AI Camp

FUNDAMENTALS OF NLP

NLP101

PRESENTED BY

CHAMEL NADIR BOUACHA

school of ai Algiers

AI Camp

WORKSHOP PLAN

- INTRODUCTION
- TEXT PREPROCESSING
- WORD EMBEDDINGS
  - WORD2VEC,
  - GLOVE
  - FASTTEXT CONCEPTS
- APPLICATION :
  - NLTK
  - SPACY

school of ai Algiers

AI Camp

INTRODUCTION

ELIZA	CHATGPT
1966	2022
REGEX	TRANSFORMERS

school of ai Algiers

AI Camp

INTRODUCTION

HOW REGEX CAN SIMULATE A CHATBOT ?

school of ai Algiers

AI Camp

TEXT PREPROCESSING

school of ai Algiers

AI Camp

TEXT PREPROCESSING

ELIZA

PATTERN MATCHING + SUBSTITUTION = CONVERSATION.

```
User:   I am unhappy.
ELIZA:  DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
User:   I need some help, that much seems certain.
ELIZA:  WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
User:   Perhaps I could learn to get along with my mother.
ELIZA:  TELL ME MORE ABOUT YOUR FAMILY
User:   My mother takes care of me.
ELIZA:  WHO ELSE IN YOU FAMILY TAKES CARE OF YOU
User:   My father.
ELIZA:  YOUR FATHER
User:   You are like my father in some ways.
```

Weizenbaum (1966)

school of ai Algiers

AI Camp

TEXT PREPROCESSING

ELIZA

PATTERN MATCHING + SUBSTITUTION = CONVERSATION.

```
s/.* YOU ARE (depressed|sad) .*/I AM SORRY TO HEAR YOU ARE \1/
s/.* YOU ARE (depressed|sad) .*/WHY DO YOU THINK YOU ARE \1/
s/.* all .*/IN WHAT WAY/
s/.* always .*/CAN YOU THINK OF A SPECIFIC EXAMPLE/
```

school of ai Algiers

AI Camp

TEXT PREPROCESSING

SIMPLE WORD PROCESSING TECHNIQUE COULD CREATE A DECENT RESULT

BUT

WHAT COUNTS AS A WORD?

school of ai Algiers

AI Camp

TEXT PREPROCESSING

SPEECH RECOGNITION

WE CONSIDER PUNCTUATION AS A WORD

TEXT FOCUSED TASKS

DEPENDS ON THE TASK

school of ai Algiers

AI Camp

TEXT PREPROCESSING

CORPORA

TEXT

SPEECH

school of ai Algiers

AI Camp

TEXT PREPROCESSING

CORPORA

WORD TYPES: 3 **WORD INSTANCES:** 4

school of ai Algiers

AI Camp

TEXT PREPROCESSING

LEMMAS VS. WORDFORMS

LEMMA: THE BASE FORM OF A WORD

CAT FOR CATS

WORDFORMS: FULL INFLECTED OR DERIVED FORMS

CATS FROM CAT

school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD TOKENIZATION

- PREPARES TEXT FOR ANALYSIS
- EASIER TO PROCESS TEXT
- FOUNDATION FOR TEXT REPRESENTATION (E.G WORD EMBEDDINGS)

school of ai Algiers

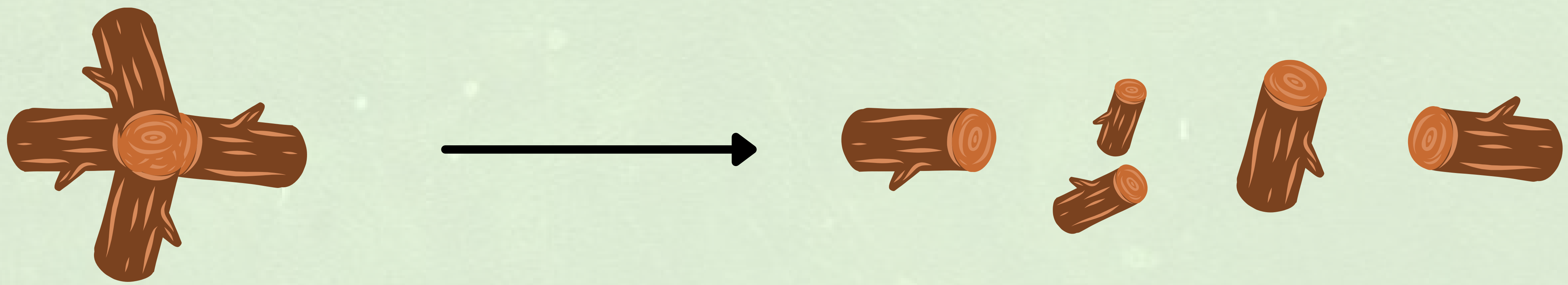
AI Camp

TEXT PREPROCESSING

WORD TOKENIZATION

1-TOP-DOWN (RULE-BASED) TOKENIZATION

DEFINING RULES (MAINLY REGEX) THAT WILL SPLIT CORPORA INTO TOKENS



school of ai Algiers

AI Camp

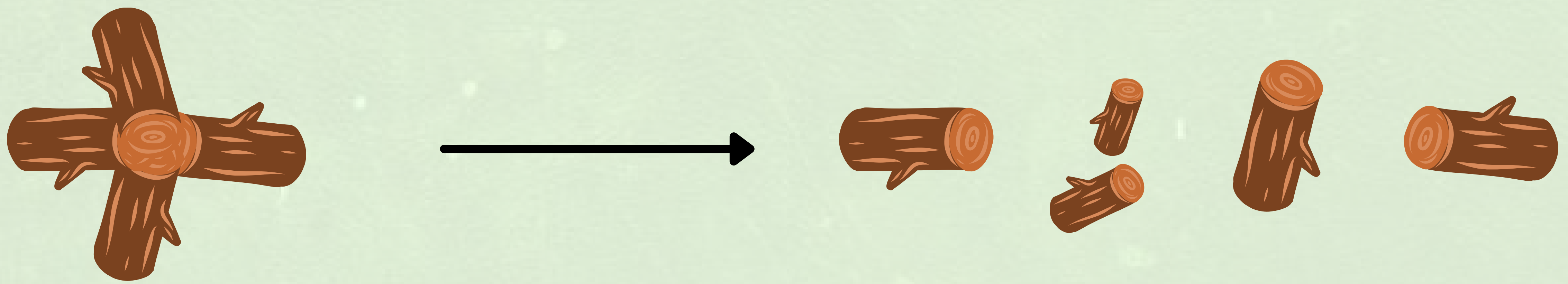
TEXT PREPROCESSING

WORD TOKENIZATION

1-TOP-DOWN (RULE-BASED) TOKENIZATION

```
Input: "The San Francisco-based restaurant," they said, "doesn't charge $10".

Output: ['The', 'San', 'Francisco-based', 'restaurant', ',', '"', 'does', "n't",
'charge', '$', '10', '.']
```



school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD TOKENIZATION

2-BOTTOM-UP TOKENIZATION

```
5 l o w
2 l o w e s t
6 n e w e r
3 w i d e r
2 n e w
```

→

```
l, o, w, e, r, s, t, er, ne, low, newer.
```

school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD NORMALIZATION

RAW CORPORA → NORMALIZED CORPORA IN STANDARD FORMAT

school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD NORMALIZATION

CASE FOLDING

CONVERTS ALL CHARACTERS TO LOWERCASE.

EFFECTIVE FOR INFORMATION RETRIEVAL AND SPEECH RECOGNITION

school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD NORMALIZATION

CHOOSE A SINGLE STANDARD FORM FOR VARIATIONS

EXAMPLE: USA AND US, UH-HUH AND UHHUH.

USEFUL IN TASKS LIKE INFORMATION RETRIEVAL, WHERE WE WANT TO RETRIEVE DOCUMENTS MENTIONING EITHER FORM.

school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD NORMALIZATION

LEMMATIZATION

IDENTIFIES THE ROOT FORM (LEMMA) OF A WORD

`am`, `are`, and `is` → `be`.

`dinner` and `dinners` → `dinner`.

school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD NORMALIZATION

LEMMATIZATION

HELPS IN WEB SEARCH OR INFORMATION RETRIEVAL BY TREATING VARIATIONS OF A WORD AS EQUIVALENT:

Example: A query for `woodchucks` should also return results for `woodchuck`.

AI Camp

TEXT PREPROCESSING

WORD NORMALIZATION

LEMMATIZATION

COMPLICATED TO DETECTS LEMMA BECAUSE IT USES SOPHISTICATED METHODS USING

MORPHEMES **AFFIXES**

Example: Spanish word `amaren` (`if in the future they would love`) →

- `amar` (to love) + morphological features (third person plural, future subjunctive).

school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD NORMALIZATION

STEMMING

- NAIVE WAY COMPARED TO LEMMATIZATION
- CHOPPING OFF WORD ENDINGS (AFFIXES) WITHOUT UNDERSTANDING THE MORPHOLOGICAL STRUCTURE.

school of ai Algiers

AI Camp

TEXT PREPROCESSING

WORD NORMALIZATION

STEMMING

Original text:

> This was not the map we found in Billy Bones's chest, but an accurate copy, complete in all things-names and heights and soundings-with the single exception of the red crosses and the written notes.

→

Stemmed text:

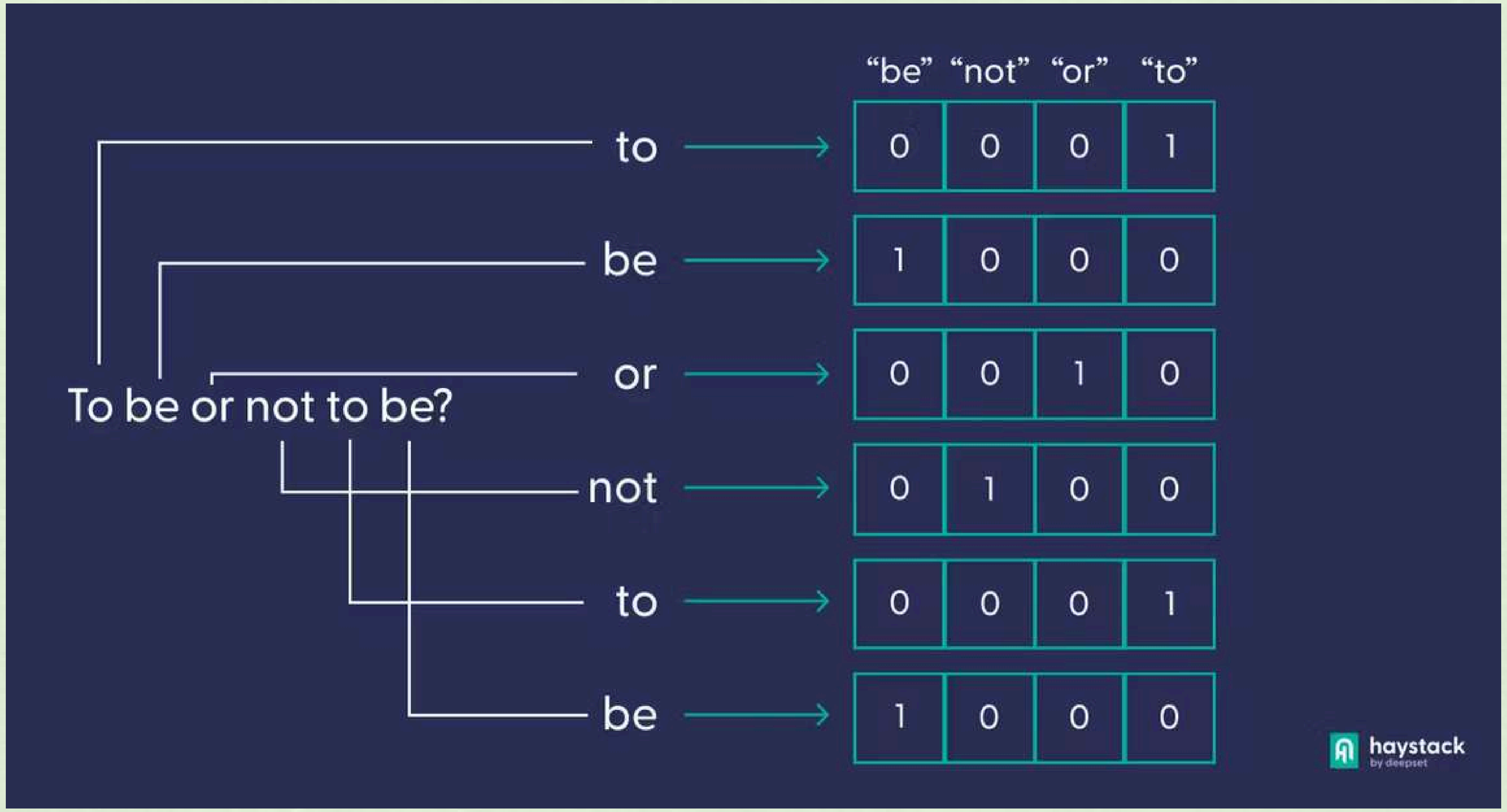
> Thi wa not the map we found in Billi Bone s chest but an accur copi complet in all thing name and height and sound with the singl except of the red cross and the written note.

school of ai Algiers

AI Camp

WORD EMBEDDINGS

TEXT EMBEDDINGS ARE NUMERICAL REPRESENTATIONS OF WORDS,



school of ai Algiers

WORD EMBEDDINGS

WHY ?

AI Camp

WORD EMBEDDINGS

CONTEXTUAL UNDERSTANDING

GENERALIZATION

ENHANCED MACHINE LEARNING TASKS

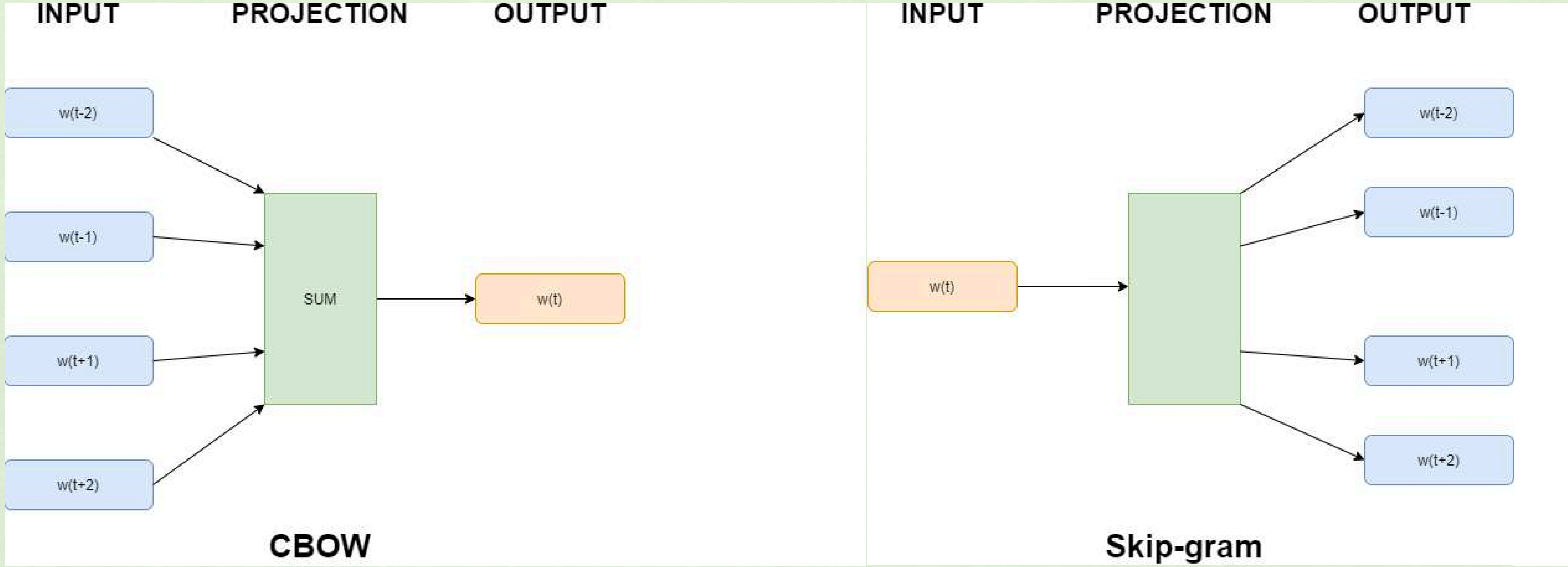
school of ai Algiers

AI Camp

WORD EMBEDDINGS

WORD2VEC

LOCAL CONTEXT ONLY



school of ai Algiers

AI Camp

WORD EMBEDDINGS

GLOVE

LOCAL CONTEXT + WORD CO-OCCURRENCE

school of ai Algiers

AI Camp

WORD EMBEDDINGS

FASTTEXT

- CHARACTER N-GRAMS (E.G., "PLAYING" → "PLAY", "ING", "LAY").
- THESE SUBWORDS ARE TRAINED TO GENERATE EMBEDDINGS

school of ai Algiers

AI Camp

WORD EMBEDDINGS

WORD2VEC

GLOVE

FASTTEXT

Feature	Word2Vec	GloVe	FastText
Developer	Google	Stanford	Facebook AI
Core Idea	Context-based	Co-occurrence	Subword-based
Context	Local	Local + Global	Local
Handles OOV?	No	No	Yes
Efficiency	High	Moderate	Moderate
Polysemy	No	No	Partially
Applications	General NLP	Large Corpora	Morphologically rich languages, OOV handling

AI Camp

APPLICATION TIME

school of ai Algiers