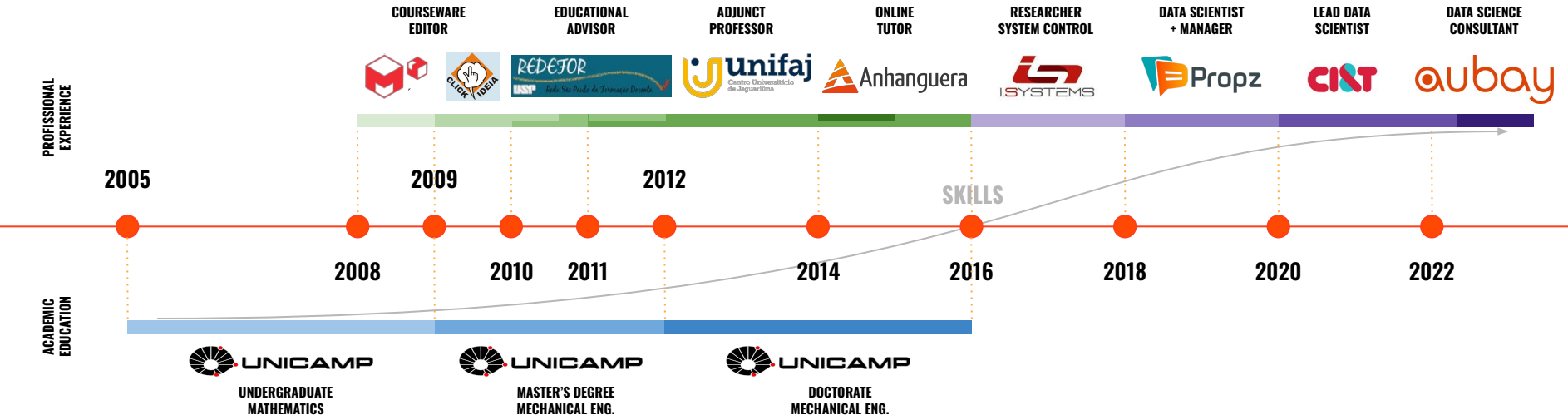
Unlocking text data to leverage business decision making through Natural Language Processing systems

Luis Rodrigues
Data Science Consultant

luis-rodrigues-phd

PROFESSIONAL TRAJECTORY OF LUIS RODRIGUES



PROJECT EVOLUTION

Propz

- RaiaDrogasil — MARKETING/CRM PROJECT
- Carrefour — AI PROJECT

CI&T

- vivo — INNOVATION PROJECT
- CENTAURO — INNOVATION PROJECT
- Nestlé — AI PROJECT

aubay

- accenture — NLP PROJECT
- BANCO DE PORTUGAL EUROSISTEMA

AGENDA

01 › 02 › 03

Motivation

NLP Theory

Application

01 › 02 › 03

Motivation | NLP Theory | Application

THE MOST AMAZING INNOVATION IN NATURAL LANGUAGE PROCESSING

Give me a text for a talk describing what is the most amazing Innovation in Natural Language Processing today and why. The answer must be a maximum of 100 words.

The most amazing innovation in Natural Language Processing today is GPT-3, an AI model capable of generating human-like text. With its 175 billion parameters, GPT-3 can understand and respond to a wide range of prompts, from answering questions to creative writing. Its ability to mimic human language patterns, comprehend context, and generate coherent responses is revolutionary. GPT-3's potential to enhance various industries, such as customer service, content creation, and language translation, makes it a truly remarkable advancement in NLP.

Source: OpenAI's chatGPT (2023), GPT-3: Revolutionary NLP Innovation, accessed June 2023

NLP APPLICATIONS AND USE CASES

Spam Classification & Priority Inbox

Event Detection & Financial Sentiment Analysis

Calendar Event Extraction

@congratsbot

Legal Entity Extraction

Template Filling (2019 UK elections)

Review Analysys Keyword Extraction

Credit Risk Assessment

Text Classification

Information Extraction

Chatbot & Conversational Agent

Information Retrieval

IBM Watson Assistant

OpenAI chatGPT

Google

Bing

Questions Answering

Amazon Alexa

Microsoft Cortana

Apple Siri

IBM Watson

Google

LangChain

Elasticsearch

01 › **02** › 03

Motivation

NLP Theory

Application

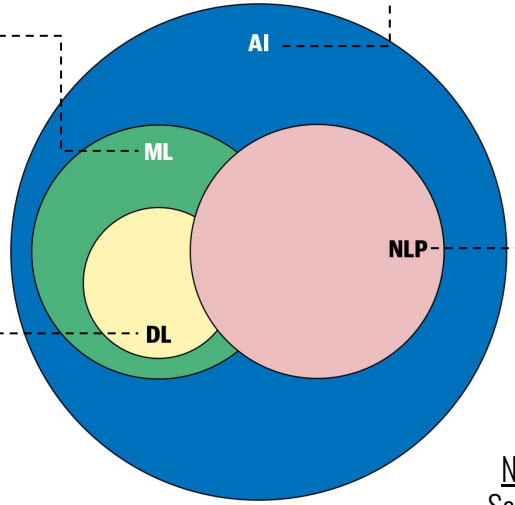
HOW AI, NLP, ML, AND DL ARE RELATED

“ Artificial Intelligence is the science and engineering of making intelligent machines. ”

JOHN MCCARTHY (1956)
One of the "founding fathers" of AI

“ Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed. ”

ARTHUR SAMUEL (1959)
Pioneer in computer gaming and AI



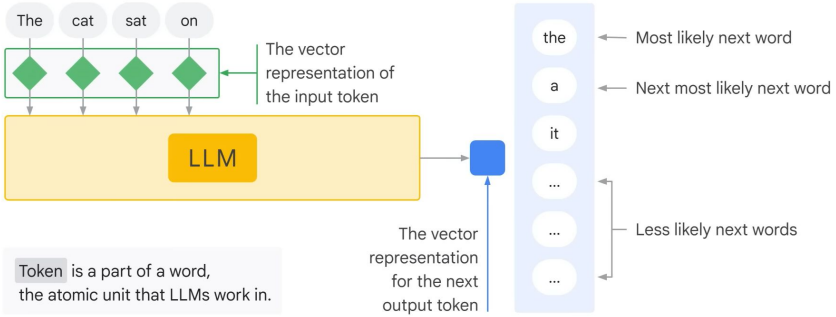
Deep Learning is a subset of machine learning based on (deep) artificial neural networks, usually applied to unstructured data.

Natural Language Processing is an area of Computer Science that deals with methods to analyze, model, and understand human language.

Source: Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. 2020. Practical Natural Language Processing. O'Reilly Media, Inc.

NLP TASKS | LANGUAGE MODELING

This is the task of predicting what the next word in a sentence will be based on the history of previous words.



Source: Google Cloud Tech (2023). Introduction to Large Language Models, accessed June 2023

The goal of this task is to learn the probability of a sequence of words appearing in a given language.

Language modeling is useful for building solutions for a wide variety of problems, such as text generation, machine translation, speech recognition, optical character recognition, and spelling correction.

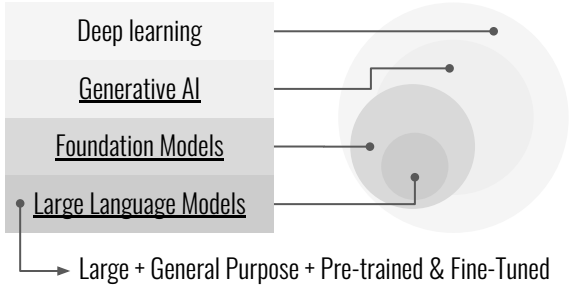
Language models do not require labelled data once they are trained in a self-supervised manner predicting a word based on the previous context, which a process called *autoregression*.

There is also Masked Language Modeling, which is a technique used for pretraining language models, specifically for tasks like fine-tuning and transfer learning. It involves masking certain words or tokens in a given text and training the language model to predict those masked words based on the surrounding context.

Masked Language Modeling with HuggingFace Transformers.ipynb

```
from transformers import pipeline

model = pipeline("fill-mask",
model="distilroberta-base")

text = "The capital of France is
<mask>."

output = model(text)
masked_word = output[0]["token_str"]

print(masked_word)

>>> Paris
```

LARGE LANGUAGE MODELS



Two main types of LLMs

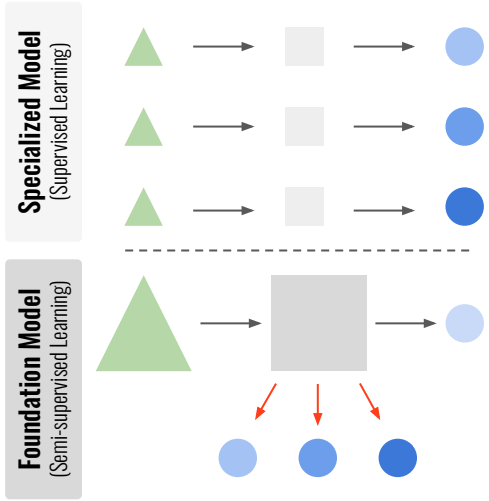
Base (Generic or Raw) LLM - Predicts next the next word (token)
- ✔ Once upon a time, there was a unicorn
- ✘ What is the capital of France?

Instruction Tuned LLM - Tries to follow instructions
(fine-tuned on instruction-response pairs)

Pros and Cons

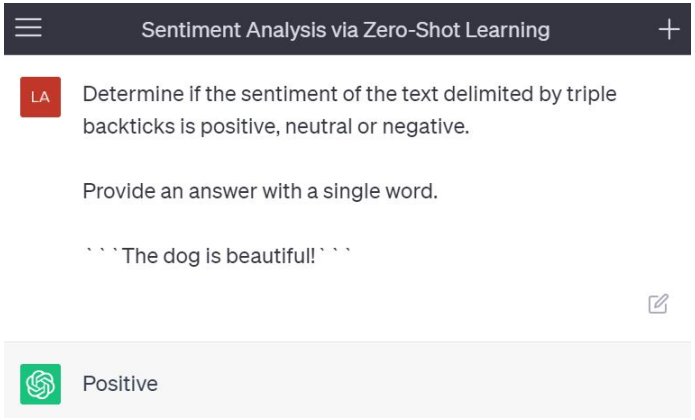
Productivity	Cost
Performance	Trust

Paradigm Shift



→ Prompt Engineering

(if you have no data or have few domain data)

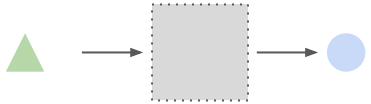


Source: OpenAI's chatGPT (2023). Sentiment Analysis Example via Zero-Shot Learning, accessed June 2023

→ Fine-tuning

(if you have domain data)

Fine-tuning

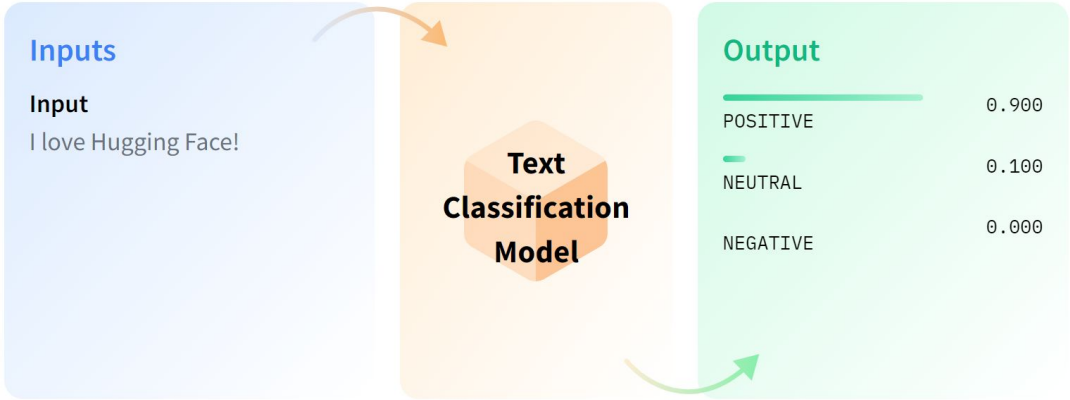


Transfer Learning

qubay

NLP TASKS | TEXT CLASSIFICATION

This is the task of bucketing the text into a known set of categories based on its content.



Source: Hugging Face (2023). Text Classification, accessed June 2023

Text classification is by far the most popular task in NLP and is used in a variety of tools.

Some famous task variants include *Sentiment Analysis*, *Natural Language Inference* (NLI) and *Question NLI*.

Beyond building your own model or using some pre-trained one, you can use existing APIs from cloud service providers (eg, Google, Amazon, Azure, IBM), that provide off-the-shelf content classification models.

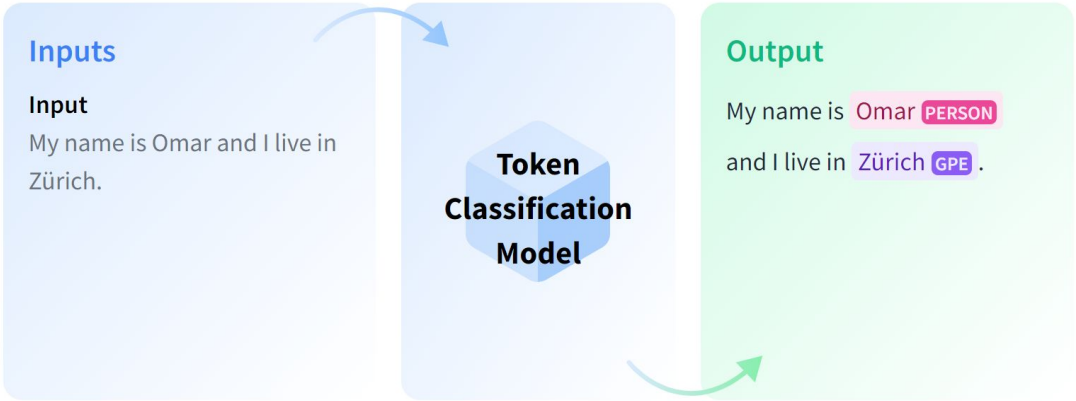
More recently, with the emergence of Large Language Models, there are also zero-shot learning, *one-shot learning* and *few-shot learning* (classification), that can be thought of as an instance of transfer learning.

Zero-Shot Classification with HuggingFace Transformers.ipynb

```
from transformers import pipeline
import numpy as np

model =
pipeline("zero-shot-classification",
model="facebook/bart-large-mnli")

text = "I have a problem with my iphone
that needs to be resolved asap!"
candidate_labels = ["urgent", "not
urgent", "phone", "tablet"]

output = model(text, candidate_labels)

index_best_label =
np.argmax(output["scores"])
best_label =
output["labels"][index_best_label]
print(best_label)
>>> urgent
```

NLP TASKS | INFORMATION EXTRACTION

As the name indicates, this is the task of extracting relevant information from text, such as calendar events from emails or the names of people mentioned in a social media post.



Source: Hugging Face (2023). Token Classification, accessed June 2023

Some specific Information Extraction tasks include *keyword and phrase extraction*, *pattern-based extraction* (e.g., addresses, phone numbers, dates, ...), *named entity recognition* (e.g., names of persons, locations, and organizations, dates, amounts, and so on), *named entity disambiguation and linking* (e.g., Albert Einstein, Einstein, ...), and *relationship extraction* (e.g., connect Steve Jobs to Apple with the relationship of co-founder).

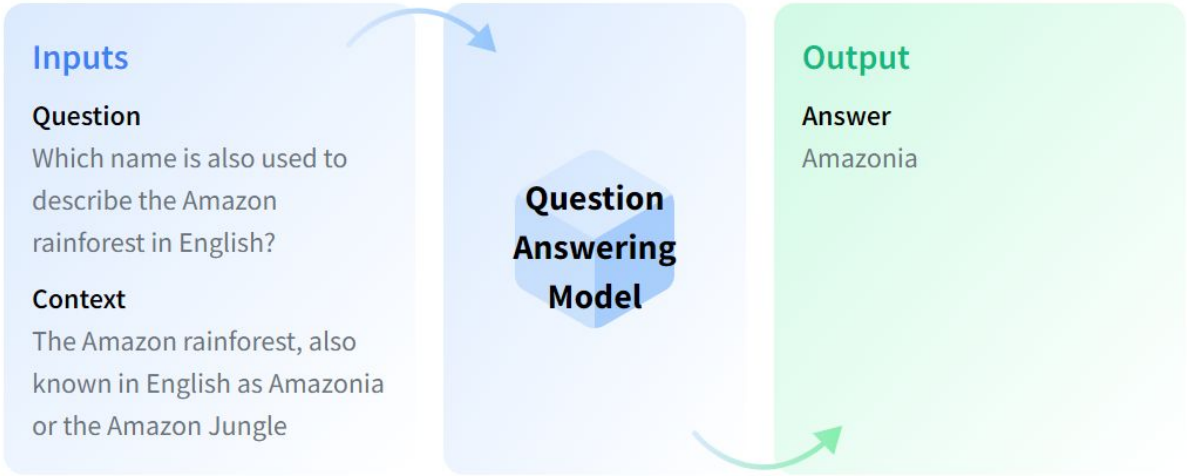
It's useful for several downstream NLP tasks, such as information retrieval, automatic document tagging, recommendation systems, text summarization, question-answering systems, etc.

Named Entity Recognition (NER) with Spacy.ipynb

```
import spacy

nlp = spacy.load("en_core_web_sm")

doc = nlp("""Apple is looking at buying U.K. startup for $1 billion, the company's chief executive Tim Cook has revealed, on his trip to Brazil on Tuesday.""")

for ent in doc.ents:
    print(ent.text, ent.label_)

>>> Apple ORG
>>> U.K. GPE
>>> $1 billion MONEY
>>> Tim Cook PERSON
>>> Brazil GPE
>>> Tuesday DATE
```

NLP TASKS | QUESTIONS ANSWERING

This is the task of building a system that can automatically answer questions posed in natural language.



Source: Hugging Face (2023). Question Answering, accessed June 2023

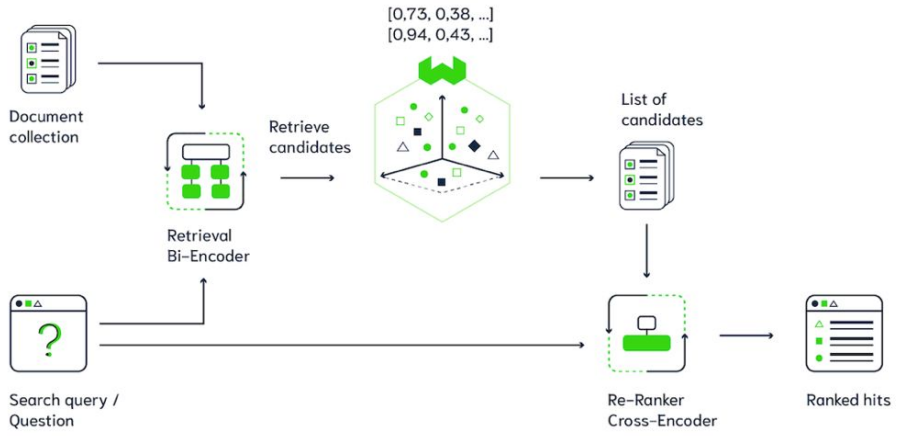
QA models differ in the way answers are created (*extractive* or *generative*) and in where answers are taken from (*open*, from a provided context, or *closed*, answer is completely generated by a model).

Question Answering (QA) with HuggingFace Transformers.ipynb

```
from transformers import pipeline

model = pipeline(model =
"distilbert-base-cased-distilled-squad)"

question = "Which name is also used to
describe the Amazon rainforest in English?"

# Example with 150+ words and 1000+ characters
context = """The Amazon rainforest
(Portuguese: Floresta Amazônica or Amazônia;
Spanish: Selva Amazónica, Amazonía or usually
Amazonia; French: Forêt amazonienne; Dutch:
Amazoneregenwoud), also known in English as
Amazonia or the Amazon Jungle, is a moist
broadleaf forest that covers most  ..."""

output = model(question = question, context =
context)
answer = output["answer"]

print(answer)

>>> Amazonia
```

NLP TASKS | INFORMATION RETRIEVAL

This is the task of finding documents relevant to a user query from a large collection.



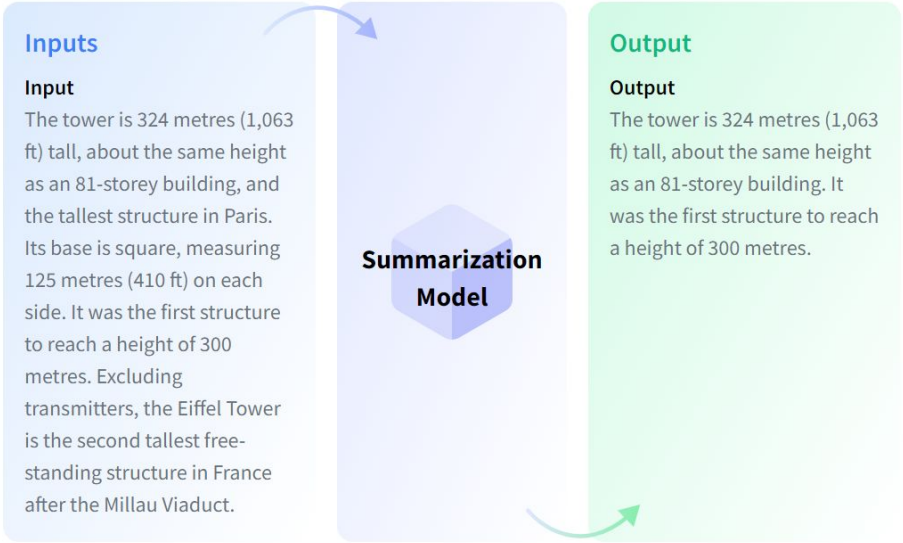
Source: Weaviate (2022). Using Cross-Encoders as reranker in multistage vector search, accessed June 2023

Applications like Google Search are well-known use cases of information retrieval.
For the retrieval, we can use either lexical search (eg, ElasticSearch) or semantic search (eg, FAISS, Chroma).
Lexical search is a fast and efficient way to find exact matches (eg, people, places, things) in large databases.
Semantic search overcomes the shortcomings of keyword-based search by looking at the meaning of the text.
For complex tasks, as question answering, the search can significantly be improved using Retrieve & Re-Rank.

```
Sentence Similarity with Sentence
Transformers and HuggingFace
Transformers.ipynb

from sentence_transformers import
SentenceTransformer, util

model =
SentenceTransformer('sentence-transf
ormers/all-MiniLM-L6-v2')

sentences = ["what is happiness?",
"Happiness is a state of the
spirit."]
embeddings = model.encode(sentences)

similarity =
util.pytorch_cos_sim(embeddings[0],
embeddings[1]).numpy()[0][0]
print(similarity)

>>> 0.7928971
```

NLP TASKS | TEXT SUMMARIZATION

This task aims to create short summaries of longer documents while retaining the core content and preserving the overall meaning of the text.



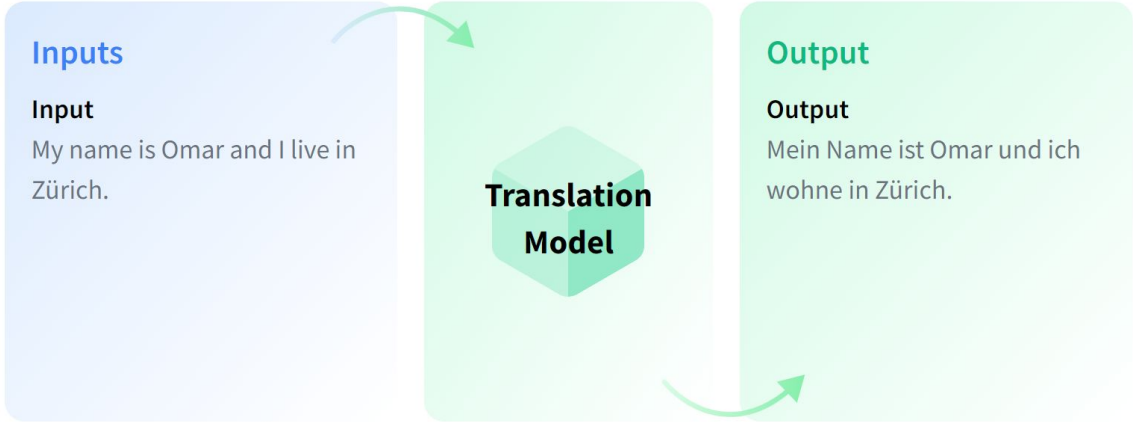
Source: Hugging Face (2023). Summarization, accessed June 2023

Summarization models differ in the way answers are created (*extractive* or *generative*).

```
Summarization with HuggingFace
Transformers.ipynb

from transformers import pipeline

model = pipeline (model =
"sshleifer/distilbart-cnn-12-6" )

text = """The tower is 324 metres (1,063 ft)
tall, about the same height as an 81-storey
building, and the tallest structure in Paris.
Its base is square, measuring 125 metres (410
ft) on each side. During its construction, the
Eiffel Tower surpassed the Washington Monument
to become the tallest man-made structure in
the world, a title it held for 41 years until
the Chrysler Building in New York City was
finished in 1930. It was the first structure
to reach a height of 300 metres. Due to the
addition of a broadcasting aerial at the top
of the tower in 1957, it is now taller than
the Chrysler Building by 5.2 metres (17 ft).
Excluding transmitters, the Eiffel Tower is
the second tallest free-standing structure in
France after the Millau Viaduct."""

output = model (text)
summary = output [0]["summary_text" ]

print (summary)
```

NLP TASKS | MACHINE TRANSLATION

This is the task of converting a piece of text from one language to another.



Source: Hugging Face (2023). Translation, accessed June 2023

Tools like Google Translate are common applications of this task.
Translation models can be used to build conversational agents across different languages.

Translation with HuggingFace Transformers.ipynb

```
from transformers import pipeline

model_checkpoint = "t5-base"
model =
pipeline("translation_en_to_fr",
model=model_checkpoint)

text = "My name is Imar and I live
in Zürich."

output = model(text)
translation =
output[0]["translation_text"]

print(translation)

>>> Je m'appelle Imar et j'habite à
Zürich.
```

01 › 02 › **03**

Motivation

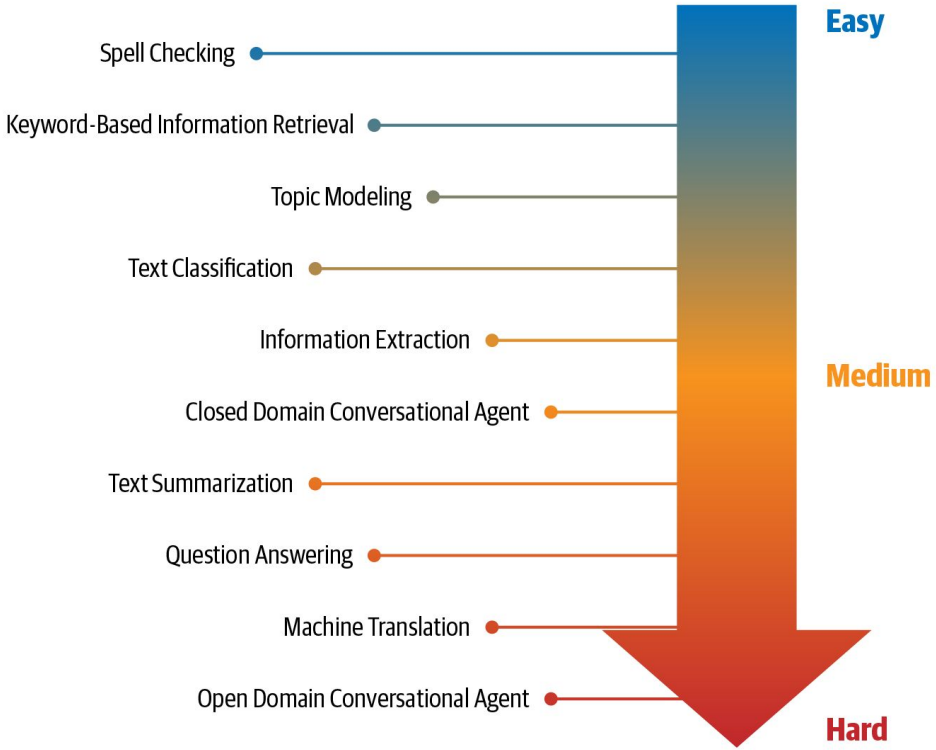
NLP Theory

Application

WHERE TO START?

NLP TASKS ORGANIZED ACCORDING TO THEIR RELATIVE DIFFICULTY

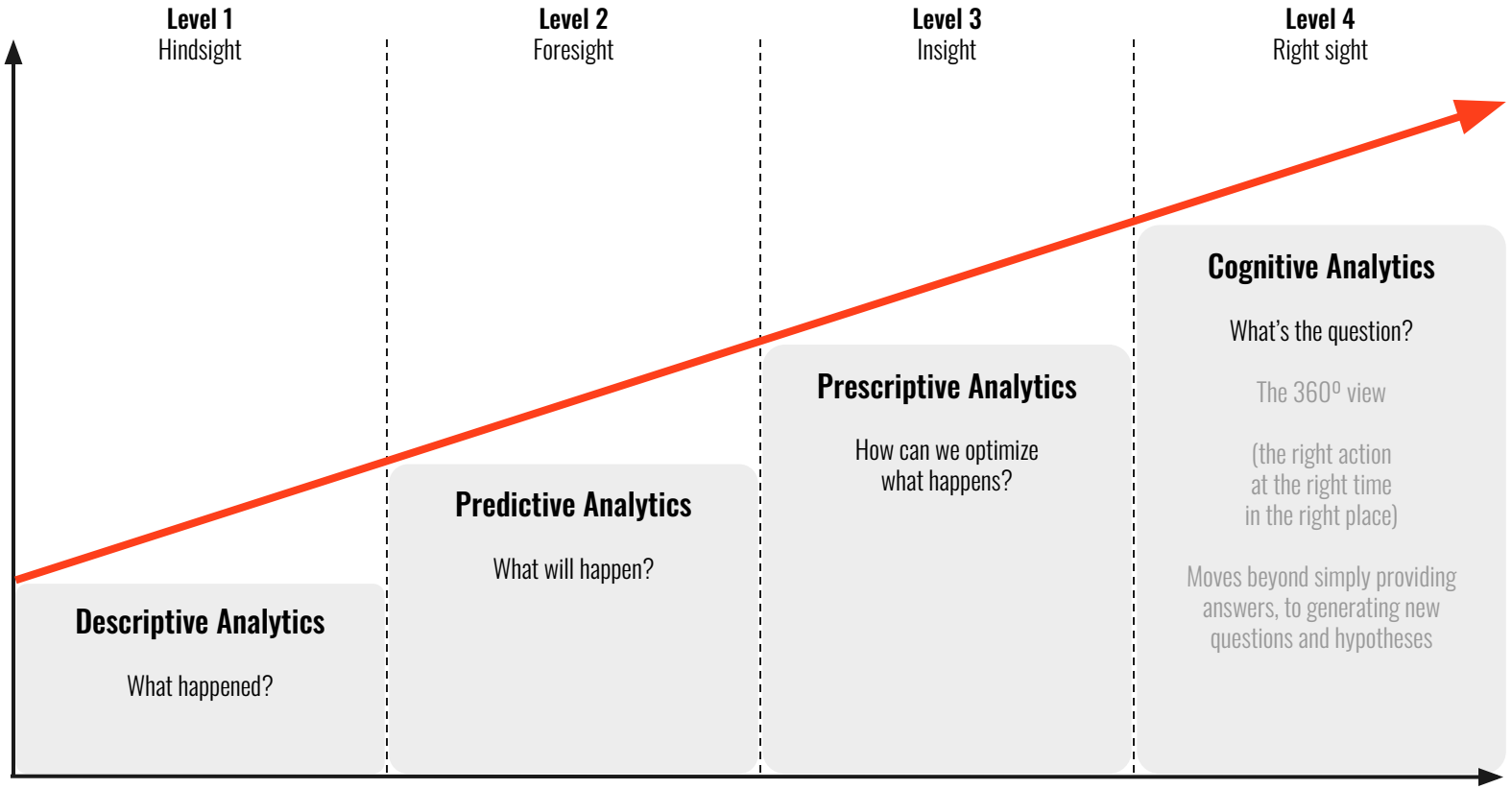
Relative Difficulty in Terms of Developing Comprehensive Solutions



Source: Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. 2020. Practical Natural Language Processing. O'Reilly Media, Inc.

BIG DATA **ANALYTICS MATURITY**

Kirk Borne - Chief Science Officer at DataPrime, Inc.



Source: http://kirkborne.net/ & https://twitter.com/KirkDBorne/status/1062470770427879425 & https://www.superdatascience.com/podcast/the-importance-of-data-science-literacy

qubay

START WITH THE BUSINESS PROBLEM

Case 1

Analyze customers' perception of a product or service based on their reviews on e-commerce websites

Example of a review that praises some aspects and criticizes few

Reviewed 21 February 2013

Good food service lacking

The cafe in the Jordaan is lovely in a cold Winter day or outside in the summer. Food and drinks great and a very good menu.

Service though is abysmal. There is a fine line between over eager and good service but here I had to get up and ask for everything I needed. Shame but still worth the visit.

Source: Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. 2020. Practical Natural Language Processing. O'Reilly Media, Inc.

Example of possible NLP solution

- E-commerce website scraping
- Text extraction and cleaning
- Topic modeling or category classification
- Sentiment analysis
- Score regression / XAI

Other Similar Applications

- Identify the social media post that brands (customer support team) must respond
- Screen customer complaints and analyze criticality for queue prioritization
- Authorship attribution, or identifying the unknown authors of texts from a pool of author
- News sentiment analysis for financial market forecast
- Triage of posts in an online support forum for mental health services
- Segregate fake news from real news

START WITH THE BUSINESS PROBLEM

Case 2

Speed up the time to insight of business people in document analysis

Document View

What is the document about?
- Categories or Topics

Who is the document about?
- Named Entities

What are the key information?
- Keywords and domain-specific words

What is the emotion tone?
- Sentiment of the document
- Sentiment per paragraph-topic pair

What are the main takeaway?
- Summary and translation (if in another language)

Related Document View

Are there documents of the same category/topic?
- Retrieve documents by the category/topic

Are there documents of the same entities?
- Retrieve documents by the entity

Are there similar documents?
- Retrieve documents by semantic similarity

Is it possible to talk to the database?
- Question answering over documents

Corpus View

What are the statistics of the corpus?
- Total number of documents, categories, ...

What are the most important keywords, entities, topics, ...?
- Word cloud (frequently) or importance (YAKE, ...)

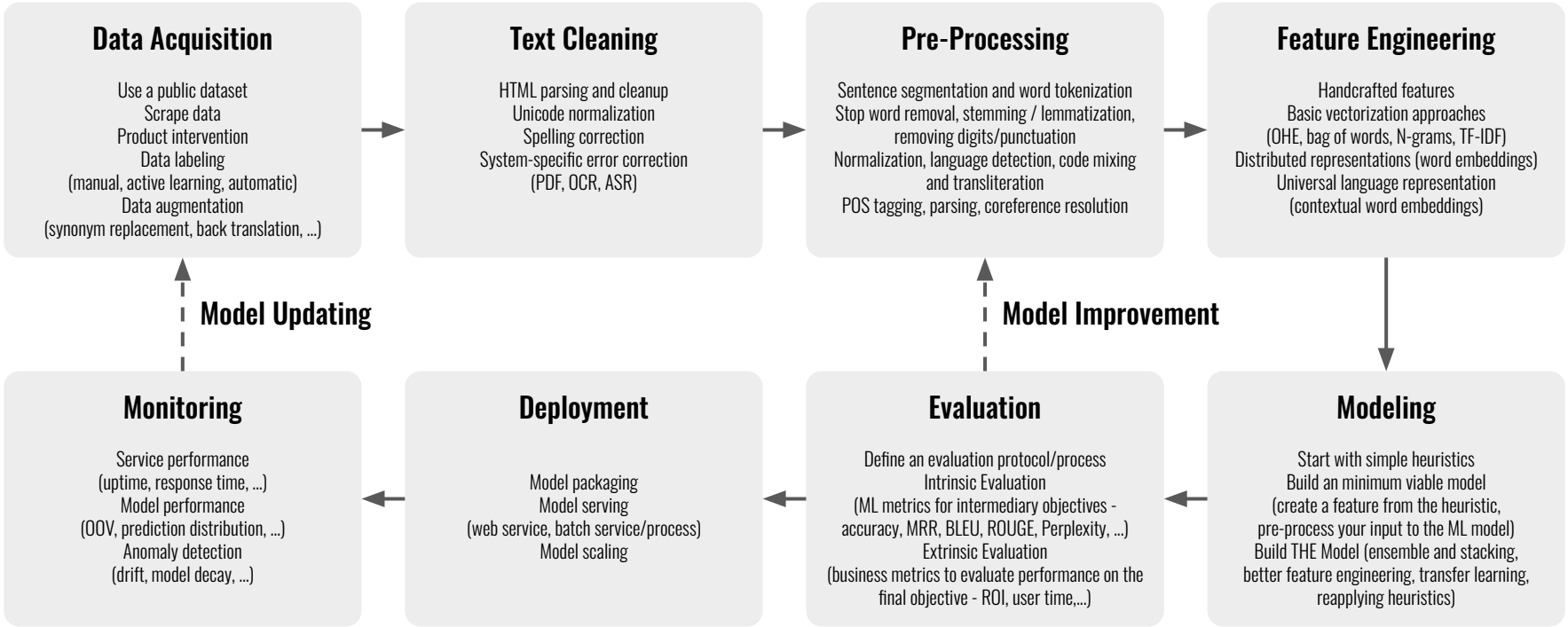
What are the associations between keywords, entities, ...?
- Co-occurrence between keywords
- Lift between entities and topics
- Total amount per topic
- Average sentiment per region

What are the time series trends of keywords, entities, ...?
- Sentiment evolution over time
- Topic trend classification (growth, fade-out)

Possible Solutions

- **Self-Service User Interface** (for web services - API)
- **Interactive Dashboard** (for batch services and also web services)

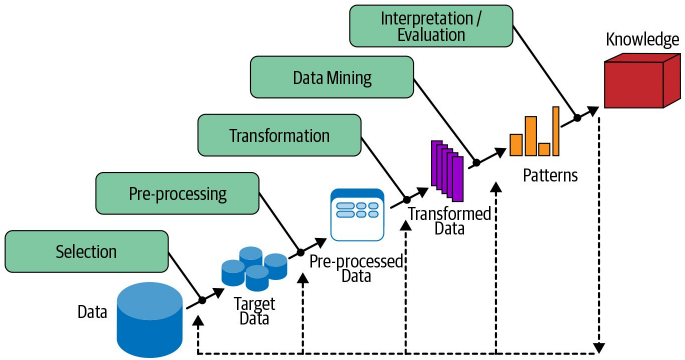
GENERIC NLP PIPELINE



THE DATA SCIENCE PROCESS

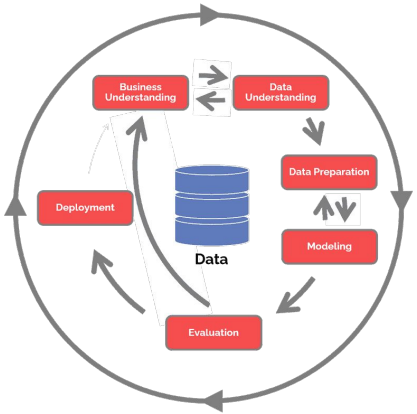
Data science is a broad term describing the algorithms and processes used to extract meaningful information and actionable insights from all forms of data. Thus, all NLP work in the industry can be categorized under the data science umbrella.

KDD
(1996)



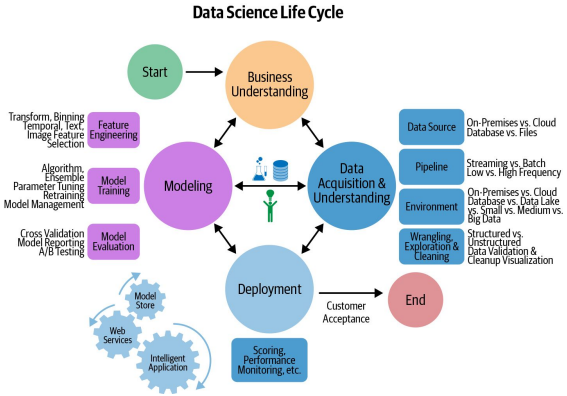
Source: Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. 2020. Practical Natural Language Processing. O'Reilly Media, Inc.

CRISP-DM
(1999)



Source: Data Science Process Alliance - What is CRISP-DM?

TDSP
(2017)



Source: Microsoft Learn - What is the TDSP?

MAKING ARTIFICIAL INTELLIGENCE SUCCEED

Planning

Team

1. Scientists who build models
   a. (should have) Scientists who have worked in industry after graduate school: Solve an AI problem is very different from academia, where data is clean is techniques are SOTA
2. Engineers who operationalize and maintain models
   a. (should have) Engineers who understand scale and data pipelines
3. Leaders who manage AI teams and strategize
   a. (should have) Leaders who have also been individual contributor scientists in the past: AI is fundamentally different from software engineering, from defining the problem statement to planning project timelines

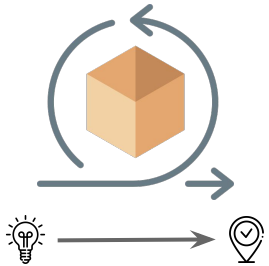
Right Problem and Right Expectations

1. Work with the stakeholders first to clearly define the task
   a. A great way to do this is to take a set of diverse example inputs that include edge cases and ask the stakeholders to write down the desired output
2. Set up right expectations about AI:
   a. AI system (as of today) will give wrong output for a subset of inputs.
   b. It's better to treat AI as augmented intelligence to support human efforts rather than artificial intelligence to replace human efforts.
   c. Beyond a point, model performance stagnates and doesn't continue rising with time.

Data and Timing

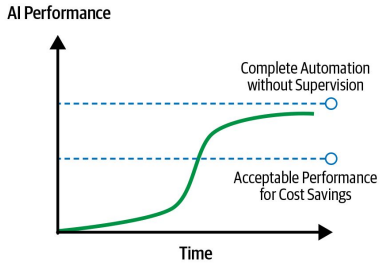
1. Quality and quantity of data
   a. What does high quality mean? Data that is structured, homogenous, cleaned, and free of noise and outliers.
   b. How much data is enough? This is a hard question to answer, but there are some rules of thumb. For instance, for sentence classification, using baseline algorithms such as Naive Bayes or random forest, at least, 2.000 data points per class.
2. Data labeling
   a. This is often a time-consuming and expensive process, specially to get bulk labeled data as in the beginning
   b. After stabilized into production, getting the production data annotated is a continuous process from there on.

Execution



Process

1. Follow a standard product development processes for Data Science
   - AI projects and different from software projects
2. Start simple, establish strong baselines
   - A SOTA technique might only give us marginal improvement over a rule-based system.
3. Make it work, make it better
   - Building a model is often only 5–10% of most AI projects (complete one full cycle asap)
4. Keep shorter turnaround cycles
   - Build models quickly (experimentation) and present the results to stakeholders frequently.



Evaluation

1. Set up the right metrics
   a. AI/ML metrics (accuracy, precision, recall, etc)
   b. Business metrics (according to the business problem that is being solved)
2. Define levels for complete automation and acceptable performance.
   a. Beyond a point, model performance stagnates and doesn't continue rising with time.
   b. Analyse if the current performance justifies the investment in compute-intensive methods that spend huge amounts on GPU and cloud services

THANK YOU