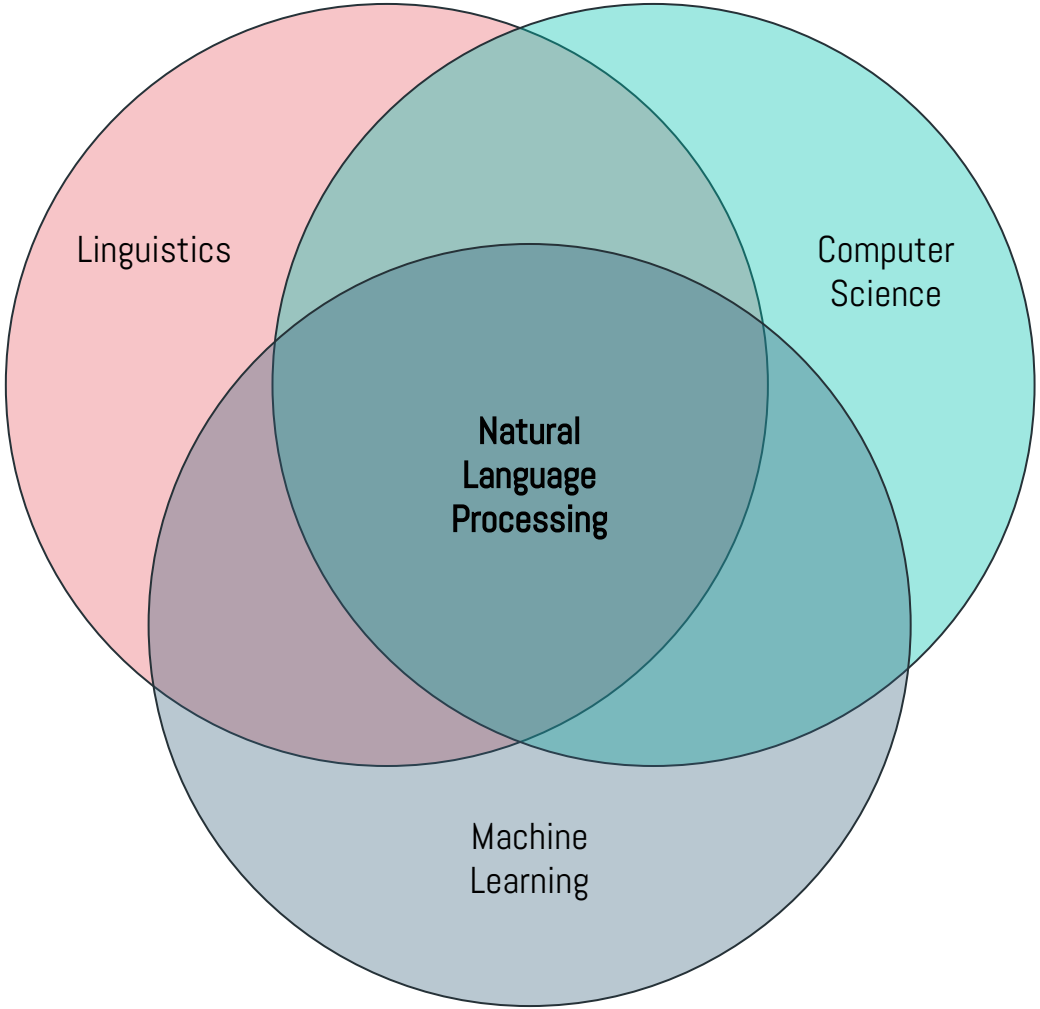
Teaching Computers to Read (Part 1)

The history and basics of Natural Language Processing

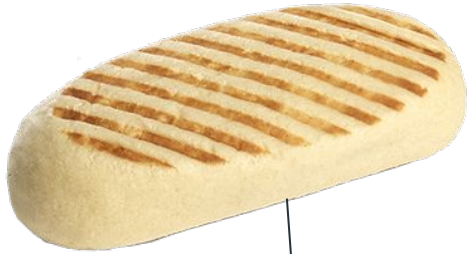
What is Natural Language Processing?



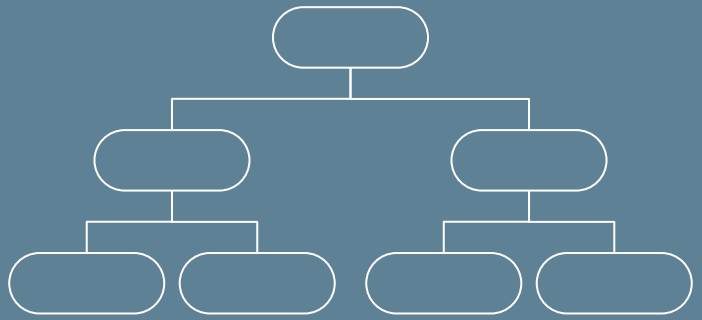
Linguistics

The scientific study of language

- Earliest record from 6th century BC
- Ferdinand de Saussure (1857 - 1913)
- Languages have two components:
  1. A system of signs
  2. A social phenomenon

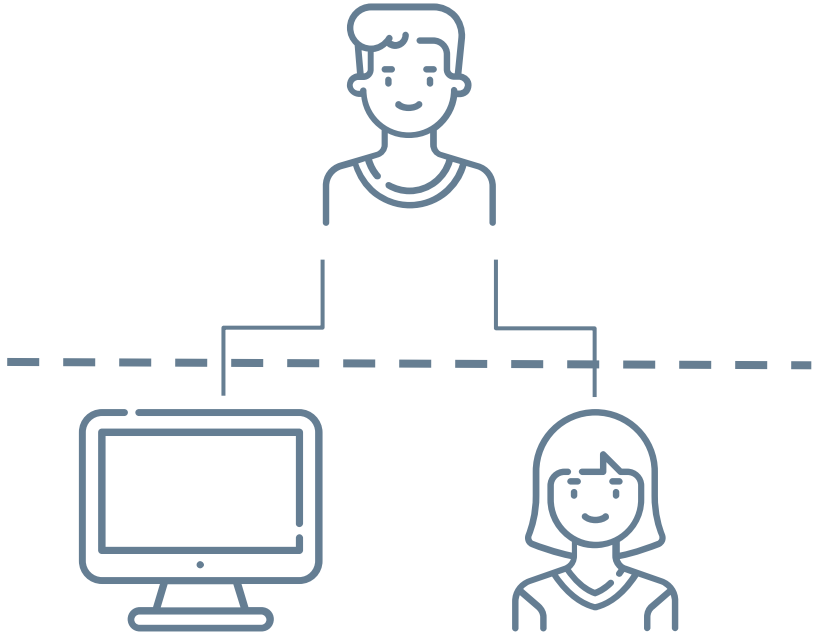


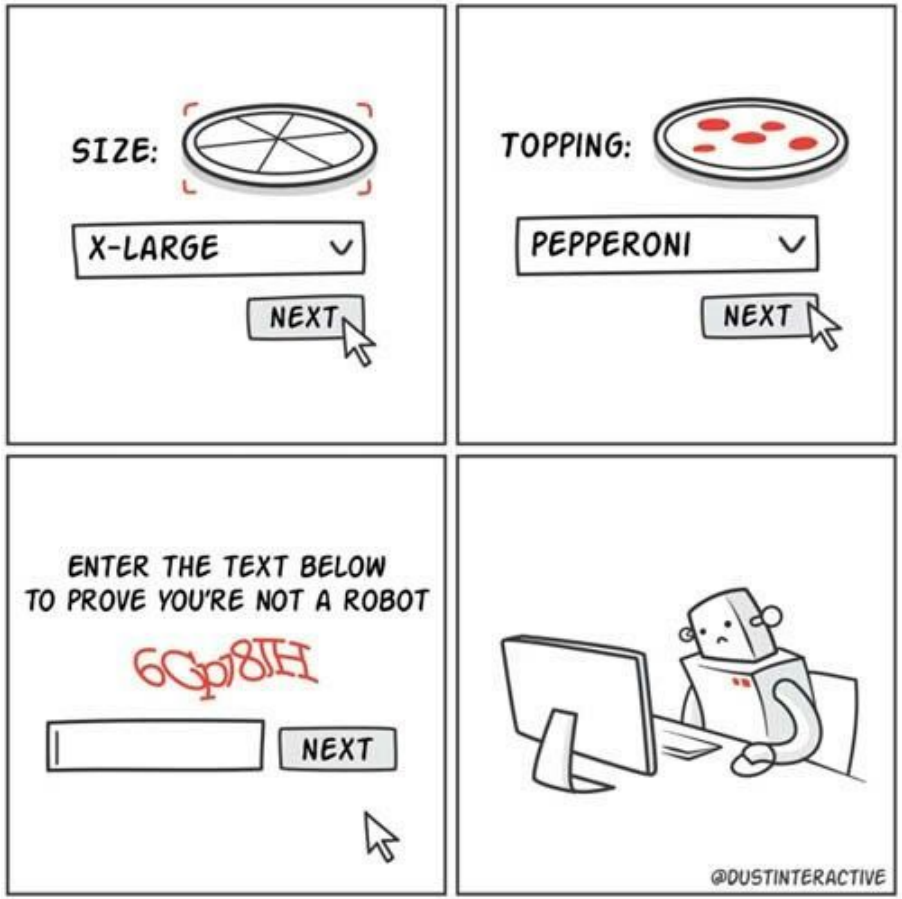
1900–1950



Linguists focus on “The structuralist approach” trying to model languages as systems

1950: Turing Test



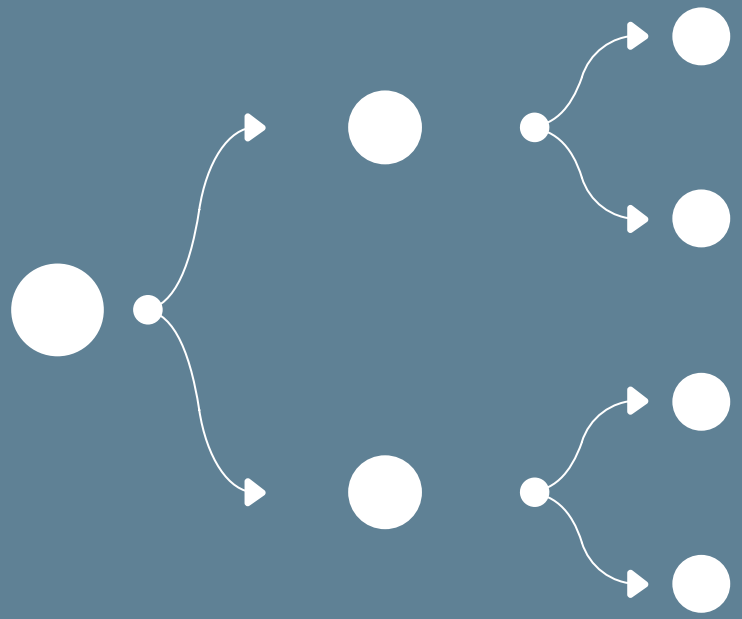

SIZE:
X-LARGE
NEXT
TOPPING:
PEPPERONI
NEXT
ENTER THE TEXT BELOW
TO PROVE YOU'RE NOT A ROBOT
NEXT
@DUSTINTERACTIVE

1950s - 1990s

NLP focuses on building

Symbolic

Systems



1954: Georgetown-IBM

Experiment to tackle **Machine Translation**

- An exact equivalent for the Russian language item exists
- Rearrange the position of words to find translations
- If there are several choices, the result is based on the consecutive words (maximum of three).
- Several choices but the result depends on the previous words (maximum of three).
- The item is not present in the input language
- The item is not present in the output language

1966: ELIZA

ELIZA: How do you do. Please tell me your problem.

ME: I am trying to give a presentation at a conference.

ELIZA: Is it because you are trying to give a presentation at a conference that you came to me?

ME: Yes, that is why.

ELIZA: I understand.

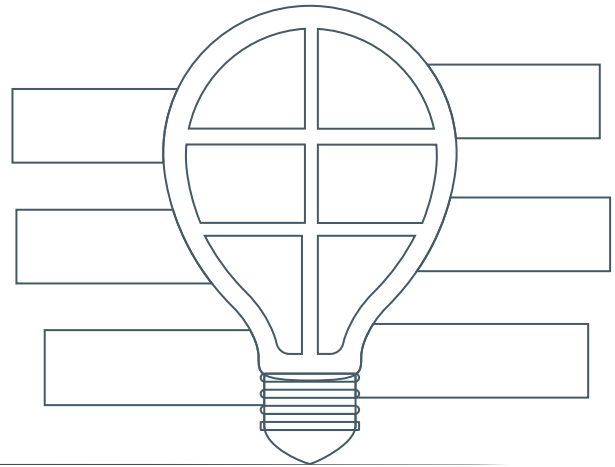
ME: Are you able to offer me advice?

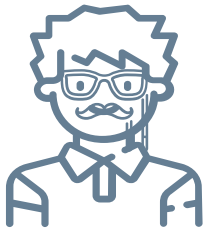
ELIZA: Why are you interested in whether I am able to offer you advice or not?

ME: Because it might help

ELIZA: Is that the real reason?

ME: No... I'm lonely





I saw a girl beside an owl with glasses

I saw a girl beside an owl with glasses

I saw a girl beside an owl with glasses

I saw a girl beside an owl with glasses

“A language is not just words. It’s a culture, a tradition, a unification of a community, a whole history that creates what a community is. It’s all embodied in a language.”

—Noam Chomsky

1986: Lesk algorithm

I found it hard to fit the nut around the end of the bolt.

A fruit consisting of a hard or tough shell around an edible kernel.

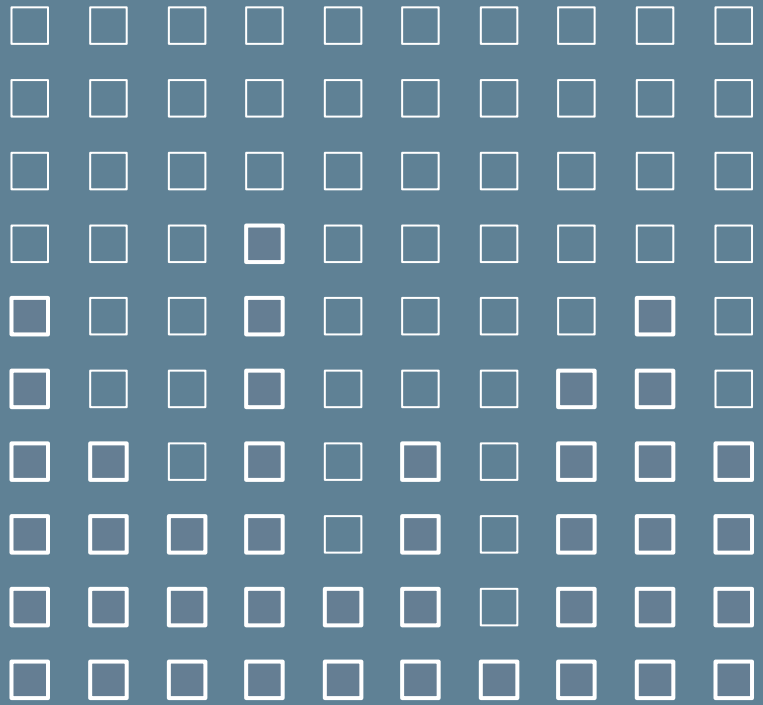
A small flat piece of metal or other material, typically square or hexagonal, with a threaded hole through it for screwing on to a bolt as a fastener.

1990s - 2010s

NLP focuses on building

Statistical

Systems



Key differences

Statistical Inference of Rules

Rules are no longer created by the human designers of the system but chosen by the system itself based on statistical outcomes

Large corpora are required

The machine-learning method requires large amounts of sample text in what is called a “corpus”

Multiple outputs with probabilities

Rather than having a discreet answer, the systems can provide a range of outputs along with their statistical likelihood of being correct

Example: Markov Chain Prediction

Predicts

The next word in a chain based on the current word and preceding words

Trained

On a corpus of text, ideally as similar as possible to the domain being predicted

Uses

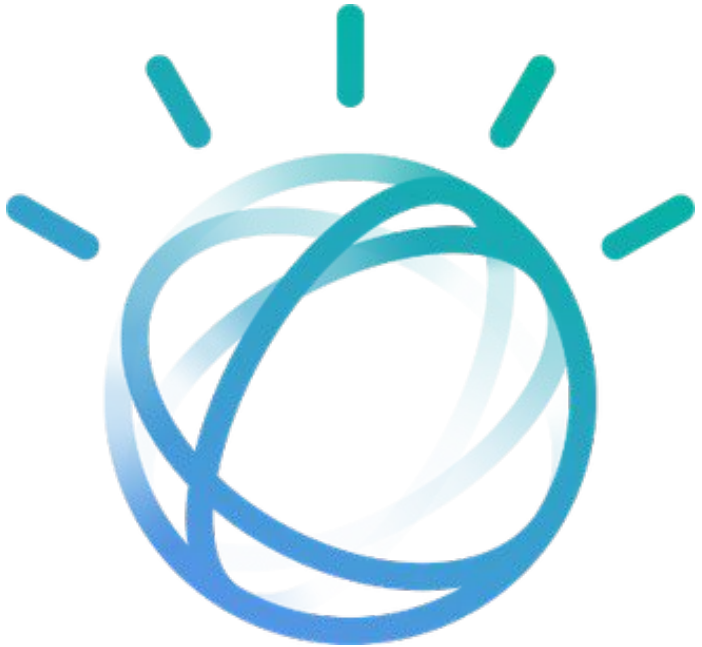
An n-gram model for training and prediction. Higher n-grams are more accurate but harder to train.

I was eating a key lime ...

2006: IBM Watson

Question answering system

- Uses hundreds of NLP algorithms
- Selects answers where the algorithms agree
- Competed on the US game show **Jeopardy**
- In **2011** won $1m on the show, defeating previous champions



IBM Watson

Question: In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

Evidence: In May, Craig arrived in India after he celebrated his anniversary in Portugal.

Keyword matching:
Celebrated
Arrival
May
Anniversary
India
Portugal

The explorer must be Craig

IBM Watson

Question: In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

Evidence: On the 27th of May 1498, Vasco da Gama landed in Kappad Beach

Temporal Reasoning: 400th anniversary in May 1898 = May 1498

Statistical Paraphrasing: arrival in = landed in

Geospatial Reasoning: India = Kappad Beach

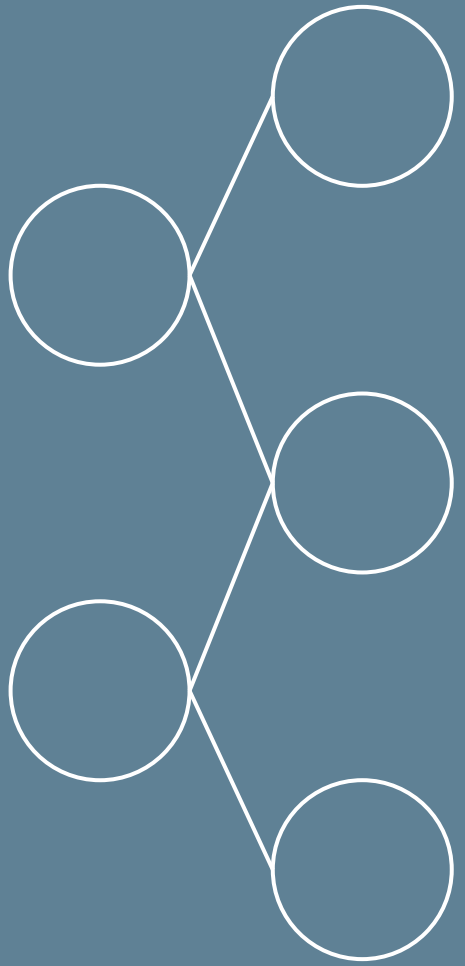
The explorer must be Vasco da Gama

2010s – Now

NLP focuses on building

Neural Network

Systems



Deep Learning

The rise of deep learning in general plus some good results using them for NLP

01

02

Cloud Computing

Previously the hardware required was a big limitation

Word Embeddings

The rise of pre-trained embedding models like word2vec

03

04

RNNs

Developments in deep learning to allow networks to carry memory

Modern NLP uses

Language Translation

Speech Transcription

Sentiment Analysis

Voice Assistants

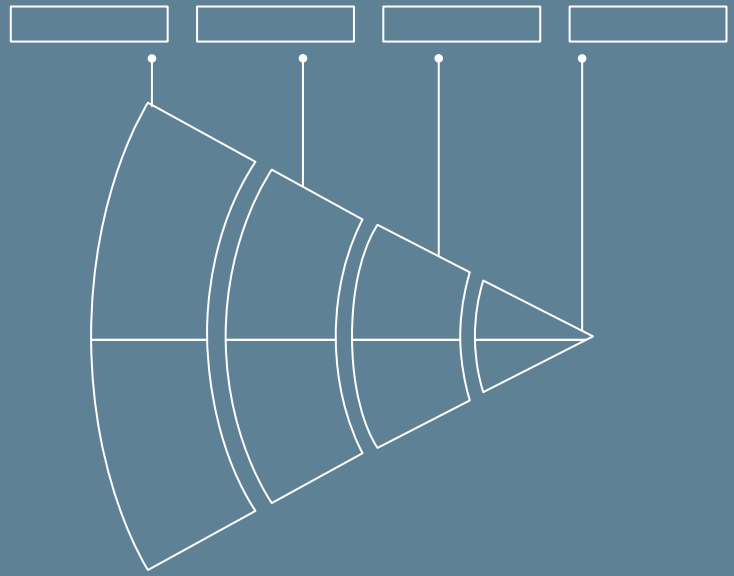
Autocompletion

Spam Detection

NLP Tasks

Text data is difficult to work with. It requires a lot of

Pre Processing



Regular Expressions

Before

<h1>Donald J. Trump - @realDonaldTrump</h1>
Despite the constant negative press covfefe

After

Donald J. Trump
Despite the constant negative press covfefe

Tokenization

Before

The greatest wealth is to live content with little

After

['The',
'greatest',
'wealth',
'is',
'to',
'live',
'content',
'with',
'little']

Stop-Words

Before

Be kind, for everyone you meet is fighting a hard battle.

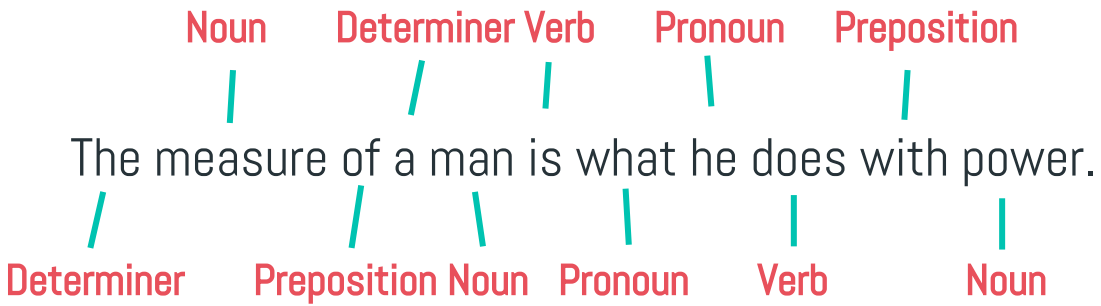
After

Be kind, everyone meet fighting hard battle.

Stemming

Before	After
Magical, Magician, Magically	Magic
Hunters, Hunting, Hunted	Hunt
Airline, Airliner, Airlines	Airlin

Part of Speech (PoS) Tagging



Lemmatization

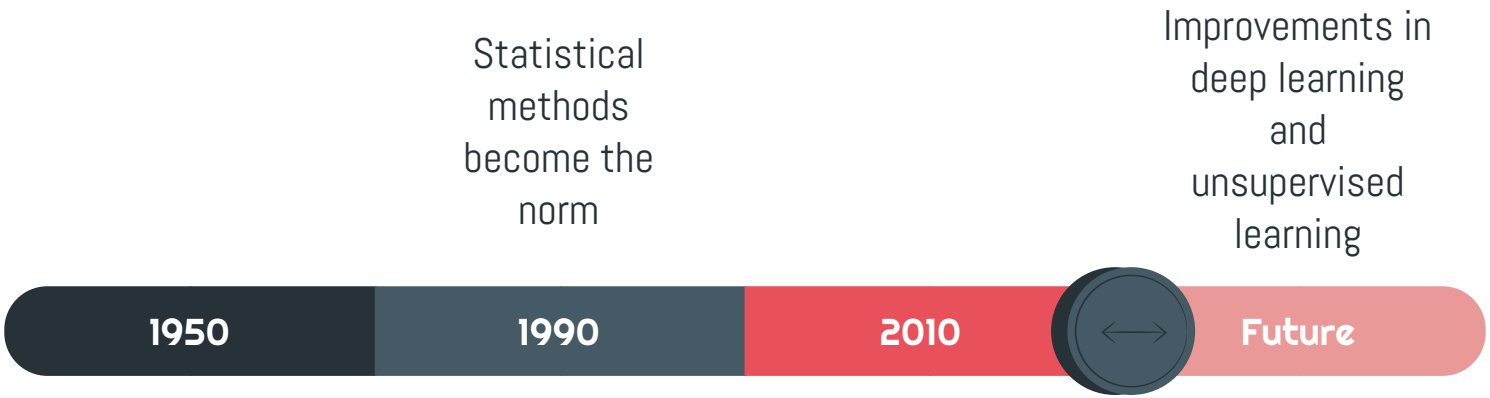
Before	After
Was, is, are	Be
Meeting, meet, met (verb)	Meet
Meeting (noun)	Meeting

N-grams

[Only [the [dead] [have] seen] the end of war.

(Only, the)
(the, dead)
(dead, have)
(have, seen)
(seen, the)
(the, end)
(end, of)
(of, war)

Summary


1950
1990
2010
Future
NLP as a discipline is born. Efforts are rule-based.
Statistical methods become the norm
Neural Networks provide better results
Improvements in deep learning and unsupervised learning

How do I get started?

Watson NLU (Natural Language Understanding)

IBM Developer - developer.ibm.com

Thanks!

Do you have any questions?

edmundshee@uk.ibm.com

@ukcloudman