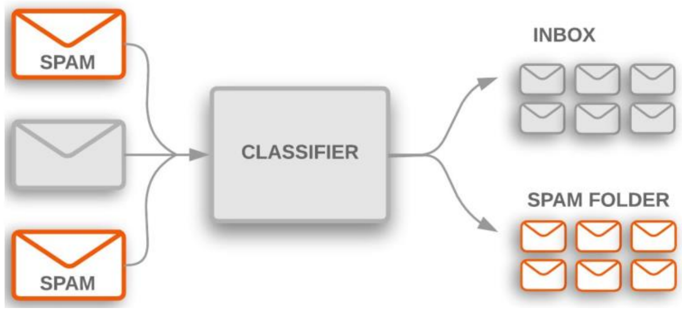
Machine Learning for NLP

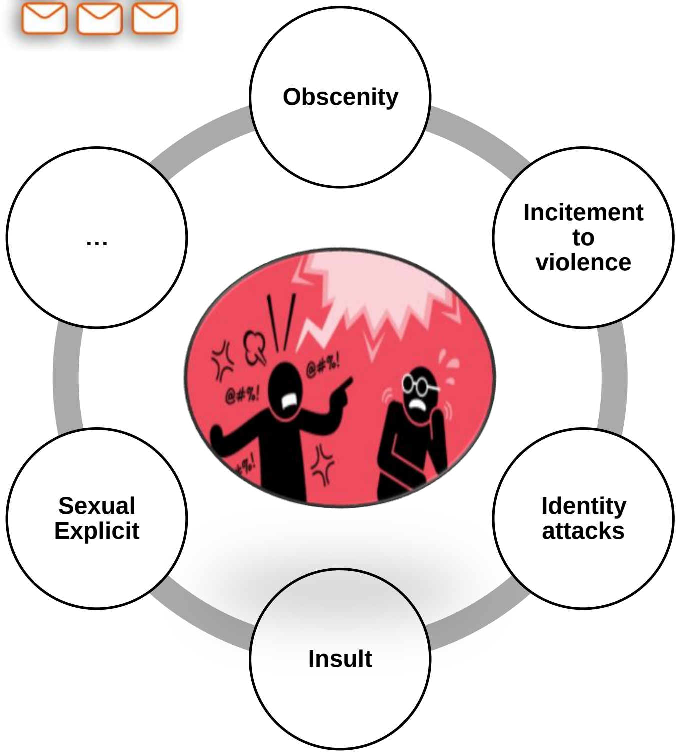
Mouhcine MENDIL

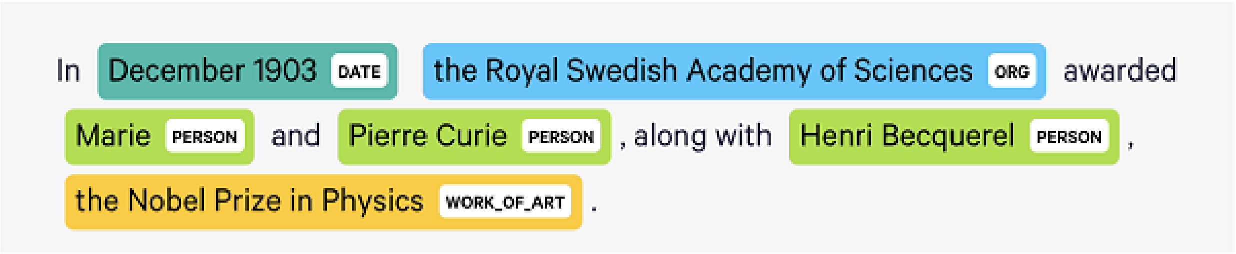
Some Tasks in NLP: Classification

Assigning predefined labels or categories to text:

- Hate speech detection in social media
- Spam detection
- Topic categorization
- Named Entity Recognition
- …





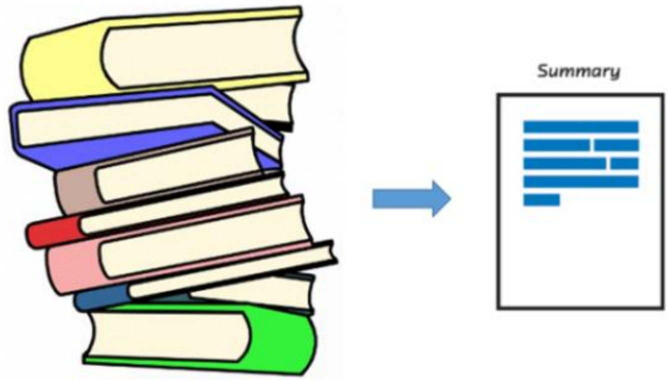


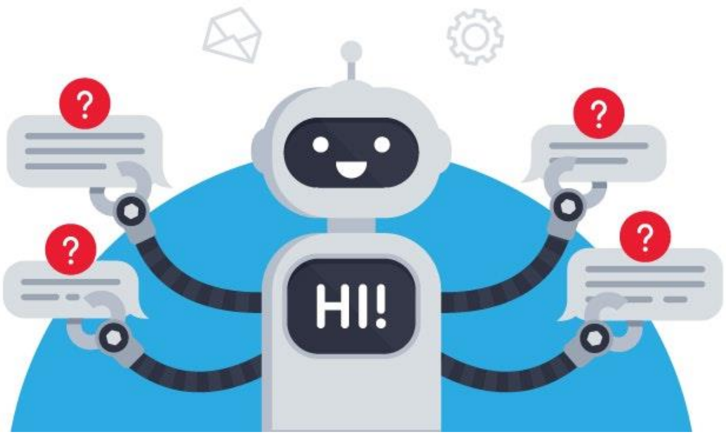
FRENCH INSTITUTES OF TECHNOLOGY

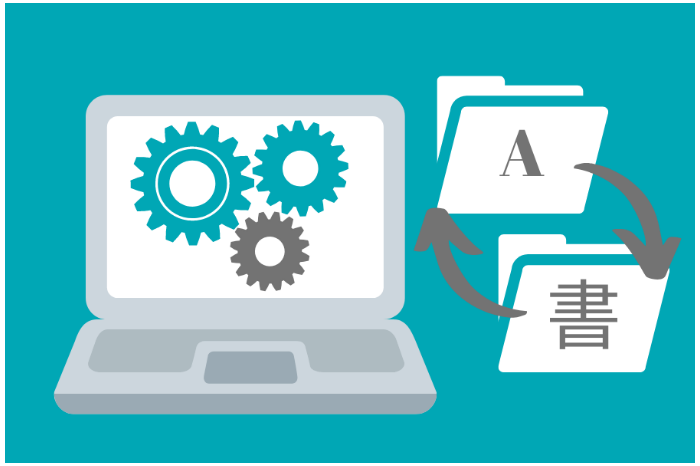
Some Tasks in NLP: Text Generation

Producing coherent and contextually relevant text based on given input:

- Machine Translation
- Text Summarization
- Paraphrasing
- Chatbots
- …





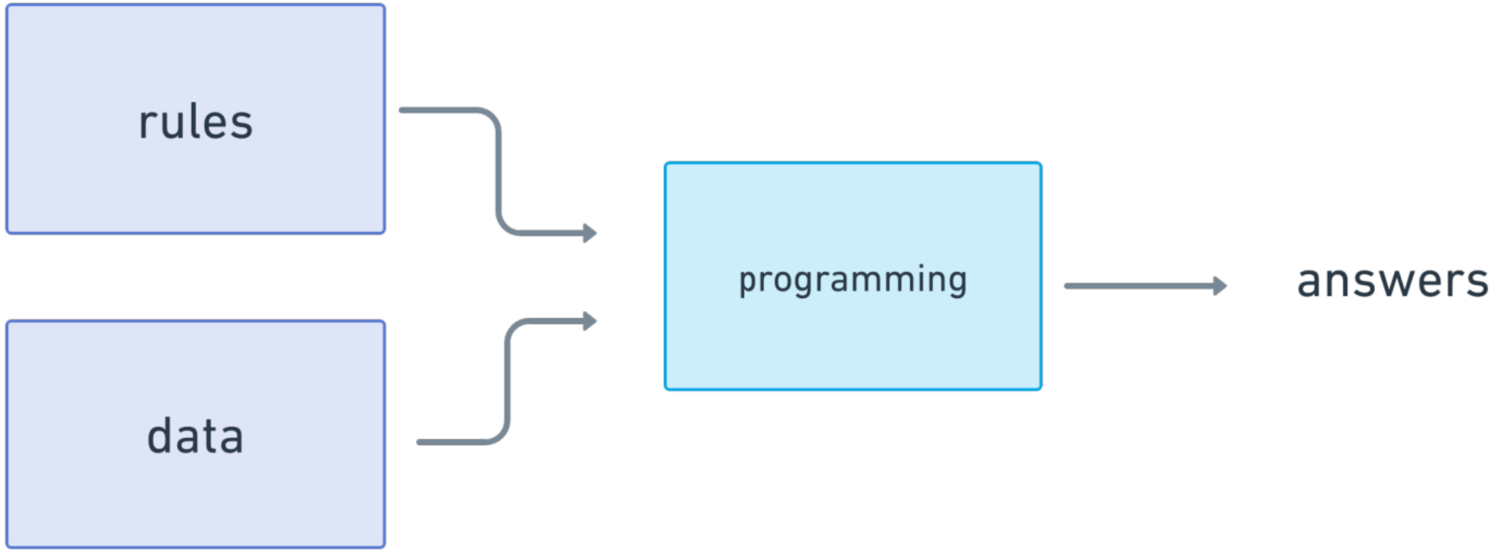


DeepL

FRENCH INSTITUTES OF TECHNOLOGY

Rules-based Systems

- Early approaches (60’s): symbolic methods, hand-written rules



Rules-based Systems

- Regular expressions: language for specifying a matching pattern in text search

example@gmail.com

([a-zA-Z0–9_.+-]+)@[a-zA-Z0–9_.+-]+\.[a-zA-Z0–9_.+-]

Rules-based Systems

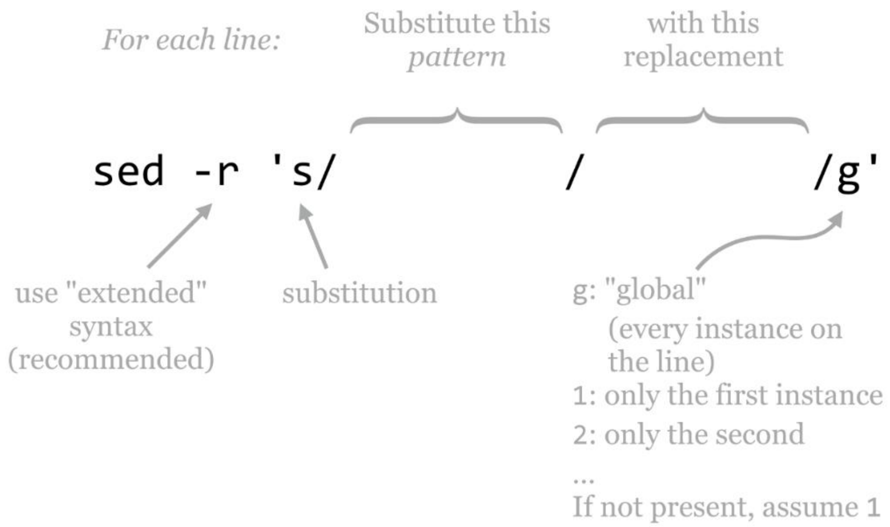
- Regular expressions: language for specifying a matching pattern in text search
- Useful for text normalization and substitution

In [3]:
```python
import re

text = "i love python and PYTHON is great for many tasks"
pattern = r'\bpython\b'
repl = "Python"
result = re.sub(pattern, repl, text, flags=re.IGNORECASE)
```

In [4]:
```python
print(result)
```

```
i love Python and Python is great for many tasks
```



Rules-based Systems

- ELIZA simulates a Rogerian psychologist

https://web.njit.edu/~ronkowit/eliza.html

FRENCH INSTITUTES OF TECHNOLOGY

Rules-based Systems

- ELIZA simulates a Rogerian psychologist
- Hard-coded rules based on pattern matching and substitution

https://web.njit.edu/~ronkowit/eliza.html

FRENCH INSTITUTES OF TECHNOLOGY

Rules-based Systems

- ELIZA simulates a Rogerian psychologist
- Hard-coded rules based on pattern matching and substitution

Example of a condition and possible answers:

"?*x I want ?*y": [

"What would it mean if you got ?y?",

"Why do you want ?y?",

"Suppose you got ?y soon.“

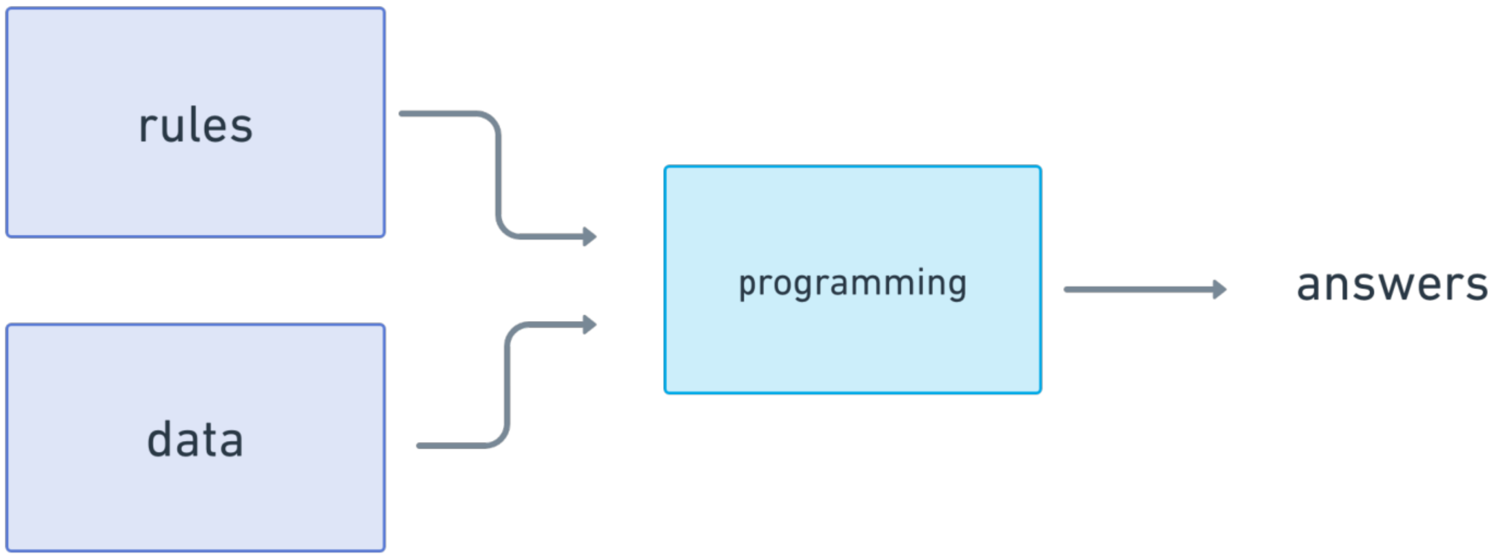
…

]

https://web.njit.edu/~ronkowit/eliza.html

Rules-based Systems

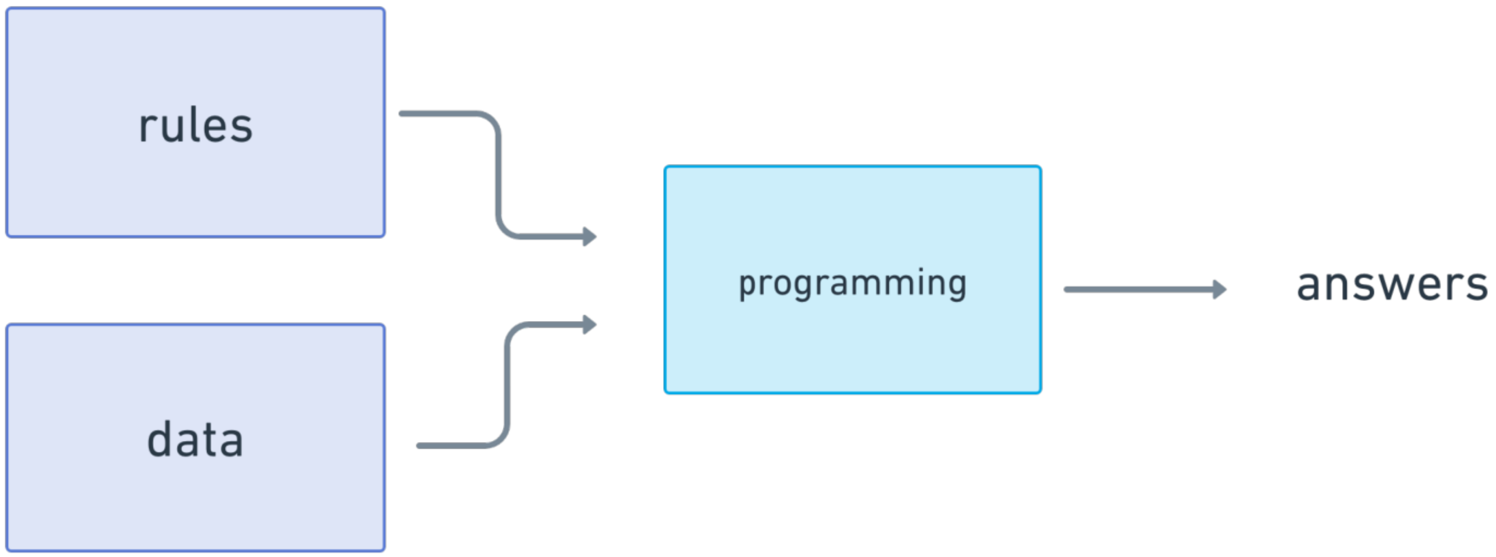
- Early approaches (60’s): symbolic methods, hand-written rules
- Advantages: based on expert knowledge, precise



FRENCH INSTITUTES OF TECHNOLOGY

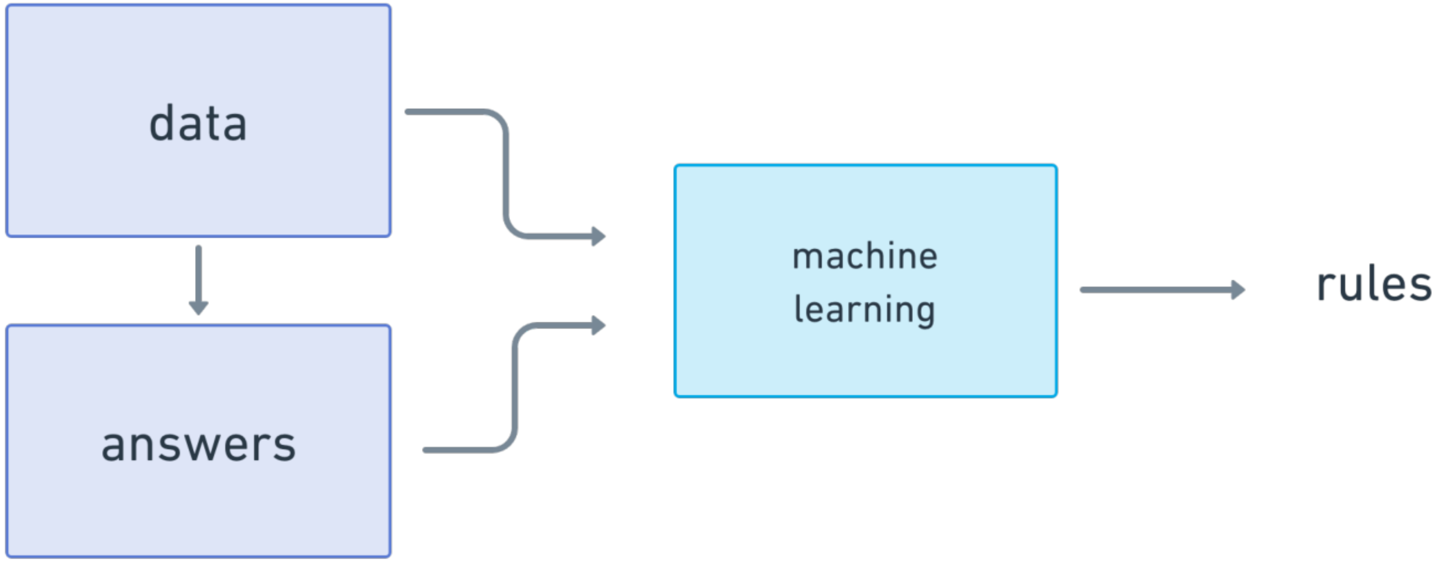
Rules-based Systems

- Early approaches (60's): symbolic methods, hand-written rules
- Advantages: based on expert knowledge, precise
- Downsides: lack of coverage, expensive to build and maintain



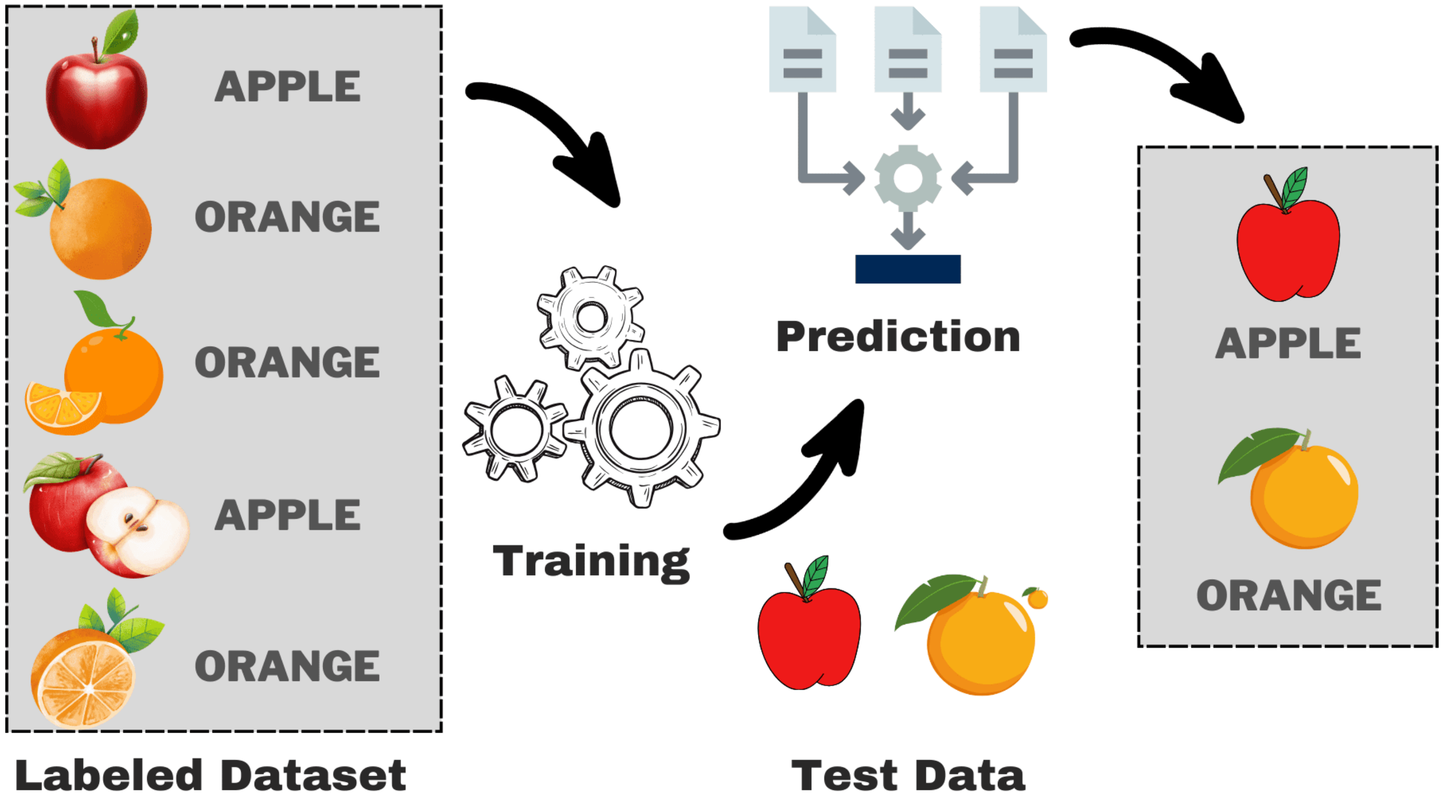
Learning from Data (Supervised)

- Learn rules automatically: machine learning (90’s), neural methods (2010’s)



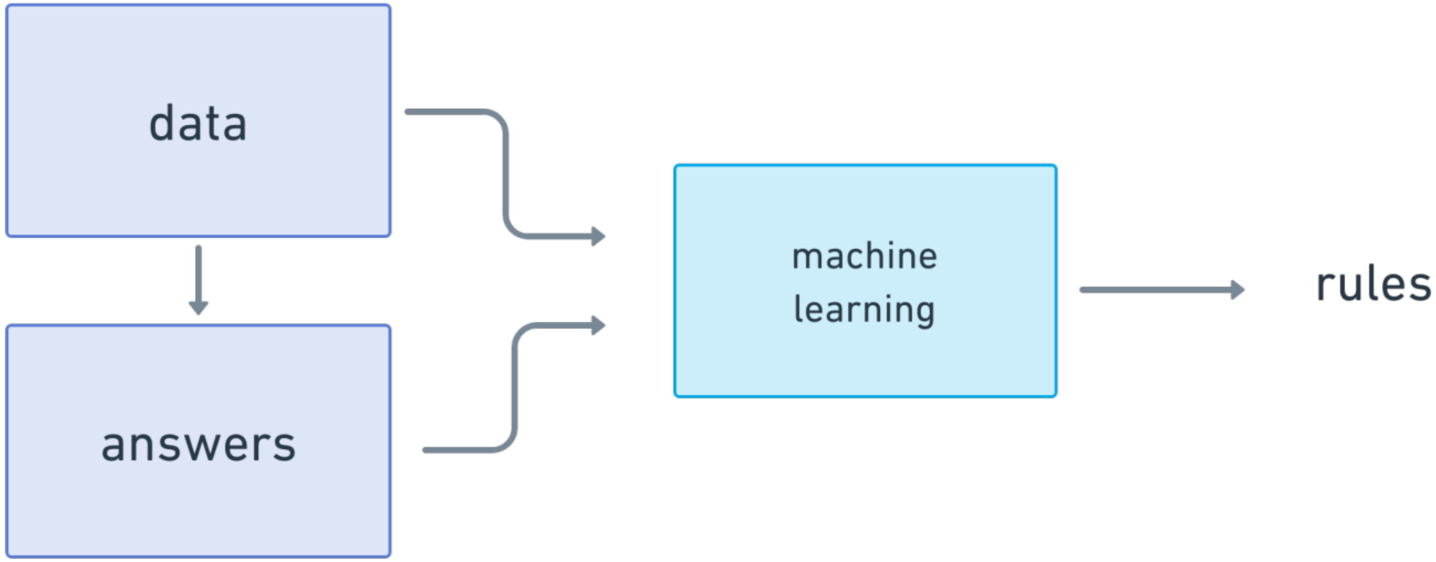
Learning from Data (Supervised)

- Learn rules automatically: machine learning (90's), neural methods (2010's)



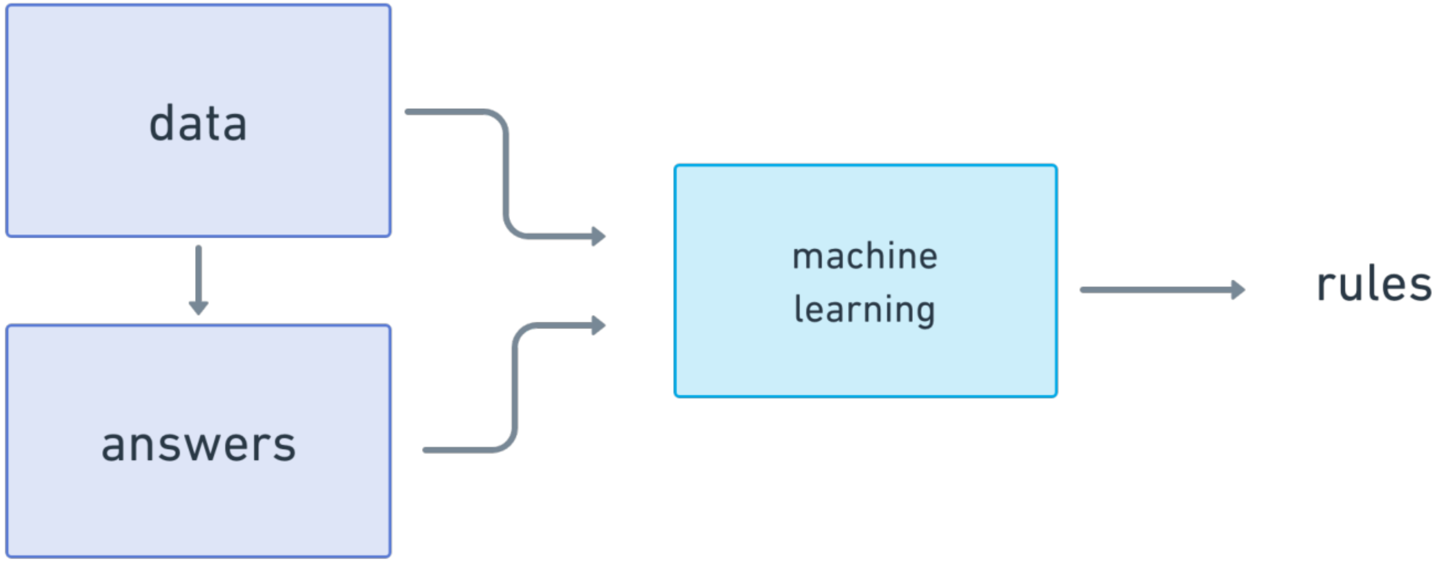
Learning from Data (Supervised)

- Learn rules automatically: machine learning (90’s), neural methods (2010’s)
- Advantages: improved performance and generalization, fast

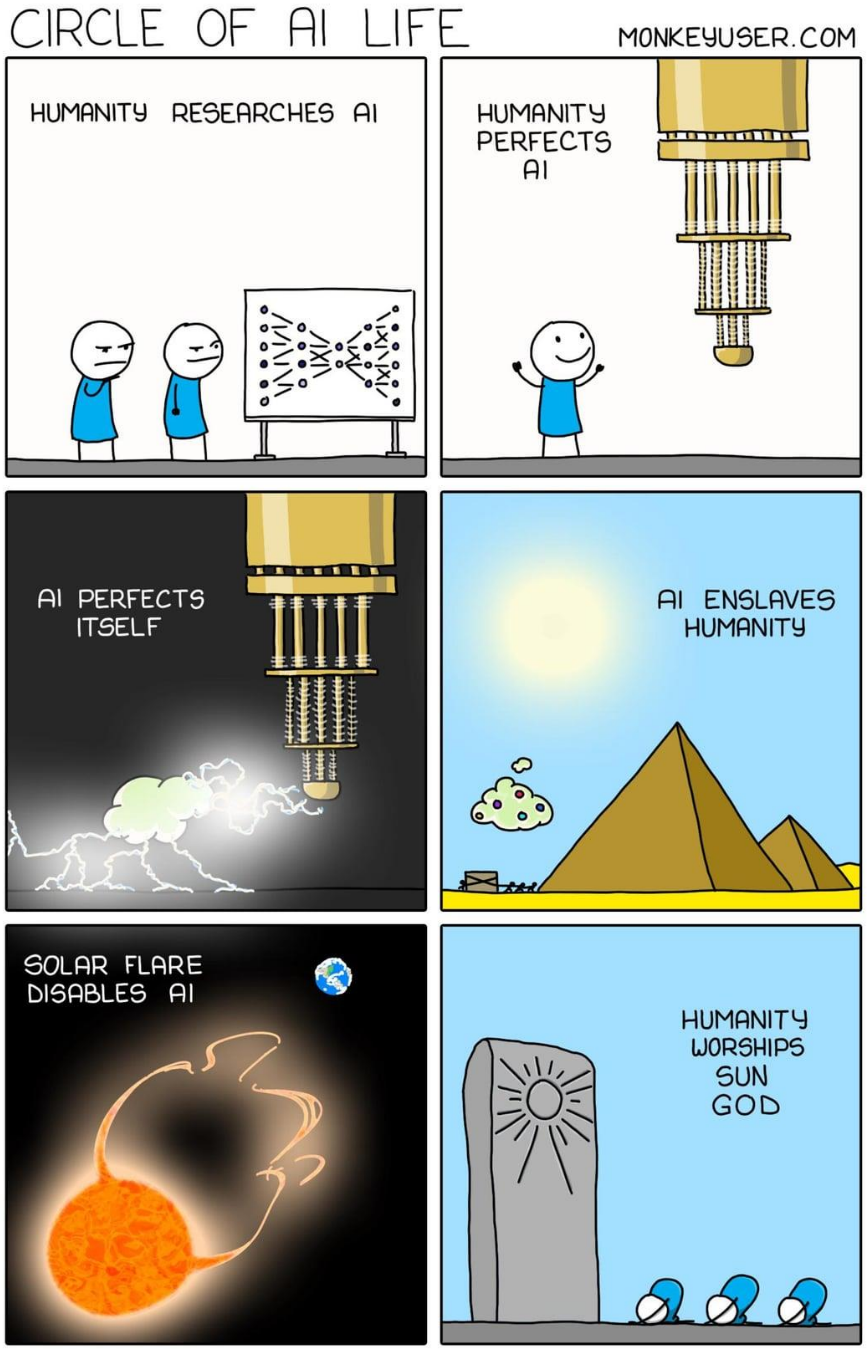


Learning from Data (Supervised)

- Learn rules automatically: machine learning (90’s), neural methods (2010’s)
- Advantages: improved performance and generalization, fast
- Downsides: complex, hard to interpret



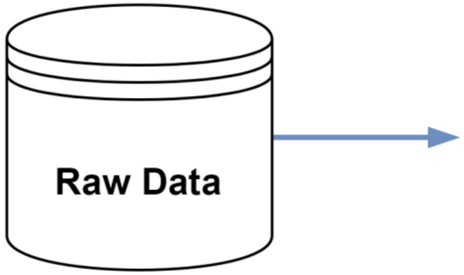
FRENCH INSTITUTES OF TECHNOLOGY



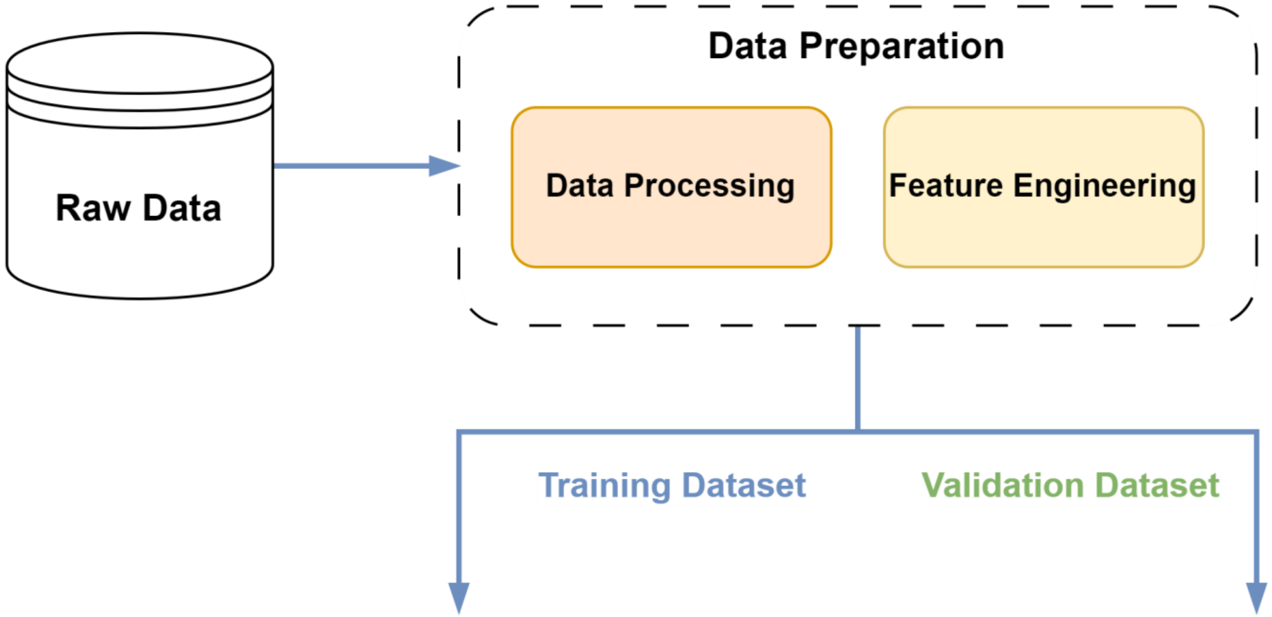
Questions so far ?

FRENCH INSTITUTES OF TECHNOLOGY

ML Workflow (Supervised)

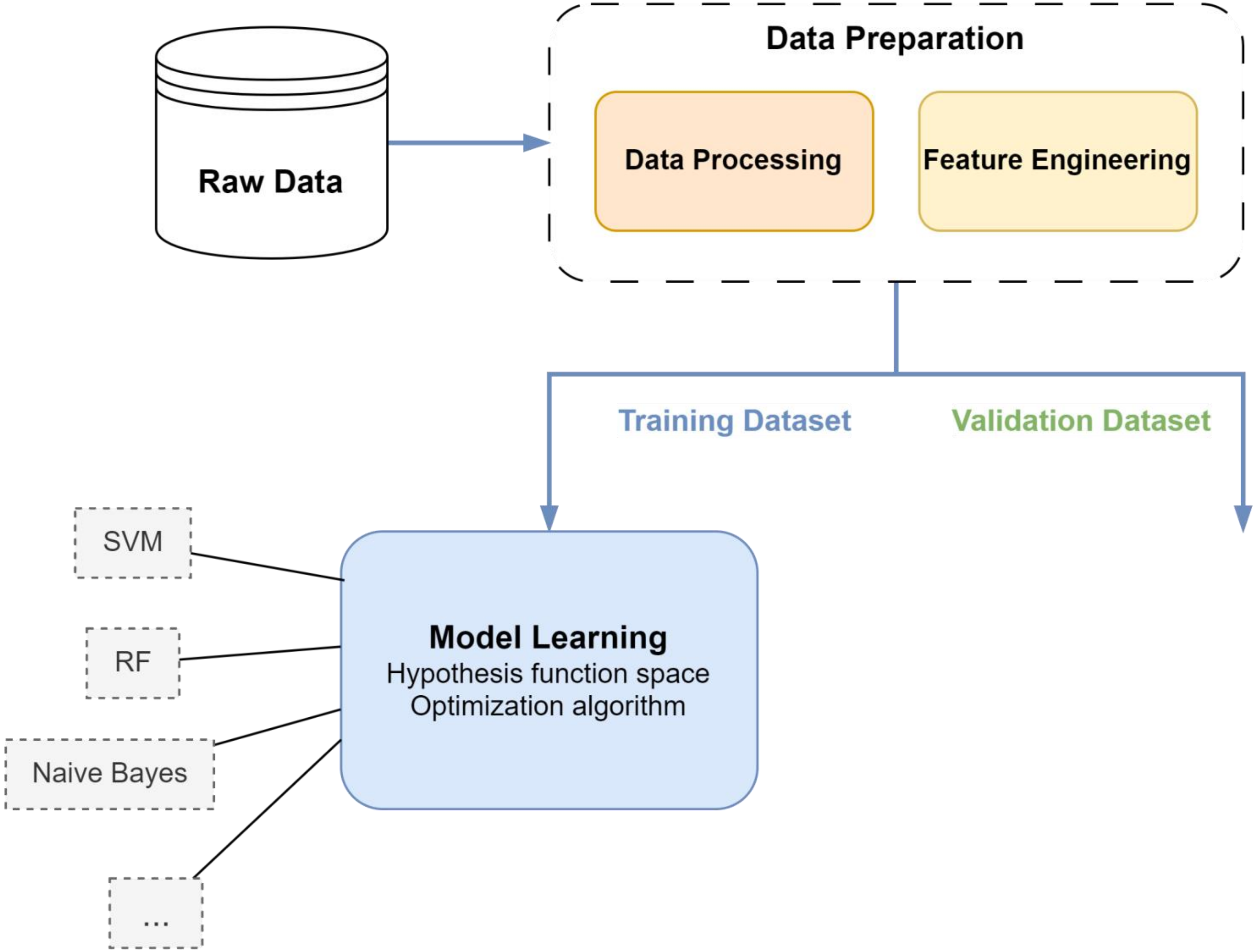


ML Workflow (Supervised)

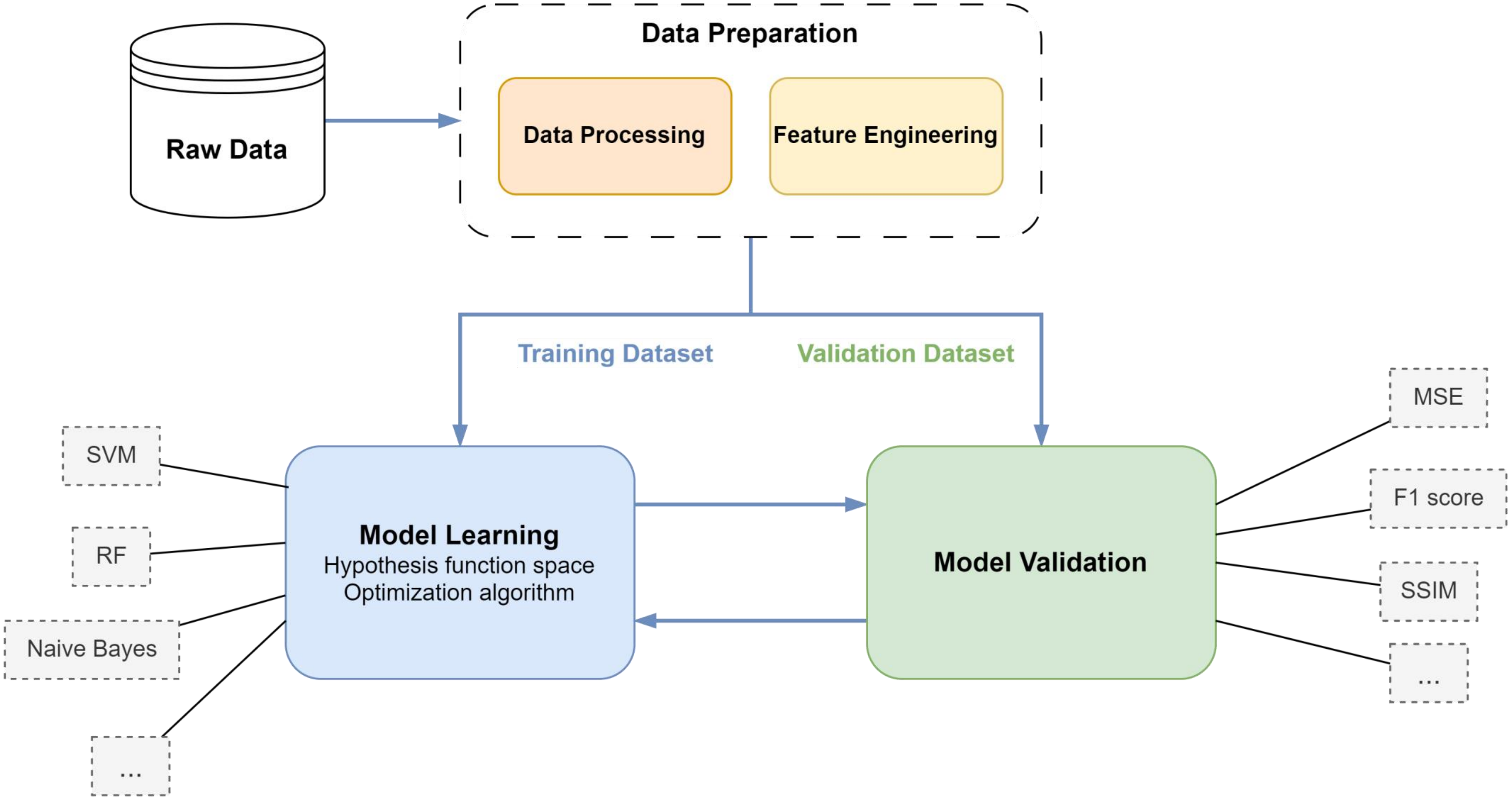


FRENCH INSTITUTES OF TECHNOLOGY

ML Workflow (Supervised)



ML Workflow (Supervised)



ML Workflow for NLP

- How to split data into train/val/test subsets ?

FRENCH INSTITUTES OF TECHNOLOGY

ML Workflow for NLP

- How to split data into train/val/test subsets ?

- Models need for numerical features. How to go from sentences/words/tokens to vector representation?

FRENCH INSTITUTES OF TECHNOLOGY

ML Workflow for NLP

- How to split data into train/val/test subsets ?

- Models need for numerical features. How to go from sentences/words/tokens to vector representation?

- Are there suitable models for learning ?

ML Workflow for NLP

- How to split data into train/val/test subsets ?

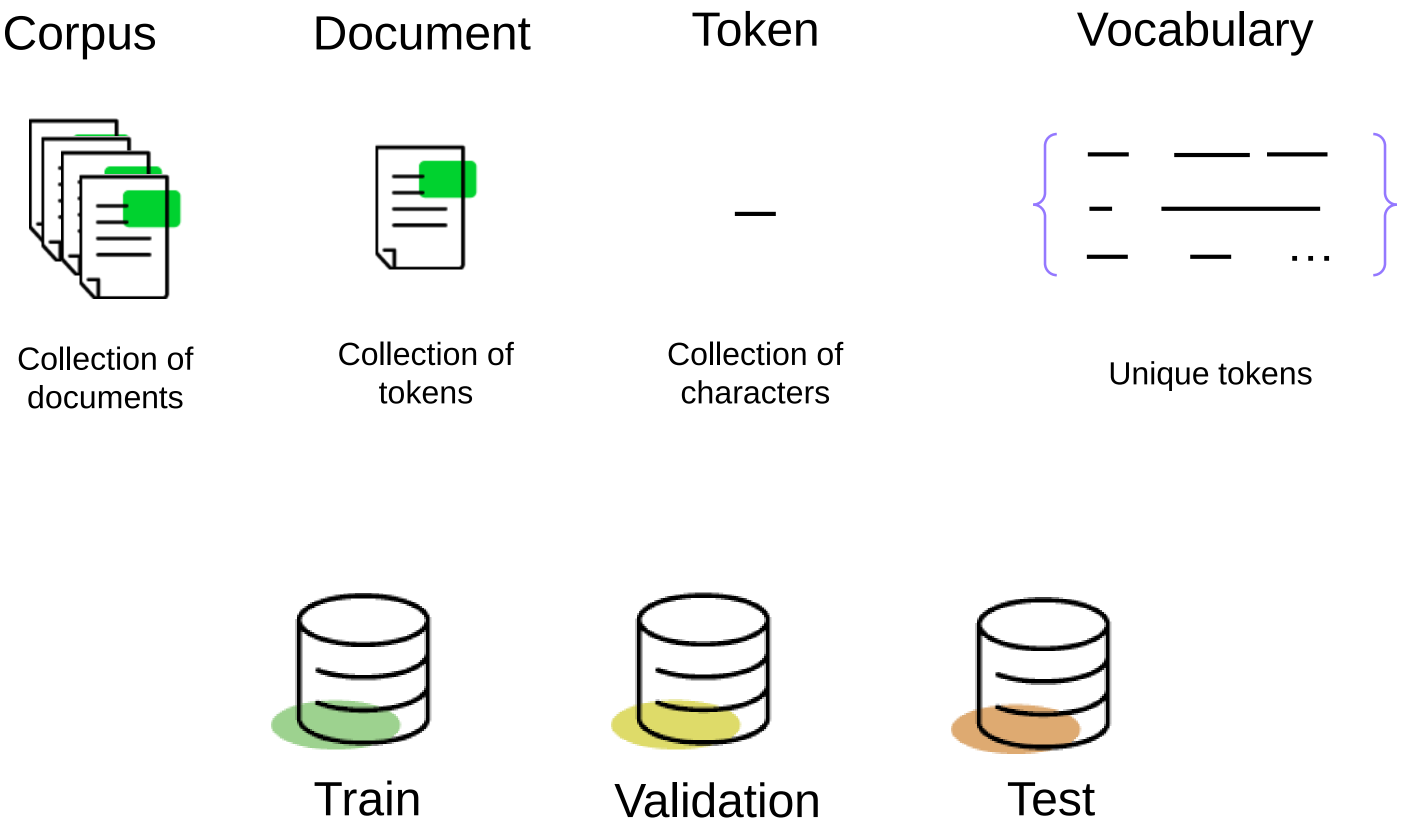
- Models need for numerical features. How to go from sentences/words/tokens to vector representation?

- Are there suitable models for learning ?

- Metrics for evaluation ?

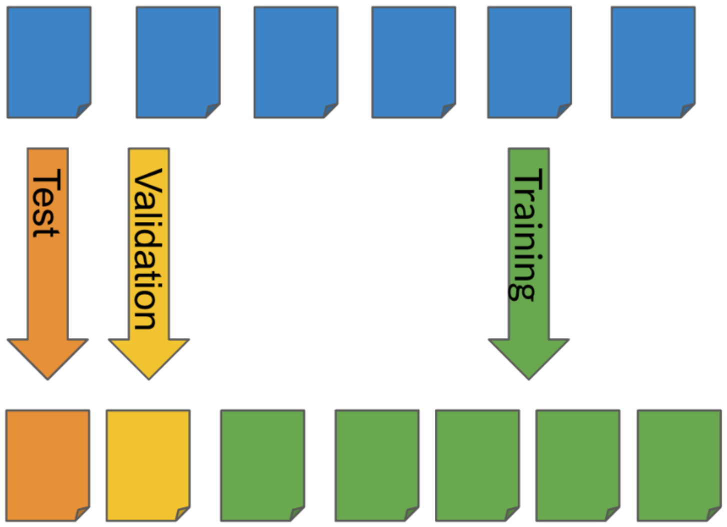
FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Splitting the data

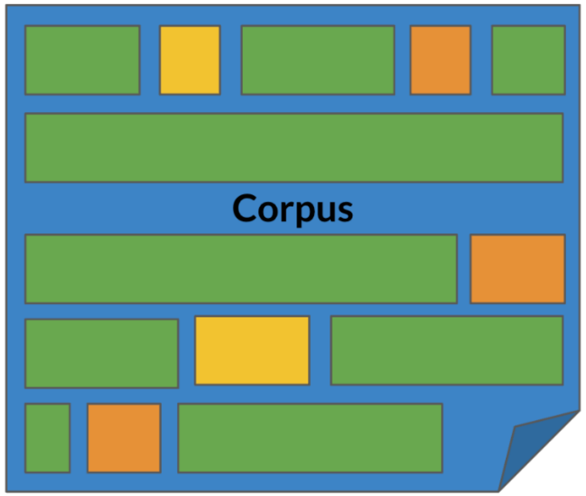


ML Models for NLP: Splitting the data

- Continuous text



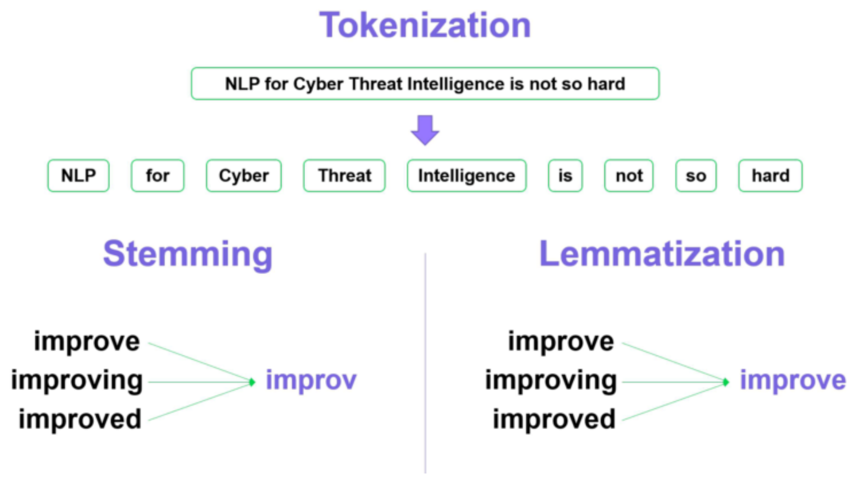
- Random short sequences



ML for NLP: From language to Vectors

Data Preparation

- Preprocessing & normlalization: lower case, special characters, stop words, stemming, tokenizing, …



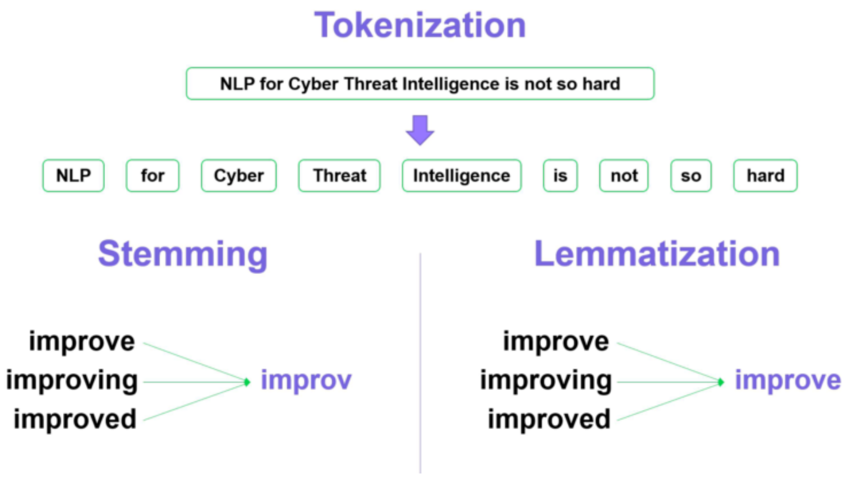
FRENCH INSTITUTES OF TECHNOLOGY

ML for NLP: From language to Vectors

Data Preparation

- Preprocessing & normlalization: lower case, special characters, stop words, stemming, tokenizing, …

- Numerical feature vectors: ML methods requires numerical features



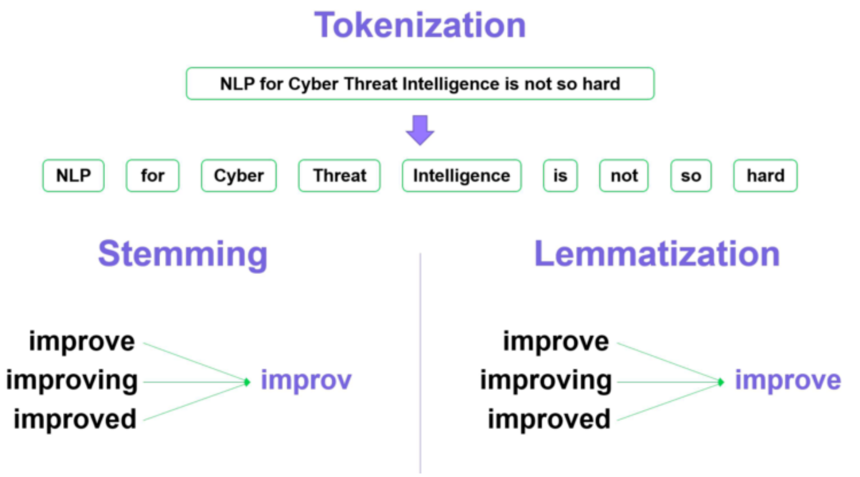
FRENCH INSTITUTES OF TECHNOLOGY

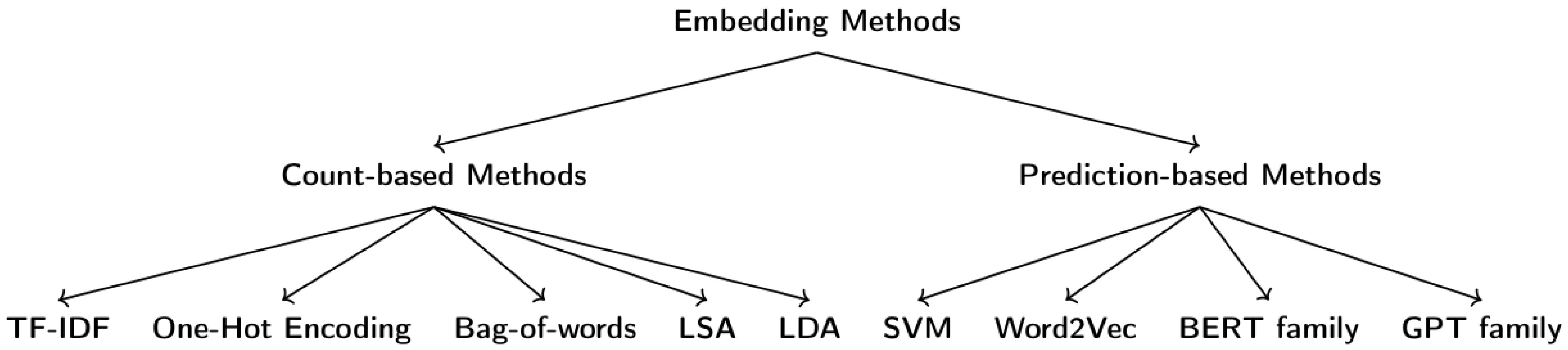
ML for NLP: From language to Vectors

Data Preparation

- Preprocessing & normlalization: lower case, special characters, stop words, stemming, tokenizing, …

- Numerical feature vectors: ML methods requires numerical features



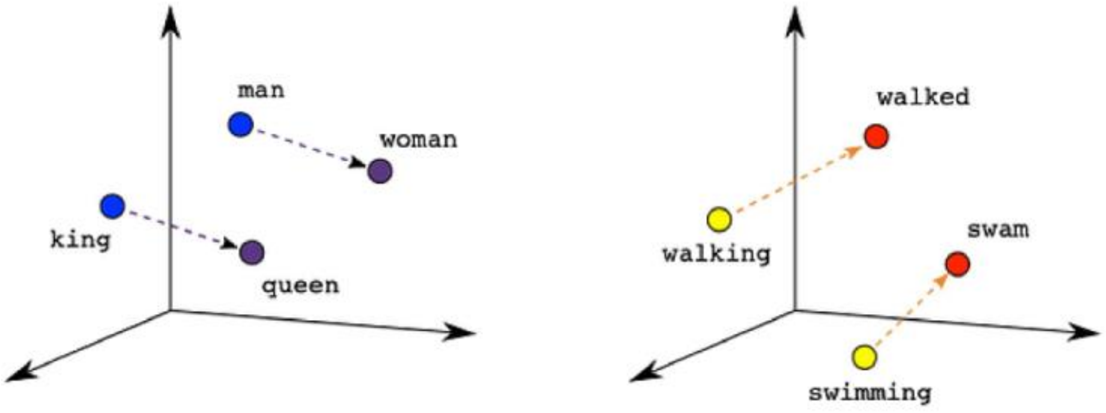


FRENCH INSTITUTES OF TECHNOLOGY

ML for NLP: From language to Vectors

Data Preparation

- Preprocessing & normlalization: lower case, special characters, stop words, stemming, tokenizing, …
- Numerical feature vectors: ML methods requires numerical features



Corpus

Document 1: "Cyber threats pose a risk to organizations."

Document 2: "Inadequate cyber defenses expose organizations to cyber risks."

Vocabulary (after lemmatization and sorting by alphabetical order)

[cyber, defense, expose, inadequate, organization, pose, risk, threats]

One-hot Encoding of Document 1 (with the Vocabulary)

Doc1\Voc	cyber	defense	expose	inadequate	organization	pose	risk	threats
Cyber	1	0	0	0	0	0	0	0
threats	0	0	0	0	0	0	0	1
pose	0	0	0	0	0	1	0	0
risk	0	0	0	0	0	0	1	0
organization	0	0	0	0	1	0	0	0

BoW(with the Vocabulary)

Doc\Voc	cyber	defense	expose	inadequate	organization	pose	risk	threats
Document 1	1	0	0	0	1	1	1	1
Document 2	2	1	1	1	1	0	1	1

TF-IDF of Document 2 (with the Corpus & Vocabulary)

	cyber	defense	expose	inadequate	organization	pose	risk	threats
# docs w/ the token	2	1	1	1	2	1	2	1
IDF	$\log(\frac{2}{2})$	$\log(\frac{2}{1})$	$\log(\frac{2}{1})$	$\log(\frac{2}{1})$	$\log(\frac{2}{2})$	$\log(\frac{2}{1})$	$\log(\frac{2}{2})$	$\log(\frac{2}{1})$
TF(*,Doc2)	$\frac{2}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$
TF - IDF(*,Doc2)	0	0.043	0.043	0.043	0	0.043	0	0

ML Models for NLP: Classification

- *Input:*
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \dots, c_K\}$
  - A training set of n hand-labeled documents $(\mathrm{d}_1, \mathrm{c}_1), \dots, (d_n, c_n)$
- *Output:*
  - a learned classifier $\hat{f}: d \rightarrow c \in C$

FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Classification

- *Input:*
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_K\}$
  - A training set of n hand-labeled documents $(\mathrm{d}_1, \mathrm{c}_1), \ldots, (d_n, c_n)$
- *Output:*
  - a learned classifier $\hat{f}: d \to c \in C$
  - Actually $\hat{f}: d \to (\pi_1, \ldots, \pi_k) \in [0,1]^K$

FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Classification

- Any kind of classifier can be used:

  - Naïve Bayes
  - Logistic regression
  - Support-vector machines
  - Random Forest
  - k-Nearest Neighbors
  - …

FRENCH INSTITUTES OF TECHNOLOGY

Metrics for Classification:

- Usual classification metrics can be used:

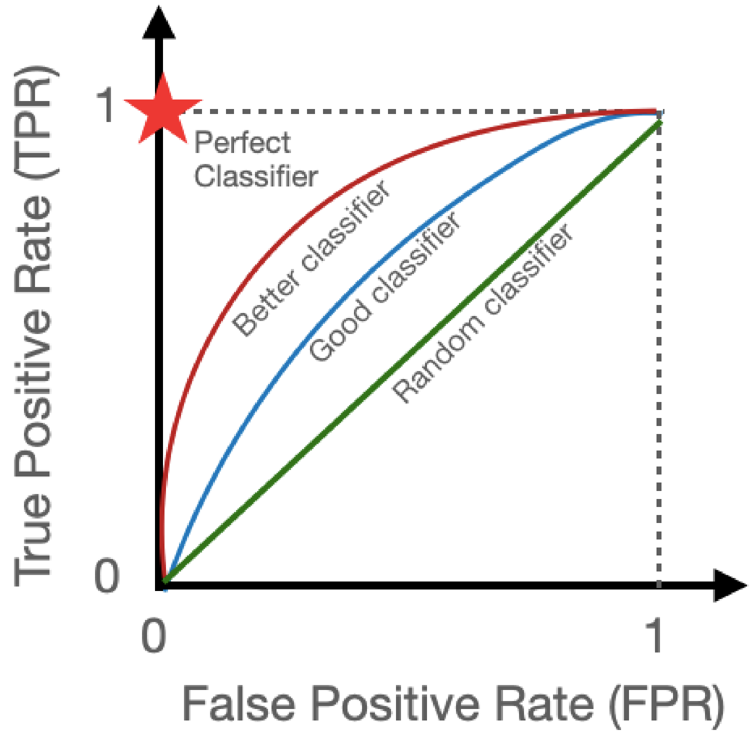
		ACTUAL: Positive	ACTUAL: Negative
PREDICTED	Positive	True Positive (TP) 5	False Positive (FP) 10
	Negative	False Negative (FN) 15	True Negative (TN) 70

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 0.75$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.25$$

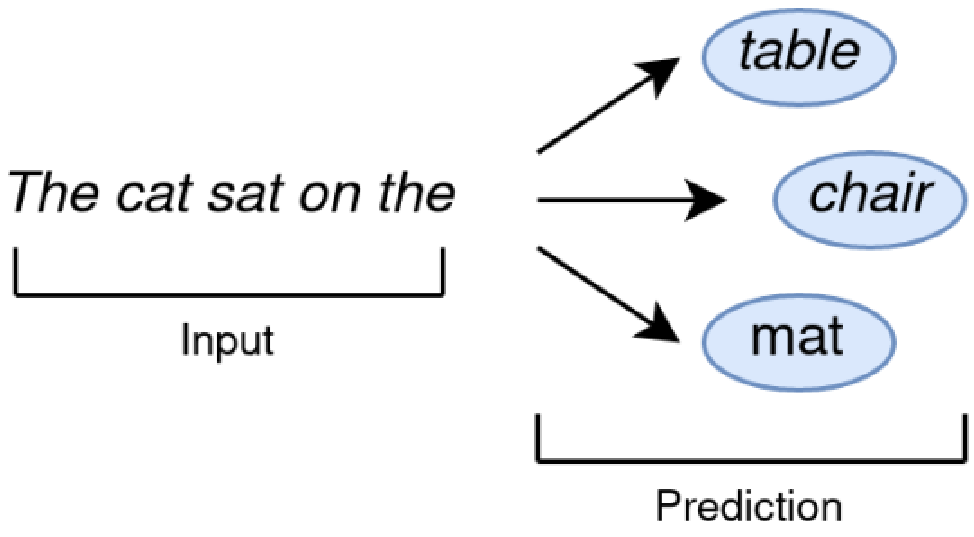
$$\text{Precision} = \frac{TP}{TP + FP} = 0.33$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = 0.28$$

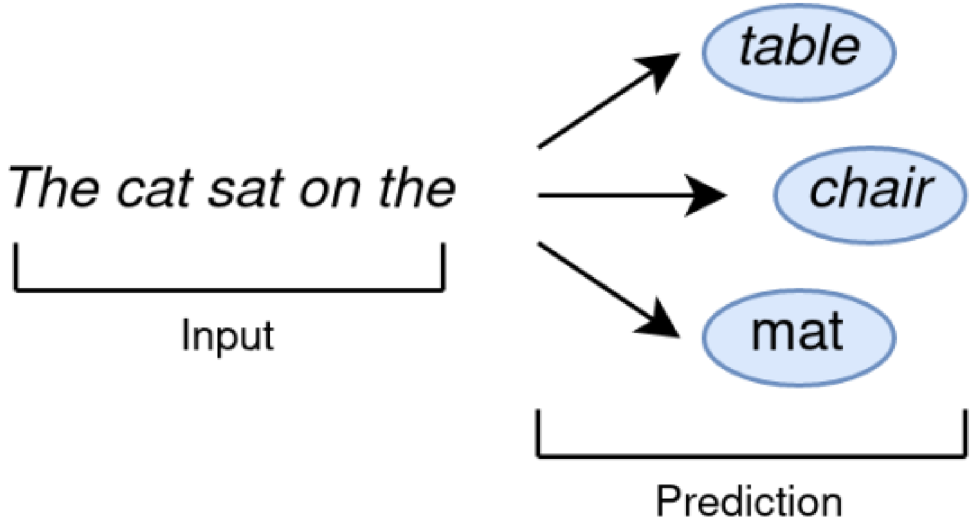


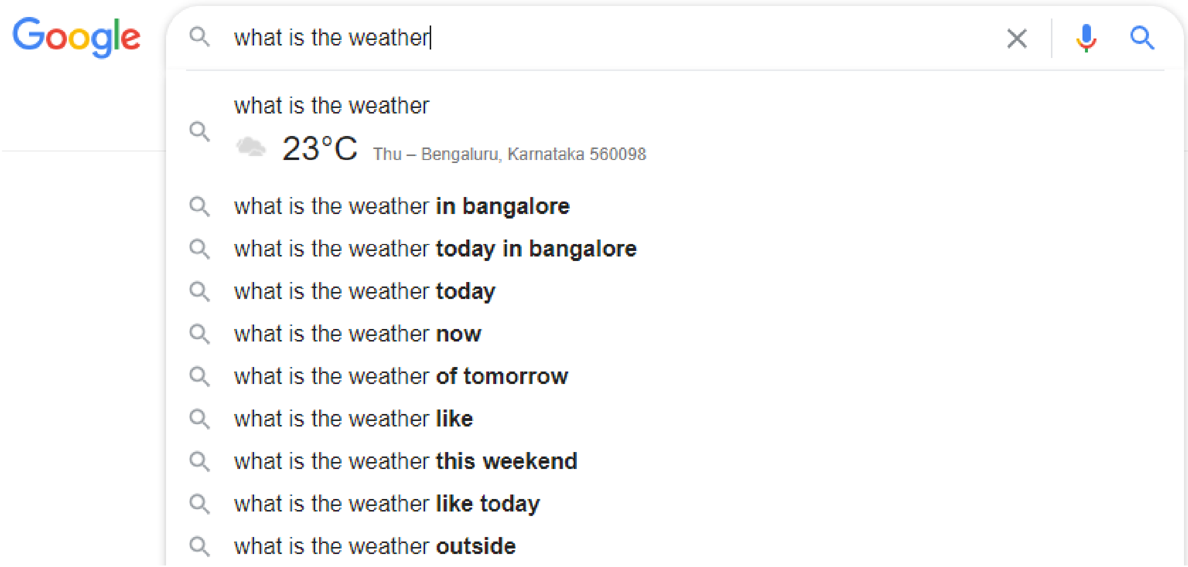
FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Language Models



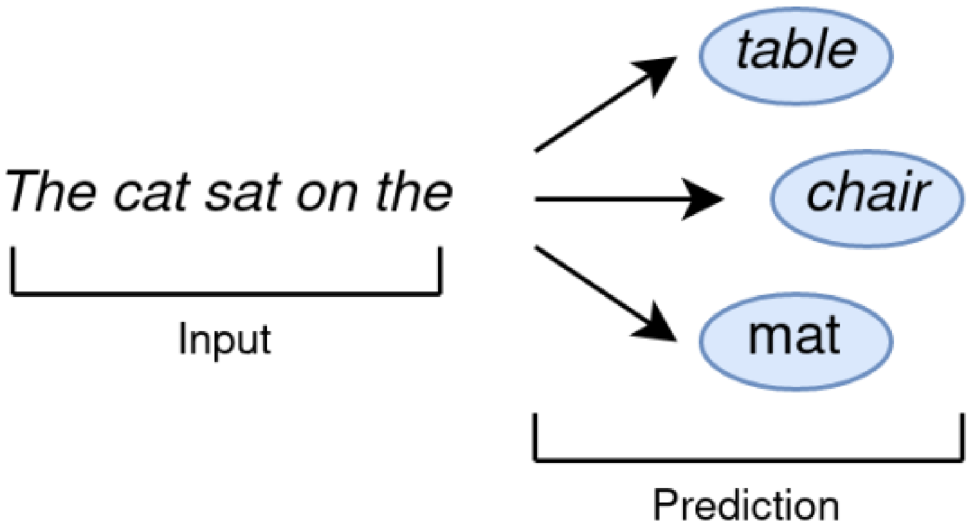
ML Models for NLP: Language Models





FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Language Models



$P(\text{table} \mid \text{The cat sat on the})$

$P(\text{chair} \mid \text{The cat sat on the})$

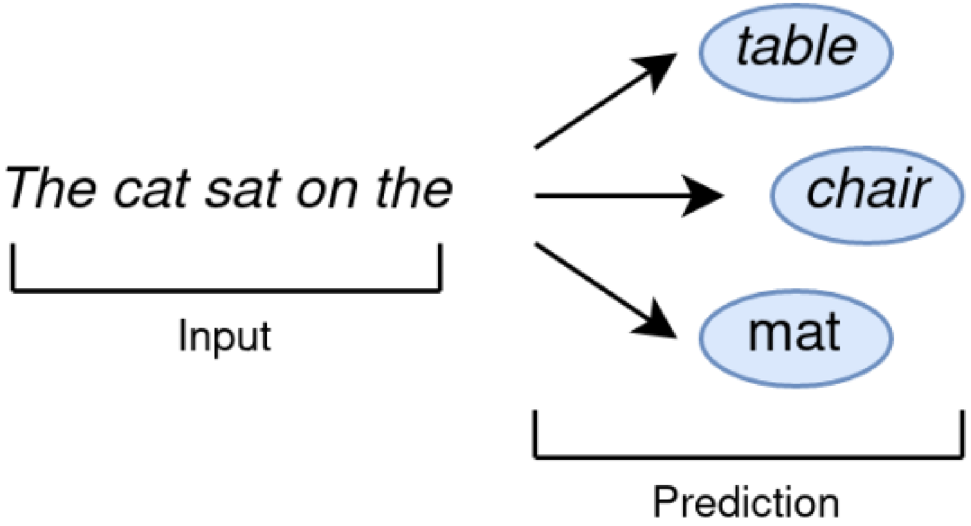
$P(\text{mat} \mid \text{The cat sat on the})$

…

$$P(\text{table} \mid \text{The cat sat on the}) = ?$$

FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Language Models



$P(\text{table} \mid \text{The cat sat on the})$

$P(\text{chair} \mid \text{The cat sat on the})$

$P(\text{mat} \mid \text{The cat sat on the})$

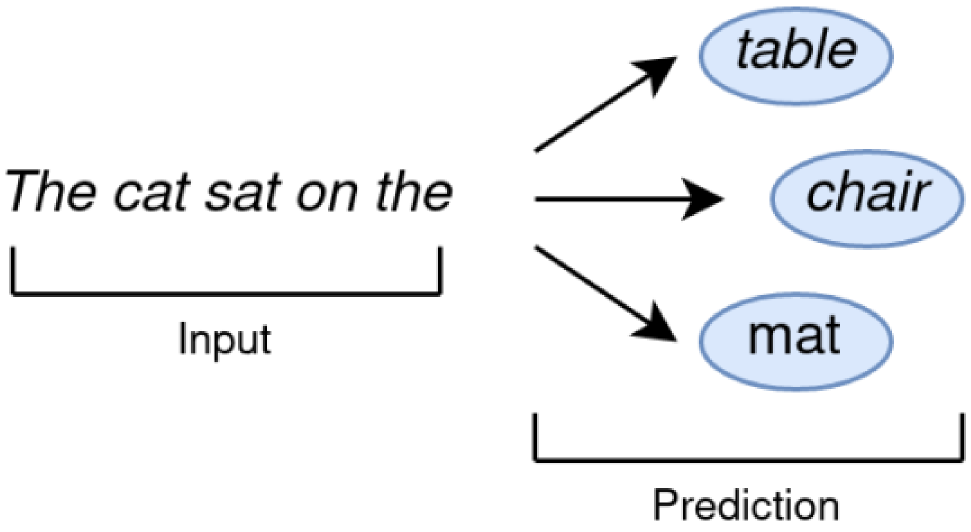
…

$$P(\text{table} \mid \text{The cat sat on the}) = ?$$

To be estimated from data

FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Language Models



$P(\text{table} \mid \text{The cat sat on the})$

$P(\text{chair} \mid \text{The cat sat on the})$

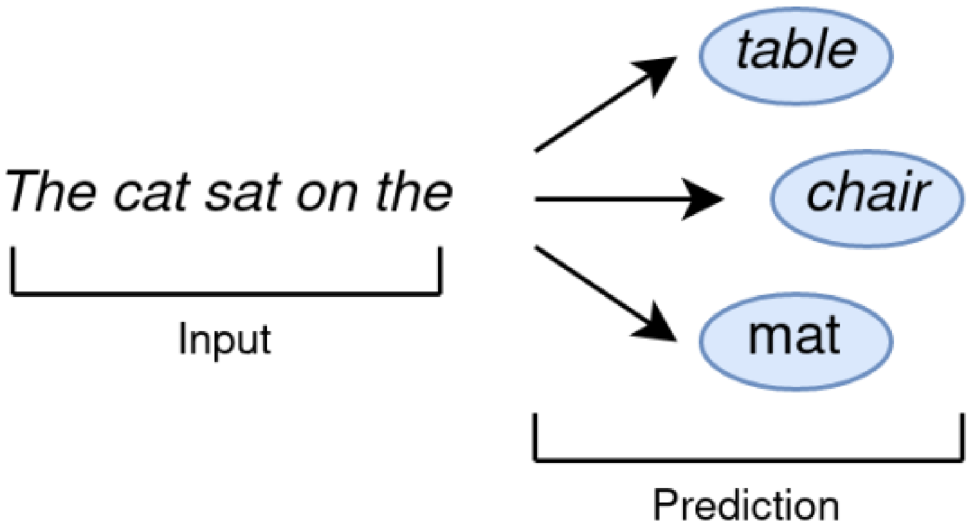
$P(\text{mat} \mid \text{The cat sat on the})$

…

$$P(\text{table} \mid \text{The cat sat on the}) = \frac{\#\,(The\ cat\ sat\ on\ the\ table)}{\#\,(The\ cat\ sat\ on\ the)}$$

FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Language Models



$P(\text{table} \mid \text{The cat sat on the})$

$P(\text{chair} \mid \text{The cat sat on the})$

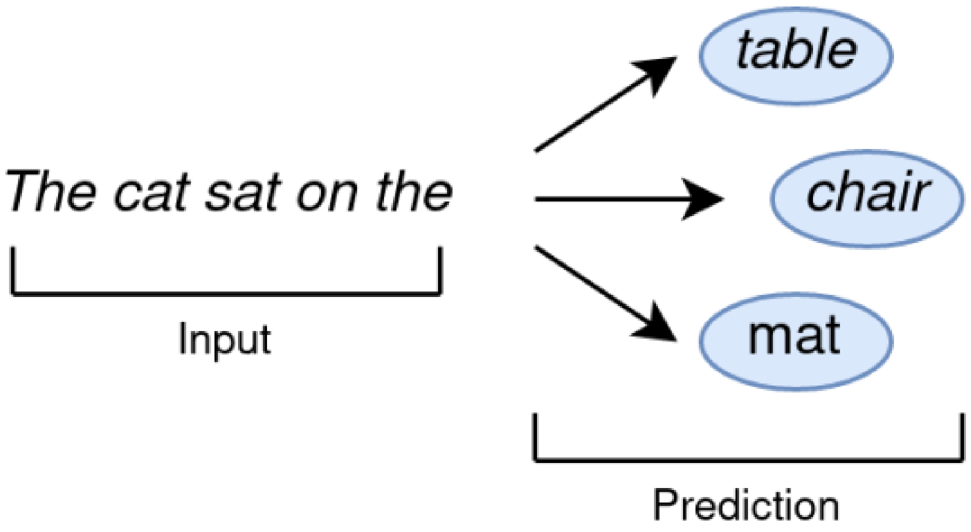
$P(\text{mat} \mid \text{The cat sat on the})$

…

$$P(\text{table} \mid \text{The cat sat on the}) = \frac{\#\,(The\ cat\ sat\ on\ the\ table)}{\#\,(The\ cat\ sat\ on\ the)}$$

We'll hardly see enough data for estimating these

ML Models for NLP: Language Models



$P(\text{table} \mid \text{The cat sat on the})$

$P(\text{chair} \mid \text{The cat sat on the})$

$P(\text{mat} \mid \text{The cat sat on the})$

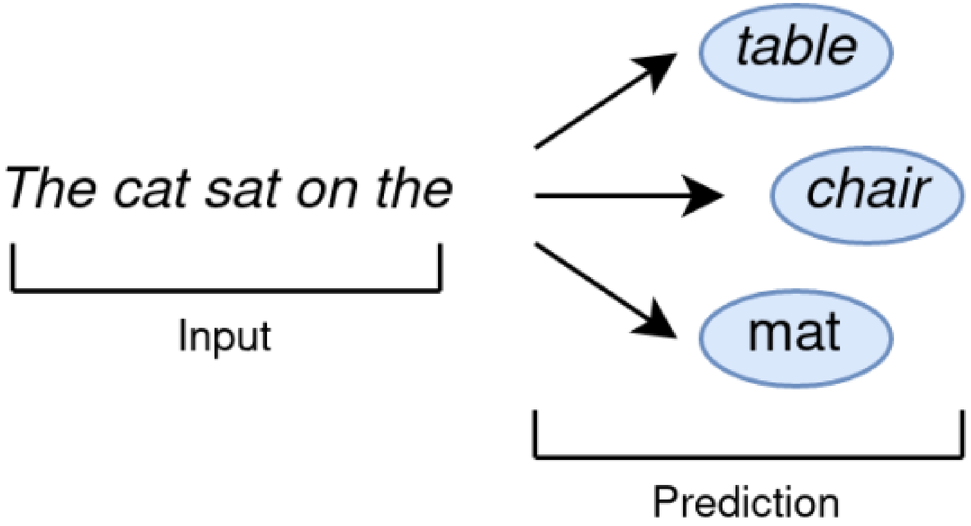
…

$$P(\text{table} \mid \text{The cat sat on the}) \approx \begin{cases} P(table)\ or \\ P(table \mid the)\ or \\ P(table \mid on\ the)\ or \\ P(table \mid sat\ on\ the)\ or \\ \ldots \\ P(w_i \mid w_1 w_2 \ldots w_{i-1}) \end{cases}$$

Markov Assumption

FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Language Models



$P(\text{table} \mid \text{The cat sat on the})$

$P(\text{chair} \mid \text{The cat sat on the})$

$P(\text{mat} \mid \text{The cat sat on the})$

...

$$P(\text{table} \mid \text{The cat sat on the}) \approx \begin{cases} P(table)\ or \\ P(table \mid the)\ or \\ P(table \mid on\ the)\ or \\ P(table \mid sat\ on\ the)\ or \\ \dots \\ P(w_i \mid w_1 w_2 \dots w_{i-1}) \end{cases}$$

Markov Assumption

Increasing context and complexity

ML Models for NLP: Language Models

Training corpus

<s> I am a Student </s>
<s> Student I am </s>
<s> I do like Machine Learning </s>

$$P(w_i \mid w_{i-1}) = \frac{\#\,(w_{i-1}, w_i)}{\#\,(w_{i-1})}$$

$P(I \mid <s>) =$

$P(Student \mid <s>) =$

$P(am \mid I) =$

$P(</s> \mid Student) =$

$P(Student \mid am) =$

$P(do \mid I) =$

FRENCH INSTITUTES OF TECHNOLOGY

ML Models for NLP: Language Models

$$P(w_i \mid w_{i-1}) = \frac{\#\,(w_{i-1}, w_i)}{\#\,(w_{i-1})}$$

Training corpus

<s> I am a Student </s>
<s> Student I am </s>
<s> I do like Machine Learning </s>

$$P(I \mid <s>) = \frac{2}{3} \qquad P(Student \mid <s>) = \frac{1}{3} \qquad P(am \mid I) = \frac{2}{3}$$

$$P(</s> \mid Student) = \frac{1}{2} \qquad P(Student \mid am) = \frac{1}{2} \qquad P(do \mid I) = \frac{1}{3}$$

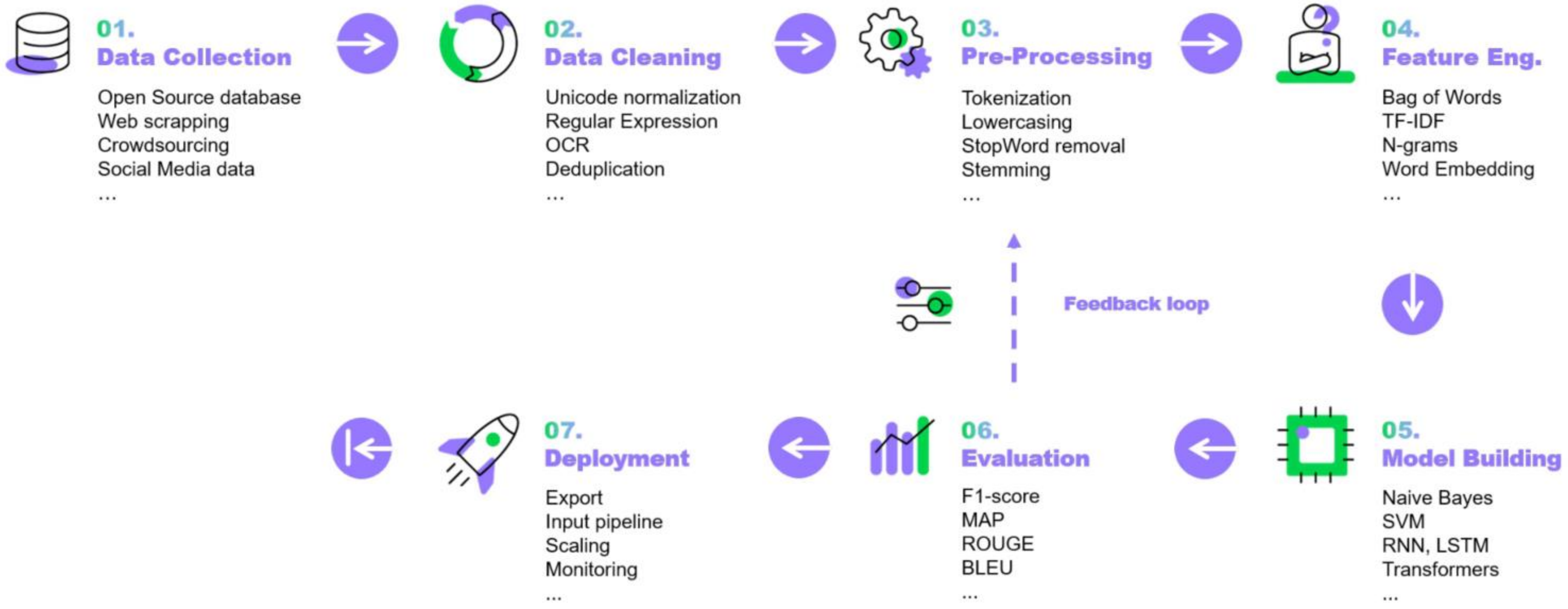
FRENCH INSTITUTES OF TECHNOLOGY

Metrics for text generation

- **Perplexity**: Measures how well the model predicts a (test) sequence (used for language models)
- **BLEU (Bilingual Evaluation Understudy)**: Measures n-gram overlap between generated and reference text (used for machine translation).
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**: Focuses on content overlap (used for summarization).
- …

FRENCH INSTITUTES OF TECHNOLOGY

NLP Workflow



FRENCH INSTITUTES OF TECHNOLOGY

FRENCH INSTITUTES OF TECHNOLOGY