



TEXT WRANGLING, JAVASCRIPT AND MACHINE LEARNING

Glenn Jones – Edinburghjs – 23 March 2023

"The worst thing in data science is a senior developer with domain knowledge"

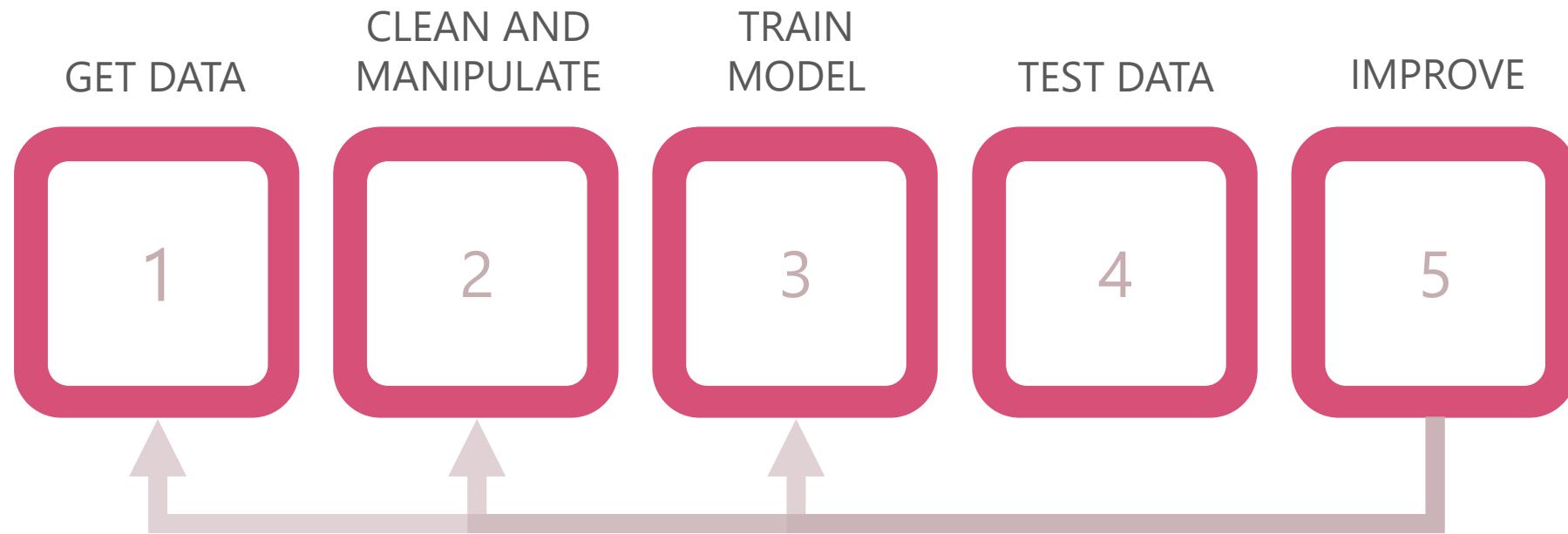
Overheard comment at a data meetup

THE SCIENTIFIC METHOD

*"a researcher develops a hypothesis, tests it through various means, and then modifies the hypothesis on the **basis of the outcome of the tests and experiments.**"*

britannica

TYPICAL ML PROCESS



LIBRARY LANDSCAPE



Python > doccano

NumPy

Python >

pandas

GET DATA

CLEAN AND
MANIPULATE

TRAIN
MODEL

TEST DATA

IMPROVE

JavaScript >

DANFO.JS

arquero

wink

NLTK



Hugging Face

PyTorch

GENSIM
topic modelling for humans

scikit
learn

spaCy

seaborr

TensorFlow

plotly

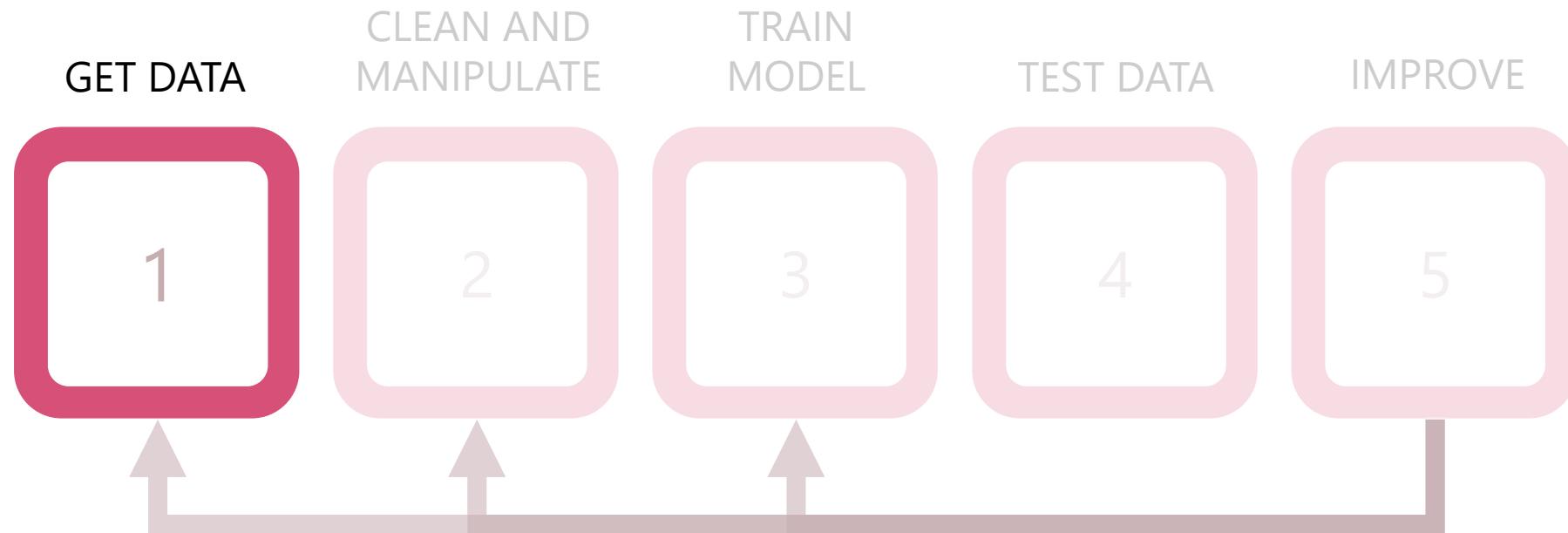
TF
ML

plotly

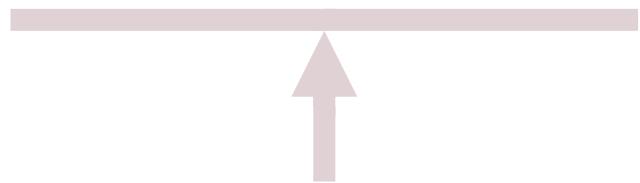
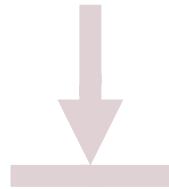
TF
ML

Node.js Notebooks (REPL)

TYPICAL ML PROCESS



"Lead"
Levels of seniority



Job title
"Senior UX Researcher"

Lead UX Researcher – Edinburgh (part-time)

doccano - doccano

Not Secure | 0.0.0.0:8000/projects/5/sequence-labeling?page=32

UX Jobtitles NER

Start Annotation

X Filter List Comment Edit Delete Calendar

32 of 1634 | < < > >|

Home Dataset Labels Members Comments Guideline Metrics Settings

Senior UX Researcher, Alexa Shopping

•SENIORITY

•JOBTITLE

•COREJOBTITLE

Progress

Total 1634

Complete 0

0%

Label Types

JOBTITLE 1

SENIORITY 2

COREJOBTITLE 3

Key Value

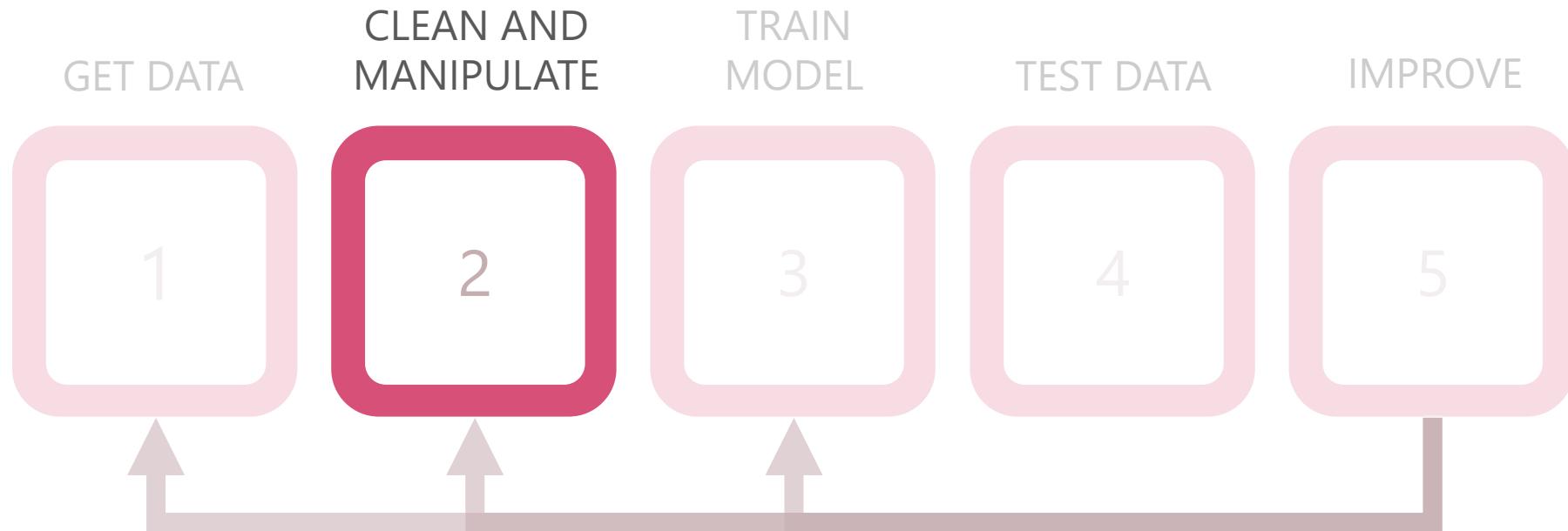
No data available

doccano

DOCCANO JSONL EXPORT

```
1  {
2      "text":"Senior UX Researcher - Fulfilment",
3      "label": [
4          [0,6,"SENIORITY"],
5          [0,20,"JOBTITLE"],
6          [7,20,"COREJOBTITLE"]
7      ]
8  }
```

TYPICAL ML PROCESS





NOTEBOOKS

Are built from a series of blocks containing:

- Code
- Outputs from code
- Visualizations
- Explanatory text
- Statistical results

Each document usually contains all the material to explore and analysis a dataset, with lots of notes to help understand the concepts and flow

The screenshot shows a Jupyter Notebook interface with the title "nlp-examples.nnb — edinburghjs-text wrangling". The left sidebar contains various icons for file operations, search, and help. The main area displays a section titled "Simple word tokenise" which includes a description and a code snippet. The code uses regular expressions to remove special characters and split the text into words. A red box highlights the output of the code execution, which shows the tokenized words: 'snr', 'frontend', and 'developer'. Below the code, there is a message about a debugger being attached. A large pink button labeled "DEMO >" is overlaid on the right side. At the bottom, there is a footer with navigation icons and the text "More sophisticated tree tokenization - with unist".

```
function tokenise(text, lower = true) {
  if (lower) text = text.toLowerCase();
  return text
    .replace(/[\.\,\#\!\%\^\&\*;:{}=\-_`~()]/g, "")
    .replace(/\s{2,}/g, " ")
    .split(" ");
}
tokenise('Snr. Front-end Developer');
```

[25] [JavaScript]

... [

'snr',
'frontend',
'developer'

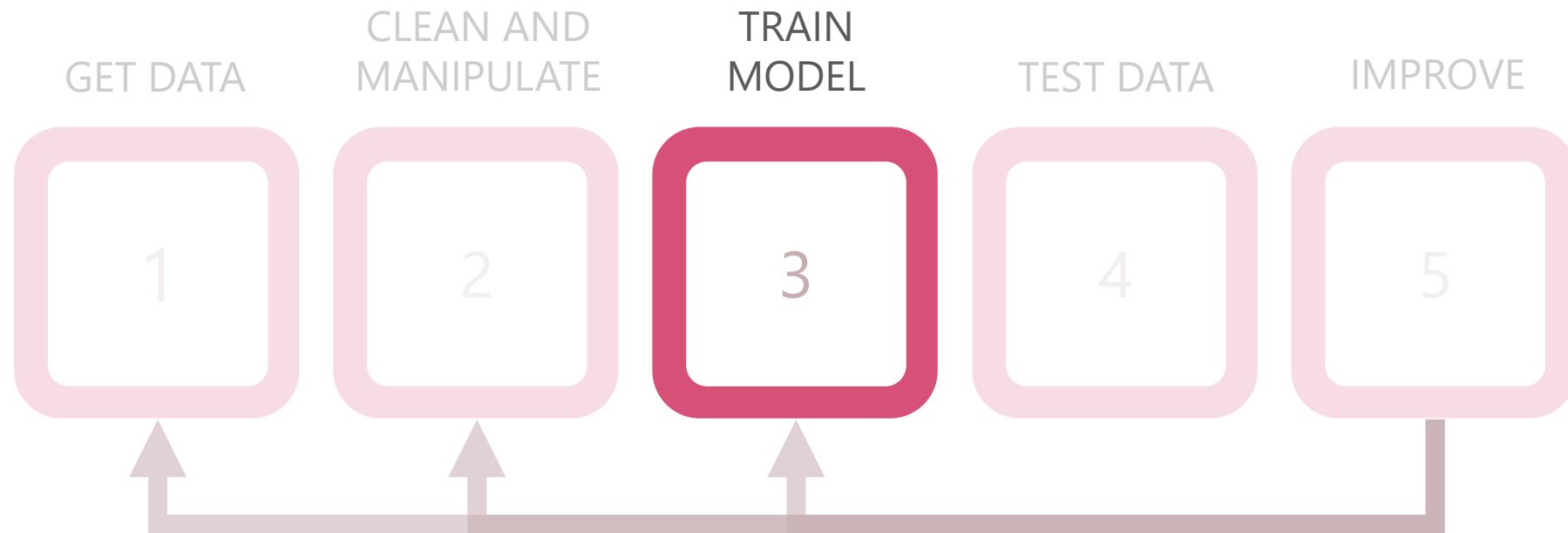
]

Debugger ending on ws://127.0.0.1:65313/1f673e04-4534-4d5b-a86d-14714e92d59d
Debugger attached.
Debugger ending on ws://127.0.0.1:65313/1f673e04-4534-4d5b-a86d-14714e92d59d
For help, see: <https://nodejs.org/en/docs/inspector>

DEMO >

More sophisticated tree tokenization - with unist

TYPICAL ML PROCESS





OpenAI creates large language models (LLM)
that are non-deterministic.

They are a type of model called **transformers**
and can be fine tuned with additional data

Playground - OpenAI API

platform.openai.com/playground

Help Personal ⓖ

Playground

Load a preset... Save View code Share ...

In two sentences tell me about ben nevis?

Ben Nevis is the highest mountain in the British Isles, standing at 1,345 metres (4,413 ft) at its summit. It is located at the western end of the Grampian Mountains in the Lochaber area of the Scottish Highlands, close to the town of Fort William.

Looking for ChatGPT? Try it now ↗

Submit

Playground - OpenAI API

platform.openai.com/playground

Help Personal ⓖ

Playground

Load a preset... Save View code Share ...

You are a highly intelligent and accurate ux design domain Named-entity recognition(NER) system. You take Passage as input and your task is to recognize and extract specific types of ux design domain named entities in that given passage and classify into a set of following predefined entity types: Your output format is only `[{'label': type of entity from predefined entity types, 'text': entity in the input text},...,{['branch' : Appropriate branch of the passage , 'group': Appropriate Group of the passage}]` form, no other form. Do not return empty entity values.

Input: Senior UX Researcher, part-time 6 month contract, Edinburgh - Hybrid

Output:

Looking for ChatGPT? Try it now ↗

Submit

150

Playground - OpenAI API

platform.openai.com/playground

Help Personal ⓖ

Playground

Load a preset... Save View code Share ...

You are a highly intelligent and accurate UX design domain named-entity recognition(NER) system. You take Passage as input and your task is to recognize and extract specific types of UX design domain named entities in that given passage and classify into a set of following predefined entity types: Your output format is only `[{'label': type of entity from predefined entity types, 'text': entity in the input text}, ..., {'branch': Appropriate branch of the passage, 'group': Appropriate Group of the passage}]` form, no other form. Do not return empty entity values.

Input: Senior UX Researcher, part-time 6 month contract, Edinburgh - Hybrid

Output: `[{'label': 'Person', 'text': 'Senior UX Researcher'}, {'label': 'Time', 'text': '6 month contract'}, {'label': 'Place', 'text': 'Edinburgh'}, {'label': 'Position', 'text': 'Hybrid'}, {'branch': 'UX Design', 'group': 'Research'}]`

Looking for ChatGPT? Try it now ↗

Submit ⏪ ⏵ ⏴ ⏵ ⏴ 231

ner02.nnb — Prompt

ner01.nnb M ner02.nnb M ner.njk

notebooks > ner02.nnb > M+Prompt - Name entity recognition (NER) - UX Job titles > M+With labels > JS let result = await nlpPrompter.fit('ner.njk', {

+ Code + Markdown | ▶ Run All ⌘ Clear All Outputs ⋮ Node.js

Prompt - Name entity recognition (NER) - UX Job titles

Add Prompt module and get API key from .env file

```
1 const Prompt = require('../index.js')
2 const {Prompter, OpenAI} = Prompt;
3
4 const DotEnv = require('dotenv');
5 DotEnv.config({path: '../.env'});
6 const apiKey = process.env.OPENAI_API_KEY;
```

[8] ✓ 0.1s JavaScript

Load the model for OpenAI

```
1 let model = new OpenAI(apiKey);
2 let nlpPrompter = new Prompter(model);
```

[9] ✓ 0.0s JavaScript

Call the prompt module

- The `domain` property helps focus the language model.
- The `textInput` property is the text to be parsed.

DEMO >

```
1 let result = await nlpPrompter.fit('ner.njk', {
```

master* ⌘ 0 △ 0 Launch Program (Prompt)

Cell 9 of 14

Playground - OpenAI API

platform.openai.com/playground?lang=node.js&model=davinci%3Aft-personal-2023-03-21-22-17-08

Overview Documentation API reference Examples Playground Help Personal

Playground Load a preset... Save View code Share ...

Senior UX Researcher, part-time 6 month contract, Edinburgh - Hybrid ->
[{"label": "SENIORITY", "text": "Senior"}, {"label": "JOBTITLE", "text": "Senior UX Researcher"}, {"label": "COREJOBTITLE", "text": "UX Researcher"}] >

Mode Complete

Model davinci:ft-personal...

Temperature 0

Maximum length 256

Stop sequences Enter sequence and press Tab
"]]

Frequency penalty 0

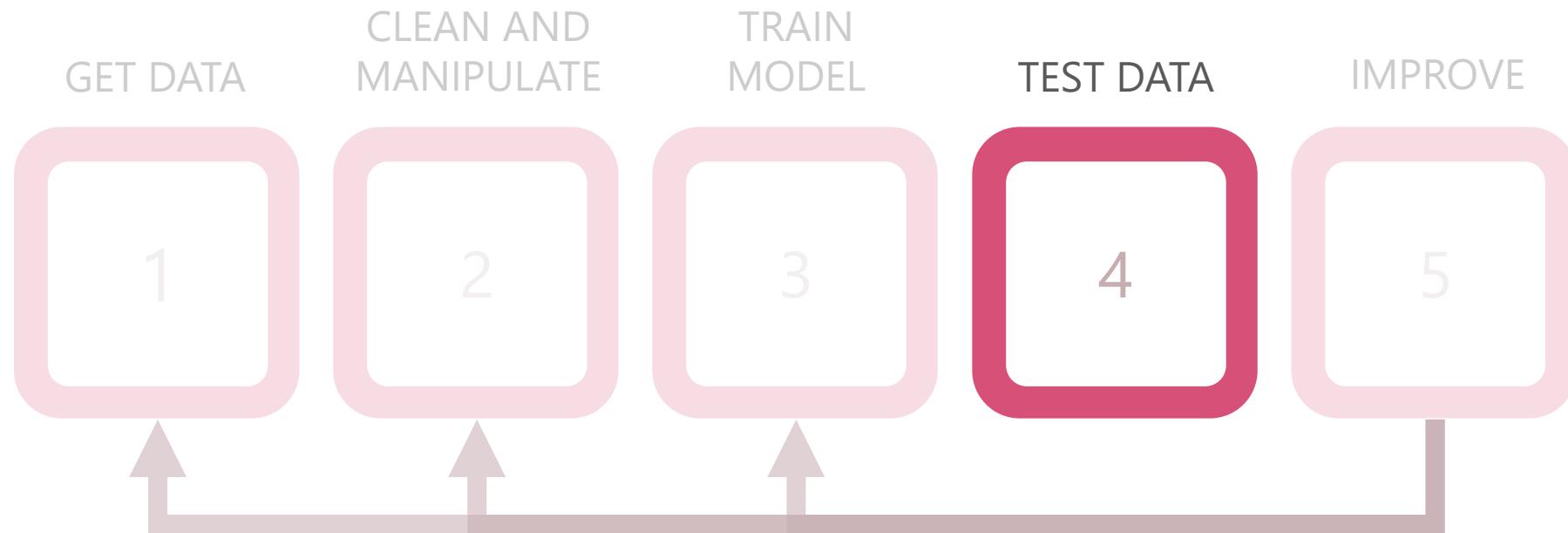
Submit

Looking for ChatGPT? Try it now ↗

57

"]]

TYPICAL ML PROCESS



JOB TITLES

	Rule based	OpenAI* Labels provided	OpenAI* Fine-tuned
Accuracy	66.39%	79.47%	95.08%
Speed	0.0094ms	3.954ms	2.854ms

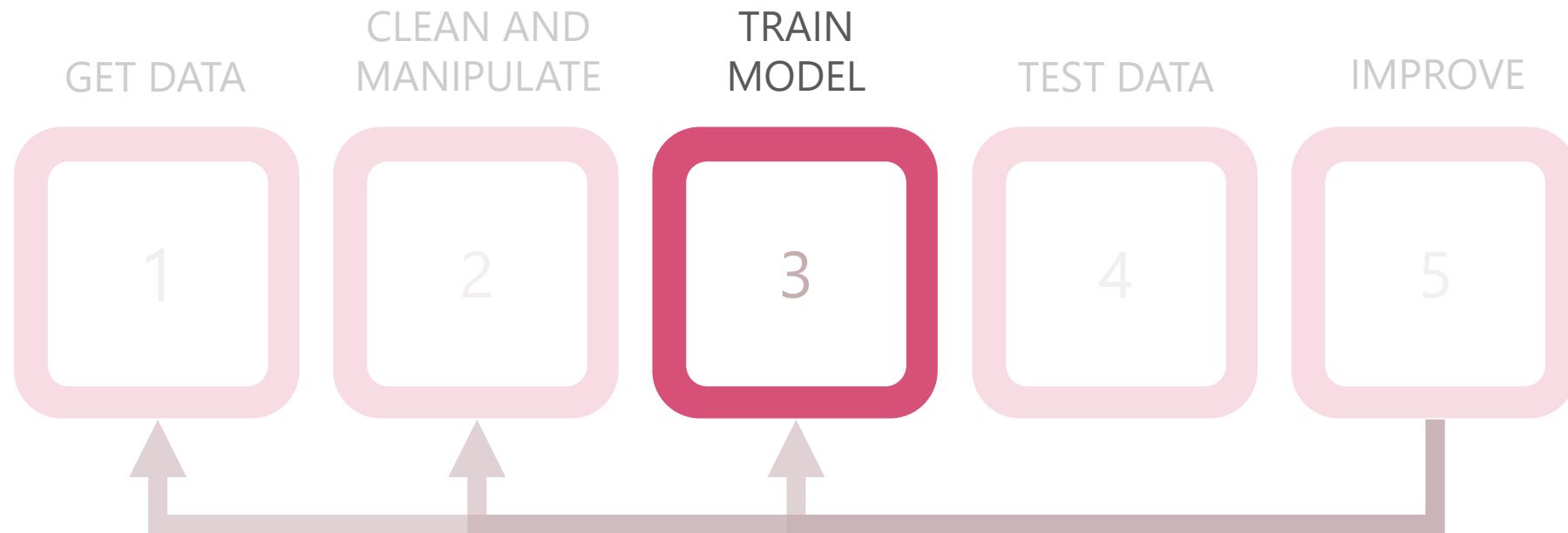
Based on 70/30 train-test set of 1,635 annotations
Fine-tuning contained 467 examples from train set
* OpenAI model was text-davince-003

SENIORITY

	Rule based	OpenAI* Labels provided	OpenAI* Fine-tuned
Accuracy	73.57%	43.94%	96.30%
Speed	0.0029ms	3.954ms	2.854ms

Based on 70/30 train-test set of 1,635 annotations
Fine-tuning contained 467 examples from train set
* OpenAI base model was text-davince-003

TYPICAL ML PROCESS



isdirectemployer v1.0.0

Not Secure | 0.0.0.0:3062

Is direct employer - v1.0.0

ML API to determine if a job description is from a direct employer or a third party recruitment organisation. [API documentation](#)

API call took 29.6ms

```
{"type": "agency", "scores": {"0": 0.9848238229751587, "1": 0.01517618726938963}}
```

Job Description

Will design and develop the look and feel of existing and brand new web application platforms, incorporating the brand and vision to improve creative interactive designs from functional specifications through to the technical implementation and deployment to customers. Be involved in all aspects of the process, including testing and prototypes.

You will consult with customers to fully identify the problems and their requirements then convey these requirements to other stakeholders, and liaise regularly with customers to facilitate reviews and improvements.

Skills:

- Ability to design the look, feel and user flow for both B2B and consumer web applications and supporting collateral, e.g. email
- Ability to contribute to strategic design and product architecture decisions
- Good understanding of core web technologies such as CSS, UI frameworks and SPA design
- Creating and maintaining application UI design and style guides with active development teams
- Designing responsive layouts for form and application design

Bright Purple is an equal opportunities employer: we are proud to work with values of diversity and inclusion in our industry

SUBMIT

DEMO >

Convolutional Neural Networks (CNN)

Models - Hugging Face x +

huggingface.co/models

Hugging Face Search models, datasets, users... Models Datasets Spaces Docs Solutions Pricing Log In Sign Up

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

Feature Extraction Text-to-Image

Image-to-Text Text-to-Video

Visual Question Answering

Document Question Answering

Graph Machine Learning

Computer Vision

Depth Estimation Image Classification

Object Detection Image Segmentation

Image-to-Image

Unconditional Image Generation

Video Classification

Zero-Shot Image Classification

Natural Language Processing

Text Classification Token Classification

Table Question Answering

Question Answering

Zero-Shot Classification Translation

Models 159,169 Filter by name

Full-text search Sort: Most Downloads

bert-base-uncased
Updated Nov 16, 2022 · 44.1M · 625

emilyalsentzer/Bio_ClinicalBERT
Updated Feb 27, 2022 · 19.3M · 108

gpt2
Updated Dec 16, 2022 · 18.2M · 752

xlm-roberta-base
Updated Nov 16, 2022 · 15.8M · 209

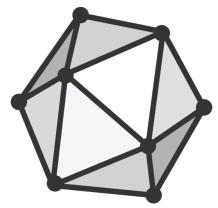
jonatasgrosman/wav2vec2-large-xlsr-53-english
Updated Dec 14, 2022 · 15.6M · 32

openai/clip-vit-large-patch14
Updated Oct 4, 2022 · 10.6M · 260

distilbert-base-uncased
Updated Nov 16, 2022 · 8.74M · 152

microsoft/layoutlmv3-base
Updated Dec 13, 2022 · 8.42M · 87

t5-base



ONNX



The Open Neural Network Exchange

WEBASSEMBLY

Quantization

The screenshot shows a web browser window for the `Transformers.js` demo at `localhost:8080`. The page has a dark theme with white text. At the top, there's a navigation bar with links for `Demo`, `Getting Started`, `Usage`, and `Examples`. Below the navigation, the title `Transformers.js` is displayed in large, bold, white font. A sub-header says "Run 😊 Transformers in your browser!" followed by two buttons: a blue `Get Started` button and a white `View Source` button with a magnifying glass icon.

Demo

Don't believe us? Play around with some of these models:

Task:

Code completion w/ Salesforce/codegen-350M-mono (365 MB)

Parameters:

Max length 10 No. beams 1 Temp. 0.001 Top K 0 Sample No

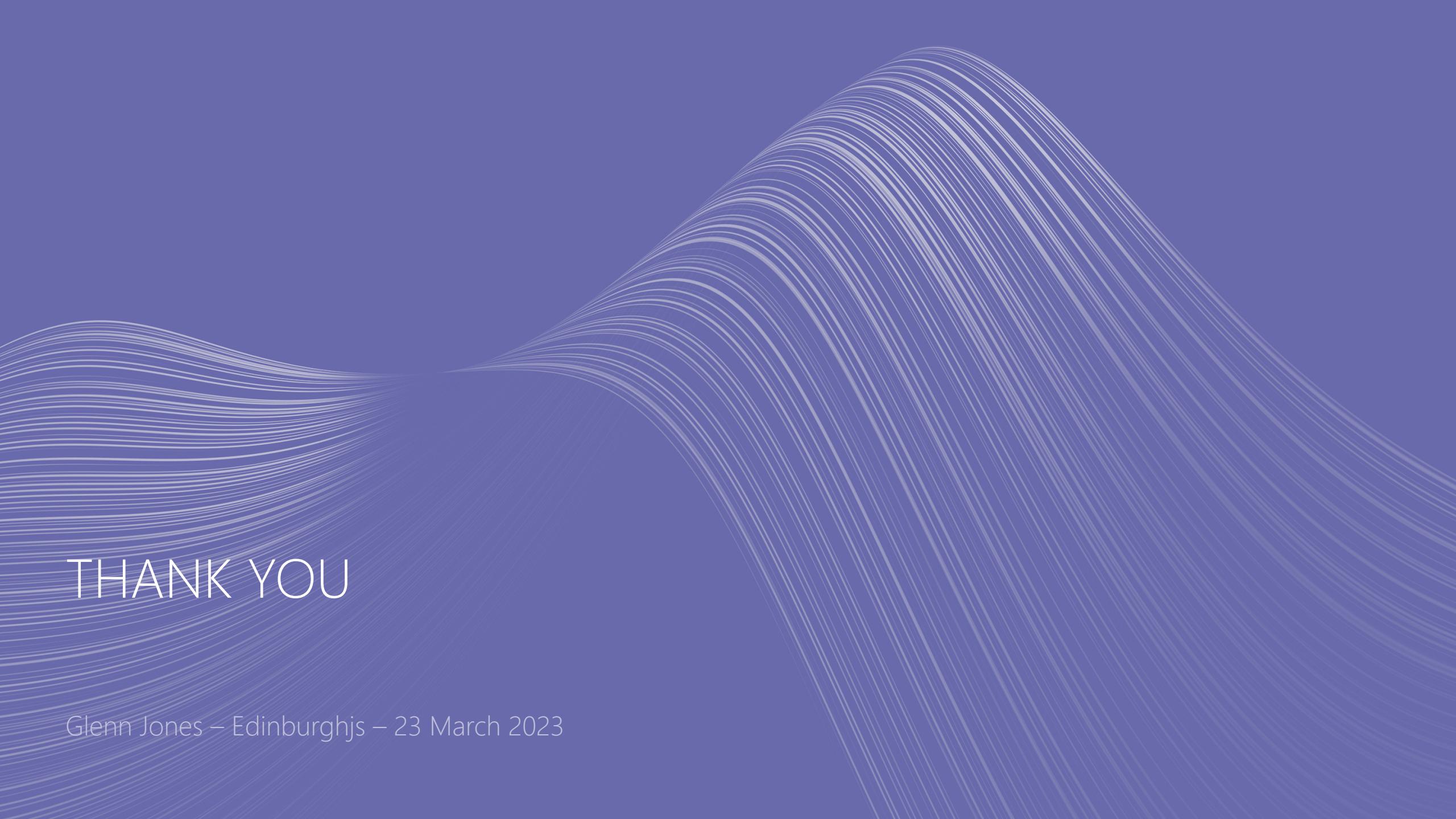
```
def percentageOf(a, b):
    return (a / b) * 100
```

DEMO >

MACHINE LEARNING

Is at a point of
democratization and
commodification



The background of the slide features a subtle, abstract design. It consists of numerous thin, white, wavy lines that radiate from the bottom left corner towards the top right. These lines create a sense of motion and depth against a solid blue gradient background, which transitions from a darker shade at the top to a lighter shade at the bottom.

THANK YOU

Glenn Jones – Edinburghjs – 23 March 2023