NLP 4 HEALTH APPLICATIONS

Alessandro Bondielli

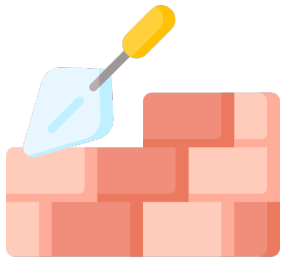
alessandro.bondielli@unipi.it

Assistant Professor

Dept. of Computer Science

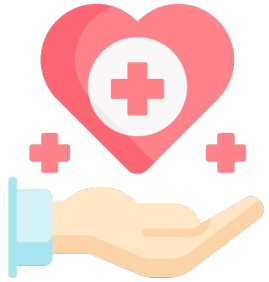
& Dept. Of Philology, Literature and Linguistics

Slides and Materials



Part I

NLP fundamentals



Part II

NLP for Health

Part III

Hands on

Part I

NLP fundamentals

Human language *is* hard

"*She saw the man with a telescope*"

- Who has the telescope?

"*Time flies like an arrow; fruit flies like a banana*"

- Multiple interpretation due to structure and wordplay

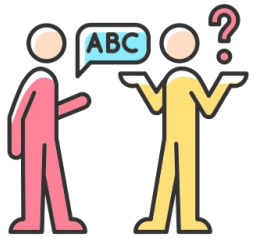
"*I'm feeling blue today*"

- Semantic ambiguity due to idiomatic expressions

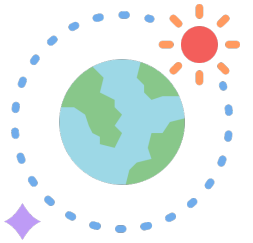
Human language *is* hard

Multiple language phoenomena, each with its own complexity

- Syntax, lexicon, semantics, pragmatics…



Language changes over time and space



Natural Language

- Language is the most distinctive feature of human intelligence
  - Orangutan's intelligence:
    - vision, use of tools, making plans, but they lack language to express and to communicate
- Language shapes thought
- Emulating language capabilities is a scientific challenge

Natural Language

The use of a complex language is a **distinctive trait of humans**

- Thousands of symbols
- Complex syntax
- Semantic is *mostly* compositional
- Potentially **ambiguous**
- Relies on **shared world knowledge**
- Learned from scratch **along life**
- It **evolves** along our lives

James R Hurford, *"Human uniqueness, learned symbols and recursive though"*

Thomas C. Scott-Phillips, Richard A. Blythe, "Why is combinatorial communication rare in the natural world, and why is language an exception to this trend?

Structured vs Unstructured Data

- **85%** of business-relevant information originates in unstructured form, primarily text
- Information is mostly communicated by reading or writing e-mails, reports, or articles and the like, in conversations, or by listening/watching media
- Challenges:
  - Universal agreed ontologies
  - Commonsense knowledge
  - Grounding

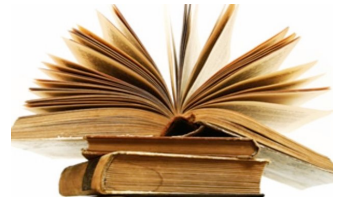
Natural language

- Unstructured data with a high level of variability
- Implicit information content
- Information extraction requires a linguistic understanding of the text
- Rich in intra- and extra-textual entities, events, and relationships
- Very heterogeneous sources







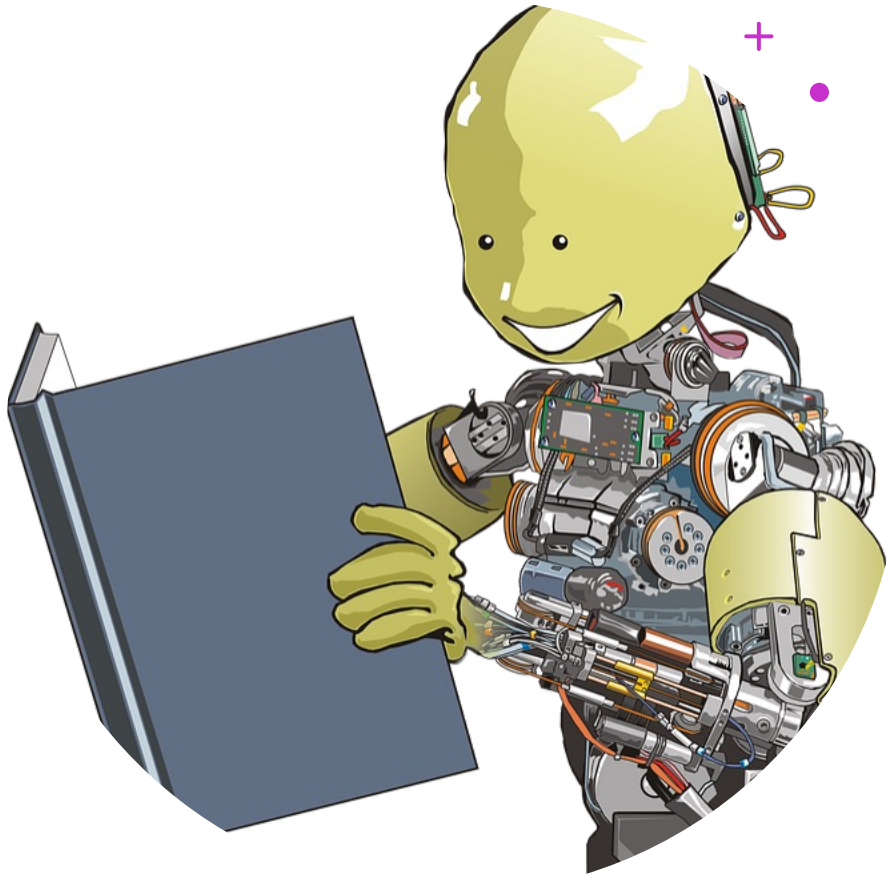



Natural Language Understanding (by machines)

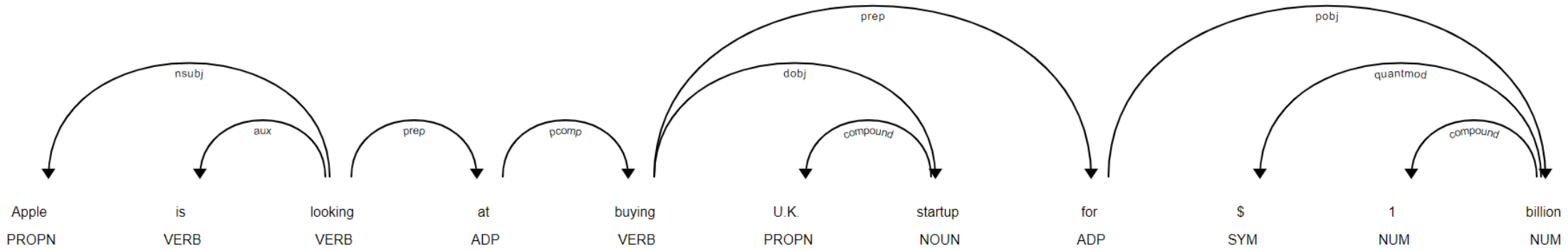
Natural Language Understanding aims at building machines that are able to **receive** and **give** information using natural language, like humans do

From a computational point of view, natural language understanding is considered to be an AI-complete problem

Practical NLP tasks are simplifications of NLU that make the problem easier to solve



Complex syntax



ambiguity

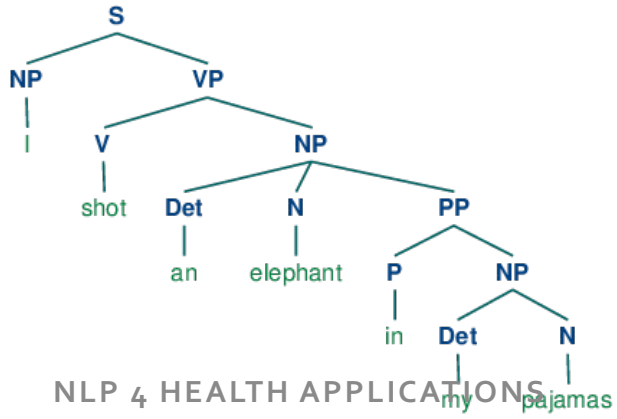
word	context	PoS
beat	Listen to this cool beat	Noun
beat	I'll beat you at checkers!	Verb

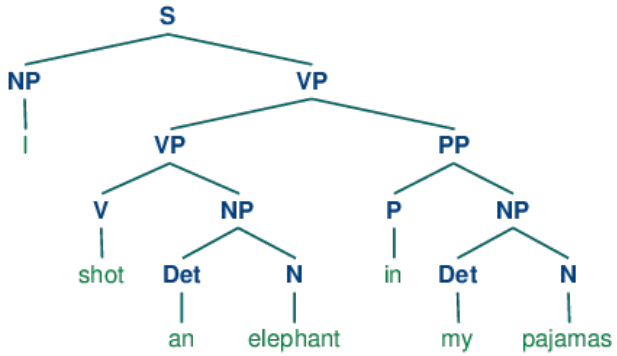
Part-of-Speech
The same word means two different things depending on its part-of-speech (POS)

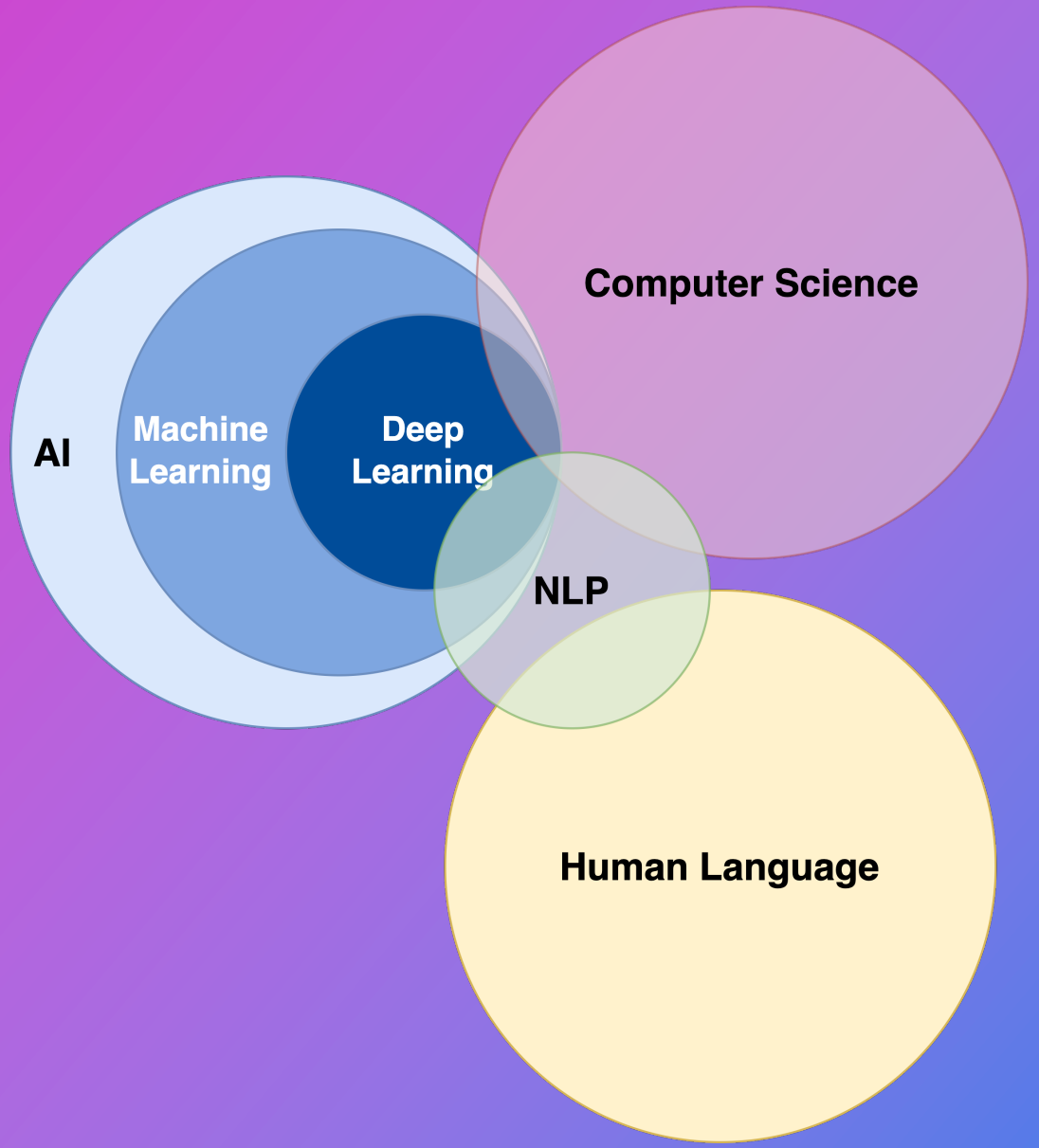
word	context	Word Sense
interest	There is interest on this topic	attention
interest	The bank raised the interest rate	money

Word Sense
The same word means two different things depending on the context in which is used

Syntax
Changes the meaning of the sentence







NLP

Involves computer scientists, linguists, and cognitive scientists

Studies **algorithms for processing natural language**

WHY?

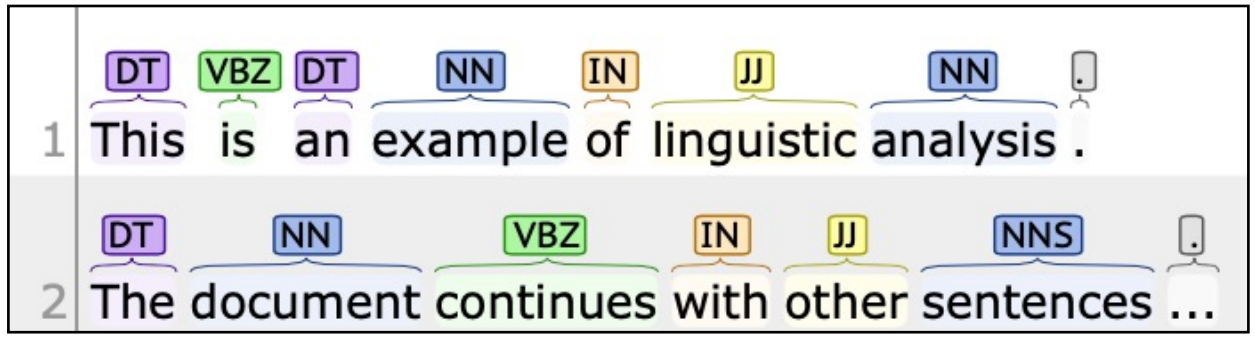
- Massive amounts of textual data
- Humans cannot process them
- Importance of Information Extraction

Linguistic knowledge

This is an example of linguistic analysis. The document continues with other sentences...

This is an example of linguistic analysis. The document continues with other sentences...

This is an example of linguistic analysis . The document continues with other sentences ...



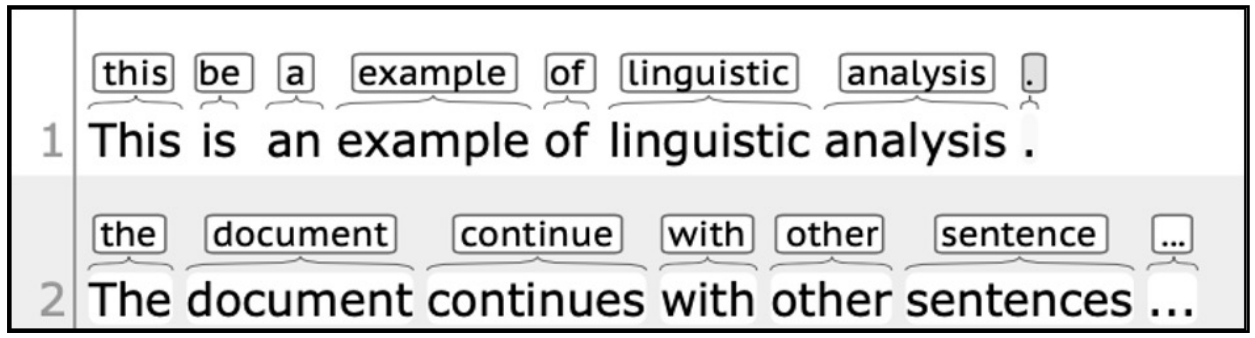
- Sentence Splitting
- Tokenization
- PoS-tagging
- Lemma
- Parsing

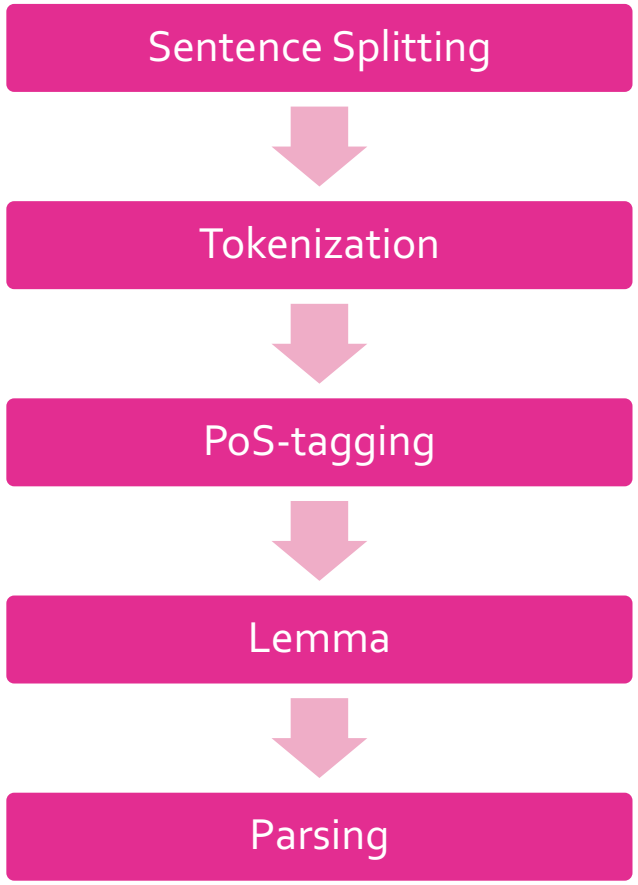
Linguistic knowledge

This is an example of linguistic analysis. The document continues with other sentences...

This is an example of linguistic analysis. The document continues with other sentences...

This is an example of linguistic analysis . The document continues with other sentences ...



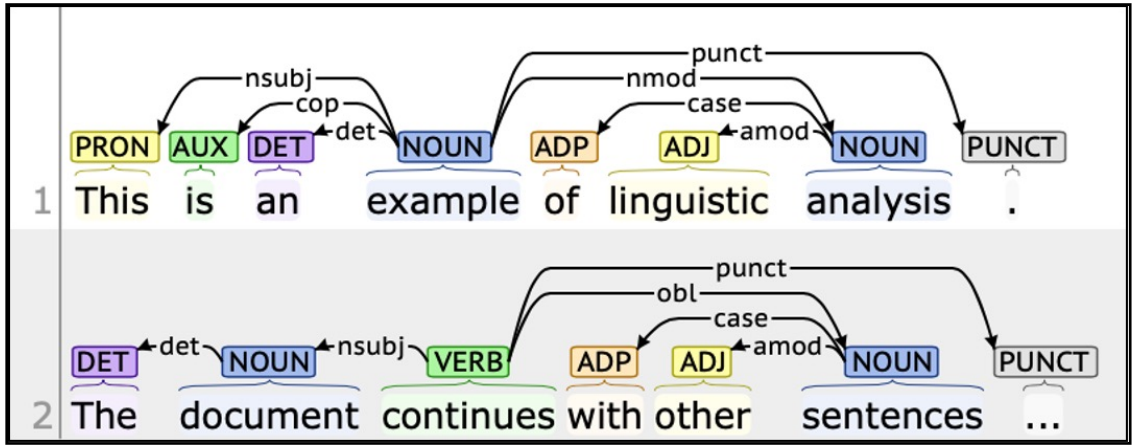


Linguistic knowledge

This is an example of linguistic analysis. The document continues with other sentences...

This is an example of linguistic analysis. The document continues with other sentences...

This is an example of linguistic analysis . The document continues with other sentences ...



- Sentence Splitting
- Tokenization
- PoS-tagging
- Lemma
- Parsing

Traditional Information extraction tasks

Proper nouns (entities)

named entity recognition

Relations

relation extraction

Terms

term extraction

Topics

topic modeling

Temporal expressions

temporal expressions

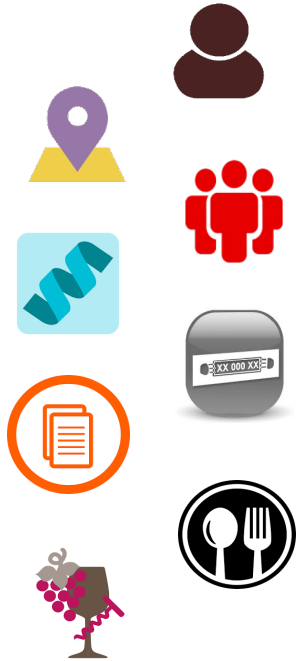
Events

event extraction

Named entity recognition and classification

- Named Entity recognition and classification refers to the task of **recognizing** and **classifying** proper nouns

- **Traditional classes**
  - PERSON
  - LOCATION
  - ORGANIZATION
- But **not limited to**…
  - Protein
  - Car license plate
  - Document
  - Food ingredient
  - Wine



Relations among entities

Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

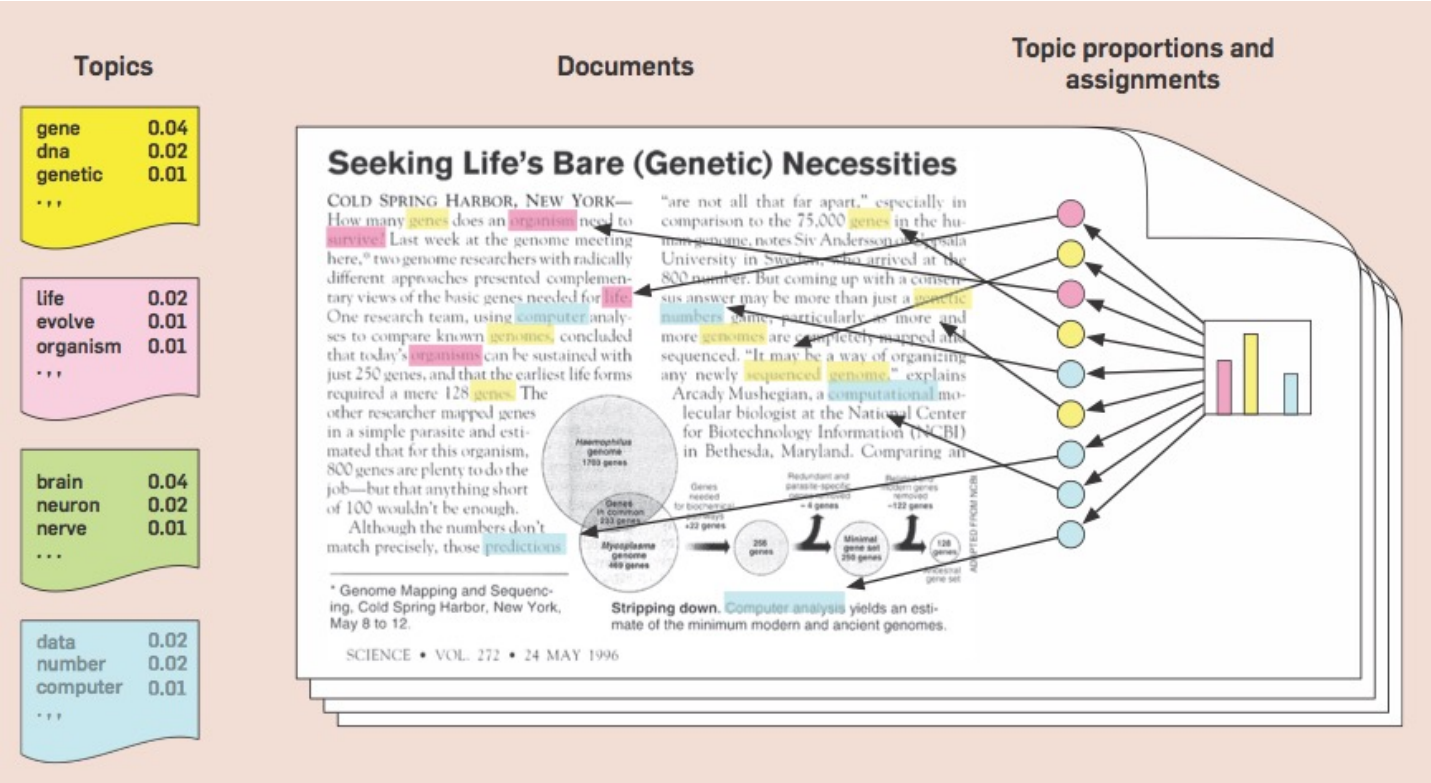
Figure 17.10 Semantic relations with examples and the named entity types they involve.

Term extraction

- Terms can be either simple (i.e. single tokens) or complex (i.e. larger information cores)
- Typically, complex terms are less polysemous and less ambiguous
- **Technical and domain terms**
  - operating system, periodic table
- **Idiomatic constructions**
  - kick the bucket (to die), break a leg (good luck)
- **Support verb construction**
  - pay attention, give support, …

Topic [Modelling | mining | Extraction]

- Topic Models are Bayesian generative models that identify the *topics* of a text
- A topic is a latent «theme» represented by a probability distribution of the words



NLP yesterday and Today

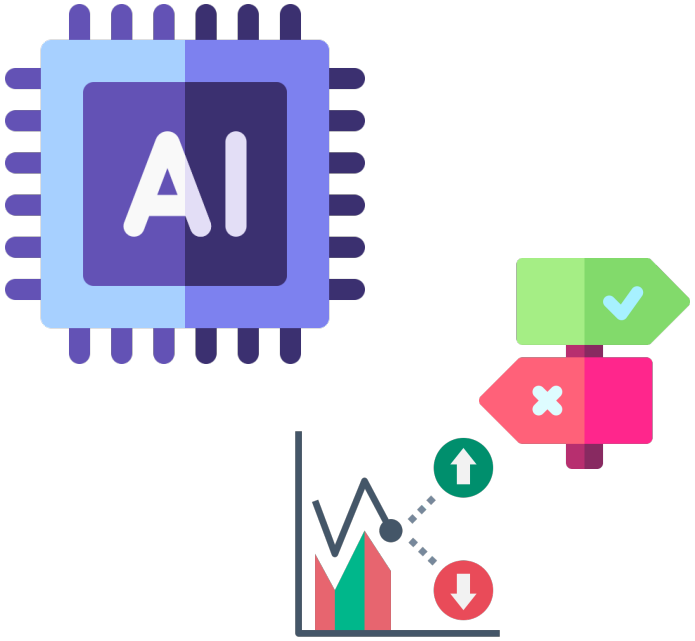
Yesterday

- Rule based learning
- Humans "teach" machines how to encode information from NL

Today

- Machine Learning (AI)
- Rules are automatically extracted from examples
  - Many, *many* examples…

What is AI good at?



- AI systems are incredibly good at
  - Recognizing patterns of information in the data
  - Making decisions & predictions based on such patterns

What Language Models are

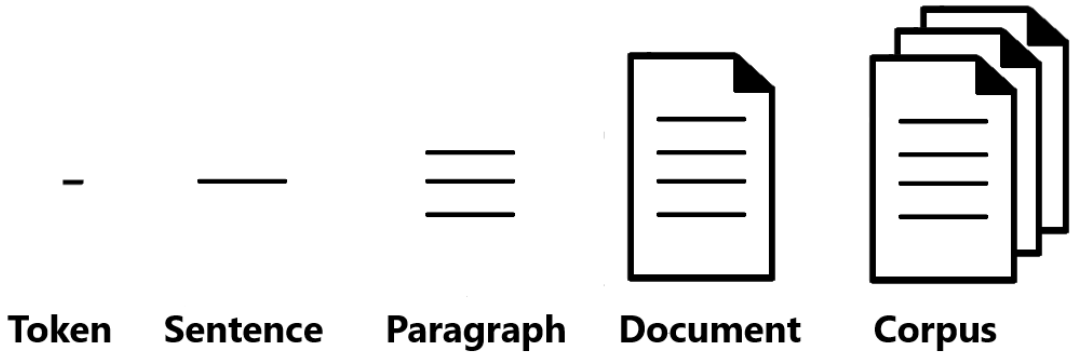
A Language Model (LMs) is a computational model designed to understand and generate human language.

How can we build a Language Model?

Two possible ways:

- Explicitly encode linguistic theories and language rules in an algorithm
- Study the statistical properties of language use and exploit them to build a model
  - Let's take a look at a couple of intuitions

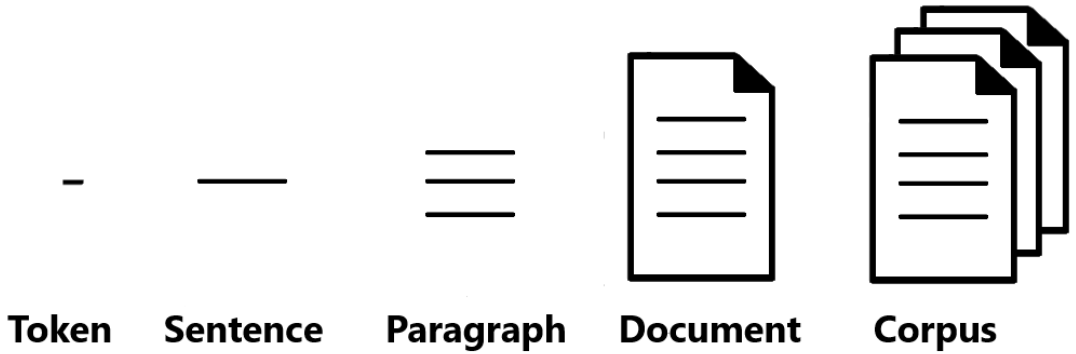
Words and probabilities



$$p(token) = \frac{f(token)}{len(Document)}$$

$$p(word|context) = \frac{p(context|word) * p(word)}{p(context)}$$

Sequences of tokens



We could compute the probability of a sequence with a Hidden Markov Model:

- oth order model: $P(w_1, \ldots, w_n) = P(w_1) * P(w_2) * P(w_3) * \cdots * P(w_n)$
- 1st order model: $P(w_1, \ldots, w_n) = P(w_1) * P(w_2|w_1) * P(w_3|w_2) * \cdots * P(w_n|w_{n-1})$
- 2nd order model: $P(w_1, \ldots, w_n) = P(w_1) * P(w_2|w_1) * P(w_3|w_2, w_1) * \cdots * P(w_n|w_{n-1}, w_{n-2})$
- …

The most basic form of a LM

The distributional hypothesis

«You shall know a word by the company it keeps»

J. R. Firth (1957)

Idea: Semantically similar words tend to occur in similar contexts

The meaning of a word can be inferred from the distributional patterns of other words that frequently appear nearby in a given corpus

Distributional semantics – an example

Sentences:

1. The curious cat chases the ball eagerly.
2. The playful dog fetches the ball eagerly.

Word-Context Matrix:

	the	curious	chases	ball	playful	fetches	eagerly
cat	1	1	1	1	0	0	1
dog	1	0	0	1	1	1	1

Distributional semantics – an example

Sentences:

1. The curious cat chases the ball eagerly.
2. The playful dog fetches the ball eagerly.

Word-Context Matrix:

	the	curious	chases	ball	playful	fetches	eagerly
cat	1	1	1	1	0	0	1
dog	1	0	0	1	1	1	1

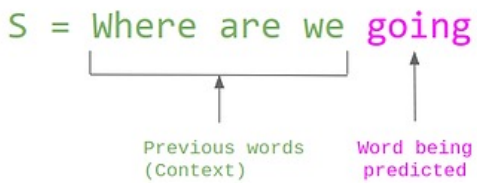
Training a Language Model

"A Language Model (LMs) is a computational model designed to *understand* and *generate* human language."

- Understand words with distributional semantics
- Generate words based on previous ones
- Do the above based on language data

Training a Language Model

- Collect **enough language data** to train a ML/DL model
  - Textual corpora representative of a language
- Train the model (a neural net) to predict words based on their contexts
  - A [MASK]ed word in a sentence (*Masked LM*)
  - The next word given previous ones (*Causal LM*)
- The model learns the statistical distribution of words
  - Latent representations that approximate words' meanings in their contexts, aka ***embeddings***
  - A probability distribution over words based on the surrounding/preceding ones



P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

Embeddings

- N-dimensional latent representations of words/sentences that encode their meaning
  - Latent → The "meaning" of each dimension is not explicitly modelled but is learned implicitly by the model
    - Like weights of a neural net: the specific weight's "meaning" is obscure
  - Similar embeddings represent linguistic events that have similar meanings
    - Similar = close in the n-dimensional space

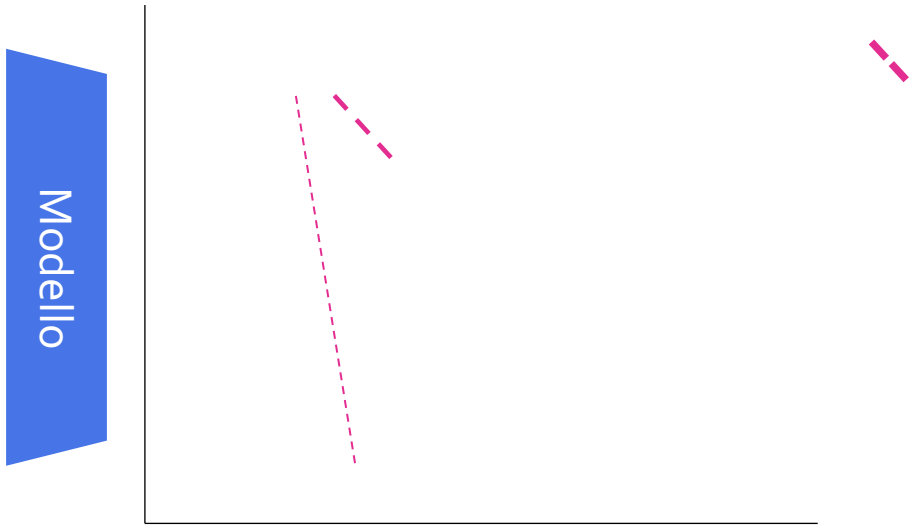
AI is cool

Math is hard

I love AI

I'd rather be on the beach

I study math



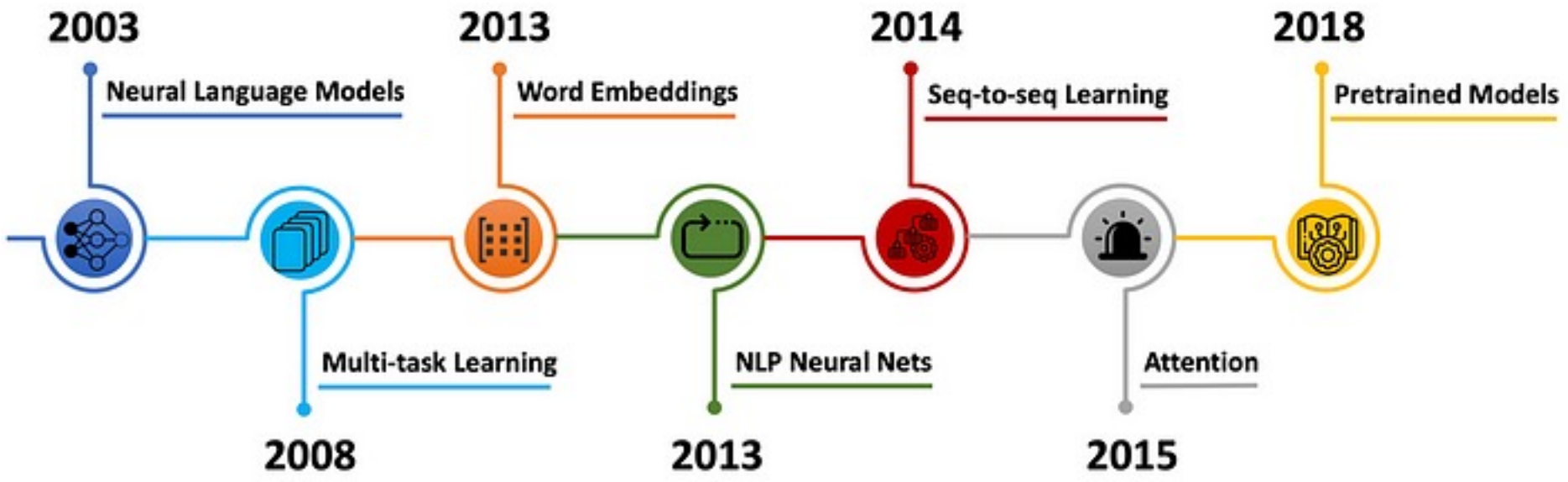
Text generation

- Given our learned probability distribution and an input sequence, we can try to predict what is the next most likely token of the sequence
  - *I'd love to see the new star wars* […] → Movie / Tv Show / …

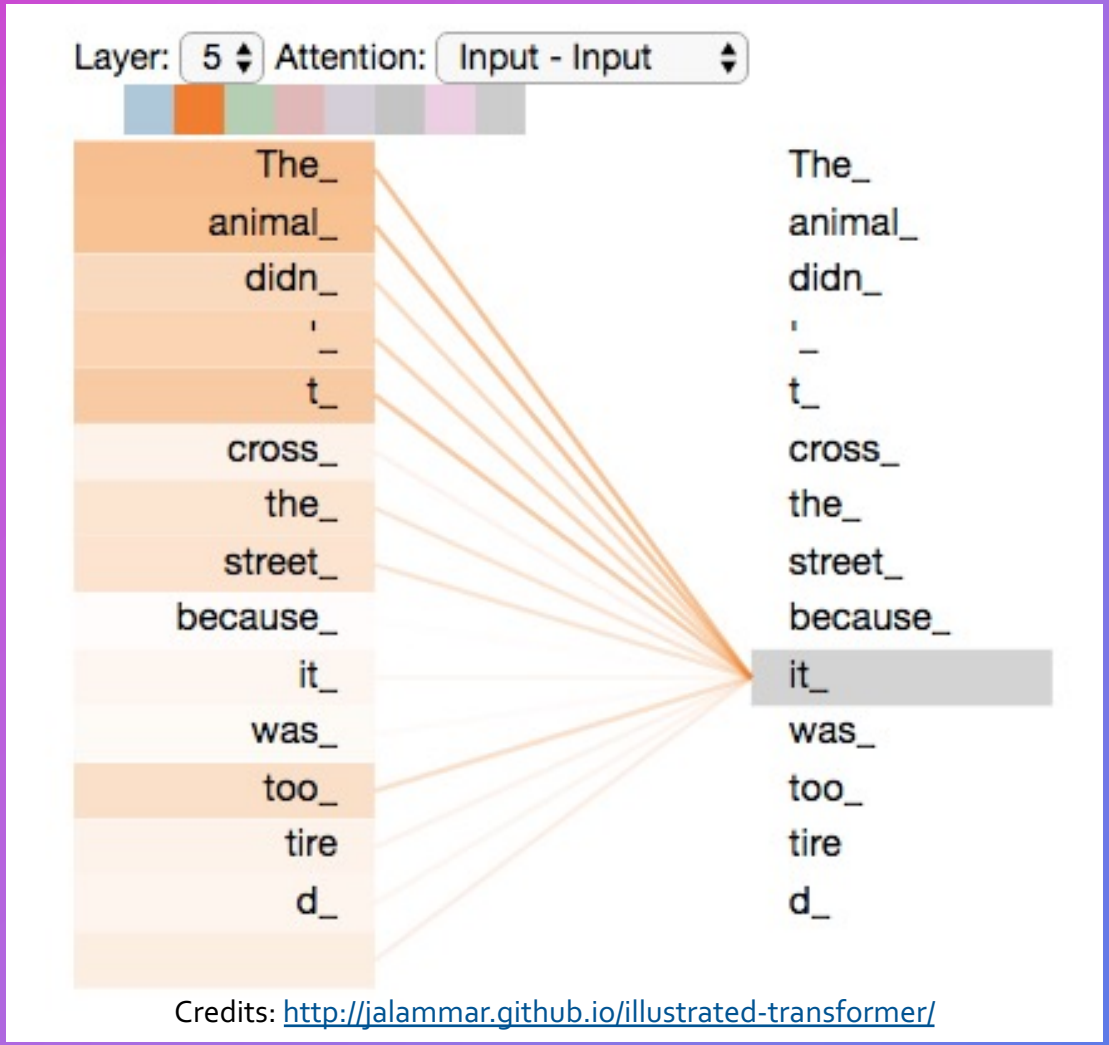
- And we can take our input sequence + its most likely next token and use this as input, and so on (until we reach the model's maximum input size) *autoregressively*

Evolution of Language Models (pre GPTs)

Models are tranied so solve a specific task (or set of tasks)
e.g., seq. classification, tranlslation, ecc…



Credits: https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-2-f5e575e8e37



Credits: http://jalammar.github.io/illustrated-transformer/

Attention

Idea: prioritise relevant information

- Within a sequence (self-attention)
- Across sequences (global)

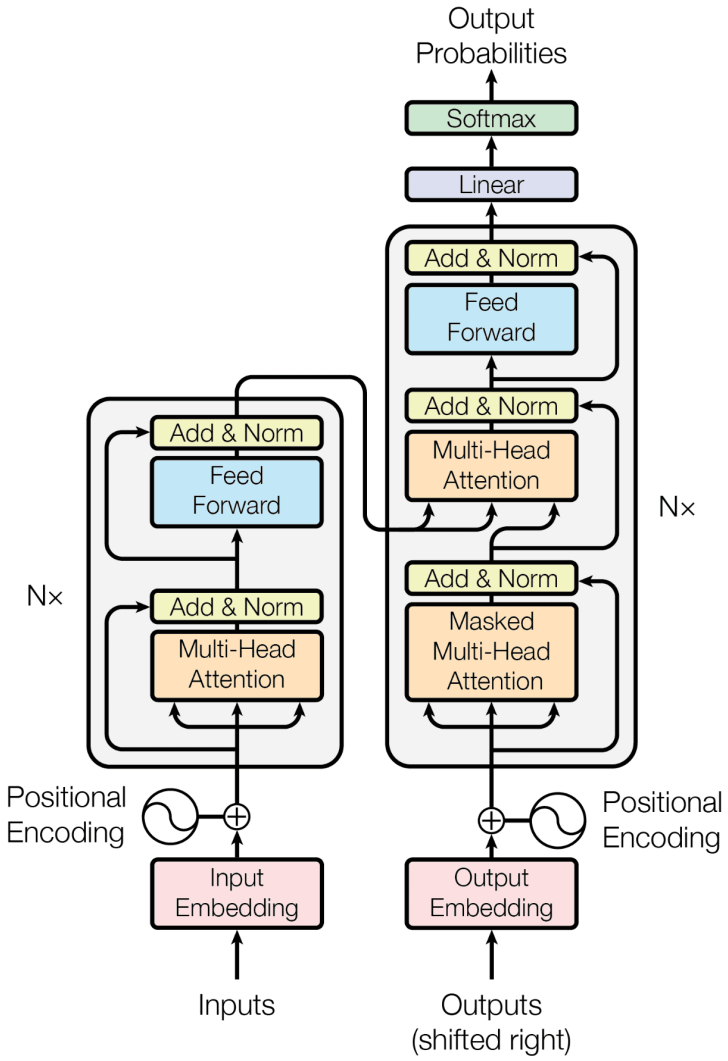
How: Assign importance to connections based on key, queries and values

Why: capture dependencies, parallelise process

The Transformer

Encoder

Captures input information



Decoder

Generates output based on encoded information

Three kinds of Transformer

Encoder only

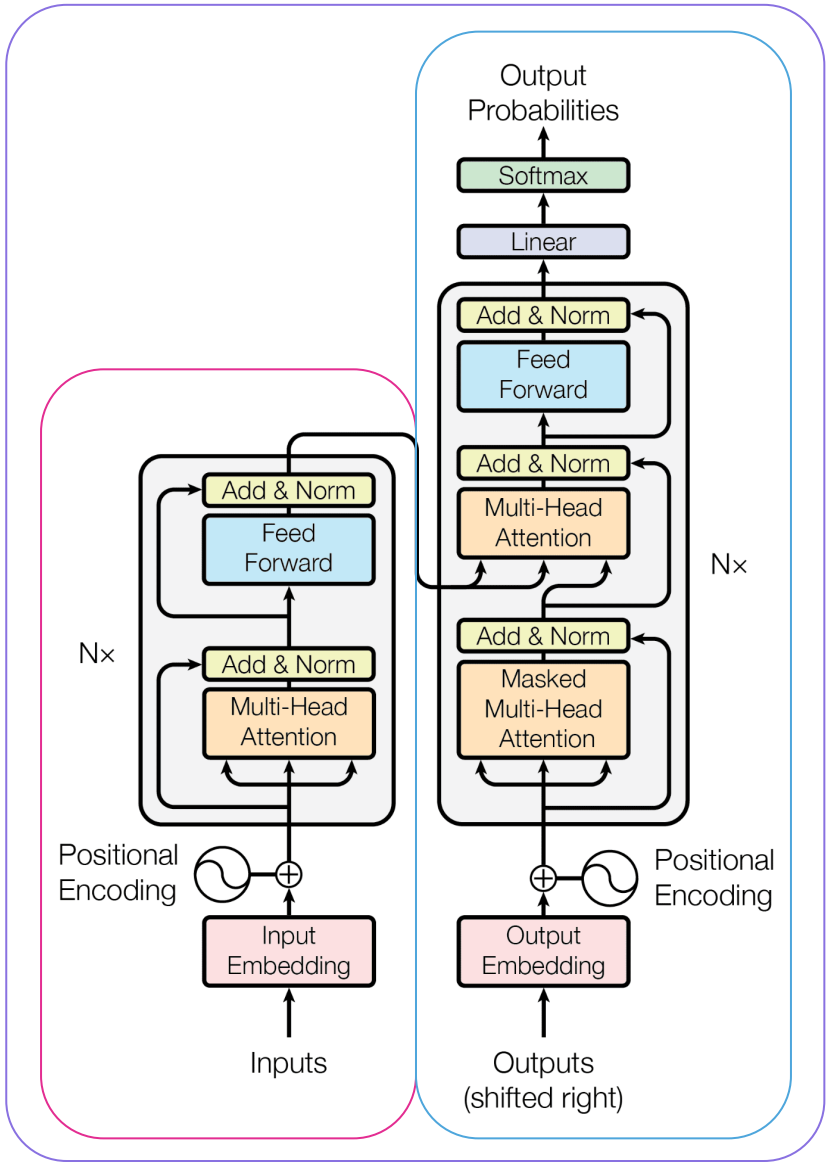
Used for NLU tasks where the input sequence is processed to extract useful representations or embeddings like token-level or document-level classification.
Example: BERT

Decoder only

Designed for NLG tasks where the model generates a sequence based on a given context or input.
Example: GPTs

Encoder-Decoder

Used for tasks that involve transforming an input sequence into an output sequence, where the output depends on the input sequence like Machine Translation
Example: T5



Pre-training & Fine-tuning

Pre-training: Train the model on 1+ languages

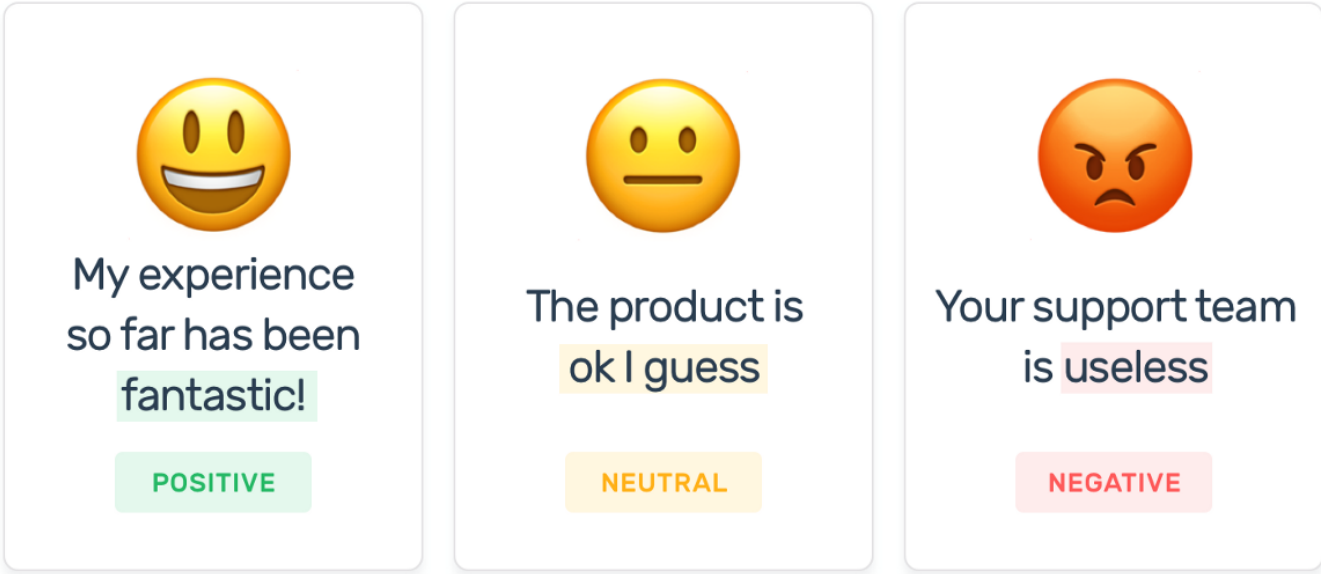
- Learn the distribution/representations based on data (unsupervised)
- A model that "knows" linguistic info

Fine tuning: Adapt the model to a specific downstream task

- Using labelled data (supervised)
- Model the relation between inputs (embeddings) and the task by further training e.g. the final few layers of the Transformer (+ an additional clasifier)

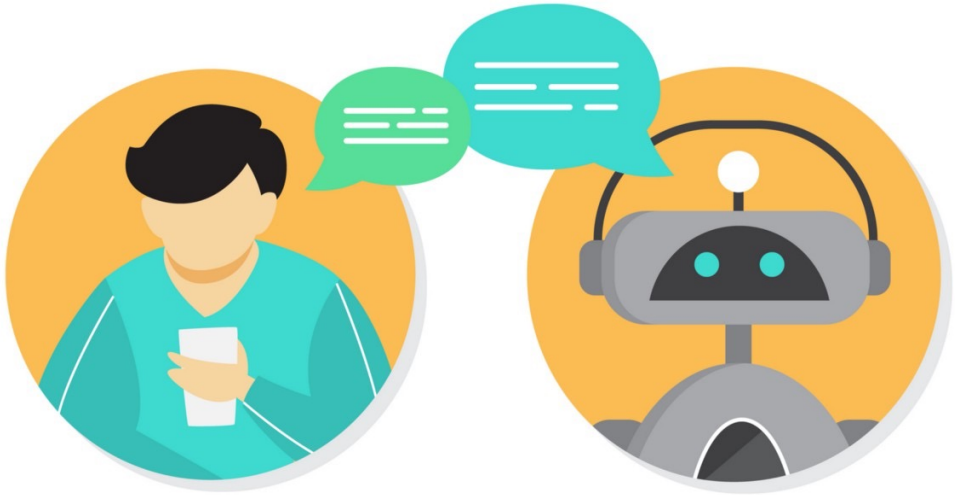
Text classification

- Given a text, and a range of categories to which it may belong, identify the appropriate category
  - **Affect**
    - Sentiment Analysis
    - Emotion Detection
    - Hate speech detection
    - Abusive language detection
    - …
  - Topics
  - Spam Detection
  - …



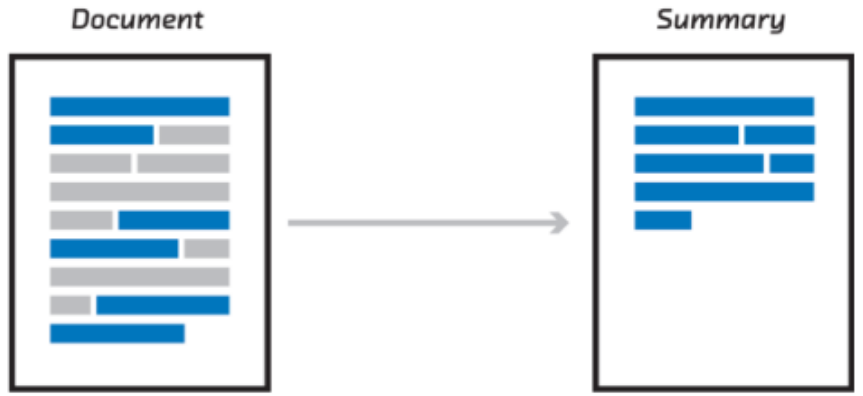
Question Answering

- Allows to answer questions of general understanding concerning a collection of documents
- **Various levels of difficulty:**
  - "What is the capital of France?"
  - "How many years have passed since Dante's birth?"
  - "What were the social effects of the pandemic?"

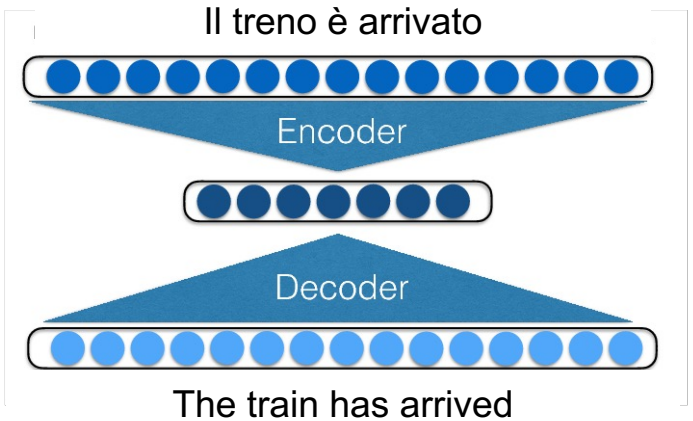


Text summarization

- Given a text, automatically create its summary
- **Extractive summarization**
  - Extracts relevant sentences from the original document
- **Abstractive summarization**
  - Generates a new, shorter text



Machine translation

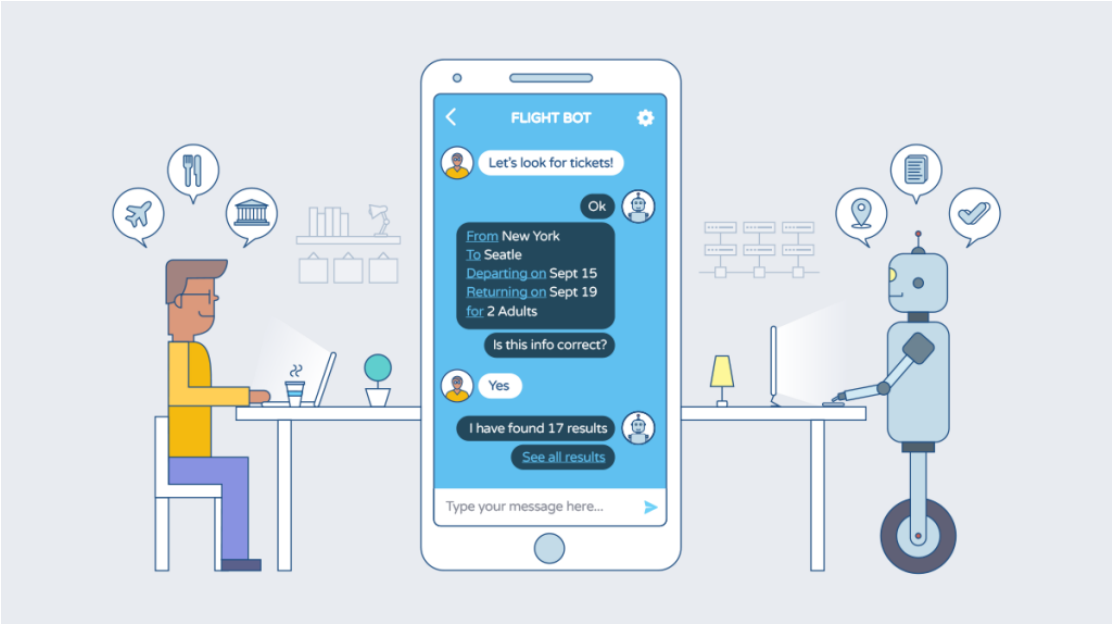


È arrivato un treno.

A train has arrived.

Natural Language Generation (NLG)

- Starting from a ***prompt***, the system generates a text in natural language (e.g. its follow up)
- Useful in many contexts
  - Chatbots
  - Virtual assistants
  - Summarization
  - Automatic content creation
  - …



Part II

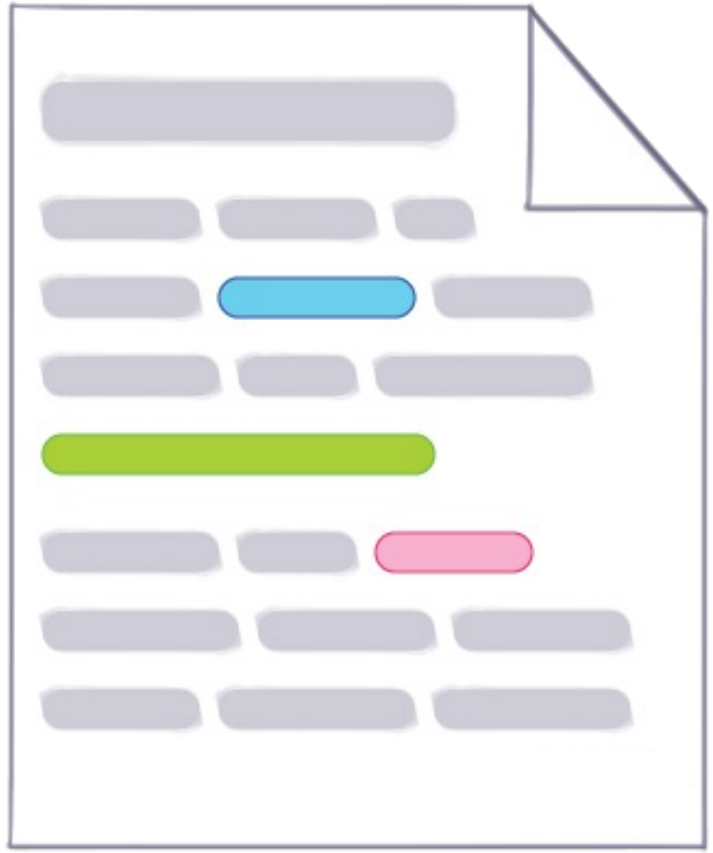
NLP for health

What can NLP do for healthcare?

- Both *traditional* and more modern NLP solutions have their place in a healthcare-centered pipeline
- Key aspects
  - Extract information from medical texts to obtain structured knowledge
  - Categorize/cluster clinical narratives
  - NLU and NLG for patient monitoring and telemedicine
  - Foster research in medical and biomedical fields
- Let's see a few examples

Clinincal text processing

- **Goal**: Extract structured data/information from unstructured clinical texts
- **How**: Process clinical documents with specialized NLP pipelines
- **Example**: Entity Linking and Named Entity Recognition (NER) for identifying medical concepts in text



Entity Linking

Entity mentions are detected and classified in the Unified Medical Language System (UMLS)

Spinal and bulbar muscular atrophy (SBMA) is an inherited motor neuron disease caused by the expansion of a polyglutamine tract within the androgen receptor (AR). SBMA can be caused by this easily.

Spinal ENTITY and bulbar muscular atrophy ENTITY (SBMA ENTITY) is an inherited ENTITY motor neuron disease ENTITY caused by the expansion ENTITY of a polyglutamine tract ENTITY within the androgen receptor ENTITY (AR ENTITY). SBMA ENTITY can be caused by this easily.

	text	Canonical Name	Concept ID	TUI(s)	Score	start	end
0	Spinal	spinal	C0521329	T082	1	0	1
1	bulbar muscular atrophy	Bulbo-Spinal Atrophy, X-Linked	C1839259	T047	0.909614	2	5
2	SBMA	Bulbo-Spinal Atrophy, X-Linked	C1839259	T047	1	6	7
3	inherited	Hereditary	C0439660	T169	1	10	11
4	motor neuron disease	Motor Neuron Disease	C0085084	T047	1	11	14
5	expansion	cell growth	C0007595	T043	0.864297	17	18
6	androgen receptor	AR gene	C1367578	T028	1	24	26
7	AR	AR gene	C1367578	T028	1	27	28
8	SBMA	Bulbo-Spinal Atrophy, X-Linked	C1839259	T047	1	30	31

Specialized NER

A specialized NER model trained on specific NEs is used annotatate tokens (and spans of tokens) in the text

Spinal and bulbar muscular atrophy (SBMA) is an inherited motor neuron disease caused by the expansion of a polyglutamine tract within the androgen receptor (AR). SBMA can be caused by this easily.

Spinal and bulbar muscular atrophy **DISEASE** (SBMA **DISEASE**) is an inherited motor neuron disease caused by the expansion of a polyglutamine tract within the androgen receptor (AR). SBMA **DISEASE** can be caused by this easily.

This model recognizes diseases and chemical compounds

Specialized NER

A specialized NER model trained on specific NEs is used annotatate tokens (and spans of tokens) in the text

Spinal and bulbar muscular atrophy (SBMA) is an inherited motor neuron disease caused by the expansion of a polyglutamine tract within the androgen receptor (AR). SBMA can be caused by this easily.

Spinal and bulbar muscular atrophy (SBMA) is an inherited motor neuron disease caused by the expansion of a polyglutamine tract within the androgen receptor **PROTEIN** (AR **PROTEIN**). SBMA can be caused by this easily.

This model is specialized on proteins, cells and DNA entitiies

Clinical Decision Support Systems

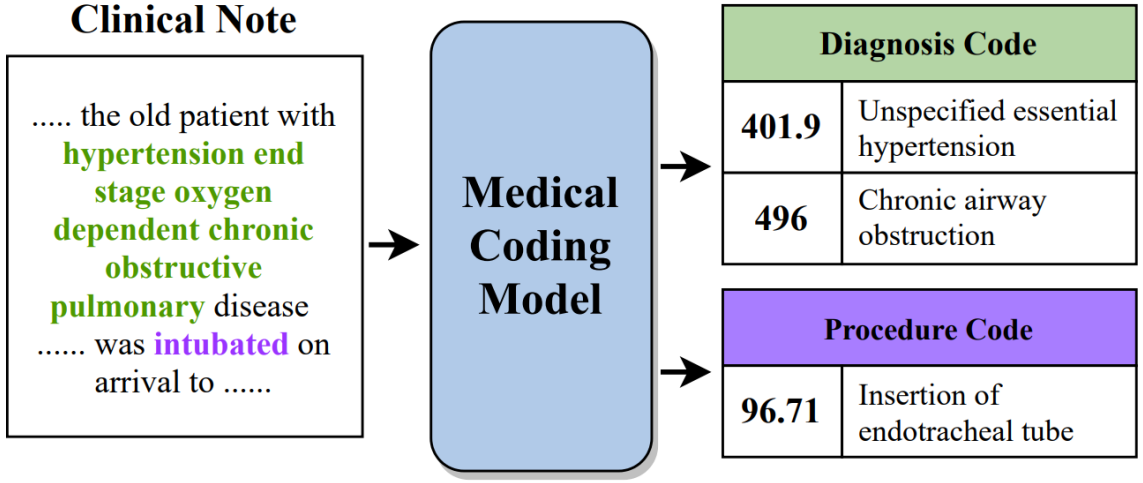
Goal: analyze patients' data to provide real-time recommendations for clinicians

How:

- Specialized NLU models trained on medical literature and procedures
- Predictive models based on patients' clinical history, treatment and outcomes

Medical coding (and billing)

- **Goal**: Automatically convert clinical narratives into standardized medical codes
- **How**: Language Model fine-tuned on classification
  - Clinical note to codes/procedures
- Can be used for:
  - Faster and more efficient processing in adiminstrative/management softwares within medical structures
  - Medical billing/insurance ☹



Pharmacovigilance and Adverse Drug Event Detection

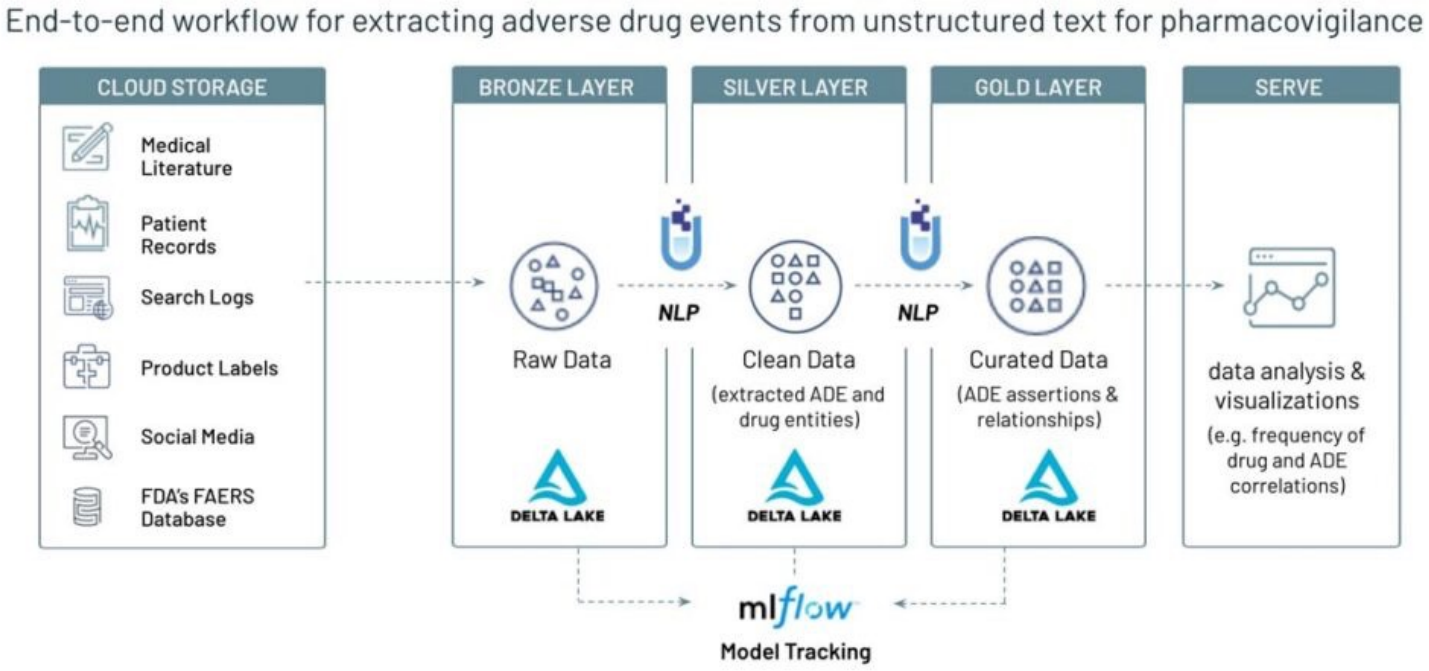


Image from Databricks

- **Goal**: monitor and identify possible adverse drug events during patient's treatment
- **How**: use NLP pipelines to identify drug interactions and side effects mining electronic health records of patients + literature, product labels etc.

Telemedicine & Remote monitoring

Telemedicine

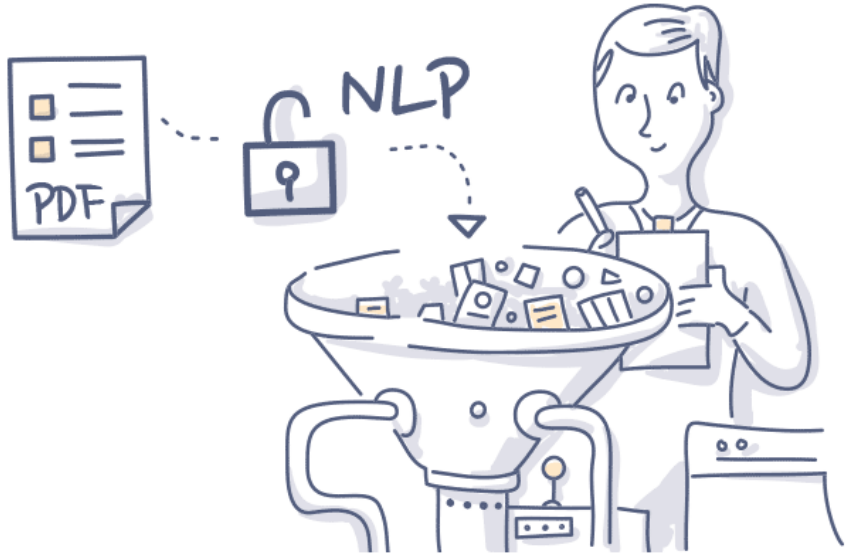
- NLU models to analyze medical data and conversations from virtual visits
- Conversational AI agents for Q&A and recommendations

Remote Monitoring

- Example: Analyze web data like social media and online forums for signs of mental health issues
  - Using Sentiment Analysis and Emotion Detection techniques

Fostering Research

- NLP has drastically improved biomedical research
- Provides tool for mining biomedical literature for relevant health information
  - Insights for drug discovery, disease understanding, and treatment options
- The Transfomer architecture is in some capacity behind:
  - **Alpha Fold (DeepMind):** predict 3D structure of proteins, a 50-years old unsolved challenge
  - Key advancements of last 3-4 years in the **Human Genome Project** to map the human genome



Part III

Hands on (with Jupyter Notebooks)

NLP pipelines

- Several NLP pipelines for linguistic annotation and Information Extraction developed through the years

NLTK

Natural Language Toolkit

The Old School, each step in the pipeline is handled separately.
Not many extensions.

Relevant mostly for NLP students.

Stanza

Stanford NLP group

Python version of the very popular CoreNLP.
Best performances, many languages, not so many add-ons.

Relevant mostly for academics.

SpaCy

"Industrial-strenght NLP"

Simple and fast, an industry standard and many available plug-ins for ML and custom workflows + easy visualization.

Relevant for everyone.

SpaCy

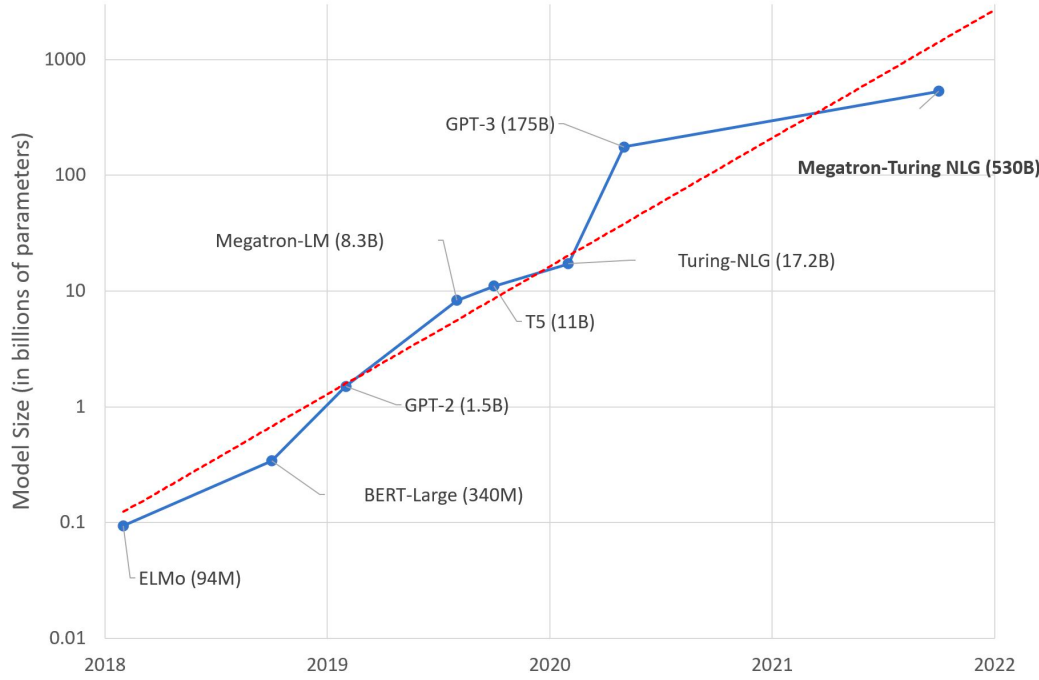
Demo Notebook
(including SpaCy usage)

- SpaCy with custom pipelines and models for scientific docs
  - Tokenizer with tokenization rules on top of spaCy's rule-based tokenizer
  - POS tagger and syntactic parser trained on biomedical data
  - Entity span detection model
  - NER models for more specific tasks

Model	Description
en_core_sci_sm	A full spaCy pipeline for biomedical data with a ~100k vocabulary.
en_core_sci_md	A full spaCy pipeline for biomedical data with a ~360k vocabulary and 50k word vectors.
en_core_sci_lg	A full spaCy pipeline for biomedical data with a ~785k vocabulary and 600k word vectors.
en_core_sci_scibert	A full spaCy pipeline for biomedical data with a ~785k vocabulary and `allenai/scibert-base` as the transformer model. You may want to use a GPU with this model.
en_ner_craft_md	A spaCy NER model trained on the CRAFT corpus.
en_ner_jnlpba_md	A spaCy NER model trained on the JNLPBA corpus.
en_ner_bc5cdr_md	A spaCy NER model trained on the BC5CDR corpus.
en_ner_bionlp13cg_md	A spaCy NER model trained on the BIONLP13CG corpus.

Neural Language Models

- Transformer-based LLMs are extremely popular
  - Almost a new model every other day
  - Different sized classes of LLMs – 1.5 to 175+ B parametres
    - Roughly 2x the number of Billion of parameters in GPU memory to run a model locally in half precision (e.g., a 1.5B model runs on a 3GB GPU, a 7B model requires ~15 GB memory and so on…)
  - **Bigger is not always better** for narrower-focused domains such as health
  - Chose the best kind of model for the problem
    - Is it Classification? Text generation? Translation? …



Fantastic LLMs and where to find them

HuggingFace

Code libraries in Python for running/training most LLMs (except commercial ones)

- Transformers, Diffusers, Accelerate, Tokenizers, and more

Models repository

- Pre-trained & fine-tuned models ready to run in the code libraries

Datasets repository

- Open sourced datasets to train LLMs

Guides, tutorials, code samples and more…

https://huggingface.co/

Transformers & Pipelines

- **Flexible Configuration**: Customize models to suit specific tasks and requirements
- **Cross-Platform Support**: Compatible with various programming languages and environments
- **Simple Interface**: Intuitive APIs for model loading, training, and inference (with pipelines)

What are Pipelines?

- **High-level interfaces** for common NLP tasks
  - Abstract away complex implementation details
- Execute tasks such as text/token classification, generation, summarization etc. with minimal code
  - Literaly, no more than 5-10 lines of code including imports…

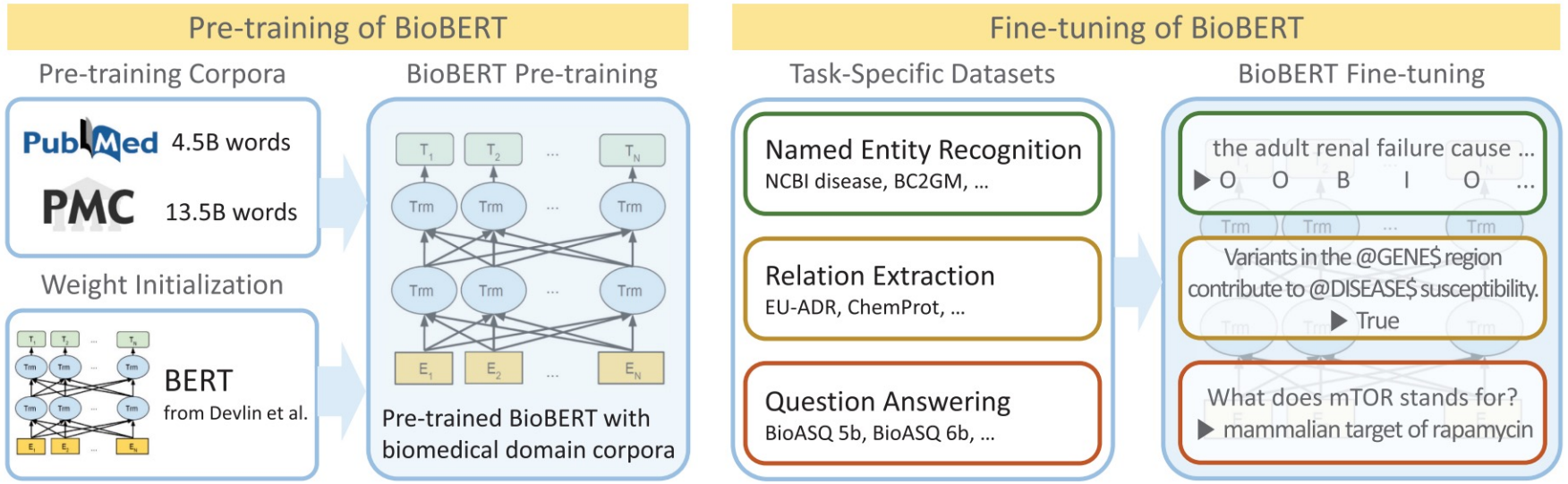
```
from transformers import pipeline

mask_filler = pipeline("fill-mask", "username/my_awesome_eli5_mlm_model")
mask_filler(text, top_k=3)
```

Example: BioBERT

Demo Notebooks for BioBERT (fill mask + NER)
& other Medical NER
(with HuggingFace Pipelines)

- A BERT model pre-trained on biomedical data
- Fine-tuned checkpoints for different tasks in the biomedical domain



Lee et al., 2019

To sum up

Staggering evolution of NLP in the past 10 years

From count-based and Markov-chain based LMs to ChatGPT

NLP-based systems enable a wide array of tasks that involve natural language

Linguistic analysis but also information extraction, text categorization, text generation

Healthcare can strongly benefit from the implementation of NLP techniques

For helping both doctors, patients, and researchers

Many open models for biomedical and scientific data

Still some open challenges, both technological and most importantly ethical

THANK YOU

Questions?

If you are curious about NLP feel free to drop me a line at

alessandro.bondielli@unipi.it