



AIEI007: Natural Language Processing

L1: Introduction to NLP

Autumn 2024

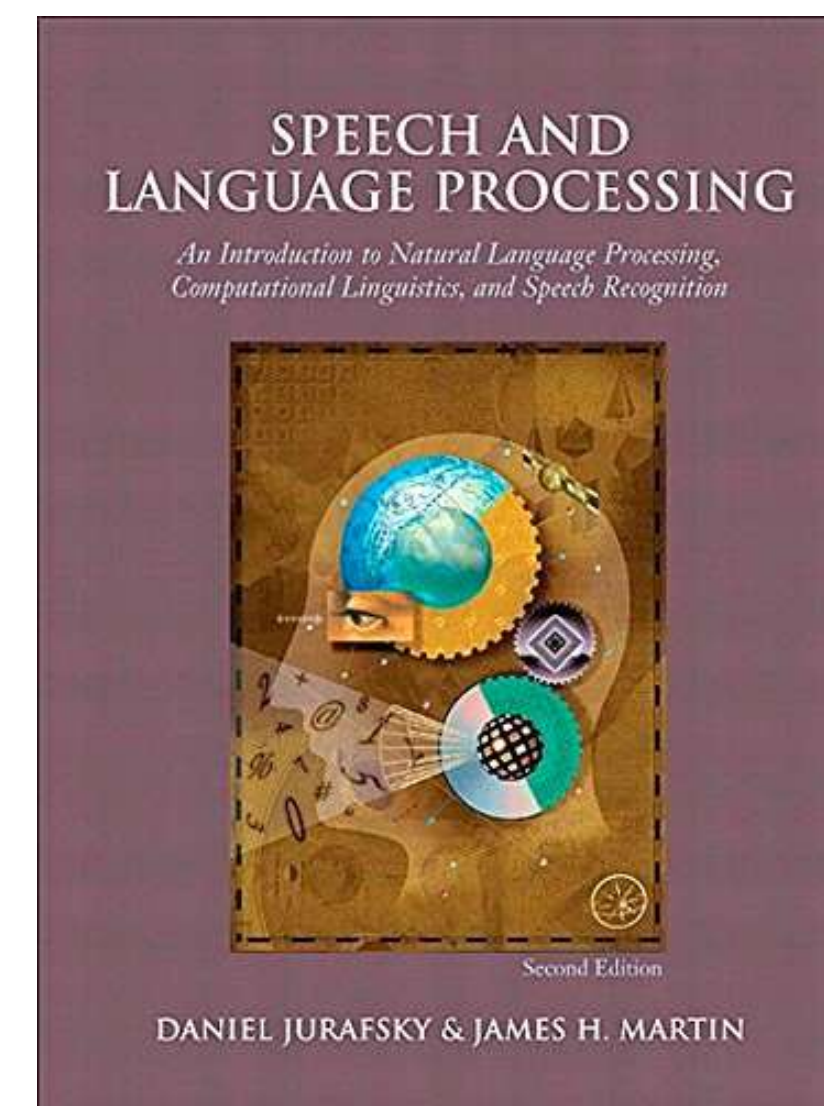
Textbooks

<https://web.stanford.edu/~jurafsky/slp3/>

Speech and Language Processing (3rd ed. draft)

[Dan Jurafsky](#) and [James H. Martin](#)

A good way to learn about state-of-the-art NLP concepts is through **research papers** and **blog posts**



Course goals



- Gain an understanding of the **fundamentals** of different sub-fields within NLP
- Understand **theoretical** concepts and algorithms
- **Hands on experience** building statistical models for language processing
- Carry out an **independent research project** at the end

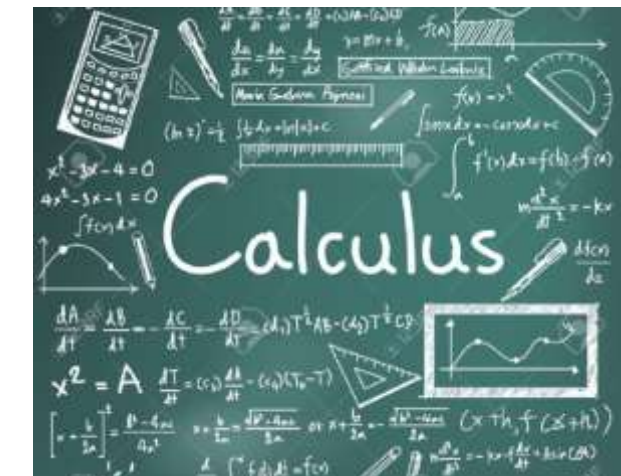
Prerequisites

- **Required:** AIE1001, AIE1006, knowledge of probability, linear algebra, calculus (A0 will give you a sense)
- Be ready to pick up new ML concepts
- Proficiency in Python: programming assignments and projects will require use of Python, Numpy and PyTorch.

Q. Why are AIE1001, AIE1006 prerequisites?

We assume you have learned the following concepts already:

- Language models
- Logistic regression w/ regularization
- Unsupervised vs supervised learning
- Feedforward neural networks, convolutional neural networks
- PyTorch programming
- (A little bit of reinforcement learning)



What is NLP and what is the course about

Natural Language processing

- NLP = building **computer programs** to analyze, understand and generate **human language - either spoken or written** (informal)
- NLP is an interdisciplinary field



Natural Language processing

- NLP = building **computer programs** to analyze, understand and generate **human language - either spoken or written** (informal)

Communication with humans (ex. personal assistants, customer service)

Access the wealth of information about the world — crucial for AI systems



Banking assistant

ONLINE

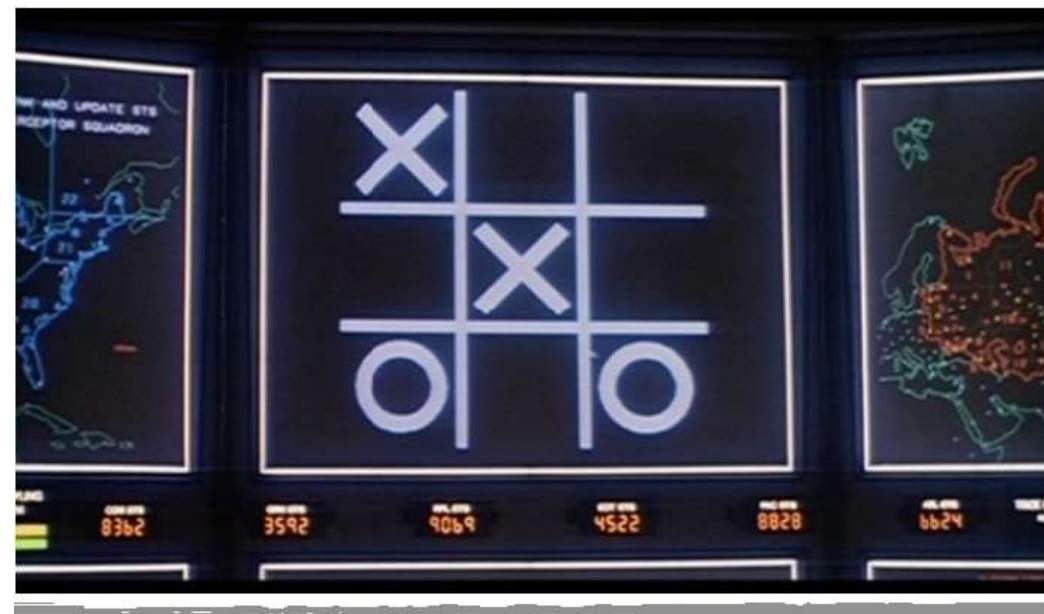


OFFLINE

Computer learns to play Civilization by reading the instruction manual

By Matthew Rogers on July 14, 2011 at 5:03 pm [16 Comments](#)

回 国 日 令 人 532 SHARES

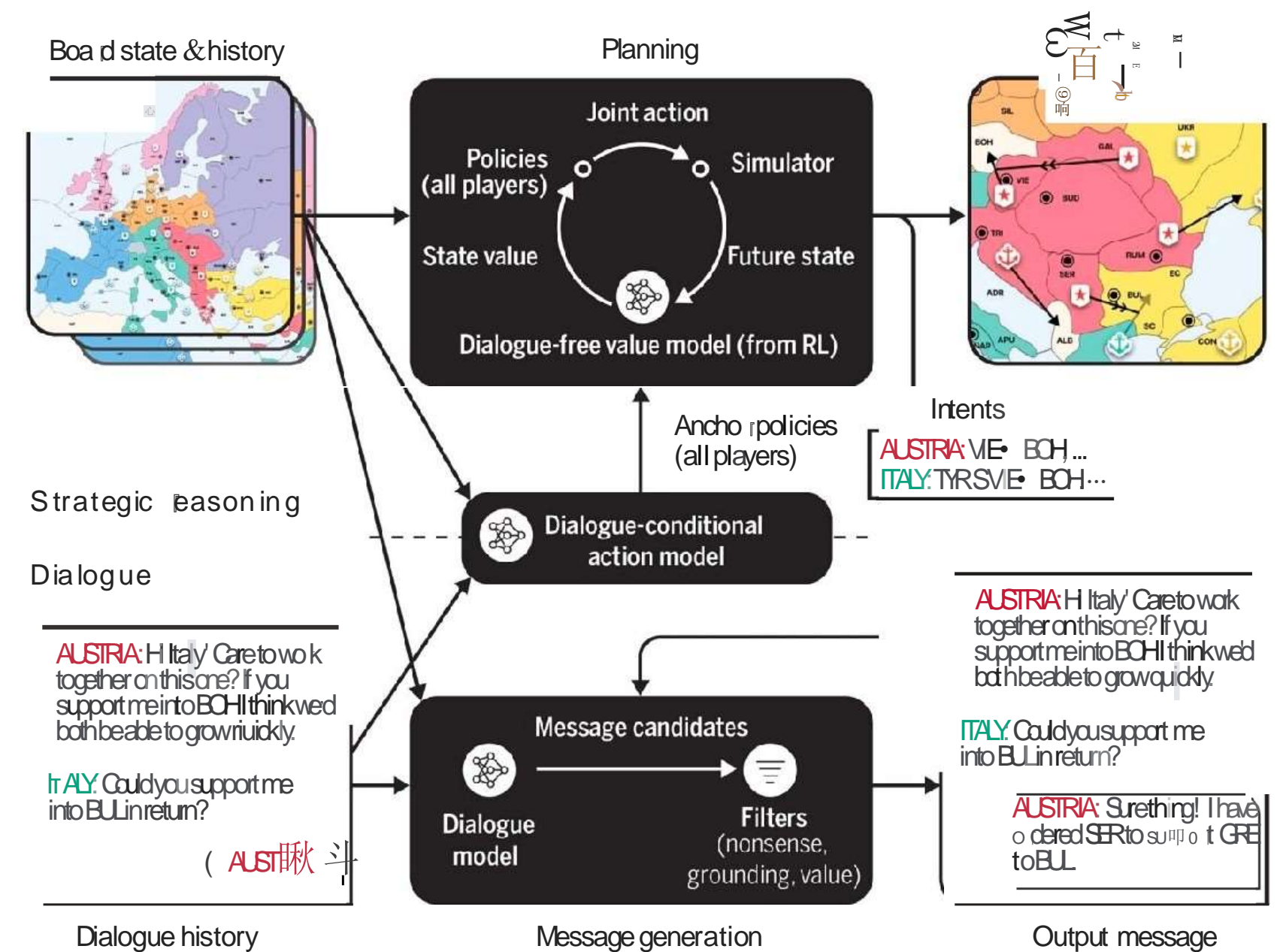


MIT researchers just got a computer to accomplish yet another task that most humans are incapable of doing: It learned how to play a game by reading the instruction manual.

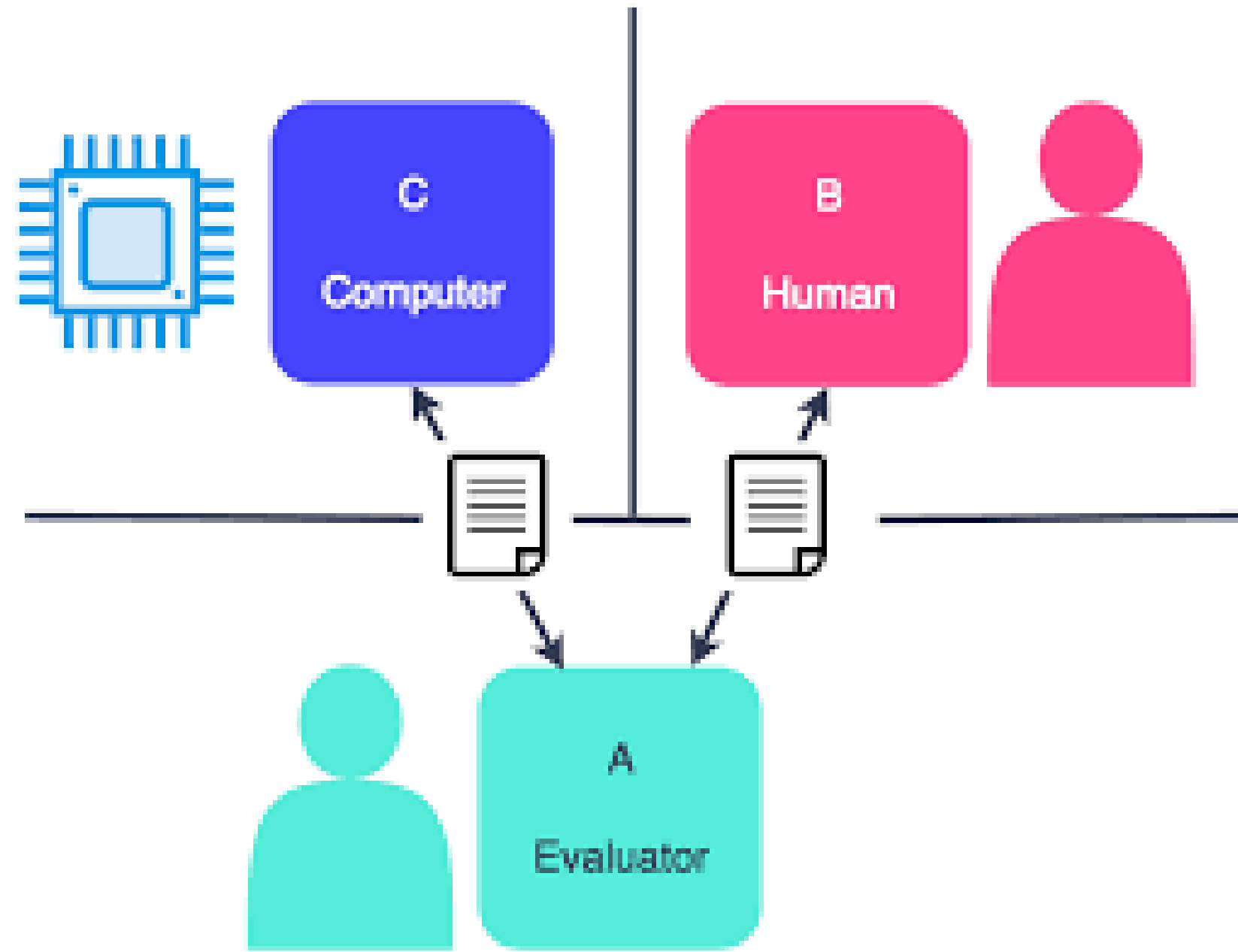
The MIT Computer Science and Artificial Intelligence lab has a computer that now plays Civilization

Meta's New AI Ranked in the Top 10% at the Game 'Diplomacy' and Human Players were None the Wiser

每静消 dd Gent > November 28, 2022



Turing Test

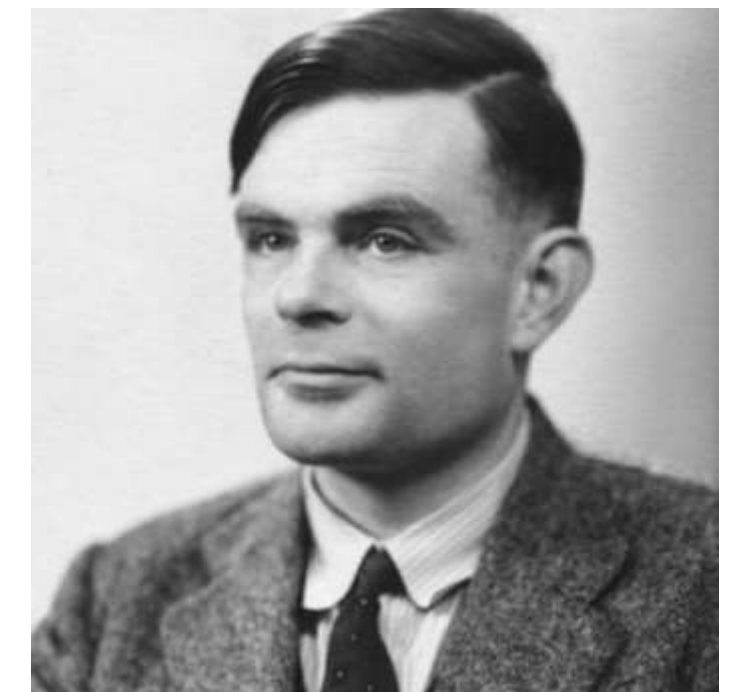


A. M. Turing (1950) *Computing Machinery and Intelligence*. *Mind* 49: 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

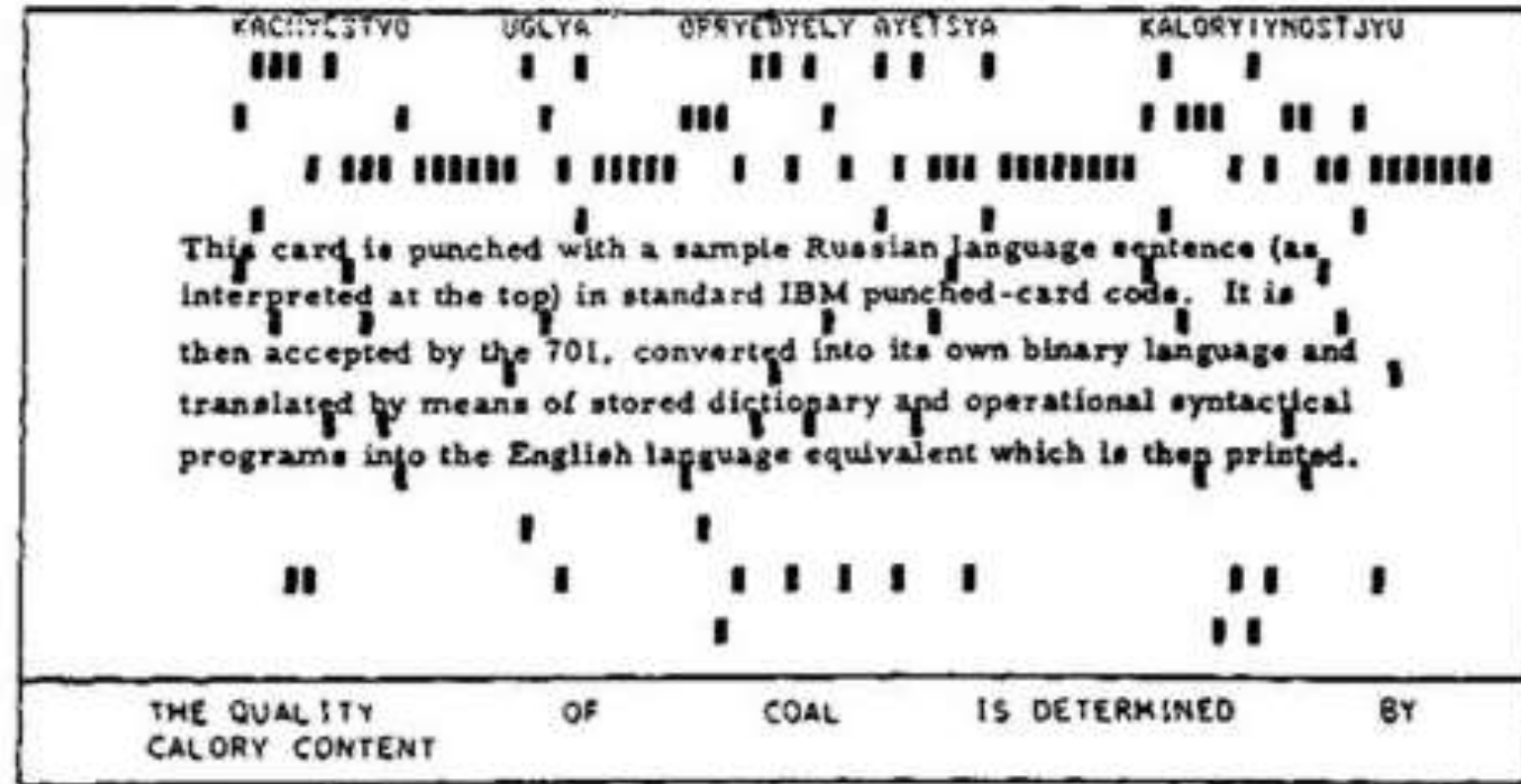
By A. M. Turing

1. The Imitation Game



Ability to understand and generate language ~ intelligence

How it started



Specimen punched card and below a strip with translation, printed within a few seconds

Georgetown experiment 1954

“Within three or five years, machine translation will be a solved problem”

How it's going

DETECT LANGUAGE YORUBA **ENGLISH** MALAY

The man shot the elephant while wearing his pyjamas.

52 / 5000

HINDI YORUBA ENGLISH

उस आदमी ने हाथी को अपना पजामा पहनाया।

us aadamee ne haathee ko apana pajaama pahanaaya.

DETECT LANGUAGE **HINDI** ENGLISH YORUBA

उस आदमी ने हाथी को अपना पजामा पहनाया।

us aadamee ne haathee ko apana pajaama pahanaaya.

37 / 5000 अ

HINDI YORUBA **ENGLISH**

The man put his pajamas on the elephant.

Text Documents Websites

DETECT LANGUAGE HINDI **ENGLISH** KOREAN

↔ ENGLISH **HINDI** KOREAN

The man shot the elephant while wearing his pajamas

आदमी ने अपना पजामा पहने हुए हाथी को गोली मार दी

aadamee ne apana pajaama pahane hue haathee ko golee maar dee

51 / 5,000

Text Documents Websites

DETECT LANGUAGE ENGLISH **HINDI** KOREAN

↔ HINDI **ENGLISH** KOREAN

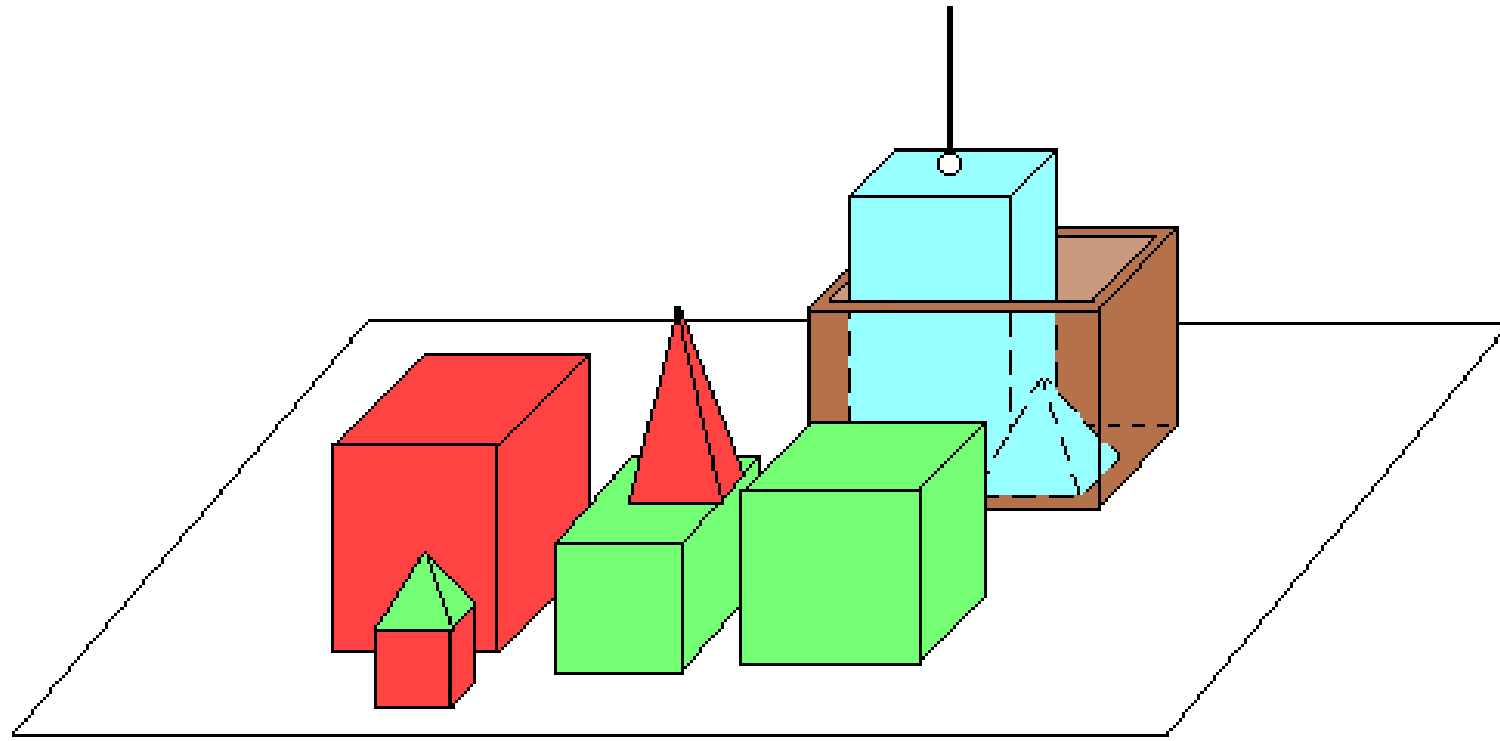
आदमी ने अपना पजामा पहने हुए हाथी को गोली मार दी

Man shoots elephant wearing his pajamas

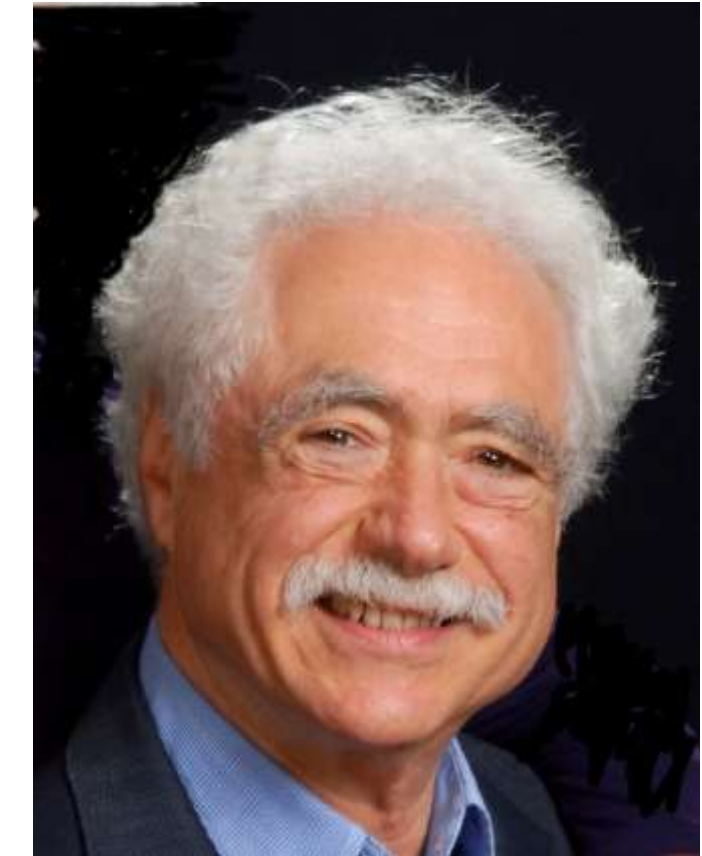
aadamee ne apana pajaama pahane hue haathee ko golee maar dee

47 / 5,000 अ

(Based on Google Translate result in 2023-01)



SHRDLU,
1968



> How many red
blocks are there?

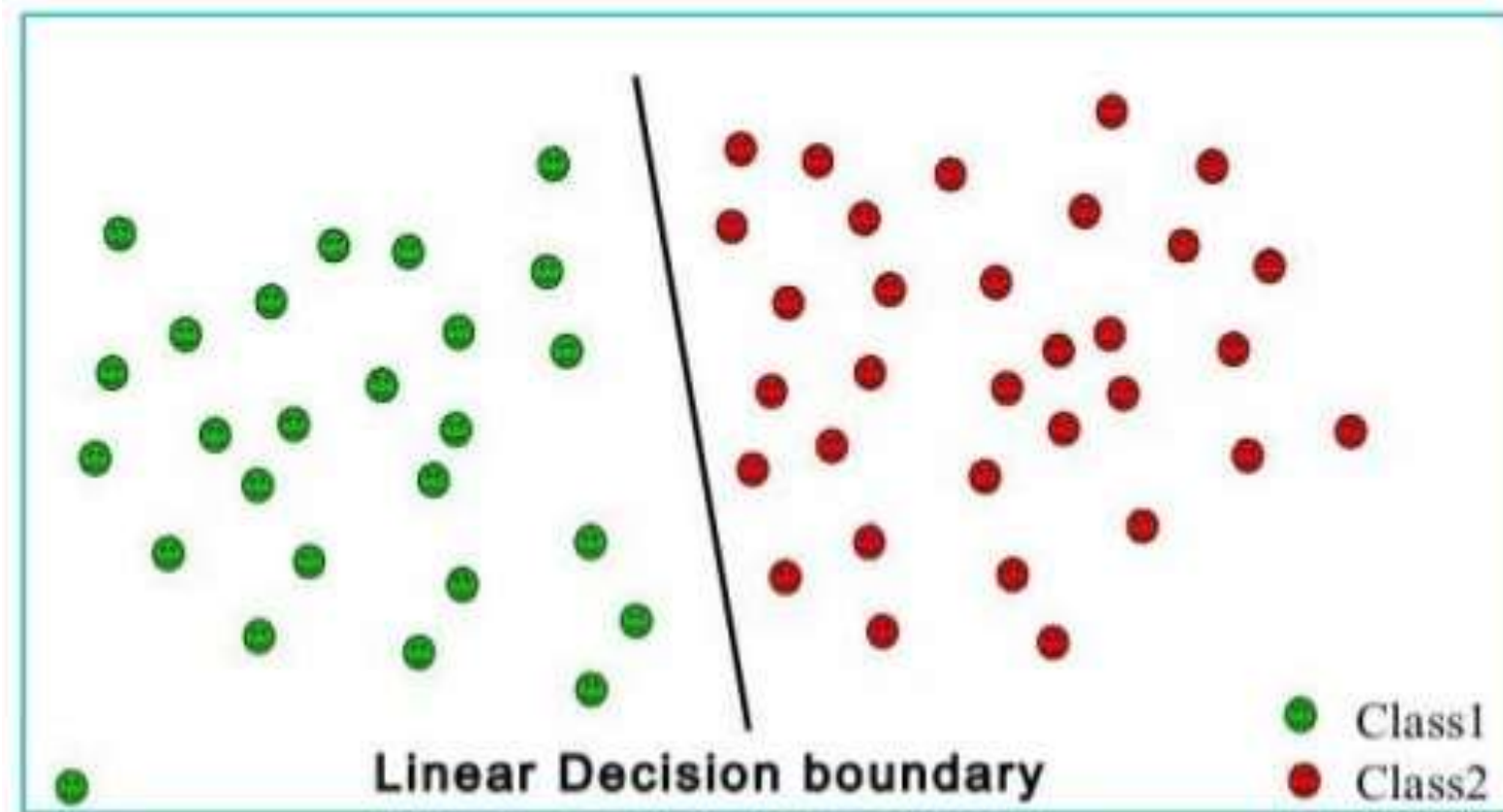
- THREE OF THEM

> Pick up the red
block on top of a
green one

OK.

- Rule-based, requiring extensive programming
- Limited domain

Statistical learning

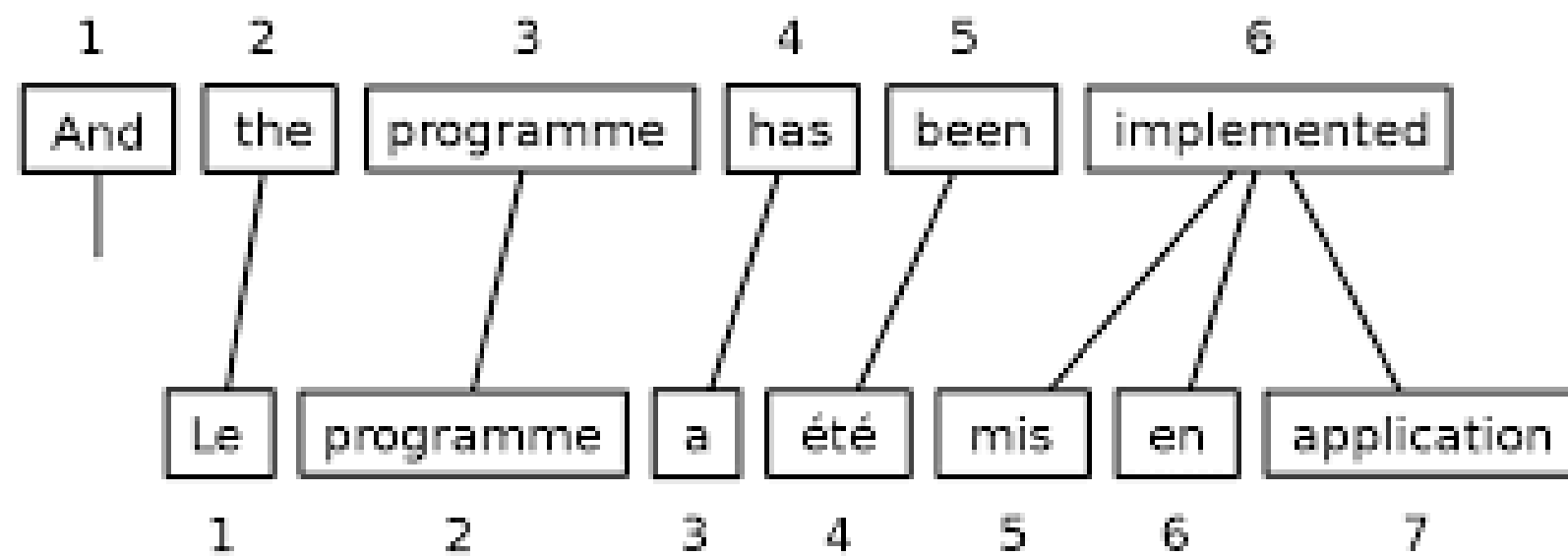


- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora

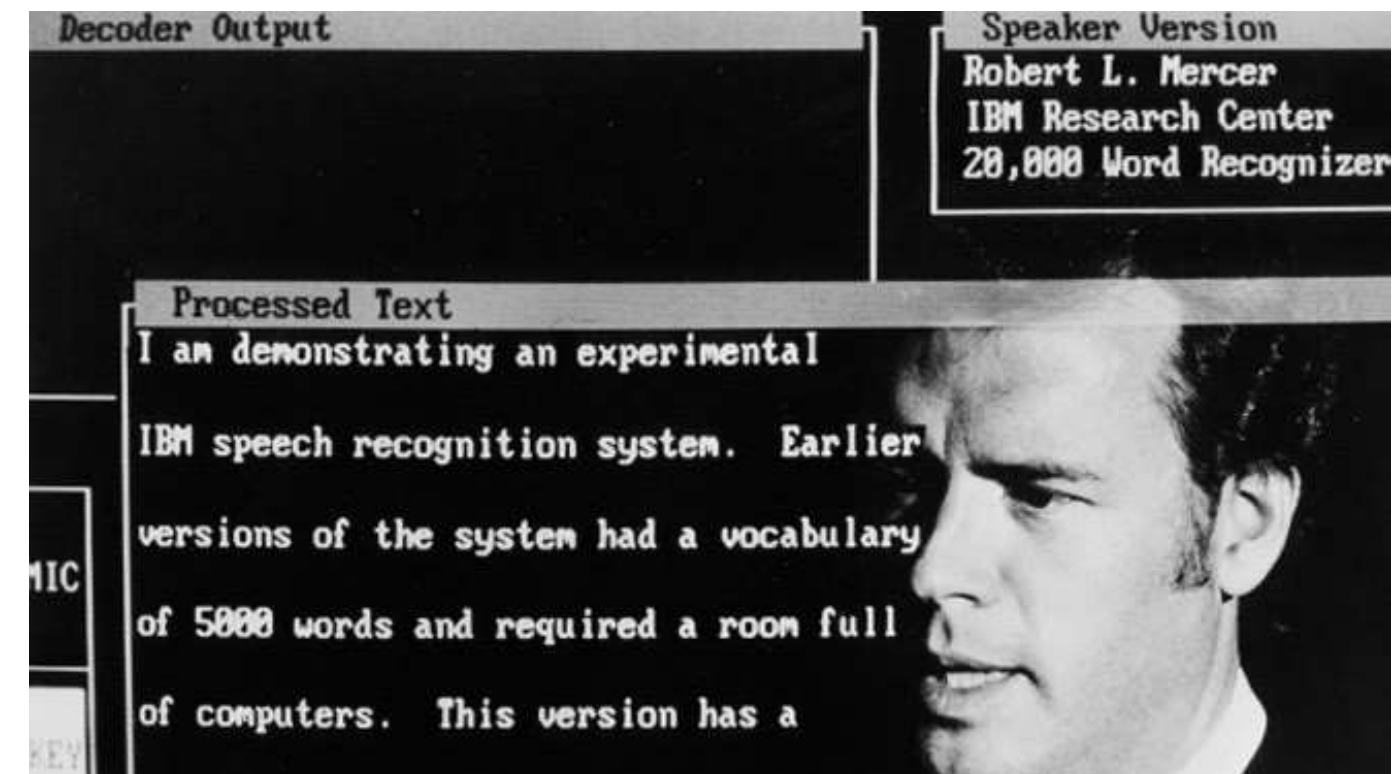
Unsupervised vs. supervised?

Statistical learning

IBM translation models



Speech recognition



Anytime a linguist leaves the group the (speech) recognition rate goes up
- Fred Jelinek 1998

The era of deep learning

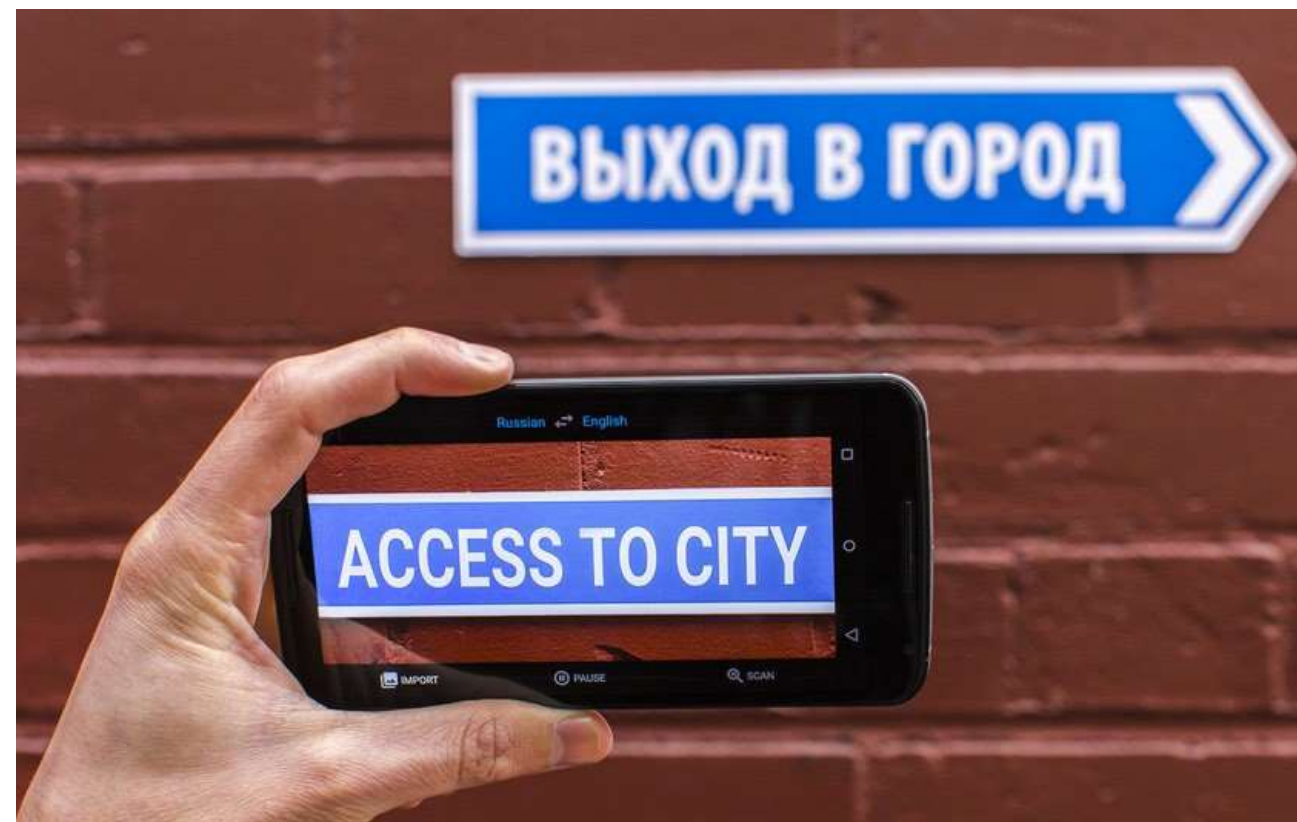
- Significant advances in core NLP technologies
- **Essential ingredient:** large-scale supervision, lots of compute
- Reduced manual effort - less/zero **feature engineering**



GPU



TPU



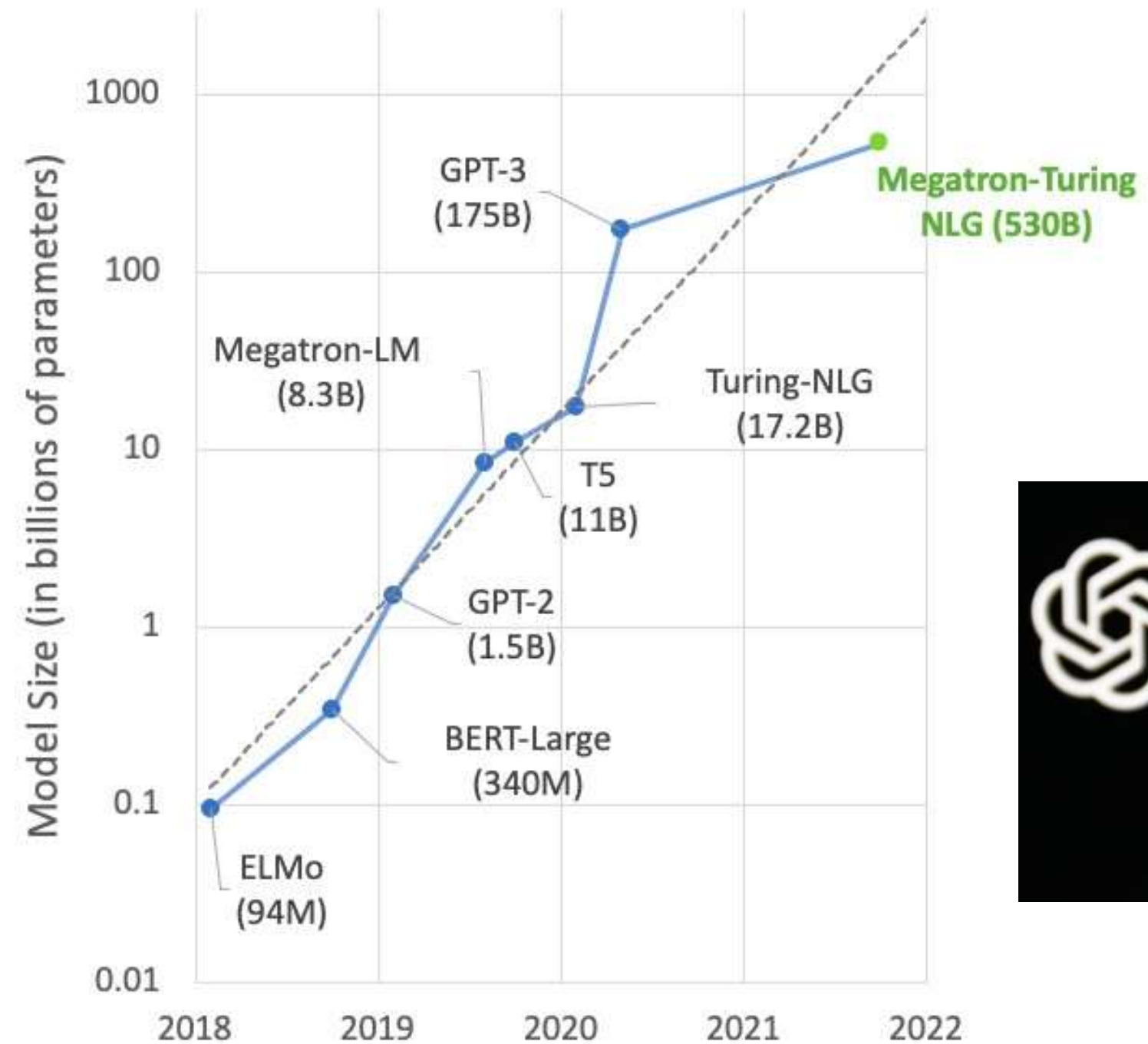
36M sentence pairs

Russian: Машинный перевод - это круто!



English: Machine translation is cool!

The era of pre-training / LLMs



BERT, ELMo, ERNIE...

- Leverages a lot of unlabeled text
- Model size increased by $10^3 - 10^5$ x in parameters

ChatGPT examples



What NLP topics should I teach in 2023?



Here are some NLP topics that are popular and relevant in 2023:



1. Transformers (BERT, GPT, etc.)
2. Transfer Learning in NLP
3. Multi-lingual NLP
4. Attention Mechanisms
5. Adversarial Training in NLP
6. NLP for Question Answering
7. Sentiment Analysis
8. Text Generation (summarization, translation, etc.)
9. Named Entity Recognition
10. Dialogue Systems

Note: The field of NLP is constantly evolving, so it's important to stay up-to-date with recent advances and adjust your teaching content accordingly.

ChatGPT examples



what do you know about the Princeton NLP group?



The Princeton NLP group is a research group at Princeton University that focuses on the study of natural language processing (NLP) and its applications. The group is led by Professor Christopher Manning, who is a leading expert in the field of NLP. The group has published numerous research papers and has developed several widely-used NLP tools, including the Stanford Parser and Stanford Named Entity Recognizer. The group also collaborates with other research groups and organizations, such as the Allen Institute for Artificial Intelligence and Google, on NLP research projects.



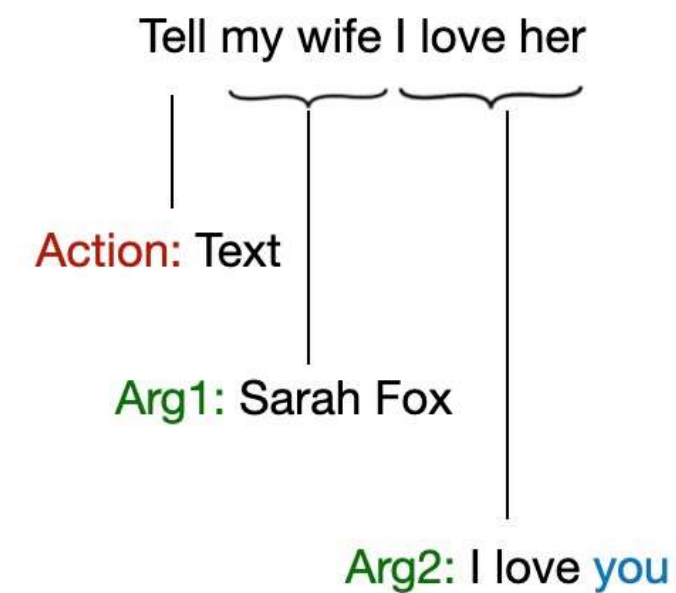
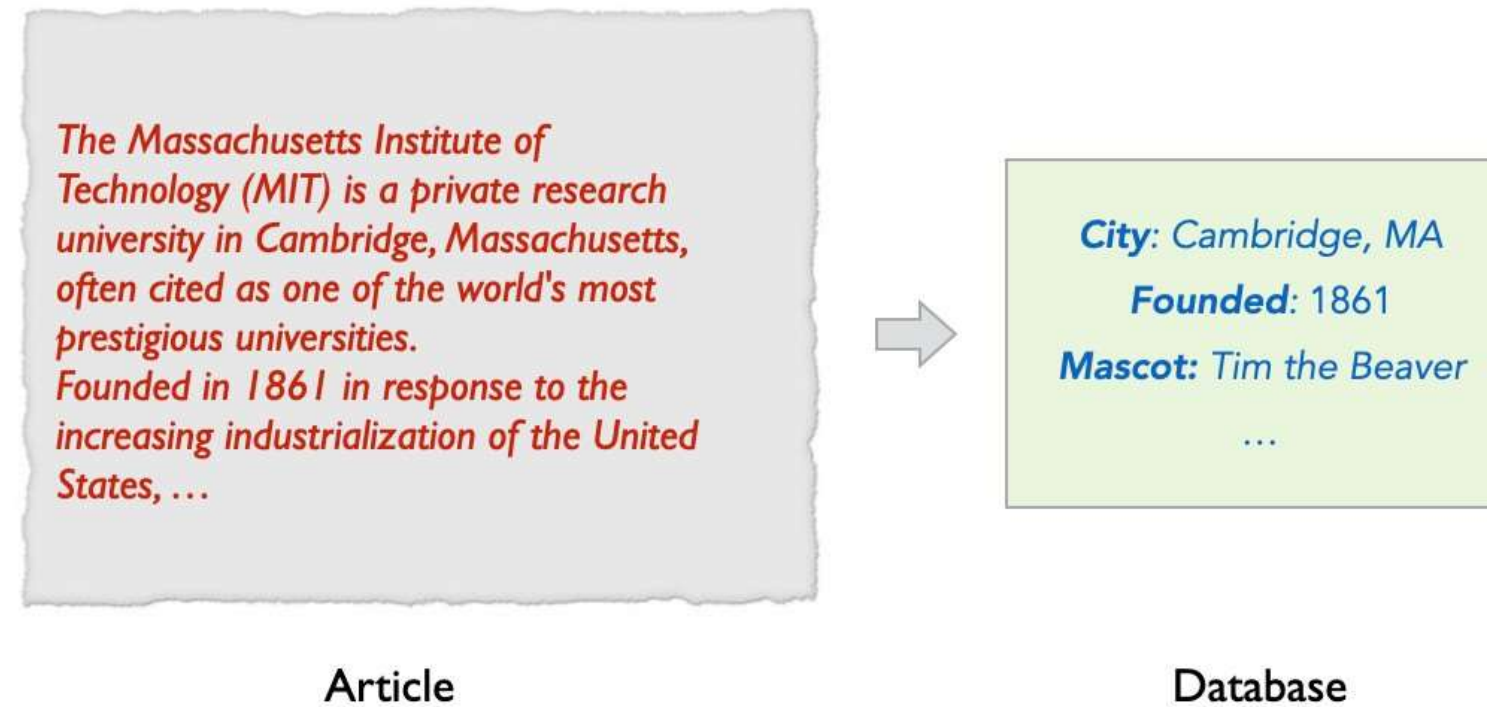
Why do you hallucinate so much?



As an AI language model, I don't have the capacity to hallucinate as I don't possess emotions, sensations, or consciousness. I can only generate responses based on patterns in the text I was trained on.

NLP subfields

- Machine translation
- Question answering
- Information extraction
- Text summarization
- Dialogue
- Semantic parsing
- Syntactic parsing
- Language grounding
-



Topics we will cover in this class

- We will focus on fundamentals, important concepts rather than concrete applications
- The class will be a mix of statistical and neural NLP approaches (>70% neural)
 - N-gram language models
 - Text classification
 - Word embeddings × 2
 - Sequence models × 2
 - Neural networks for NLP
 - Recurrent neural networks × 2

- (neural) machine translation × 2
- **Transformers** × 2
- BERT, ELMo, ...
- **Large language models**

+ guest
lectures!

Before midterm

After midterm

Why is language difficult to understand?

Why is language difficult to understand?

- Ambiguous
- Dialects
- Accents
- listener has to infer - pragmatics
- humor, sarcasm, irony
- context, dependencies

Lexical ambiguity

The fisherman went to the *bank*.

bank¹

/bɑŋk/ 

noun

plural noun: **banks**

1. the land alongside or sloping down to a river or lake.

"willows lined the bank"

synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe; [More](#)

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.

"I paid the money straight into my bank"

synonyms: financial institution, merchant bank, savings bank, finance company, trust company,

One word can mean several different things

Lexical ambiguity

The fisherman went to the *bank*. He deposited some money.

bank¹

/bɑŋk/ 

noun

plural noun: **banks**

1. the land alongside or sloping down to a river or lake.

"willows lined the bank"

synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe; [More](#)

1. a financial establishment that invests money deposited by customers, pays it out when required, makes loans at interest, and exchanges currency.

"I paid the money straight into my bank"

synonyms: financial institution, merchant bank, savings bank, finance company, trust company,

Word sense disambiguation

Lexical variations

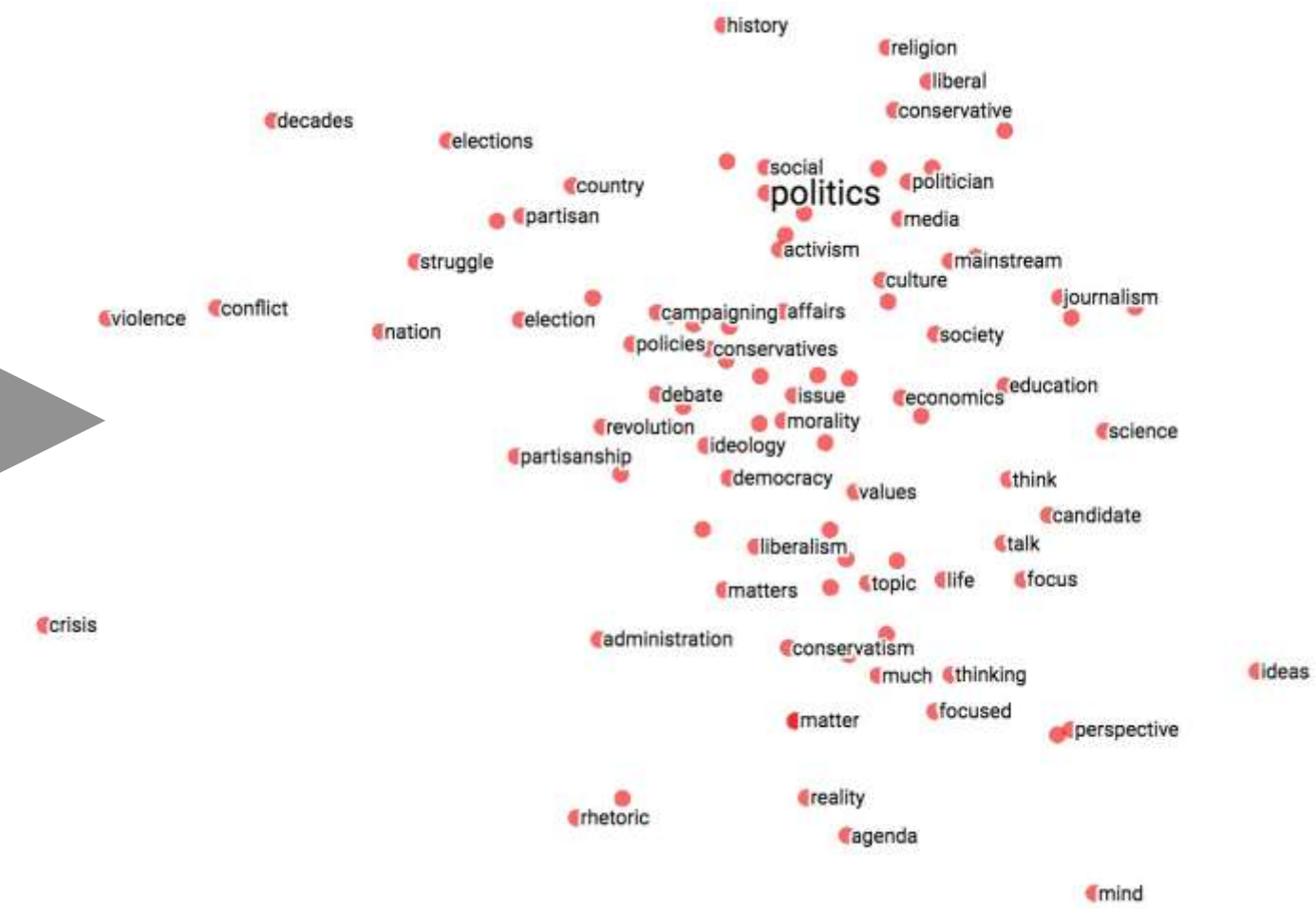


**ACCORDING TO THE THESAURUS,
"THEY'RE HUMID, PREPOSSESSING
HOMOSAPIENS WITH FULL SIZED AORTIC
PUMPS" MEANS "THEY'RE WARM, NICE
PEOPLE WITH BIG HEARTS."**

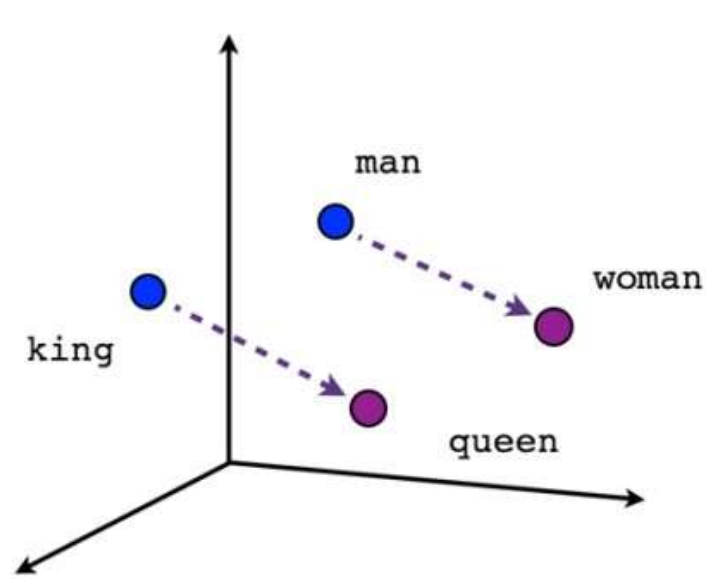
Several words can mean the same thing!

Distributed representations

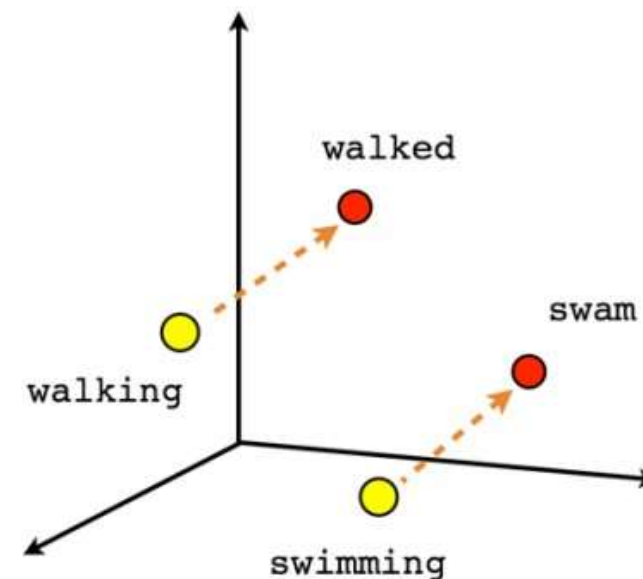
Project words onto a continuous vector space



Similar words closer to each other



Male-Female



Verb tense

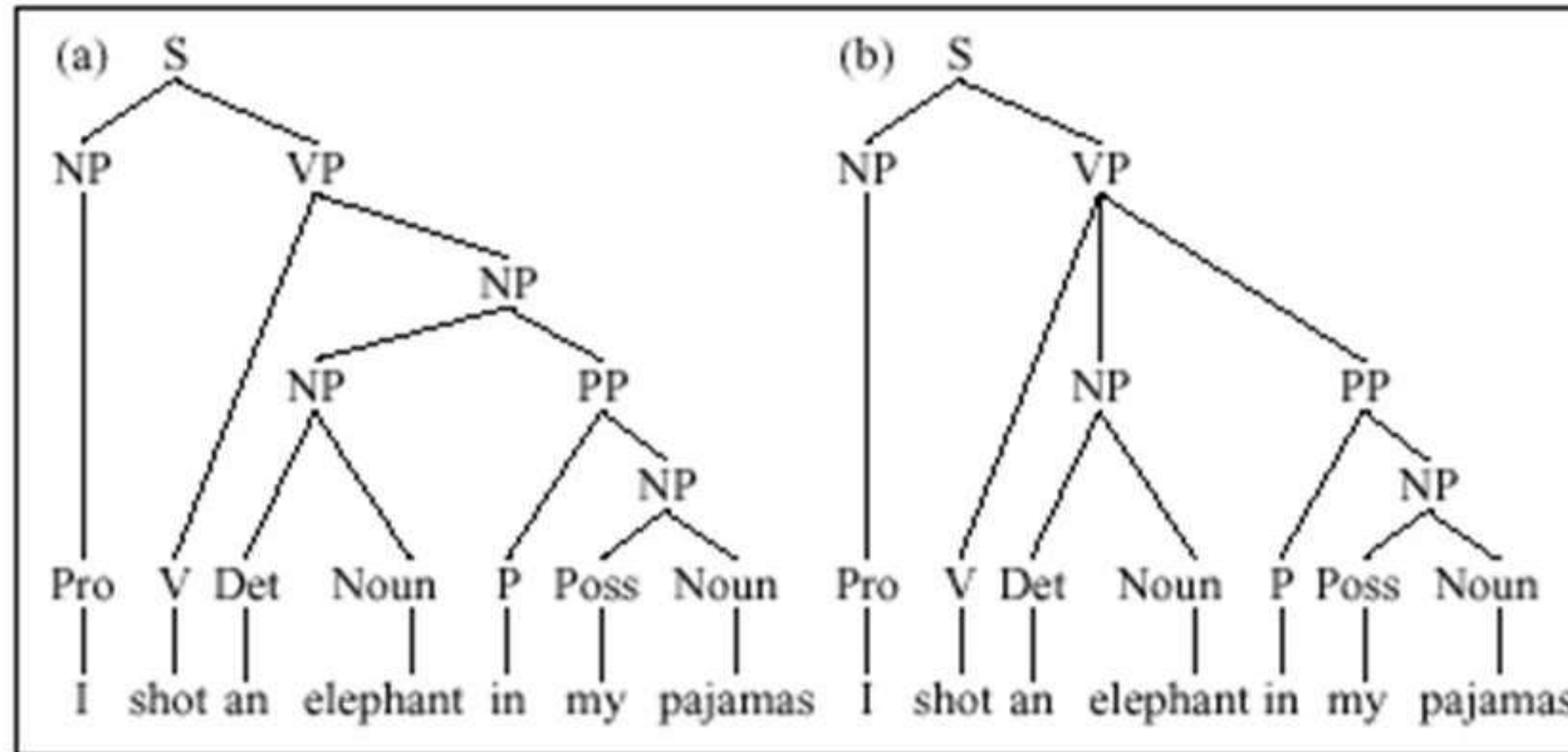
$$v(\text{king}) + v(\text{man}) - v(\text{woman}) = v(\text{queen})$$

Comprehending word sequences

- My brother went to the park near my sister's house
 - Park my went house near to sister's my brother the
 - "My brother went park near sister's house"?
 - The old man the boat
 - Fat people eat
accumulates **Garden**
Path sentence
- Implicit structure in all languages
 - Coarse-to-fine levels (recursive)
 - What are some good data structures to represent this?

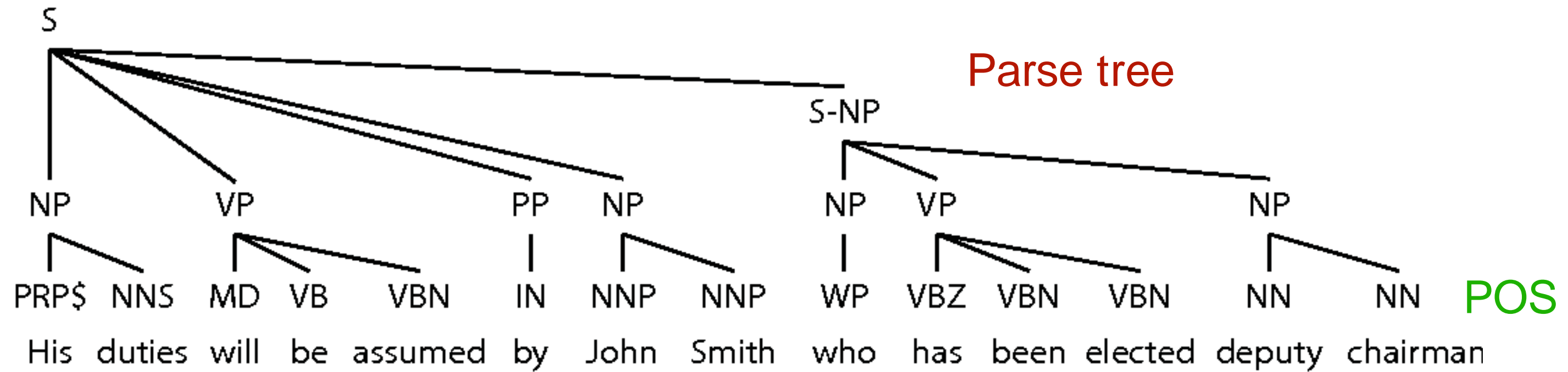
Syntactic ambiguity

I shot an elephant in my pajamas



Human language is full of such examples!

Syntactic parsing



Online tools: <http://nlp.stanford.edu:8080/corenlp/>

Discourse ambiguity

- The man couldn't lift his son because **he** was so **heavy**.
- The man couldn't lift his son because **he** was so **weak**.

What does "he" refer to?

- The city councilmen refused the demonstrators a permit because **they** **feared** violence.
- The city councilmen refused the demonstrators a permit because **they** **advocated** violence.

What does "they" refer to?

Anaphora resolution

Next: n-gram language models