Introduction: AI and NLP, history and resources (PART ONE)

Linguistic Resources for Natural Language Processing
LM Language Technologies and Digital Humanities
2024-25

Cristina Bosco

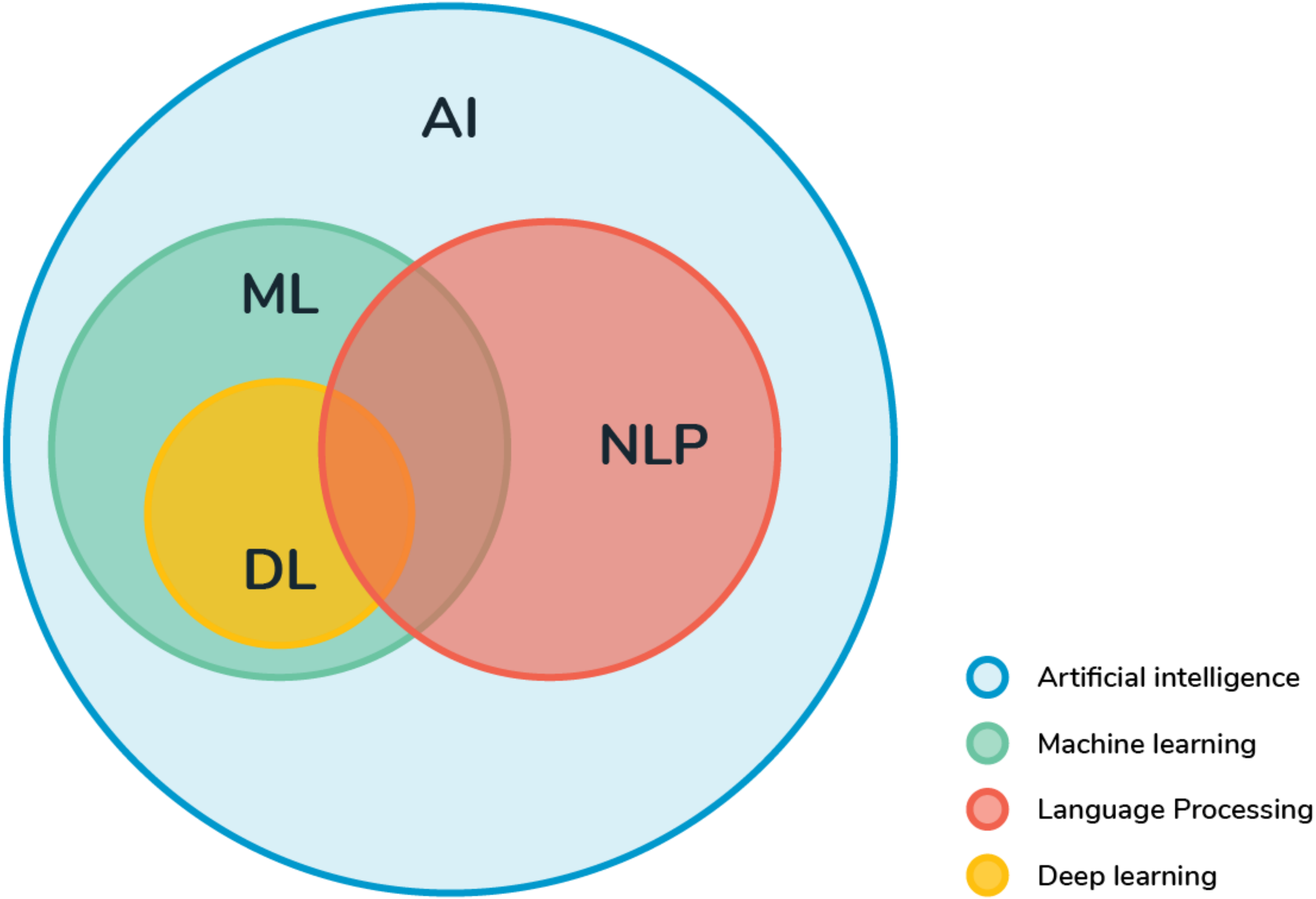
Overview

- What is NLP?
- A little bit of history of Artificial Intelligence: the notion of intelligence and the emergence of NLP
- Early NLP, Machine Translation and Dialog
- New ideas in NLP: Knowledge representation
- From symbolic to statistical processing

What is NLP?

NLP is the acronym of
Natural Language Processing
and the synonym of
Computational Linguistics.

NLP is the area of the Artificial Intelligence (AI)
that deals with
human language


AI
ML
DL
NLP
Artificial intelligence
Machine learning
Language Processing
Deep learning

What is the goal of the NLP?

What is the goal of the NLP?

NLP's main aim is at
creating machines able to speak and understand and
behave like humans using natural languages,
machine capable of simulating all
the linguistic behaviour of a human being.

It addresses therefore challenges related to
generating and understanding language.

AI and NLP

Natural language processing (i.e. understanding, generating, translating, …) is historically one of the first tasks set by scientists working in the field of **AI**.

But NLP has proven to be one of the most difficult tasks in AI !

NLP meaningfully changed during time and the notion of RESOURCE also.

AI and NLP

AI was born in the early 1950s immediately after the end of the II World War.

AI was necessarily **modest** in its goals, constrained as it was by:

- limitations of **hardware and software** (machines with very limited computing power)

- **memories** with limited storage capacity and realised on devices with slow access time

- the unavailability of high-level **programming languages**.

AI and NLP

In summer 1956 a 2-month workshop, the **Dartmouth Summer Research Project on Artificial Intelligence**, was organised which is the founding event of AI and the place where the term Artificial Intelligence was used for the first time.

AI scientists were convinced that any intelligent behavior, such as the use of human language, could be simulated by a machine

The goal of the workshop is

> *"… to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."*

https://en.wikipedia.org/wiki/Dartmouth_workshop

The meaning of *intelligence*

The possibility of building intelligent machines opened a debate about the meaning of *intelligence* … a debate that is still ongoing!

In 1950, in the paper "COMPUTING MACHINERY AND INTELLIGENCE" [1] Alan Mathison Turing (1912 - 1954) proposed an operational definition of intelligence with the famous Turing test, also called i*mitation game*.

[1]: https://academic.oup.com/mind/article/LIX/236/433/986238

Turing test

The imitation game is played with three people:

- a man (A)
- a woman (B)
- an interrogator (C) who may be of either sex.

The **object** of the game **for C is to determine which between A and B is the man and which is the woman**, by asking questions.

Therefore C stays in a room apart from the other two and knows A and B only by the labels X and Y.
C's questions and A's and B's answers are written and displayed on a monitor, in order that tones of voice may not help C.

Turing test

At the end of the game C must say either "X is A and Y is B" or "X is B and Y is A”.

C is allowed to put questions to X and Y thus:

"Will X please tell me the length of his or her hair?”

It is **A's object** in the game to try and **cause C to make the wrong identification**.

The answer about the hair might therefore be: "*My hair is shingled, and the longest strands are about nine inches long.*”

Turing test

The **object** of the game for the third player **B is** instead **to help the interrogator C**.

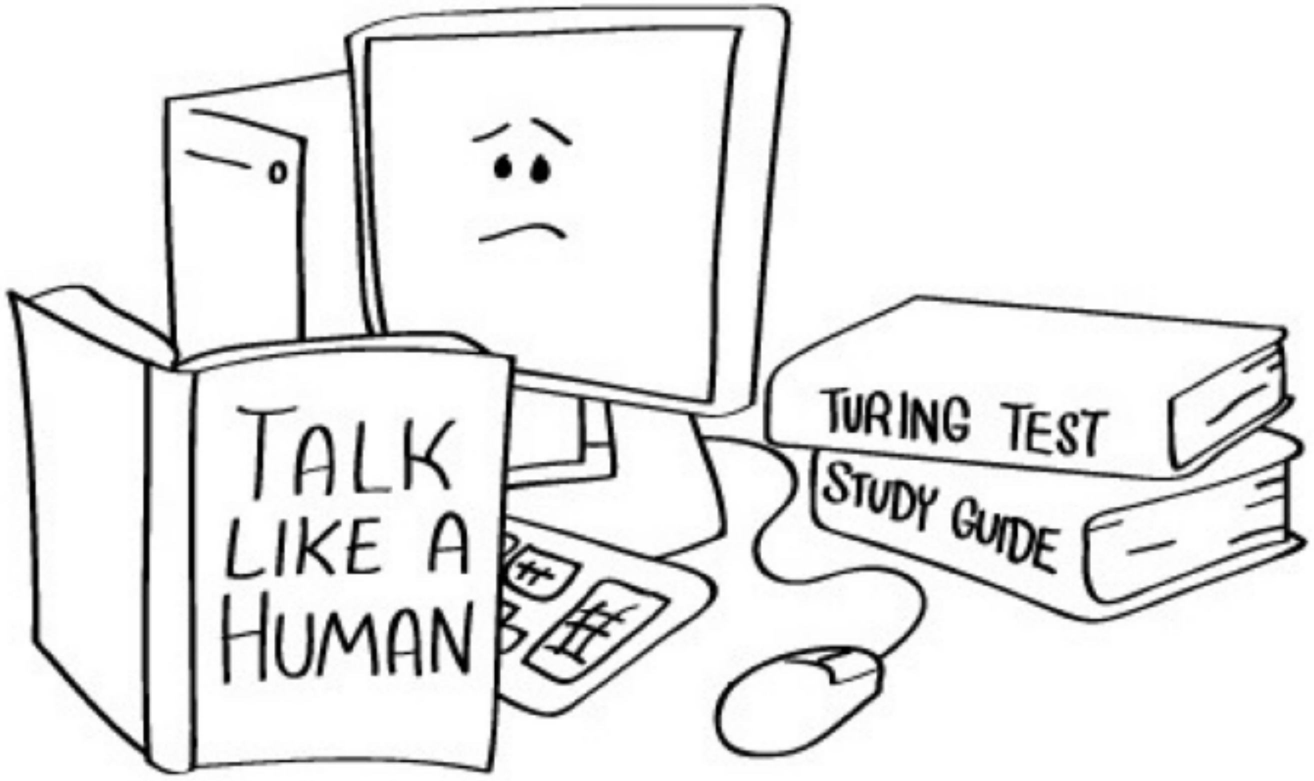
The best strategy for her is probably to give truthful answers.

She can add such things as "I am the woman, don't listen to him!" to her answers, but it will avail nothing as the man can make similar remarks.

Turing test

But what will happen when a machine takes the part of A or B in this game?

If the interrogator decide wrongly as often when the game is played by humans, can we consider that machine *intelligent*?



Turing test

TURING TEST EXTRA CREDIT:
CONVINCE THE EXAMINER THAT HE'S A COMPUTER.
YOU KNOW, YOU MAKE SOME REALLY GOOD POINTS.
I'M ... NOT EVEN SURE WHO I AM ANYMORE.

Turing test

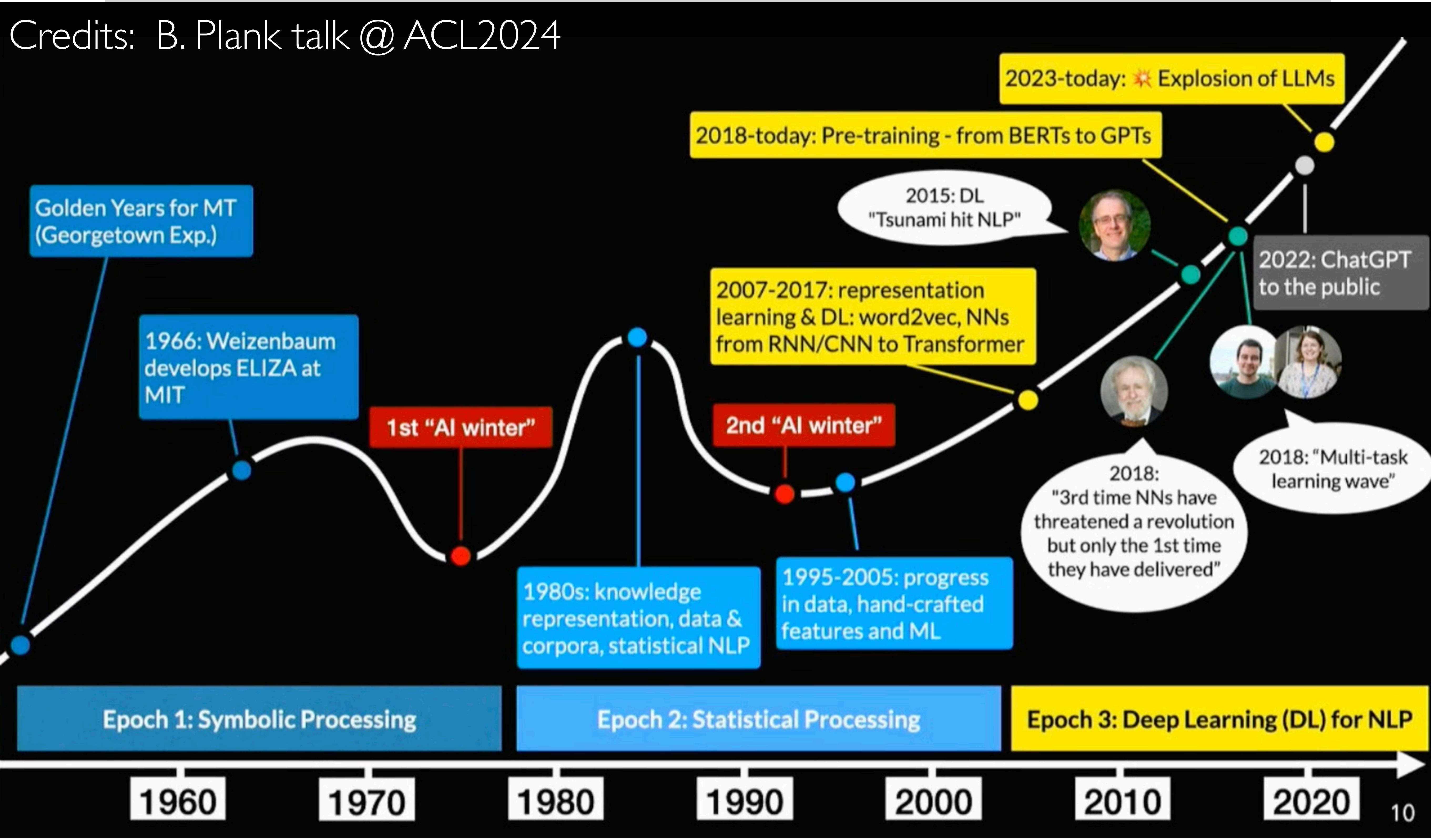
A definitive answer to the question "what is intelligence?" has not been given until now.

What is particularly interesting, however, is that the Turing test has given us an operational **definition of intelligence based** strictly **on linguistic ability**.

The fact that only intelligent beings, like humans, can use language led to a particular focus of AI on language
and
the emergence of NLP.

NLP: history at glance

Credits: B. Plank talk @ ACL2024



Golden Years for MT (Georgetown Exp.)

1966: Weizenbaum develops ELIZA at MIT

1st "AI winter"

1980s: knowledge representation, data & corpora, statistical NLP

2nd "AI winter"

1995-2005: progress in data, hand-crafted features and ML

2007-2017: representation learning & DL: word2vec, NNs from RNN/CNN to Transformer

2015: DL "Tsunami hit NLP"

2018: "3rd time NNs have threatened a revolution but only the 1st time they have delivered"

2018-today: Pre-training - from BERTs to GPTs

2018: "Multi-task learning wave"

2022: ChatGPT to the public

2023-today: 💥 Explosion of LLMs

Epoch 1: Symbolic Processing

Epoch 2: Statistical Processing

Epoch 3: Deep Learning (DL) for NLP

NLP history versus tasks

We can observe the history of NLP at glance on the perspective of tasks:

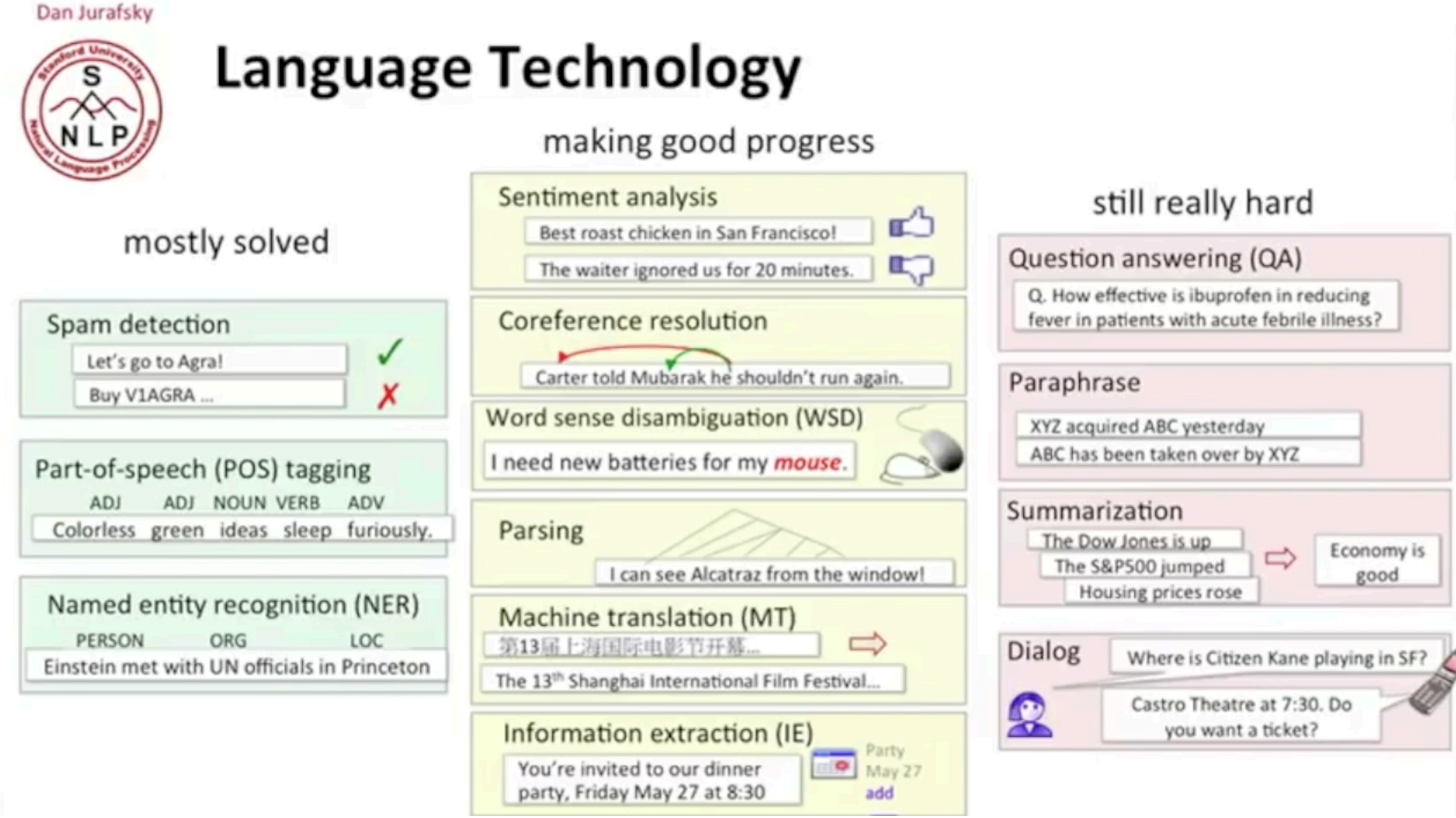
- In 1950-1960 NLP focuses on **Machine Translation**
- In 1960-1970 NLP focuses on **Dialog**
- In 1970-1990 NLP focuses on **Knowledge representation**
- In 1990- 2000 NLP focuses on **Parsing** (morphological and syntactic analysis)
- From 2000 NLP opens to **several different tasks**

NLP history versus tasks

… today NLP is focused on several different tasks …

Which ones?

NLP: today tasks



Dan Jurafsky

Language Technology

mostly solved

Spam detection

- Let's go to Agra! ✓
- Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON			ORG		LOC
Einstein	met	with	UN	officials in	Princeton

making good progress

Sentiment analysis

- Best roast chicken in San Francisco!
- The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

⇨ The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

- XYZ acquired ABC yesterday
- ABC has been taken over by XYZ

Summarization

- The Dow Jones is up
- The S&P500 jumped
- Housing prices rose

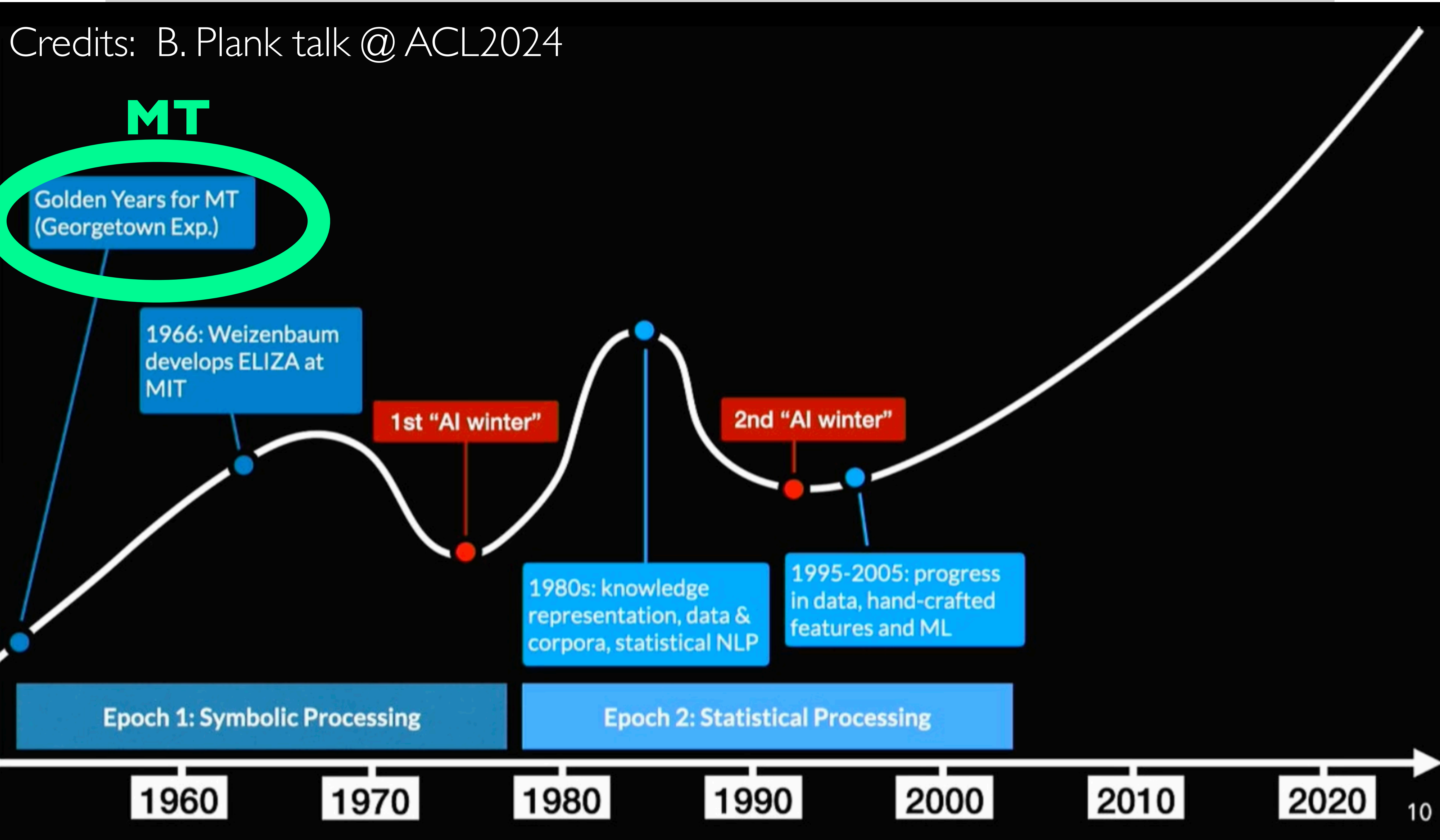
⇨ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

NLP: Machine Translation



Credits: B. Plank talk @ ACL2024

NLP and MT

When NLP was in its infancy, it was only focused on **Machine Translation** (MT).

The major stimulus to the development of MT was the memorandum "*On translation*" published in July 1949 by **Warren Weaver**. A widespread optimism stemmed from the war-time success in code-breaking, from developments by Shannon in information theory and from speculations about universal principles underlying natural languages.

Only English and Russian were addressed at that time, i.e. the languages of the two countries mostly engaged in the cold war.

NLP and MT

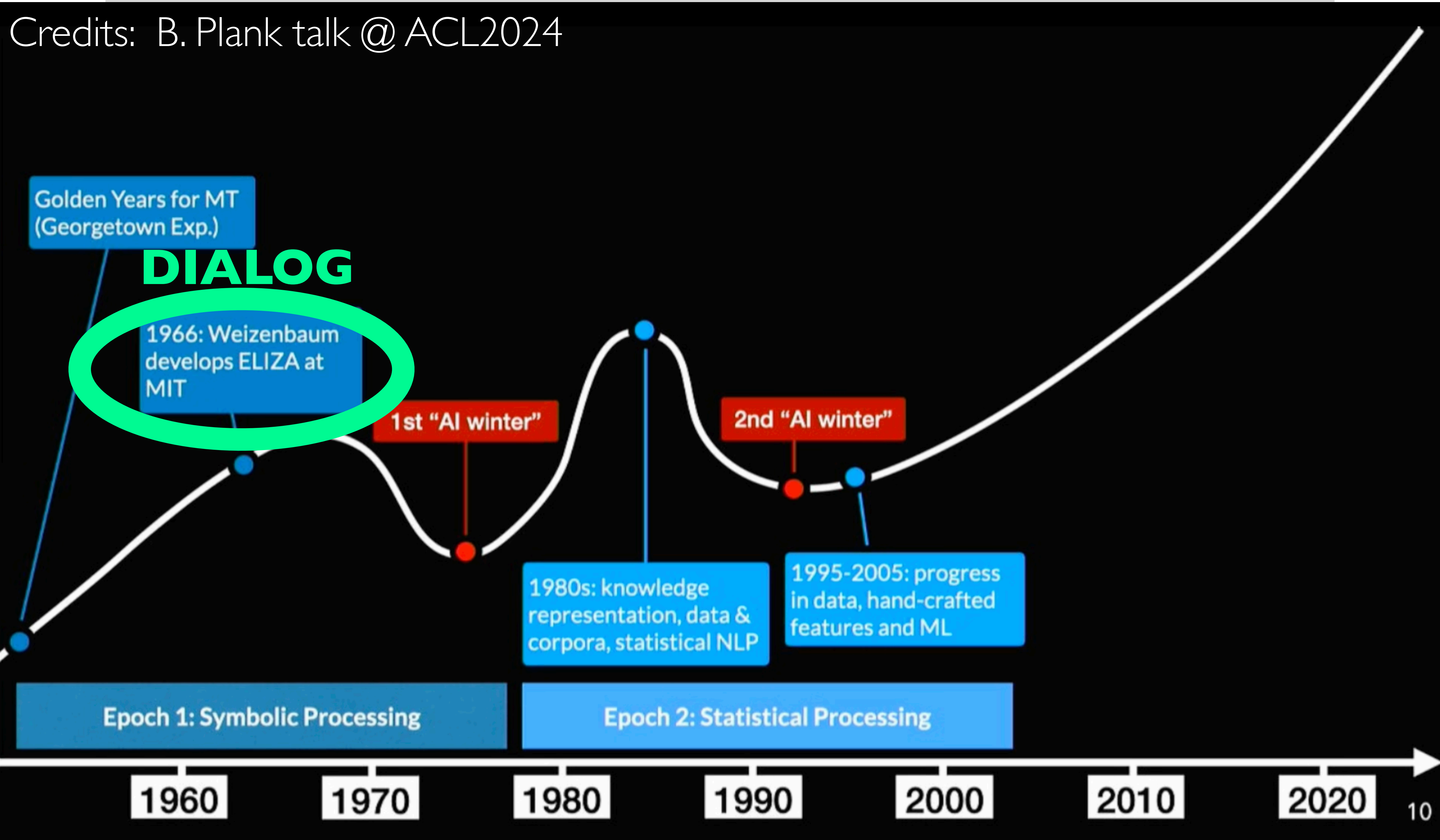
The first MT systems apply crude **dictionary based approaches**:
- word-for-word translation
- statistical methods advocated alongside cryptography.

The earliest systems consisted primarily of large **bilingual dictionaries** where entries for words of the source language gave one or more equivalents in the target language and some rules for producing the correct word order in the output.

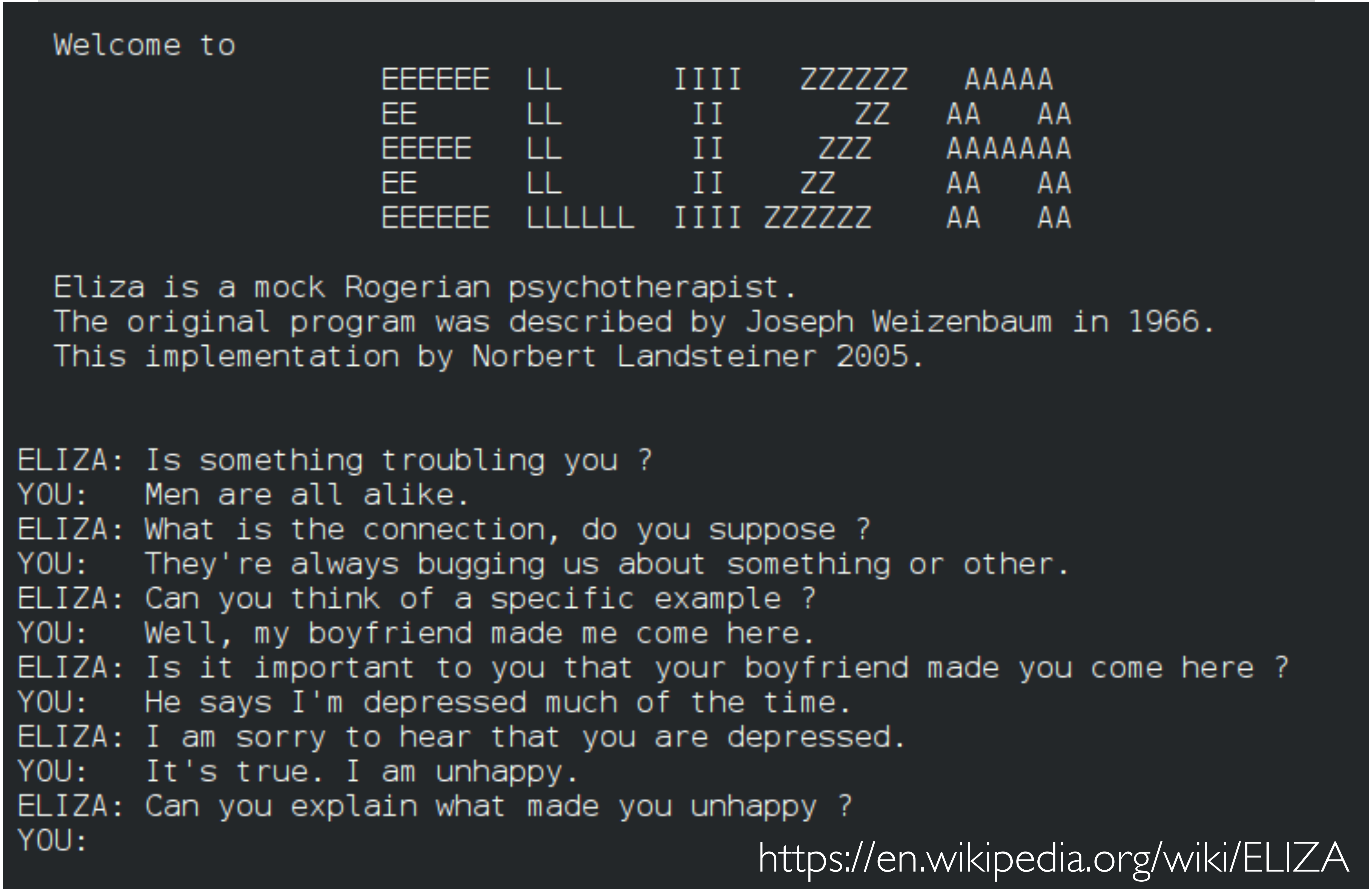
Dictionaries are the first form of linguistic resources used in NLP, but they were only used to store knowledge and make it accessible to MT tools.

NLP: Dialog

Credits: B. Plank talk @ ACL2024


Golden Years for MT (Georgetown Exp.)
DIALOG
1966: Weizenbaum develops ELIZA at MIT
1st "AI winter"
1980s: knowledge representation, data & corpora, statistical NLP
2nd "AI winter"
1995-2005: progress in data, hand-crafted features and ML
Epoch 1: Symbolic Processing
Epoch 2: Statistical Processing
1960
1970
1980
1990
2000
2010
2020

NLP: Dialog



```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA    AA
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    EE      LL       II    ZZ      AA    AA
                    EEEEEE  LLLLLL  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:
```

https://en.wikipedia.org/wiki/ELIZA

NLP: early failures

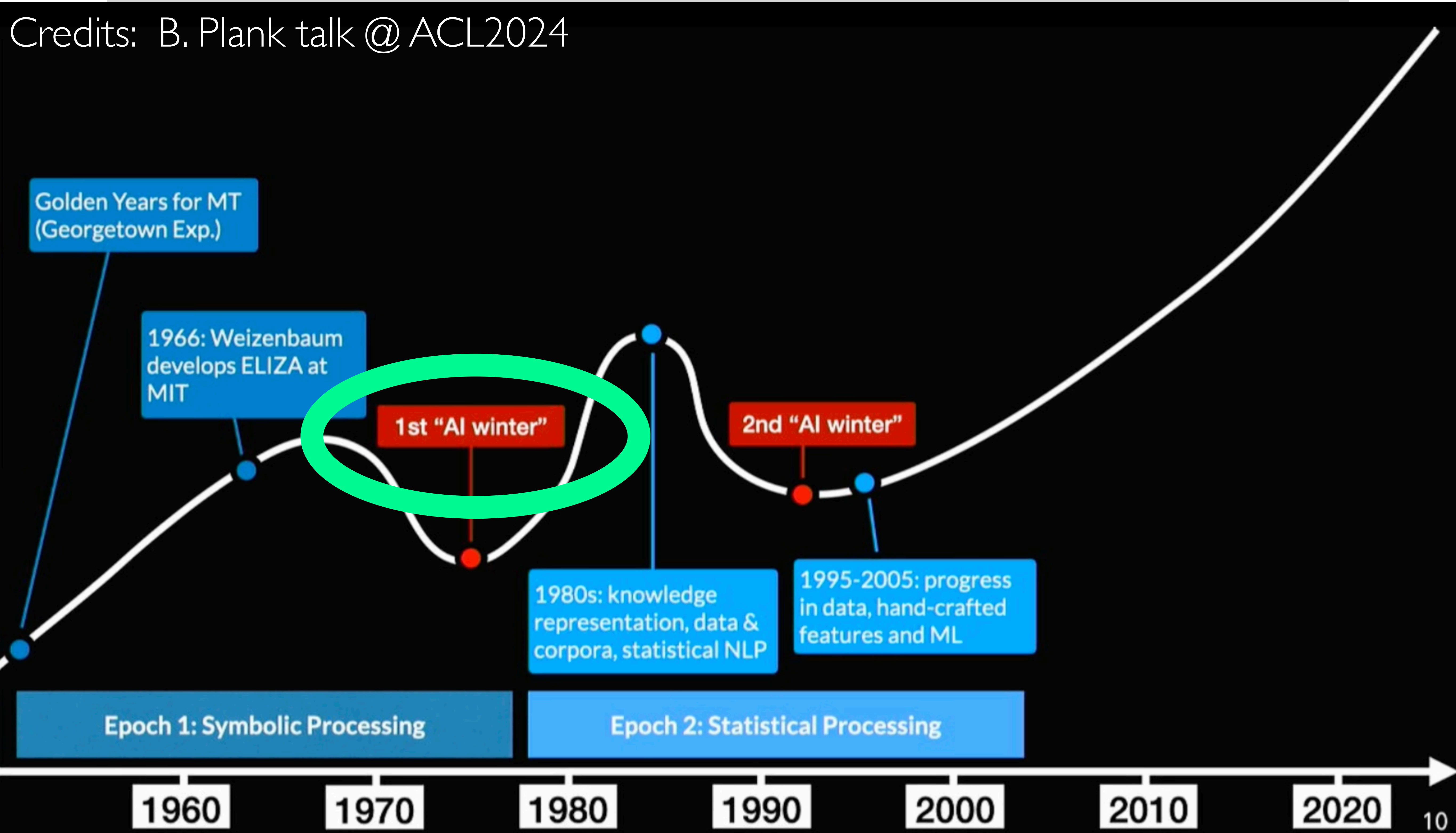
The task of simulating language comprehension and generation (in MT) or dialog (such as ELIZA) proved to be very complicated, regardless of the limits of the machines in terms of computational power and memory.

The **results** of the first generations of NLP systems were indeed **very poor,** they gave rise to the **1st NLP and AI winter.**

The NLP became the subject of **criticism**, while **funding**, which had been plentiful up to that point, was severely **cut**.

NLP: the 1st winter

Credits: B. Plank talk @ ACL2024



Criticism of MT

In 1960, the mathematician and philosopher professor Joshua Bar-Hillel, who worked at MIT until a few years ago, published the article "*A Demonstration of the Non feasibility of Fully Automatic High Quality Translation*" [2].

He provided a demonstration that the automatic translation of a short and simple sentence can be very hard, indeed impossible.

[2]: https://aclanthology.org/www.mt-archive.info/Bar-Hillel-1959-App4.pdf

The example of Bar-Hillel against MT

The sentence in bold is the focus of the argumentation of Bar-Hillel

Little John was looking for his toy box. Finally he found it.
The box was in the pen*. John was very happy.*

The word ‘*pen*' has two meanings and related translations:
• writing tool > in Italian ‘*penna*’, in French '*stylo*'
• fence, playpen for babies > in Italian ‘*box*’, in French ‘*parc d’enfants*'

A human translator can easily select the correct meaning, but how can do that a machine?

The example of Bar-Hillel against MT

?

What kind of knowledge is necessary to translate the sentence in bold?

Little John was looking for his toy box. Finally he found it.
***The box was in the pen**. John was very happy.*

It is necessary to have knowledge of the different dimensions of objects and the ability to compare them.

Criticism of MT

Criticism of MT grew in the following decade, prompting the US government to form a committee to evaluate the results of MT. (Automatic Language Processing Advisory Committee).

The very negative results of the evaluation were published in 1966 in the ALPAC report [3]
It led to the US government cutting funding and promoted **a new perspective in the NLP**.

[3]: https://nap.nationalacademies.org/resource/alpac_lm/ARC000005.pdf

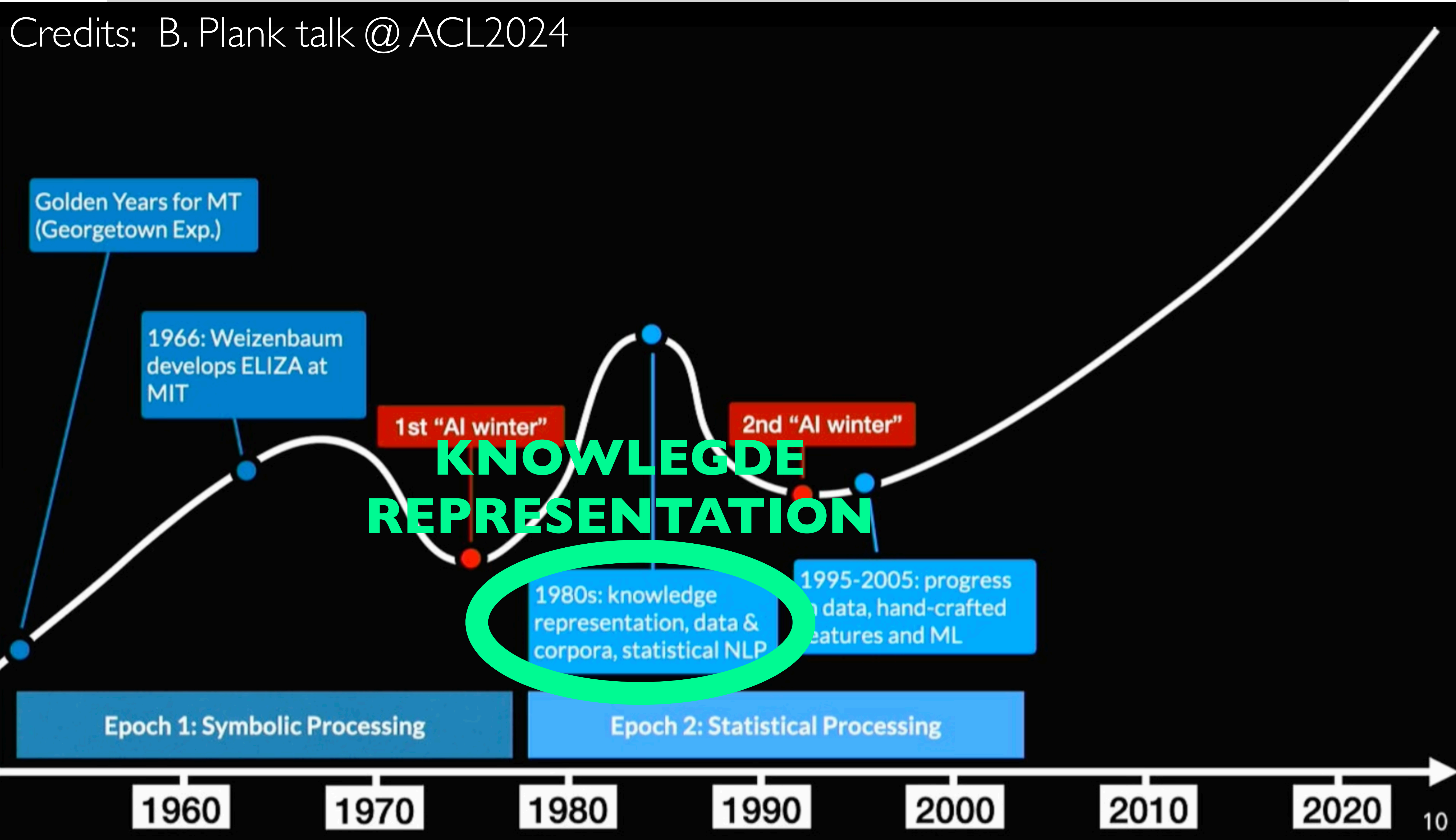
NLP: from NLP failures to new ideas

The failures of the early NLP led to the development of various novel ideas !

These ideas led
to the development approaches for
knowledge representation
but
they also pave the way for the
shift from symbolic to statistical processing
in NLP

NLP: Knowledge representation

Credits: B. Plank talk @ ACL2024



NLP: new ideas

When we use language, we are forced to:

- **perform** several **mental activities and operations**, most of which we are not aware of

- **access to** a wide variety of linguistic (and non-linguistic) **knowledge**

NLP: new ideas

When we use language, we are forced to:

- **perform** several **mental activities and operations**, most of which we are not aware of

carefully investigate activities involved in linguistic behaviour

- **access to** a wide variety of linguistic (and non-linguistic) **knowledge**

NLP: new ideas

When we use language, we are forced to:

- **perform** several **mental activities and operations**, most of which we are not aware of

carefully investigate activities involved in linguistic behaviour

take inspiration from different disciplines

- **access to** a wide variety of linguistic (and non-linguistic) **knowledge**

NLP: new ideas

When we use language, we are forced to:

- **perform** several **mental activities and operations**, most of which we are not aware of

carefully investigate activities involved in linguistic behaviour

take inspiration from different disciplines

- **access to** a wide variety of linguistic (and non-linguistic) **knowledge**

study method for formalising and storing knowledge

NLP: new ideas

When we use language, we are forced to:

- **perform** several **mental activities and operations**, most of which we are not aware of

carefully investigate activities involved in linguistic behaviour

take inspiration from different disciplines

- **access to** a wide variety of linguistic (and non-linguistic) **knowledge**

study methods for formalising and storing knowledge

study methods for dealing with complex knowledge

NLP: new ideas

carefully investigate activities involved in linguistic behaviour

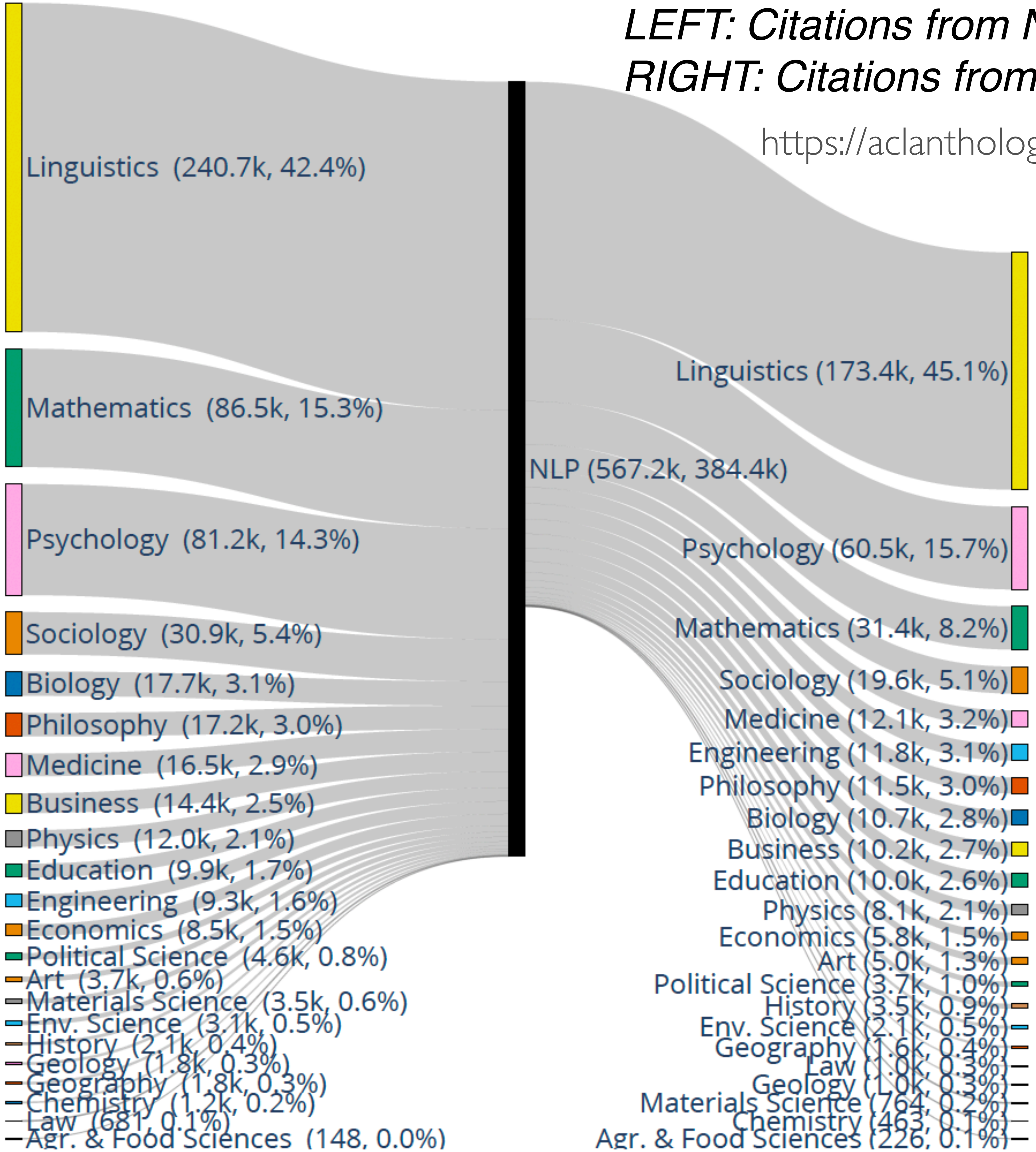
take inspiration from different disciplines

As far as the **mental operations and activities** necessary to understand language, a phase of NLP started where they have been careful studied by computer scientists, psychologists, linguists.

Also today **NLP builds on** techniques and insights from a number of **different disciplines**:
theoretical linguistics and computer science,
mathematical logic,
psychology and cognitive science …

LEFT: Citations from NLP to other disciplines
RIGHT: Citations from other disciplines to NLP

https://aclanthology.org/2023.emnlp-main.797/



NLP: problem-solving

study methods for dealing with complex knowledge

MT and dialog are too complex tasks.

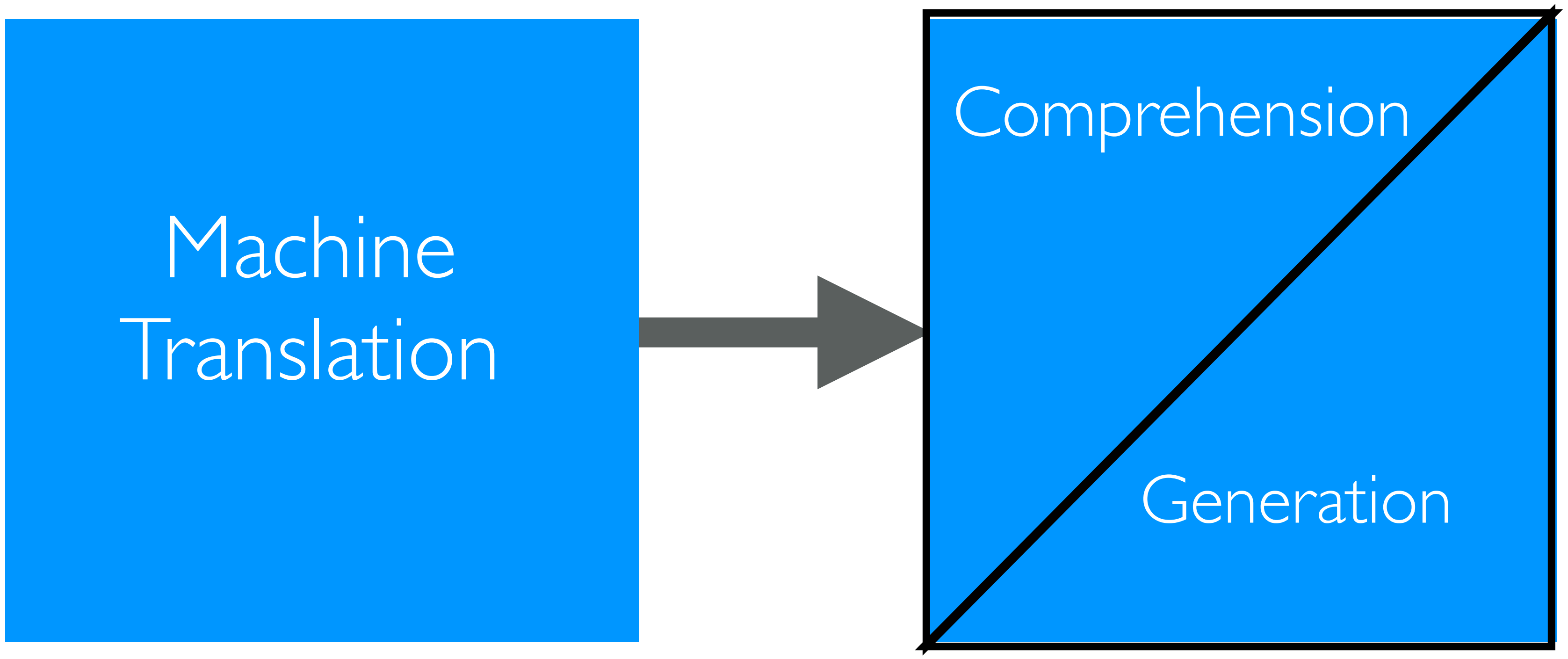
Can they be treated as a **composition of** simpler **subtasks**?
YES, applying a **problem-solving** approach!

Machine Translation

NLP: problem-solving

MT and dialog are too complex tasks.

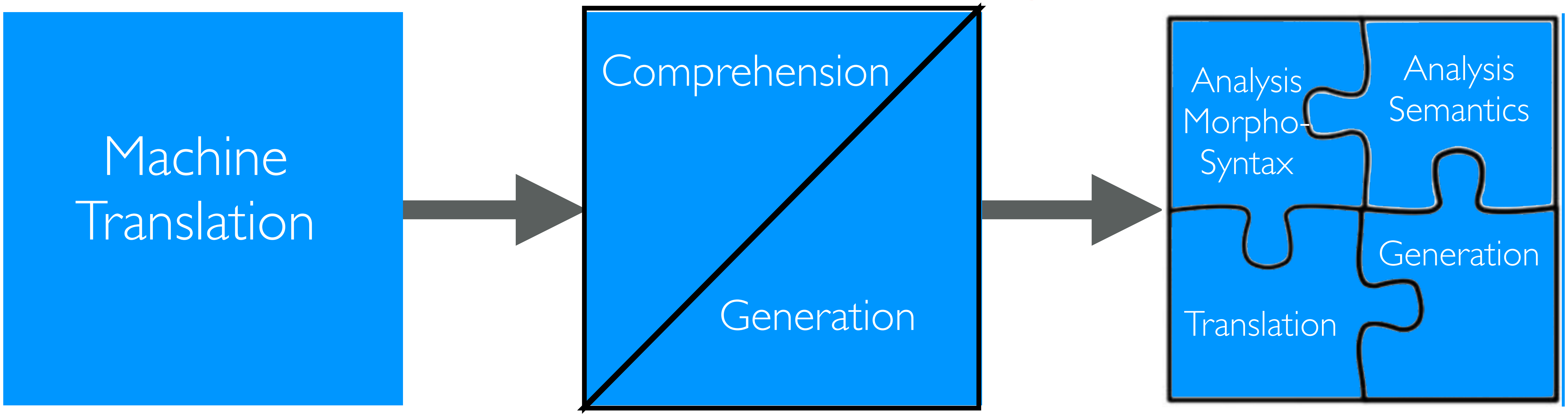
Can they be treated as a **composition of** simpler **subtasks**?
YES, applying a **problem-solving** approach!



NLP: problem-solving

MT and dialog are too complex tasks.

Can they be treated as a **composition of** simpler **subtasks**?
YES, applying a **problem-solving** approach!



NLP: focus on knowledge

study methods for formalising and storing knowledge

It soon became clear to researchers working at NLP that they were needed:

- **organized collections of linguistic knowledge** to cope with language comprehension and generation tasks

NLP: focus on knowledge

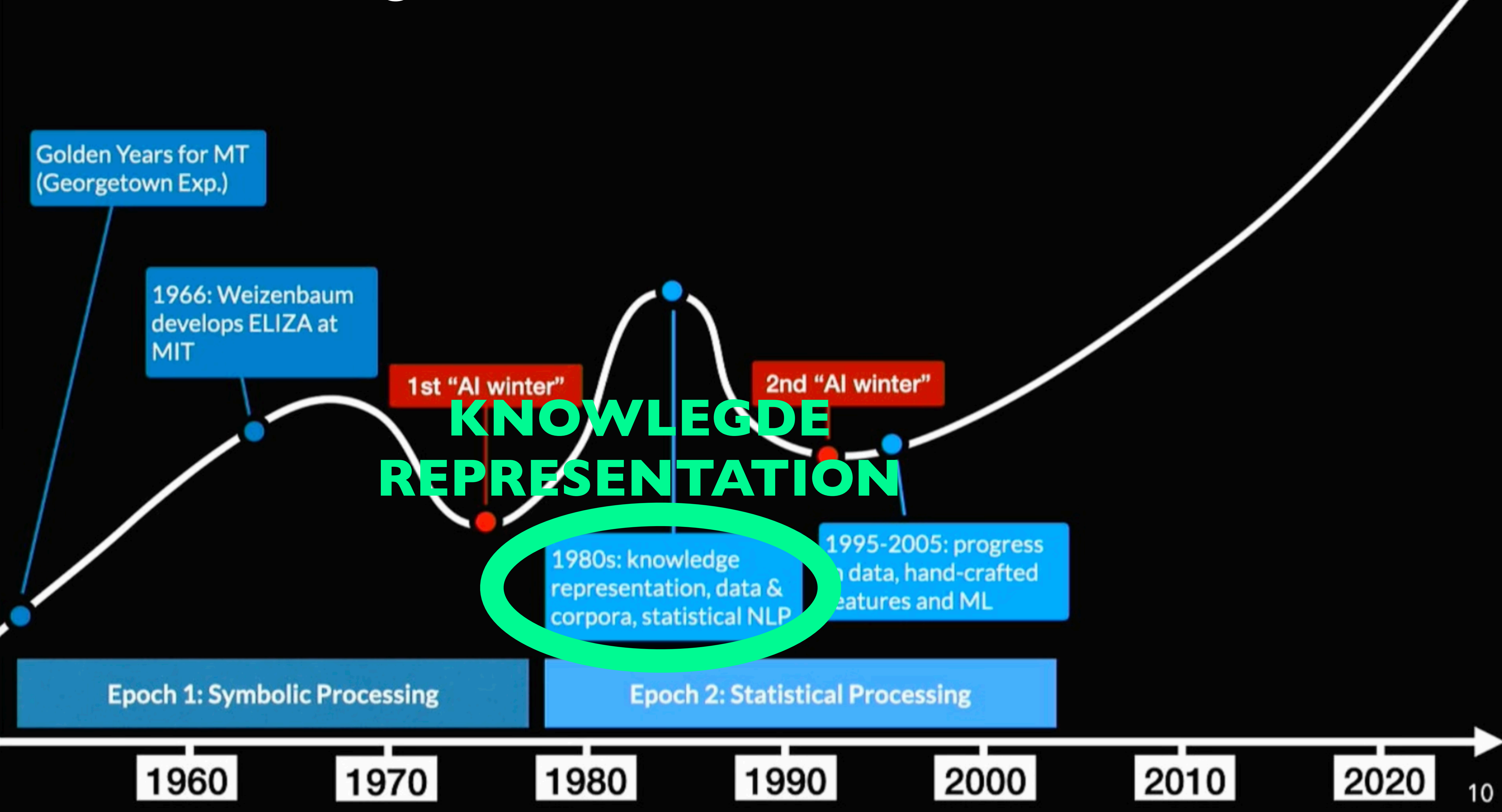
After the failure of the first generation NLP, AI scientists started to focus on the **collection and formal representation of linguistic knowledge**, that is information associated with language use.

A number of NLP projects were inspired by contemporary developments in theoretical linguistics.

While remaining in the context of symbolic processing, growing interest in knowledge representation prepares the transition to statistical approaches

NLP: Knowledge representation

Credits: B. Plank talk @ ACL2024



Golden Years for MT (Georgetown Exp.)

1966: Weizenbaum develops ELIZA at MIT

1st "AI winter"

2nd "AI winter"

KNOWLEGDE REPRESENTATION

1980s: knowledge representation, data & corpora, statistical NLP

1995-2005: progress data, hand-crafted eatures and ML

Epoch 1: Symbolic Processing

Epoch 2: Statistical Processing

NLP and syntactic knowledge

In 1957 Noam Chomsky published *Syntactic structures*, a book that paved the way for the approach whereby human language can be formalised as a (large) set of rules.

The main focus of NLP and related studies at that time was on grammatical structure, which was viewed as a kind of algebraic framework of grammatical rules.

The solution to the NLP problem seemed to be sought in the formalization of linguistic knowledge and especially syntactic knowledge**.**

NLP and formalisation

Chomsky's work introduced a methodology that was to **dominate theoretical linguistics** for decades to come and also to decisively support the development of NLP.

Research focused on the **formalisation of language**, i.e. linguists postulated **formal grammar rules** that were tested against their own intuition or that of native speakers of other languages.
Language is seen as a set of symbols.

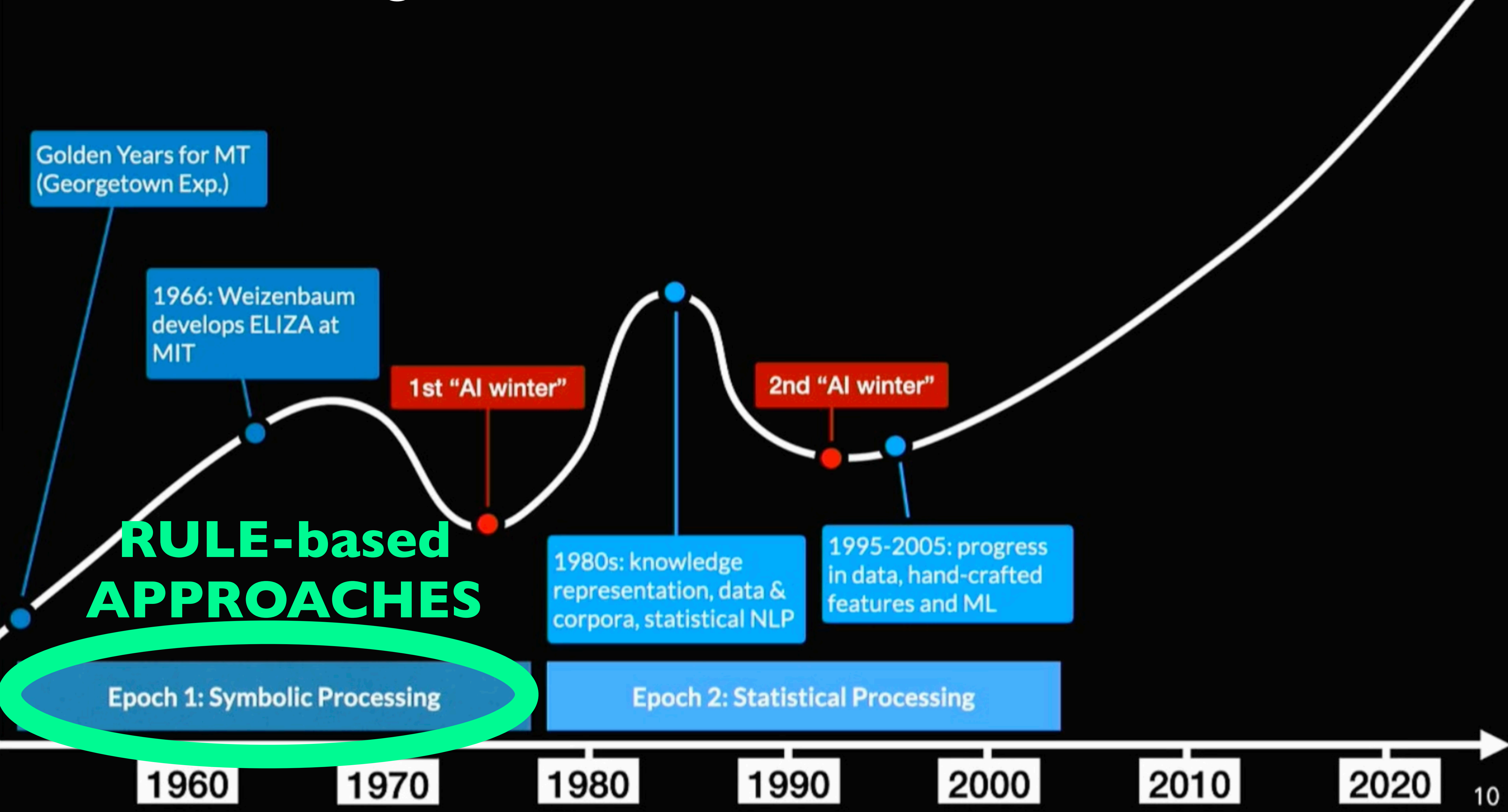
NLP and formalisation

Underlying Chomskian theories is the assumption that native speakers have a specific ***competence*** in their native language, an internalized knowledge of which they are not always aware.

Native speakers are able to recognise whether a sequence of words is grammatically correct and forms an acceptable sentence in their language, even if they have never seen those words in that order.

NLP: symbolic processing

Credits: B. Plank talk @ ACL2024



RULE-based APPROACHES

Symbolic and Rule-based NLP

The focus on knowledge representation led also to the development of systems based on set of rules in which linguistic knowledge is carefully encoded.

The assumptions behind this approach are:

- **ALL the linguistic knowledge can be expressed as formal rules**
- **A finite amount of knowledge allows us to cope with all the linguistic productions**

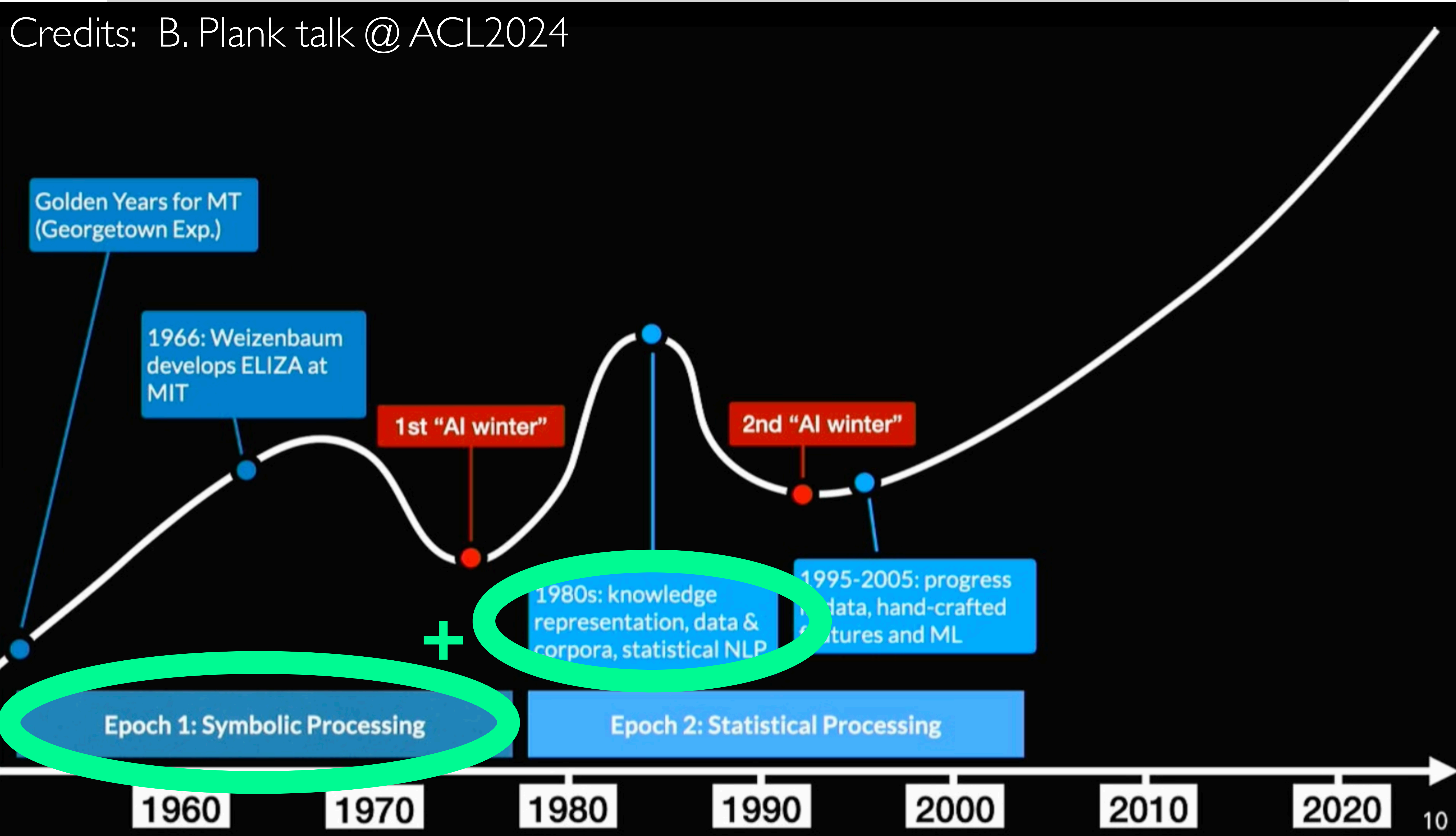
Symbolic and Rule-based NLP

Language is seen as a set of symbols (words and punctuation marks) and rules for composing and manipulating them.

On the one hand, techniques for representing knowledge have led to increasingly refined and precise formalisms for expressing linguistic information associated to words and structures.

On the other hand, algorithms and procedures are developed that use knowledge and apply them rules to simulate language competence.

NLP: symbolic processing



Credits: B. Plank talk @ ACL2024

Symbolic and Rule-based NLP

Grammar rules are a crucial part of the **linguistic knowledge** required for coping with language comprehension and generation.

In the first decades of NLP, several efforts have been made to manually create grammars and **collections of grammatical rules** that can be used by NLP tools.
The amount of knowledge that goes into an NLP task is huge and also difficult to describe.

How can knowledge be formalised? **How much** knowledge should be formalised?

Rules: an example

Rules for adjective associated with noun in English:

- **the adjective precedes the noun** > ‘*blue dogs*’ / ‘*the red cat*’

Rules for adjective associated with noun in English:

- **the adjective precedes the noun** > ‘*blue dogs*’ / ‘*the red cat*’ / **red the cat*

Rules for adjective associated with noun in English:

- **the adjective precedes the noun** > ‘*blue dogs*’ / ‘*the red cat*’ / ******red the cat*

Rules must be precise

- **the adjective immediately precedes the noun**, so if the noun is associated also with an article then the adjective is placed between article and noun > ‘*the red cat*’

Rules for adjective associated with noun in English:

- **the adjective precedes the noun** > '*blue dogs*' / '*the red cat*' / **red the cat*

- **the adjective immediately precedes the noun**, so if the noun is associated also with an article then the adjective is placed between article and noun > '*the red cat*' / **the my cat*

Rules for adjective associated with noun in English:

- **the adjective precedes the noun** > ‘_blue dogs_’ / ‘_the red cat_’ / _*red the cat_

- **the adjective immediately precedes the noun**, so if the noun is associated also with an article then the adjective is placed between article and noun > ‘_the red cat_’ / _*the my cat_

Rules must be detailed

- **when adjective is possessive, it does not admit the presence of the article** > ‘_my cat_’ / ‘_your dogs_'

Are there other exceptions to these rules?

Symbolic and Rule-based NLP

How much knowledge we need?
An enormous amount of knowledge is necessary to cope also with small pieces of a natural language

What kind of knowledge we need?
Also non-linguistic knowledge is involved in the understanding of text (see the example provided by Bar-Hillel for MT).

People became more and more skeptical about the possibility of having good NLP tools in a few years, despite the overfunding of this research area.