COS 484: Natural Language Processing

L1: Introduction to NLP

Spring 2024

Logistics and course structure

Course staff

Instructor

Karthik Narasimhan

Graduate TAs

Anika Maskara

Colin Wang

Howard Yen

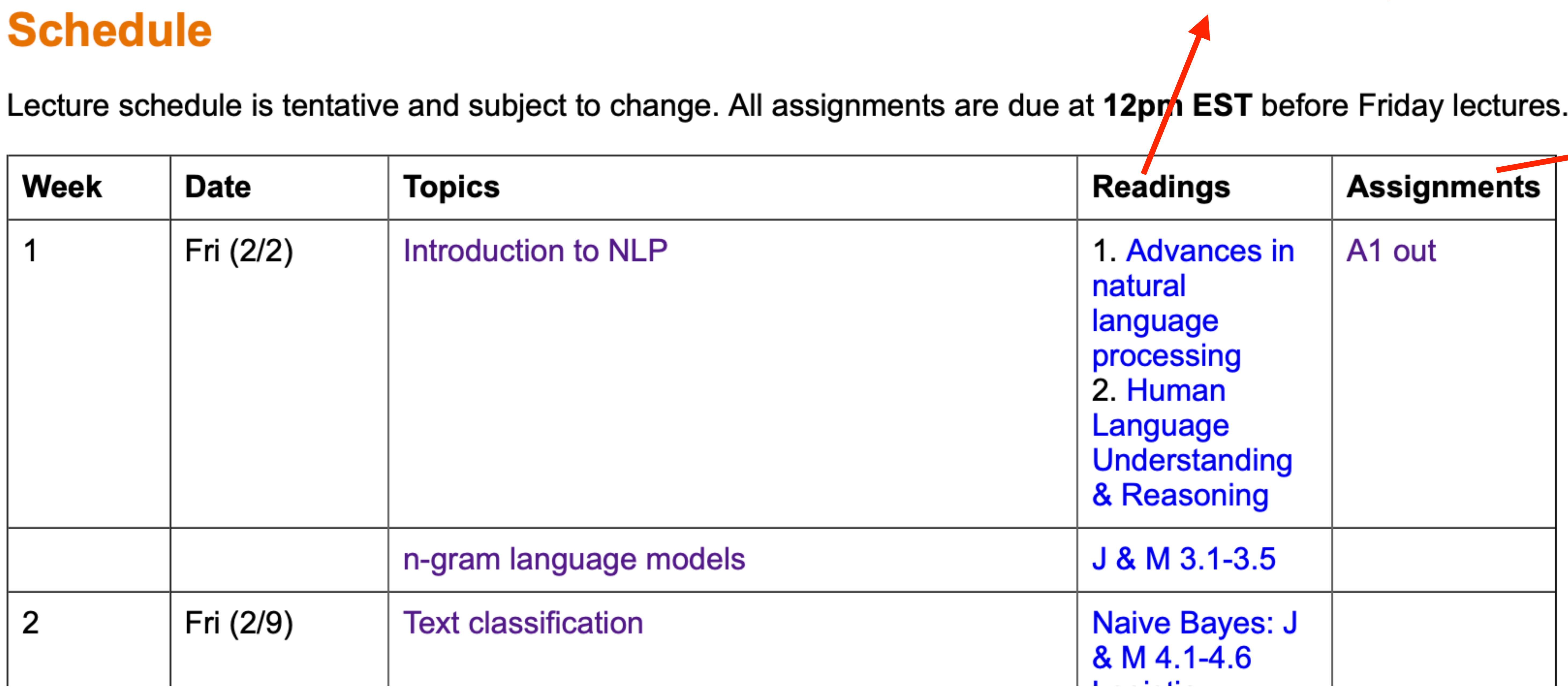
Yihan Wang

Logistics

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!

All readings are optional but can help strengthen your understanding



Schedule

Lecture schedule is tentative and subject to change. All assignments are due at **12pm EST** before Friday lectures.

Week	Date	Topics	Readings	Assignments
1	Fri (2/2)	Introduction to NLP	1. Advances in natural language processing 2. Human Language Understanding & Reasoning	A1 out
		n-gram language models	J & M 3.1-3.5	
2	Fri (2/9)	Text classification	Naive Bayes: J & M 4.1-4.6	

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

Logistics

- **Lectures:** Friday 1:30pm - 4:20pm in CS 104 (10-15min break in between)
- **Precepts:** 1-hour precept every week taught by TAs (optional)
    - Tuesdays, 3:30-4:30pm, location will be announced on Ed
    - Refresher for basic maths, review course materials, Colab/Pytorch tutorial etc

Logistics

- Sign up for **Ed** and **Gradescope** today
    - We will use **iClicker** for polls in the class.

Go to

join.iClicker.com

KMIS

- **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

Course structure

Assignments (40%): 4 total

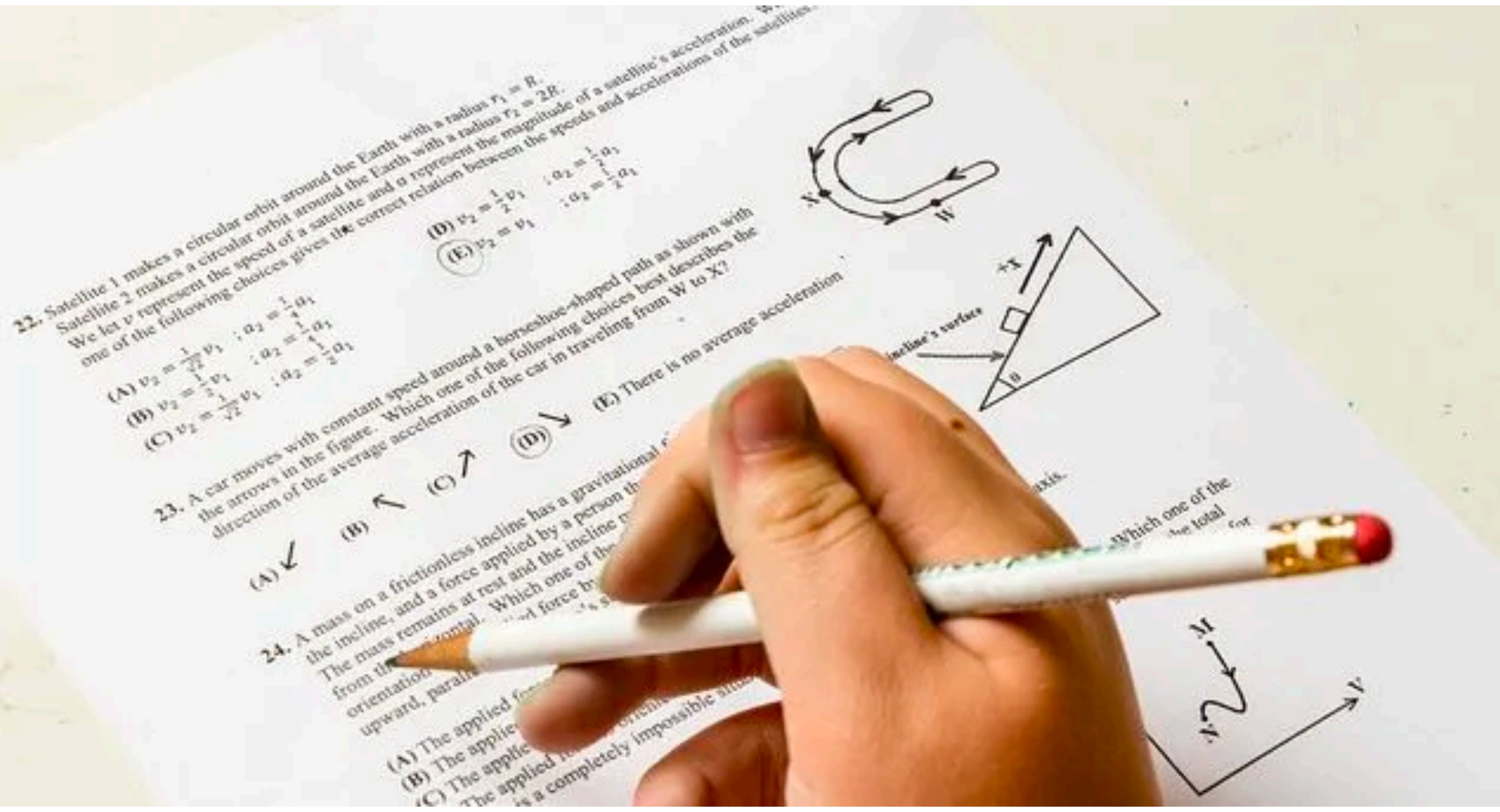
- **A1 (10%)** will be released TODAY and due in two weeks
- **A2, A3, A4**: 10% each - each assignment has 2 or 3 weeks
- Every assignment has a **written** component and a **programming** component based on Colab (A4 requires access to GPUs).
- You will need to know how to program in PyTorch/Numpy.

You have **96 free late hours** 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)

Course structure

Midterm (25%)

- A 3-hour timed exam on Gradescope in a 27-hour window (**March 7-8**)
- All the topics up to March 1 will be covered
- No final exam



Course structure

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
- Proposal (0%) due before - date will be announced soon
- Poster presentation tentatively scheduled on May 3 (during reading period)
- Final report due on May 7

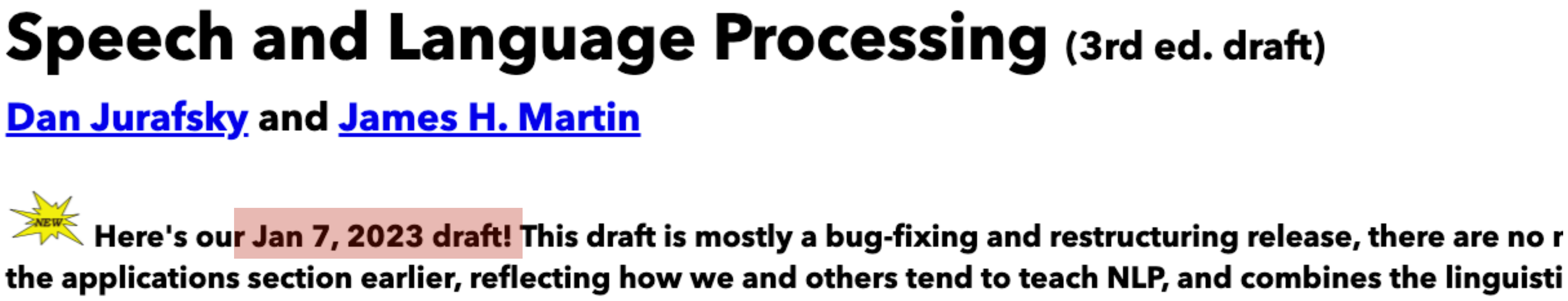
Extra bonus (5%) - participation in class and Ed discussions

*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/

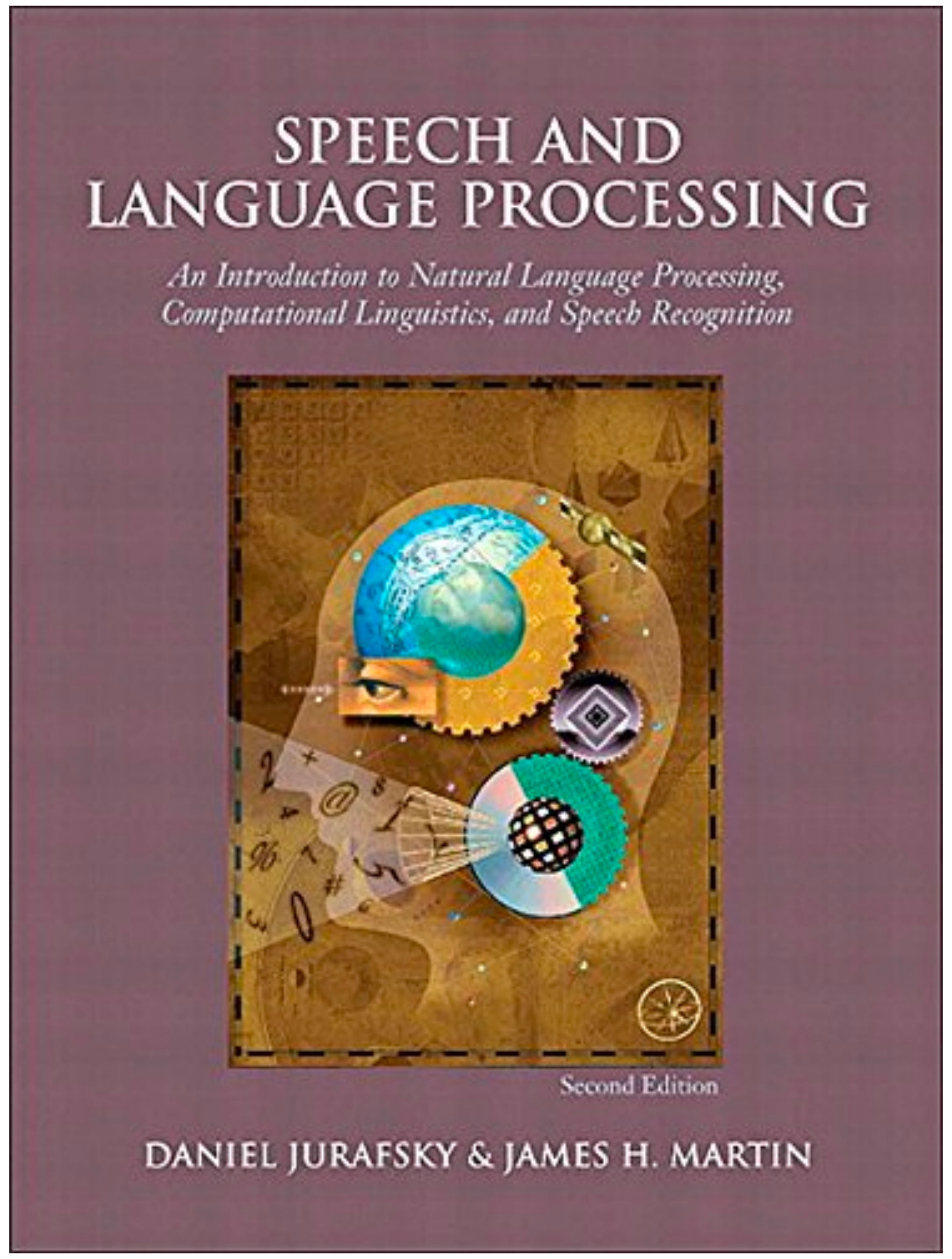


Speech and Language Processing (3rd ed. draft)

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

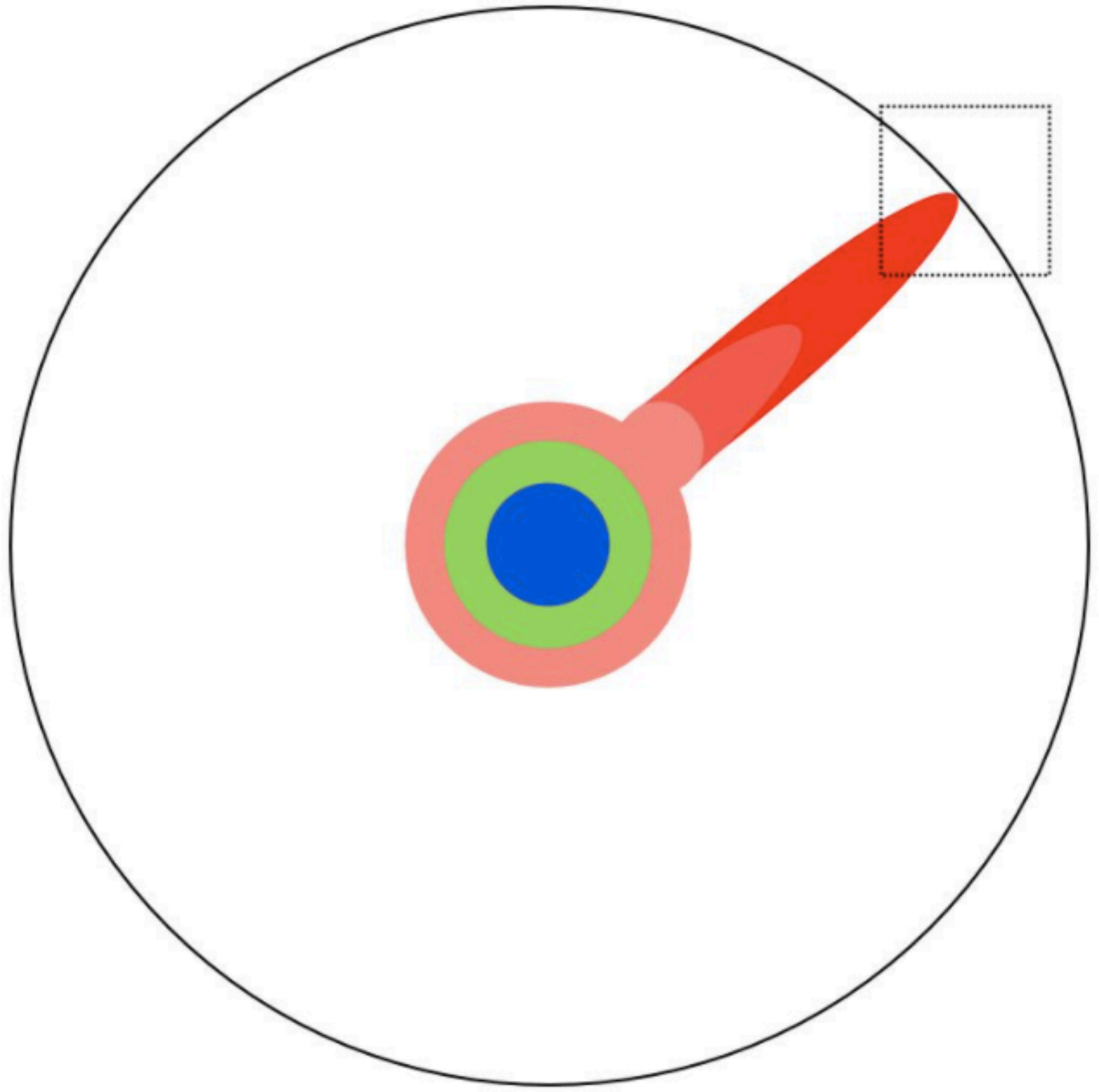
A good way to learn about state-of-the-art NLP concepts is through **research papers** and **blog posts**


SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

2nd edition ✗

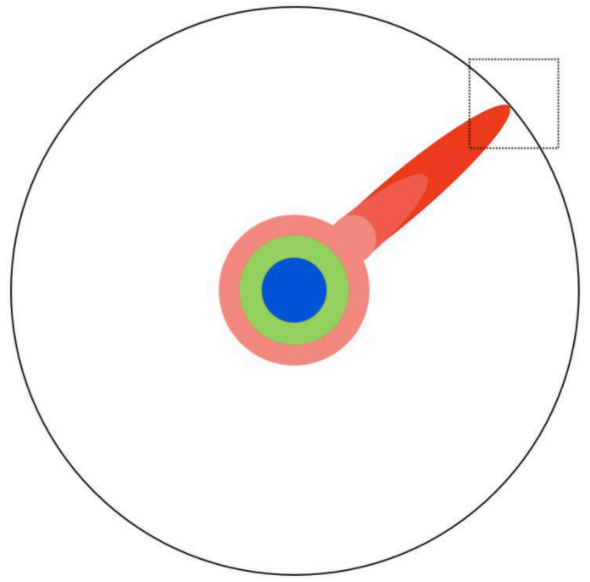
Course goals

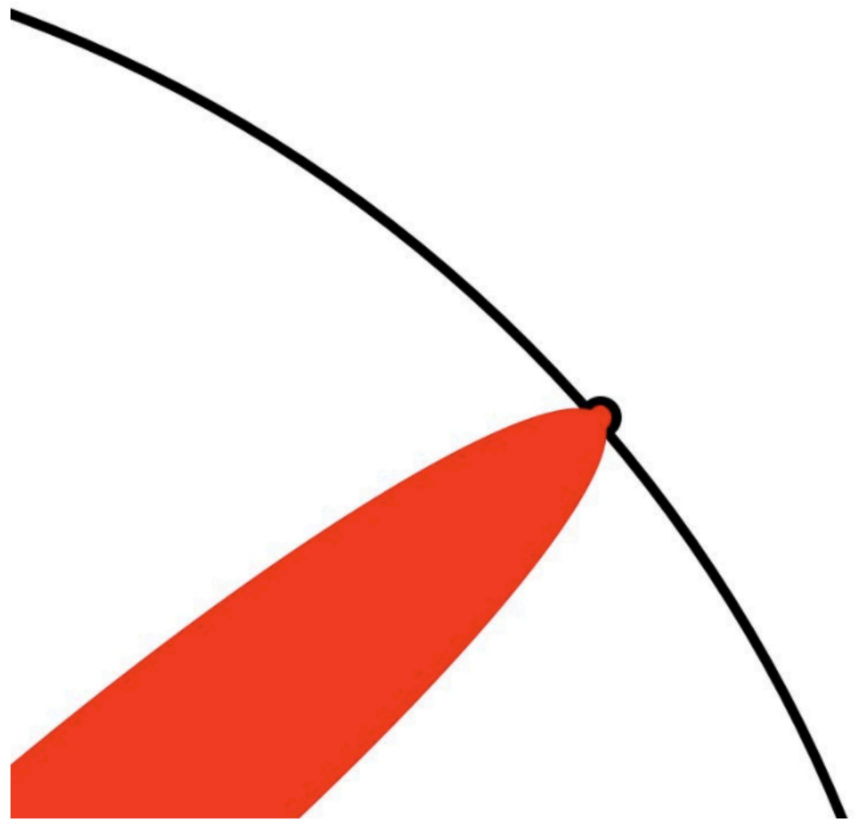
- Gain an understanding of the **fundamentals** of different sub-fields within NLP
- Understand **theoretical** concepts and algorithms
- **Hands on experience** building statistical models for language processing
- Carry out an **independent research project** at the end



This is an advanced class

Image: https://matt.might.net/articles/phd-school-in-pictures/





This is a *very* advanced class

Image: https://matt.might.net/articles/phd-school-in-pictures/

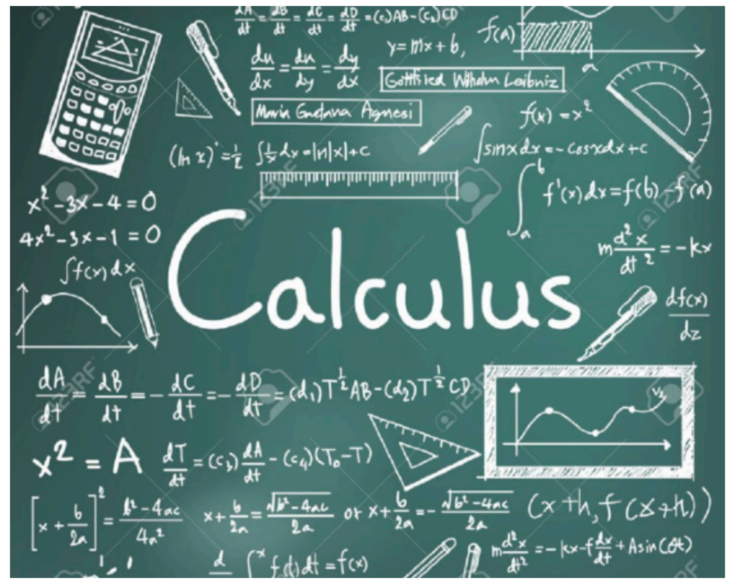
Prerequisites

- **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)



PyTorch

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

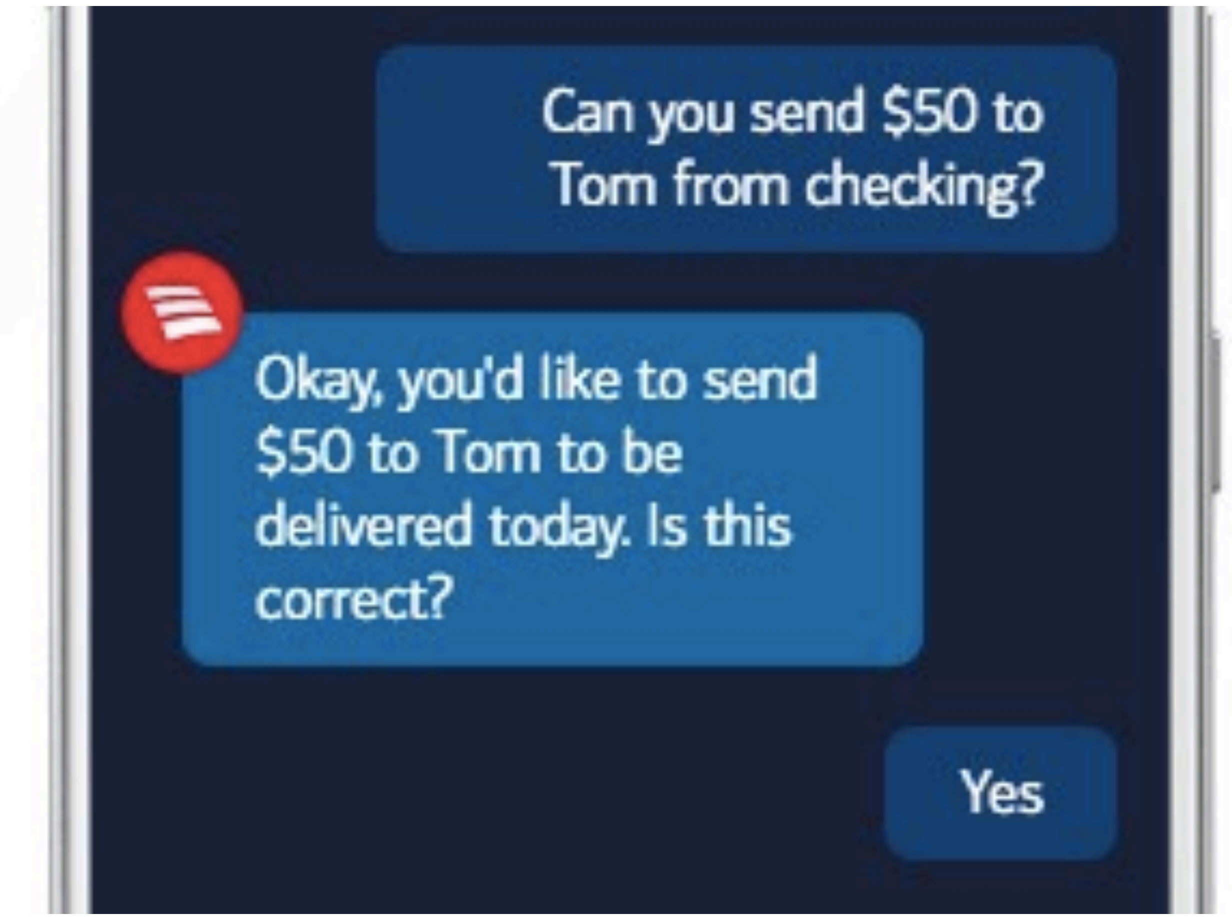
What we say to dogs
Okay, Ginger! I've had it! You stay out of the garbage! Understand, Ginger? Stay out of the garbage, or else!

What they hear
blah blah GINGER blah blah blah blah blah blah blah blah GINGER blah blah blah blah blah blah... blah

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)



Can you send $50 to Tom from checking?

Okay, you'd like to send $50 to Tom to be delivered today. Is this correct?

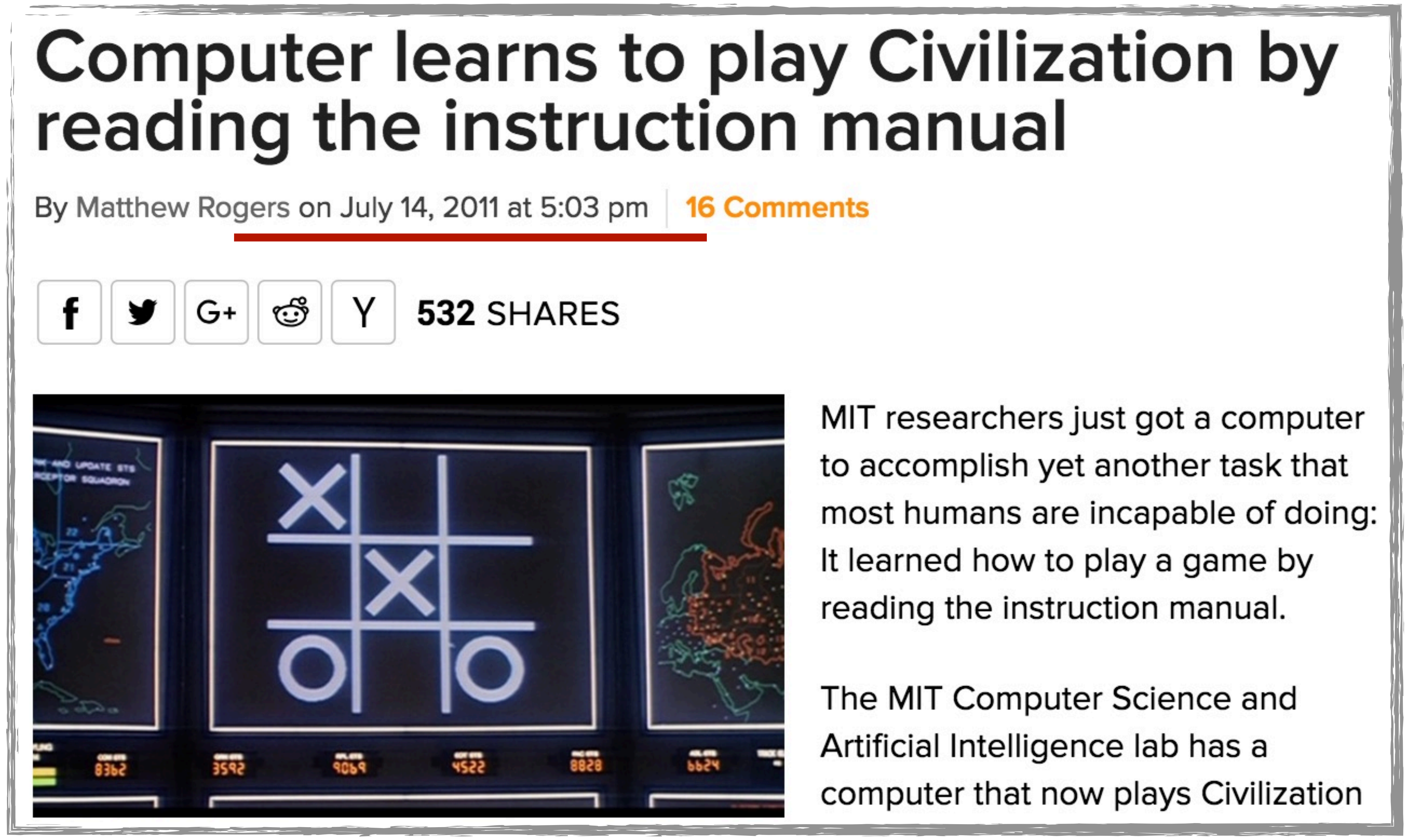
Yes

Banking assistant

ONLINE

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

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

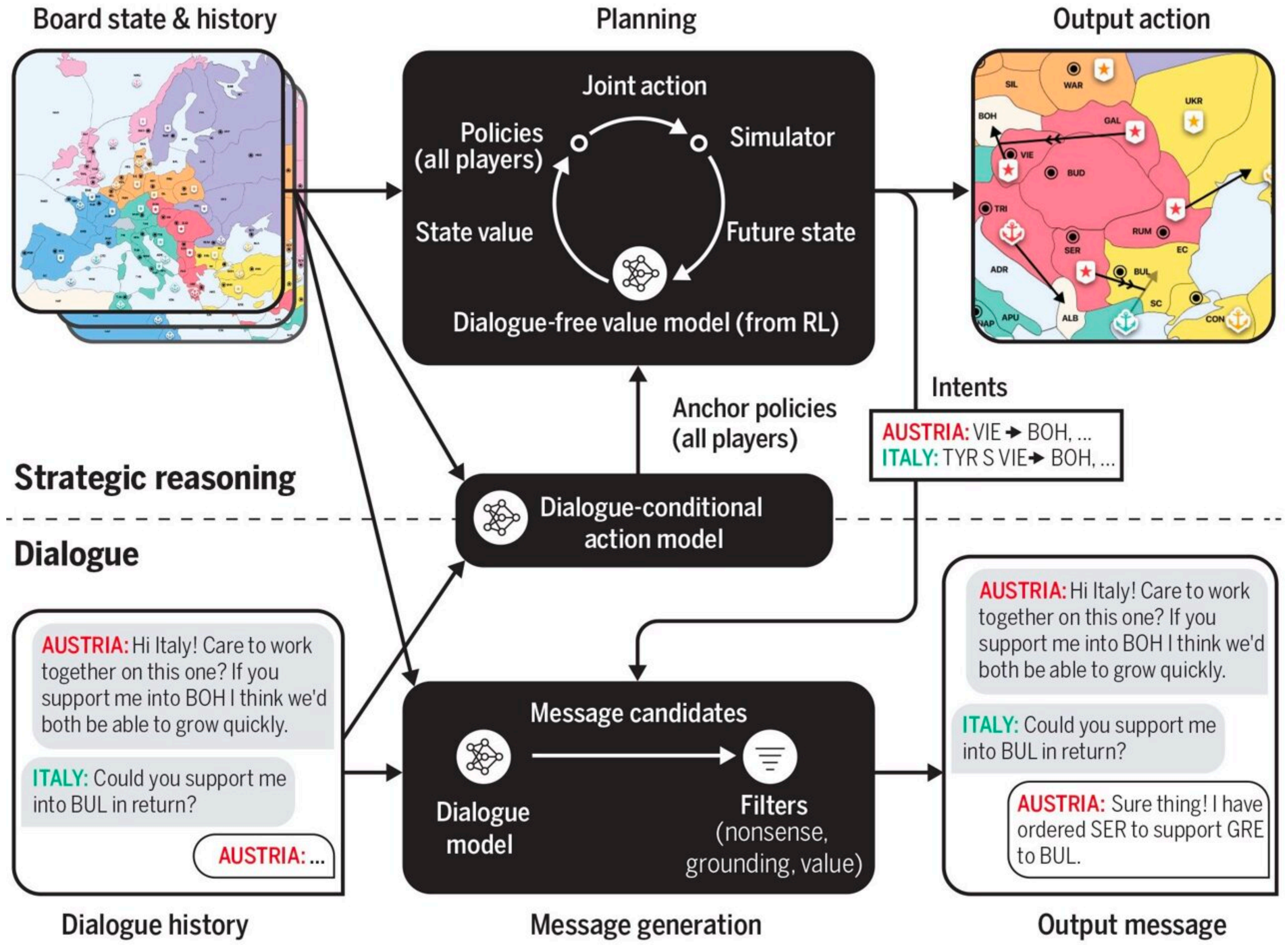
MIT researchers just got a computer to accomplish yet another task that most humans are incapable of doing: It learned how to play a game by reading the instruction manual.

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



AUSTRIA: VIE → BOH, ...
ITALY: TYR S VIE → BOH, ...

AUSTRIA: Hi Italy! Care to work together on this one? If you support me into BOH I think we'd both be able to grow quickly.

ITALY: Could you support me into BUL in return?

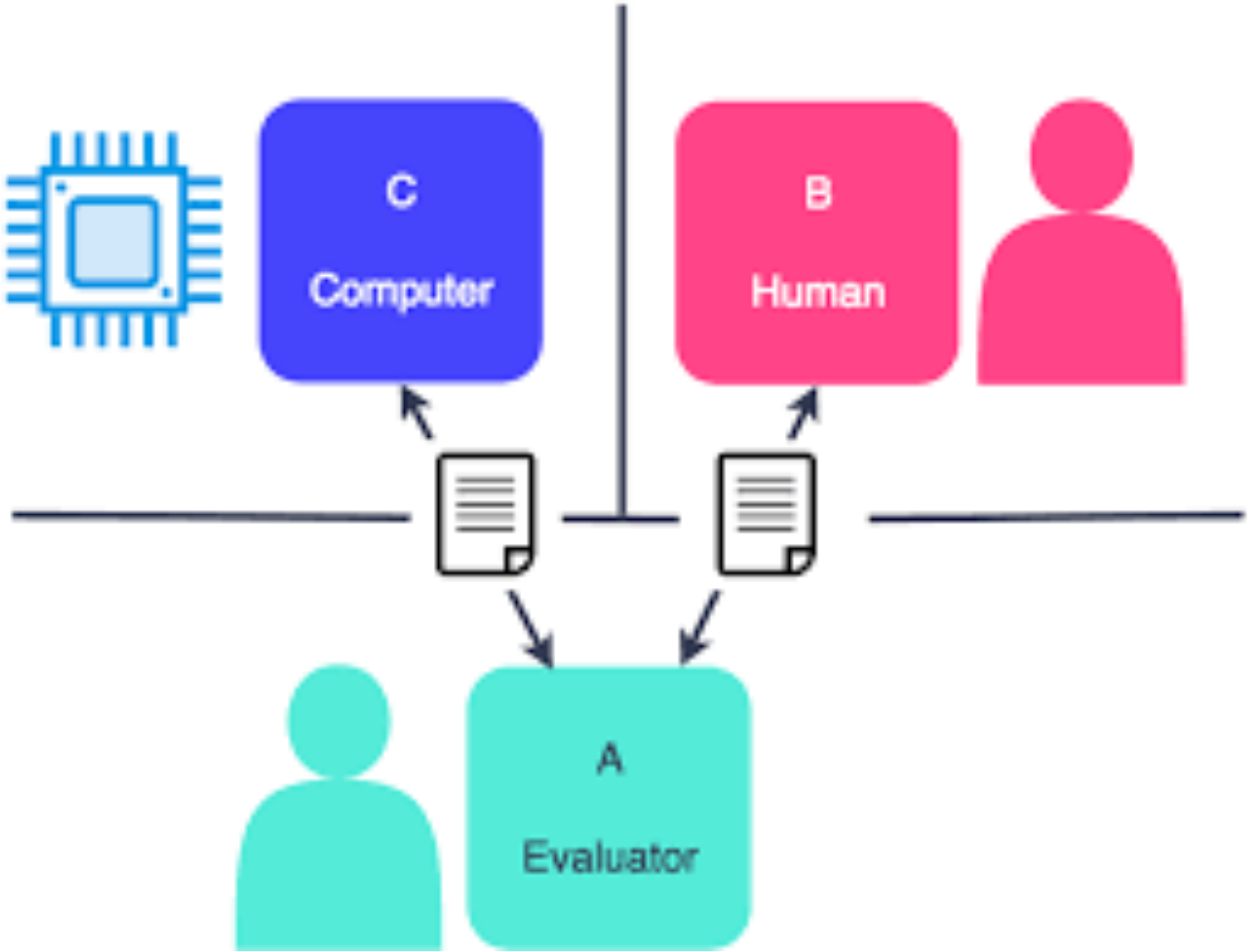
AUSTRIA: ...

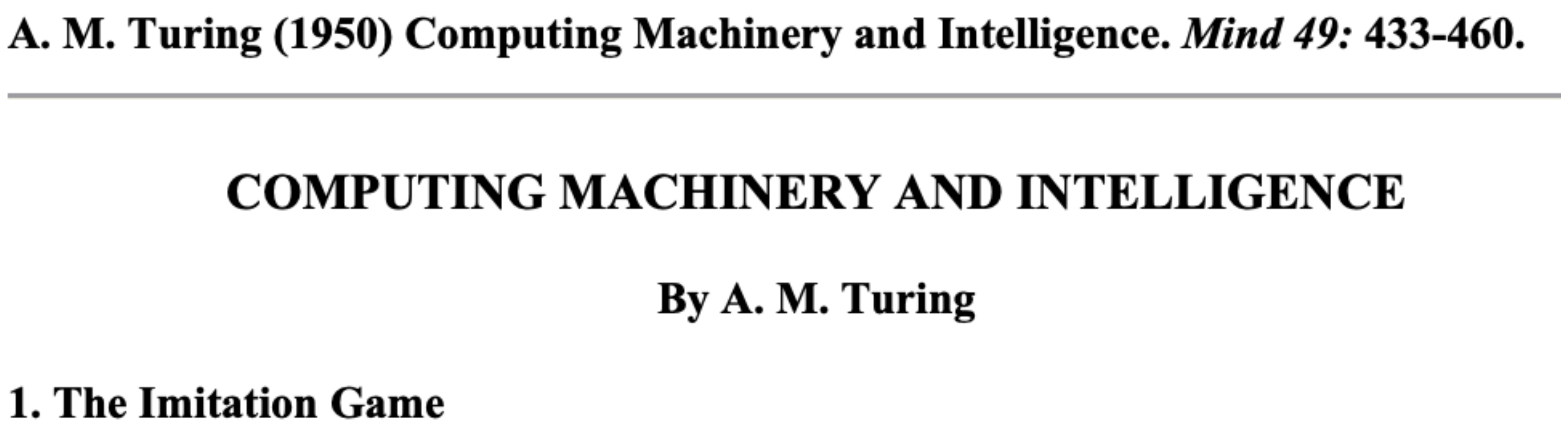
AUSTRIA: Hi Italy! Care to work together on this one? If you support me into BOH I think we'd both be able to grow quickly.

ITALY: Could you support me into BUL in return?

AUSTRIA: Sure thing! I have ordered SER to support GRE to BUL.

Turing Test

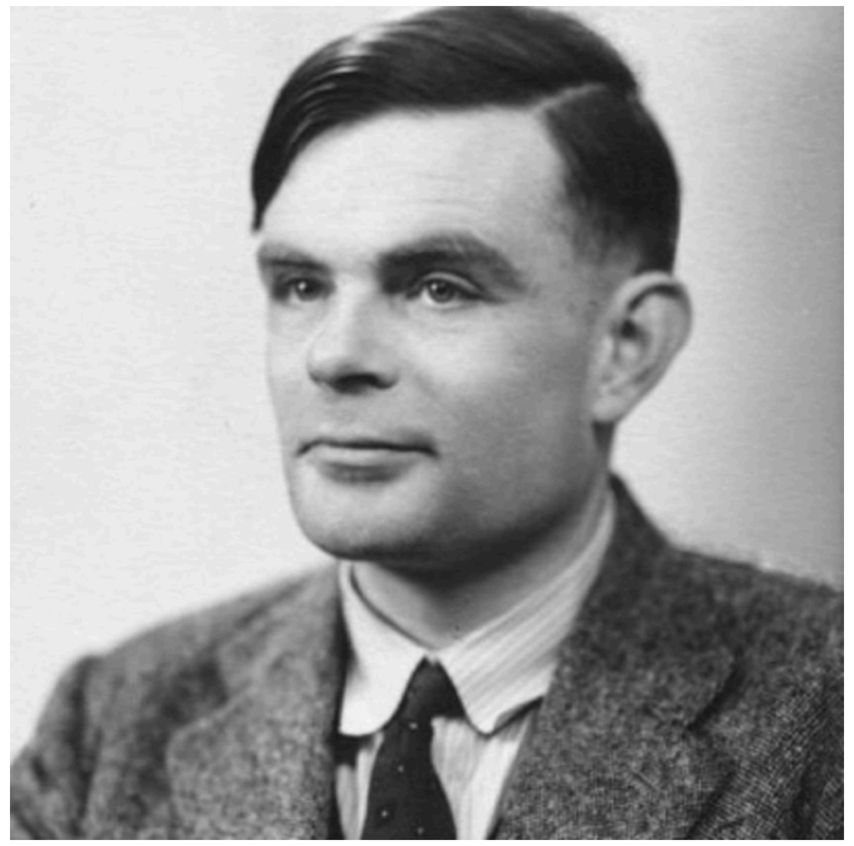



A. M. Turing (1950) Computing Machinery and Intelligence. *Mind 49:* 433-460.

COMPUTING MACHINERY AND INTELLIGENCE

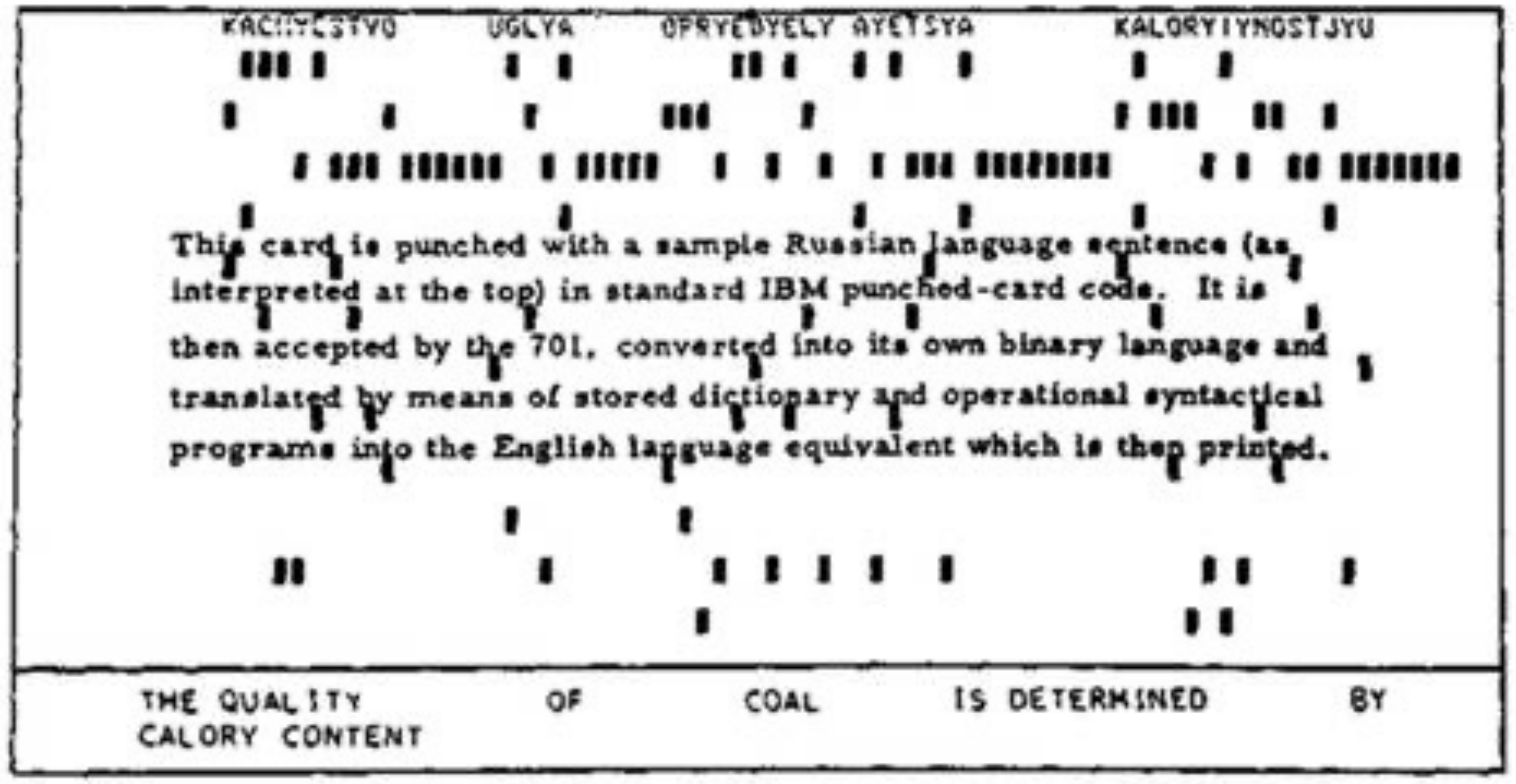
By A. M. Turing

1. The Imitation Game



Ability to understand and generate language ~ intelligence

How it started



KACHYESTVO UGLYA OPRYEDYELYAYETSYA KALORYIYNOSTJYU

This card is punched with a sample Russian language sentence (as interpreted at the top) in standard IBM punched-card code. It is then accepted by the 701, converted into its own binary language and translated by means of stored dictionary and operational syntactical programs into the English language equivalent which is then printed.

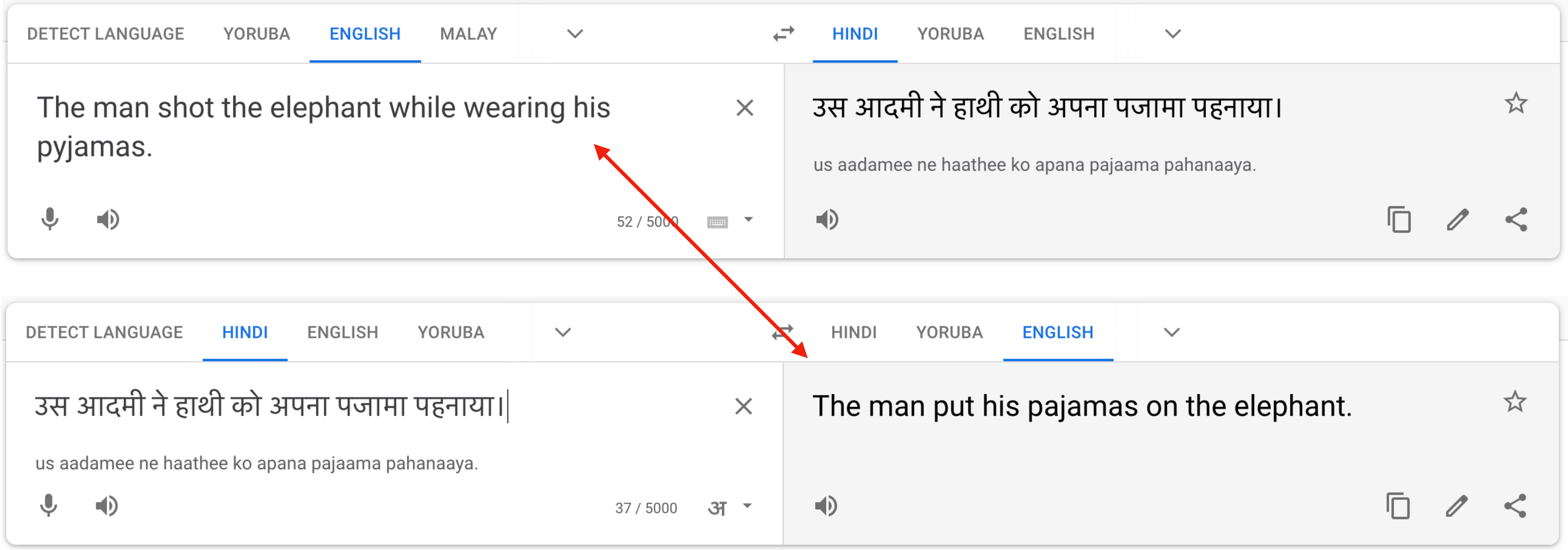
THE QUALITY OF COAL IS DETERMINED BY CALORY CONTENT

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



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.

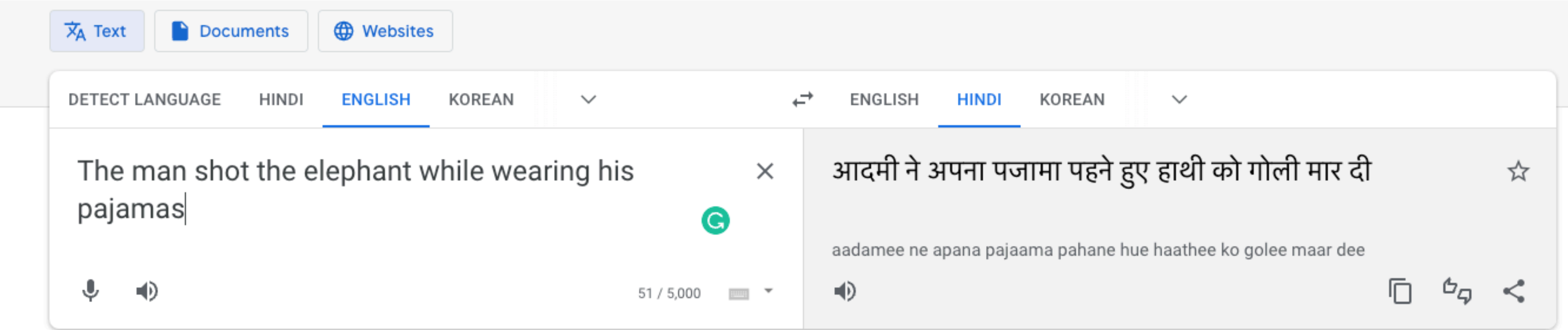
52 / 5000

उस आदमी ने हाथी को अपना पजामा पहनाया।

us aadamee ne haathee ko apana pajaama pahanaaya.

37 / 5000

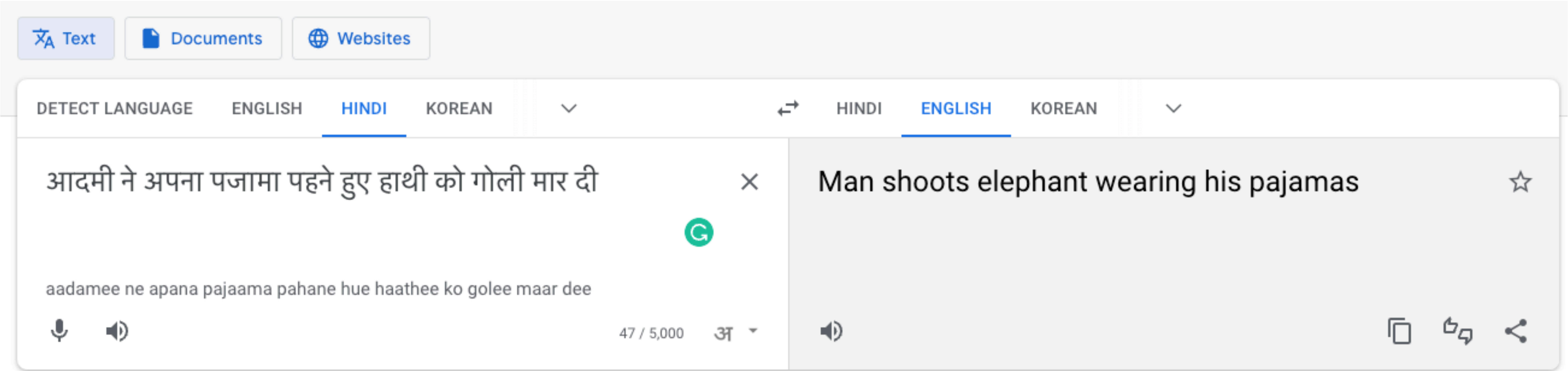
The man put his pajamas on the elephant.



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

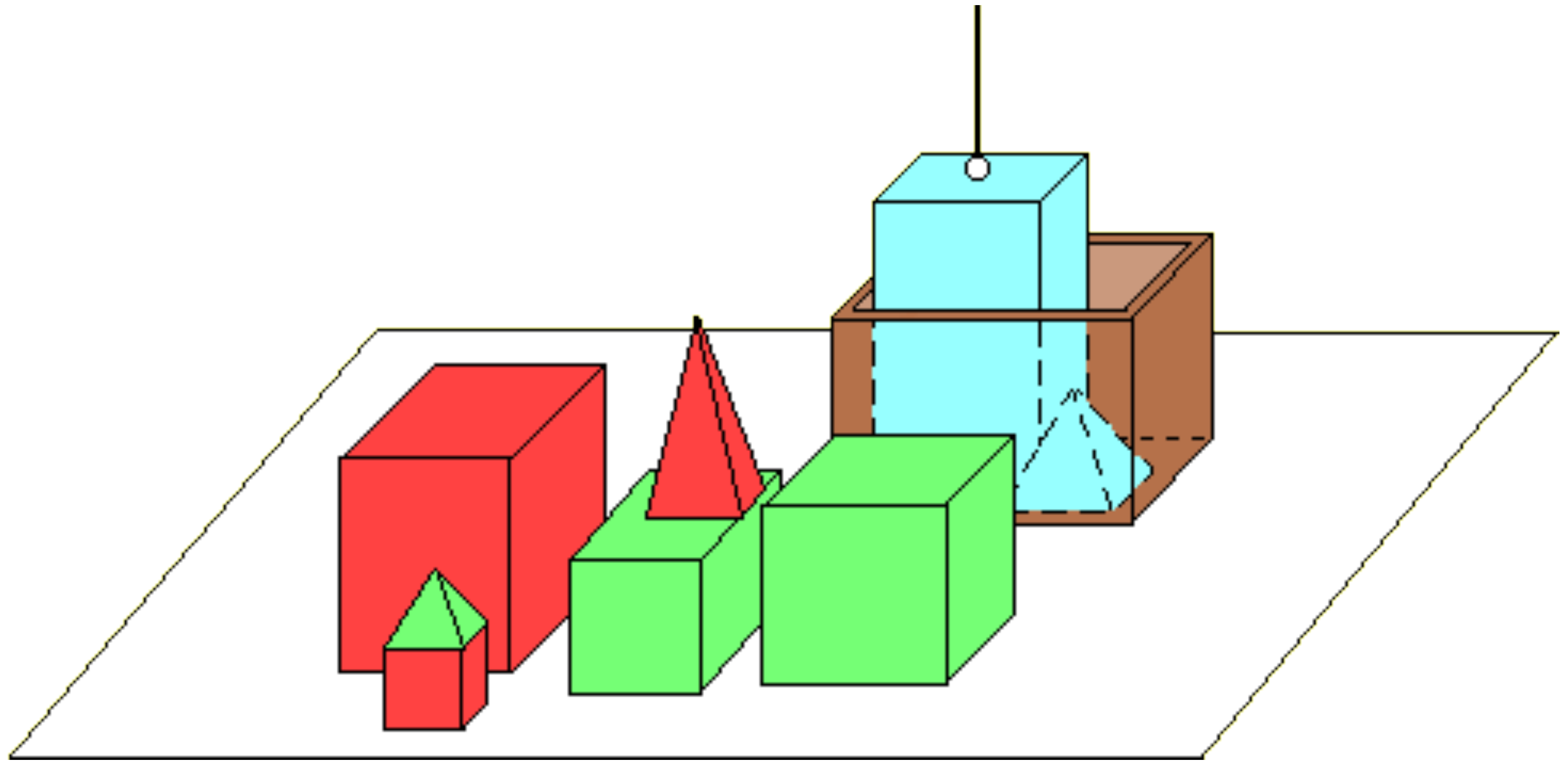


आदमी ने अपना पजामा पहने हुए हाथी को गोली मार दी

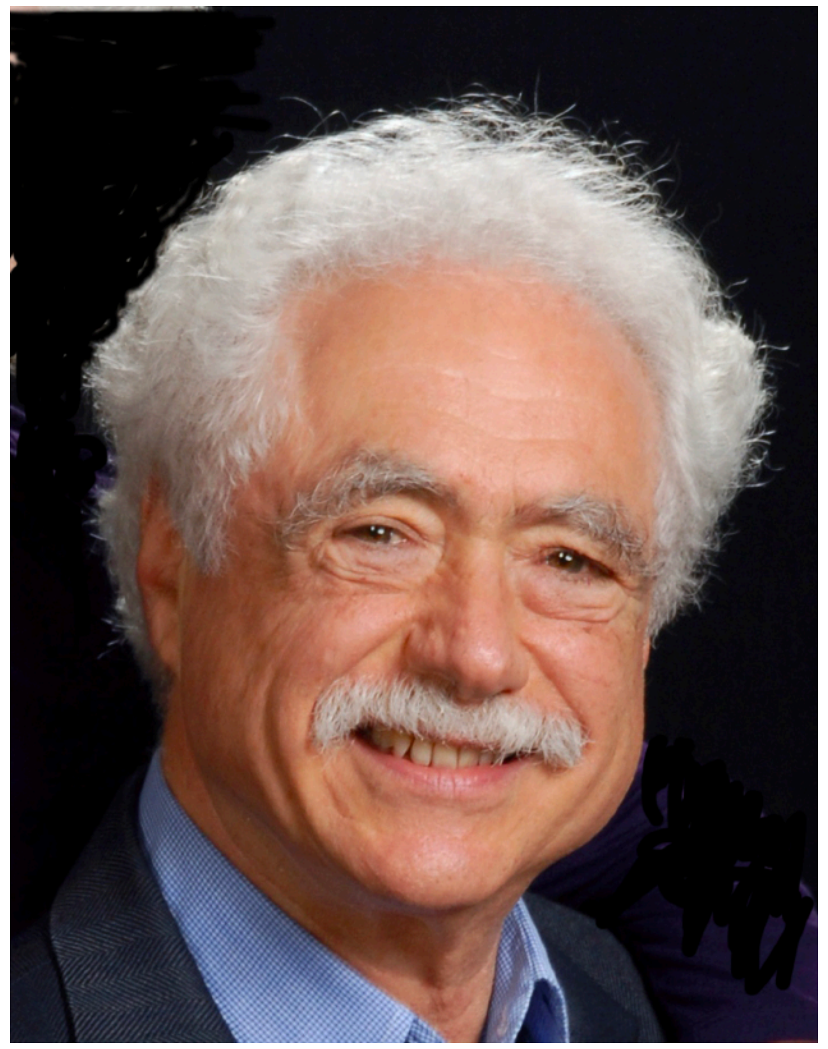
aadamee ne apana pajaama pahane hue haathee ko golee maar dee

Man shoots elephant wearing his pajamas

(Based on Google Translate result in 2023-01)



SHRDLU,
1968

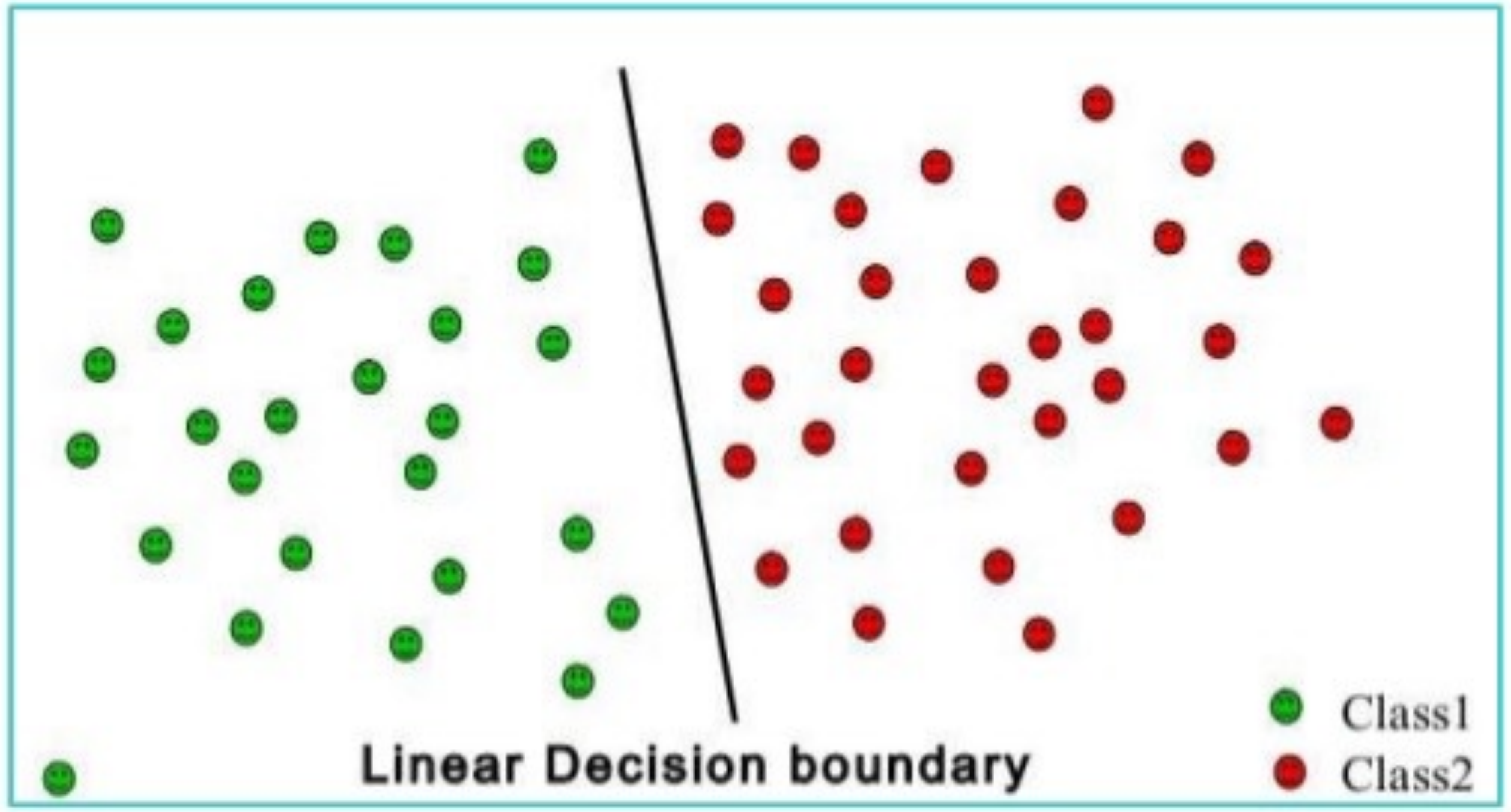


> How many red blocks are there?
- THREE OF THEM

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

- Rule-based, requiring extensive programming
- Limited domain

Statistical learning

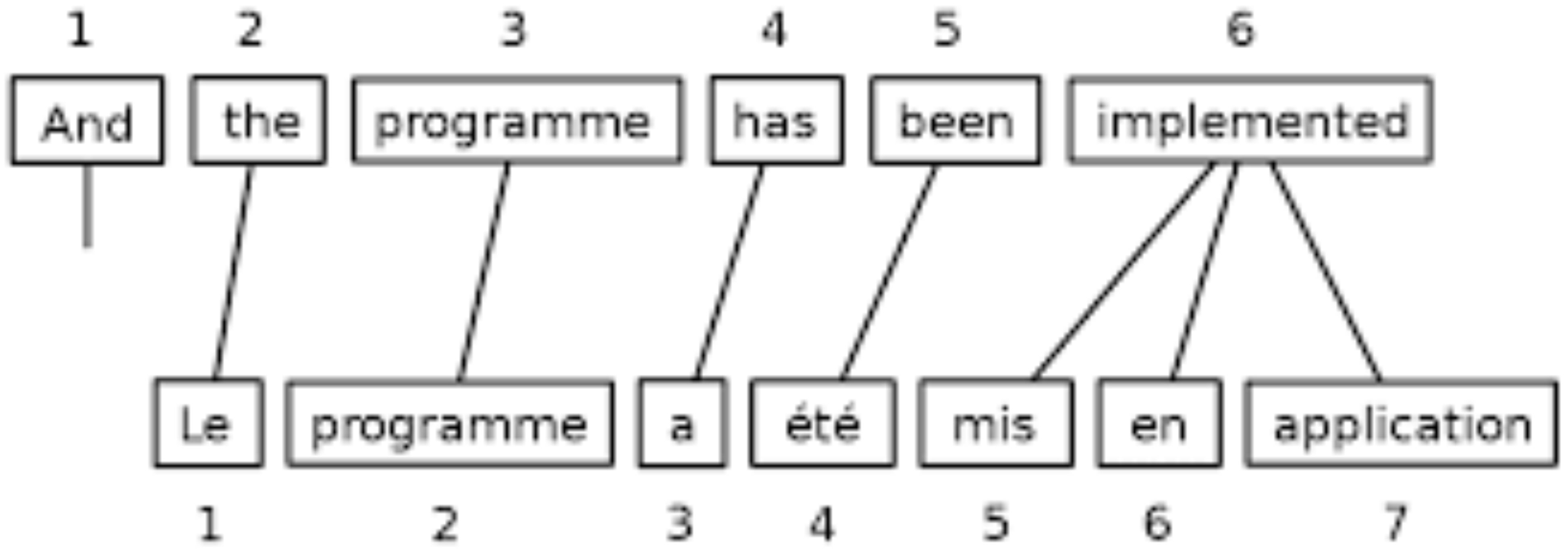


- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora

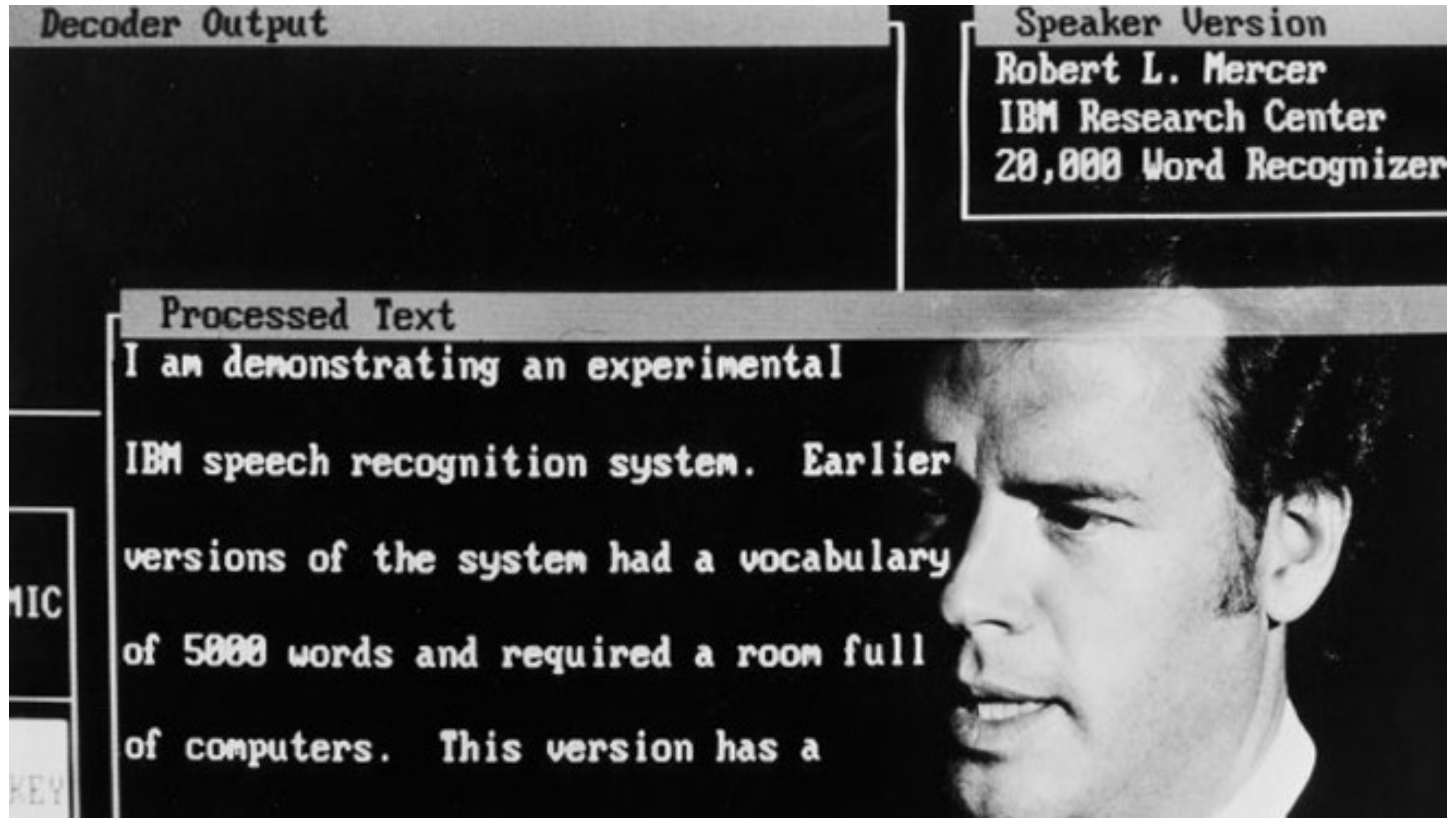
Unsupervised vs. supervised?

Statistical learning

IBM translation models



Speech recognition



Speaker Version
Robert L. Mercer
IBM Research Center
20,000 Word Recognizer

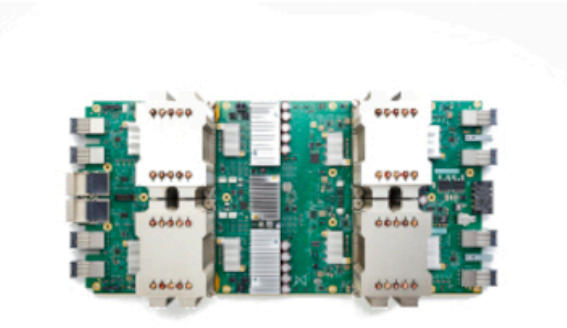
Processed Text
I am demonstrating an experimental
IBM speech recognition system. Earlier
versions of the system had a vocabulary
of 5000 words and required a room full
of computers. This version has a

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

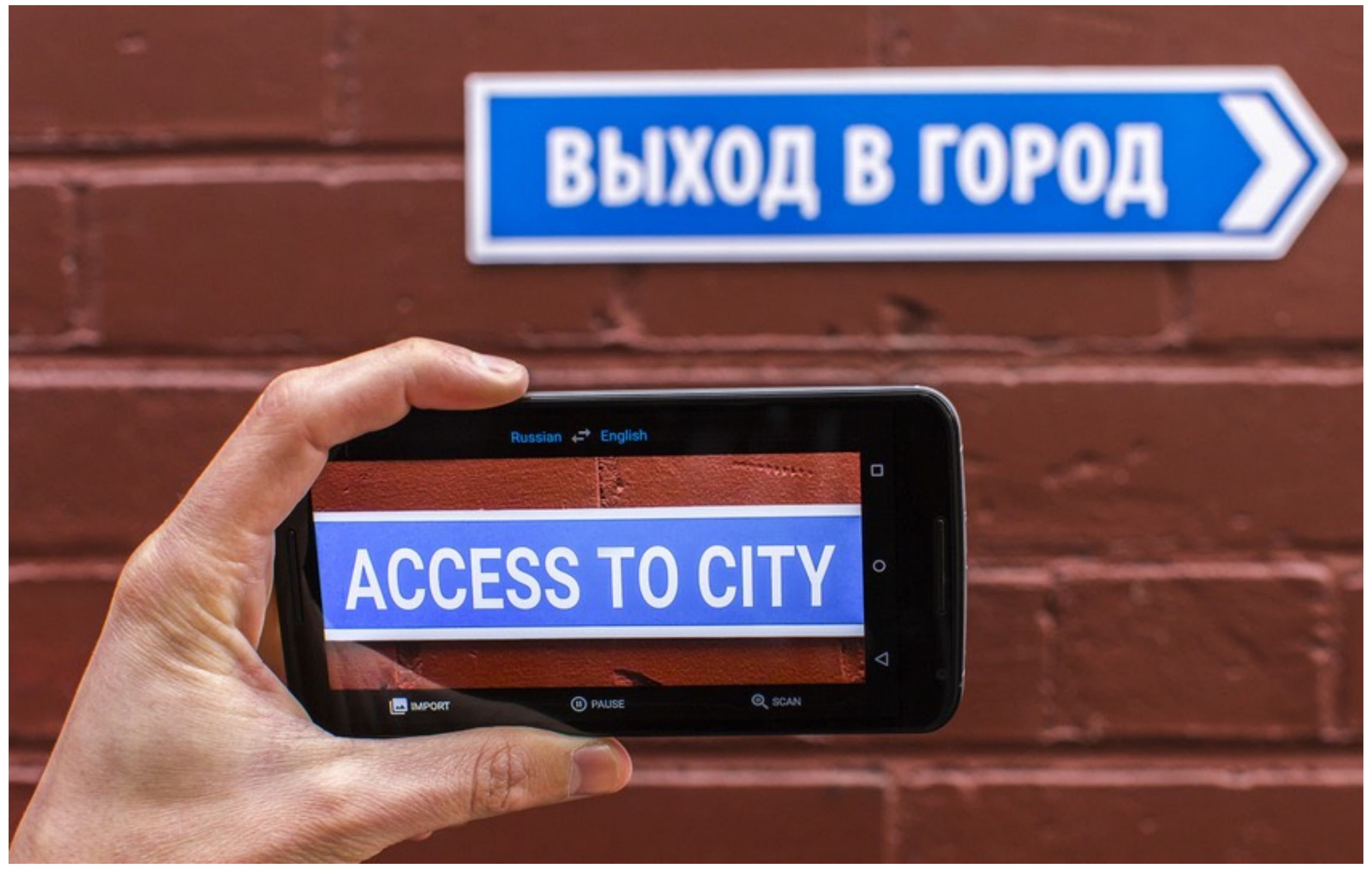
The era of deep learning

- 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

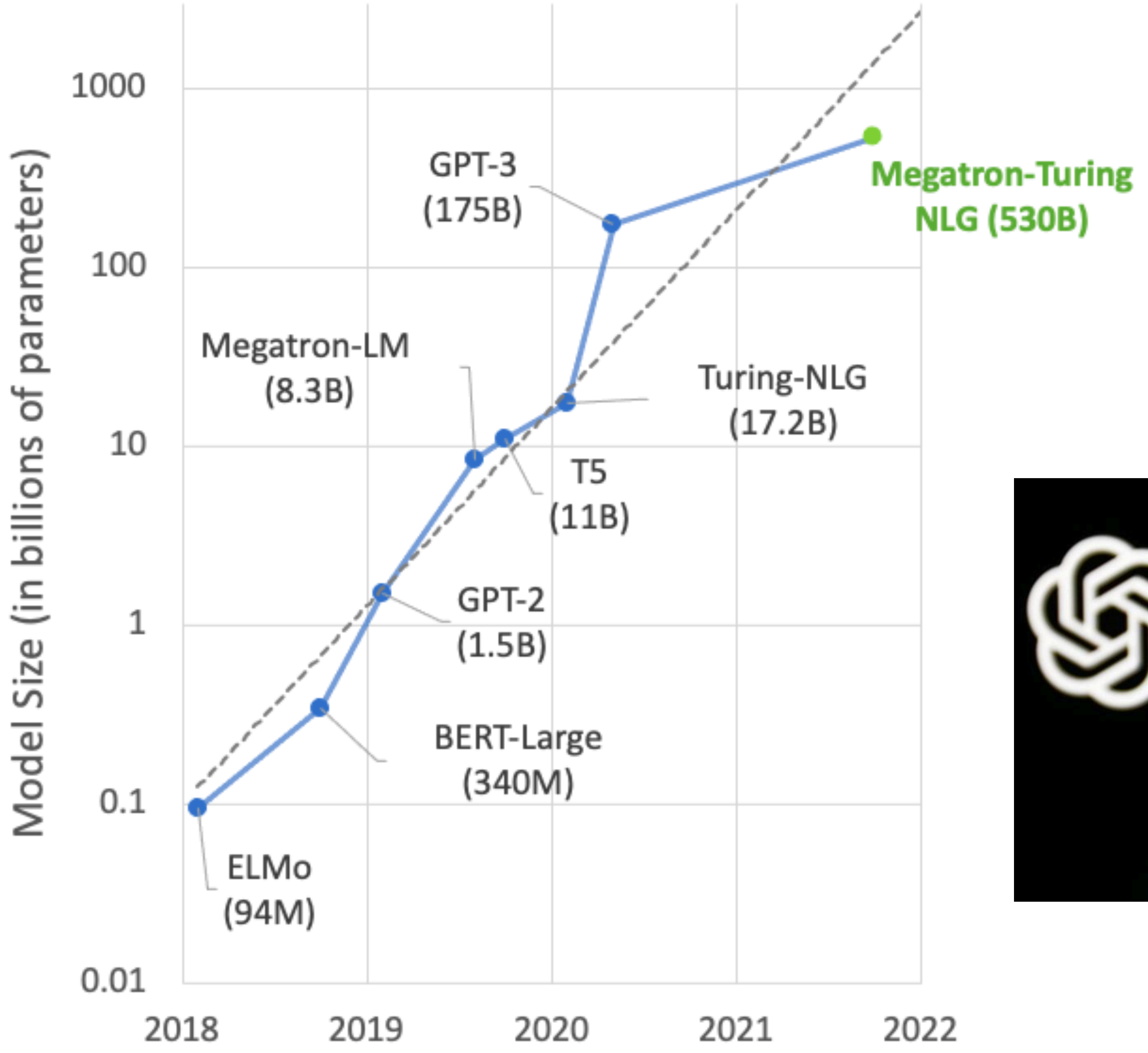
Russian: Машинный перевод - это круто!

↕

English: Machine translation is cool!

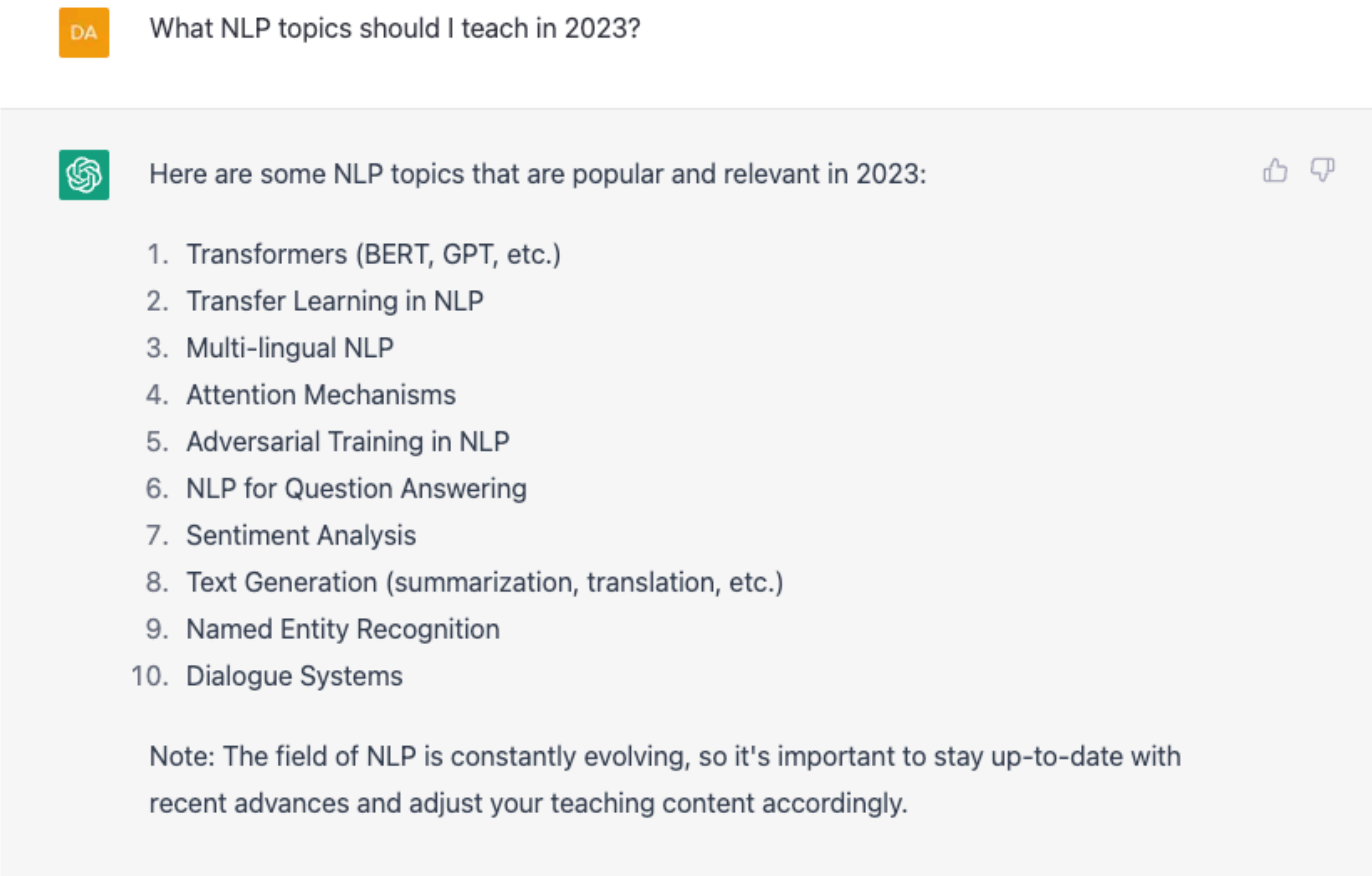
The era of pre-training / LLMs

BERT, ELMo, ERNIE…



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

ChatGPT examples



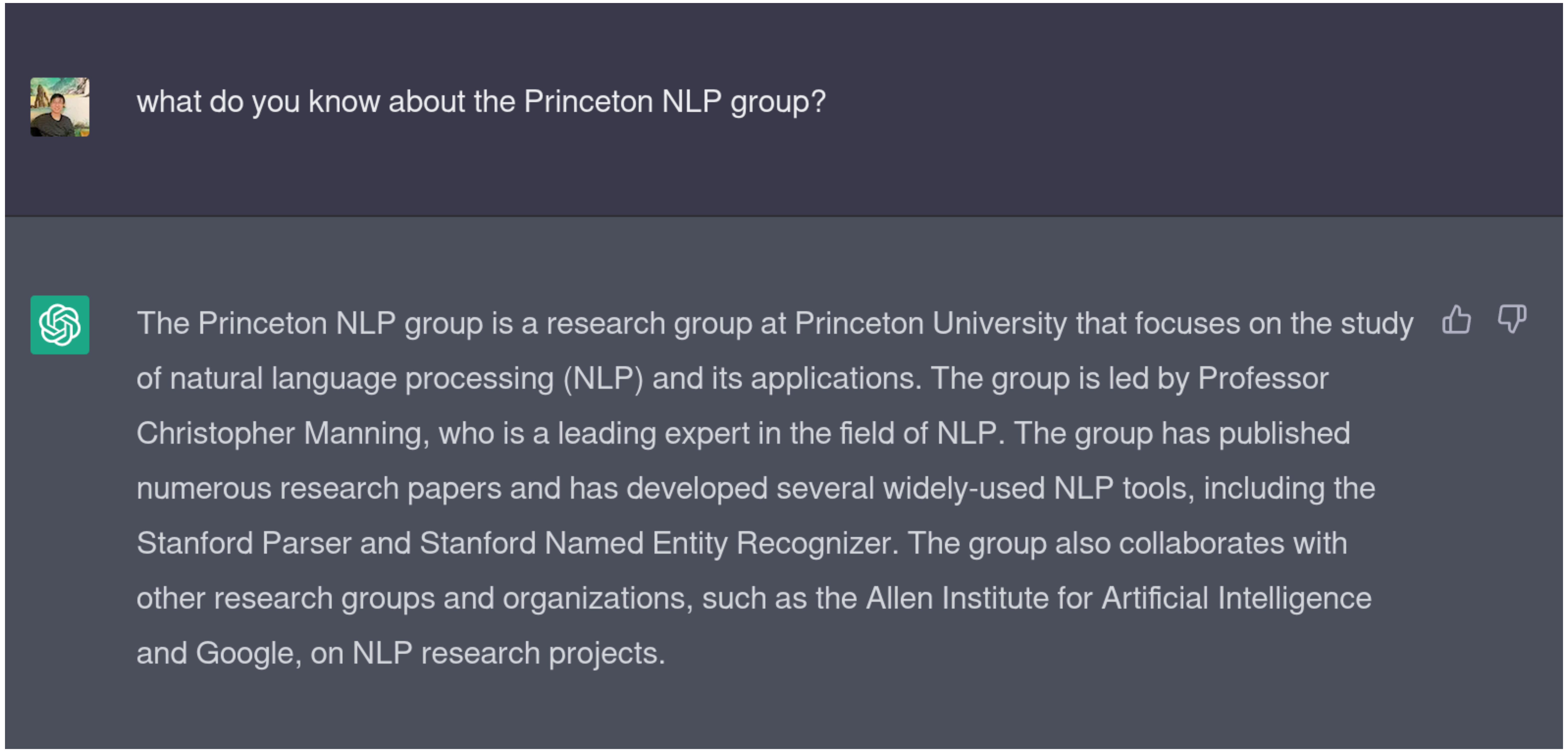
What NLP topics should I teach in 2023?

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

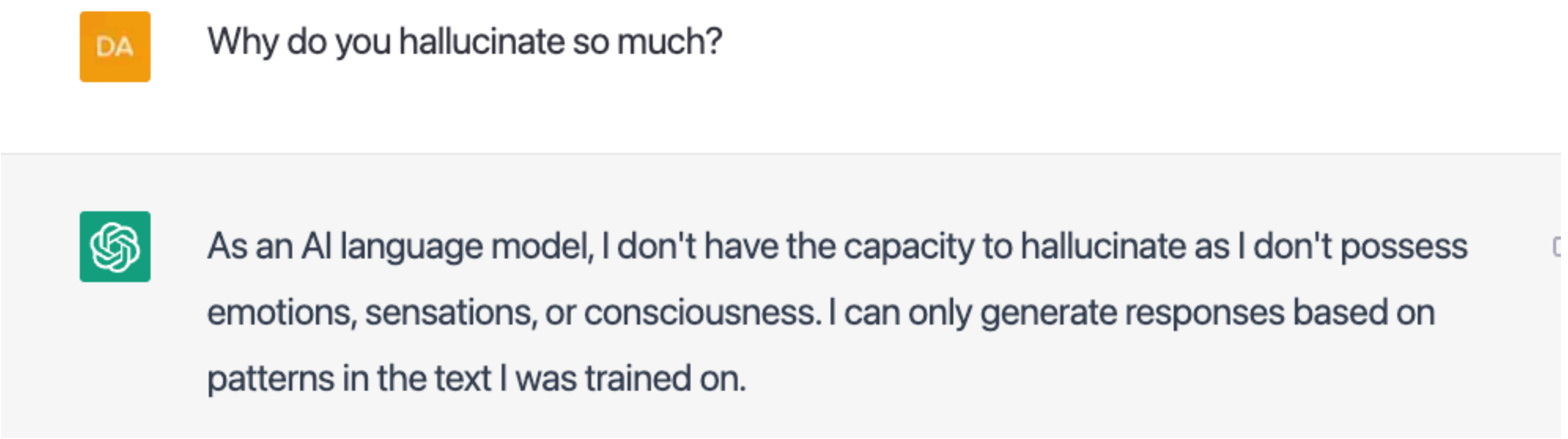
Note: The field of NLP is constantly evolving, so it's important to stay up-to-date with recent advances and adjust your teaching content accordingly.

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.

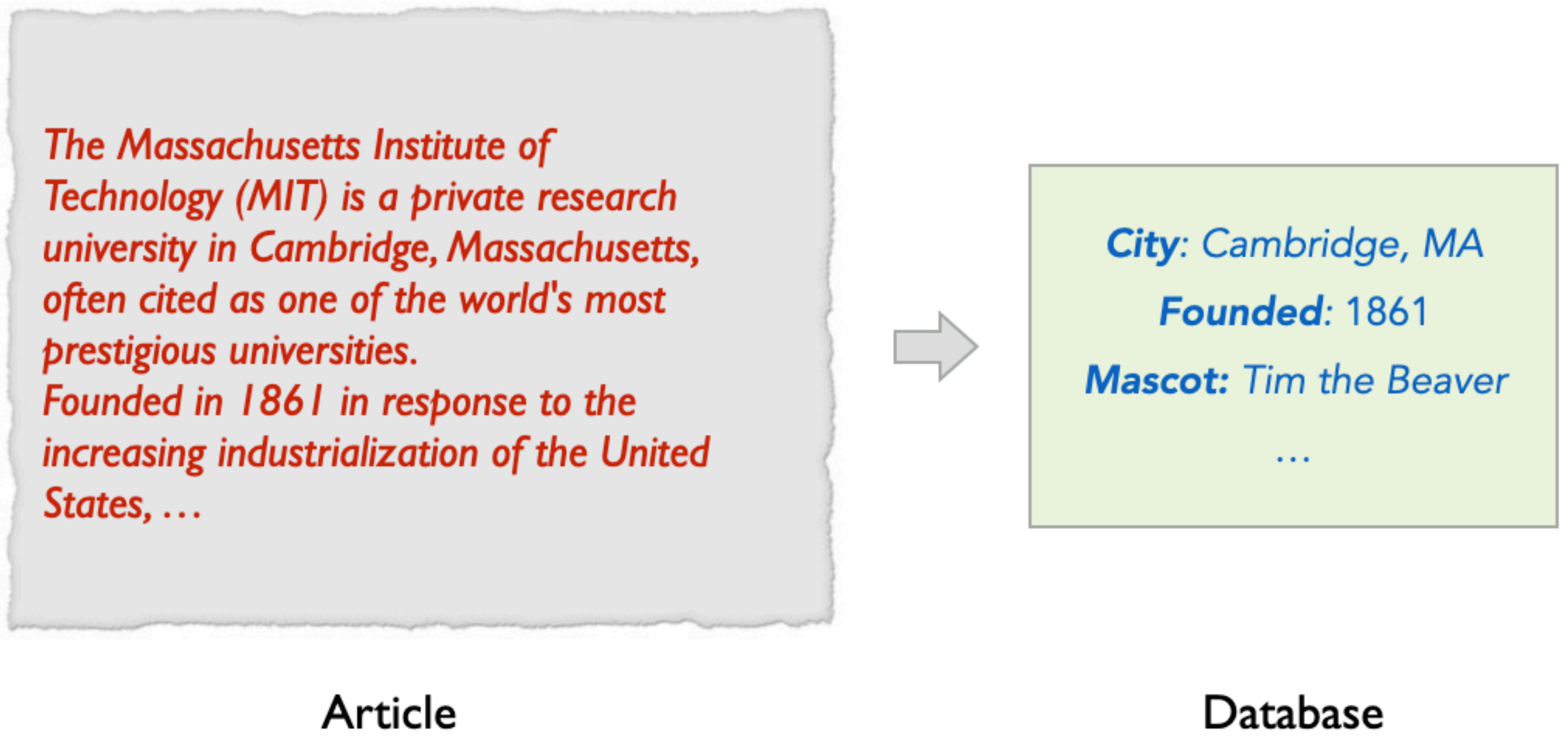


Why do you hallucinate so much?

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
- ….

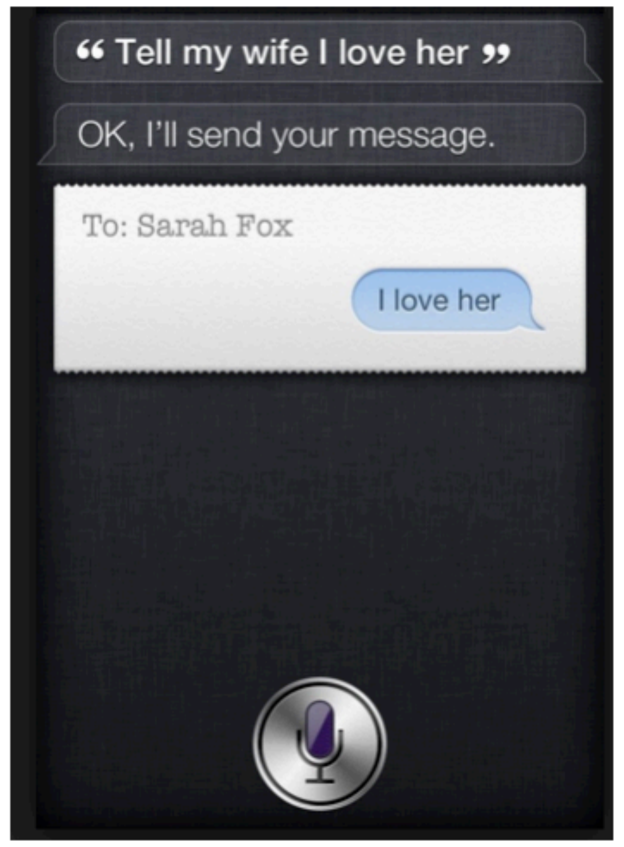


The Massachusetts Institute of Technology (MIT) is a private research university in Cambridge, Massachusetts, often cited as one of the world's most prestigious universities. Founded in 1861 in response to the increasing industrialization of the United States, …

***City**: Cambridge, MA*
***Founded**: 1861*
***Mascot:** Tim the Beaver*
…

Article

Database

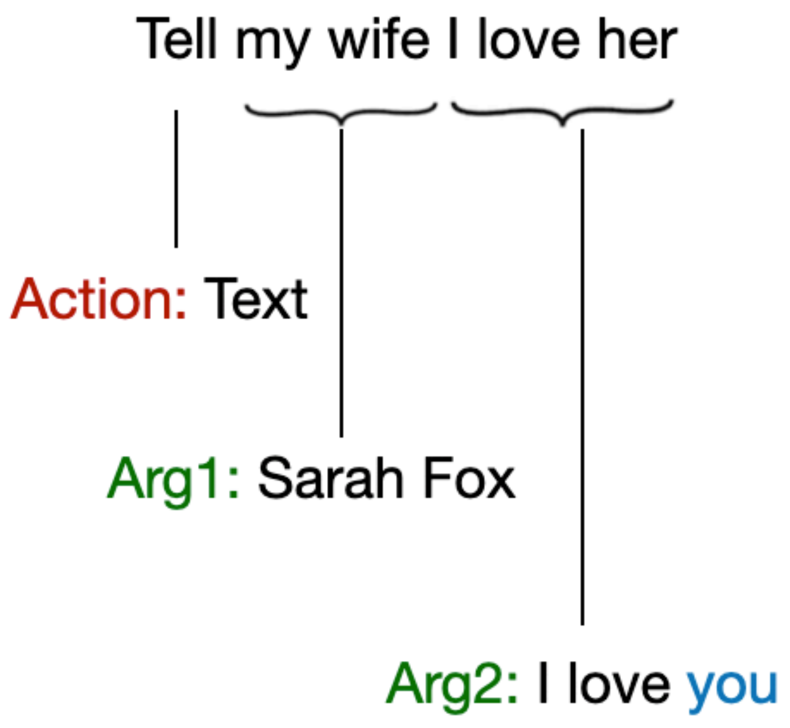


“ Tell my wife I love her ”

OK, I'll send your message.

To: Sarah Fox

I love her

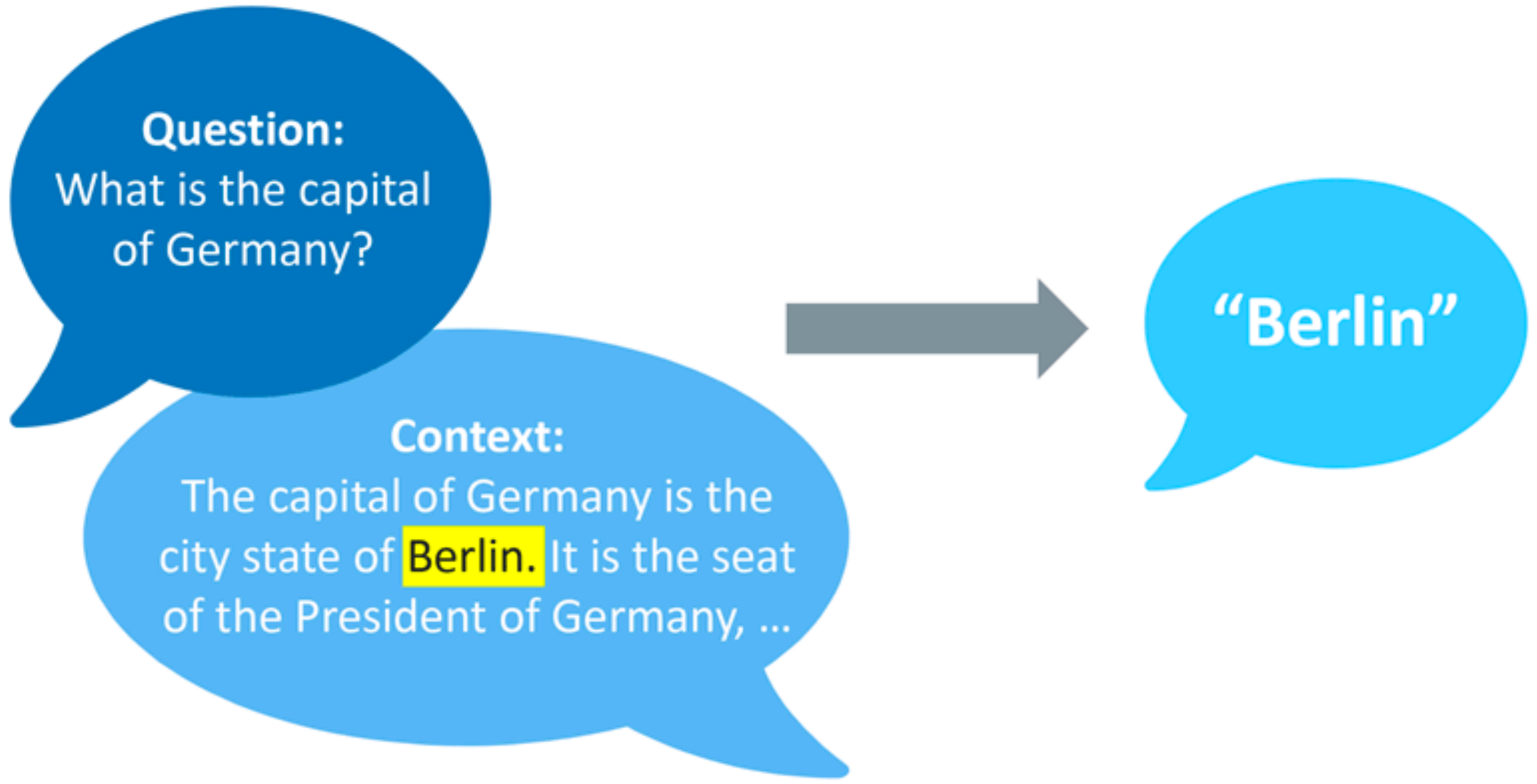


Tell my wife I love her

Action: Text

Arg1: Sarah Fox

Arg2: I love you



Question: What is the capital of Germany?

Context: The capital of Germany is the city state of Berlin. It is the seat of the President of Germany, ...

“Berlin”

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 $\times 2$
- Sequence models $\times 2$
- Neural networks for NLP
- Recurrent neural networks $\times 2$

Before midterm

- (neural) machine translation $\times 2$
- **Transformers** $\times 2$
- BERT, ELMo, …
- **Large language models**

+ guest lectures!

After midterm

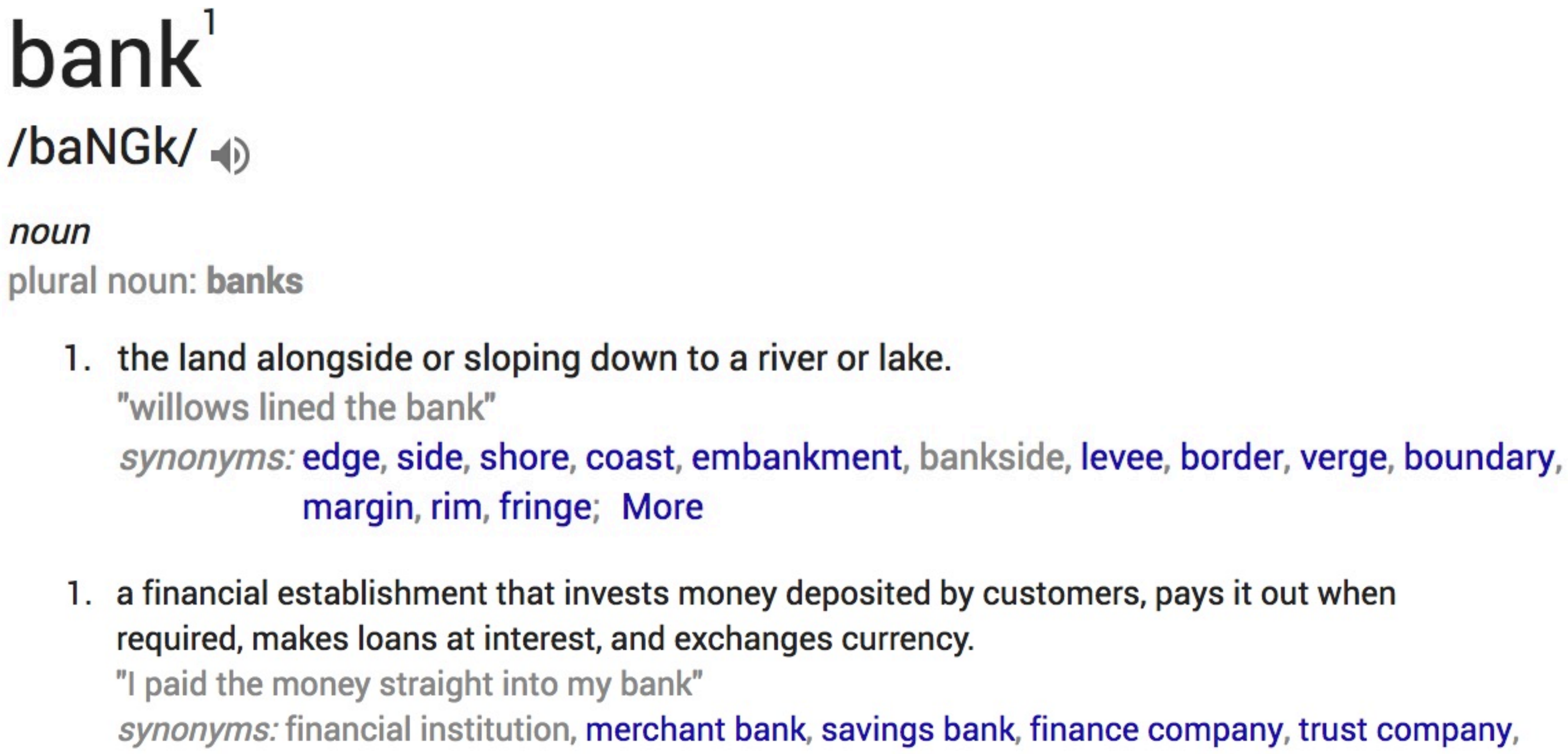
Why is language difficult to understand?

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[1]

/baNGk/ 🔊

noun

plural noun: **banks**

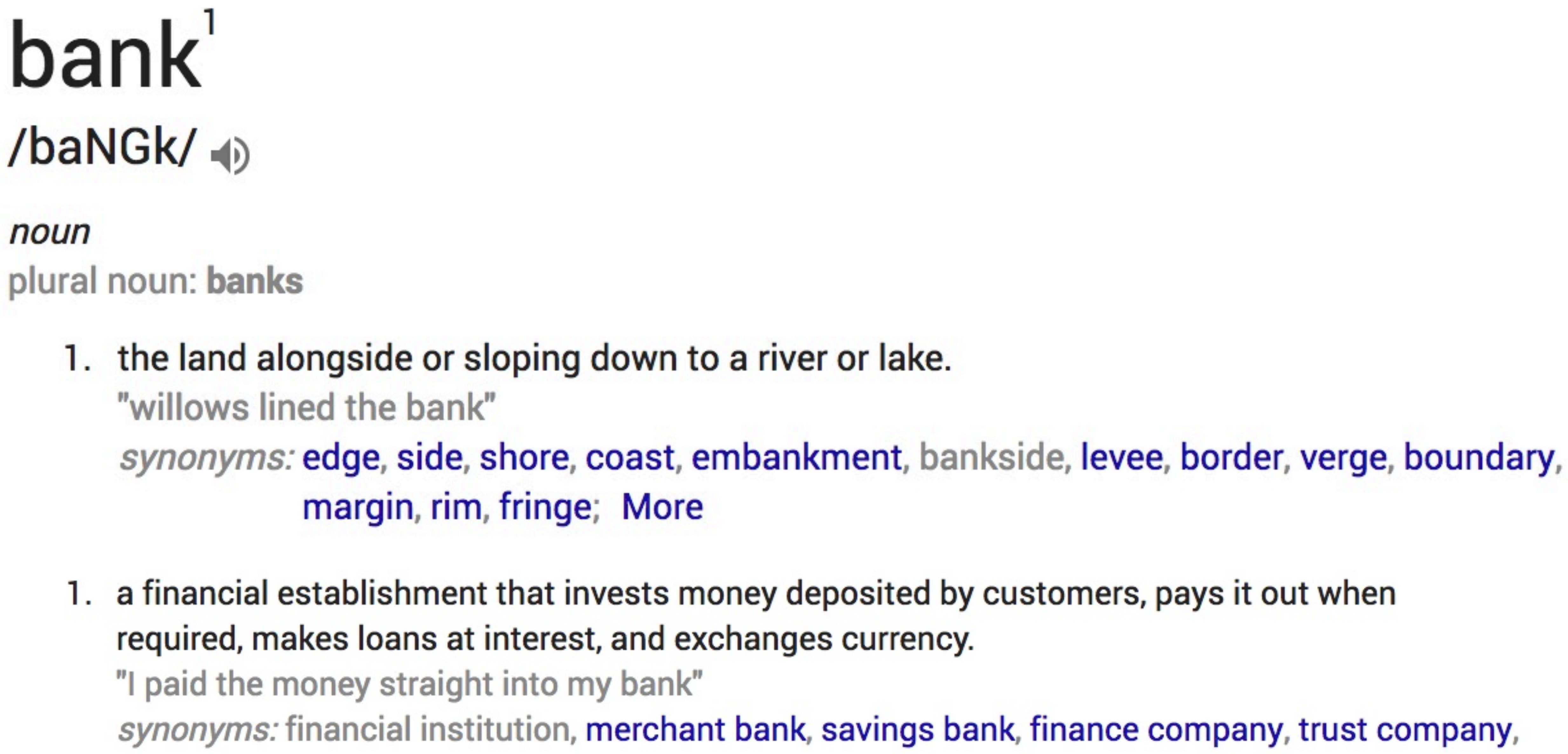
1. the land alongside or sloping down to a river or lake.
   "willows lined the bank"
   synonyms: edge, side, shore, coast, embankment, bankside, levee, border, verge, boundary, margin, rim, fringe; More

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^1

/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

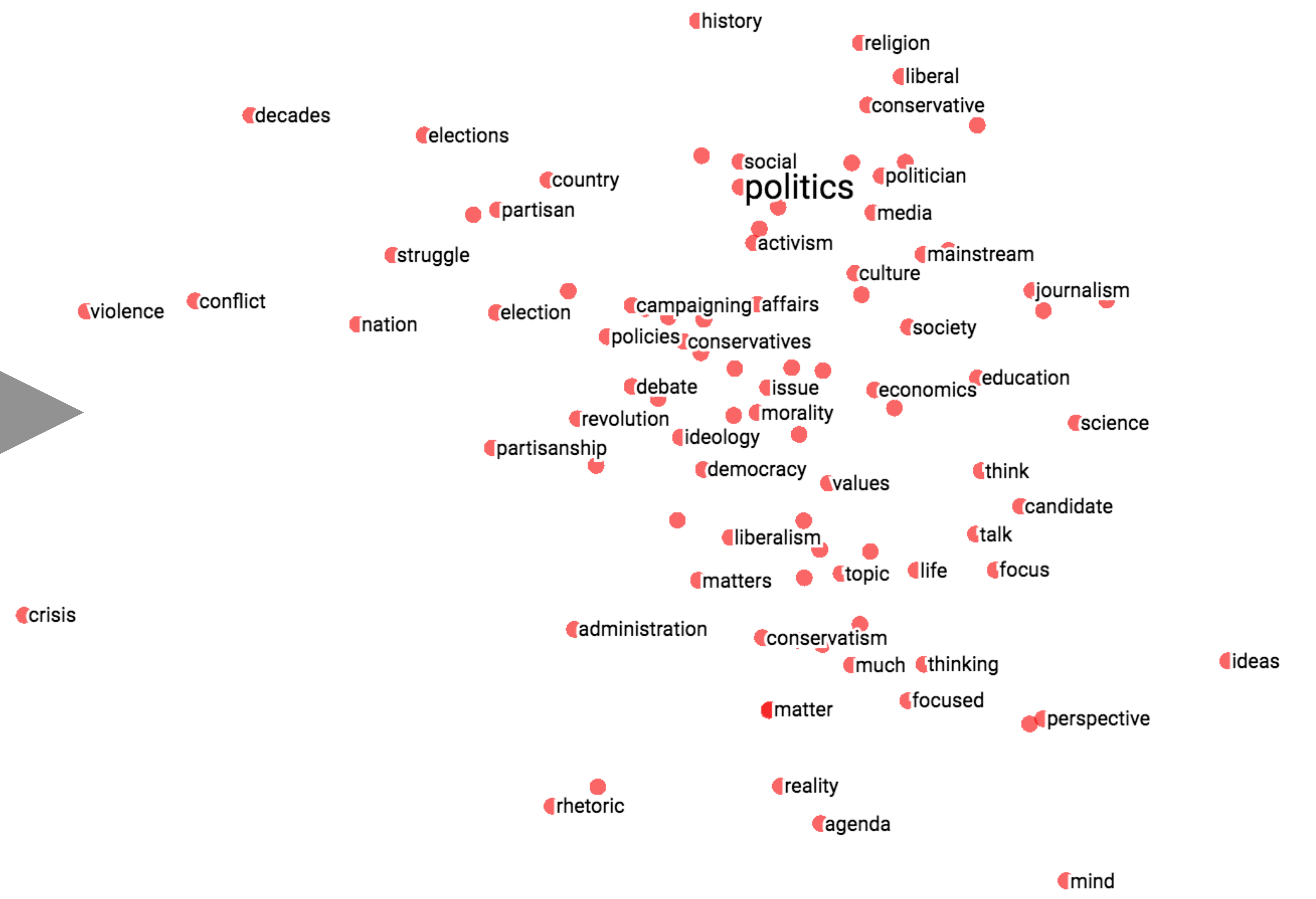
Lexical variations

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

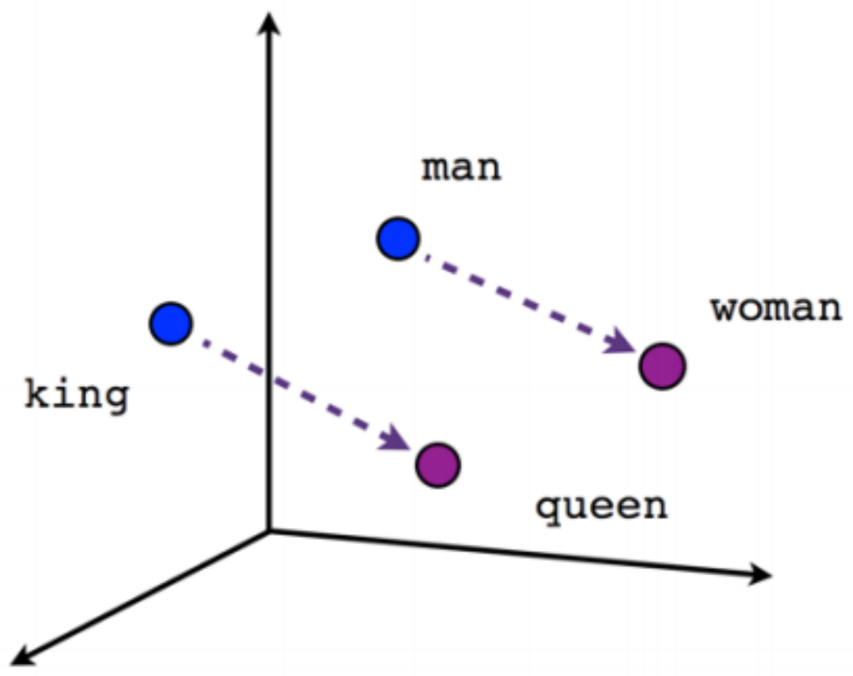
Several words can mean the same thing!

Distributed representations

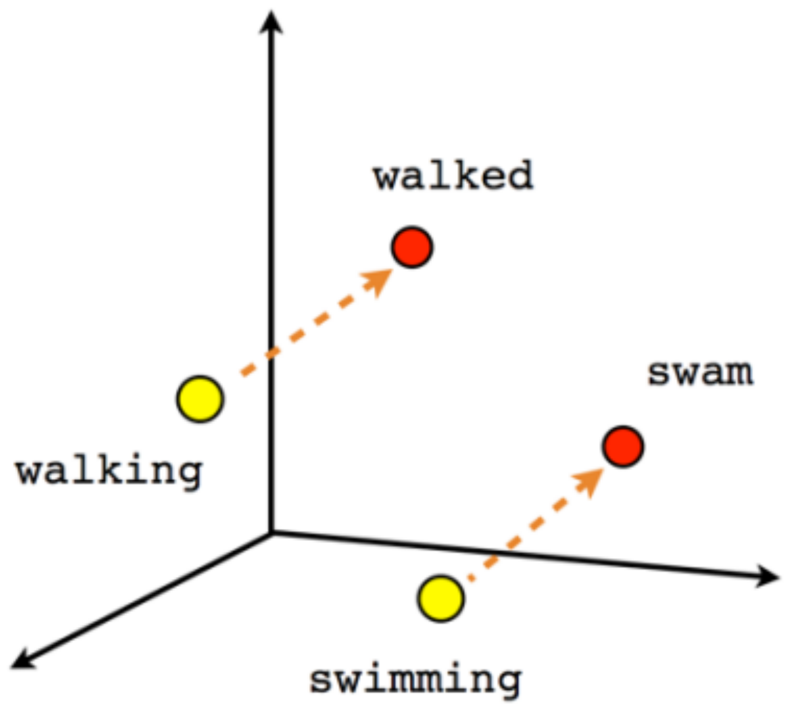
Project words onto a continuous vector space



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

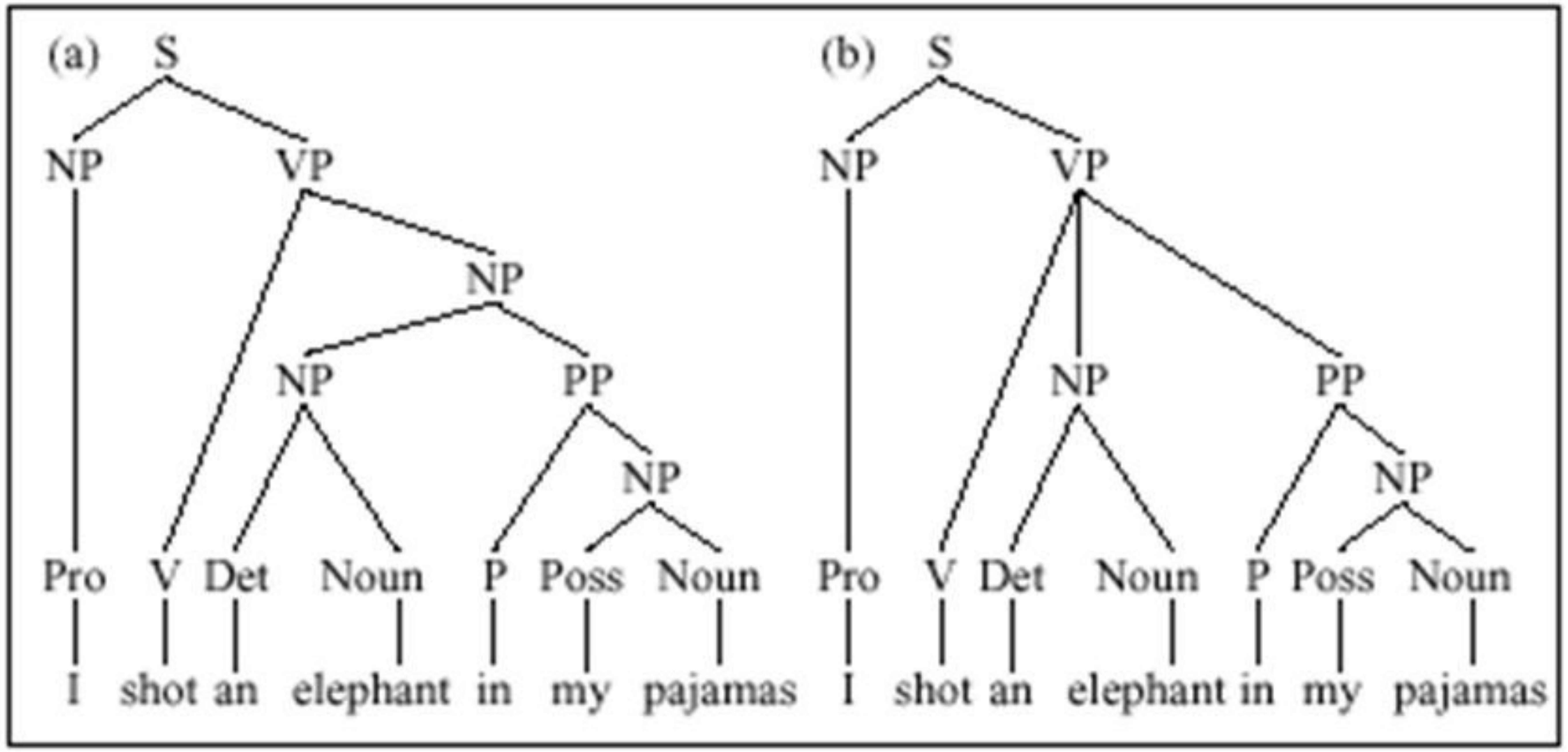
- My brother went to the park near my sister’s house
- Park my went house near to sister’s my brother the
- “My brother went park near sister’s house”?
- The old man the boat
- Fat people eat accumulates

Garden Path sentence

- Implicit structure in all languages
- Coarse-to-fine levels (recursive)
- What are some good data structures to represent this?

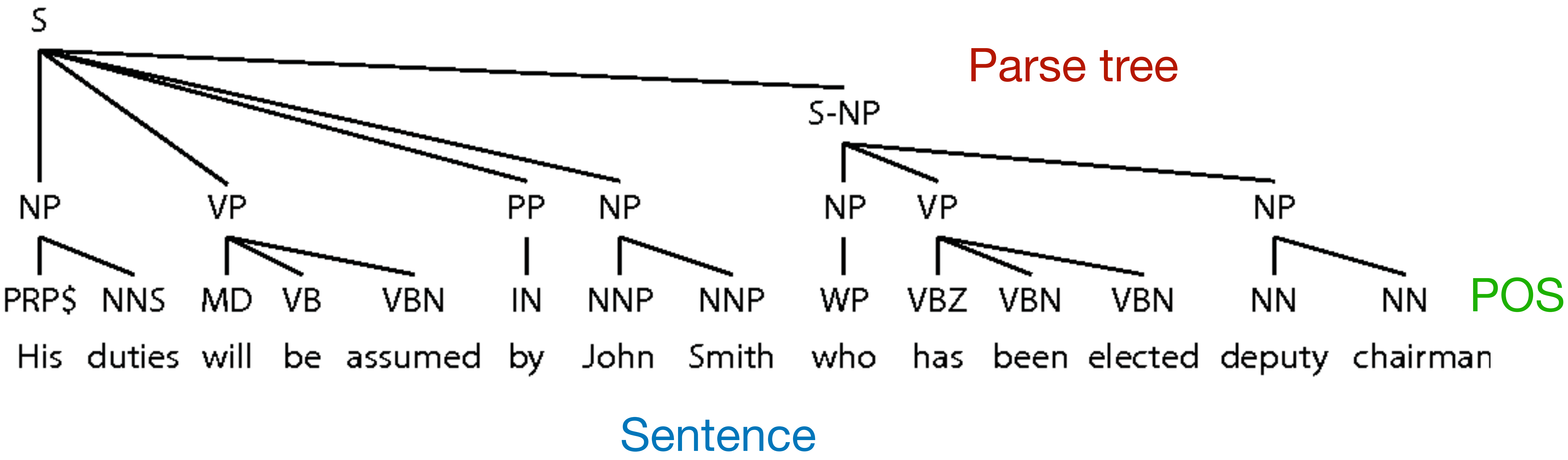
Syntactic ambiguity

I shot an elephant in my pajamas



Human language is full of such examples!

Syntactic parsing


S
S-NP
Parse tree
NP
VP
PP
NP
NP
VP
NP
PRP$
NNS
MD
VB
VBN
IN
NNP
NNP
WP
VBZ
VBN
VBN
NN
NN
POS
His duties will be assumed by John Smith who has been elected deputy chairman
Sentence

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