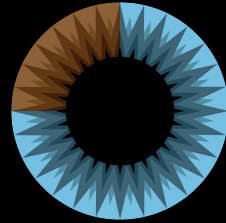


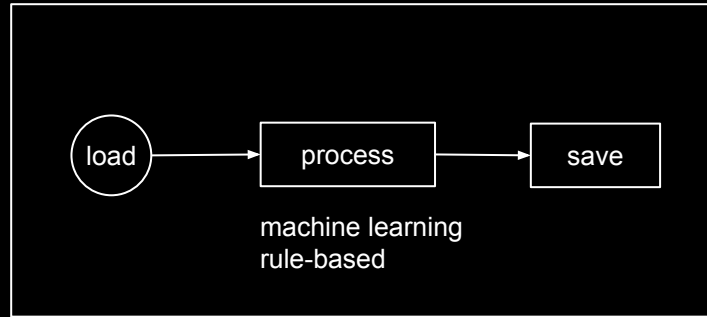
# MACHINE LEARNING

Highly recommended for  
background information



3Blue1Brown's YouTube Course on Neural  
Networks and Deep Learning

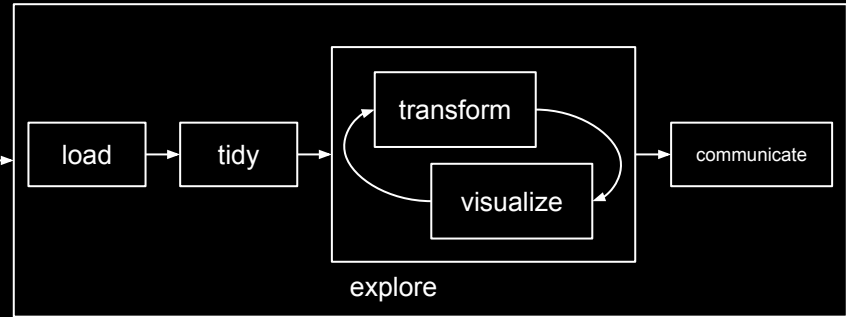
pre-process  
unstructured data



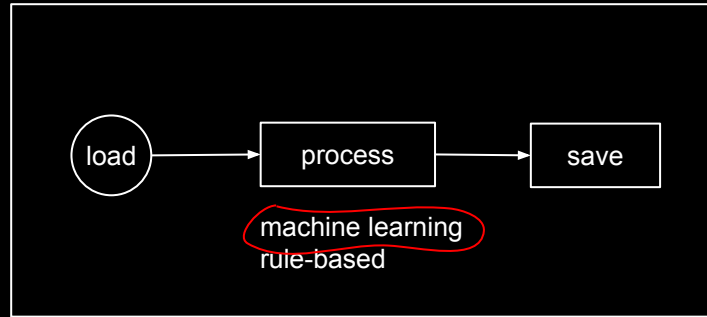
program



exploratory data  
analysis



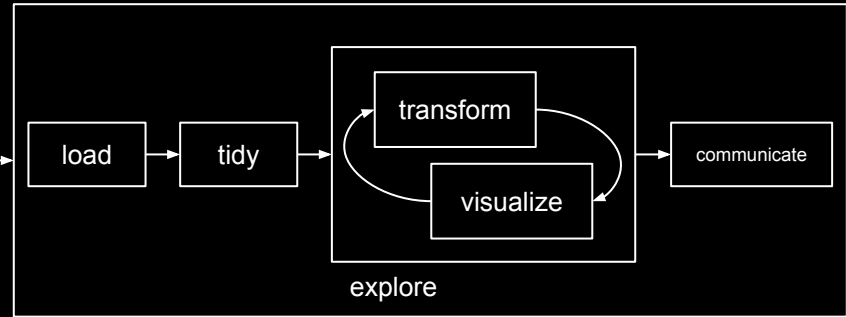
pre-process  
unstructured data



program



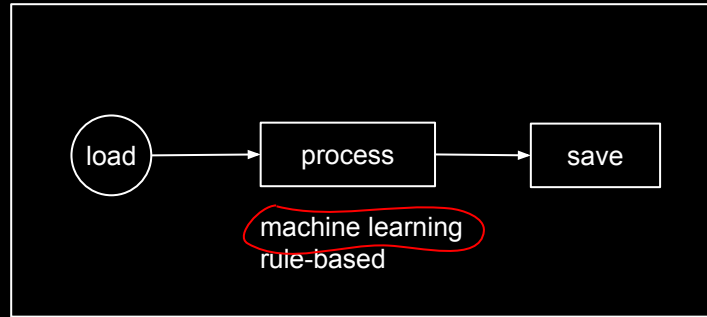
exploratory data  
analysis



program



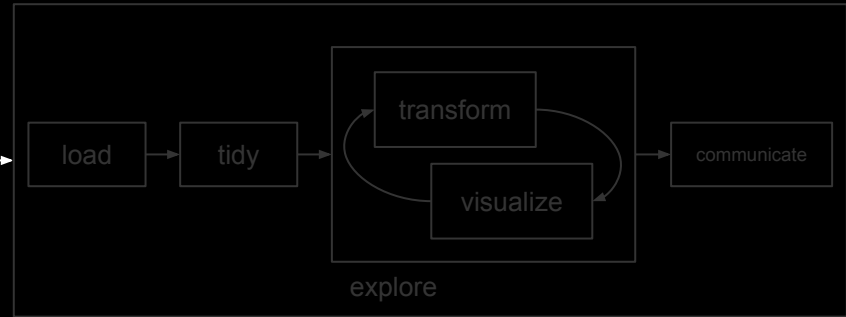
pre-process  
unstructured data



program



exploratory data  
analysis



program





machine learning

program



YouTube



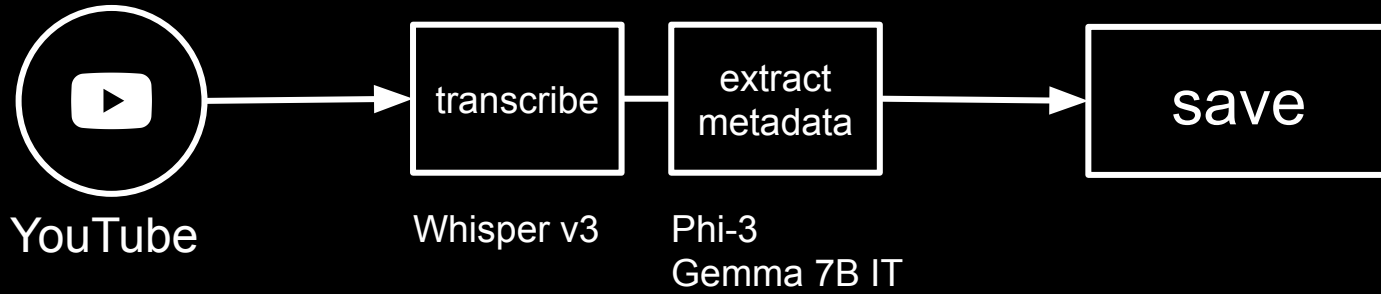
process

machine learning



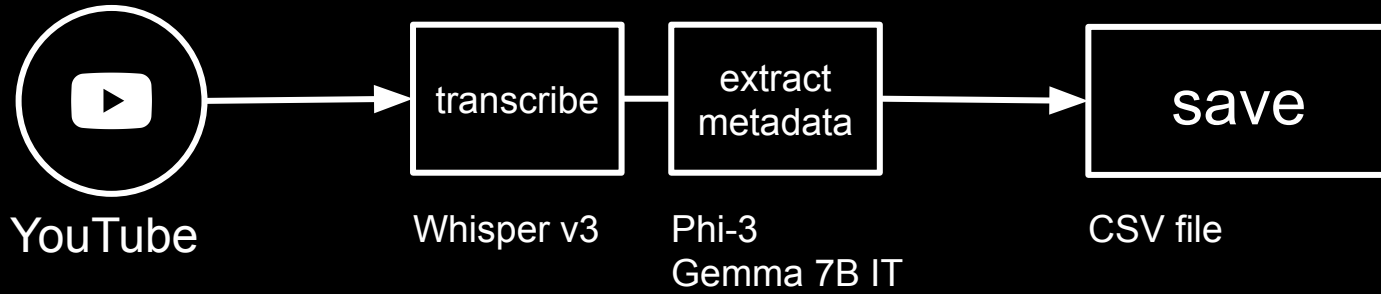
save

program



program





program

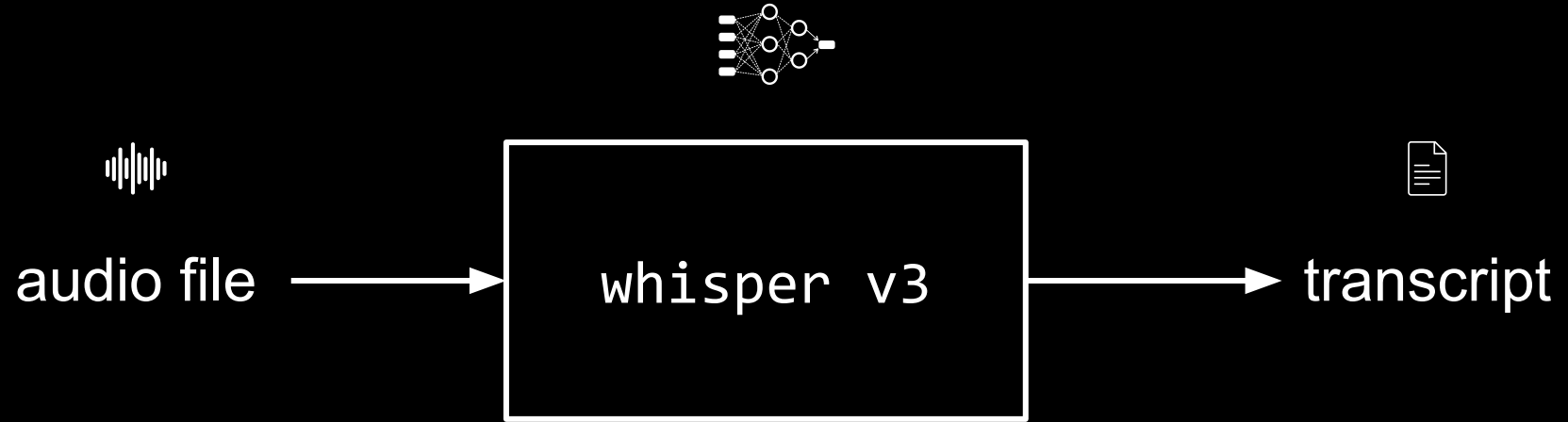
# YouTube API

# Whisper v3

<https://arxiv.org/abs/2212.04356>



<https://huggingface.co/openai/whisper-large-v3>



# Large Language Models (LLM)

what has been said so far?  
(*prompt / context*)

what has been said so far?  
(*prompt / context*)



prediction of next token based on  
learnt probability distribution



what has been said so far?  
(*prompt / context*)



prediction of next token based on  
learnt probability distribution

+

(randomness)

what has been said so far?  
(*prompt / context*)



prediction of next token based on  
learnt probability distribution

+

(randomness)

+

(filter)

(*discriminating, insulting content*)

what has been said so far?  
(*prompt / context*)



prediction of next token based on  
learnt probability distribution

+

(randomness)

+

(filter)

(*discriminating, insulting content*)

next word (*token*)



what has been said so far?  
(*prompt / context*)



prediction of next token based on  
learnt probability distribution

+

(randomness)

+

(filter)

(*discriminating, insulting content*)



next word (*token*)



# PROMPTING

<https://www.promptingguide.ai/>



elements of a prompt

<instruction>

<context>

<input data>

<output indicator>

elements of a prompt

<instruction>

<context>

<input data>

<output indicator>

example prompt

Explain the binary number system.



## elements of a prompt

<instruction>

<context>

<input data>

<output indicator>

## example prompt

Explain the binary number system.

start simple

## elements of a prompt

<instruction>

<context>

<input data>

<output indicator>

## example prompt

You are a friendly tutor and your task is to explain complex concepts as simple as possible.

Explain the binary number system.

## elements of a prompt

<instruction>

<context>

<input data>

<output indicator>

## example prompt

You are a friendly tutor and your task is to explain complex concepts as simple as possible.

Your answers are never longer than 10 sentences.

Explain the binary number system.

# ZERO-SHOT PROMPTING

## elements of a prompt

<instruction>

<context>

<input data>

<output indicator>

## example prompt

Classify the text into neutral,  
negative or positive.

Text: "What a great dinner!"

Sentiment:

## elements of a prompt

<instruction>

<context>

<input data>

<output indicator>

## example prompt

Classify the text into neutral,  
negative or positive.

Text: "What a great dinner!"

Sentiment:

this will be replaced with  
data later...

# FEW-SHOT PROMPTING

IN-CONTEXT LEARNING

## examples in the context to learn from

Extract all references to countries and their continent in the following text using the format from the examples below.

Example 1: "They played the team called 'Die Mannschaft' in the world cup final"

Correct answer: Germany, Europe

Example 2: "The Three Lions once again lost to Germany in a semi final"

Correct answer: England, Europe, Germany, Europe

Text: "The Selecao was destroyed 1:7 by the DFB selection in their home stadium."

Answer:



## examples in the context to learn from

Extract all references to countries and their continent in the following text using the format from the examples below.

Example 1: "They played the team called 'Die Mannschaft' in the world cup final"

Correct answer: Germany, Europe

Example 2: "The Three Lions once again lost to Germany in a semi final"

Correct answer: England, Europe, Germany, Europe

Text: "The Selecao was destroyed 1:7 by the DFB selection in their home stadium."

Answer:

more prompting strategies

chain-of-thought (CoT)

self-consistency

generate knowledge prompting

prompt chaining (subtasks)

tree-of-thoughts (ToT)

retrieval-augmented-generation (RAG)

...

# Phi-3

<https://arxiv.org/abs/2404.14219>



~~<https://huggingface.co/microsoft/Phi-3-mini-128k-instruct>~~

<https://huggingface.co/microsoft/Phi-3-medium-128k-instruct>

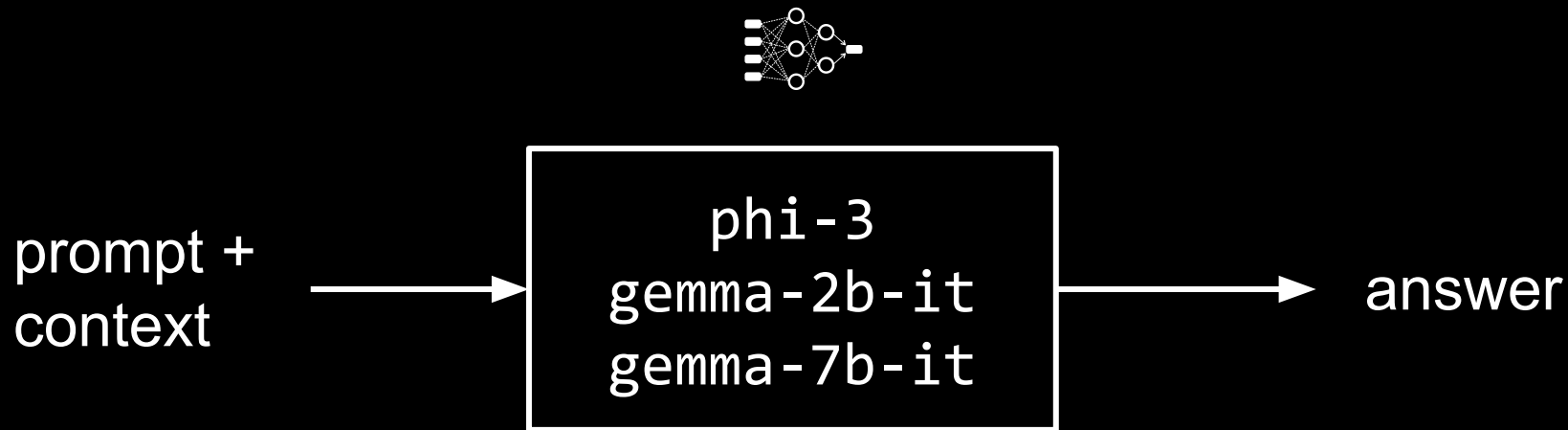
# Gemma 2B / 7B Instruct

<https://arxiv.org/abs/2403.08295>



<https://huggingface.co/google/gemma-2b-it>

<https://huggingface.co/google/gemma-7b-it>



# OpenAI's GPT-4.1 family

<https://platform.openai.com/docs/models>



4.1 nano

GPT-4.1 nano

Default

Fastest, most cost-effective GPT-4.1 model

Compare

Try in Playground

INTELLIGENCE

Average

SPEED

Very fast

PRICE

\$0.1 • \$0.4

Input • Output

INPUT

Text, image

OUTPUT

Text

GPT-4.1 nano is the fastest, most cost-effective GPT-4.1 model.

✦ 1,047,576 context window

↪ 32,768 max output tokens

🗓 Jun 01, 2024 knowledge cutoff

Pricing

Pricing is based on the number of tokens used. For tool-specific models, like search and computer use, there's a fee per tool call. See details in the [pricing page](#).

Text tokens

Per 1M tokens • Batch API price

Input

\$0.10

Cached input

\$0.025

Output

\$0.40

<https://platform.openai.com/docs/models/gpt-4.1-nano>

1 mio. token →  
~2500 pages

**4.1 nano** **GPT-4.1 nano** Default ⓘ ⓘ  
Fastest, most cost-effective GPT-4.1 model

Compare Try in Playground

INTELLIGENCE	SPEED	PRICE	INPUT	OUTPUT
● ● Average	⚡ ⚡ ⚡ ⚡ ⚡ Very fast	\$0.1 • \$0.4 Input • Output	📄 🖼️ 🗑️ Text, image	📄 🗑️ 🗑️ Text

GPT-4.1 nano is the fastest, most cost-effective GPT-4.1 model.

- ✦ 1,047,576 context window
- ↪ 32,768 max output tokens
- 🗓 Jun 01, 2024 knowledge cutoff

**Pricing**

Pricing is based on the number of tokens used. For tool-specific models, like search and computer use, there's a fee per tool call. See details in the [pricing page](#).

Text tokens Per 1M tokens • Batch API price ☐

Input	Cached input	Output
\$0.10	\$0.025	\$0.40

<https://platform.openai.com/docs/models/gpt-4.1-nano>

1 mio. token →  
~2500 pages

roughly 10 cents  
as input

The screenshot shows the OpenAI GPT-4.1 nano model page. It features a header with the model name, a 'Default' dropdown, and a 'Try in Playground' button. Below the header is a comparison table with columns for Intelligence, Speed, Price, Input, and Output. The 'Price' column shows '\$0.1 • \$0.4' for 'Input • Output'. The 'Input' column shows 'Text, image' and the 'Output' column shows 'Text'. A blue arrow points from the text '1 mio. token → ~2500 pages' to the '1,047,576 context window' value in the 'Input' column. Another blue arrow points from the text 'roughly 10 cents as input' to the '\$0.10' value in the 'Input' column of the pricing section.

**GPT-4.1 nano** Default

Fastest, most cost-effective GPT-4.1 model

Compare Try in Playground

INTELLIGENCE	SPEED	PRICE	INPUT	OUTPUT
Average	Very fast	\$0.1 • \$0.4 Input • Output	Text, image	Text

GPT-4.1 nano is the fastest, most cost-effective GPT-4.1 model.

- ✦ 1,047,576 context window
- ↪ 32,768 max output tokens
- 📅 Jun 01, 2024 knowledge cutoff

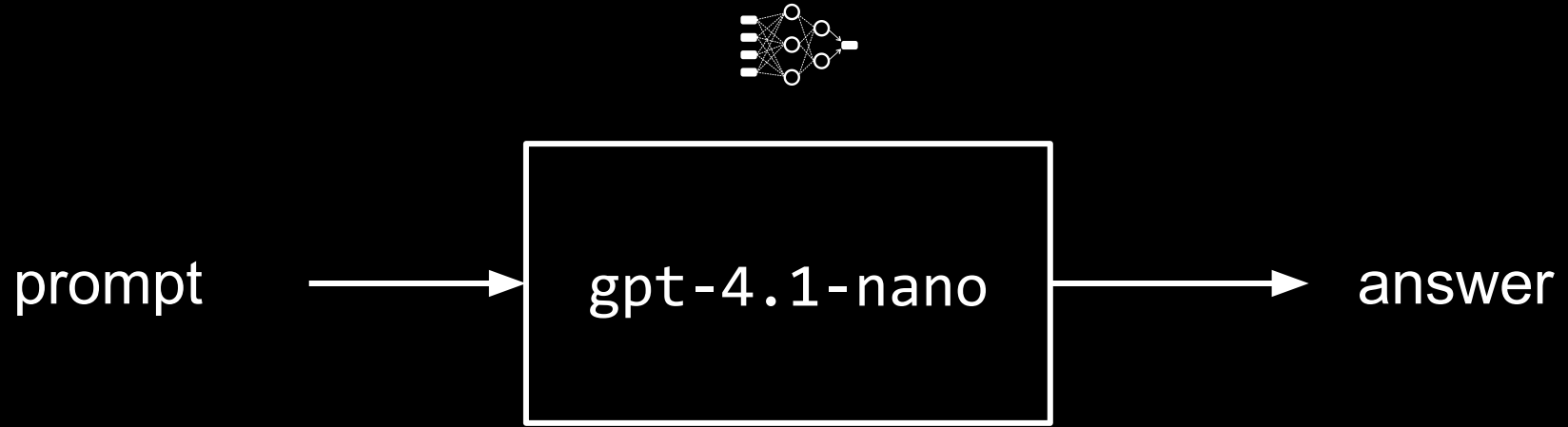
**Pricing**

Pricing is based on the number of tokens used. For tool-specific models, like search and computer use, there's a fee per tool call. See details in the [pricing page](#).

Text tokens Per 1M tokens • Batch API price ☐

Input	Cached input	Output
\$0.10	\$0.025	\$0.40

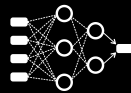
<https://platform.openai.com/docs/models/gpt-4.1-nano>



prompt +  
image



answer



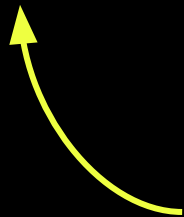
prompt +  
image



gpt-4.1-nano



answer



multimodality