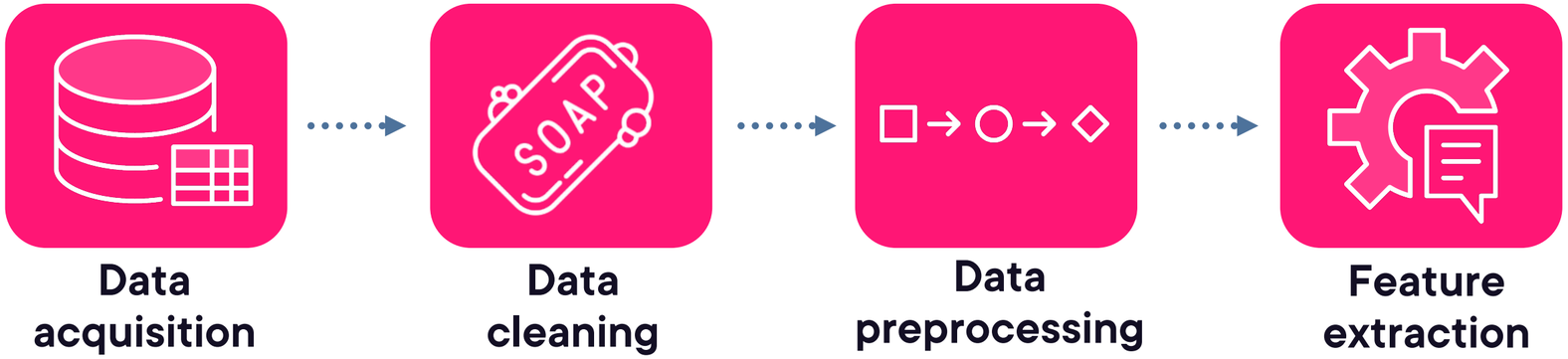
Exploring the NLP Pipeline

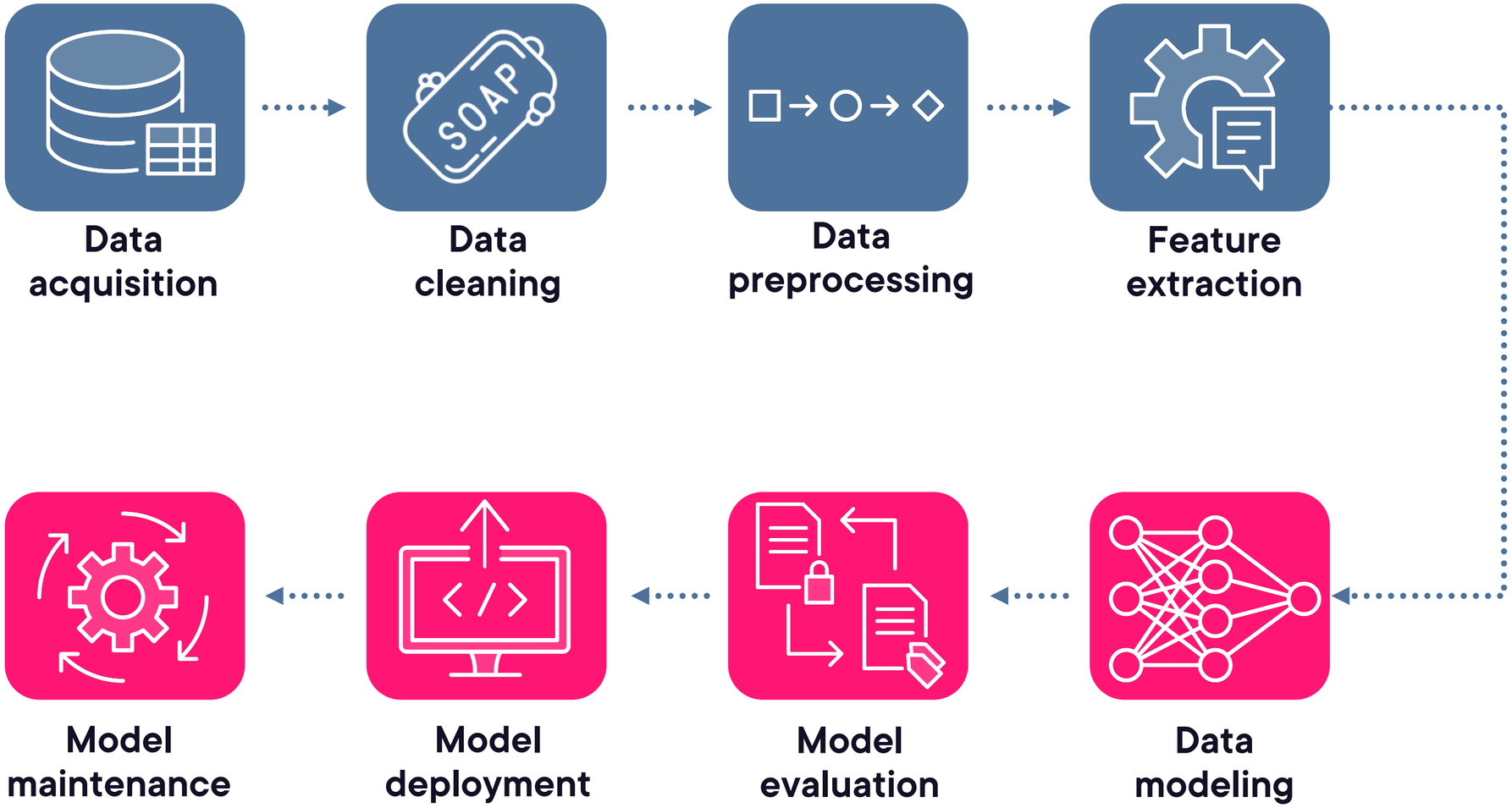
Mihaela Danci
Data Analyst

linkedin.com/in/mihaela-danci/

NLP Pipeline



NLP Pipeline



Overview

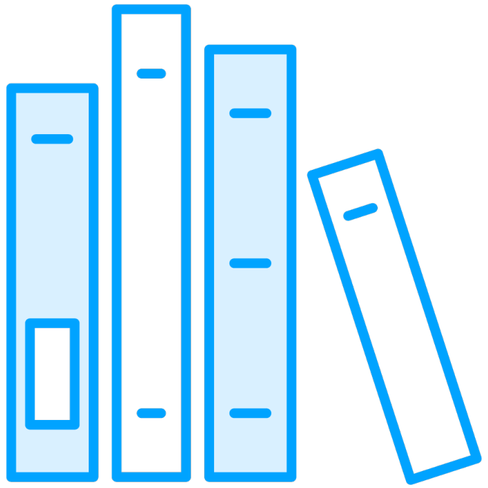
Explore each NLP pipeline stage

Hands-on learning

- Classification problem
- Categorize news articles

Data Acquisition

Data Acquisition



Digital libraries

Websites

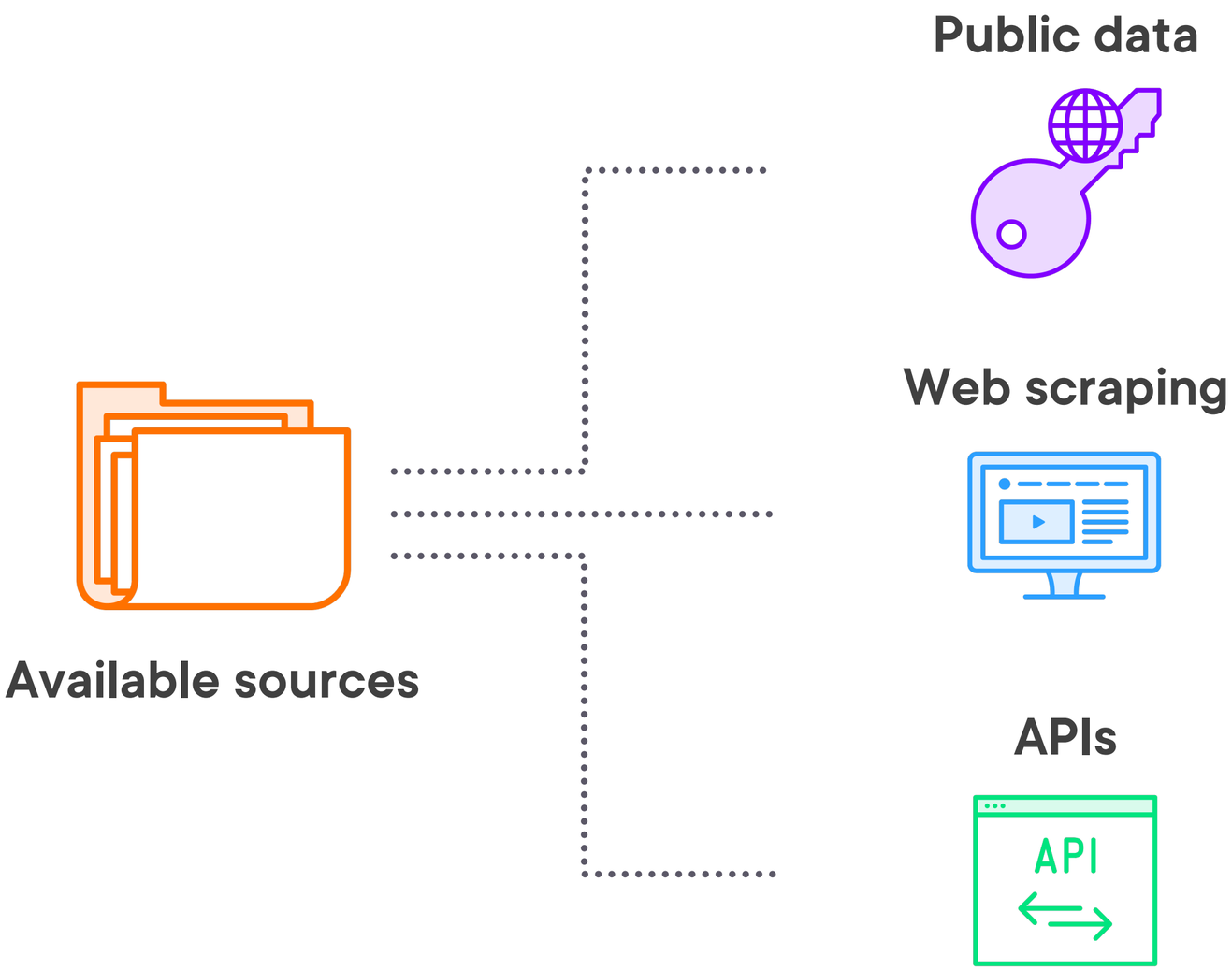
Media transcripts

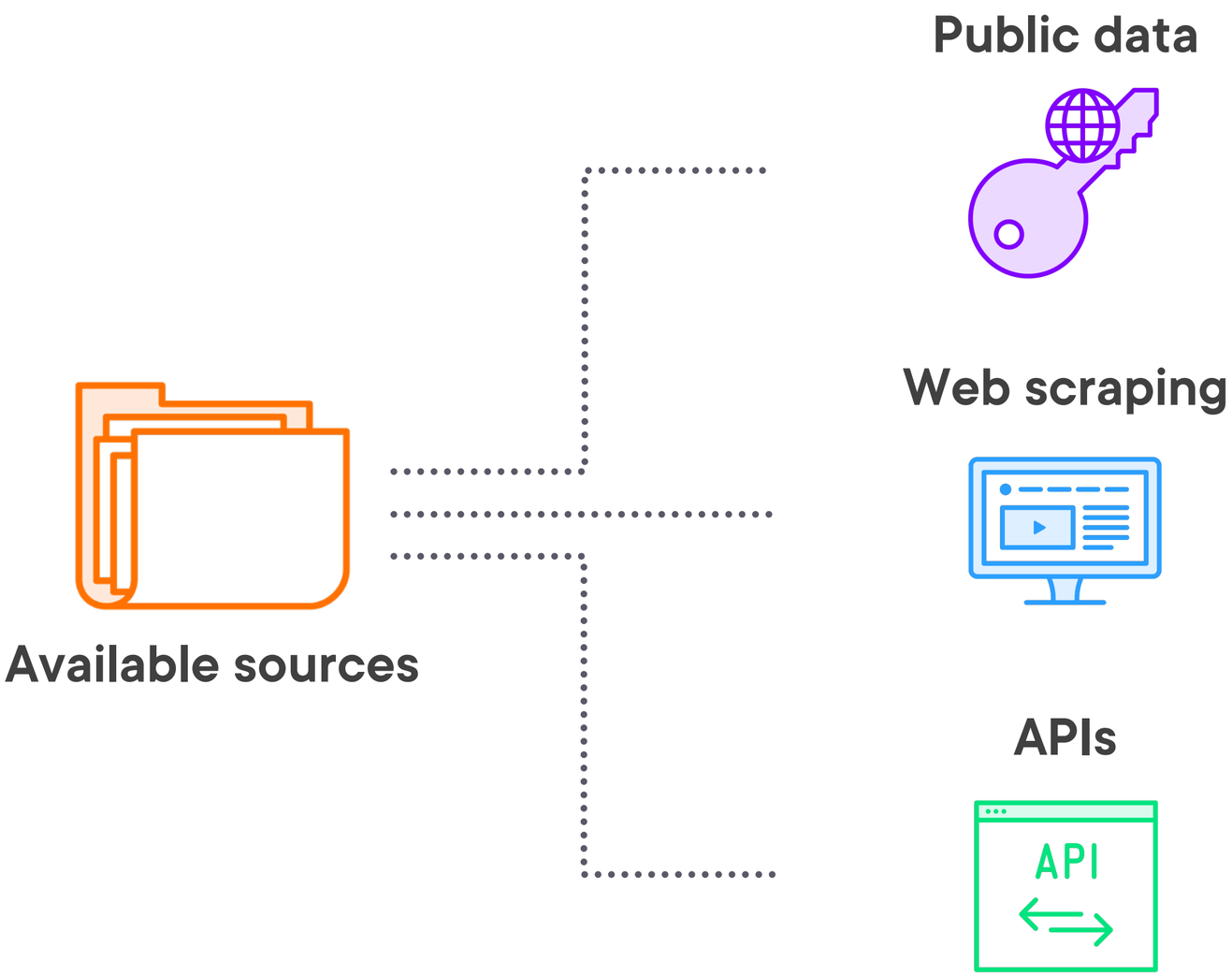
Data Acquisition

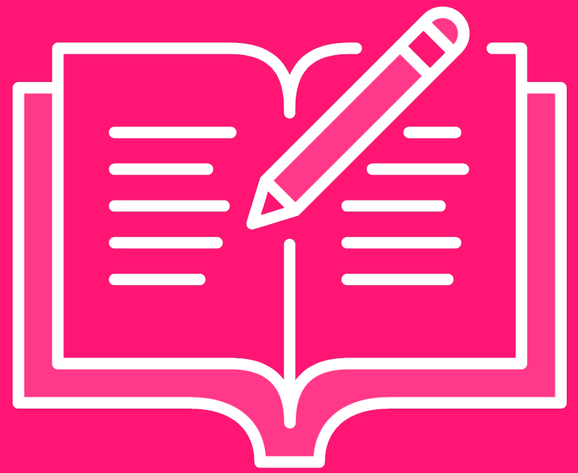
Volume

Variety

This is often not the case.


Public data
Web scraping
Available sources
APIs
API

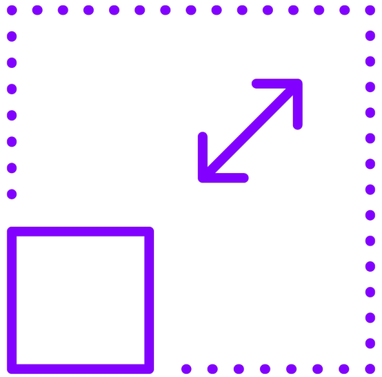




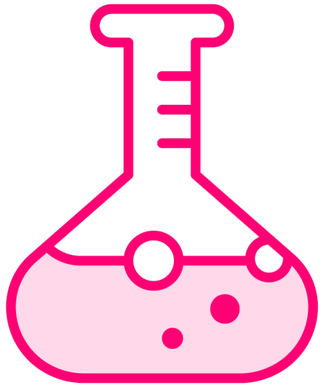
Manual collection

Data Acquisition

Use available sources

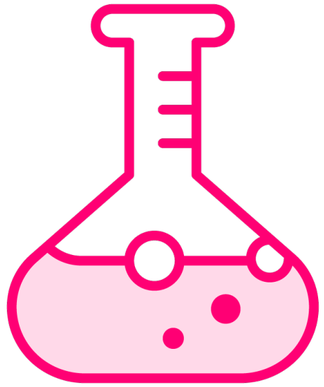


Data augmentation



Synthetic data creation

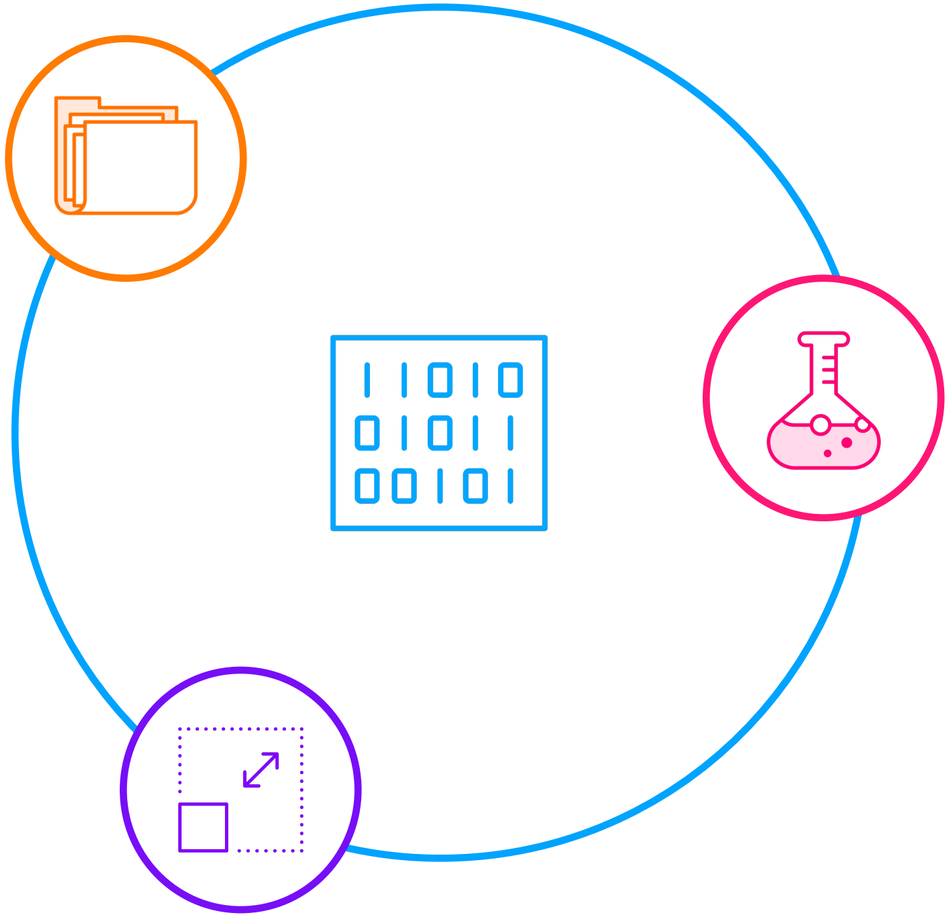
Data Acquisition



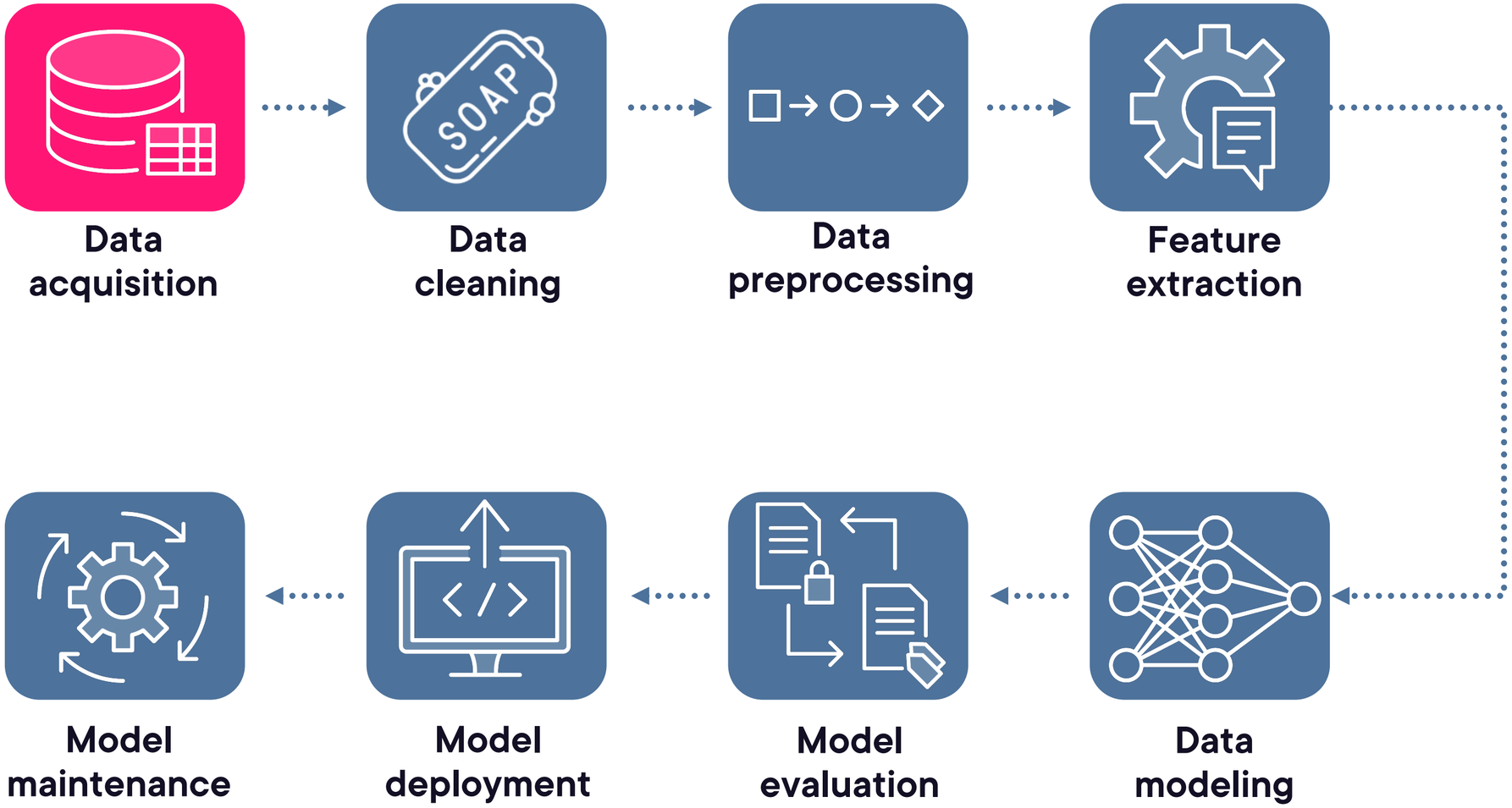
Synthetic data creation

- Privacy protection
- Minimize bias
- Large volumes

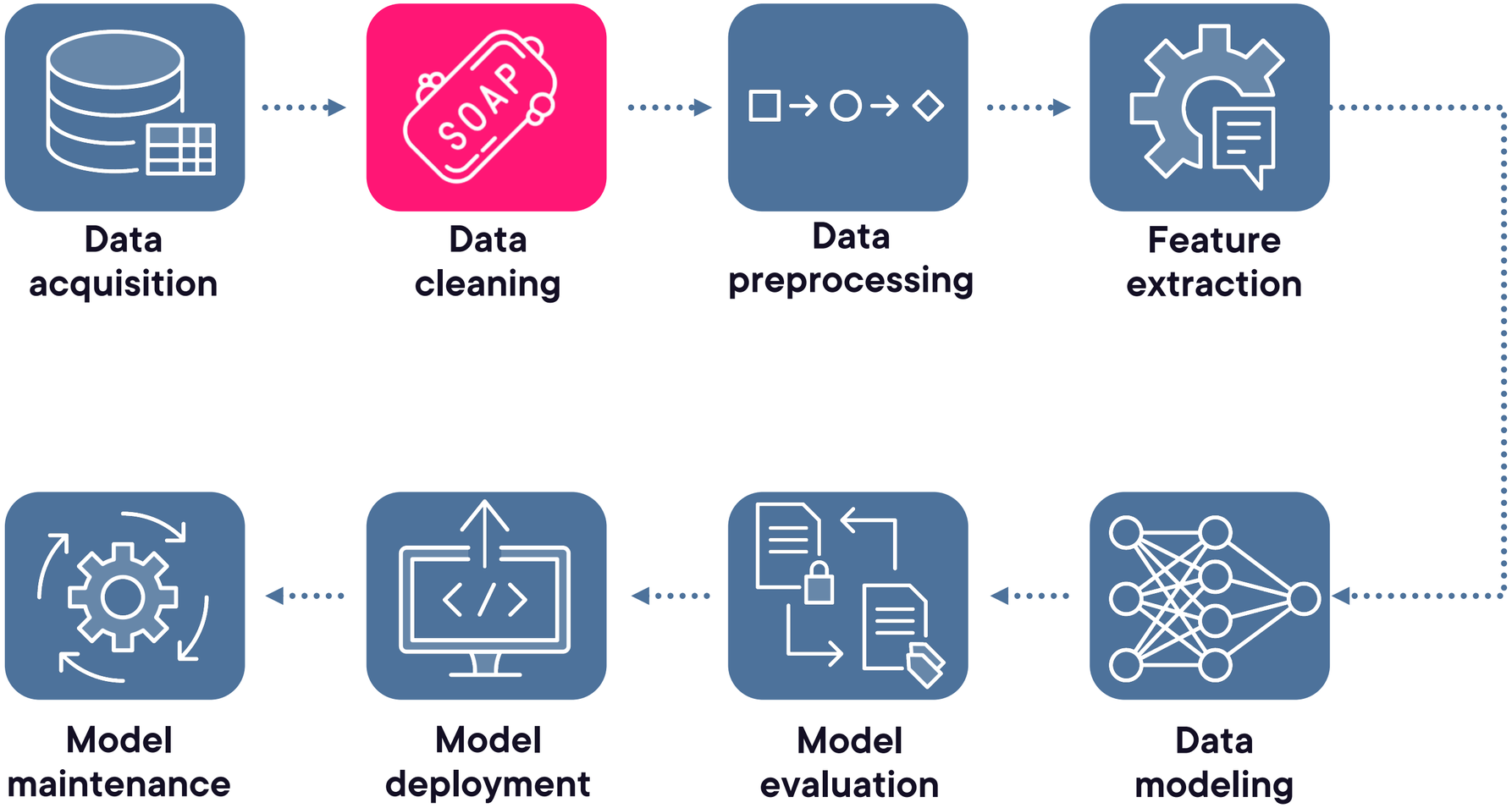
Data Acquisition



NLP Pipeline



NLP Pipeline



Demo

Import the data into Jupyter Notebook

Dataset

- A collection of newsgroup documents

Objective

- Build a model capable of classifying newsgroup posts by topic

Data Cleaning and Preprocessing

Data

Unneeded formatting

Mistakes

Irrelevant details

Step 1: Data Cleaning and Standardization

Common Actions

Addressing issues like spelling, grammar, typographical errors, and whitespaces

Standardizing capitalization, date formats, and encoding

Discarding irrelevant elements, like metadata or non-textual content

Step 2: Deconstruct Natural Language

Deconstruct Natural Language

Tokenization

Remove stop words

Stemming

Lemmatization

POS tagging

NER

Demo

Clean and preprocess the articles

- Remove headers, whitespaces, and punctuation
- Perform tokenization, and lemmatization

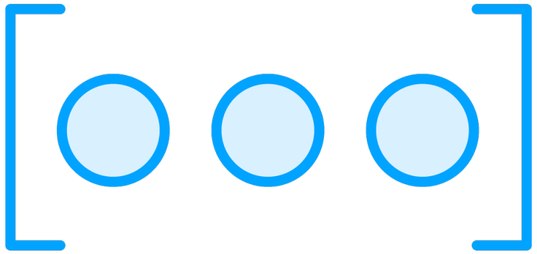
Feature Extraction

Feature Extraction

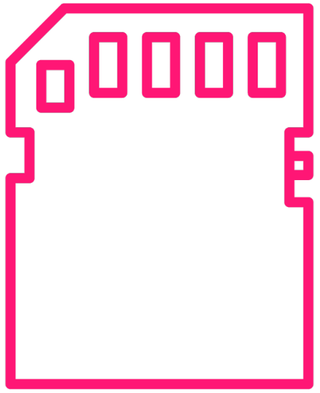
Feature Extraction Techniques

	One-hot encoding			
Sentence:	I	have	a	cat
Vector:	[1 0 0 0]	[0 1 0 0]	[0 0 1 0]	[0 0 0 1]

Drawbacks



Sparsity



Computationally inefficient

Feature Extraction Techniques

I have a cat. My cat likes to play.

Bag-of-words

”cat”: 2

“I” : 1

“a”: 1

“my”: 1

“likes” : 1

“to”: 1

“play”: 1

The more frequent a word is the more important it might be to the text

Drawbacks

Cat sat on mat.

Mat sat on cat.

These sentences appear the same in the model.

Feature Extraction Techniques

One-hot encoding

Bag-of-words

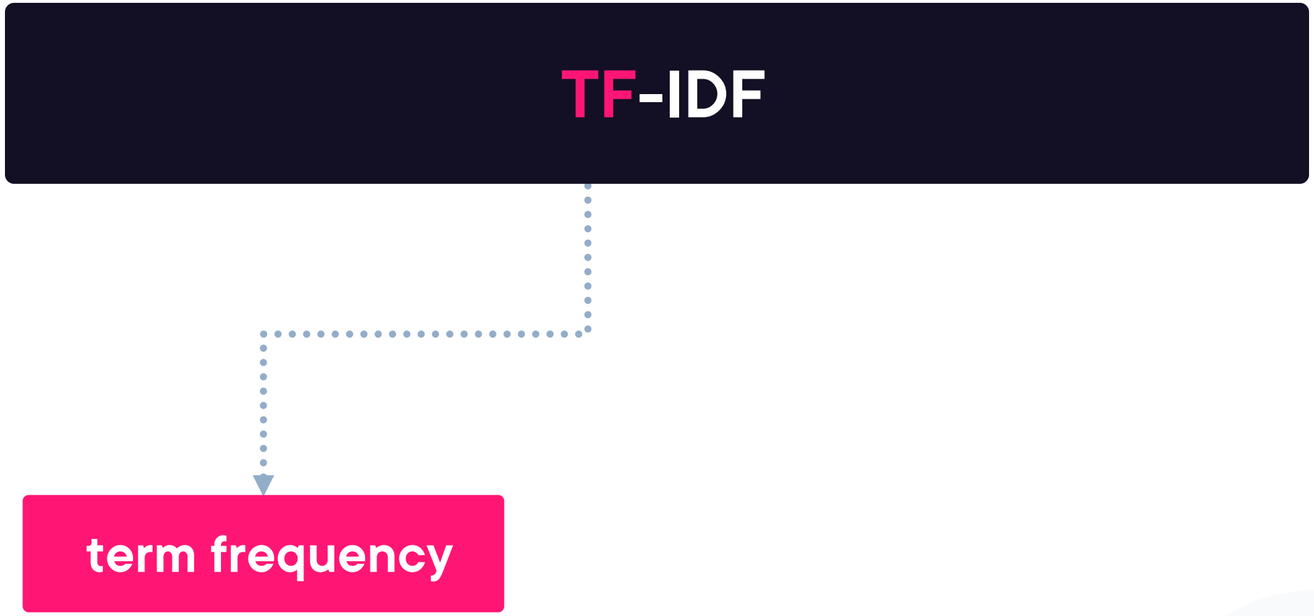
Bag-of-n-grams

TF-IDF

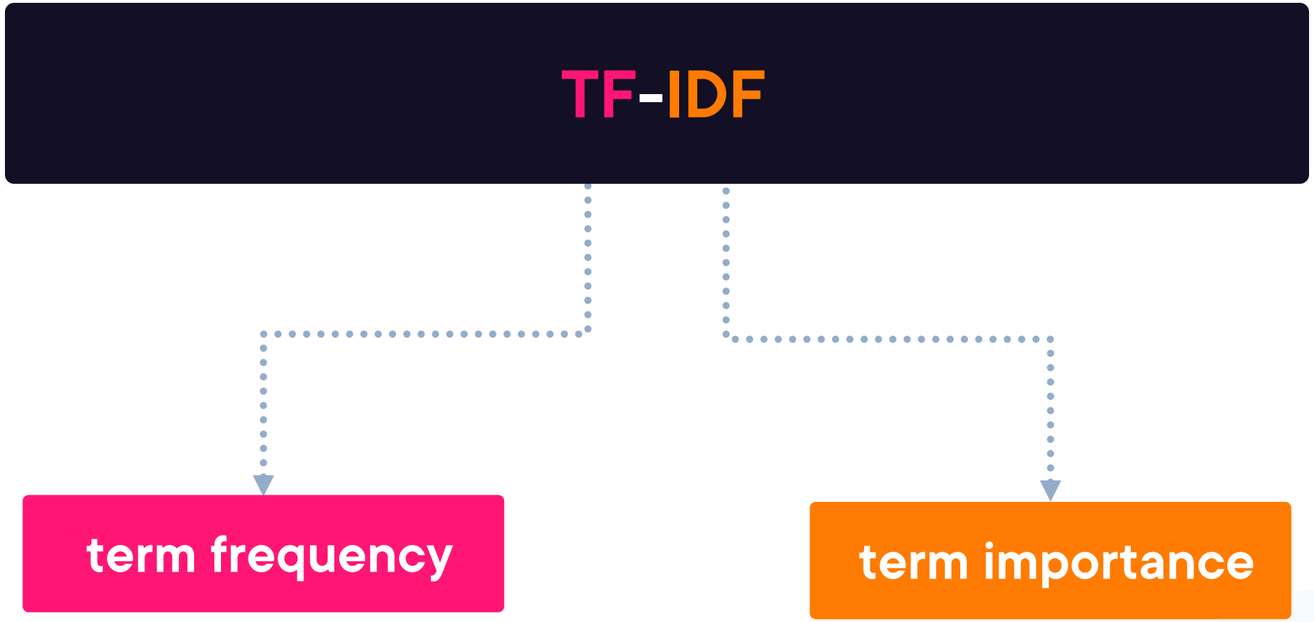
Feature Extraction Techniques

TF-IDF

Feature Extraction Techniques



Feature Extraction Techniques



TF-IDF reflects a word's importance in a document relative to a collection of documents.

TF-IDF

A higher TF-IDF score suggests greater relevance

IDF suggests rarity

Common words are less informative

Feature Extraction Techniques

One-hot encoding

Bag-of-words

Bag-of-n-grams

TF-IDF

Word embeddings

Word Embeddings

$$\begin{bmatrix} 4 & 7 \\ 1 & 5 \end{bmatrix}$$

Dense vector

Large corpora

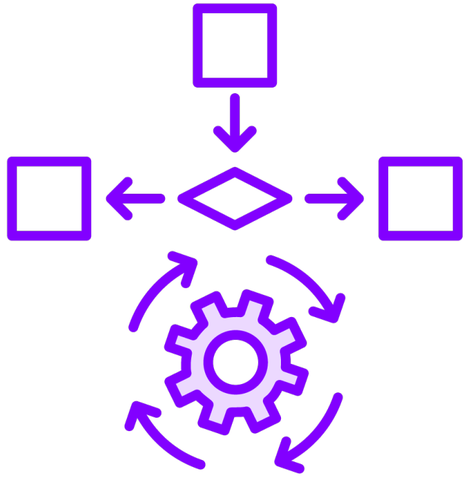
Demo

Convert the preprocessed text into a numerical format

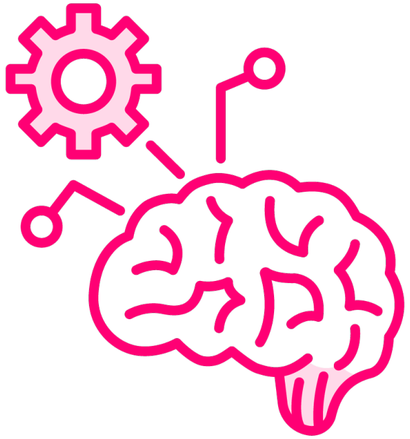
- Bag-of-words
- TF-IDF

Data Modeling and Evaluation

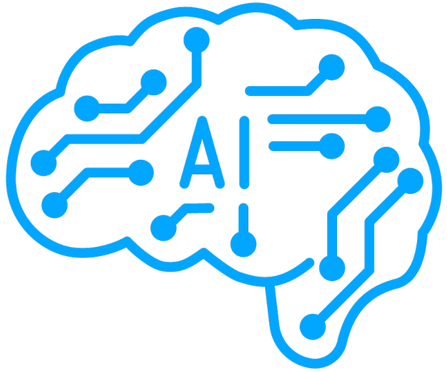
Natural Language



Rule-based

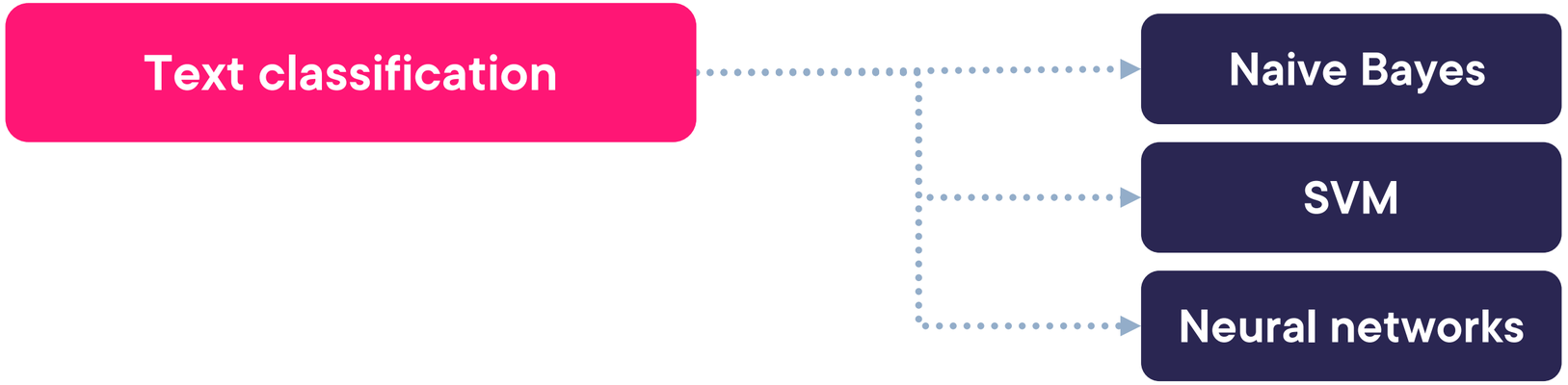


Machine learning

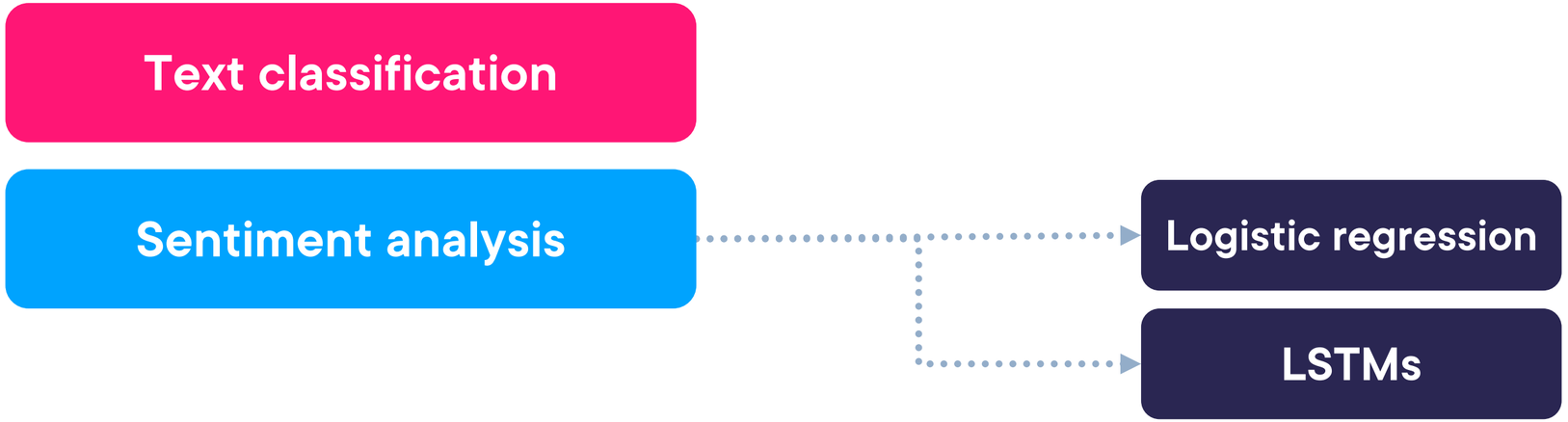


Deep learning

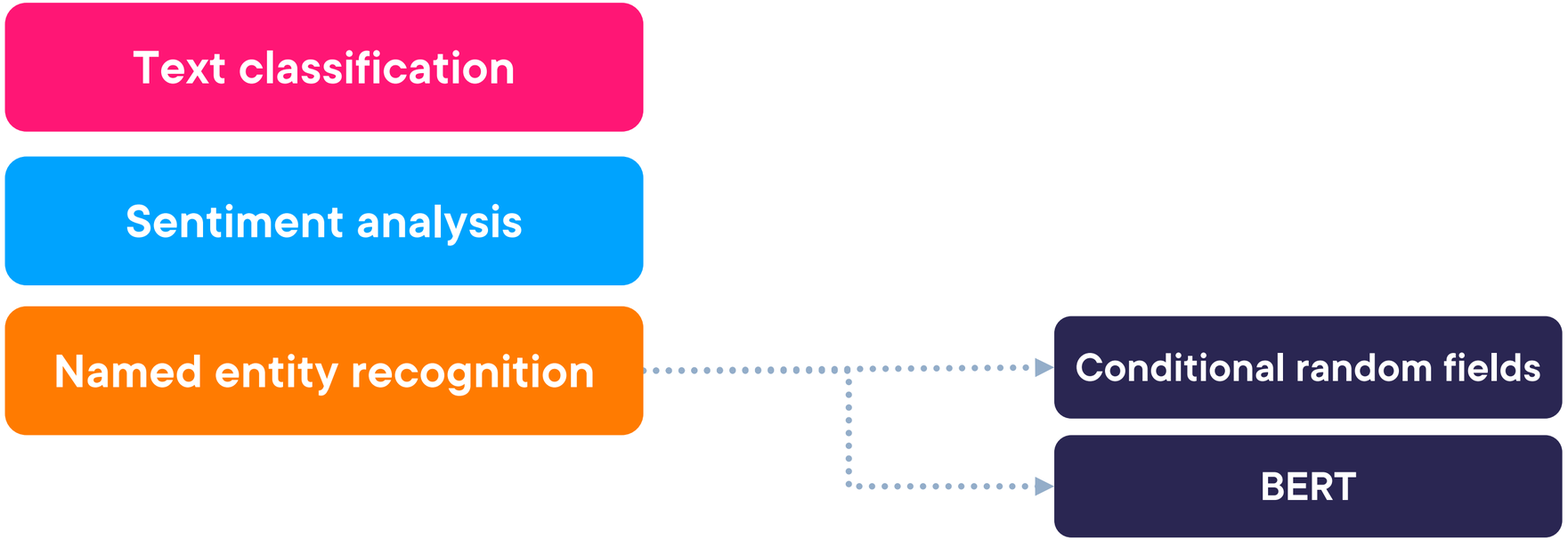
Models



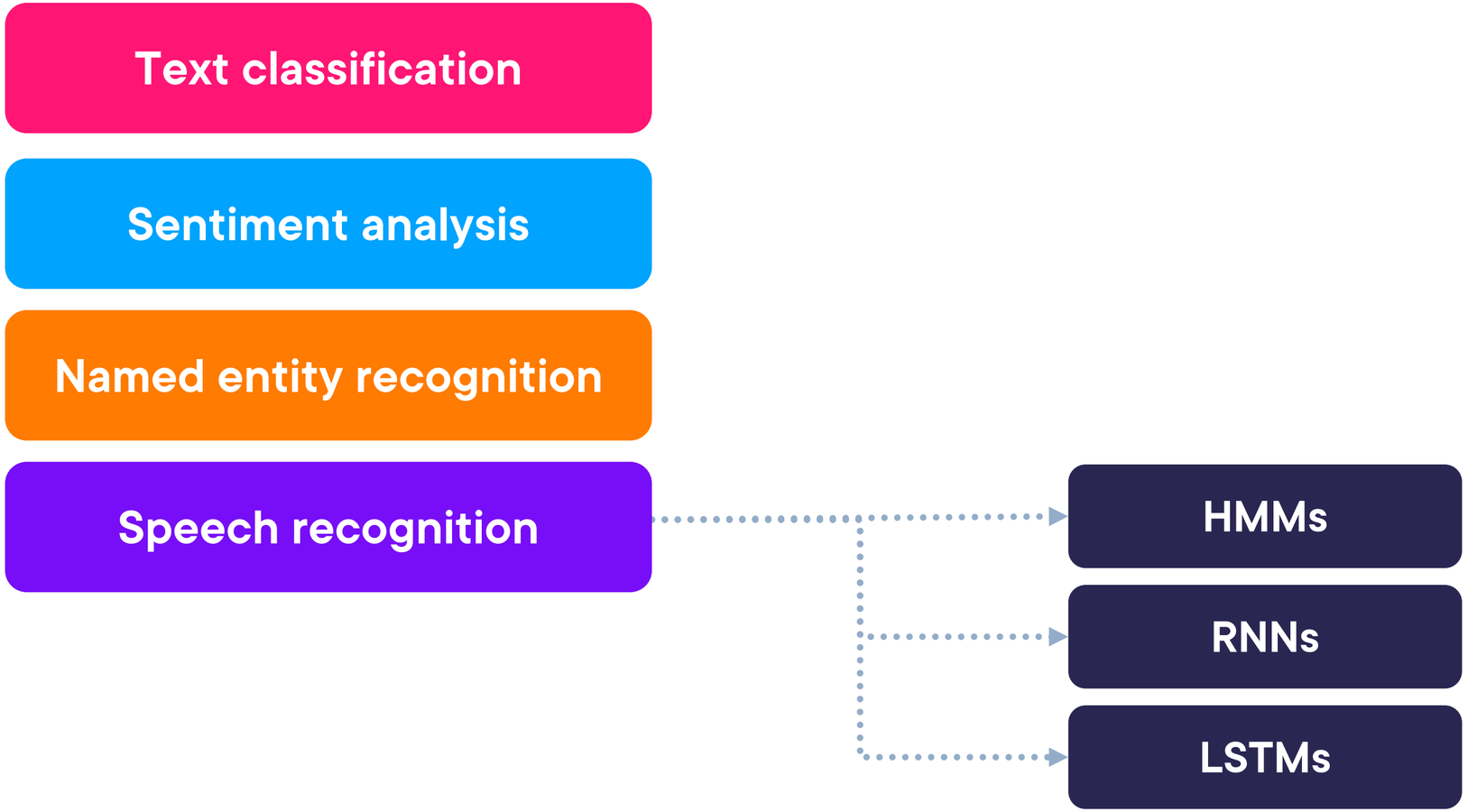
Models


Text classification
Sentiment analysis
Logistic regression
LSTMs

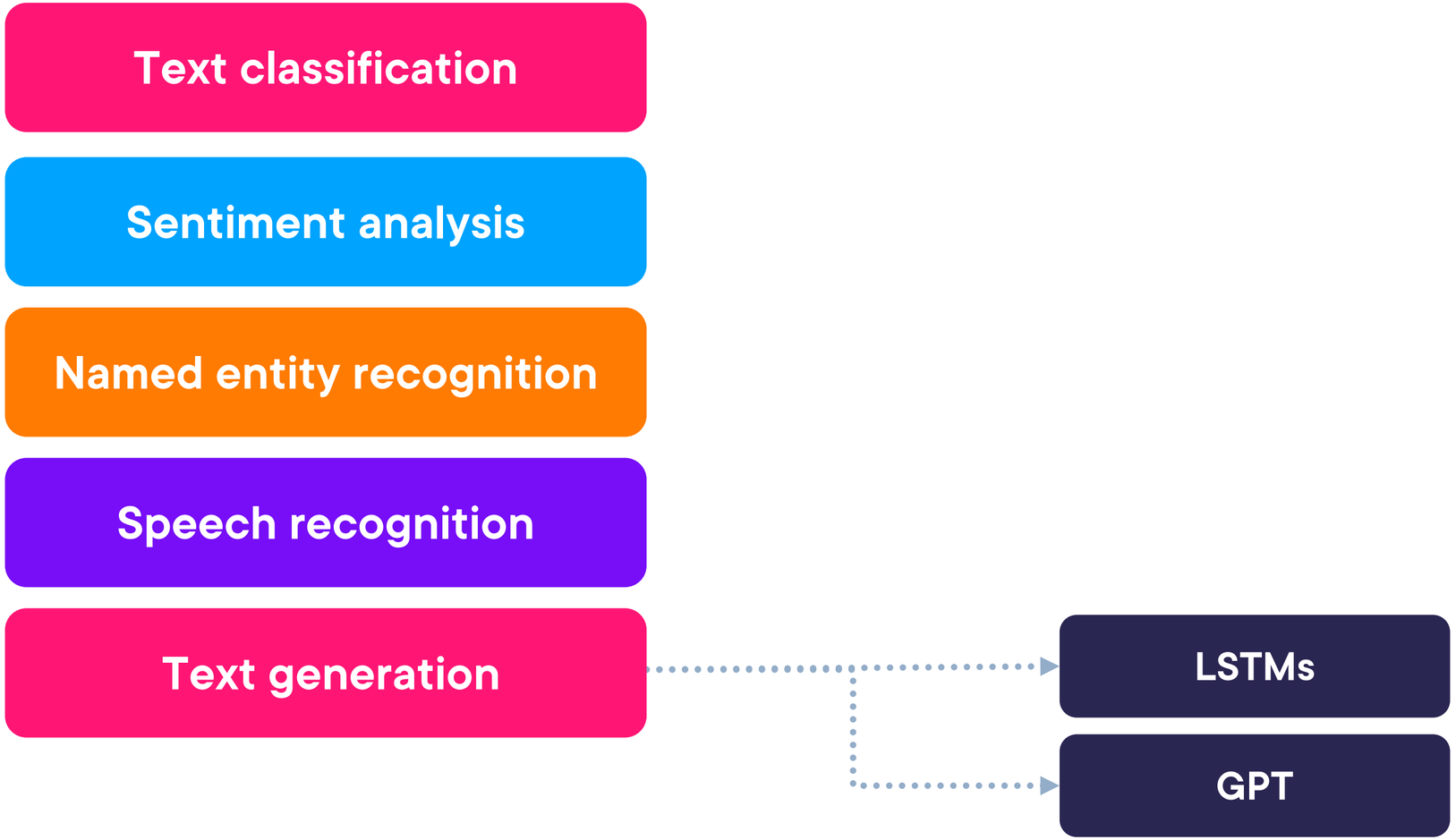
Models


Text classification
Sentiment analysis
Named entity recognition
Conditional random fields
BERT

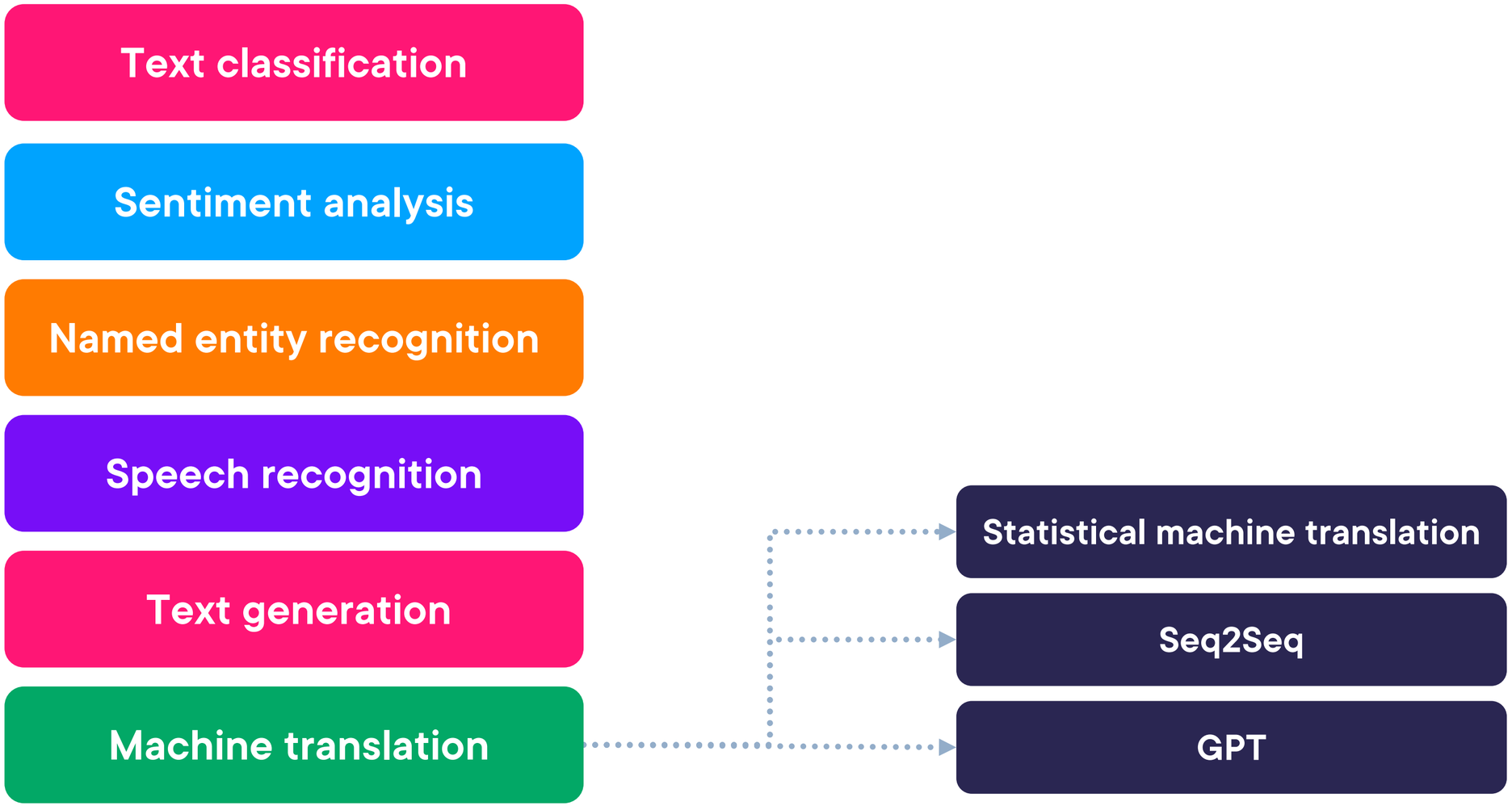
Models


Text classification
Sentiment analysis
Named entity recognition
Speech recognition
HMMs
RNNs
LSTMs

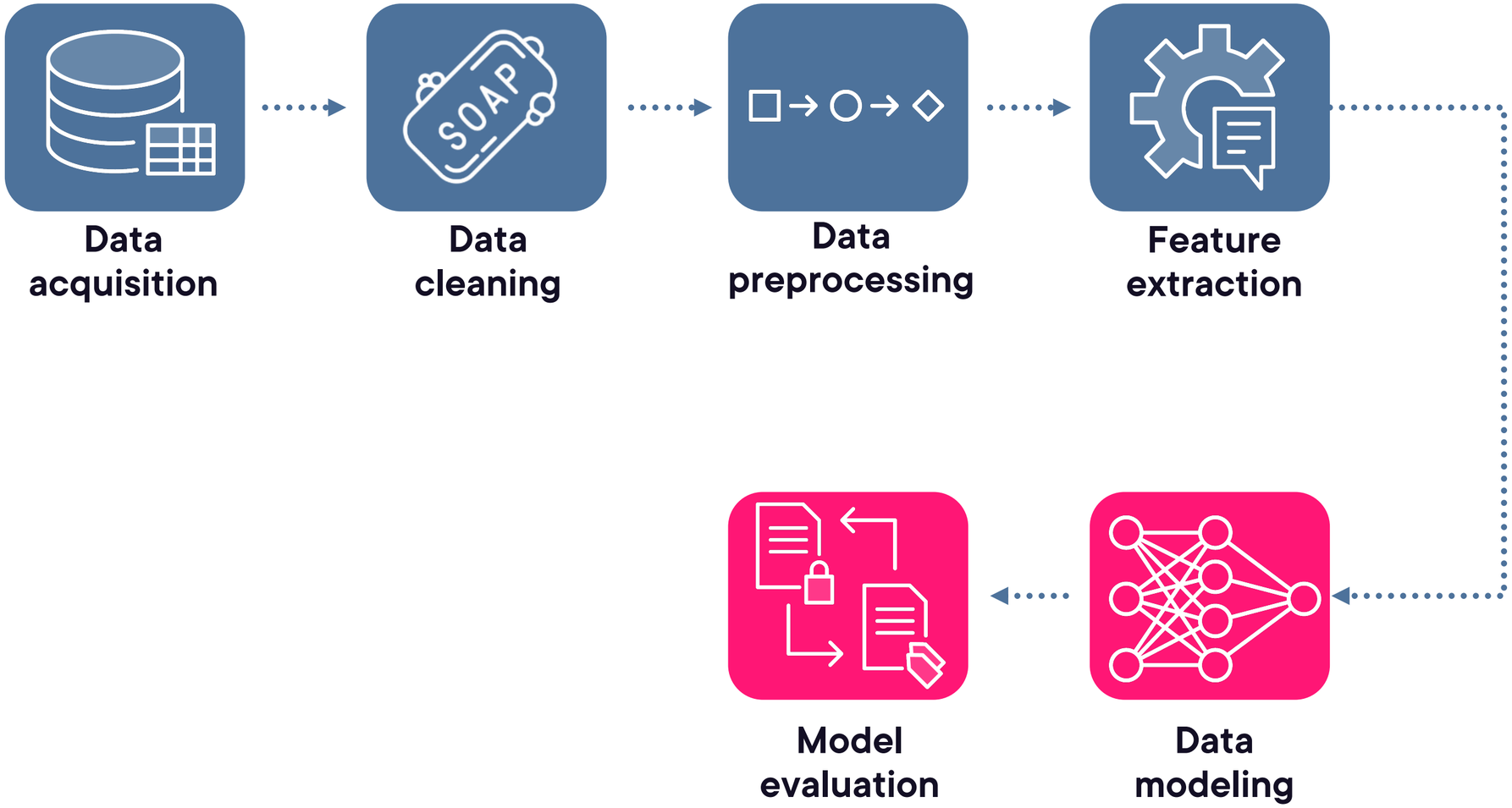
Models


Text classification
Sentiment analysis
Named entity recognition
Speech recognition
Text generation
LSTMs
GPT

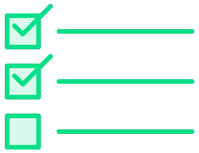
Models


Text classification
Sentiment analysis
Named entity recognition
Speech recognition
Text generation
Machine translation
Statistical machine translation
Seq2Seq
GPT

NLP Pipeline

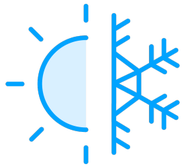


Evaluation Metrics



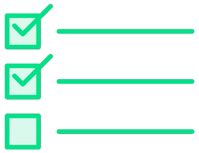
Accuracy

Precision



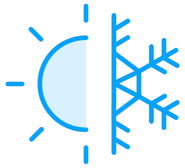
Recall

Evaluation Metrics



Accuracy

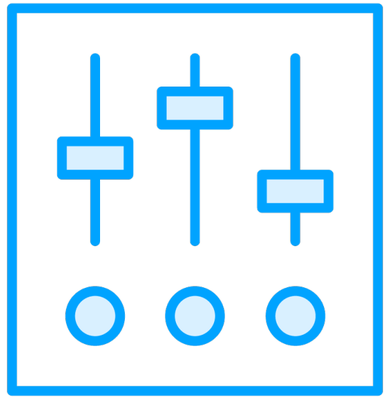
Precision



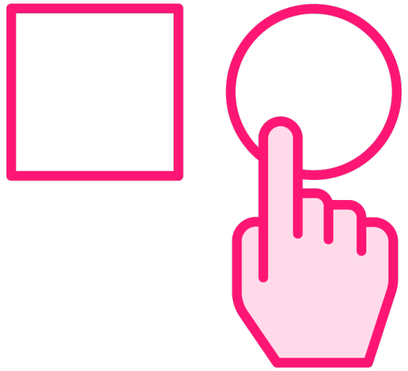
Recall

The evaluation phase is iterative.

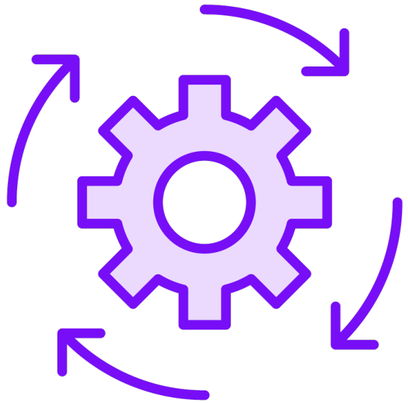
Evaluation



Adjusting parameters

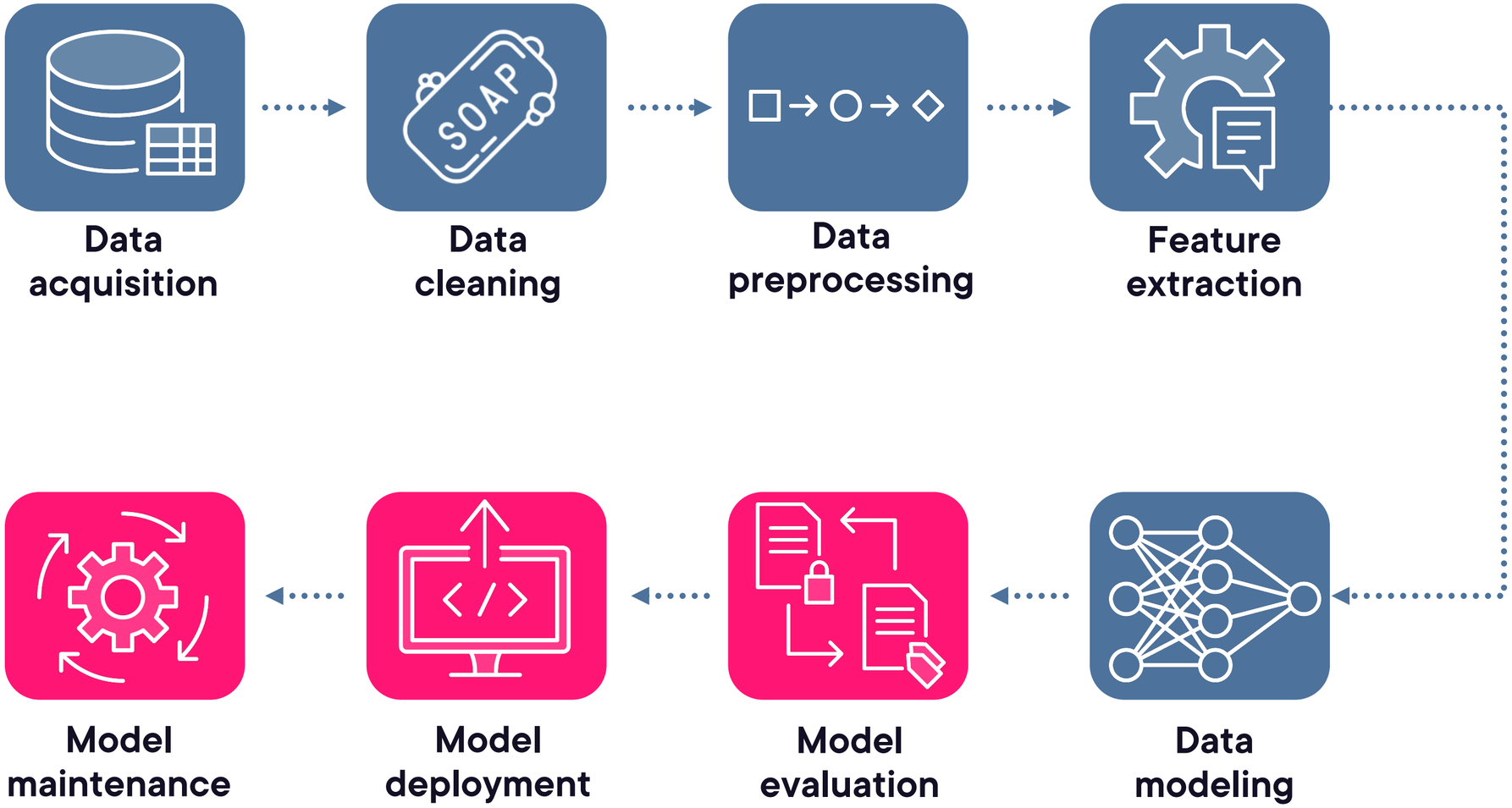


Choosing different algorithms



Revising data cleaning methods

NLP Pipeline



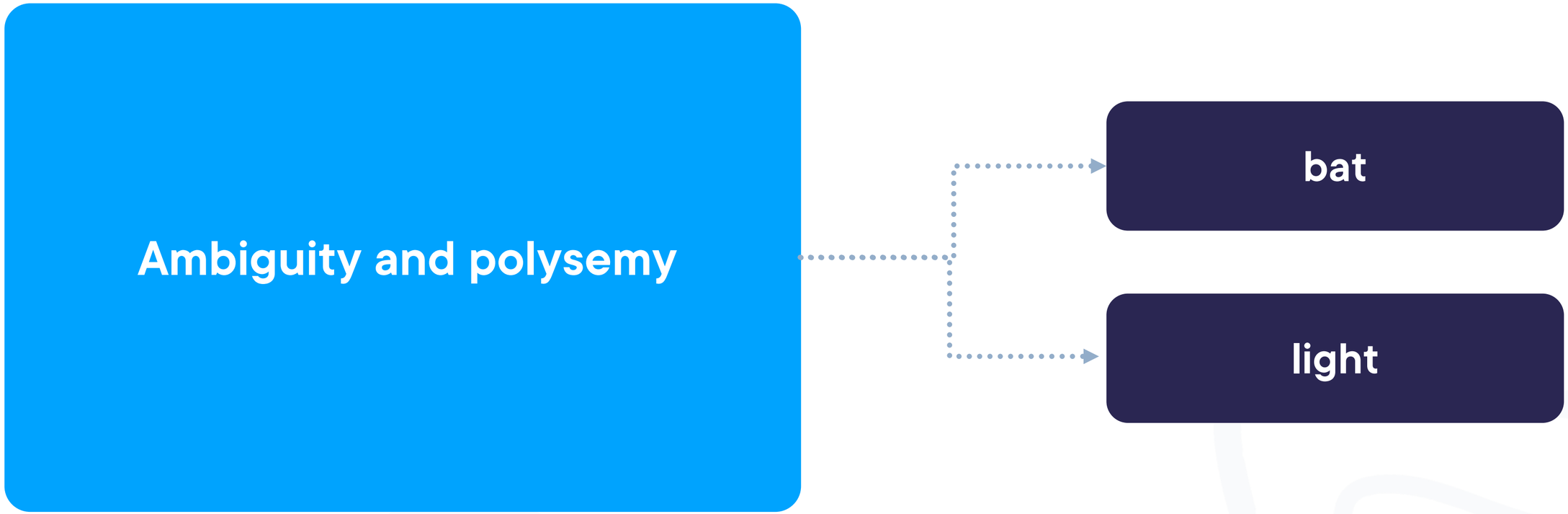
Demo

Train and evaluate a classifier

- Naive Bayes

Challenges in NLP

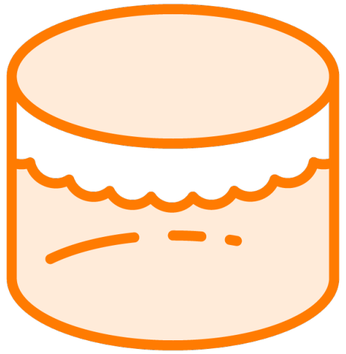
Challenges



NLP uses various disambiguation techniques.

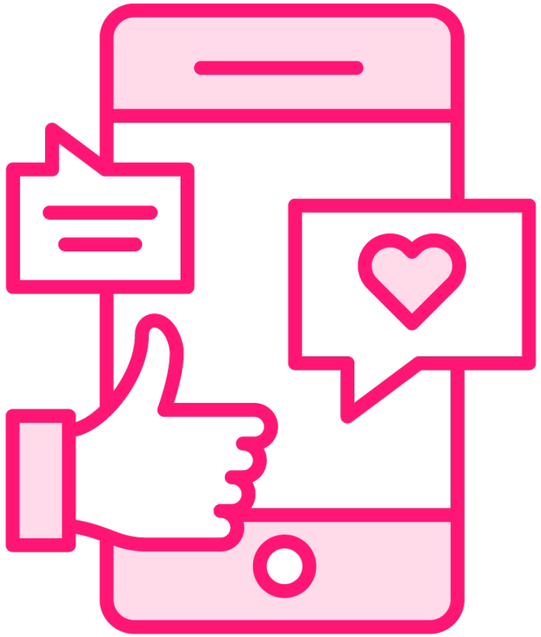
Disambiguation Techniques

Use context to determine a word's meaning



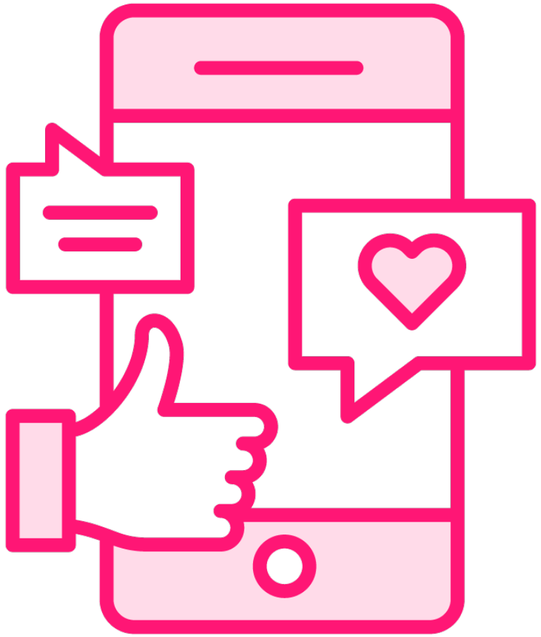
Perform training on large datasets

Sentiment Analysis



The movie was incredibly light.

Sentiment Analysis



The movie was incredibly light.

positive

Sarcasm and irony

Sarcasm and irony

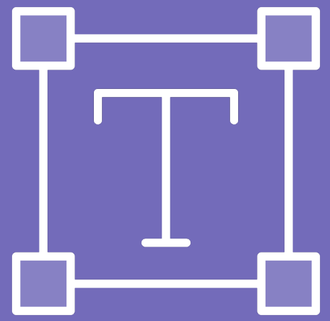
Multiple languages and dialects

Sarcasm and irony

Multiple languages and dialects

Cultural and linguistic diversity

Sarcasm and irony



Multiple languages and dialects

Cultural and linguistic diversity

Ethical issues

Summary

NLP Pipeline

- Data acquisition
- Data cleaning and preprocessing
- Feature extraction
- Data modeling and evaluation

Thank you!

linkedin.com/in/mihaela-danci/