



# COS 484: Natural Language Processing

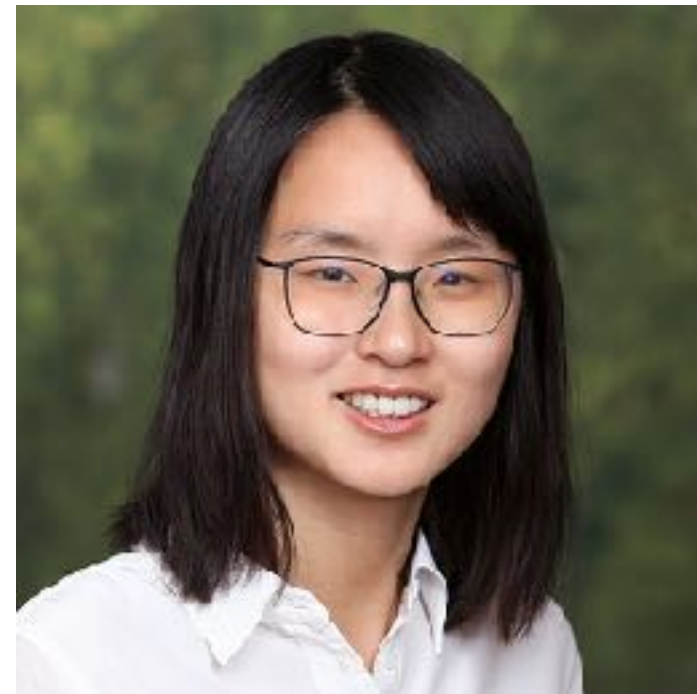
## LI: Introduction to NLP

Spring 2025

# Logistics and course structure

# Course staff

## Instructors



Danqi Chen



Tri Dao



Vikram Ramaswamy

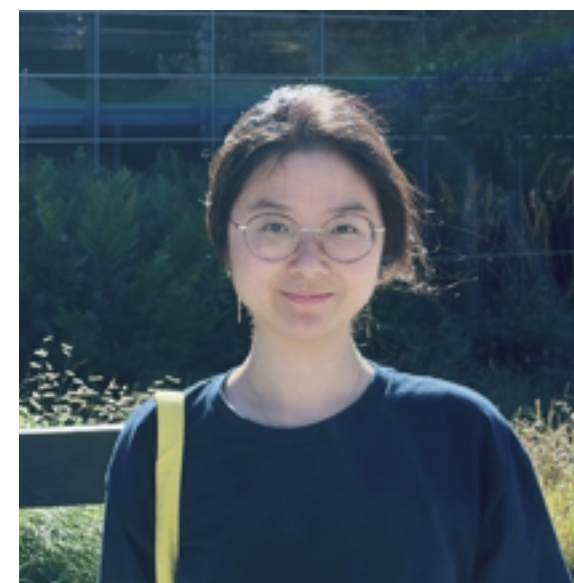
## Graduate TAs



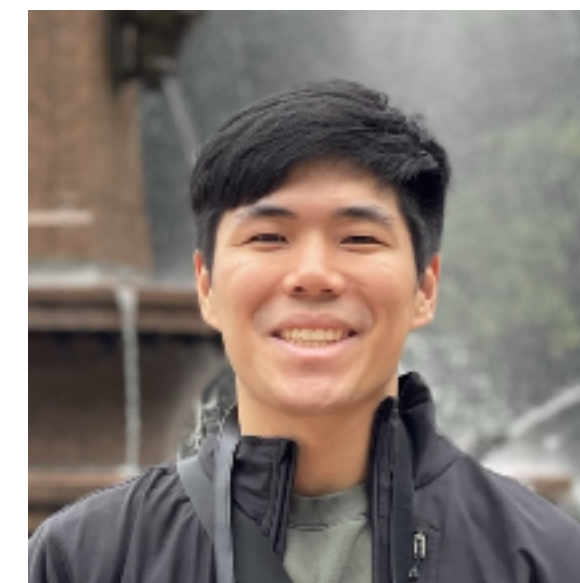
Adithya  
Bhaskar



David Braun



Catherine  
Cheng



Simon Park



Colin Wang



Tyler Zhu

# Website

**Course webpage:** <https://nlp.cs.princeton.edu/cos484/>

- Contains all the detailed information about the course (slides, reading lists, assignments, policy, office hours, etc).
- Ed will be used for all announcements - make sure you have **notifications** turned on!

## Schedule

Lecture schedule is tentative and subject to change. All assignments are due at **11:59pm EST** before the lectures.

All readings are optional but can help strengthen your understanding

Week	Date	Topics	Readings	Assignments
1	Mon (1/27)	Introduction to NLP	1. <a href="#">Advances in natural language processing</a> 2. <a href="#">Human Language Understanding &amp; Reasoning</a>	A0 out
	Wed (1/29)	n-gram language models	<a href="#">J &amp; M 3.1-3.5</a>	
	Fri (1/31)	Precept 1		
2	Mon (2/3)	Text classification	<a href="#">Naive Bayes: J &amp; M 4.1-4.6</a> <a href="#">Logistic regression: J &amp; M 5.1-5.8</a>	A0 due, A1 out

All assignments will be due on Monday at 12pm before the lecture

# Course Meetings

- **Lectures:** Monday/Wednesday 3 - 4:20pm at Friend 101
- **Precepts:** 1-hour precept every week taught by TAs (optional)
  - Fridays, either 10 - 10:50am or 11-11:50am at CS105
  - Refresher for basic maths, review course materials, Colab/Pytorch tutorial etc



# Contact Policy

- Sign up for Ed and Gradescope today!
- We'll use iClicker for in class polls
- Ed will be the main (and only) forum for all class-related questions and discussion.
- You can use private Ed posts for personal matters.
- No email please
- We would like to help in our office hours!
- All the information will be posted this week on the website. We will also have a number of undergraduate TAs' office hours

Go to

[join.iClicker.com](https://join.iClicker.com)

**HXFY**



# Assignments

## Assignments (40%): 5 total

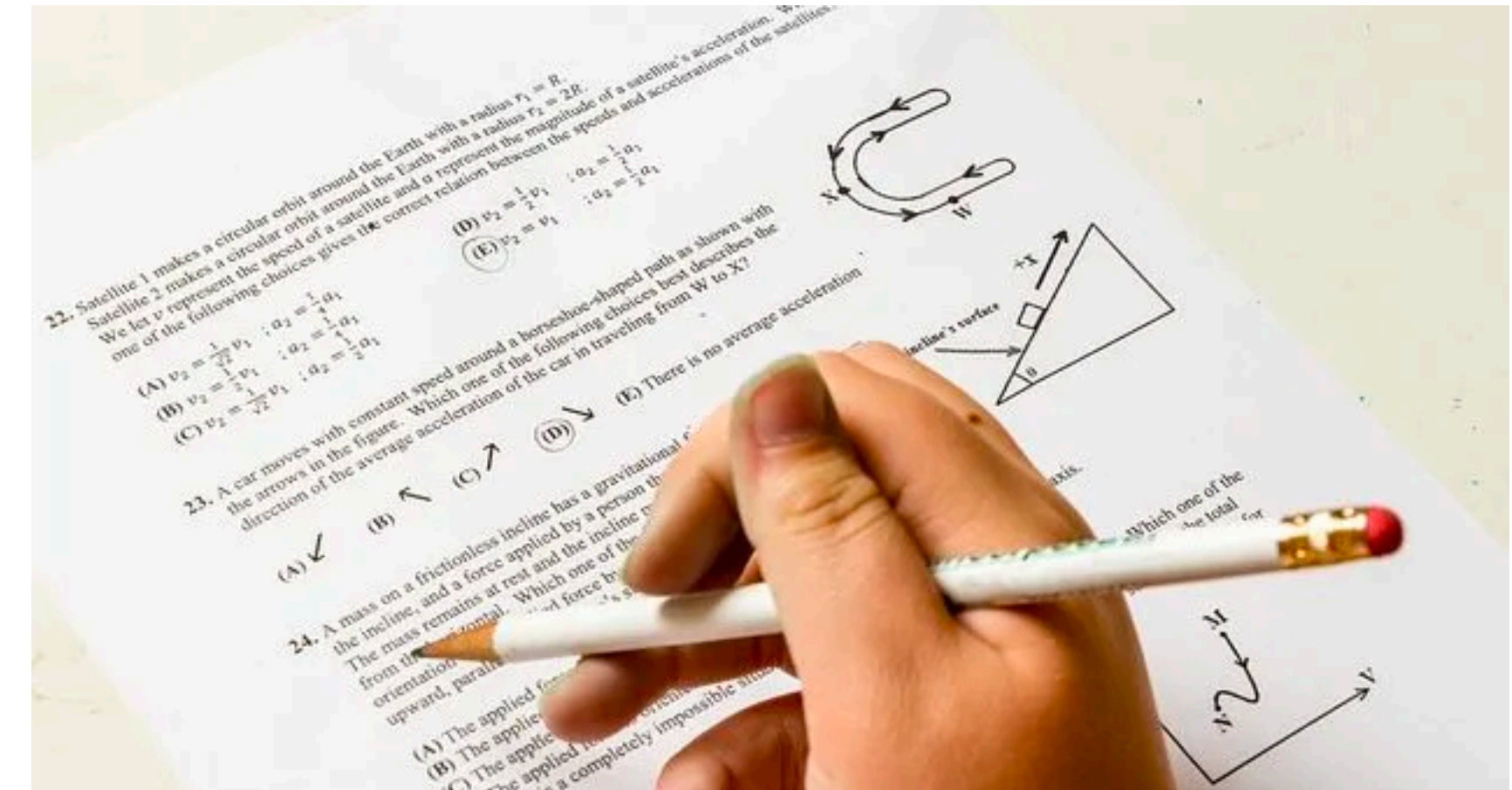
- **A0 (4%)** will be released TODAY and due next Monday
- **A1, A2, A3, A4**: 9% each - each assignment has 2 or 3 weeks
- Every assignment has a **written** component and a **programming** component based on Colab (A4 requires access to GPUs).
- You will need to know how to program in PyTorch/Numpy.

You have **4 free late days** for all assignments; After that, 10% penalty for each late day (up to a maximum of 3 days beyond which submissions will not be accepted)



# Midterm

- Midterm (25%)
  - A 3-hour timed exam on Gradescope in a 36-hour window (**March 5-6**)
  - All the topics up to and including Feb 26 will be covered
  - No final exam





# Final Project

## Final project (35%)

- Complete in a team of 3
- Two options: (a) reproducing a state-of-the-art NLP paper (ACL/EMNLP/NAACL 2021-2022) and coming up with your analysis, ablations or innovations (**strongly encouraged**) (b) competing a research project (get permission)
- Proposal (0%) due before - date will be announced soon
- Poster presentation tentatively scheduled on May 5 (during reading period)
- Final report due on May 11, 3:30 pm

**Extra bonus (5%)** - participation in class and Ed discussions

*No **pre-determined cut-offs for final grades**, will be decided at the end taking into account the performance of the entire class and will be fairly assigned to measure your level of understanding of the subject.*

# Textbooks

(NLP is a rapid-moving field...)

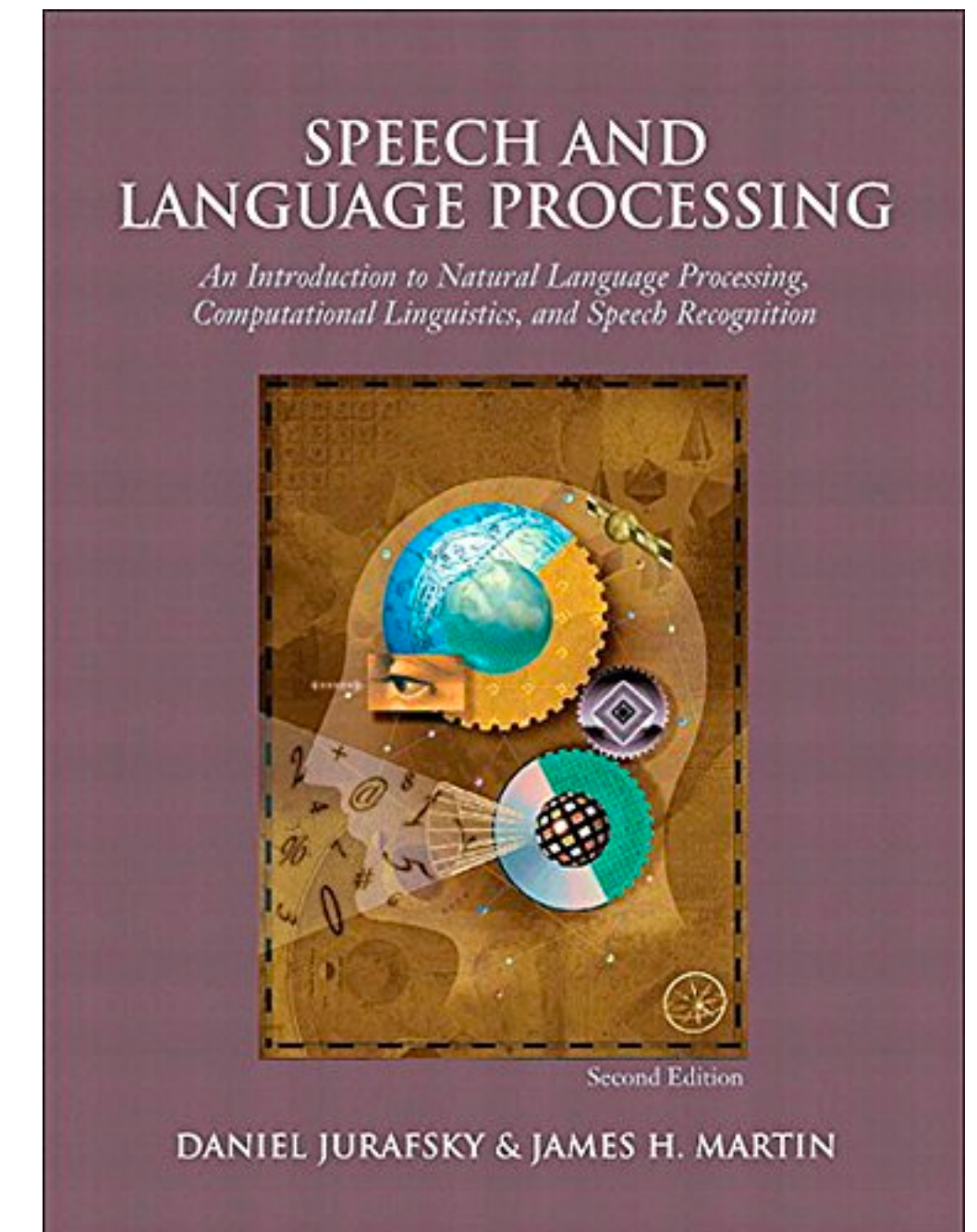
<https://web.stanford.edu/~jurafsky/slp3/>

## Speech and Language Processing (3rd ed. draft)

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

 Here's our **Jan 7, 2023 draft!** This draft is mostly a bug-fixing and restructuring release, there are no r the applications section earlier, reflecting how we and others tend to teach NLP, and combines the linguisti

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

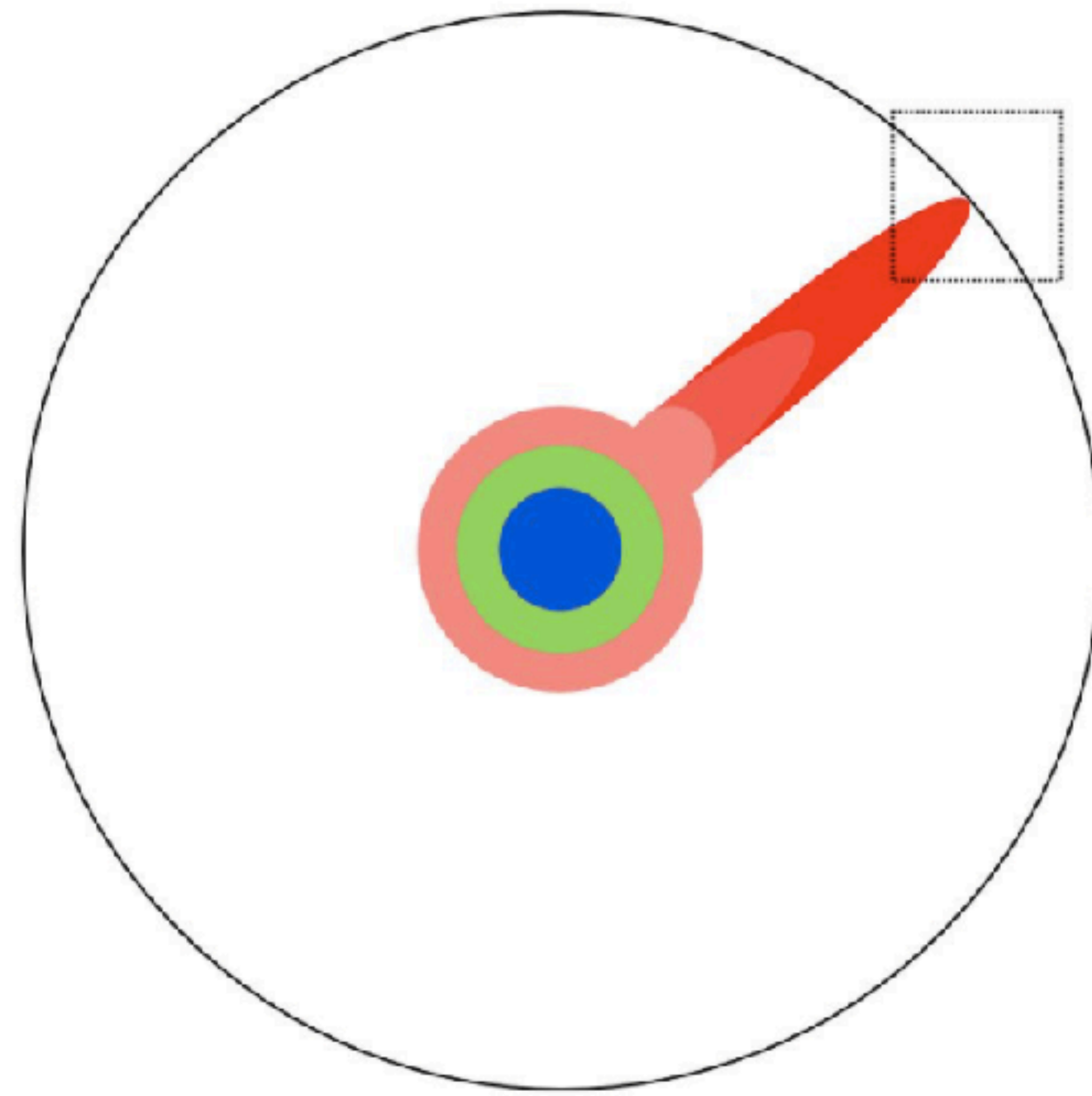


  
2nd edition ✘

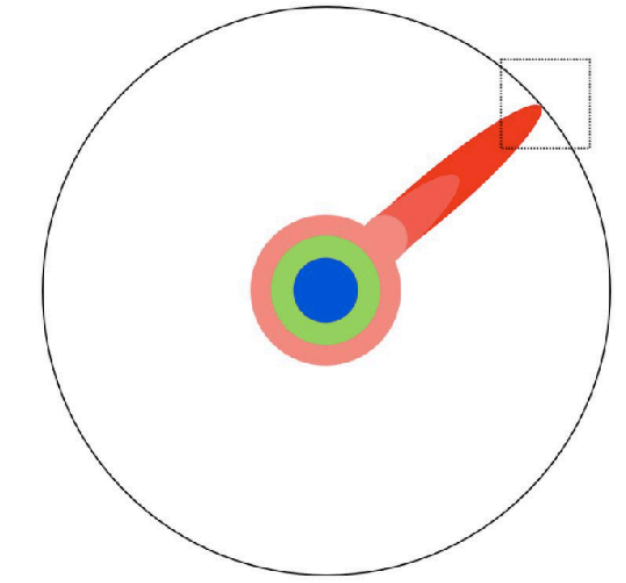
# Course goals



- 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



This is an advanced class



This is a *very* advanced class

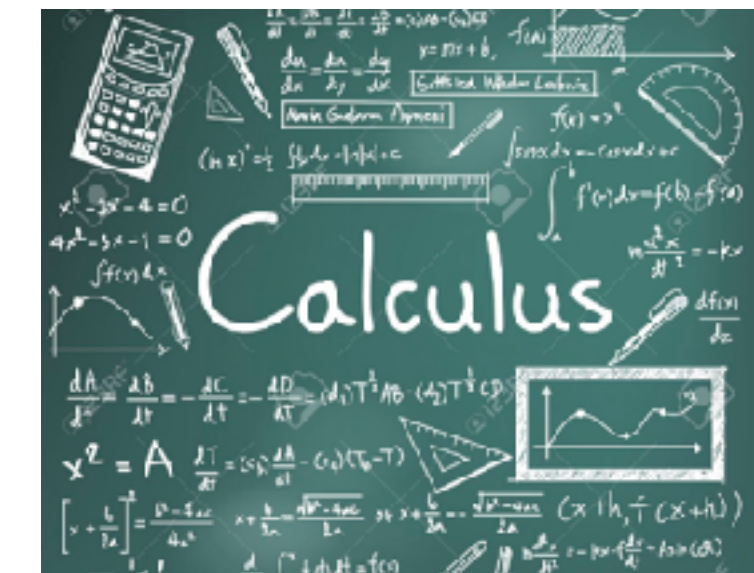
# Prerequisites

- **Required: COS324**, 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 is COS324 a prerequisite?

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 anyway?

# Natural Language Processing

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



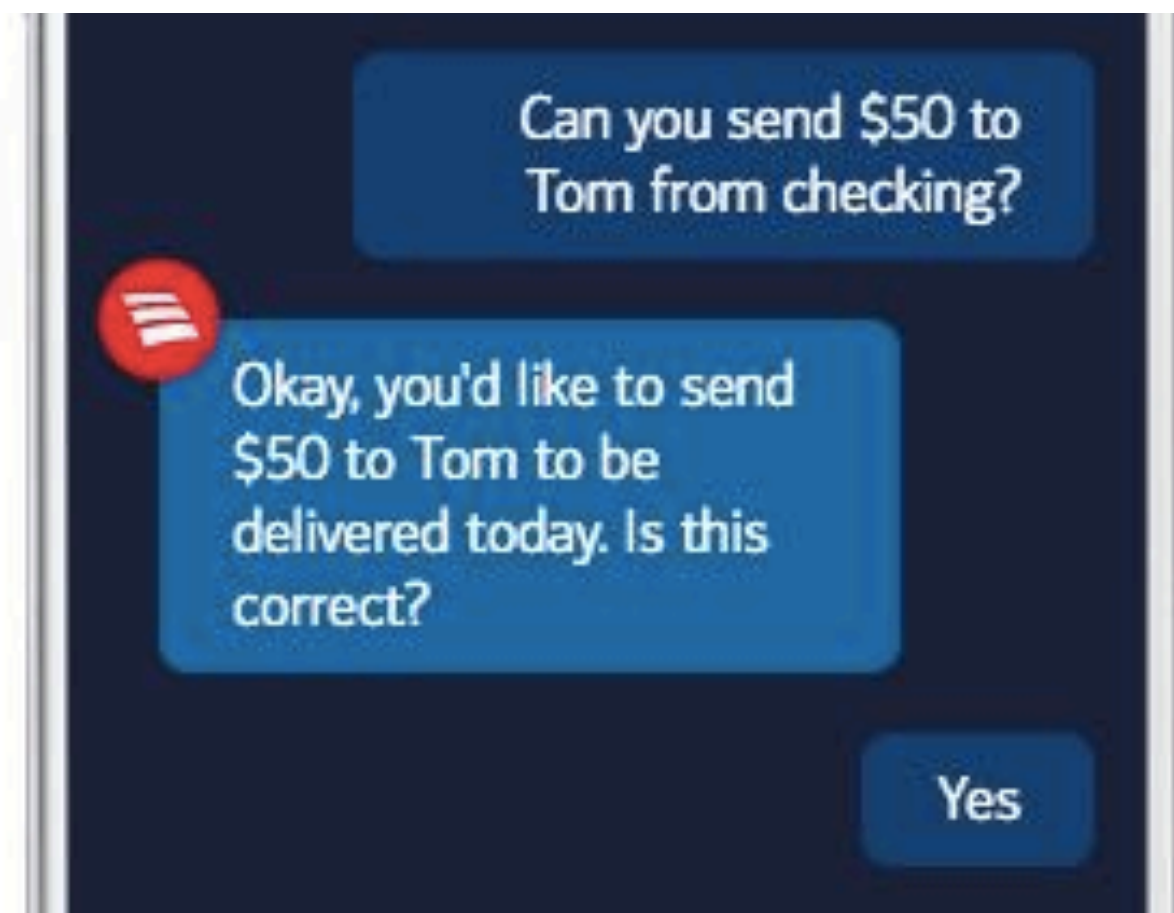


# Natural Language Processing

- 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**

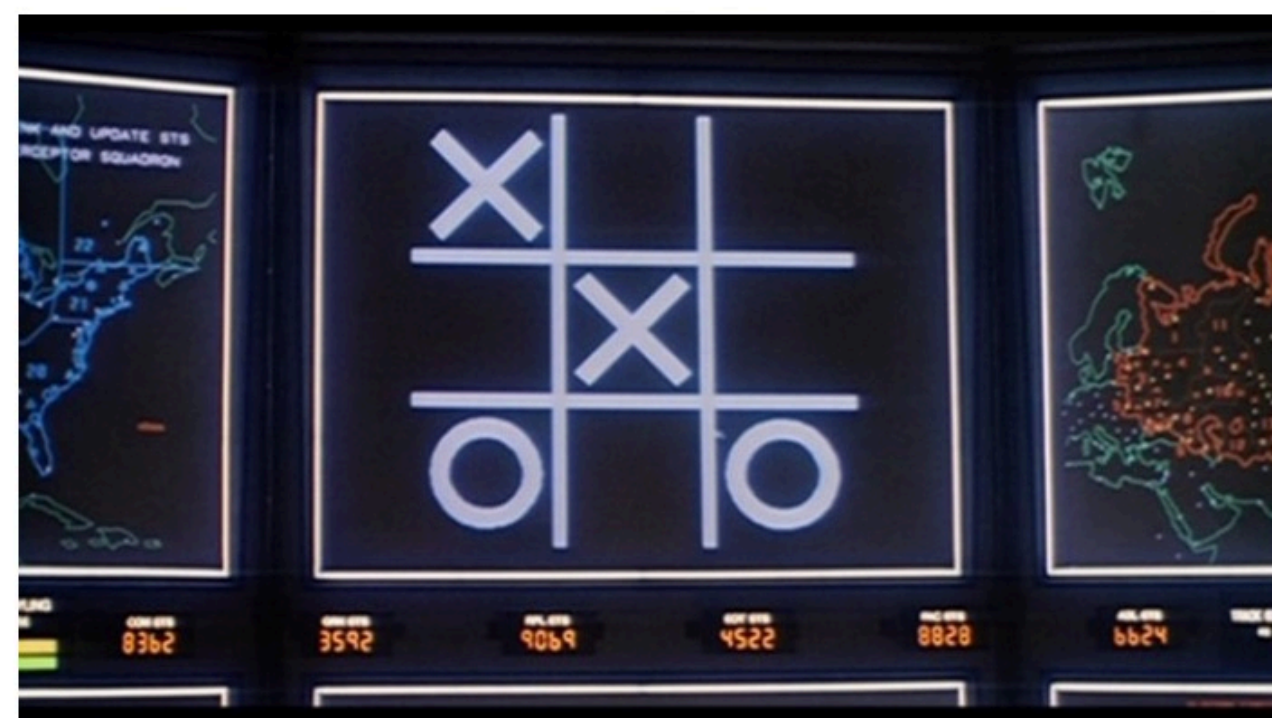


**OFFLINE**

# Computer learns to play Civilization by reading the instruction manual

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

     **532 SHARES**

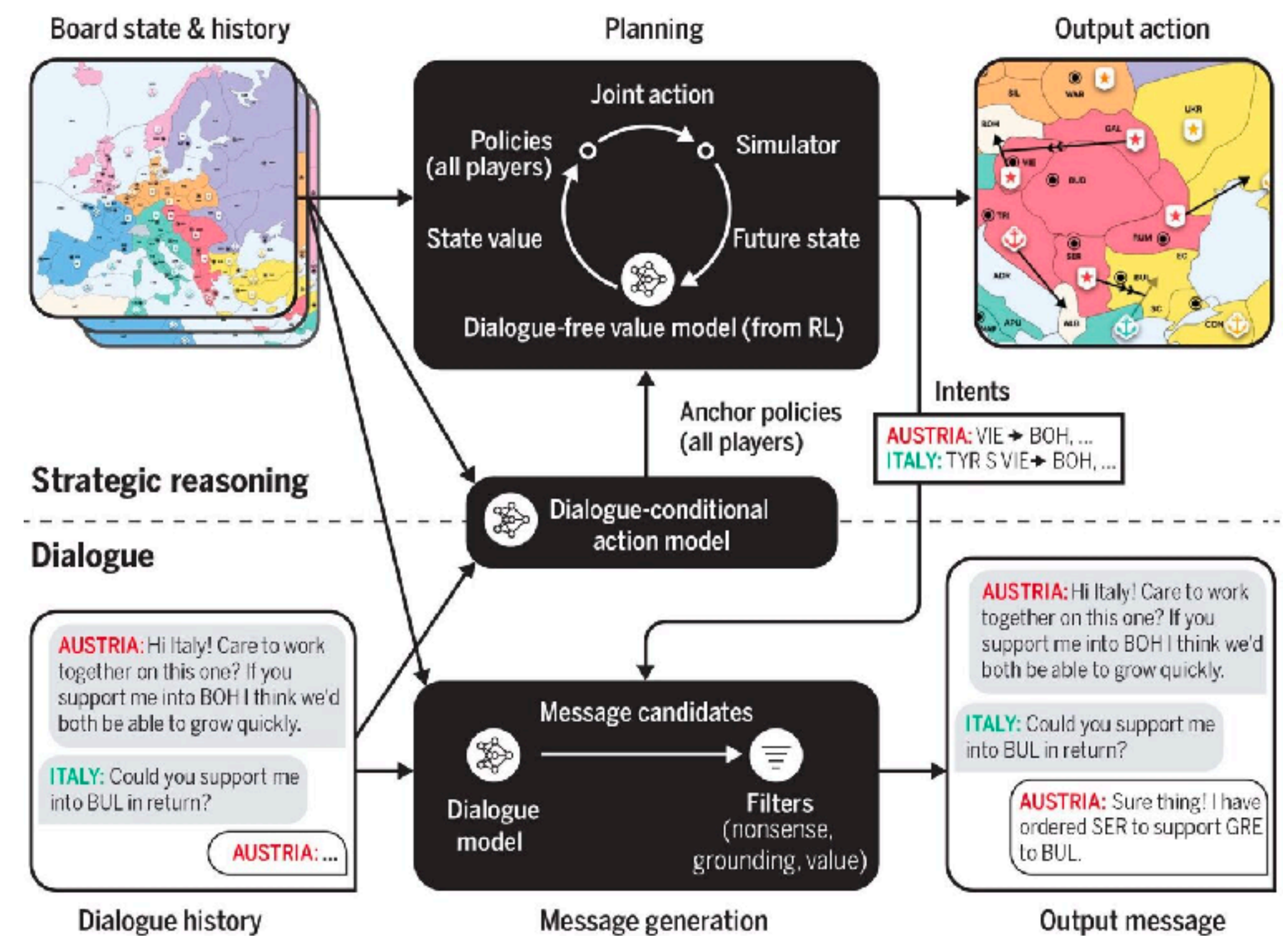


MIT researchers just got a computer to accomplish yet another task that most humans are incapable of doing: It learned how to play a game by reading the instruction manual.

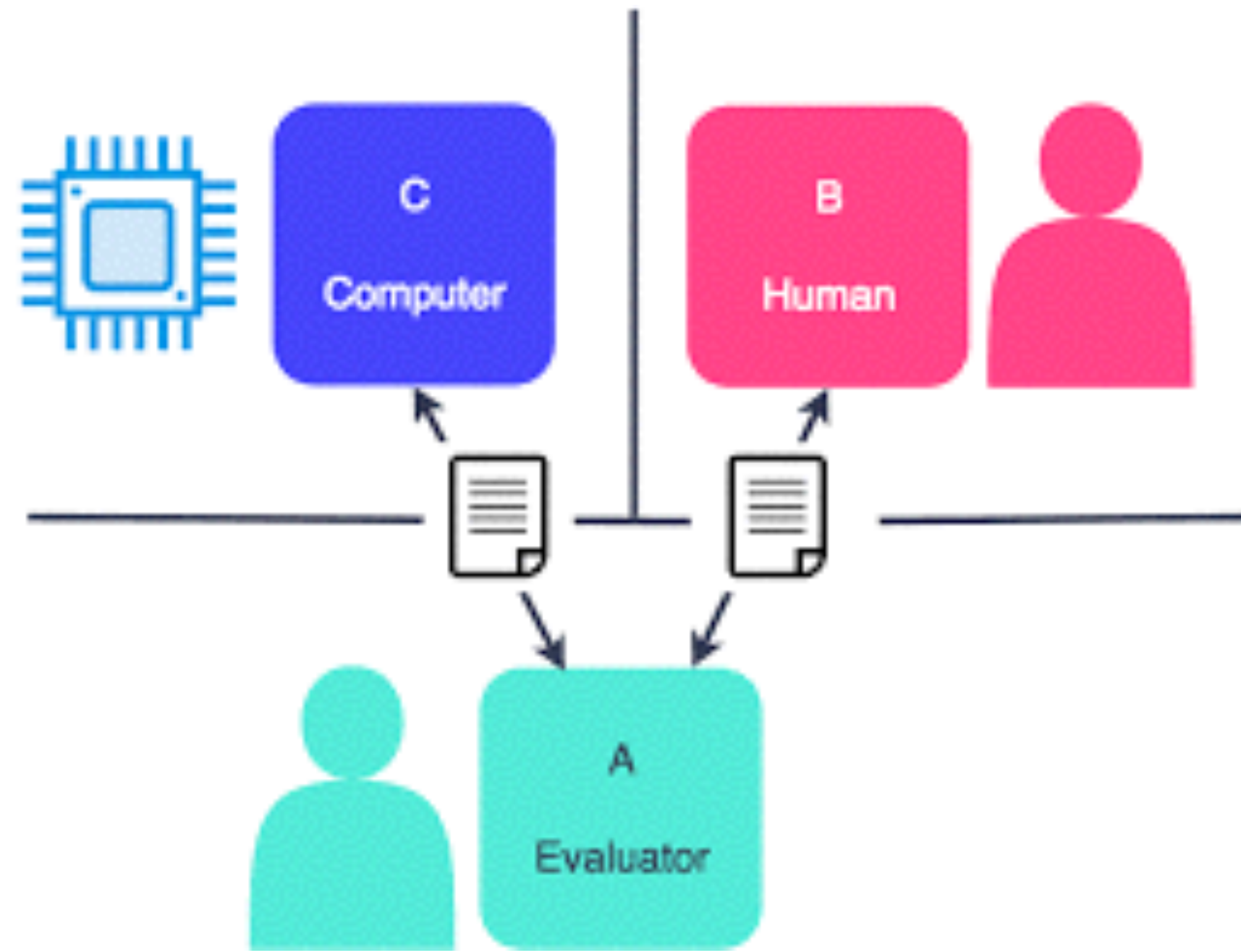
The MIT Computer Science and Artificial Intelligence lab has a computer that now plays Civilization

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

By [Edd Gent](#) > November 28, 2022



# Turing Test

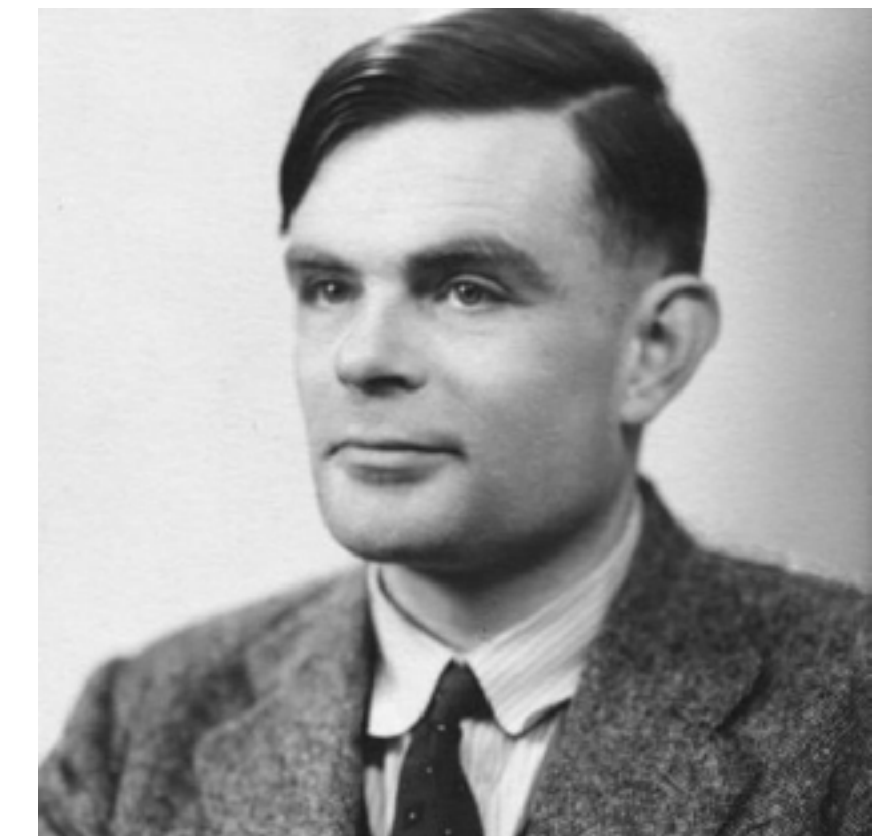


A. M. Turing (1950) *Computing Machinery and Intelligence*. *Mind* 49: 433-460.

## COMPUTING MACHINERY AND INTELLIGENCE

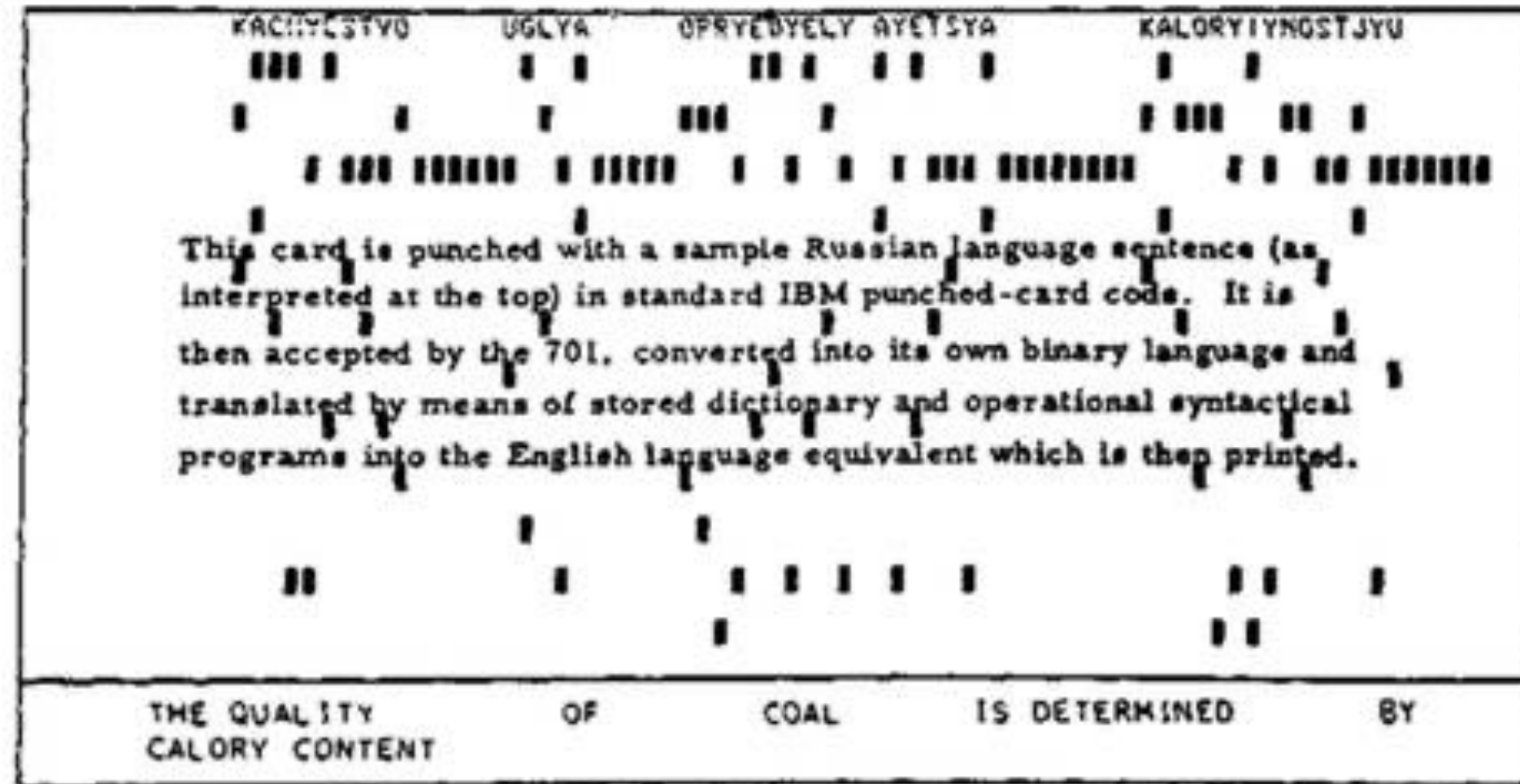
By A. M. Turing

### 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

↔ **HINDI** YORUBA ENGLISH

The man shot the elephant while wearing his pyjamas.

उस आदमी ने हाथी को अपना पजामा पहनाया।

us aadamee ne haathee ko apana pajaama pahanaaya.

52 / 5000

DETECT LANGUAGE **HINDI** ENGLISH YORUBA

↔ HINDI YORUBA **ENGLISH**

उस आदमी ने हाथी को अपना पजामा पहनाया।

The man put his pajamas on the elephant.

us aadamee ne haathee ko apana pajaama pahanaaya.

37 / 5000 अ

# It's getting better!

Detect language **English** Maori Spanish ▾ ↔ **Hindi** Maori English ▾

The man shot the elephant while wearing pyjamas ×

आदमी ने पायजामा पहने हुए हाथी को गोली मार दी ☆

aadamee ne paayajaama pahane hue haathee ko golee maar dee

47 / 5,000 🗣️ 🔊 📄 🗨️ 📧

[Send feedback](#)

Detect language **Hindi** English Maori ▾ ↔ Hindi Maori **English** ▾

आदमी ने पायजामा पहने हुए हाथी को गोली मार दी ×

Man shoots elephant while wearing pyjamas ☆

aadamee ne paayajaama pahane hue haathee ko golee maar dee

44 / 5,000 🗣️ 🔊 📄 🗨️ 📧

[Send feedback](#)

# It's getting better! (Kind of)

Detect language **English** Maori Spanish ▾ ↔ **Maori** English Spanish ▾

The man shot the elephant while wearing pyjamas. ×

I pupuhi te tangata i te arewhana i a ia e mau kakahu ana. ☆

48 / 5,000 ▾

Send feedback

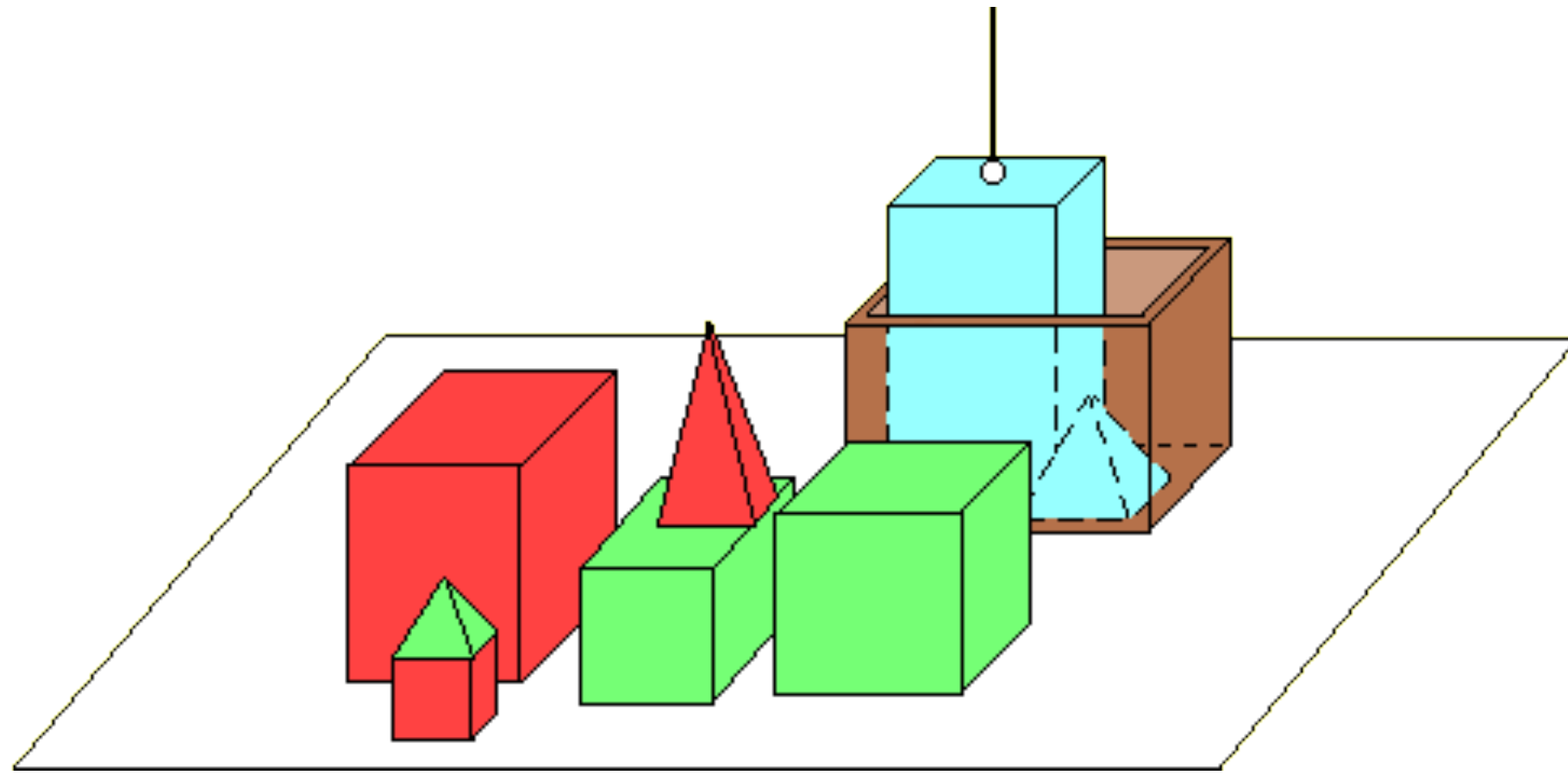
Detect language English **Maori** Spanish ▾ ↔ Maori **English** Spanish ▾

I pupuhi te tangata i te arewhana i a ia e mau kakahu ana. ×

The man shot the elephant while it was wearing clothes. ☆

58 / 5,000 ▾

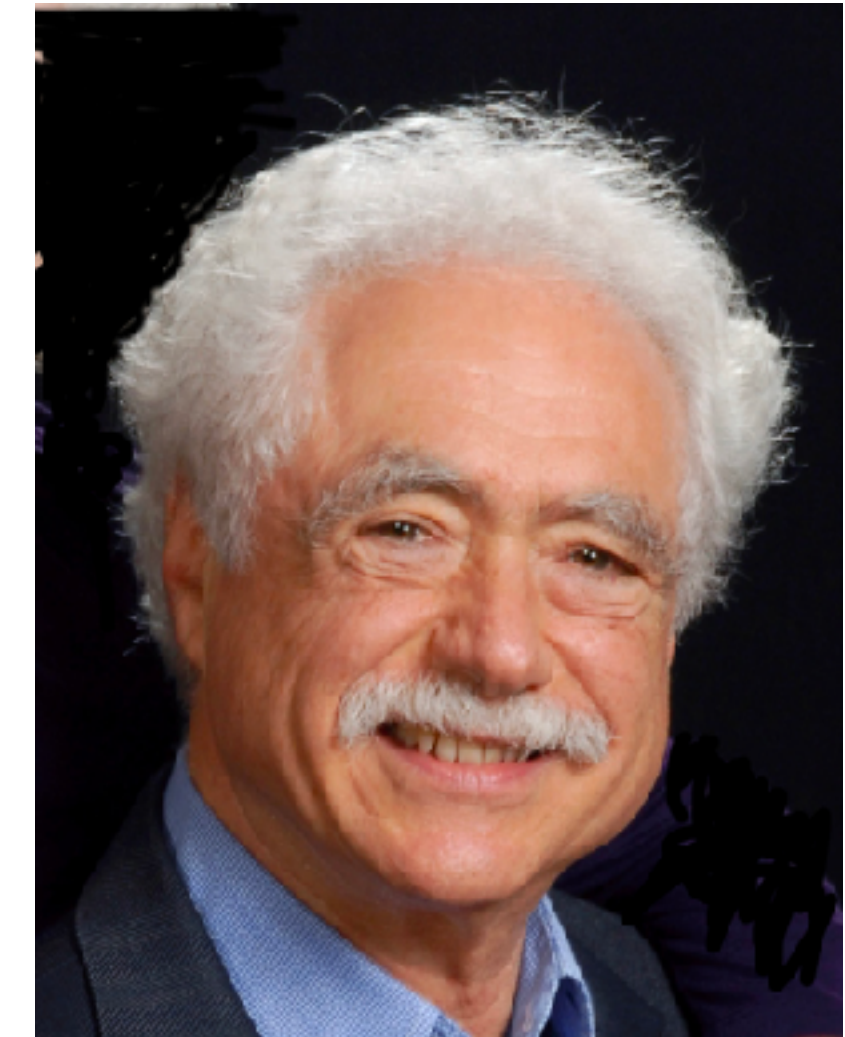
Send feedback



SHRDLU,  
1968

> How many red  
blocks are there?  
- **THREE OF THEM**

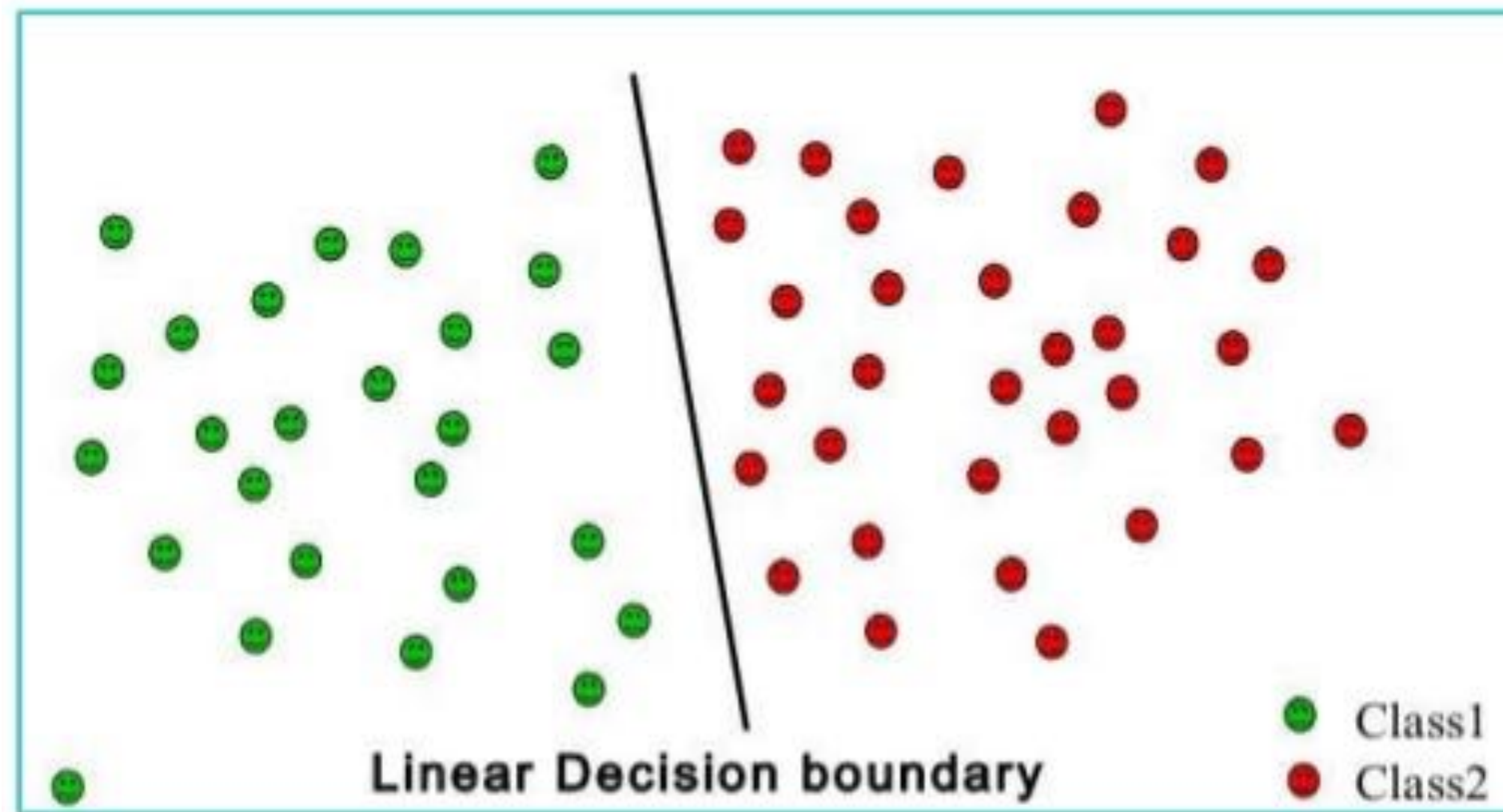
> Pick up the red  
block on top of a  
green one  
**OK.**



Terry Winograd

- 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?



# The era of deep learning

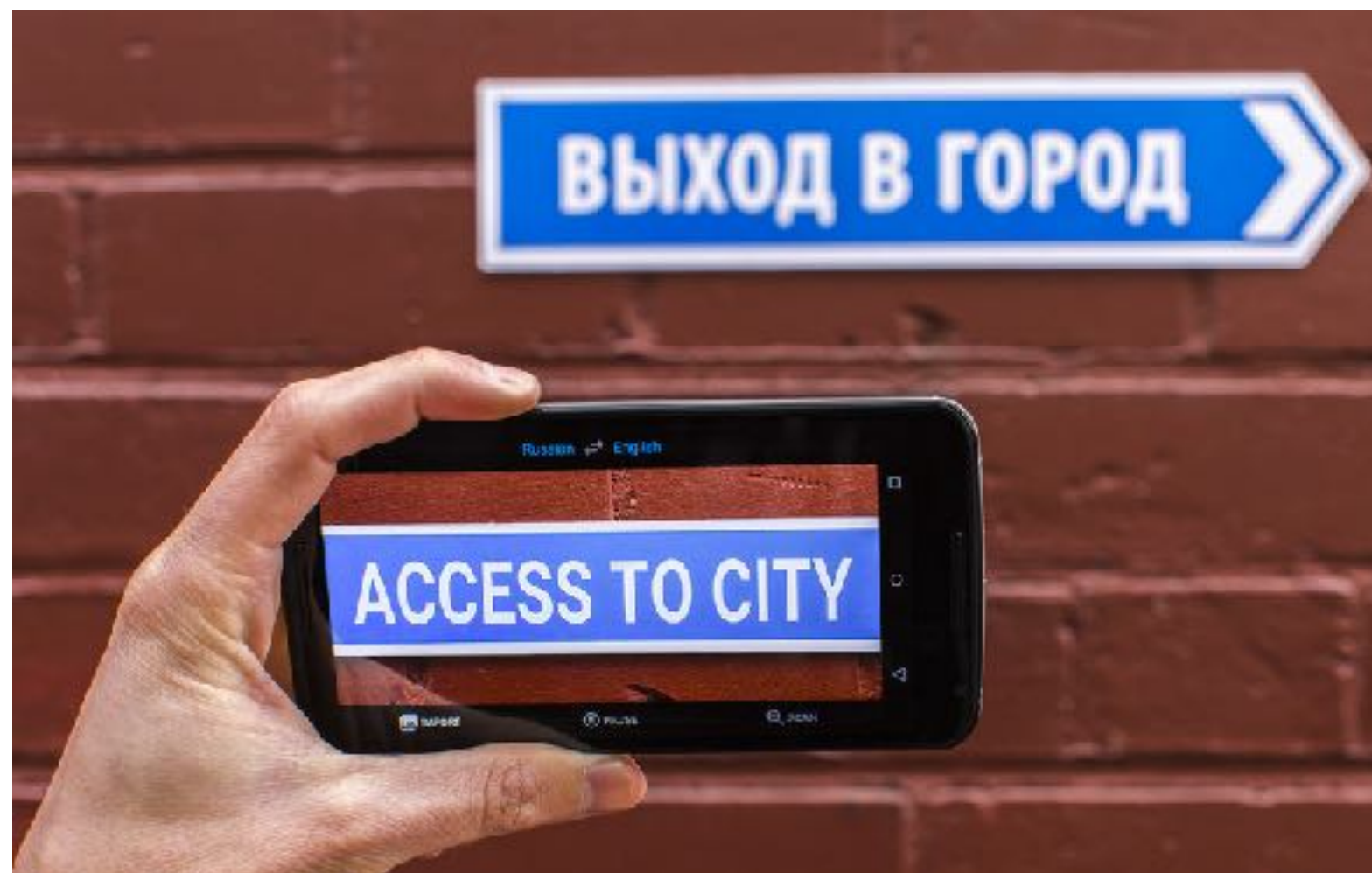
- Significant advances in core NLP technologies
- **Essential ingredient:** large-scale supervision, lots of compute
- Reduced manual effort - less/zero **feature engineering**



GPU



TPU



36M sentence pairs

*Russian: Машинный перевод - это круто!*

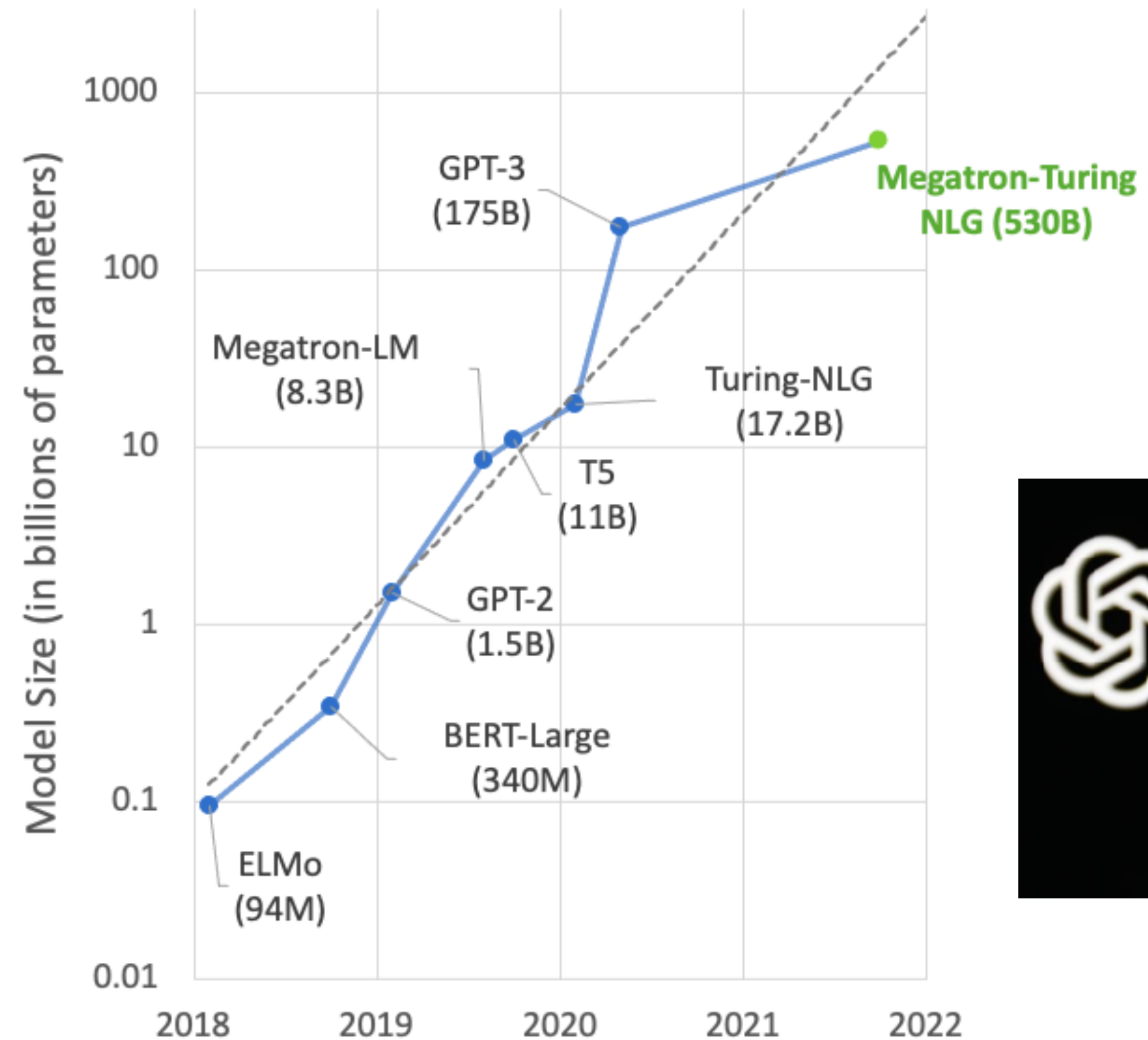


*English: Machine translation is cool!*

# 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?



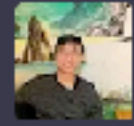
Here are some NLP topics that are popular and relevant in 2023:



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.



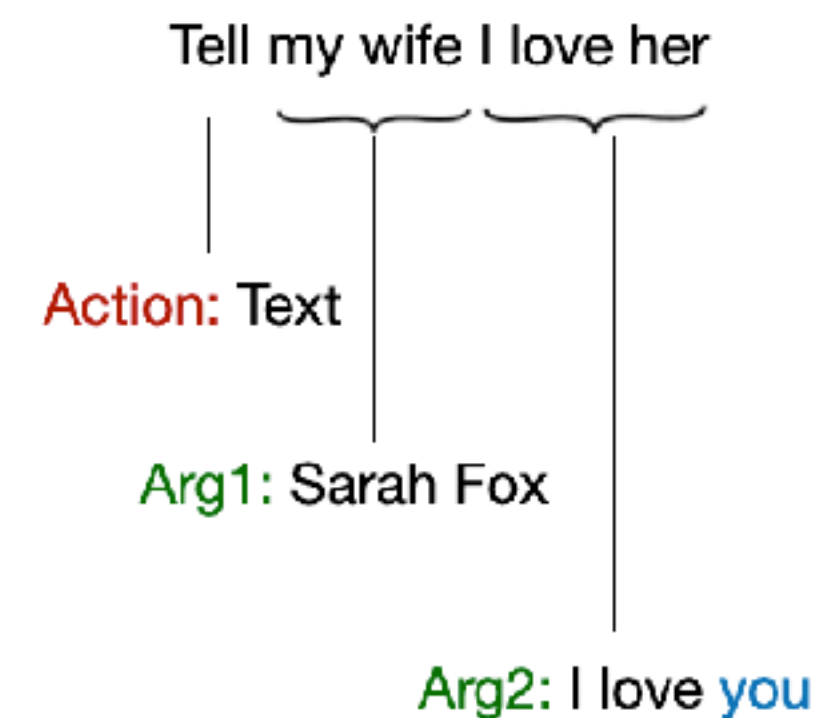
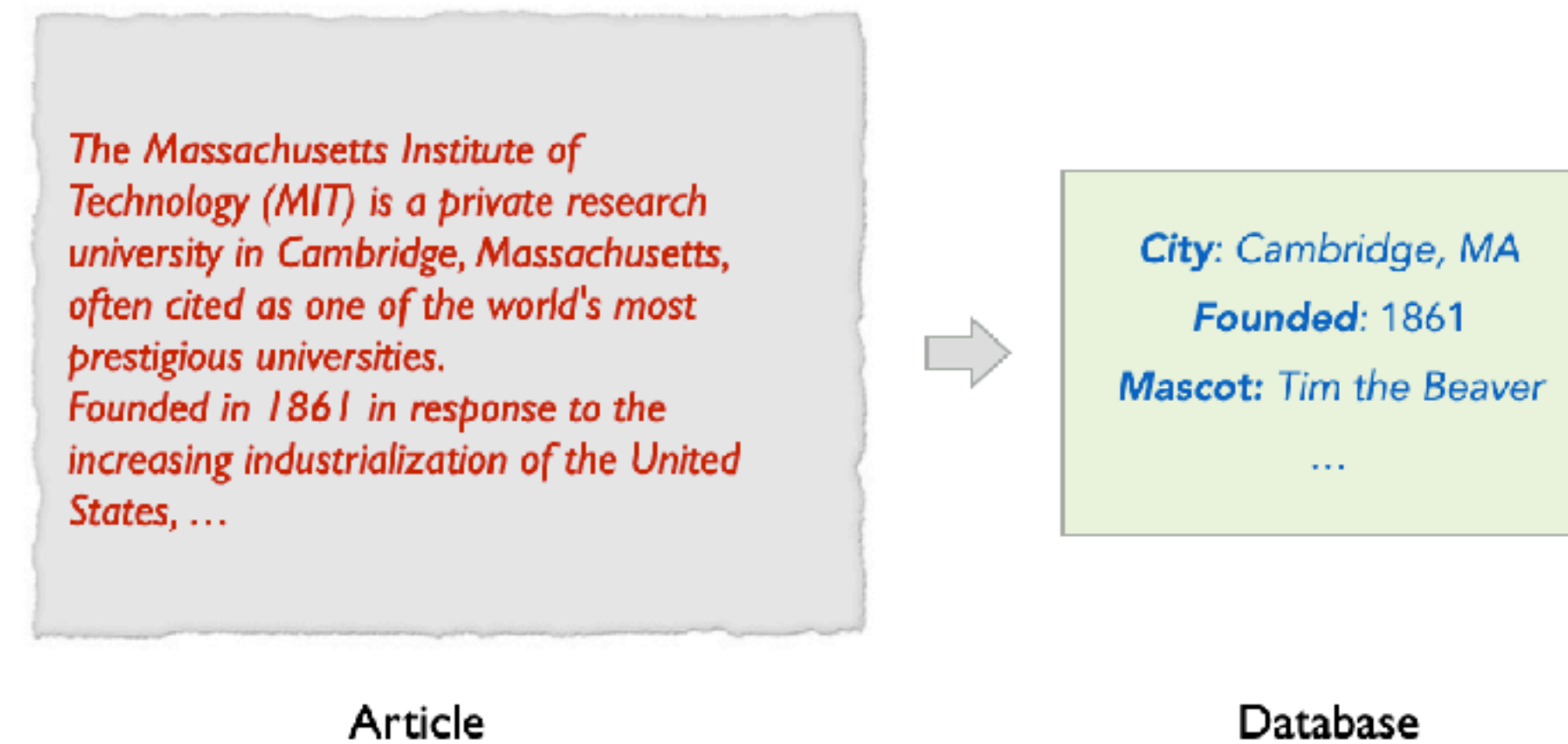
Why do you hallucinate so much?



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

Before midterm

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

+ guest  
lectures!

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**.

**bank**<sup>1</sup>

/baNGk/ 

*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.

**bank**<sup>1</sup>

/baNGk/ 

*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

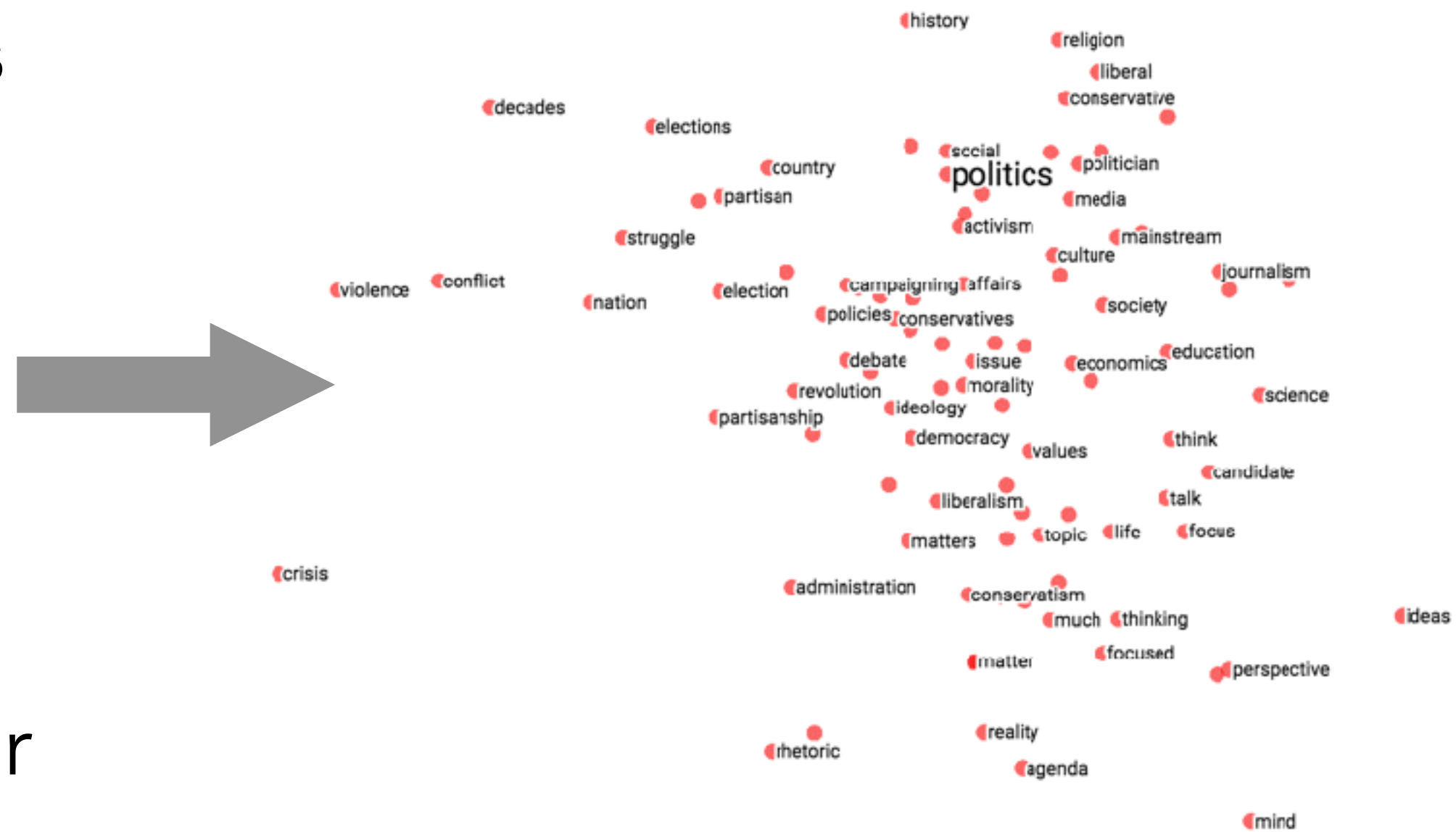


**ACCORDING TO THE THESAURUS,  
"THEY'RE HUMID, PREPOSSESSING  
HOMOSAPIENS WITH FULL SIZED AORTIC  
PUMPS" MEANS "THEY'RE WARM, NICE  
PEOPLE WITH BIG HEARTS."**

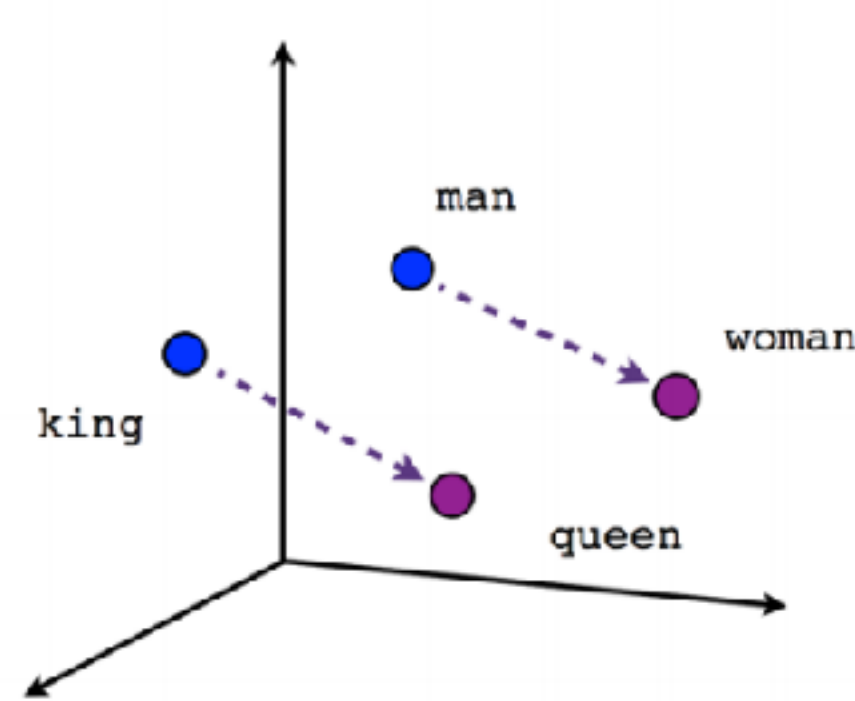
Several words can mean the same thing!

# Distributed representations

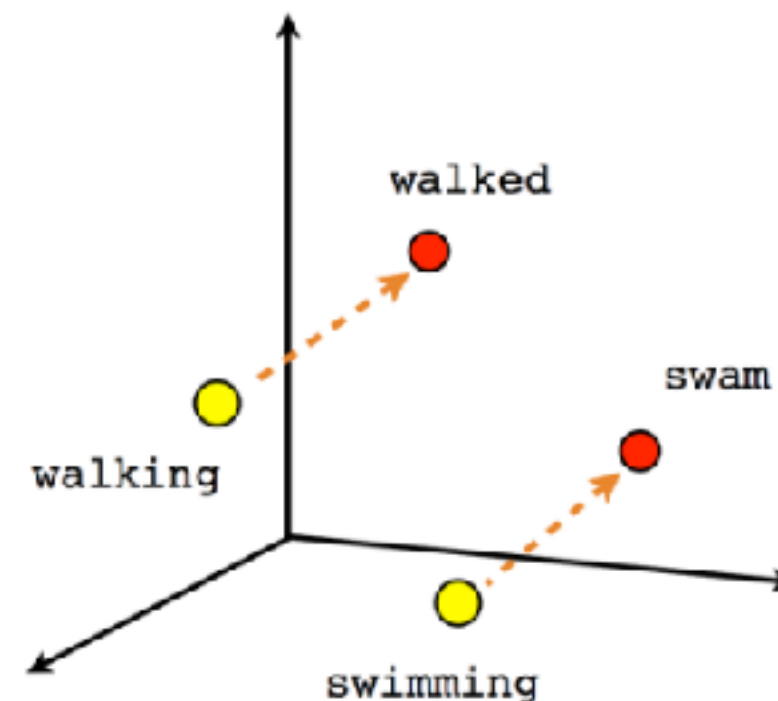
Project words onto a continuous vector space



Similar words closer to each other



Male-Female



Verb tense

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

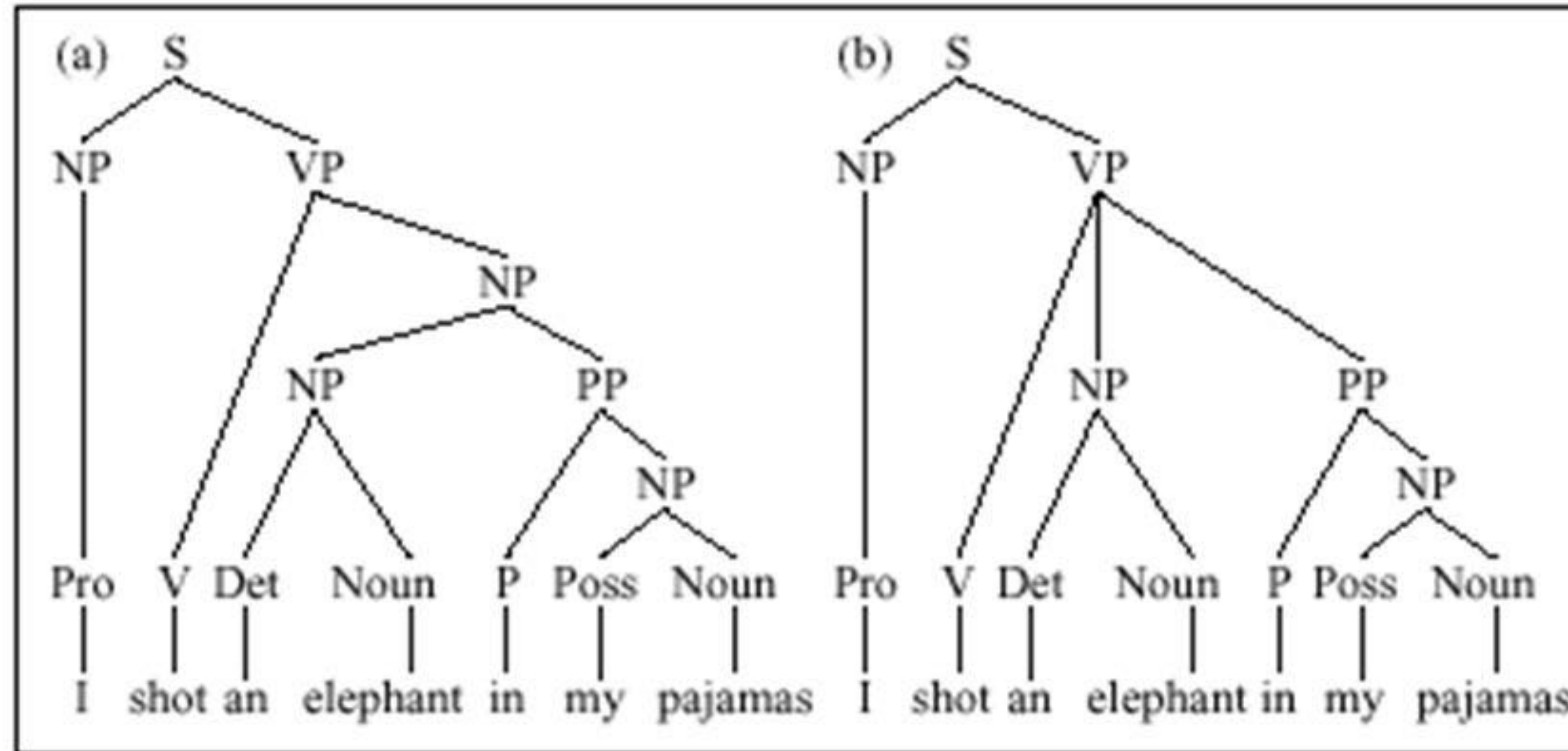
# Comprehending word sequences

- 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
  - The cotton clothing is made of grows in Mississippi
- Implicit structure in all languages
  - Coarse-to-fine levels (recursive)
  - What are some good data structures to represent this?

Garden Path sentence

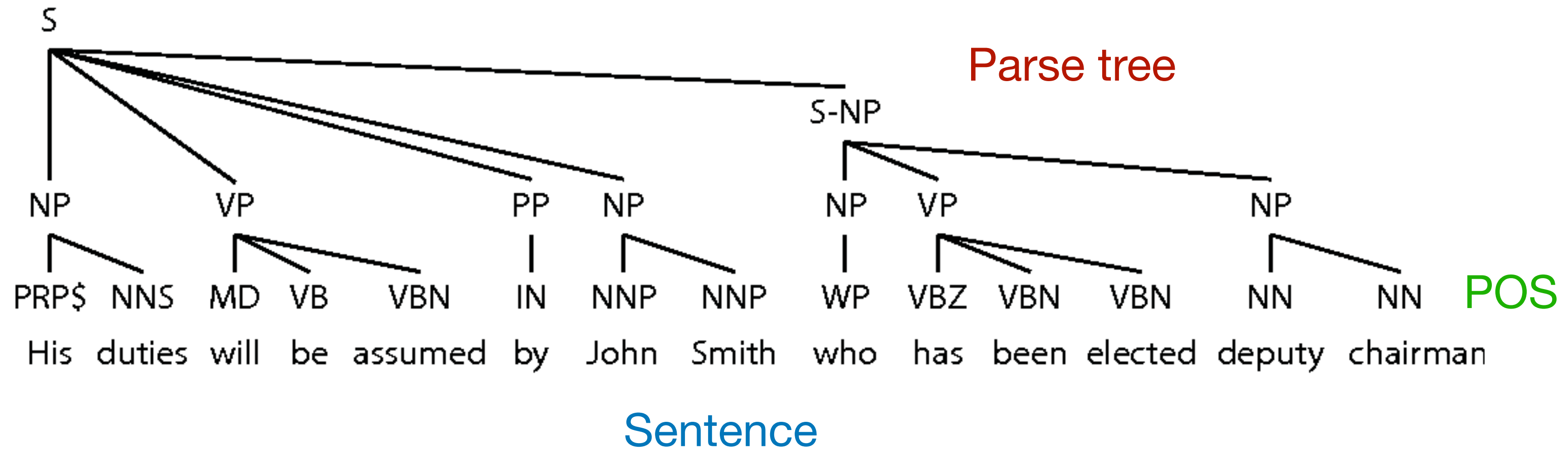
# Syntactic ambiguity

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