Courses 7 - Production ML Systems

Module 5: Hybrid ML Systems

Lesson Title: **Introduction**

Format: Presenter

Presenter: Val

Video Name: T-PSML-O_5_I1_introduction

Google Cloud

Hybrid ML Systems

Lak Lakshmanan

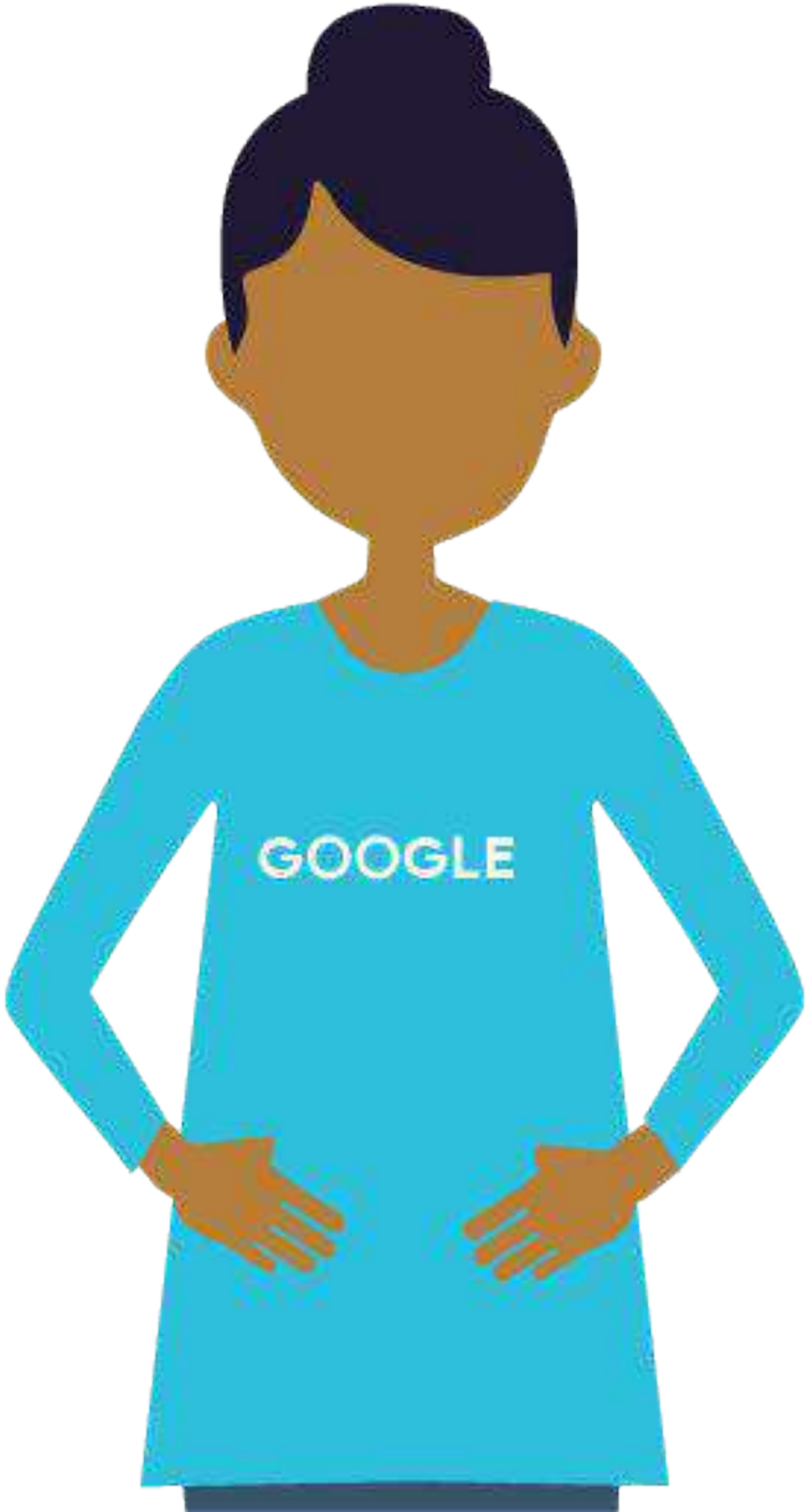
Title Safe >

< Action Safe

Learn how to...

Build hybrid cloud machine learning models

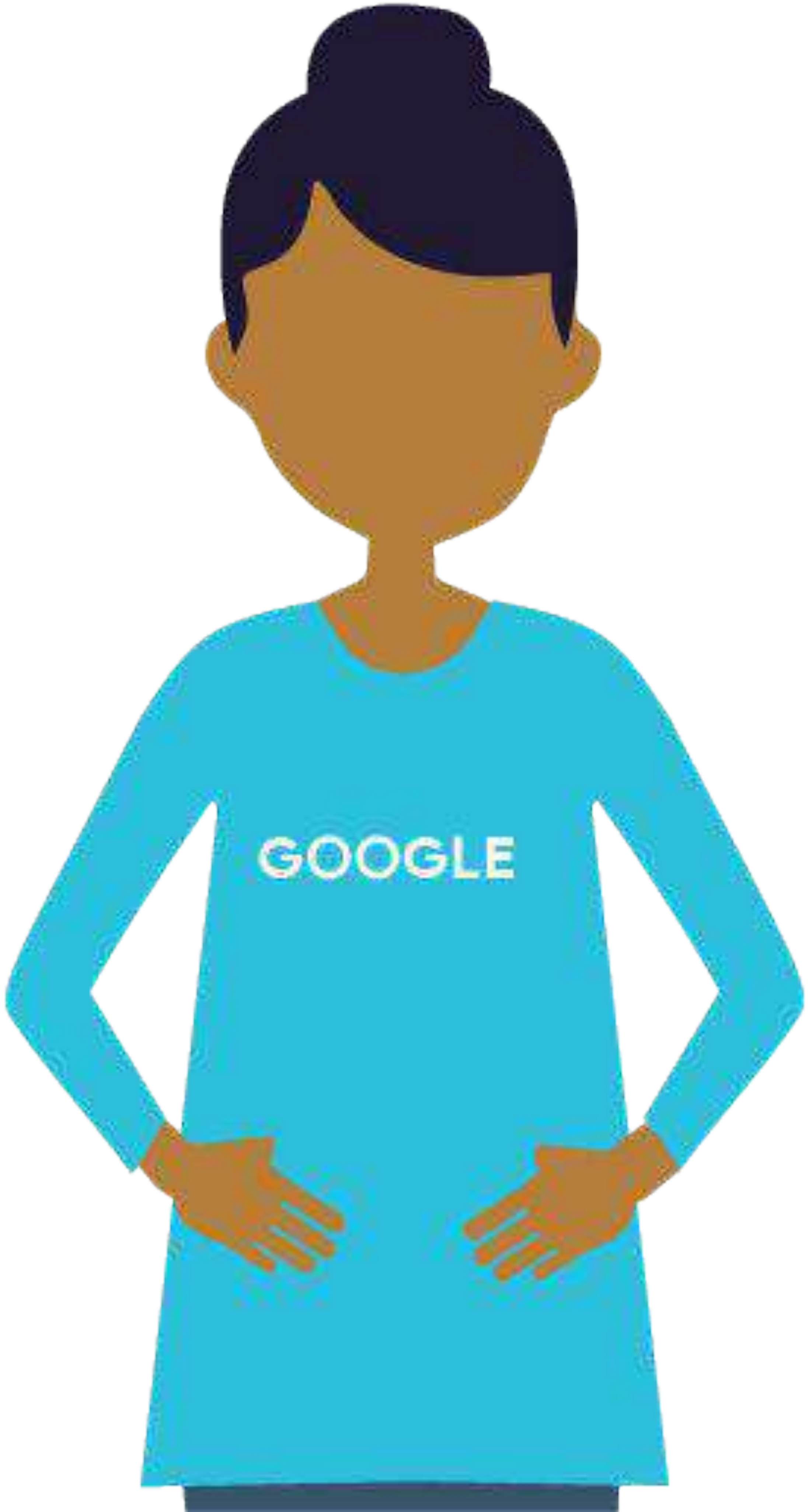
Optimize TensorFlow graphs for mobile

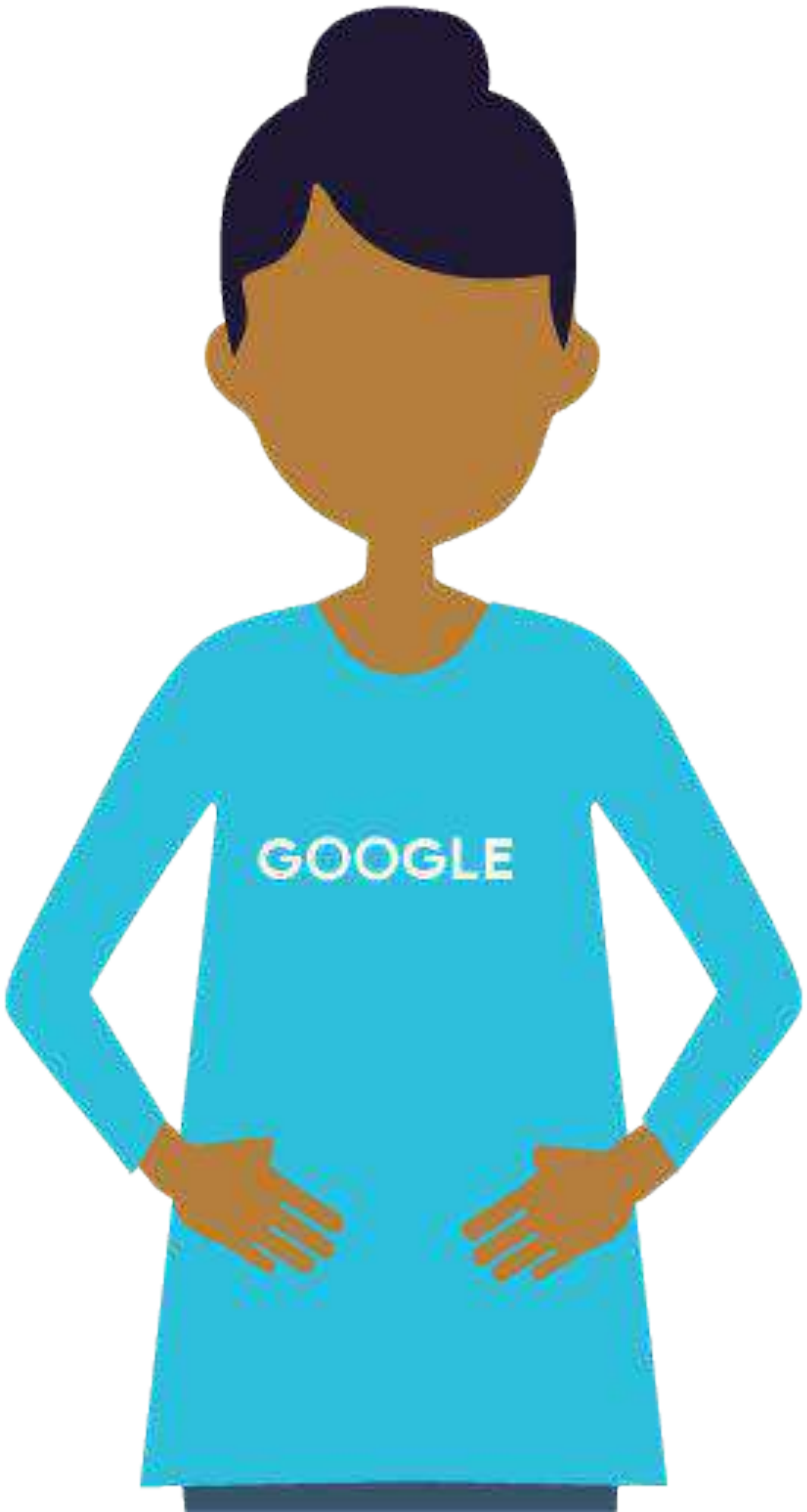


Agenda

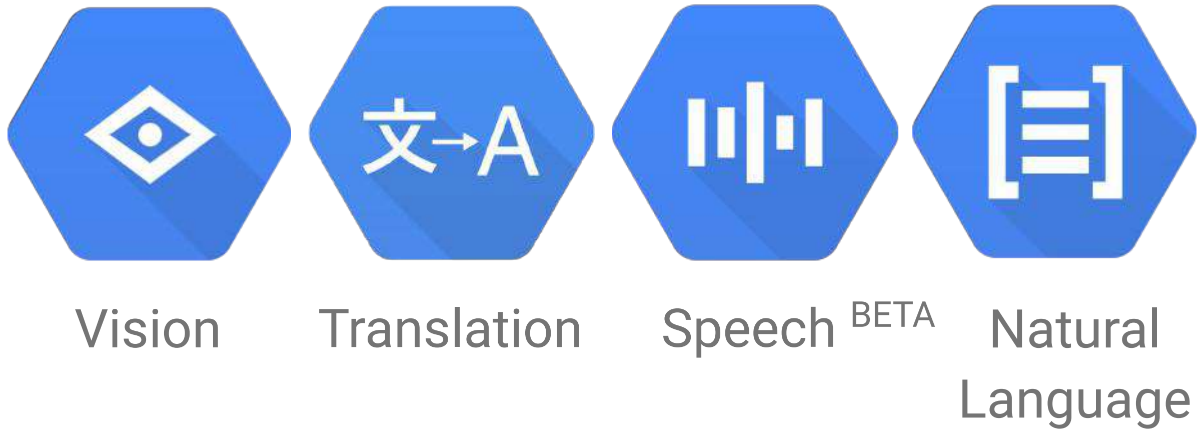
Kubeflow for hybrid cloud

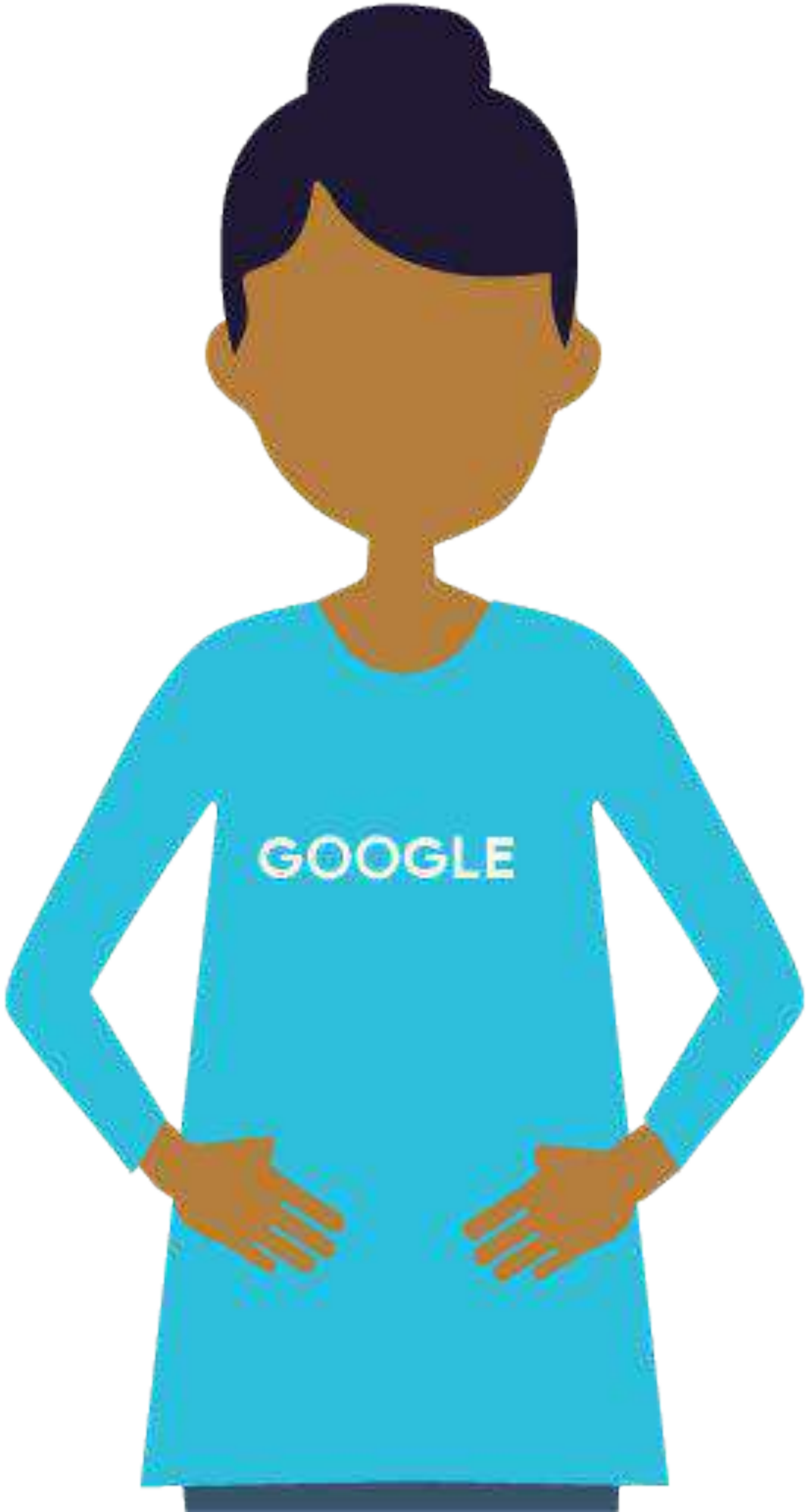
Optimizing TensorFlow for mobile





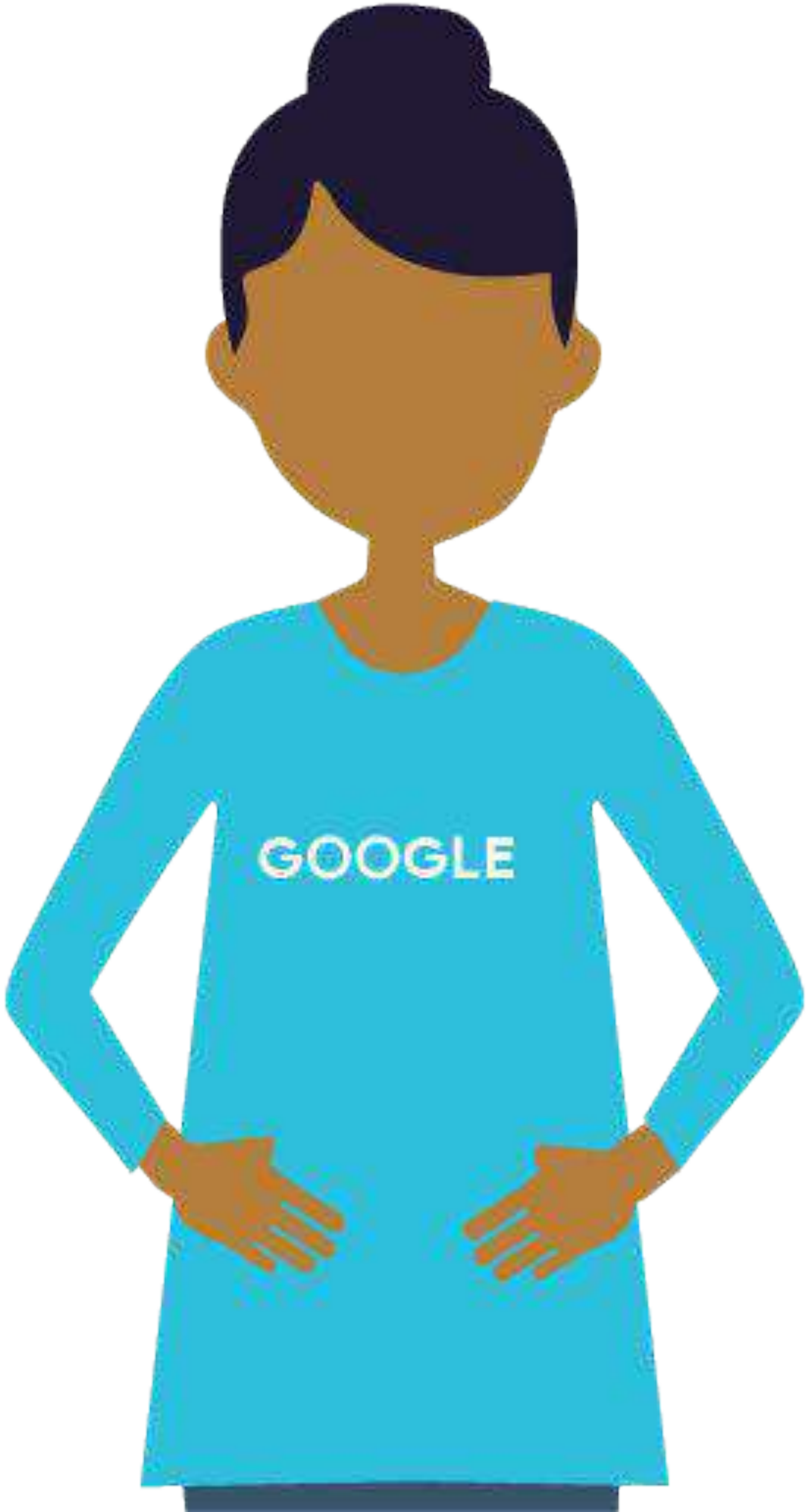
Choose from ready-made ML models



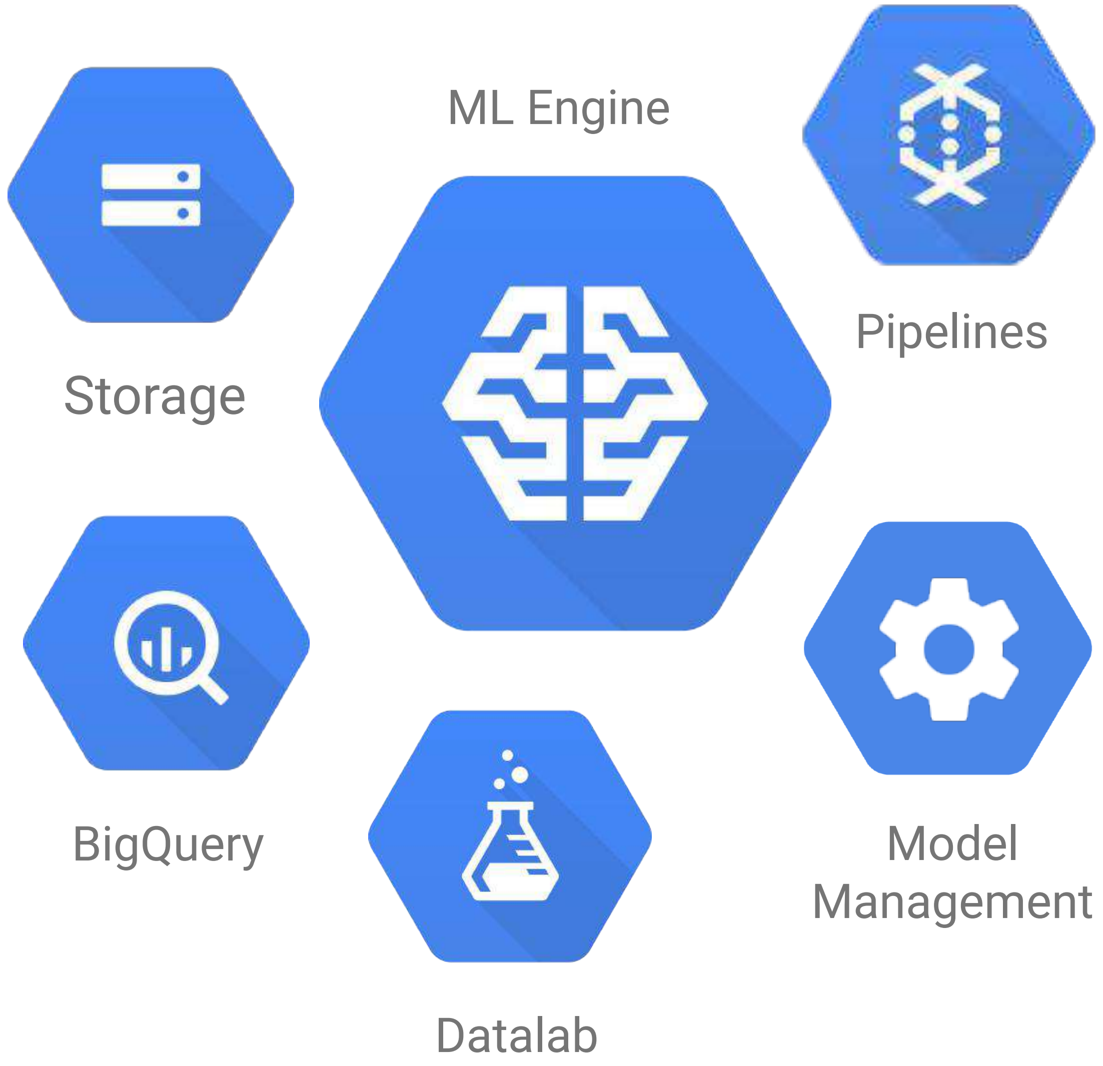


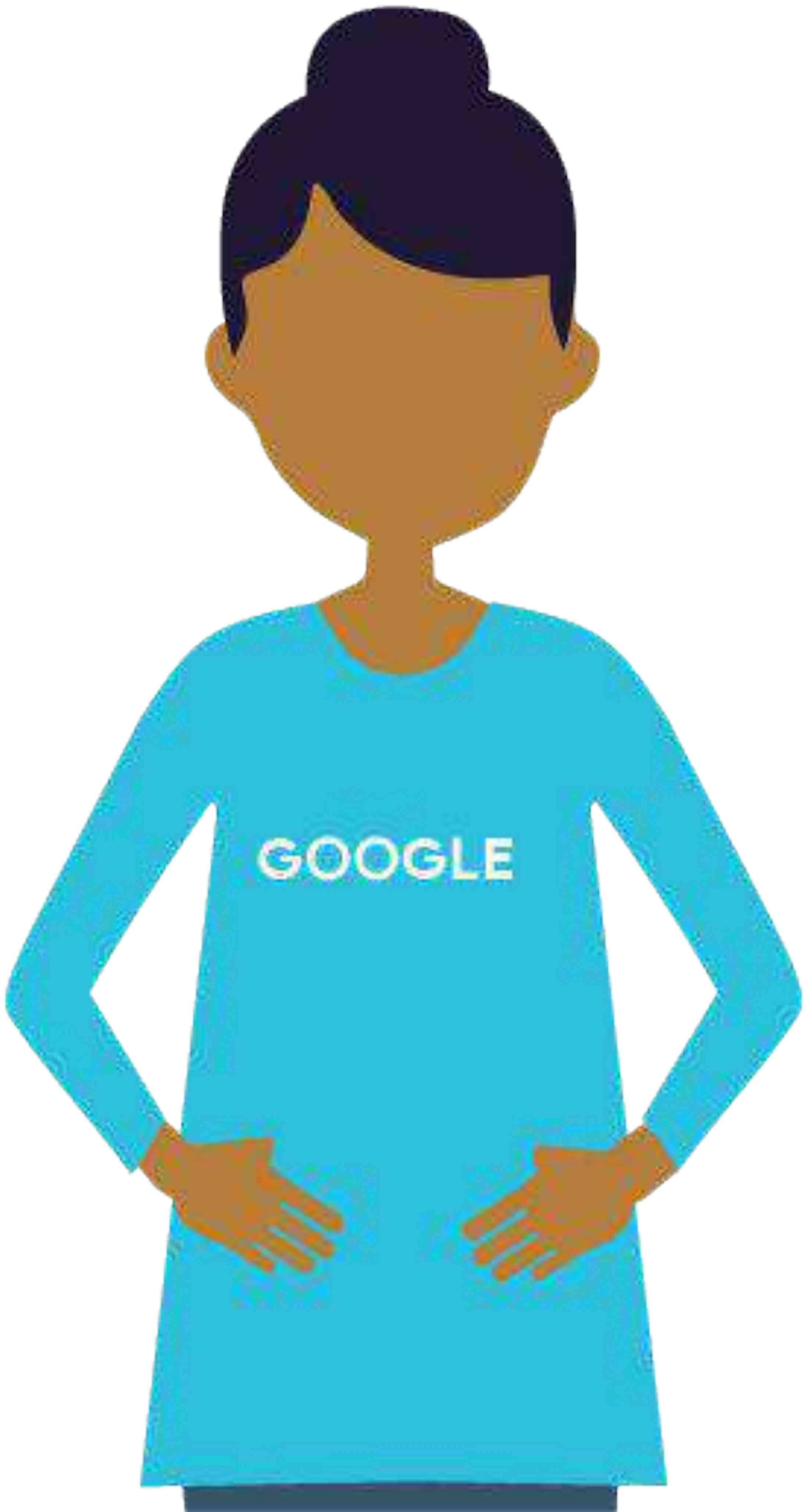
Customize ready-made ML models

Auto-ML



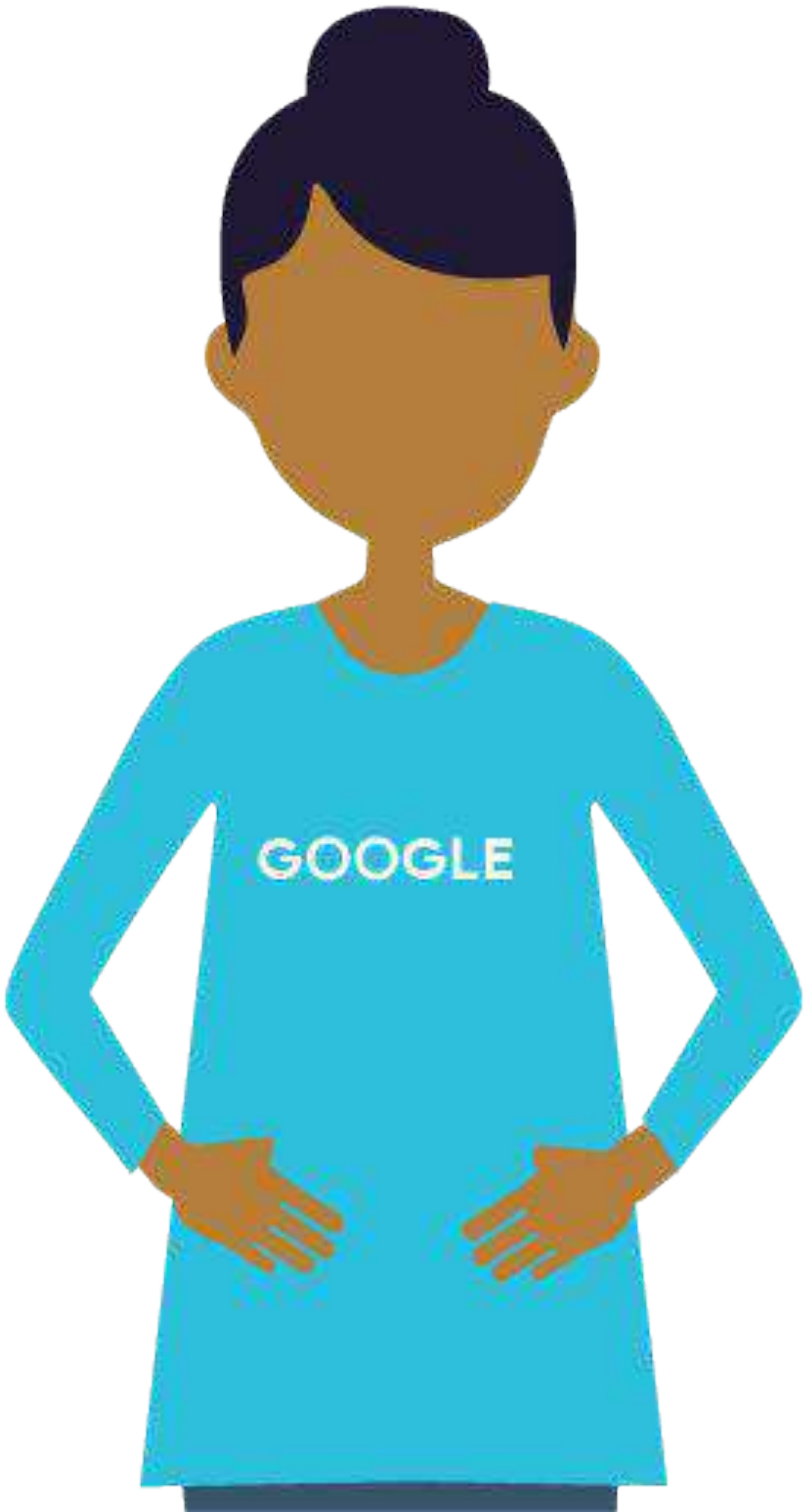
Build, train, and serve, your own custom ML Models





ML runtimes in a cloud-native environment

1. Prototype with Cloud Datalab or Deep Learning Image



ML runtimes in a cloud-native environment

1. Prototype with Cloud Datalab or Deep Learning Image

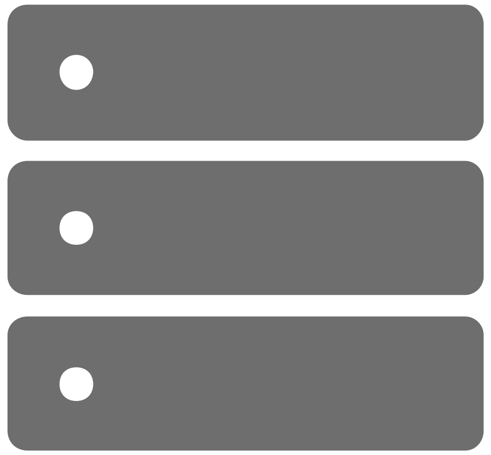
2. Distribute and autoscale training and predictions with Cloud ML Engine

You may not be able to do machine learning solely on Google Cloud

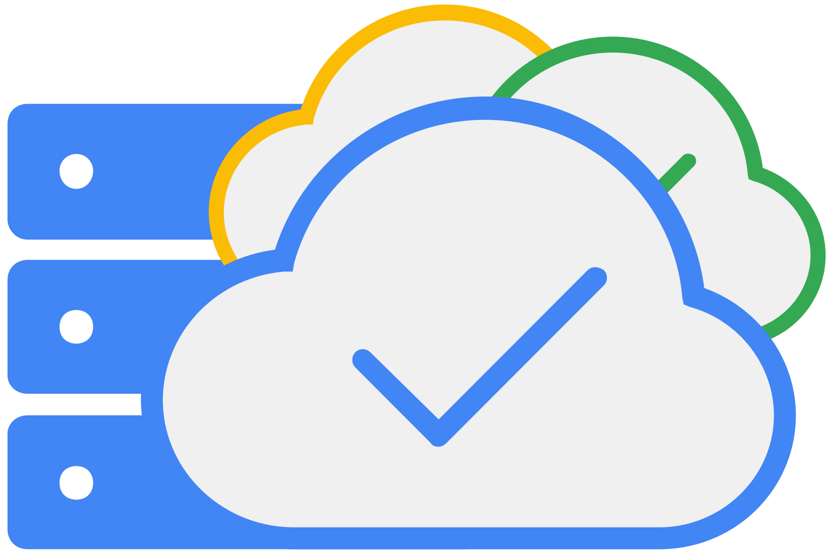
Tied to On-Premise Infrastructure

You may not be able to do machine learning solely on Google Cloud

Tied to On-Premise Infrastructure

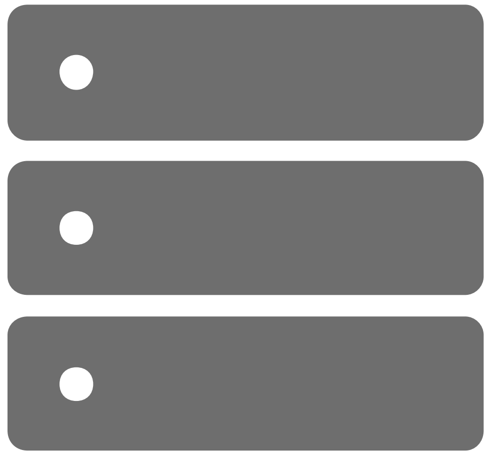


Multi Cloud System Architecture

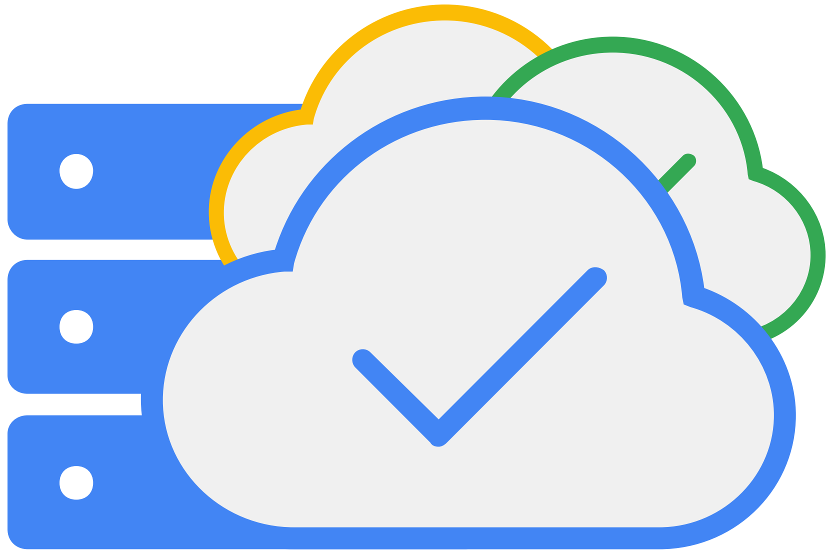


You may not be able to do machine learning solely on Google Cloud

Tied to On-Premise Infrastructure



Multi Cloud System Architecture



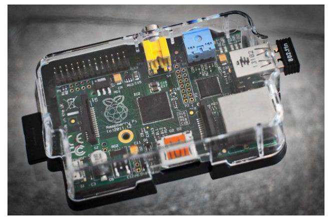
Running ML on the edge

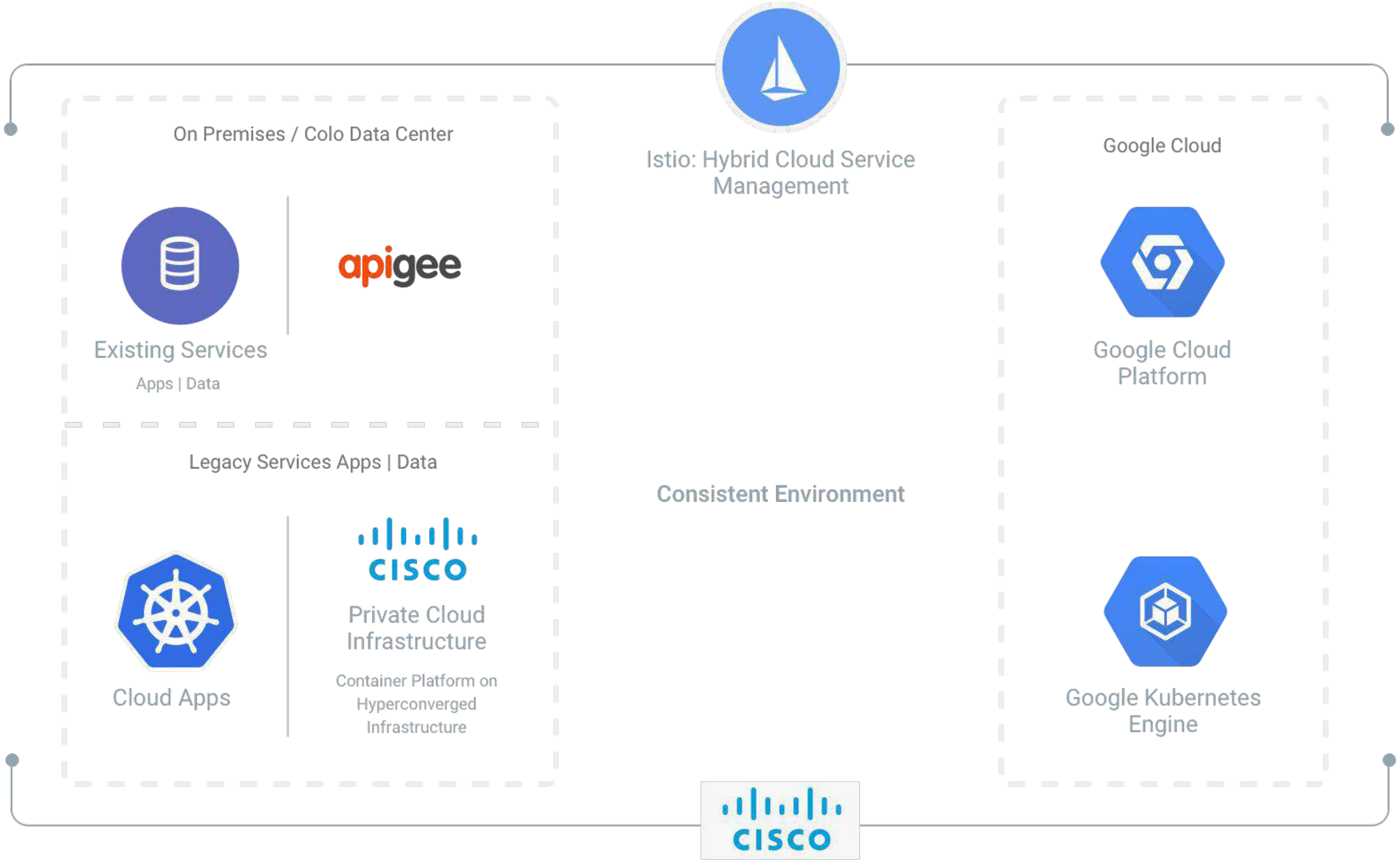
Android / iOS

15% Tabby
14% Lynx
12% Tiger Cat
10% Egyptian Cat

Edge TPU

Raspberry Pi





Istio: Hybrid Cloud Service Management

On Premises / Colo Data Center

Existing Services

Apps | Data

apigee

Legacy Services Apps | Data

Cloud Apps

cisco

Private Cloud Infrastructure

Container Platform on Hyperconverged Infrastructure

Consistent Environment

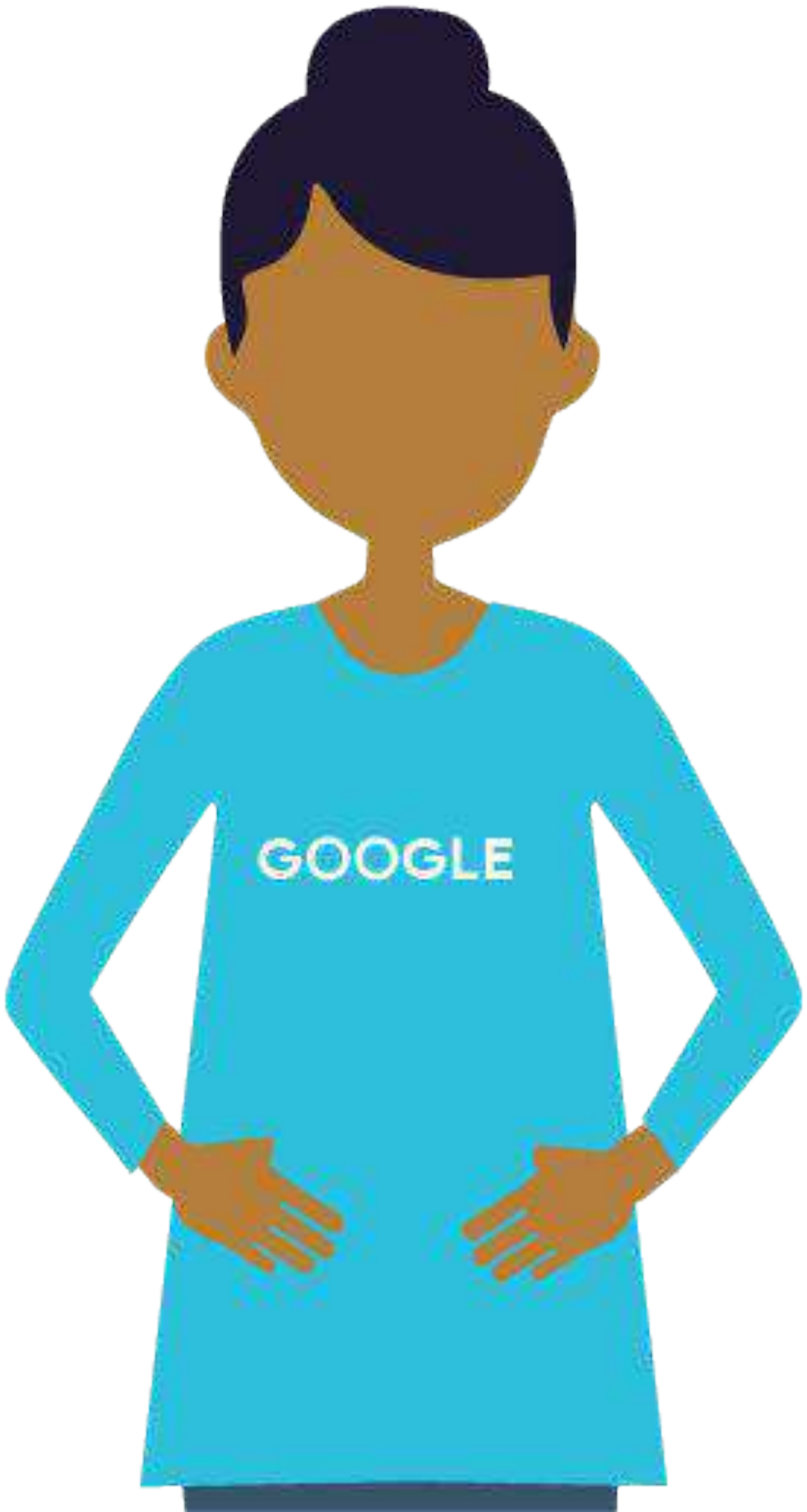
Google Cloud

Google Cloud Platform

Google Kubernetes Engine

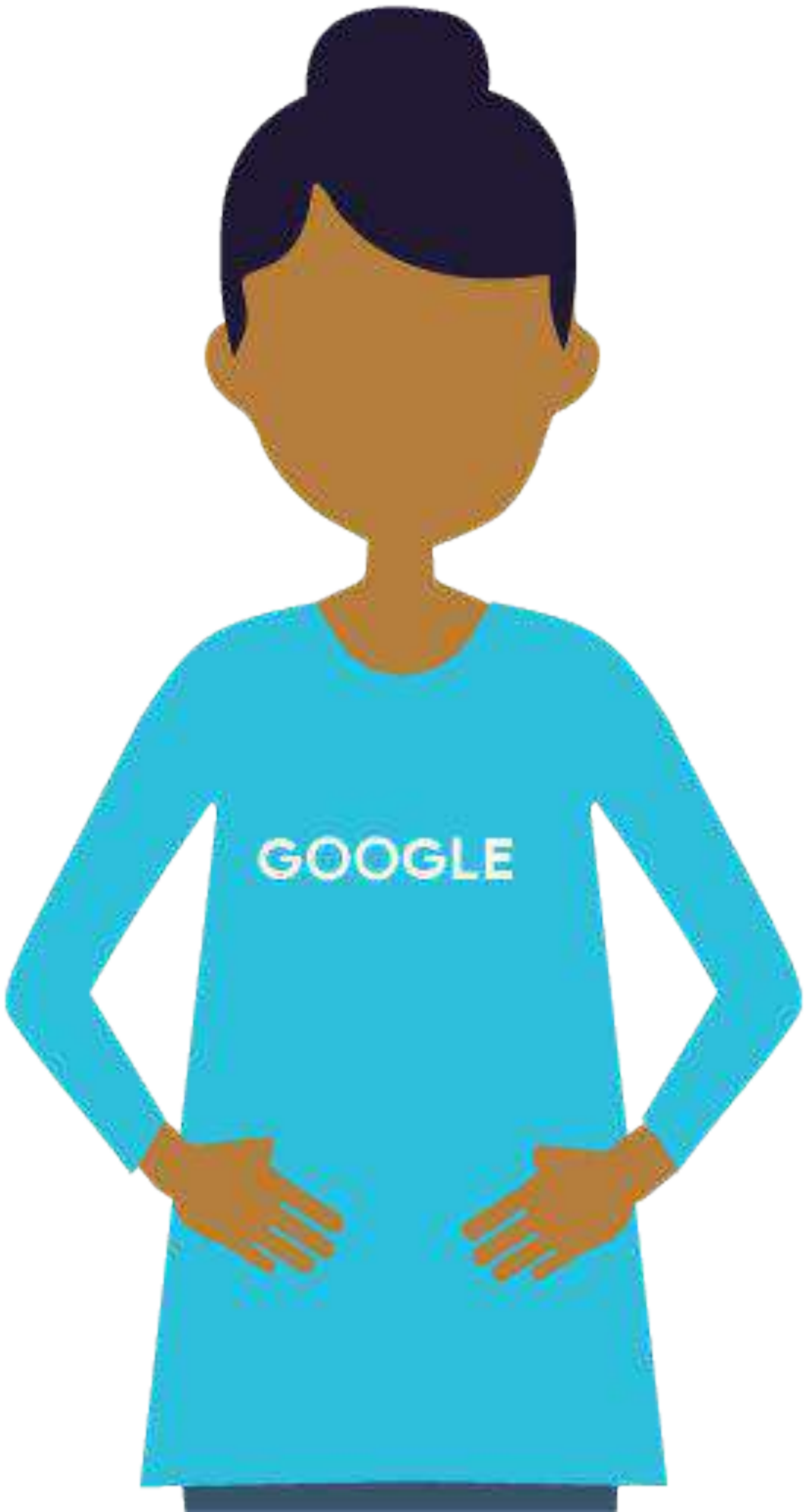
cisco

Networking | Security | Cloud Native Infrastructure | Consumption Management

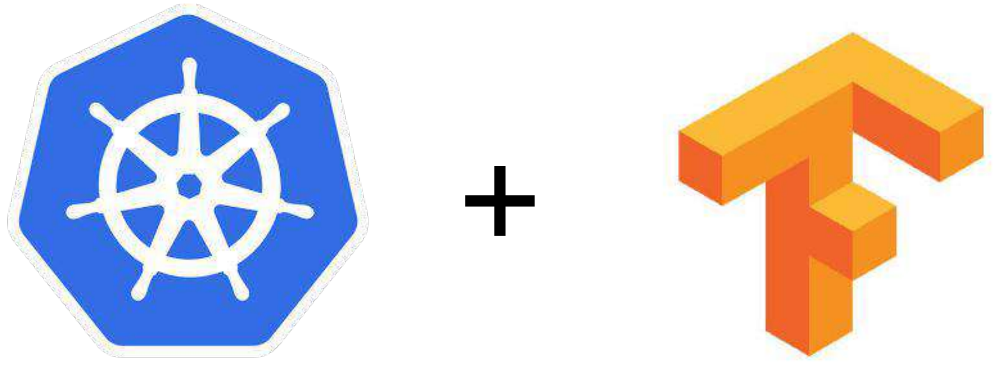


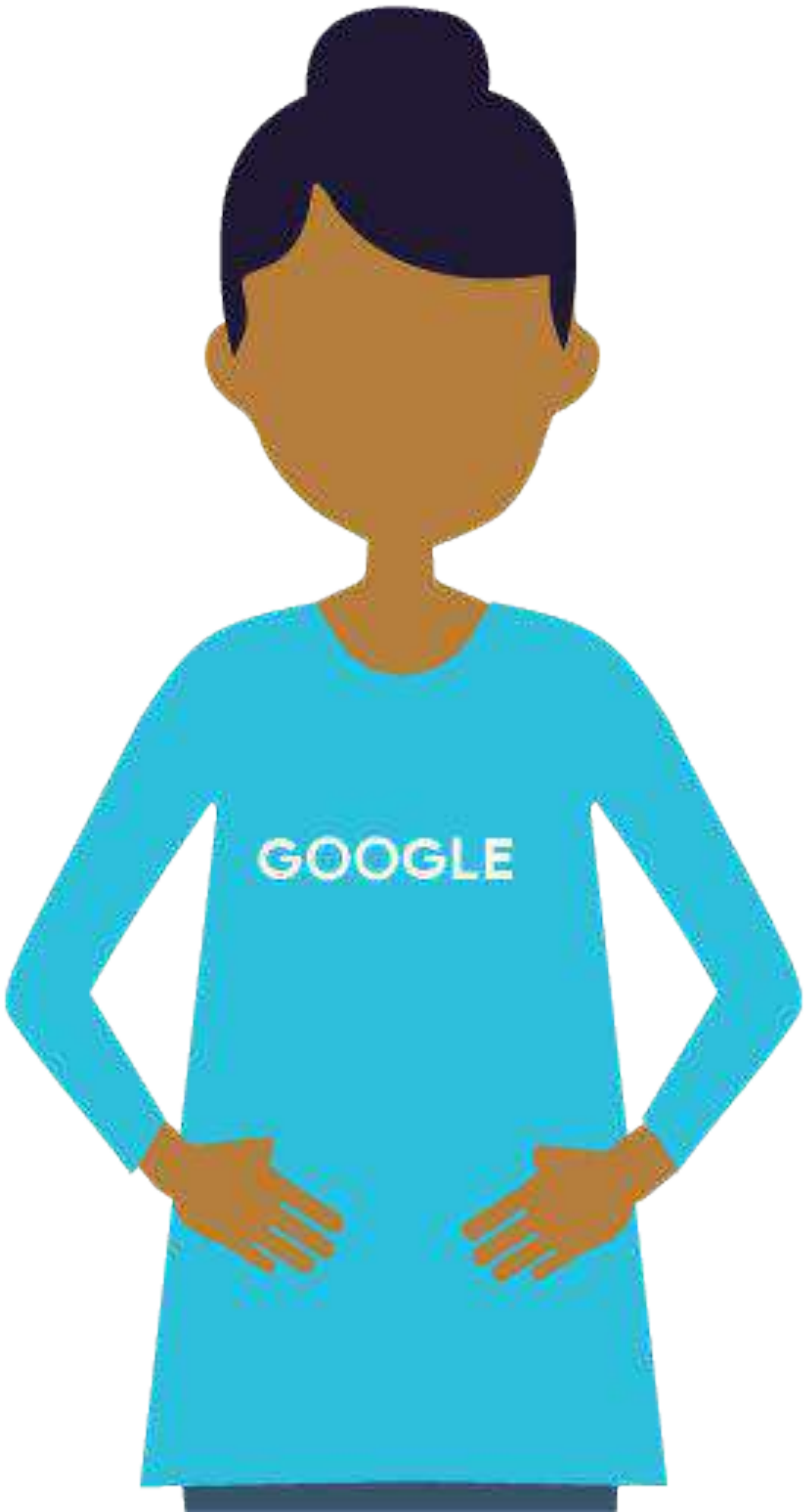
Kubernetes minimizes infrastructure management

kubernetes



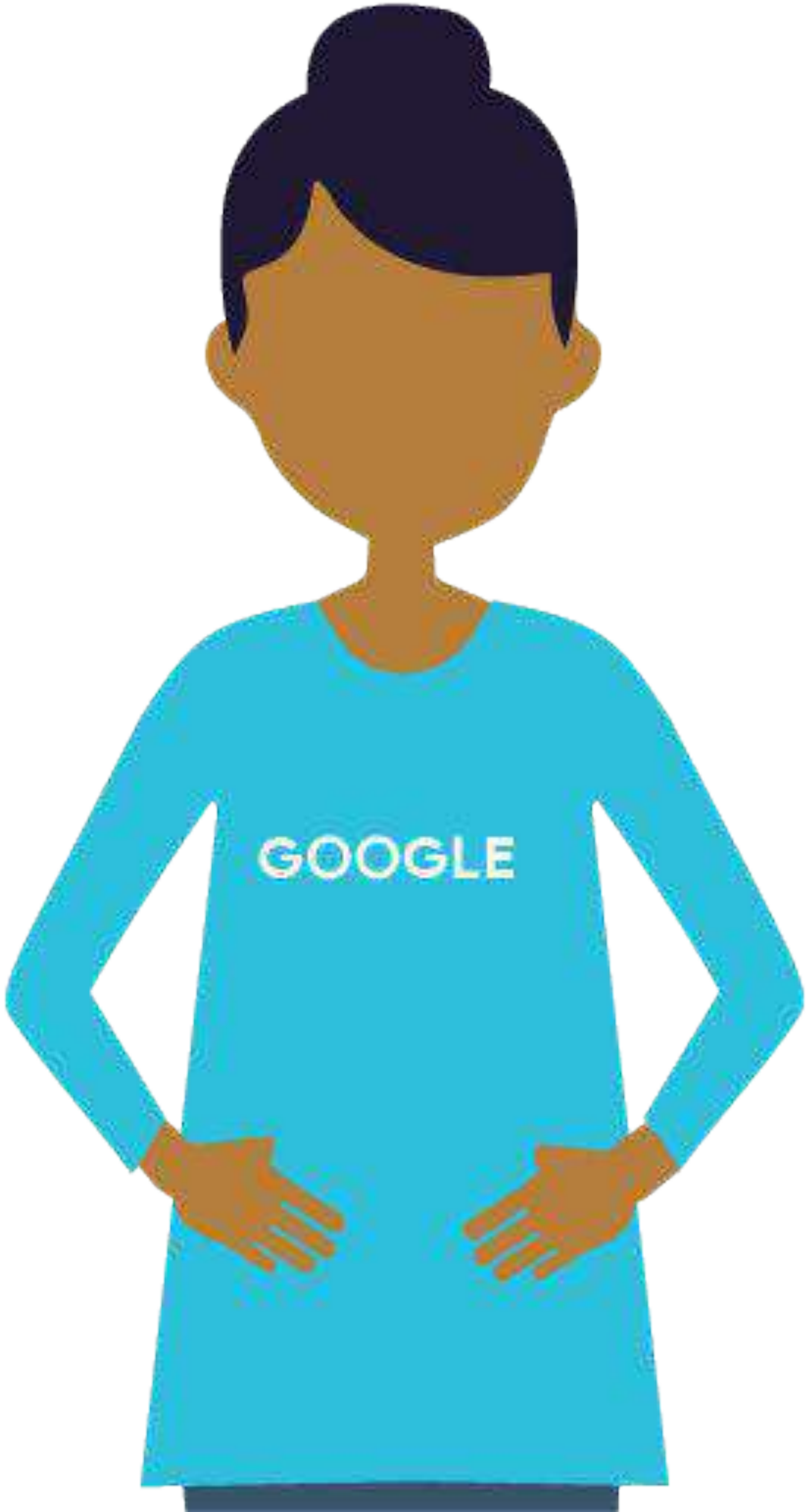
Kubeflow enables hybrid machine learning





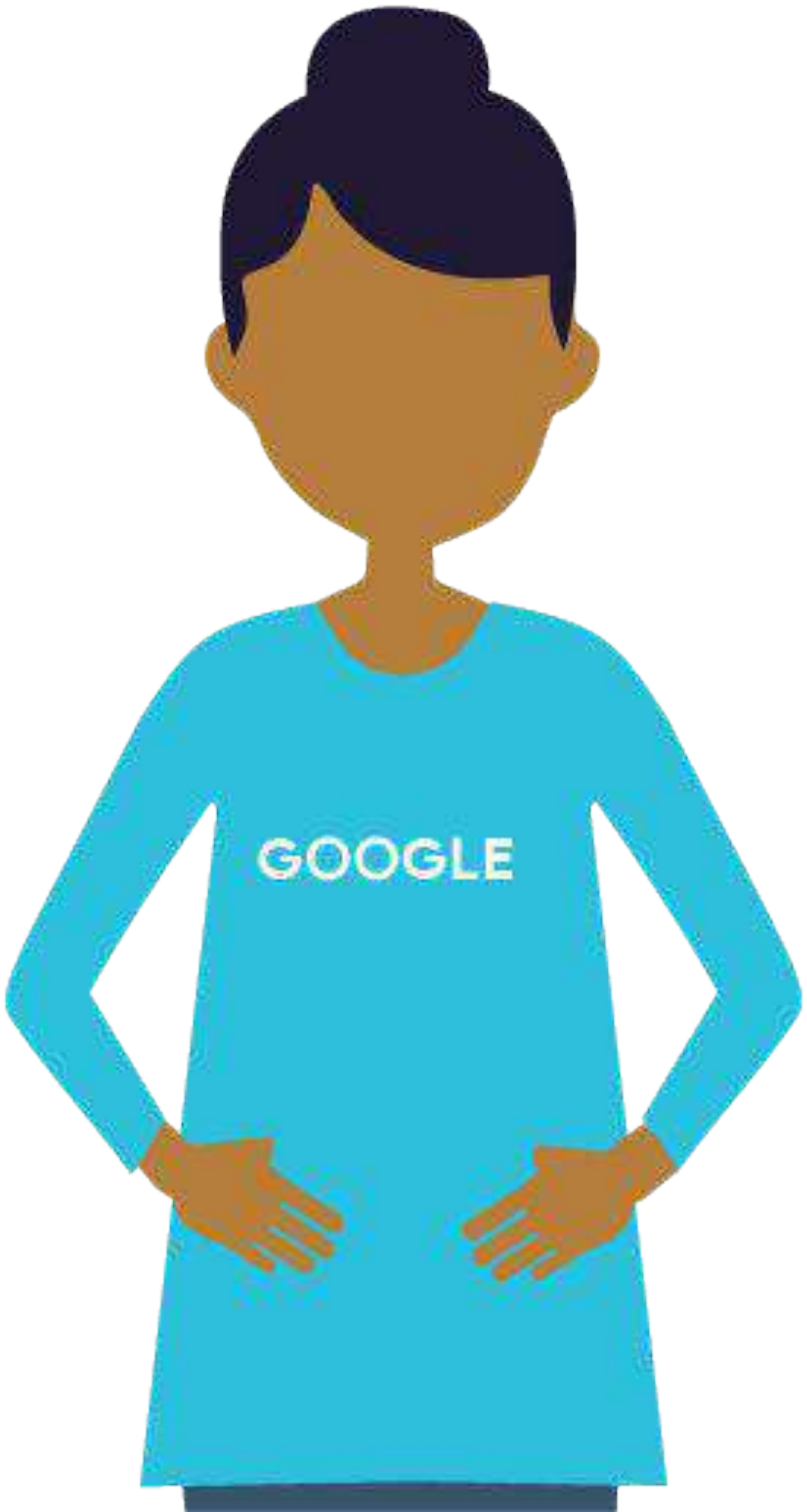
Kubeflow enables hybrid machine learning

Kubeflow



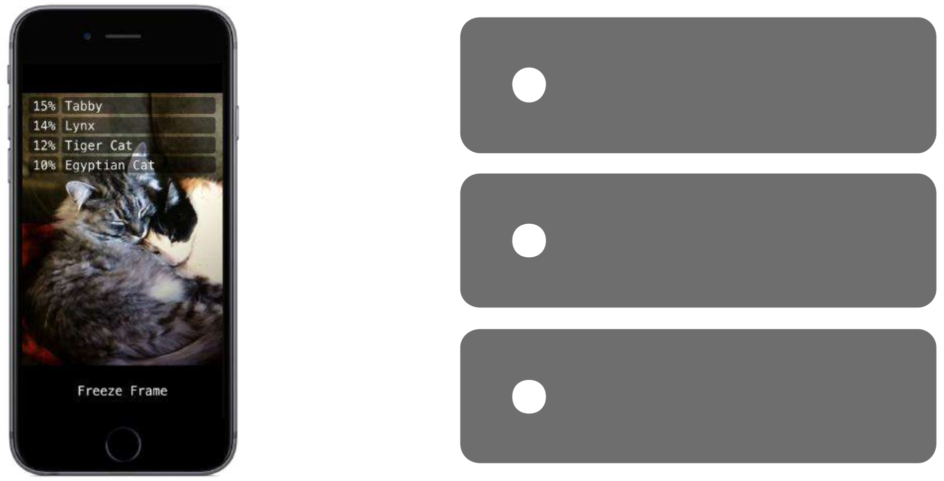
Kubeflow enables hybrid machine learning

Kubeflow on Kubernetes Engine



Kubeflow enables hybrid machine learning

Kubeflow



Courses 7 - Production ML Systems

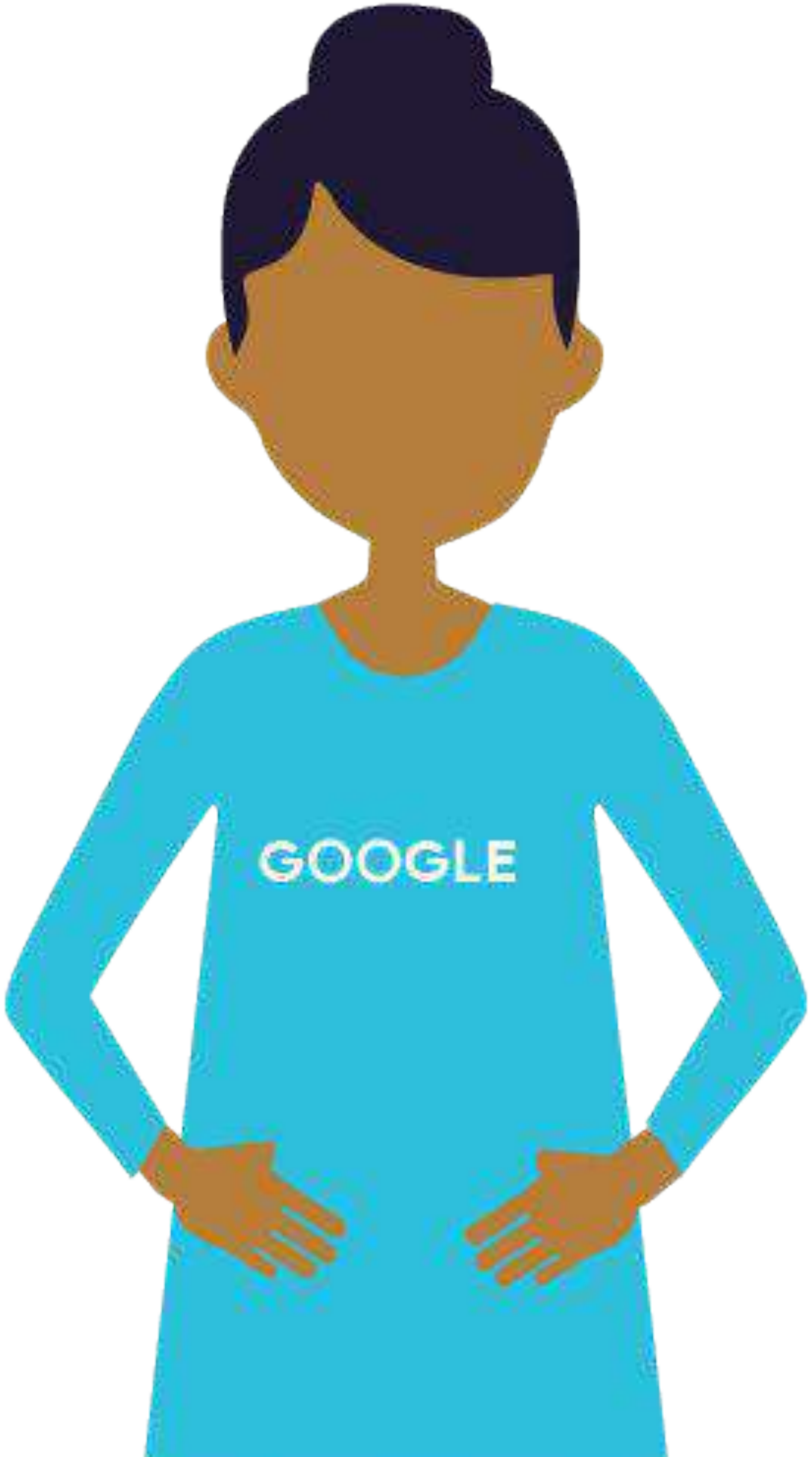
Module 5: Hybrid ML Systems

Lesson Title: **Machine learning on hybrid cloud**

Format: Presenter

Presenter: Val

Video Name: T-PSML-O_5_l2_machine_learning_on_hybrid_cloud

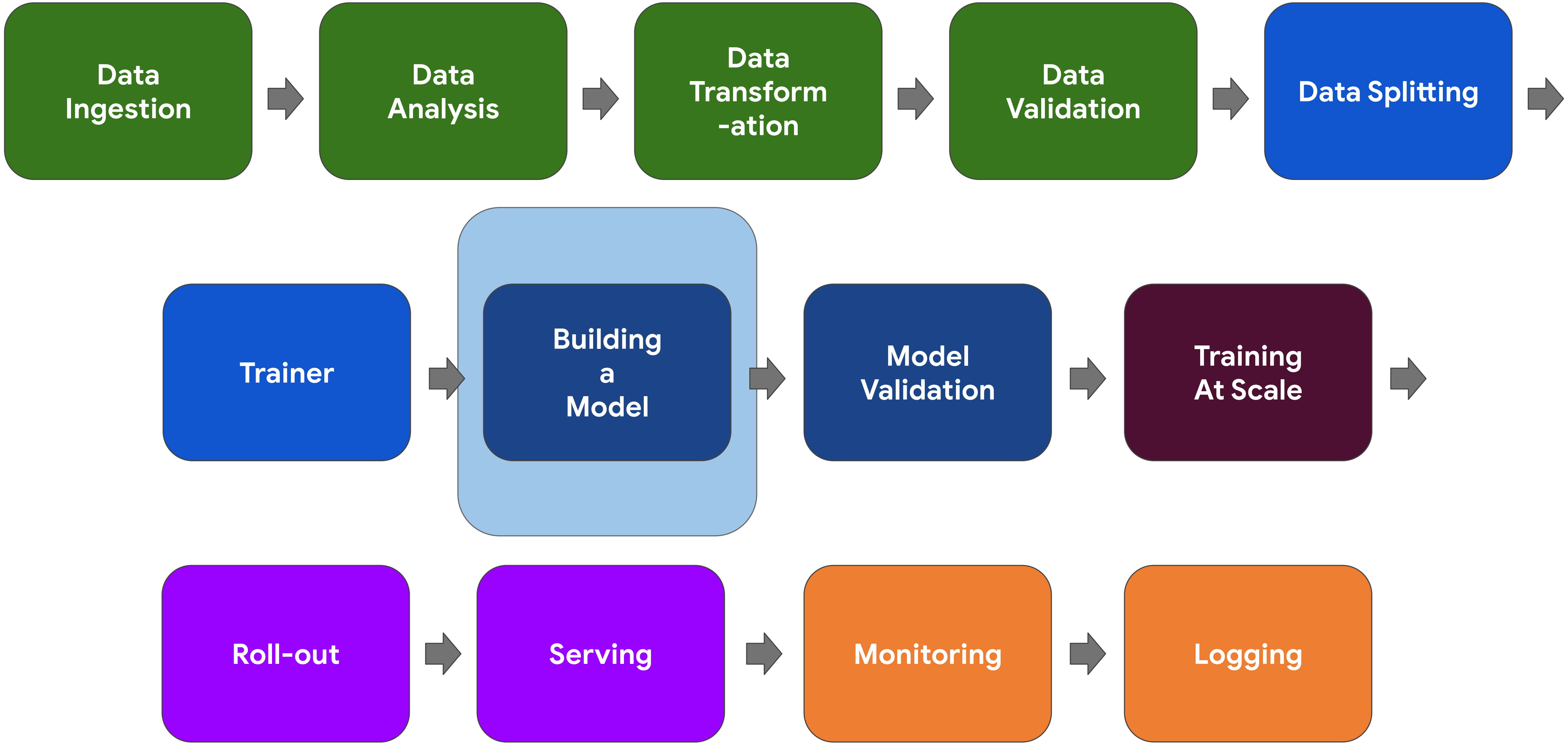


- Composability
- Portability
- Scalability

Composability

Building
a
Model

Building a model is only one part of the entire system



Data Ingestion

Data Analysis

Data Transformation

Data Validation

Data Splitting

Trainer

Building a Model

Model Validation

Training At Scale

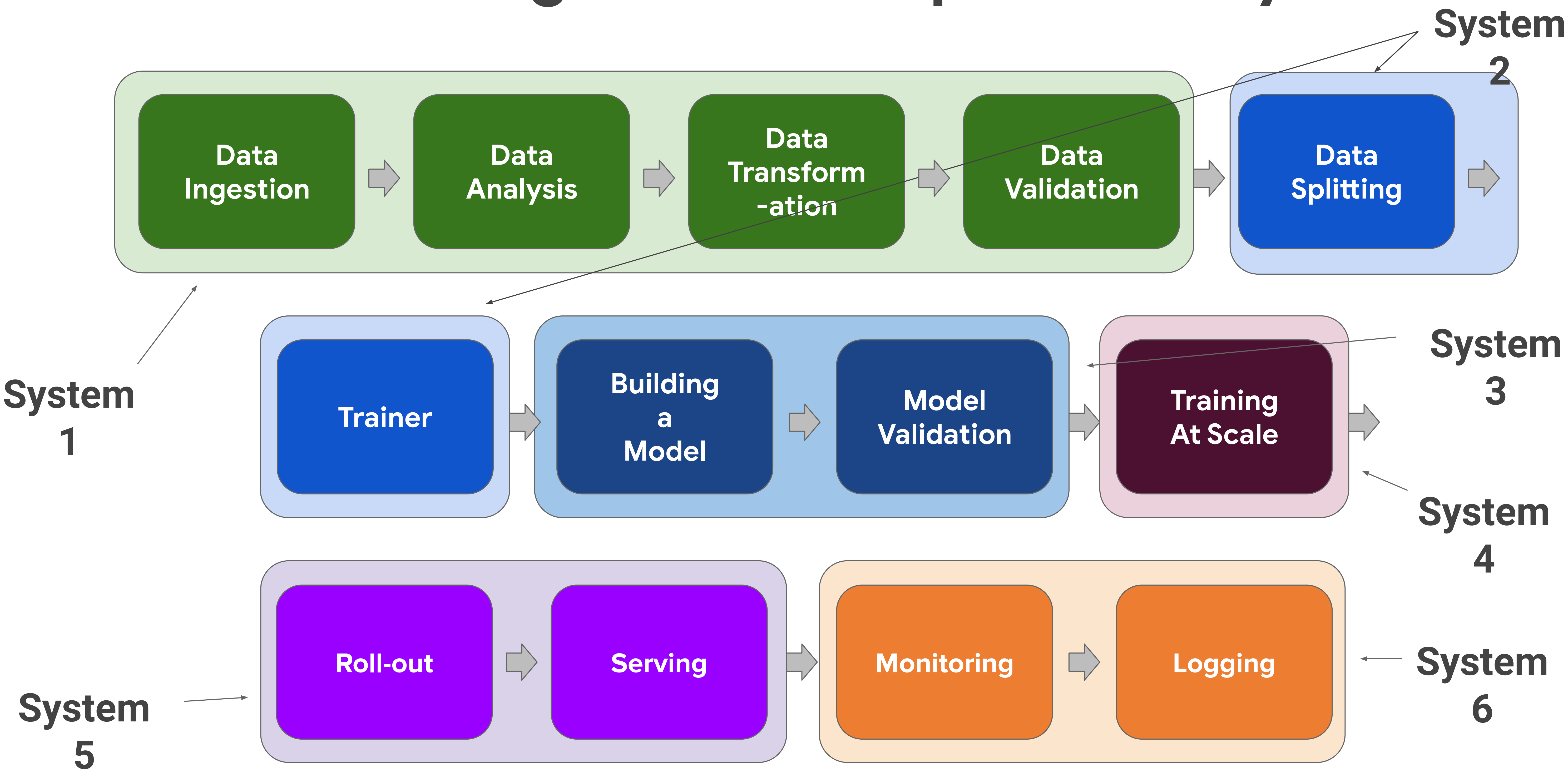
Roll-out

Serving

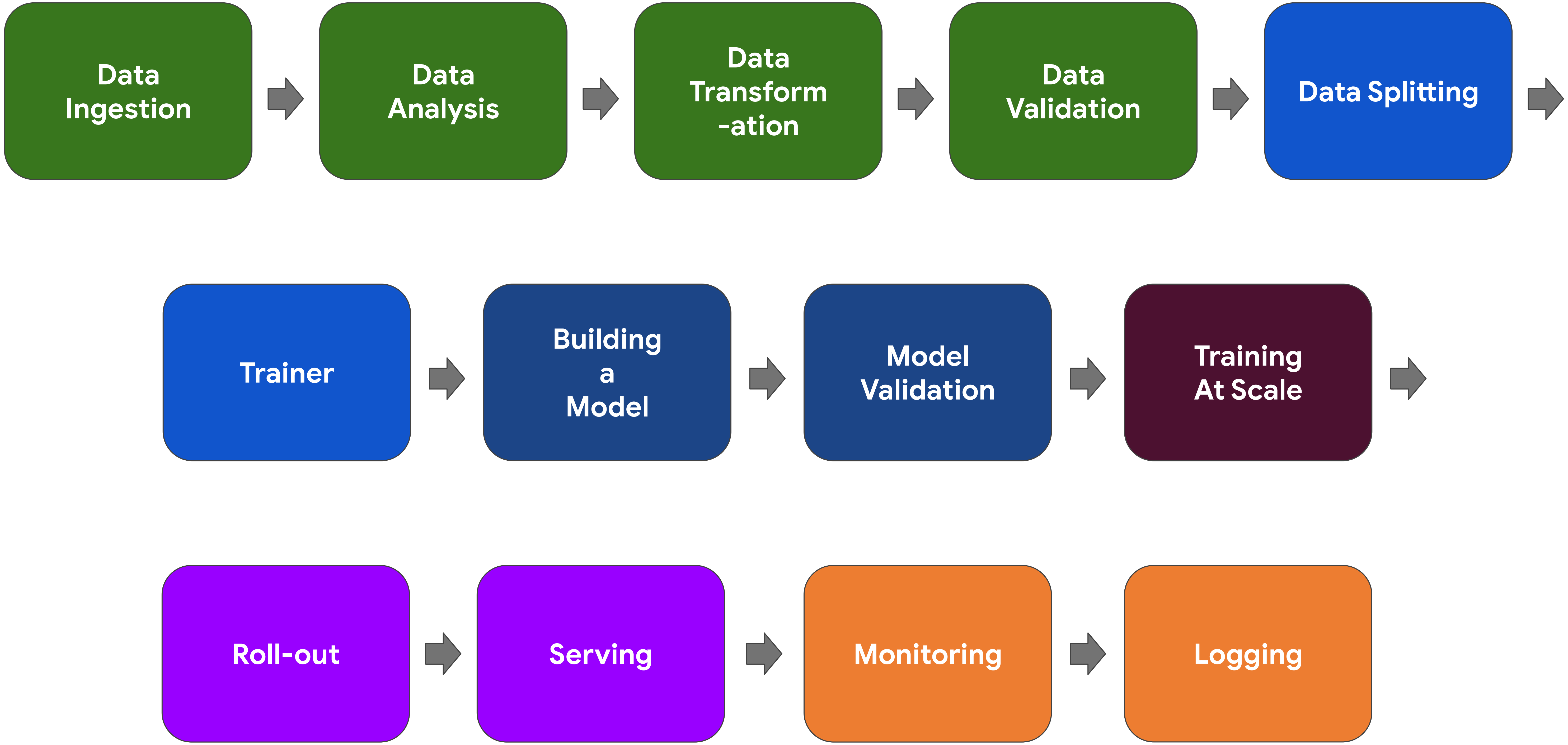
Monitoring

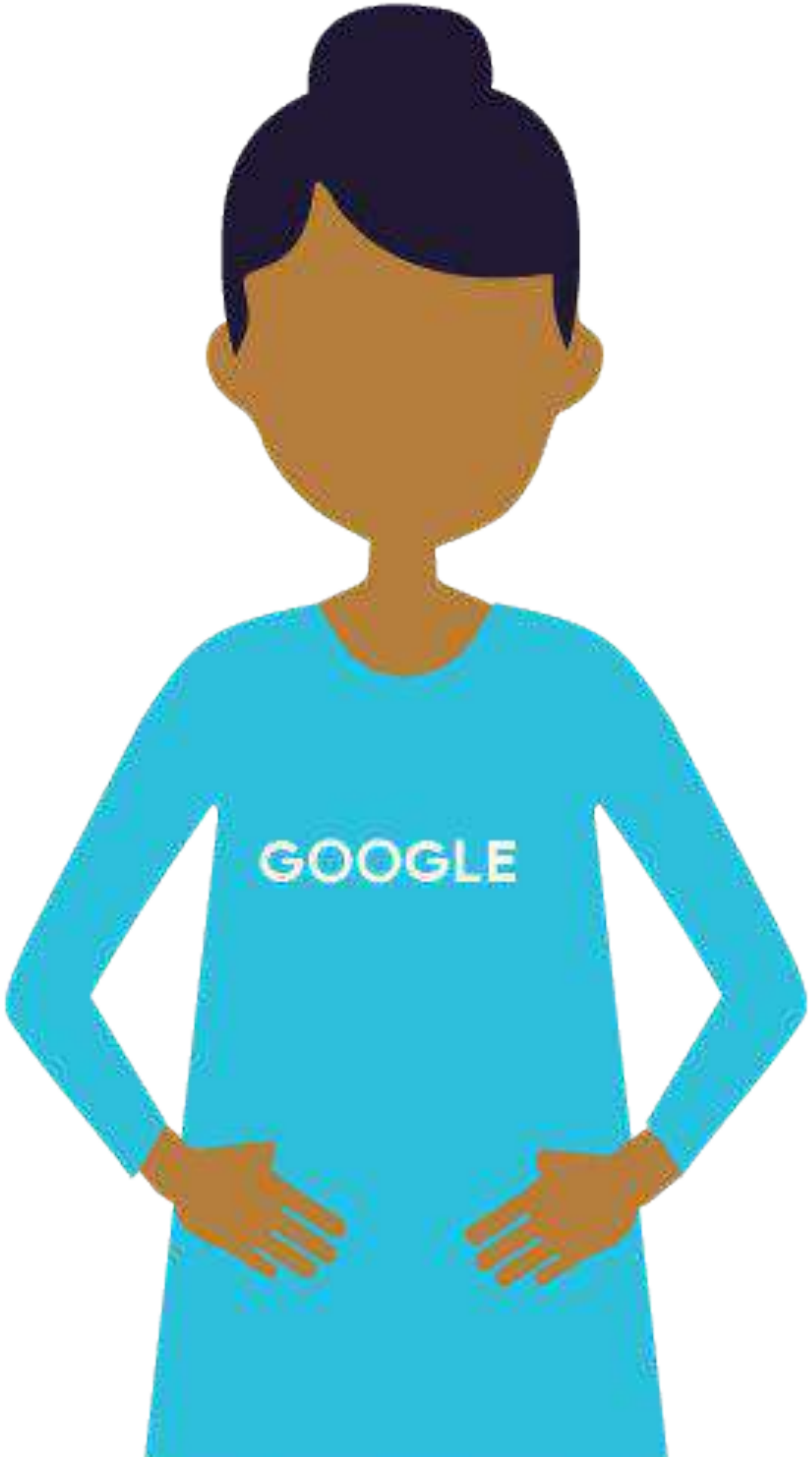
Logging

Each ML Stage is an Independent System



Composability is about microservices

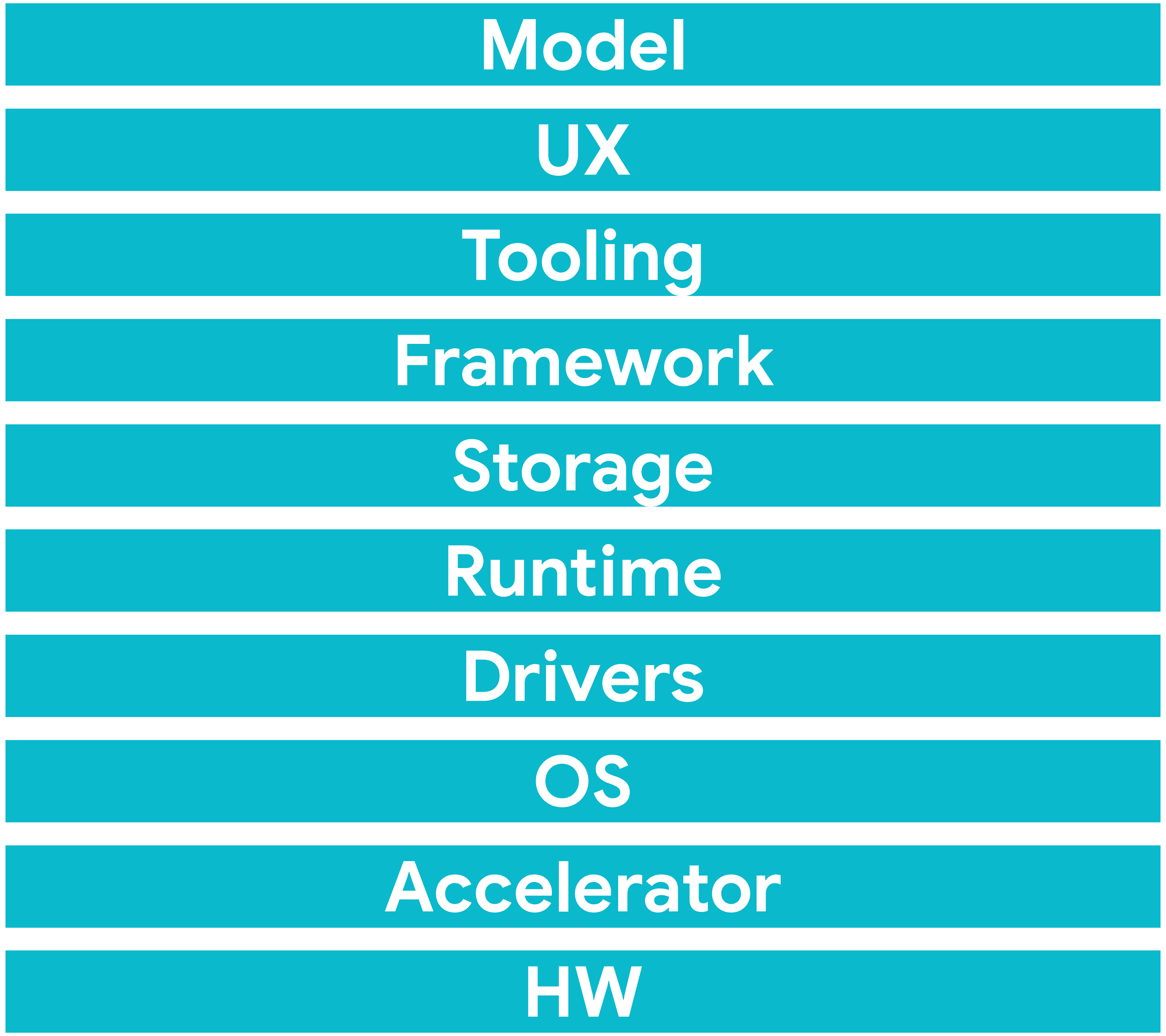




Portability

Portability

Experimentation



Model

UX

Tooling

Framework

Storage

Runtime

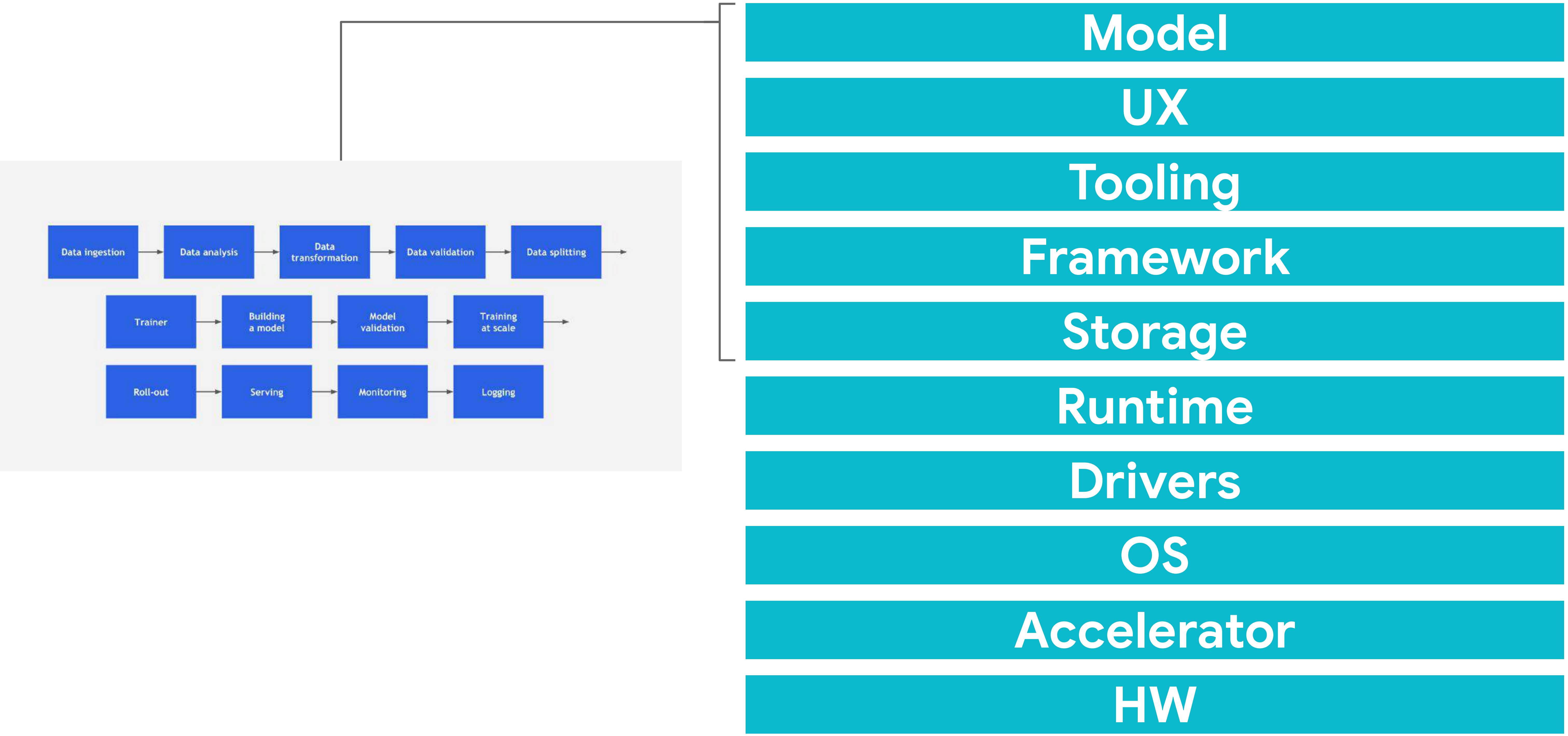
Drivers

OS

Accelerator

HW

Portability



Experimentation

Model

UX

Tooling

Framework

Storage

Runtime

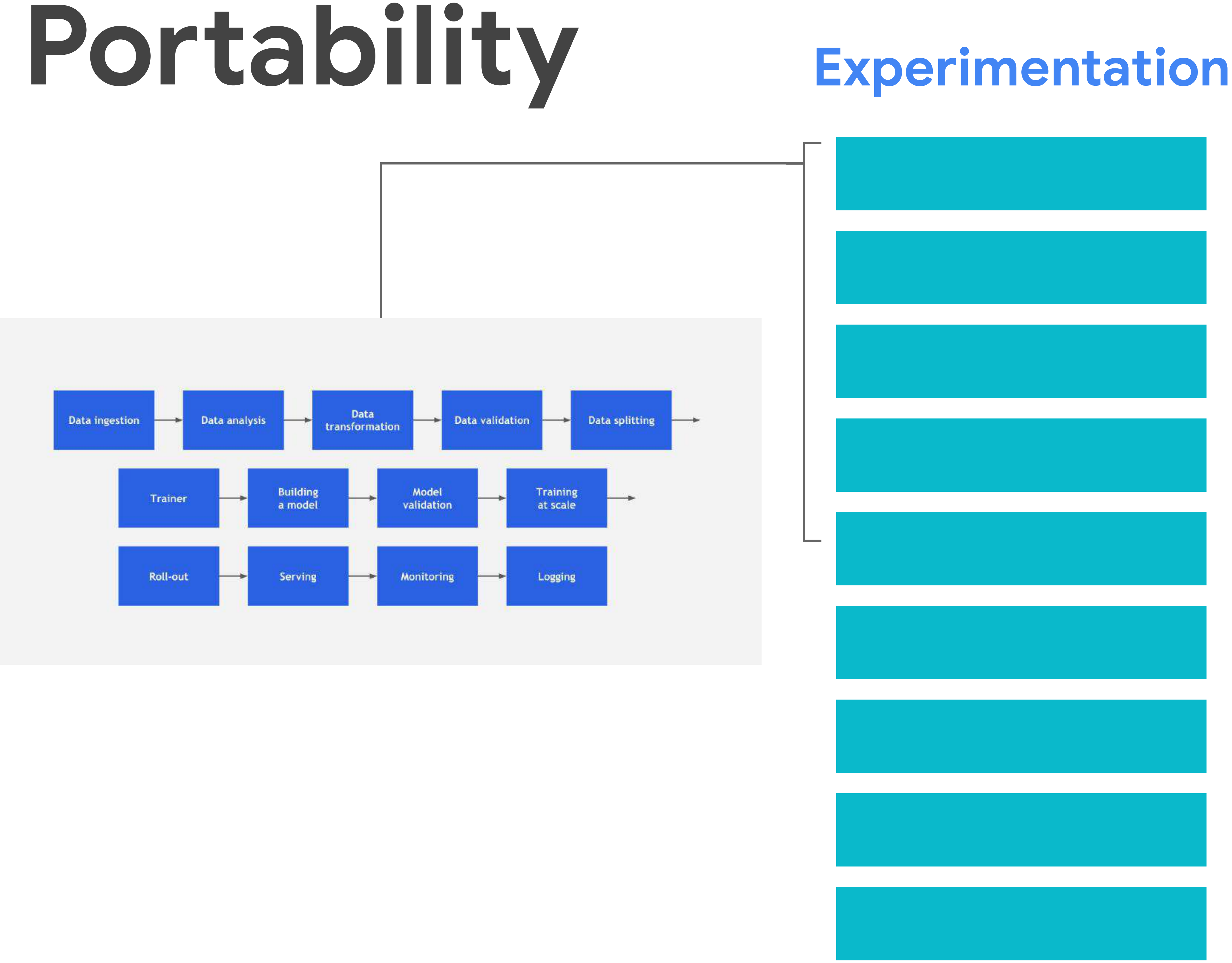
Drivers

OS

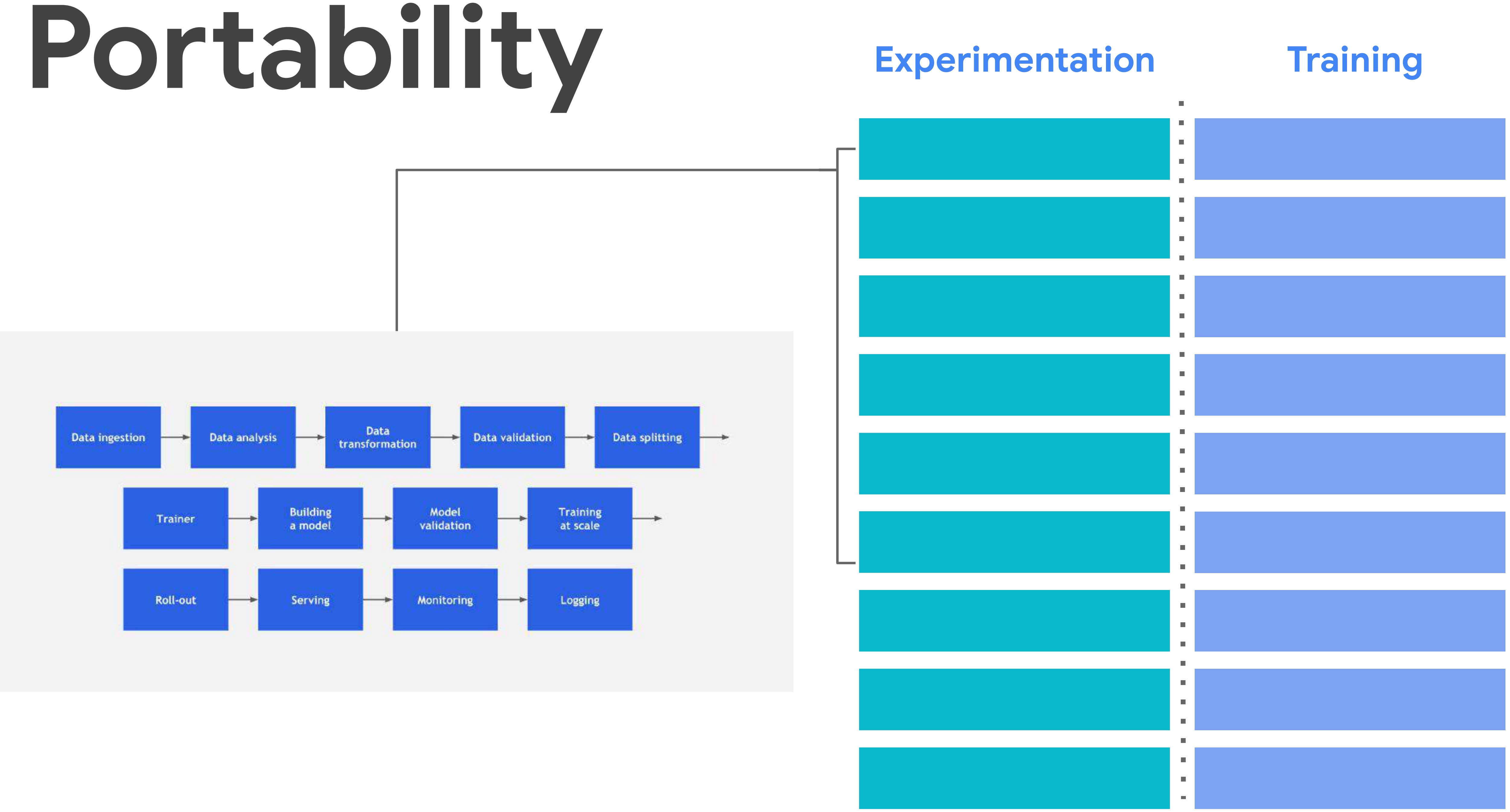
Accelerator

HW

Portability



Portability



Experimentation

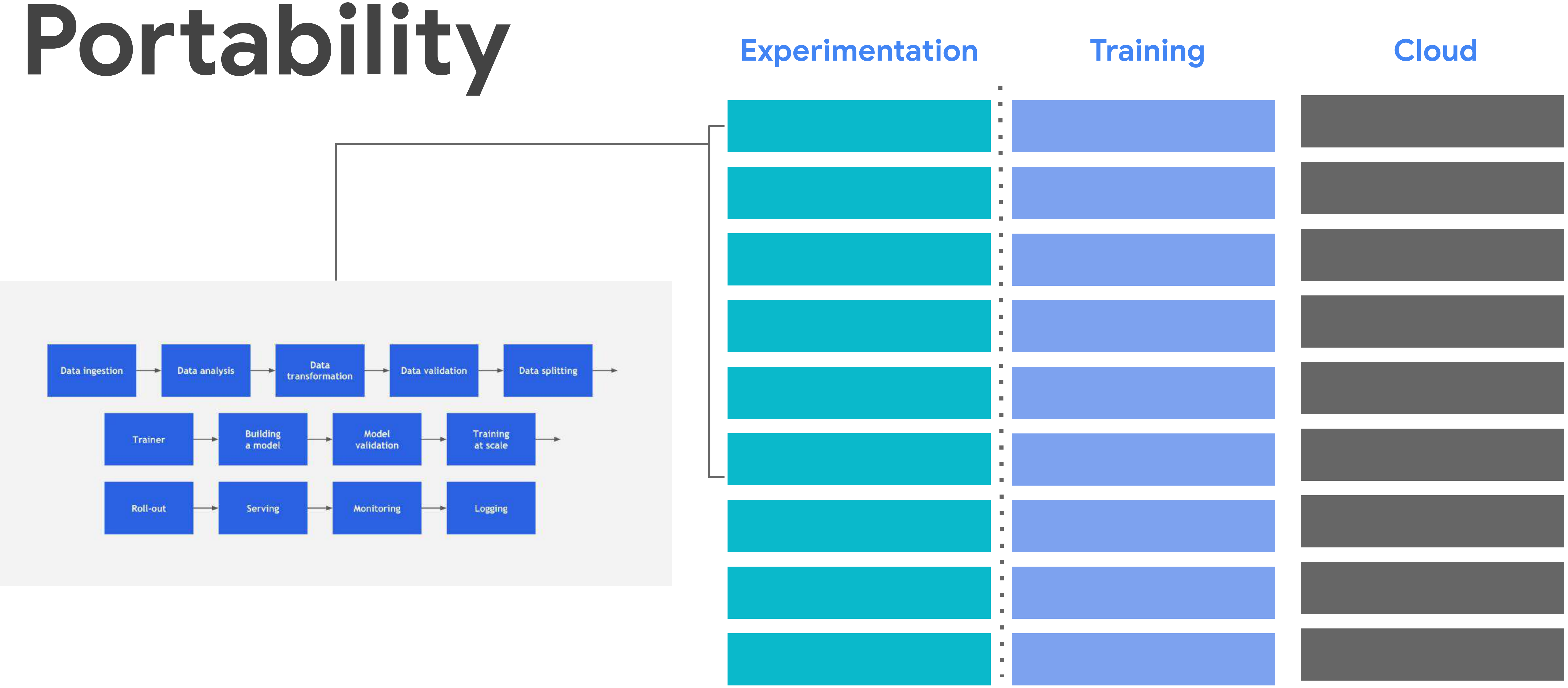
Training

Data ingestion → Data analysis → Data transformation → Data validation → Data splitting →

Trainer → Building a model → Model validation → Training at scale →

Roll-out → Serving → Monitoring → Logging

Portability



Experimentation

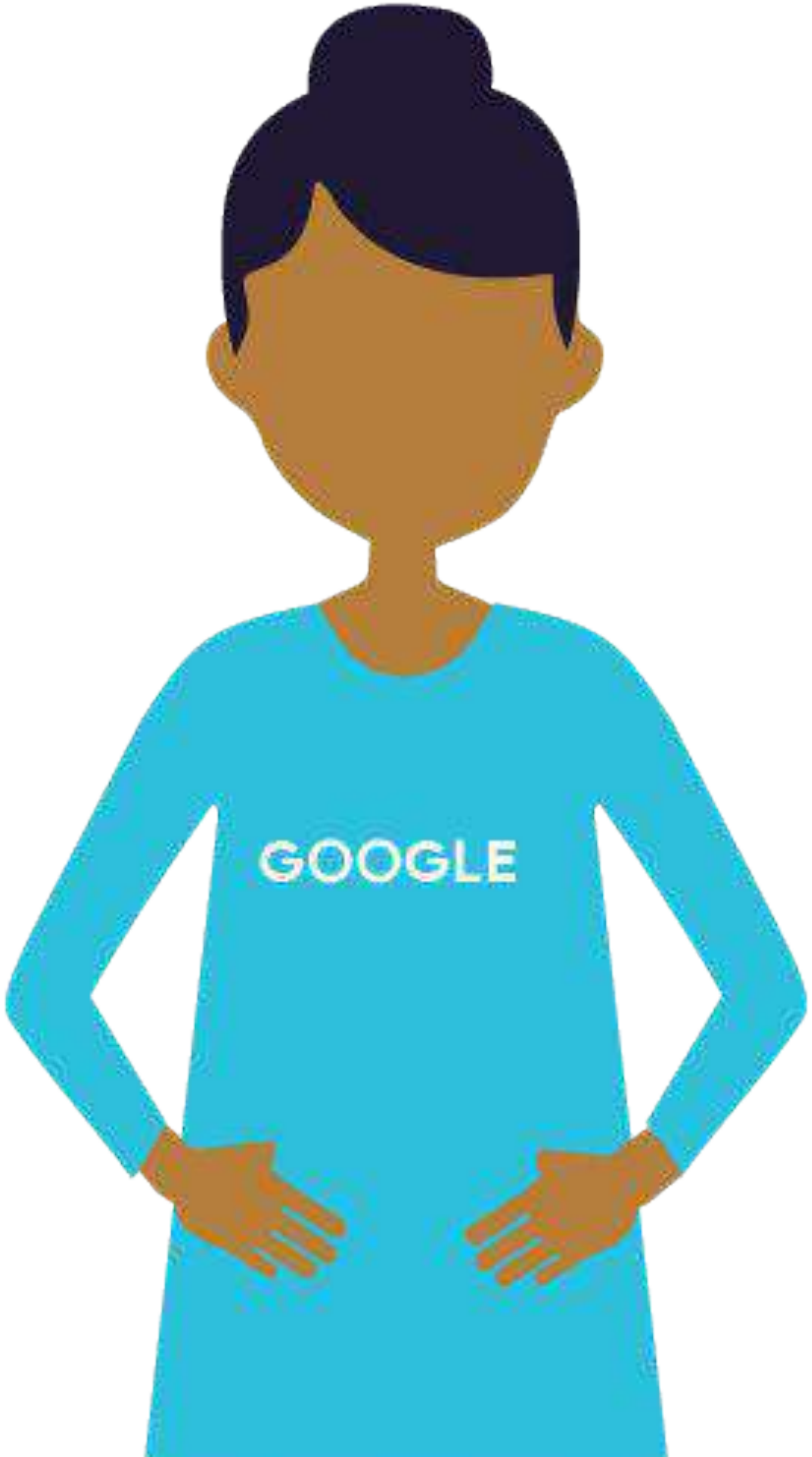
Training

Cloud

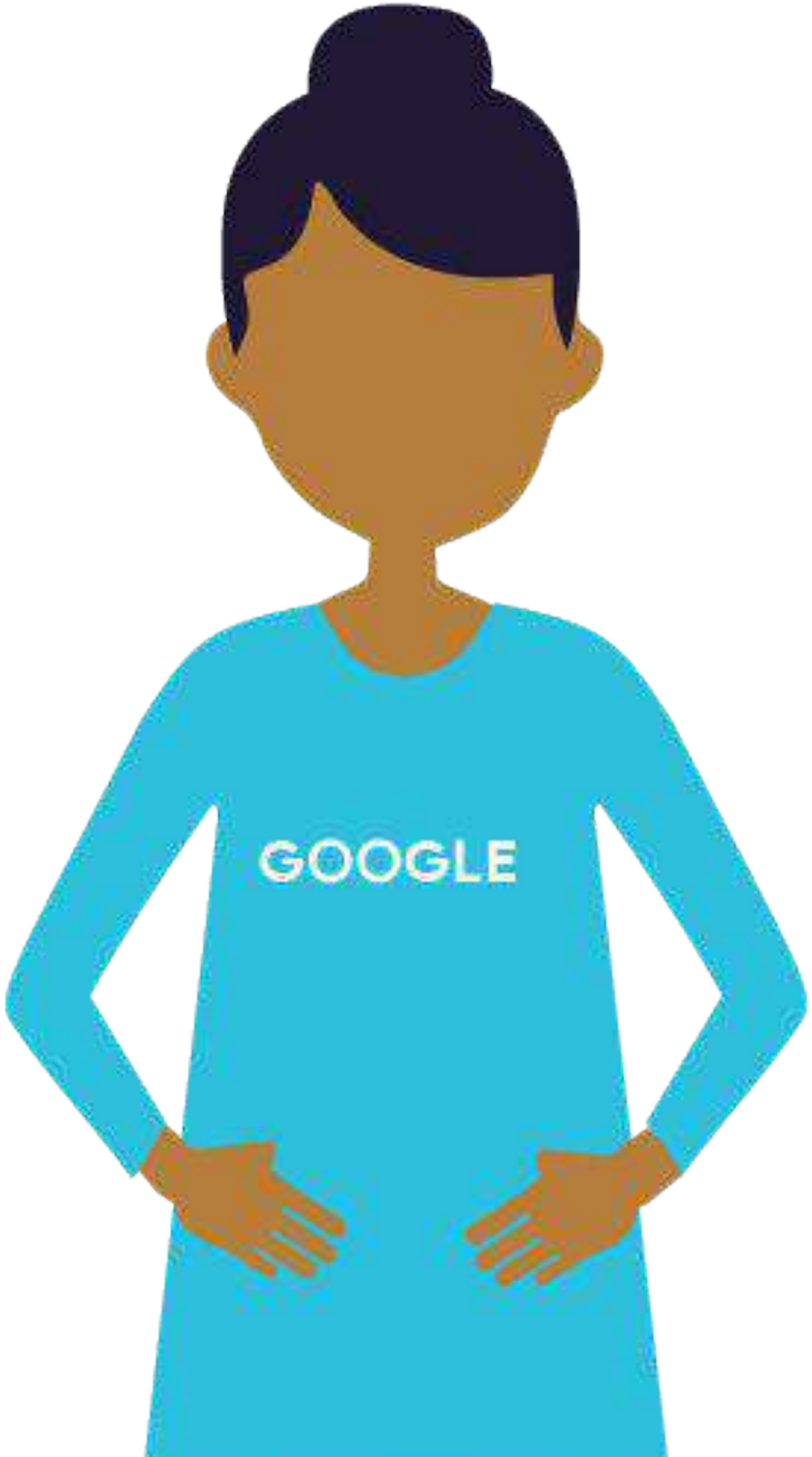
Data ingestion → Data analysis → Data transformation → Data validation → Data splitting →

Trainer → Building a model → Model validation → Training at scale →

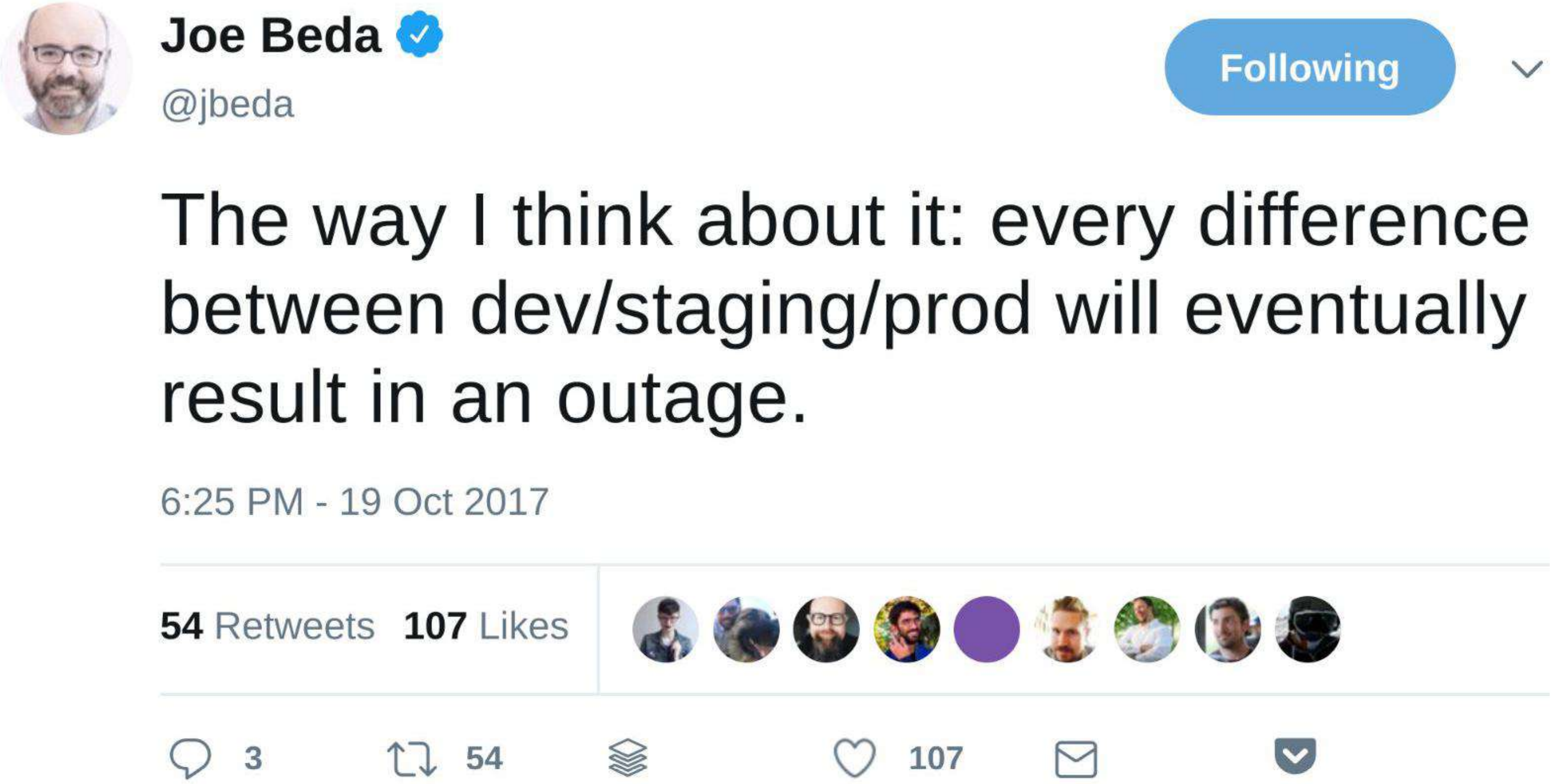
Roll-out → Serving → Monitoring → Logging



“Portability doesn’t matter to me”



Wrong!



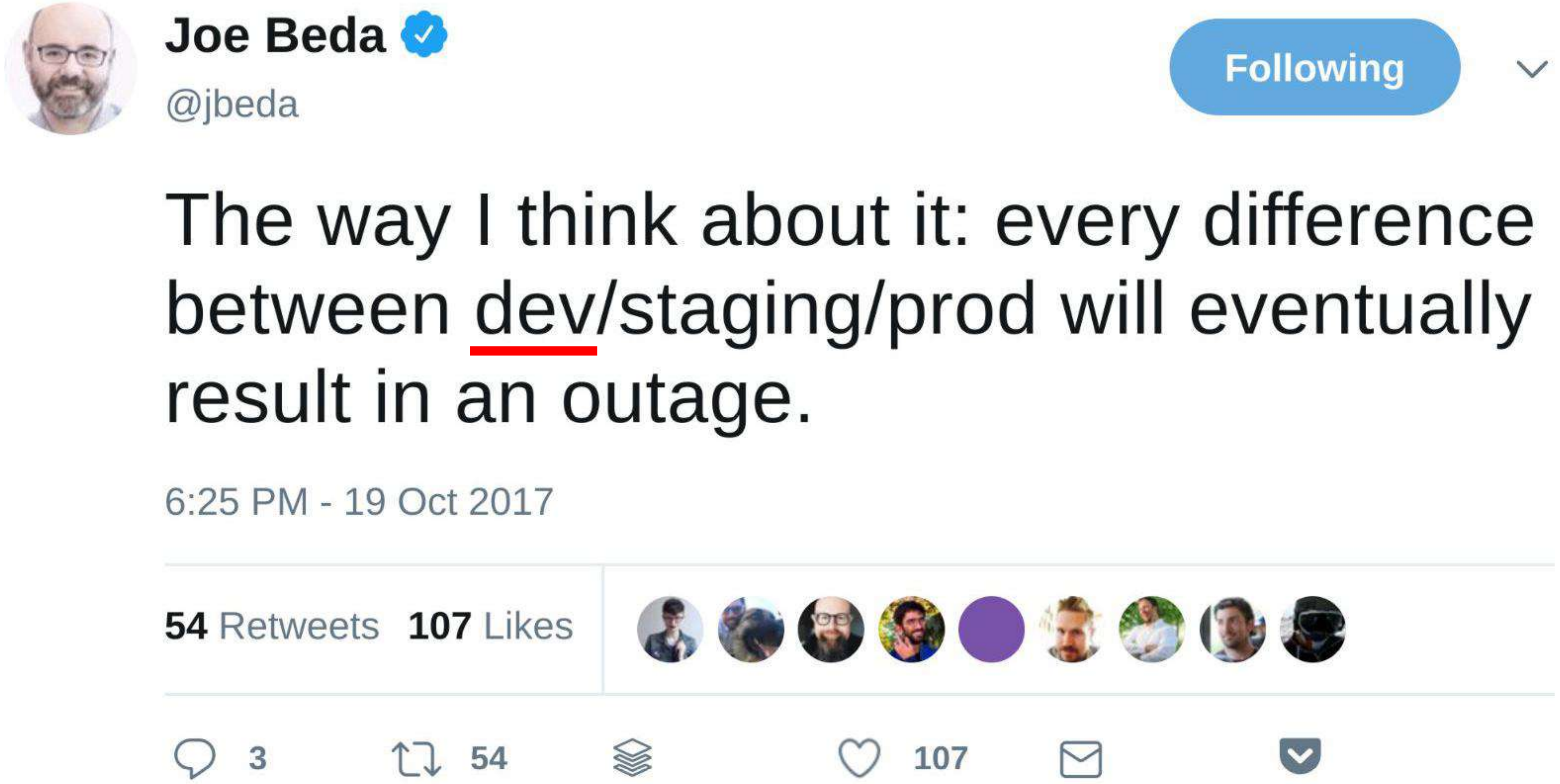
Joe Beda

@jbeda

The way I think about it: every difference between dev/staging/prod will eventually result in an outage.

6:25 PM - 19 Oct 2017

54 Retweets **107** Likes



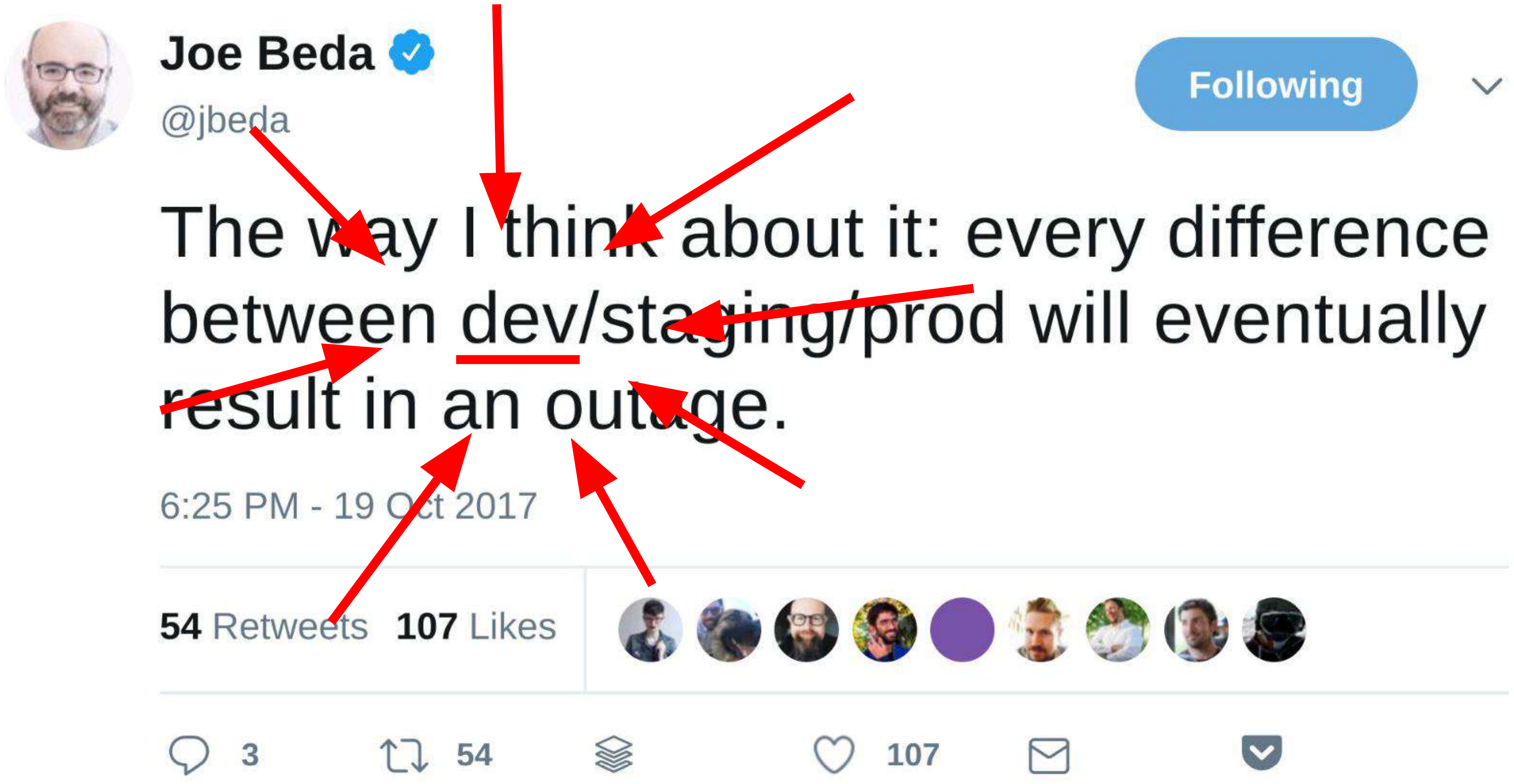
Joe Beda

@jbeda

The way I think about it: every difference between dev/staging/prod will eventually result in an outage.

6:25 PM - 19 Oct 2017

54 Retweets **107** Likes



Joe Beda

@jbeda

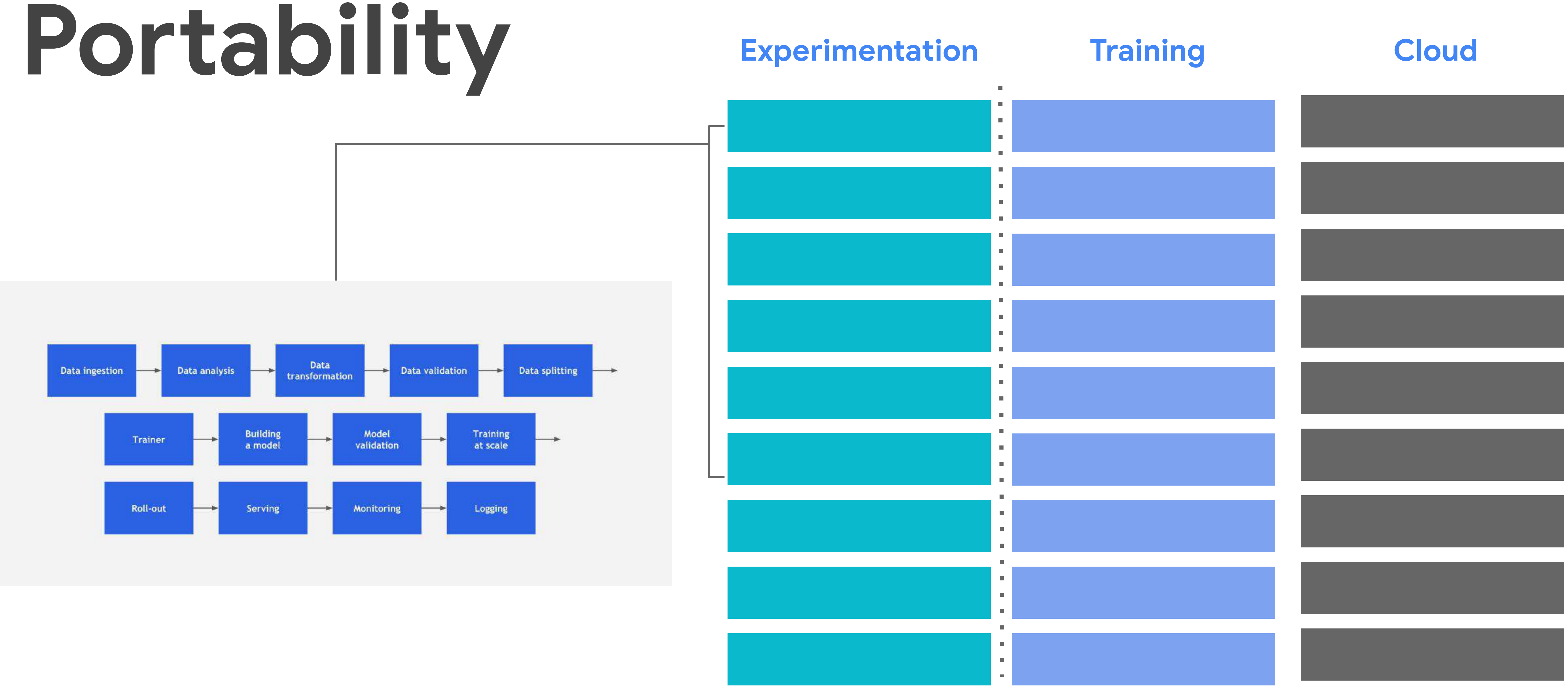
The way I think about it: every difference between dev/staging/prod will eventually result in an outage.

6:25 PM - 19 Oct 2017

54 Retweets **107** Likes

3 54 107

Portability



Data ingestion → Data analysis → Data transformation → Data validation → Data splitting →

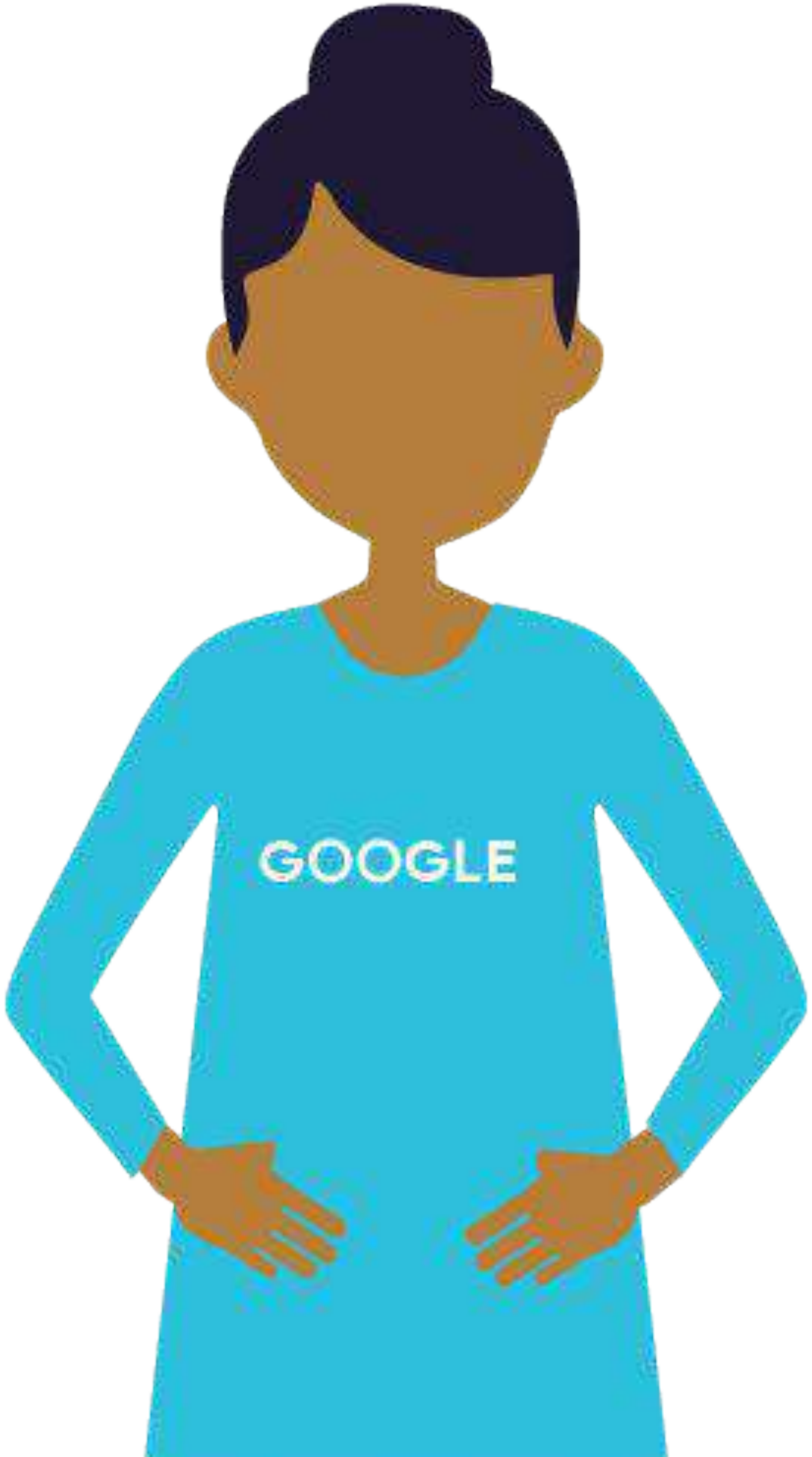
Trainer → Building a model → Model validation → Training at scale →

Roll-out → Serving → Monitoring → Logging

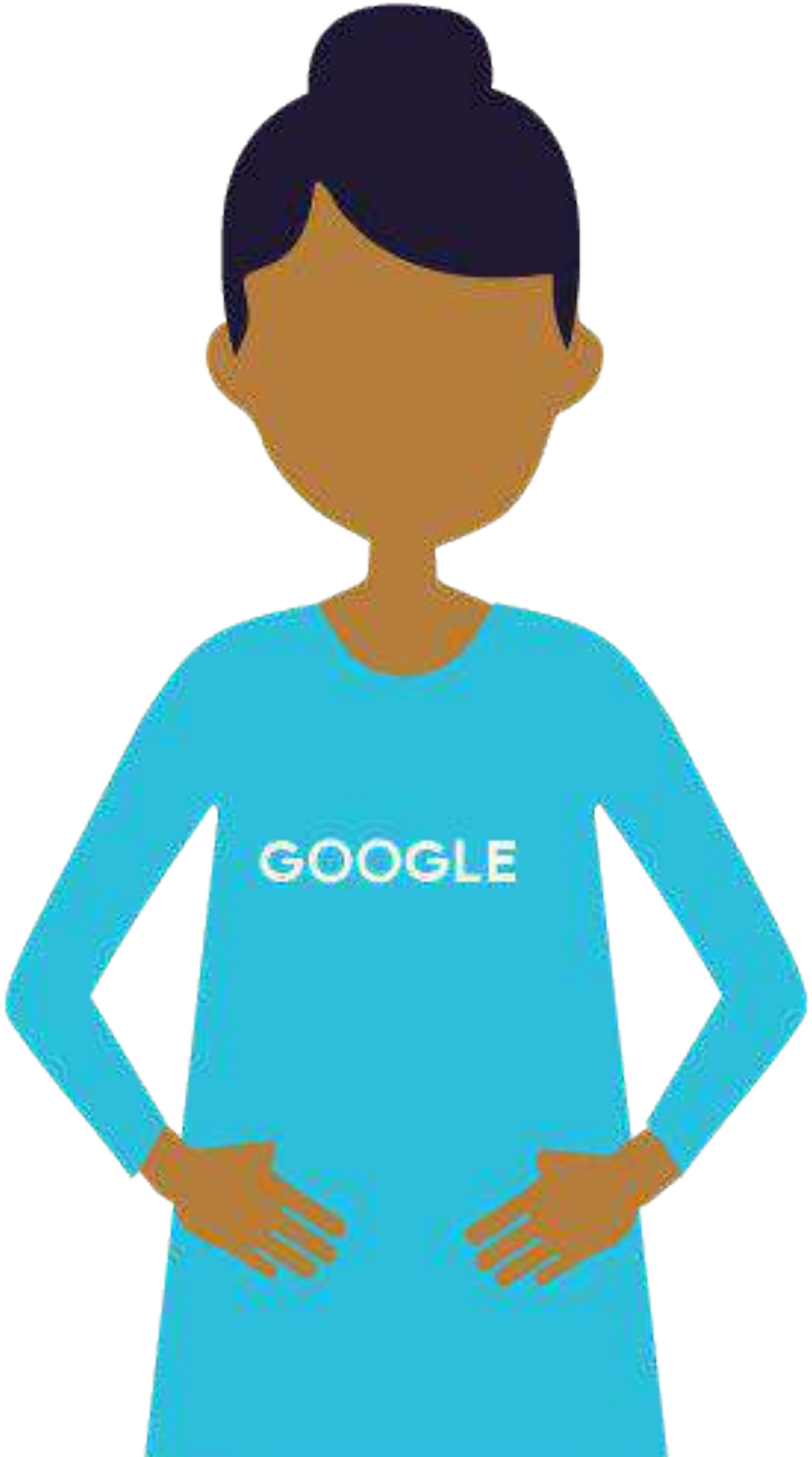
Experimentation

Training

Cloud

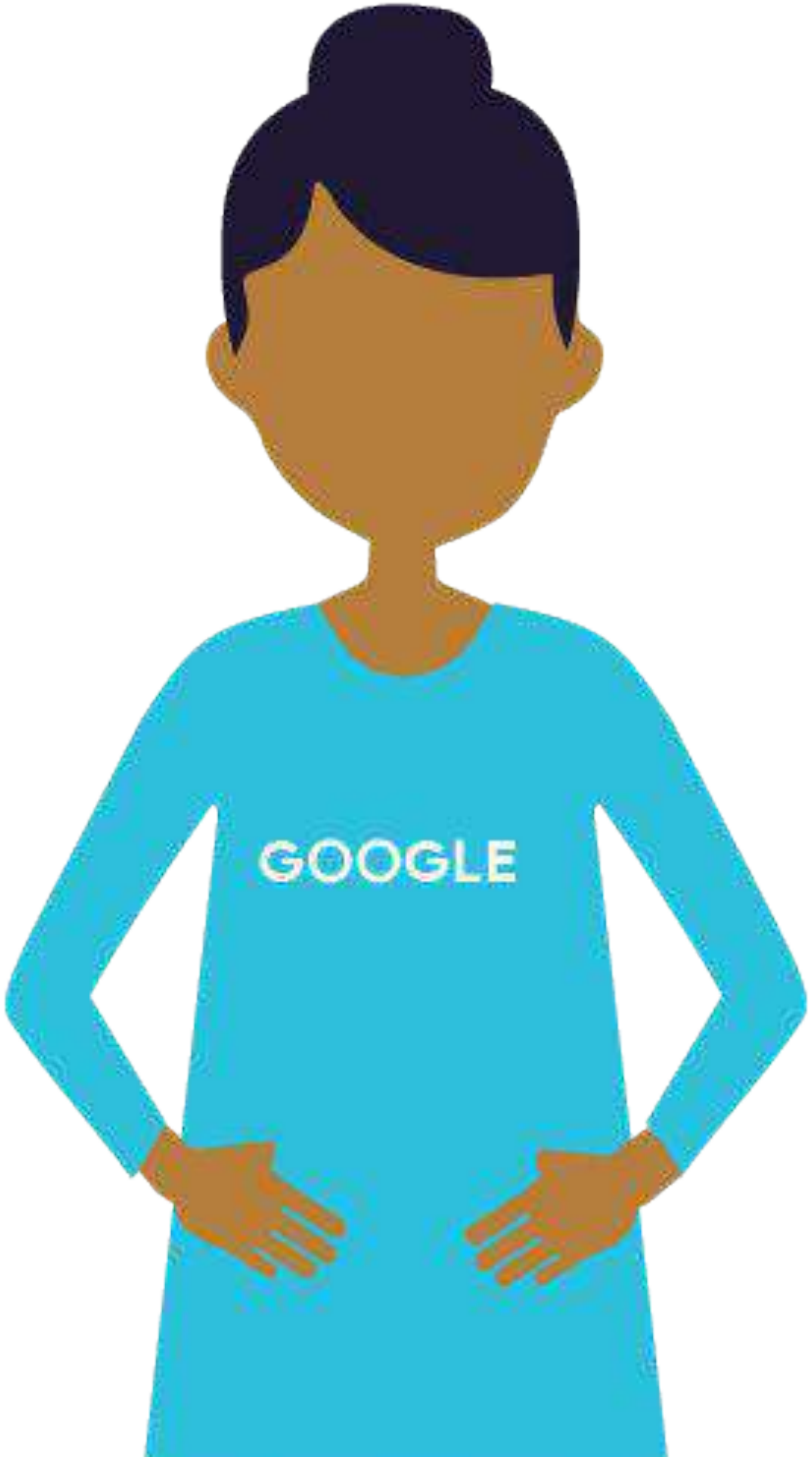


Your Laptop Counts.



Scalability

- **More** accelerators (GPU, TPU)
- **More** CPUs
- **More** disk/networking
- **More** skillsets (data engineers, data scientists)
- **More** teams
- **More experiments**



- Composability
- Portability
- Scalability

Courses 7 - Production ML Systems

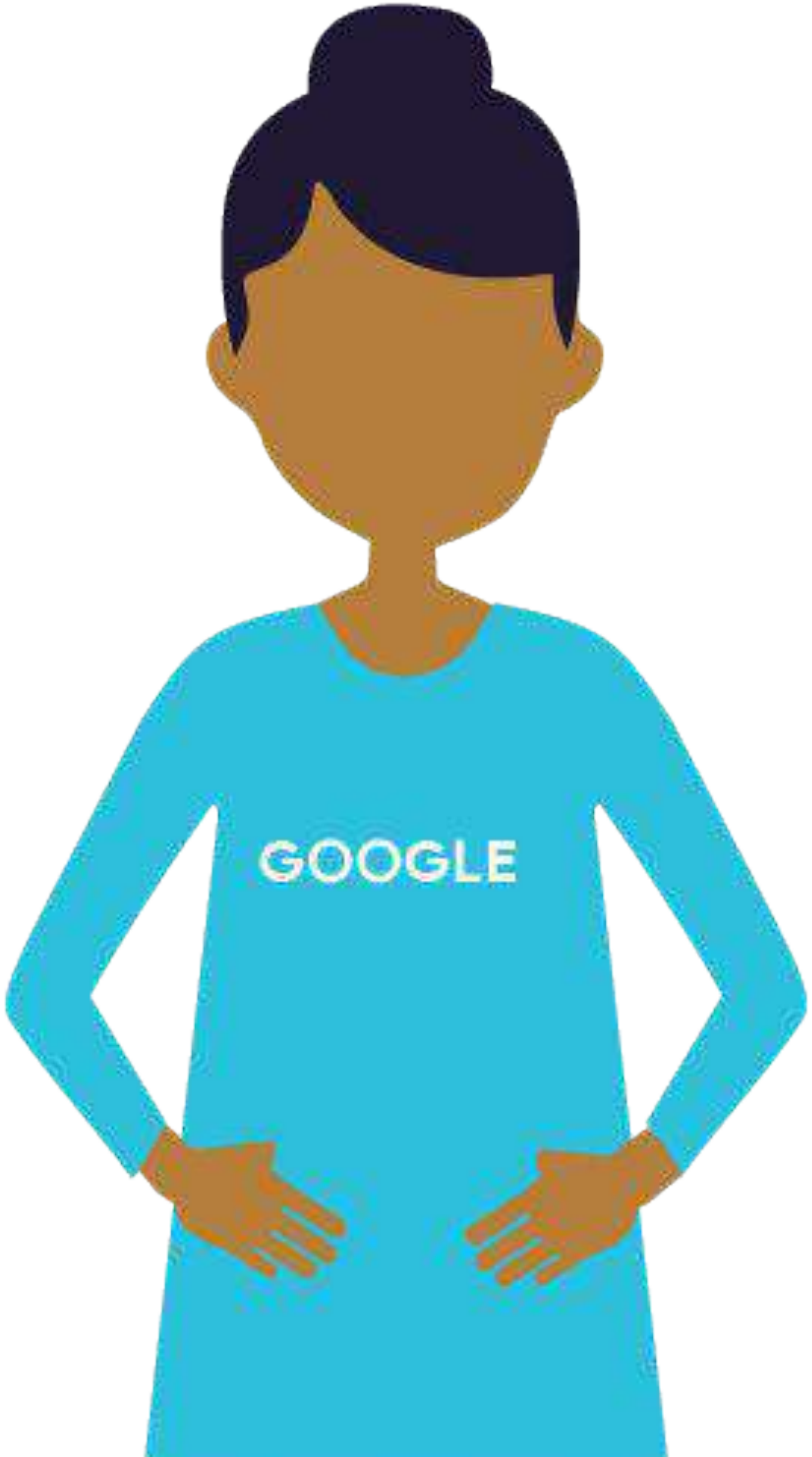
Module 5: Hybrid ML Systems

Lesson Title: **Kubeflow**

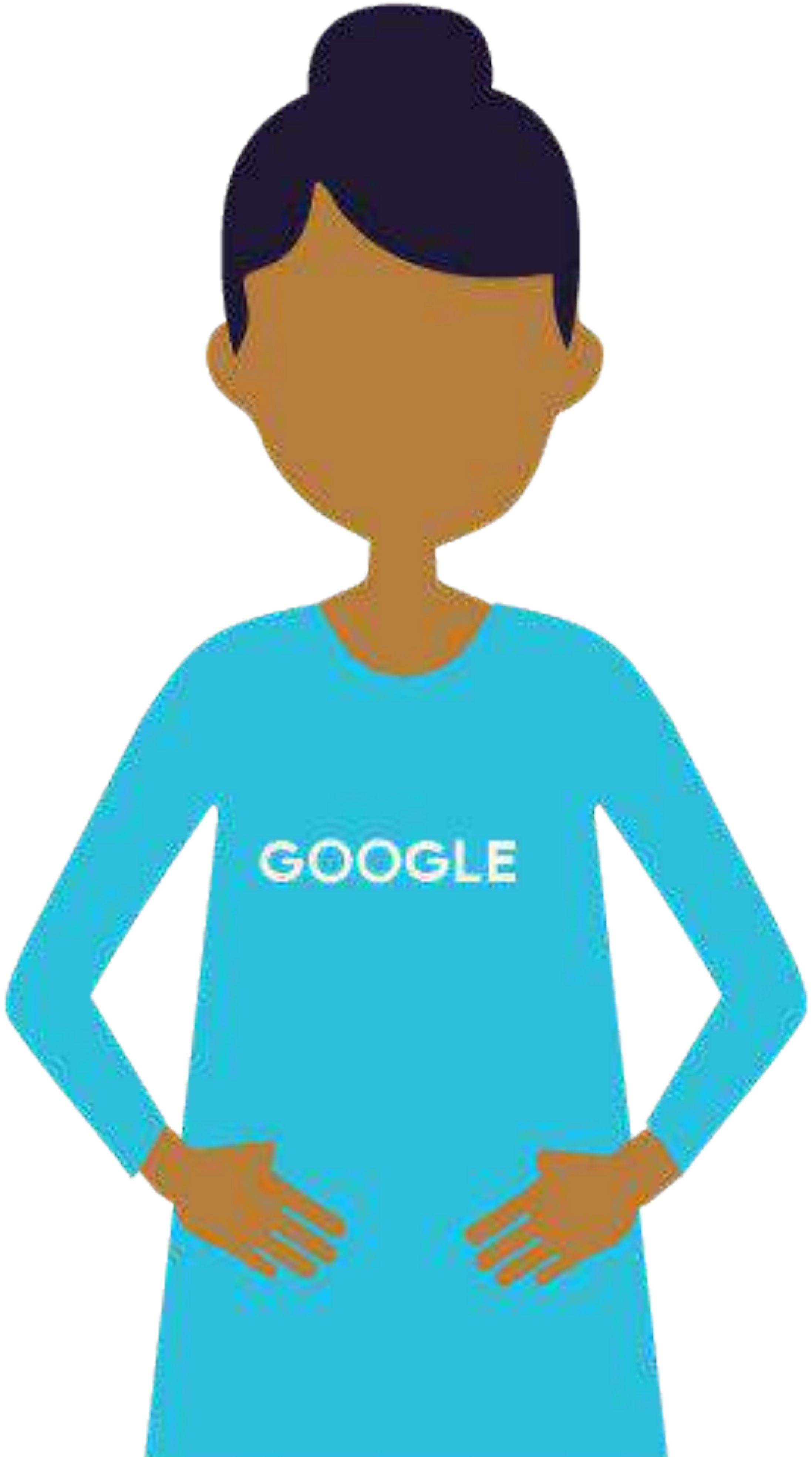
Format: Presenter

Presenter: Val

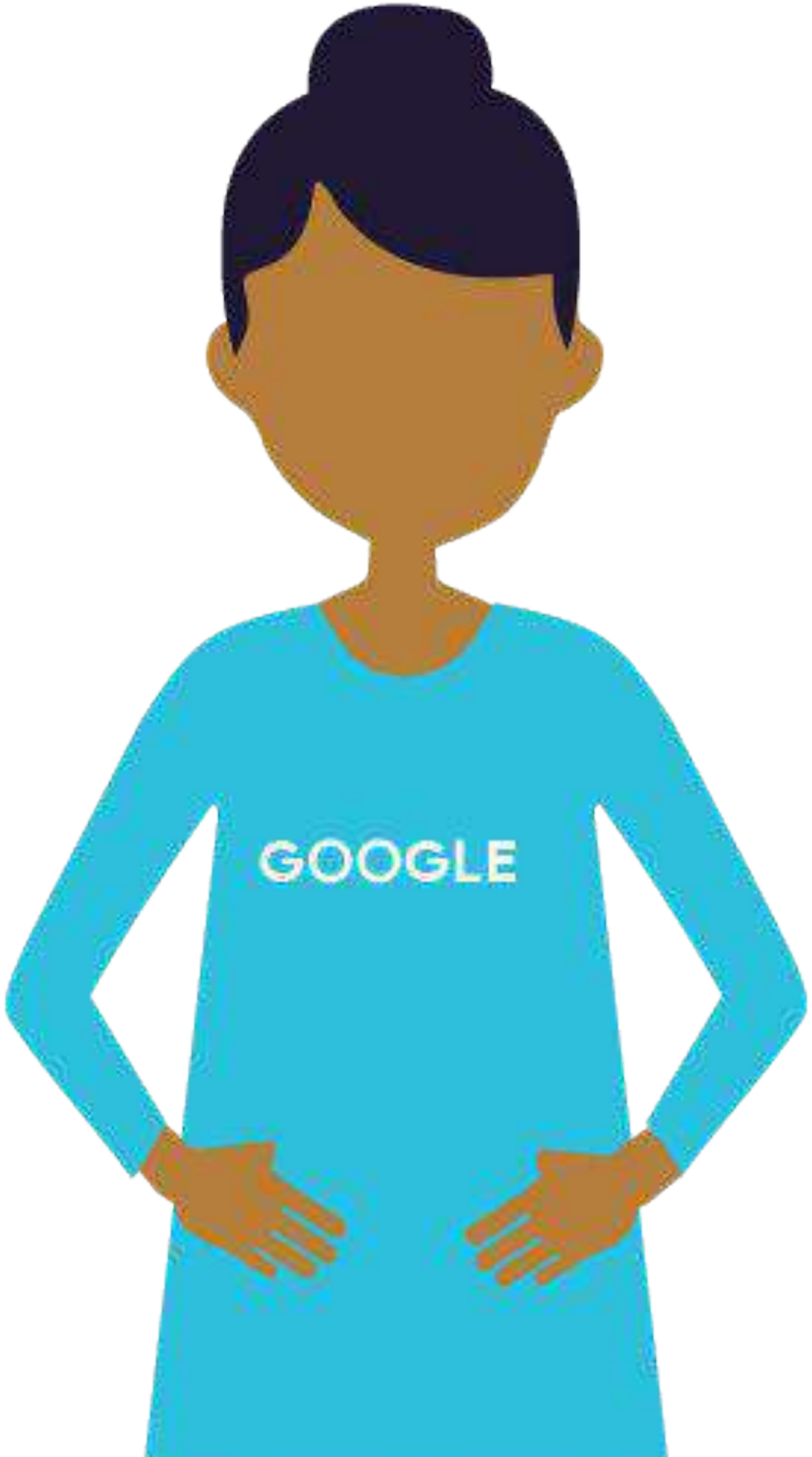
Video Name: T-PSML-O_5_l3_kubeflow



You know what’s really good at composability, portability, and scalability?

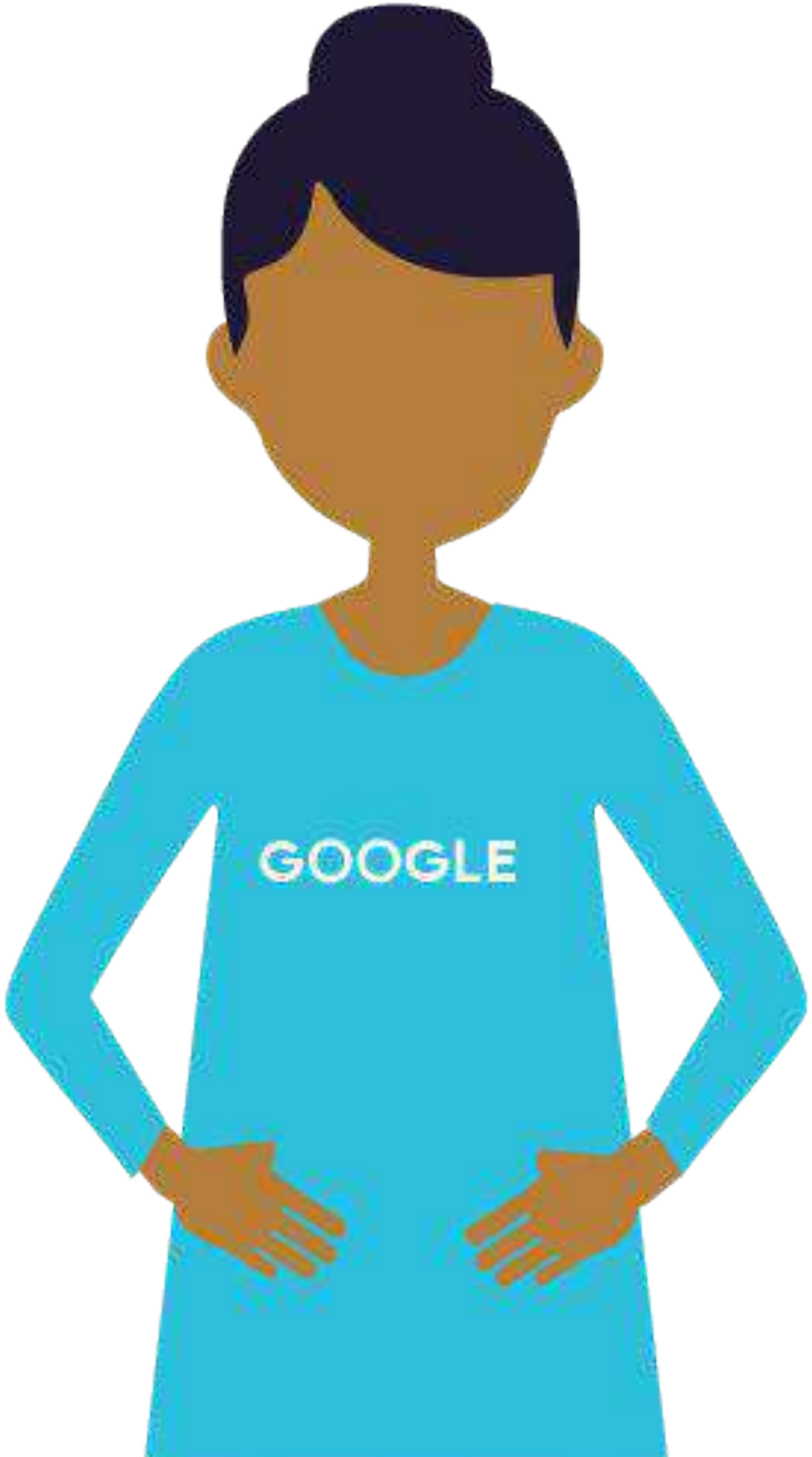


Containers & Kubernetes



Containers & Kubernetes

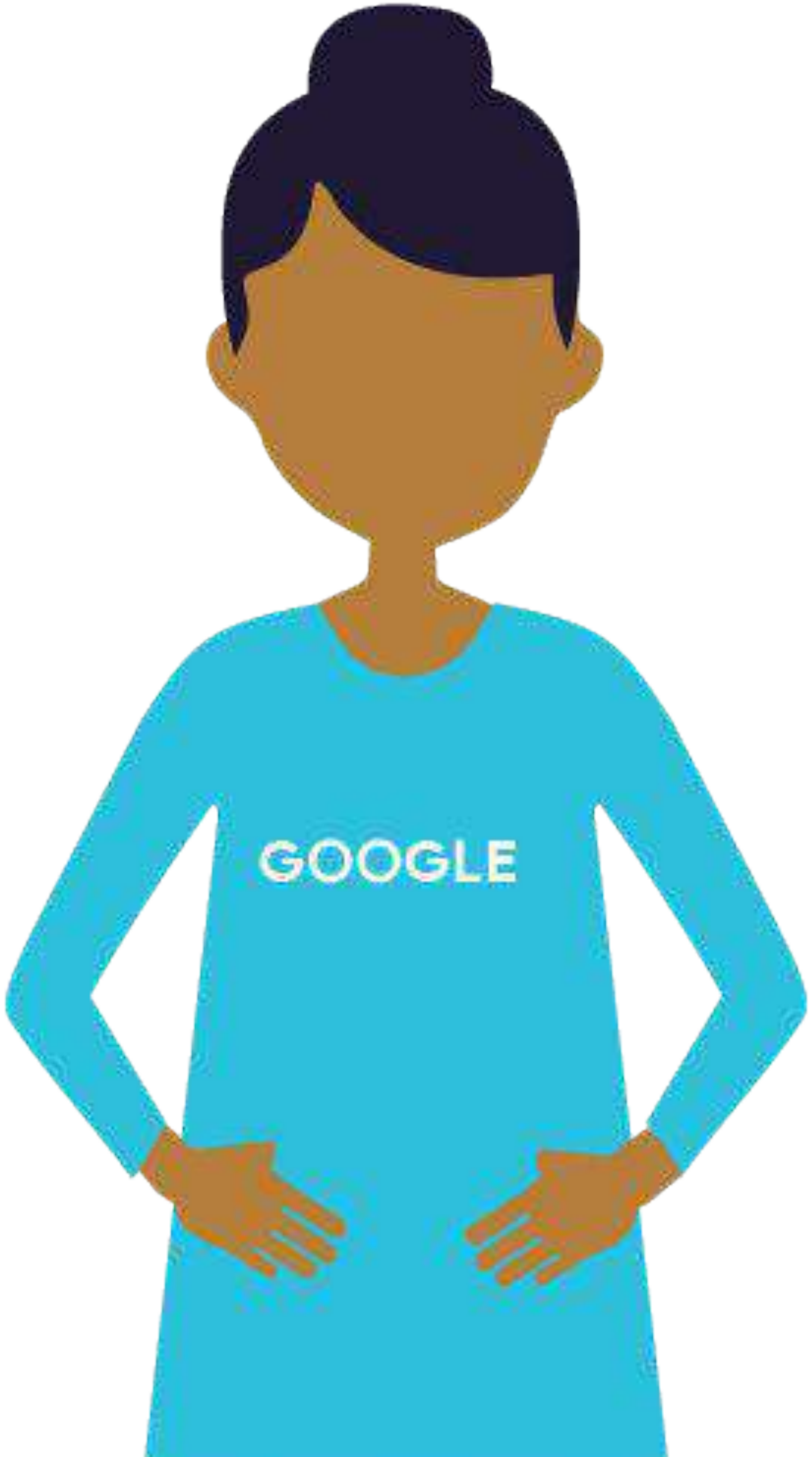
except



Oh, you want to use ML on K8s?

First become an expert in:

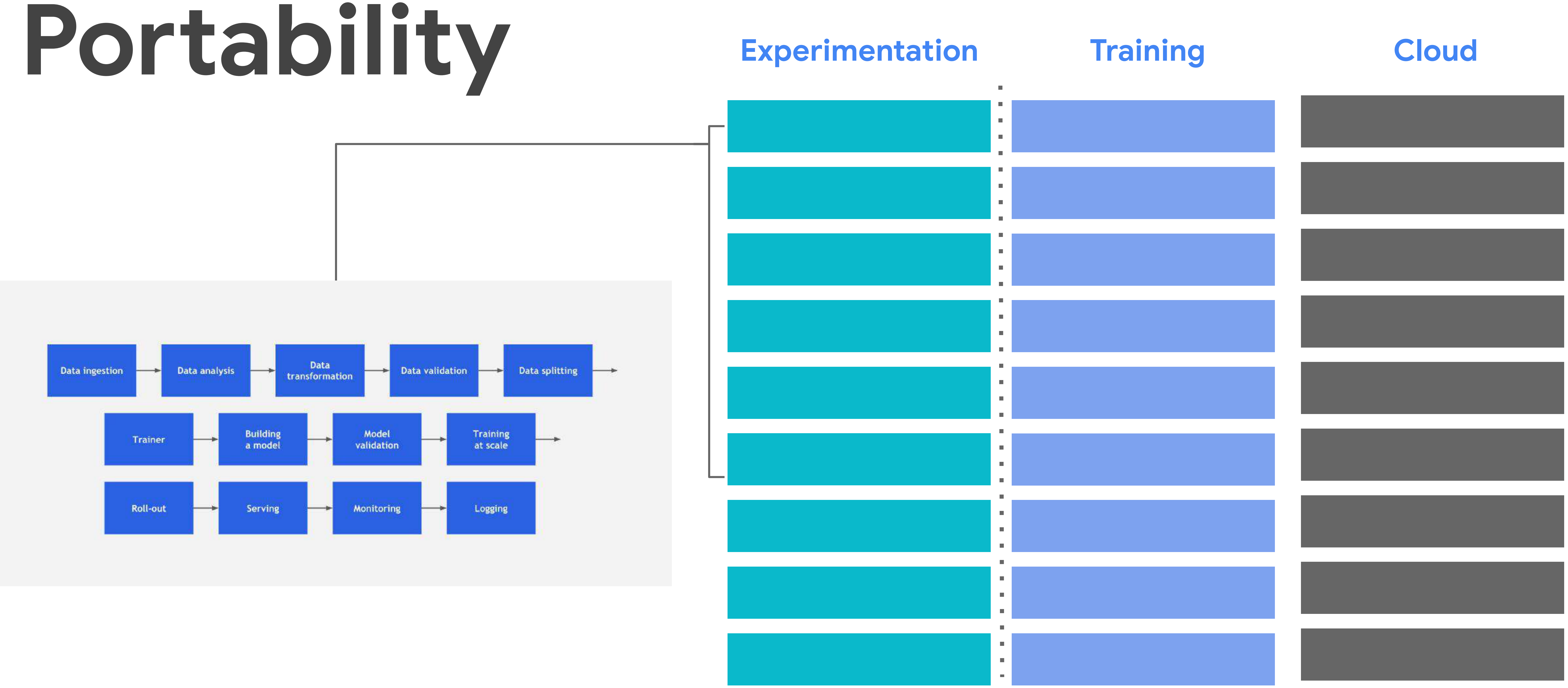
- Containers
- Packaging
- Kubernetes service endpoints
- Persistent volumes
- Scaling
- Immutable deployments
- GPUs, Drivers & the GPL
- Cloud APIs
- DevOps



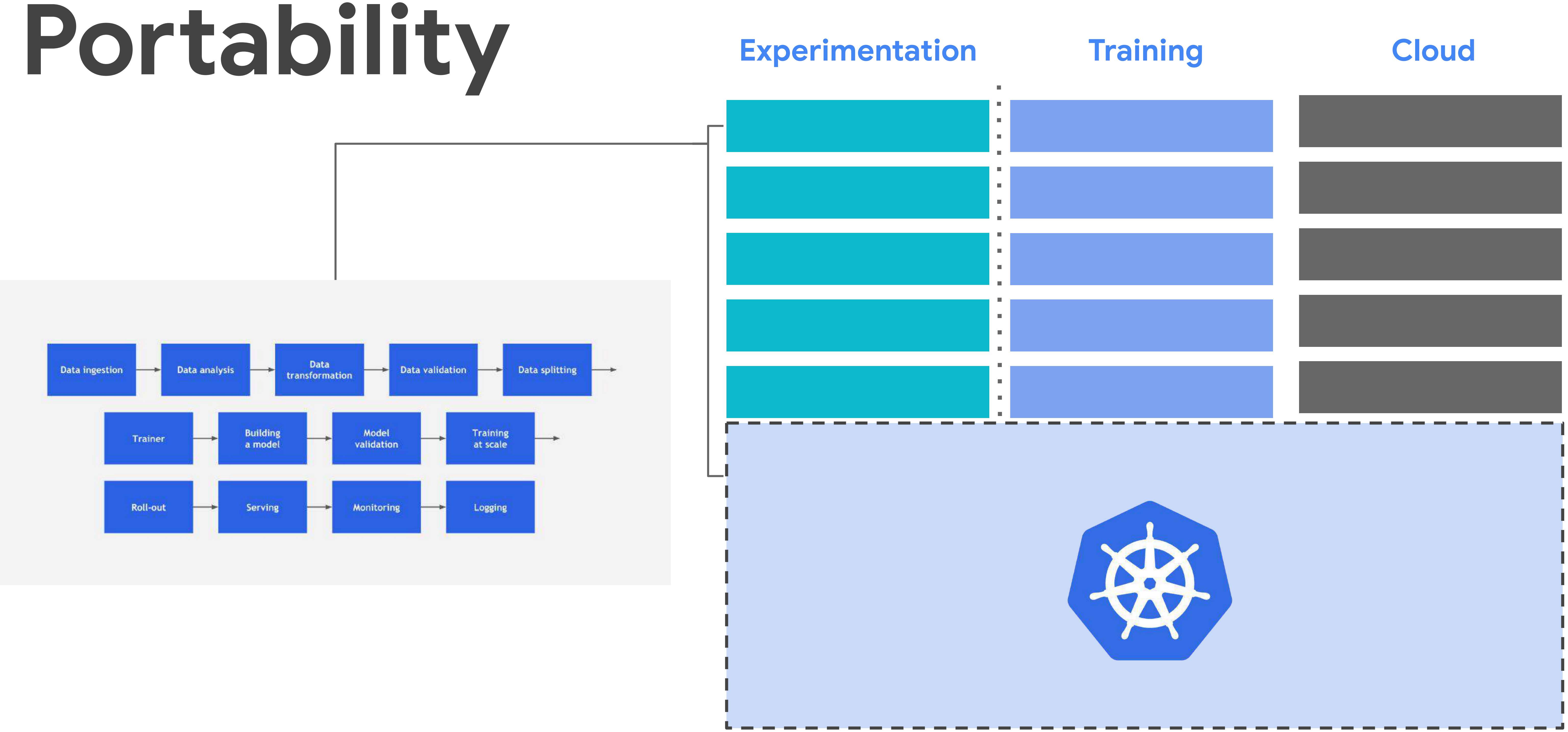
Oh, you want to use ML on K8s?

Make it Easy for Everyone
to Develop, Deploy and Manage
Portable, Distributed ML
on Kubernetes

Portability



Portability



Experimentation

Training

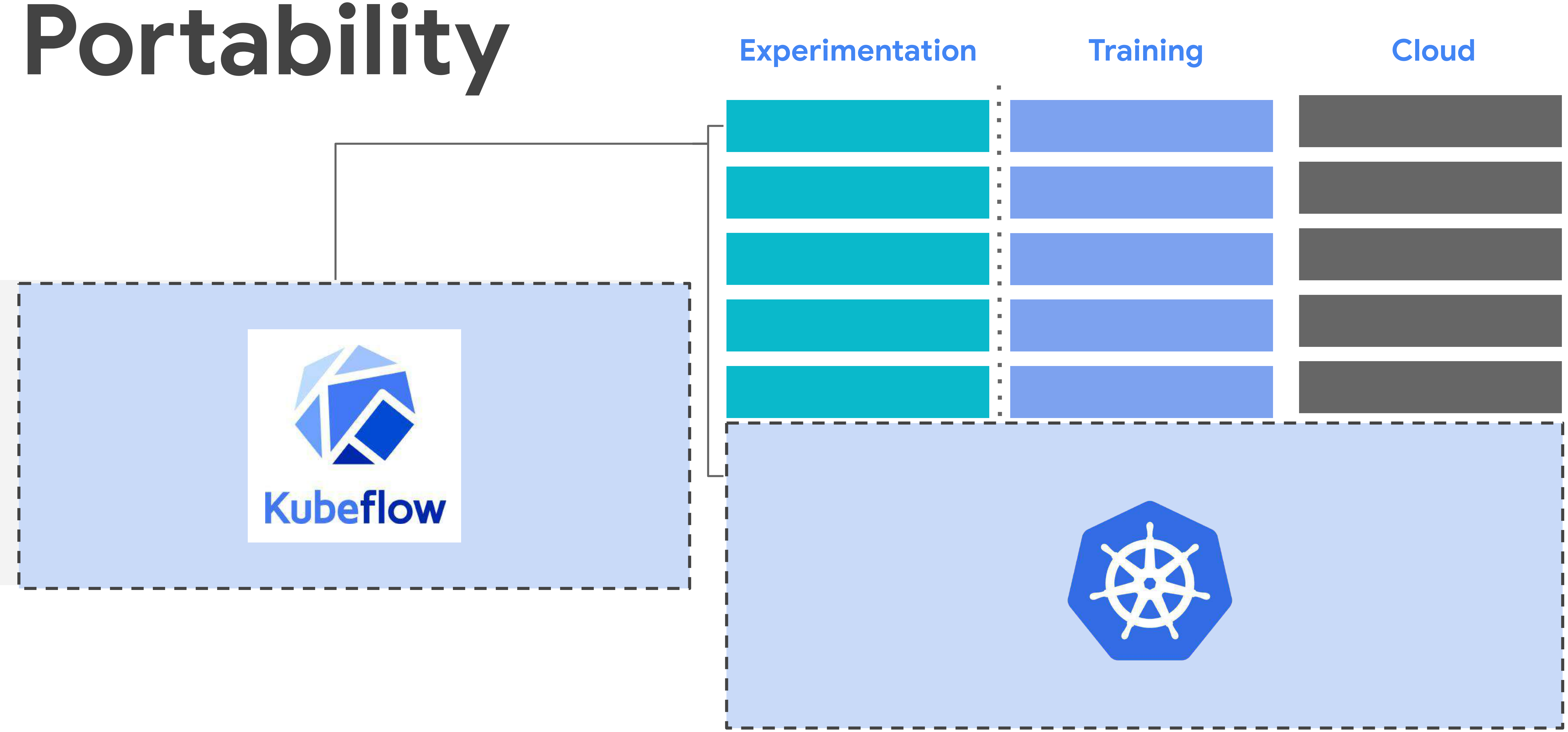
Cloud

Data ingestion → Data analysis → Data transformation → Data validation → Data splitting →

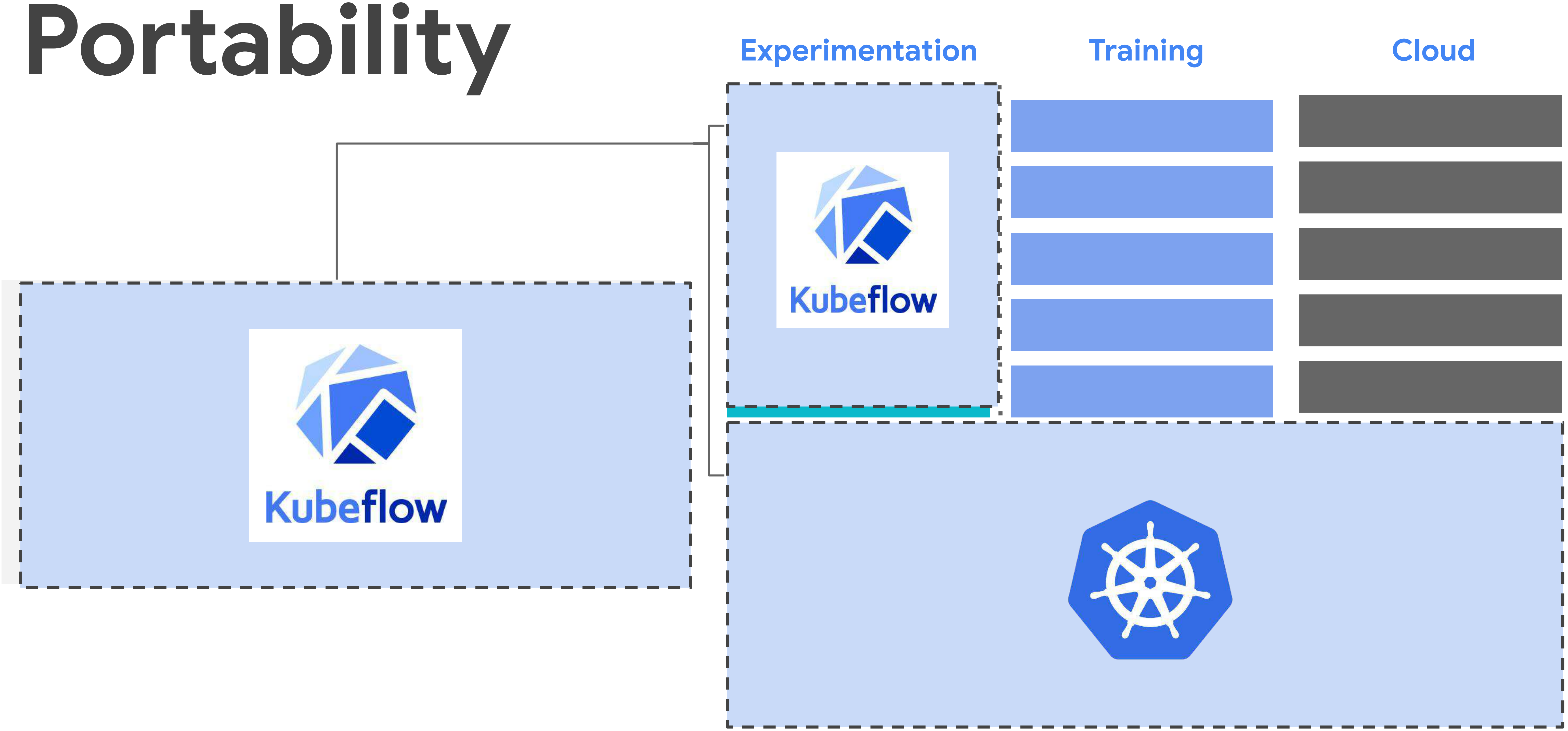
Trainer → Building a model → Model validation → Training at scale →

Roll-out → Serving → Monitoring → Logging

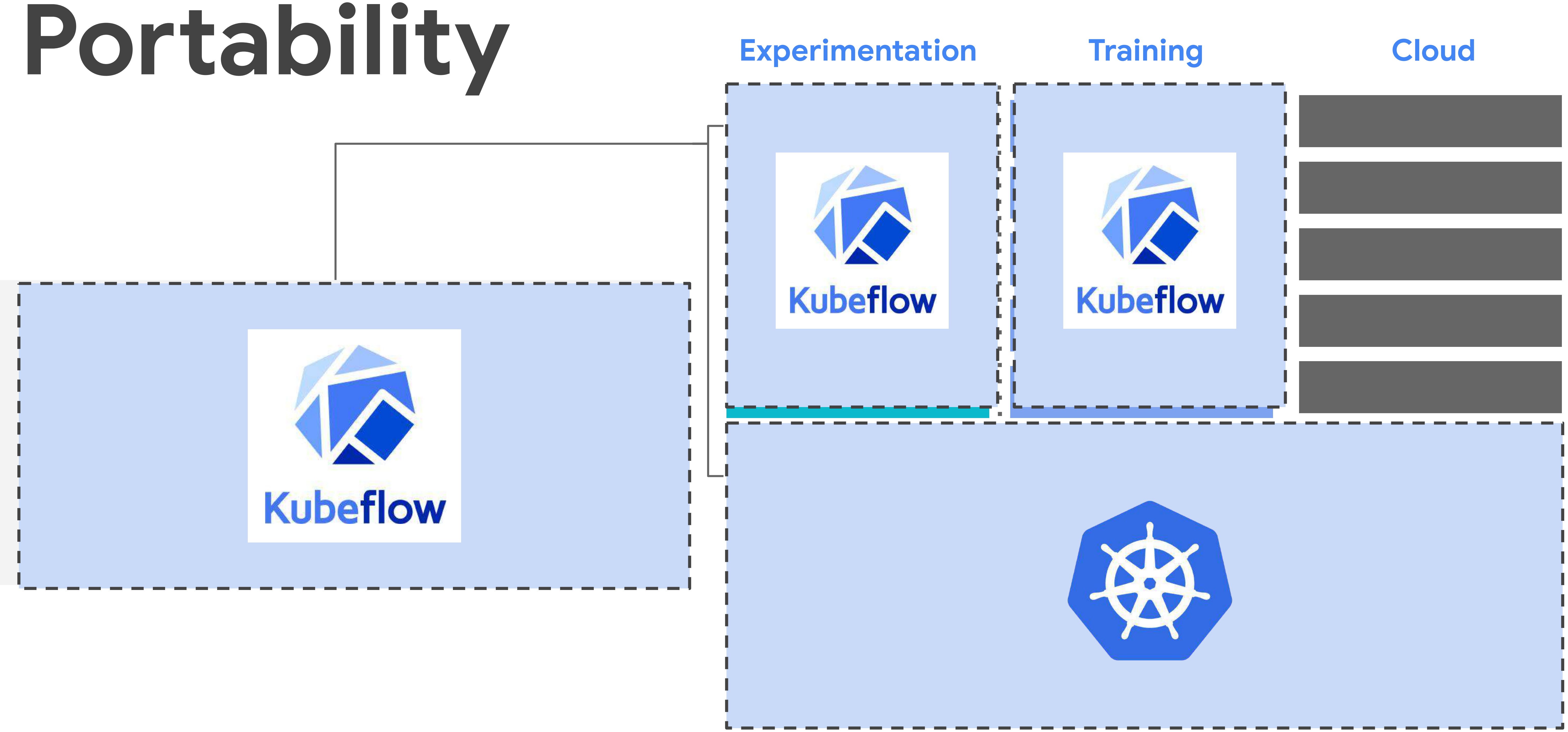
Portability



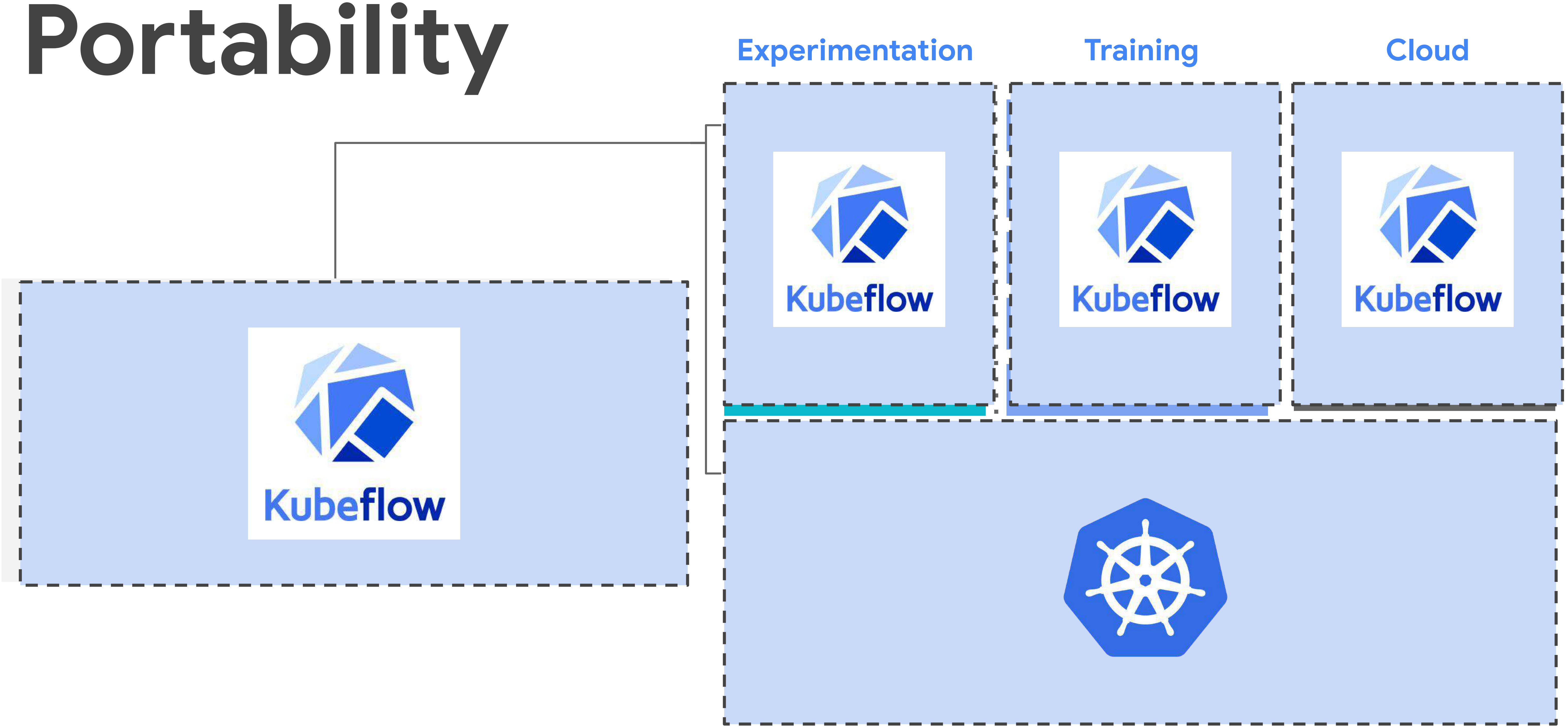
Portability

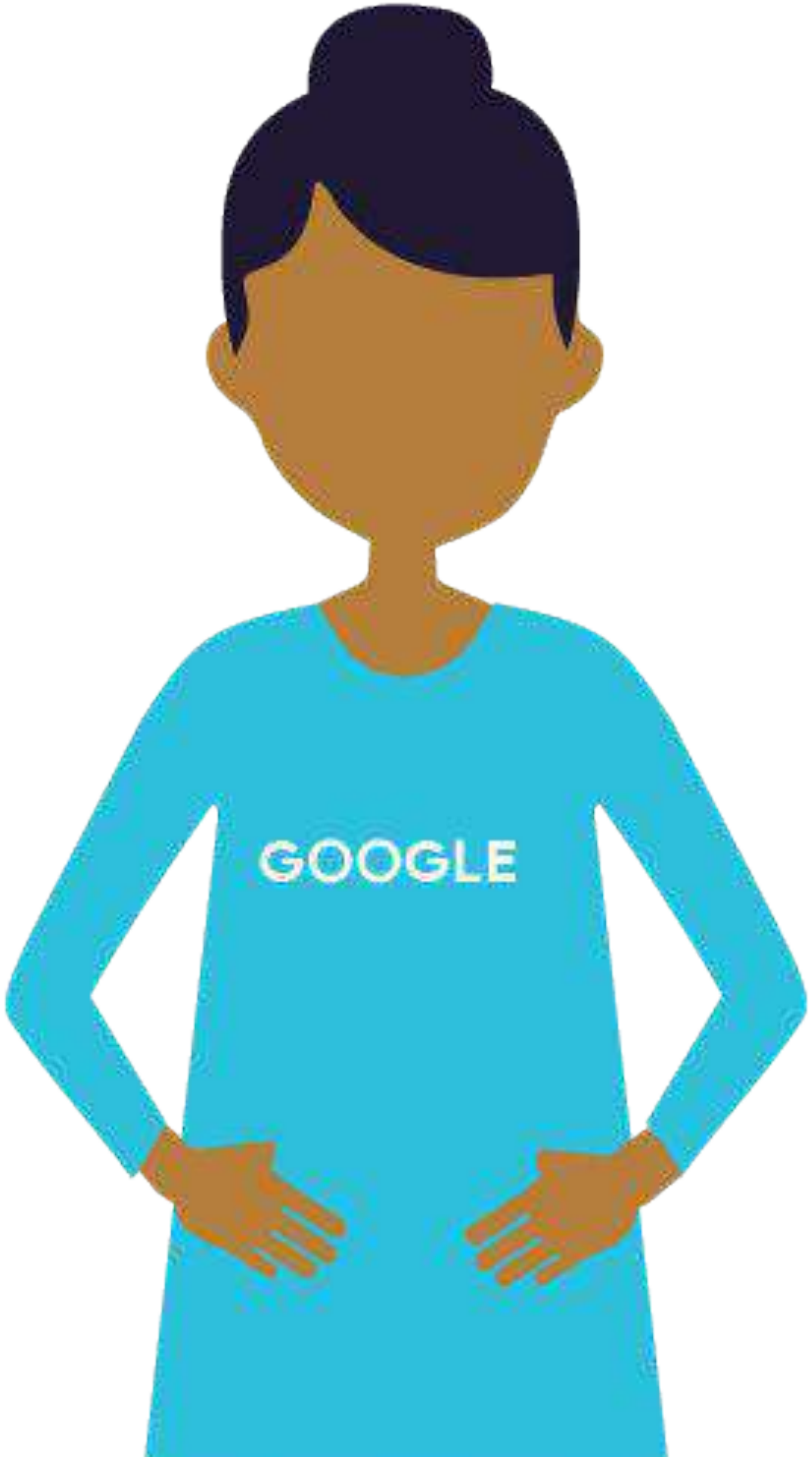


Portability

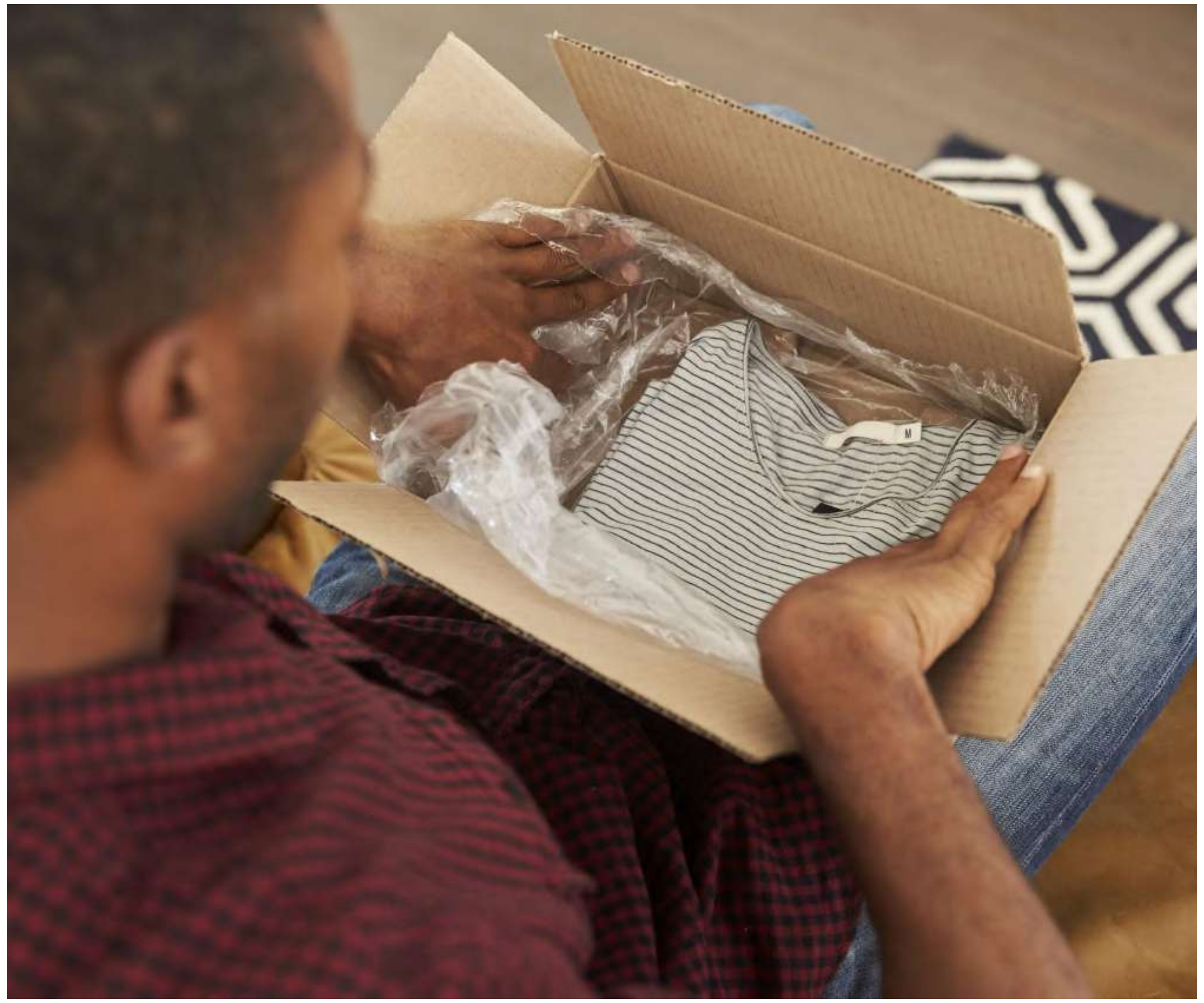


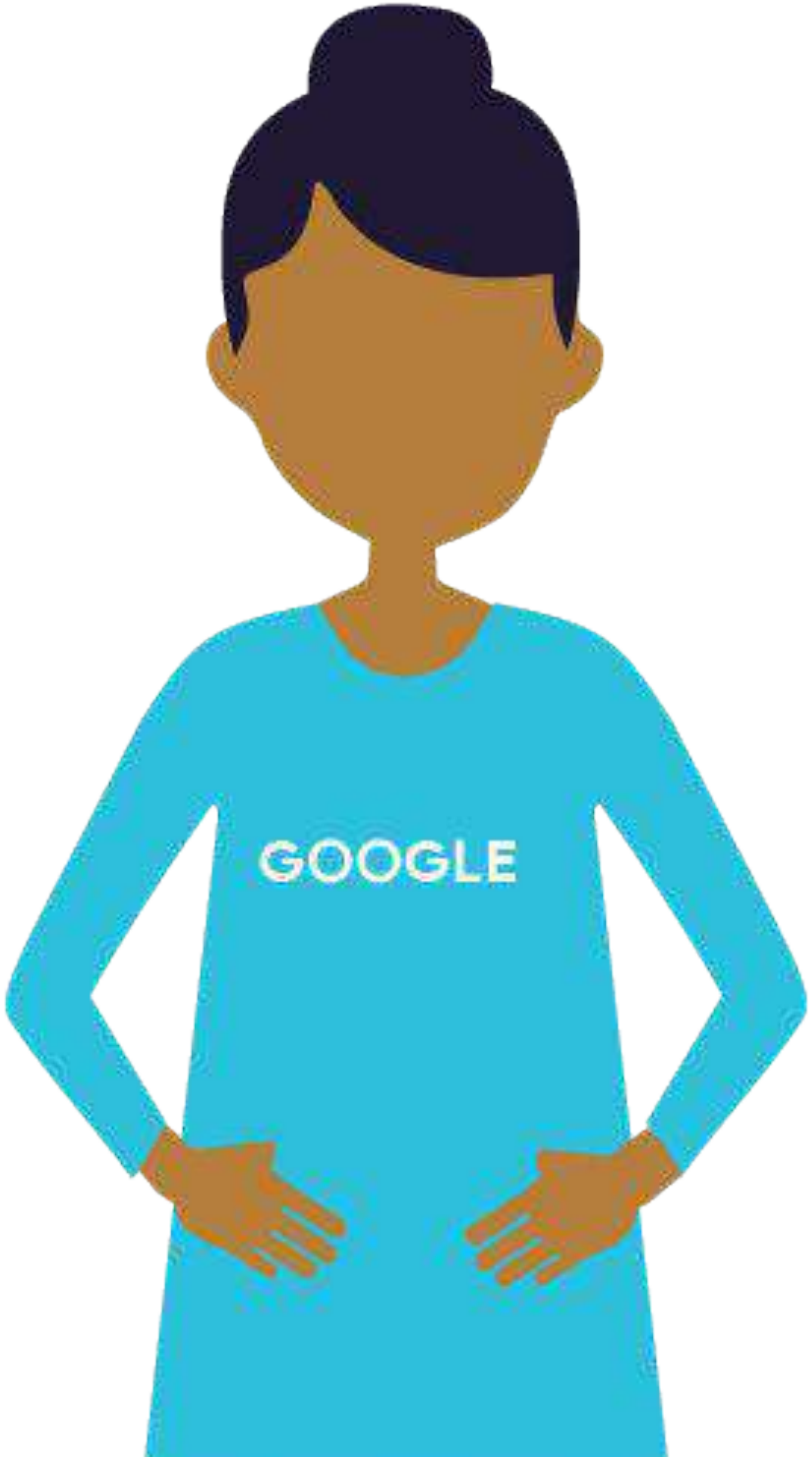
Portability





What's in the box?

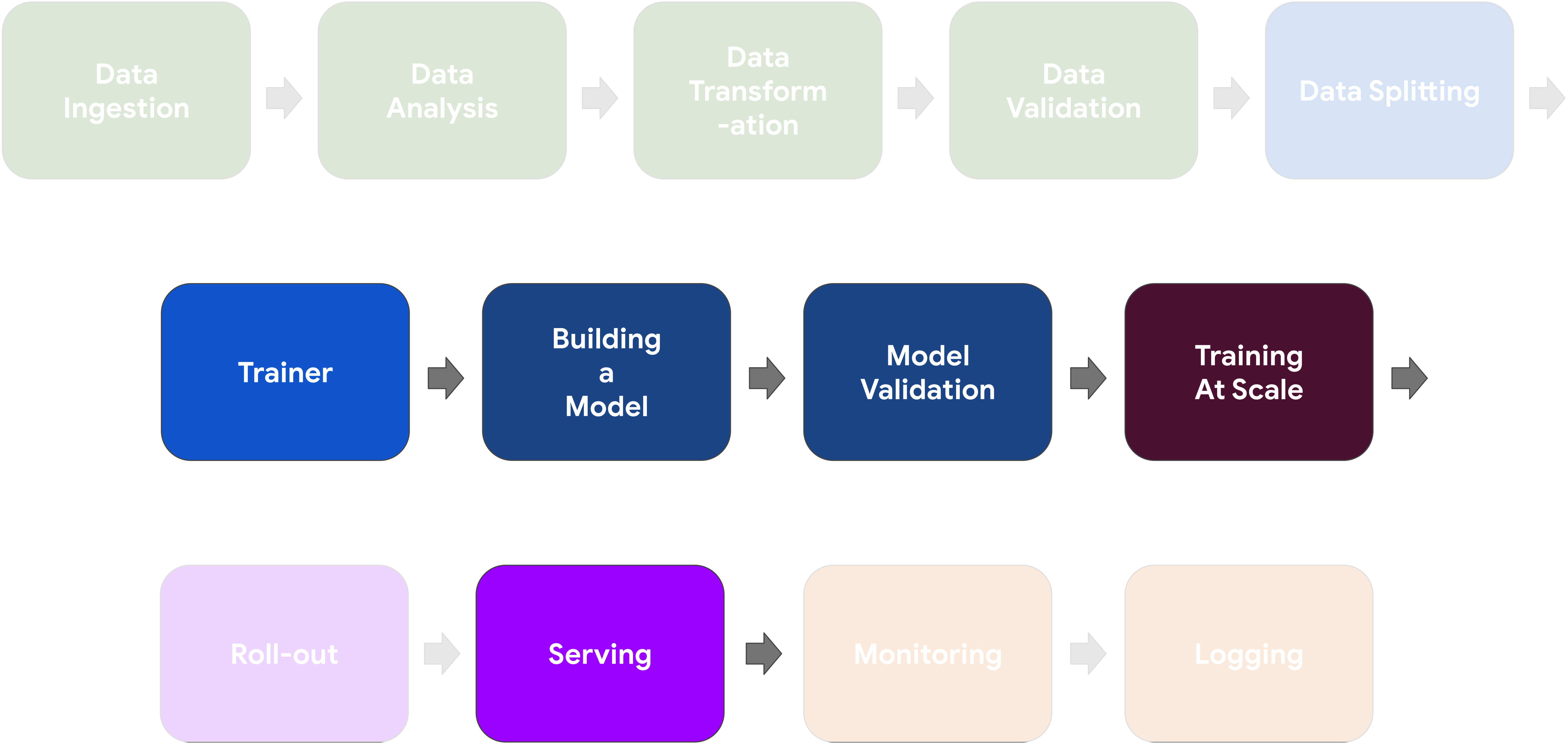




What's in the box?

- **Jupyter notebook**
- Multi-architecture, **distributed training**
- Multi-framework **model serving**
- Examples and walkthroughs for getting started
- **Ksonnet packaging for customizing it yourself!**

What’s in the box?



Data Ingestion

Data Analysis

Data Transform-ation

Data Validation

Data Splitting

Trainer

Building a Model

Model Validation

Training At Scale

Roll-out

Serving

Monitoring

Logging

Courses 7 - Production ML Systems

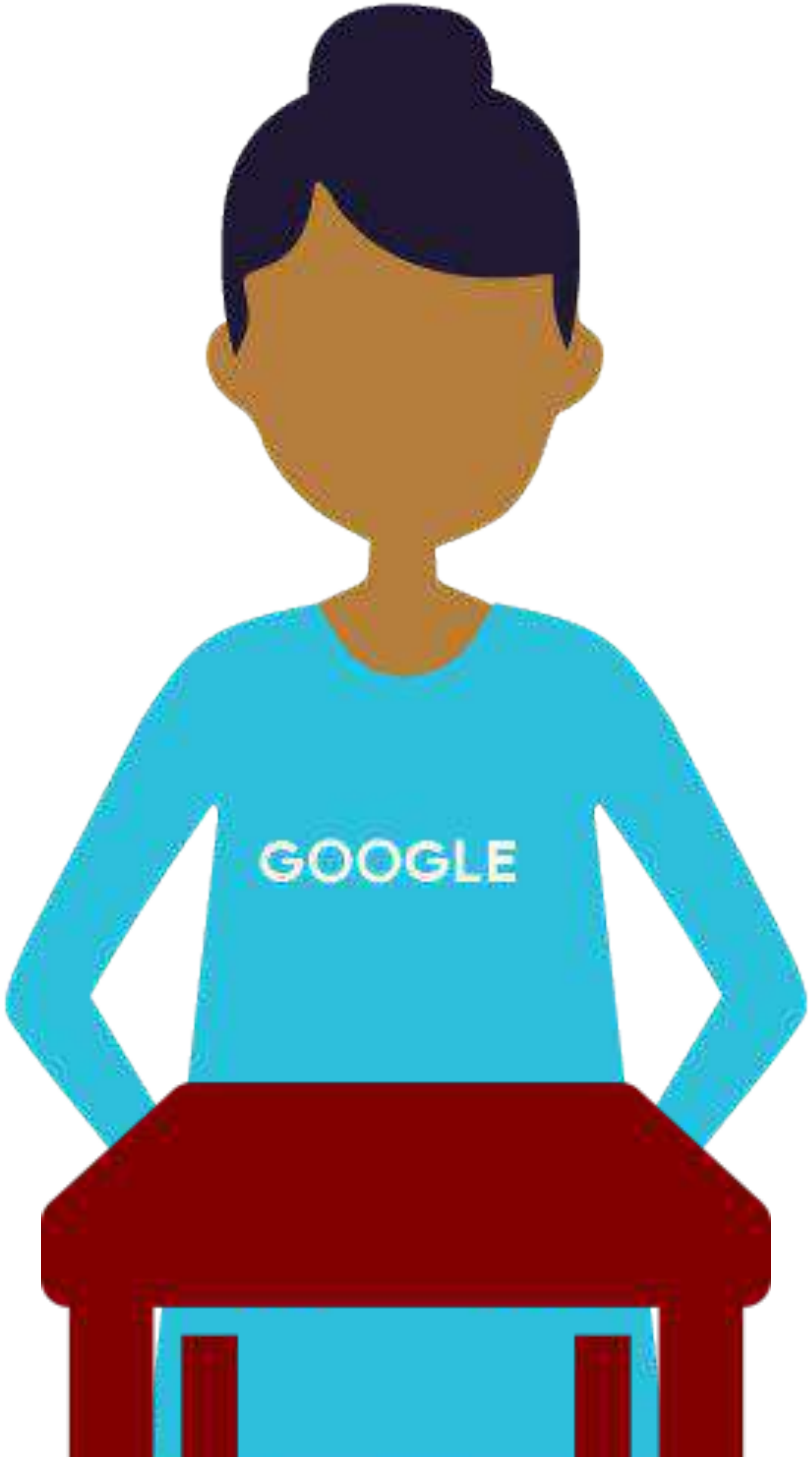
Module 5: Hybrid ML Systems

Lesson Title: **Kubeflow Demo**

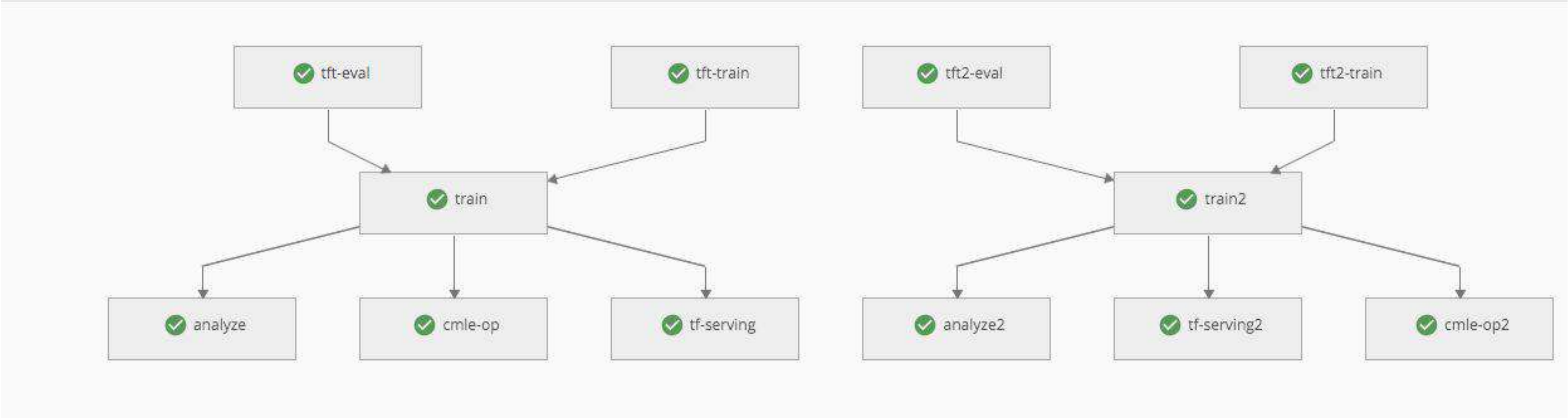
Format: On-Camera Screencast

Presenter: Amy Unruh

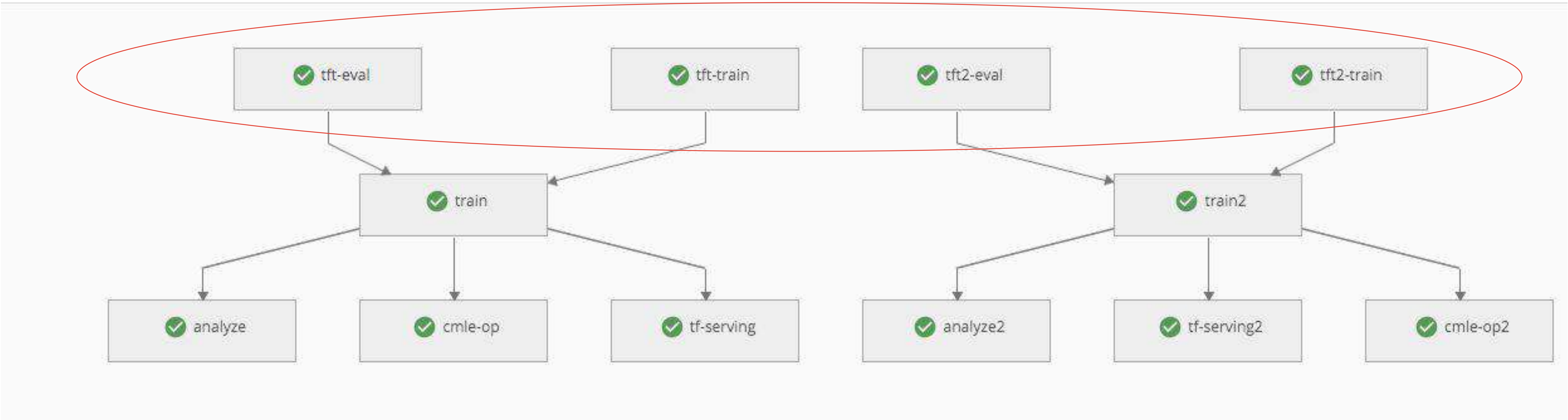
Video Name: T-PSML-O_5_l4_kubeflow_demo

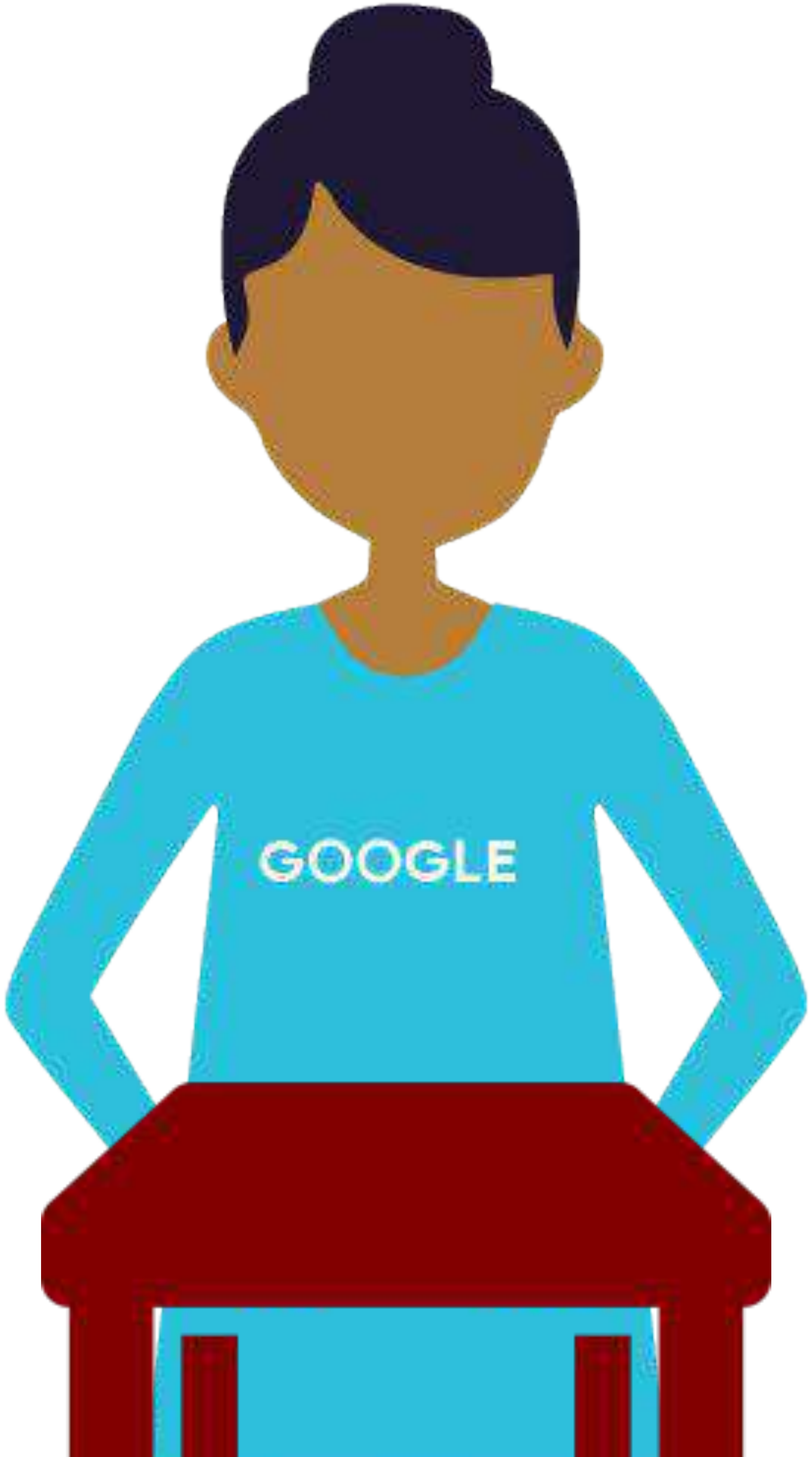

GOOGLE

The ML workflow we're going to run



The ML workflow we're going to run



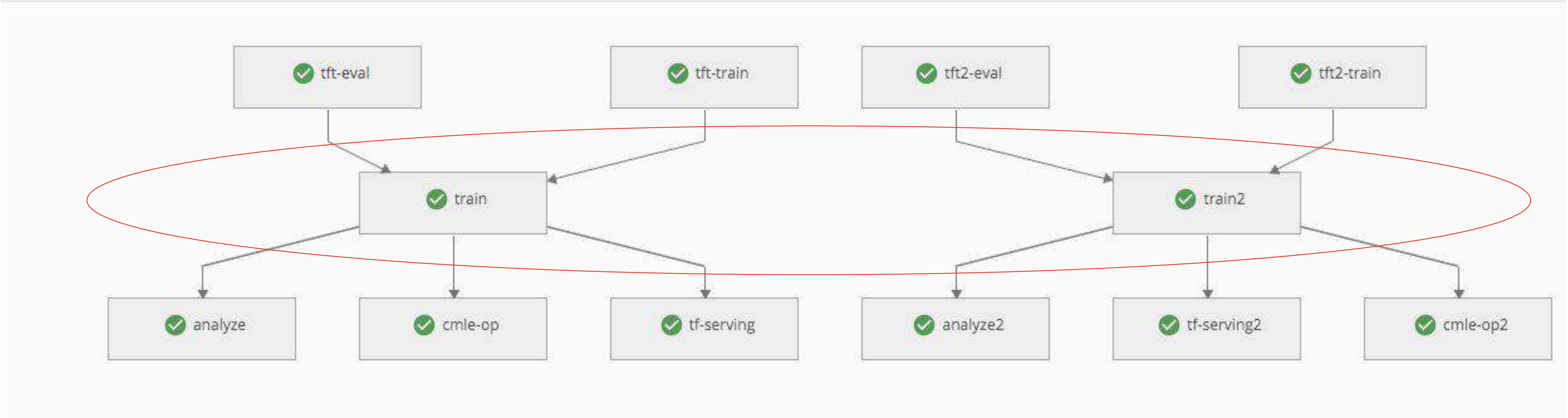


Feature Engineering + Model Analysis

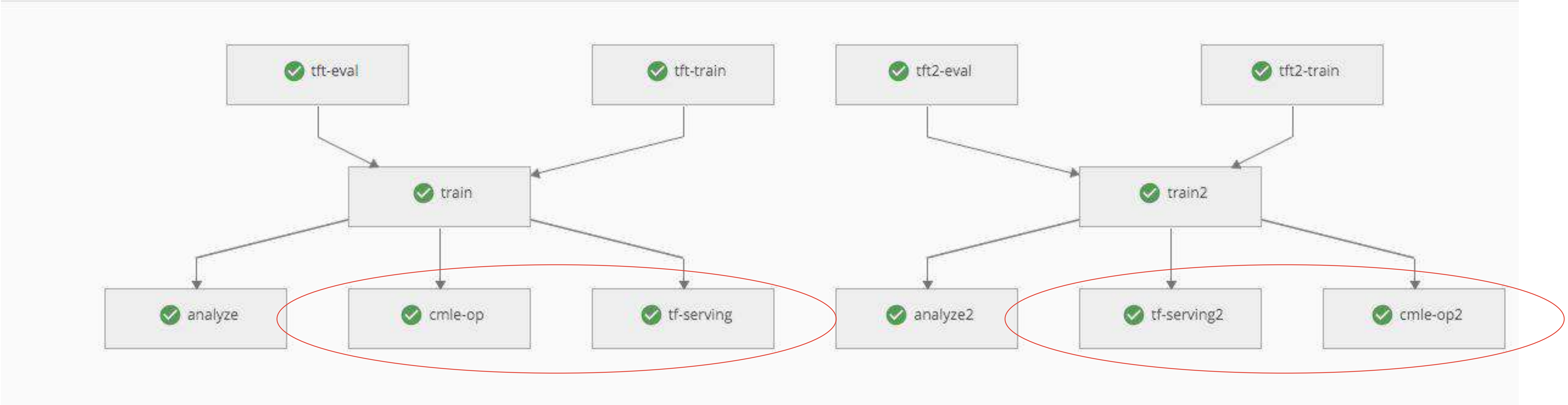
TensorFlow

tf.transform

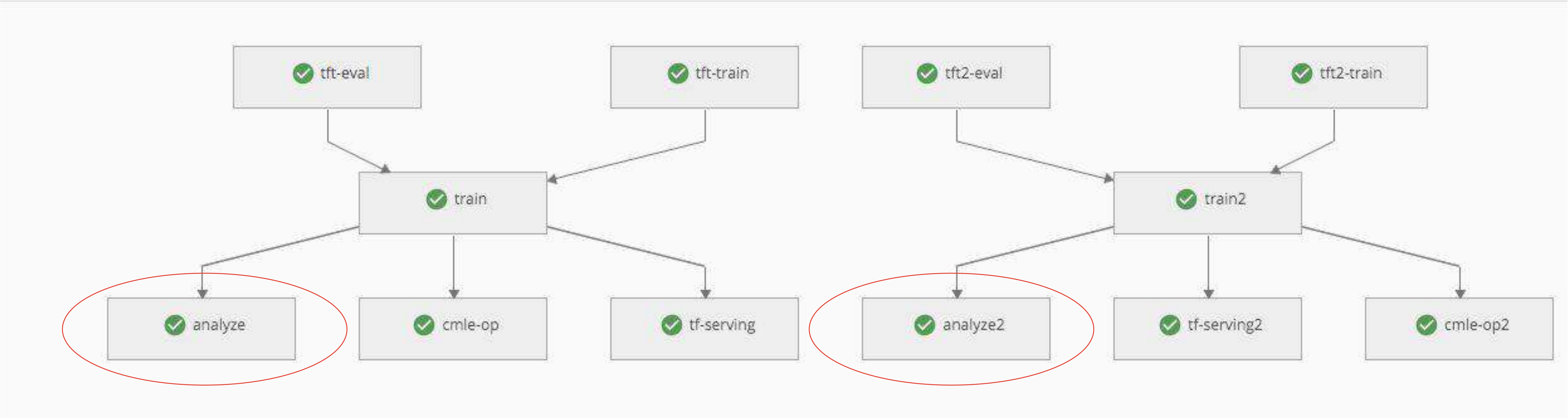
The ML workflow we're going to run

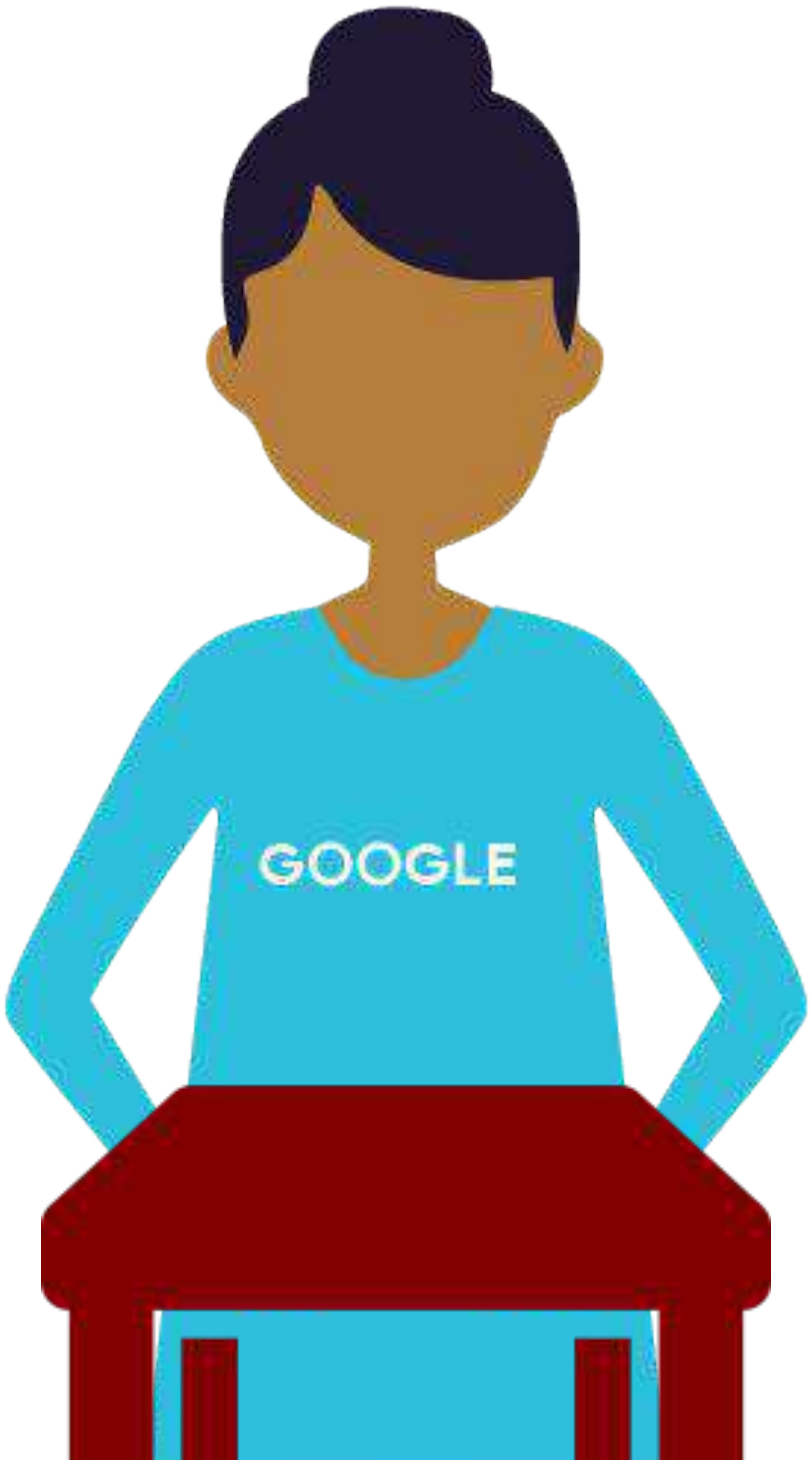


The ML workflow we're going to run



The ML workflow we’re going to run



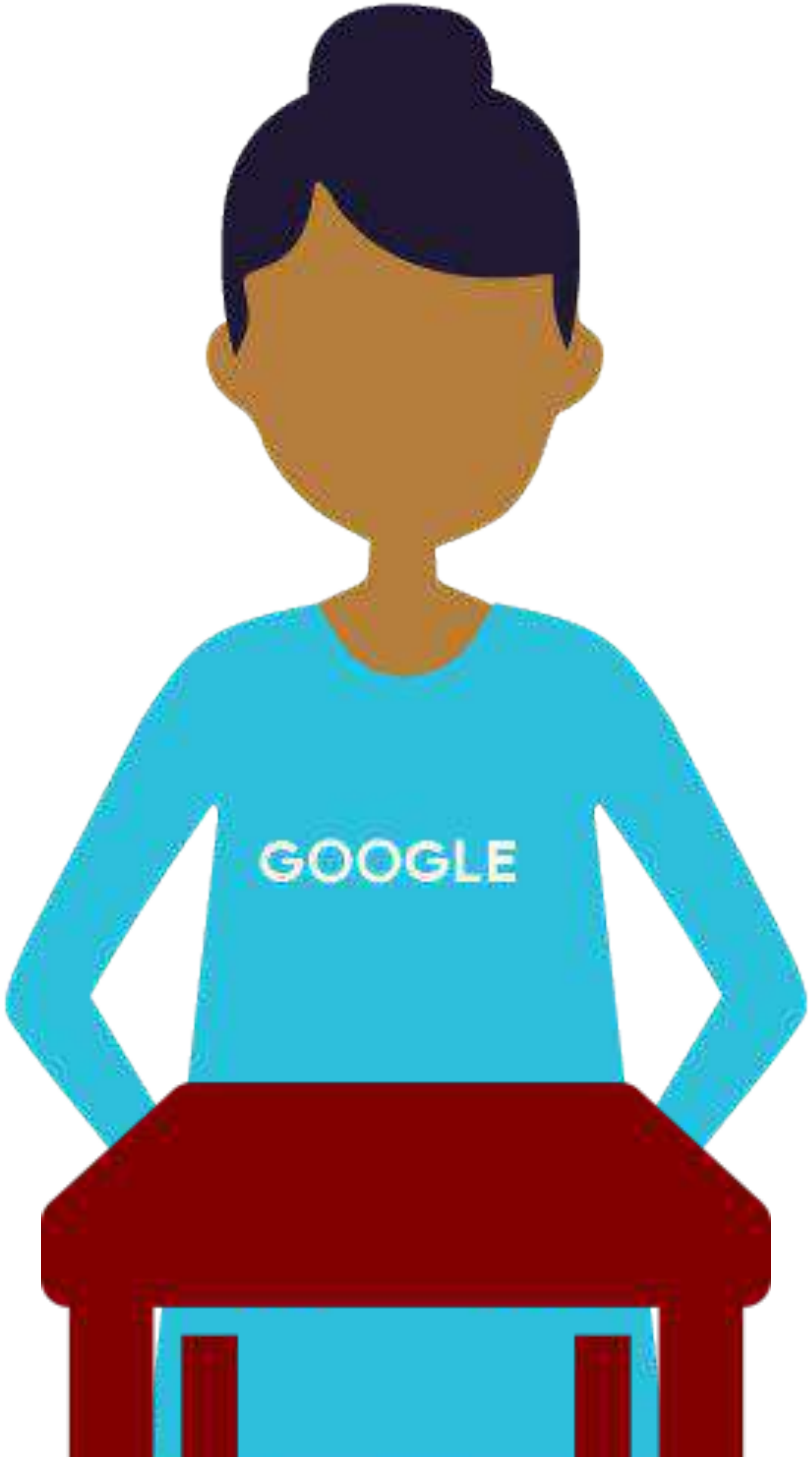


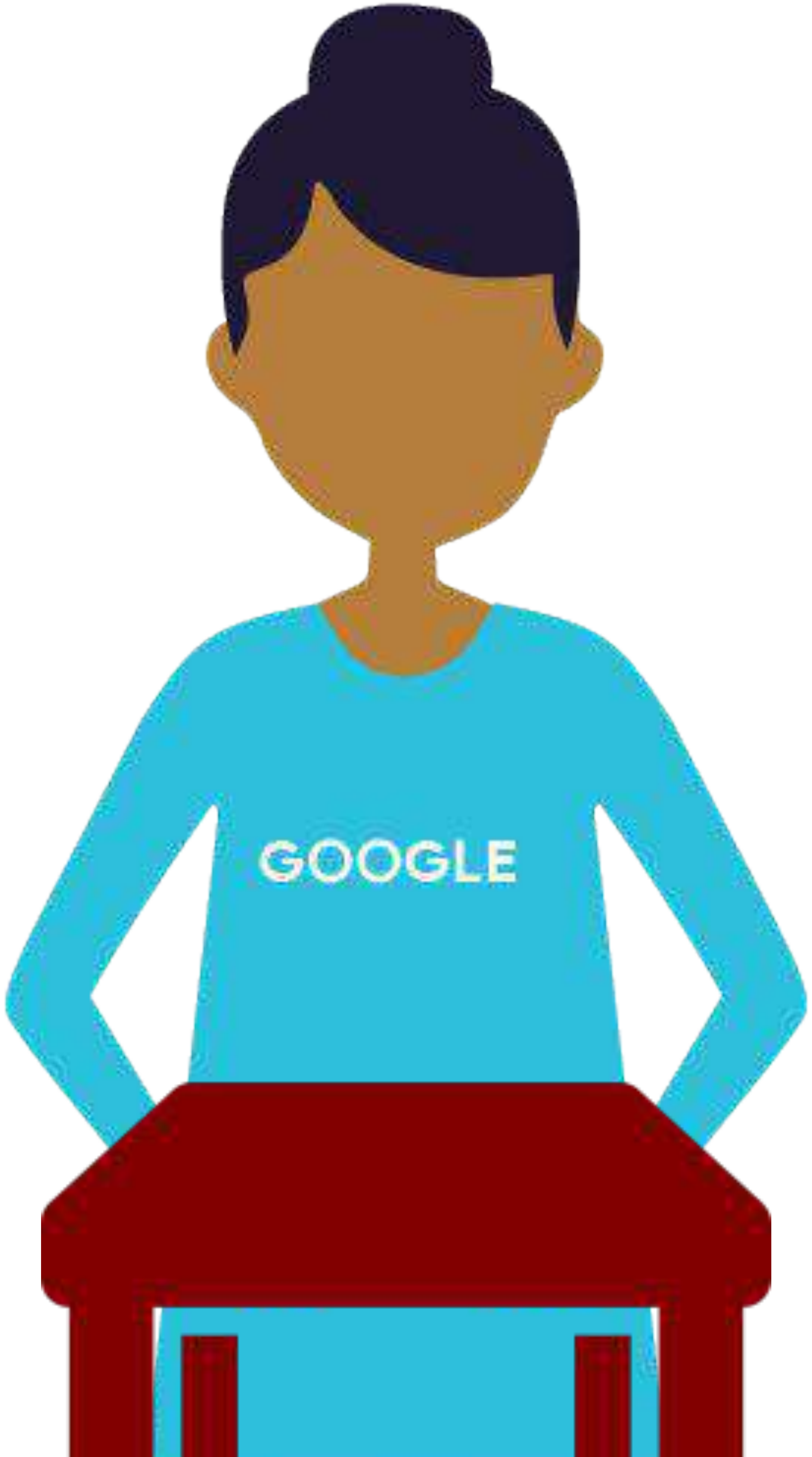
Feature Engineering + Model Analysis

TensorFlow

tf.transform

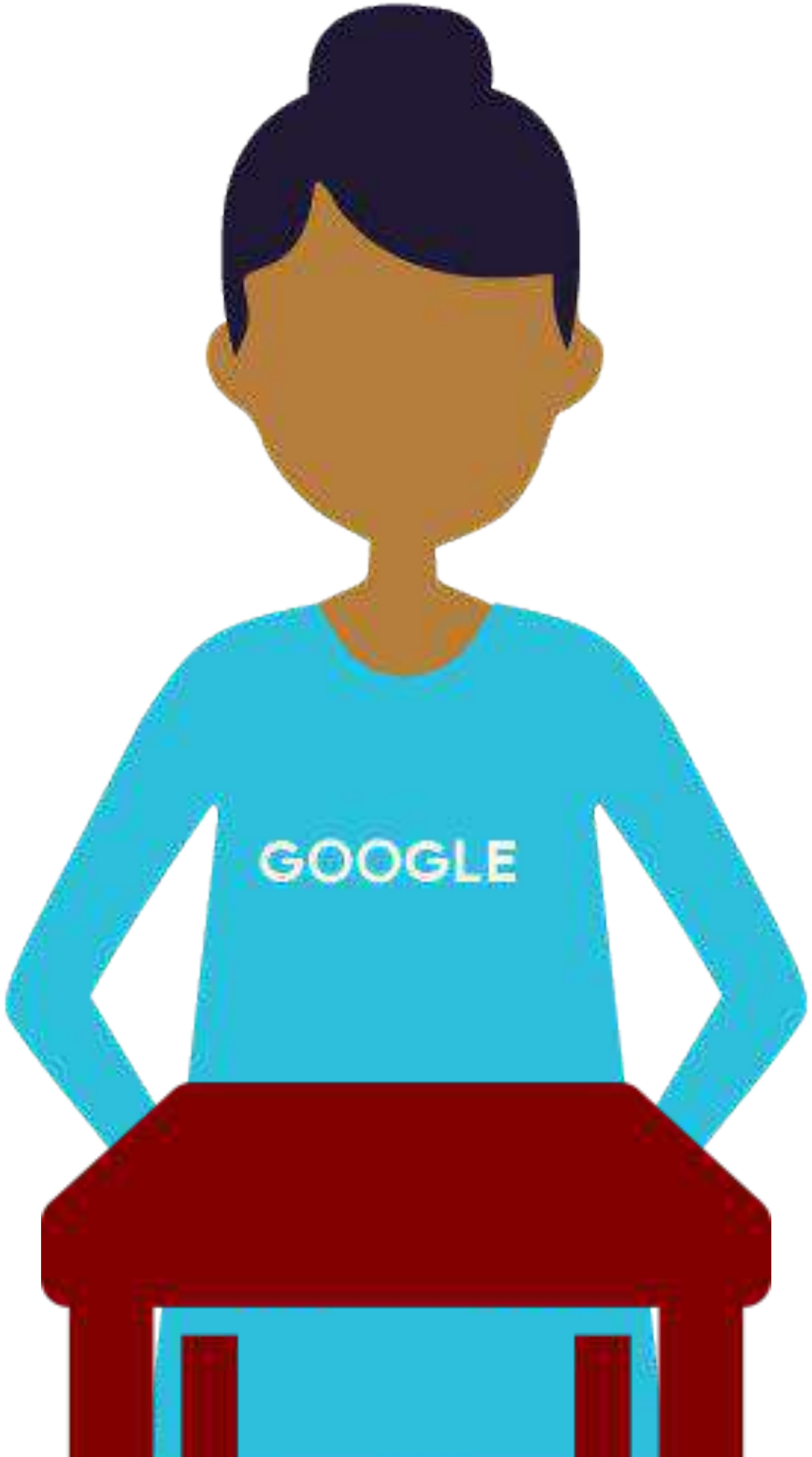
TensorFlow Model Analysis


GOOGLE



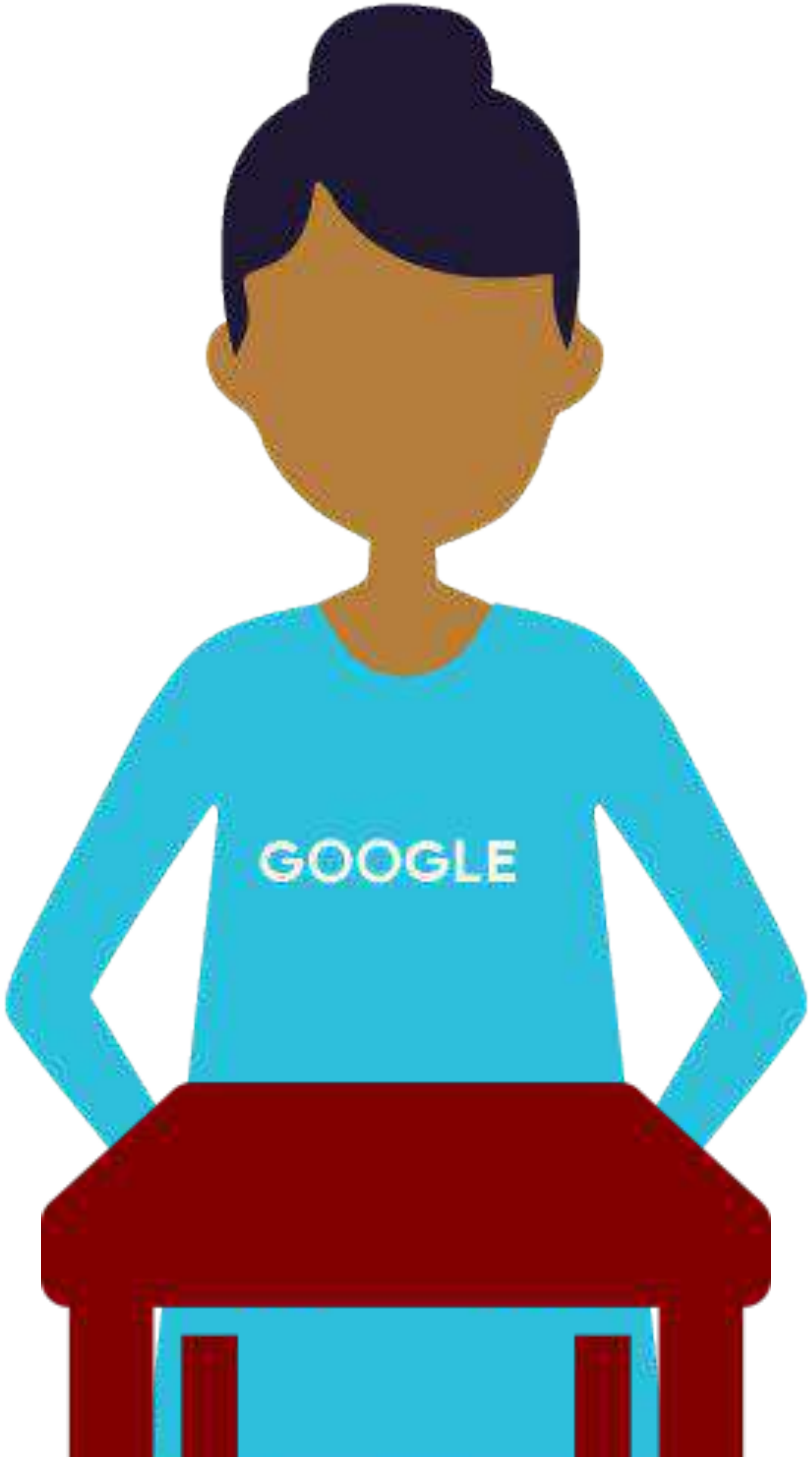
Kubeflow Benefits

- Portability
- Composability and Reproducibility
- Scalability
- Visualization and Collaboration



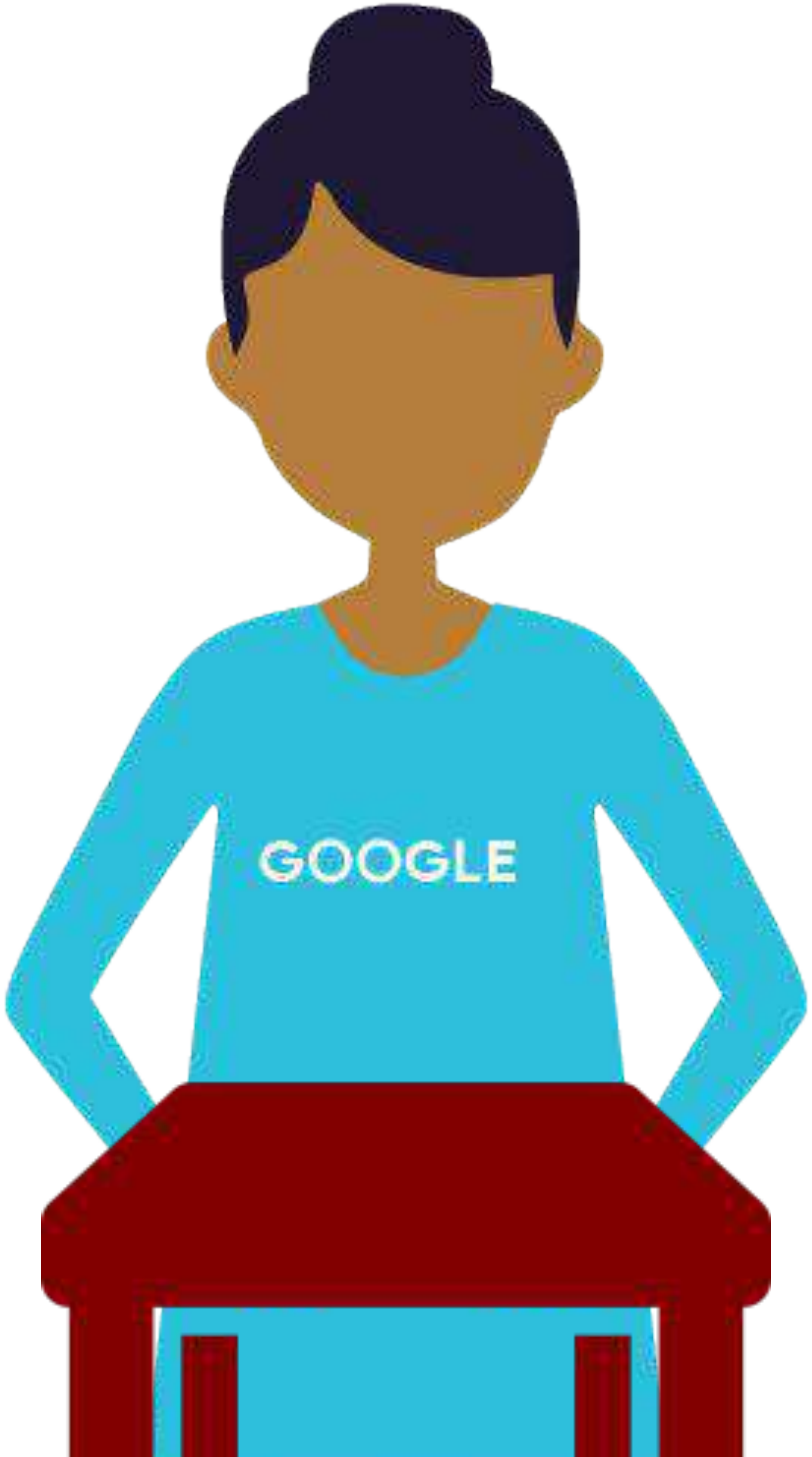
Kubeflow Benefits

- **Portability**
- Composability and Reproducibility
- Scalability
- Visualization and Collaboration



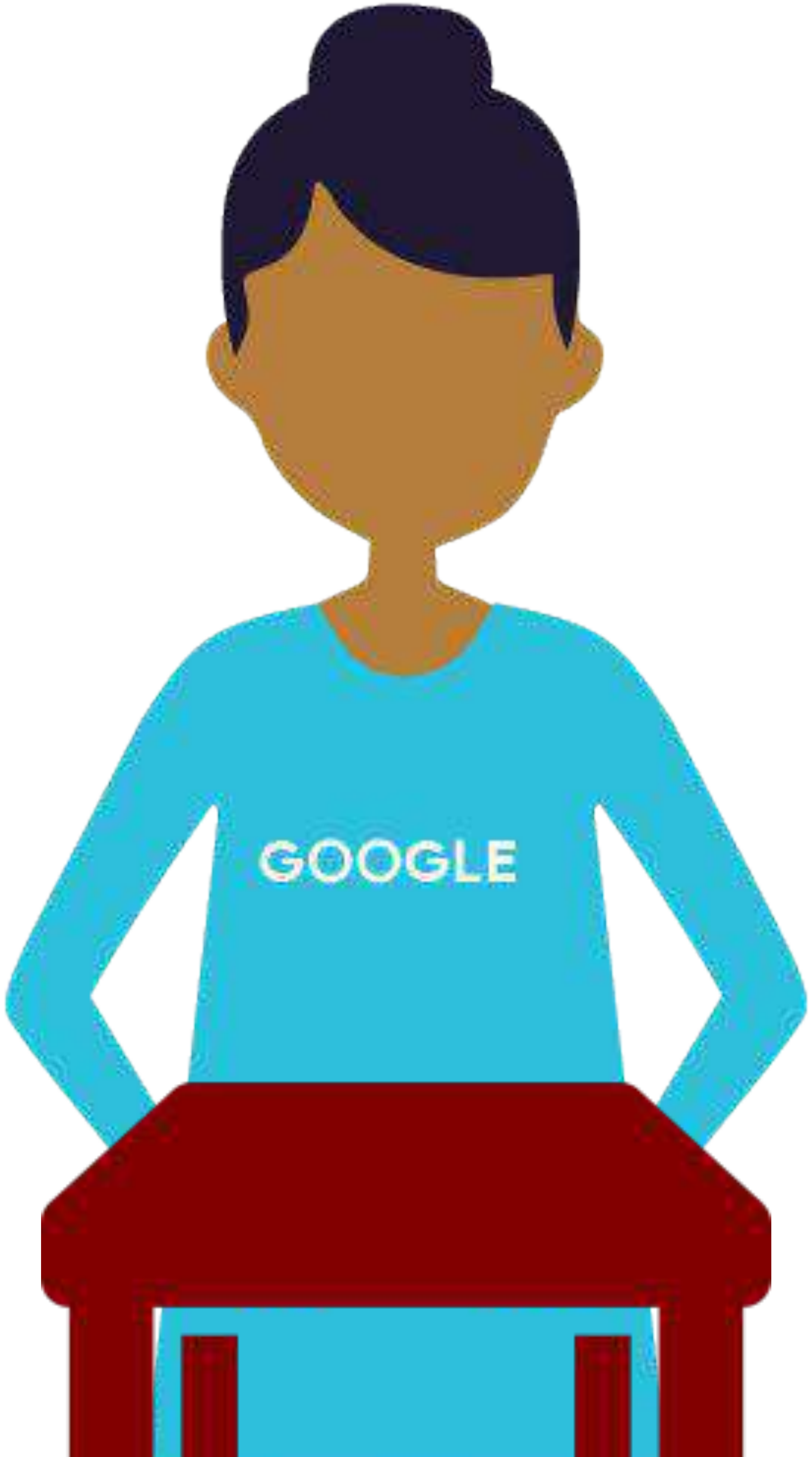
Kubeflow Benefits

- Portability
- **Composability and Reproducibility**
- Scalability
- Visualization and Collaboration



Kubeflow Benefits

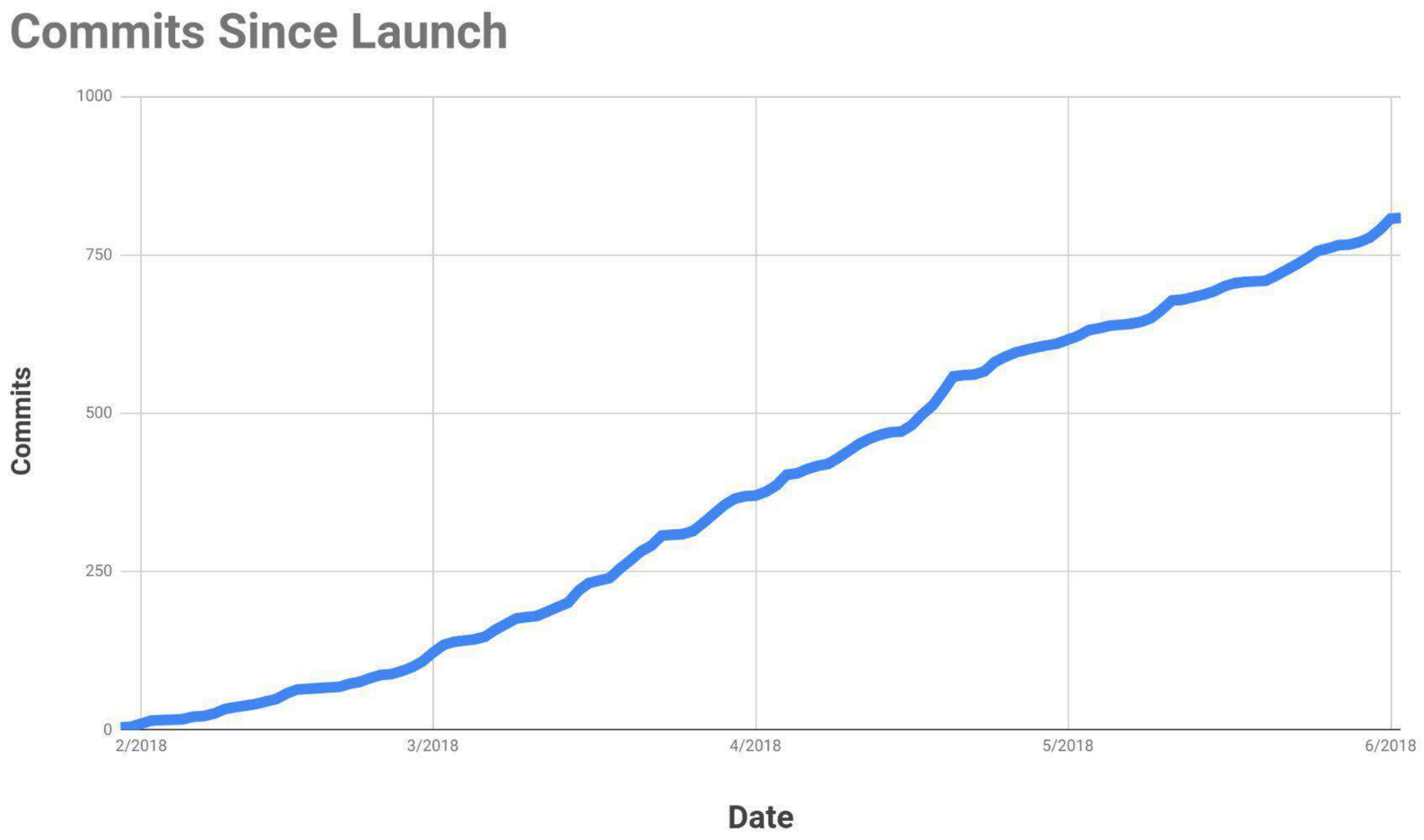
- Portability
- Composability and Reproducibility
- **Scalability**
- Visualization and Collaboration



Kubeflow Benefits

- Portability
- Composability and Reproducibility
- Scalability
- **Visualization and Collaboration**

Momentum!



- 800+ commits
- 70+ Community contributors
- 17+ Companies

Alibaba Cloud 蚂蚁金服 ANT FINANCIAL Caicloud CANONICAL

CISCO DATAWIRE DELL GitHub

Google heptio HUAWEI intel

Microsoft MOMENTA Pachyderm jupyter

redhat Seldon Technologies UBER weaveworks

Courses 7 - Production ML Systems

Module 5: Hybrid ML Systems

Lesson Title: **Embedded Models**

Format: Presenter

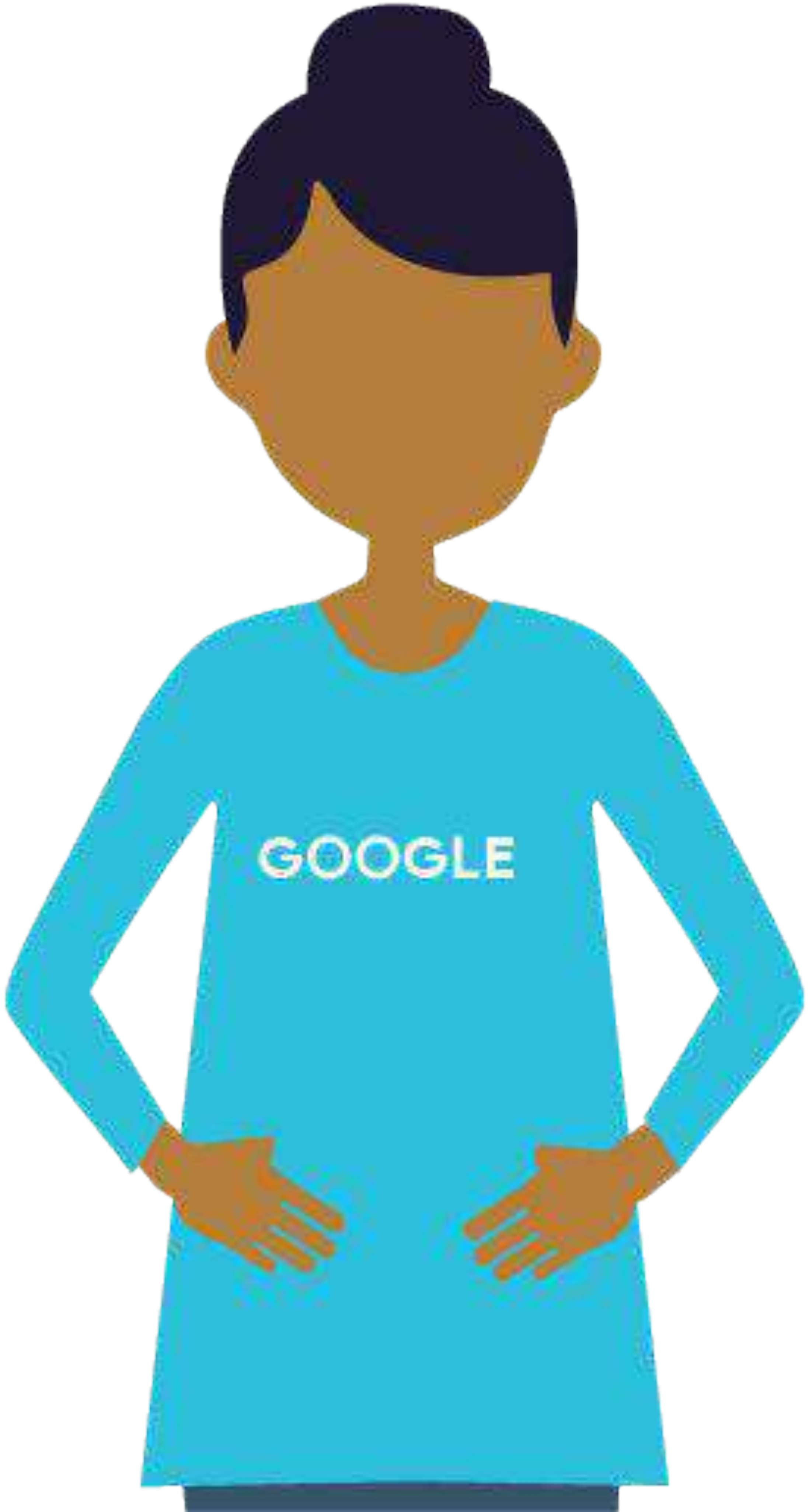
Presenter: Val

Video Name: T-PSML-O_5_l5_embedded_models

Agenda

Kubeflow for hybrid cloud

Optimizing TensorFlow for mobile



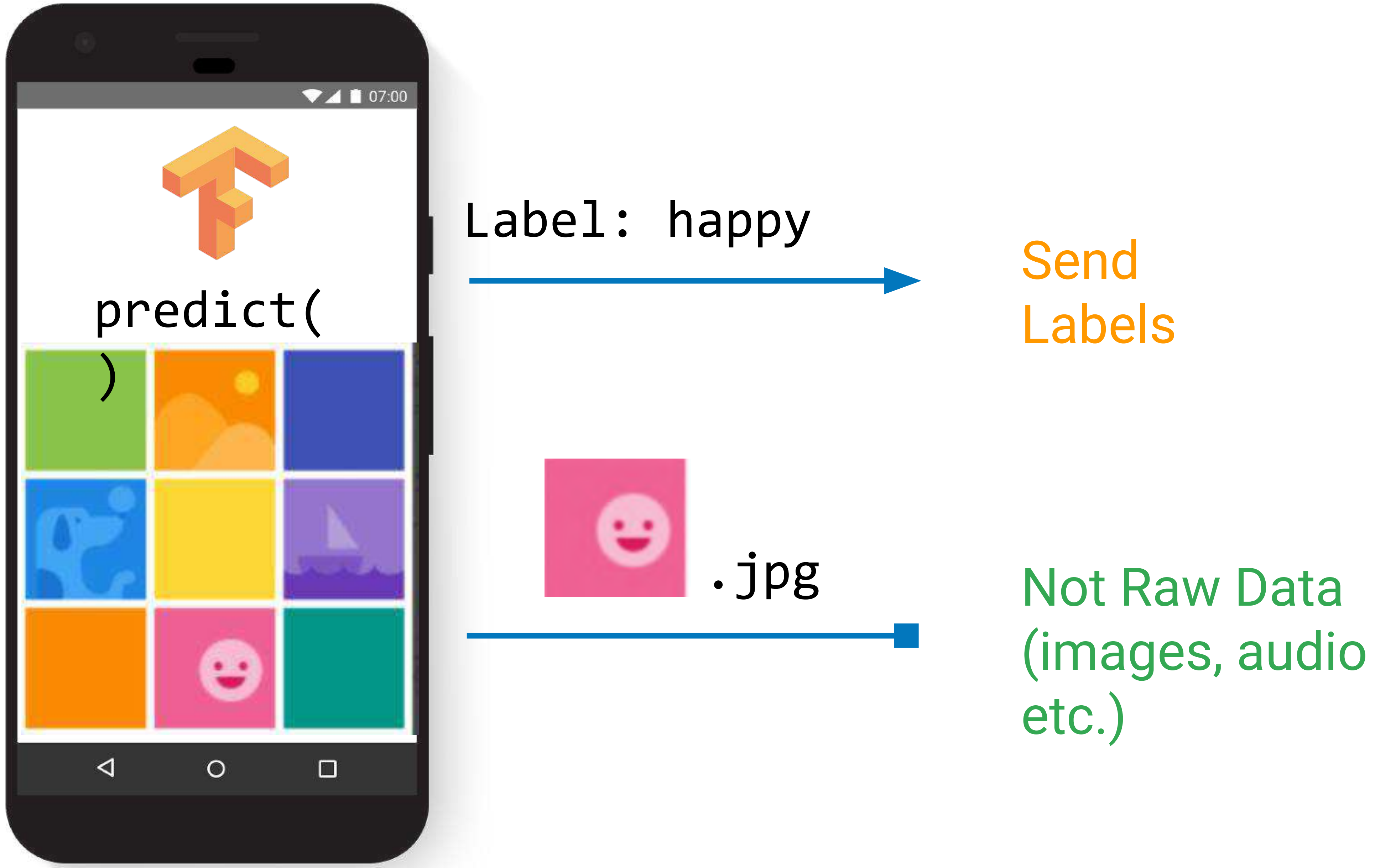
Increasingly, applications are combining ML with mobile apps

- Image/OCR
- Speech ⇔ Text
- Translation

ML models can help extract meaning from raw data, thus reducing network traffic

- Image recognition: send raw image v. send detected label

- Motion detection: send raw motion v. send feature vector

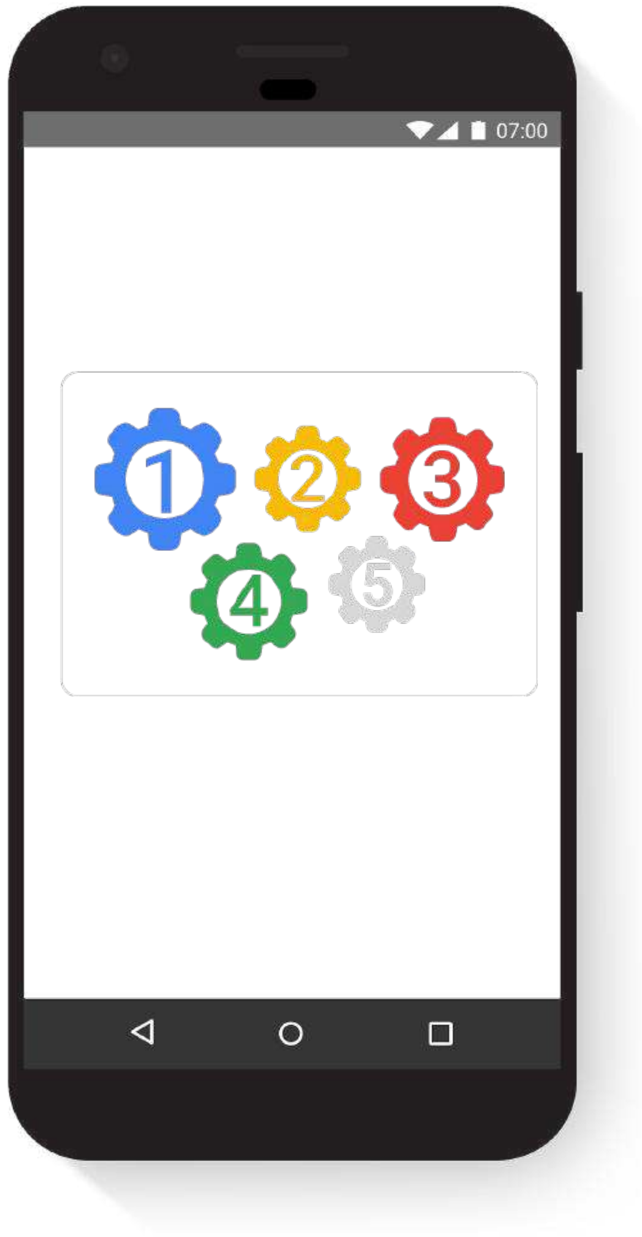


Send Labels

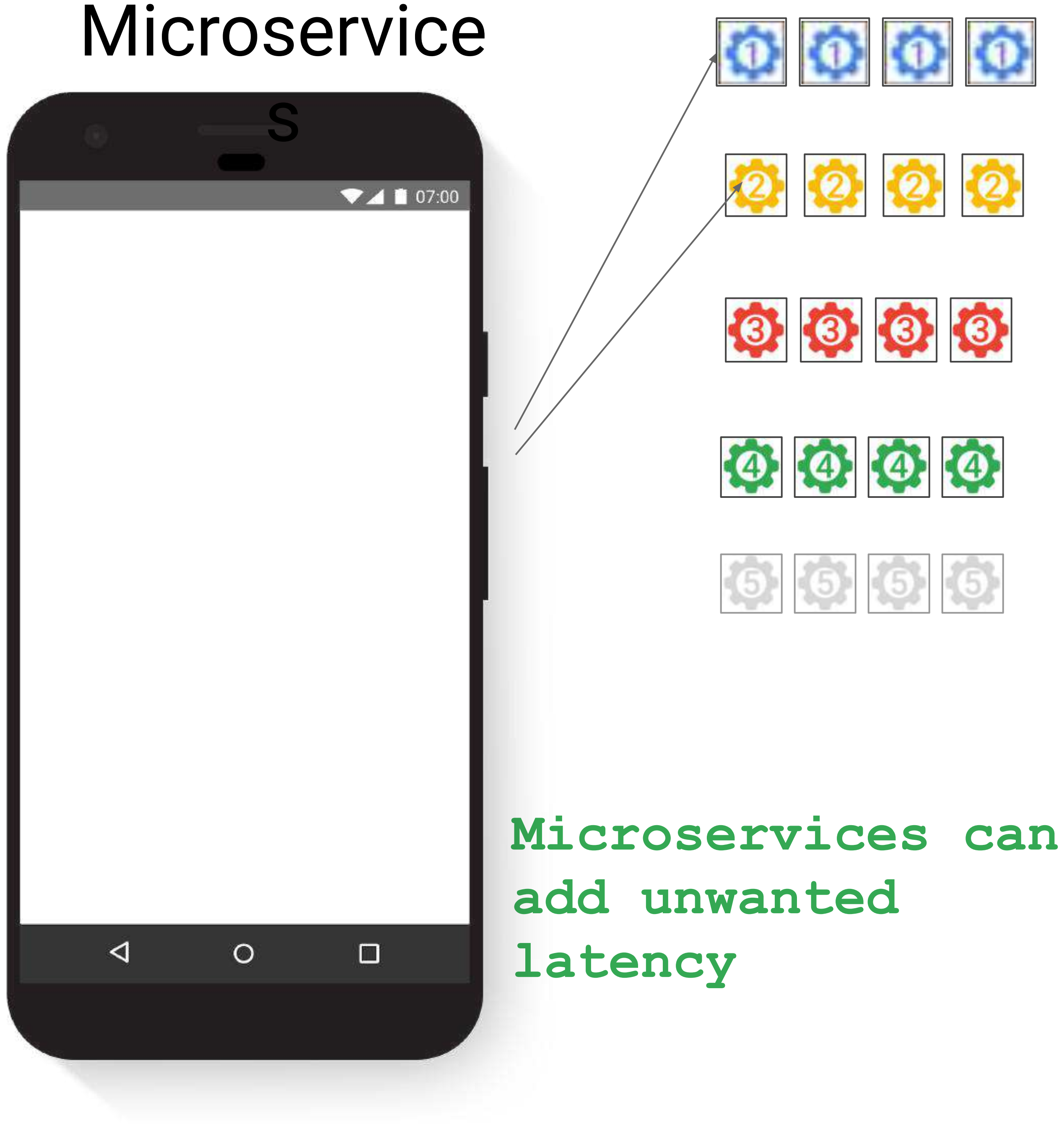
Not Raw Data (images, audio etc.)

From mobile devices, we often can't use the microservices approach

Monolithic Service

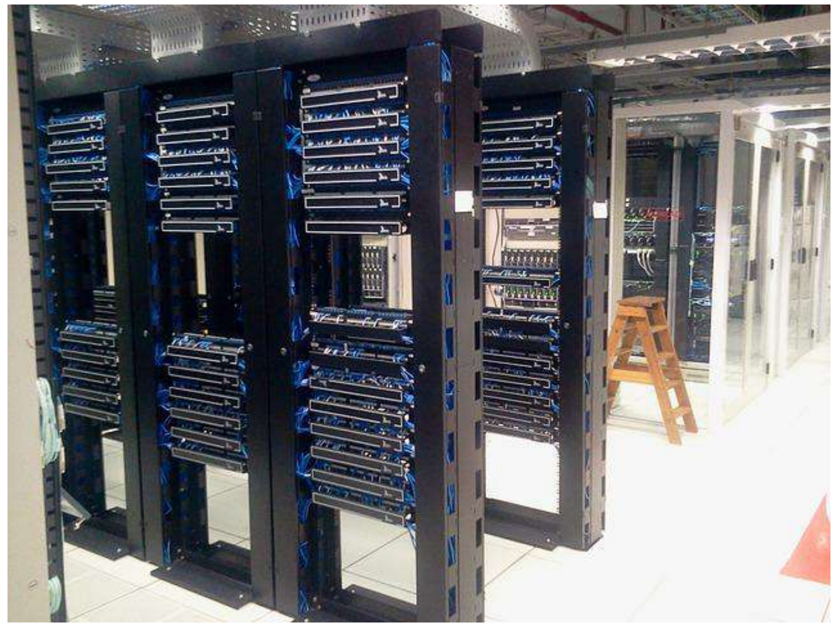


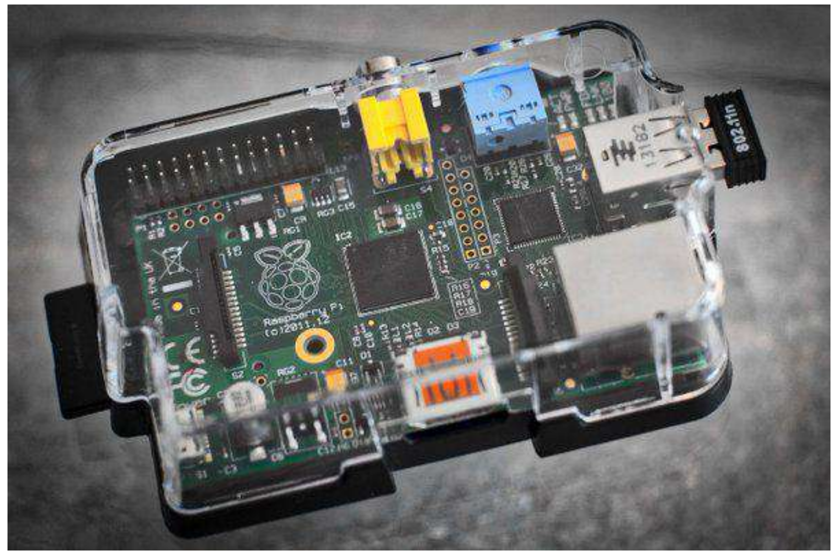
Microservice

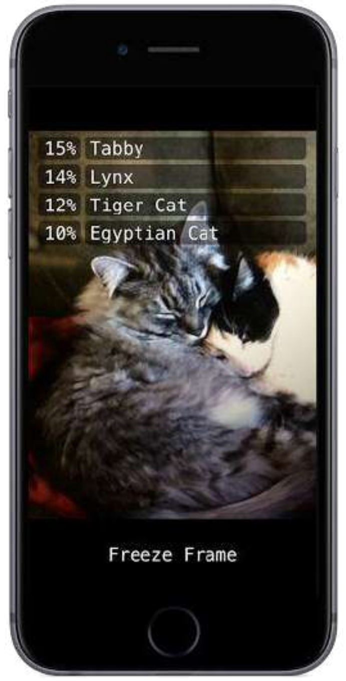


Microservices can add unwanted latency

In these situations, we'd like to train on the cloud, predict on device







```
15% Tabby
14% Lynx
12% Tiger Cat
10% Egyptian Cat
```

Courses 7 - Production ML Systems

Module 5: Hybrid ML Systems

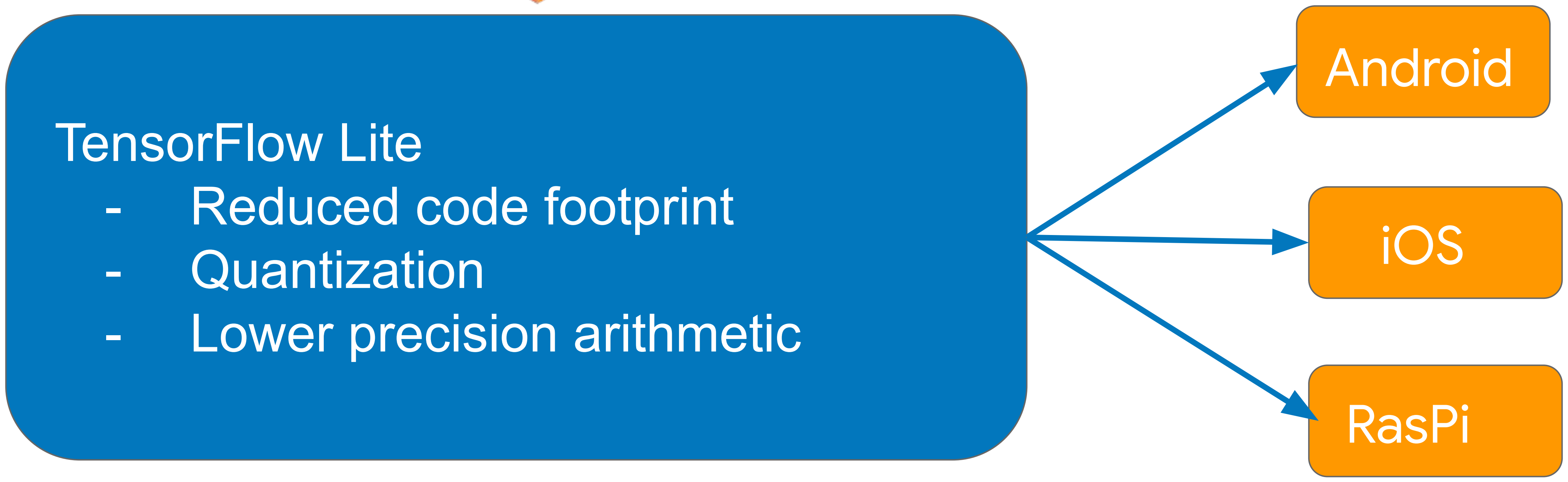
Lesson Title: **TensorFlow Lite**

Format: Presenter

Presenter: Val

Video Name: T-PSML-O_5_l6_tensorflow_lite

TensorFlow supports multiple mobile platforms



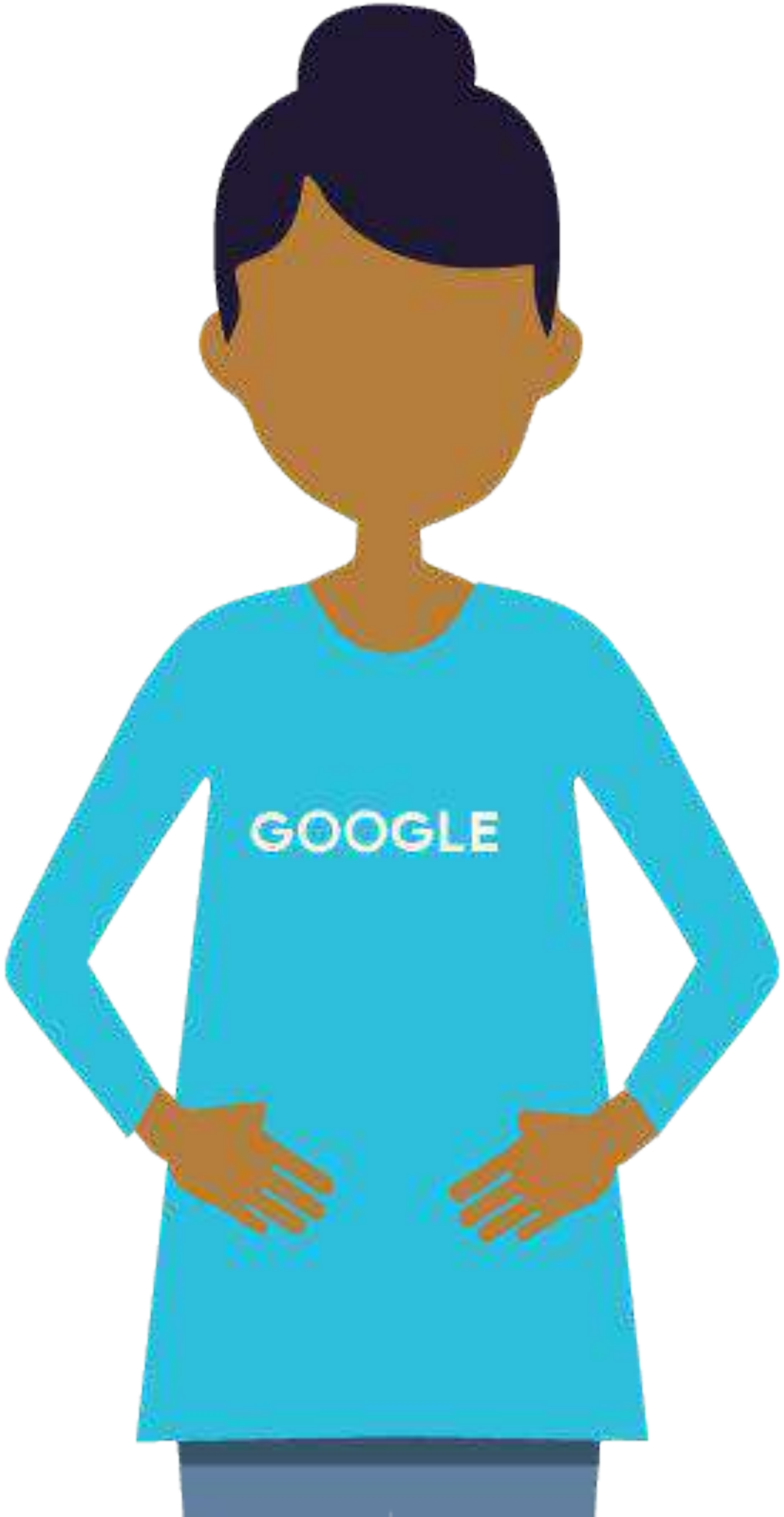
TensorFlow Lite

- Reduced code footprint
- Quantization
- Lower precision arithmetic

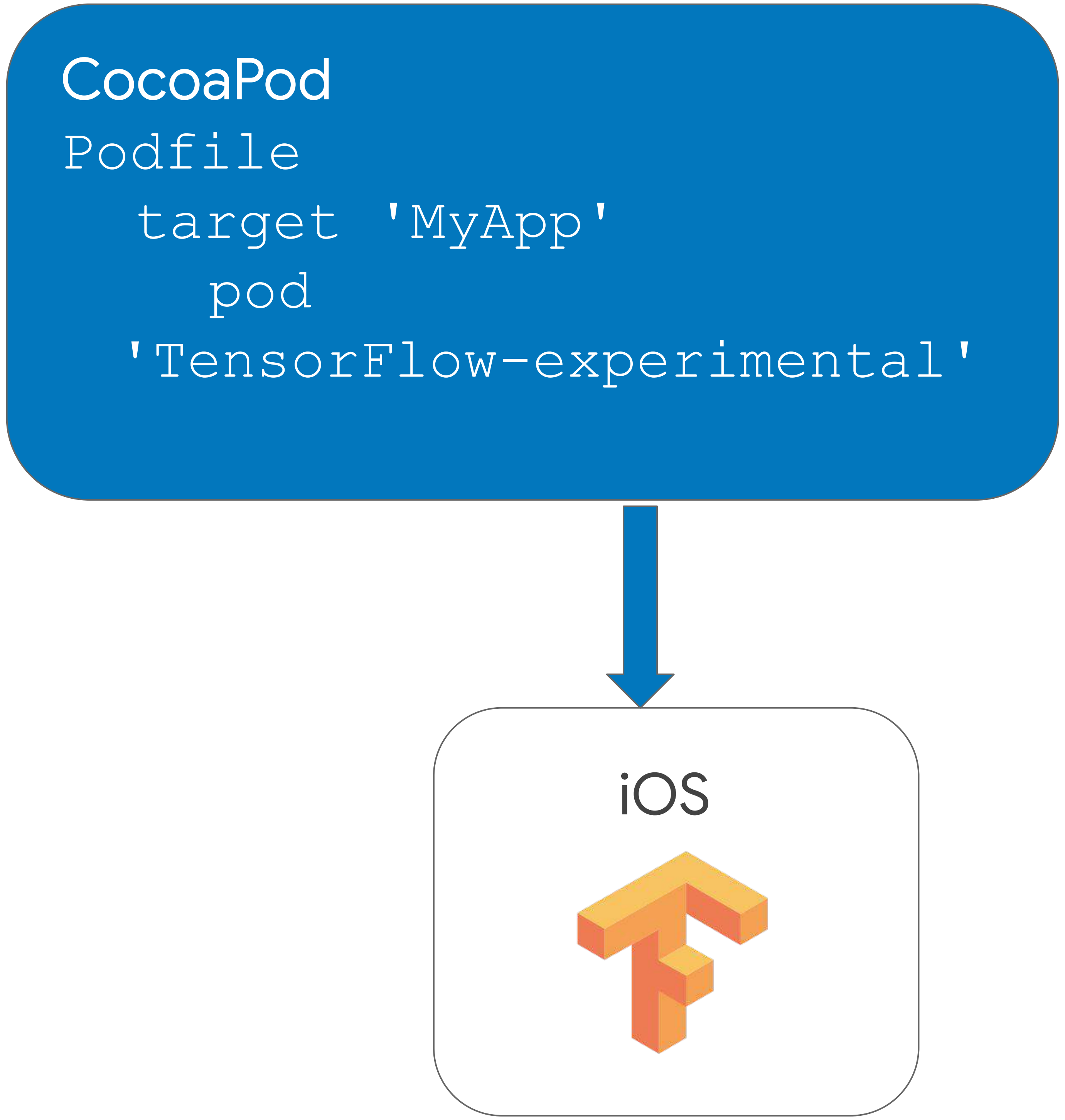
Build with Bazel by starting with a git clone

```
Install:
     TensorFlow
     Bazel
     Android Studio
(optional)
     Android SDK
     Android NDK

Config:
     Edit
tensorflow/WORKSPACE
```

```
android_sdk_repository(
    name = "androidsdk",
    api_level = 23,
    build_tools_version = "25.0.2",
    path =
    "<path-to-android-sdk>",
)
android_ndk_repository(
    Name = "androidndk",
    Path = "<path-to-android-ndk>",
    api_level=14
)
```



Cocoapods support for iOS



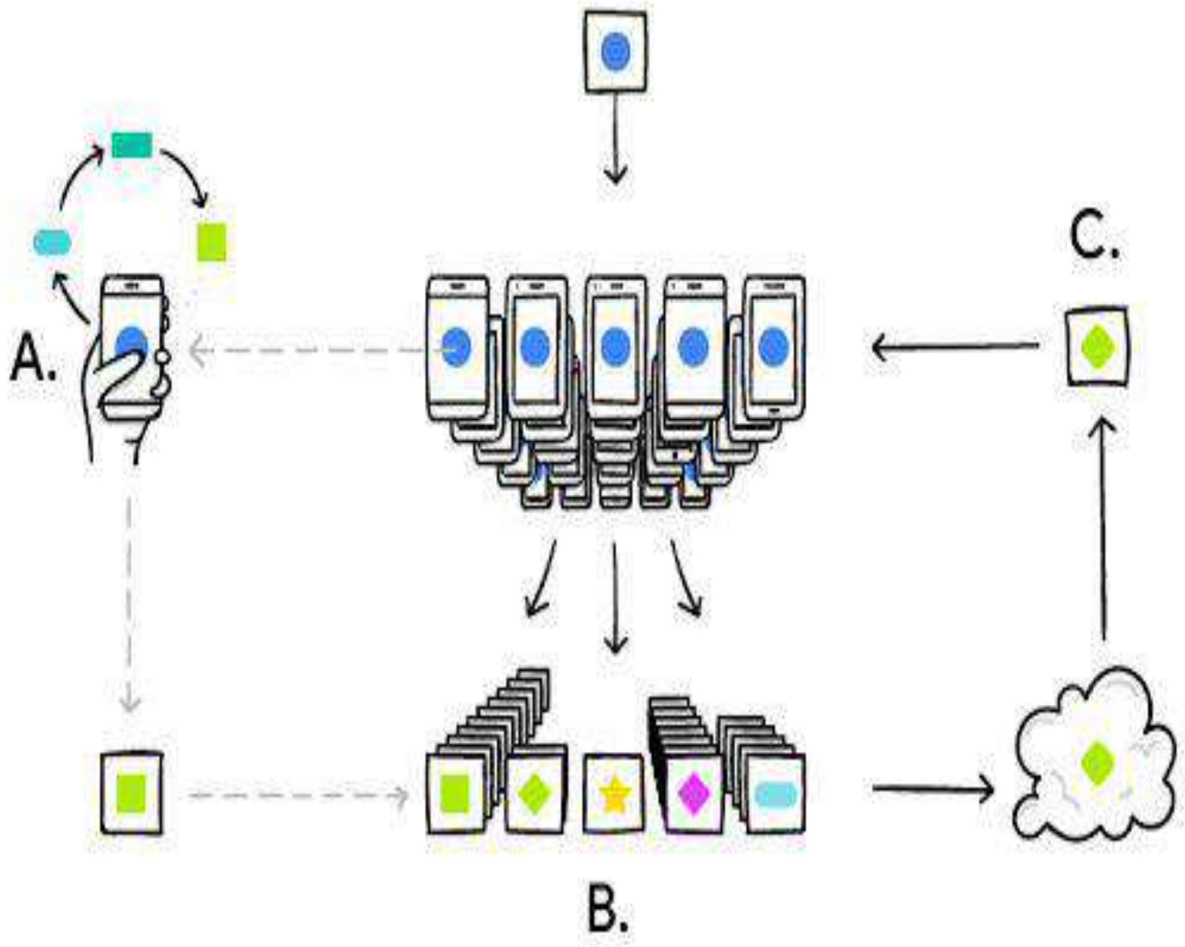
CocoaPod

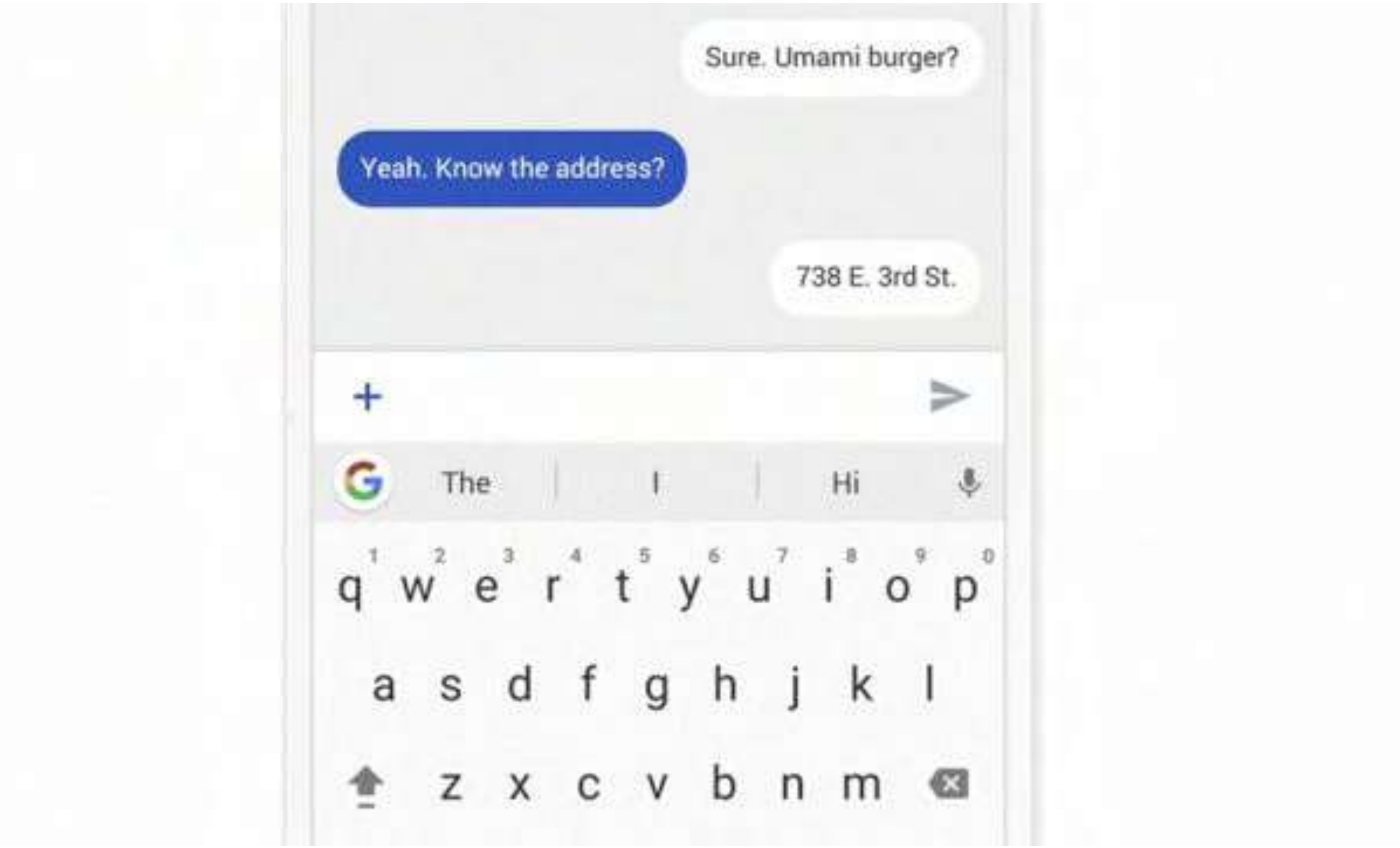
```
Podfile
  target 'MyApp'
    pod
   'TensorFlow-experimental'
```

Understand how to Code with the API

```
c.inferenceInterface =
    new TensorFlowInferenceInterface(assetManager, modelFilename);

// Copy the input data into TensorFlow.
inferenceInterface.feed(inputName, floatValues, 1, inputSize, inputSize, 3);

// Run the inference call.
inferenceInterface.run(outputNames, logStats);

// Copy the output Tensor back into the output array.
inferenceInterface.fetch(outputName, outputs);
```

Even though we have talked primarily about prediction on mobile, a new frontier is federated learning





Sure. Umami burger?

Yeah. Know the address?

738 E. 3rd St.

Federated learning in Google Keyboard
https://research.googleblog.com/2017/04/federated-learning-collaborative.html

Courses 7 - Production ML Systems

Module 5: Hybrid ML Systems

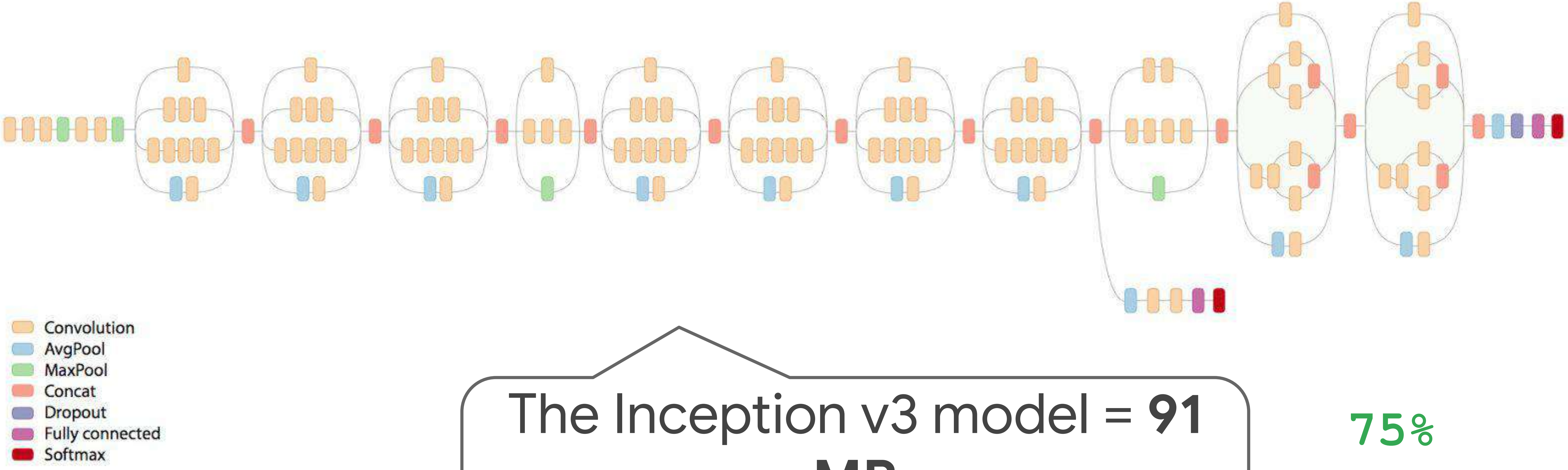
Lesson Title: **Optimizing for Mobile**

Format: Presenter

Presenter: Val

Video Name: T-PSML-O_5_l7_optimizing_for_mobile

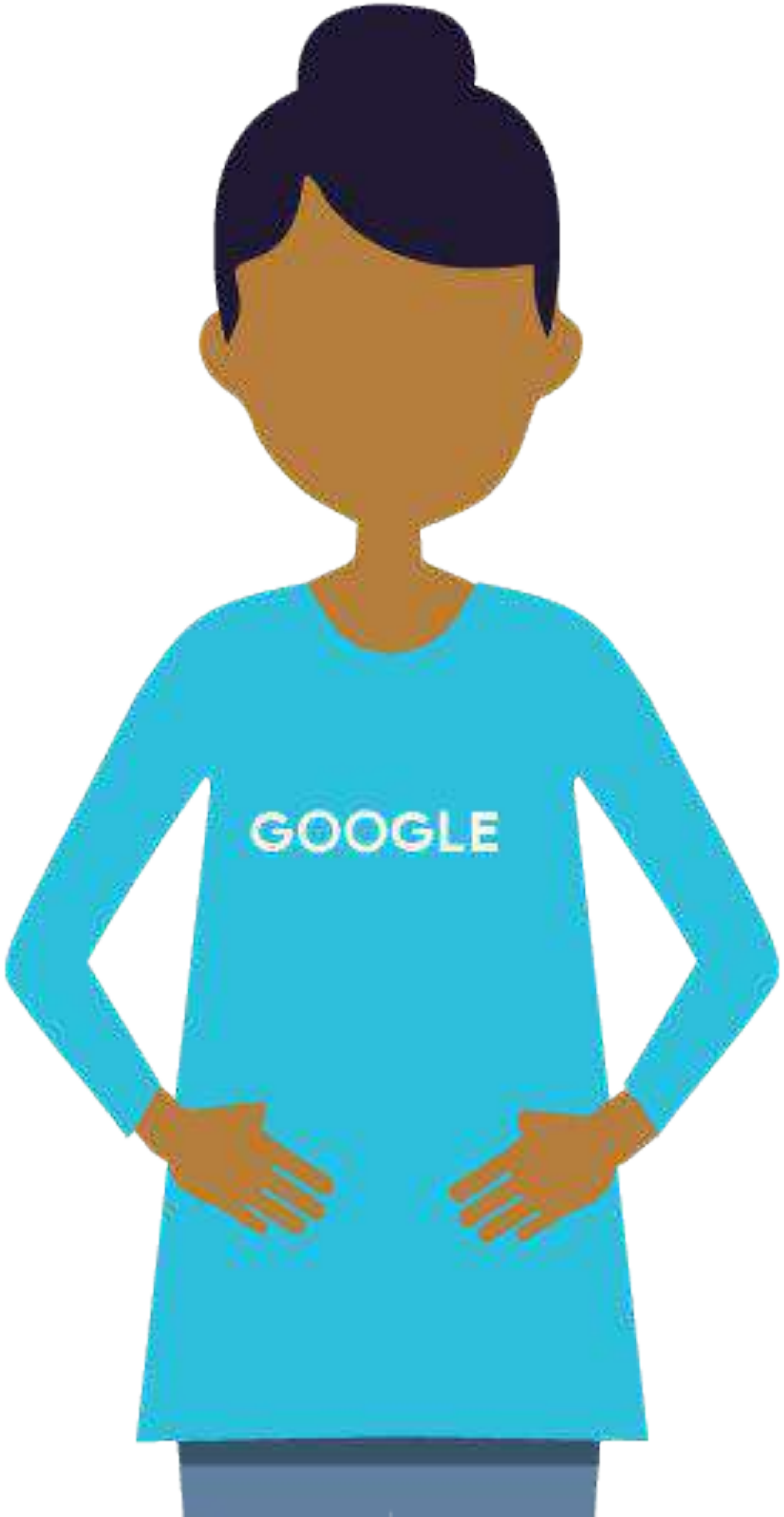
Large neural networks can be compressed



The Inception v3 model = **91 MB**

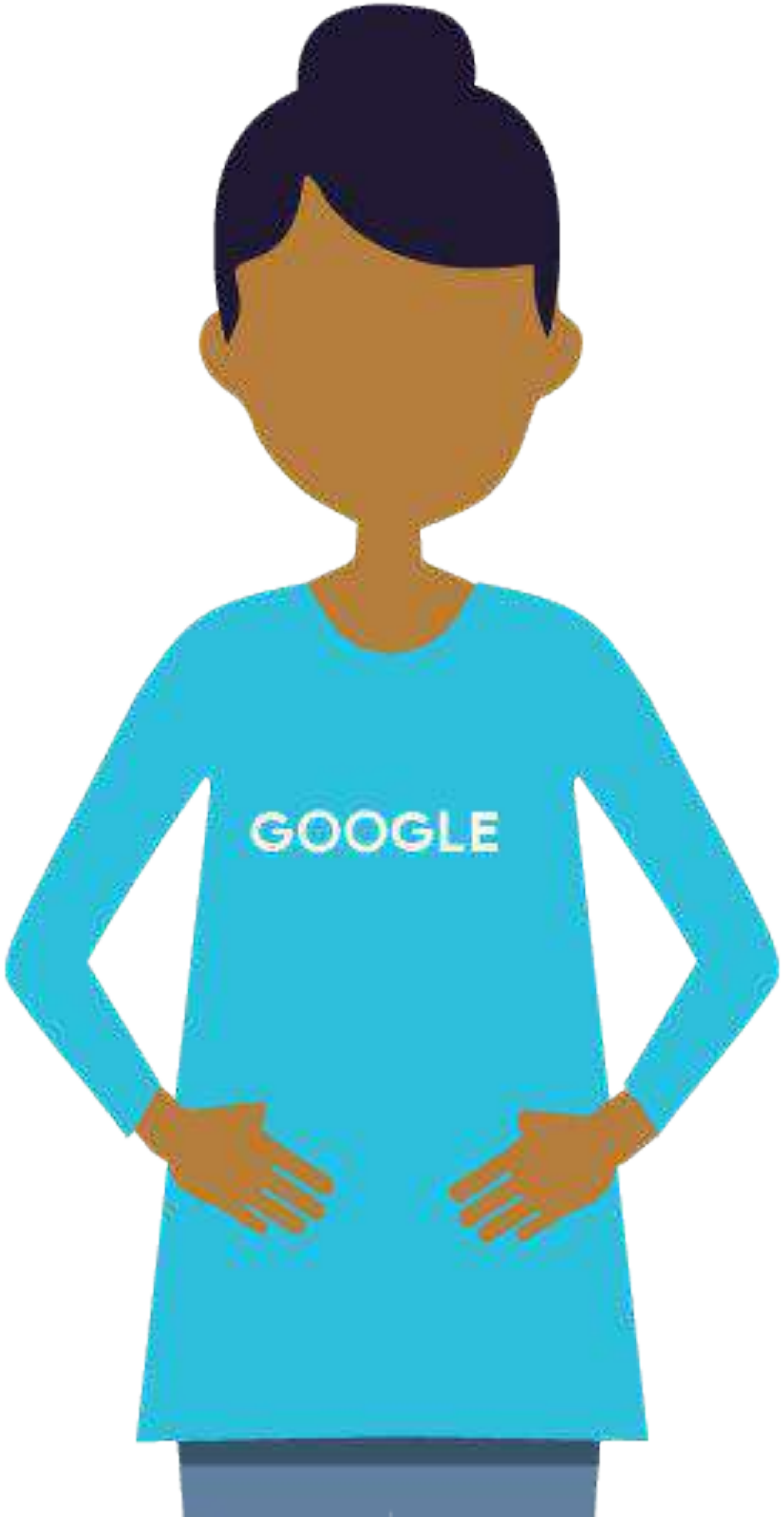
TensorFlow binary = **12 MB**

75% smaller

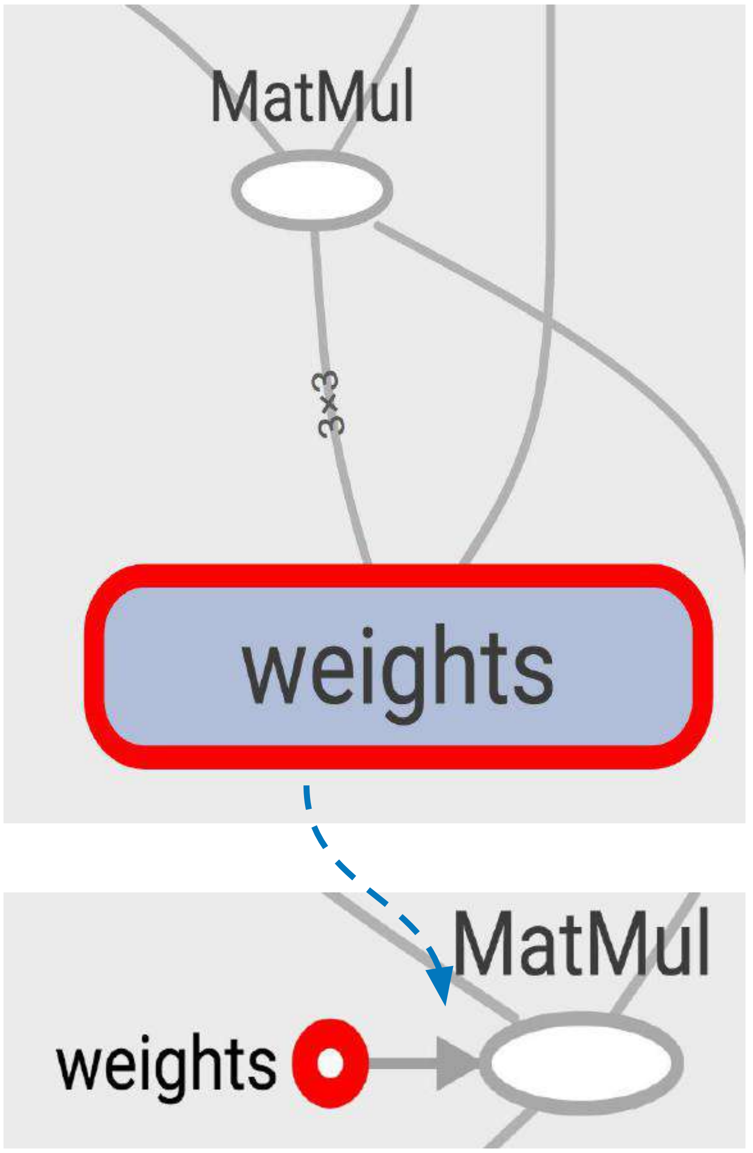


There are several methods to reduce model size

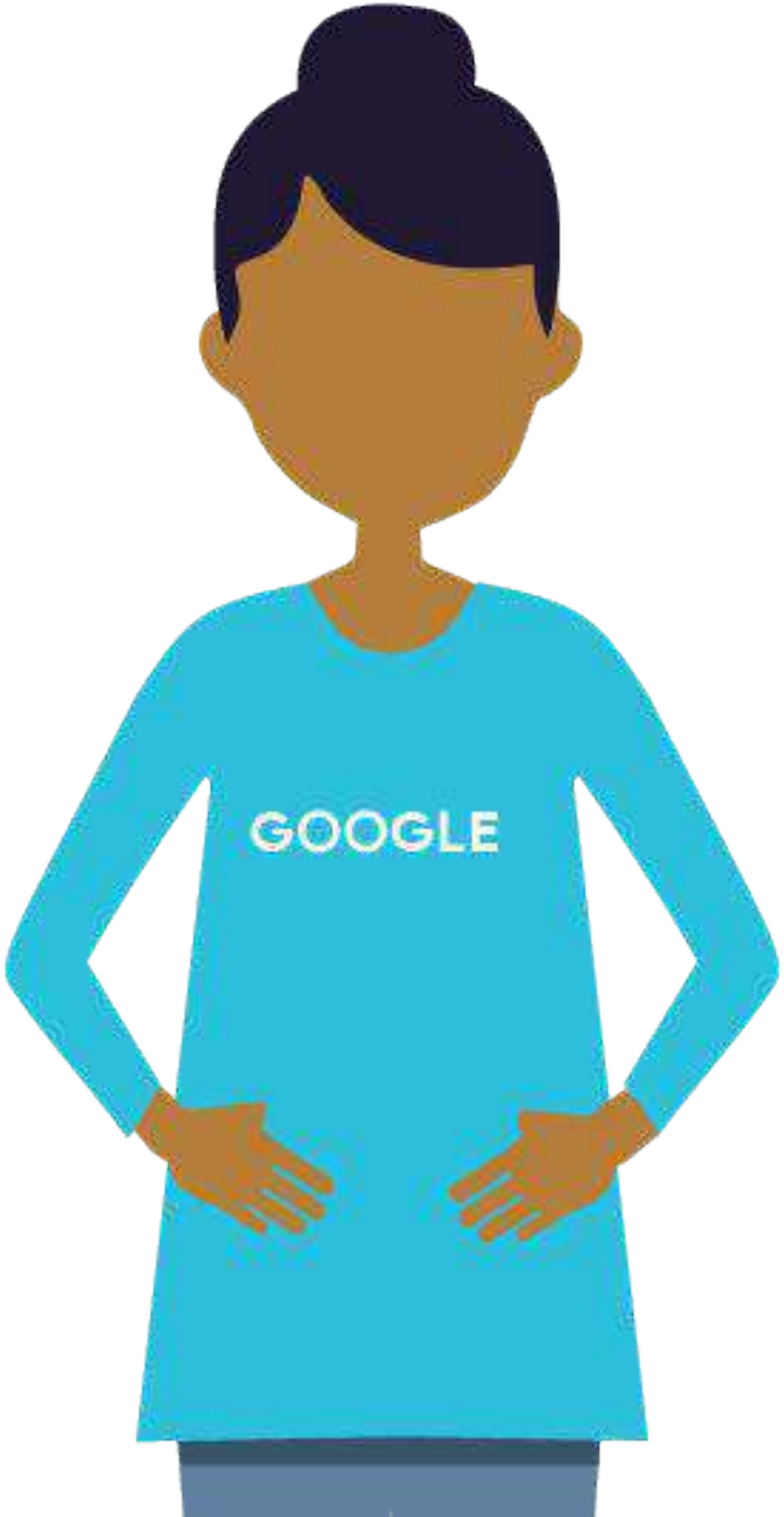
- Freeze graph
- Transform the graph
- Quantize weights and calculations



Freezing a graph can do load time optimization



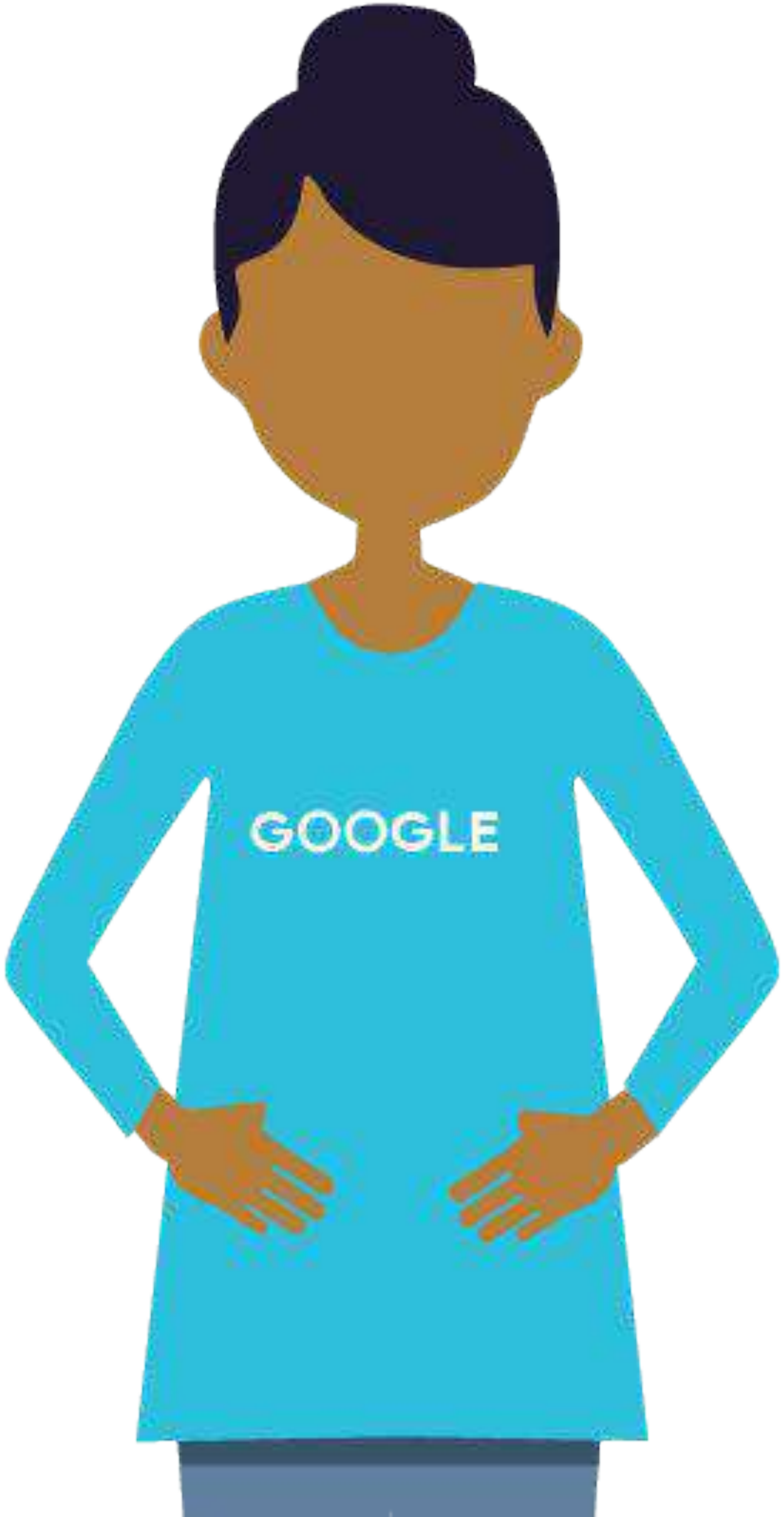
Converts variables to constants and removes checkpoints



Transform your graph to remove nodes you don’t use in prediction

strip_unused_nodes:

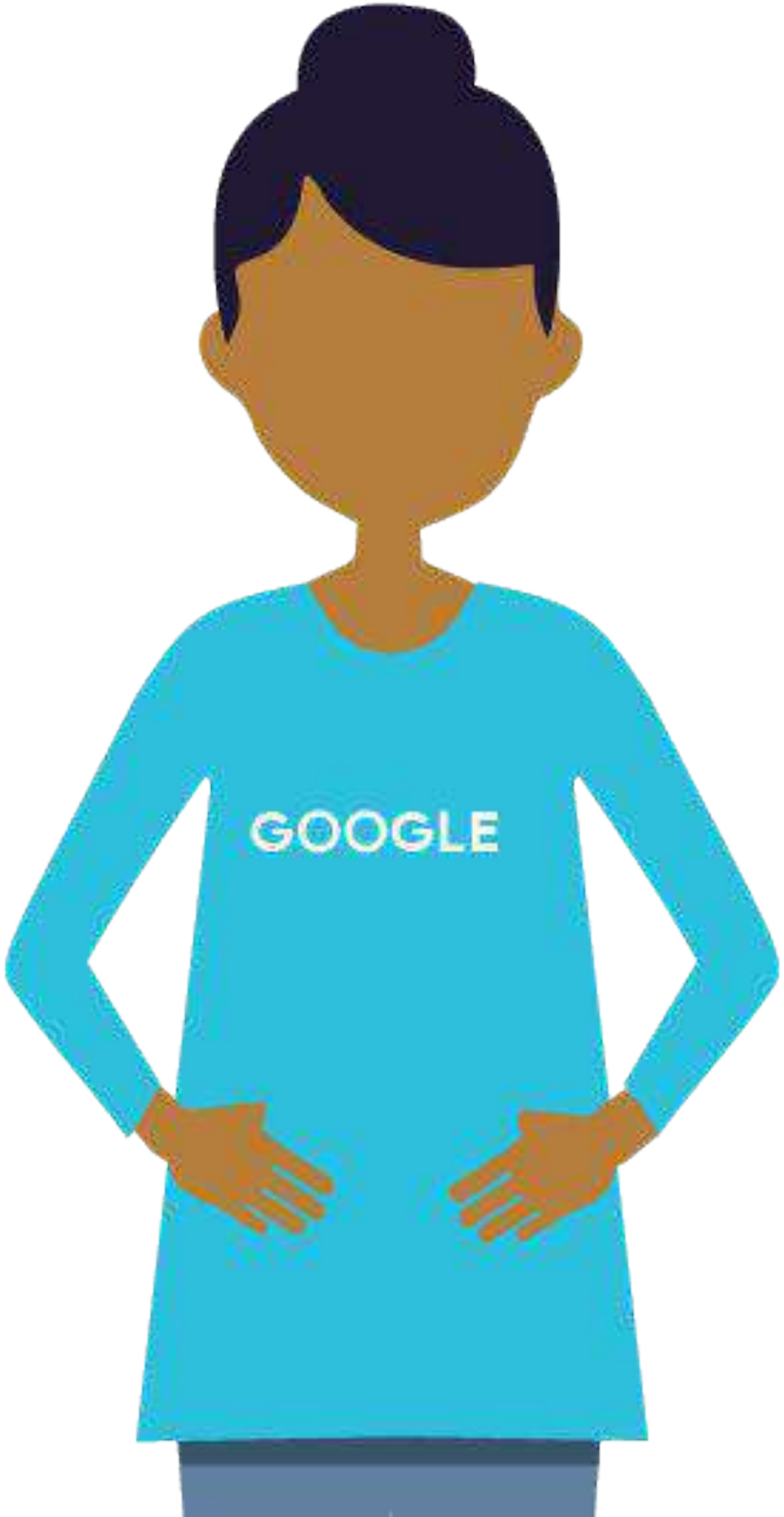
Remove training-only operations



Transform your graph to remove nodes you don't use in prediction

remove_nodes:

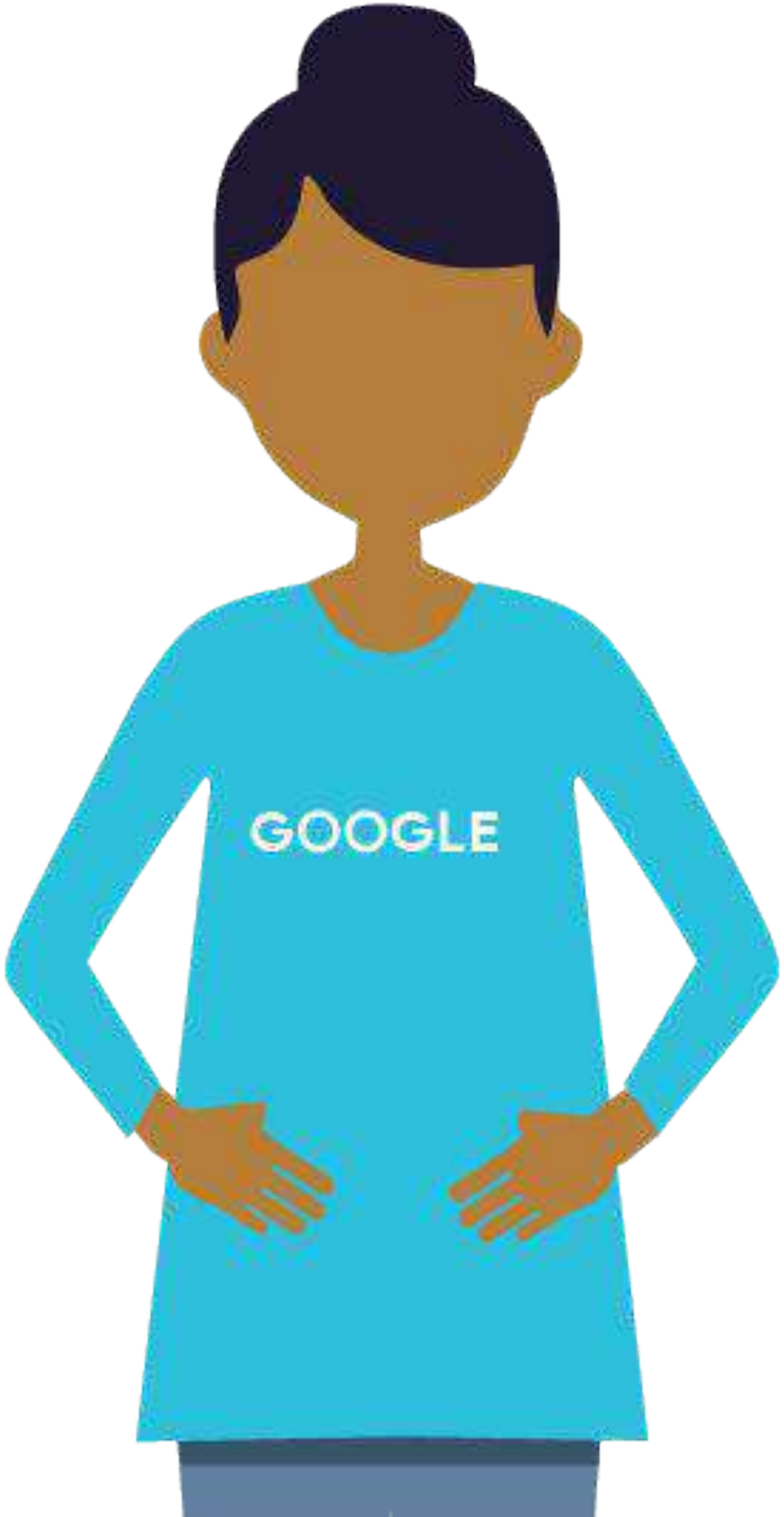
Remove debug nodes



Transform your graph to remove nodes you don't use in prediction

fold_batch_norms:

Remove Muls for batch norm



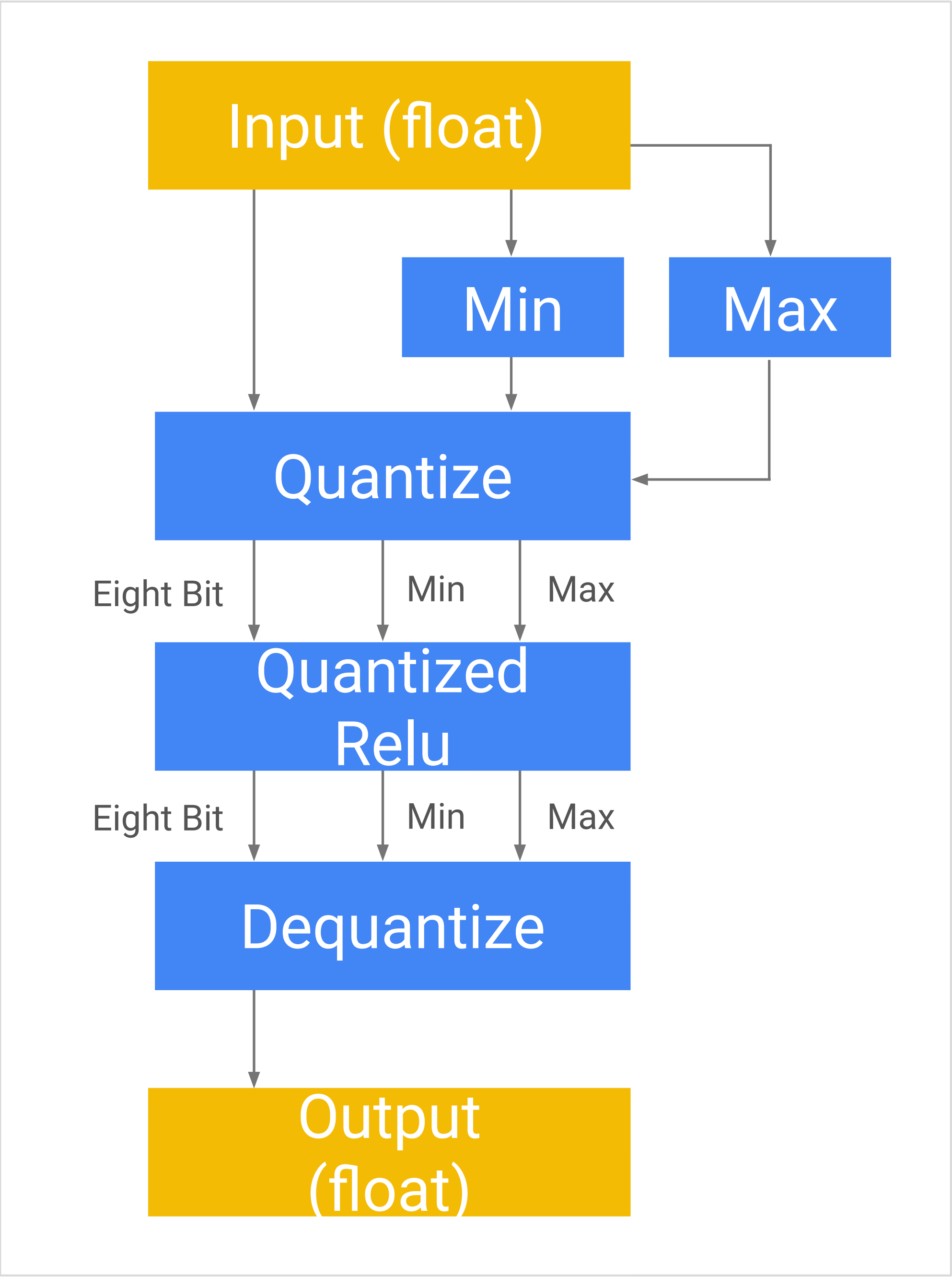
Transform your graph to remove nodes you don't use in prediction

quantize_weights
quantize_nodes

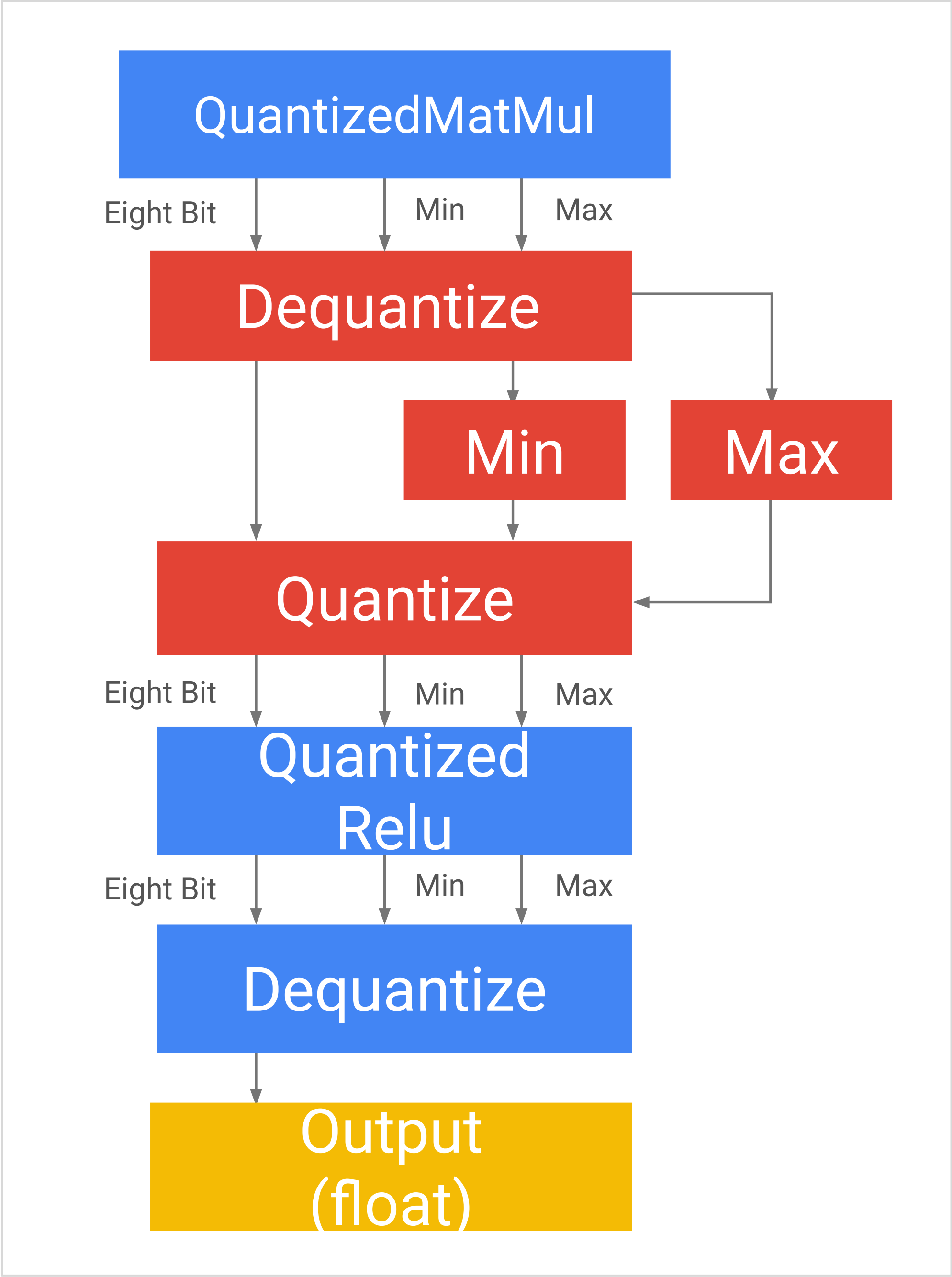
Add quantization

Quantizing weights and calculations boosts performance

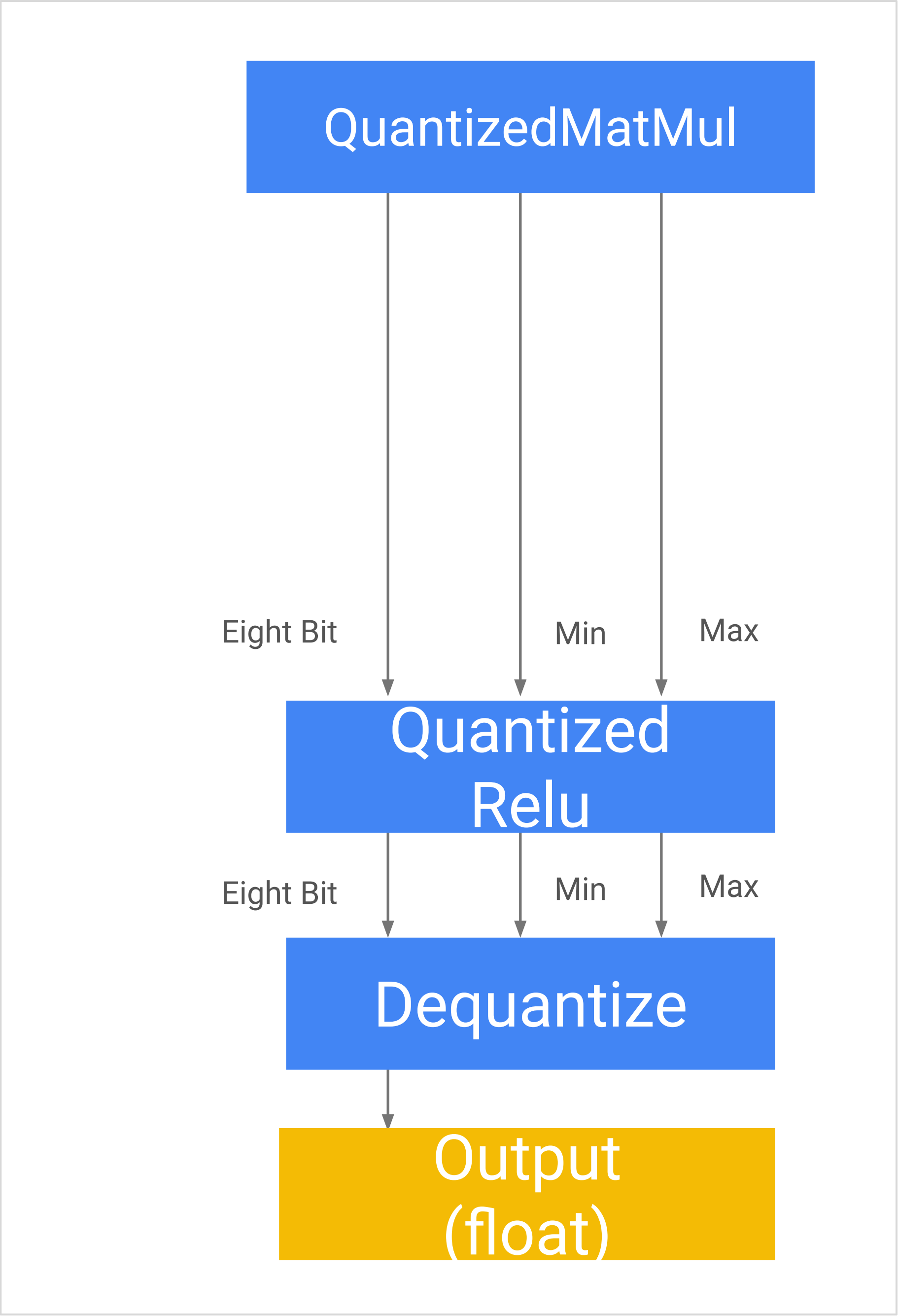
Stage



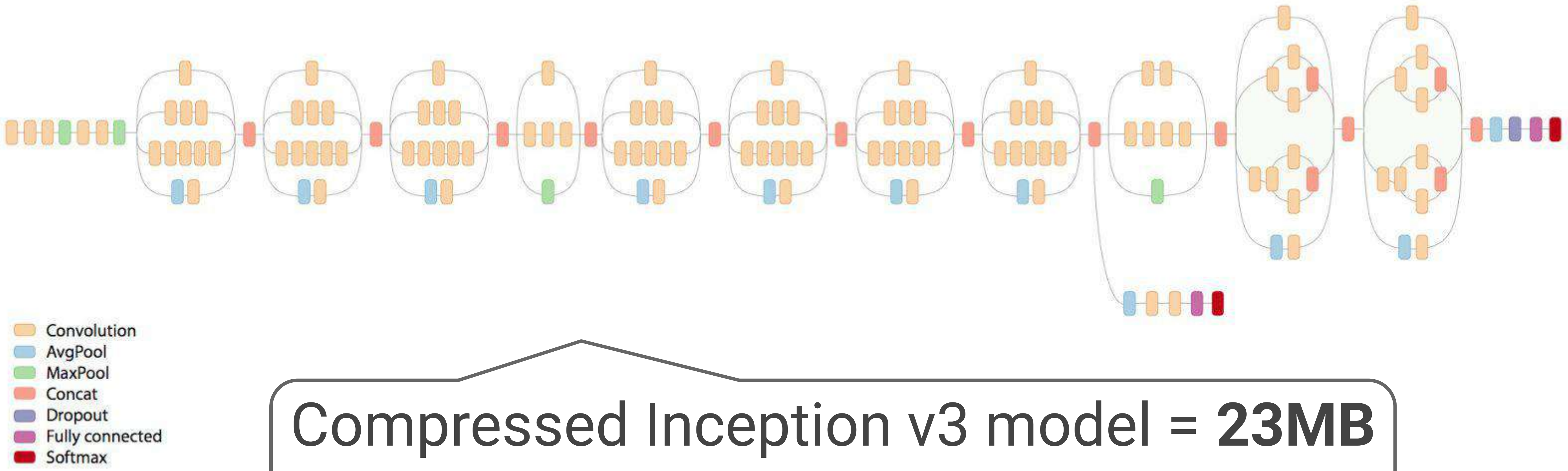
Stage



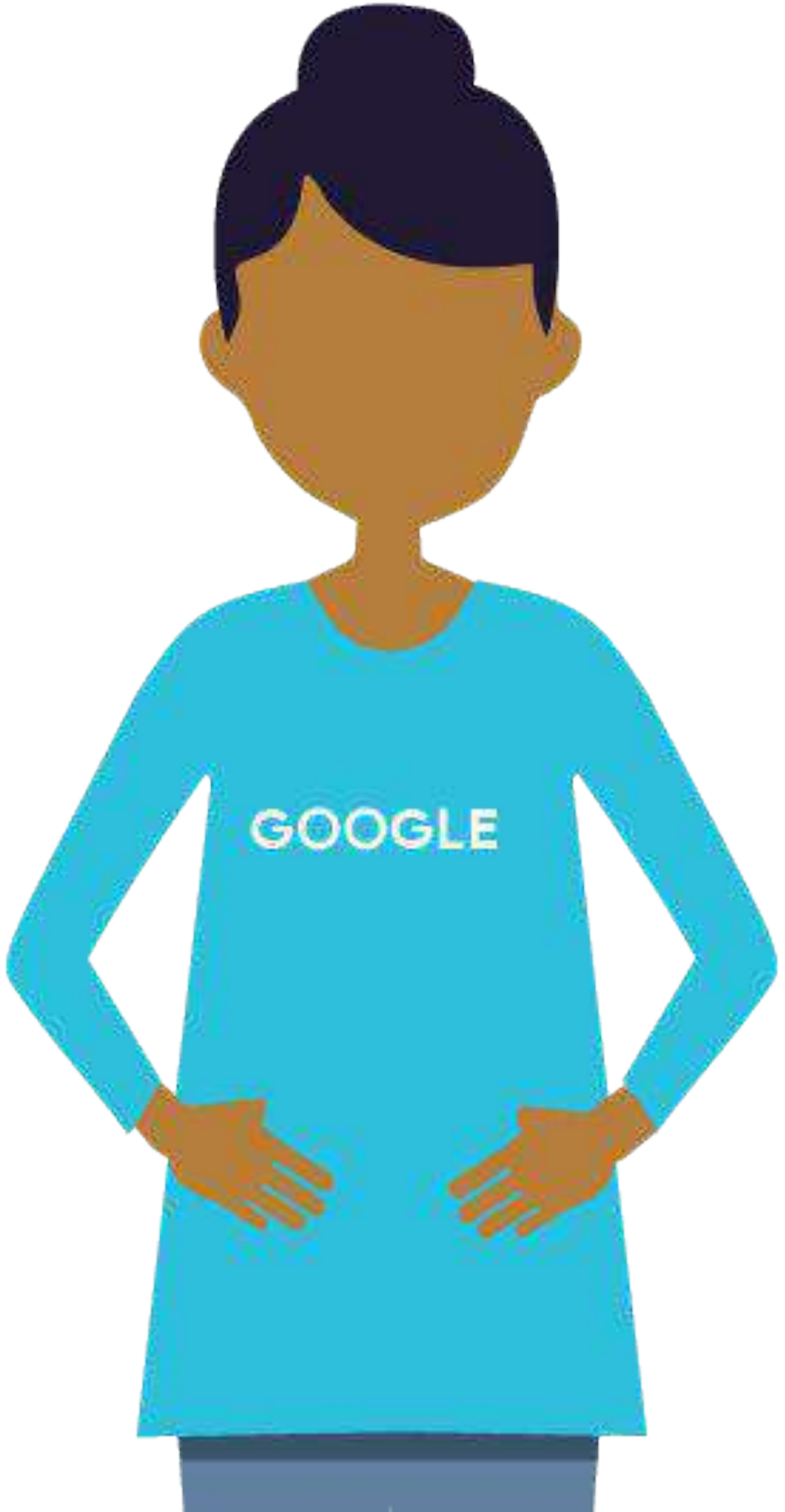
Stage



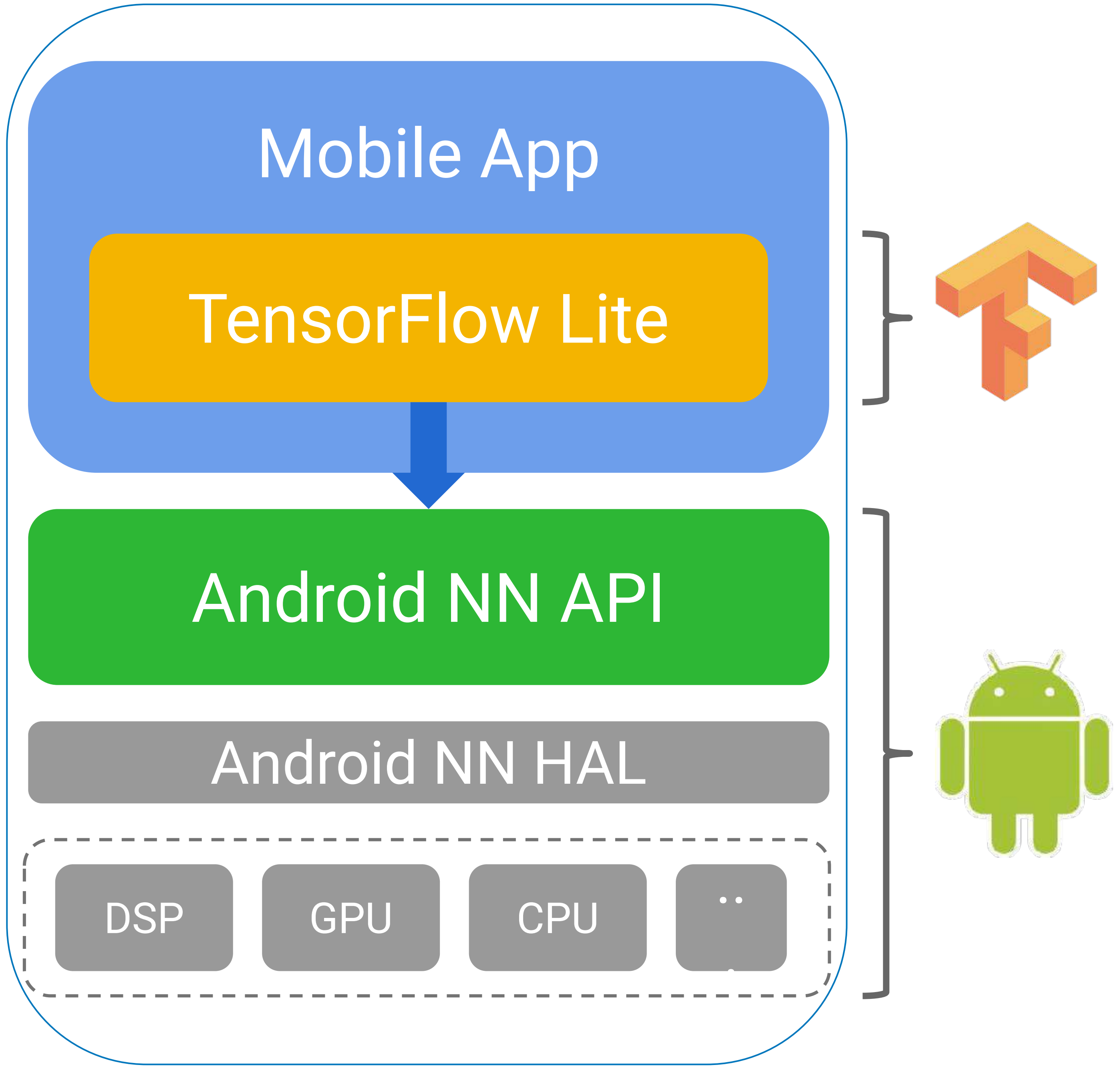
After these optimizations, the neural network is 75% smaller



Compressed Inception v3 model = **23MB**
TensorFlow binary = **1.5MB**



TensorFlow Lite is optimized for mobile apps



Courses 7 - Production ML Systems

Module 5: Hybrid ML Systems

Lesson Title: **Summary**

Format: Presenter

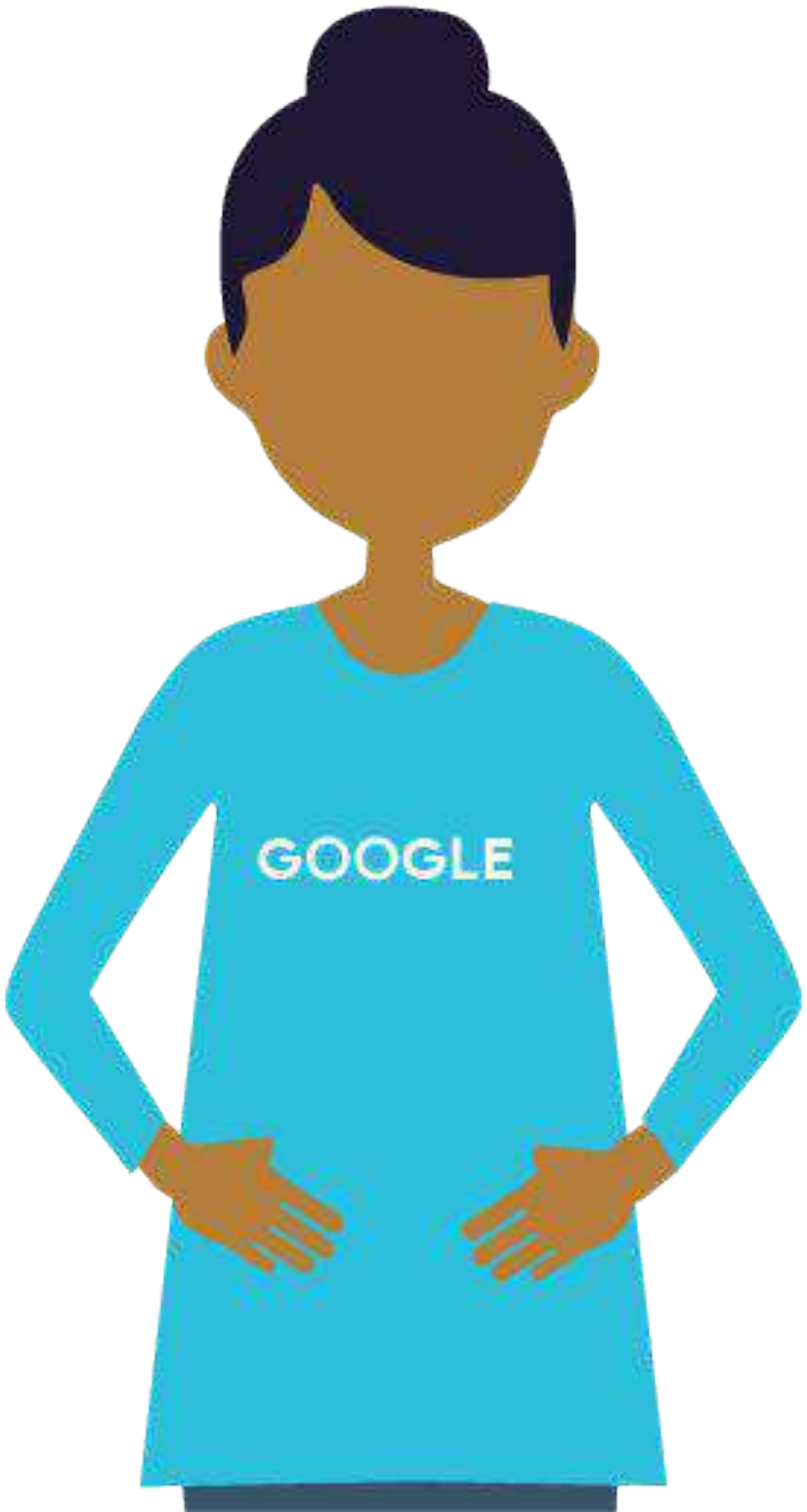
Presenter: Val

Video Name: T-PSML-O_5_l8_summary

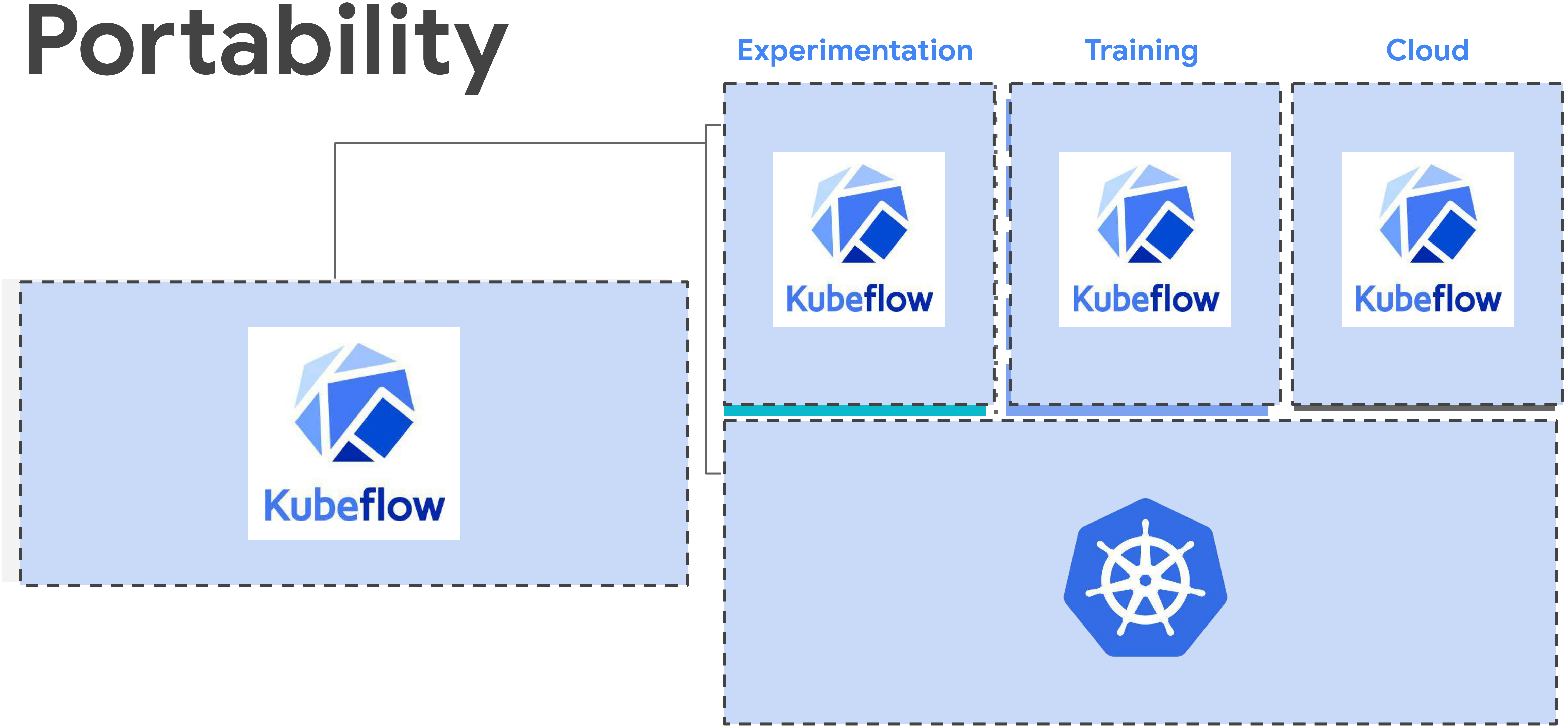
Summary

Build hybrid cloud machine learning models

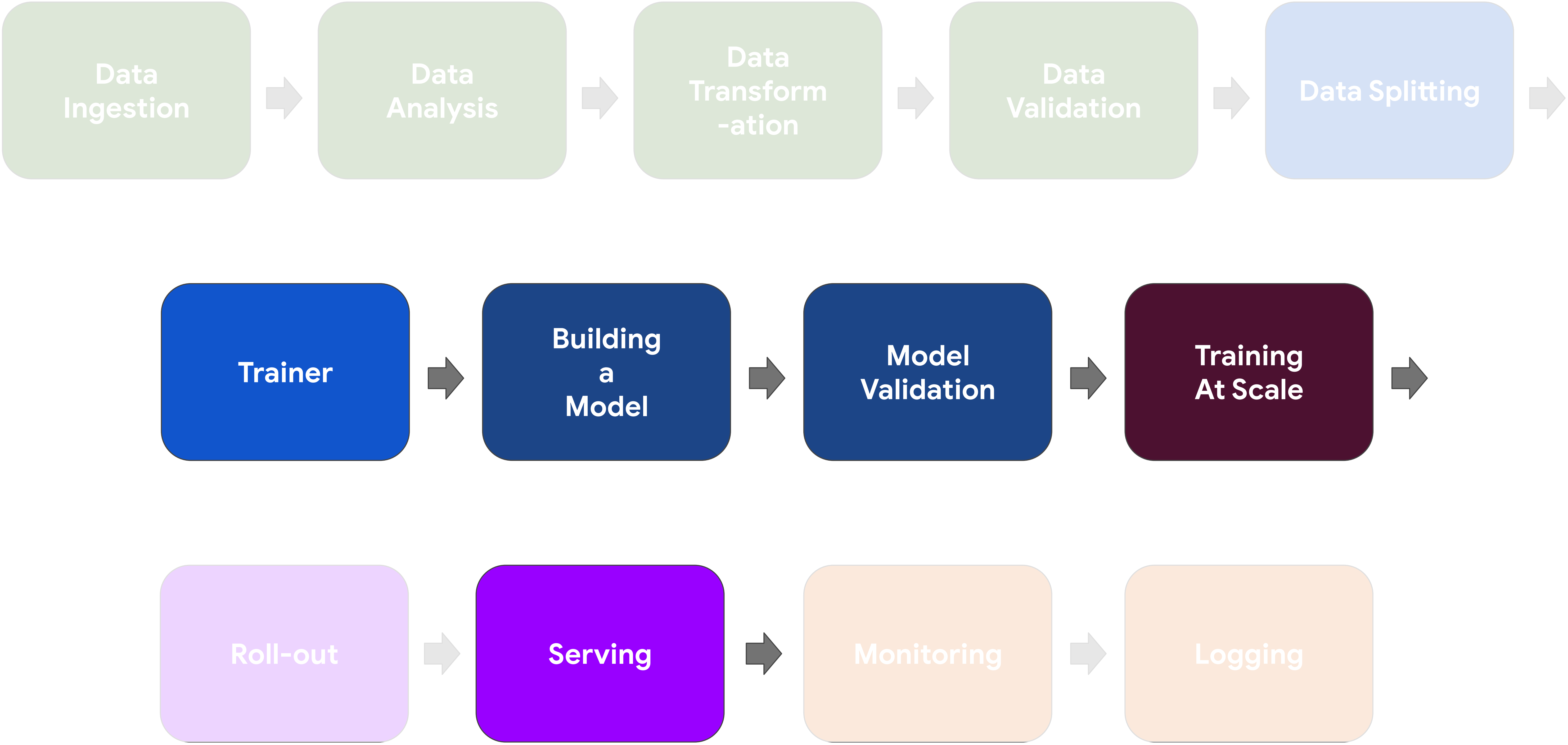
Optimize TensorFlow graphs for mobile

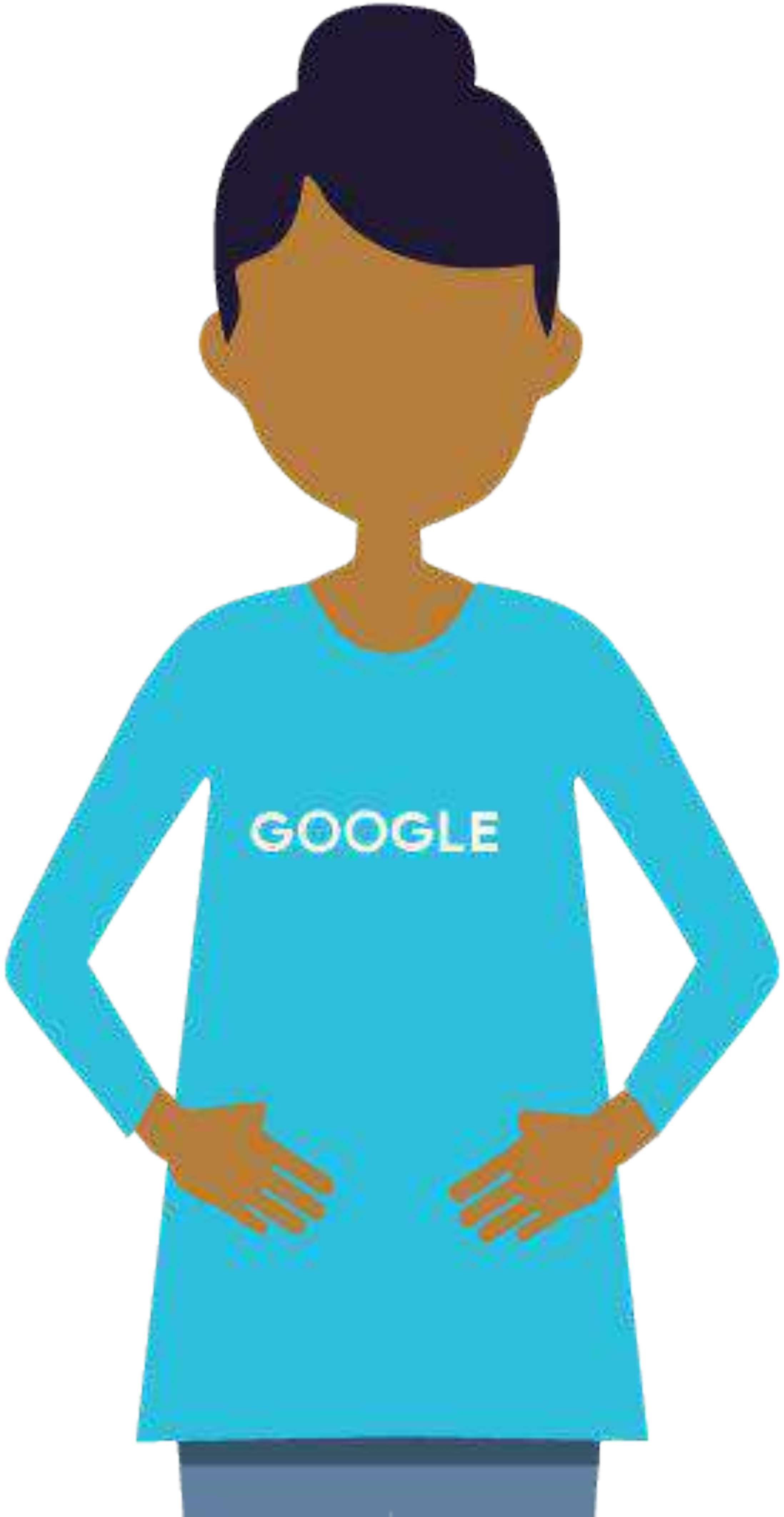


Portability

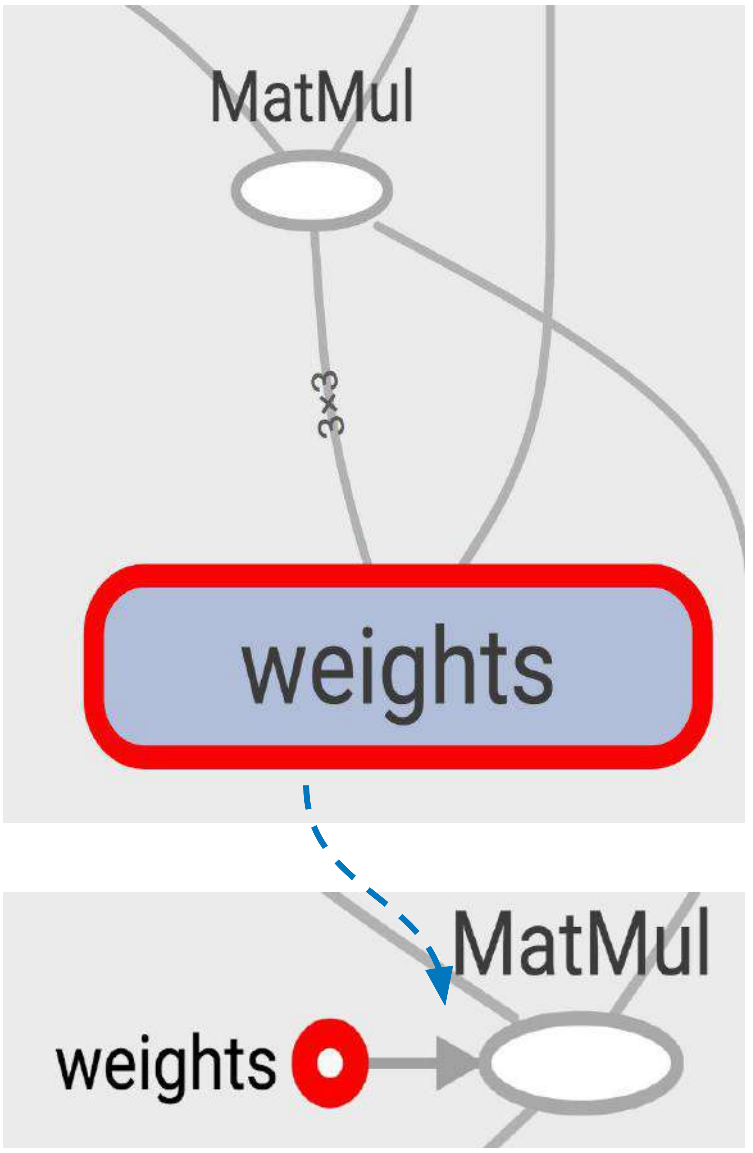


What's in the box?

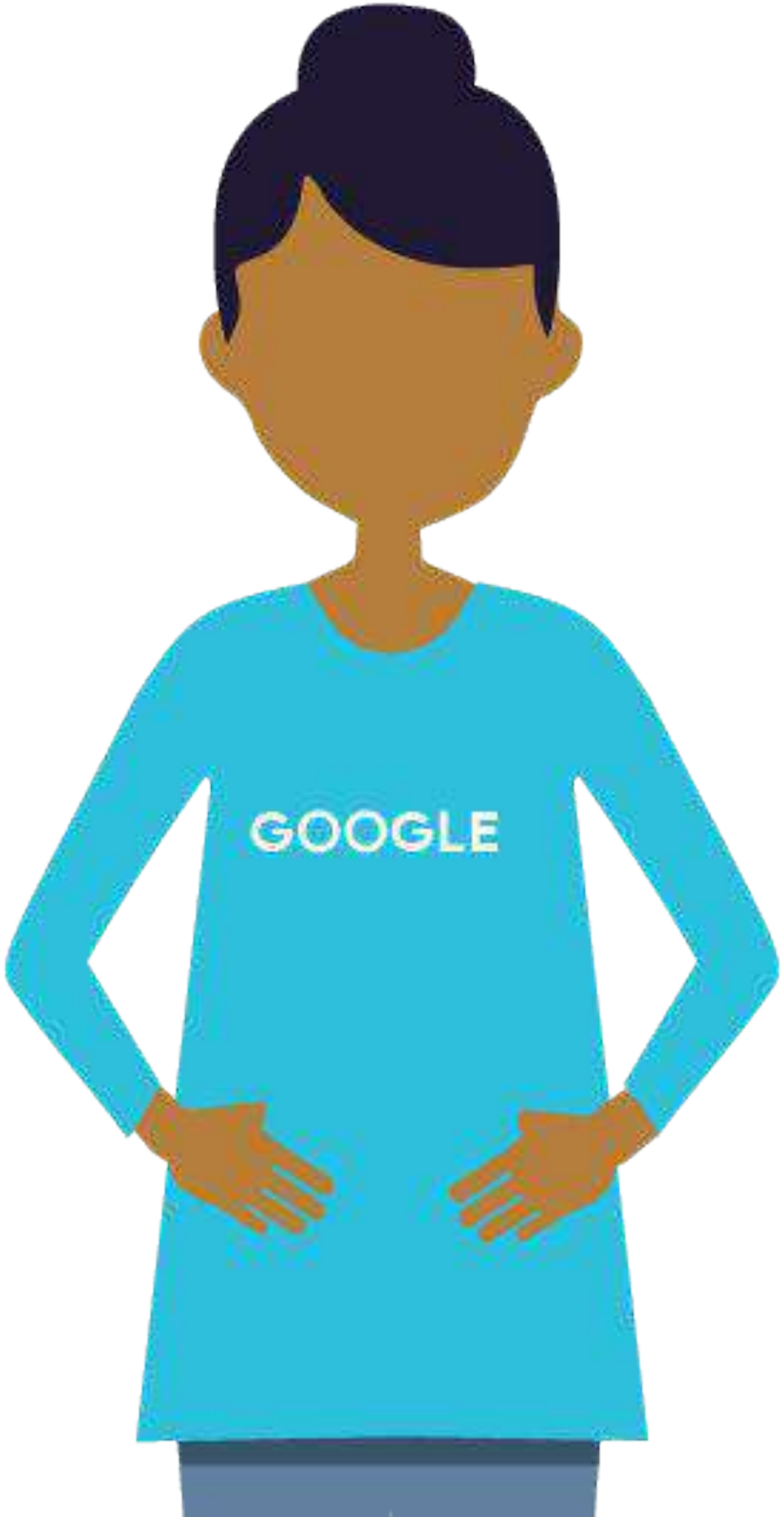




Freezing a graph can do load time optimization



Converts variables to constants and removes checkpoints



Transform your graph to remove nodes you don't use in prediction

quantize_weights
quantize_nodes

Add quantization