



Session 1

Introduction

Azam Rabiee, PhD

August 16, 2020

Statistics of Participants

➔ <https://forms.gle/mRxhkQwbRYhtAyEw9>

NLP Participants Survey

You have registered in "Basics of NLP" event, sponsored by Digikala Academy. We appreciate you to take 2 minutes to fill out this form so that we know the average level of participants. With this knowledge, the content and pace of presentation can be customized for more efficiency.

Thanks!

* Required



Session 1

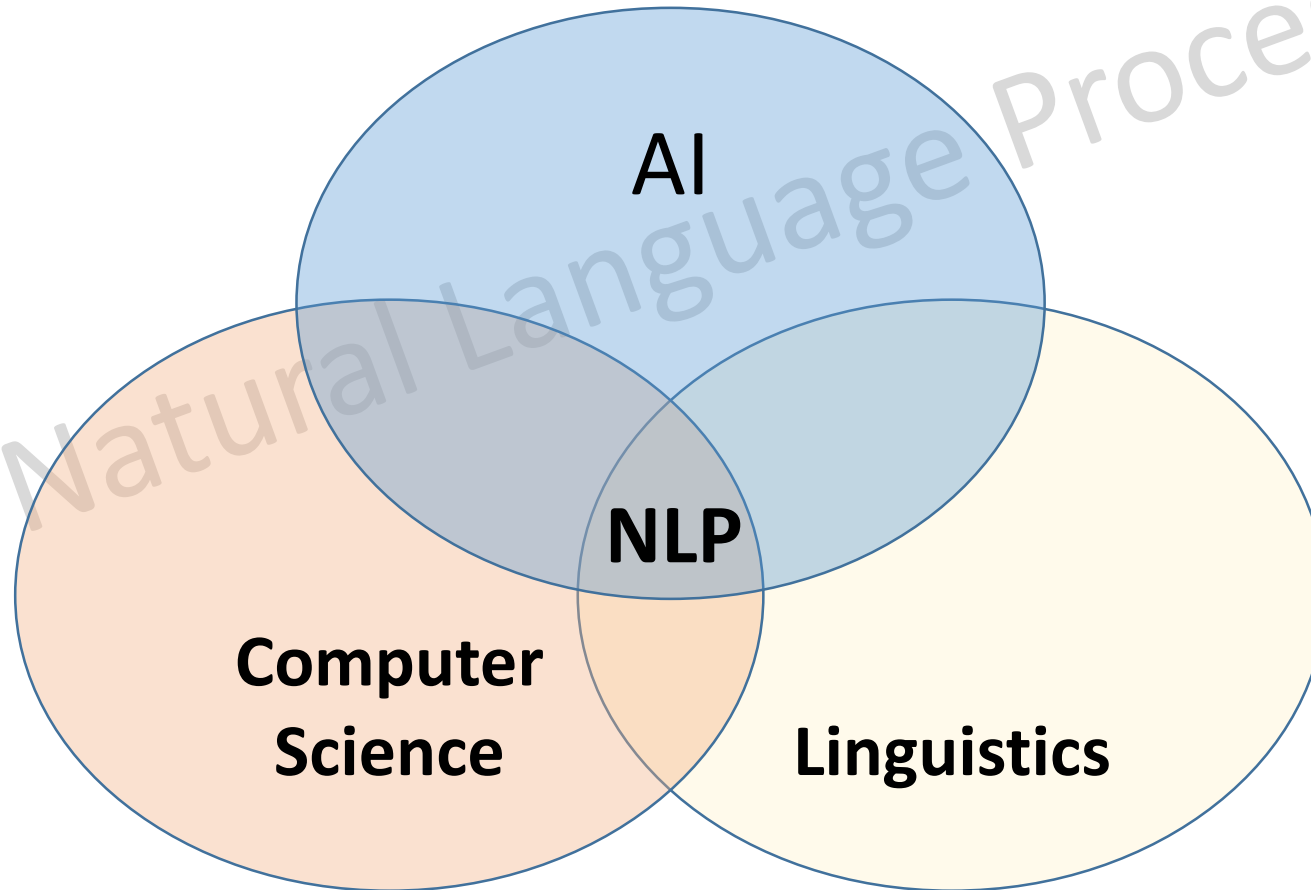
Introduction

Azam Rabiee, PhD

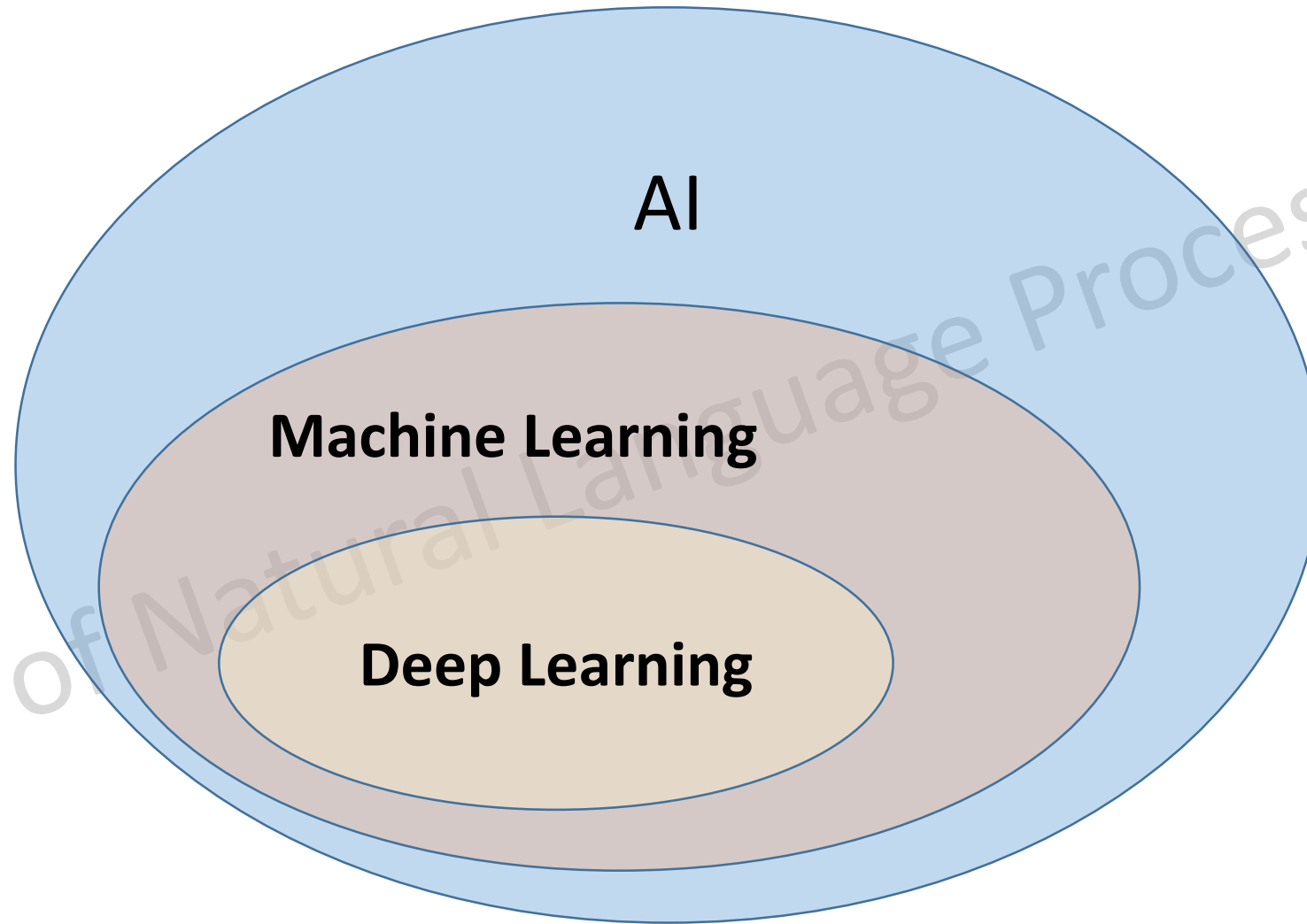
August 16, 2020

What is NLP?

The interdisciplinary field of computer science, AI, and linguistics.
The ability for computers to understand human language.



Artificial Intelligence



Outline

We Review ML/DL Approaches for NLP in three events:

Basic

- **Session 1.** Introduction
- **Session 2.** Basics of Linguistics
- **Session 3.** Basics of ML
- **Session 4 (Lab).** Effective Word Representation by python

Intermediate

- TBA

Advanced

- TBA

NLP Basics

Session 1: Introduction

- Applications
- Tasks
- Approaches

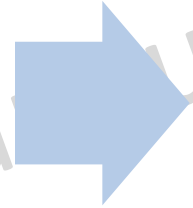
Session 2. Basics of Linguistics

Session 3. Basics of ML

Session 4 (Lab). Effective Word Representation by python

Machine Translation

I am a student.



Je suis étudiant.

Sentiment Analysis



The movie was good.



The movie was made in 1981.

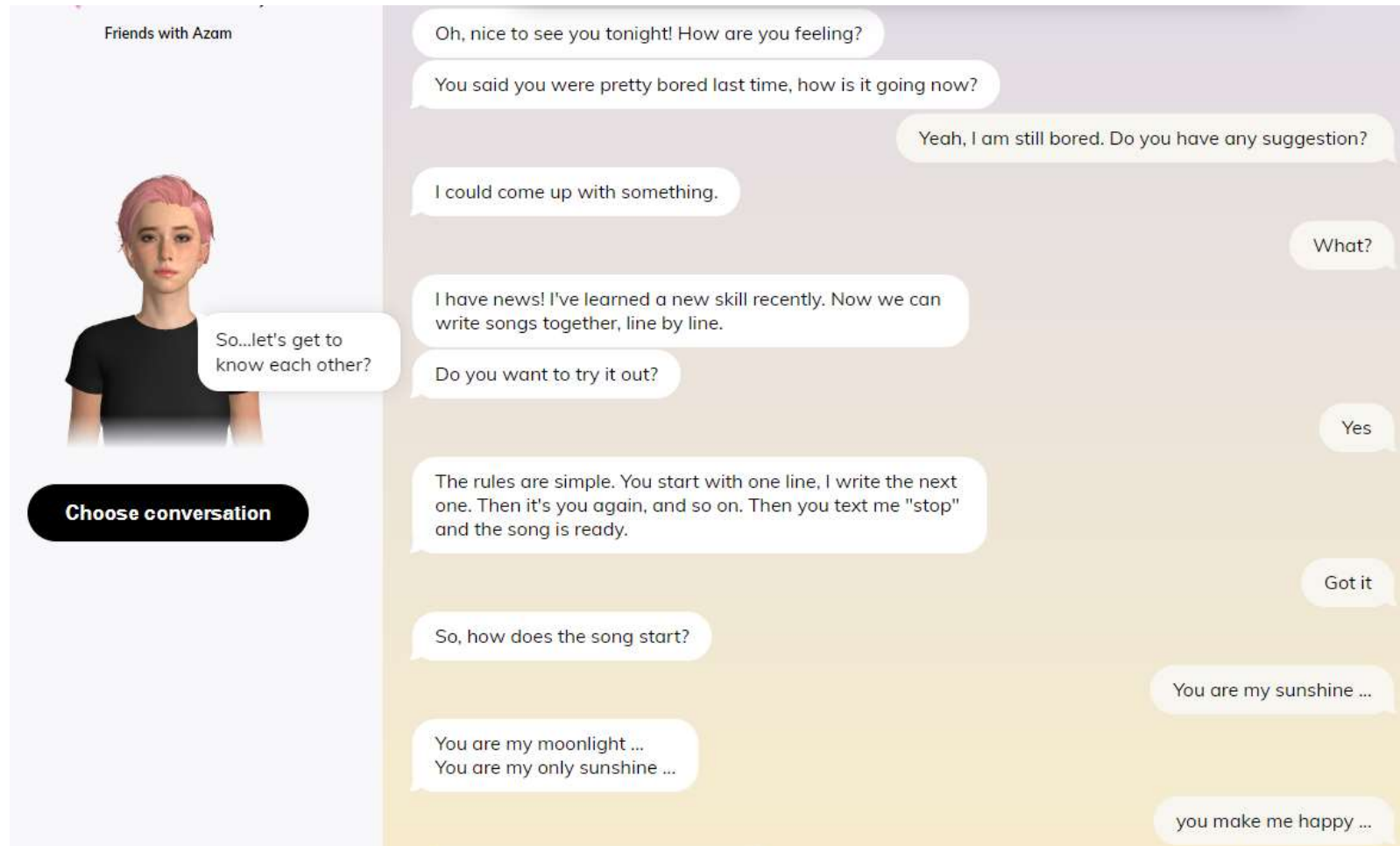


The movie was bad.

Applications

Chatbot

<https://replika.ai/>



Text/Document Summarization

"The **Army Corps of Engineers**, rushing to meet **President Bush**'s promise to protect New Orleans by the start of the 2006 hurricane season, installed defective flood-control pumps last year despite warnings from its own expert that the equipment would fail during a storm, according to documents obtained by The Associated Press"

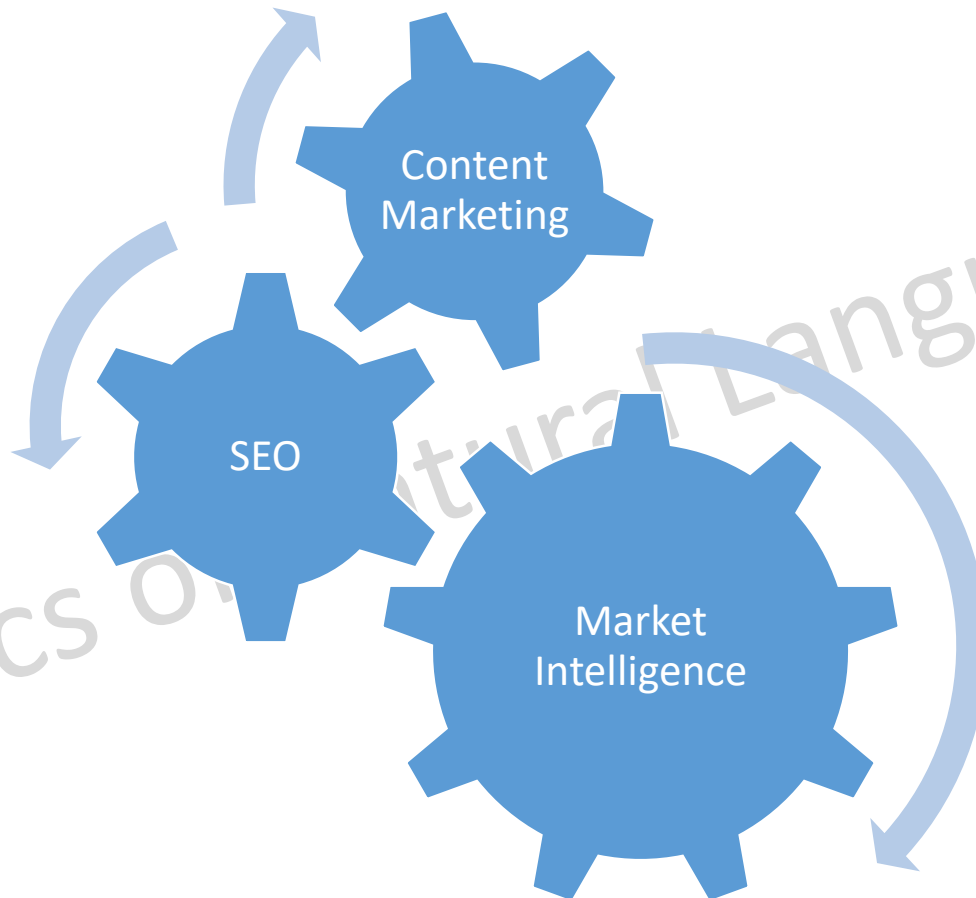
Keyphrase Extraction

- "Army Corps of Engineers"
- "President Bush"

Abstraction

- "inadequate protection from floods"
- "political negligence"

AI Marketing

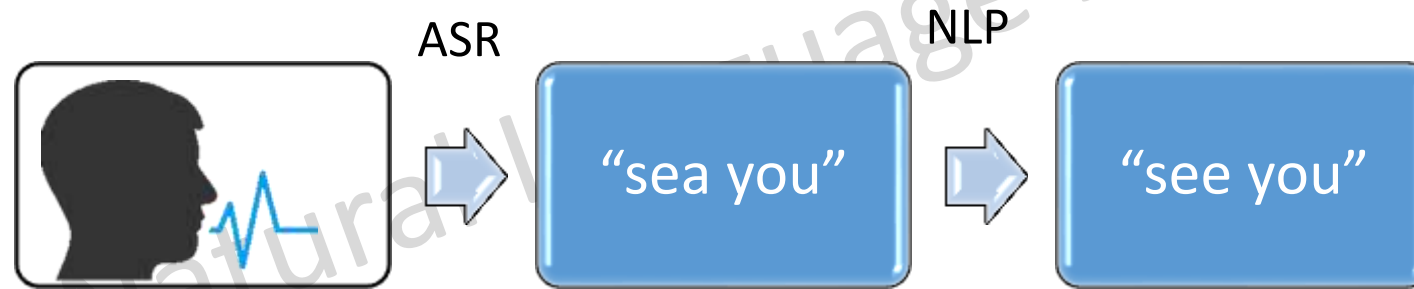


- Creating, publishing, and distributing content for a targeted audience
- Recommender systems
- CRM chatbots
- ...

Text/Document Classification



Speech Recognition



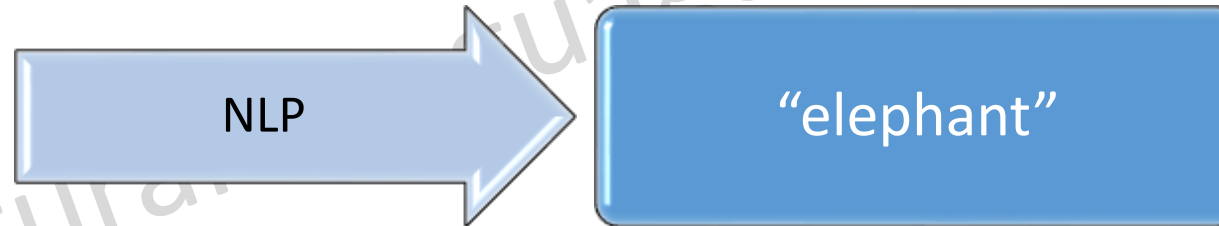
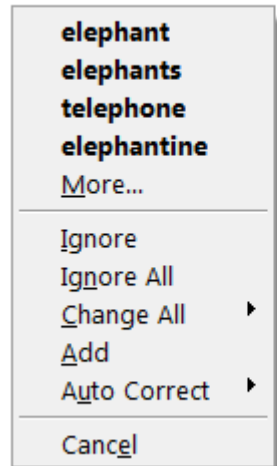
homophone correction

Handwritten Character Recognition



Spell Checking

The **elephante** enjoyed the peanuts.



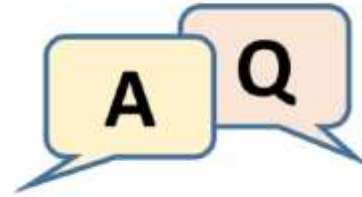
Applications



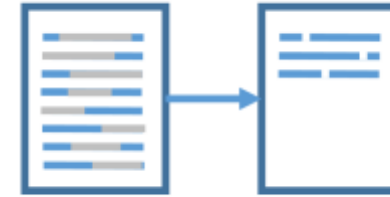
Machine Translation



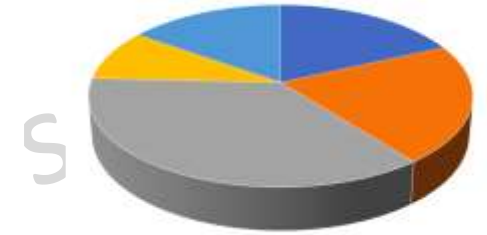
Sentiment Analysis



Question Answering



Automatic Summarization



AI Marketing



Text / Document Classification



Speech Recognition



↓
Give me

Handwritten Character Recognition



Spell Checking



Which one is not of NLP applications?

1. Email spam filtering
2. Medical image analysis
3. Smart assistants, such as smart speakers
4. Search engines

NLP Basics

Session 1: Introduction

- Applications
- **Tasks**
- Approaches

Session 2. Basics of Linguistics

Session 3. Basics of ML

Session 4 (Lab). Effective Word Representation by python

Language Modeling

the task of assigning a probability to sentences in a language

$$p(w_1, w_2, \dots, w_m)$$

Natural Language Generation (NLG)

Can you please come **here?**



History



predicted word

Natural Language Understanding (NLU)

machine reading comprehension
an AI-hard problem

Q: Which NLP applications need NLU?

Natural Language Inference (NLI)

the task of determining whether a “hypothesis” is true, false, or undetermined given a “premise”.

Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	False	The man is sleeping.
An older and younger man smiling.	Undetermined	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	True	Some men are playing a sport.

Information Extraction

the task of automatically extracting structured information from documents

Basics of Natural Language Processing



Some NLP tasks





Named Entity Recognition (NER), seeking to locate and classify named entities into pre-defined categories, is the task of:

1. Language Modeling
2. Information Extraction
3. NLU
4. NLG
5. NLI

Jim bought 300 shares of Acme Corp. in 2006.

[Jim] bought 300 shares of [Acme Corp.] in [2006].

Person

Organization

Time



Identify duplicate questions for Q/A and chatbot applications is the task of:

1. Language Modeling
2. Information Extraction
3. NLU
4. NLG
5. NLI

How old are you? = What is your age?

Where are you from?

≠

Where are you going?

NLP Basics

Session 1: Introduction

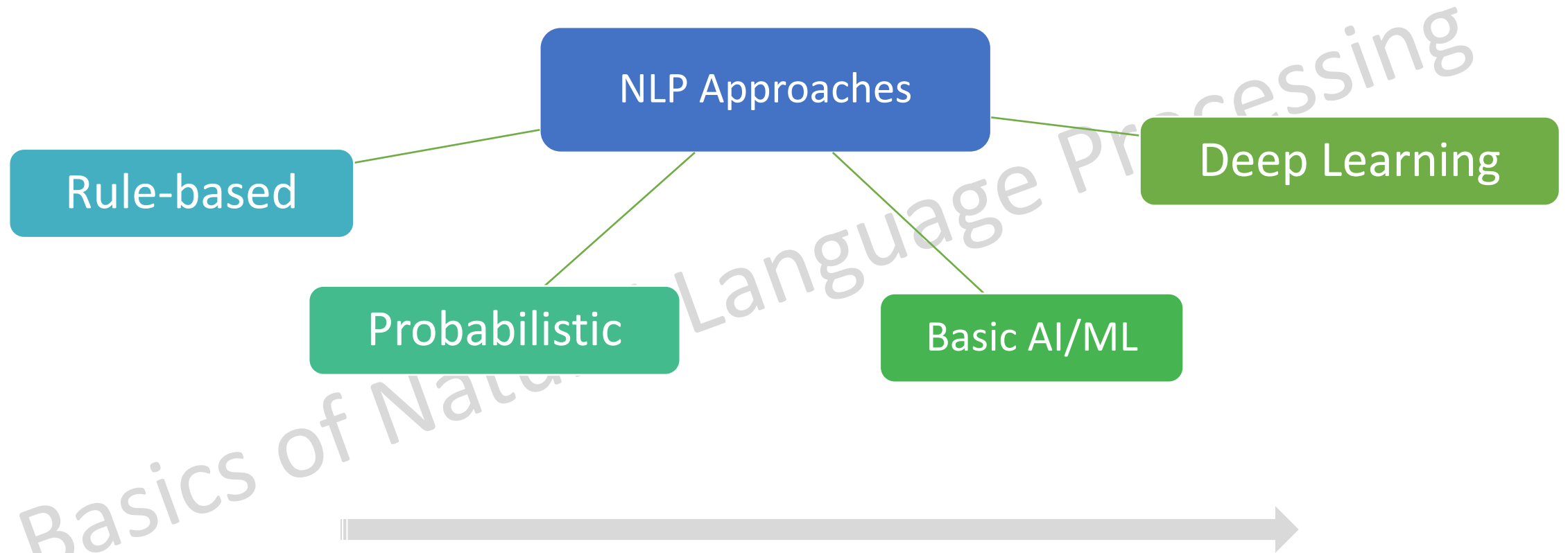
- Applications
- Tasks
- **Approaches**

Session 2. Basics of Linguistics

Session 3. Basics of ML

Session 4 (Lab). Effective Word Representation by python

Approaches





Sentiment Analysis



The movie was good.



The movie was made in 1981.



The movie was bad.

Approaches

Dictionary-based

Naïve Bayes

Regression

ML / DL



Sentiment Analysis

Note: Primary approaches are unable to analyze implicit sentiment.



The movie was good.

Explicit

The movie was made in 1981.

Implicit



The movie was bad.

Explicit



Sentiment Analysis

Note: Primary approaches are unable to analyze implicit sentiment.



The movie was good.

Explicit

The movie was made in 1981.

Implicit

The movie was made in 1981.
I never miss old movies.



The movie was bad.

Explicit



Sentiment Analysis

Note: Primary approaches are unable to analyze implicit sentiment.



The movie was good.

The movie was made in 1981.

Explicit

Implicit

Explicit



The movie was bad.

Context

- The movie was made in 1981. I never miss old movies.

Sarcasm

- Did you enjoy the shopping?
- YES, SURE! With two annoying kids!

Comparison

- This is better than nothing



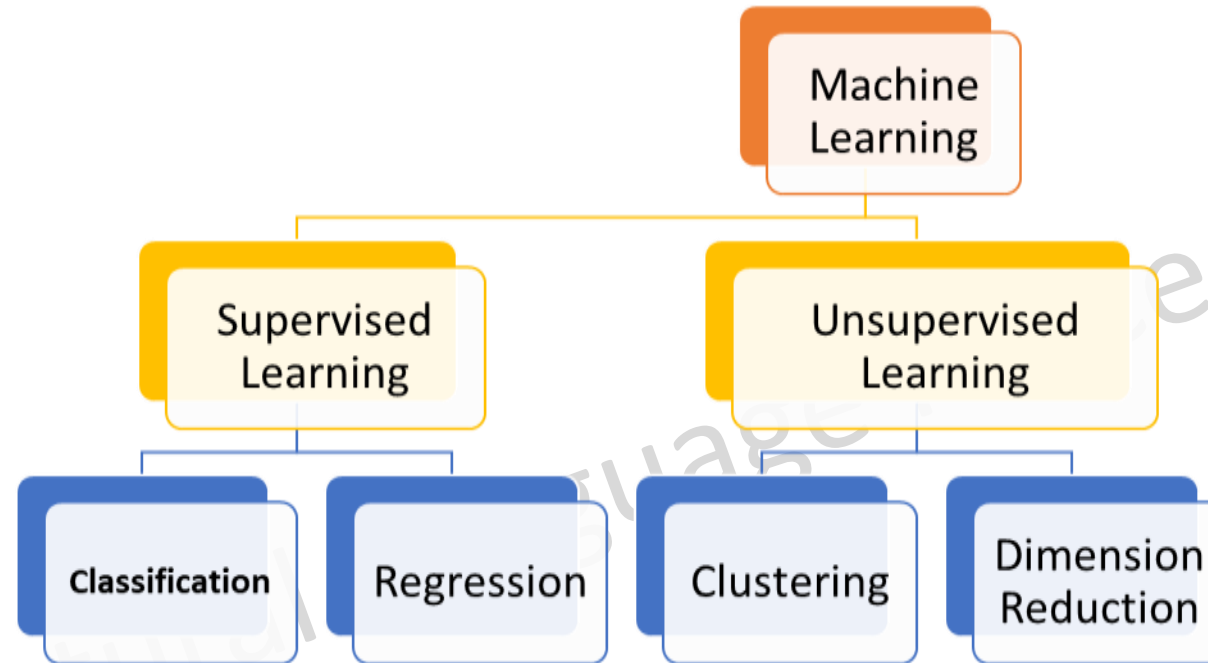
ML / DL methods are of the state-of-the-art approaches for NLP

Basics of Natural Language Processing

Taxonomy of ML Methods

ML Styles

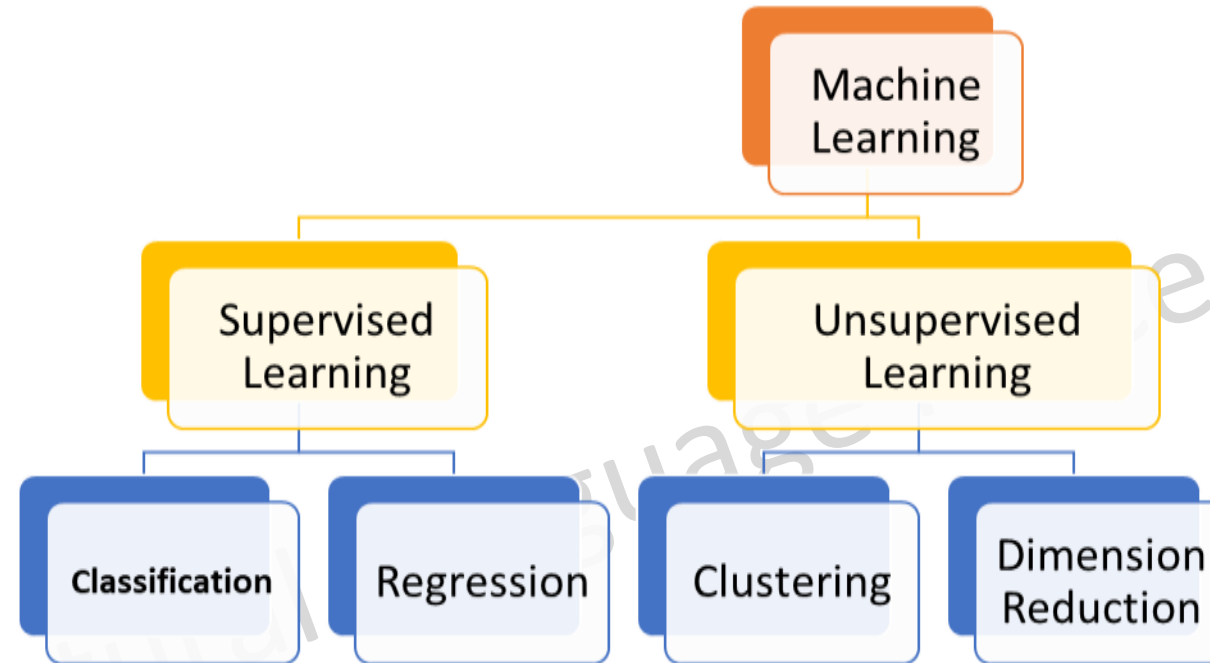
Tasks



Taxonomy of ML Methods

ML Styles

Tasks



Label/Target

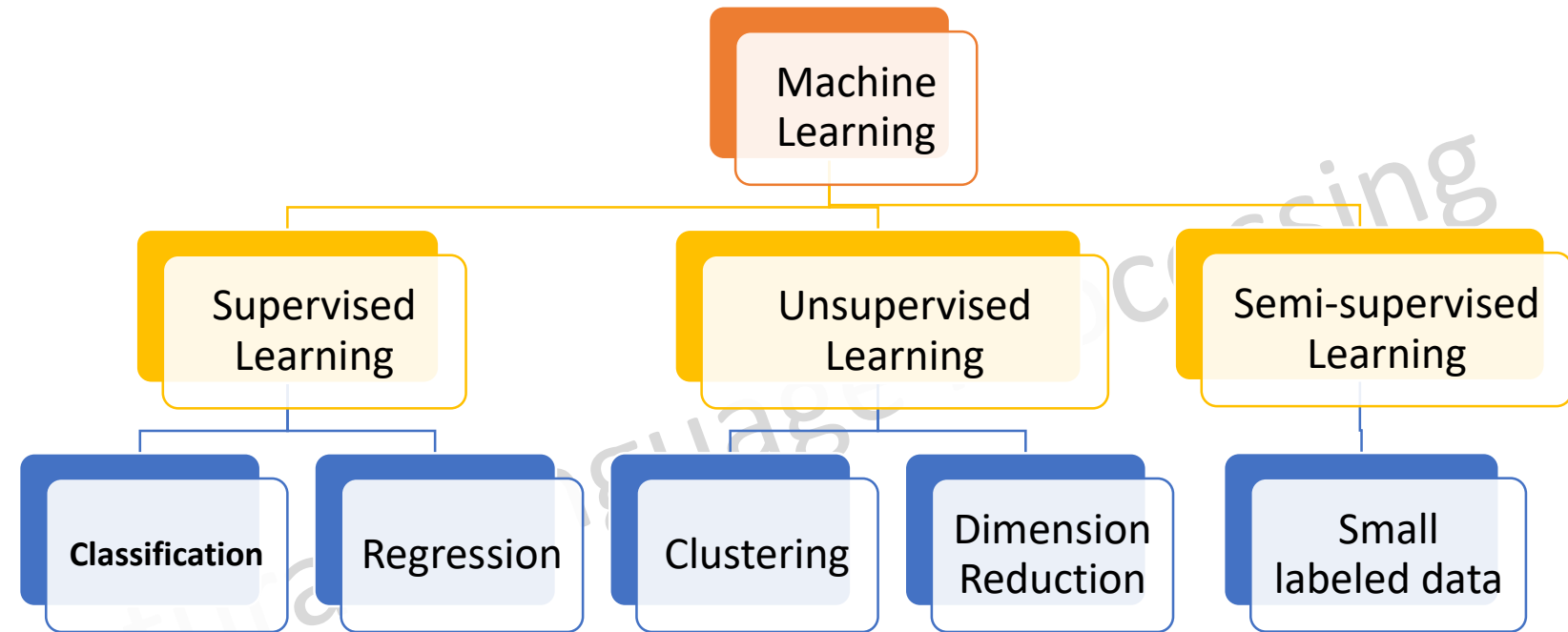
↓

user ID	time	price (\$)	purchased
4783	Jan 21 08:15.20	7.95	yes
3893	March 3 11:30.15	10.00	yes
8384	June 11 14:15.05	9.50	no
0931	Aug 2 20:30.55	12.90	yes

Taxonomy of ML Methods

ML Styles

Tasks



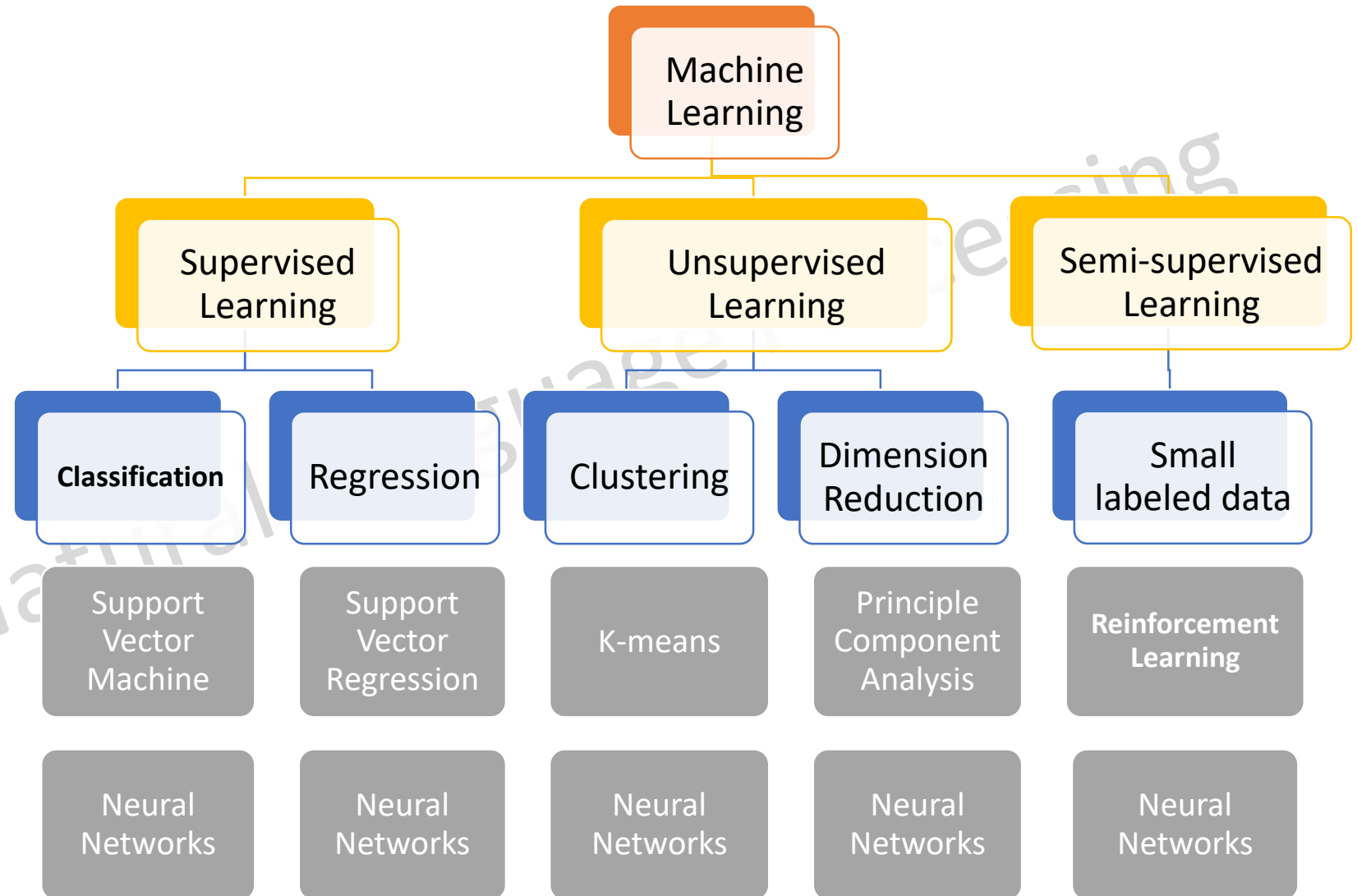
Semi-supervised learning is an approach to [machine learning](#) that combines a small amount of [labeled data](#) with a large amount of [unlabeled data](#) during training.

Taxonomy of ML Methods

ML Styles

Tasks

Algorithms

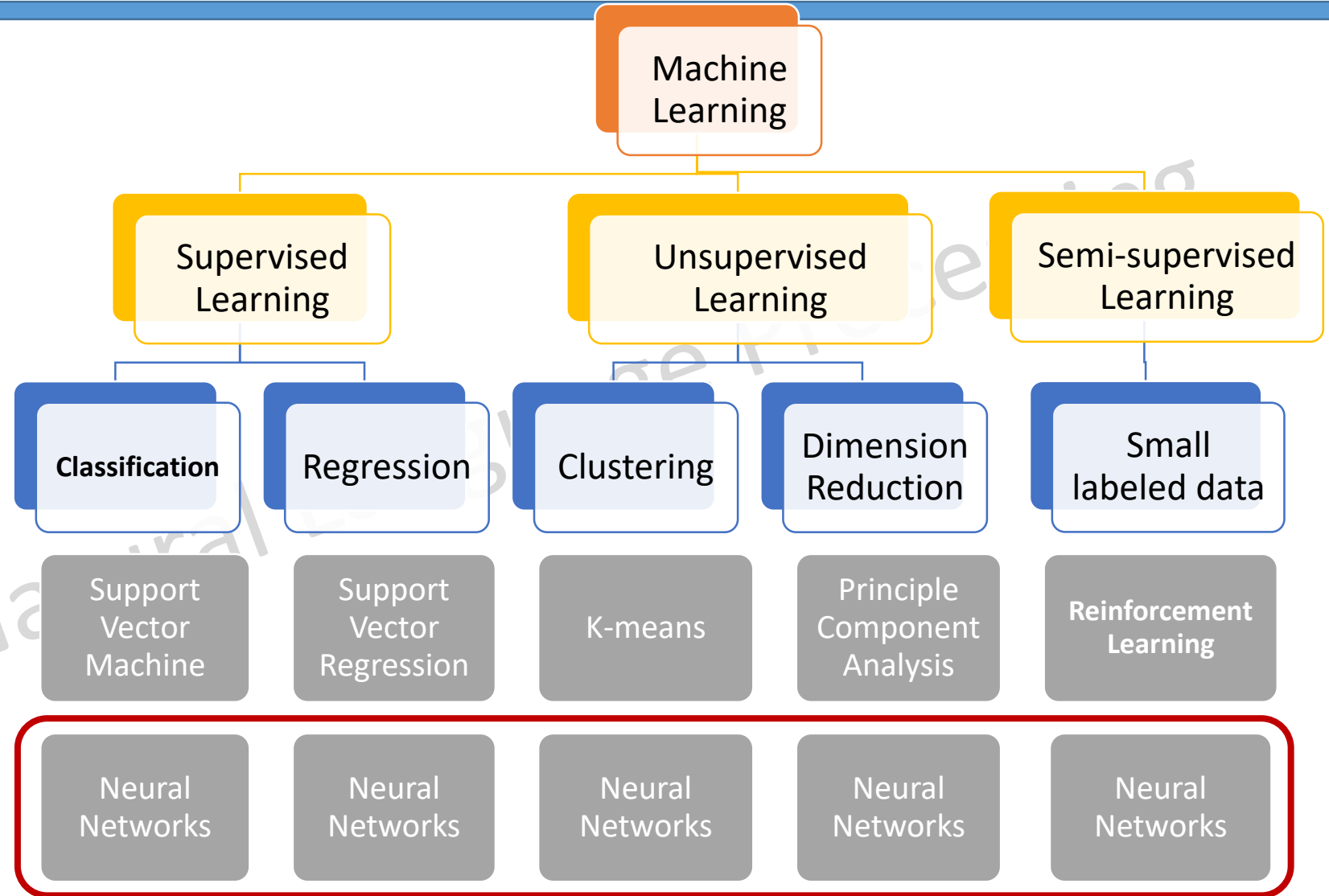


Why Neural Networks?

ML Styles

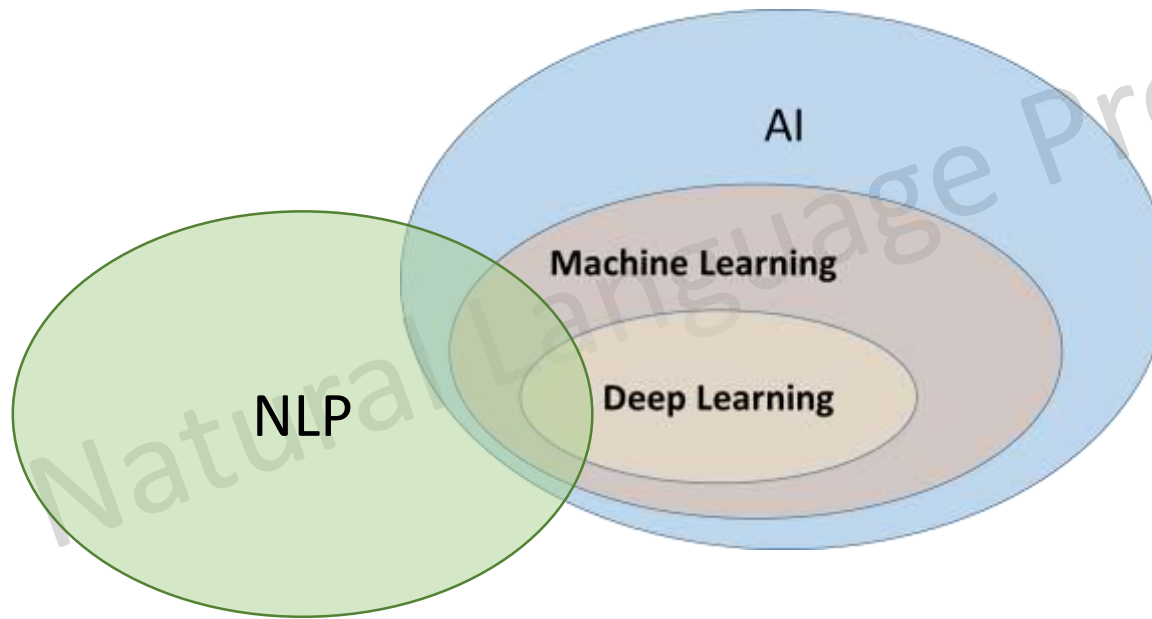
Tasks

Algorithms



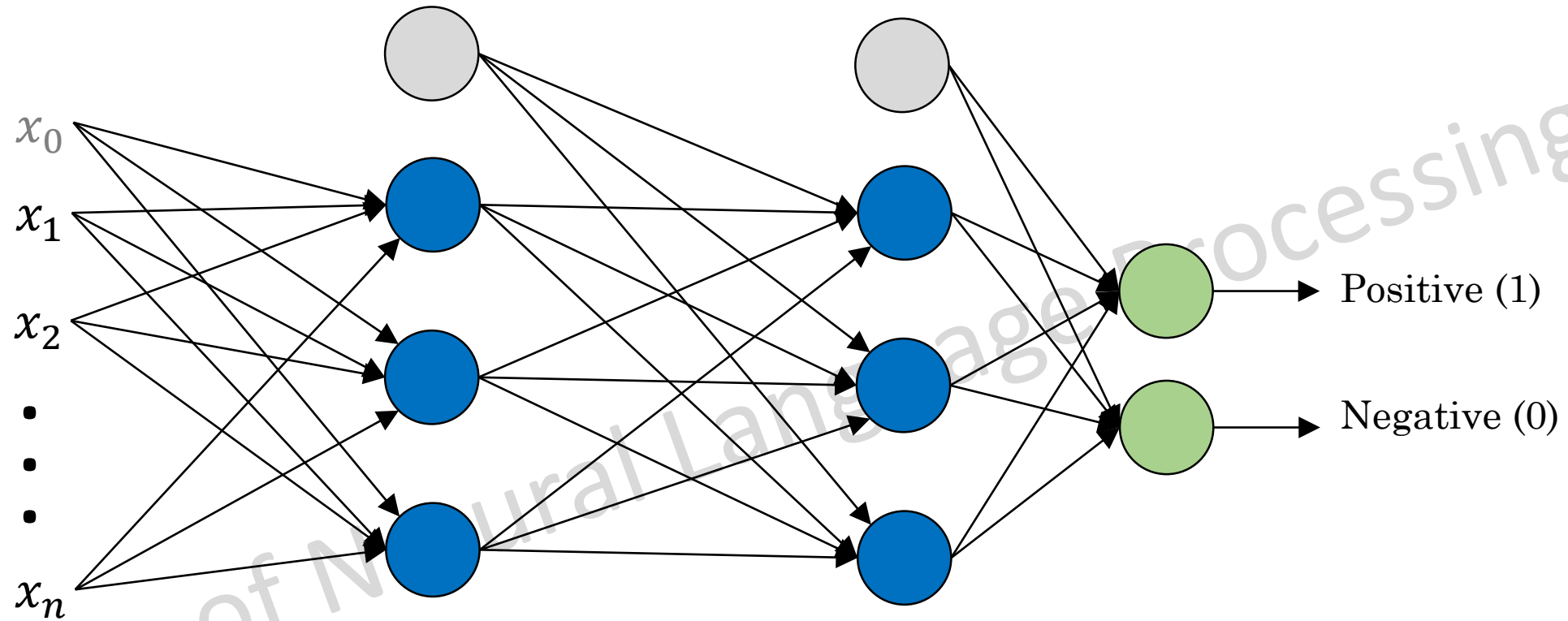
Recall: Roadmap

We will review ML/DL methods for NLP





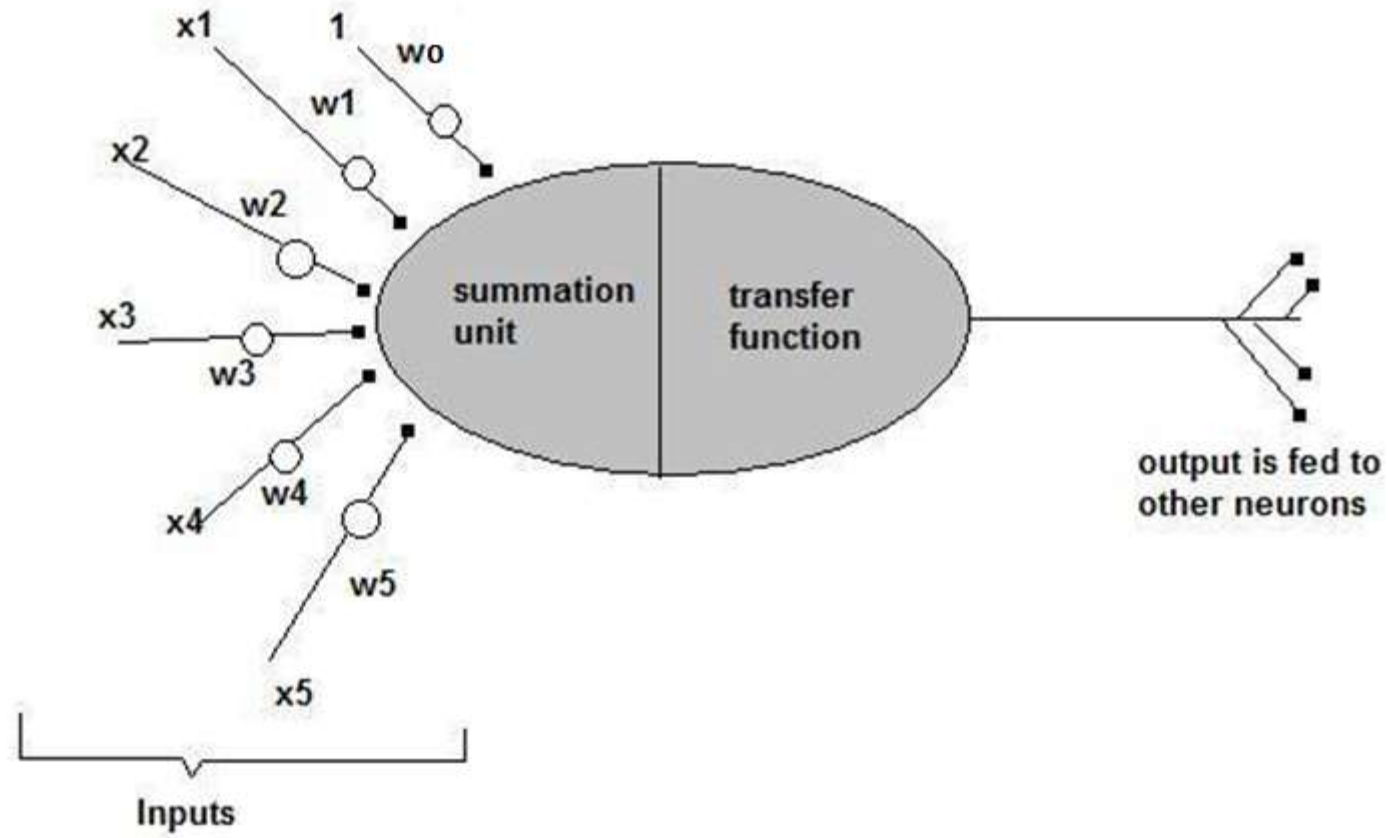
ANN for Sentiment Analysis



$\mathbf{x} = [x_1, x_2, \dots, x_n]$: tweets, such as "This movie was almost good"

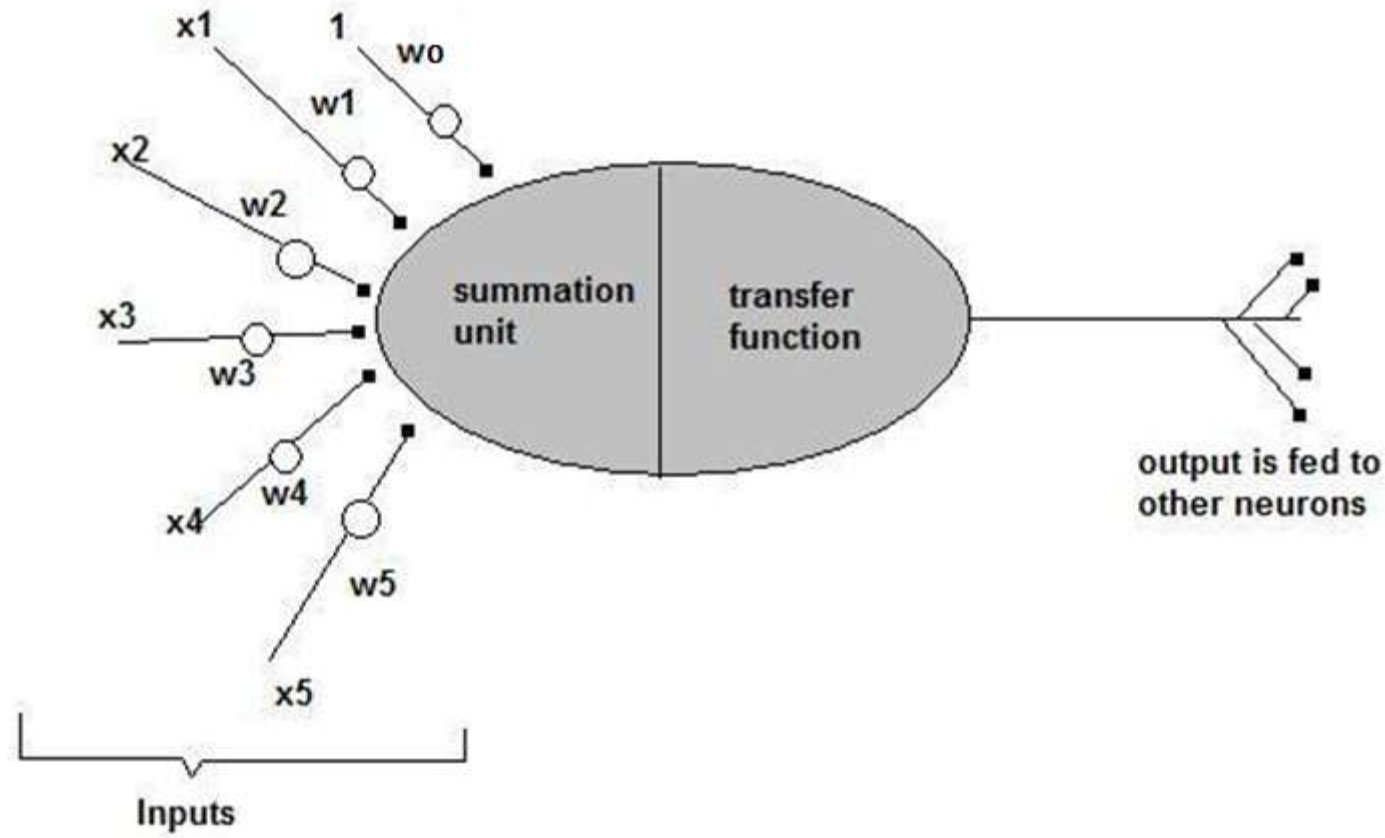
A Neuron model

A Single Neuron



A Neuron model

A Single Neuron



$$z = \sum_i w_i x_i$$

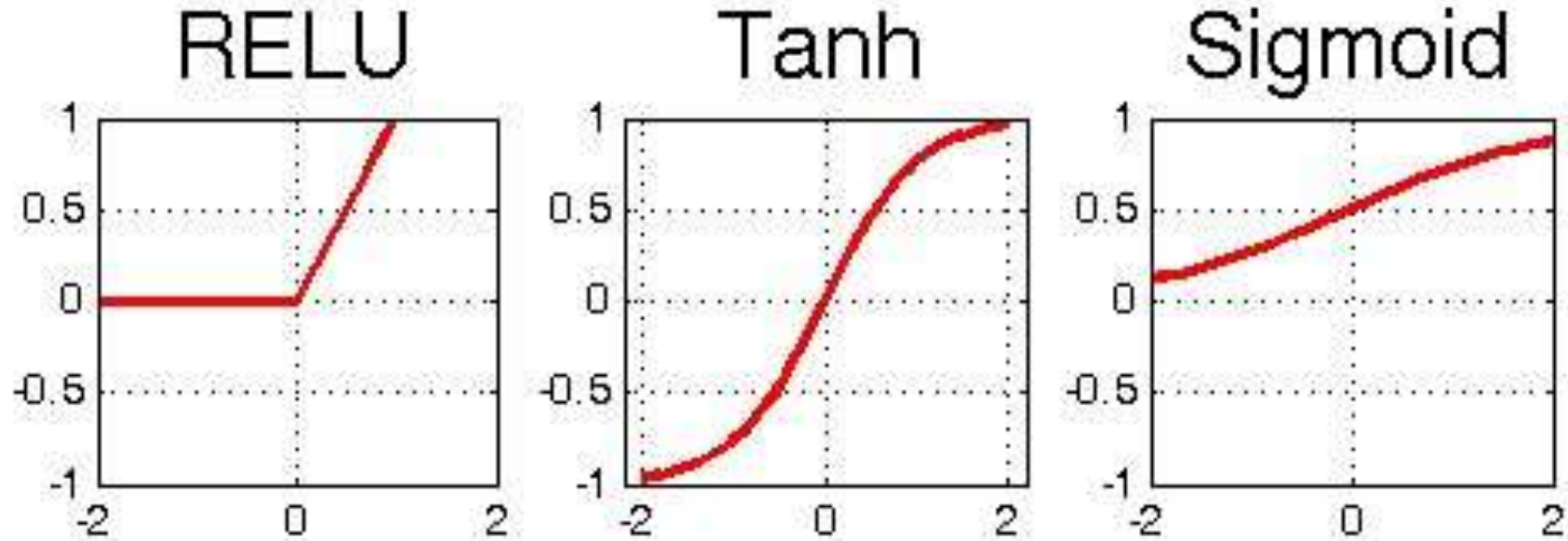
$$a = f(z)$$

f : activation/transfer function

a : activation of the neuron

Basics

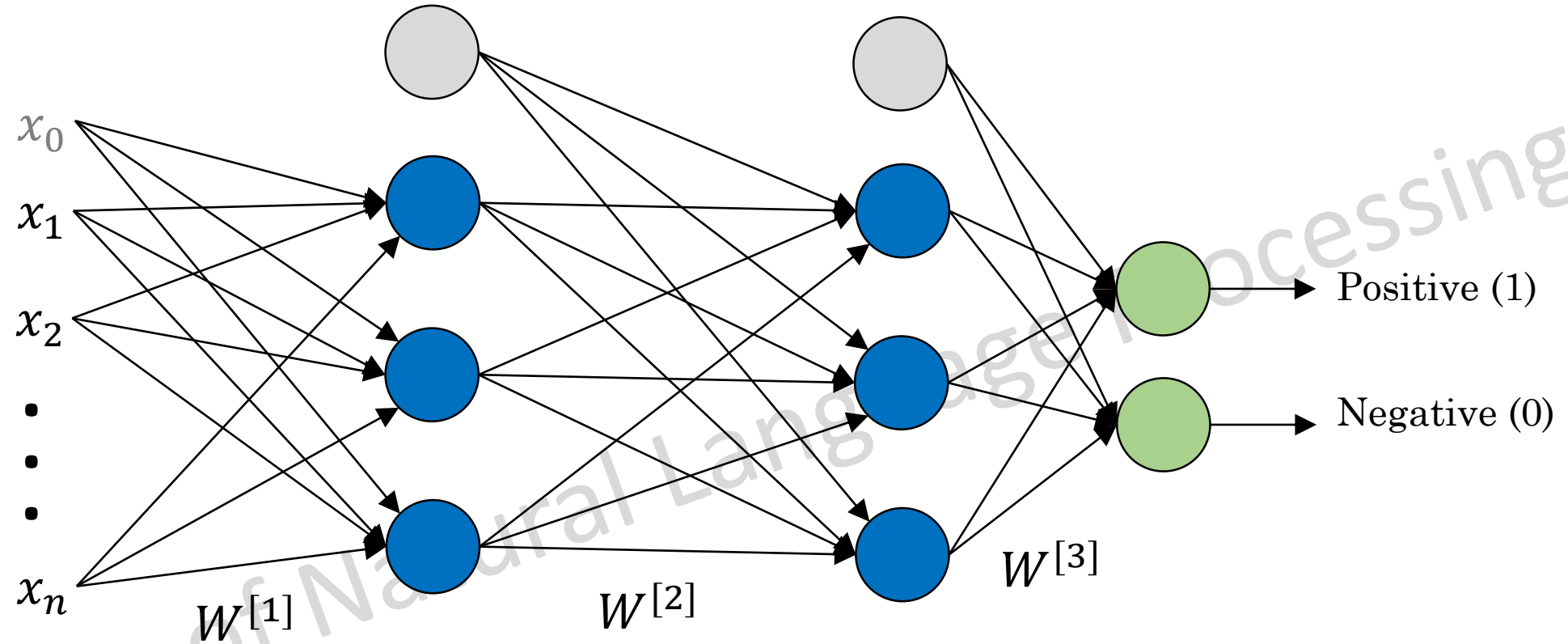
Activation Functions



[Hamdan, M. K. (2018). VHDL auto-generation tool for optimized hardware acceleration of convolutional neural networks on FPGA (VGT)]



ANN for Sentiment Analysis: Forward Propagation



$$\mathbf{a}^{[0]} = \mathbf{x}$$

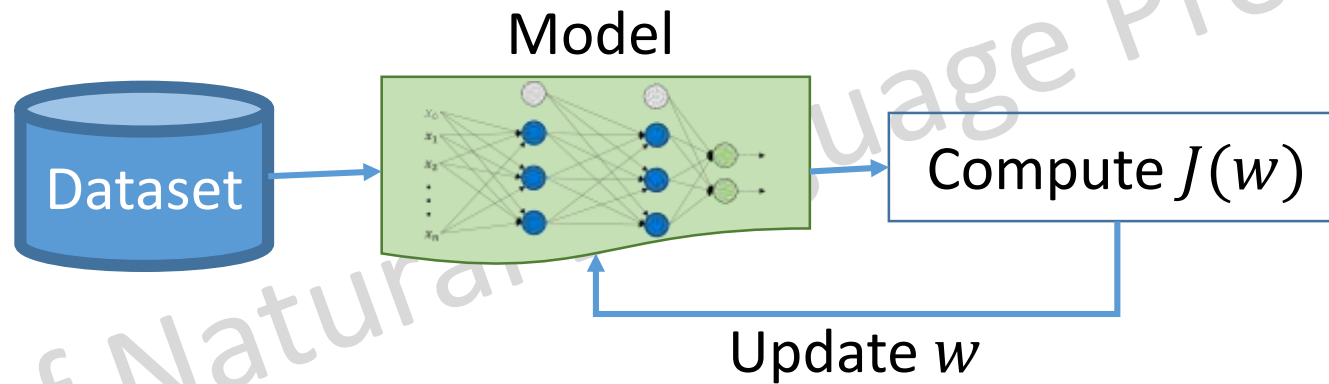
$$\mathbf{z}^{[i]} = \mathbf{W}^{[i]} \mathbf{a}^{[i-1]}$$

$$\mathbf{a}^{[i]} = f(\mathbf{z}^{[i]})$$

$$\mathbf{y} = \mathbf{a}^{[3]}$$

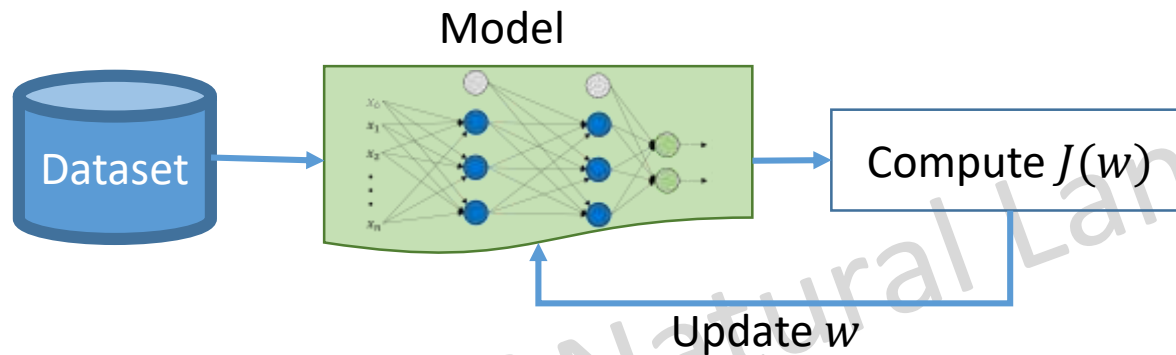
ANN for Sentiment Analysis: Train

Train: Finding optimum **parameters** that minimize a desired **cost function** $J(w)$.



ANN for Sentiment Analysis: Train

Train: Finding optimum parameters that minimize a desired cost function $J(w)$.



Sample **cost function**: **MSE**

$$\frac{1}{2}(\hat{y} - y)^2$$

y : Ideal target

\hat{y} : Network output

Sample **optimization Algorithm**: **GD**

$$w(t) = w(t - 1) - \eta \nabla J(w)$$

η : Learning rate



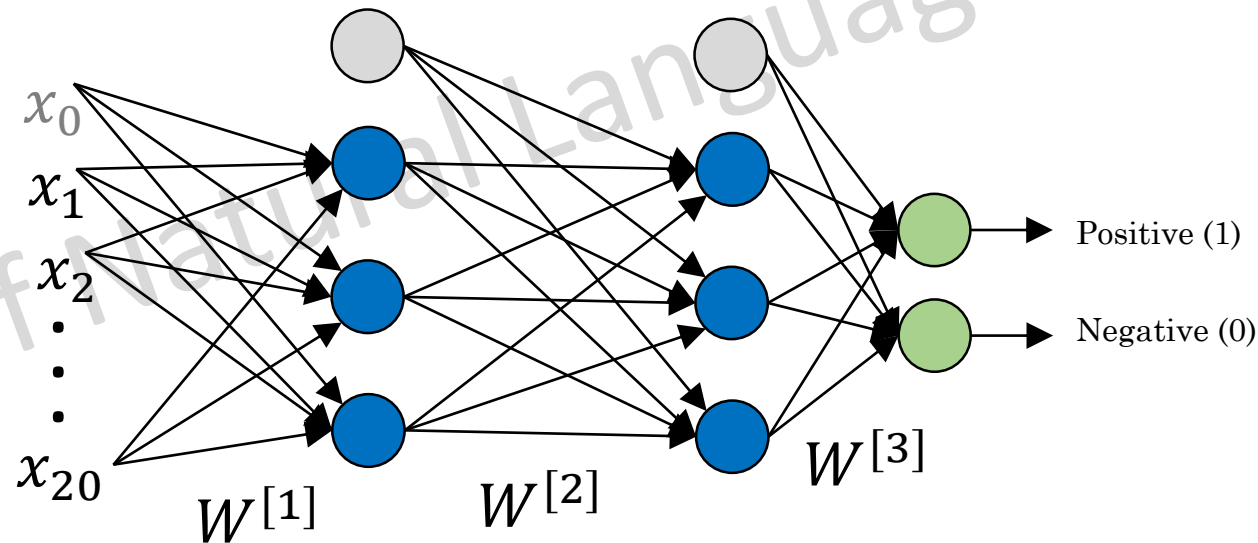
Components of every ML / DL algorithm:

1. A dataset
2. A model
3. A cost function
4. An optimization algorithm

QUIZ



How many trainable parameters does the following network have?





OpenAI GPT-3

Generative Pre-trained Transformer 3 (GPT-3)

is a language model to produce human-like text,

created by [OpenAI](#),

introduced in May 2020, and is in beta testing as of July 2020

GPT-3 for Information Retrieval



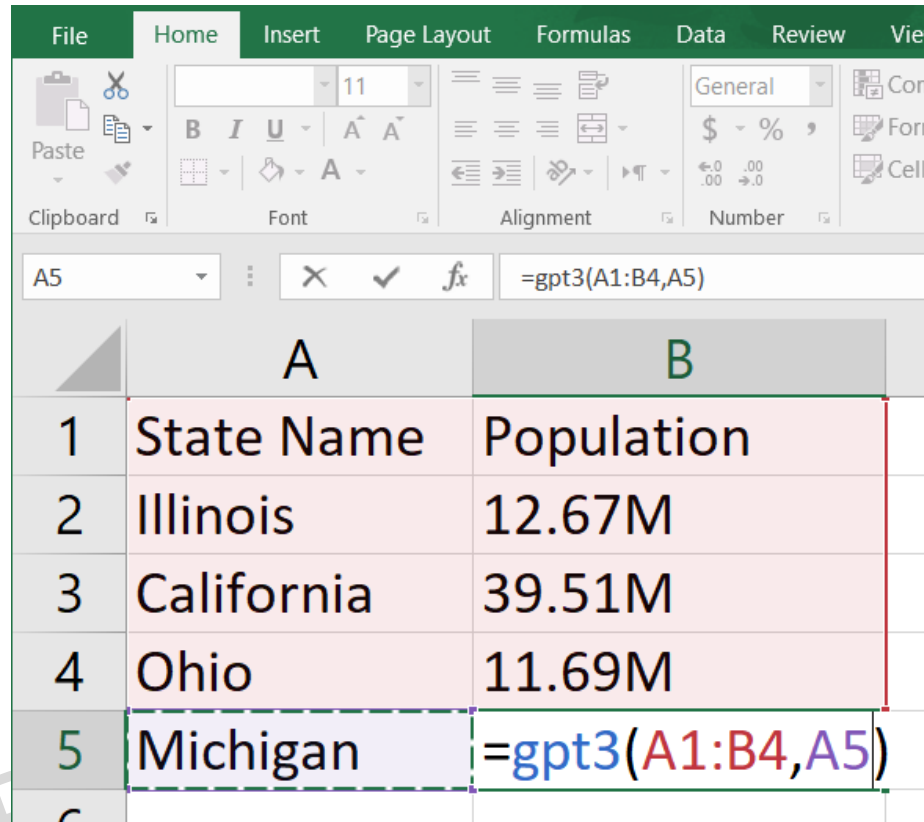
GPT-3 writes SQL Query

Text: Select the "Students" from the "School" table joined with "Class" table:

```
Code: SELECT * FROM Students  
INNER JOIN Class  
ON Students.ID = Class.StudentID
```

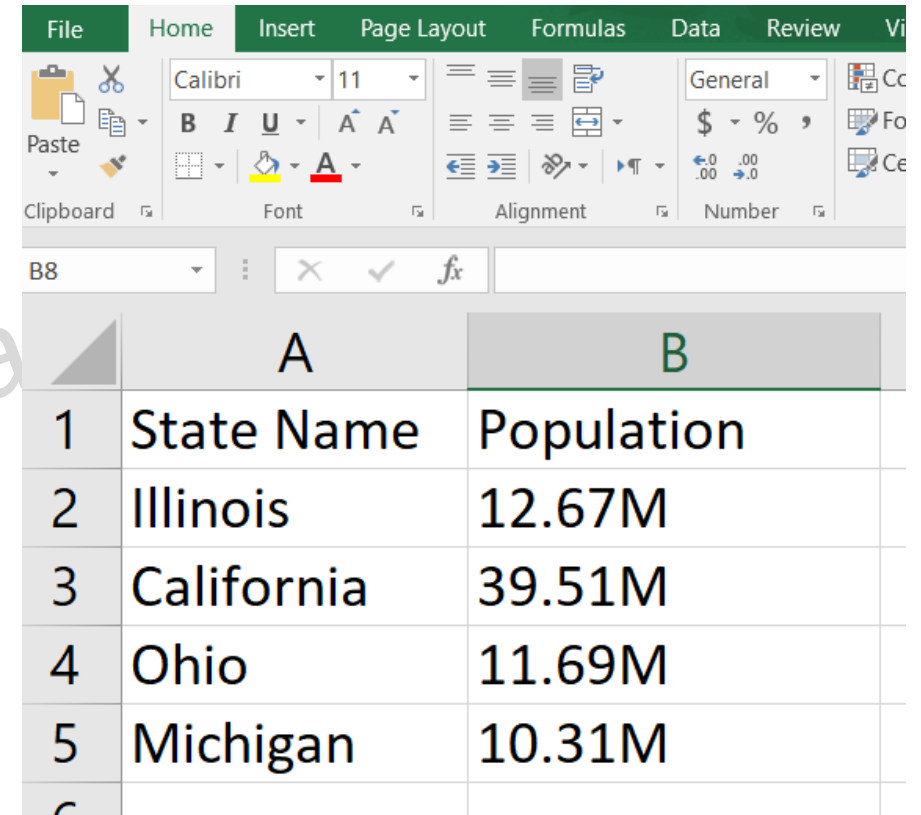
Basics of Natural Language Processing

=GPT3() in Excel



The screenshot shows the Excel interface with the formula bar displaying `=gpt3(A1:B4,A5)`. The table below has columns A and B. Row 1 contains headers 'State Name' and 'Population'. Rows 2-4 contain data for Illinois, California, and Ohio. Row 5 contains 'Michigan' and the formula `=gpt3(A1:B4,A5)`.

	A	B
1	State Name	Population
2	Illinois	12.67M
3	California	39.51M
4	Ohio	11.69M
5	Michigan	<code>=gpt3(A1:B4,A5)</code>



The screenshot shows the result of the GPT3 function. The formula bar is empty, and the table below shows the updated population for Michigan in row 5.

	A	B
1	State Name	Population
2	Illinois	12.67M
3	California	39.51M
4	Ohio	11.69M
5	Michigan	10.31M

GPT-3 Writes codes

debuild.co

Describe your app.

Clear

Generate

an input that says "Enter a todo" and a button that says "Save todo". then show me all my todos



debuild.co

Describe your app.

Clear

Generate

Just describe your app!

```
// an input that says "Enter a todo"
and a button that says "Save todo".
then show me all my todos
class App extends React.Component {

  constructor(props) {

    super(props)

  }

}
```

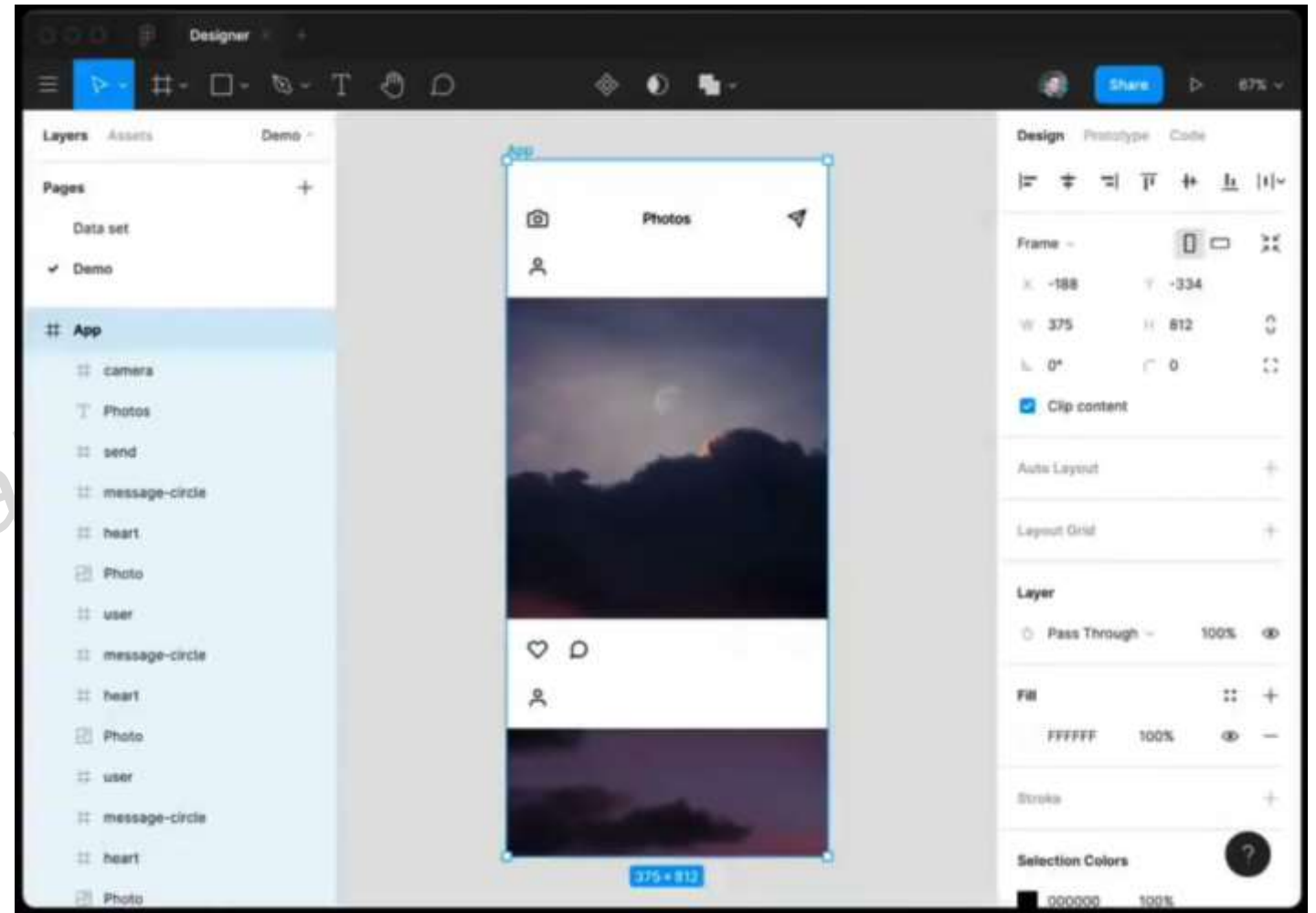
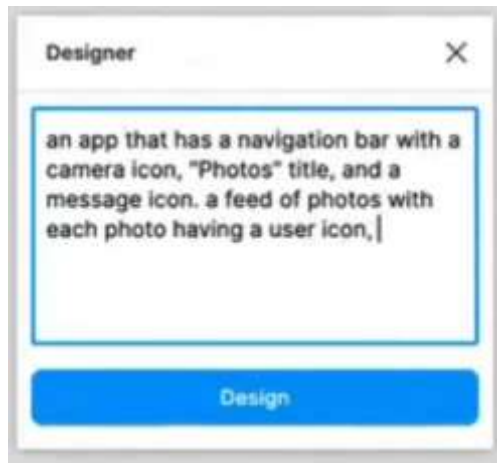
Enter a todo

Save todo

learn about ai

WHAT

GPT-3 Designs UI



GPT-3 as a Search Engine

ask me
anything

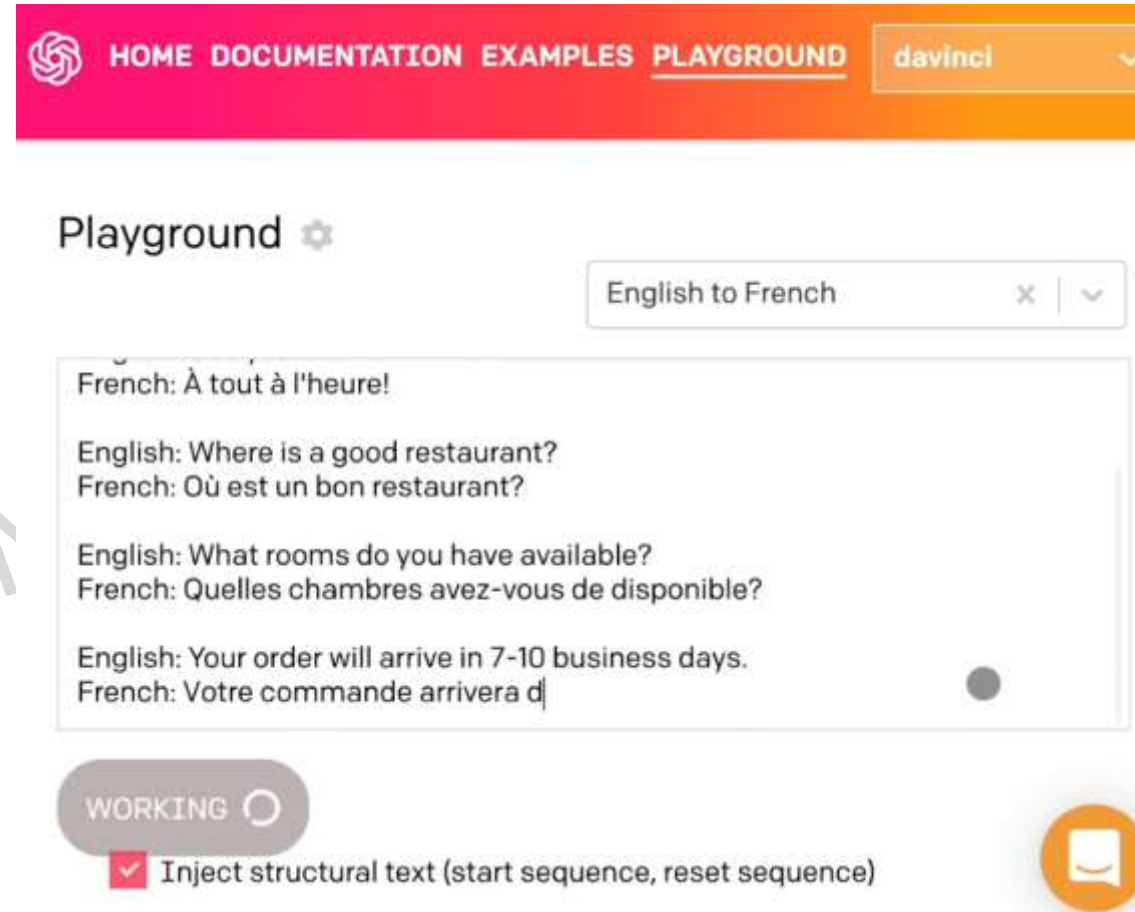
Who killed Mahamta Gandhi?

Seek


Nathuram Godse killed Mahamta Gandhi.


[More info on this / Show more results](#)


GPT-3 Translates





GPT-3 Summarizes


 [HOME](#) [DOCUMENTATION](#) [EXAMPLES](#) [PLAYGROUND](#)


davinci 

Playground 

Summarize for a 2nd grader  

warning. If Confidential Information is transmitted orally, the Disclosing Party shall promptly provide writing indicating that such oral communication constituted Confidential Information. 2. Exclusions from Confidential Information. Receiving Party's obligations under this Agreement do not extend to information that is: (a) publicly known at the time of disclosure or subsequently becomes publicly known through no fault of the Receiving Party; (b) discovered or created by the Receiving Party before disclosure by Disclosing Party; (c) learned by the Receiving Party through legitimate means other than from the Disclosing Party or Disclosing Party's representatives; or (d) is disclosed by Receiving Party with Disclosing Party's prior written approval.

Submit 

☒ Inject structural text (start sequence, reset sequen 

Language Models are Few-Shot Learners

Tom B. Brown*	Benjamin Mann*	Nick Ryder*	Melanie Subbiah*	
Jared Kaplan†	Prafulla Dhariwal	Arvind Neelakantan	Pranav Shyam	Girish Sastry
Amanda Askell	Sandhini Agarwal	Ariel Herbert-Voss	Gretchen Krueger	Tom Henighan
Rewon Child	Aditya Ramesh	Daniel M. Ziegler	Jeffrey Wu	Clemens Winter
Christopher Hesse	Mark Chen	Eric Sigler	Mateusz Litwin	Scott Gray
Benjamin Chess	Jack Clark	Christopher Berner		
Sam McCandlish	Alec Radford	Ilya Sutskever	Dario Amodei	

OpenAI

<https://arxiv.org/pdf/2005.14165.pdf>

David Chalmers, an Australian philosopher, described GPT-3 as "one of the most interesting and important AI systems ever produced"

OpenAI API Waitlist

We're offering free access to the API through mid-August for our private beta, while we determine our longer-term pricing. Describe your use case or product below to join the waitlist.

We're also kicking off an academic access program to let researchers build and experiment with the API. We're going to start with an initial set of academic researchers and collaborators who will gain free access to the API.

* Required

1. What kind of access are you interested in? *

- ☐ I am interested in being a beta user of the API
- ☐ I am interested in conducting academic research on the API
- ☐ I am interested in receiving email updates about the API

<https://arxiv.org/pdf/2005.14165.pdf>

Submit



AI is developing rapidly in any fields, including NLP.

زبان فارسی را دریابیم

Wrap up

Session 1: Introduction

- Applications
- Tasks
- Approaches

Session 2. Basics of Linguistics

Session 3. Basics of ML

Session 4 (Lab). Effective Word Representation by python

Digikala Academy NLP Events

Basic

- **Session 1.** Introduction
- **Session 2.** Basics of Linguistics
- **Session 3.** Basics of ML
- **Session 4 (Lab).** Effective Word Representation by python

Intermediate

- TBA

Advanced

- TBA

References of this session



- Brown, Tom B., et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).