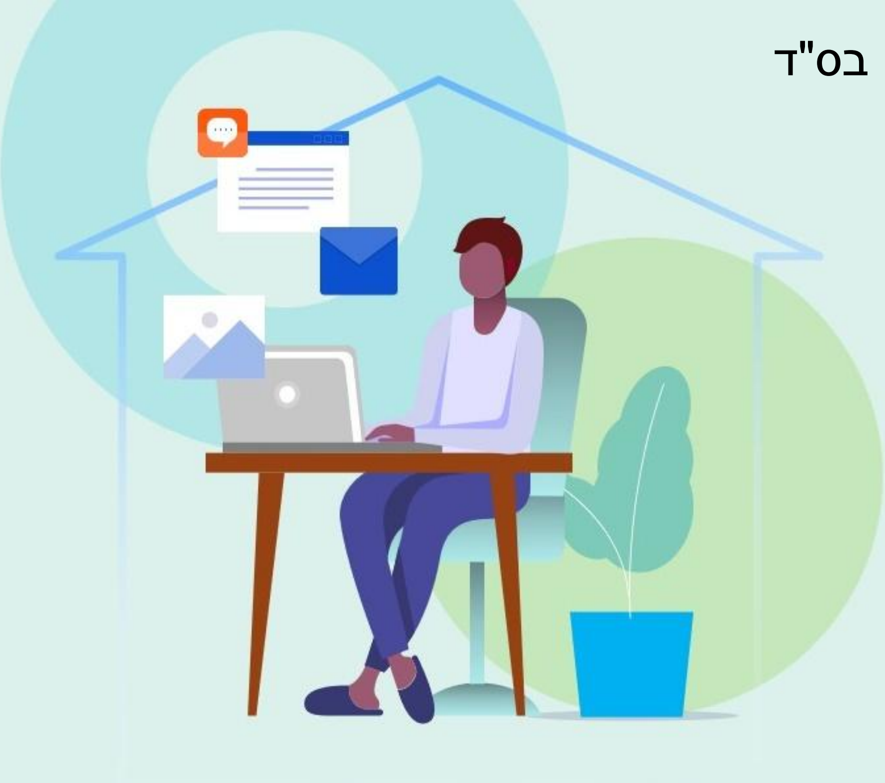
בס"ד

AI & NLP

(Natural Language Processing)

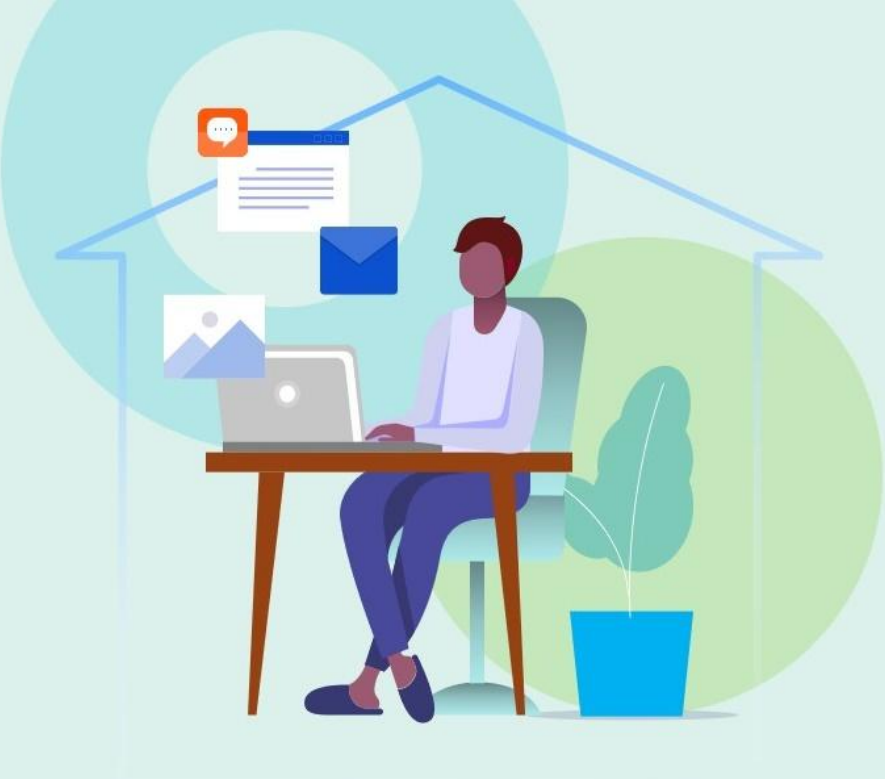
For SW Engineers



David Berger, 2025
Software Engineer

ExLibris
Part of Clarivate

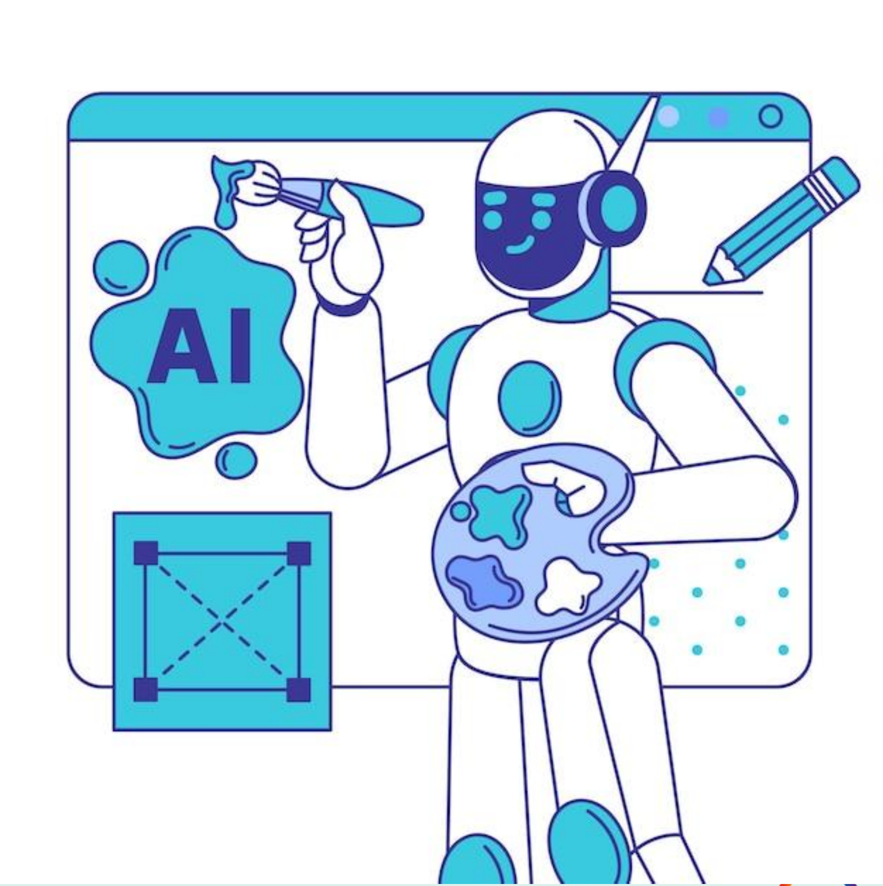
1. What is AI?
2. Using AI
3. Review – Activity!
4. What is NLP?
5. An NLP Pipeline
6. Group Activity!



ExLibris
Part of Clarivate

1. What is AI?

And will it replace me?



ExLibris
Part of Clarivate

AI vs ML vs Data Science

AI

Performing human-like operations

Machine Learning

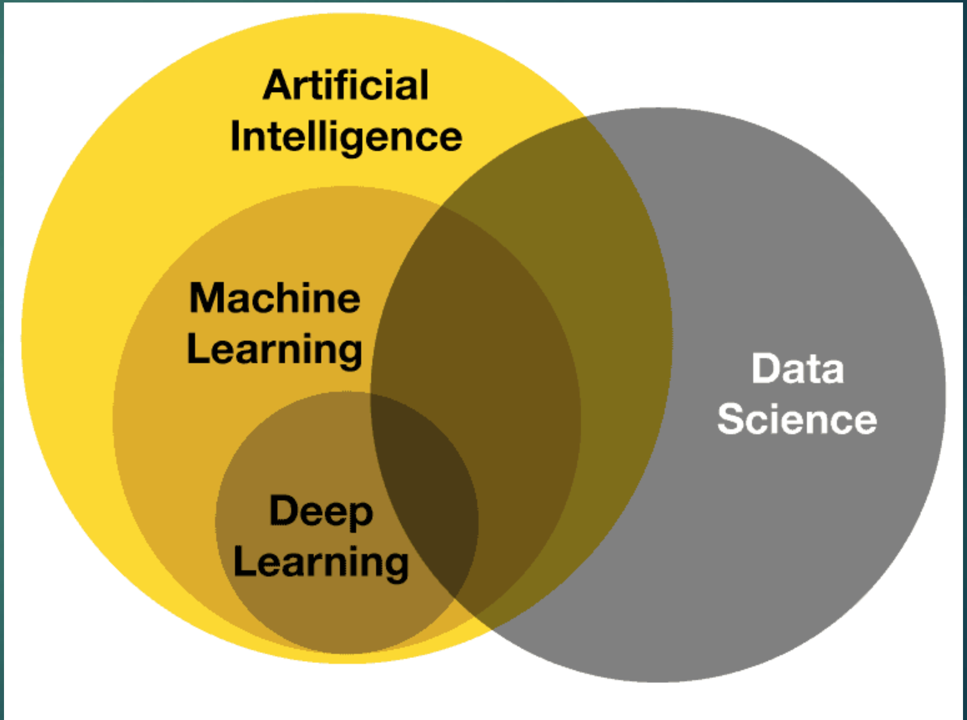
Learning patterns from data

Deep Learning

Complex, multi layered learning

Data Science

Extract insights from Data



Fields of AI/ML/Data Science

* Grouping not exact

Artificial Intelligence

- Natural Language Processing
- Generative AI
- AI Agents
- Knowledge Graphs

Machine Learning

- Recommendation
- Supervised/ Unsupservised Learning
- Anomaly Detection

Deep Learning

- Neural Networks
- Speech Recognition
- Image Recognition
- Transformers

What is an AI Developer?

Data Scientist | AI Developer:

- Mathematics
- Research
- Abstract
- Specific
- Accuracy
- Experiments

SW Developer

- Logic
- Customers, Business
- Tangible / Real World
- Flexible
- Design
- Workflow

The AI Takeover

"AI is automating tasks in many industries, replacing jobs in customer service, data entry, and even creative fields like writing and design.

As technology advances, companies increasingly rely on AI for efficiency, reducing the need for human workers in repetitive and analytical roles.“

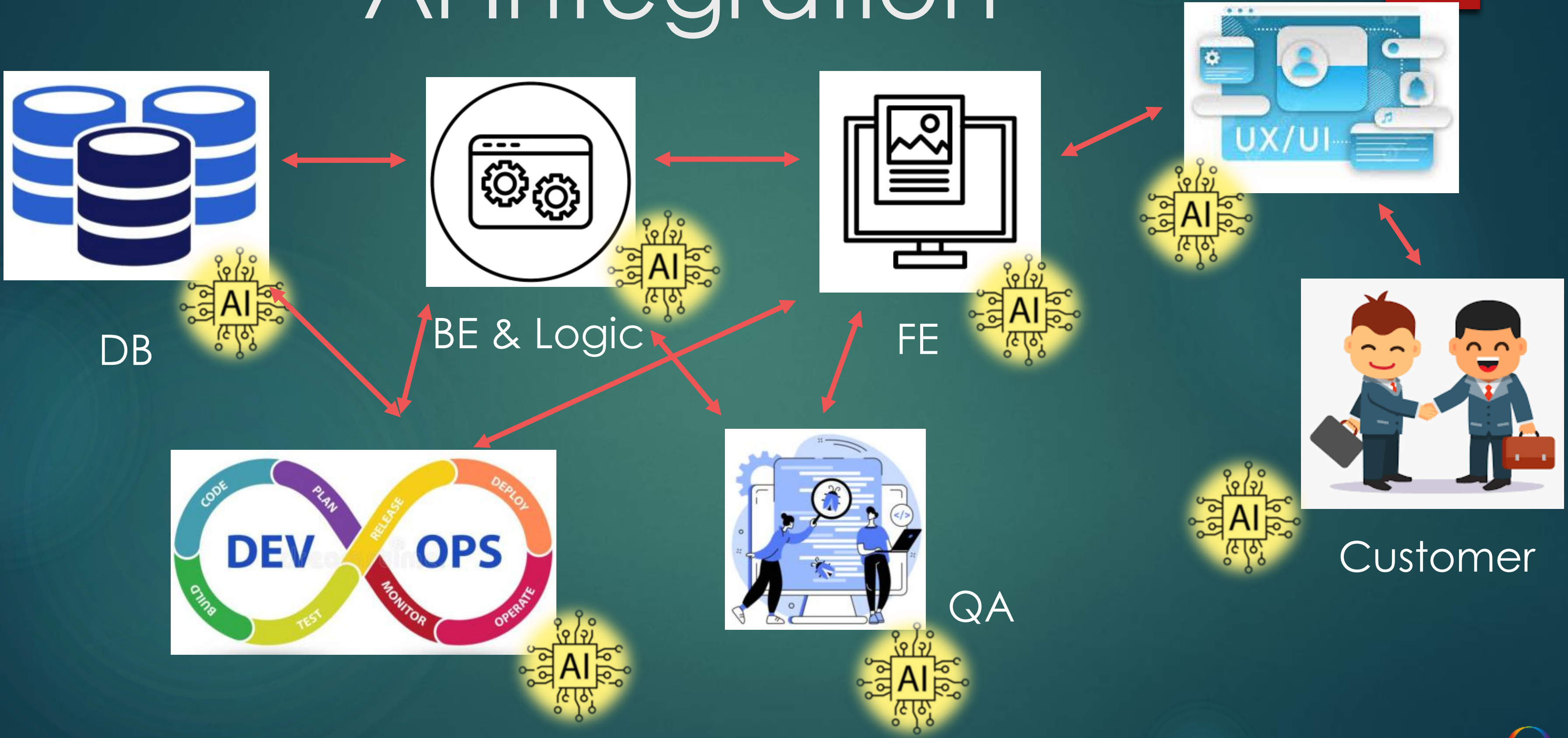
ChatGPT

What about me? (SWEs)

"Our AI model now handles about 50% of software engineering tasks, and we aim for 90% in a year."

Dario Amodei, CEO of Anthropic (2024)

Ai Integration



DB

BE & Logic

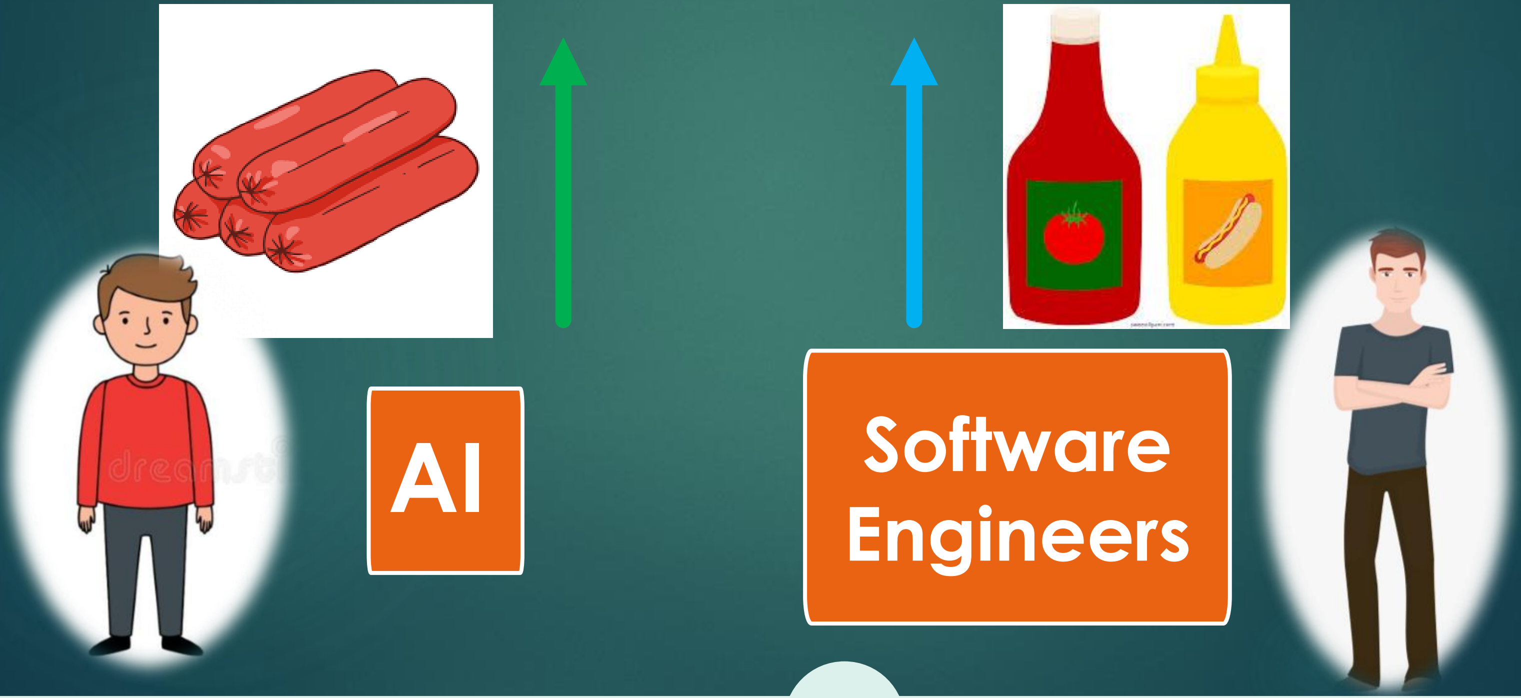
FE

QA

Customer

Complementary Goods

(מוצר משלים)



AI

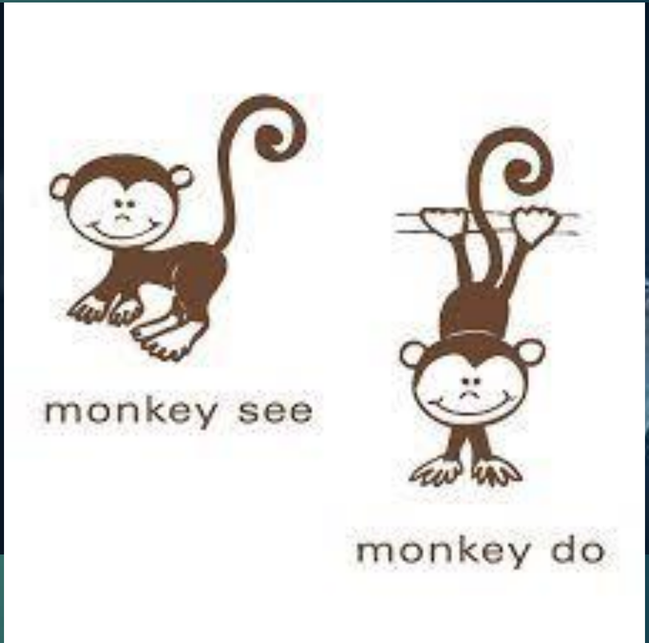
Software Engineers

2. Using AI

(and when *not* to use AI)

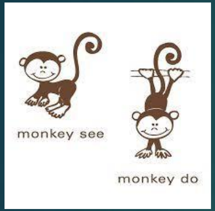
Gen AI – How it works

Data	Source (context)
Bibi is good	כאן 11
Bibi is bad	ערוץ 12–קשת
Bibi is evil	ערוץ 13–רשת
Bibi is a leader	ישראל היום
Bibi is destruction	הארץ



“Bibi is _____ ? “

Takeaways

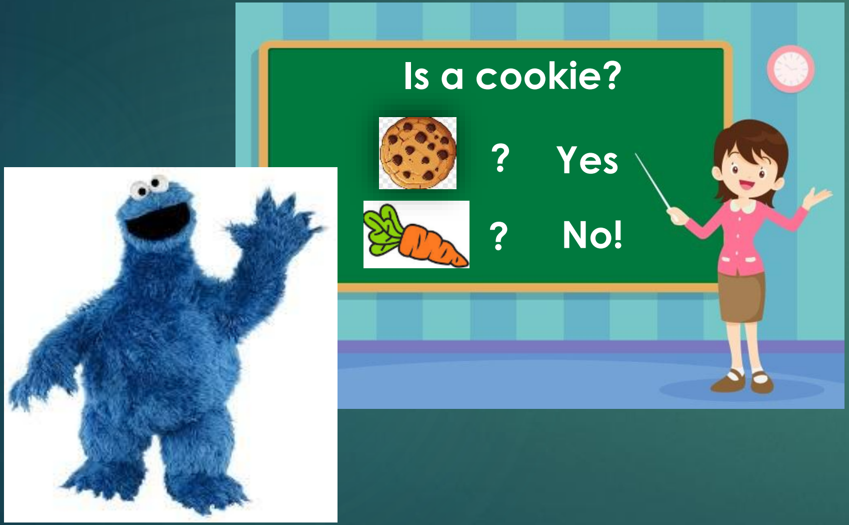


Gen AI’s output:

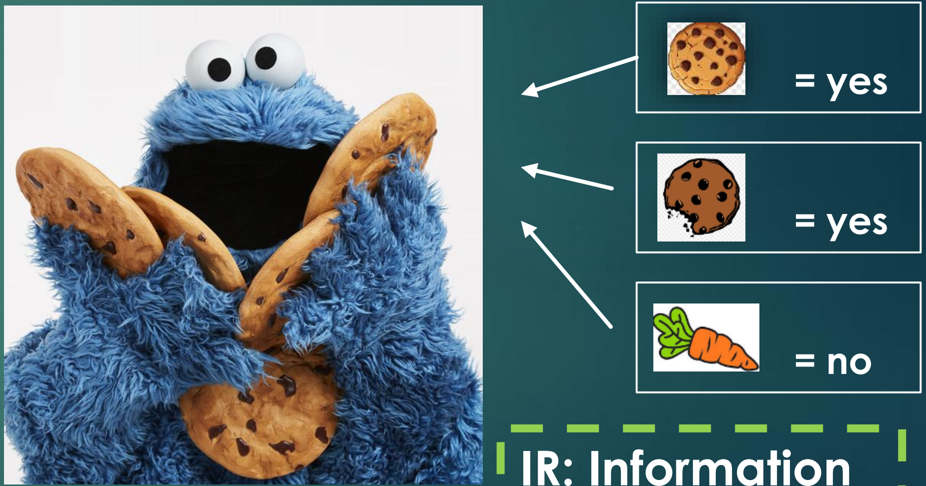
- Context
- Same Input -> Diff Results
- Inaccuracy %

Machine Learning

Supervised Learning

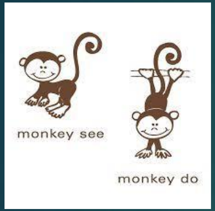


Unsupervised Learning



IR: Information Retrieval

Takeaways



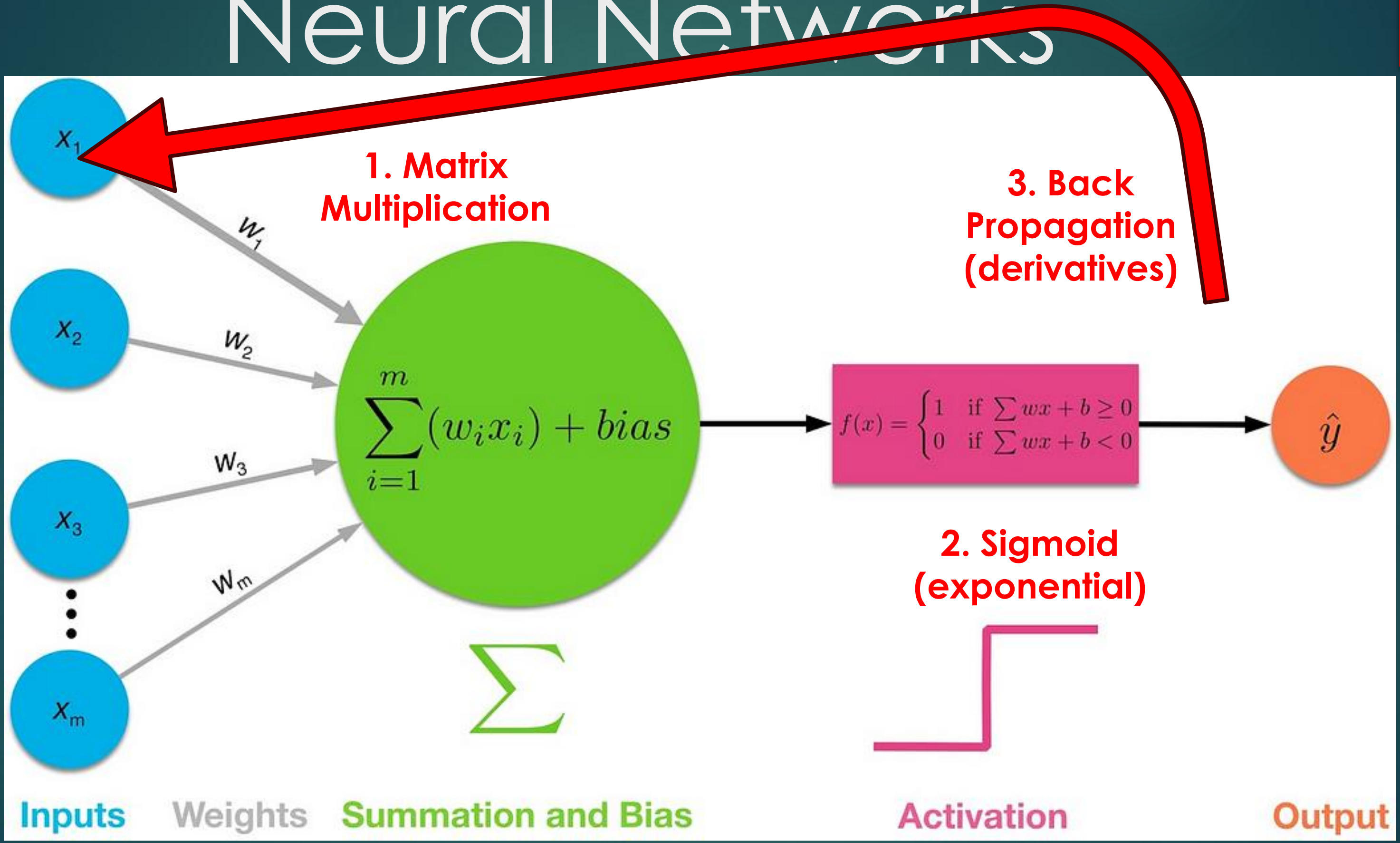
Gen AI's output:

- Context
- Same Input -> Diff Results
- Inaccuracy %

Data sets must be:

- Large
- Clean

Neural Networks

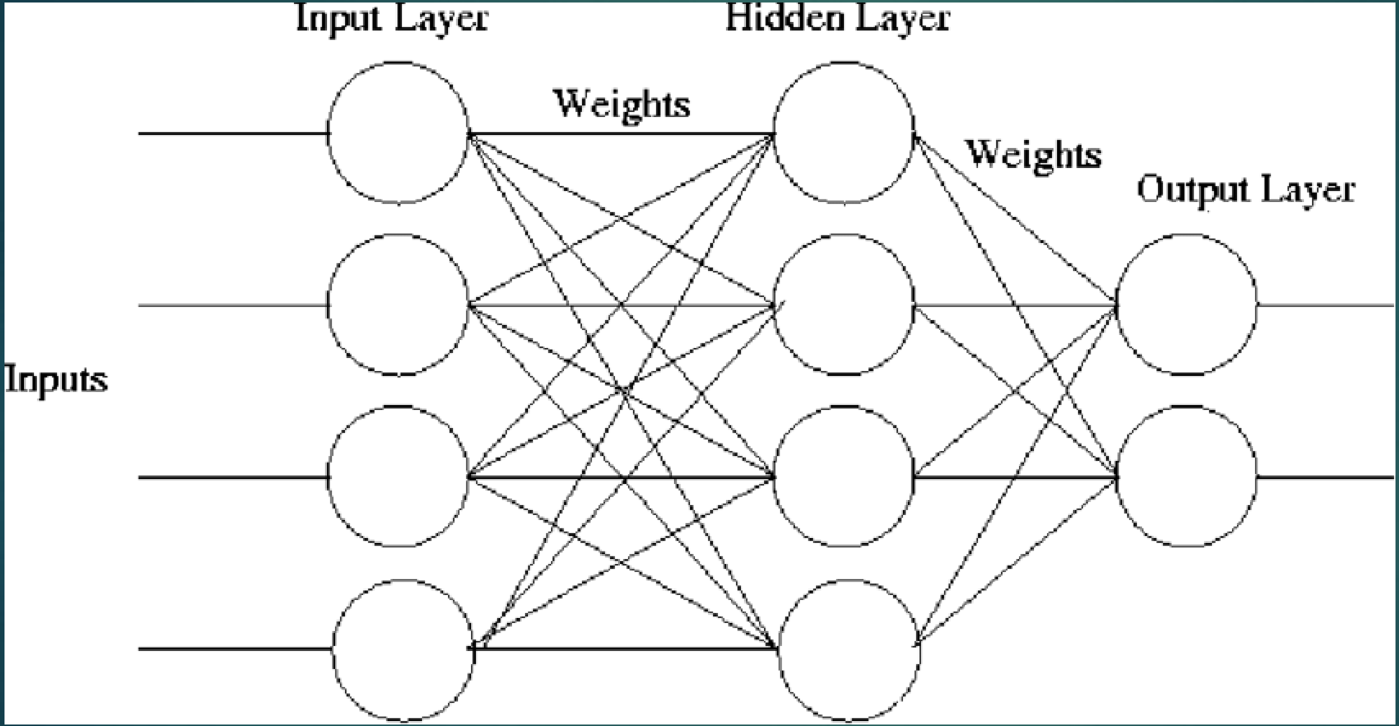

x_1
x_2
x_3
x_m
w_1
w_2
w_3
w_m
$\sum_{i=1}^{m}(w_i x_i) + bias$
$\sum$
$f(x) = \begin{cases} 1 & \text{if } \sum wx + b \geq 0 \\ 0 & \text{if } \sum wx + b < 0 \end{cases}$
$\hat{y}$
Inputs
Weights
Summation and Bias
Activation
Output

1. Matrix Multiplication

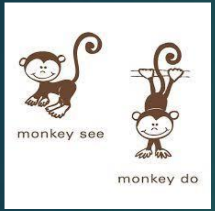
2. Sigmoid (exponential)

3. Back Propagation (derivatives)

Layered Neural Networks



Takeaways



Gen AI’s output:

- Context
- Same Input -> Diff Results
- Inaccuracy %

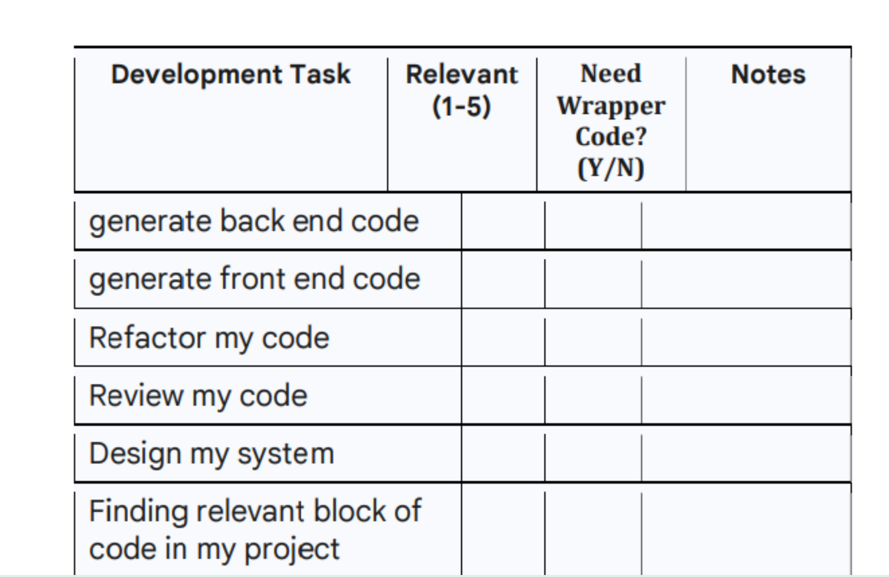
Data sets must be:

- Large
- Clean

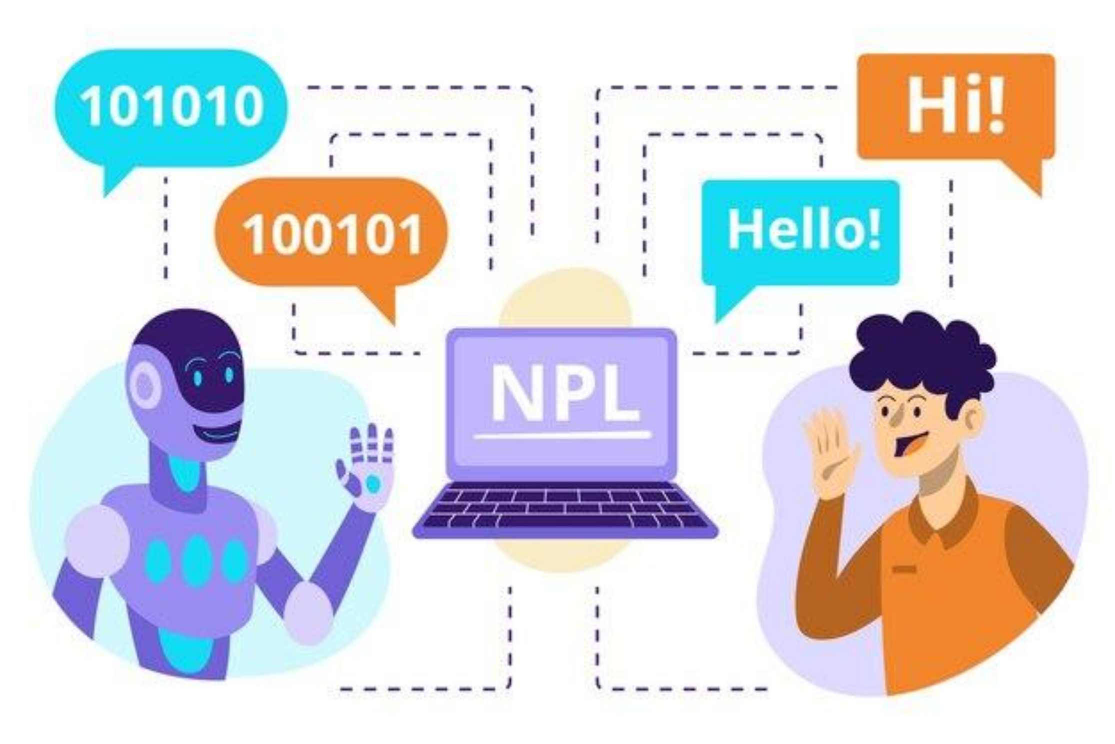
Cool AI is

- Heavy
- Slow
- Expensive

3. Review Activity!



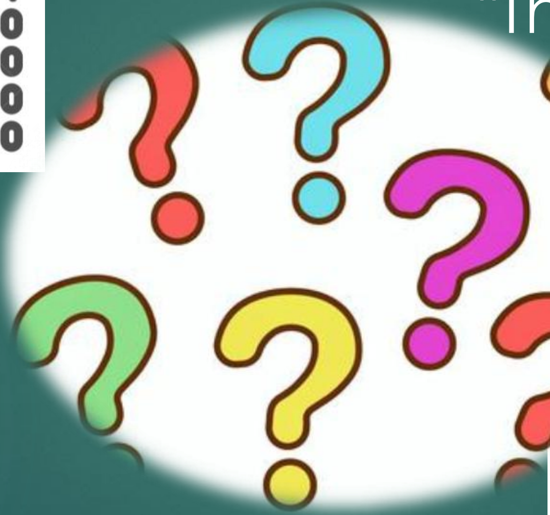
Development Task	Relevant (1-5)	Need Wrapper Code? (Y/N)	Notes
generate back end code			
generate front end code			
Refactor my code			
Review my code			
Design my system			
Finding relevant block of code in my project			



4. What is NLP?

When machines learn to talk

Math <-> Words



"The cat sat on the mat."

"The king and the queen met a wise woman in the castle."

Tokenization

["the", "king", "and", "the", "queen", "met", "a", "wise", "woman", "in", "the", "castle"]

Stop-words

["king", "queen", "met", "wise", "woman", "castle"]

Lemma

["king", "queen", **"meet"**, "wise", "woman", "castle"]

1 Word ≈ .75 token

Word Vectors

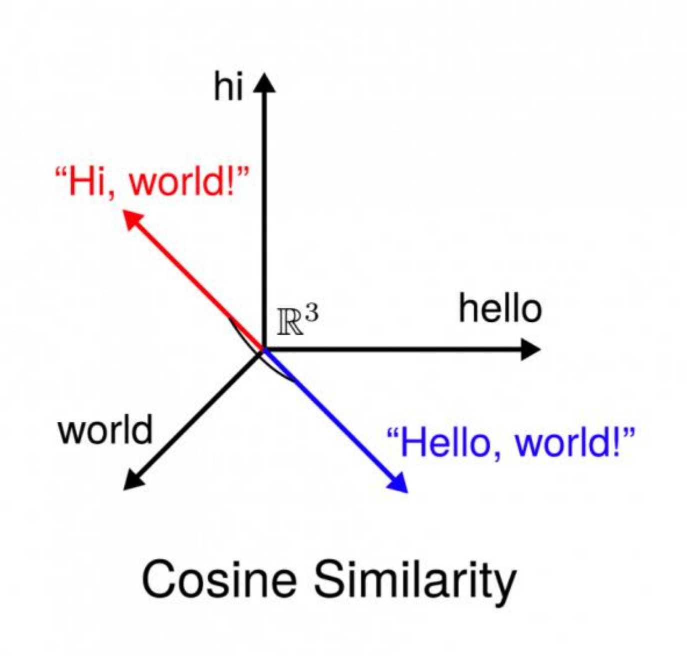
384d

["king", "queen", "meet", "wise", "woman", "castle"]

Word	Vector (x, y, z)
	[Royalty, People, Places]
king	[0.9, 0.9, 0.2]
queen	[0.9, 0.9, 0.3]
man	[0.4, 0.9, 0.2]
woman	[0.4, 0.9, 0.4]
meet	[0.2, 0.4, 0.3]
wise	[0.2, 0.3, 0.2]
castle	[0.9, 0.2, 0.7]

Vector Similarity

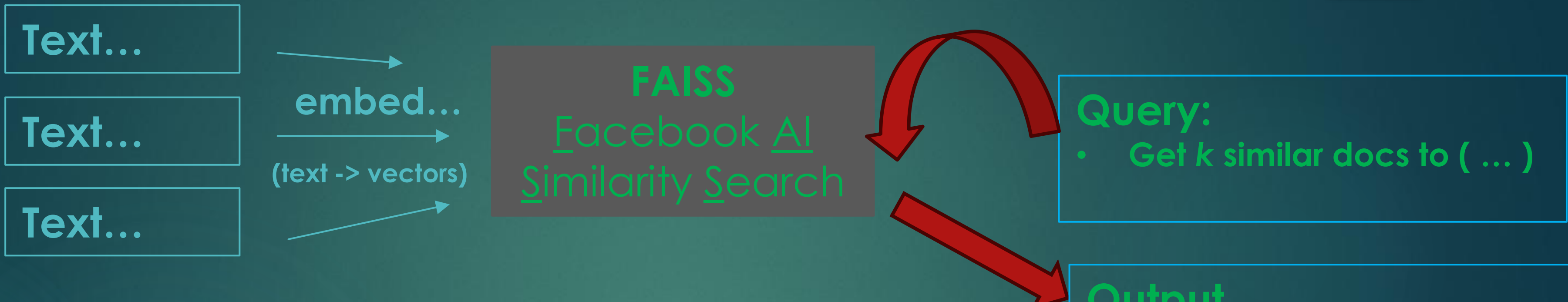
“Hi, world!” “Hello, world!”



Cosine Similarity

“The king and the queen met a wise woman in the castle.”	**1.00**
"The wise queen and king met a woman in the castle."	**~0.7-0.8**
"A castle woman met the wise king and queen."	**~0.4-0.5**

Semantic Search



Text…

Text…

Text…

embed…

(text -> vectors)

FAISS

Facebook AI Similarity Search

Query:

- Get *k* similar docs to (…)

Output

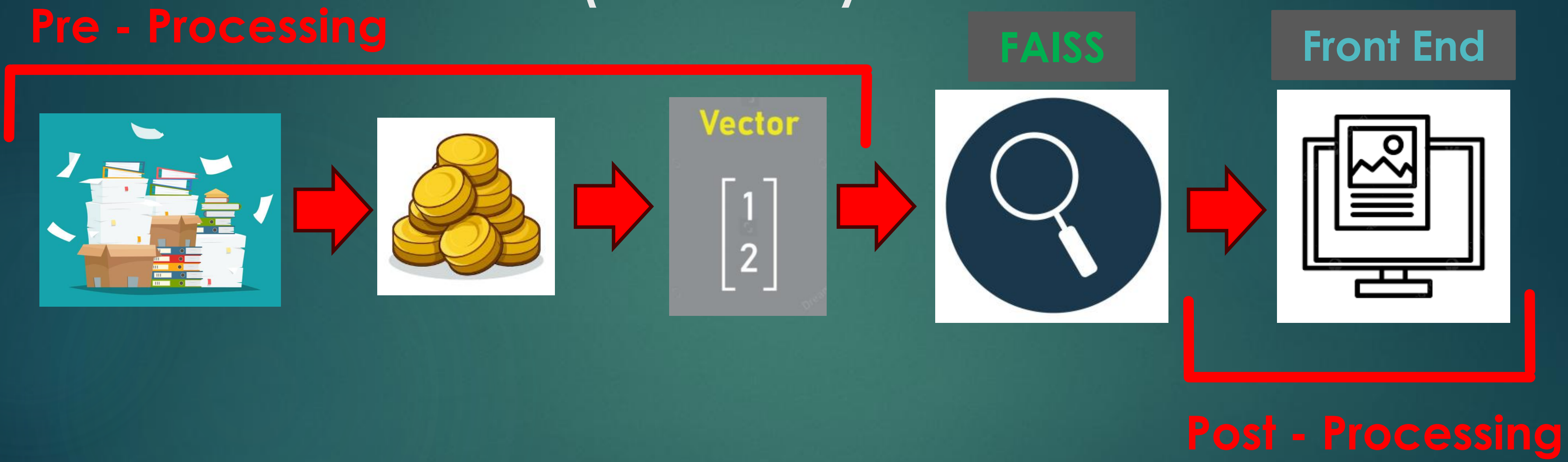
```
{
    [0] text….
    [1] text….
    [2] text…
}
```

What about Filtering?

No Metadata Parameters option…

- ***k* = total size. Filter after FAISS's result**
- **Manage Multiple FAISS indexes**

An NLP Pipeline (so far)



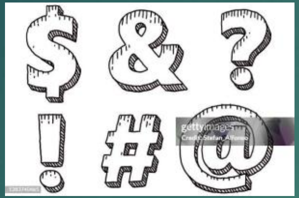
Pre - Processing

FAISS

Front End

Post - Processing

Pre Processing

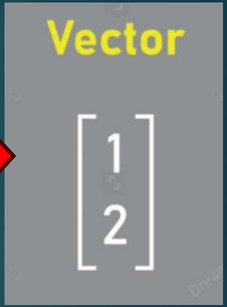


Html

<div>Learn <b>Python</b> for <i>data</i> science!</div>

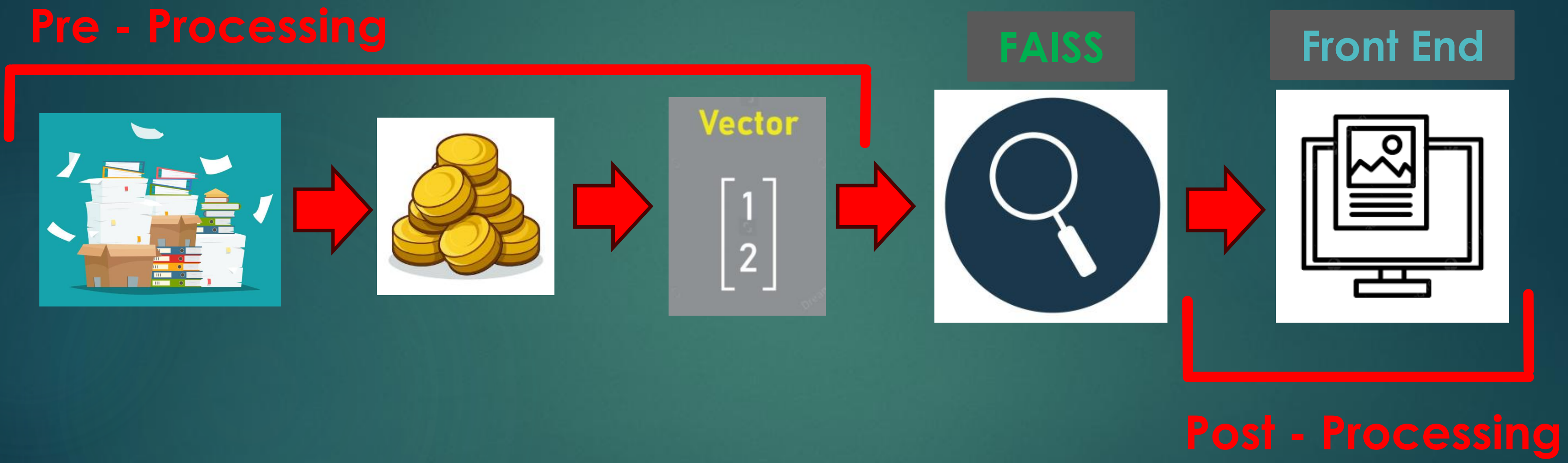
text

Learn Python for data science!



An NLP Pipeline

Html -> text



Pre - Processing

Vector

FAISS

Front End

Post - Processing

LLMs

Large Language Models

ChatGPT

- Broad Knowledge
- Multi – Lingual
- Strong Reasoning

- Open-source
- Run Locally (customizable)

Claude

- Long Context
- Good code generation
- Cheaper

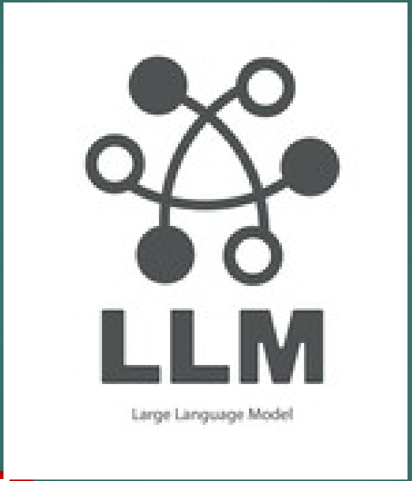
Gemini

- Integrated w/ Google system
- Accept 750k word prompt

LLMs

Parameters

Prompt	
(max length depends on Model)	
100-300 Words	1,500+ Words
₪	₪ ₪ ₪



Temperature		
0.3	0.7	1.0
Focused, Factual	Balanced	Creative, Ideas

Max_Tokens	
<100	2000+
Succinct	Large Essays
₪	₪ ₪ ₪

Pricing:

“$0.001 per 1k tokens”

1 Word ≈ .75 token

NER

Named Entity Recognition

"In September 2023, Microsoft announced a $10 billion investment in OpenAI to expand their partnership in artificial intelligence research, following similar moves by Google and Amazon in the United States."

Entity	Type
September 2023	DATE
Microsoft	ORG
$10 billion	MONEY
OpenAI	ORG
Google	ORG
Amazon	ORG
United States	GPE (Geopolitical Entity)

RE

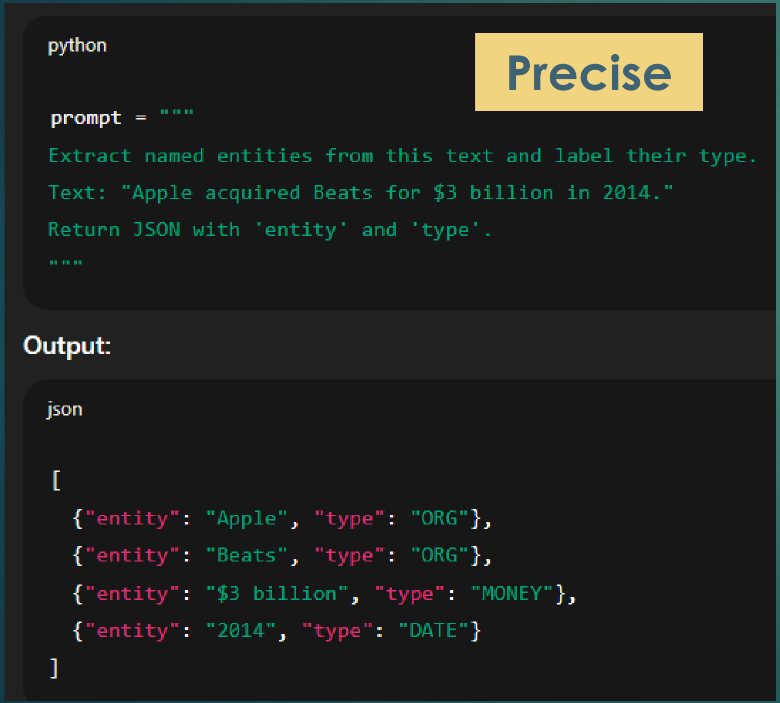
Relationship Extraction

"In September 2023, Microsoft announced a $10 billion investment in OpenAI to expand their partnership in artificial intelligence research, following similar moves by Google and Amazon in the United States."

Entity 1	**Relation**	**Entity 2**	**Example Label**
Microsoft	**invested_in**	OpenAI	INVESTMENT
Microsoft	**investment_amount**	$10 billion	FINANCIAL_VALUE
Microsoft	**investment_date**	September 2023	TIME_OF_EVENT
Microsoft	**partnered_with**	OpenAI	COLLABORATION
Google	**similar_action_to**	Microsoft	COMPARISON
Amazon	**similar_action_to**	Microsoft	COMPARISON
OpenAI	**located_in**	United States	LOCATION

NERs + RE

- LMM gets NERs



Precise

python

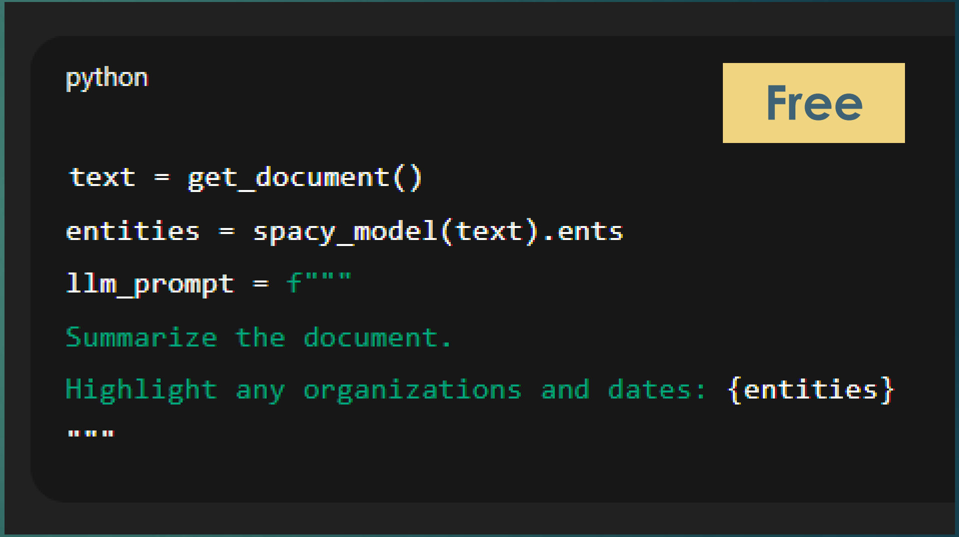
```python
prompt = """
Extract named entities from this text and label their type.
Text: "Apple acquired Beats for $3 billion in 2014."
Return JSON with 'entity' and 'type'.
"""
```

Output:

json

```json
[
  {"entity": "Apple", "type": "ORG"},
  {"entity": "Beats", "type": "ORG"},
  {"entity": "$3 billion", "type": "MONEY"},
  {"entity": "2014", "type": "DATE"}
]
```

- Give LLMs the NERs

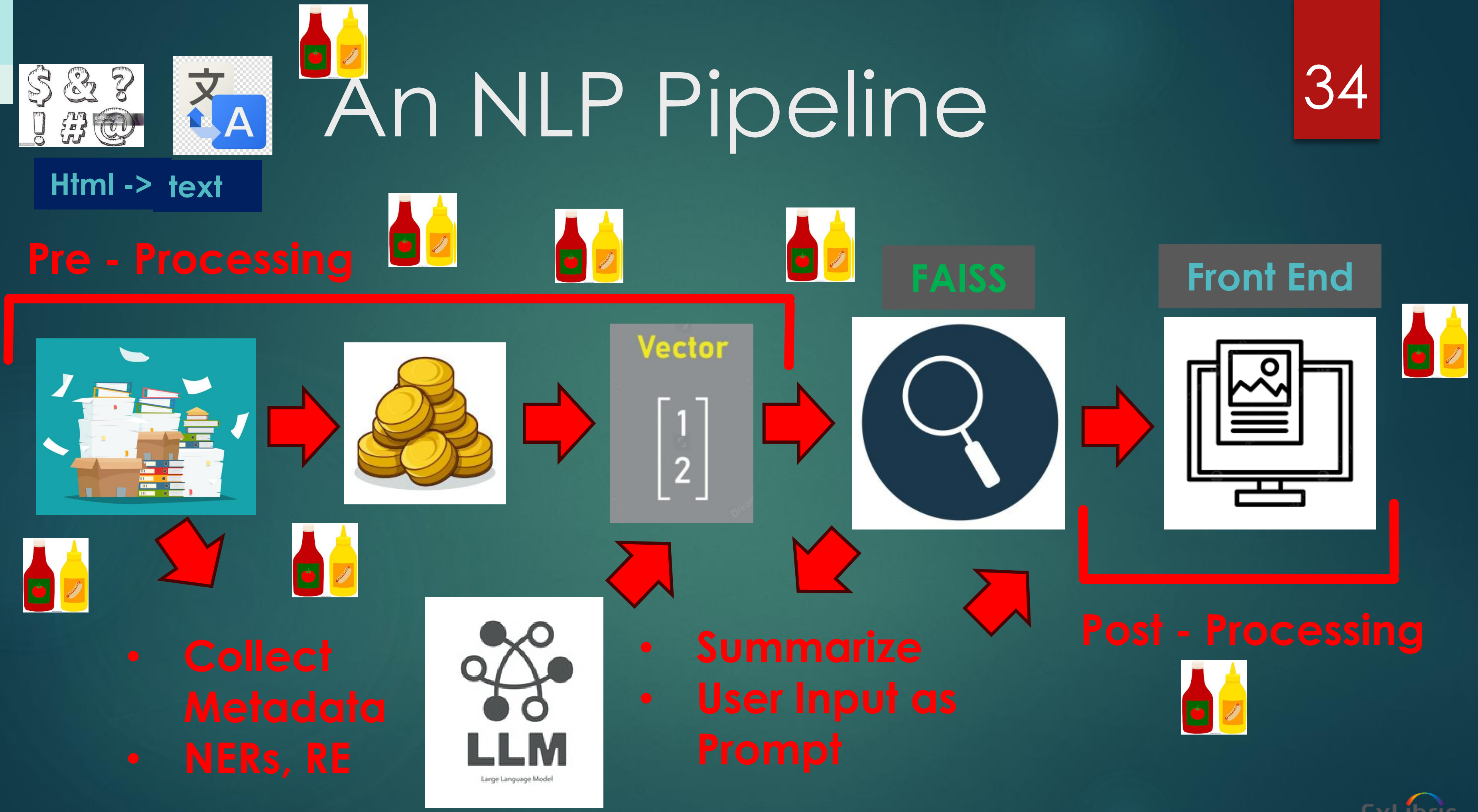


Free

python

```python
text = get_document()
entities = spacy_model(text).ents
llm_prompt = f"""
Summarize the document.
Highlight any organizations and dates: {entities}
"""
```

An NLP Pipeline



Html -> text

Pre - Processing

Vector

[1 2]

FAISS

Front End

- Collect Metadata
- NERs, RE

LLM

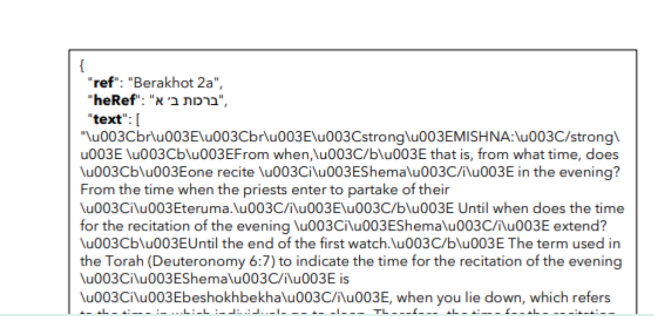
Large Language Model

- Summarize
- User Input as Prompt

Post - Processing

Group Activity

In groups of 1-3,
Design your own NLP pipeline.
See attached sheets



{
  "ref": "Berakhot 2a",
  "heRef": "ברכות ב׳ א",
  "text": [
"\u003Cbr\u003E\u003Cbr\u003E\u003Cstrong\u003EMISHNA:\u003C/strong\u003E \u003Cb\u003EFrom when,\u003C/b\u003E that is, from what time, does \u003Cb\u003Eone recite \u003Ci\u003EShema\u003C/i\u003E in the evening? From the time when the priests enter to partake of their \u003Ci\u003Eteruma.\u003C/i\u003E\u003C/b\u003E Until when does the time for the recitation of the evening \u003Ci\u003EShema\u003C/i\u003E extend? \u003Cb\u003EUntil the end of the first watch.\u003C/b\u003E The term used in the Torah (Deuteronomy 6:7) to indicate the time for the recitation of the evening \u003Ci\u003EShema\u003C/i\u003E is \u003Ci\u003Ebeshokhbekha\u003C/i\u003E, when you lie down, which refers to the time in which individuals go to sleep. Therefore, the time for the recitation

Thank you!

Questions?

David Berger, 2025
Software Engineer
Rapido, Ex libris

ExLibris Part of Clarivate