Advanced NLP with Machine Learning using Python

PIKAKSHI MANCHANDA

POST DOCTORAL RESEARCH FELLOW, VISTA AR

@itsPikakshi

p.Manchanda@Exeter.ac.uk

Previously on

Natural Language Processing with Python

Discussion on:

- What is NLP?
- Defining Text Analytics with the NLP Pipeline
  - Libraries used in Python
- Advanced Text Processing
  - Vector Space Model: TFIDF

Available on github.com/Pikakshi/NLP_Introduction

What is NLP?

- Natural Language Processing: teaching computers to understand (and generate) natural language for a range of applications by drawing insights.
- An umbrella term that describes the ability to break down the unstructured language to understand, process and generate a comprehensible unstructured output for humans.
- Draws from many disciplines including Computer Science, Computational Linguistics, Mathematics, Statistics, Artificial Intelligence, Psychology, …

Text Analytics

- Process of examining unstructured text data to extract useful information (key concepts and themes) to uncover meaningful insights.
- Helpful in tasks such as understanding customer experience, product reviews, sentiment analysis, document summarization and so on which aid in decision making processes.

Libraries available in Python

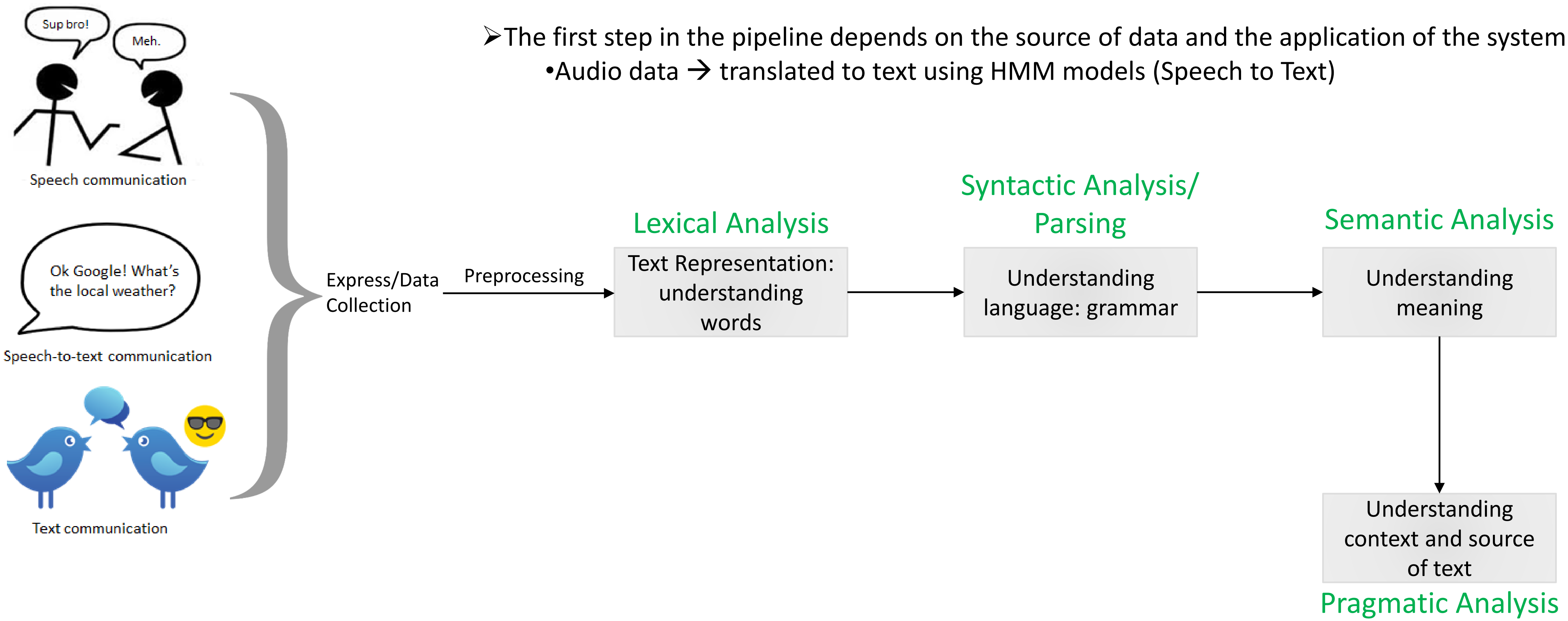
- Natural Language Toolkit (NLTK): Python Library for all NLP techniques.
- TextBlob – Easy to use NLP tools API, built on top of NLTK and Pattern.
- spaCy – Industrial strength NLP with Python and Cython.
- Gensim – Python library specialising in vector space modelling and topic modelling.
- Stanford Core NLP – A suite of NLP tools that provide POS tagging, entity recognition, sentiment analysis and many other services.
- Apache OpenNLP: Machine Learning toolkit that provides tokenizers, sentence segmentation, POS tagging, and more.
- scikit-learn: Machine learning in Python

Libraries available in Python

- Natural Language Toolkit (NLTK): Python Library for all NLP techniques.
- TextBlob – Easy to use NLP tools API, built on top of NLTK and Pattern.
- spaCy – Industrial strength NLP with Python and Cython.
- Gensim – Python library specialising in vector space modelling and topic modelling.
- Stanford Core NLP – A suite of NLP tools that provide POS tagging, entity recognition, sentiment analysis and many other services.
- Apache OpenNLP: Machine Learning toolkit that provides tokenizers, sentence segmentation, POS tagging, and more.
- scikit-learn: Machine learning in Python

The NLP Pipeline

- The first step in the pipeline depends on the source of data and the application of the system.
  - Audio data → translated to text using HMM models (Speech to Text)



Speech communication

Speech-to-text communication

Text communication

Express/Data Collection

Preprocessing

Lexical Analysis

Text Representation: understanding words

Syntactic Analysis/ Parsing

Understanding language: grammar

Semantic Analysis

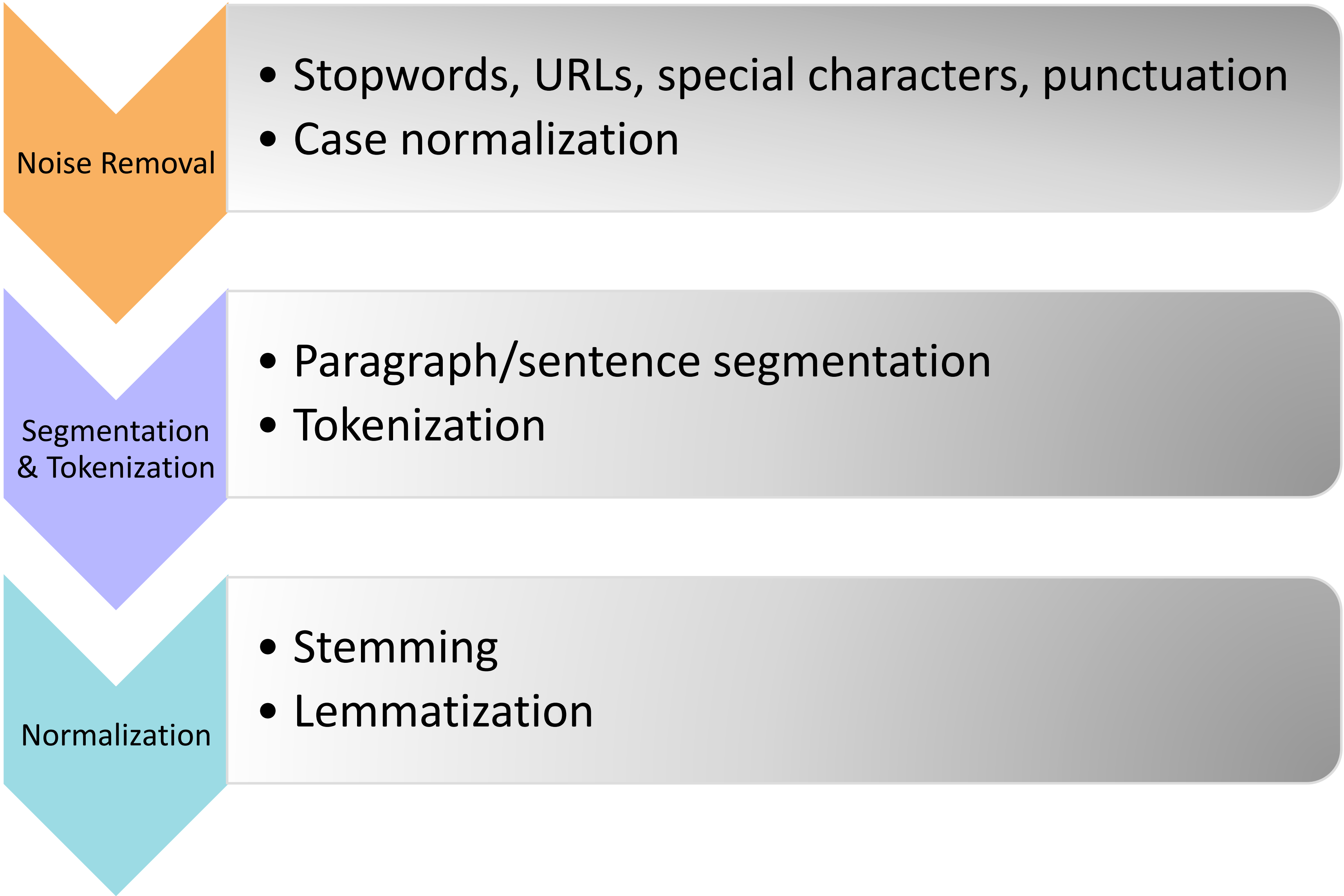
Understanding meaning

Understanding context and source of text

Pragmatic Analysis

- Different NLP systems may use different techniques, but overall, data processing steps are fairly similar.

Text Preprocessing



Noise Removal

- Stopwords, URLs, special characters, punctuation
- Case normalization

```
• set(stopwords.words('english'))
• s.translate(str.maketrans("", "", string.punctuation))
• word =  word.lower()
```

Segmentation & Tokenization

- Paragraph/sentence segmentation
- Tokenization

```
• tokenize.sent_tokenize(line)
• word_tokenize(line)
```

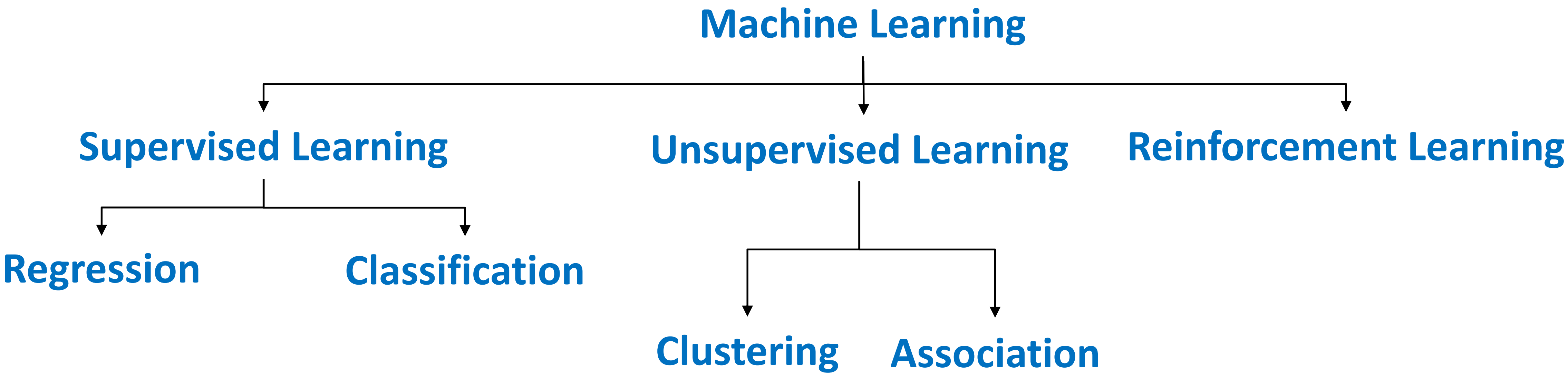
Normalization

- Stemming
- Lemmatization

```
• porter = PorterStemmer()
• lem = WordNetLemmatizer()
```

Refer to the python code for all the above steps available on github.com/Pikakshi/NLP_Introduction/TextPreprocessing.ipynb

In the morning session…



Machine Learning

Supervised Learning

Unsupervised Learning

Reinforcement Learning

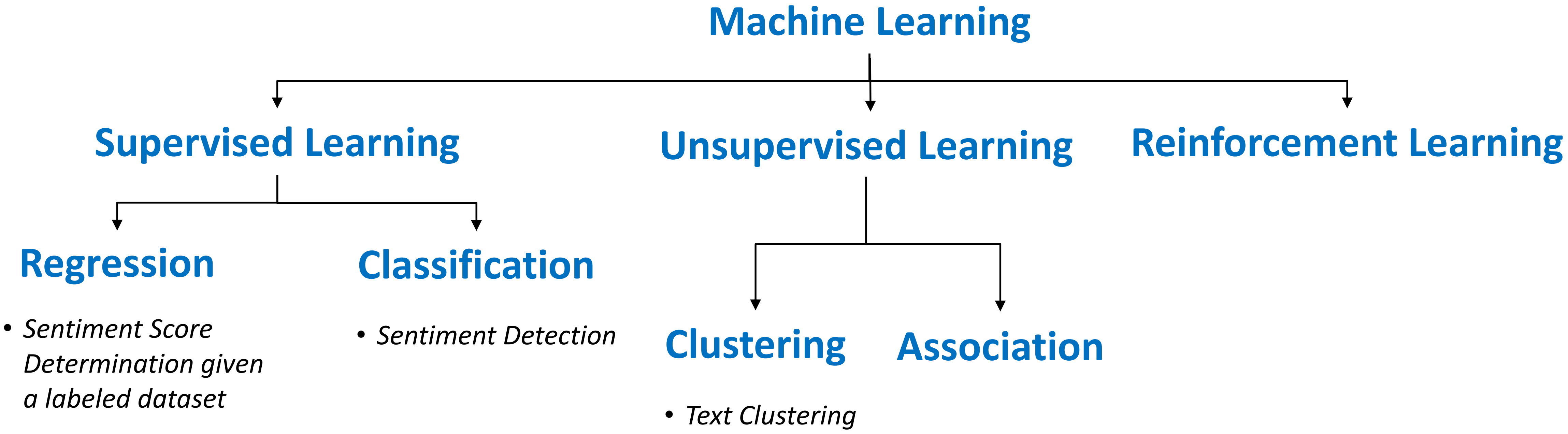
Regression

Classification

Clustering

Association

Text Processing



Machine Learning

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Regression

- *Sentiment Score Determination given a labeled dataset*

Classification

- *Sentiment Detection*

Clustering

- *Text Clustering*

Association

We also learnt that…

A Machine Learning project has a series of well known steps:

- Define the problem
- Load data
- Evaluate Algorithms
- Make Predictions

So how do we process text?

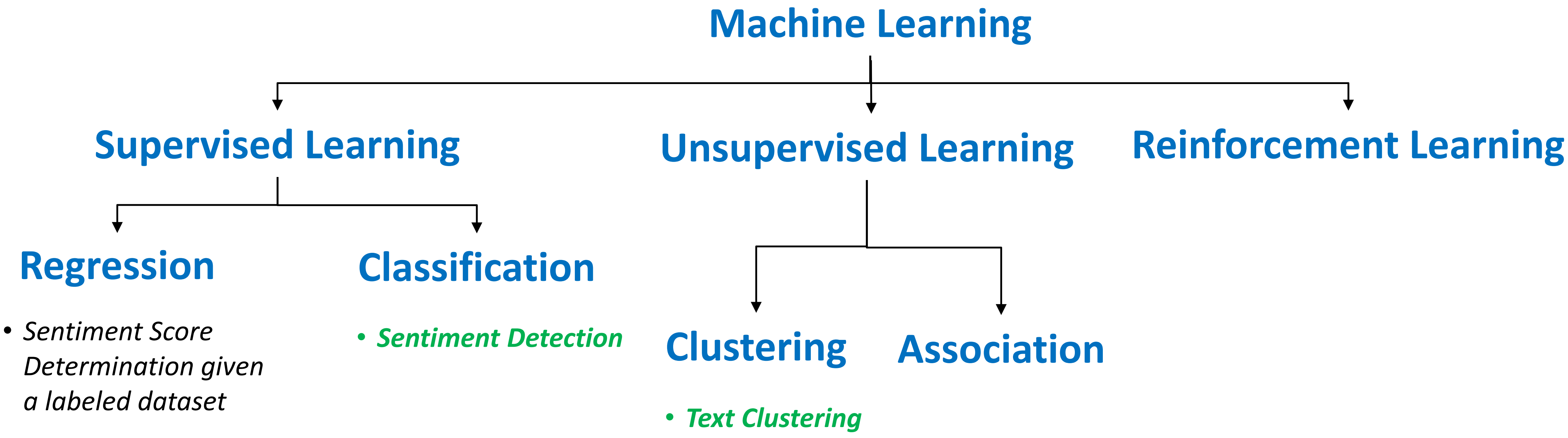
Following the same steps:

- Define the problem
- Load data
  - Loading, visualising and pre-processing for cleaning the data
  - Feature Engineering: Use of Vector space models, word embeddings, Text based features, …

Covered in previous session

- Evaluate Algorithms
  - Training on available data
  - Evaluation Metrics
  - Model Selection
- Make Predictions
  - Test selected model on unseen data

Text Processing



Machine Learning

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Regression

- *Sentiment Score Determination given a labeled dataset*

Classification

- ***Sentiment Detection***

Clustering

- ***Text Clustering***

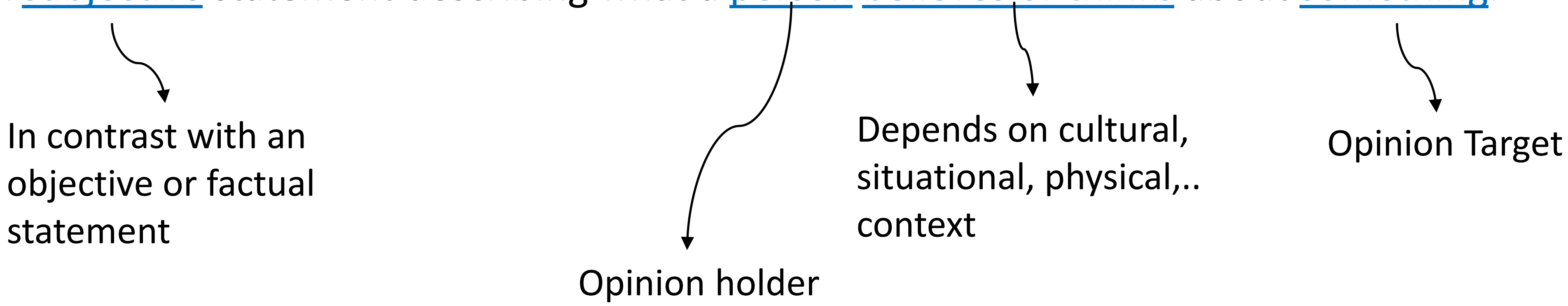
Association

1. Sentiment Classification

- Can be seen as a *Text Categorization problem*, wherein given a set of predefined categories and a training set of labeled text objects, the task is to classify a new text object into one or more of the categories.
- Learn a classifier function $f: X \rightarrow Y$, s.t. $f(x) = y \in Y$ where X = all text objects and Y = all categories
- Use of features of text objects to distinguish categories (such as use of semantic categories)
- Good performance requires: 1.) effective features and 2.) plenty of training data
- Performance is generally effected more by the effectiveness of features than by choice of a specific classifier.
- Common evaluation metrics: P/R/F1/Accuracy
- Step-by-step tutorial provided here.

What is an opinion?

- A subjective statement describing what a person believes or thinks about something.



In contrast with an objective or factual statement

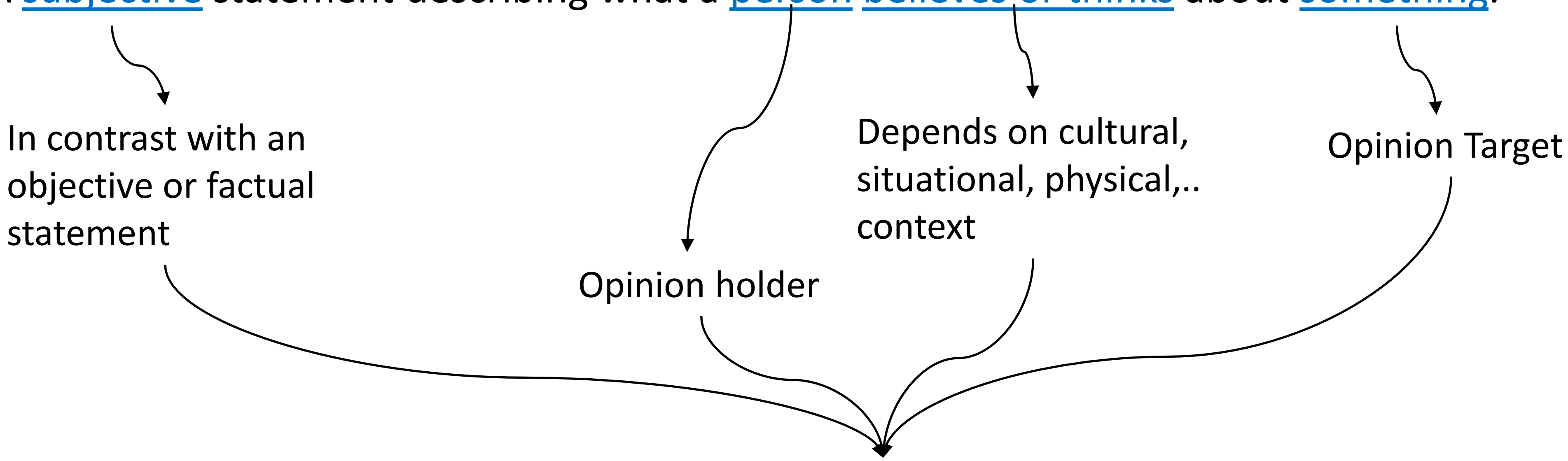
Opinion holder

Depends on cultural, situational, physical,.. context

Opinion Target

What is an opinion?

- A subjective statement describing what a person believes or thinks about something.



In contrast with an objective or factual statement

Opinion holder

Depends on cultural, situational, physical,.. context

Opinion Target

Opinion sentiment: What does the opinion tells us about the opinion holder’s feelings? Eg. Positive/negative

2. Text Clustering

- Discovering ‘natural structure’ in data and group similar objects together. Objects can be documents, terms, passages, websites and so on.

K-means Clustering:

- Represent each text object as a *term vector* and assume a similarity function defined on 2 objects.
- Start with *k* randomly selected vectors and assume they are the centroids of the *k* clusters.
- Assign every vector to a cluster whose centroid is closest to the vector (using Euclidian distance).
- Re-compute the centroid for each cluster based on the newly assigned vectors in the cluster.
- Repeat until the similarity-based objective function converges to a local minimum.

- K-means algo → easy to implement and computationally effective.
- Step-by-step tutorial provided here.

Reading Material

Online courses/Articles:

- Applied Text Mining in Python by University of Michigan at Coursera.
- Machine Learning with Text in Python – Data School

Books:

- Natural Language Processing with Python by Steven Bird, Ewan Klein and Edward Loper.
- Foundations of Statistical Natural Language Processing by Christopher Manning and Hinrich Schütze.