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

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

Cristina Bosco

RECAP

After the first 3 decades (from 1950) around, **the NLP is:**

- focused on knowledge representation, which is in turn based on more or less sophisticated theoretical frameworks

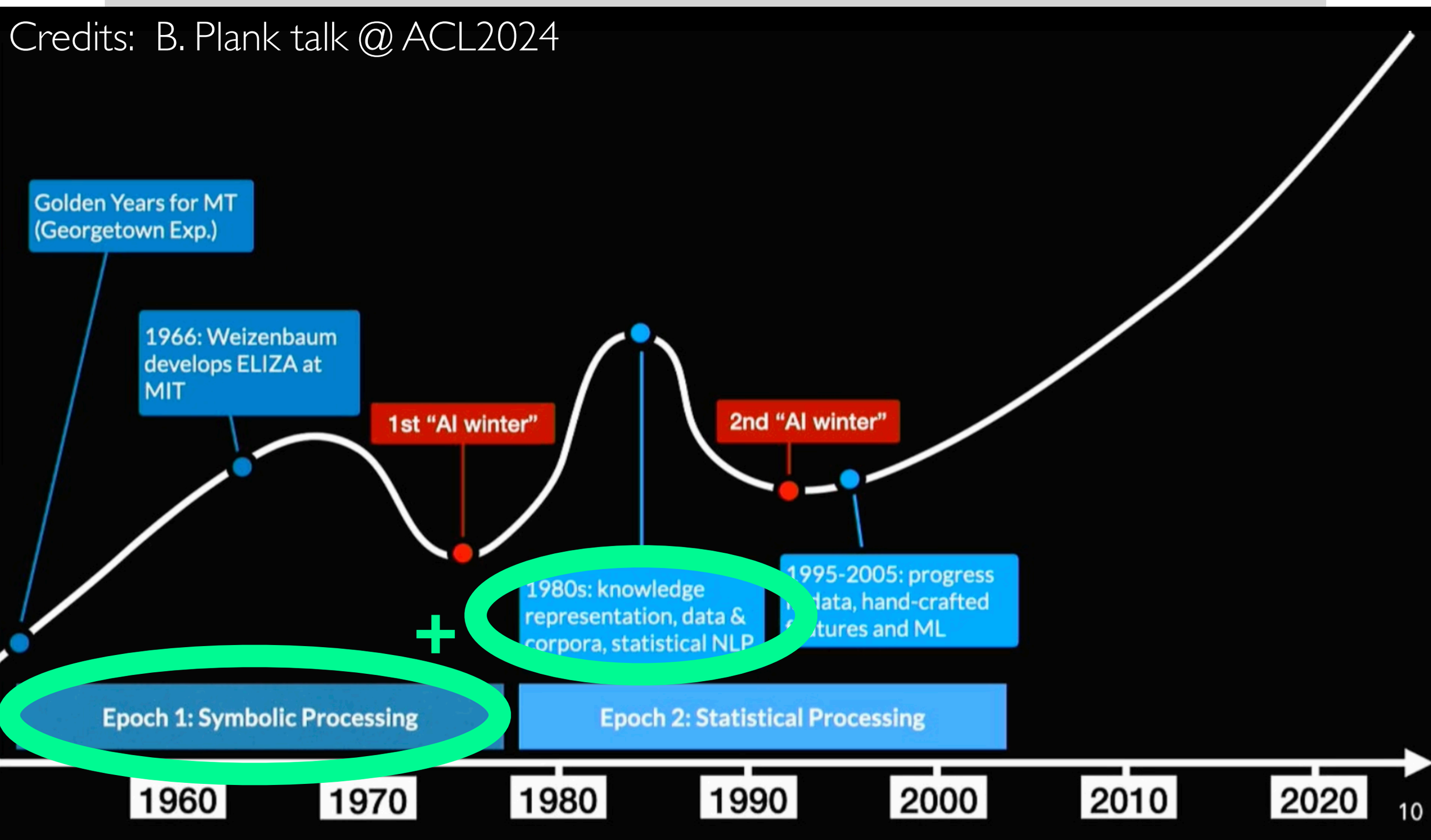
- based on rule-based approaches

- the most complex NLP tasks are broken down into various sub-tasks

- the contribution of different disciplines is considered

NLP in the 80s

Credits: B. Plank talk @ ACL2024



RECAP

During the first 3 decades of NLP **very in-depth studies** are conducted **about human language**, the most are focused on English, and on how to formally describe morphology and syntax.

RECAP

During the first 3 decades of NLP **very in-depth studies** are conducted **about human language**, the most are focused on English, and on how to formally describe morphology and syntax.

**For centuries language had been described in grammars,
in a non-formal and qualitative way.**

Traditional grammars are not suitable for NLP which is based on FORMAL GRAMMARS.

NLP can only simulate what can be described as a set of *mathematical* rules.

RECAP

On the one hand, the study of language leads to a **better knowledge** of its properties.
But it also makes us more aware of its inherent **complexity**.

On the other hand, the study of language leads to a growing awareness of what **language** is when it is **observed in the wild** and used **in different contexts and communication environments**.

The goal is becoming increasingly ambitious:
to treat language as it is used by speakers
in the real world.

Knowledge representation 🤔

Knowledge representation proved to be a very difficult task:

- It is not clear which and how much knowledge is really needed to deal with language
- Linguistic knowledge is complex and stratified
- Extra-linguistic knowledge seems to be necessary for language treatment

Where can we find the linguistic knowledge ?

Rule-based 🤔

Rule-based approaches are not suitable for coping with language in a broad sense:

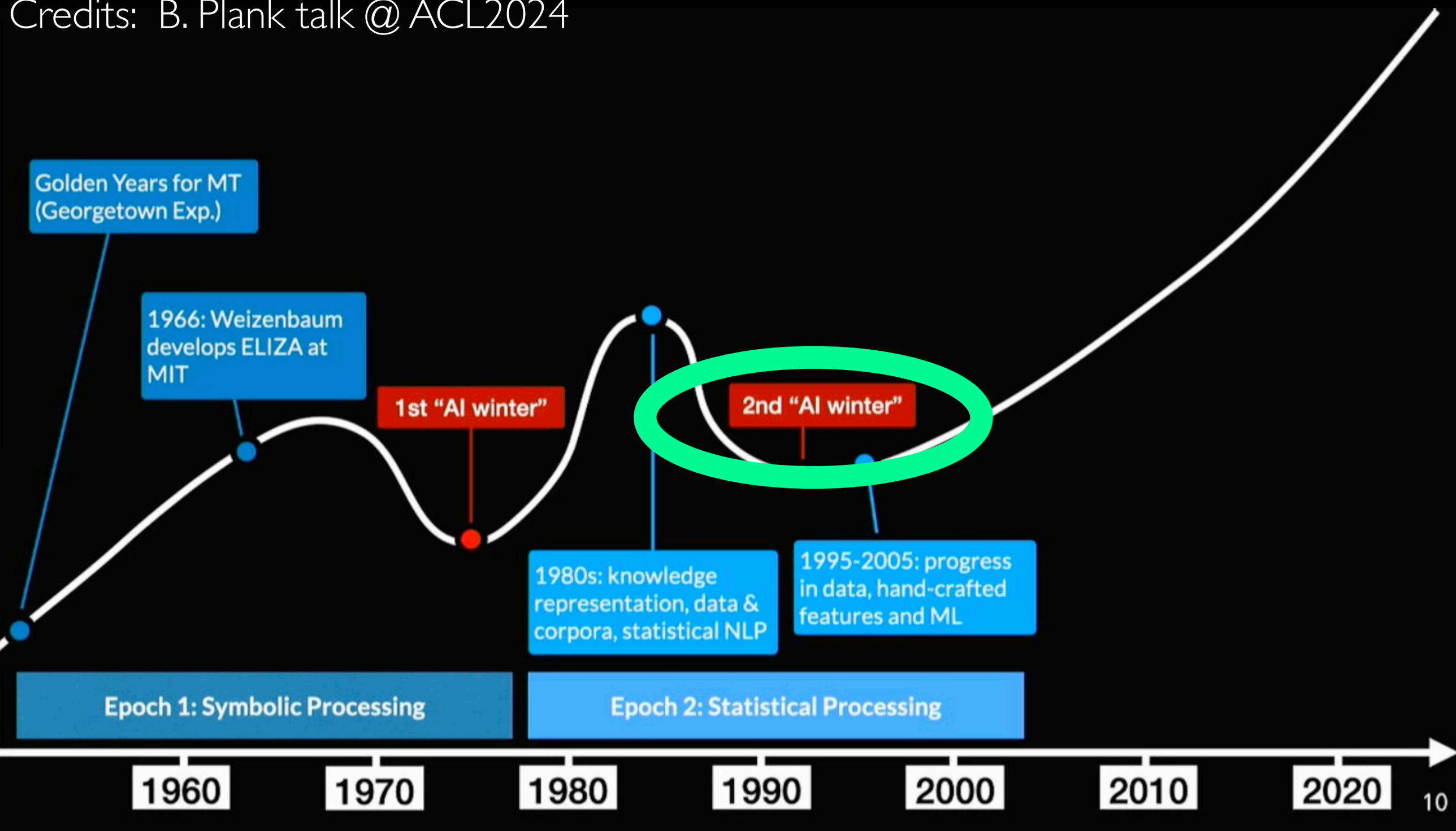
- Several rules are necessary also to deal with small subsets of a language, they can do a good work on toy domains but not on real cases

- Humans often violate grammar rules when using their language in the wild (that is in the most of cases).

How rules can deal with the exceptions ?

NLP: 2nd winter

Credits: B. Plank talk @ ACL2024



NLP: ambiguity

The formalisation of linguistic knowledge is difficult to achieve:
it is impossible to represent all knowledge.

The use of rules was difficult to implement:
it is impossible to treat exceptions using the rules.

But the situation is even worse:
the study of the last decades show that language is **inherently ambiguous**.

Ambiguity and levels

The ambiguity of natural language is probably the biggest problem for NLP.

When an object is ambiguous we have to deal with several (at least two) different *correct* interpretations of the object, without being able to choose the one that is appropriate in the specific context.

Moreover, ambiguity can occur at any level of abstraction.

Ambiguity and levels

At phonological level:

- a **phoneme** can be ambiguous, e.g. *homophon* with another term having different meaning:

English: *p*ear / *p*air (a fruit / a couple)
*h*ere / *h*ear (in this place / to listen)

Italian: *la morale* / *l'amorale* (the morality / the amoral)

Ambiguity and levels

At morphological level:
- a **lexeme** can be ambiguous, e.g. it can belong to different grammatical categories:

English: light > noun in '*light bulb*' and adjective in '*light as a feather*'

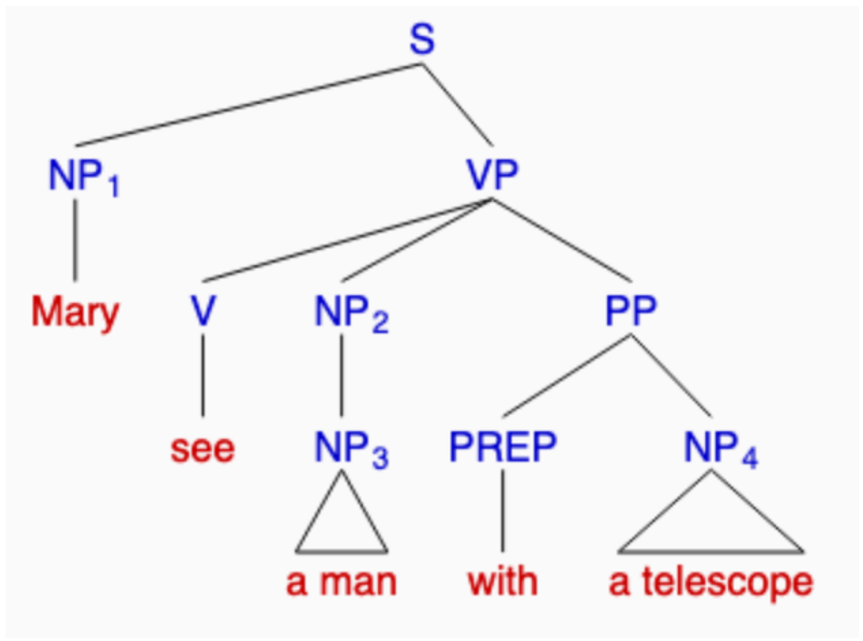
Italian: rosa > noun in 'è *fiorita una rosa*' (a rose has bloomed) and adjective in '*un abito rosa*' (a pink dress)

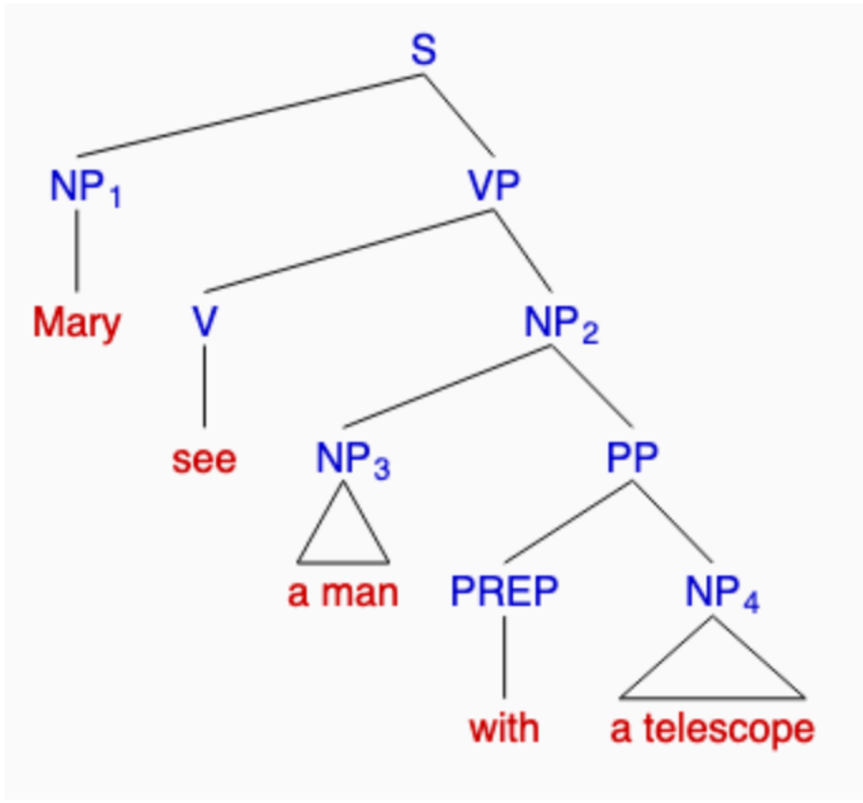
Ambiguity and levels

At syntactic level:
- a **sentence** can be ambiguous, e.g. it can have different syntactic structures:

English: 'Mary see a man with a telescope'

Is Mary using the telescope? or Is the man using the telescope?





Ambiguity and levels

At semantic level:
- the **meaning** of a sentence can be ambiguous, e.g. a sentence can have different meanings:

English: '*Every man loves a woman*' means
<for every man, there is a woman that the man loves>
(each man loves a different woman)
or
<there is one particular woman who is loved by every man>
(all the men love the same woman)
?

Ambiguity

The problem is **not how to encode the ambiguity** of a linguistic object (word, sentence or phrase) in an NLP tool and in the collection of linguistic information it uses.

For example, it is correct that a dictionary assigns more than one meaning to a single word, and that a tool generates more than one syntactic structure for a single sentence.

The problem is, **to select the correct meaning or structure for the specific context** each time the NLP performs an analysis of a text.

Ambiguity

Every time people form a written or spoken sentence, they want to convey a **single meaning** to the listener or reader.
Every time we hear or read a sentence, we understand a **single meaning**.

The interpretation of the meaning of a simple sentence may depend on the **context** of language use, but unfortunately that context may include almost all the knowledge of the community of speakers.

We humans have a very limited awareness of ambiguity !

Ambiguity

I made her duck

can be associated with 5 different meanings:

I cooked a duck for her benefit (to eat)
I cooked a duck belonging to her (to eat)
I was the one who shaped the (plaster, wood?) duck she owns
I caused her to quickly lower her head or body
I waved my magic wand and turned her into a duck

Ambiguity

I made her duck

The association with the different meanings depends on the grammatical category of the lexemes and on the syntactic structure of the sentence:

I cooked a duck for her benefit (to eat) > **her is pronoun at dative case**

I cooked a duck belonging to her (to eat) > **her is adjective**

…

I caused her to quickly lower her head or body > **her is pronoun at accusative case**

…

Ambiguity

Formalising and making all the existing knowledge available to NLP systems cannot be a solution to the problem of ambiguity, for at least two reasons:

- the knowledge is too large to be formalised

- the knowledge varies over time, language, genre, domain …

- linguistic knowledge does not provide guidance how to choose the right interpretation in contexts in which ambiguity occurs

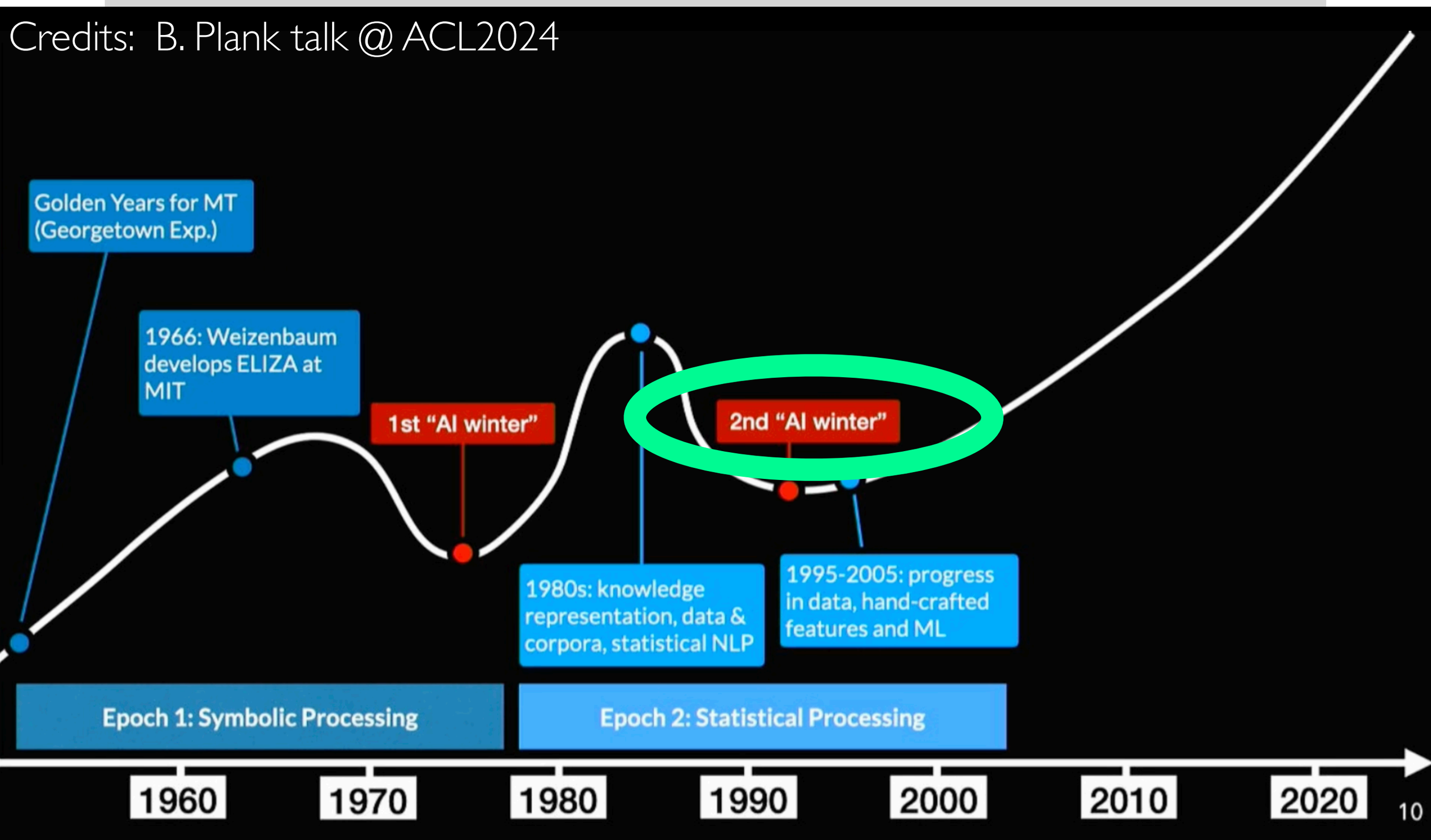
Ambiguity

No precise formalised description of all aspects of language can be complete since the knowledge required to use language includes not only the meaning of words but also the infinite grammatical structures, in which they can freely occur, and the contexts in which they must be used correctly and properly.

Natural languages are examples of open systems. Modelling them means explaining all the productions generated in their context, but also previewing all future productions that can be generated.

NLP: 2nd winter

Credits: B. Plank talk @ ACL2024



NLP as a monolithic task

How to deal with complexity?

NLP as a monolithic task

How to deal with complexity?

NLP as composition of subtasks

NLP as a monolithic task

How to deal with complexity?

Where to find knowledge?

NLP as composition of subtasks

NLP: ambiguity

There were still two main problems to solve:

- how to collect and represent all the linguistic **knowledge** needed to deal with a natural language?

- how to deal with the **ambiguity** inherent in a natural language?

The solution will only be available in the 90s … and it is the same for both!


NLP as a monolithic task
How to deal with complexity?
Where to find knowledge?
NLP as composition of subtasks
Using corpora and resources

NLP as a monolithic task

How to deal with complexity?

Where to find knowledge?

How to deal with ambiguity?

NLP as composition of subtasks

Using corpora and resources

NLP as a monolithic task

How to deal with complexity?

Where to find knowledge?

How to deal with ambiguity?

NLP as composition of subtasks

Using corpora and resources

Using corpora and resources