FourthBrain

MLE Program, Cohort 11

MLE11

Week 3: Ensuring High-Quality Data

Housekeeping

Overall:

- cameras on!
- do not leave early
- inform me if you are skipping class

Breakout Rooms and Pair Programming

- empathy
- be engaged
- collaborate

Agenda

1. Intro (5 min) - Milica
2. Pitches (30 min)
3. About the Data and Responsible AI (45 min) - Milica
4. Break (5 min)
5. API Theory + Fine-tuning live coding (45 min) - Milica
6. Break (30 min)
7. Intro to Fast API (30 min) - Chris
8. Coding Assignment (2 hrs) - Chris
    a. Intro to Fast API
    b. Breakout Room 1
    c. Recap/Discussion
    d. Breakout Room 2
9. Capstone Teams and Housekeeping Review - Wrap Up

Office Hours Reminders

Chris

Tuesdays 5 pm PT

Anna

Tuesdays 10:30 am PT

Milica

Thursdays 6 pm PT

Capstone pitches

- Glen
- Michael
- Stacey
- Arsalan
- Amit
- Adam
- Kimberly

3 min + 1 min Q/A hard cutoff

Where are we?

Our Promise to You

By the end of this course, you will be able to contribute to high-performing AI product teams by leveraging ***real-world data*** *to* ***build***, ***package***, *and* ***deploy*** *state-of-the-art ML models* as *containerized web applications* in *cloud-based production environments*.

Our Updated Curriculum!

- ML Project Scoping
- **Dealing with Real Data**
- Data Wrangling & Exploratory Analysis
- Big Data

DATA CENTRIC AI

- Supervised ML
- Deep Learning & AutoML
- Unsupervised, Semi- & Self-supervised Learning

ML MODELING

- Computer Vision
- Natural Language Processing
- Transformers & Fine Tuning Pre-Trained Networks

AI APPLICATIONS

- Building ML Web Apps
- Containerization
- Model Serving
- Machine Learning in Production

MLOps

Last Week

Concepts

- The AI Product Lifecycle & ML Project Scoping
- Data-Centric AI
- Responsible ML Principles

Hands-On Activities

- Sentiment Analysis
- Capstone Ideation & Teaming

This Week!

Concepts

- Best Practices for High-Quality Data; Establishing Data Lineage
- Labeling and Crowdsourcing
- Responsible Data
- REST **APIs** & HTTP Review

Hands-On Activities

- Hitting Twitter and Reddit **APIs** to collect real, live data
- Fine-Tuning Pre-Trained Transformer Models
- Developing a Data-Centric Proof of Concept

What questions do you have?

FourthBrain

It’s all about the Data

“Data is the ***hardest part of ML*** *and the* ***most important piece to get right****… Broken data is the most common cause of problems in production ML systems”*

- Scaling Machine Learning at Uber with Michelangelo - Uber, 2018

This Week!

Concepts

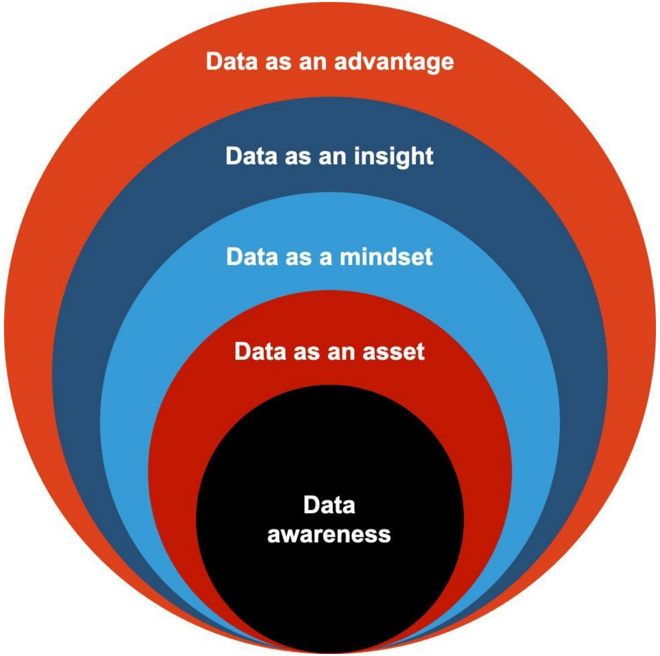
- Best Practices for High-Quality Data; Establishing Data Lineage
- Labeling and Crowdsourcing
- Responsible Data
- REST APIs & HTTP Review

Hands-On Activities

- Hitting Twitter API to collect real, live data
- Fine-Tuning Pre-Trained Transformer Models
- Developing a Data-Centric Proof of Concept with Reddit API

High-Quality Data, Six Best Practices

1. Fail to plan, plan to fail
2. User needs → Data Needs
3. Source responsibly
4. Prepare and document data
5. Design for labelers and labeling
6. Tune your model



Endava 2021

1. Fail to Plan, Plan to Fail

- **Plan to gather high-quality data from the start**

- During ML scoping, manage the entire AI product lifecycle!
    - Labeling
    - Feature space coverage
    - Minimal dimensionality
    - Maximum predictive power
    - Fairness
    - Rare events

2. User Needs ➜ Data Needs

- **User and customer needs must be translated into data needs**

- Connecting ML data + model accuracy to business KPIs is essential
- Understand **Problem**, **Why**, **Audience**, and what **Success** looks like
- During identification of datasets or planning for collection, be diligent!
  - Inspection
  - Identify potential bias
  - Design data collection methods

3. Source Responsibly

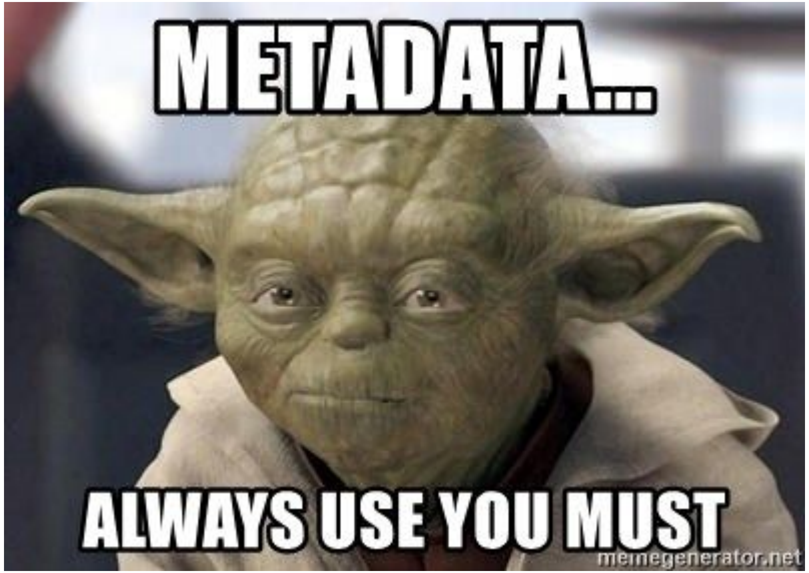
- **Consider relevance, fairness, privacy, and security**

- Many ways to collect!
  - Collect live data
  - Build your own dataset
  - Web scraping
  - Open source dataset
  - Build synthetic datasets
  - Hybrids of any of these!
- Each have challenges with biases, user privacy, and compliance

4. Prepare and Document Data

- **Prepare the dataset, and record everything about it!**

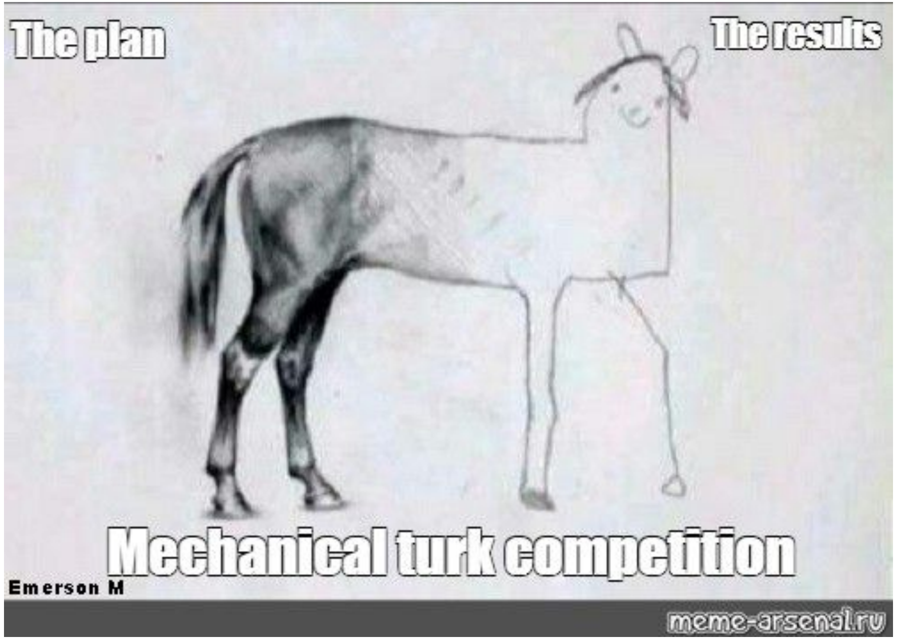
- Decision points during gathering
- Processing steps
- Metadata
- *Enough information to replicate it*



5. Design for Labelers and Labeling

- **Correct labeling of data is incredibly important**

- For projects requiring a lot of labeling work, this can get complicated
- Know your labelers and their tools
- Understand the limitations and likely types of errors that might arise



6. Tune Your Model

- **Test your model, and tune it rigorously**

- Not *just* hyperparameter tuning
- Also iterating on the underlying data, and tracing output errors
- Can mean collecting supplemental data or removing select data
- Both data and model must focus on the specific application being built

Data Lineage

- This is **just documenting your data!**

- Can be as sophisticated as [Data Version Control (DVC)]
- Can be as simple as a flowchart diagram (e.g., [draw.io])
- For your capstone project
  - Ensures other members of your team understand
  - Other contributors can easily on your project
  - Align you with industry best practices

Breakout Ideation!

5 min
(3-4 per room)

- How would you describe the "data lineage" for the project that you proposed?
- What decisions points are you likely to encounter as you dig into the details of the data?
- Designate one person to share your group's best answers!

This Week!

Concepts

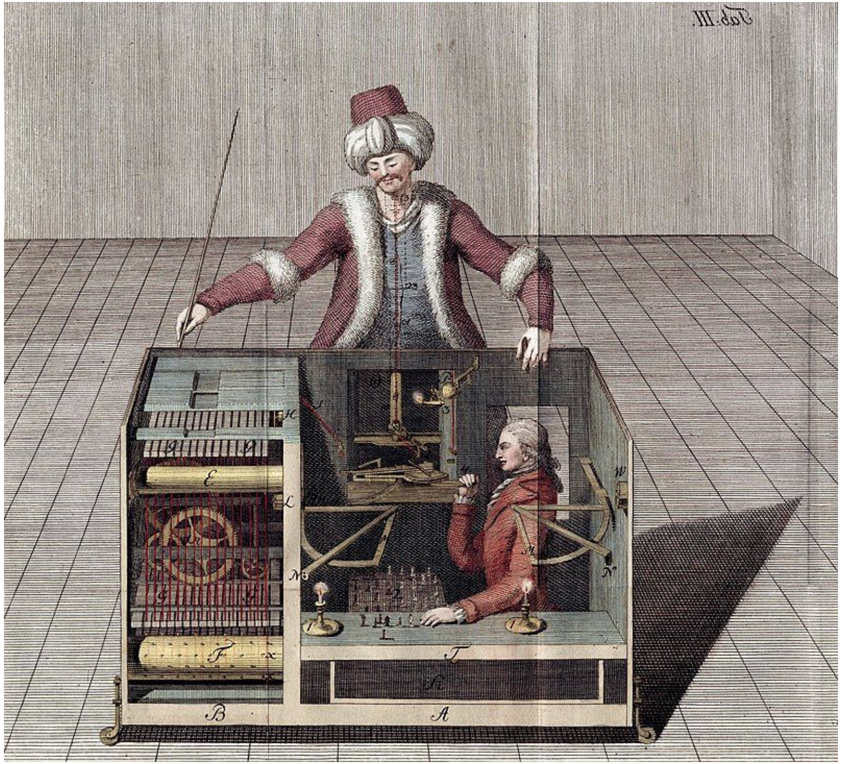
- Best Practices for High-Quality Data; Establishing Data Lineage
- Labeling and Crowdsourcing
- Responsible Data
- REST APIs & HTTP Review

Hands-On Activities

- Hitting Twitter API to collect real, live data
- Fine-Tuning Pre-Trained Transformer Models
- Developing a Data-Centric Proof of Concept with Reddit API

Crowdsourced Data Annotation

- A.k.a. “Mechanical Turk” labeling
- Often not very successful
- Even if you think it will be “easy” to label your data.
- We often overestimate how clearly we convey intent online.



Best Practice Steps to Take

1. Label many of the examples *yourself* before designing the task
2. Pay and treat workers fairly
3. Start with small pilots
4. Assume annotators are trying hard; it’s probably your communication skills
5. Provide labelers incremental feedback
6. Hire fewer people, full-time

This Week!

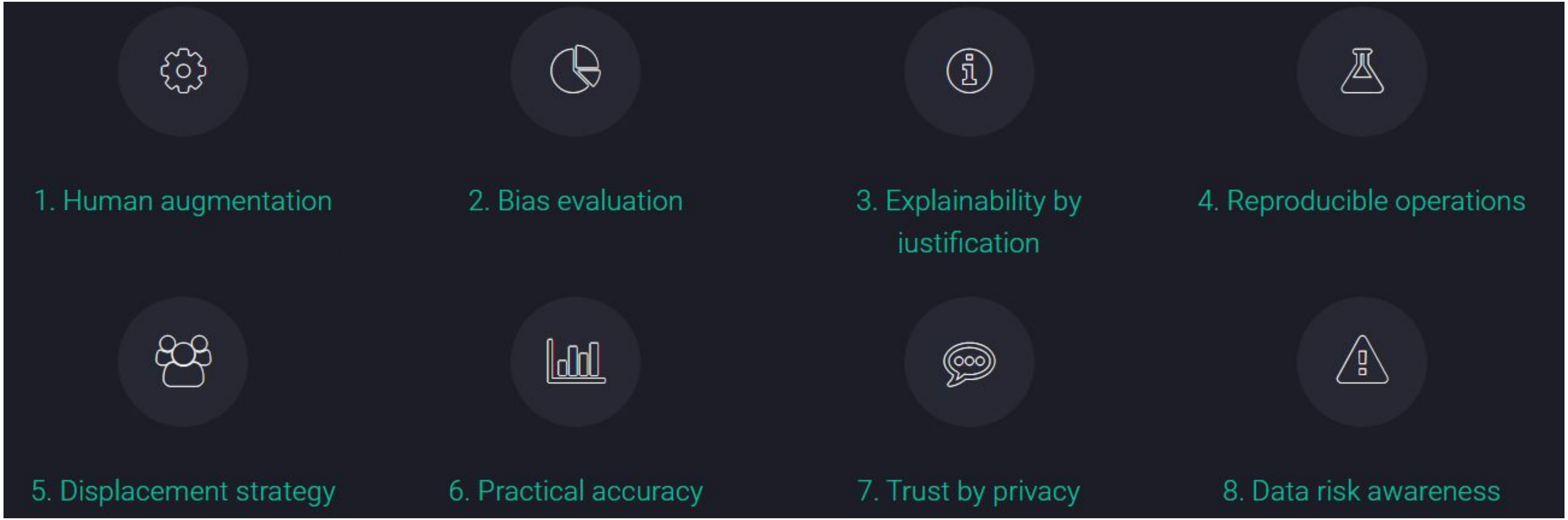
Concepts

- Best Practices for High-Quality Data; Establishing Data Lineage
- Labeling and Crowdsourcing
- Responsible Data
- REST APIs & HTTP Review

Hands-On Activities

- Hitting Twitter API to collect real, live data
- Fine-Tuning Pre-Trained Transformer Models
- Developing a Data-Centric Proof of Concept with Reddit API

Eight Principles for Responsible AI



1. Human augmentation

2. Bias evaluation

3. Explainability by iustification

4. Reproducible operations

5. Displacement strategy

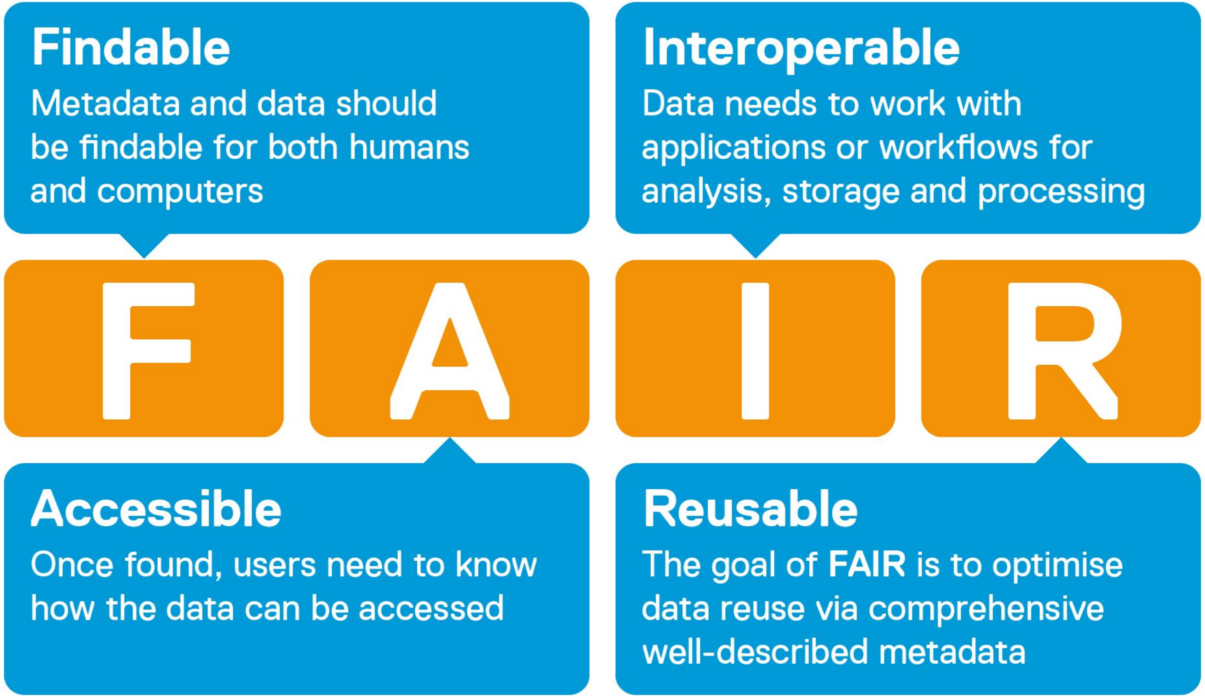
6. Practical accuracy

7. Trust by privacy

8. Data risk awareness

FAIR Principles

- Guiding principles for scientific data and stewardship
- Published in 2016
- https://www.go-fair.org/fair-principles/



Findable
Metadata and data should be findable for both humans and computers

Interoperable
Data needs to work with applications or workflows for analysis, storage and processing

F A I R

Accessible
Once found, users need to know how the data can be accessed

Reusable
The goal of **FAIR** is to optimise data reuse via comprehensive well-described metadata

Six Types of Bias

1. Confirmation Bias
    a. Of course! I knew it!
2. Selection Bias
    a. Your sample must represent your population. Statistics 101.
3. Historical Bias
    a. Socio-cultural prejudices and beliefs are represented in training data.
4. Survivorship Bias
    a. Focusing on winners; over-indexing on what survived.
5. Availability Bias
    a. Just because it was easy to get, doesn’t mean it’s the data you need.
6. Outlier Bias
    a. Averages and higher level statistics hide important information.

Discussion

Review the eight principles of Responsible AI

- **Which principle resonates with you the most?**
- **Why?**
- **What types of bias are you most likely to run into, and how can you avoid it?**
- **Capstone and biases**

Breakout Ideation!

15 min
(3-4 per room)

- Log into ChatGPT and assess it's bias
- Link

- Have the person whose idea it was take notes and share with the larger group!

Time for a Break!

5 min

This Week!

Concepts

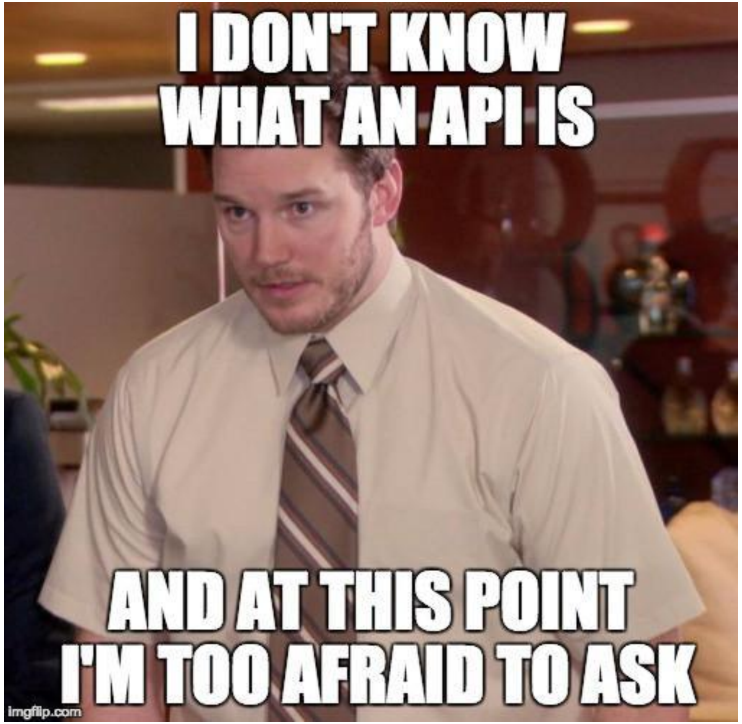
- Best practices for high-quality data
  - Data Lineage
  - Identifying, Sourcing, Collecting, Labeling, Evaluating, and Validating Data
  - Responsible Data
- REST APIs & HTTP Review

Hands-On Activities

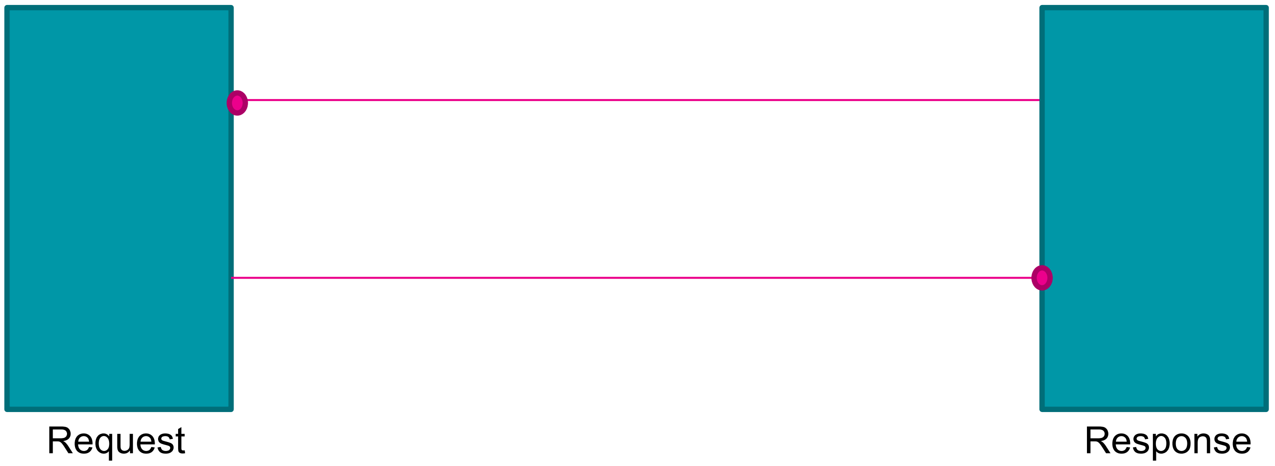
- Hitting Twitter API to collect real, live data
- Fine-Tuning Pre-Trained Transformer Models
- Developing a Data-Centric Proof of Concept with Reddit API

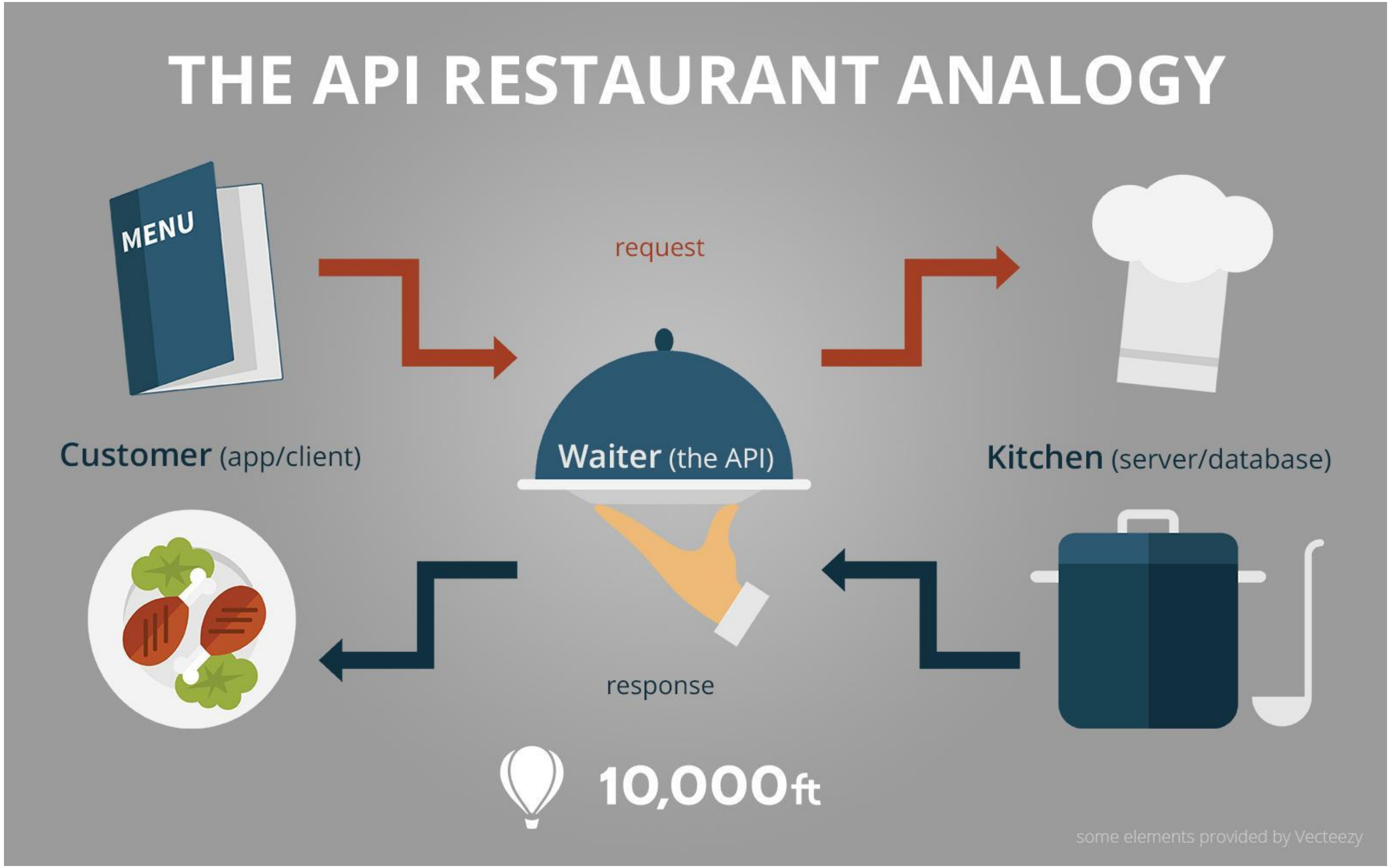
Application Programming Interfaces (APIs), an Introduction

What is an API?



API in nutshell

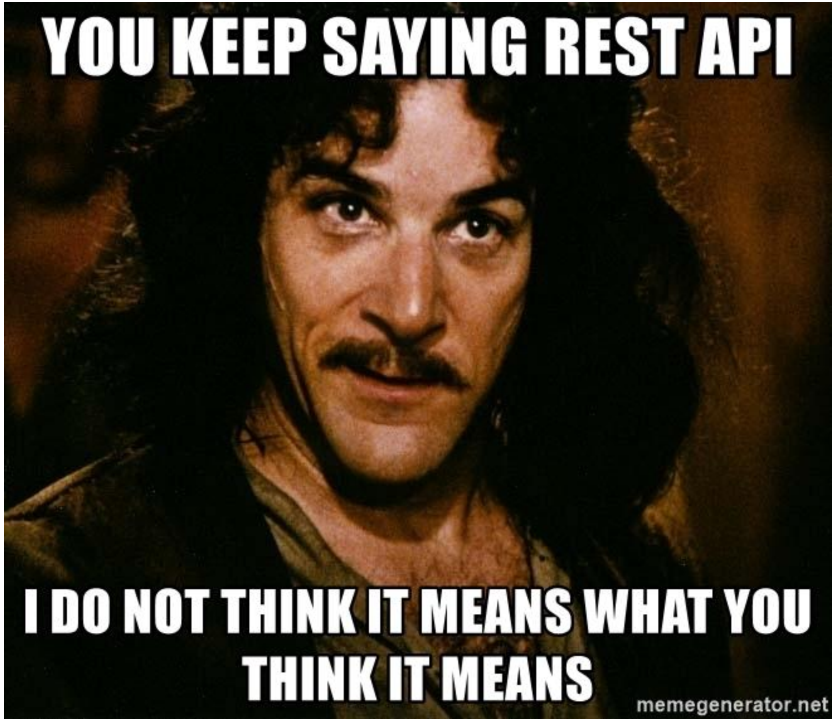



THE API RESTAURANT ANALOGY
MENU
request
Customer (app/client)
Waiter (the API)
Kitchen (server/database)
response
10,000ft
some elements provided by Vecteezy

Representational state transfer (REST)

- REST (RESTful) is a design pattern

- REST is the most common architecture used today when designing APIs



Using HTTP Methods for RESTful Services

REST (a.k.a. **CRUD**)

- **C**reate
- **R**ead
- **U**pdate/Replace | **U**pdate/Modify
- **D**elete

HTTP actions/verbs

- POST
- GET
- PUT | PATCH
- DELETE

When we "hit" APIs as MLE's, we might be doing any of these actions!

What makes a good API?

What makes a good API?

- Know it’s a real user
- Clear and understandable
- Fast and Scalable
- Robust error handling
- Reliable and stable
- Logging
- Easy-to-digest documentation
- Security
- Gentle with newbies and empowers experienced users

What questions do you have?

This Week!

Concepts

- Best practices for high-quality data
  - Data Lineage
  - Identifying, Sourcing, Collecting, Labeling, Evaluating, and Validating Data
  - Responsible Data
- REST APIs & HTTP Review

Hands-On Activities

- Hitting Twitter API to collect real, live data
- Fine-Tuning Pre-Trained Transformer Models
- Developing a Data-Centric Proof of Concept with Reddit API

FourthBrain

The Big Idea

Fine Tuning a Pre-Trained Model

Take a **model trained for one task.** (“Pre-trained”)

Tweak it to **perform a similar task.** (“Fine-tune”)

The Subtle Art of Fine-Tuning

Benefits

- Reduces computation costs (and carbon footprint)
- Use state-of-the-art models without training from scratch

Implementation

- Start with Hugging Face’s Transformers Library
- Leverage Hugging Face’s Trainer API
- Avoid headaches 🤗 . Read more here!
- *(Note: Transformers can be used for LOTS of things - think GPT)*

Hugging Face

Main features:

- Leverage **20,000+ Transformer models** (T5, Blenderbot, Bart, GPT-2, Pegasus...)
- Upload, manage and serve your **own models privately**
- Run Classification, NER, Conversational, Summarization, Translation, Question-Answering, Embeddings Extraction tasks
- Get up to **10x inference speedup** to reduce user latency
- Accelerated inference on **CPU** and **GPU** (GPU requires a Startup or Enterprise plan)
- Run **large models** that are challenging to deploy in production
- Scale to 1,000 requests per second with **automatic scaling** built-in
- **Ship new NLP features faster** as new models become available
- Build your business on a platform powered by the reference open source project in NLP

API

Fine-Tune a Pre-Trained Model Demo

Recap of This Week!

Concepts

- Best practices for high-quality data
    - Data Lineage
    - Identifying, Sourcing, Collecting, Labeling, Evaluating, and Validating Data
    - Responsible Data
- REST APIs & HTTP Review

Hands-On Activities

- Fine-Tuning Pre-Trained Transformer Models

FourthBrain

Preparing for Coding Session

Time for a Break!

30 min

FourthBrain

Coding Session

🤗 + FastAPI

Feedback on Lecture and Concepts?

To Do

- HF + Fast API assignment due December 23rd EOD
- Start working in your capstone project

Reminders!!

- NO CLASS on December 24th and December 31st
- YES to the Office Hours on December 20th and December 22nd
- YES to the Office Hours in 2023
- Look for the announcements for office hours for Dec 27th and December 29th
- **NEXT CLASS is on January 7th!!** - Initial Capstone Presentations!!!

See you next time!