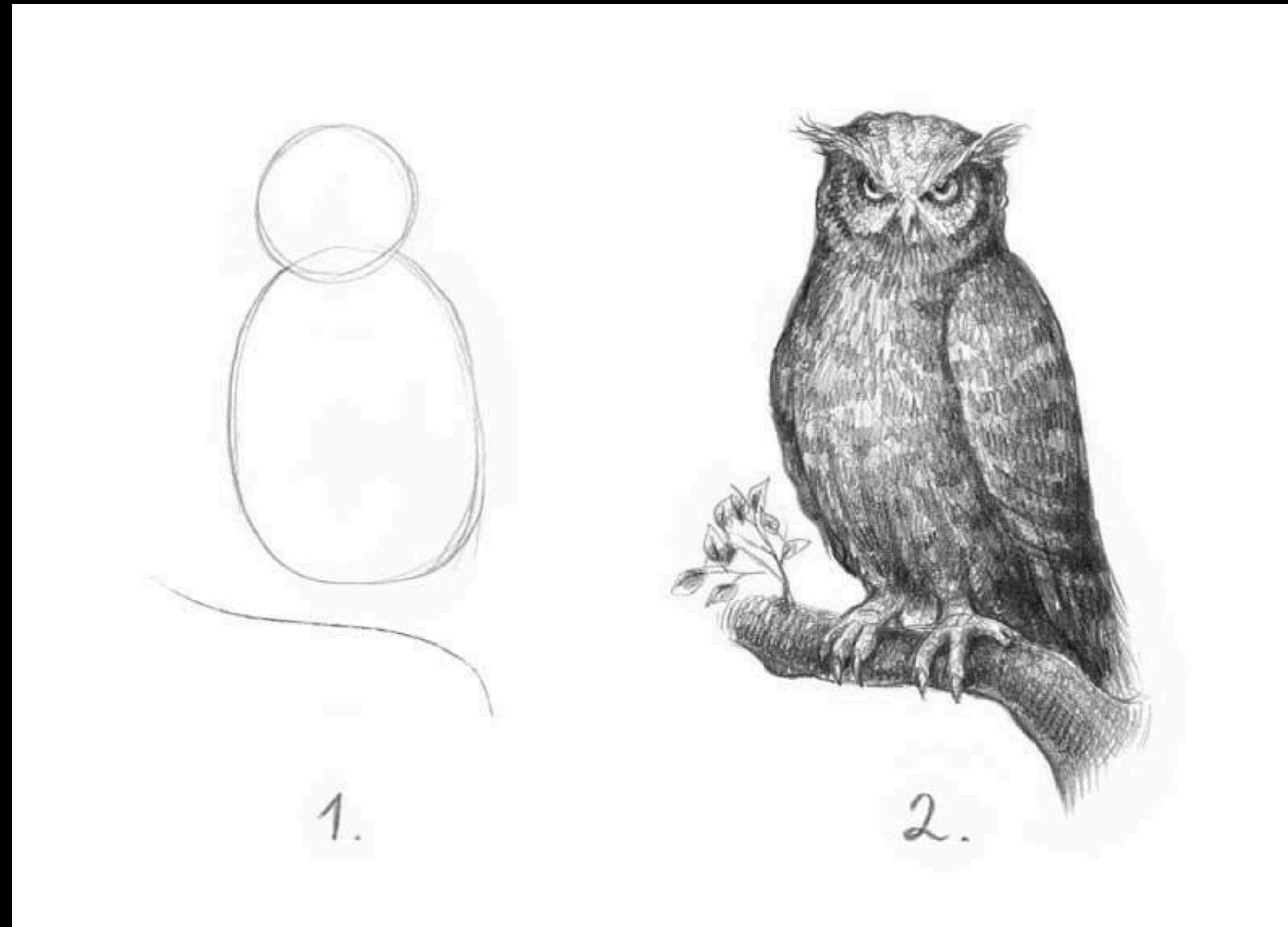


Beginning

Deploying ML models



1. Build the model

2. Deploy it



Chip Huyen
@chipro

...

Machine learning engineering is 10% machine learning
and 90% engineering.

1:40 AM · Oct 13, 2020 · Twitter Web App

593 Retweets

44 Quote Tweets

7,865 Likes



Elon Musk @elonmusk · Oct 13, 2020

...

Replying to @chipro

Yeah

109

121

5.4K



ML Engineering:

Engineering

ML

Me:

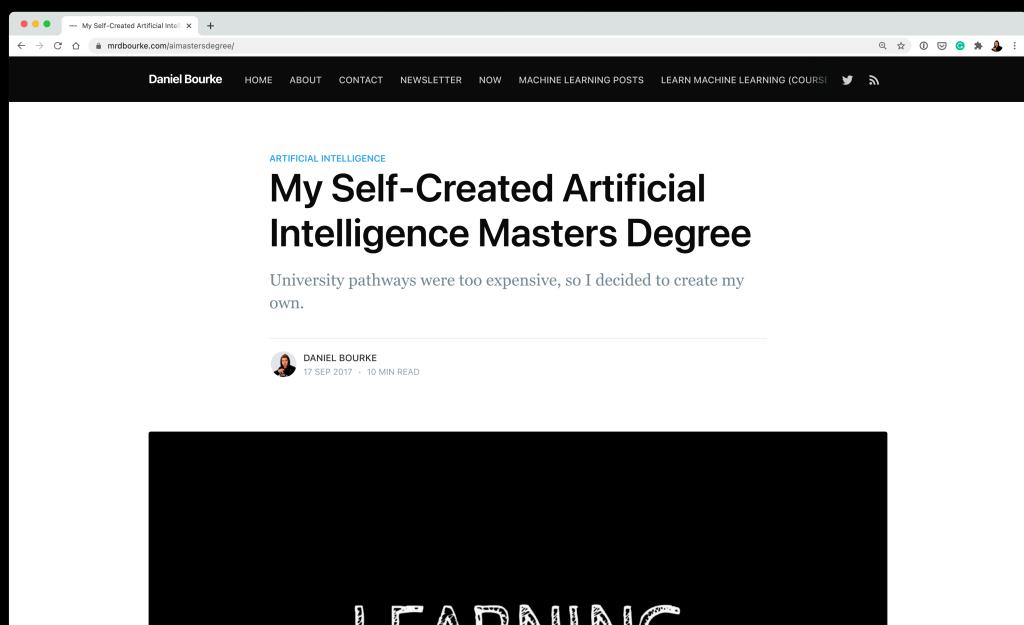
Engineering

ML

also me...



4.94 ★



Max Kelsen

small apps: yes



large apps: not yet



what I'm up to now: <https://www.mrdbourke.com/now>

Get all the stuff

<https://dbourke.link/cs329s-ml-tut>

Note: README.md still work in progress
(will fix after tutorial)

Food Vision 🍔👁 ML Deployment Recipe

(~10 mins)

Ingredients

- Data: ~10 Food Classes from Food 101
- TensorFlow/PyTorch model(s)
- Bunch of Python scripts
- Makefile/Dockerfile



Utensils

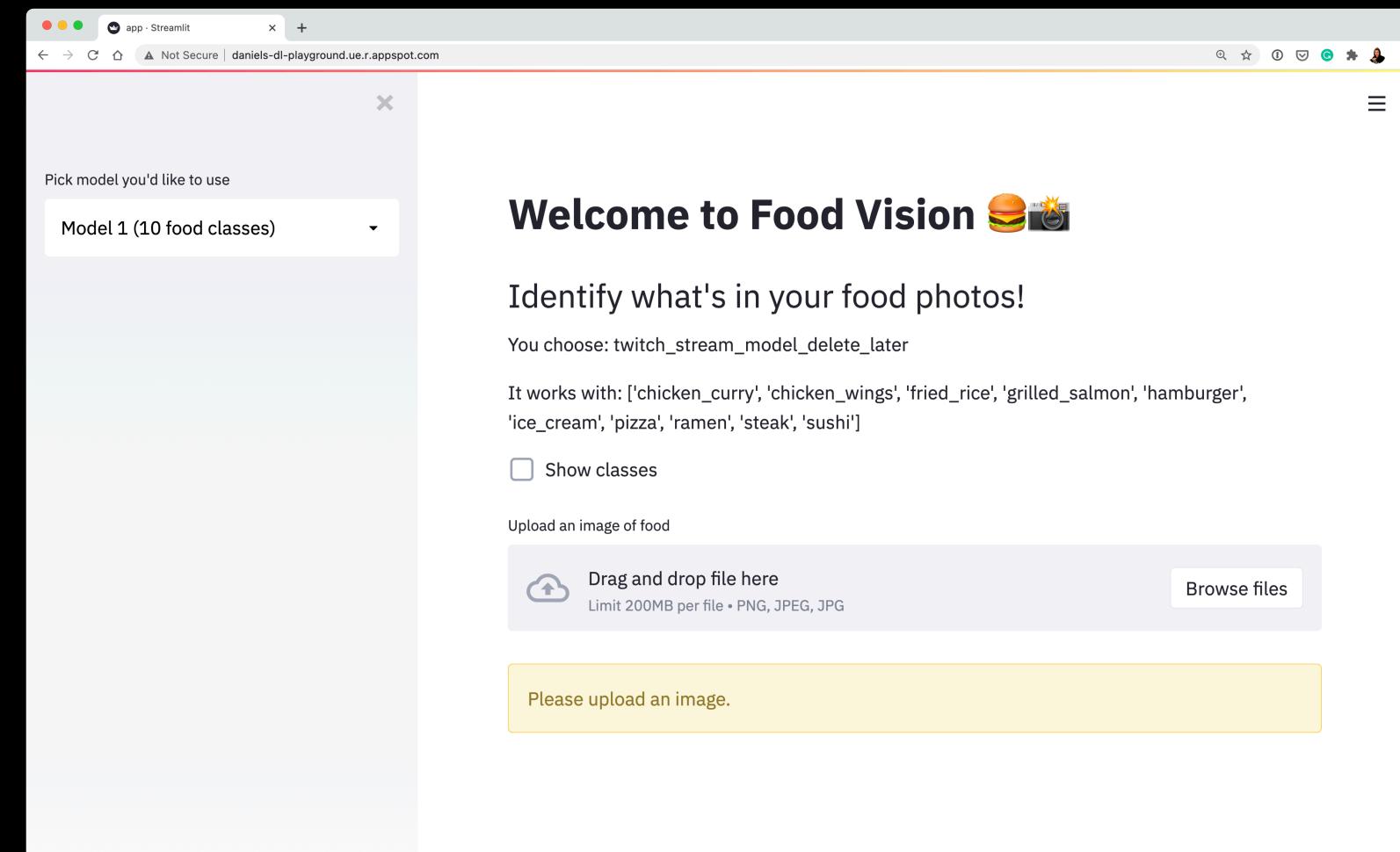
- Streamlit
- gcloud SDK (CLI tool)
- Google Cloud Project
- Google Storage
- Google AI Platform
- Docker
- Google Container Registry
- App Engine

(40-50 mins?)

Method

1. Get the app working locally
2. Deploy the model to AI Platform
3. Deploy the app to App Engine
4. ?????
5. Solve all the bugs
6. PROFIT

Outcome



+





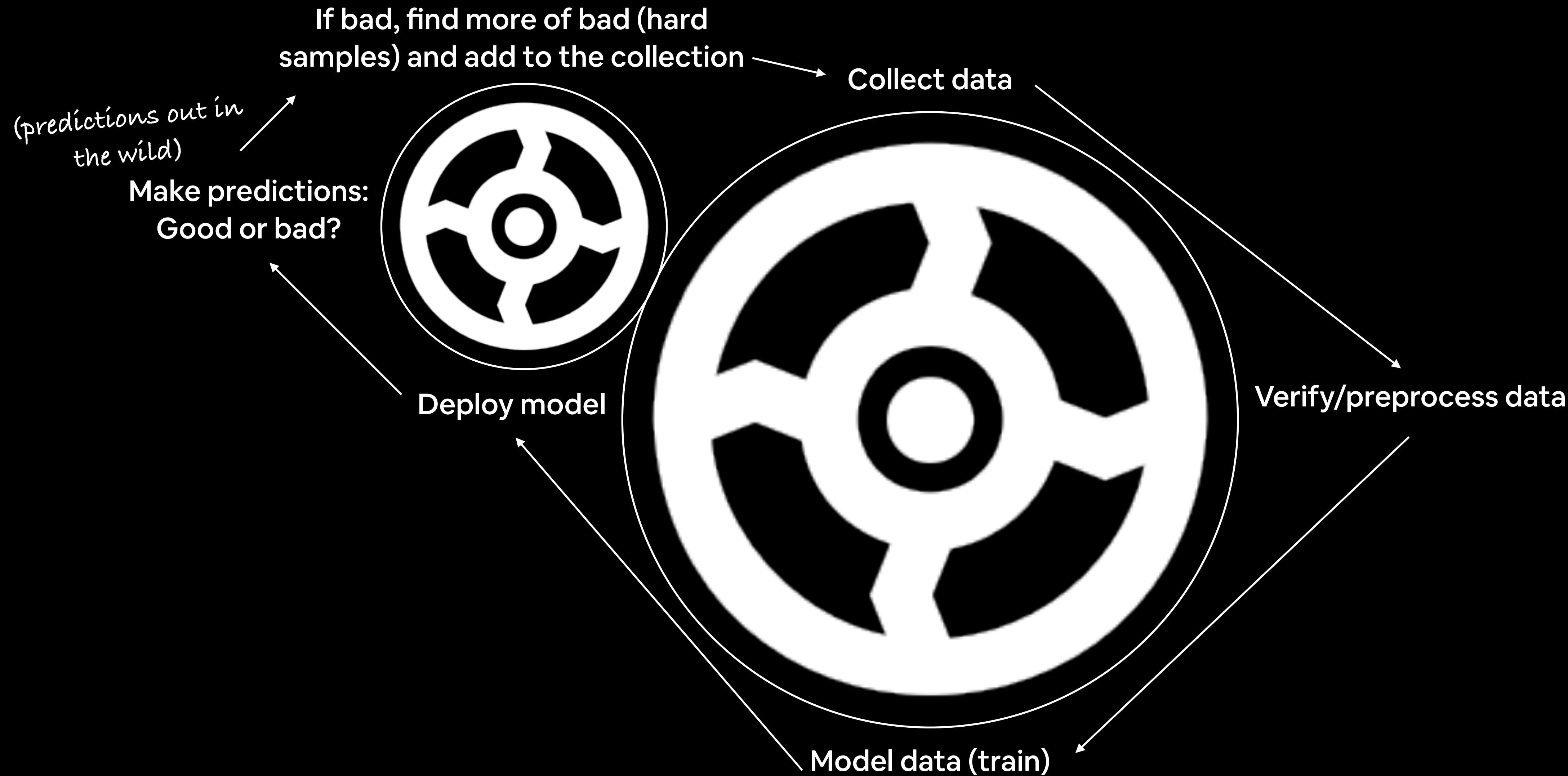
Google Cloud services cost 💰
No credits = 💸



Ask questions whenever you feel like it
(we'll do QA at the end too)

Objective: Build a data flywheel

(the holy grail of ML)

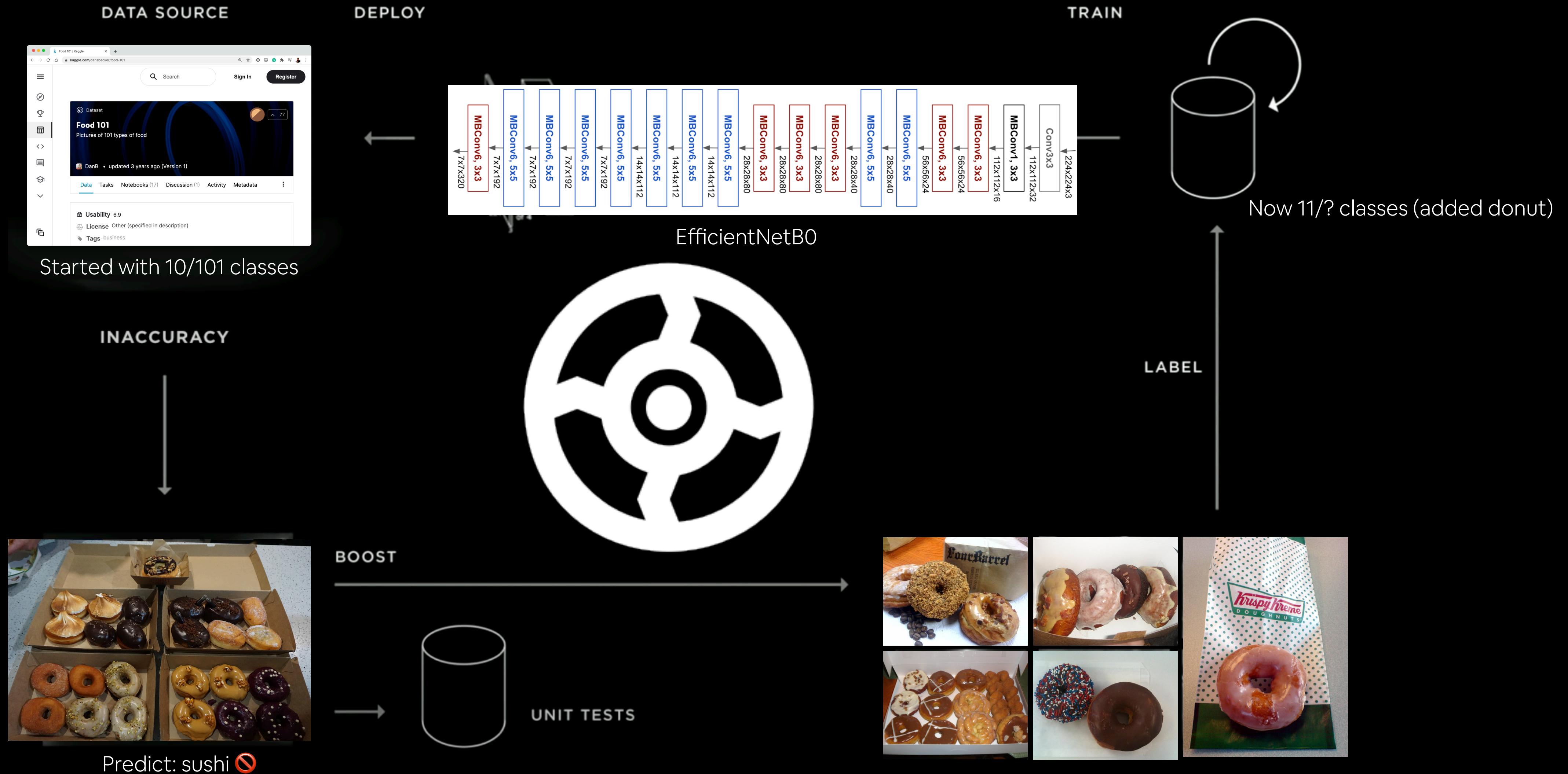


Tesla's data flywheel



Source: <https://youtu.be/Ucp0TTmvqOE>

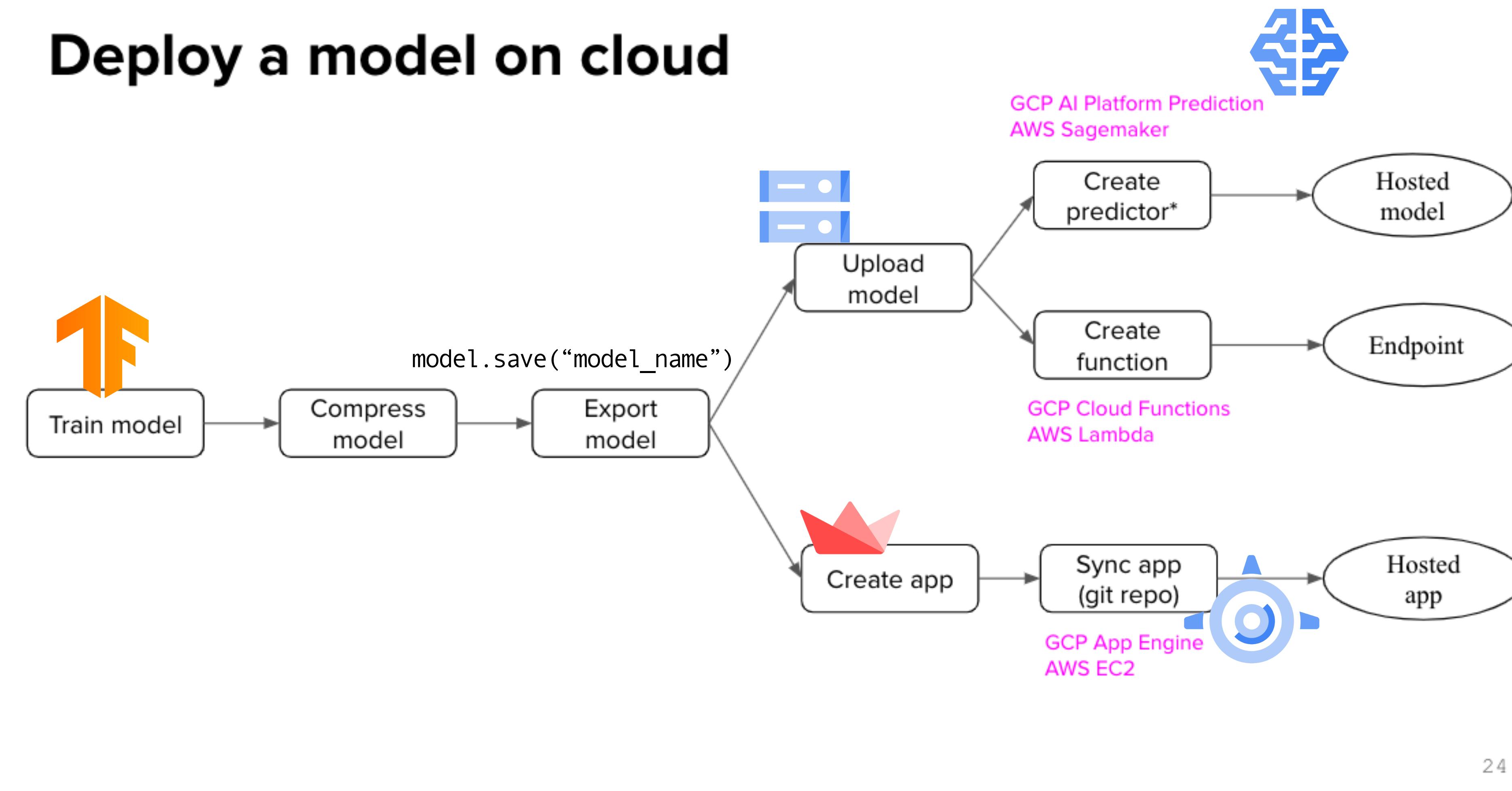
Food Vision 🍔’s data flywheel



Source: <https://youtu.be/Ucp0TTmvqOE> + my own flavouring

Middle

Deploy a model on cloud

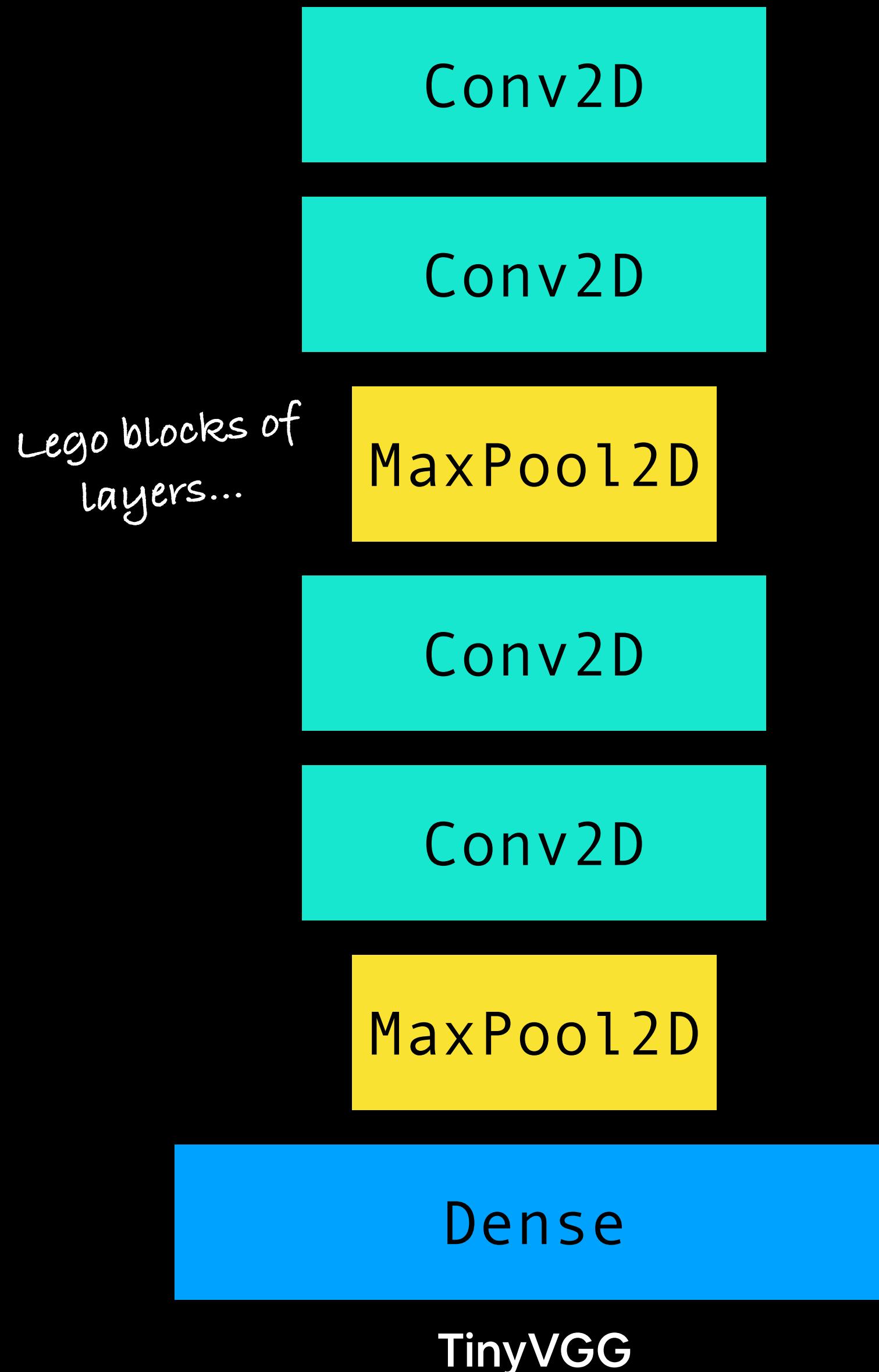


24

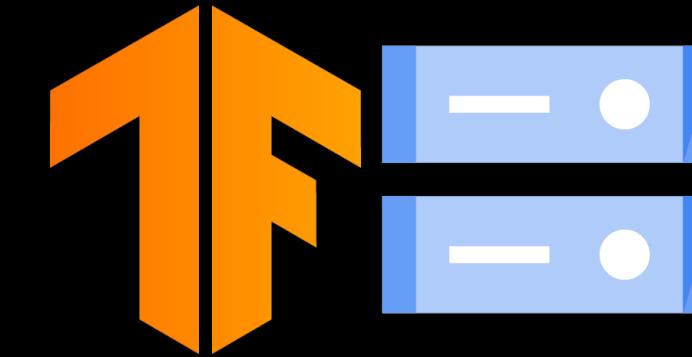
Source: Chip's Slides

model building == model deployment

(sort of)

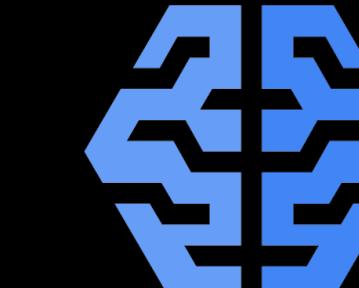


TensorFlow model on Google Storage



Lego blocks of tools...

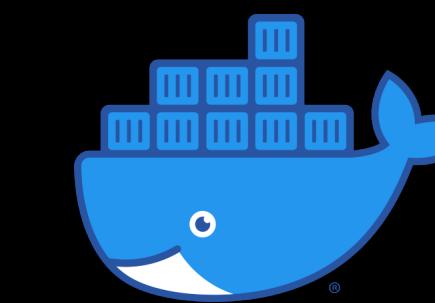
Model hosted on AI Platform



App built with Streamlit



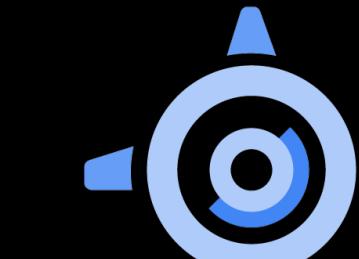
App wrapped up in Docker container



Docker container deployed to GCR



GCR hosted container deployed to App Engine



App monitored with Google Monitoring



GCR = Google Container Registry

Method

- 1. Get the app working locally**
- 2. Deploy the model to AI Platform**
- 3. Deploy the app to App Engine**
- 4. ?????**
- 5. Solve all the bugs**
- 6. PROFIT**

model building == model deployment

(sort of)

Start here



Finish here

“Everything works nicely until you deploy your model into the wild...”

- Smart Machine Learning Engineer

Problems you're going to run into...

- What do you do when your model fails (when not if)...
- What shape does your data come in? (JPEG vs PNG)
- ML Engine hard limit: 1.5MB
- What do you do when your user has a class you've never seen?
 - When do you retrain your model?
- What do you do when a user has a class not even in your scope?

MACHINE LEARNING



Problem #1 (everything looks like Daddy)

whyatt

MACHINE LEARNING



Problem #2 (everything still looks like Daddy)

End



Google Cloud services cost 💰
No credits = 💸
Shut everything down!

Learn GCP... where?

The screenshot shows a web browser window for Qwiklabs - Hands-On Cloud Training at google.qwiklabs.com. The page displays a 'Welcome, Daniel!' message and information about the service's offerings. On the left, a sidebar menu includes 'Home' (selected), 'Catalog', and 'My Learning'. A central section highlights temporary credentials and various learning options. Below this, a 'In Progress' section shows a card for a 'Machine Learning APIs' quest, which is marked as 'Advanced'.

Qwiklabs - Hands-On Cloud Training

google.qwiklabs.com

Google Cloud Training

Search

Welcome, Daniel!

We give you temporary credentials to Google Cloud Platform, so you can learn the cloud using the real thing – no simulations. From 30-minute individual labs to multi-day courses, from introductory level to expert, instructor-led or self-paced, with topics like machine learning, security, infrastructure, app dev, and more, we've got you covered.

In Progress

QUEST
Machine Learning APIs
Advanced

Extensions

- CI/CD (make a change... app updates automatically)
- Codify everything! (we did a lot of clicking around...)
- Actually log the data from your app (e.g. save logs to BigQuery/Google Storage)