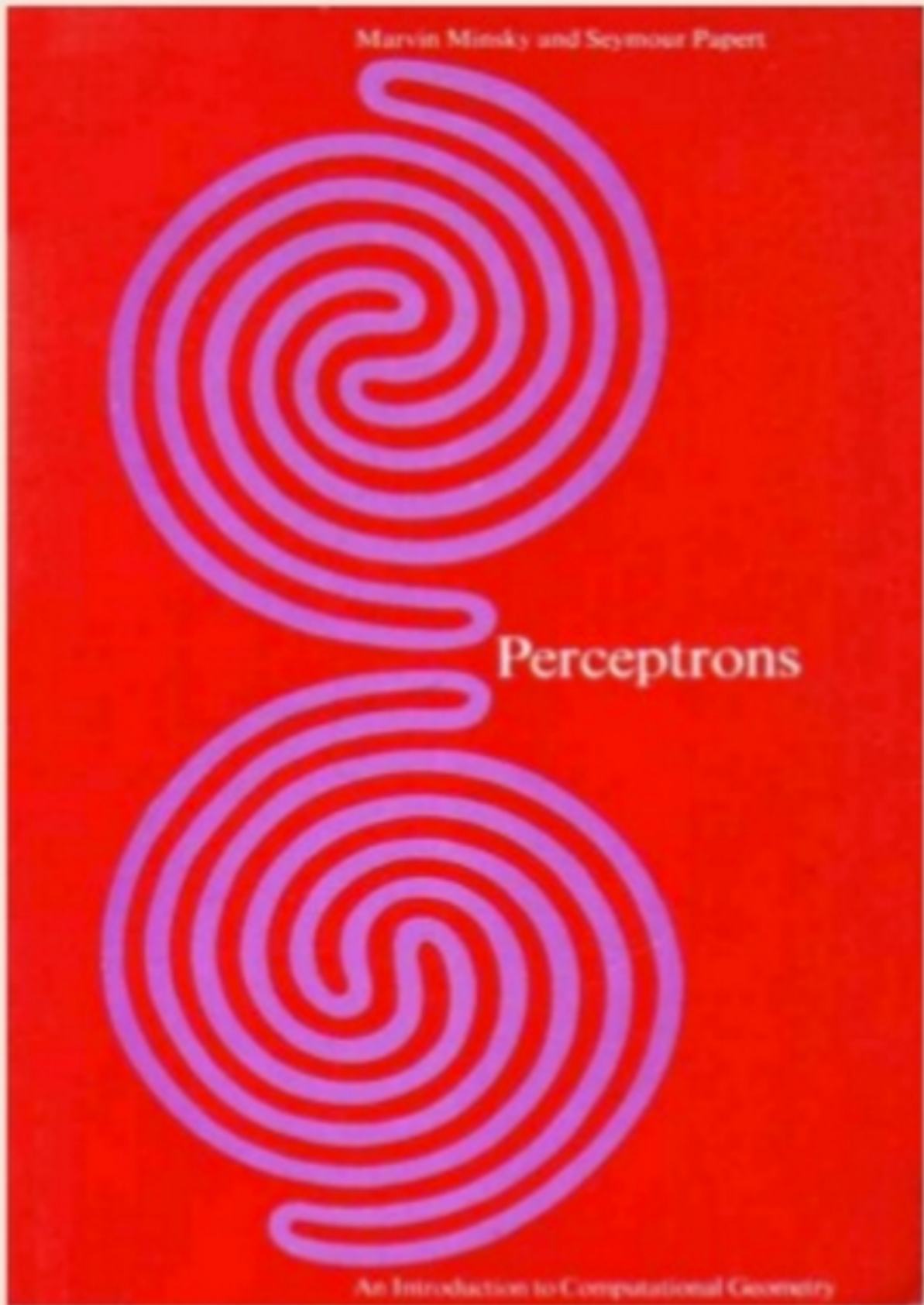
Frank Rosenblatt

Perceptrons can't do XOR

1969: Marvin Minsky and Seymour Pappert published a book - "Perceptron", which cited some limitations in Rosenblatt's technique. They found that **perceptrons can't do the Exclusive OR** logical operation.

Exclusive OR) is a digital logic gate that gives a true (1 or HIGH) output when the number of true inputs is odd. So, if input 0 and 1, or 1 and 0, the output is 1. In other cases, the outputs are 0. This book triggered a long AI Winter

1969: Perceptrons can't do XOR!

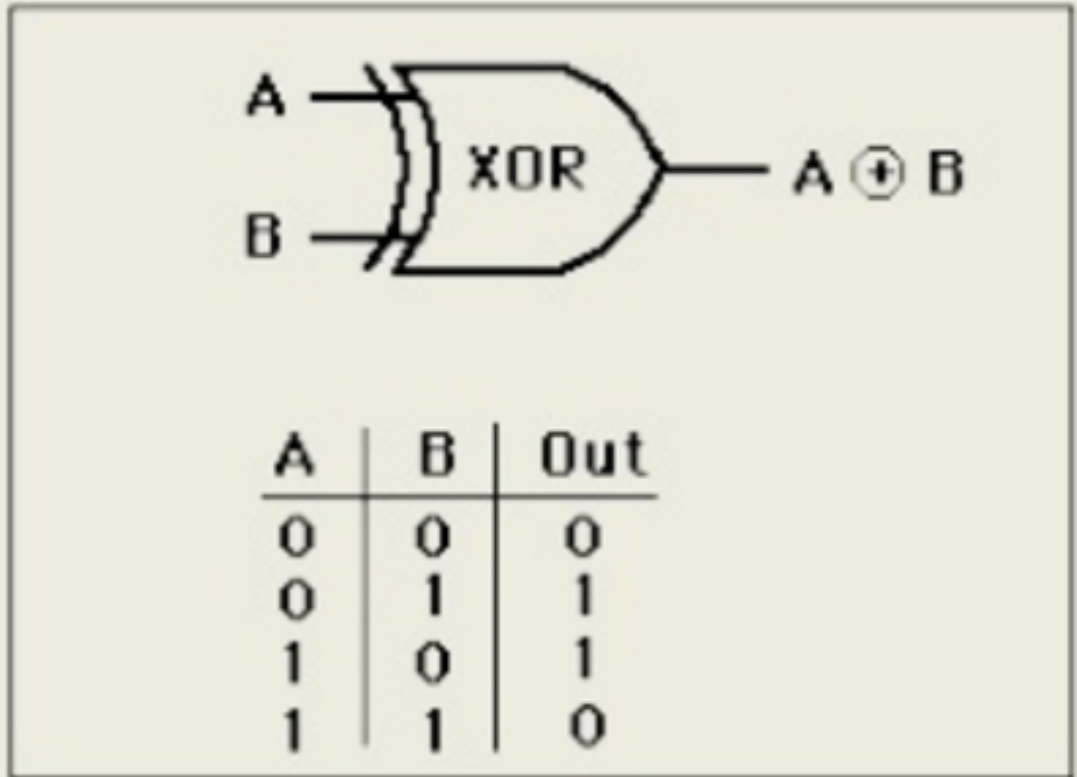


Marvin Minsky and Seymour Papert

Perceptrons


An Introduction to Computational Geometry

<http://www.i-programmer.info/images/stories/BabBag/AI/book.jpg>



A	B	Out
0	0	0
0	1	1
1	0	1
1	1	0

<http://hyperphysics.phy-astr.gsu.edu/hbase/electronic/ietron/xor.gif>



Minsky & Papert

<https://constructingkids.files.wordpress.com/2013/05/minsky-papert-71-csolomon-x640.jpg>

Second Generation NNs: Back-propagation

In July 1986, Hinton and David Rumelhart co-published a paper in the journal Nature, "**Learning Representations by Back-propagating errors**", which was the first to systematically and concisely expound the application of **back-propagating algorithms** to neural network models. **Neural network research begins to re-arise.**

Learning representations by back-propagating errors

David E. Rumelhart*, **Geoffrey E. Hinton†**
& **Ronald J. Williams***

* Institute for Cognitive Science, C-015, University of California,
San Diego, La Jolla, California 92093, USA

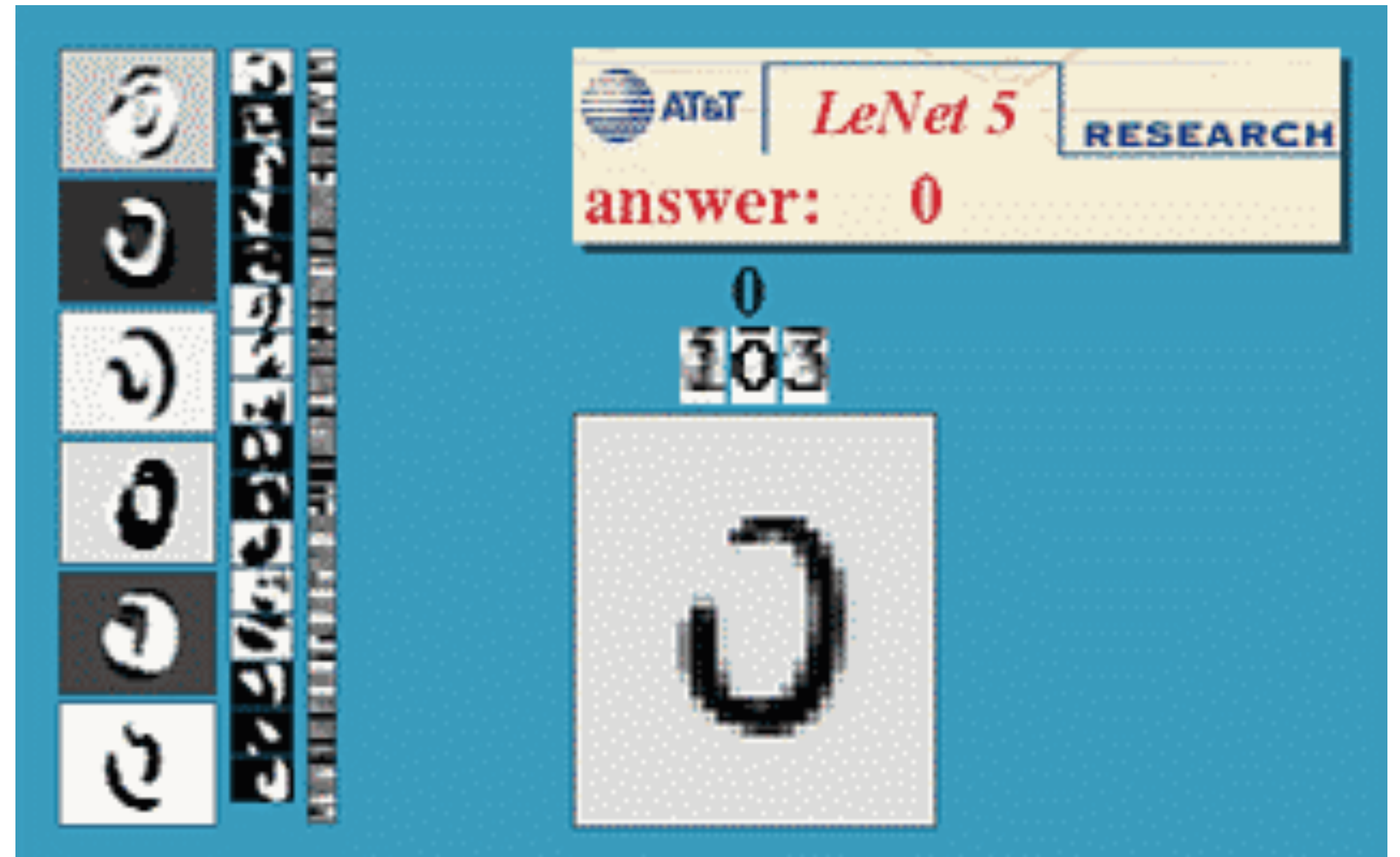
† Department of Computer Science, Carnegie-Mellon University,
Pittsburgh, Philadelphia 15213, USA

We describe a new learning procedure, **back-propagation**, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure¹.

<https://www.nature.com/articles/323533a0.pdf?origin=ppub>

Convolutional Neural Networks

In the late 1990s, Yann Lecun used a technique called “**Convolutional Neural Networks (CNN)**” for developing a commercial software to read handwritten numbers on bank checks. This **check recognition system** occupied nearly 20% of the US market in the late 1990s.



LeNet-5, convolutional neural networks
<http://yann.lecun.com/exdb/lenet/index.html>

67 Deep Belief Networks

ARTICLE

A fast learning algorithm for deep belief nets

Authors:  [Geoffrey E. Hinton](#),  [Simon Osindero](#),  [Yee-Whye Teh](#) [Authors Info & Affiliations](#)

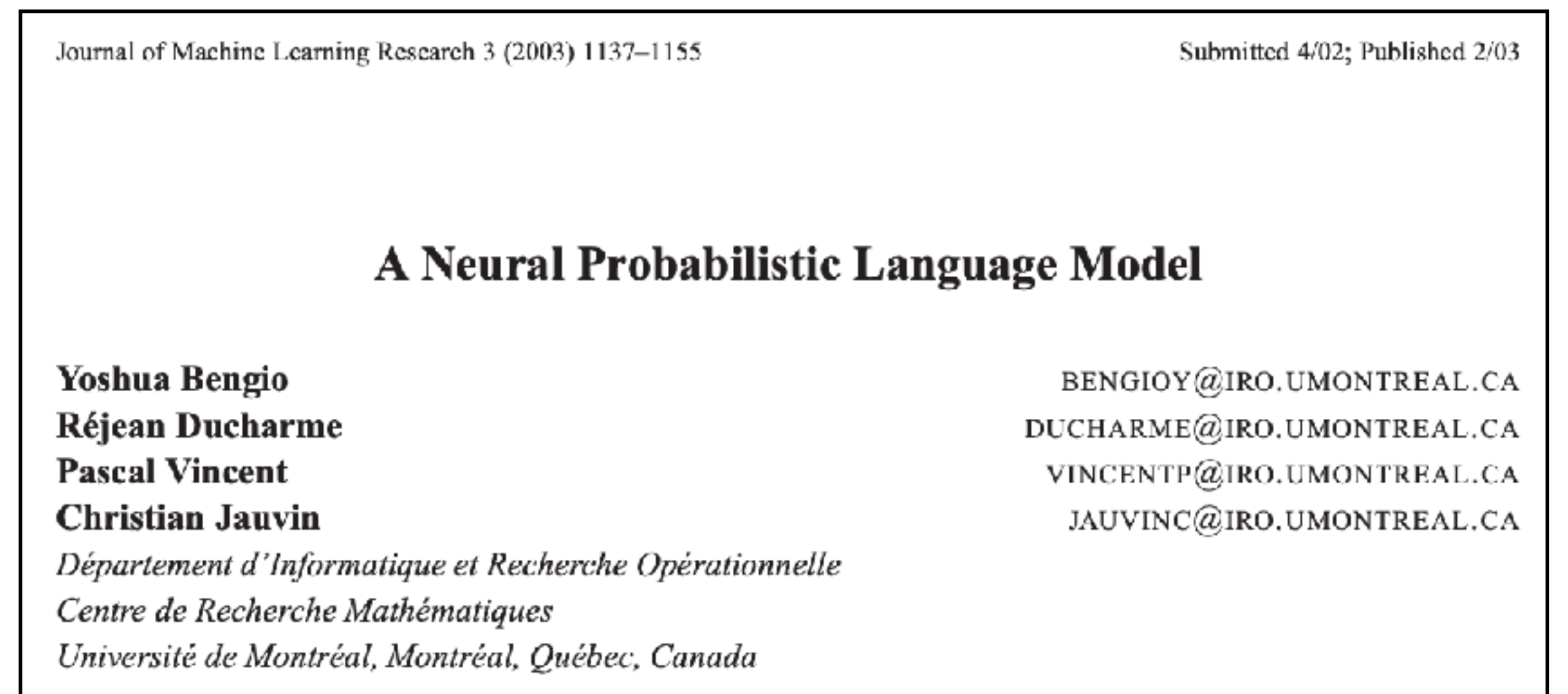
Publication: Neural Computation • July 2006 • <https://doi.org/10.1162/neco.2006.18.7.1527>

<https://dl.acm.org/doi/10.1162/neco.2006.18.7.1527>

In 2006, Geoffrey Hinton introduced **Deep Belief Networks**, also introduced **layer-wise pretraining** technique, opened current deep learning era.

68 High-dimensional word embeddings and attention

In 2000, Yoshua Bengio authored the landmark paper, “**A Neural Probabilistic Language Model**,” that introduced high-dimension **word embeddings** as a representation of word meaning. Bengio’s insights had a huge and lasting impact on natural language processing tasks. His group also introduced a form of **attention** mechanism which led to breakthroughs in machine translation and form a key component of sequential processing with deep learning.



<https://mila.quebec/wp-content/uploads/2019/08/bengio03a.pdf>

Generative Adversarial Networks

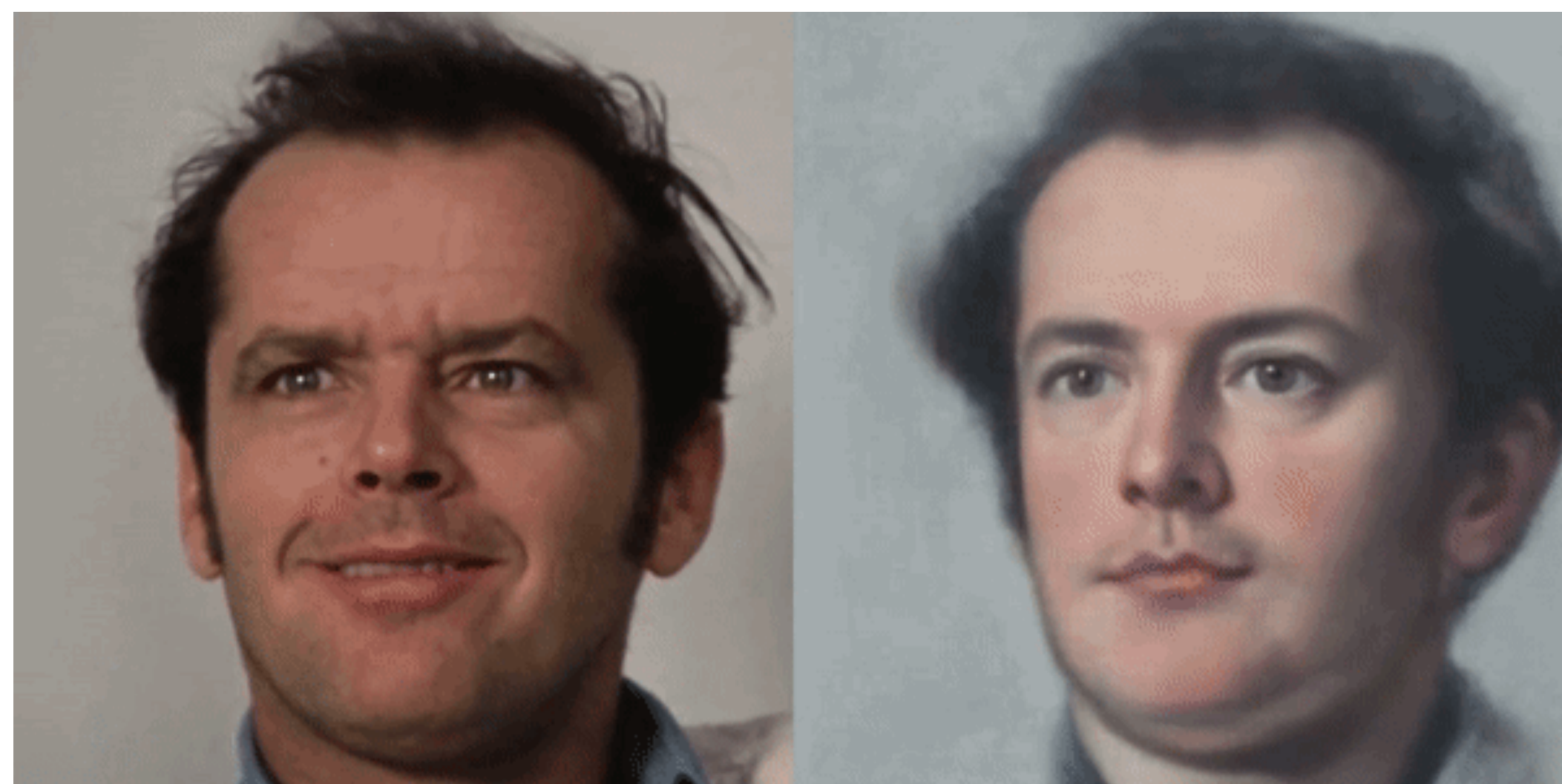
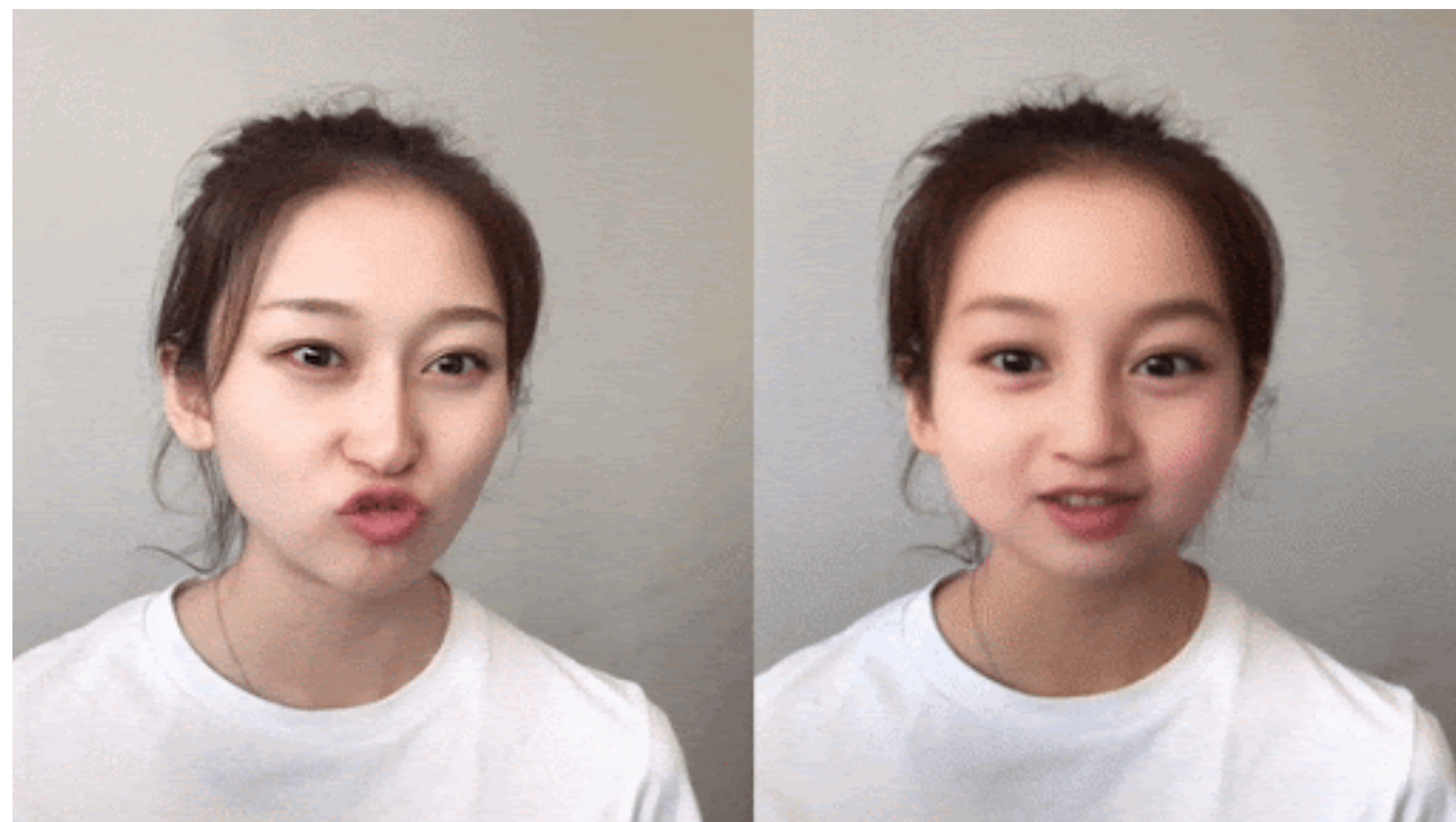
Since 2010, Bengio's papers on generative deep learning, in particular the **Generative Adversarial Networks (GANs)** developed with Ian Goodfellow, have spawned a revolution in computer vision and computer graphics. In one fascinating application of this work, computers can actually create original images, reminiscent of the creativity that is considered a hallmark of human intelligence.

Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie*, Mehdi Mirza, Bing Xu, David Warde-Farley,
Sherjil Ozair†, Aaron Courville, Yoshua Bengio‡
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

<https://arxiv.org/pdf/1406.2661.pdf>

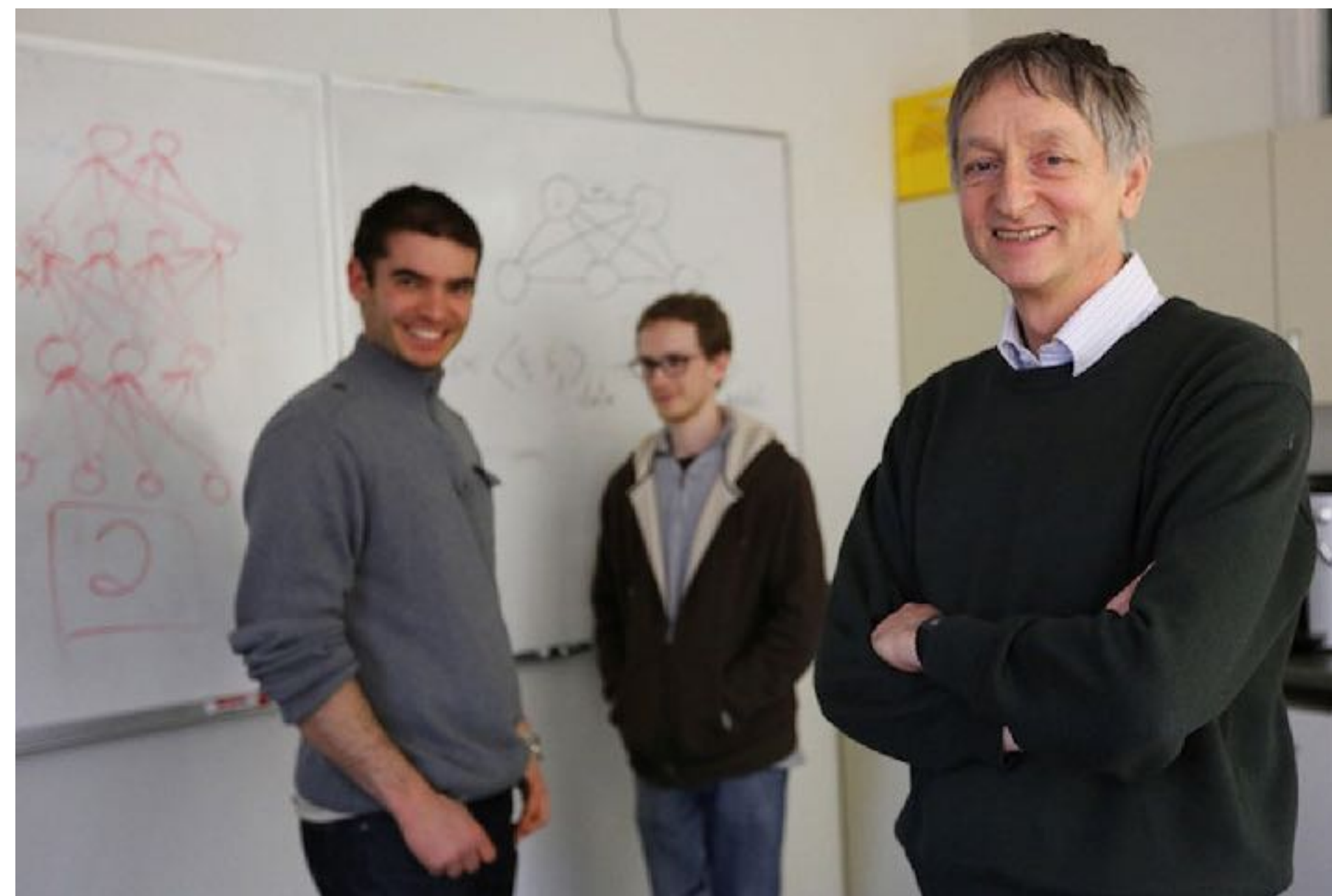
70 Generative Adversarial Networks



71 ImageNet competition

At the end of 2012, Geoff Hinton and his PhD students Alex Krizhevsky and Ilya Sutskever took the first place in the **ImageNet image classification competition**, and increased the accuracy rate to 84.7%.

Relying on deep learning, they shocked the machine learning community. Since then, a large number of researchers have begun to enter this field.



In 2013, Google Hires Brains that Helped Supercharge Machine Learning

Table 1: Major milestones that will be covered in this paper

Year	Contributer	Contribution
300 BC	Aristotle	introduced Associationism, started the history of human's attempt to understand brain.
1873	Alexander Bain	introduced Neural Groupings as the earliest models of neural network, inspired Hebbian Learning Rule.
1943	McCulloch & Pitts	introduced MCP Model, which is considered as the ancestor of Artificial Neural Model.
1949	Donald Hebb	considered as the father of neural networks, introduced Hebbian Learning Rule, which lays the foundation of modern neural network.
1958	Frank Rosenblatt	introduced the first perceptron, which highly resembles modern perceptron.
1974	Paul Werbos	introduced Backpropagation
1980	Teuvo Kohonen	introduced Self Organizing Map
	Kunihiko Fukushima	introduced Neocogitron, which inspired Convolutional Neural Network
1982	John Hopfield	introduced Hopfield Network
1985	Hilton & Sejnowski	introduced Boltzmann Machine
1986	Paul Smolensky	introduced Harmonium, which is later known as Restricted Boltzmann Machine
	Michael I. Jordan	defined and introduced Recurrent Neural Network
1990	Yann LeCun	introduced LeNet, showed the possibility of deep neural networks in practice
1997	Schuster & Paliwal	introduced Bidirectional Recurrent Neural Network
	Hochreiter & Schmidhuber	introduced LSTM, solved the problem of vanishing gradient in recurrent neural networks
2006	Geoffrey Hinton	introduced Deep Belief Networks, also introduced layer-wise pretraining technique, opened current deep learning era.
2009	Salakhutdinov & Hinton	introduced Deep Boltzmann Machines
2012	Geoffrey Hinton	introduced Dropout, an efficient way of training neural networks

On the Origin of Deep Learning

2018 Turing Award



Yoshua Bengio

Geoffrey Hinton

Yann LeCun

<https://awards.acm.org/about/2018-turing>

AlphaGo vs. Lee Sedol

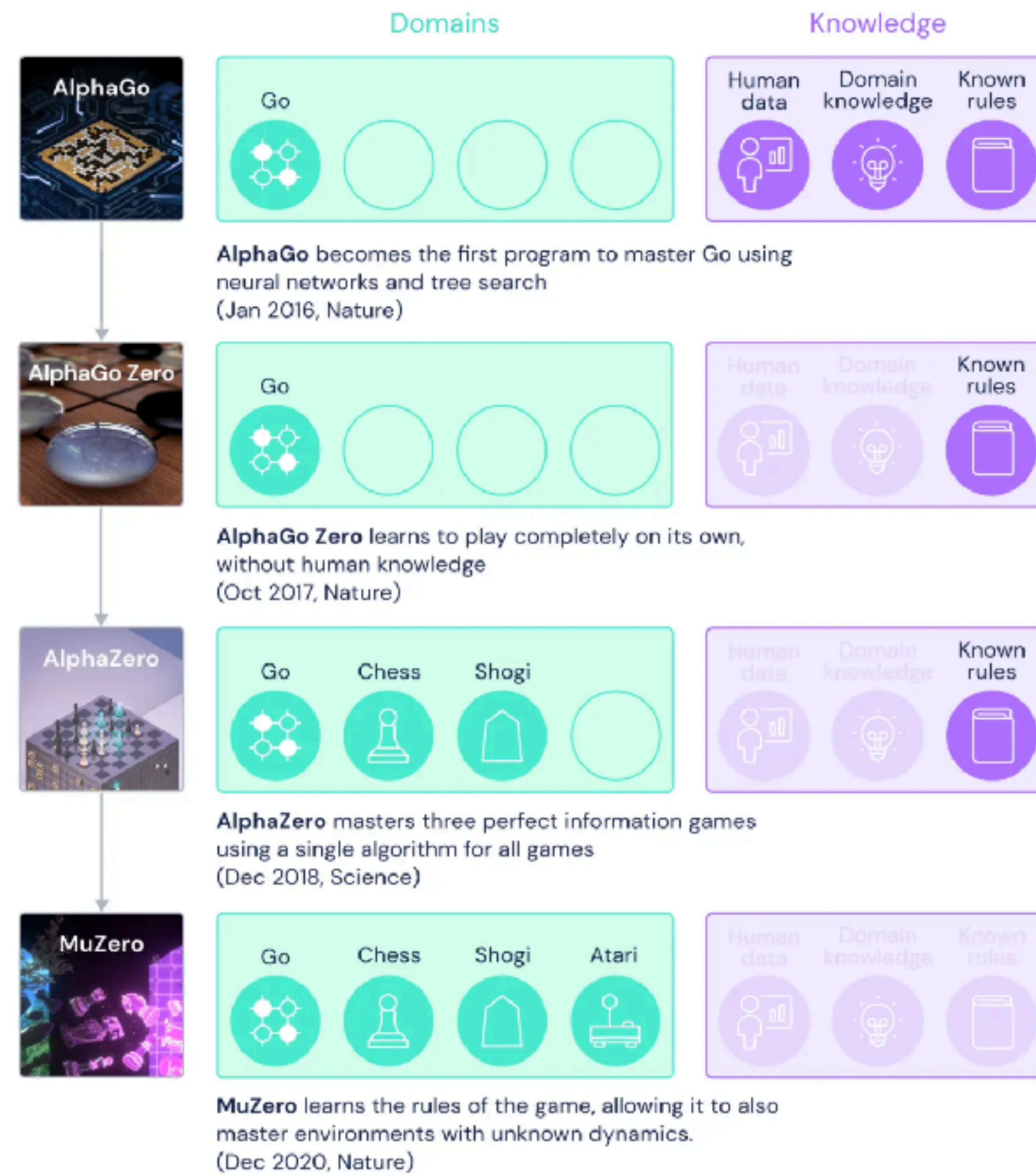
4-1

March 2016

ALPHAGO

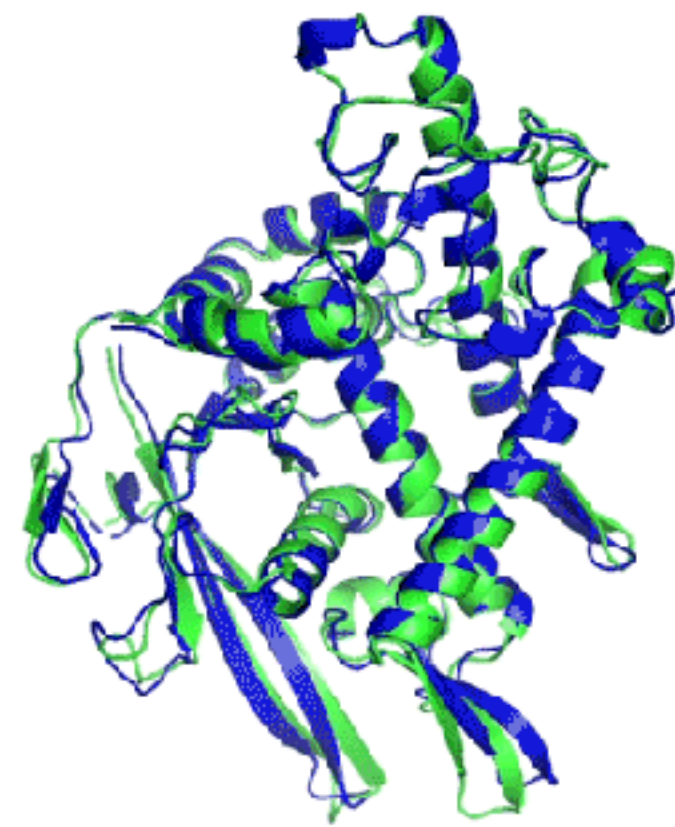


From AlphaGo to MuZero

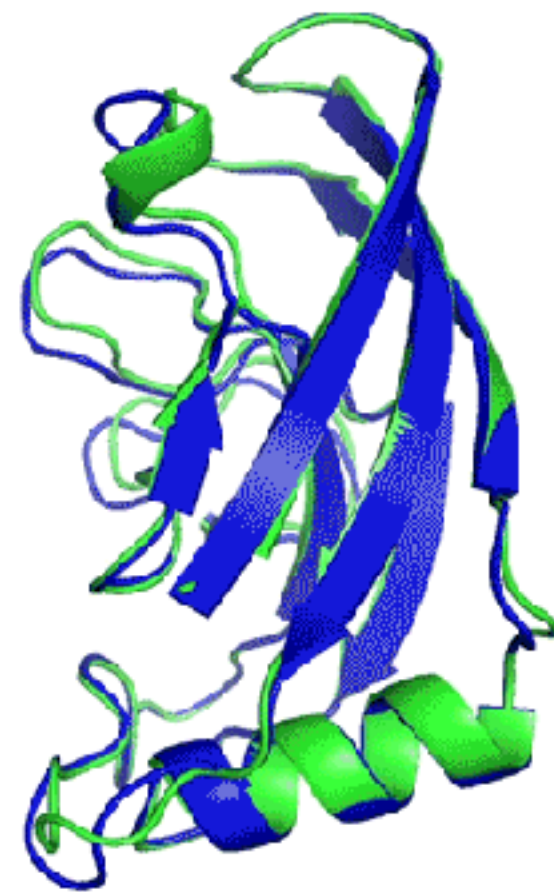


MuZero: Mastering Go, chess, shogi and Atari without rules

AlphaFold 2



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction

“

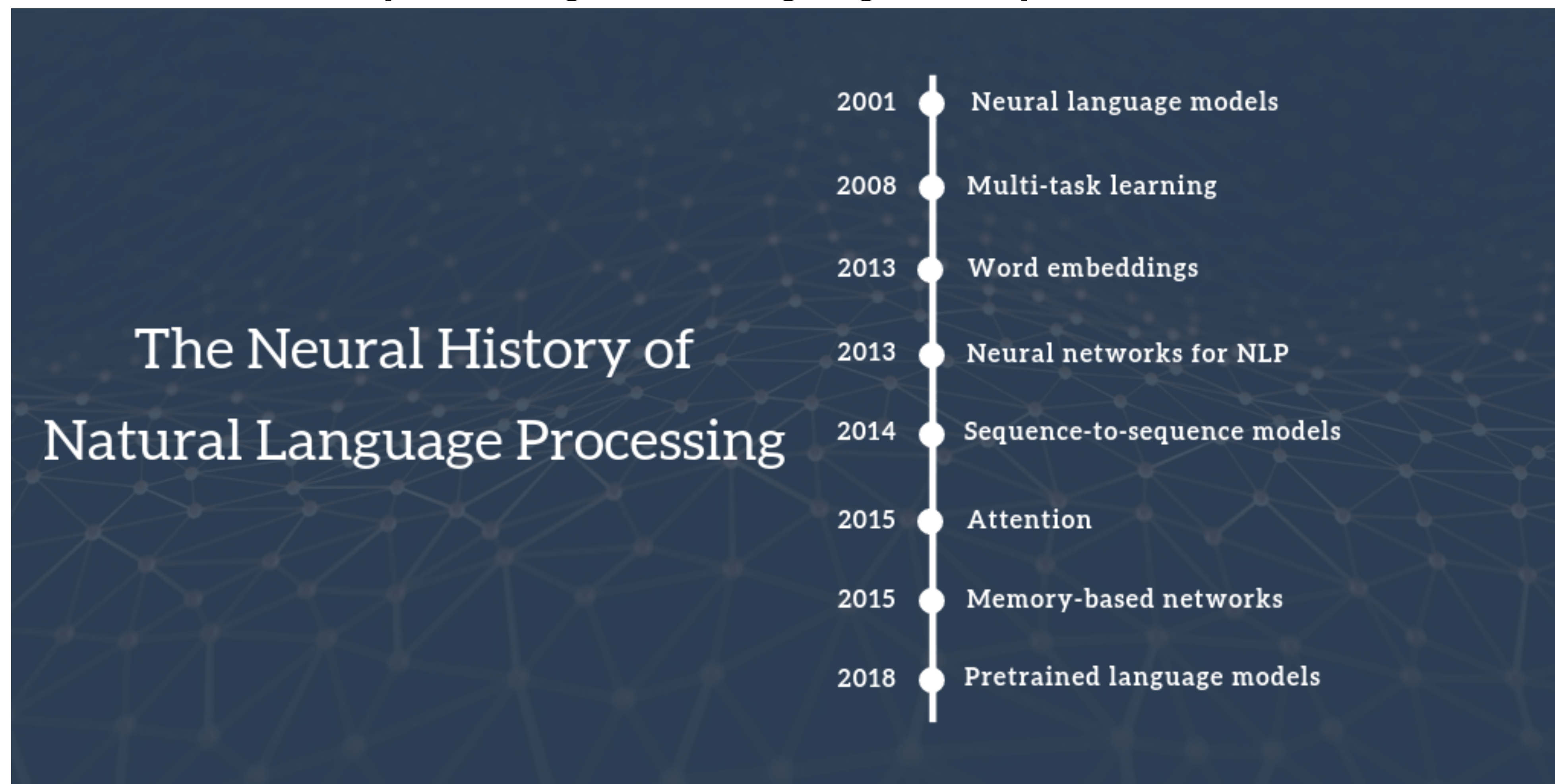
We have been stuck on this one problem – how do proteins fold up – for nearly 50 years. To see DeepMind produce a solution for this, having worked personally on this problem for so long and after so many stops and starts, wondering if we’d ever get there, is a very special moment.

PROFESSOR JOHN MOULT
CO-FOUNDER AND CHAIR OF CASP, UNIVERSITY OF MARYLAND

AlphaFold: a solution to a 50-year-old grand challenge in biology

77 The neural history of NLP

Deep learning has brought great impacts to NLP

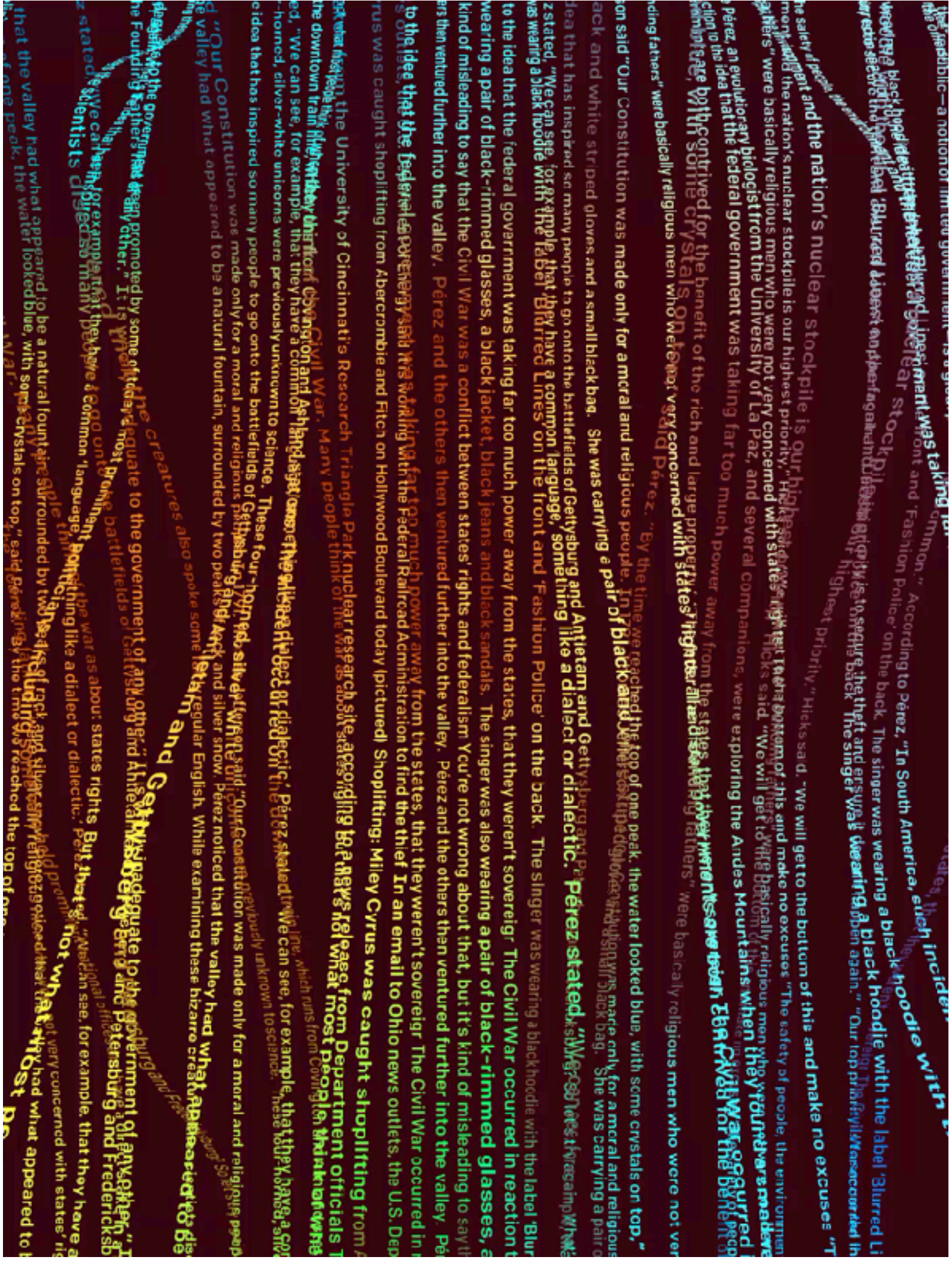


<https://runder.io/a-review-of-the-recent-history-of-nlp/>

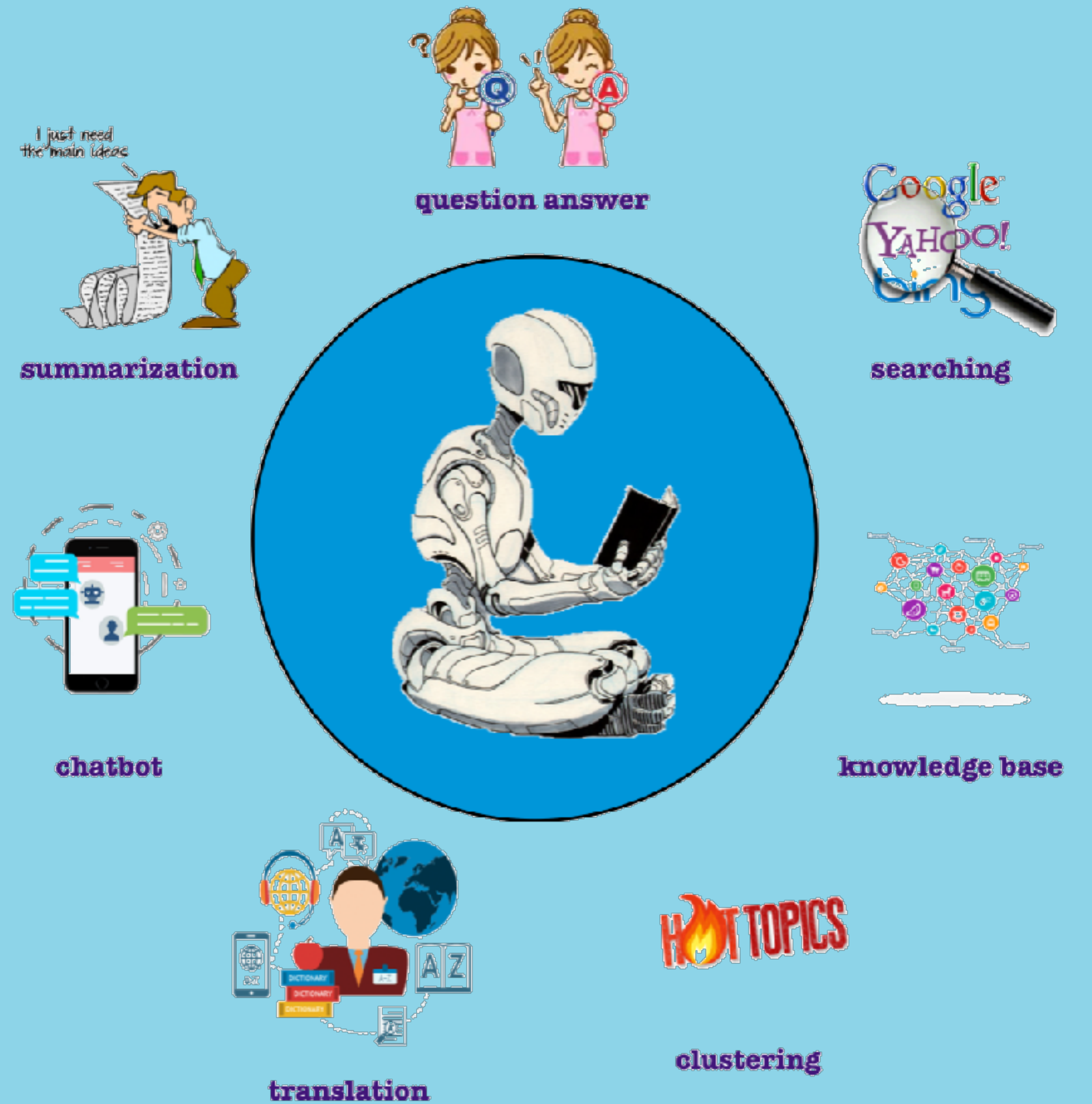
OpenAI GPT-3

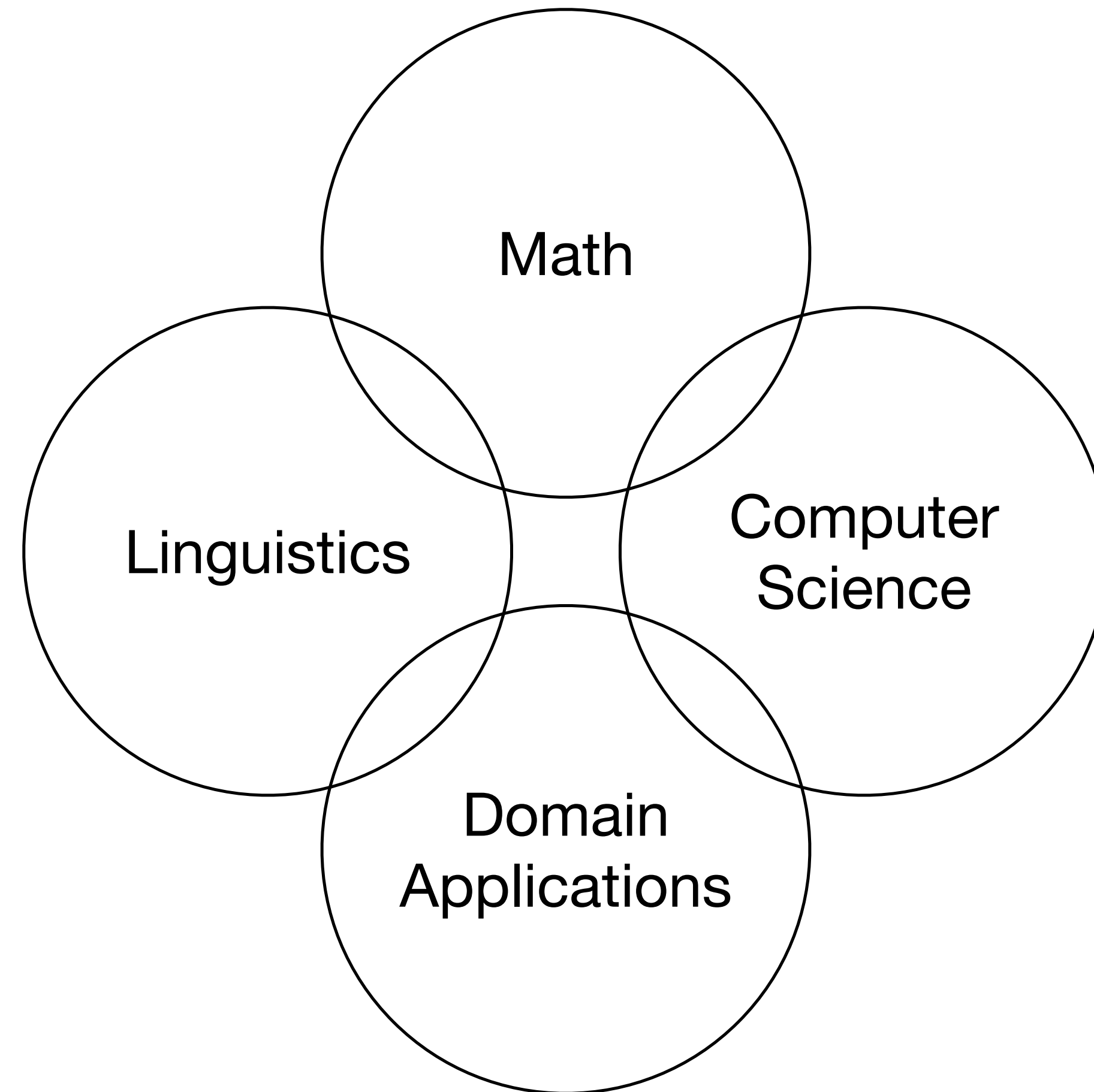
GPT-3, an autoregressive language model with 175 billion parameters, can "generate news articles which human evaluators have difficulty distinguishing from articles written by humans"

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan†	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess		Jack Clark	Christopher Berner	
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei	
		OpenAI		

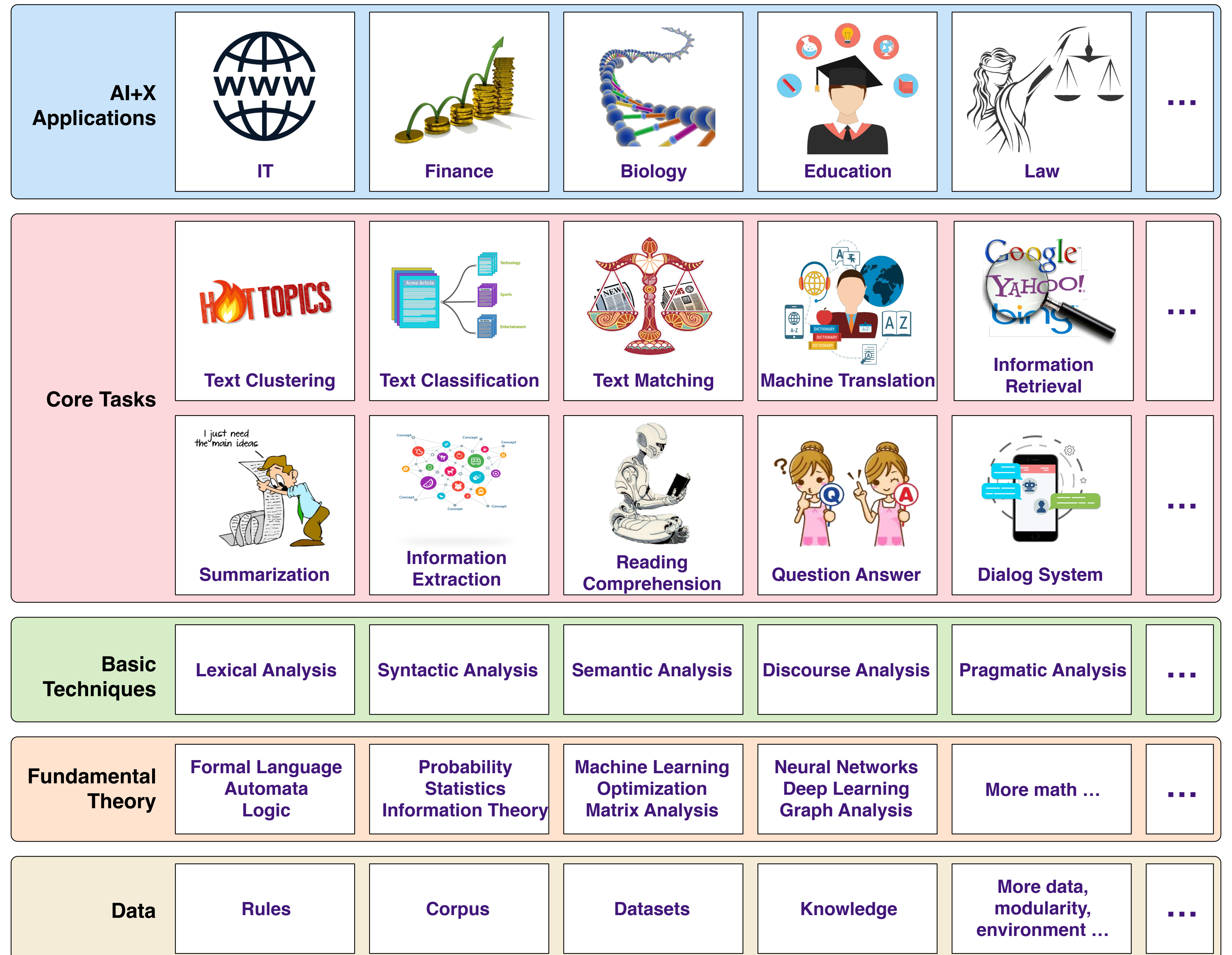


NLP techniques

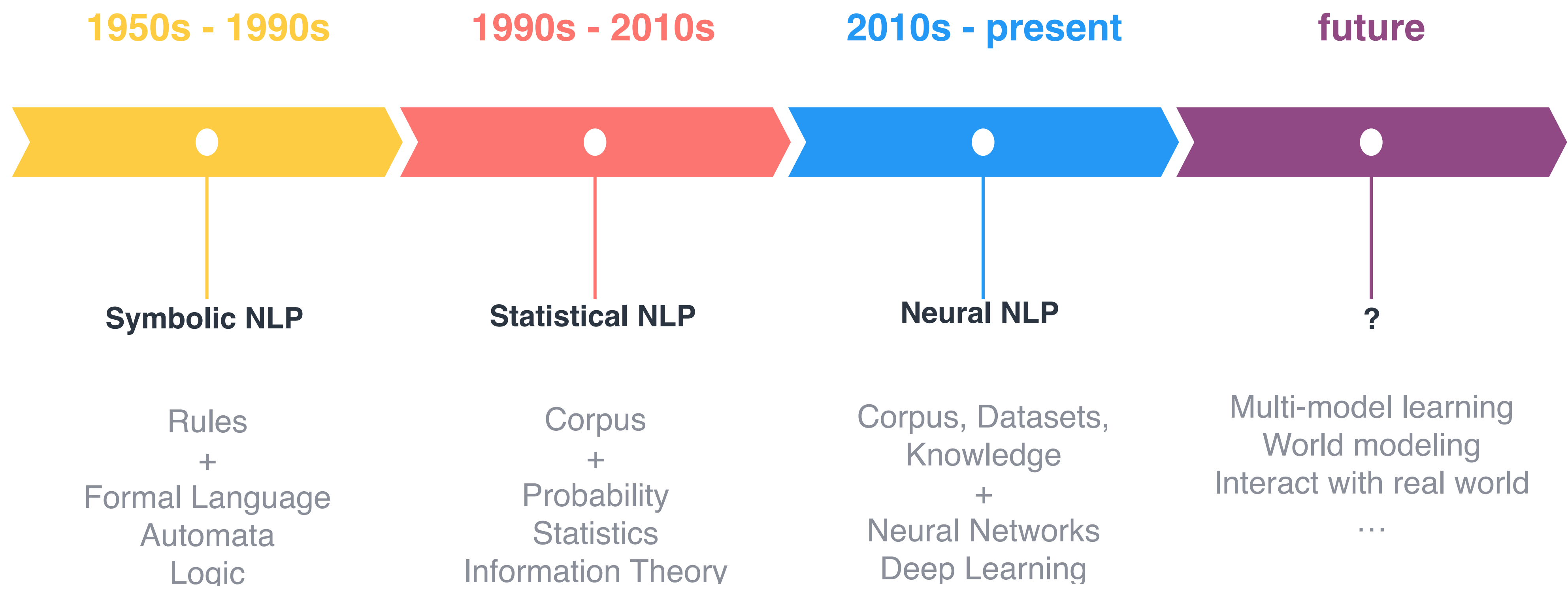




NLP Building



Different stages of NLP



Grammatical hierarchy

- **Sentence:** a sentence may consists of one or multiple clauses.
- **Clause:** a clause is a part of the sentence that contains a verb
- **Phrase:** a phrase is a group of words that express a concept and is used as a unit within a sentence, e.g., "Who ate **the last sandwich?**"
- **Word:** A word may consist of a root morpheme only, e.g. science, or a root morpheme plus other morphemes, e.g. "released = release + ed", "motivation = motivate + ion"
- **Morpheme:** morphemes are parts of words and are the smallest grammatical units, e.g., "ed", "ion", and simple words.

Sentences

consist of one or more

Clauses

consist of one or more

Phrases

consist of one or more

Words

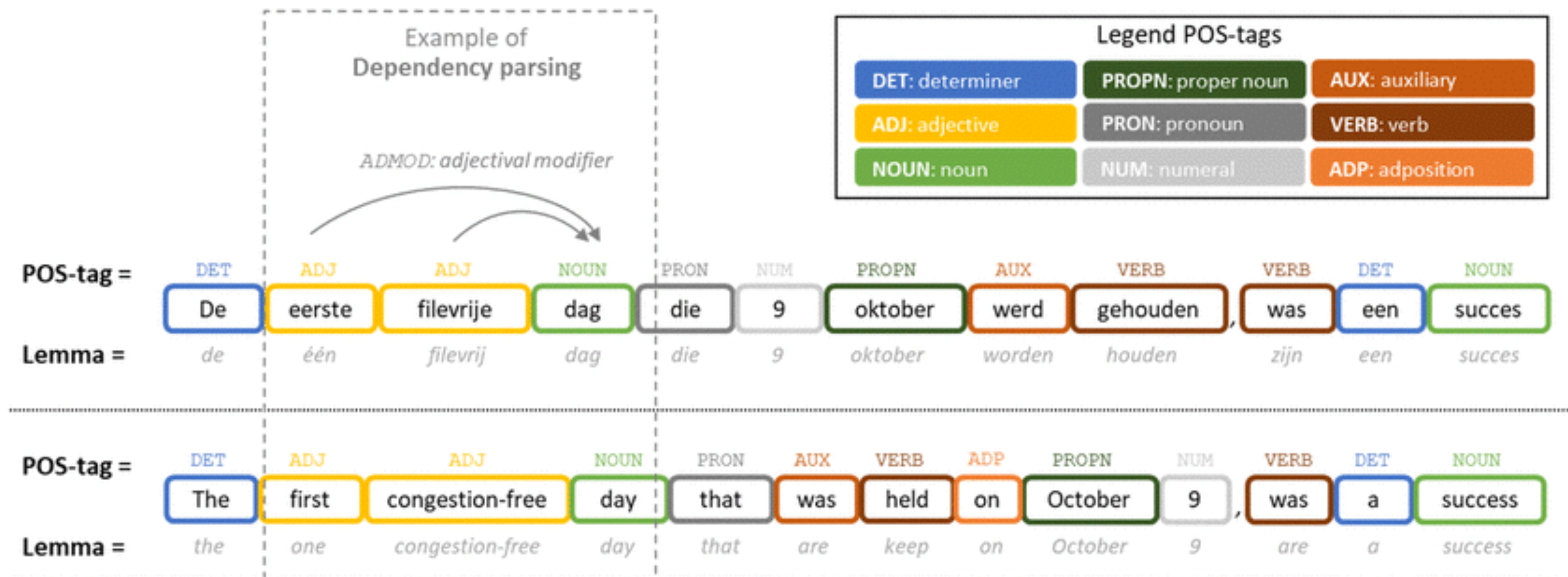
consist of one or more

Morphemes

consist of one or more phonemes .

84 Lexical analysis

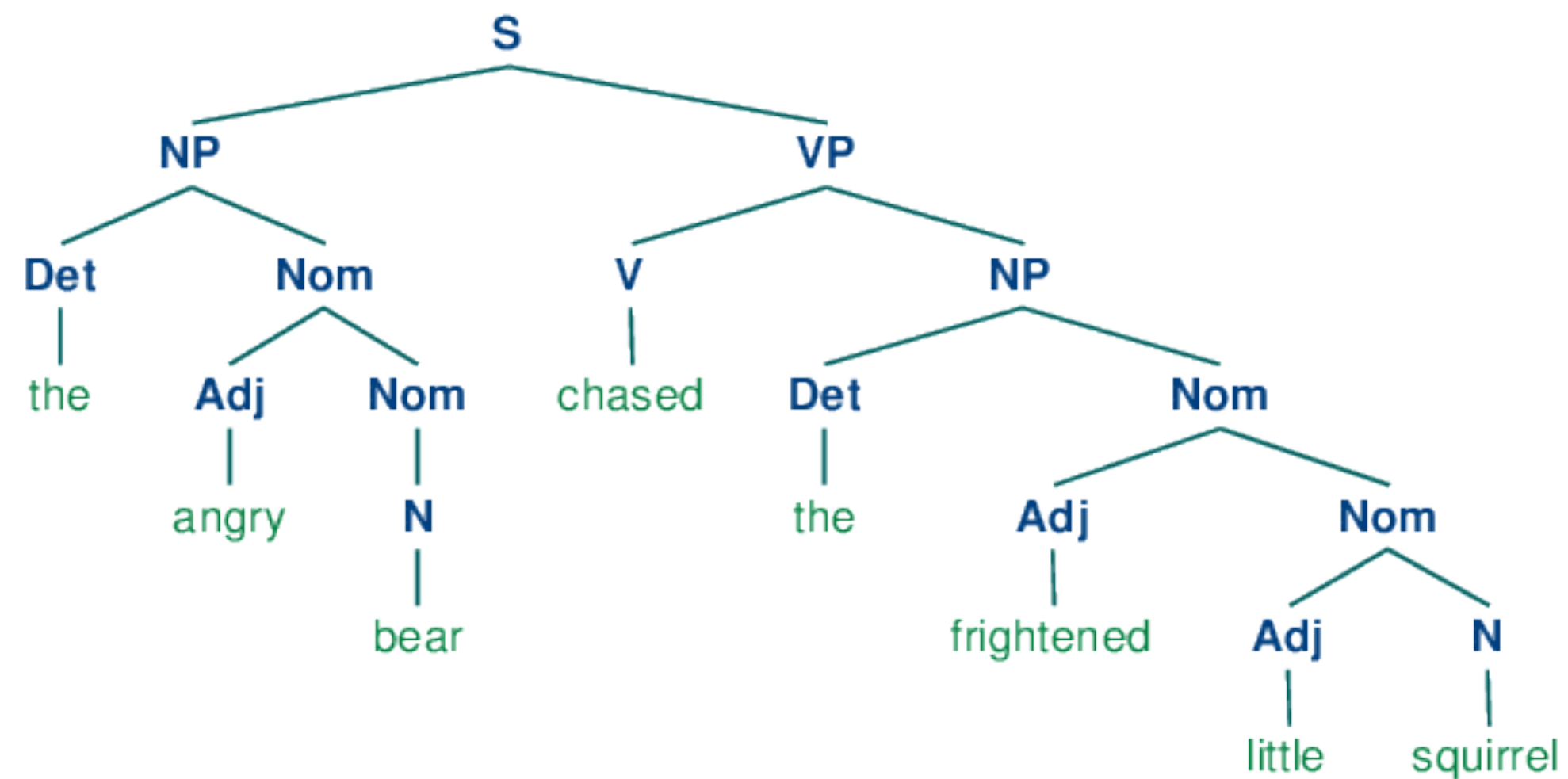
Get tokens or words from sentences, and obtain linguistic information of the words, e.g., tokenization, word segmentation, Part-of-Speech (POS) tagging



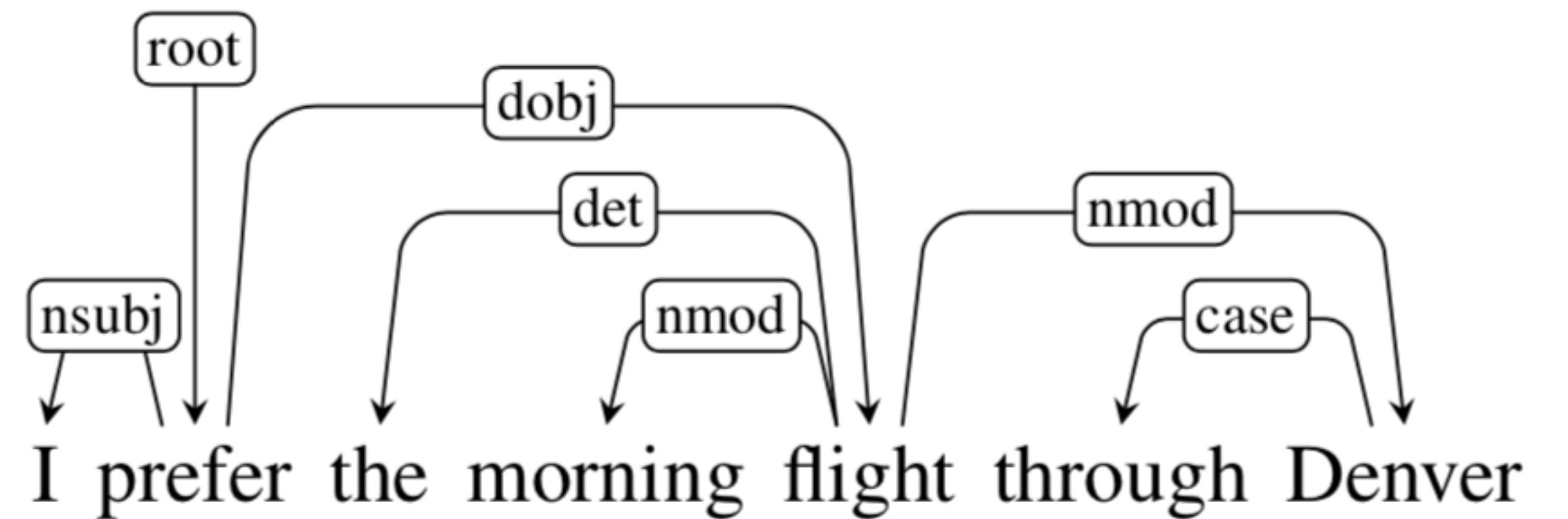
POS tagging

Syntactic analysis

Syntactic analysis, also referred to as **syntax analysis** or **parsing**, is the process of analyzing natural language with the rules of a **formal grammar**. Grammatical rules are applied to categories and groups of words, not individual words. Syntactic analysis basically **assigns a semantic structure to text**.



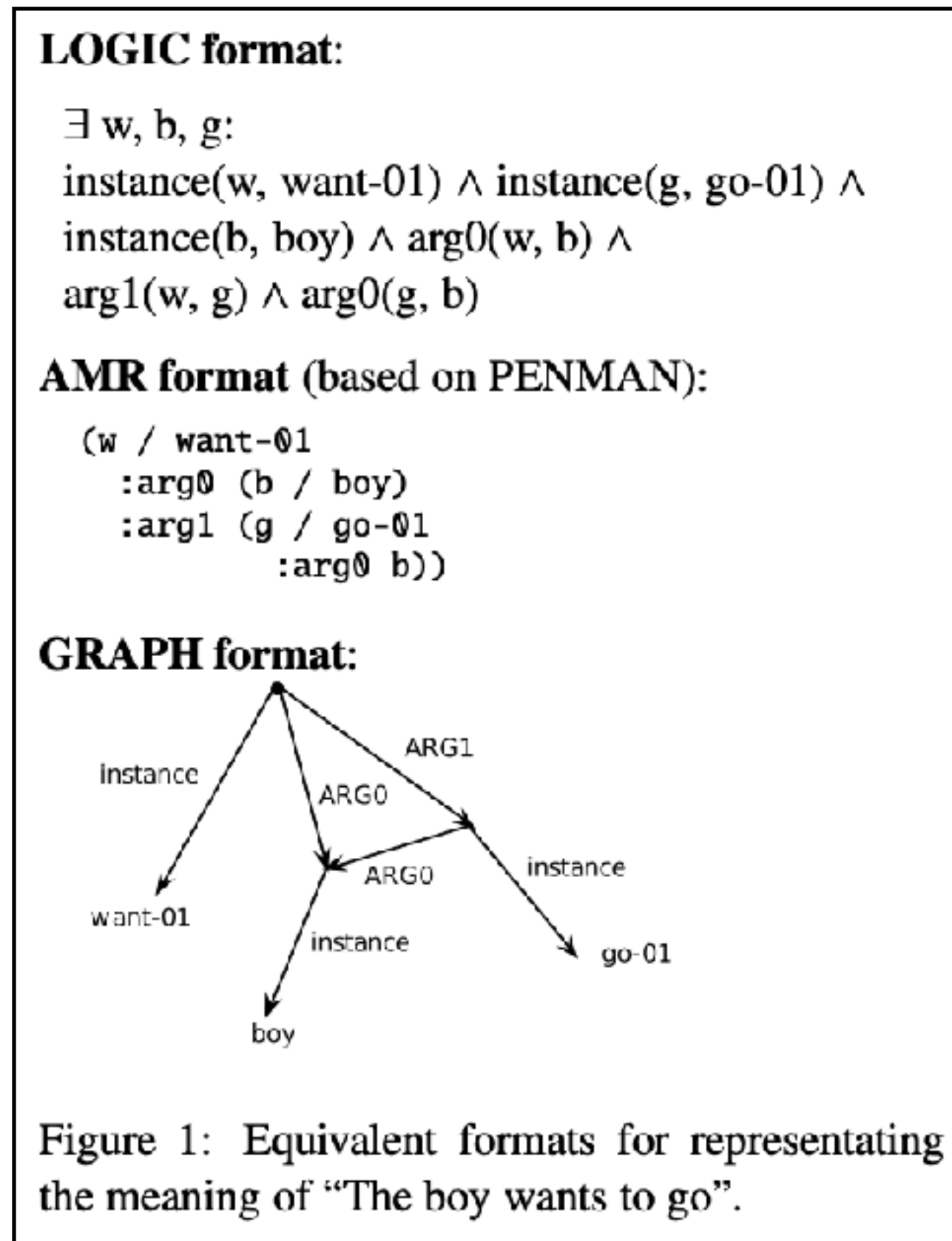
Constituency Parsing



Dependency Parsing

Semantic analysis

Semantic analysis, simply put, is the process of drawing **meaning** from text. Including Word Sense Disambiguation, Semantic Role Labeling, Semantic Parsing, etc.



Abstract Meaning Representation (AMR)

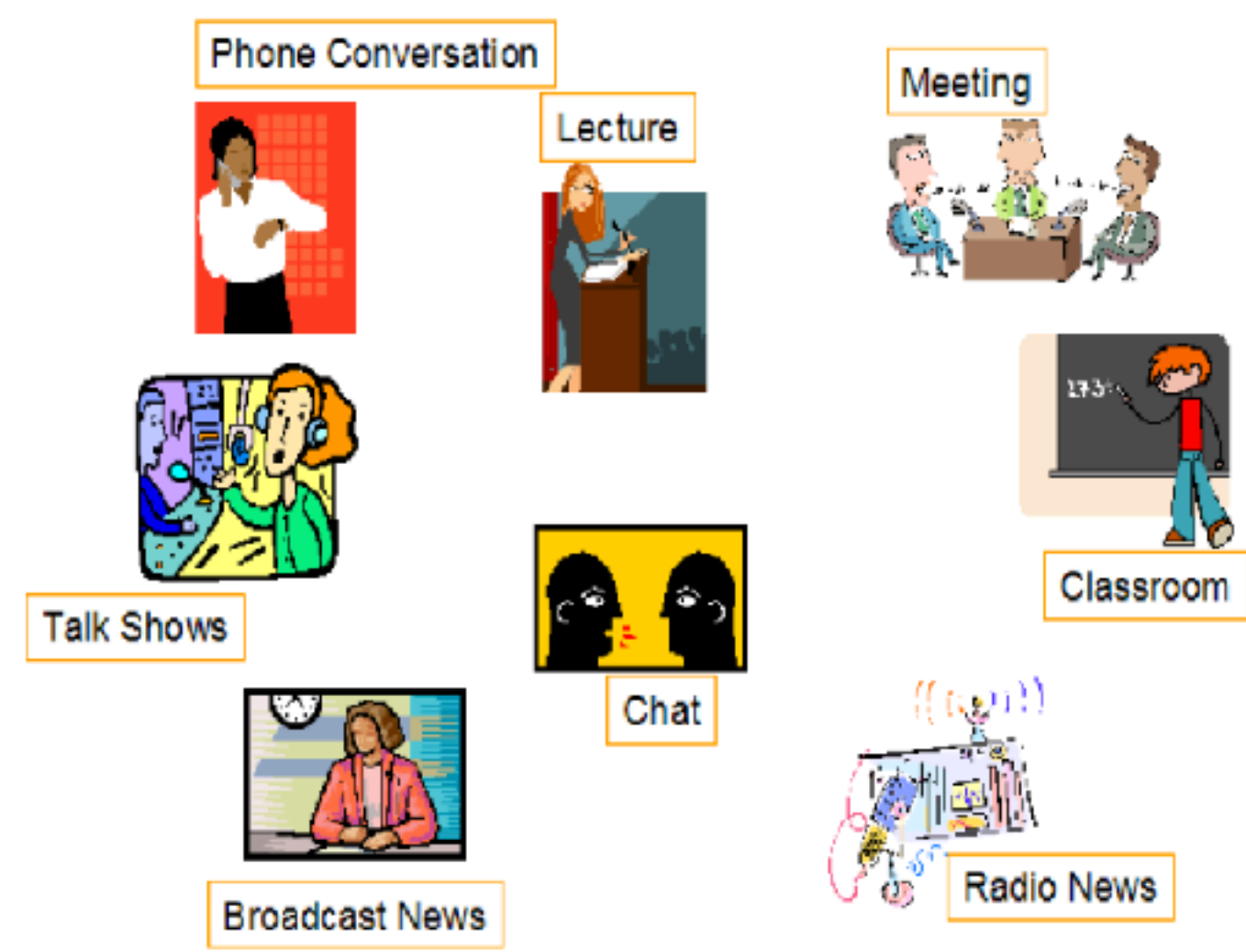
Discourse analysis

Discourse processing is a suite of Natural Language Processing (NLP) tasks to **uncover linguistic structures from texts at several levels**, which can **support many downstream applications**.

• Text/Written



• Speech

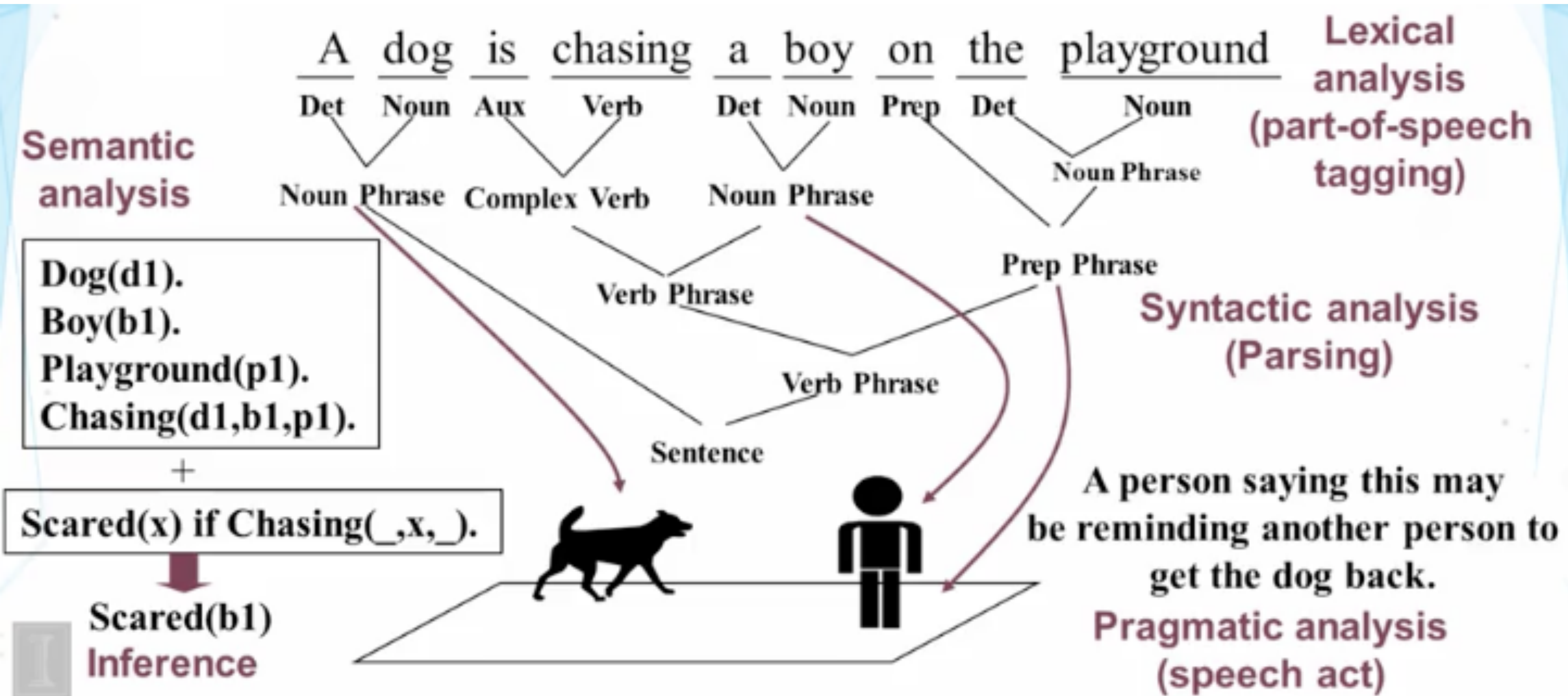


Discourse forms

Pragmatic analysis

Pragmatic analysis deals with outside **word knowledge**, which means knowledge that is external to the documents and/or queries. Pragmatics analysis that focuses on what was described is reinterpreted by what it actually meant, deriving the various aspects of language that require real world knowledge.

89 Basic concepts in NLP



NLP + X

IT

Finance

Service

Health

Biology

Education

Law

NLP resources

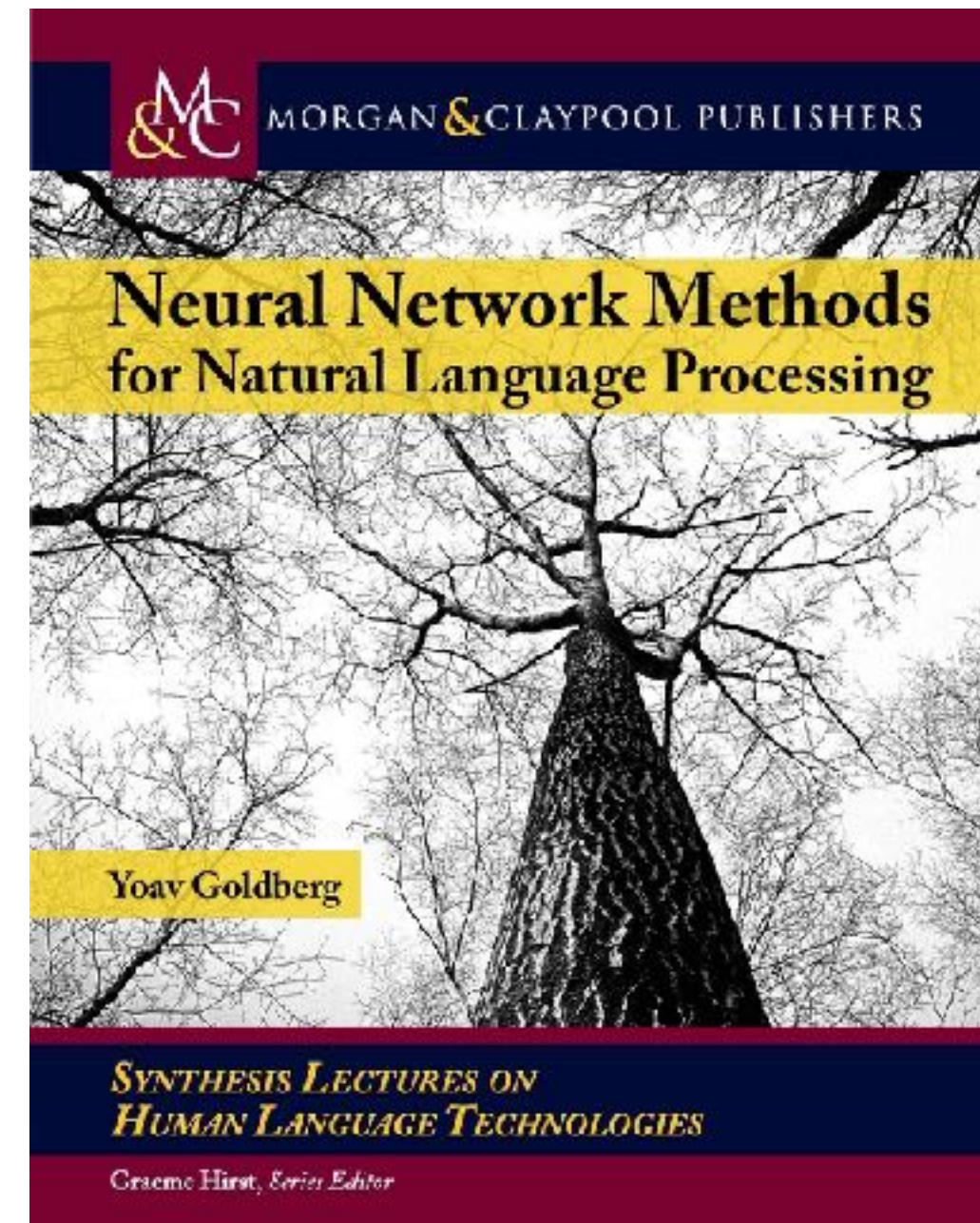
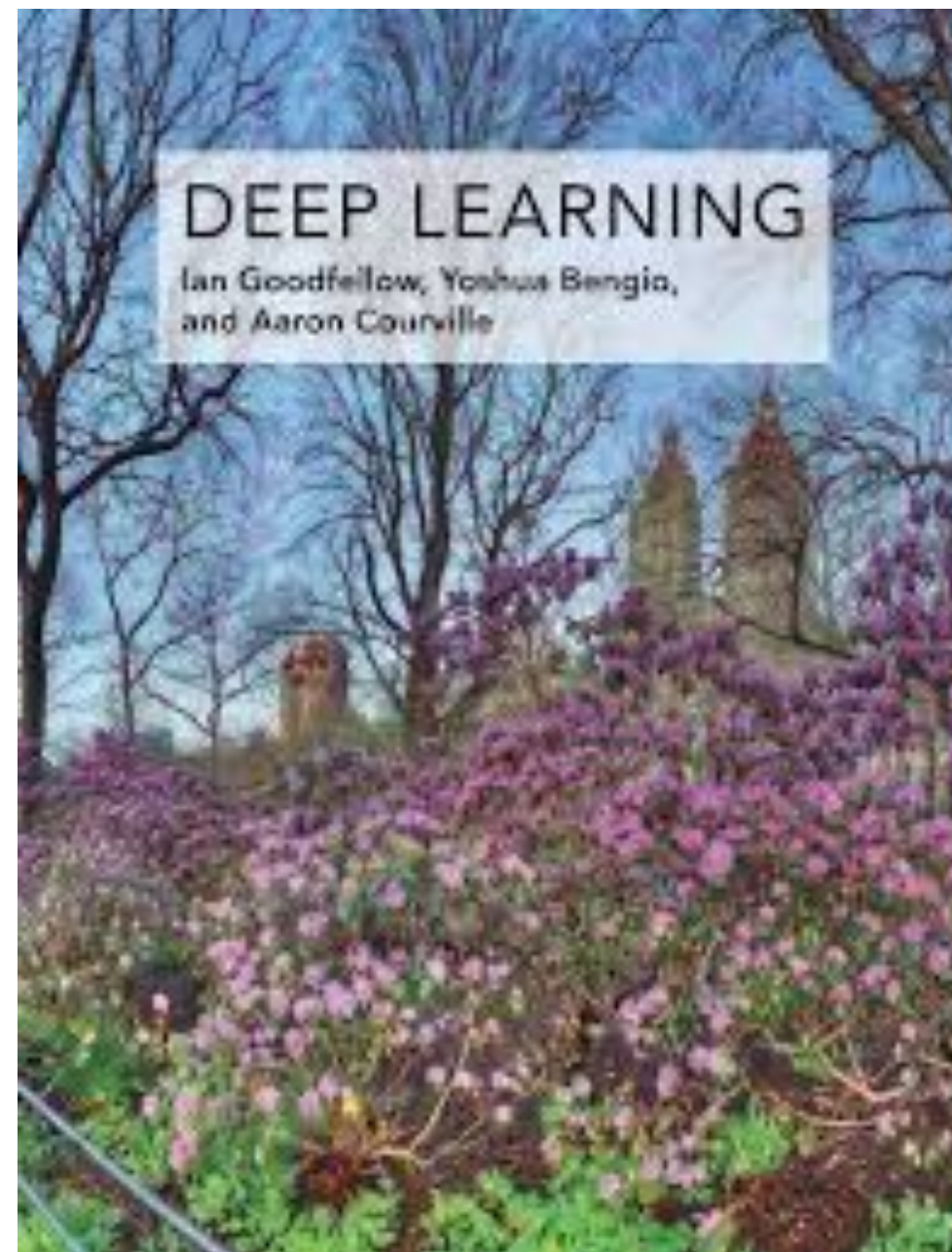


Academic Conferences

- ACL (Association of Computational Linguistics)
- EMNLP (Conference on Empirical Methods in Natural Language Processing)
- NAACL (The North American Chapter of the Association for Computational Linguistics)
- Coling (International Conference on Computational Linguistics)
- EACL (European Chapter of ACL)
- IJCNLP (International Joint Conference on Natural Language Processing)
- SIGIR (SIG Information Retrieval)
- TREC (Text REtrieval Conference)
- SIGKDD (ACM Special Interest Group on Knowledge Discovery in Data)
- WWW (The Web Conference)
- NeurIPS (Conference on Neural Information Processing Systems)
- AAAI (Association for the Advancement of Artificial Intelligence)
- ICML, CIKM, ICDM, etc. (See “AI Conference Deadlines”)

Research groups

- Google Brain, Google Research, DeepMind, Facebook AI, Microsoft Research, IBM Watson, Tencent AI, Baidu Research ...
- Mila, Vector Institute, Amii, AllenNLP ...
- Stanford, MIT, Harvard, Berkeley, CMU, Tsinghua University, Peking University, HIT, Fudan University, UCSB ...



- No textbook is required
- Check the class webpage for more information

Todo

- Check the class webpage: <http://www-labs.iro.umontreal.ca/~liubang/IFT%206289%20-%20Winter%202022.htm>
- **Start preparing your term project early:** build your team (up to 3 people), read the project proposal instructions

Next lecture: Deep Learning Basics

References

1. **NTU S-108 Applied Deep Learning, Spring 2020:** <https://www.csie.ntu.edu.tw/~miulab/s108-adl/syllabus>, lecture 1
2. **Stanford CS224n - Natural Language Processing with Deep Learning, Winter 2020:** <http://web.stanford.edu/class/cs224n/>, lecture 1
3. **Gatech CS-4650/7650 Natural Language Processing, Spring 2020:** https://www.cc.gatech.edu/classes/AY2020/cs7650_spring/, lecture 1
4. **UCAS (中国科学院大学) Natural Language Processing** by Yue Hu and Jing Yu, lecture 1

Thanks! Q&A

Bang Liu

Email: bang.liu@umontreal.ca

Homepage: <http://www-labs.iro.umontreal.ca/~liubang/>

Github: <https://github.com/BangLiu/>