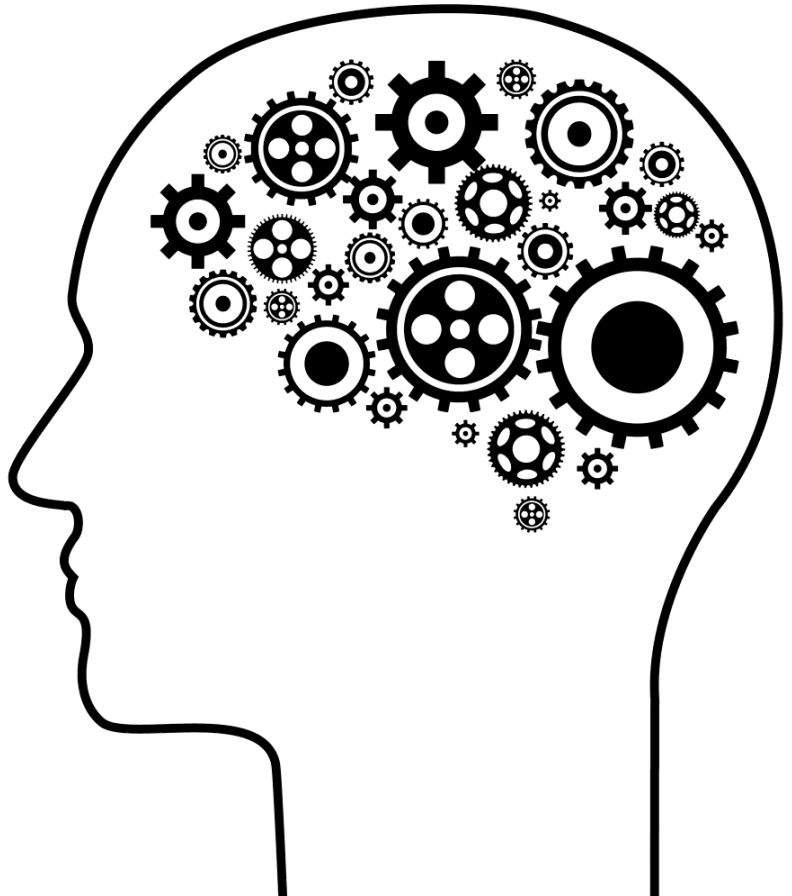
TATA CONSULTANCY SERVICES

Building on belief

Machine Learning and Artificial Intelligence

T3: T-Talk



Tarun Joseph
27 September 2024

Agenda

Part 1

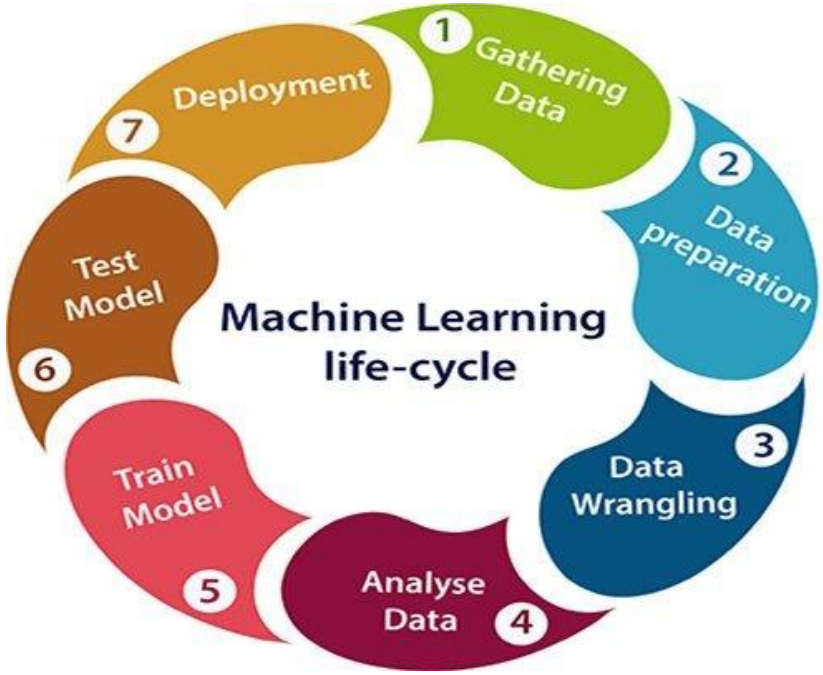
- ML Lifecycle
  - Introduction to Machine Learning
  - Supervised Learning
  - Unsupervised Learning
- Natural Language Processing
  - Machine Learning Approach
  - Deep Learning Approach

Part 2

- Practical Hands-On Walkthrough

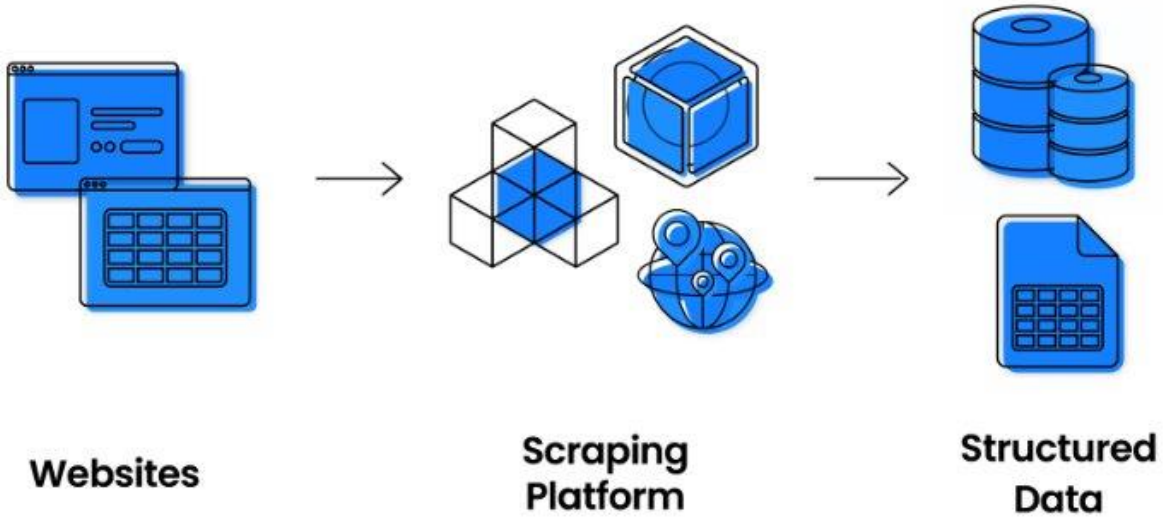
What is **ML Lifecycle**?

- **Machine Learning Lifecycle** is a cyclic process used to build an efficient ML project.
- It consists of the following 7 major steps:
    1) **Data Collection:** Web Scraping
    2) **Data Preparation:** Cleaning
    3) **Data Wrangling:** EDA
    4) **Analyze Data:** Visualization
    5) **Train the model**
    6) **Test the model**
    7) **Deployment**

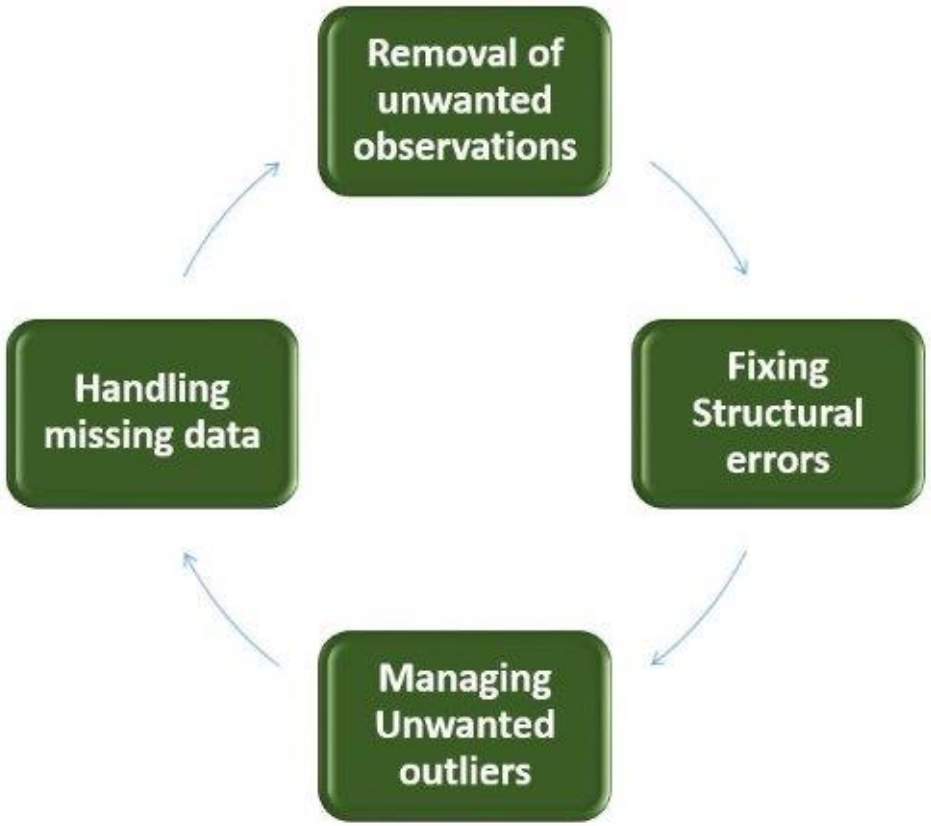


- The first underlying basic step: **Problem Understanding**/Setting Project Objectives

Data Collection: Web Scraping



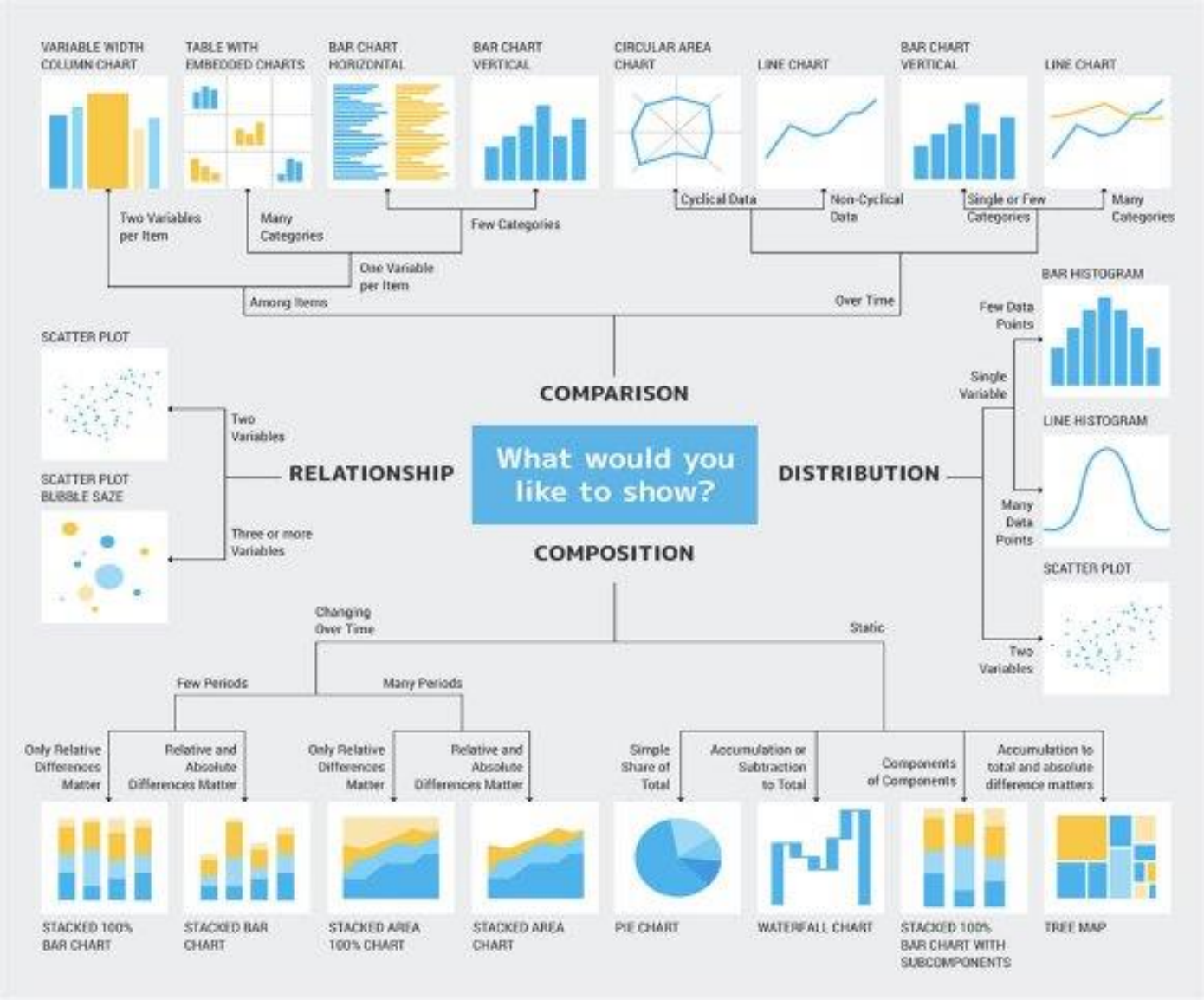
Data Preparation: Cleaning



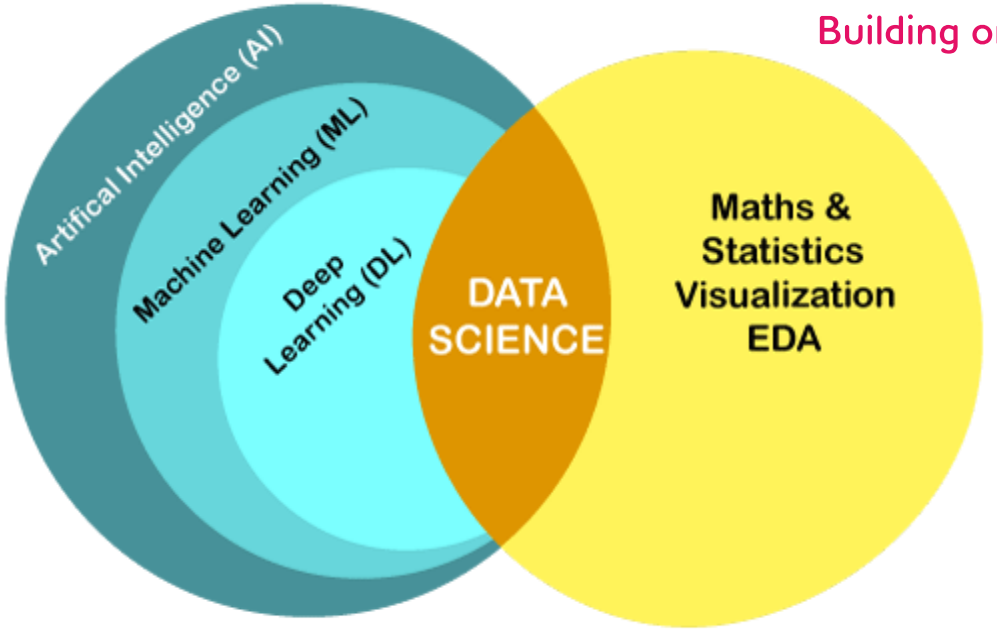
TATA CONSULTANCY SERVICES

Building on belief

Analyze Data: Visualization



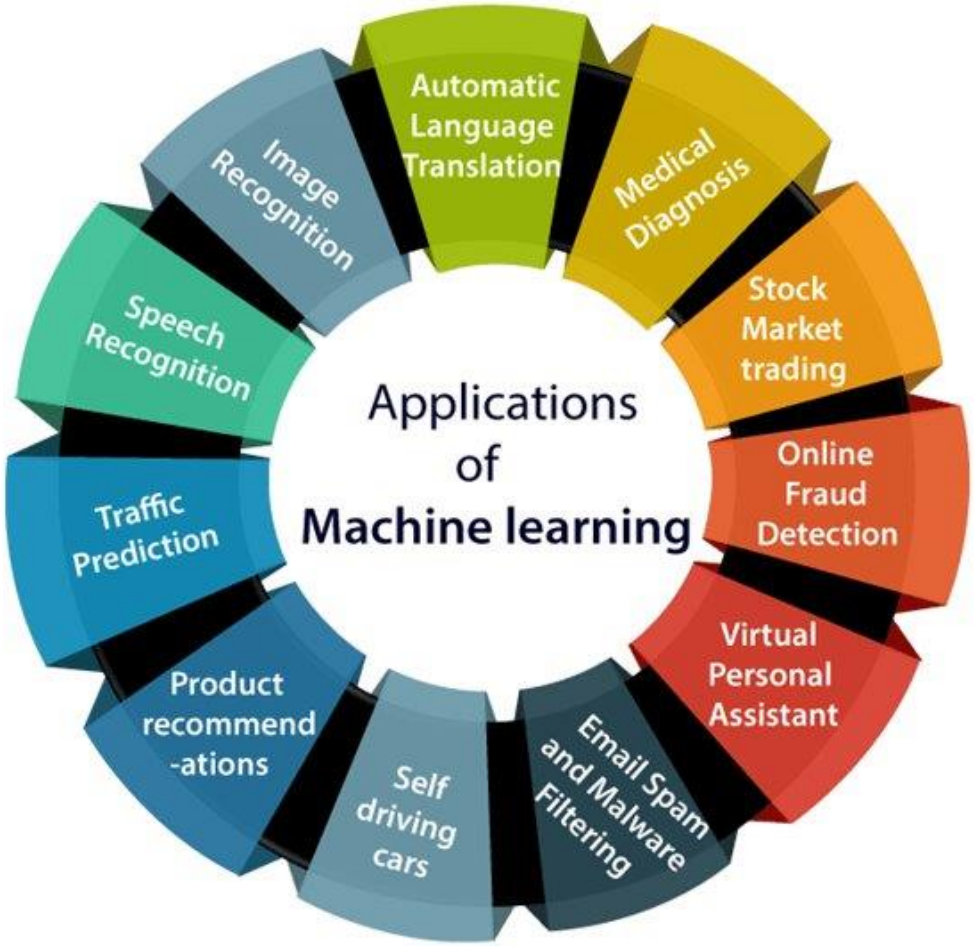
Introduction to Machine Learning



- Subset of **Artificial Intelligence** (AI)
- **Arthur Samuel**, 1959 (ML); **Geoffrey Hinton** 2006 (DL)
- **Algorithms** that enable machines to **automatically** learn from data, without being programmed.
- Algorithms build **Mathematical models** using data and help make **predictions**.
- Higher the quality of data, higher the accuracy of the model.

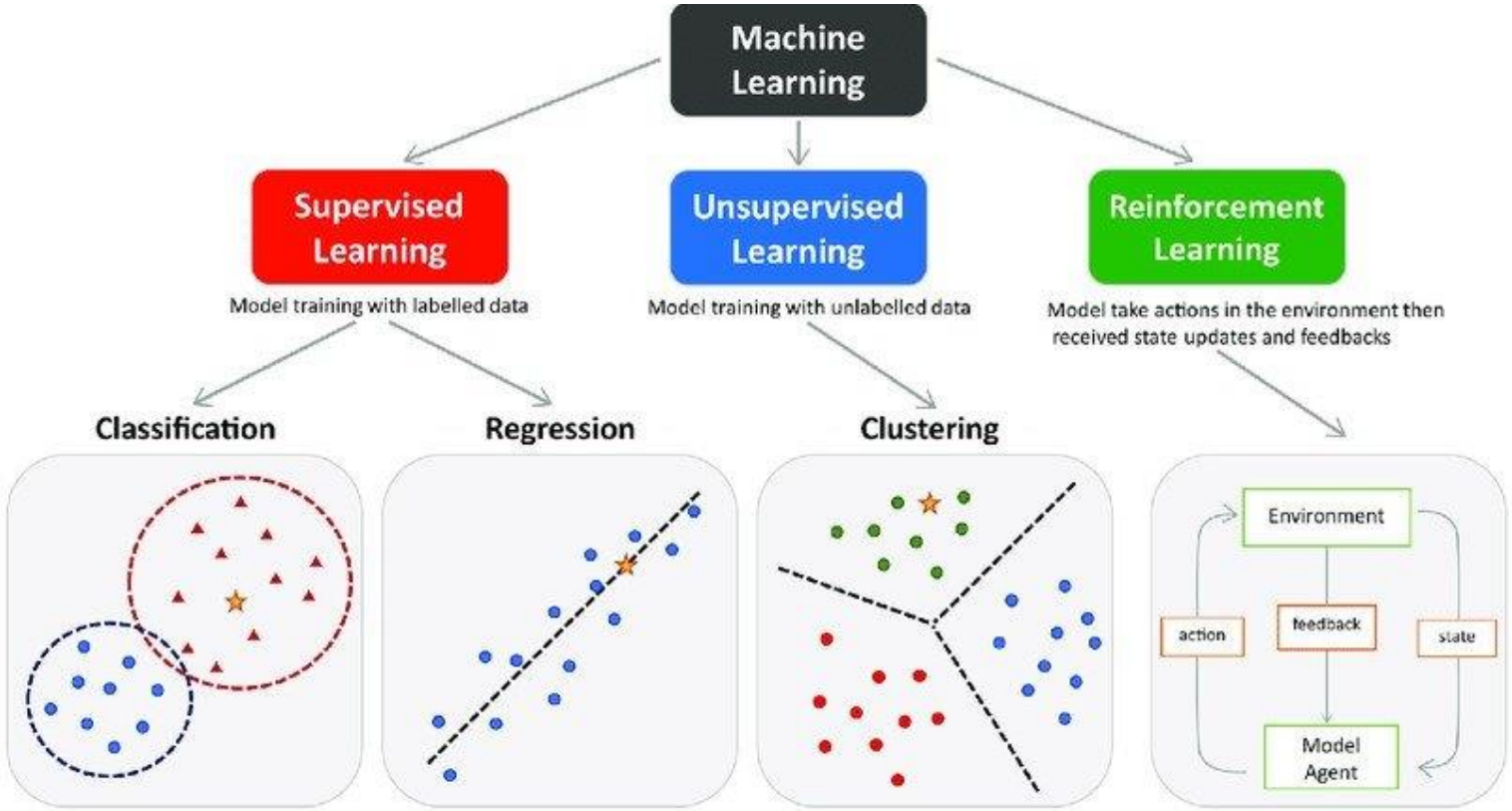
Applications of Machine Learning

- Image Recognition (Face Recognition, Robotics)
- Speech Recognition (Siri, Alexa, Google Assistant)
- Product Recommendations (Netflix, Amazon)
- Fraud Detection (Credit cards, Banks)
- Generative AI, LLM Models (Chat-GPT, LLaMA, Bard)
- Image Generation (DALL-E, Stable Diffusion, Midjourney)



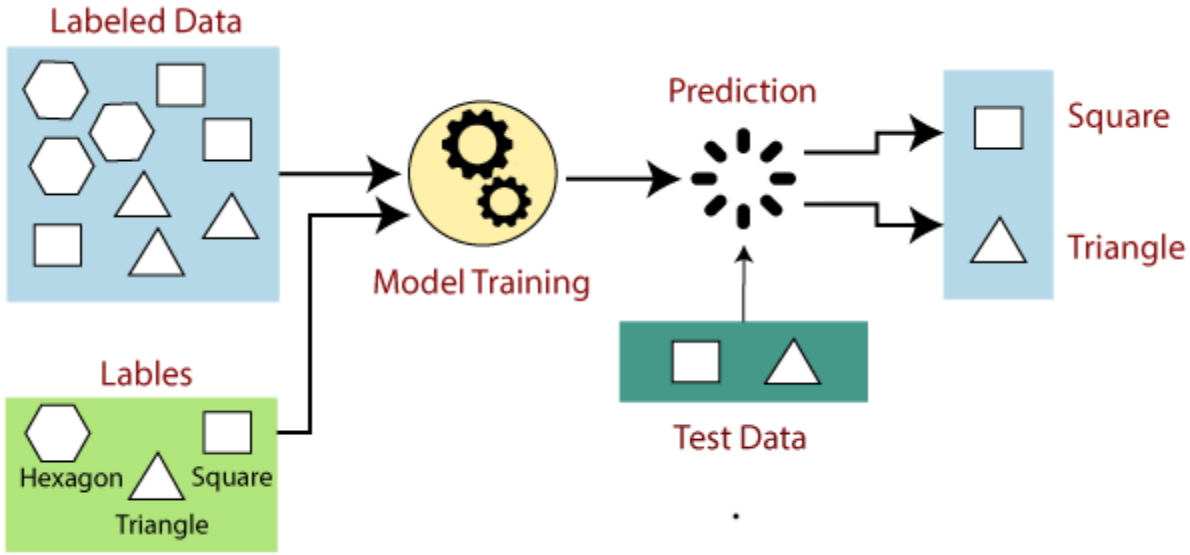
Types of Machine Learning

- Supervised Learning: Labeled data
  - Classification
  - Regression
- Unsupervised Learning: Unlabeled data
  - Clustering
- Reinforcement Learning
  - Reward-Punishment system (PPO, Q-learning)



Supervised Machine Learning

- Algorithms are trained using **labeled** input data
- Training data is already tagged with correct output
- Test data is **unlabeled** and trained model needs to **predict** the correct label.
- Labeled data **supervises** the model and teaches it to predict accurately.



- **Classification**: Spam filtering, Product categorization
- **Regression**: Price prediction, Weather Forecasting

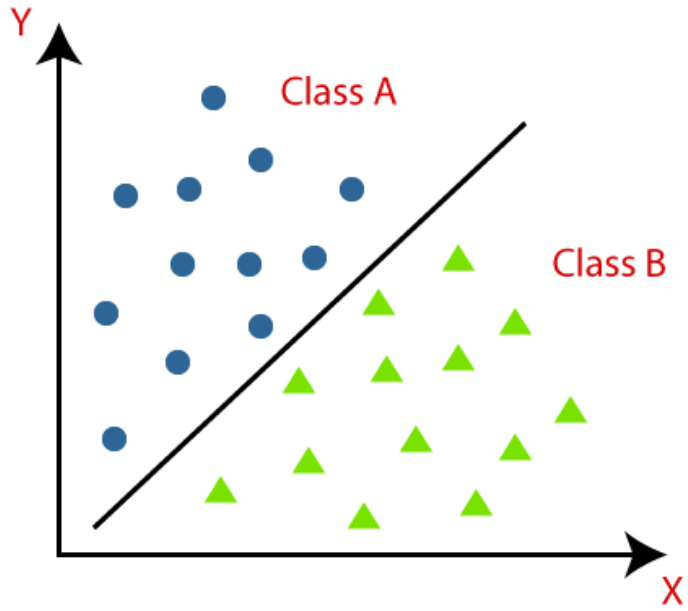
Classification Algorithms

- Used to predict **categorical** target variables using **labeled** training data as input.

- The algorithm which implements the classification on a dataset is known as a **classifier**.

- Two types of Classifications:
  1. **Binary Classification**: Two possible outcomes. E.g. Yes/No, Spam/Ham
  2. **Multi-class Classification**: More than two outcomes.
     - E.g., Product categories: Food, Clothes, Electronics etc.

- Types of Classification Algorithms:
  - **Linear Models**: Logistic Regression, SVM
  - **Non-linear Models**: KNN, Naïve Bayes, Decision Tree, Random Forest



Unsupervised Machine Learning

- Unsupervised learning in artificial intelligence uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns or data groupings without the need for human intervention.

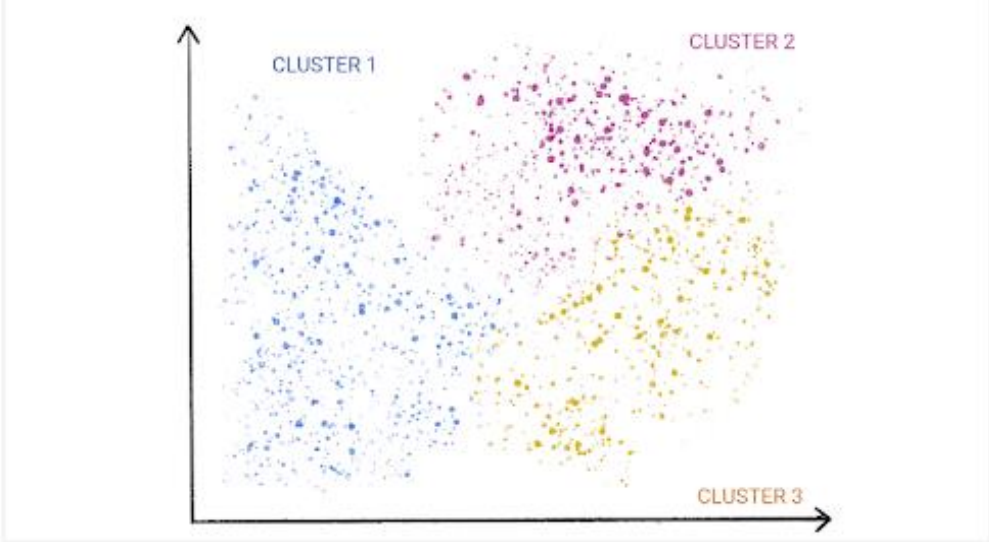


Figure 1. An ML model clustering similar data points.

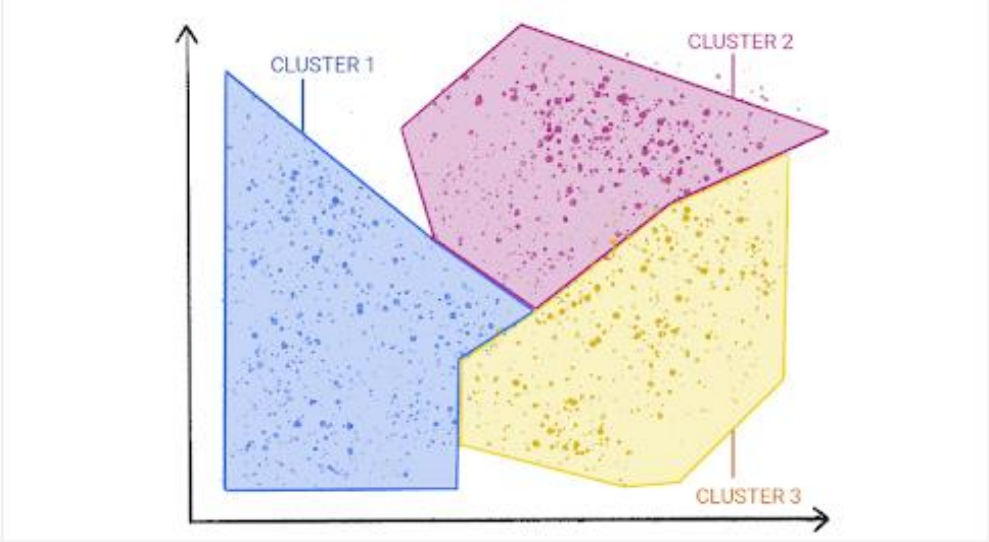
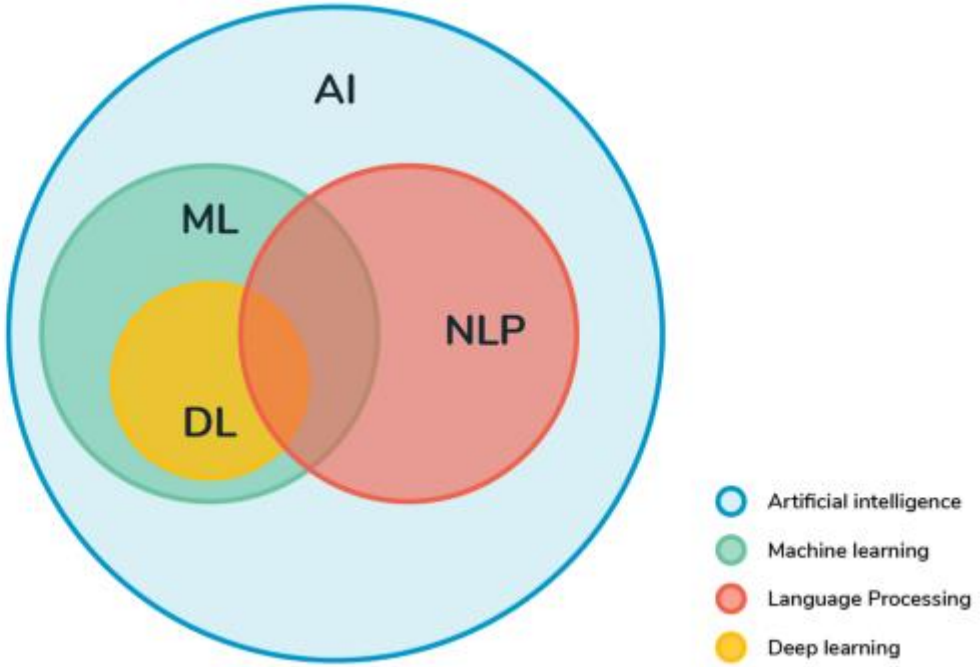


Figure 2. Groups of clusters with natural demarcations.

Natural Language Processing

- What is NLP?
- What makes dealing with NL so challenging?
- Different steps in typical NLP task: NLP Pipeline

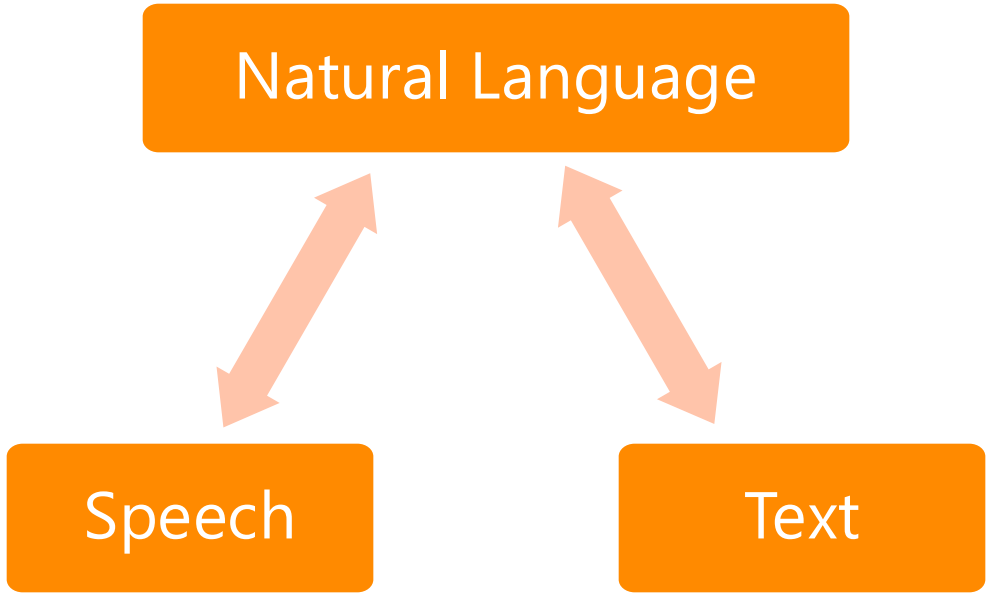
Natural Language Processing(NLP)



Natural Language Processing, or *NLP*, is the sub-field of AI that is focused on enabling computers to understand and process human languages.

https://blog.goodaudience.com/learn-natural-language-processing-from-scratch-7893314725ff

Natural Language (NL) - Refers to the way humans, communicate with each other.



- Human to human
- Audio books
- Podcasts

much more…

- Books
- Menus
- Email
- SMS
- Web Pages

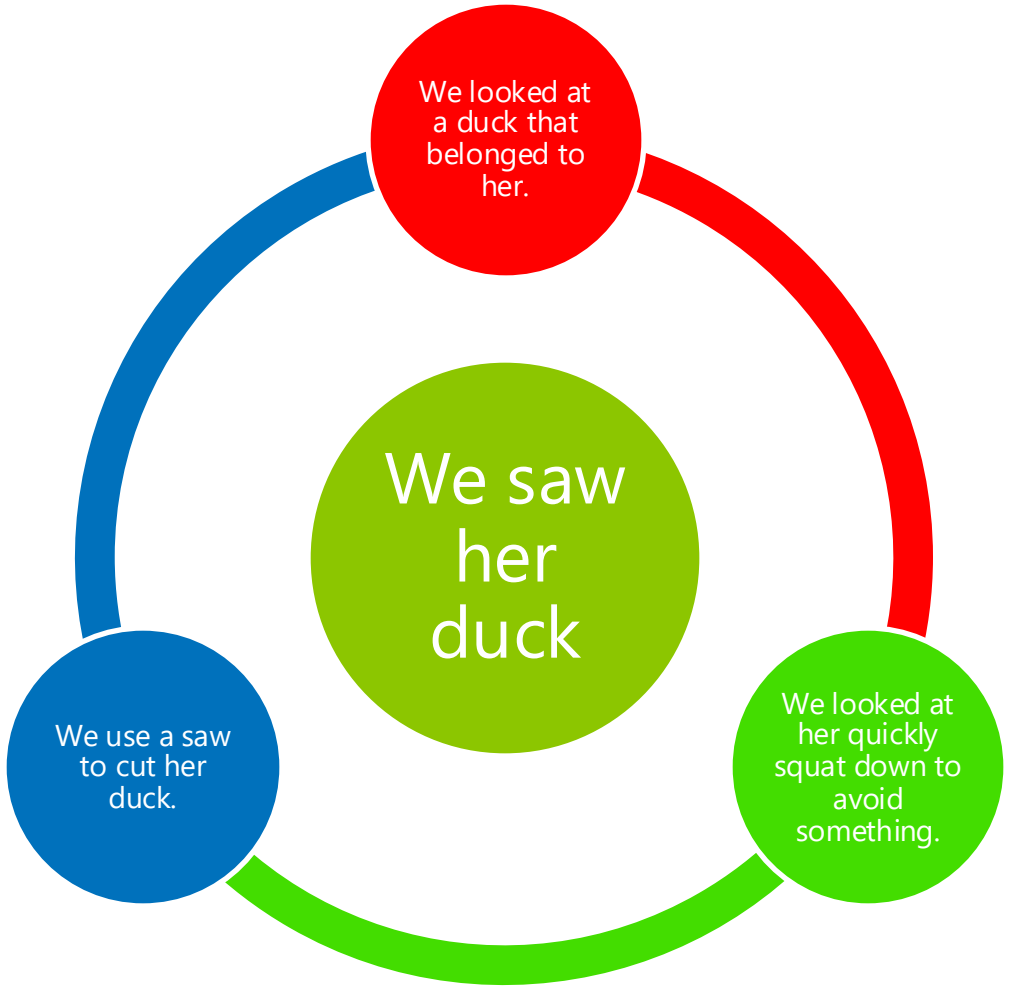
much more…

NLP Common Applications

1. Machine Translation
2. Information Extraction/Retrieval
3. Sentiment Analysis
4. Information Summarization/Analysis
5. Question Answering

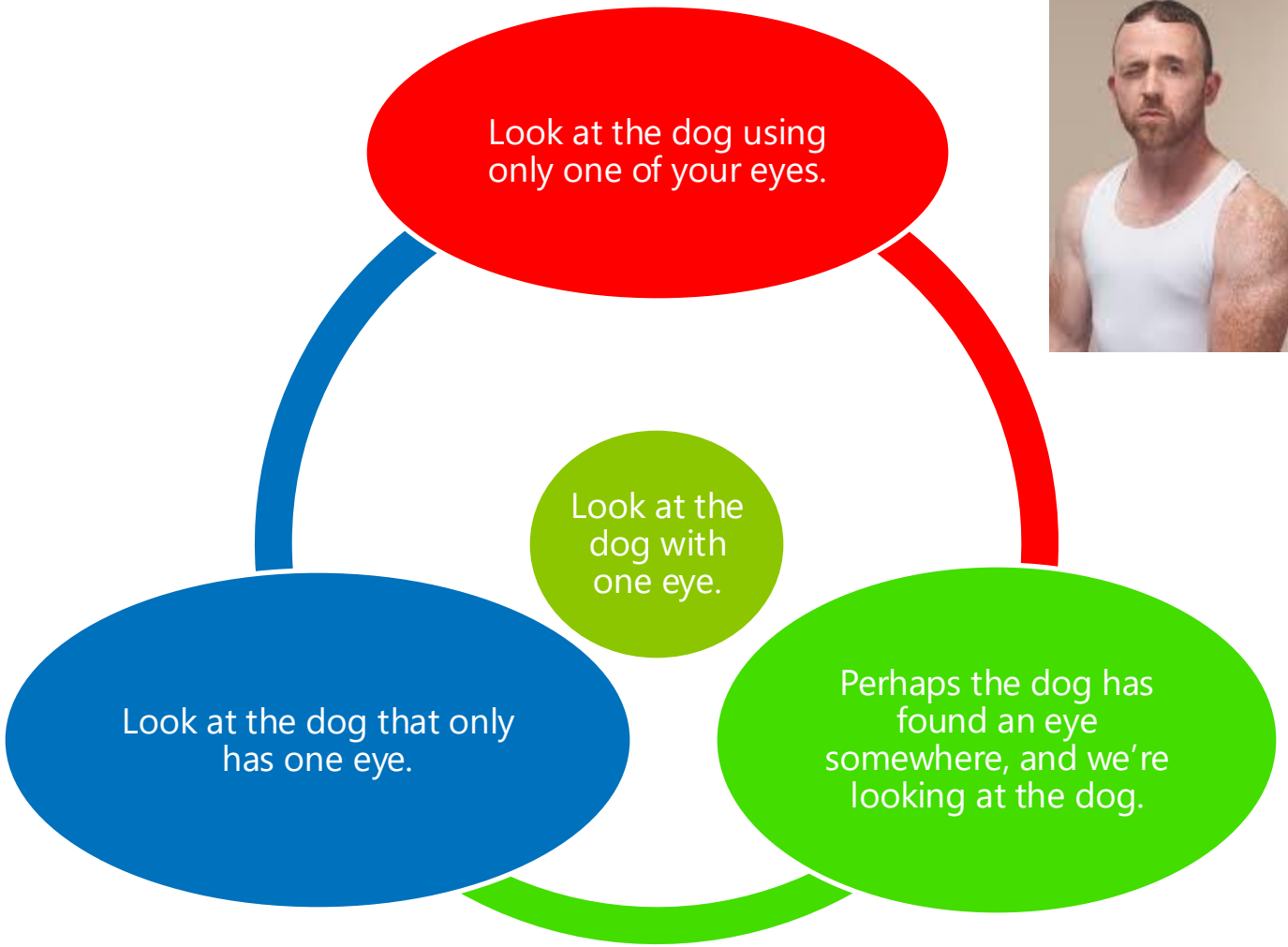
Can Computers Understand Language?

- Ambiguity: - Understanding and Modelling of elements within a **variable context**.

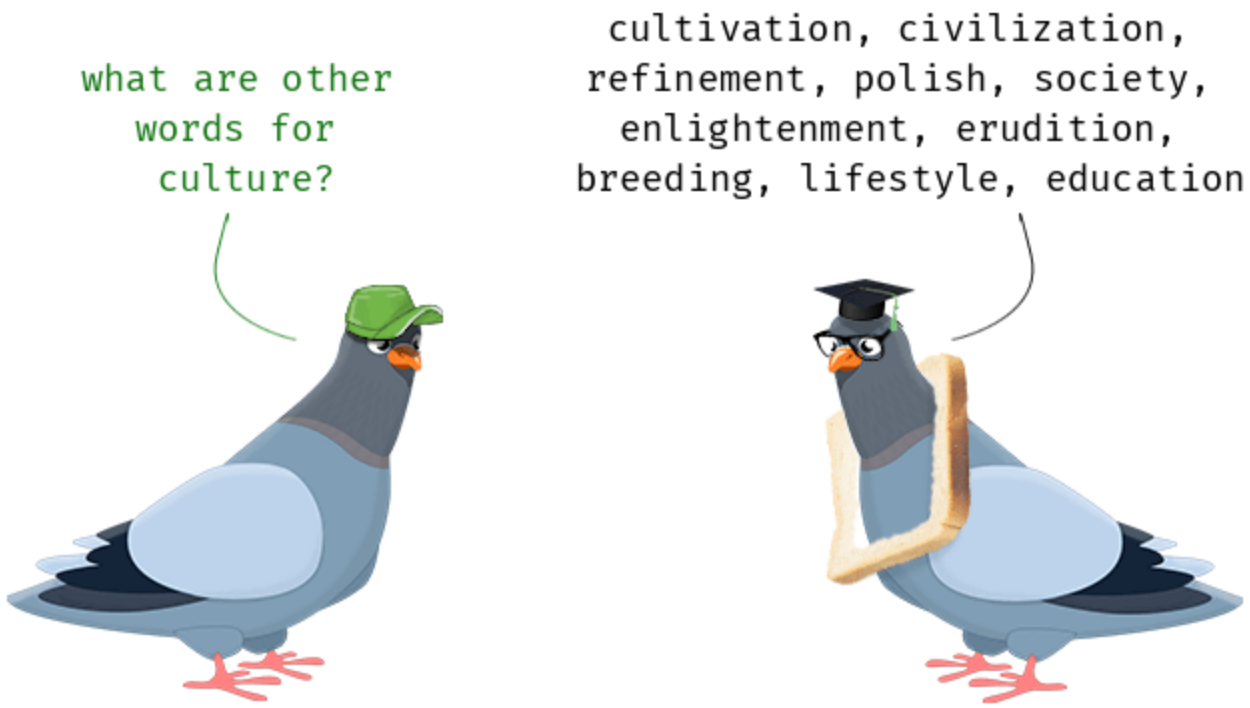


Can Computers Understand Language?

- Ambiguity: - Understanding and Modelling of elements within a **variable context**.

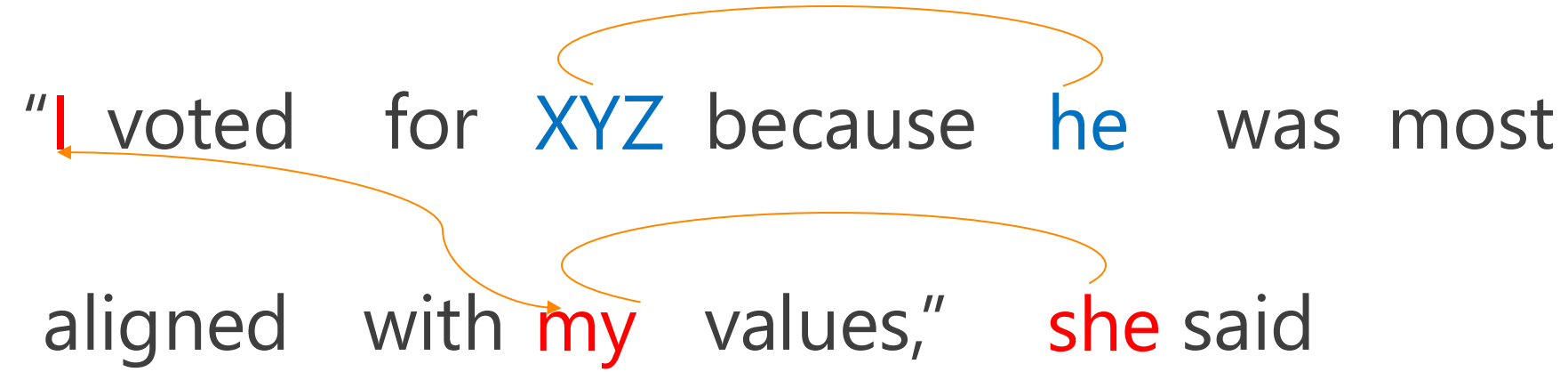


- Synonymy: Express the same idea with different terms



- **Syntax:** The structure, which takes into account several rules but also some irregularities in different cases.

- **Co-reference:**



- **Co-Reference:** Another Example

London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium. London's ancient core, the City of London, largely retains its 1.12-square-mile (2.9 km2) medieval boundaries.

Can Computers Understand Language?

Computers can’t yet truly understand English in the way that humans do

NLP is making it possible to a certain extent

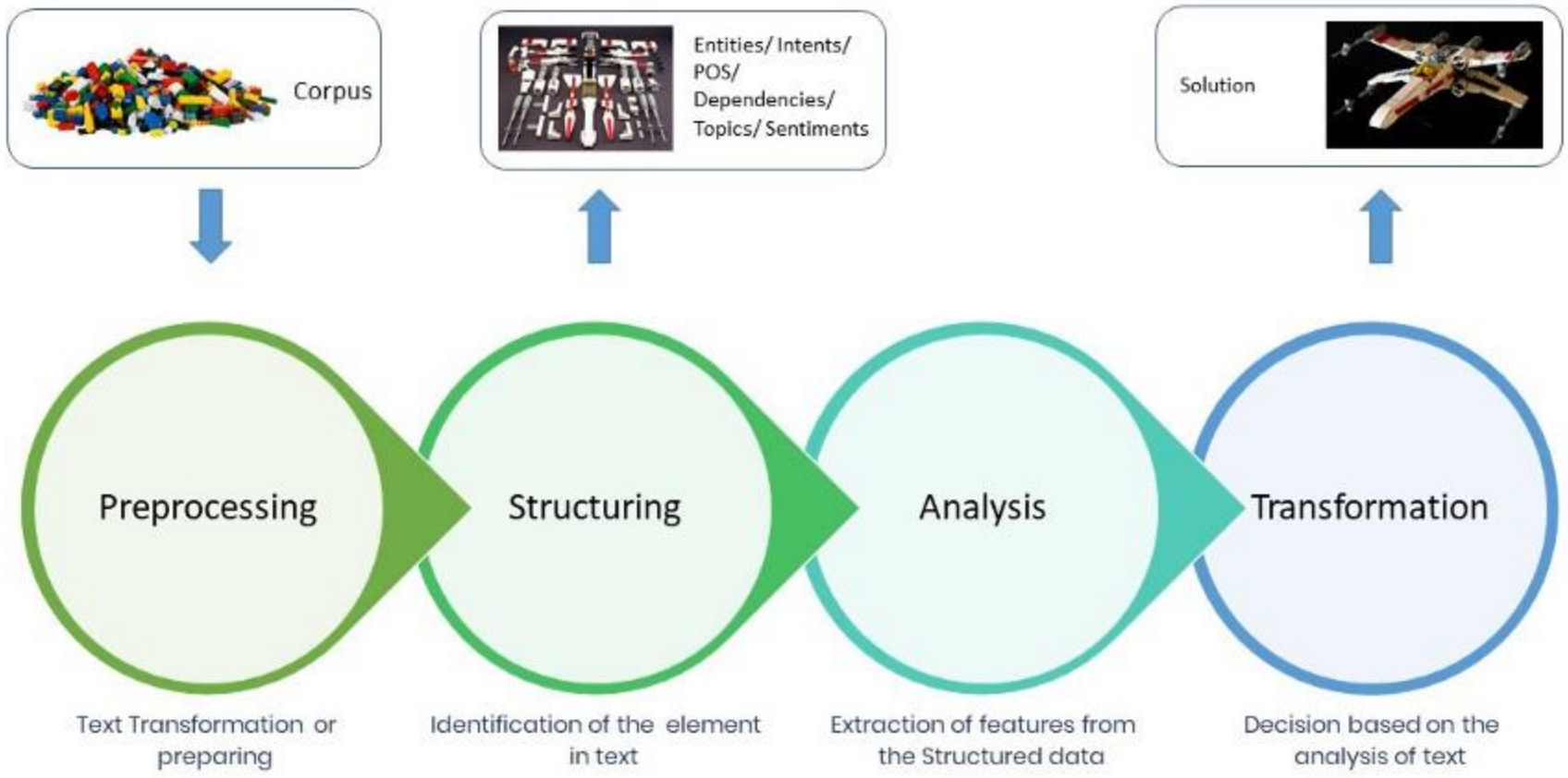
Extracting Meaning from Text is Hard

- Reading and understanding Language is very complex
- Languages doesn’t follow logical and consistent rules

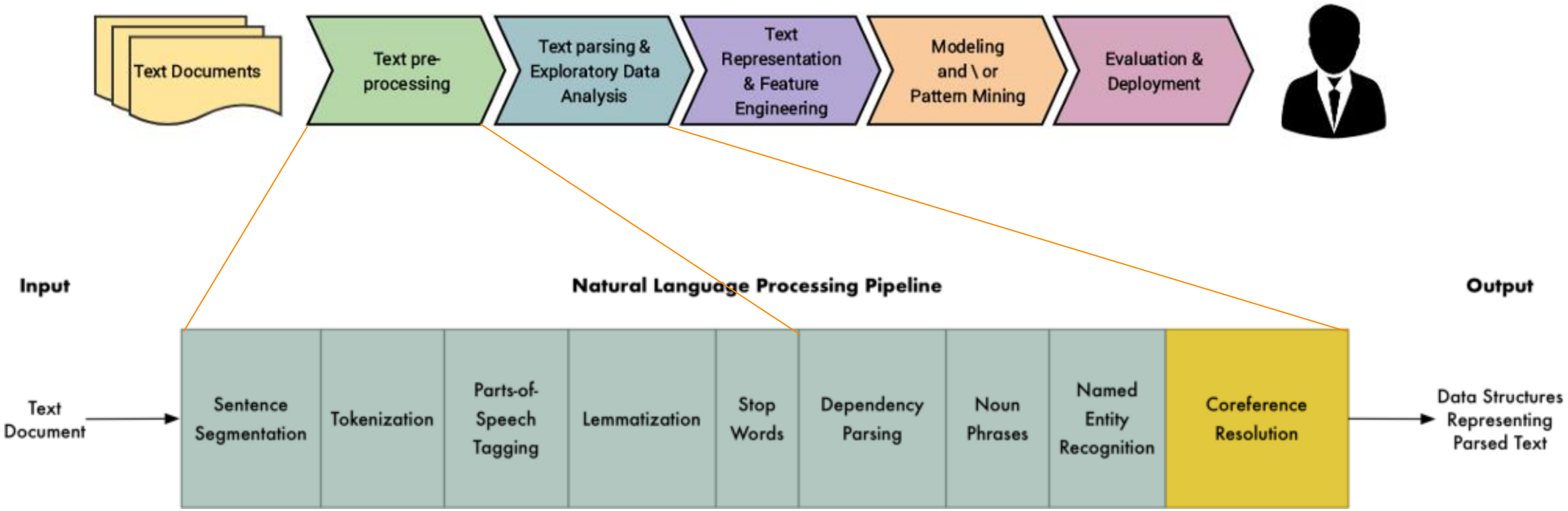
e.g. CBI grill business owner over illegal coal fires

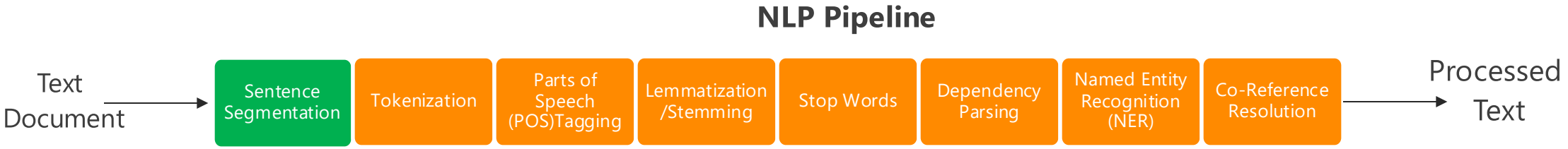
- Is CBI questioning business owner about burning coal illegally?
- Or CBI literally cooking the business owner?

Divide Complex task to manageable pieces



Pipeline - Divide Complex task to manageable pieces





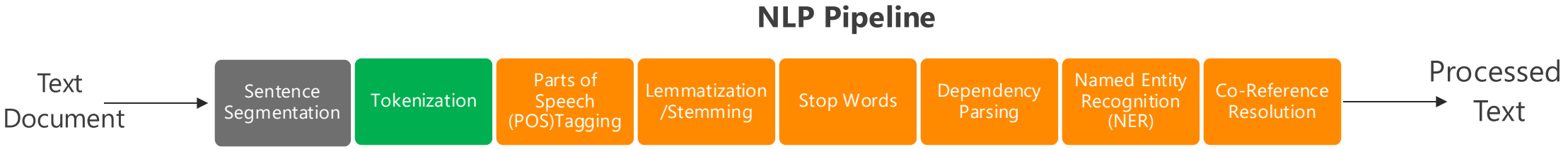
Sentence Segmentation – Breaking Text into sentences

London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.

1. "London is the capital and most populous city of England and the United Kingdom."

2. "Standing on the River Thames in the southeast of the island of Great Britain, London has been a major settlement for two millennia."

3. "It was founded by the Romans, who named it Londinium."



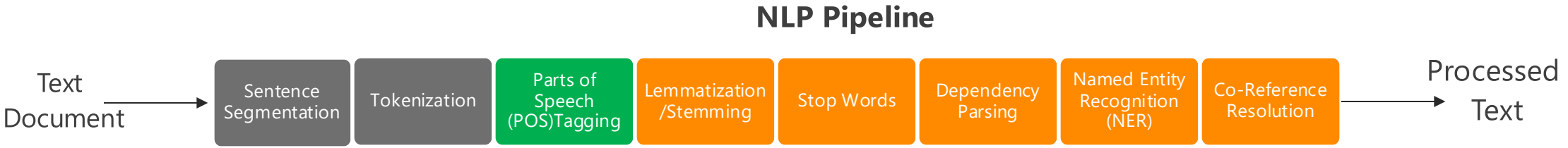
Word Tokenization – Break sentence into words

1. "London is the capital and most populous city of England and the United Kingdom."

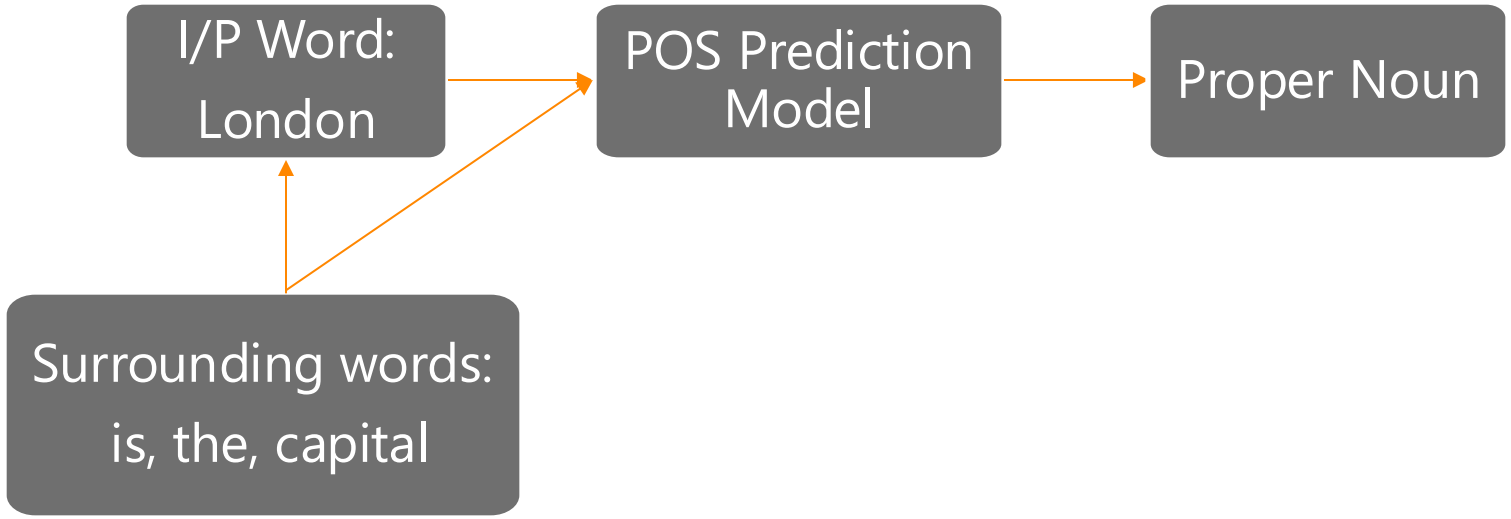
Tokenized words:

"London", "is", " the", "capital", "and", "most", "populous", "city", "of", "England", "and", "the", "United", "Kingdom", "."

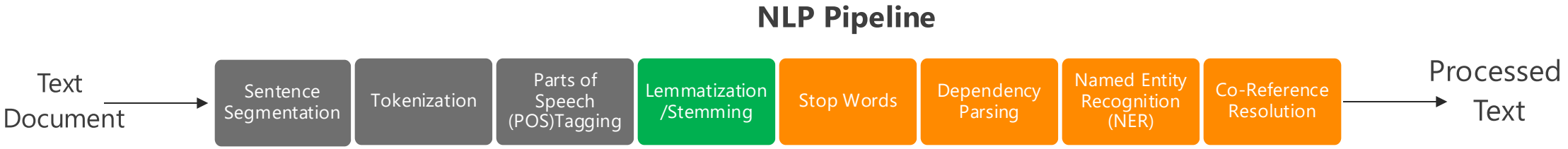
{Split at space}



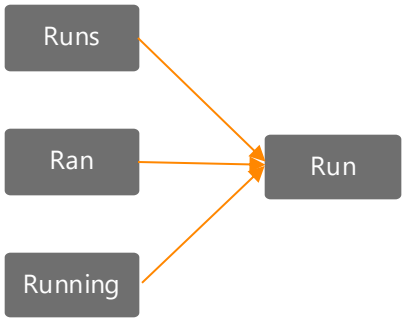
POS Tagging – Identification/Prediction of POS tags

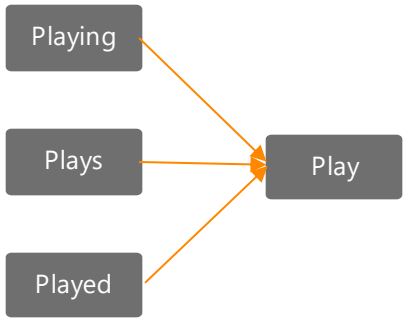


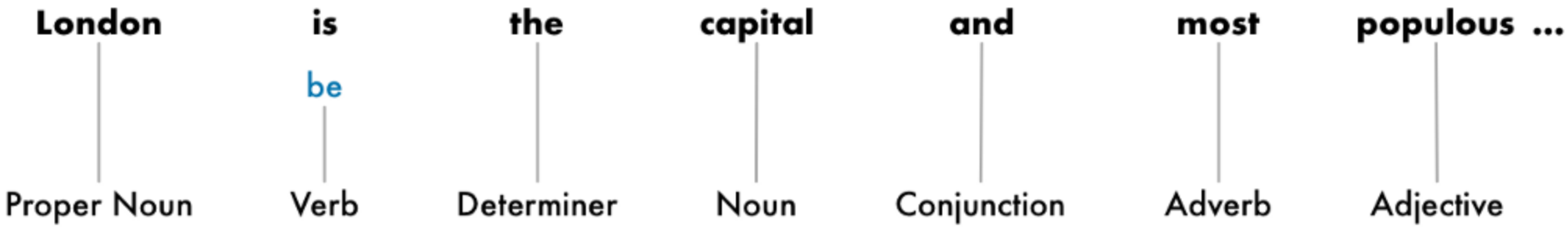
https://explosion.ai/demos/displacy-ent

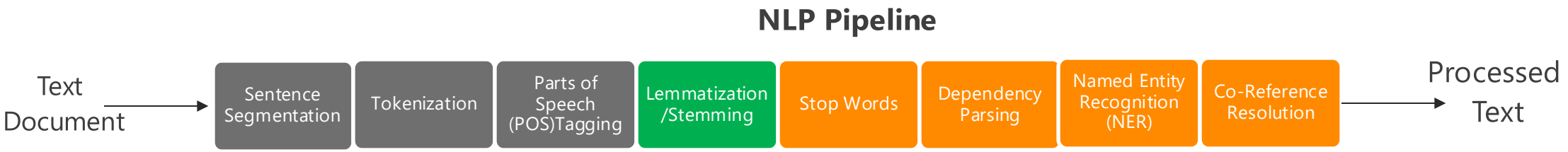


Lemmatization – Root Form or lemma of each word (Lookup Table)

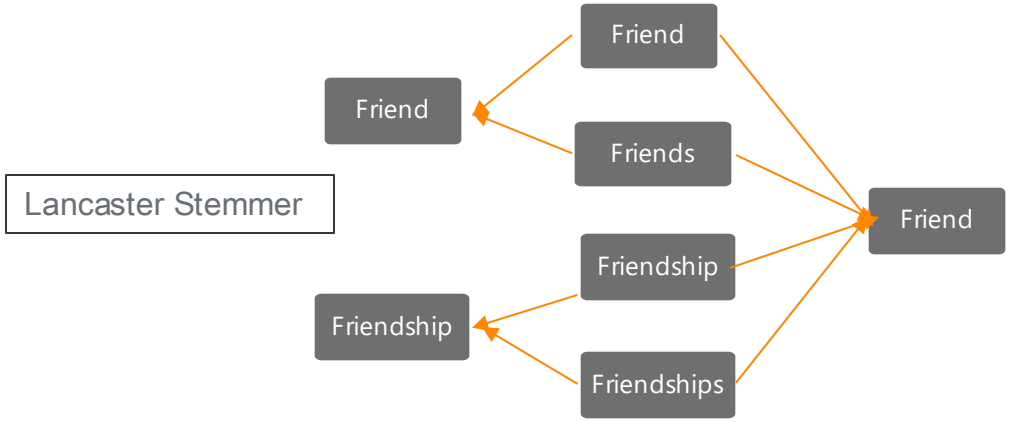


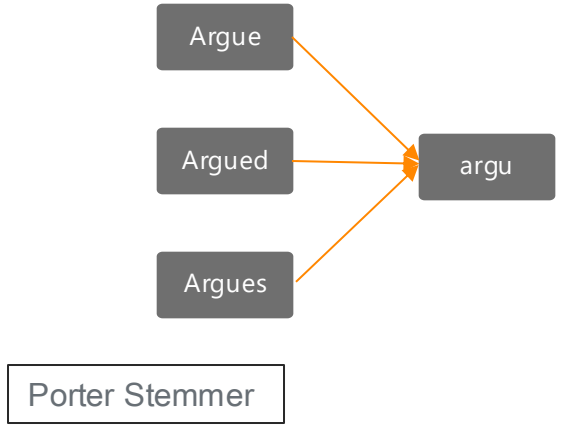


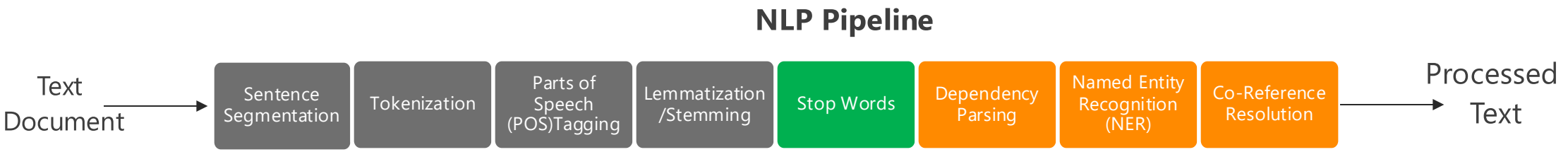




Stemming – Reduce Inflection in words by mapping to same stem



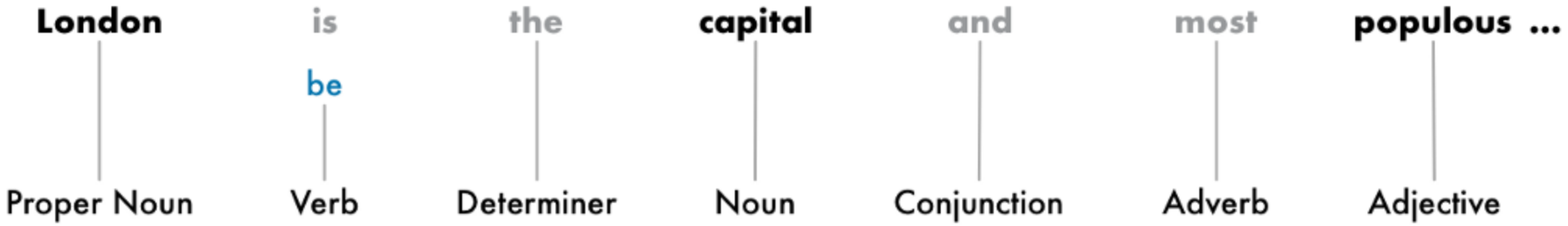




Stop Words – Words that can be filtered. (Lookup Table)

Words like - a, an, the, and , is , to ….

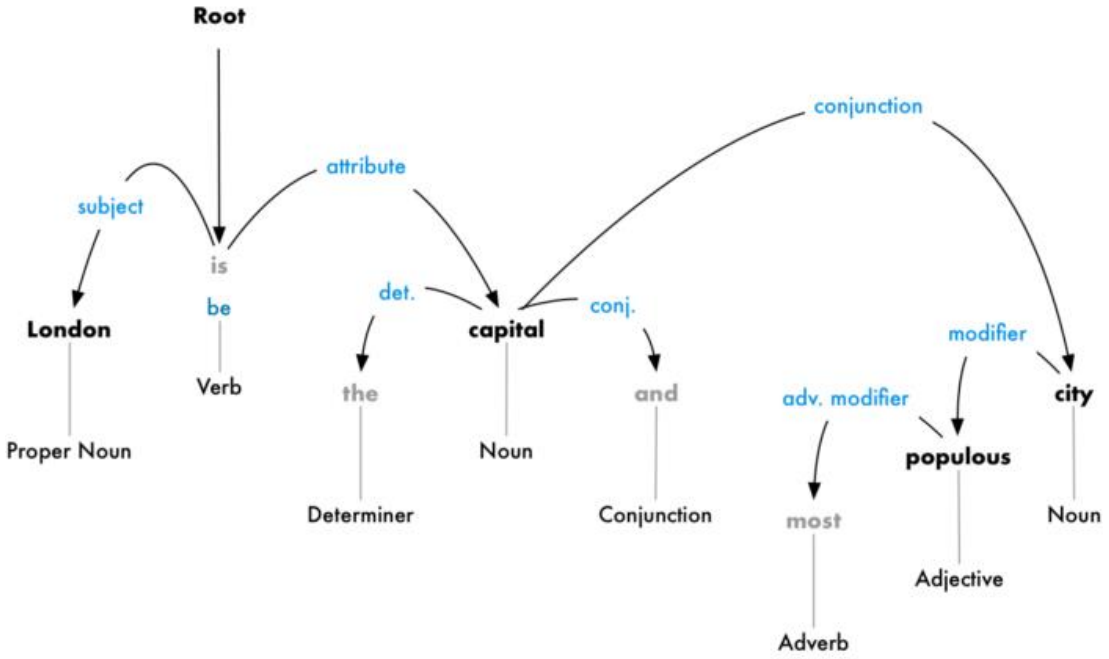
(No standard list – domain specific)



NLP Pipeline

Dependency Parsing - How words in a sentence relate

The goal is to build a tree that assigns a single **parent** word to each word in the sentence. The root of the tree will be the main verb in the sentence.



https://explosion.ai/demos/displacy?text=Its%20new&model=en_core_web_sm&cpu=1&cph=1

NLP Pipeline

Named Entity Recognition(NER) - detect nouns with the concepts that they represent

People's names
Company names
Geographic locations (Both physical and political)
Product names
Dates and times
Amounts of money
Names of events

London is the capital and most populous city of England and the United Kingdom.

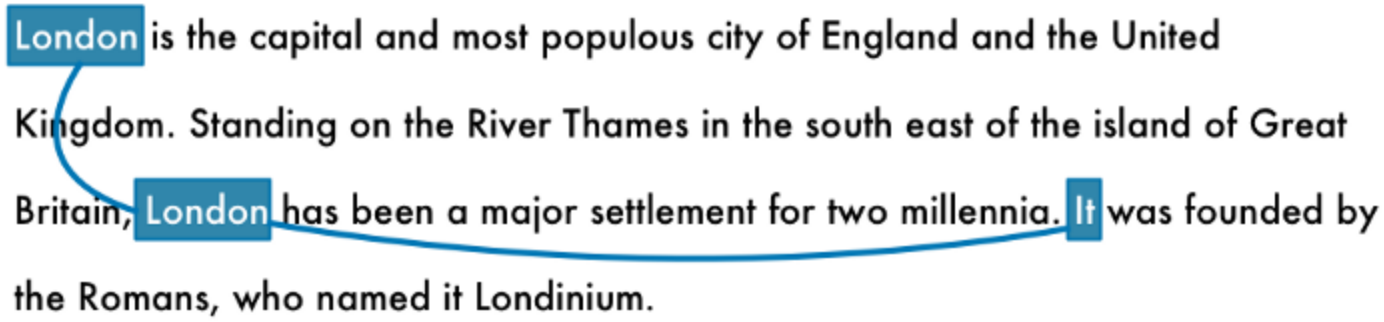
Geographic Entity *Geographic Entity* *Geographic Entity*

https://explosion.ai/demos/displacy-ent

NLP Pipeline

Co-Reference Resolution - Resolution of pronouns (shortcuts)

It was founded by the Romans, who named it Londinium



https://huggingface.co/coref/

First steps

- Computers can't directly understand text like humans can.
  - Humans automatically break down sentences into units of meaning.

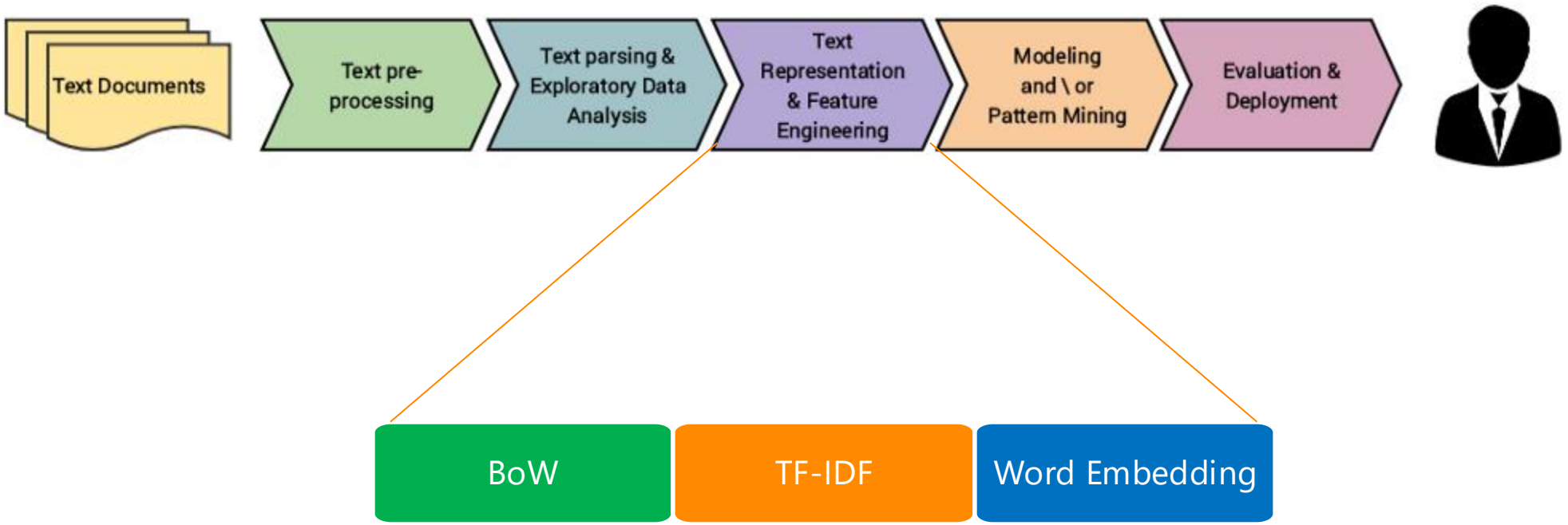
- We explicitly show the computer how to do this, in a process called text representation or features generation.

- We can convert the tokens into a matrix - (BoW, tf-idf, Word Embeddings).

- Once we have a matrix, we can use machine learning algorithm to train a model and predict scores.

https://blog.goodaudience.com/learn-natural-language-processing-from-scratch-7893314725ff

Pipeline - Divide Complex task to manageable pieces

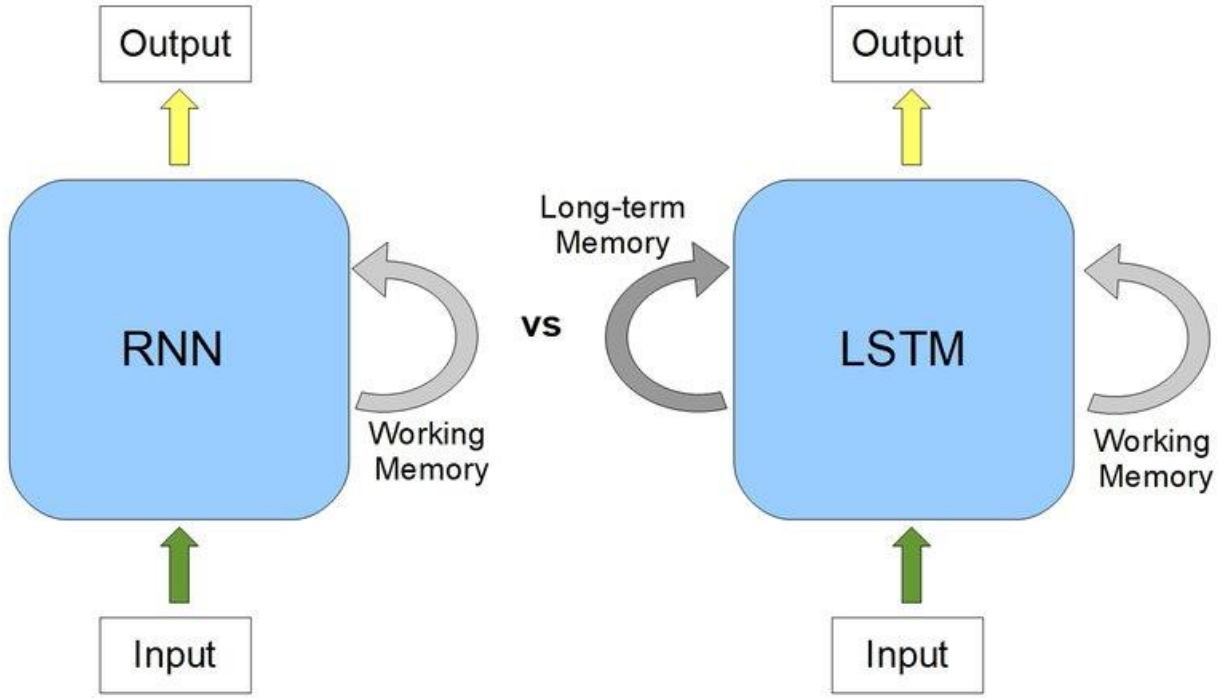


RNN (Recurrent Neural Network) :

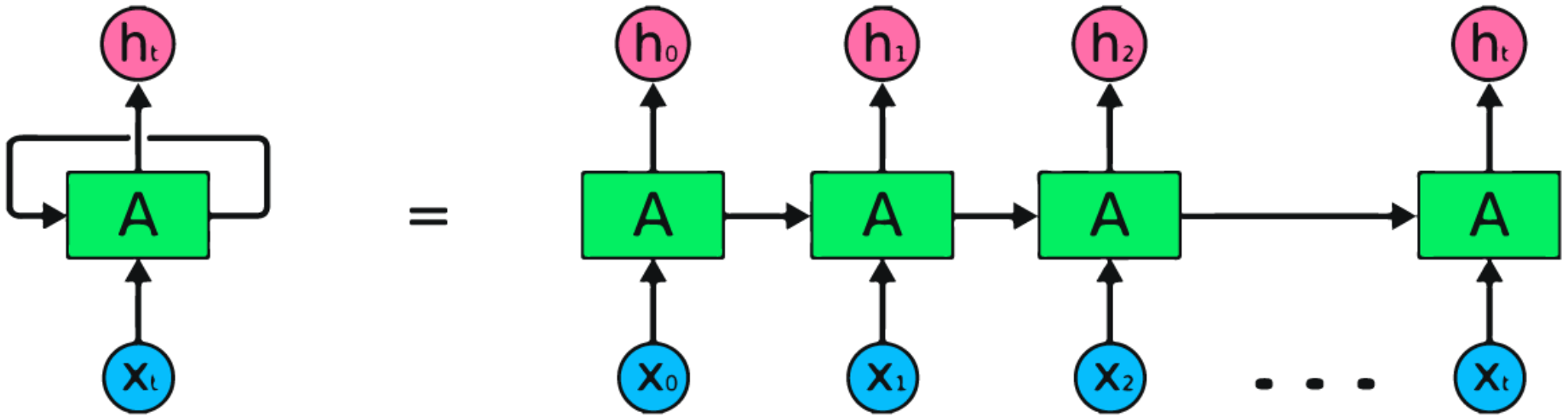
- Deep learning model.
- Sequential data input into a specific sequential data output.
- Eg: Financial data, Time series data, Text generation.
- Short-term memory.

LSTM (Long short-term Memory) :

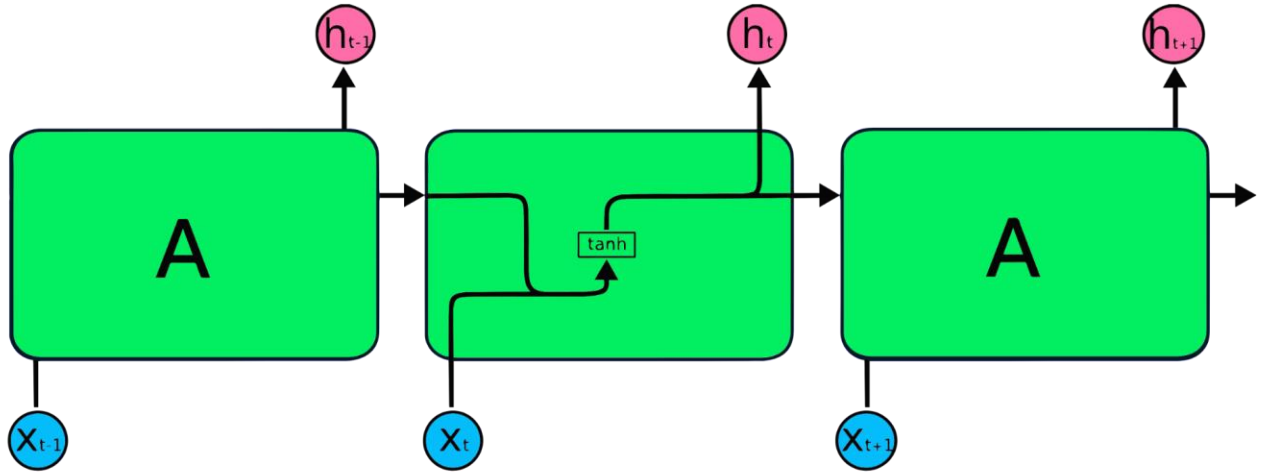
- Improved version of RNN.
- Holds an information for an extended period.
- Learning long term dependencies.
- Eg: Language translation.



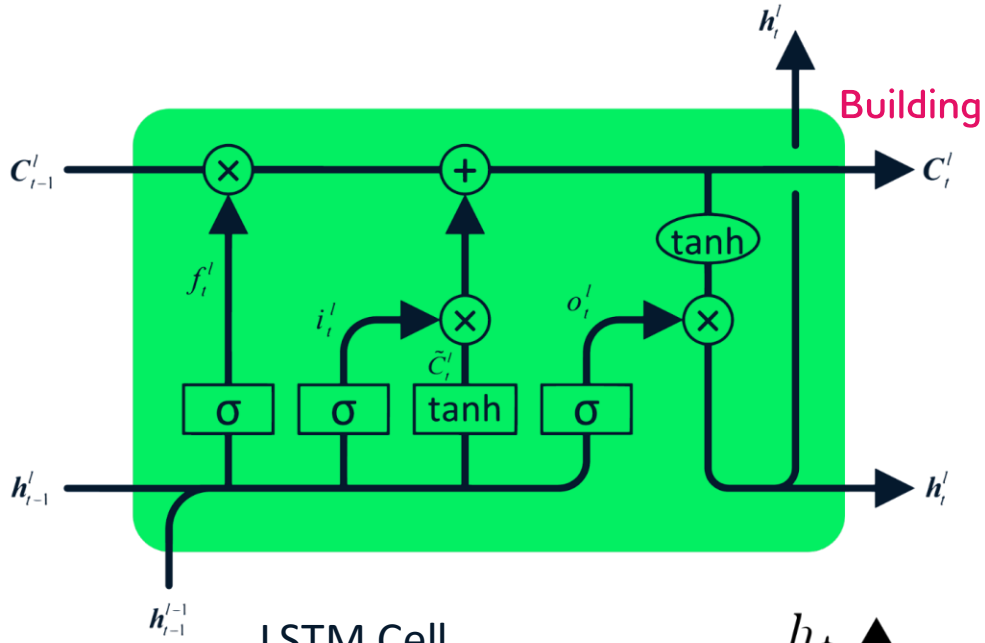
Working of RNNs



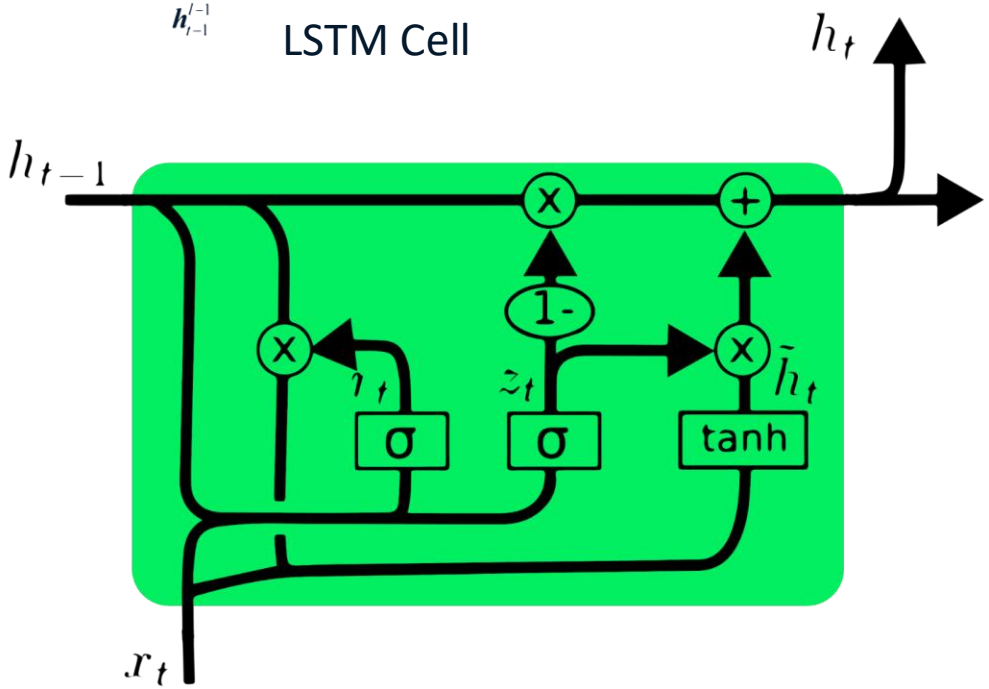
Advanced RNN Architectures: Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU)



Simple RNN Cell



LSTM Cell



GRU Cell

TATA CONSULTANCY SERVICES

Building on belief

Thank you