Introduction to NLP

Software Engineering Decision Supports Lab

University of Calgary

Table of Contents

Applications of NLP

NLP general workflow

Classification schema of NLP techniques

Textual data

Text-Preprocessing

Text parsing and exploratory data analysis

Text representation and vectorizing

Modeling and/or pattern mining

Random Forest Classifier

Goals of NLP tutorial by SEDS lab (today)

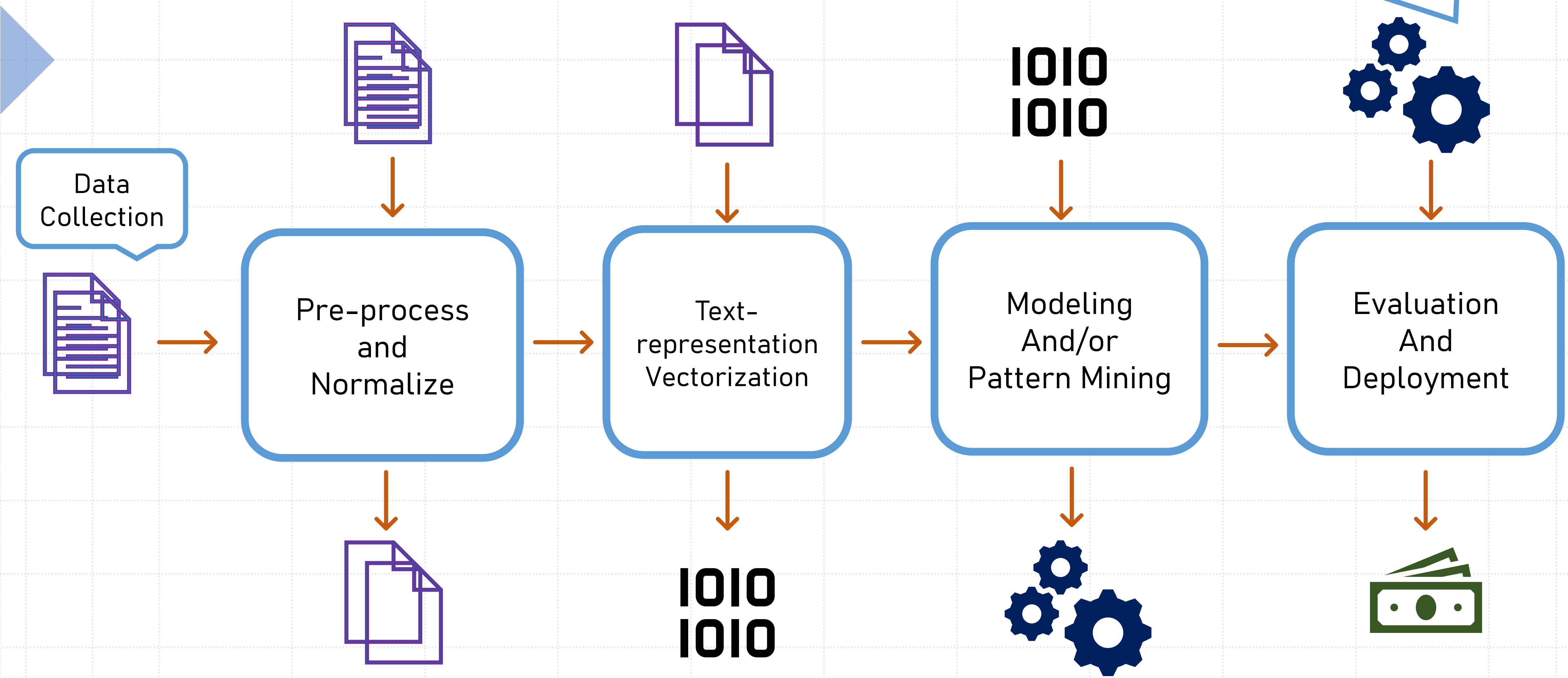
- General understanding of NLP,
  - Applications, use cases
  - Workflow
  - General tools
- Build a sentiment analysis tool Together!
  - Explaining each part
  - Implementing them

Applications of NLP

- Email Filters
- Fraud Detection
- Search Results
- Predictive Text
- Language translation
- Data and Text analysis
- More …

- Requirements Elicitation
- Sentiment Analysis
- Opinion Mining
- Text Summarization
- Context Analysis
- Decision Support

NLP general workflow



Sentiment Analysis
Classification (spam, fraud, …)
Language Generation
More …

Data Collection

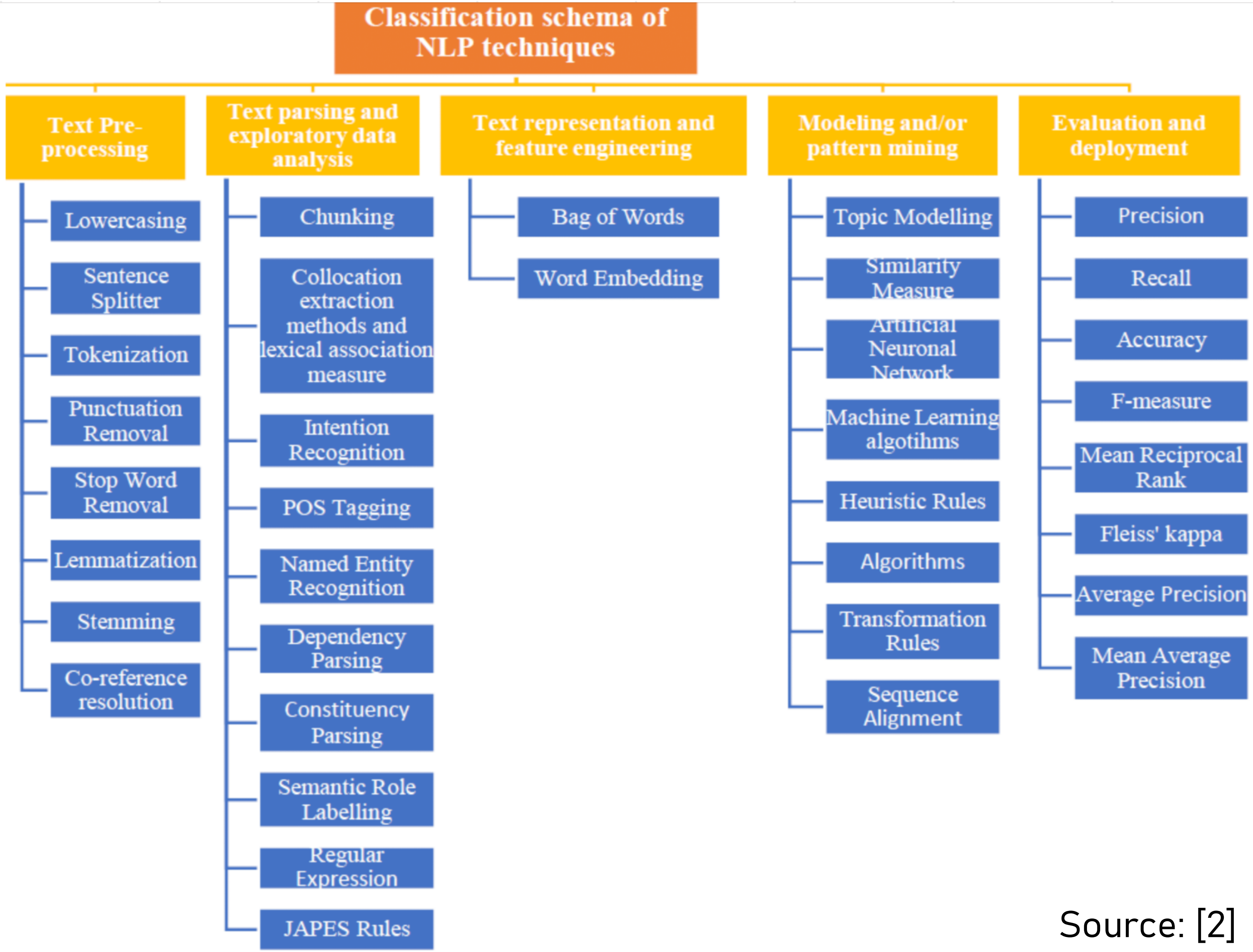
Pre-process and Normalize

Text-representation Vectorization

Modeling And/or Pattern Mining

Evaluation And Deployment

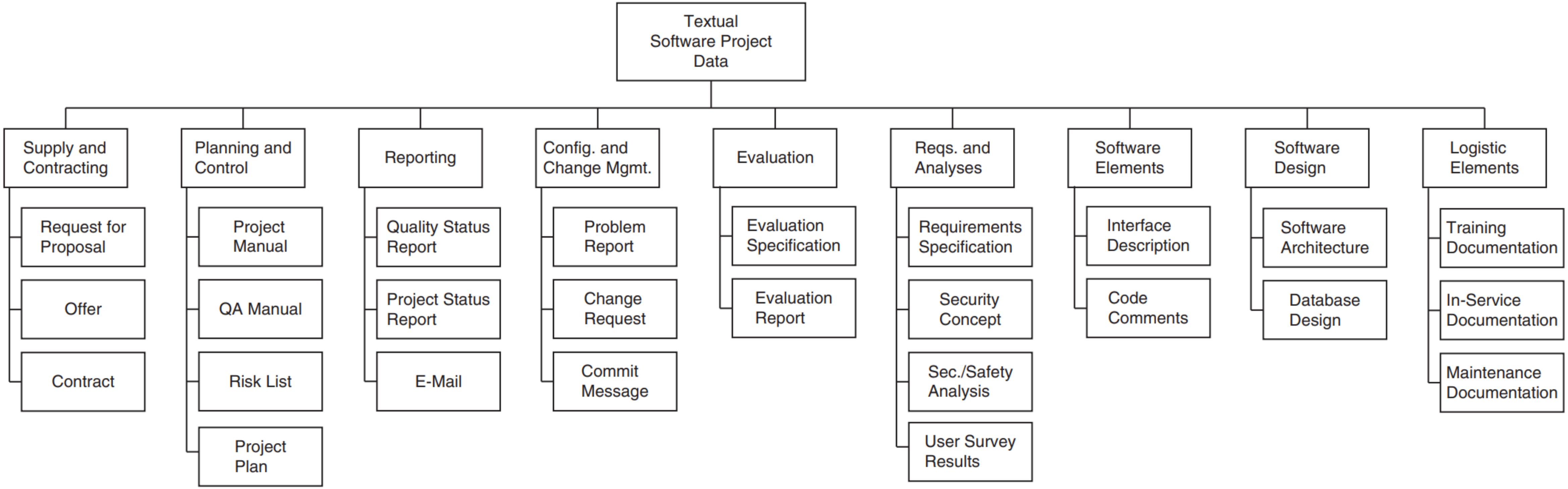
Classification schema of NLP techniques



- Classification schema of NLP techniques
  - Text Pre-processing
    - Lowercasing
    - Sentence Splitter
    - Tokenization
    - Punctuation Removal
    - Stop Word Removal
    - Lemmatization
    - Stemming
    - Co-reference resolution
  - Text parsing and exploratory data analysis
    - Chunking
    - Collocation extraction methods and lexical association measure
    - Intention Recognition
    - POS Tagging
    - Named Entity Recognition
    - Dependency Parsing
    - Constituency Parsing
    - Semantic Role Labelling
    - Regular Expression
    - JAPES Rules
  - Text representation and feature engineering
    - Bag of Words
    - Word Embedding
  - Modeling and/or pattern mining
    - Topic Modelling
    - Similarity Measure
    - Artificial Neuronal Network
    - Machine Learning algotihms
    - Heuristic Rules
    - Algorithms
    - Transformation Rules
    - Sequence Alignment
  - Evaluation and deployment
    - Precision
    - Recall
    - Accuracy
    - F-measure
    - Mean Reciprocal Rank
    - Fleiss' kappa
    - Average Precision
    - Mean Average Precision

Source: [2]

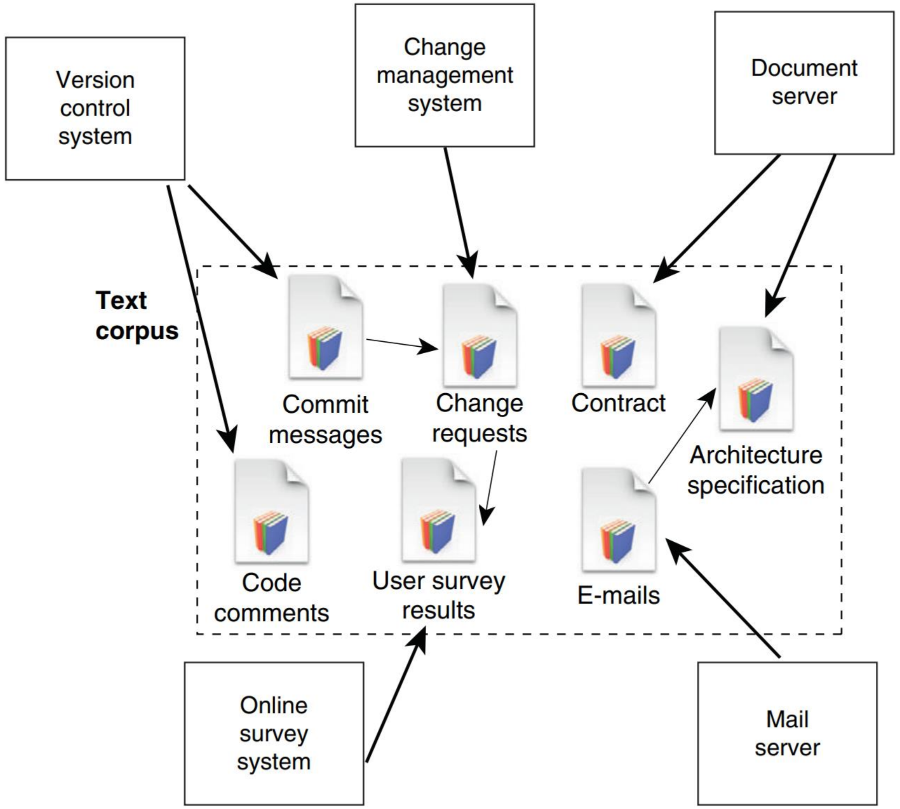
Textual Data in Software Projects



Source: [1]

Textual Data Retrieval

- API's
- Enterprise Resource Planning (ERP) systems
- Version Control history
- Communications
  - Email
  - Chat
- Surveys



Source of image: [1]

Let’s Start Coding

Text-Preprocessing

Text Parsing and Exploratory data analysis

- Named Entity Recognition
- Part-Of-Speech (POS) Tagging

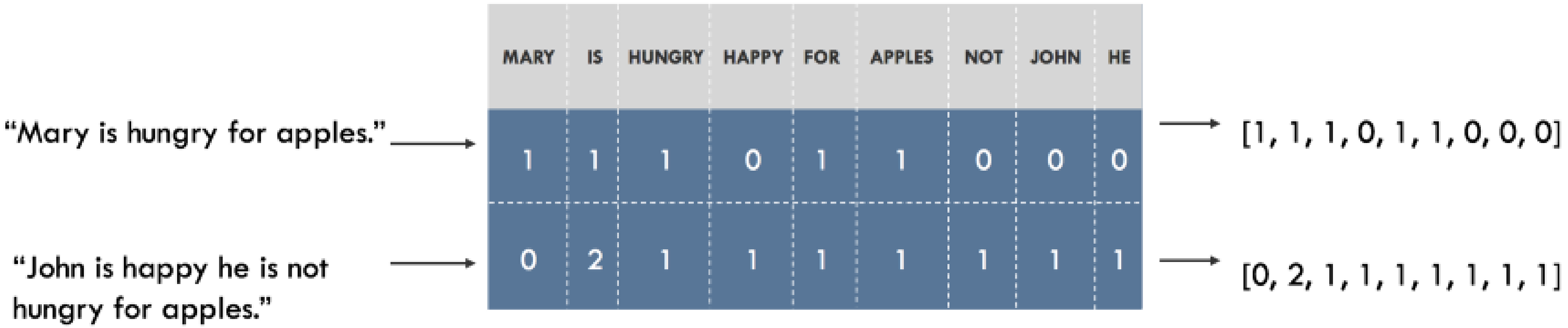
Let’s See some examples in the existing general purpose NLP tools:

- IBM’s NLP tool set called Watson (link)
- Google’s NLP API (link)

Text representation – Vectorizing

- A few Text Representation and Vectorization methods:
  - **Bag of Words (BoW)**
  - N-Gram (extension of BoW)
  - **TF-IDF**
  - Word2Vec
  - Doc2Vec
  - BERT

Bag of Words



	MARY	IS	HUNGRY	HAPPY	FOR	APPLES	NOT	JOHN	HE	
“Mary is hungry for apples.”	1	1	1	0	1	1	0	0	0	[1, 1, 1, 0, 1, 1, 0, 0, 0]
“John is happy he is not hungry for apples.”	0	2	1	1	1	1	1	1	1	[0, 2, 1, 1, 1, 1, 1, 1, 1]

Source: https://blog.insightdatascience.com/how-to-solve-90-of-nlp-problems-a-step-by-step-guide-fda605278e4e

- Why this is NOT a good text representation?

Term Frequency-Inverse Document Frequency (TF-IDF)

- Why TF-IDF?
- What is it?
  - Creates a document-term matrix; one row per document, one column per word in the corpus
  - Generates a weighting for each word/document pair intended to reflect how important a given word is to the document within the context of its frequency within a larger corpus

$$\text{M tweets}\left\{\overbrace{\begin{array}{ccccc} 0.27 & 0.23 & \dots & 0.77 & 0.68 \\ 0.86 & 0.96 & \dots & 0.3 & 0.83 \\ \dots & \dots & \dots & \dots & \dots \\ 0.18 & 0.71 & \dots & 0.87 & 0.63 \\ 0.68 & 0.29 & \dots & 0.61 & 0.92 \end{array}}^{\text{N words}}\right.$$

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i})$$

$w_{i,j}$ = weighting of word i for document j

$tf_{i,j}$ = number of times i occurs in j divided by the total number of terms in j

df_i = number of documents containing word i

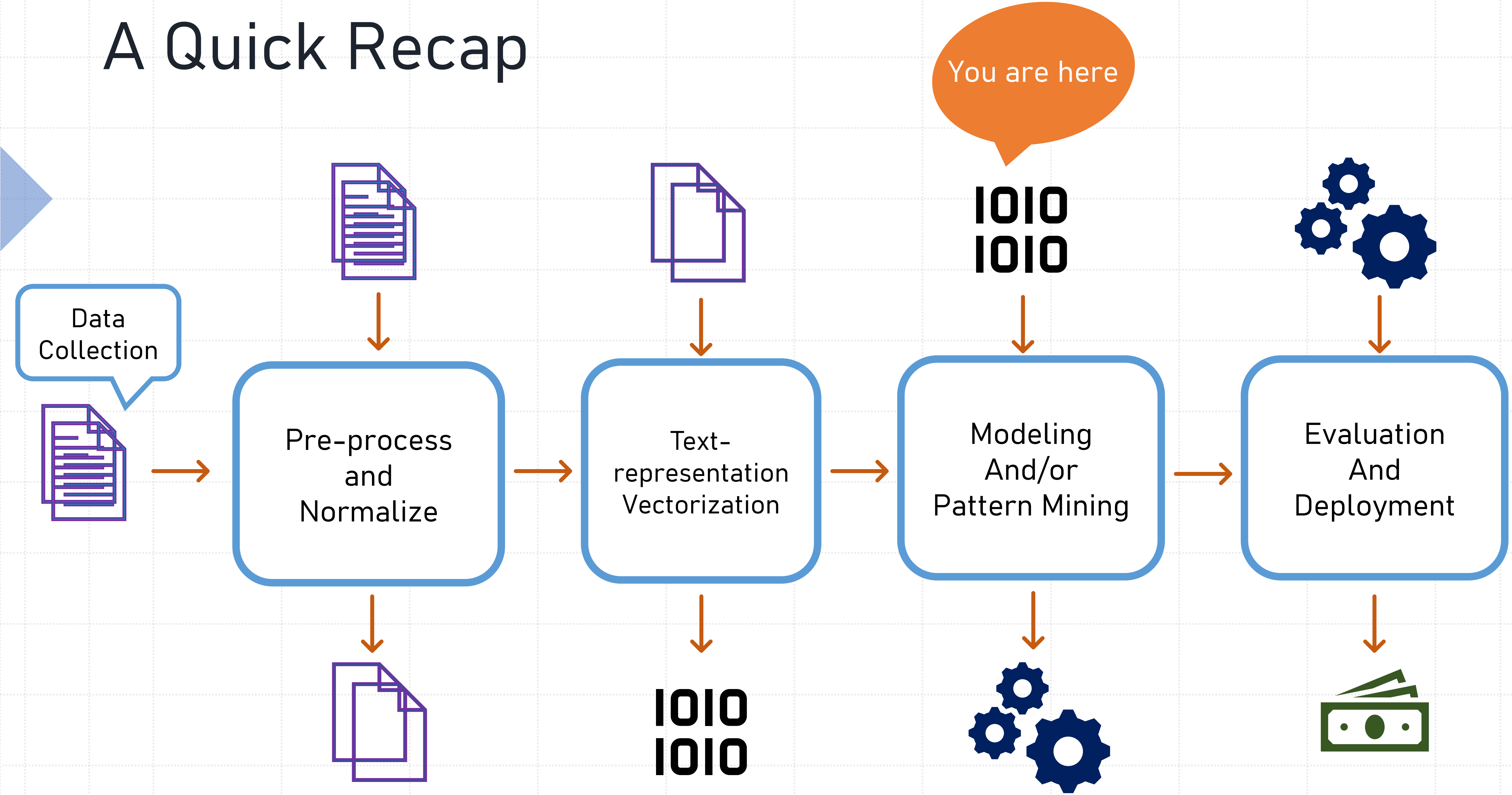
N = total number of documents

Let's Continue Coding

TF-IDF

Vectorization

A Quick Recap



You are here

Data Collection

Pre-process and Normalize

Text-representation Vectorization

Modeling And/or Pattern Mining

Evaluation And Deployment

Modeling and/or Pattern mining

- Machine Learning:
  - NVIDIA, 2016: Practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world. [3]

- Two broad types of ML
  - Supervised Learning

    Inferring a function from a labeled training data to make predictions on unseen data

    Example: Predict whether any given tweet has a positive or negative emotion

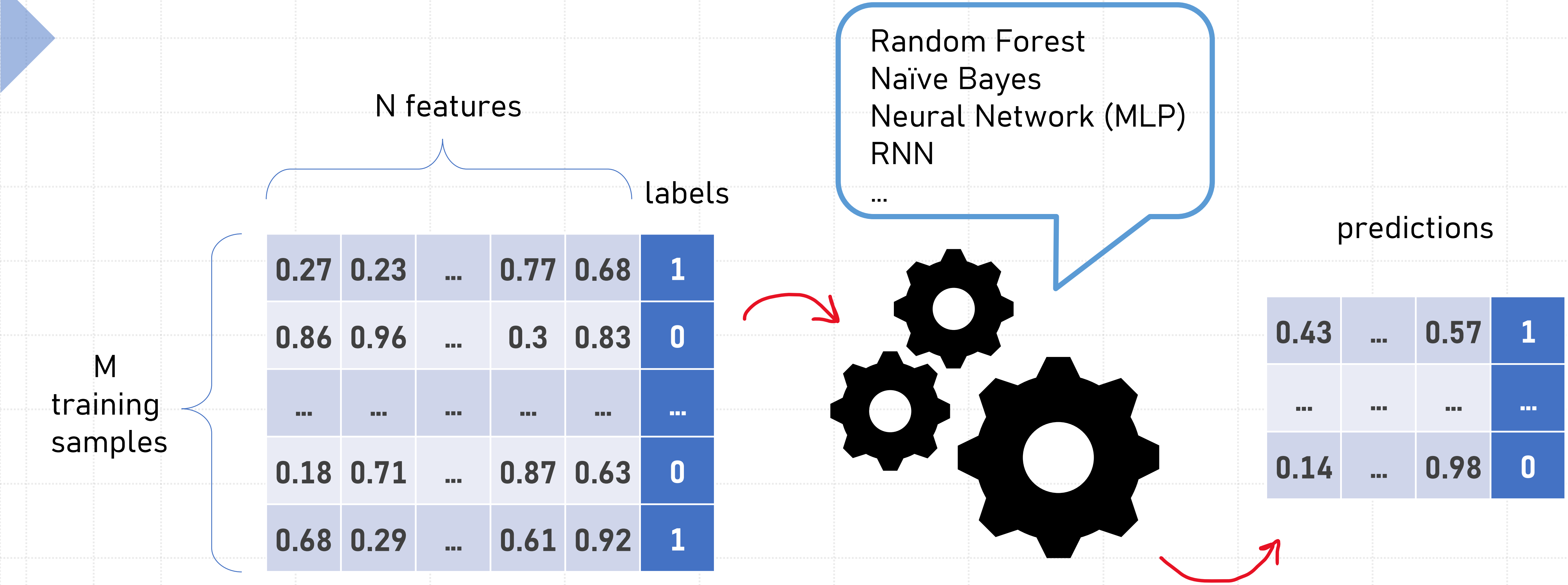
  - Unsupervised Learning

    Deriving structure from data where we don't know the effect of any of the variables

    Example: Based on the content of a tweet, group similar tweets together in distinct folders

Modeling and/or Pattern mining

Supervised Learning and pattern mining



N features

labels

M training samples

0.27	0.23	…	0.77	0.68	1
0.86	0.96	…	0.3	0.83	0
…	…	…	…	…	…
0.18	0.71	…	0.87	0.63	0
0.68	0.29	…	0.61	0.92	1

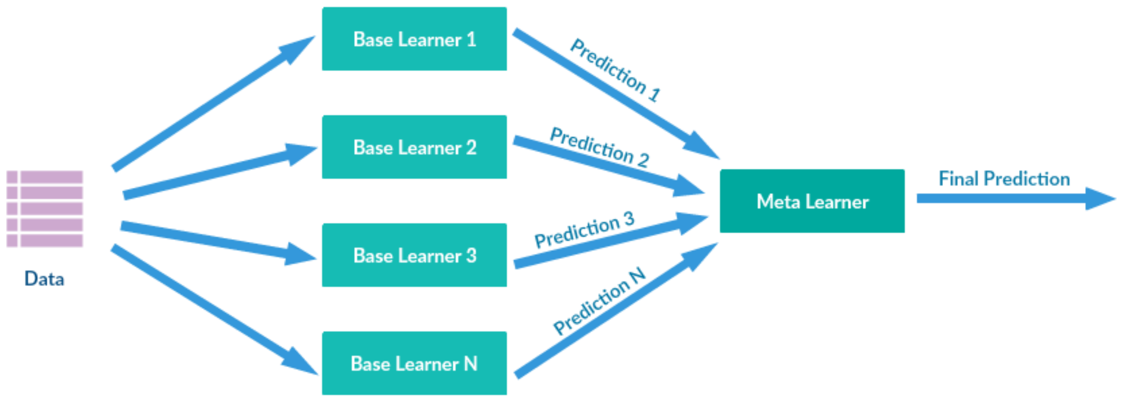
Random Forest
Naïve Bayes
Neural Network (MLP)
RNN
…

predictions

0.43	…	0.57	1
…	…	…	…
0.14	…	0.98	0

Random Forest

- Why Random Forest?
- Ensemble Methods



Source: https://medium.com/geekculture/the-power-of-ensemble-96cd2621c2de

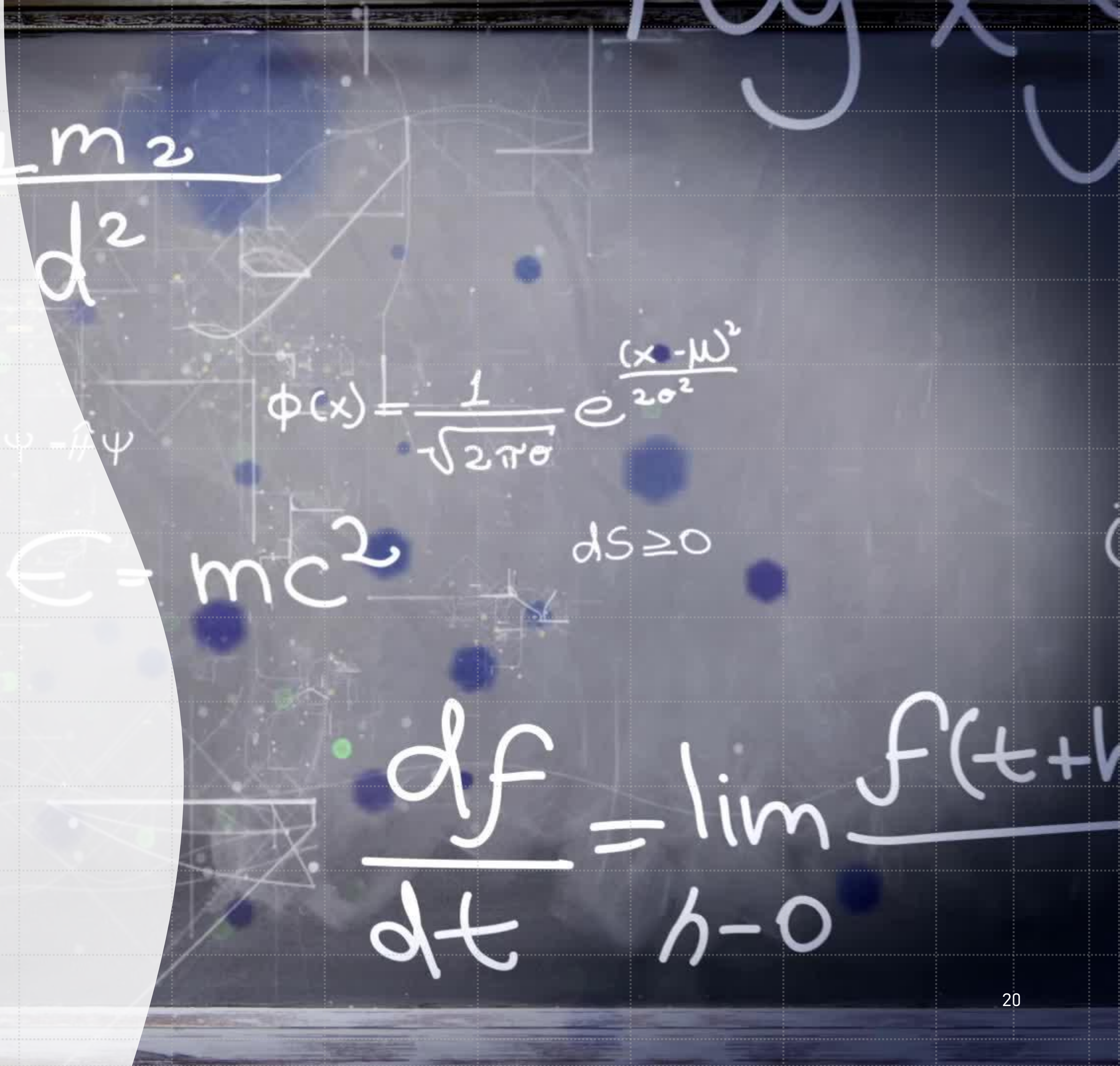
- Can be used for classification or regression
- Handles outliers, missing values, etc.
- Less likely to overfit

- Random Forest
  - **Ensemble learning method**
  - A collection of **decision trees**
  - **aggregates the predictions**
  - A simple voting method for decision trees (DT):

  70 DTs vote for **Positive** > 30 DTs for Negative

Let's Continue and Finish Coding

Modeling and Pattern Mining



References

1. Bird, Christian, Tim Menzies, and Thomas Zimmermann, eds. The art and science of analyzing software data. Elsevier, 2015.
2. Moises Gonzalez-Garcia, Speech recognition, NLP, and the use of ontologies to identify the problem-domain and solution requirements: A systematic mapping study, Information and Software Technology, 2019, In press
3. https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/#:~:text=Machine%20learning%20at%20its%20most,about%20something%20in%20the%20world.
4. https://realpython.com/nltk-nlp-python/
5. https://towardsdatascience.com/tf-idf-a-visual-explainer-and-python-implementation-on-presidential-inauguration-speeches-2a7671168550