

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

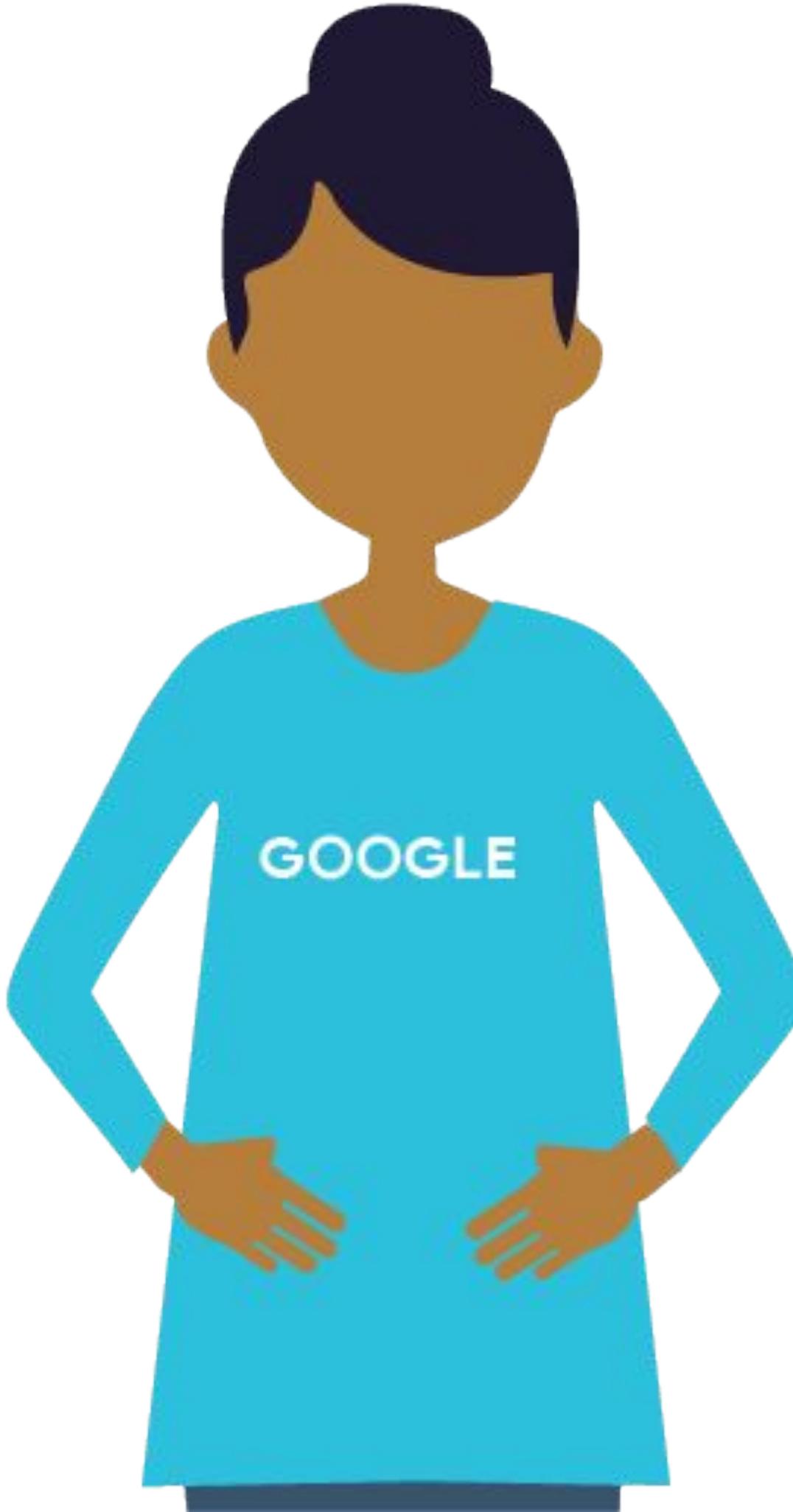
Title Safe >

< Action Safe

Learn how to...

Build hybrid cloud machine
learning models

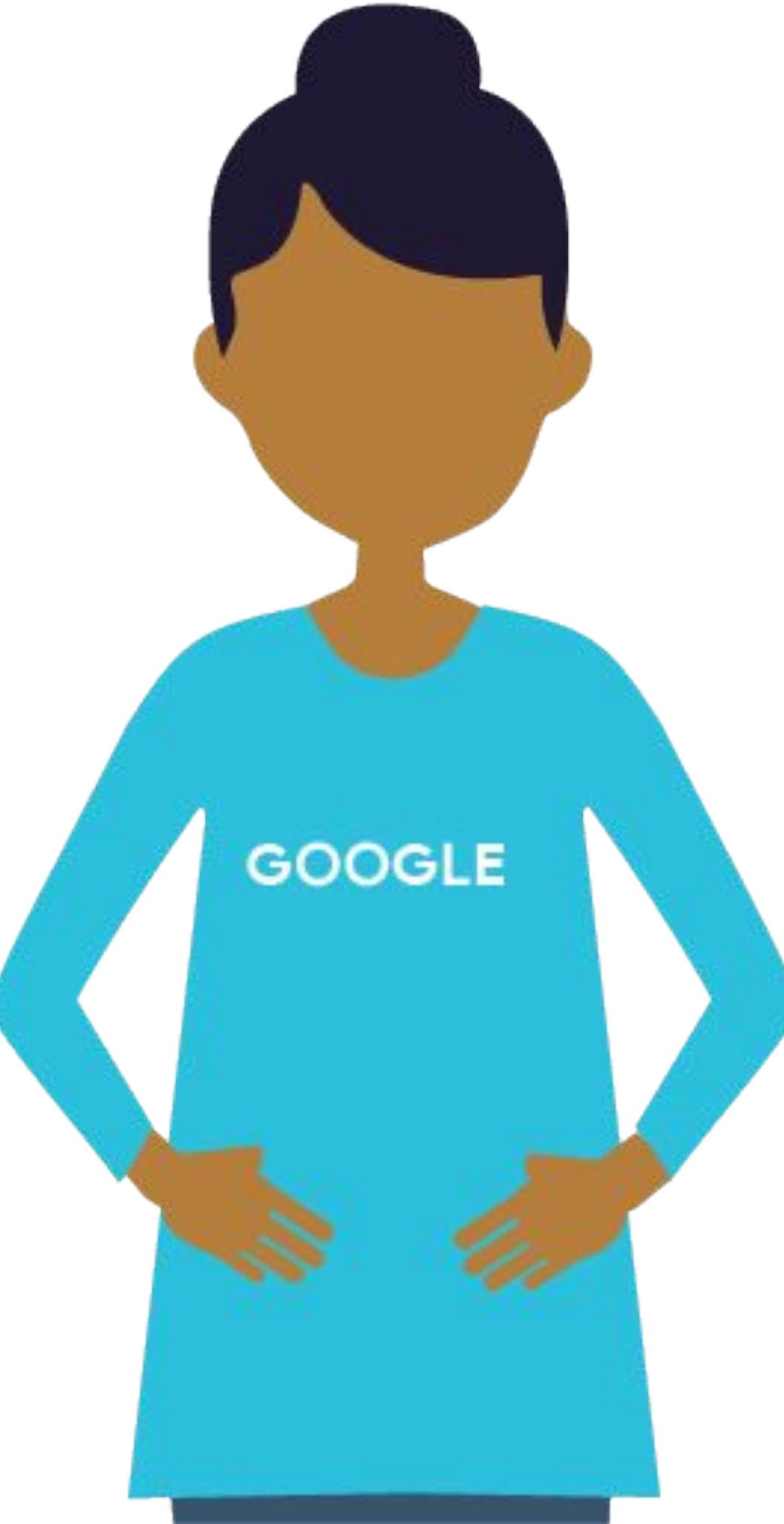
Optimize TensorFlow graphs for
mobile

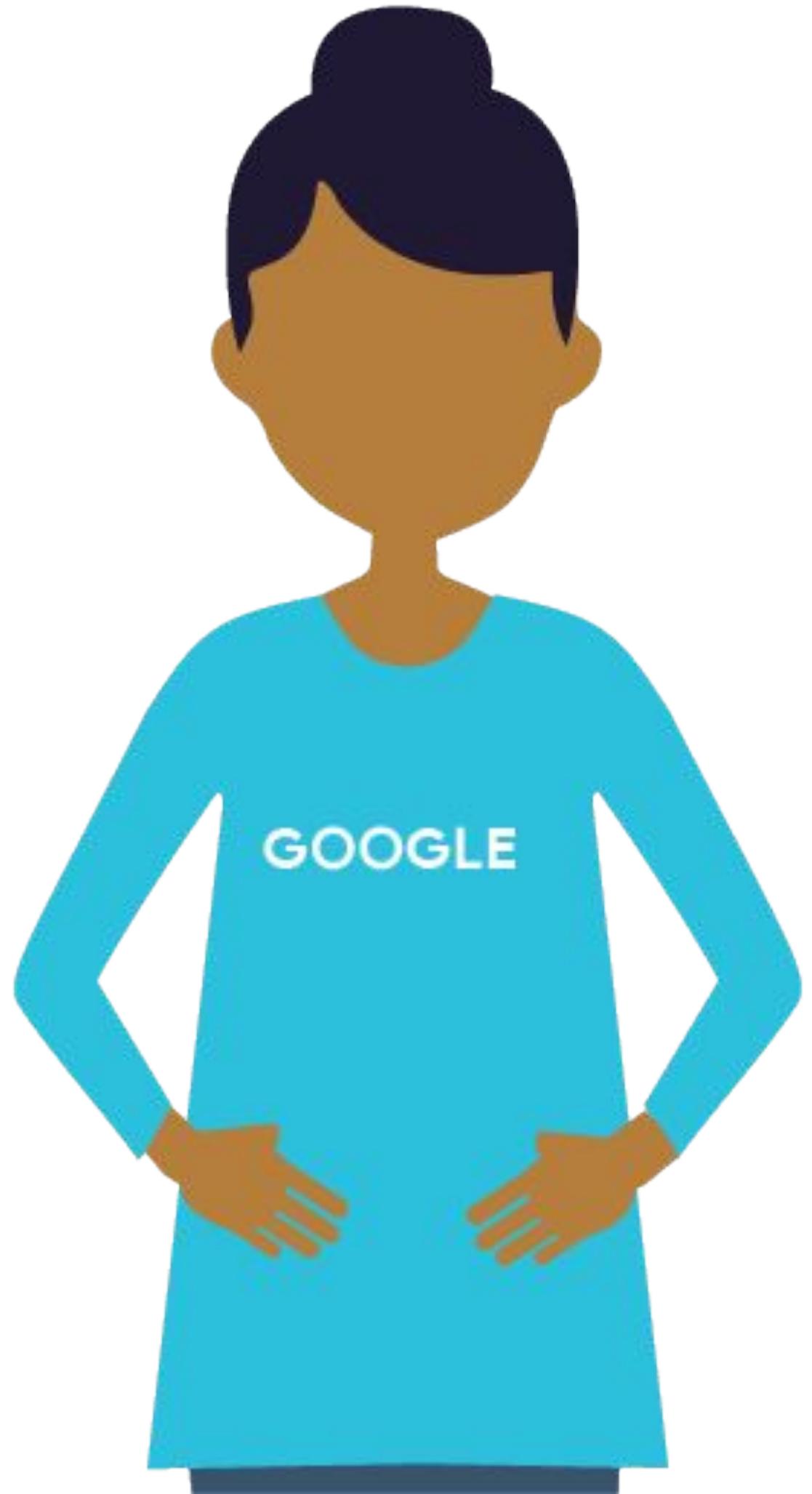


Agenda

Kubeflow for hybrid cloud

Optimizing TensorFlow for
mobile





Choose from
ready-made ML models



Vision

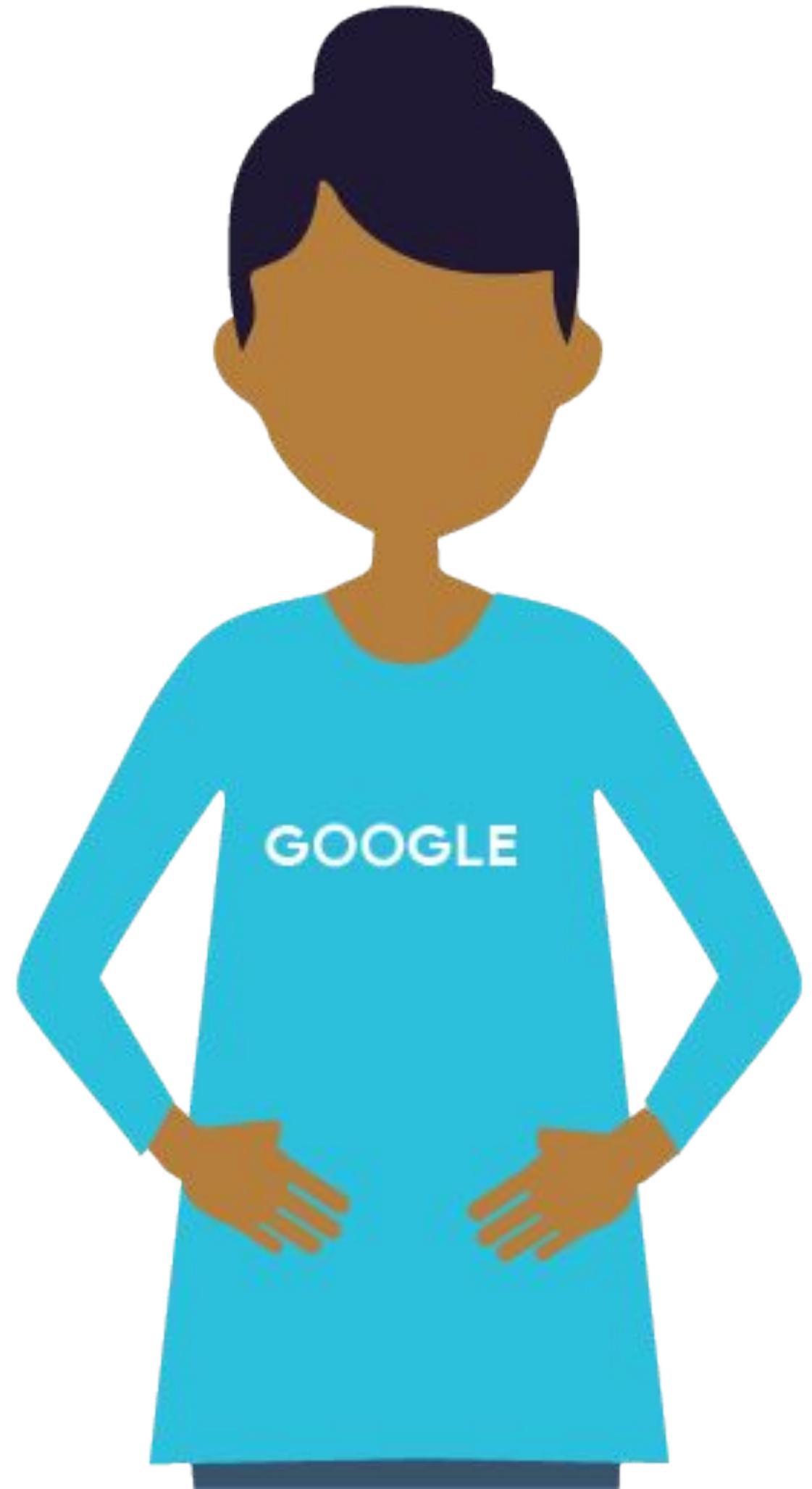


Translation

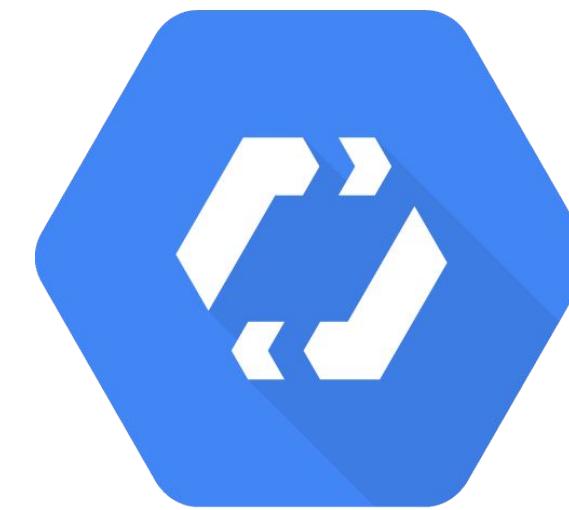


Speech BETA
Natural
Language

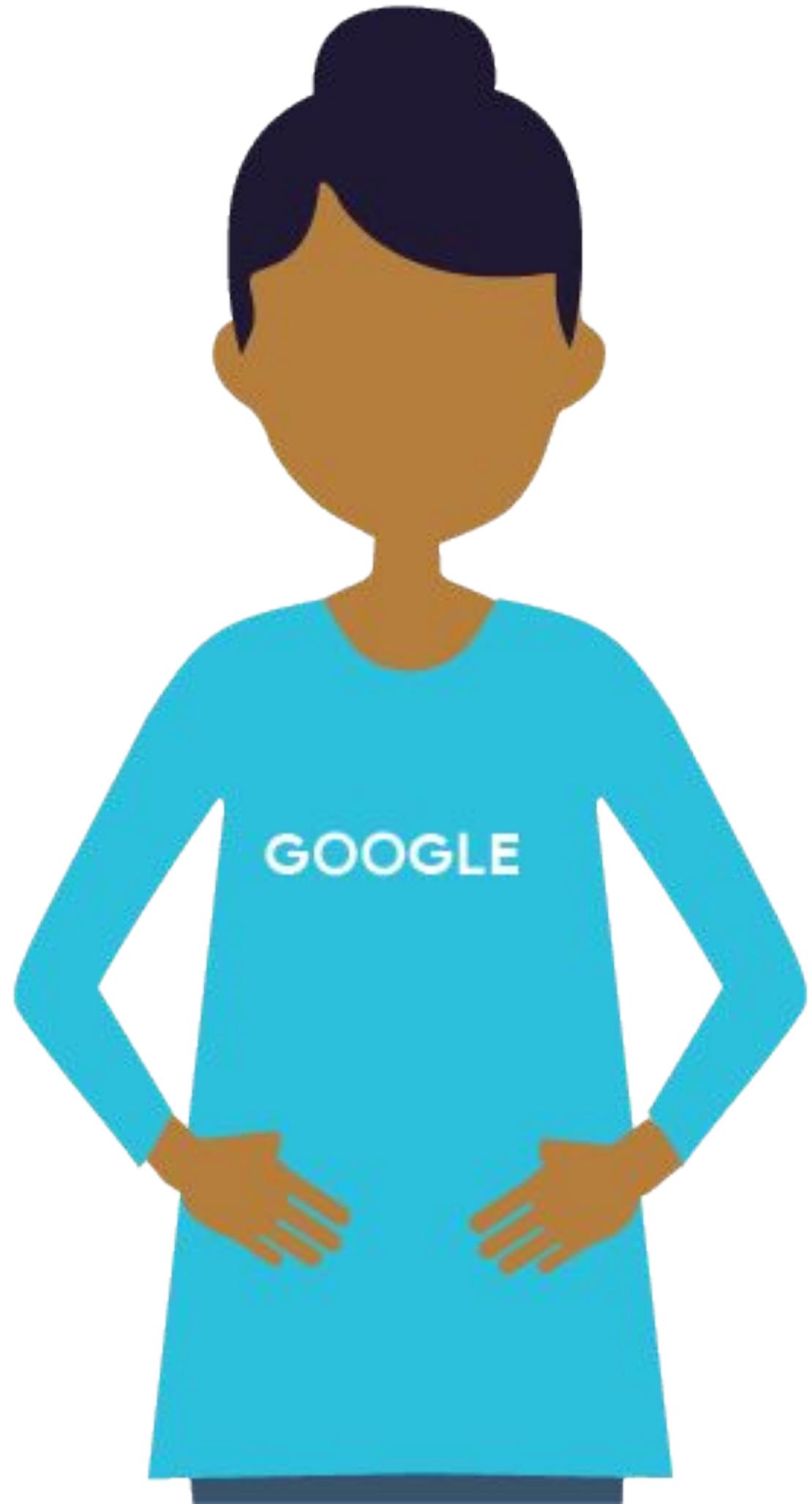




Customize ready-made
ML models

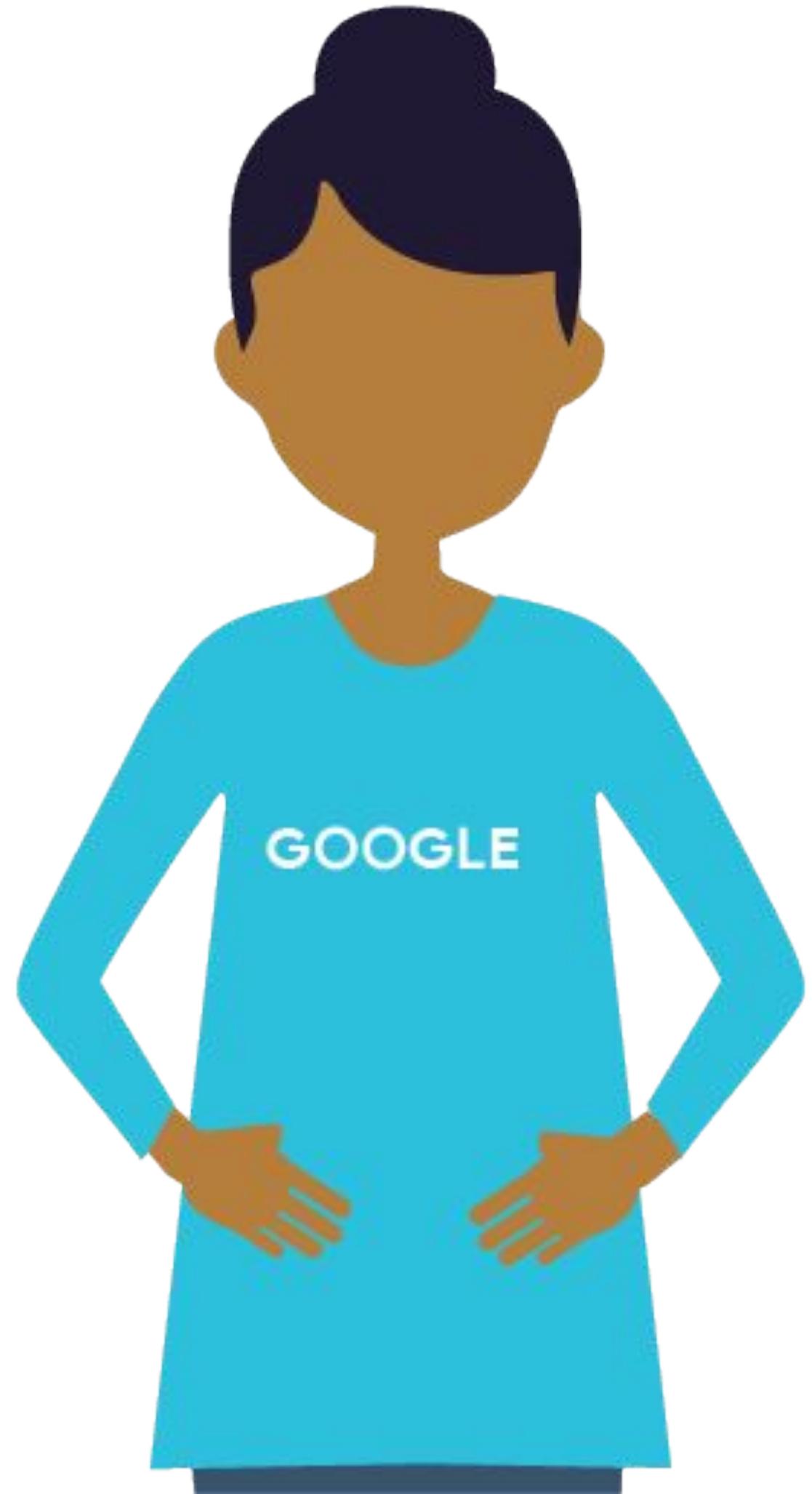


Auto-ML



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

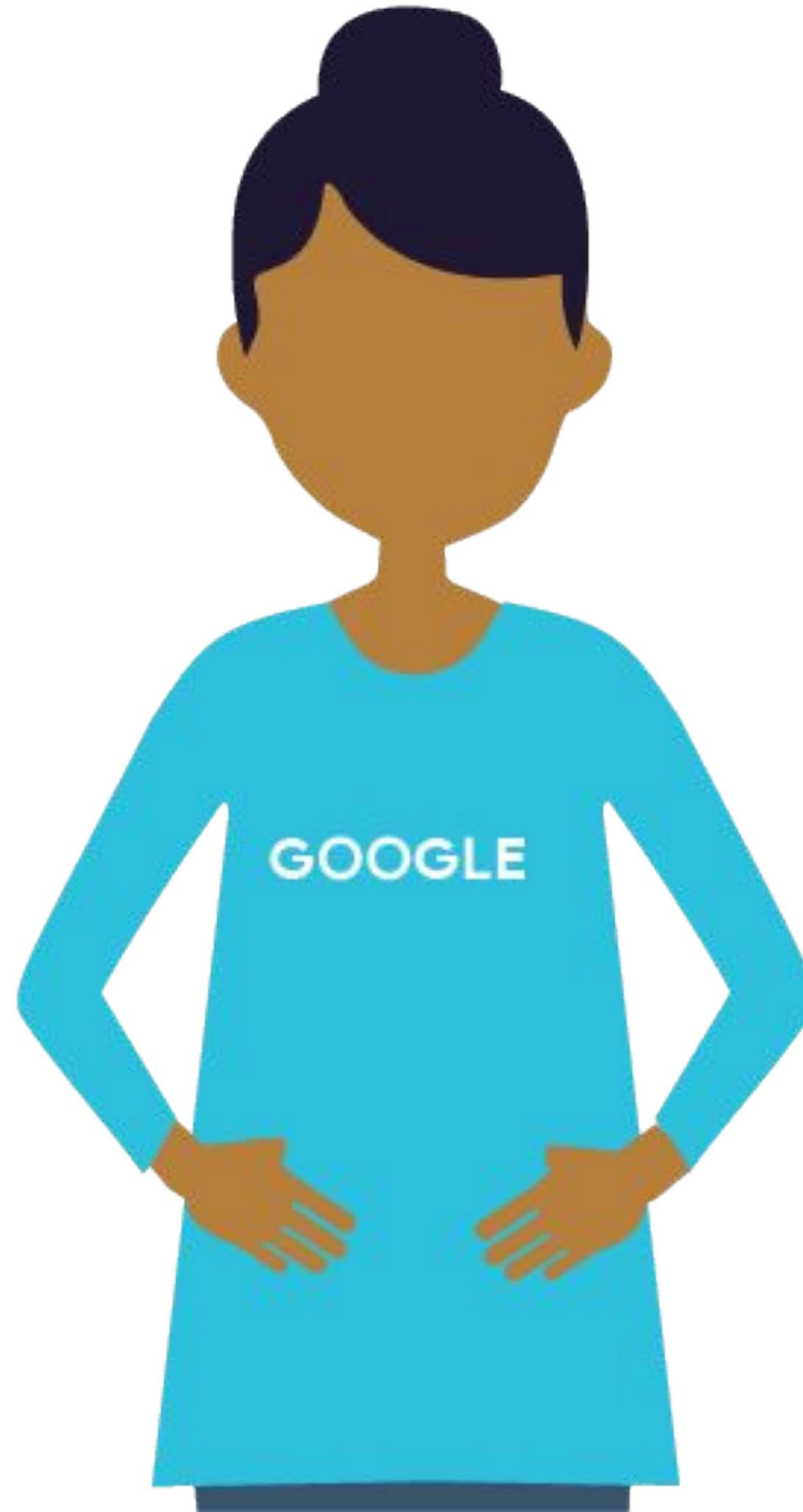




ML runtimes in a cloud-native environment

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



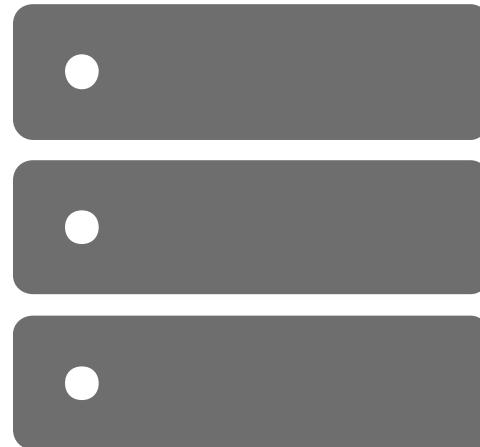
You may not be able to do machine learning
solely on Google Cloud

Tied to On-Premise
Infrastructure

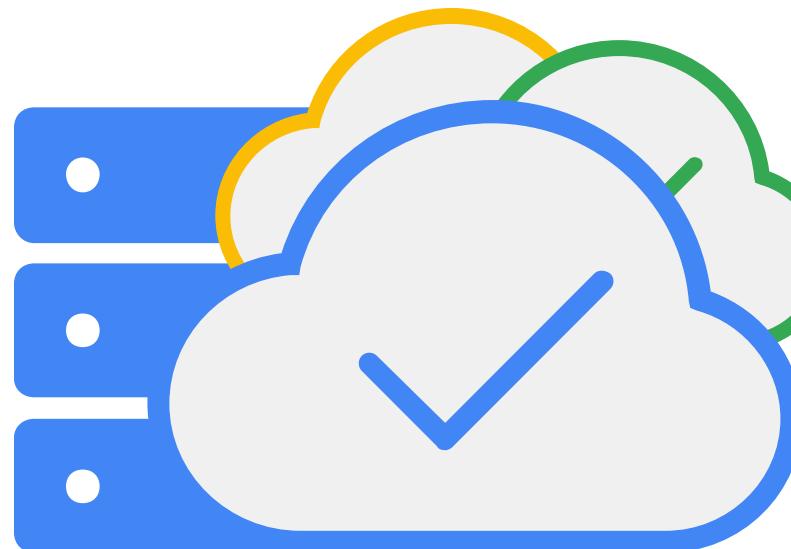
-
-
-

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

Tied to On-Premise
Infrastructure

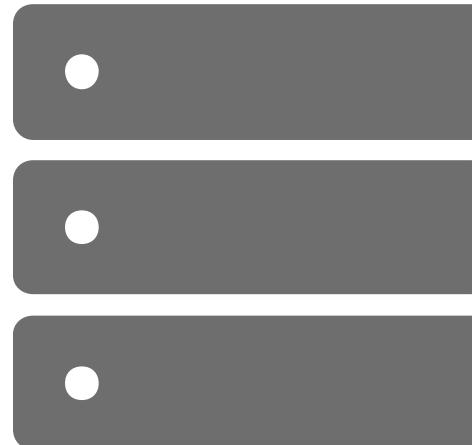


Multi Cloud System
Architecture

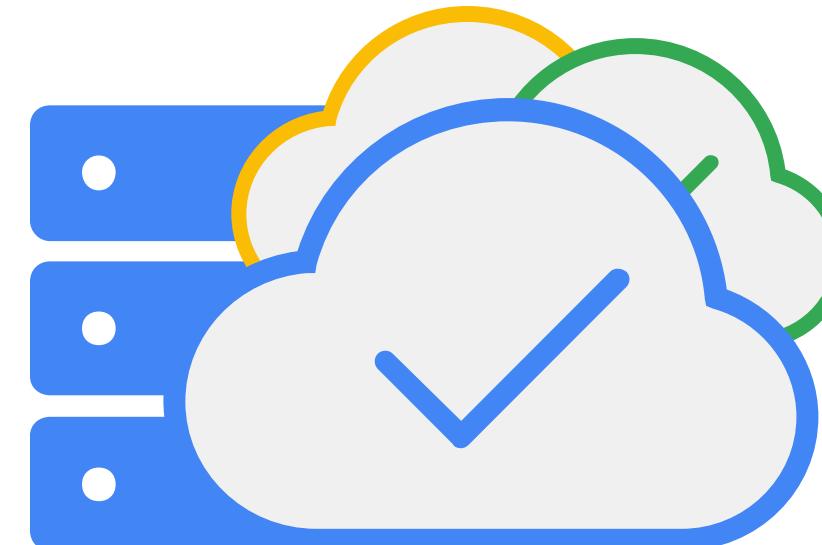


You may not be able to do machine learning solely on Google Cloud

Tied to On-Premise
Infrastructure



Multi Cloud System
Architecture



Running ML on the edge

Android / iOS

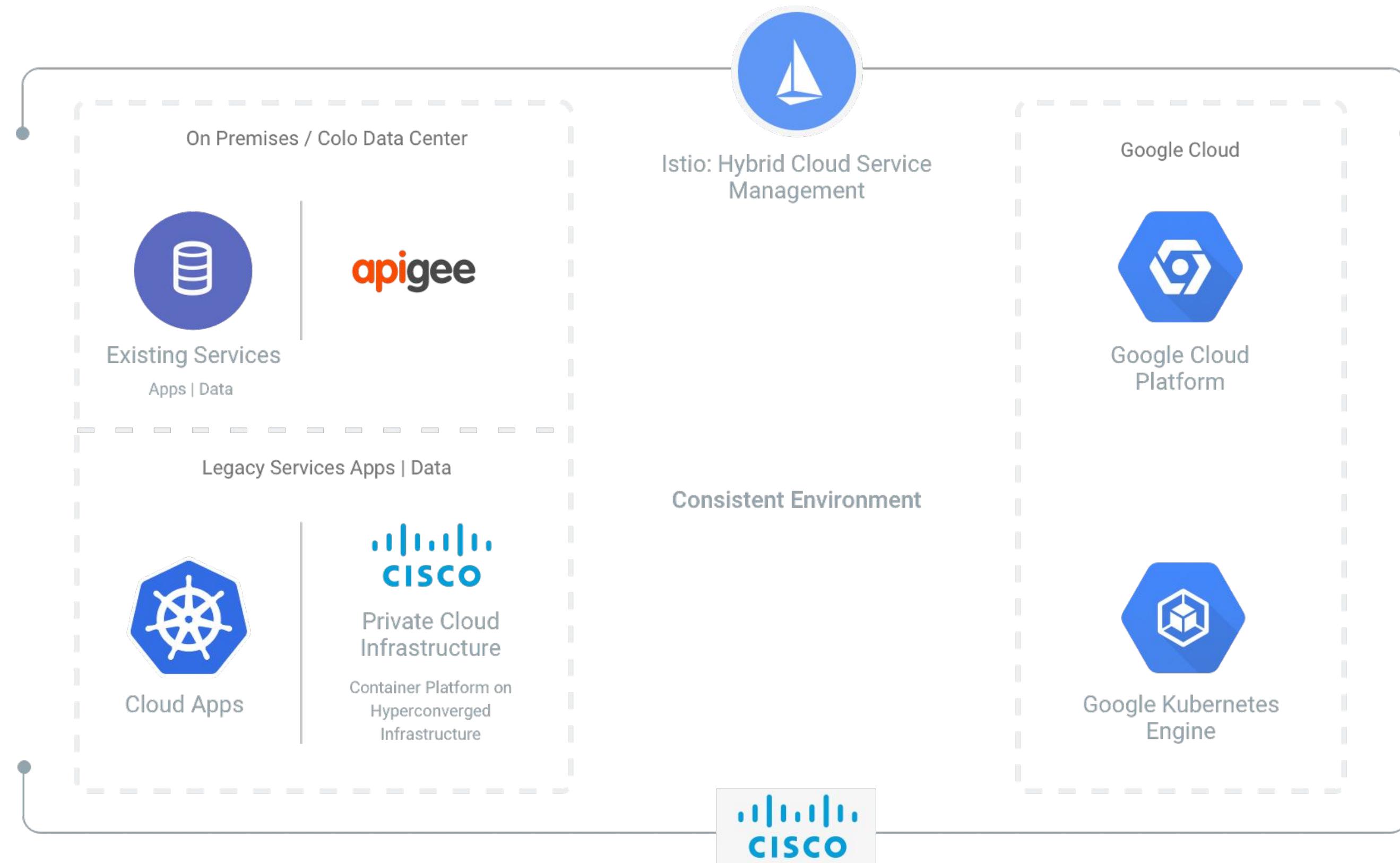


Edge TPU

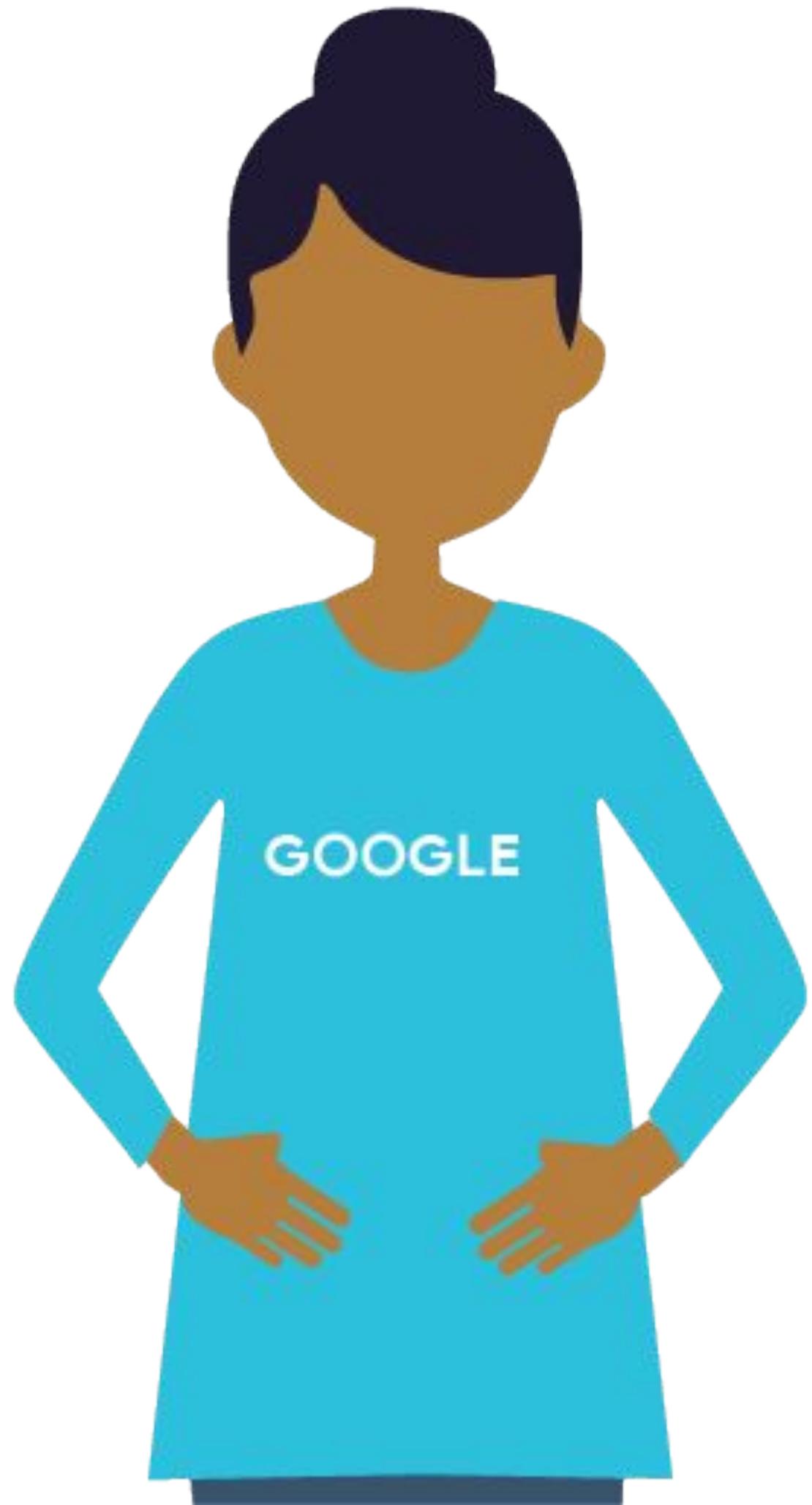


Raspberry Pi





Networking | Security | Cloud Native Infrastructure | Consumption Management

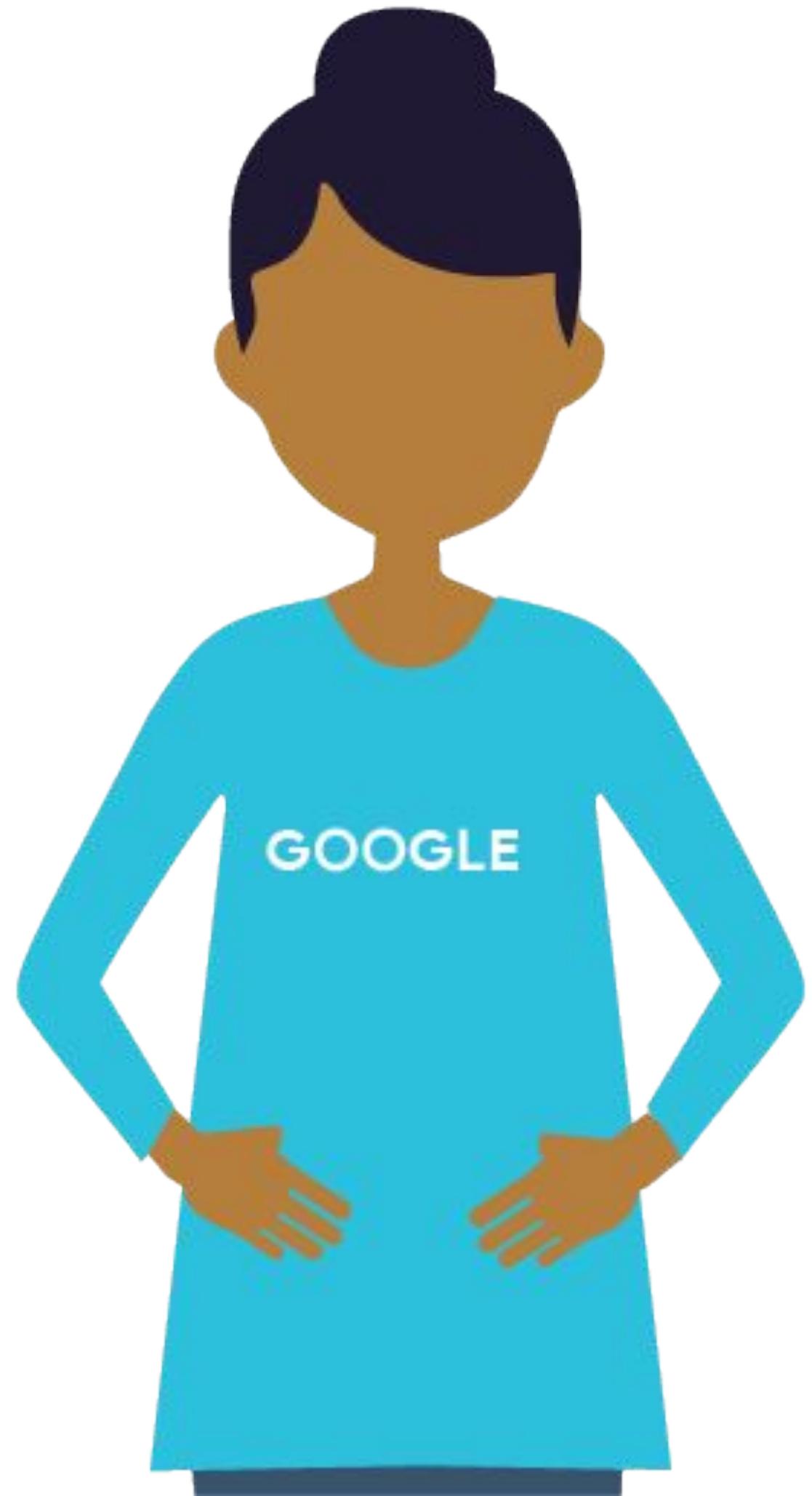


Kubernetes minimizes
infrastructure management



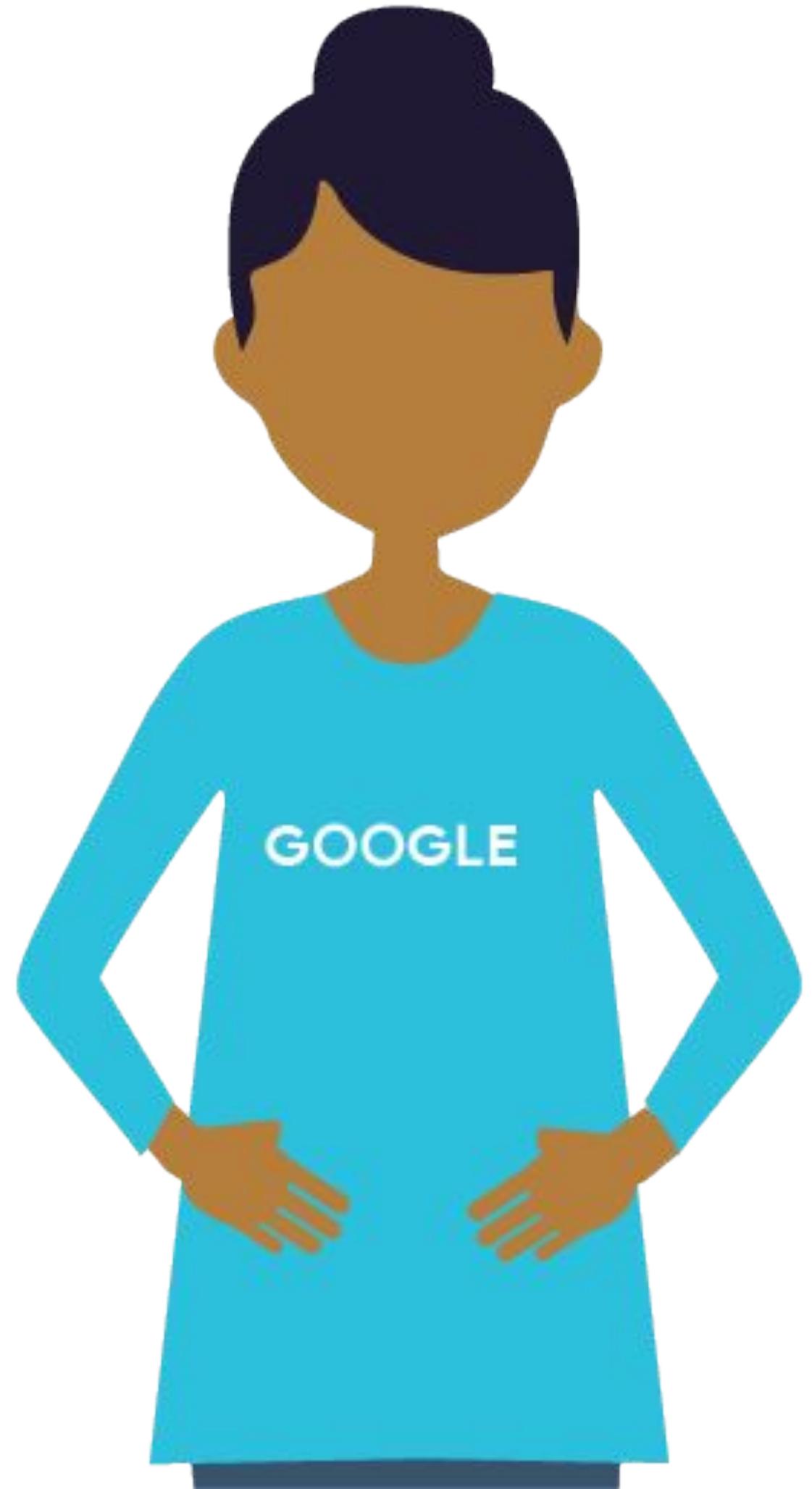
kubernetes





Kubeflow enables hybrid
machine learning

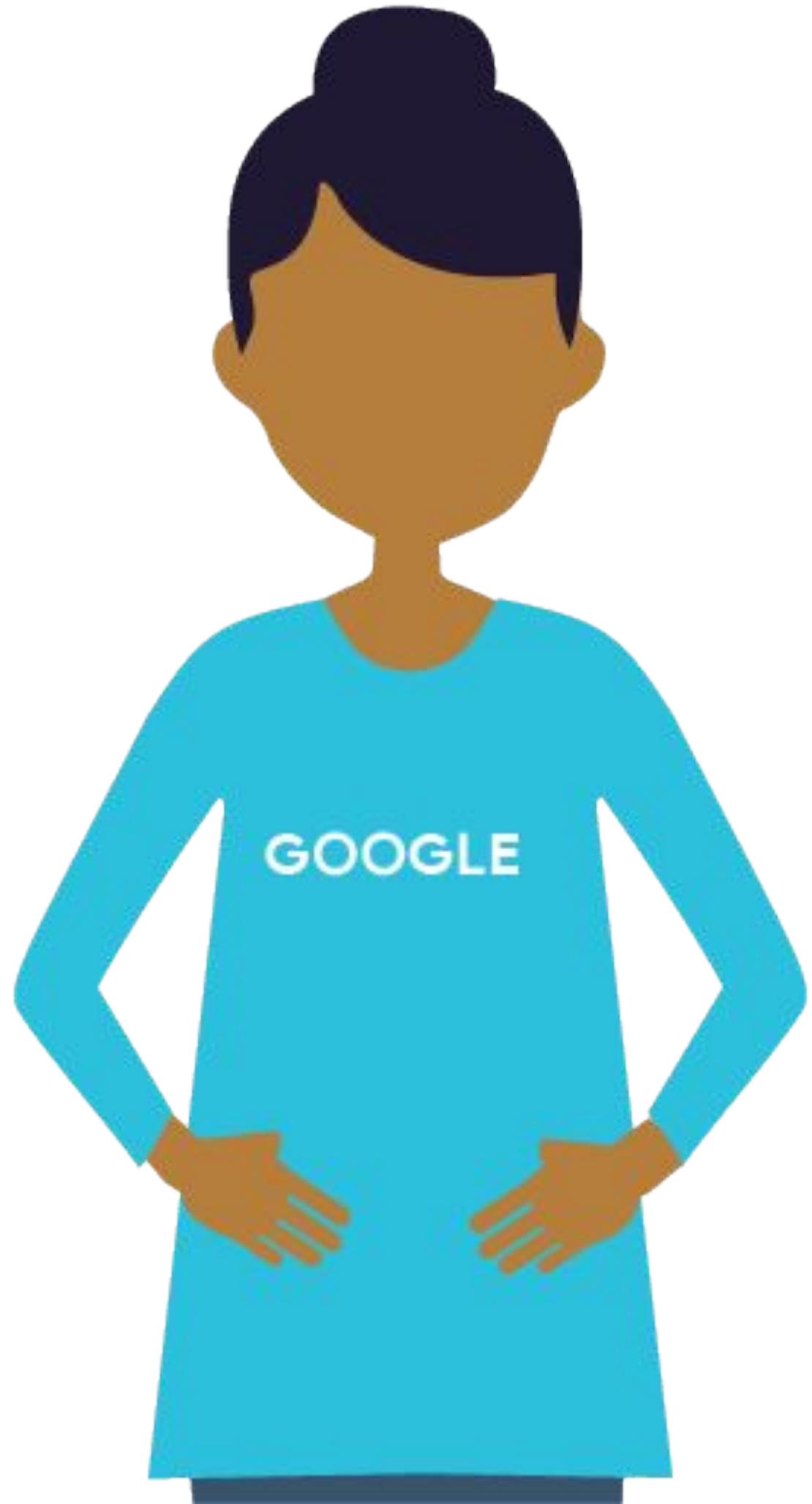




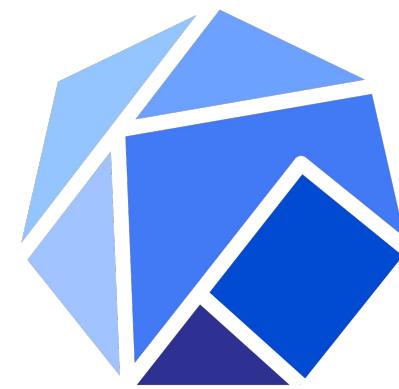
Kubeflow enables hybrid
machine learning



Kubeflow



Kubeflow enables hybrid
machine learning

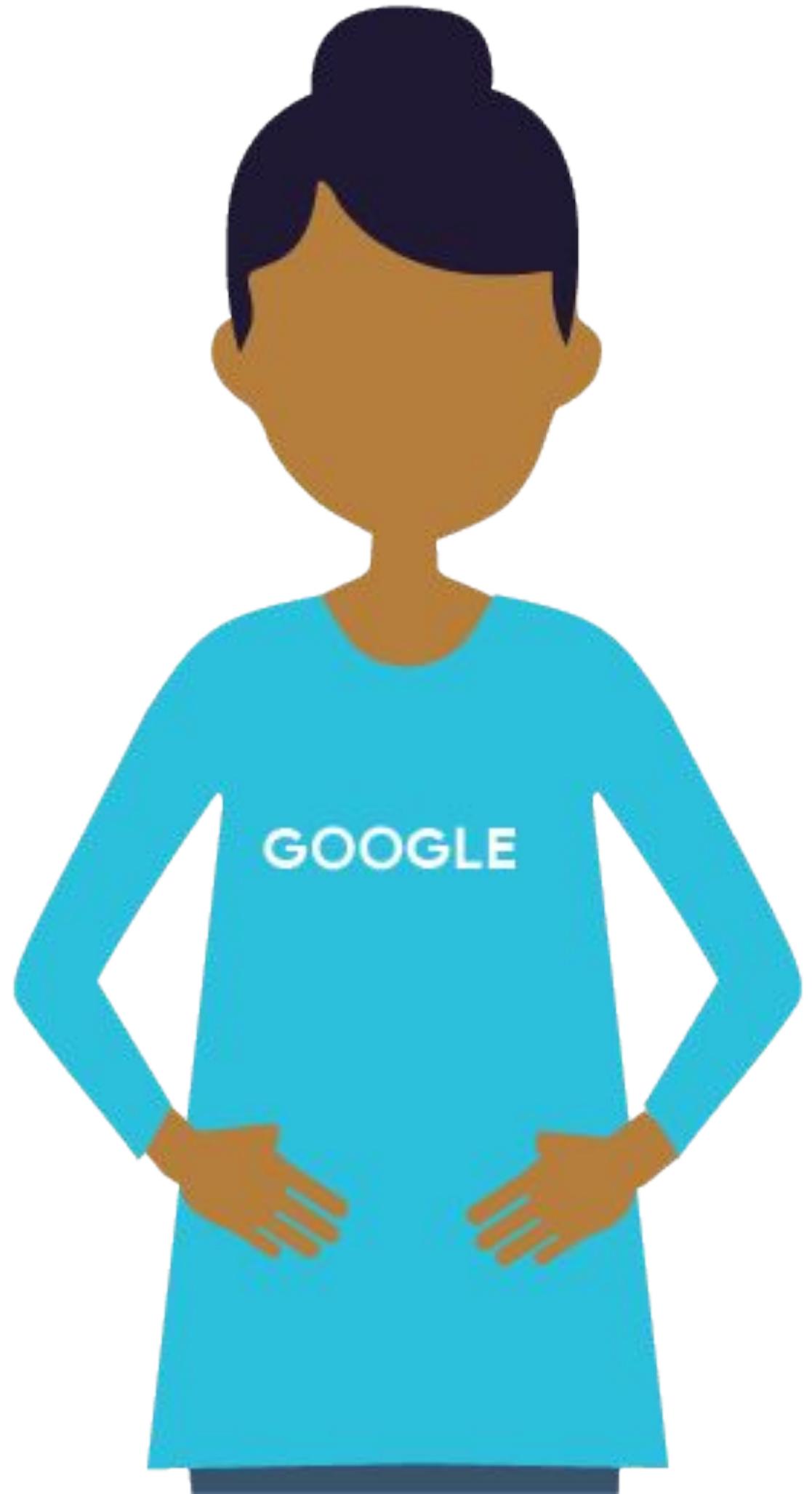


Kubeflow

on



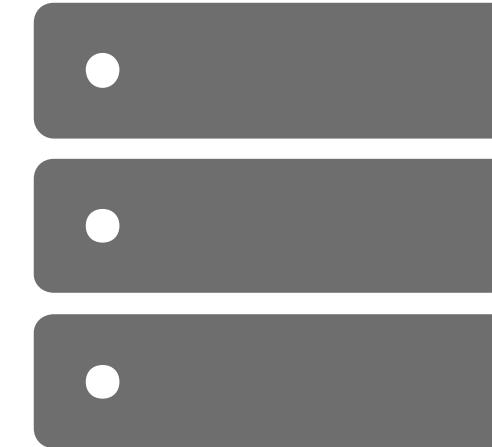
**Kubernetes
Engine**



Kubeflow enables hybrid
machine learning



Kubeflow



Courses 7 - Production ML Systems

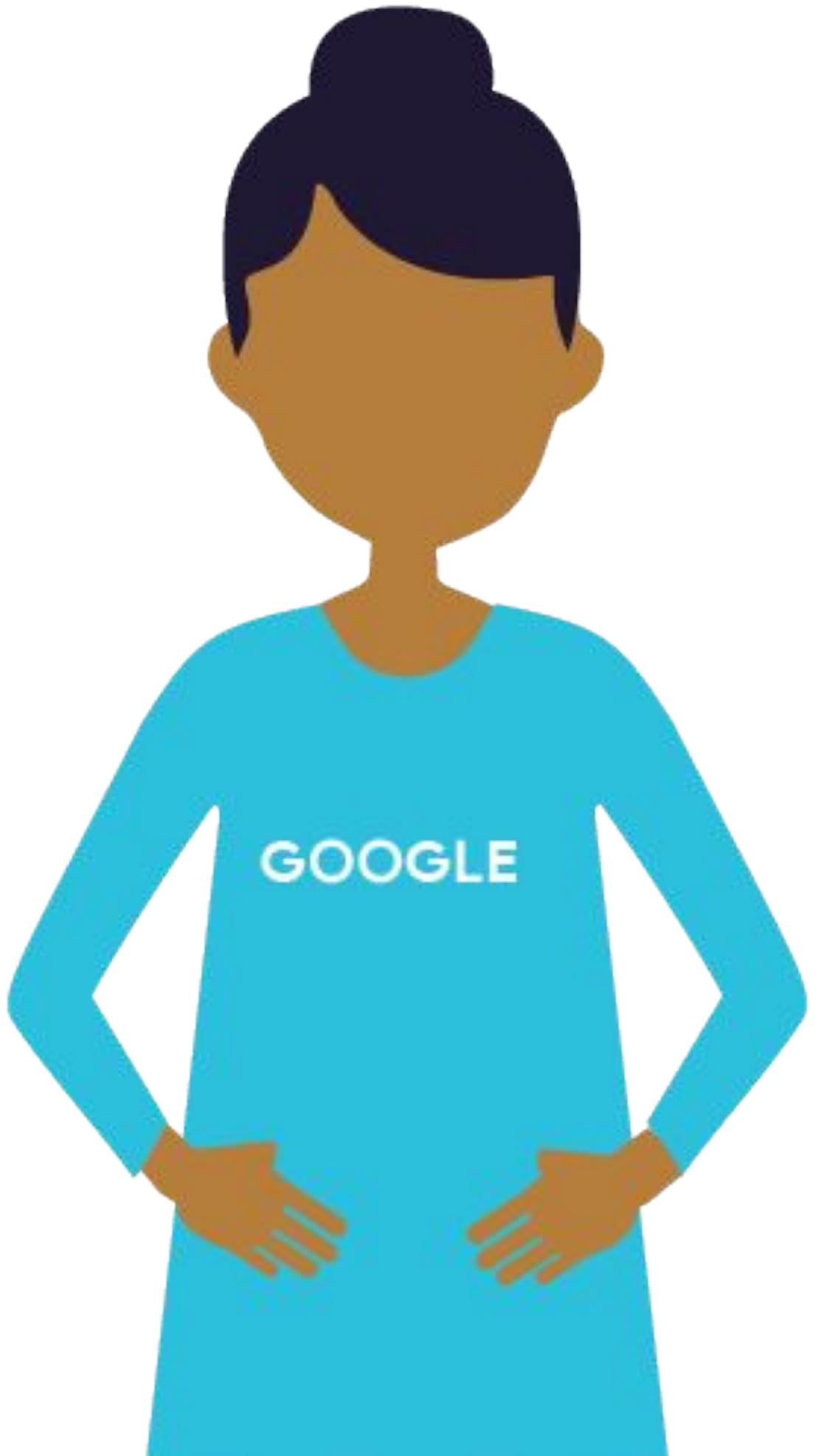
Module 5: Hybrid ML Systems

Lesson Title: Machine learning on hybrid cloud

Format: Presenter

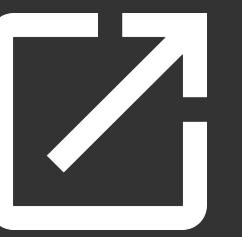
Presenter: Val

Video Name: T-PSML-O_5_I2_machine_learning_on_hybrid_cloud



 Composability

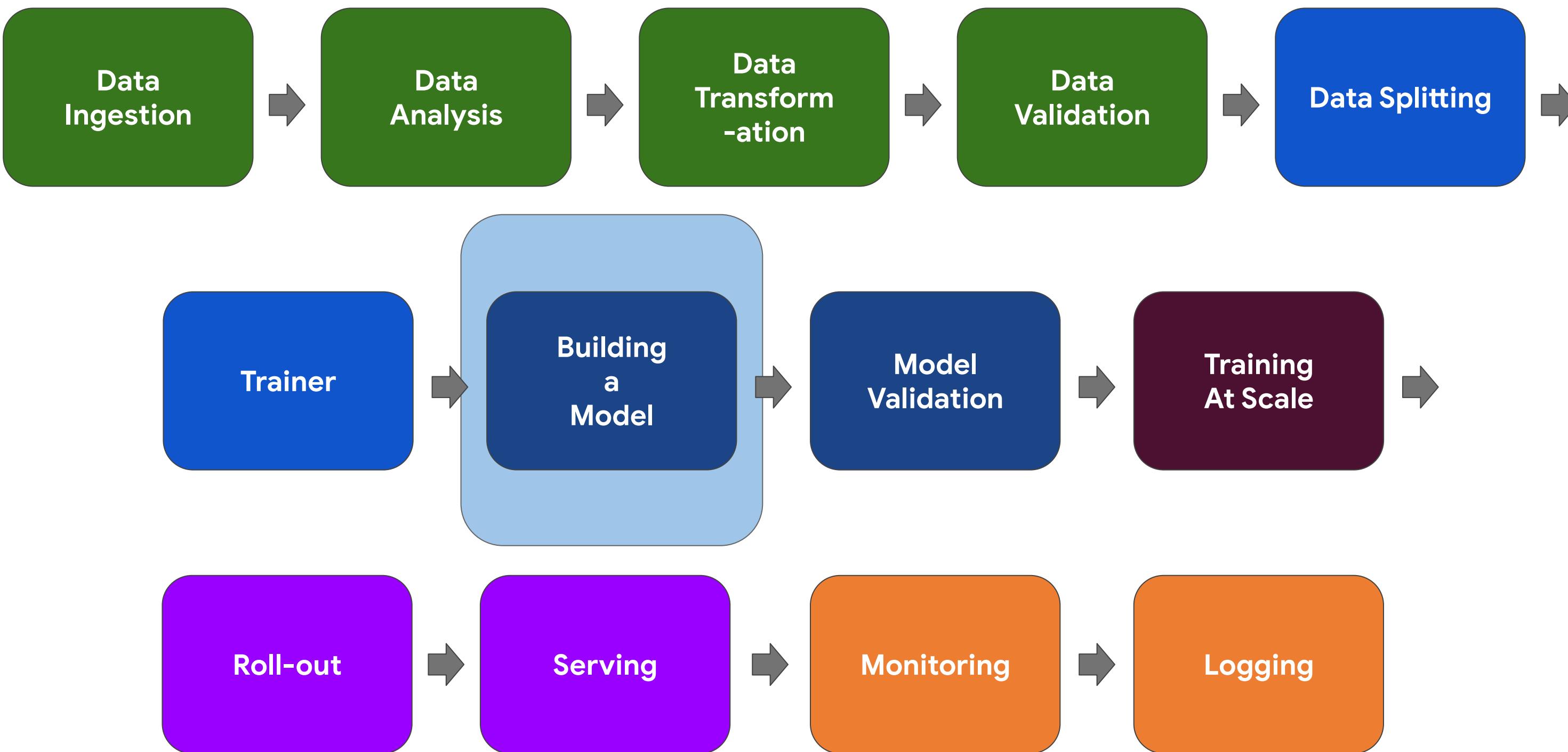
 Portability

 Scalability

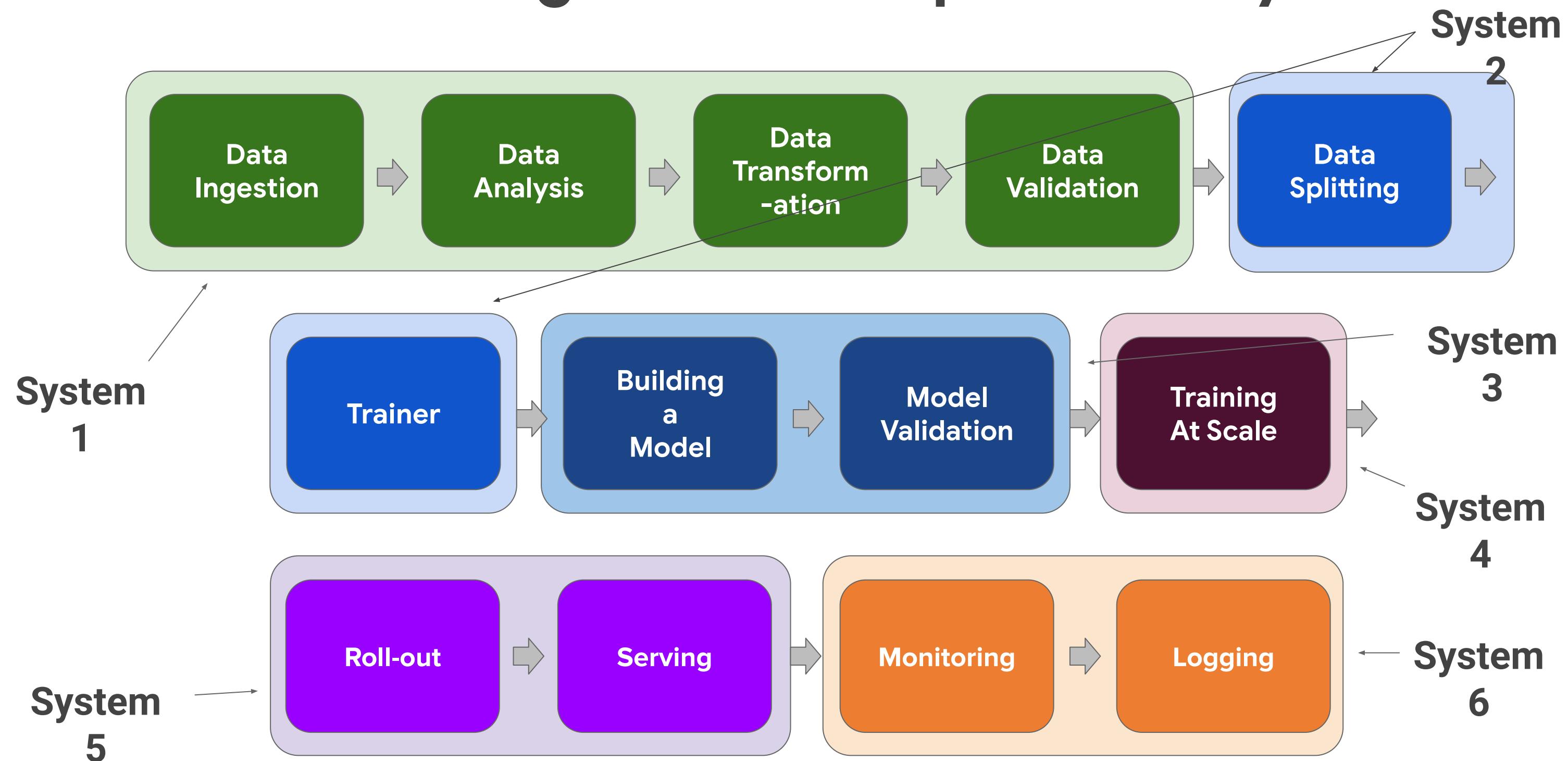
Composability

Building
a
Model

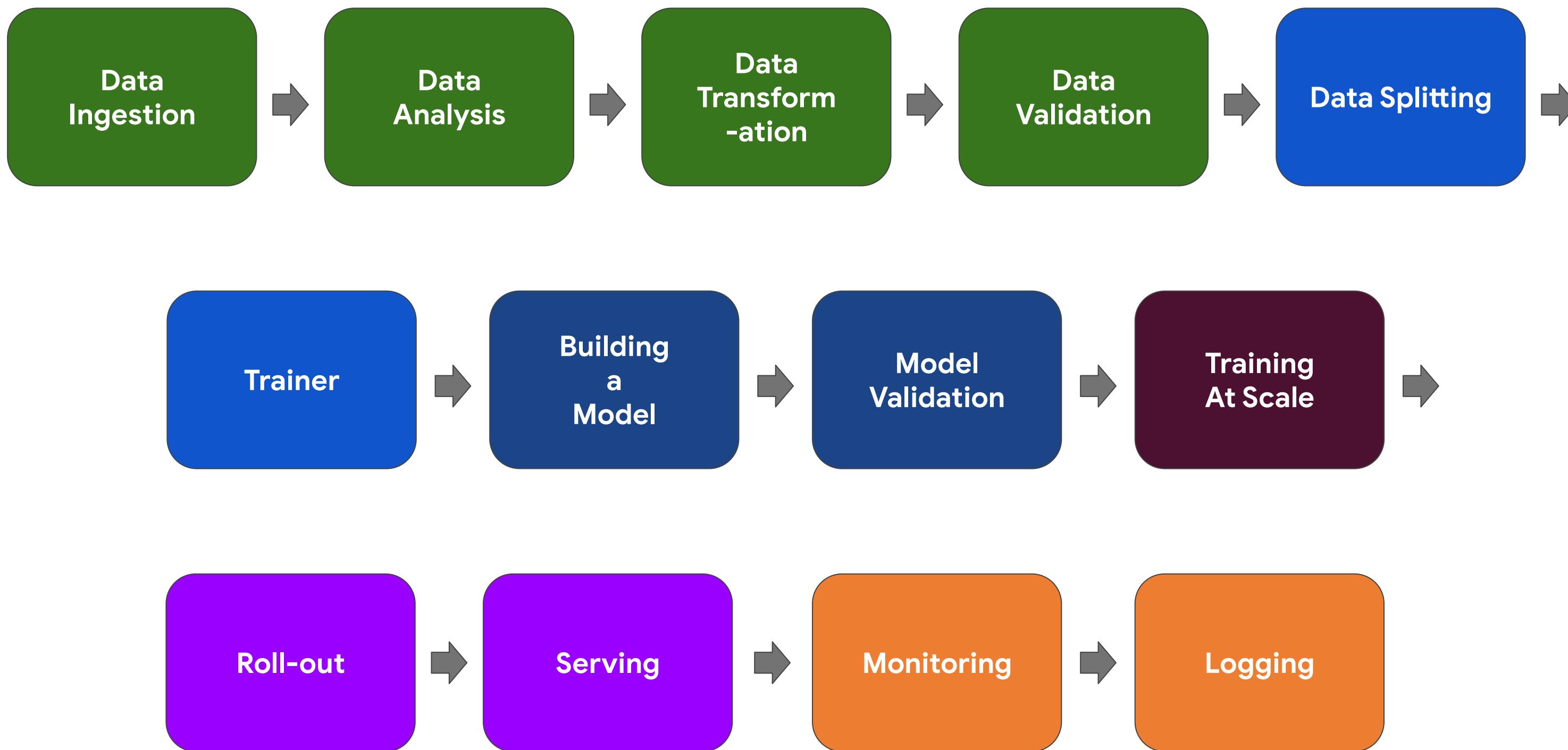
Building a model is only one part of the entire system

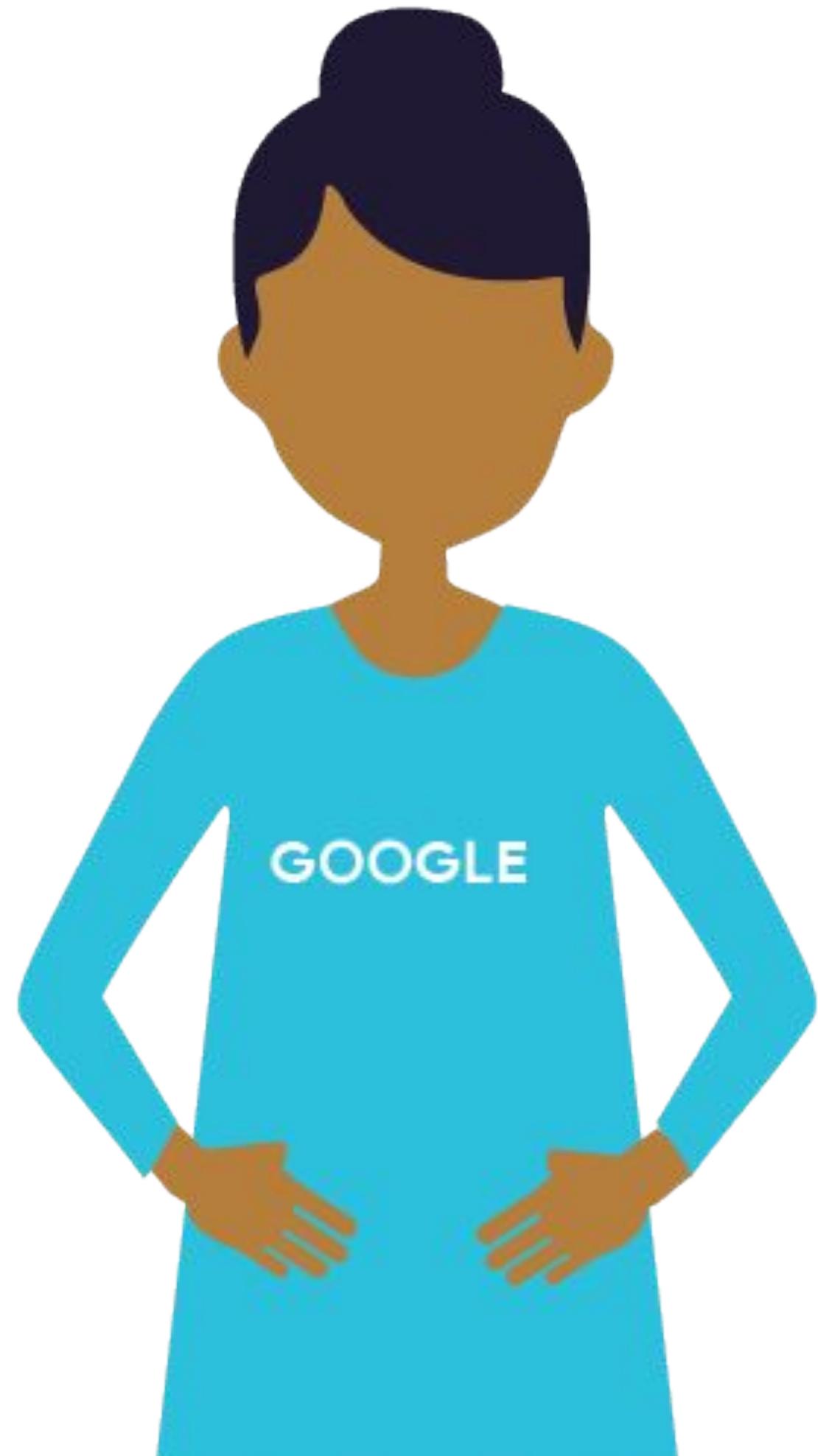


Each ML Stage is an Independent System



Composability is about microservices





Portability

Portability

Experimentation

Model

UX

Tooling

Framework

Storage

Runtime

Drivers

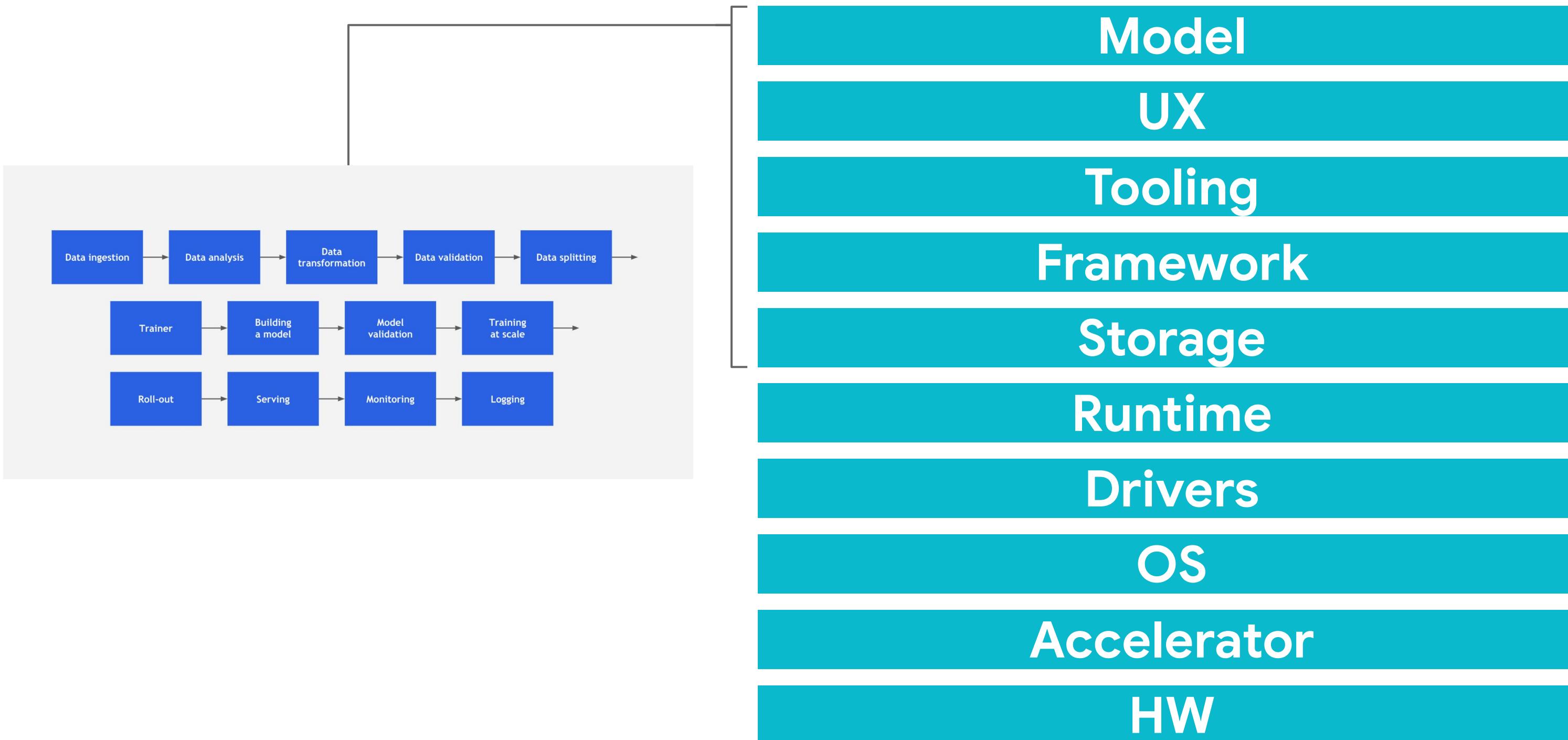
OS

Accelerator

HW

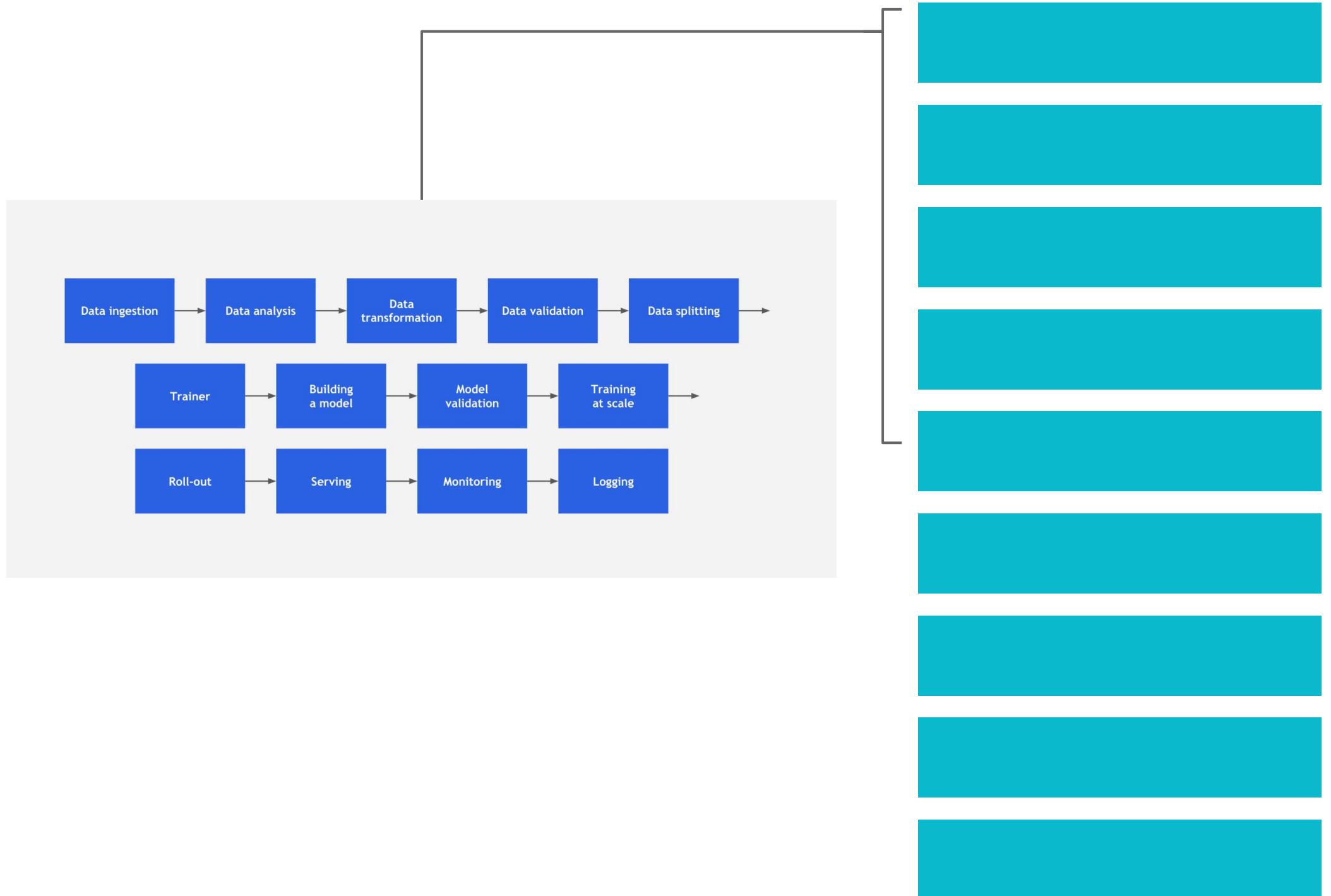
Portability

Experimentation

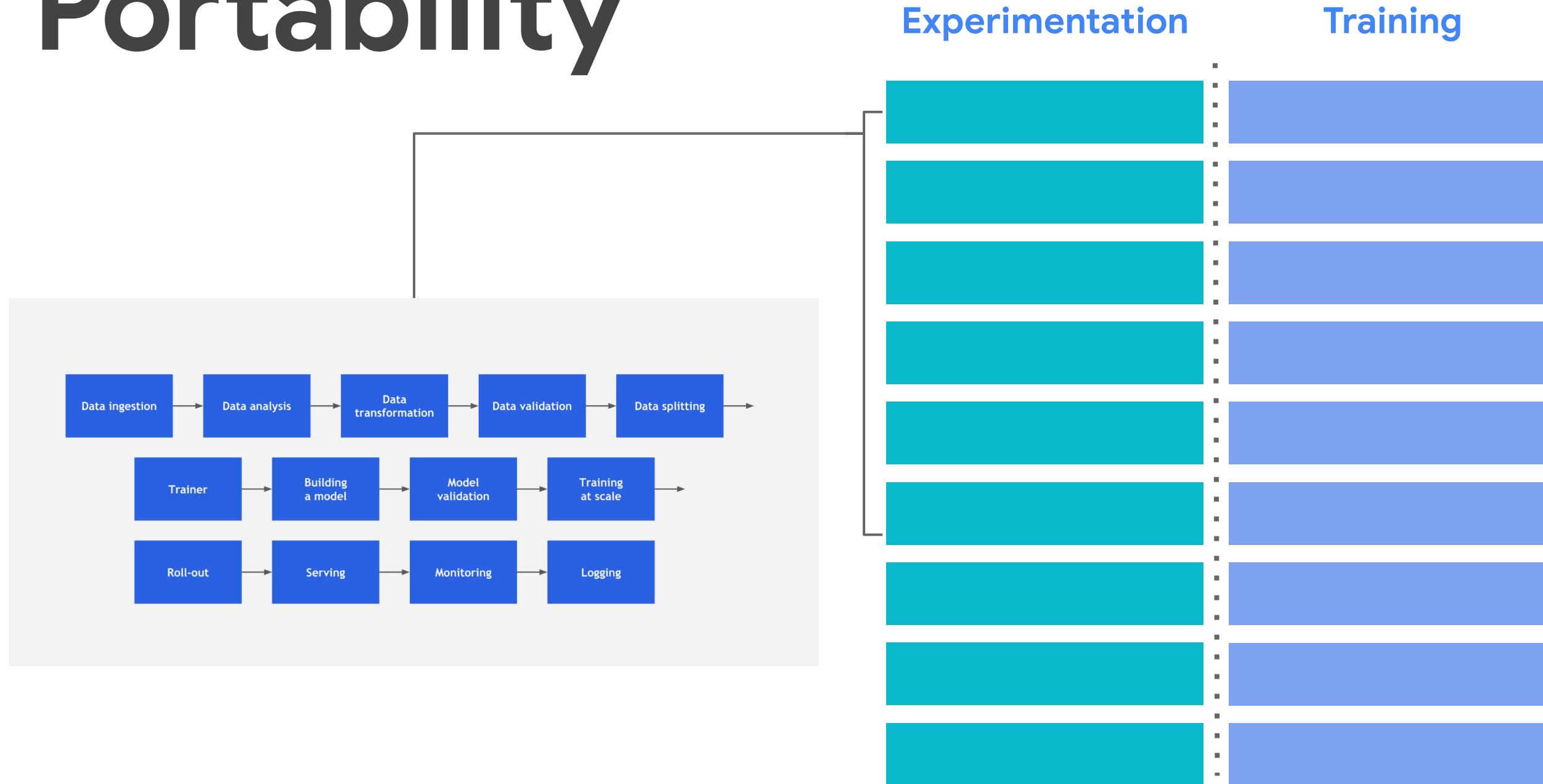


Portability

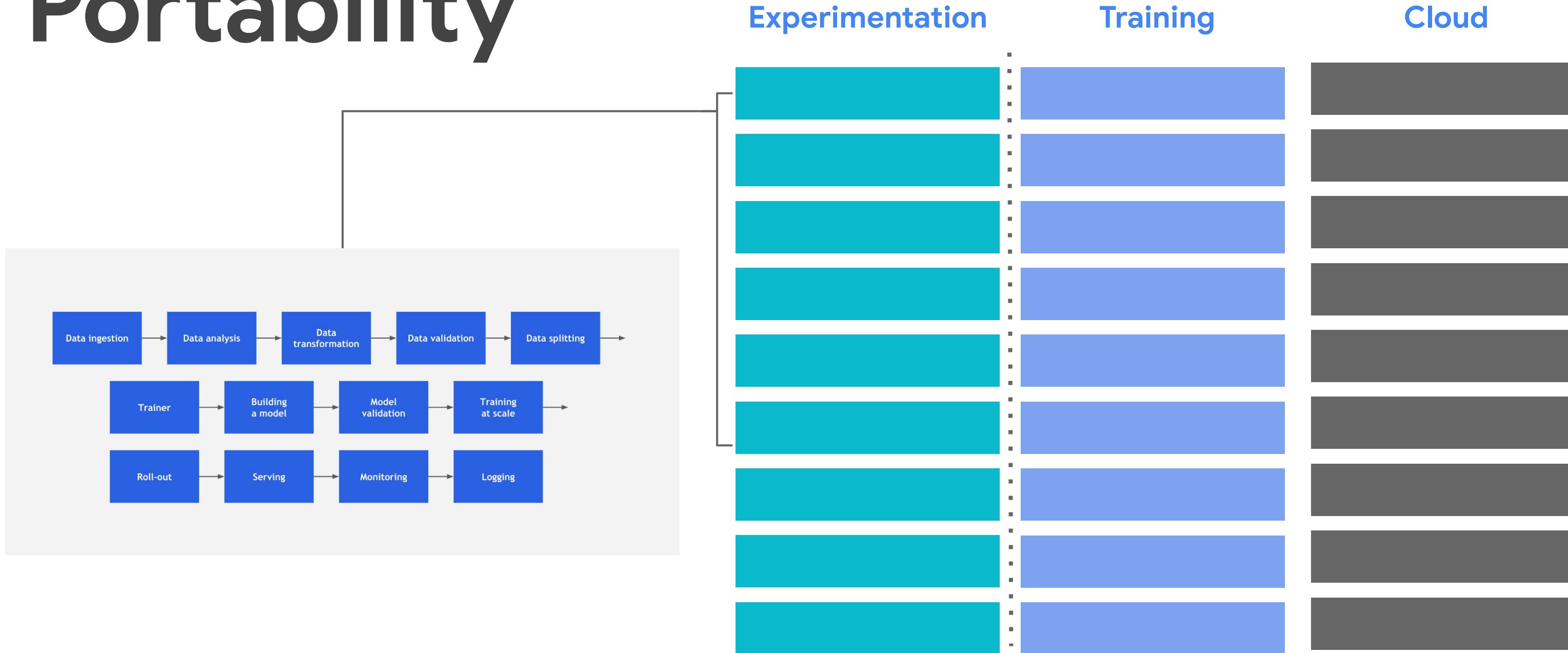
Experimentation

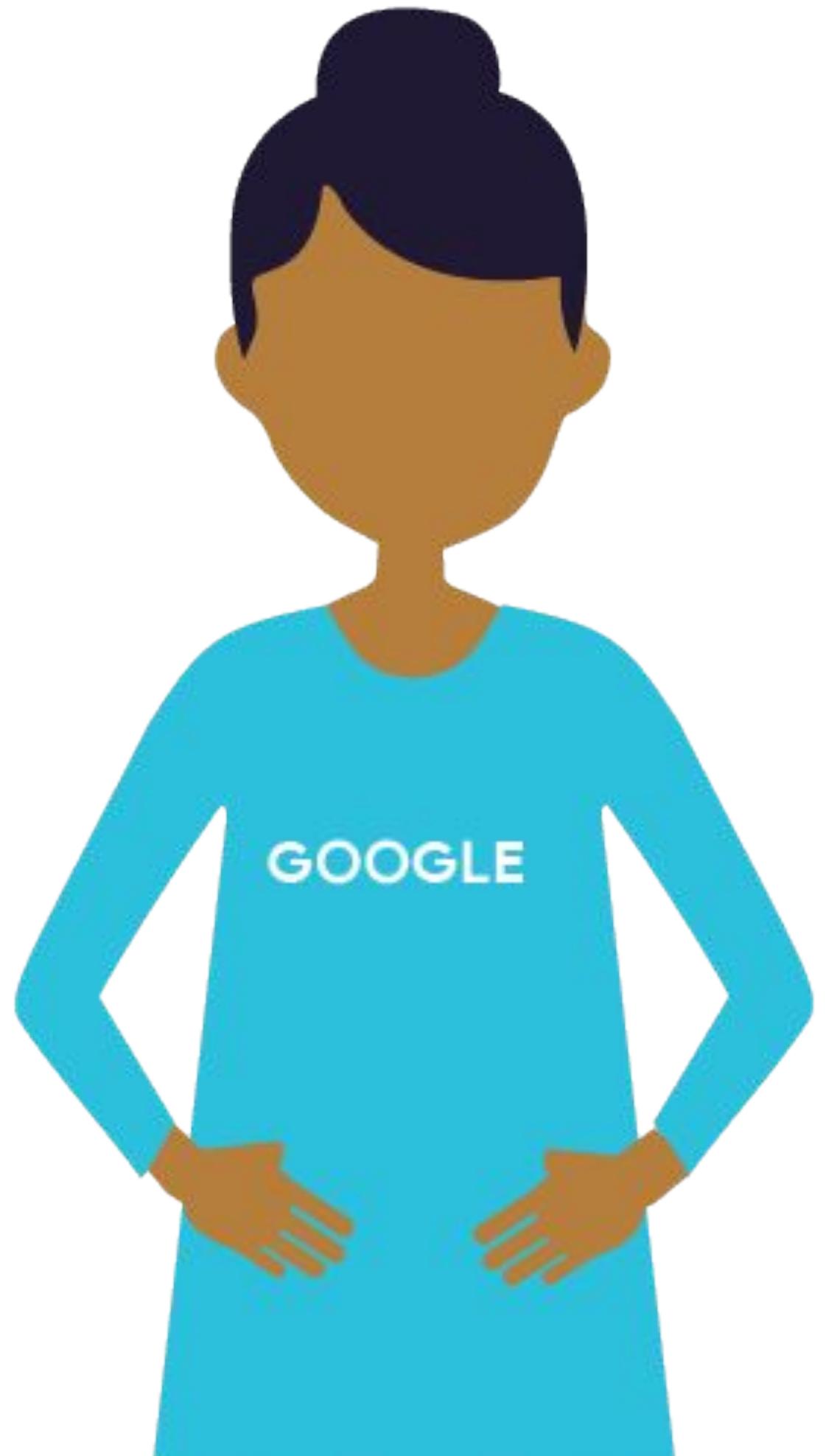


Portability

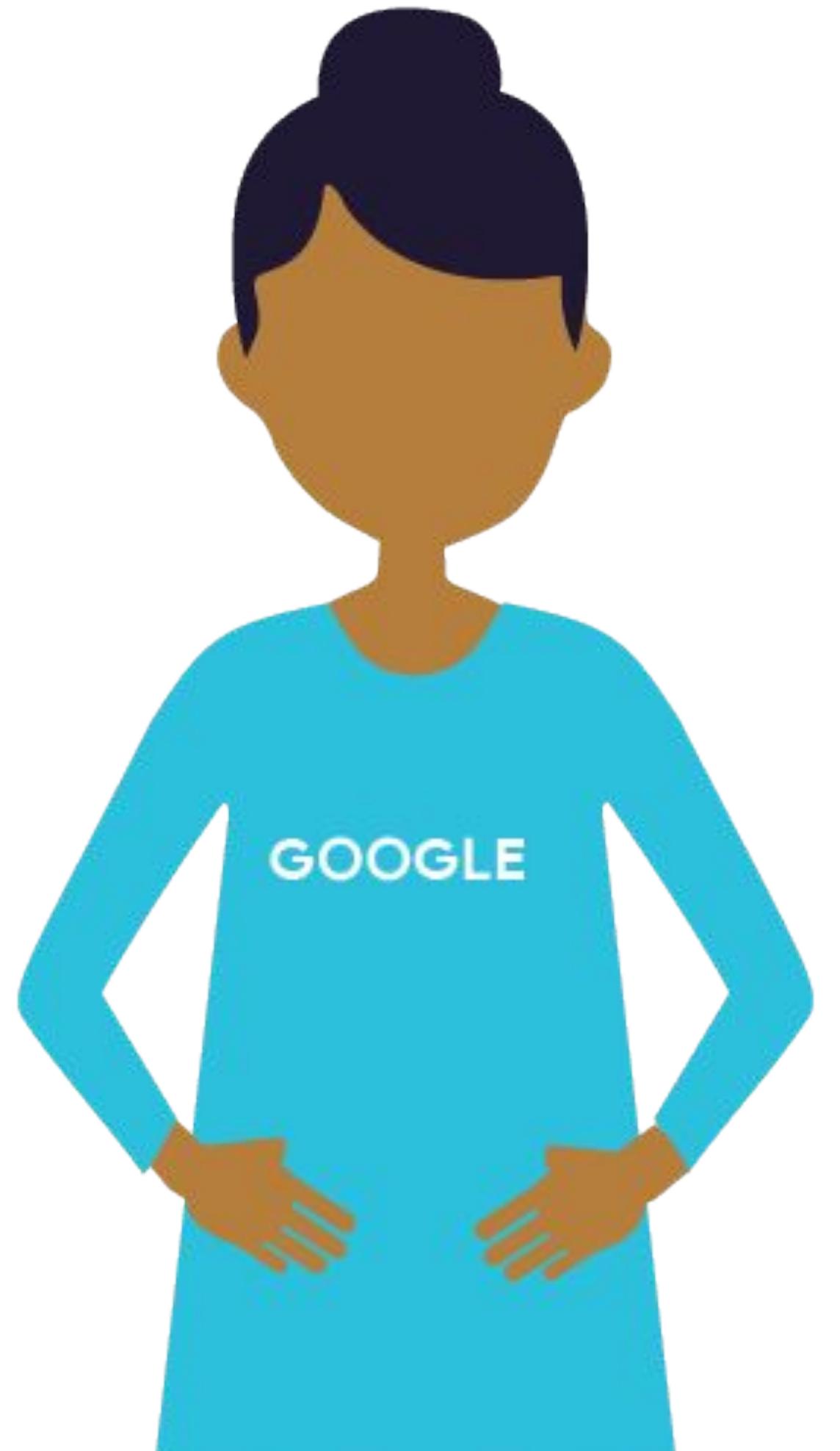


Portability





“Portability doesn’t
matter to me”



Wrong!



Joe Beda 

@jbeda

Following



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





Joe Beda

@jbeda

Following



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





Joe Beda

@jbeda

Following

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



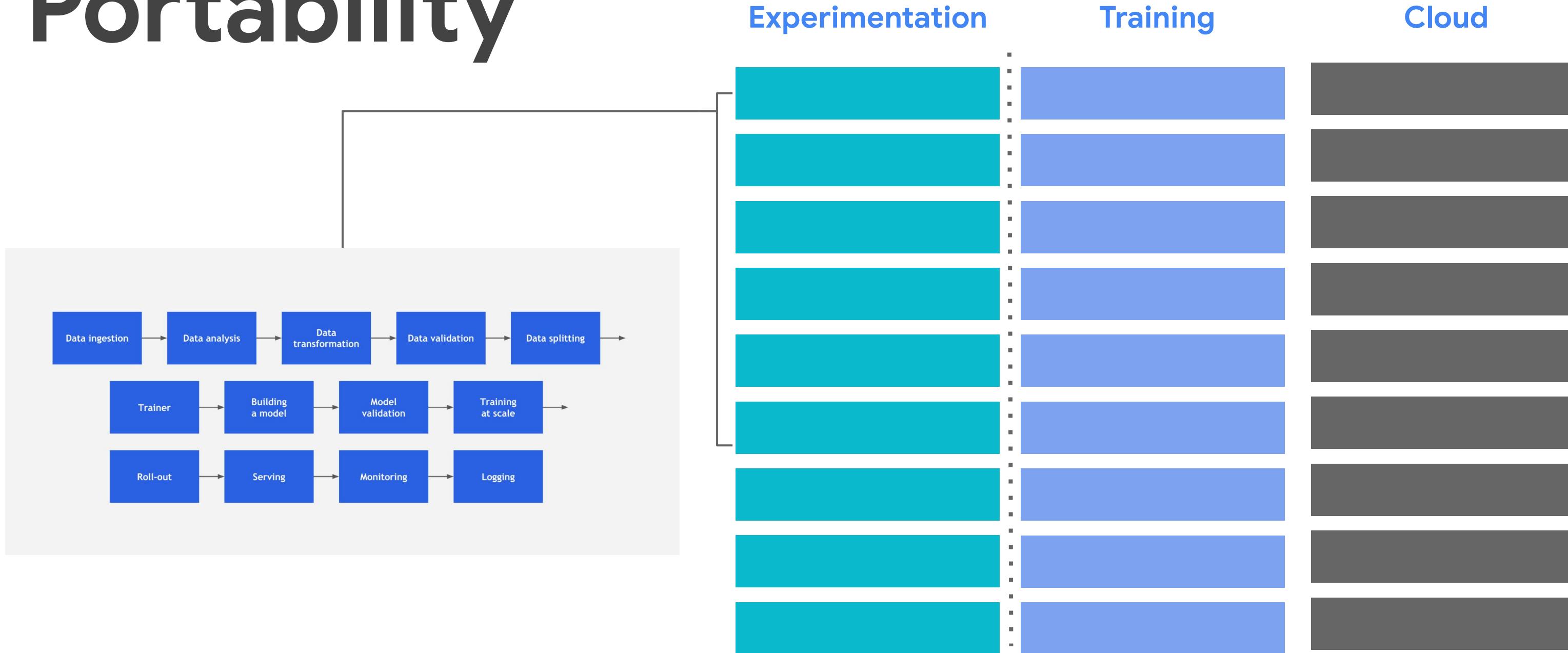
54

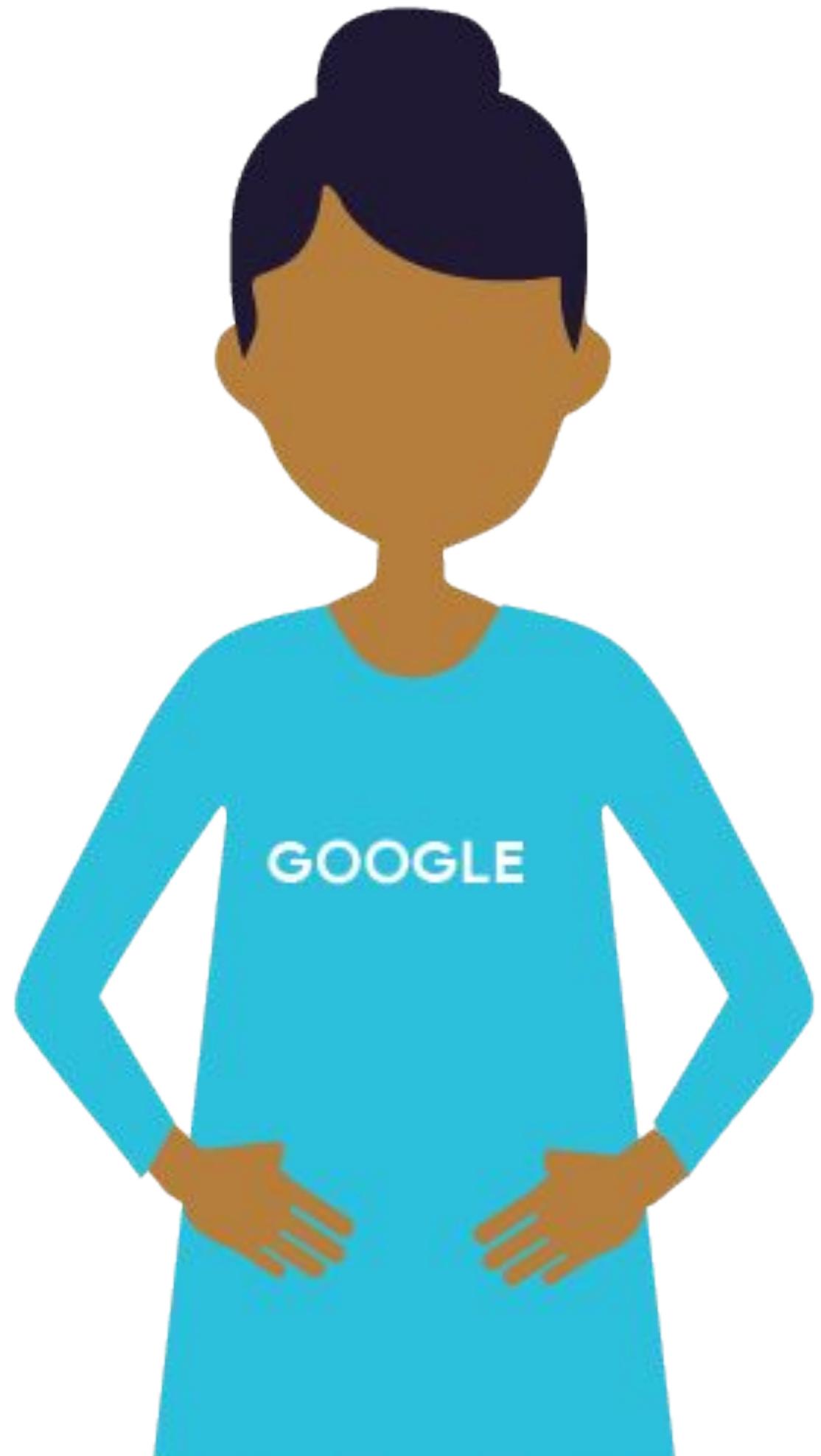


107

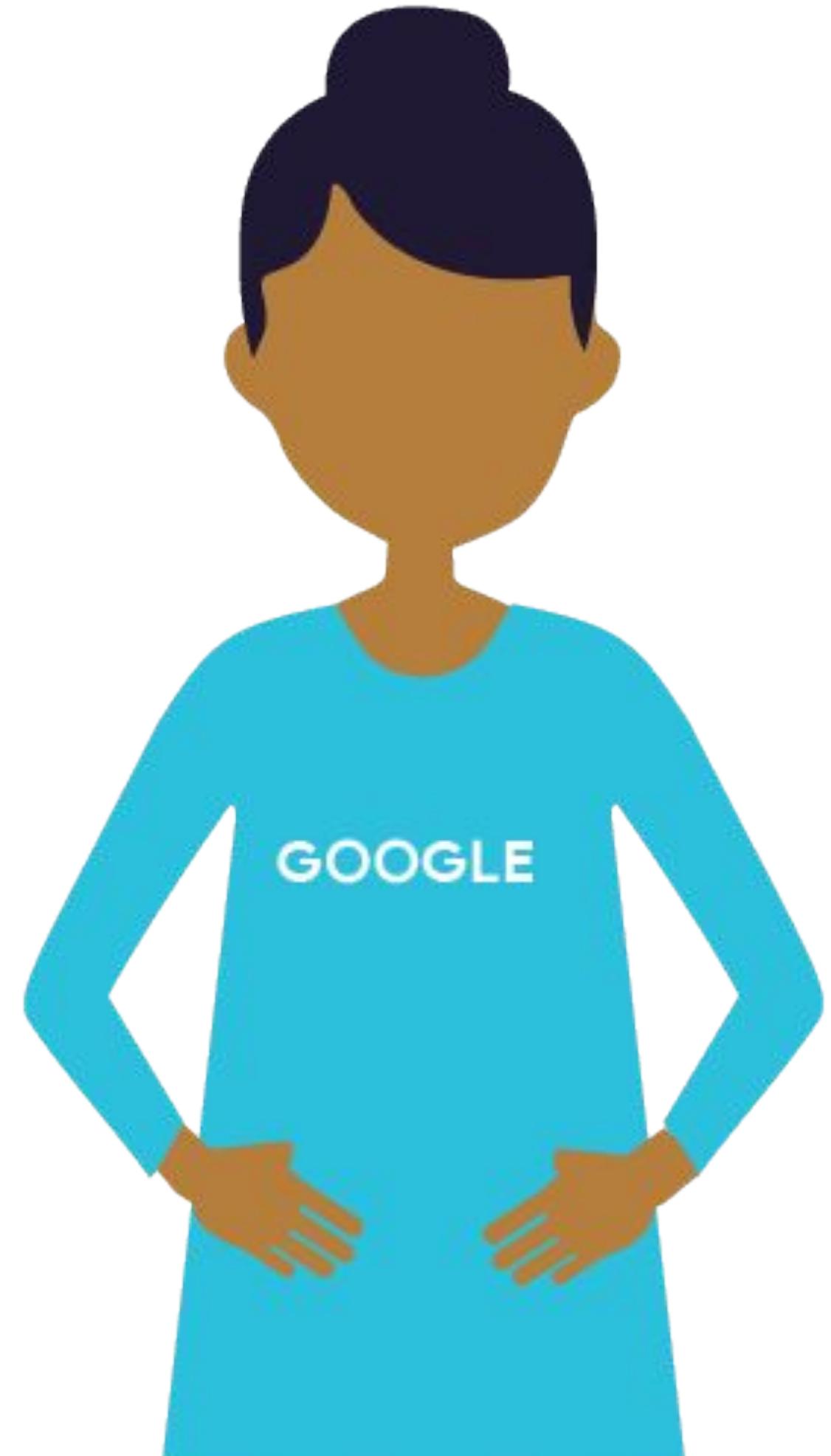


Portability



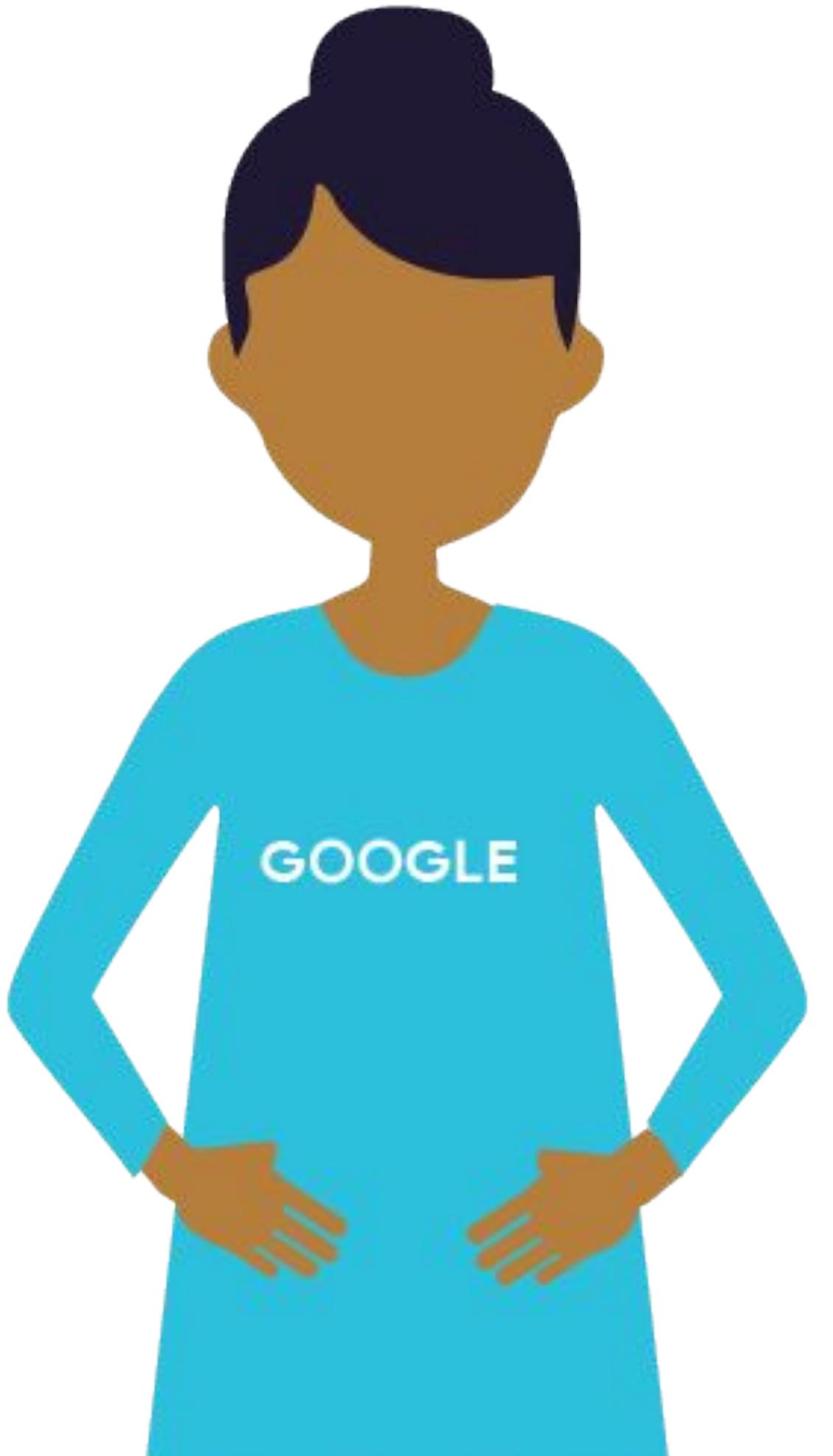


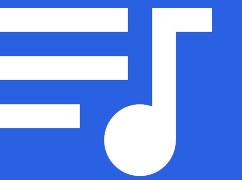
Your Laptop
Counts.



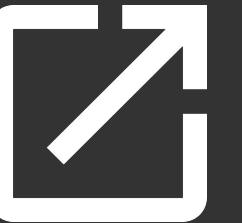
Scalability

- **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

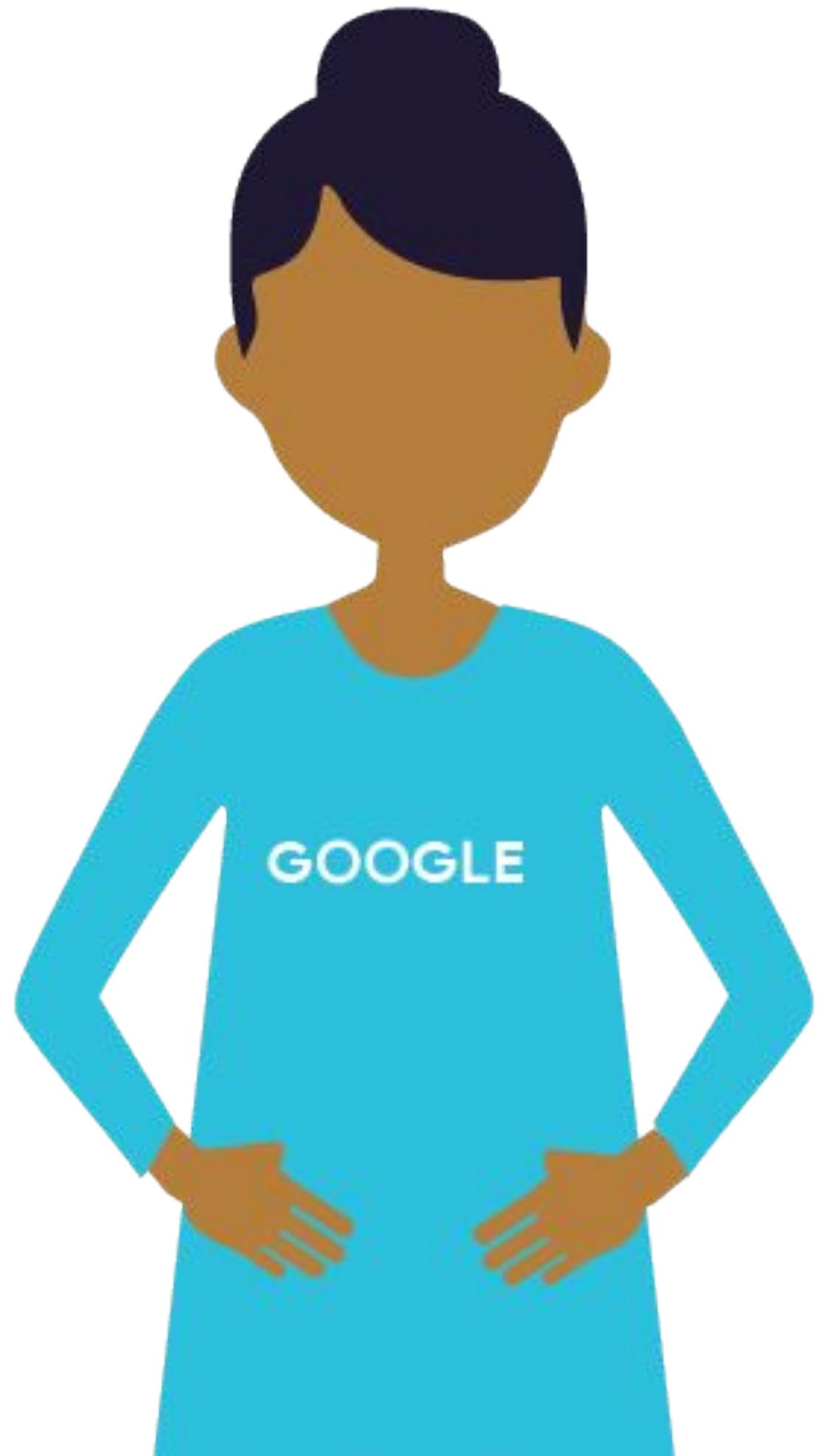
Module 5: Hybrid ML Systems

Lesson Title: **Kubeflow**

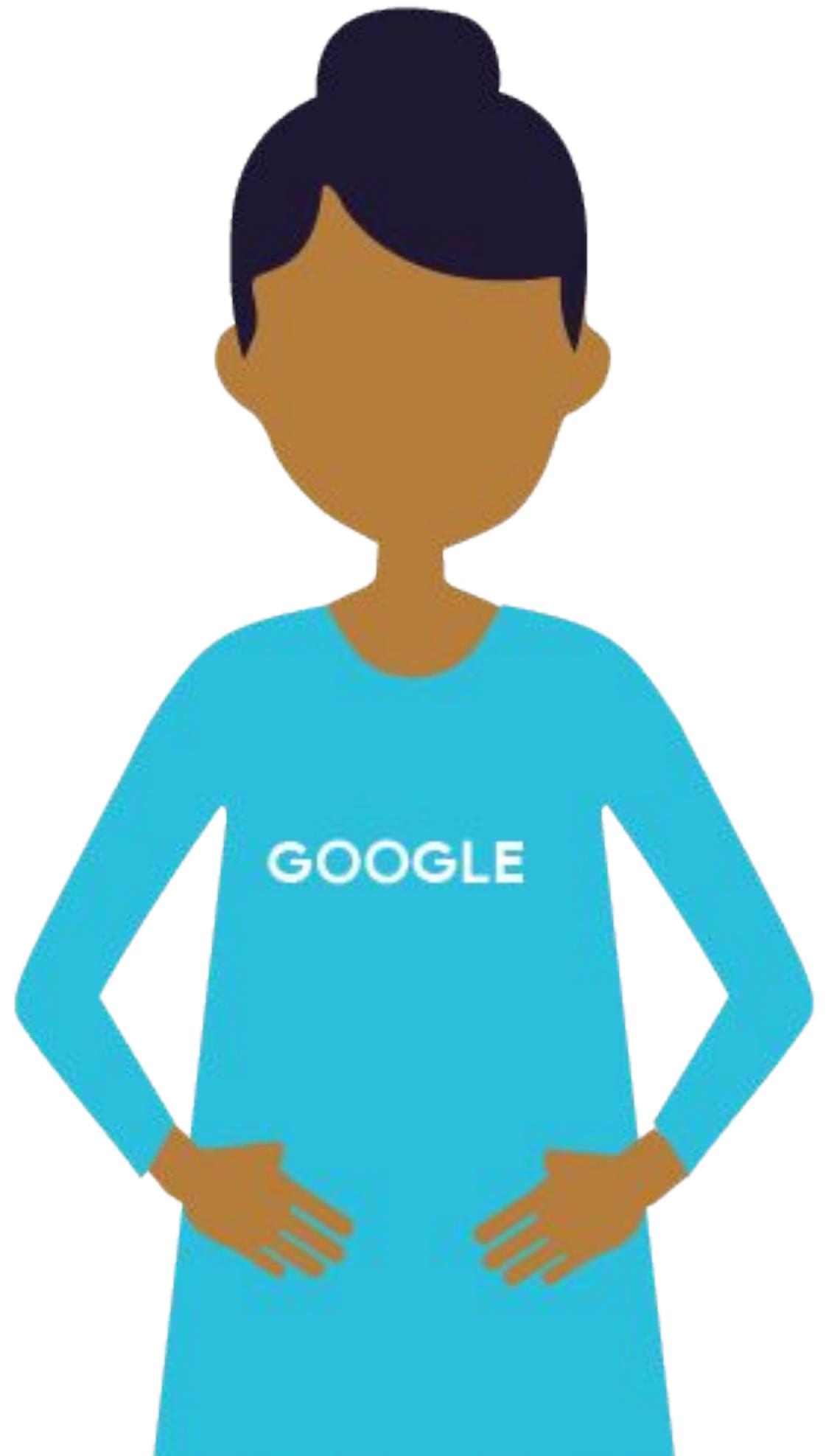
Format: Presenter

Presenter: Val

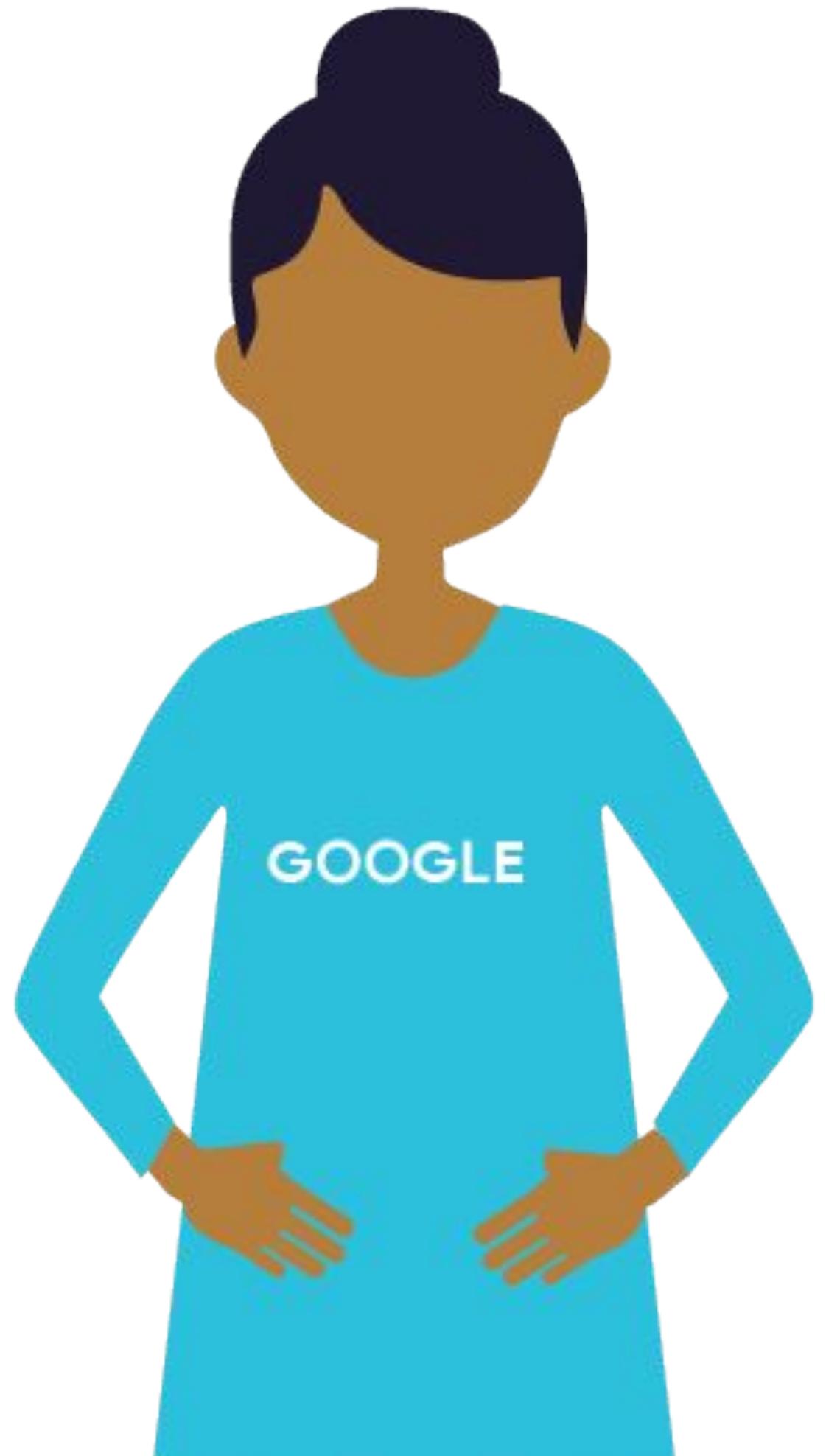
Video Name: T-PSML-O_5_I3_kubeflow



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

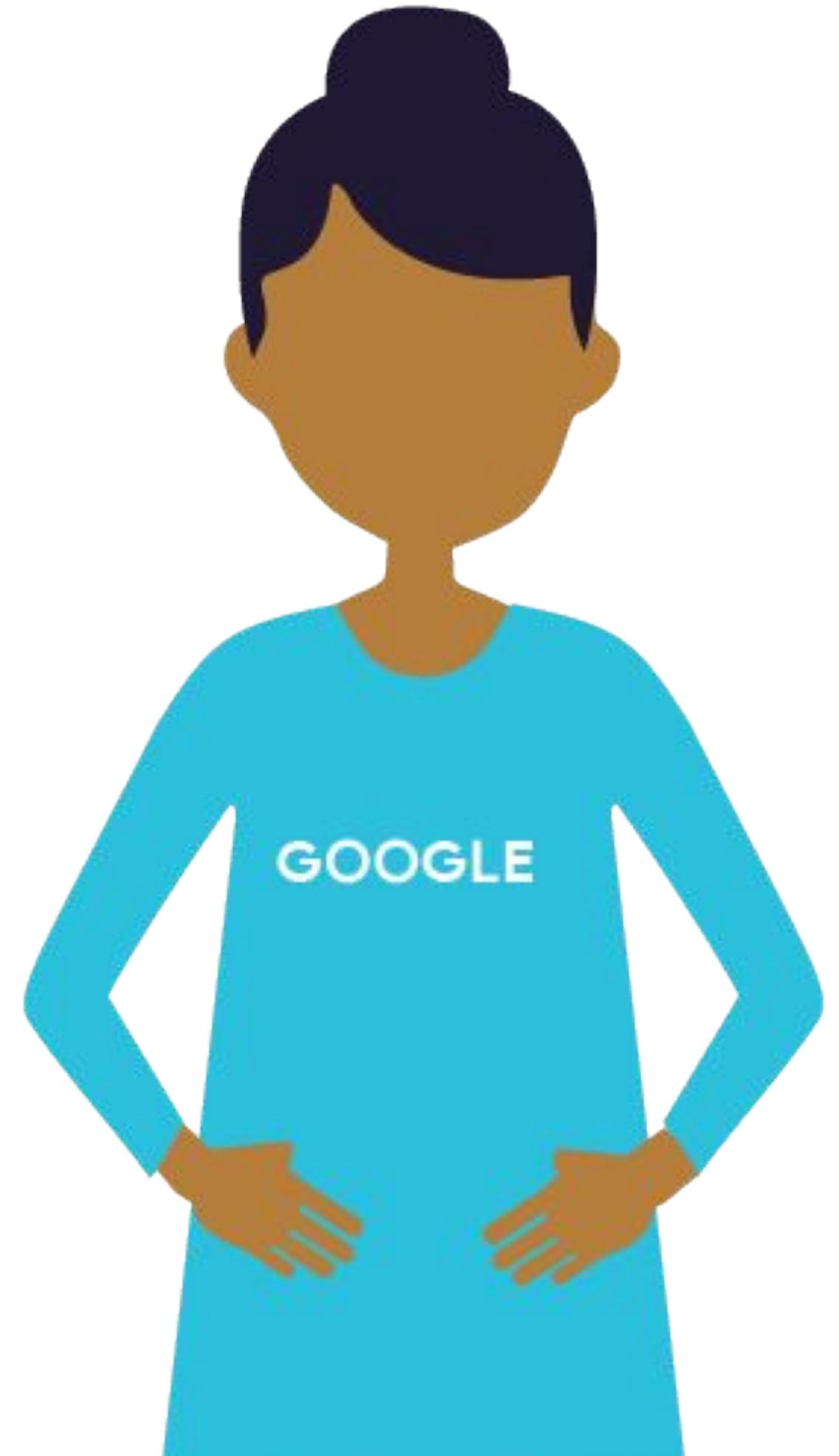


Containers & Kubernetes



Containers & Kubernetes

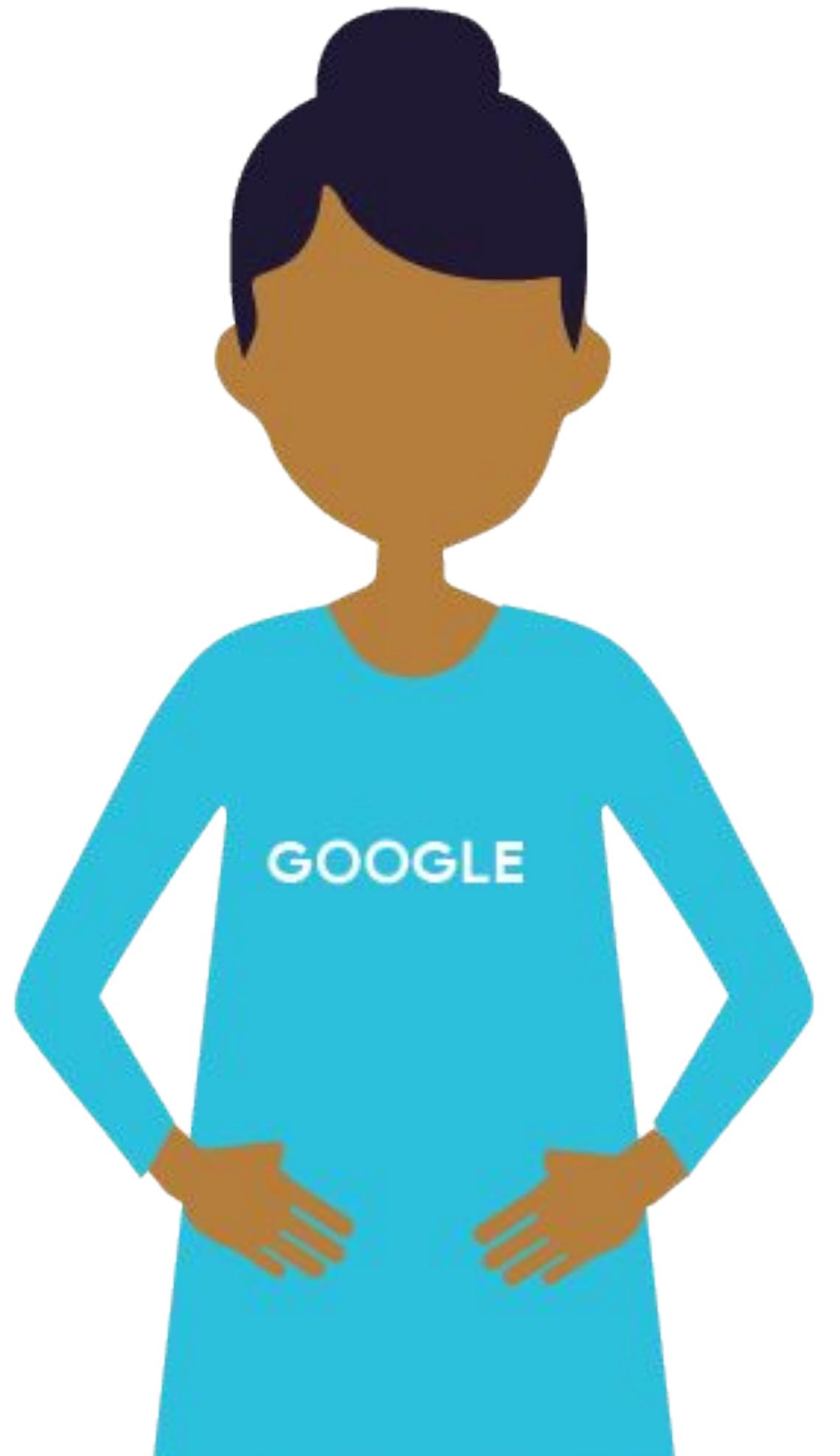
except



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

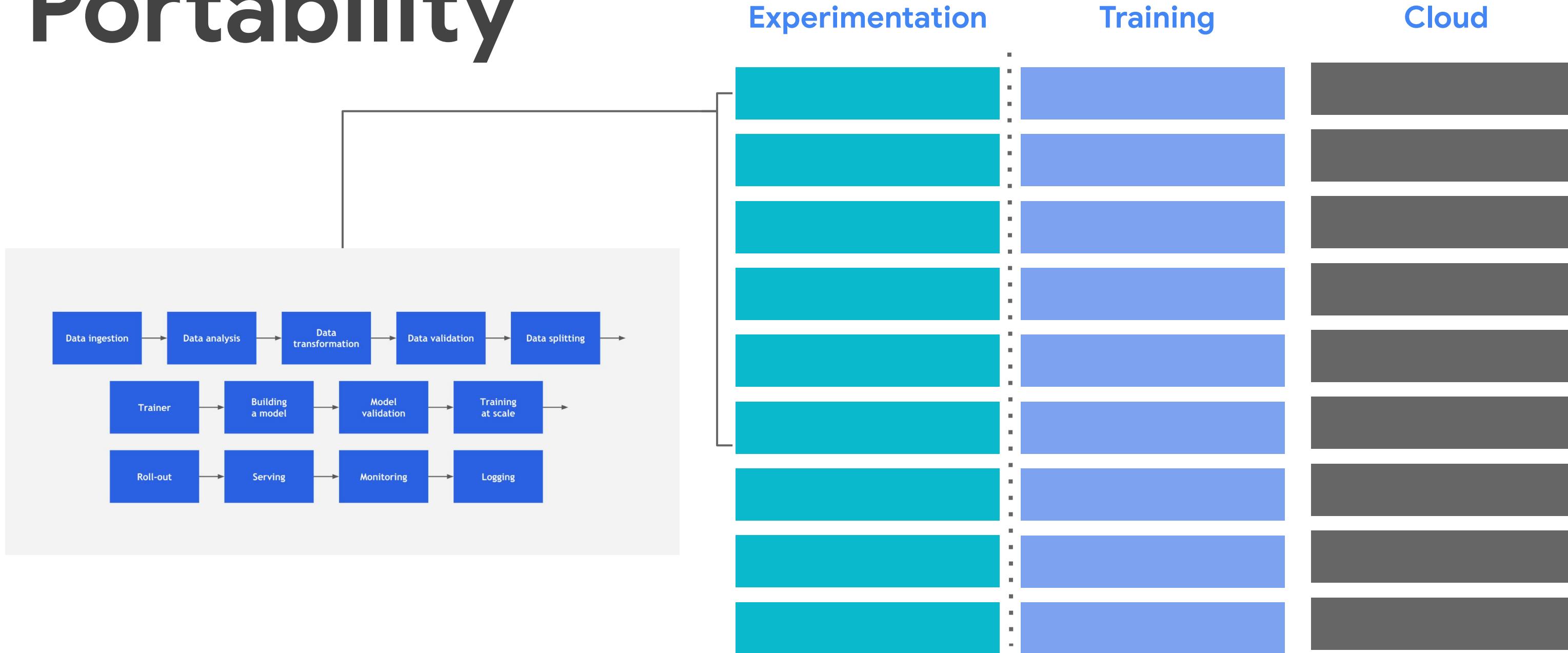


Oh, you want to
use ML on K8s?

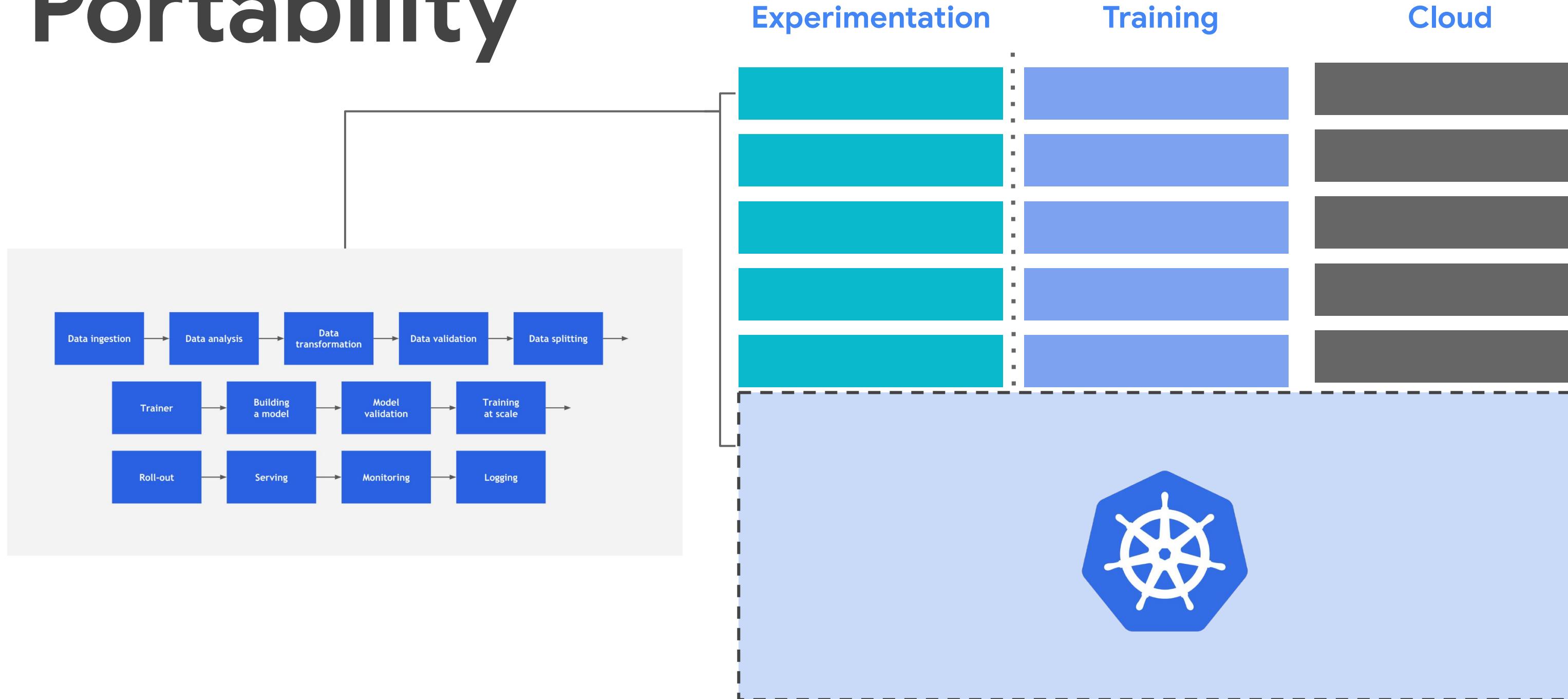


Make it Easy for Everyone
to Develop, Deploy and Manage
Portable, Distributed ML
on Kubernetes

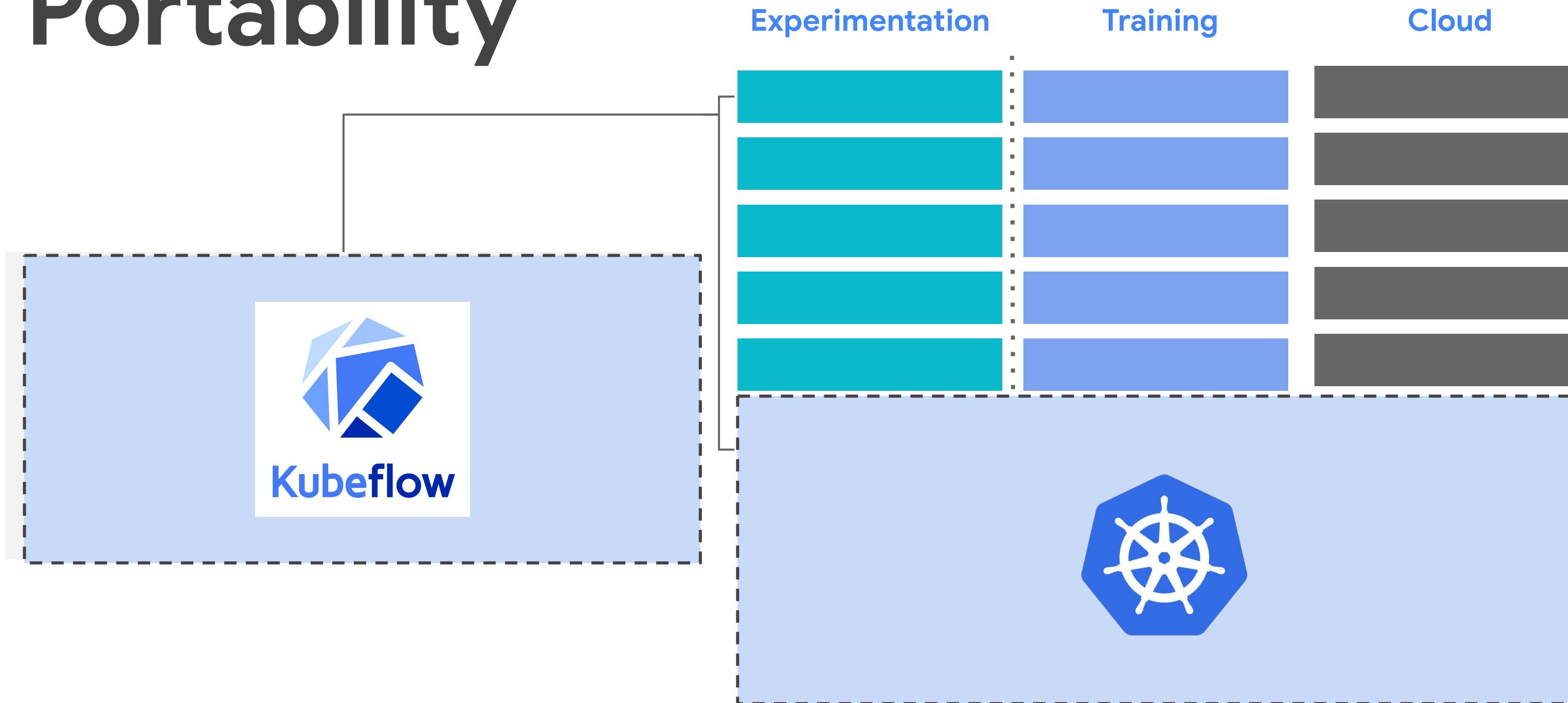
Portability



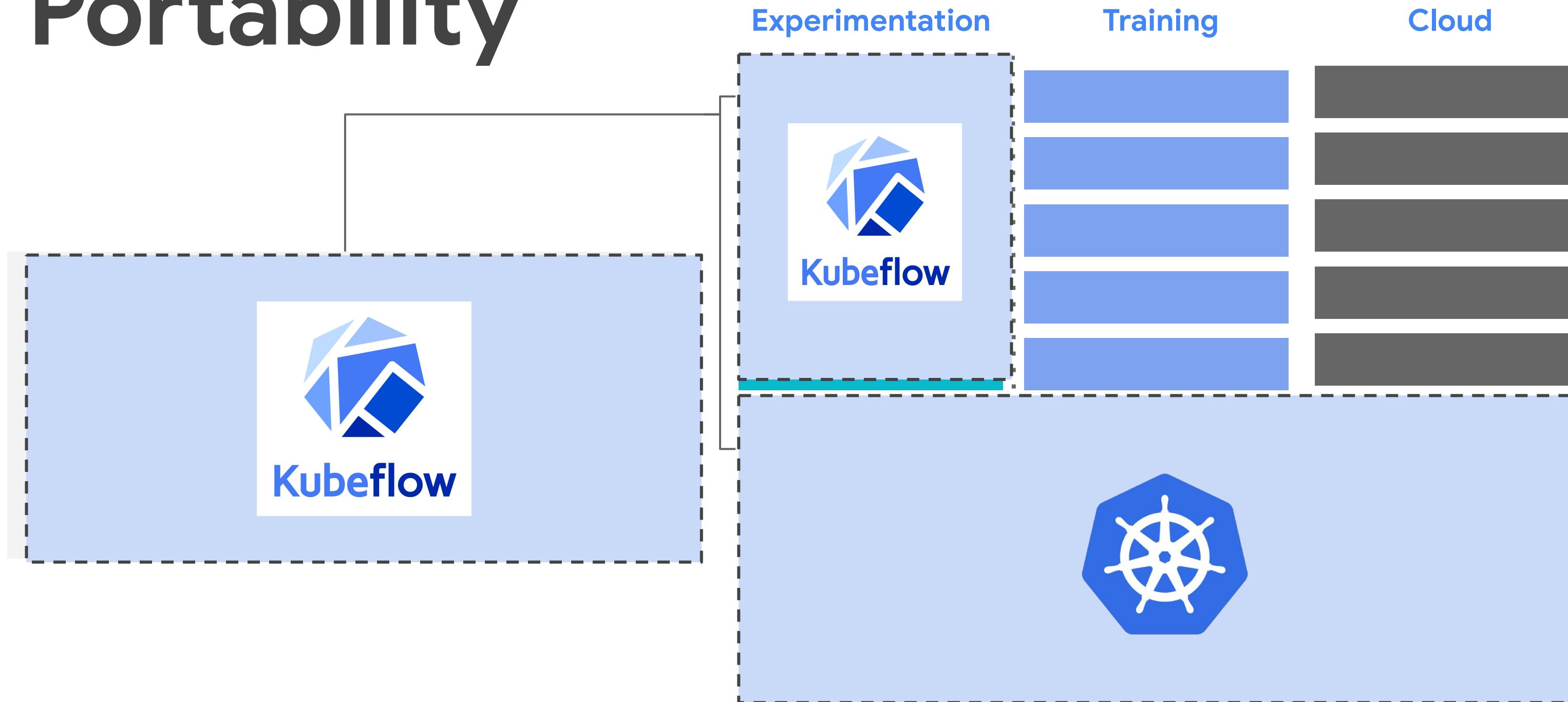
Portability



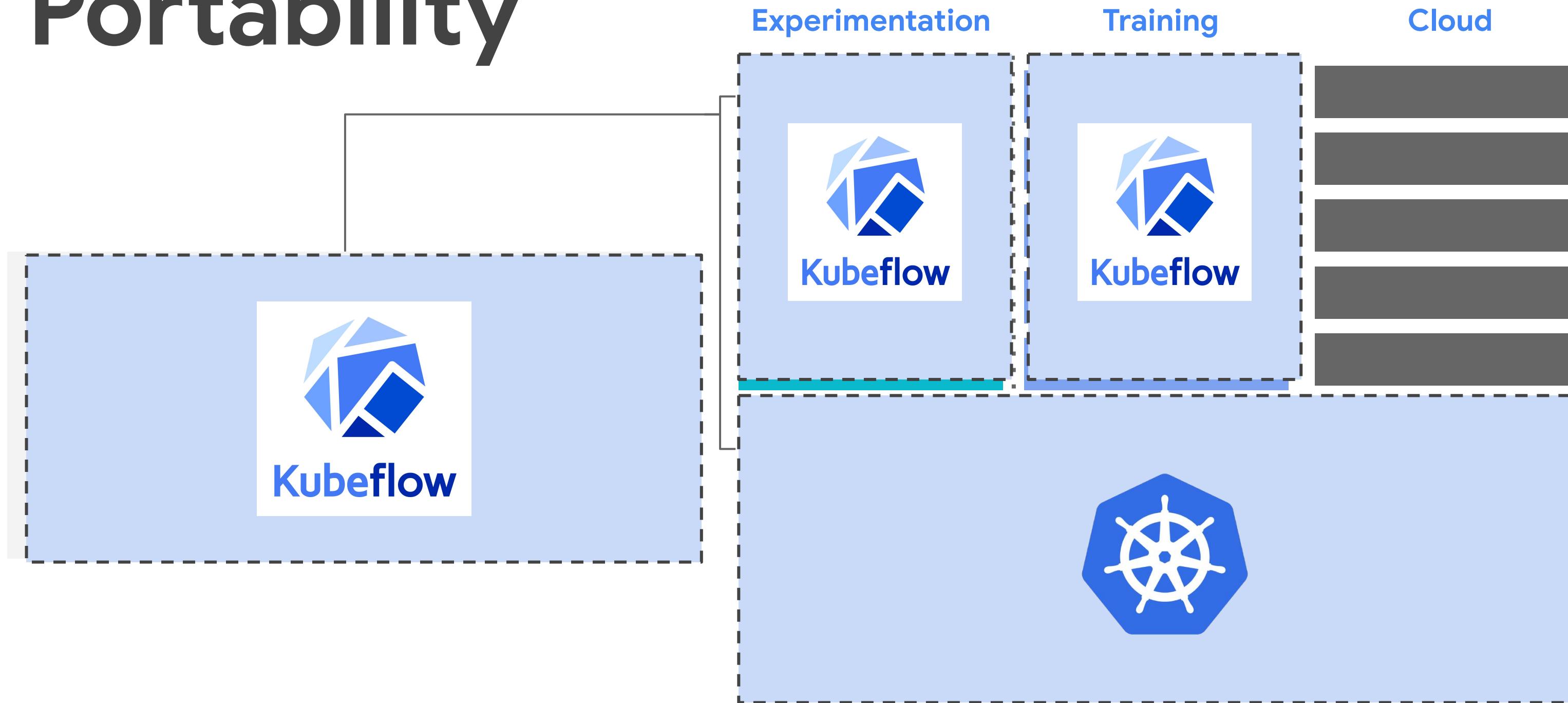
Portability



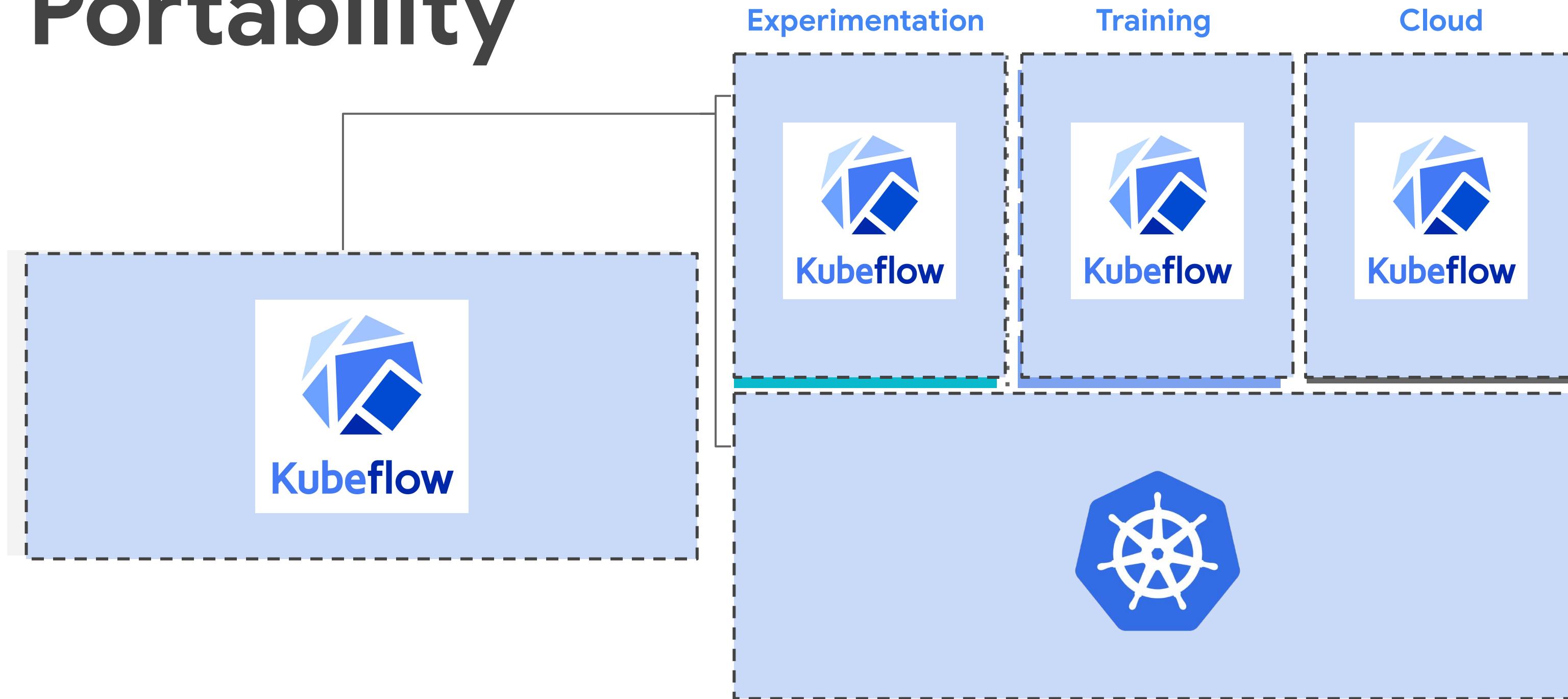
Portability

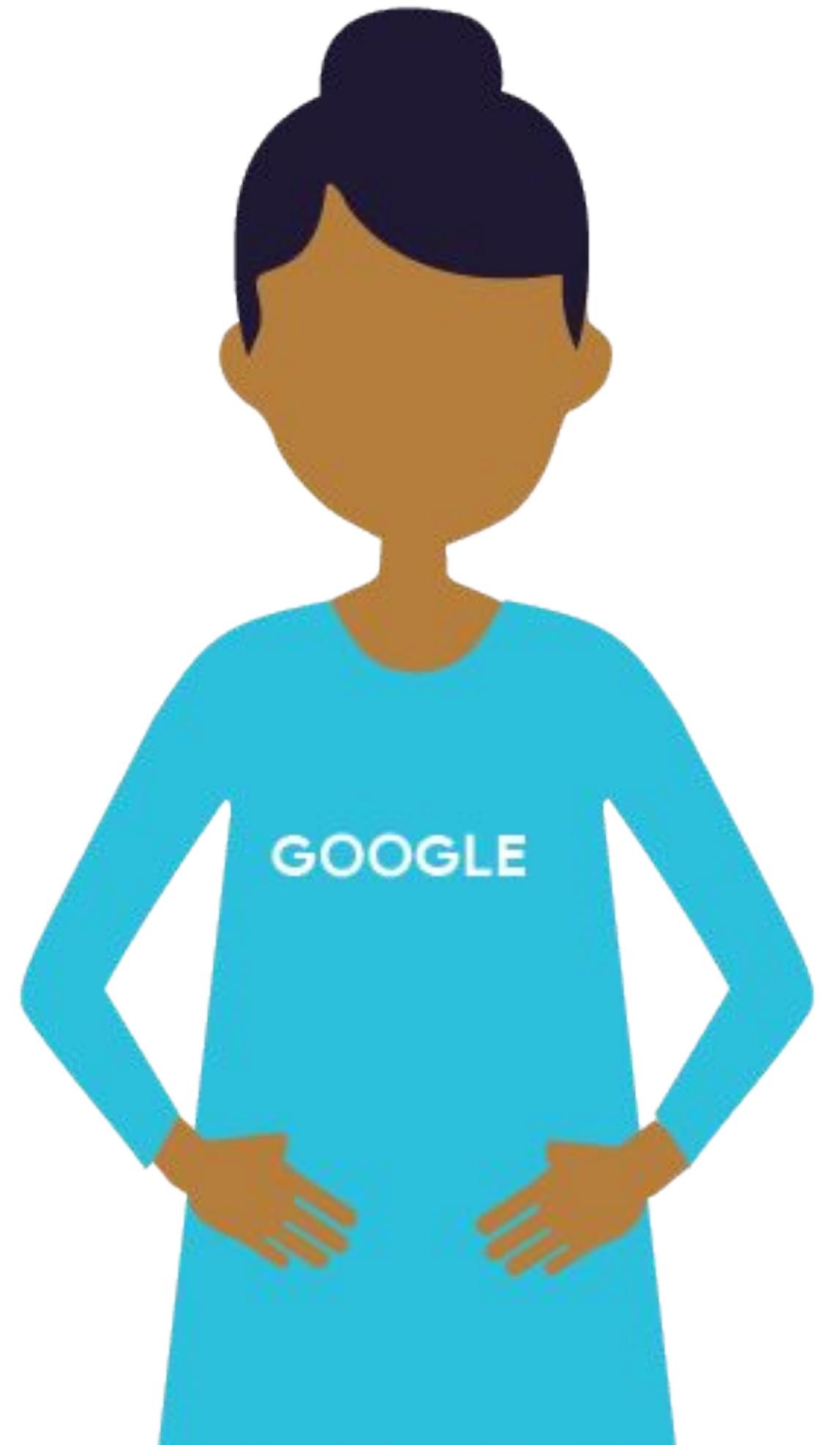


Portability

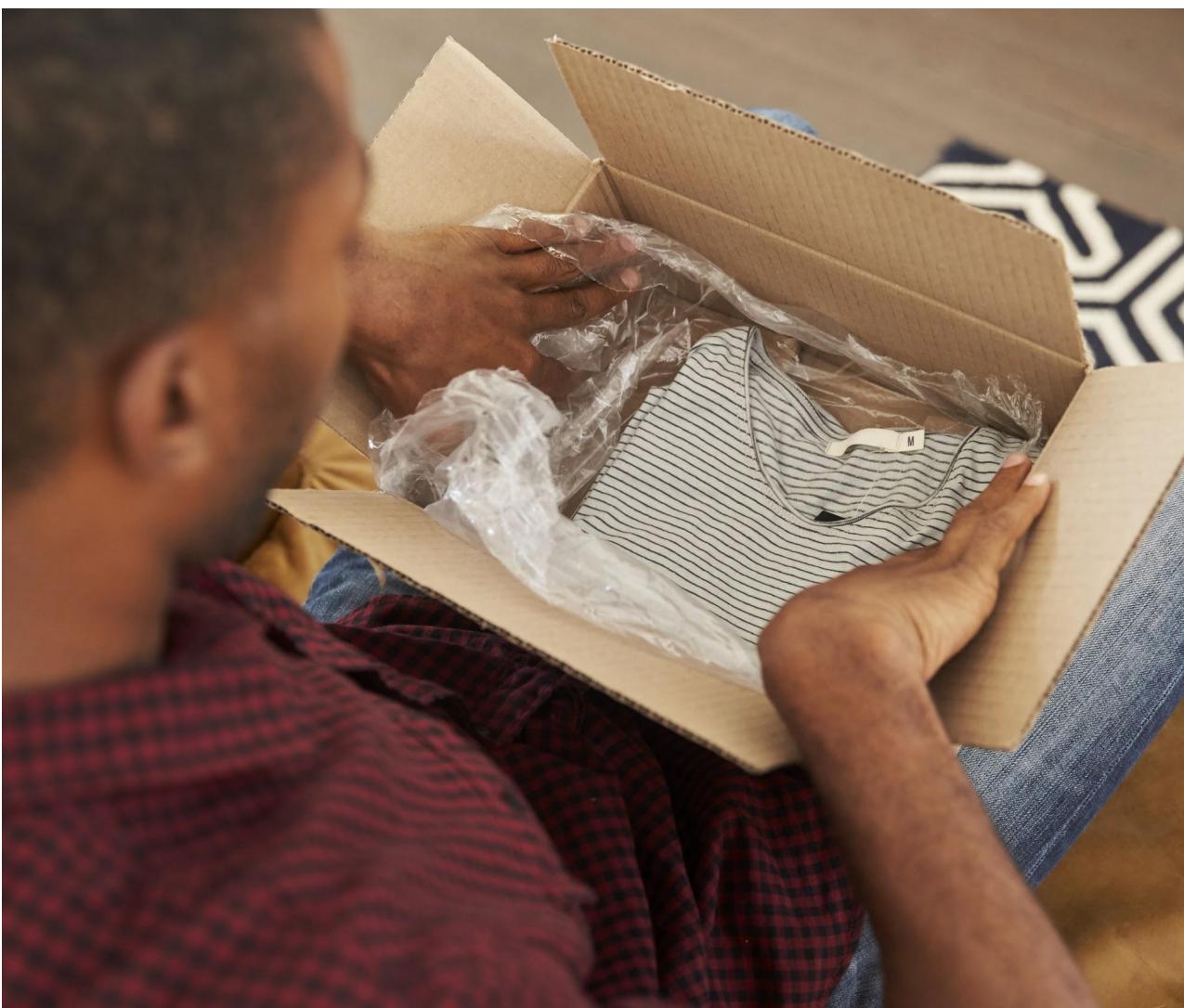


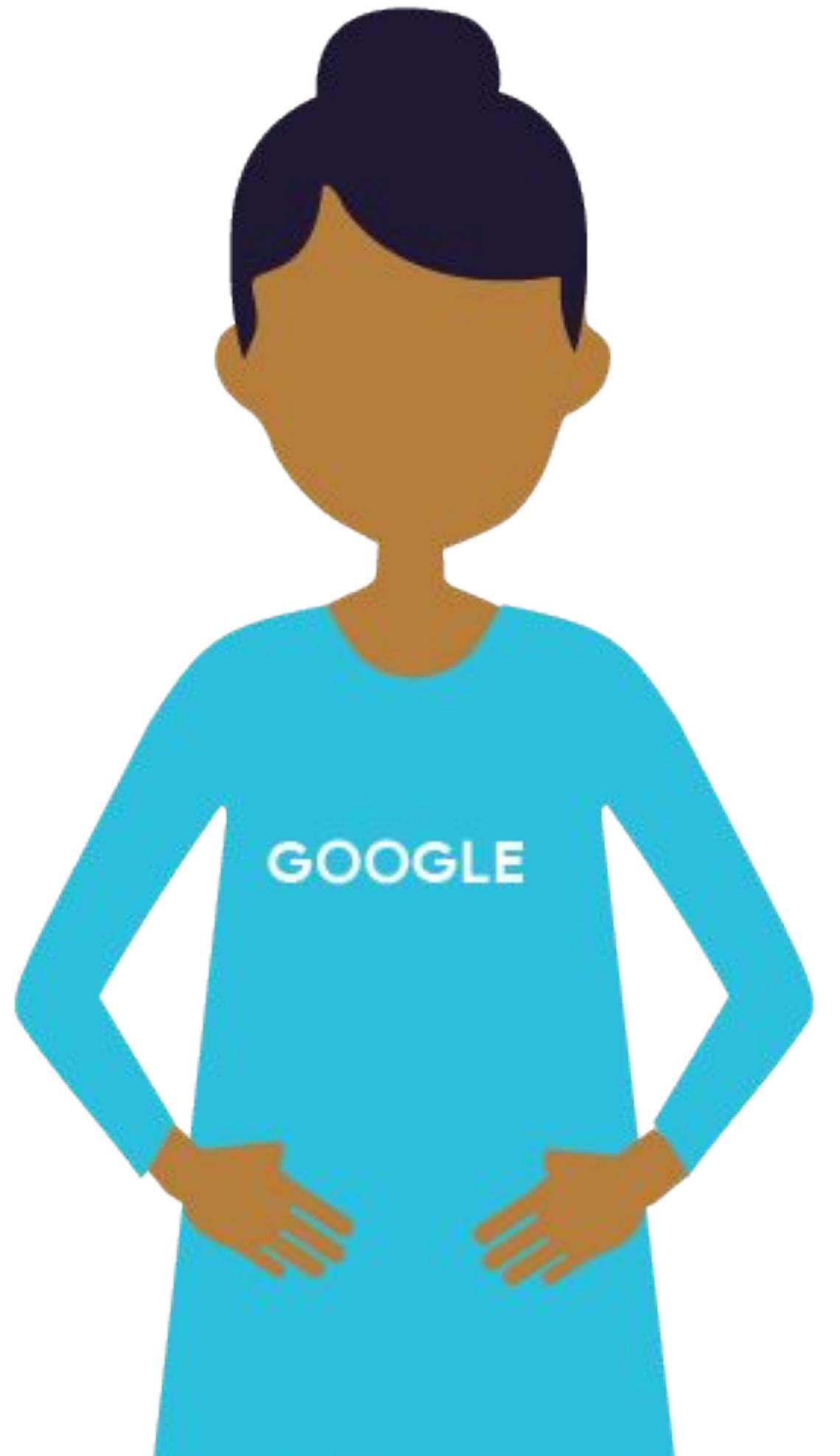
Portability





What's in the box?

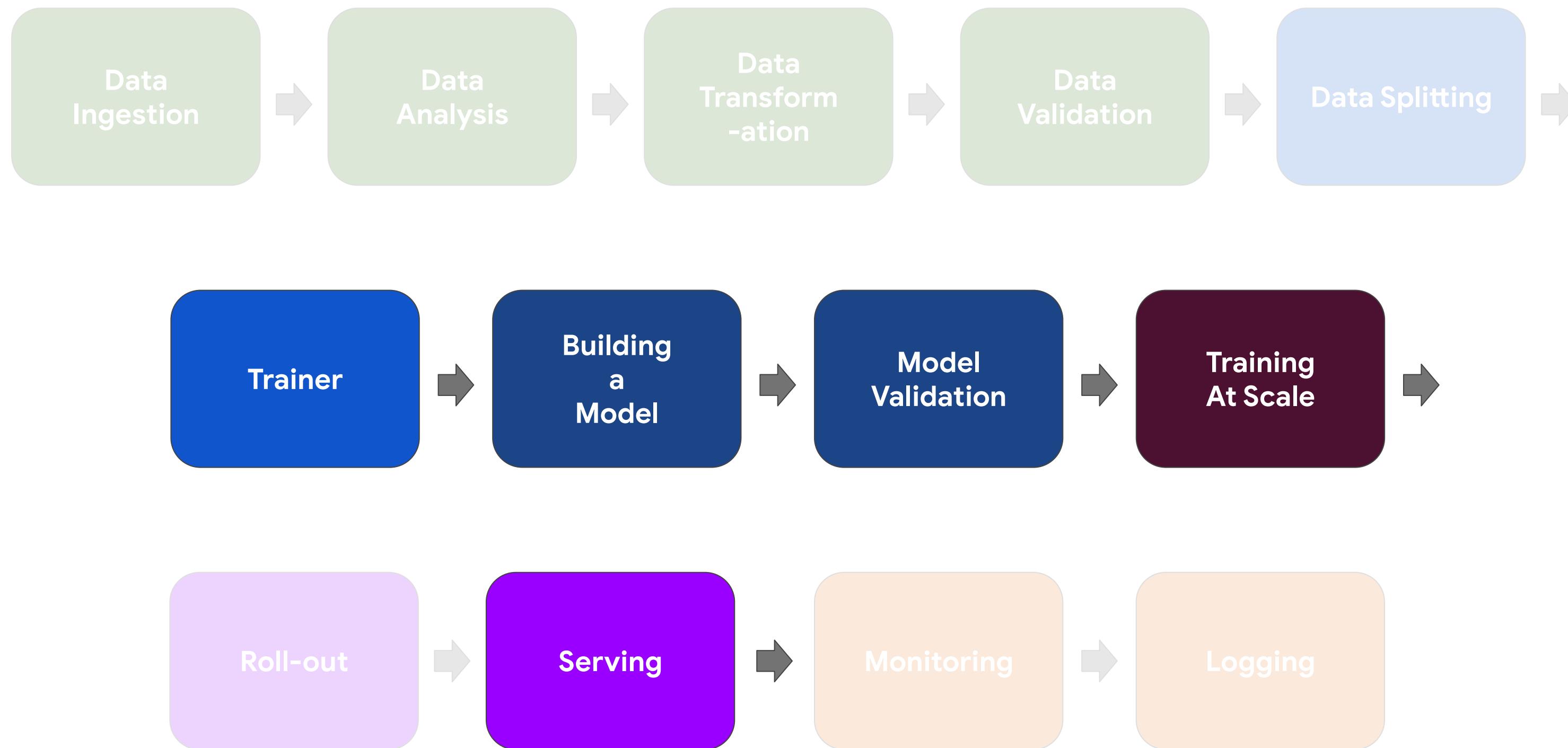




What's in the box?

- 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?



Courses 7 - Production ML Systems

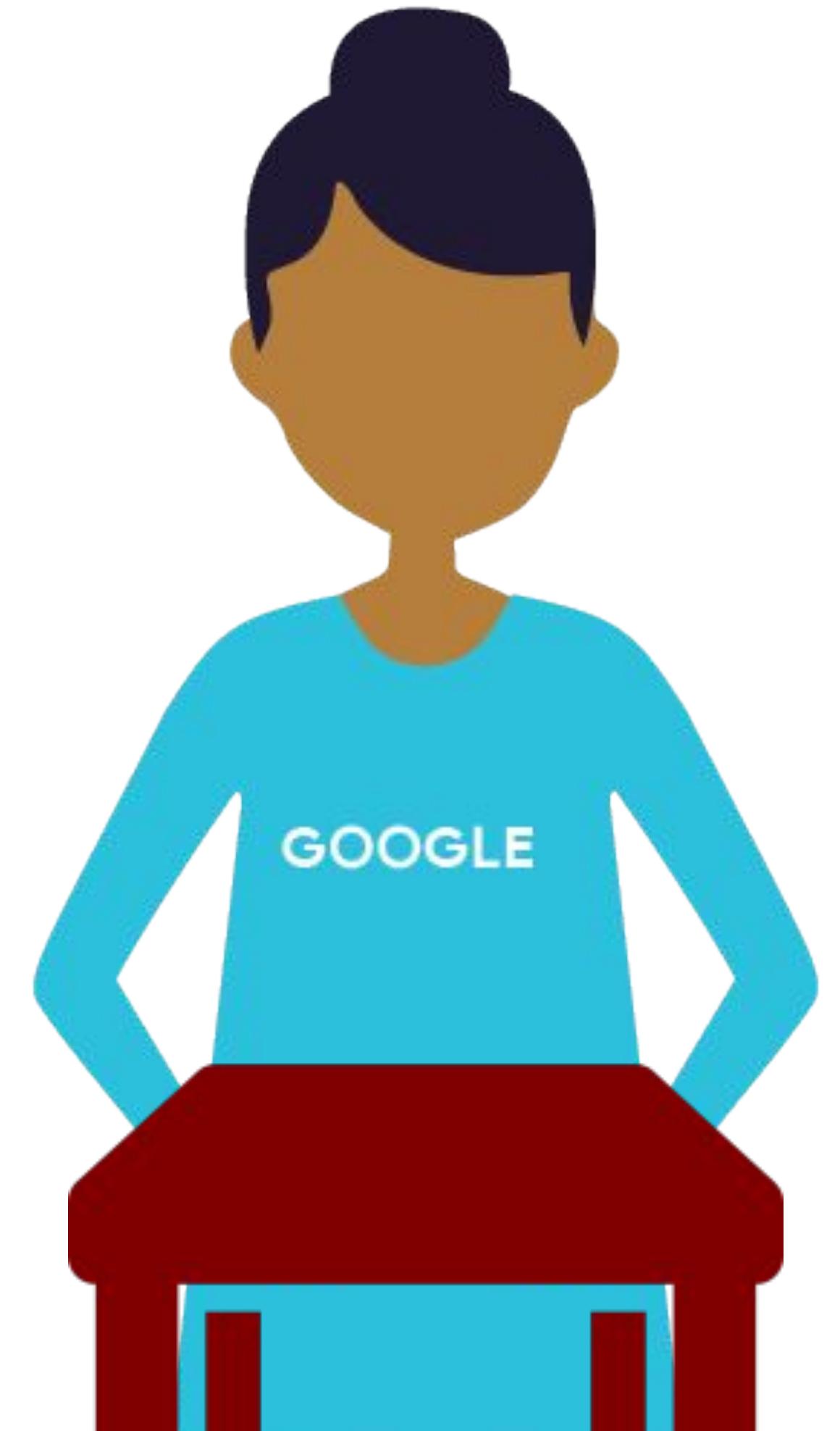
Module 5: Hybrid ML Systems

Lesson Title: **Kubeflow Demo**

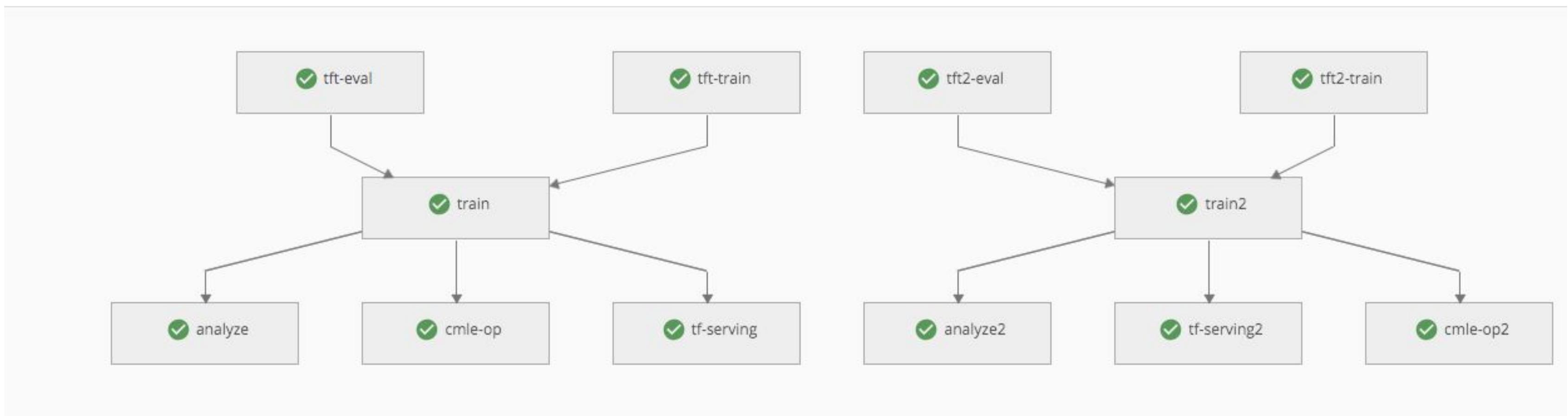
Format: On-Camera Screencast

Presenter: Amy Unruh

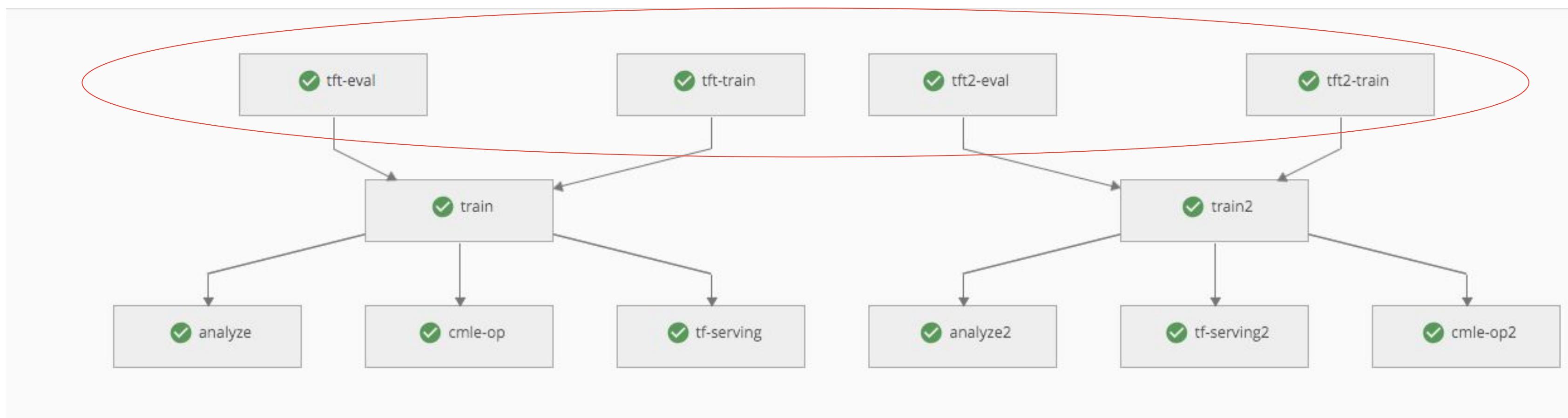
Video Name: T-PSML-O_5_I4_kubeflow_demo

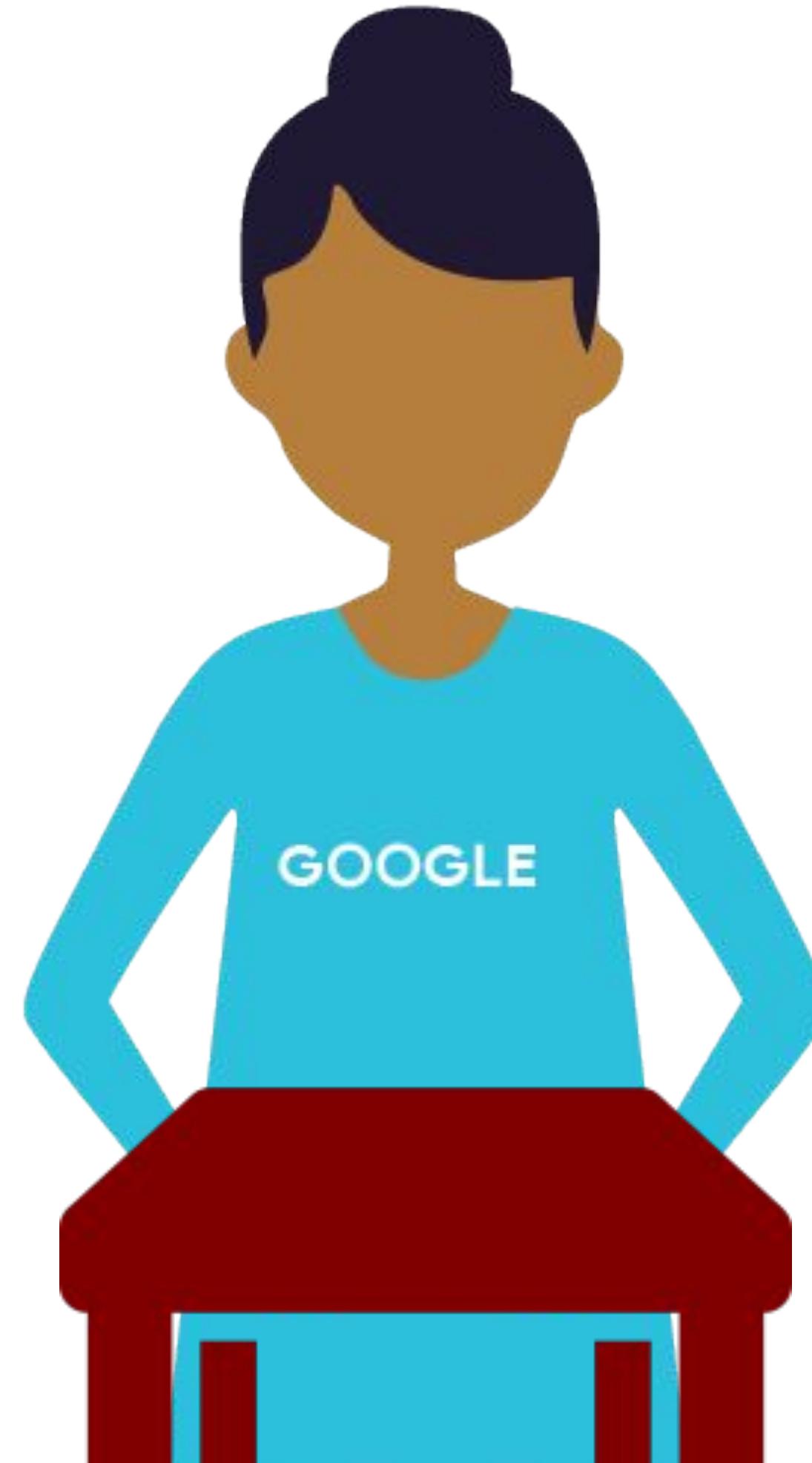


The ML workflow we're going to run



The ML workflow we're going to run





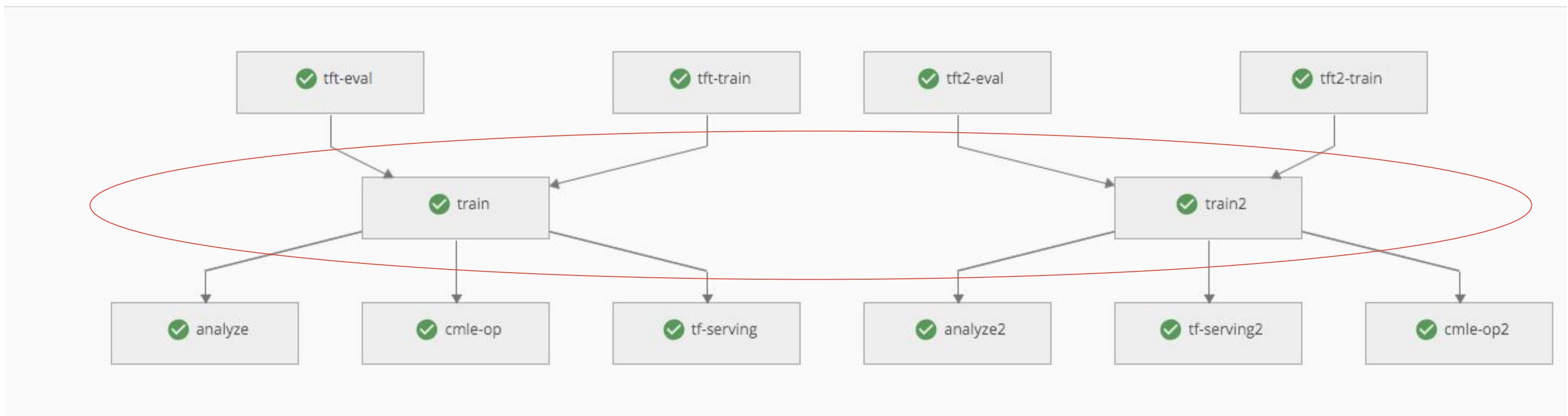
Feature Engineering + Model Analysis



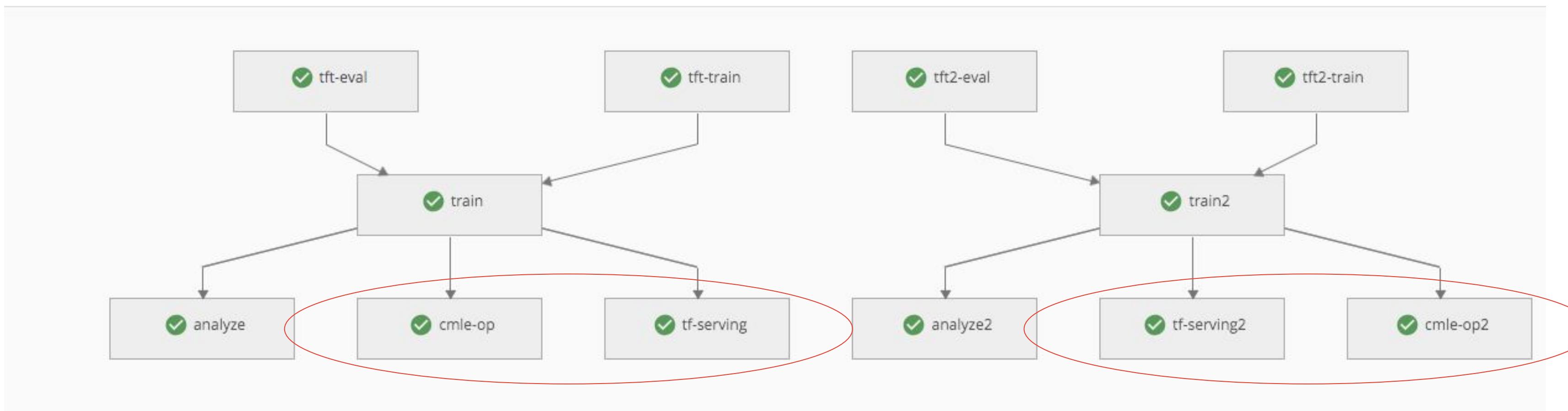
TensorFlow

`tf.transform`

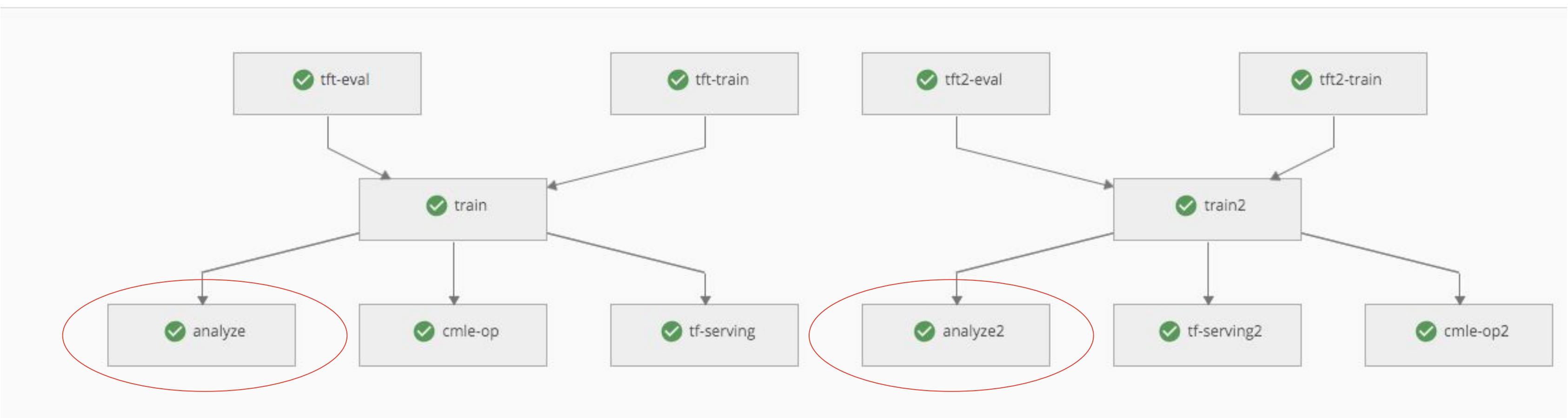
The ML workflow we're going to run

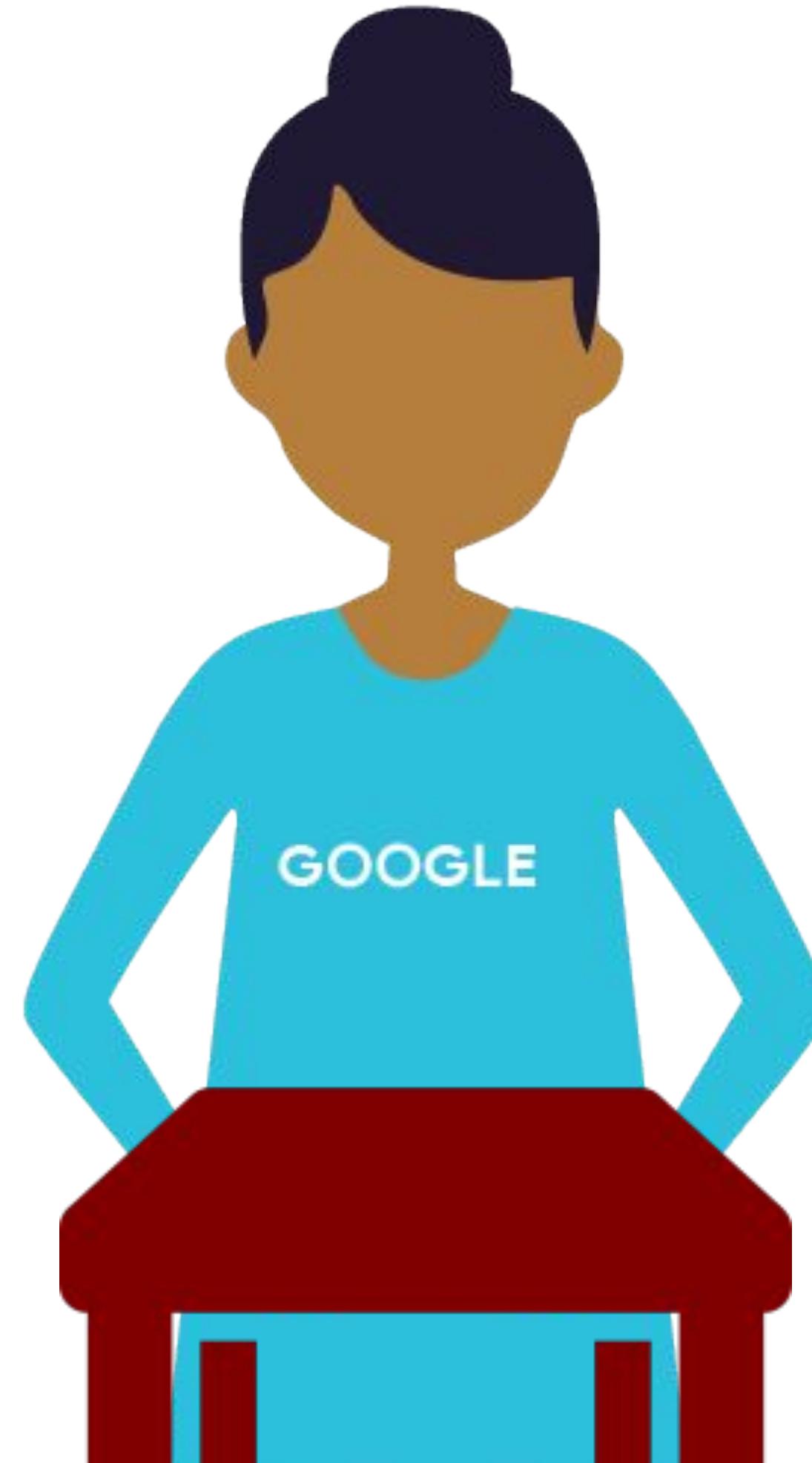


The ML workflow we're going to run



The ML workflow we're going to run

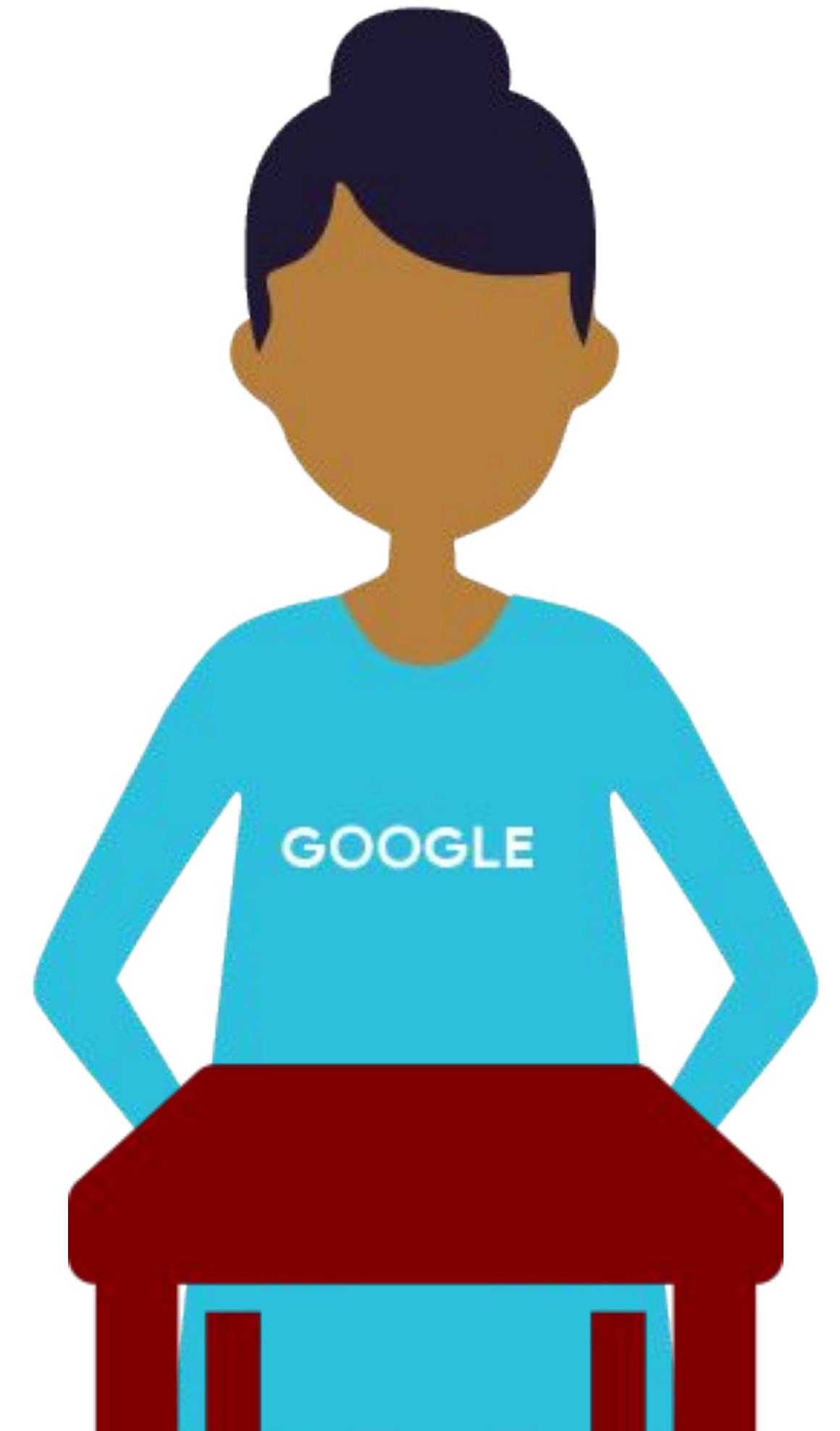


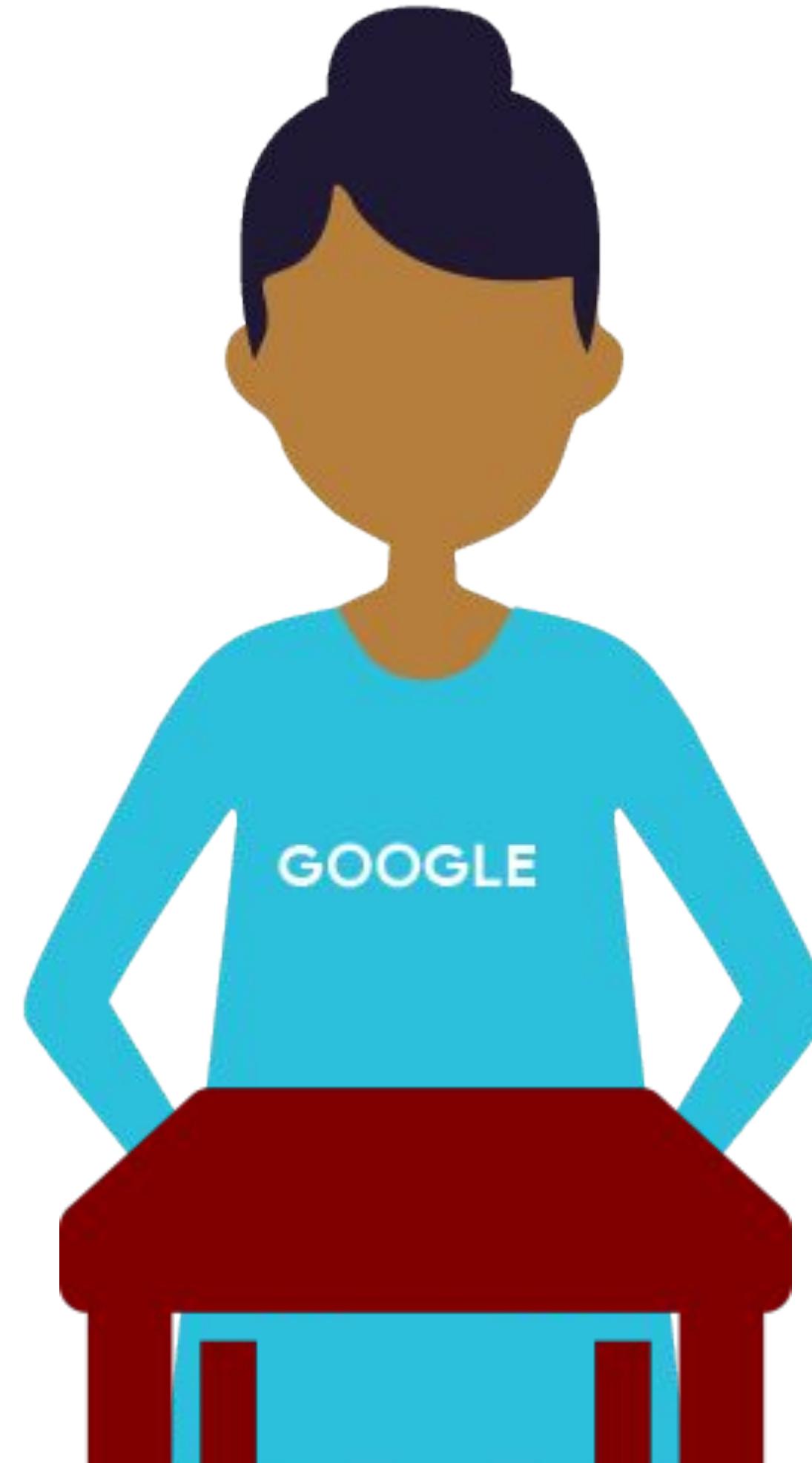


Feature Engineering + Model Analysis



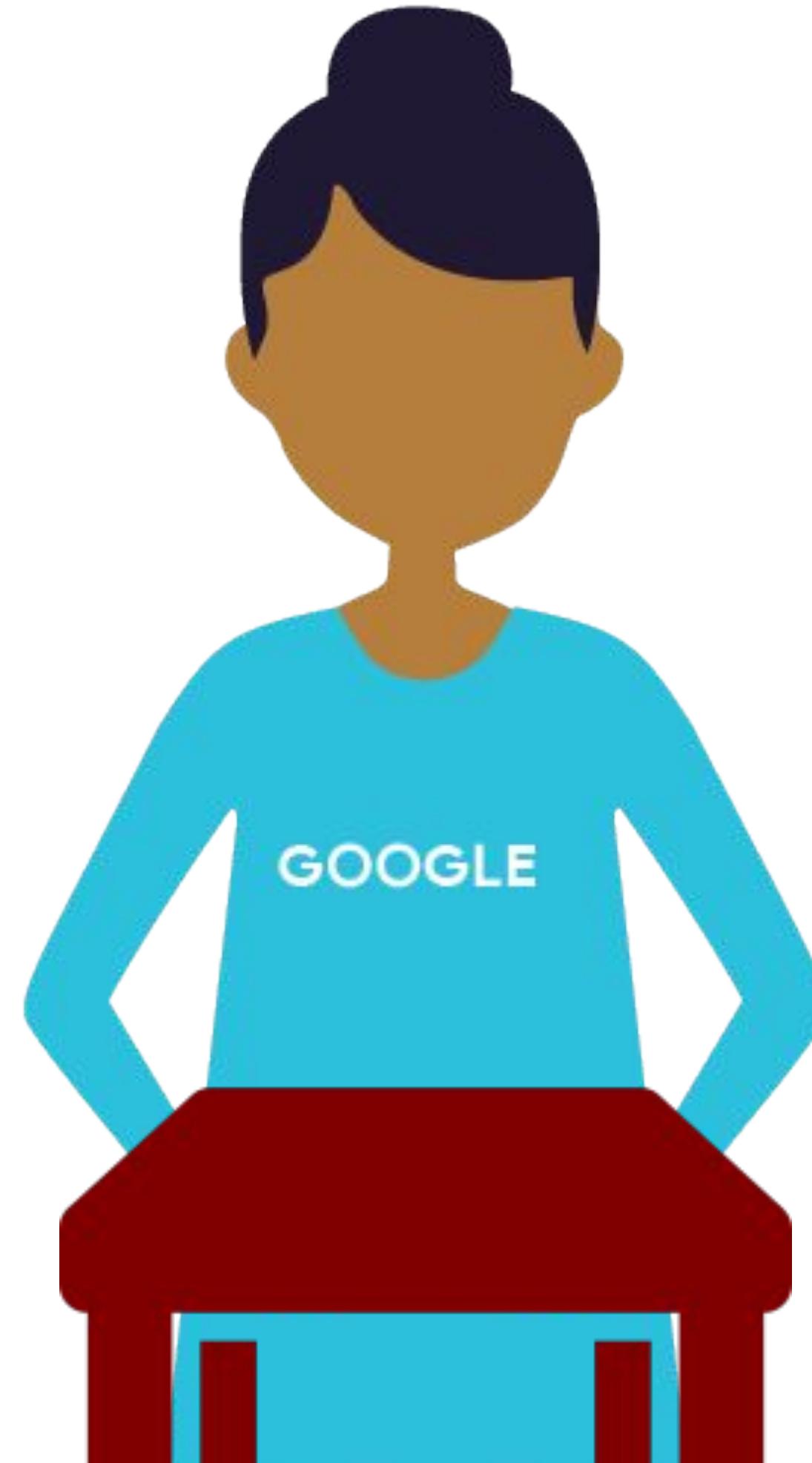
tf.transform
TensorFlow
Model
Analysis





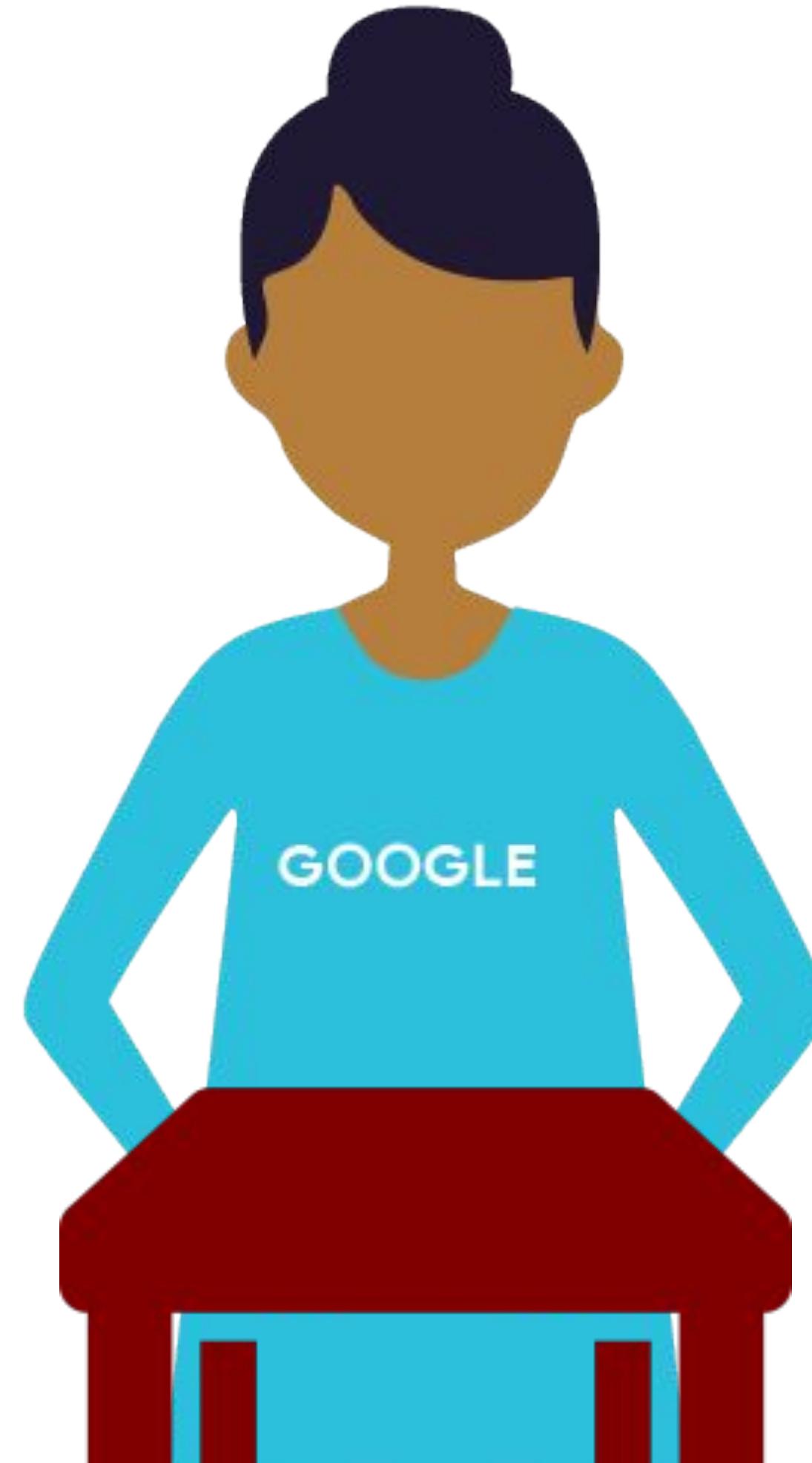
Kubeflow Benefits

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



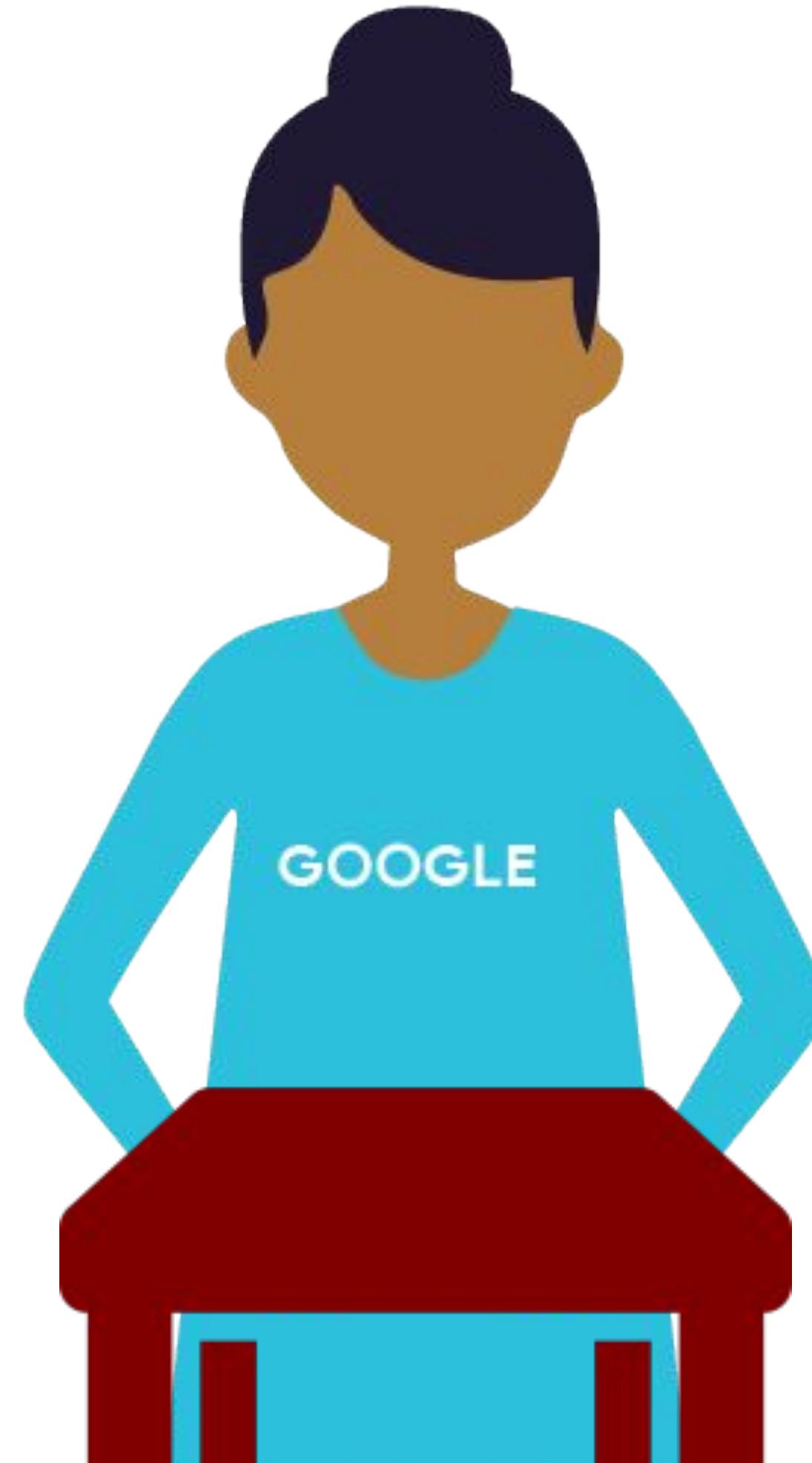
Kubeflow Benefits

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



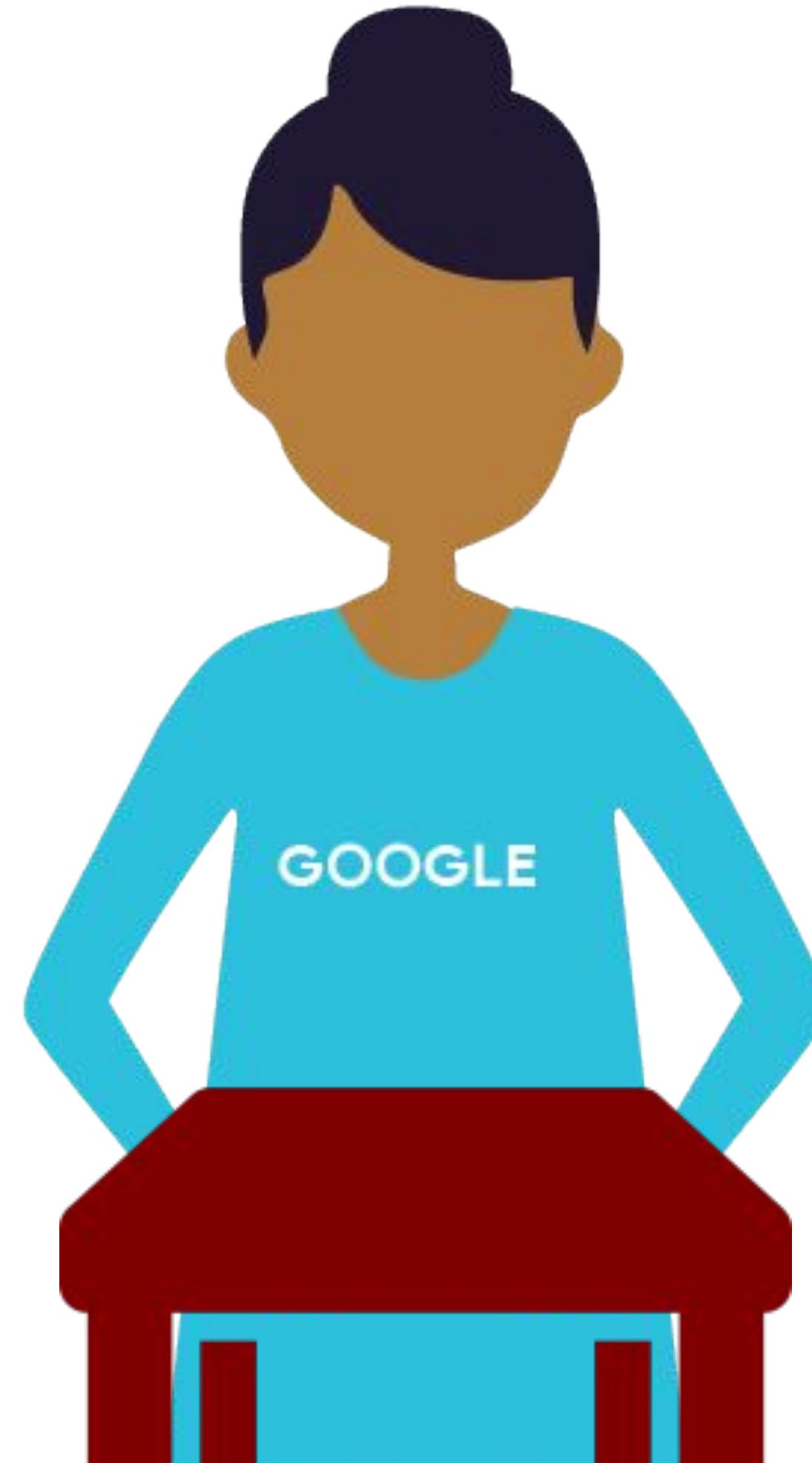
Kubeflow Benefits

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



Kubeflow Benefits

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

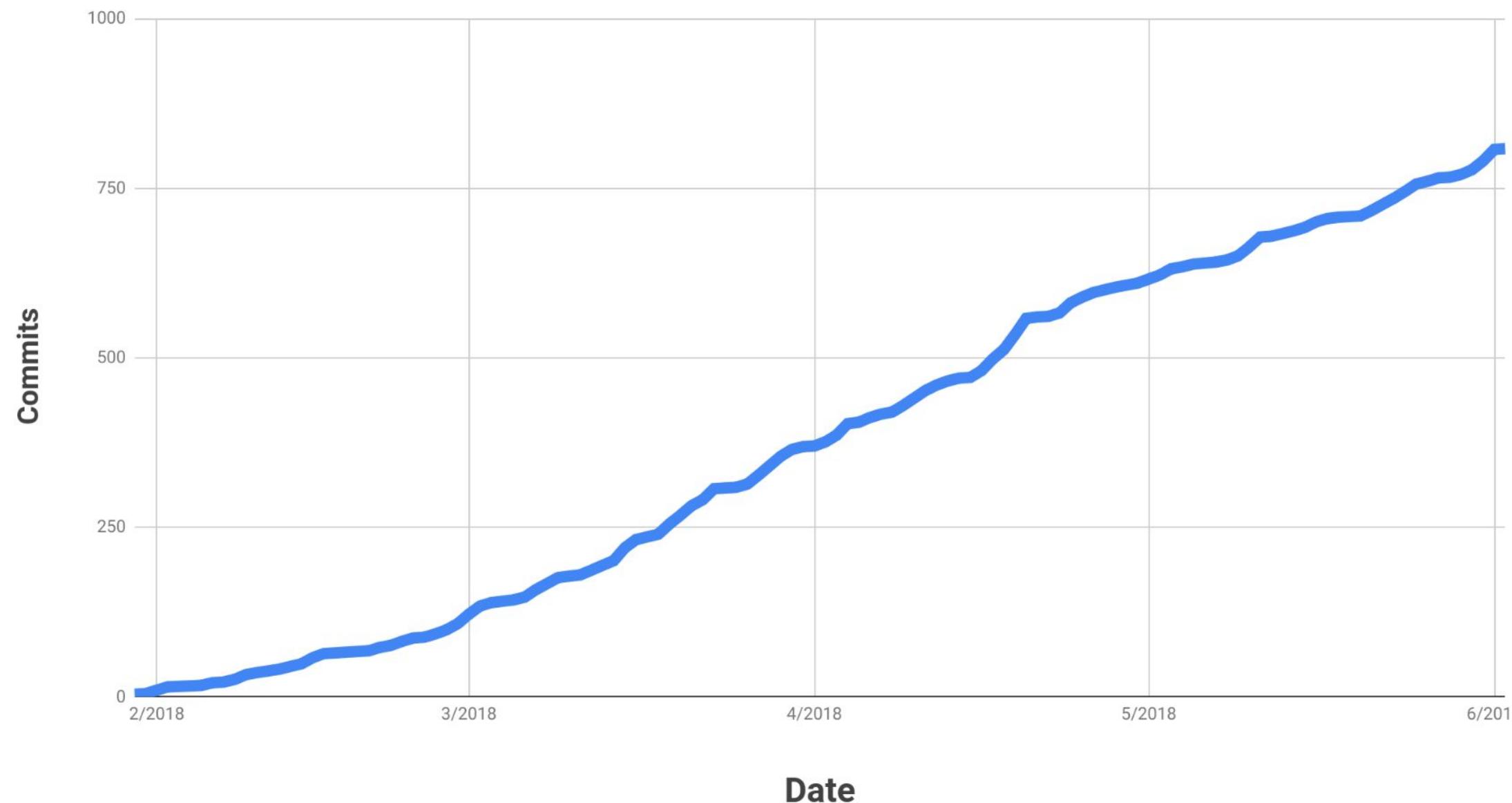


Kubeflow Benefits

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

Momentum!

Commits Since Launch



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



Courses 7 - Production ML Systems

Module 5: Hybrid ML Systems

Lesson Title: **Embedded Models**

Format: Presenter

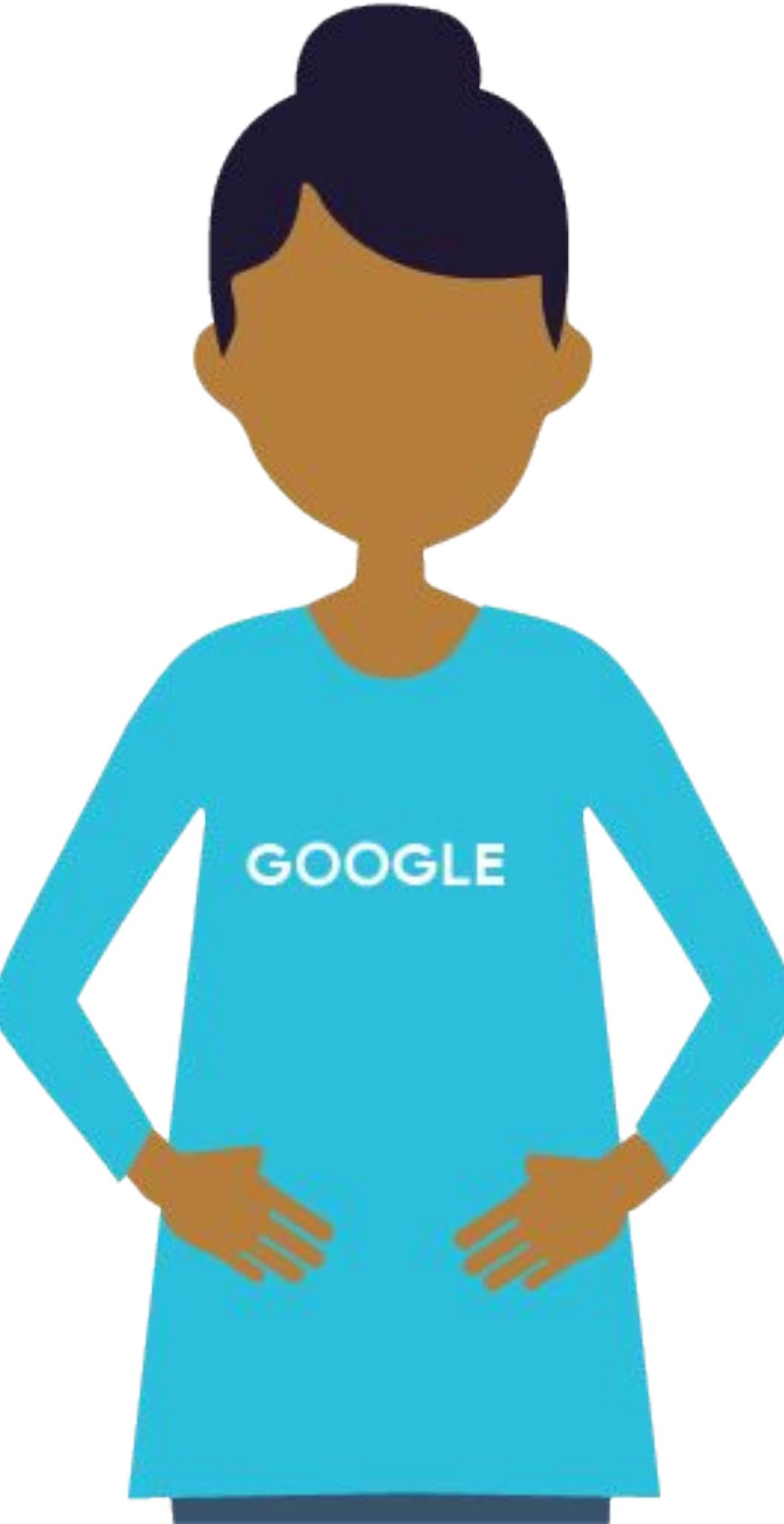
Presenter: Val

Video Name: T-PSML-O_5_I5_embedded_models

Agenda

Kubeflow for hybrid cloud

**Optimizing TensorFlow for
mobile**



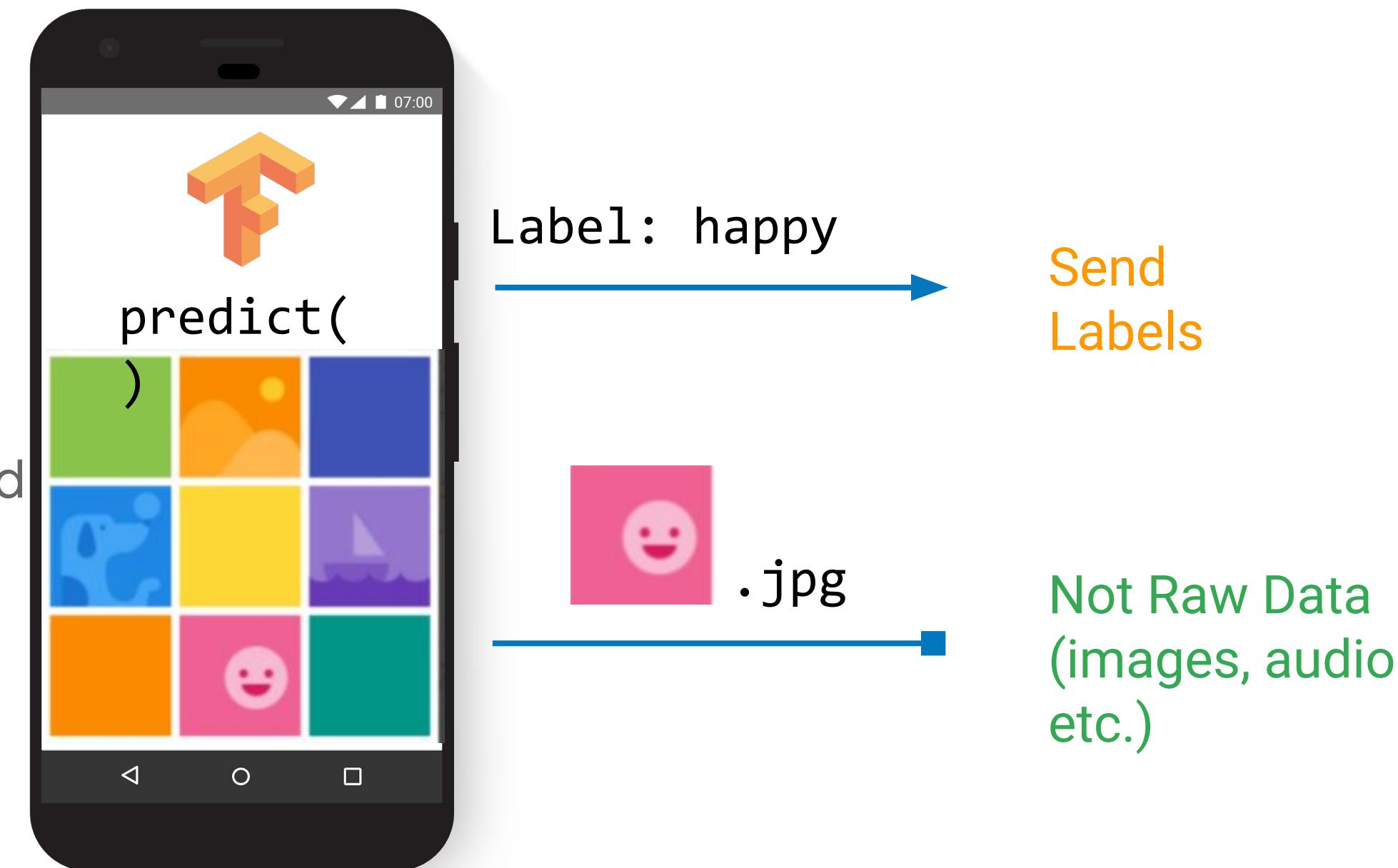
Increasingly, applications are combining ML with mobile apps



- Image/OCR
- Speech ⇔ Text
- Translation

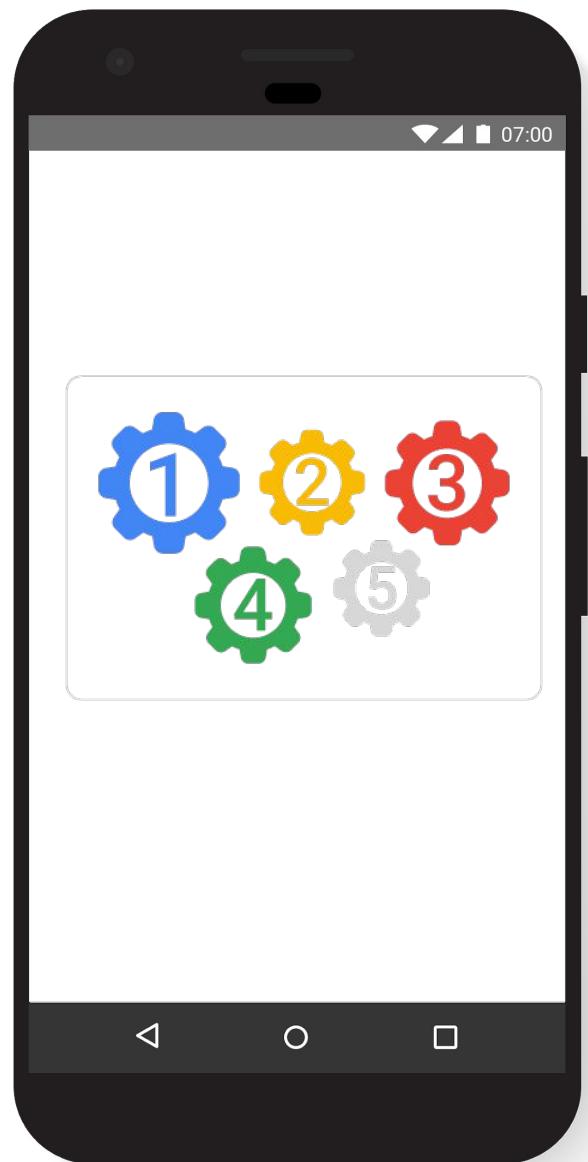
ML models can help extract meaning from raw data,
thus reducing network traffic

- Image recognition: send raw image v. send detected label
- Motion detection: send raw motion v. send feature vector

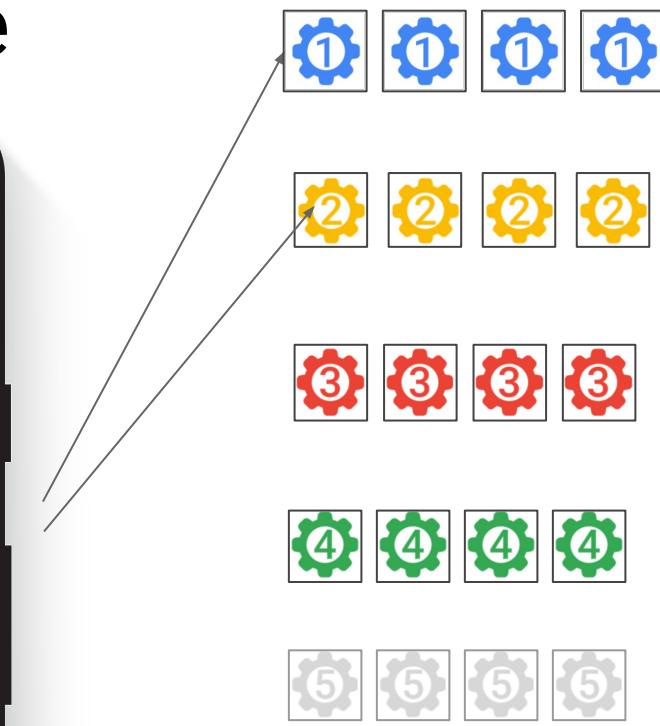
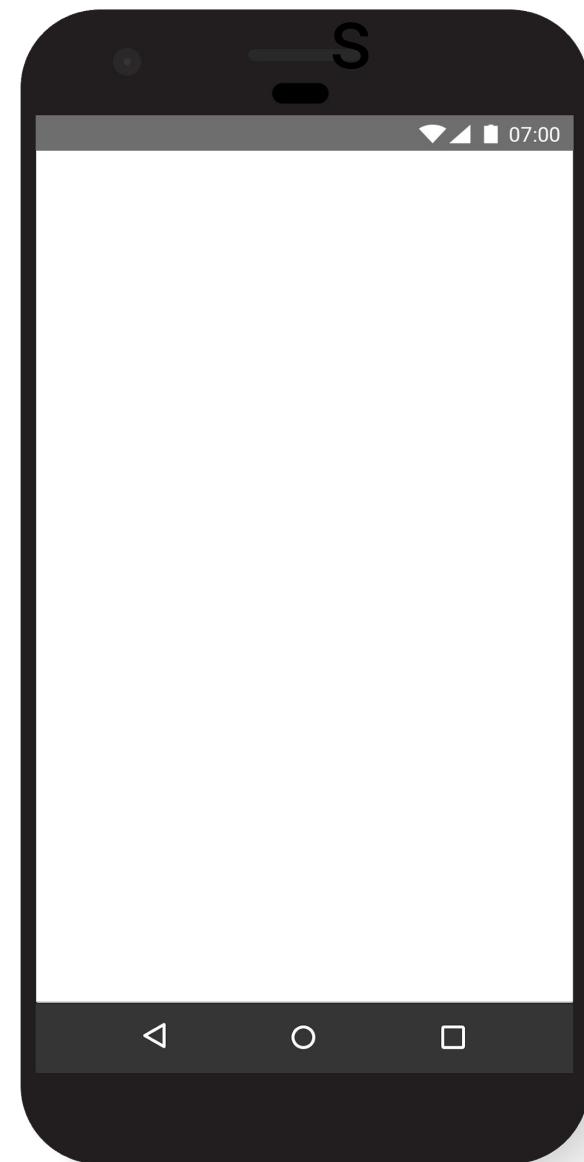


From mobile devices, we often can't
use the microservices approach

Monolithic Service

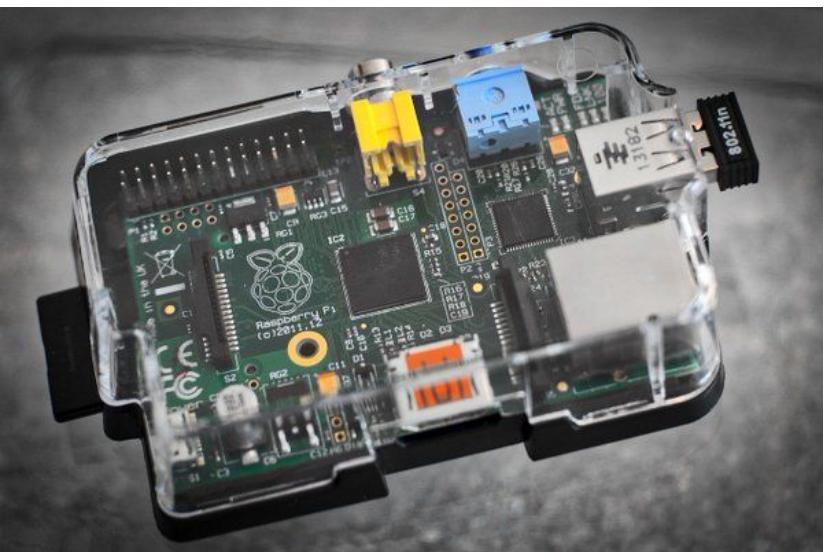


Microservice



Microservices can
add unwanted
latency

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



Courses 7 - Production ML Systems

Module 5: Hybrid ML Systems

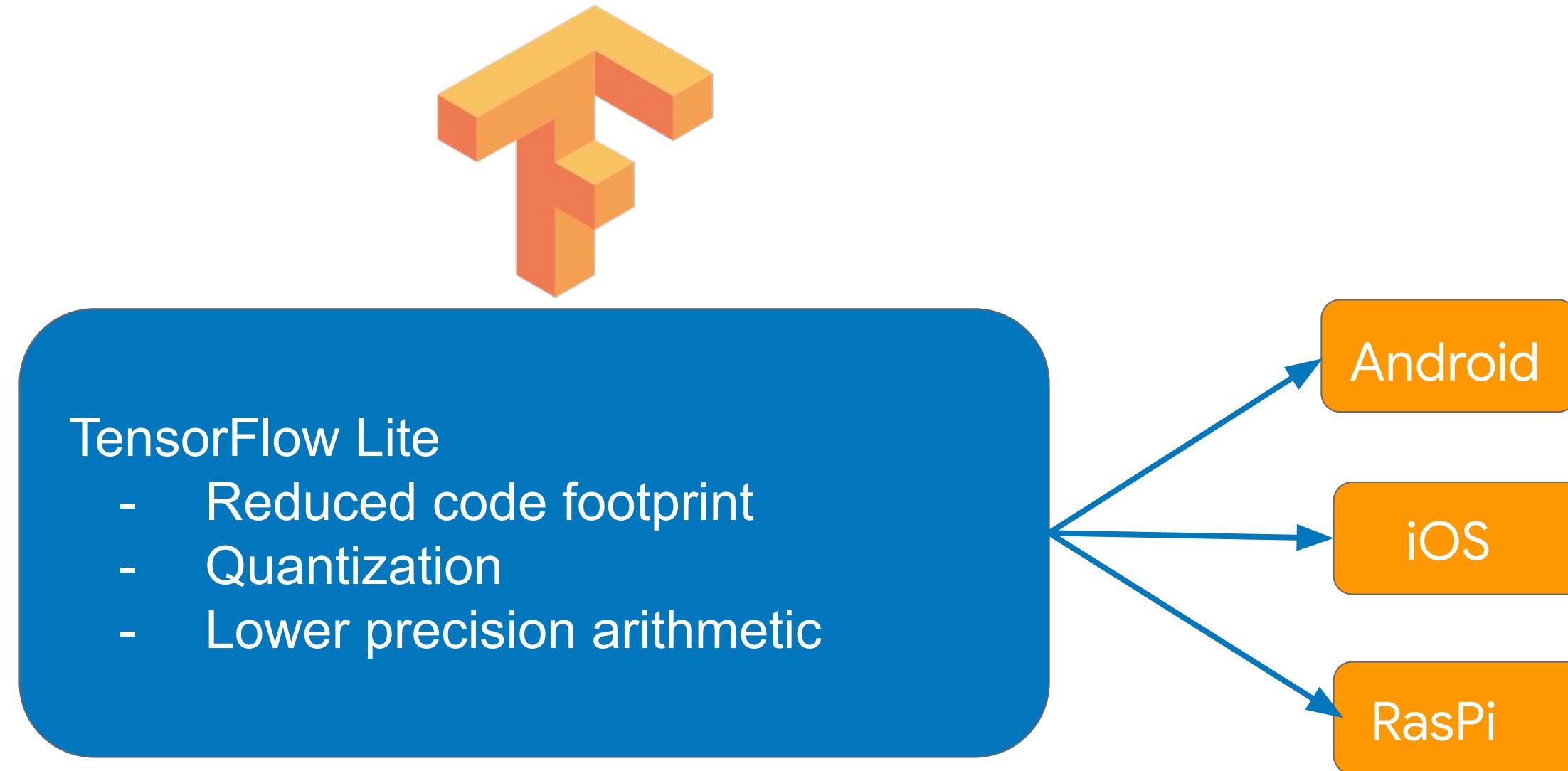
Lesson Title: **TensorFlow Lite**

Format: Presenter

Presenter: Val

Video Name: T-PSML-O_5_l6_tensorflow_lite

TensorFlow supports multiple mobile platforms



Build with Bazel by starting with a git clone

Install:

TensorFlow

Bazel

Android Studio

(optional)

Android SDK

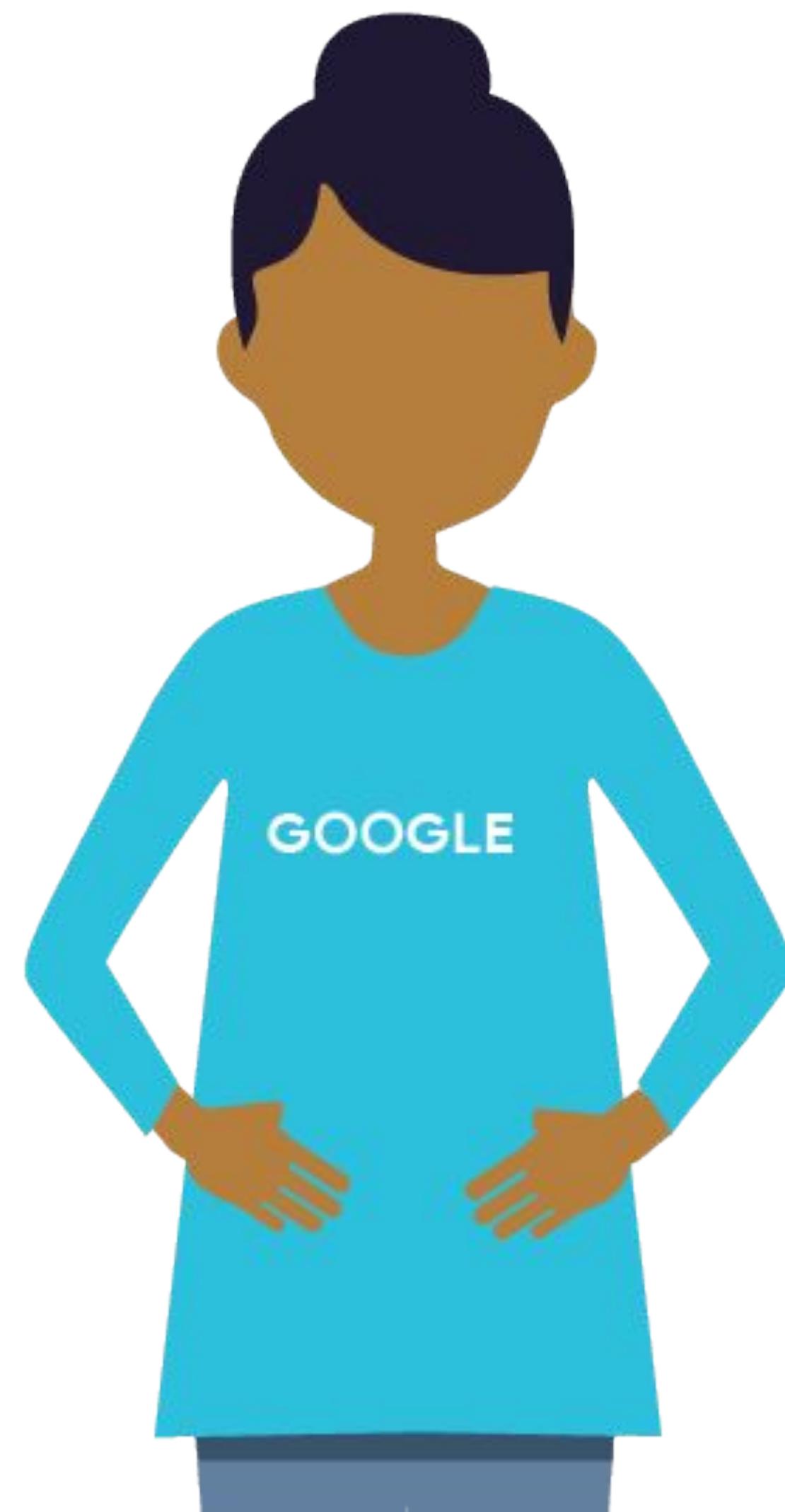
Android NDK

Config:

Edit

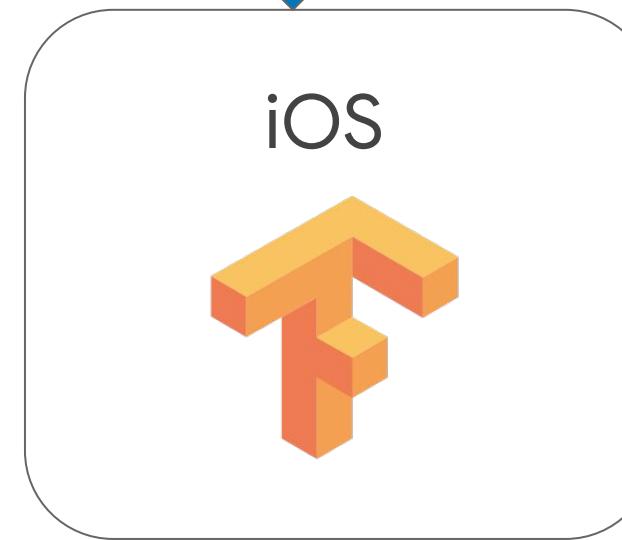
tensorflow/WORKSPACE

```
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

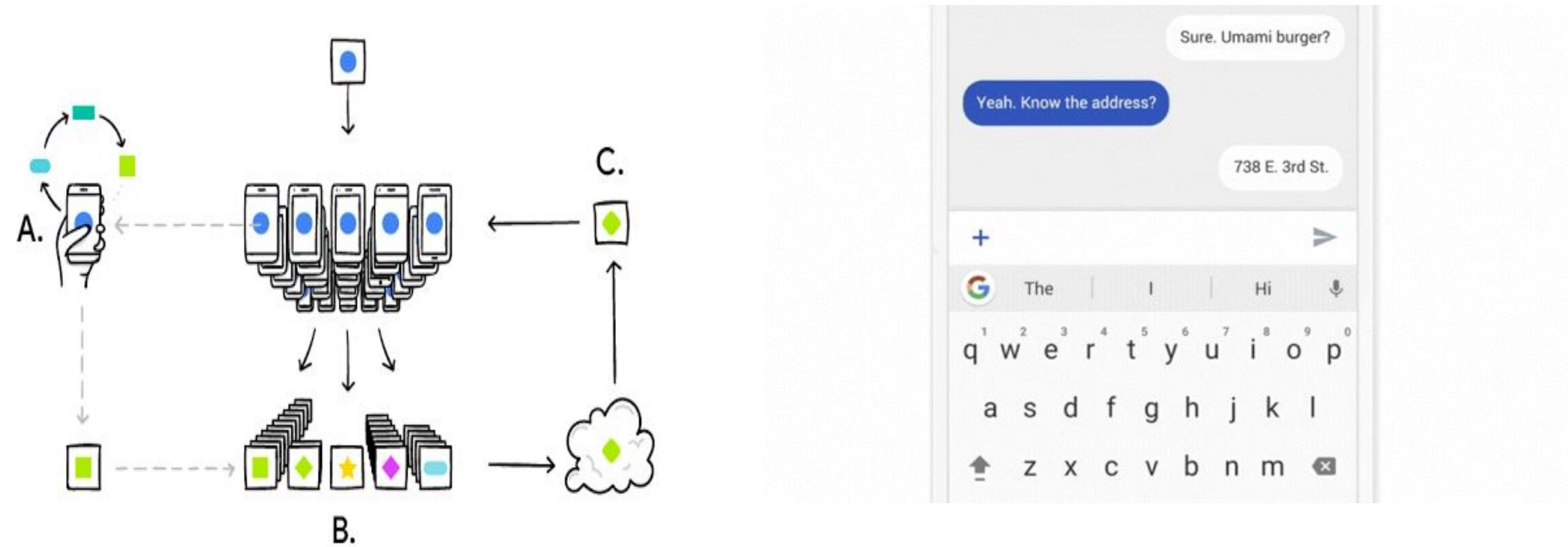
```
CocoaPod  
Podfile  
target 'MyApp'  
pod  
'TensorFlow-experimental'
```



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



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

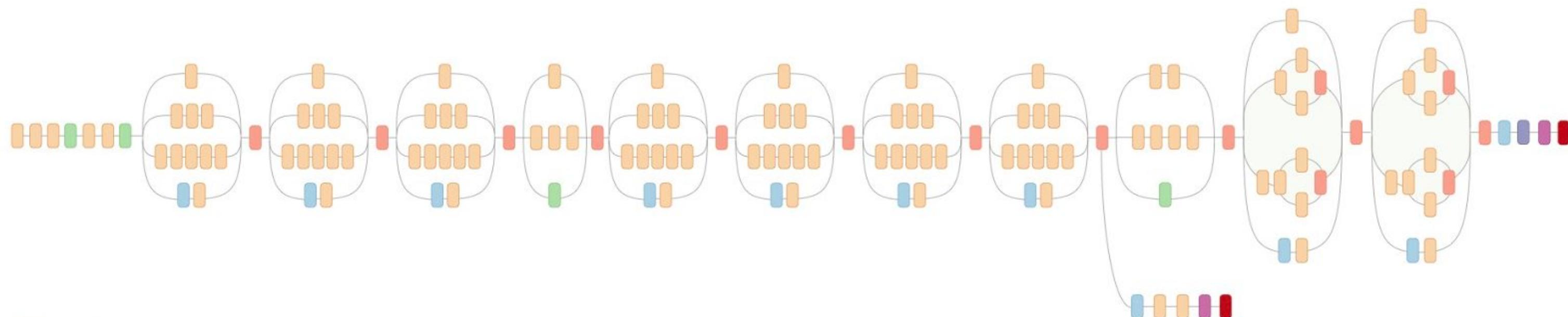
Lesson Title: Optimizing for Mobile

Format: Presenter

Presenter: Val

Video Name: T-PSML-O_5_I7_optimizing_for_mobile

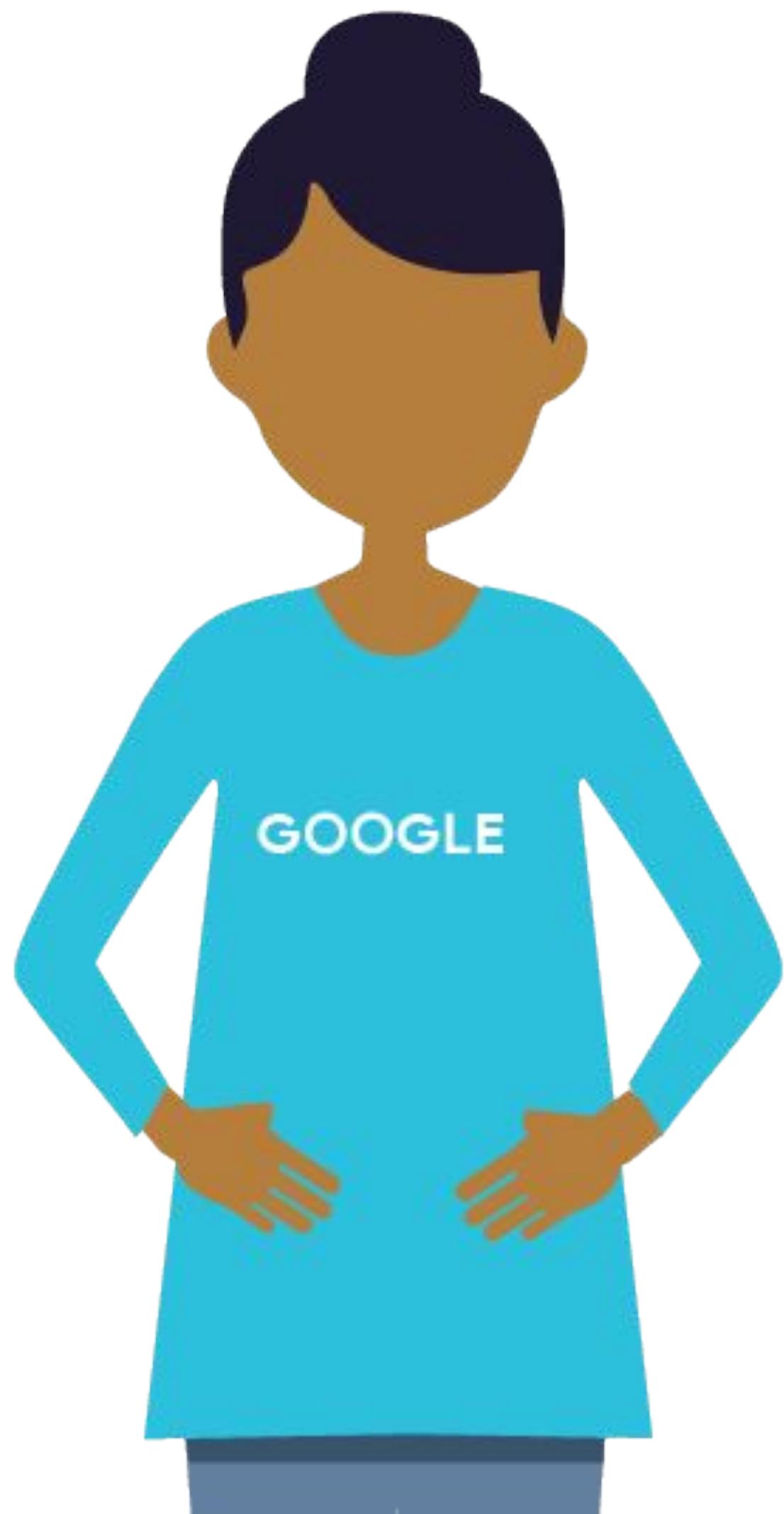
Large neural networks can be compressed



- Convolution
- AvgPool
- MaxPool
- Concat
- Dropout
- Fully connected
- Softmax

The Inception v3 model = **91 MB**
TensorFlow binary = **12 MB**

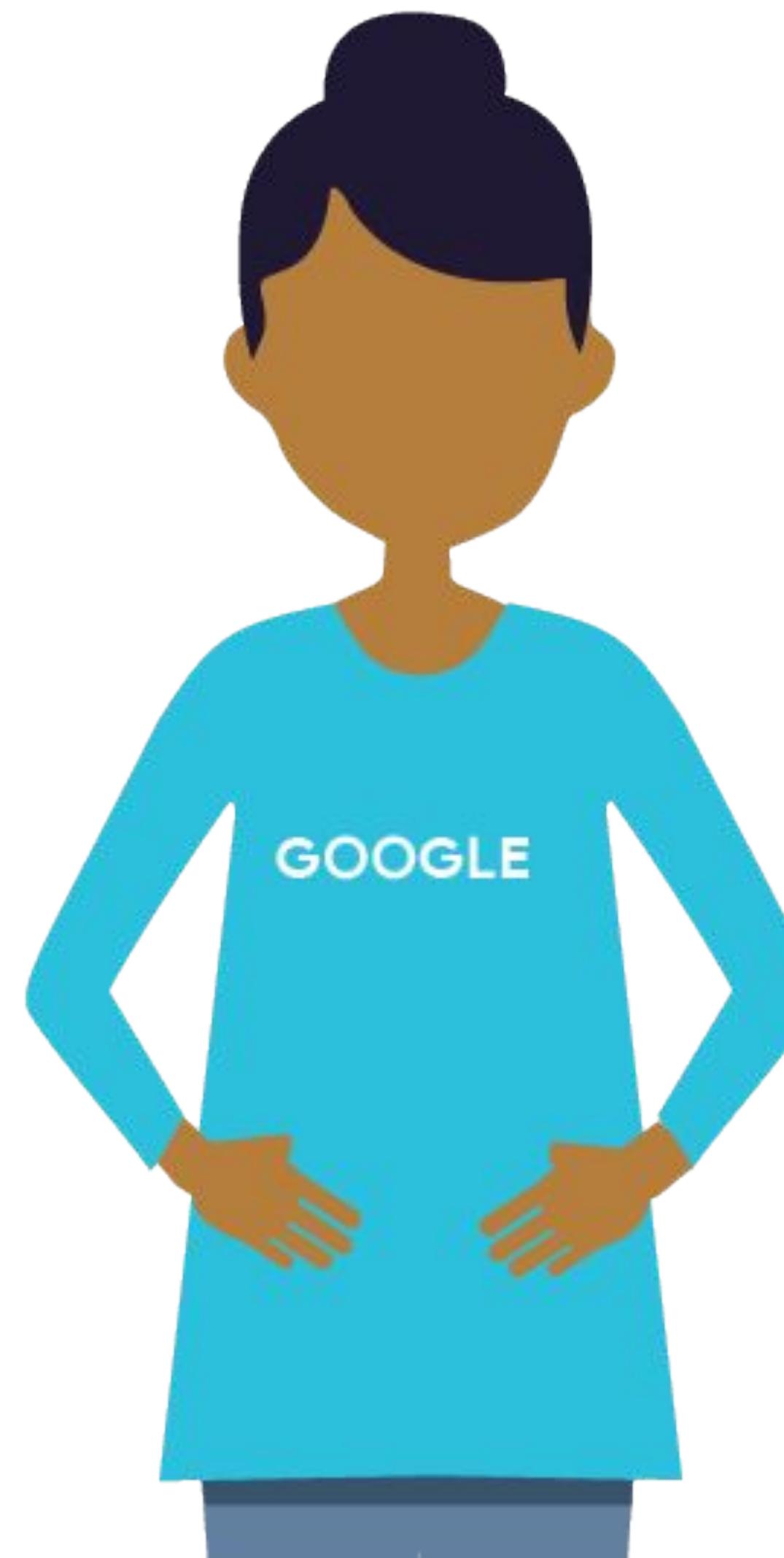
75%
smaller



There are several methods to
reduce model size



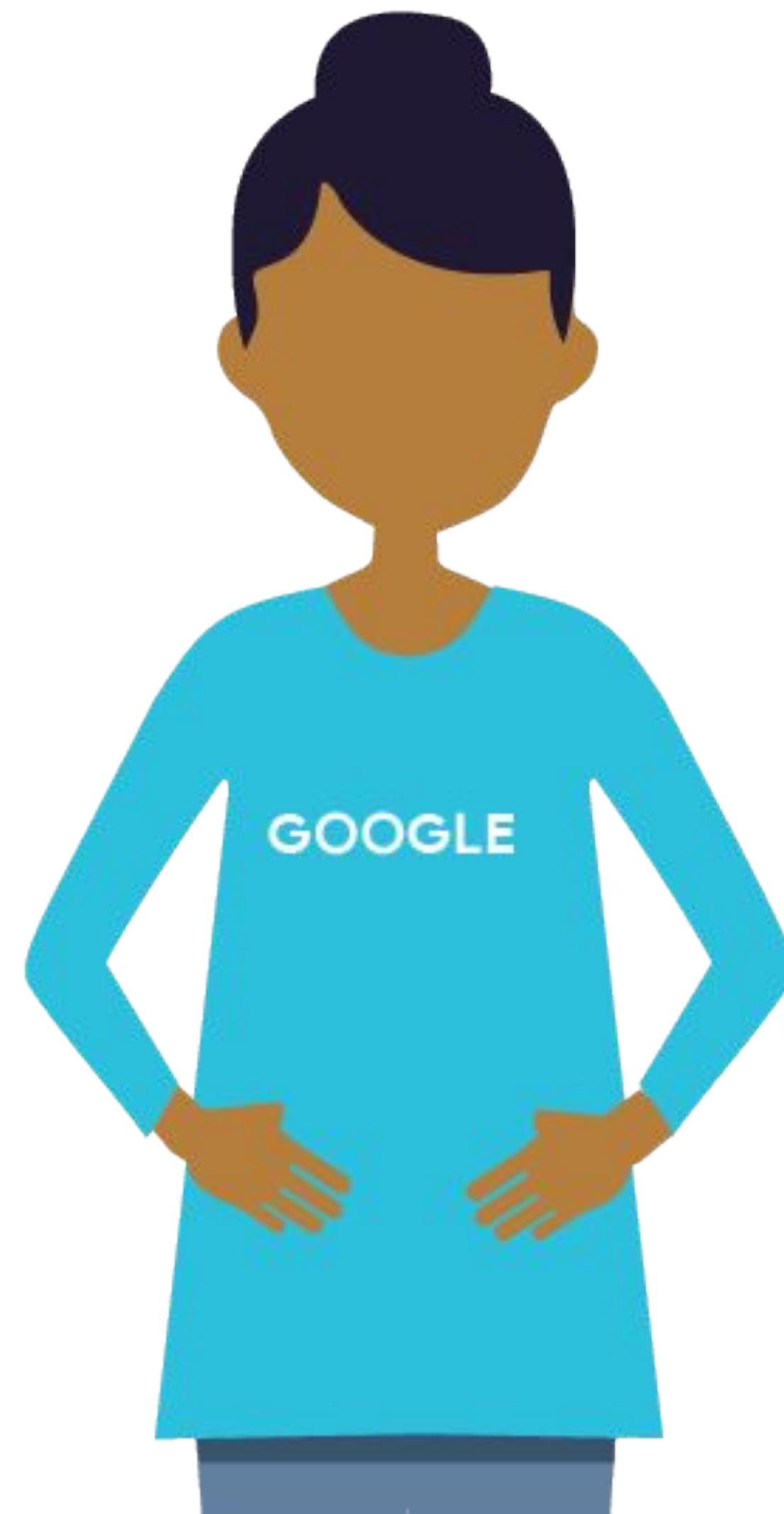
- Freeze graph
- Transform the graph
- Quantize weights and calculations



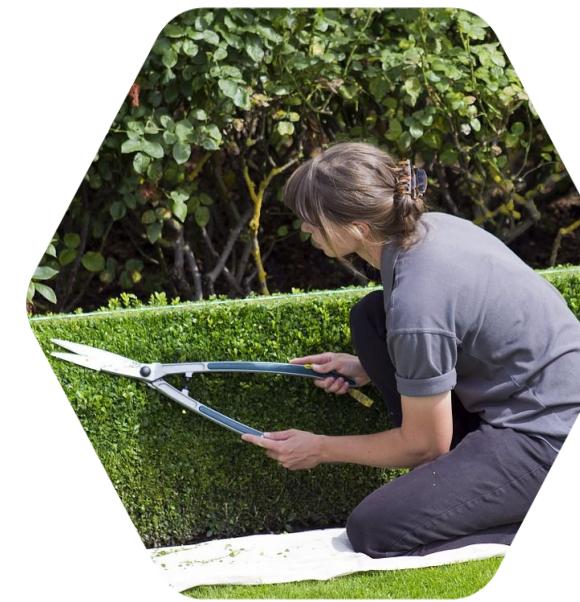
Freezing a graph can do load time optimization



Converts
variables to
constants and
removes
checkpoints

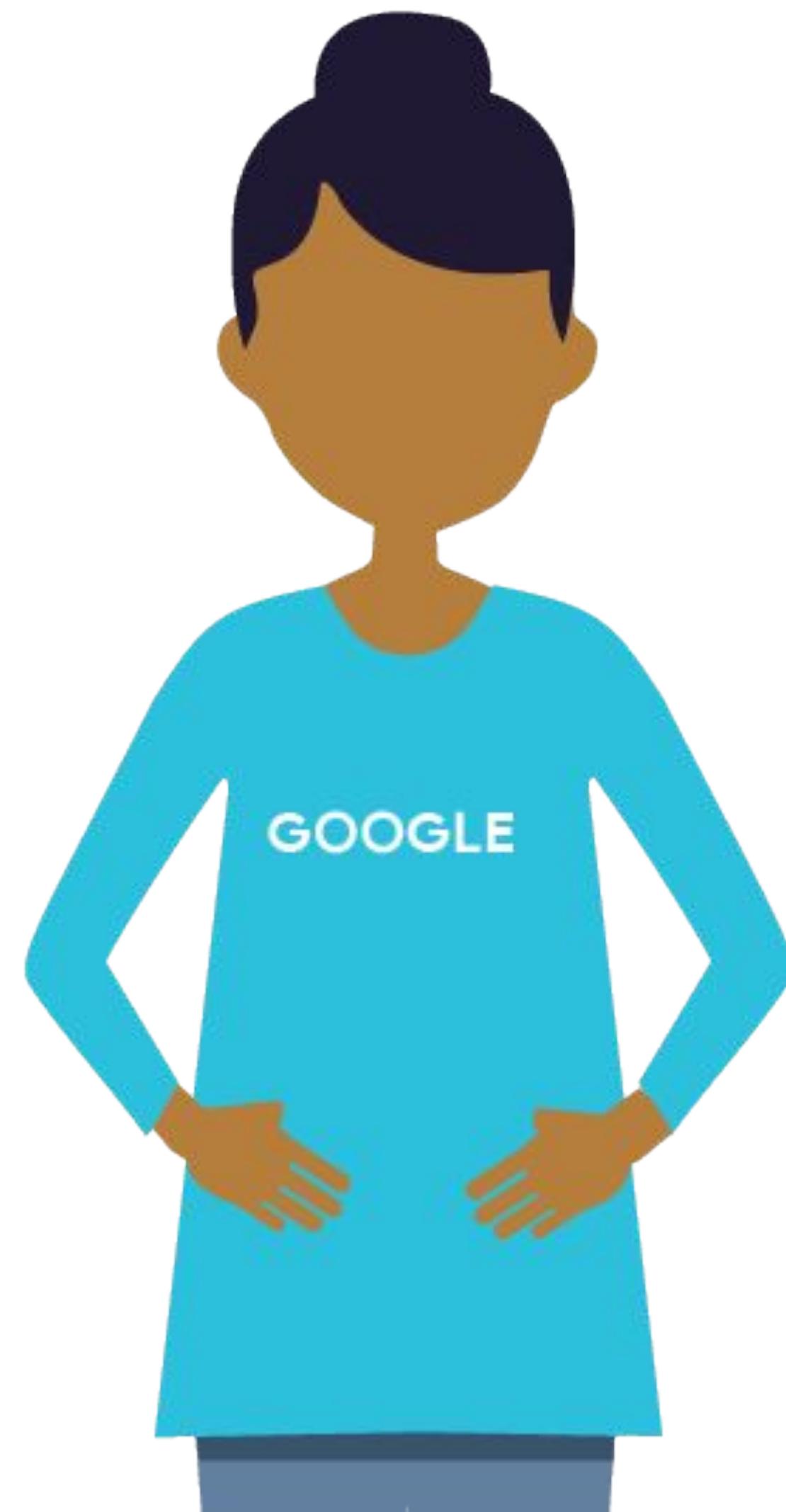


Transform your graph to
remove nodes you don't use
in prediction



`strip_unused_nodes:`

Remove training-only
operations

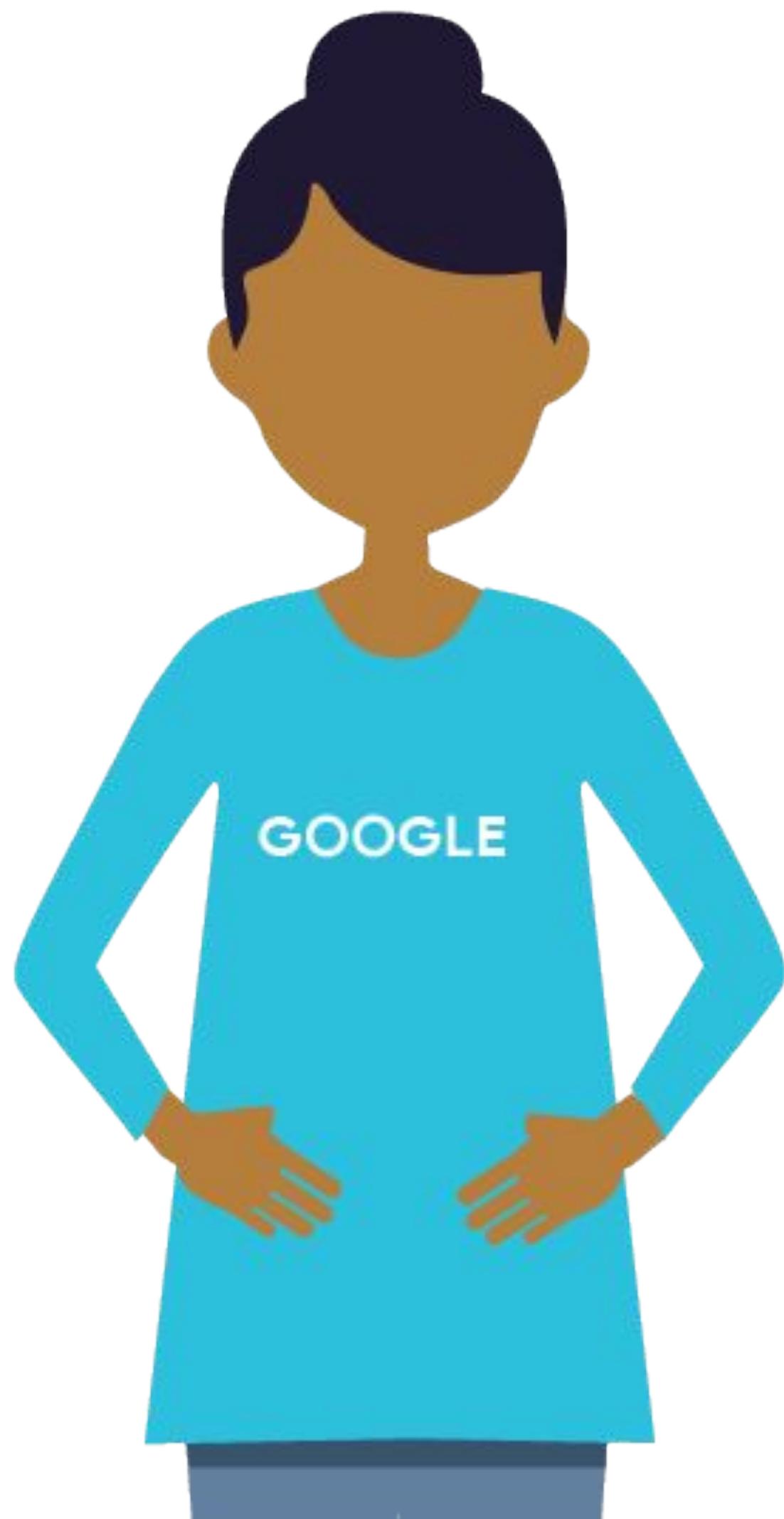


Transform your graph to
remove nodes you don't use
in prediction



remove_nodes:

Remove debug
nodes

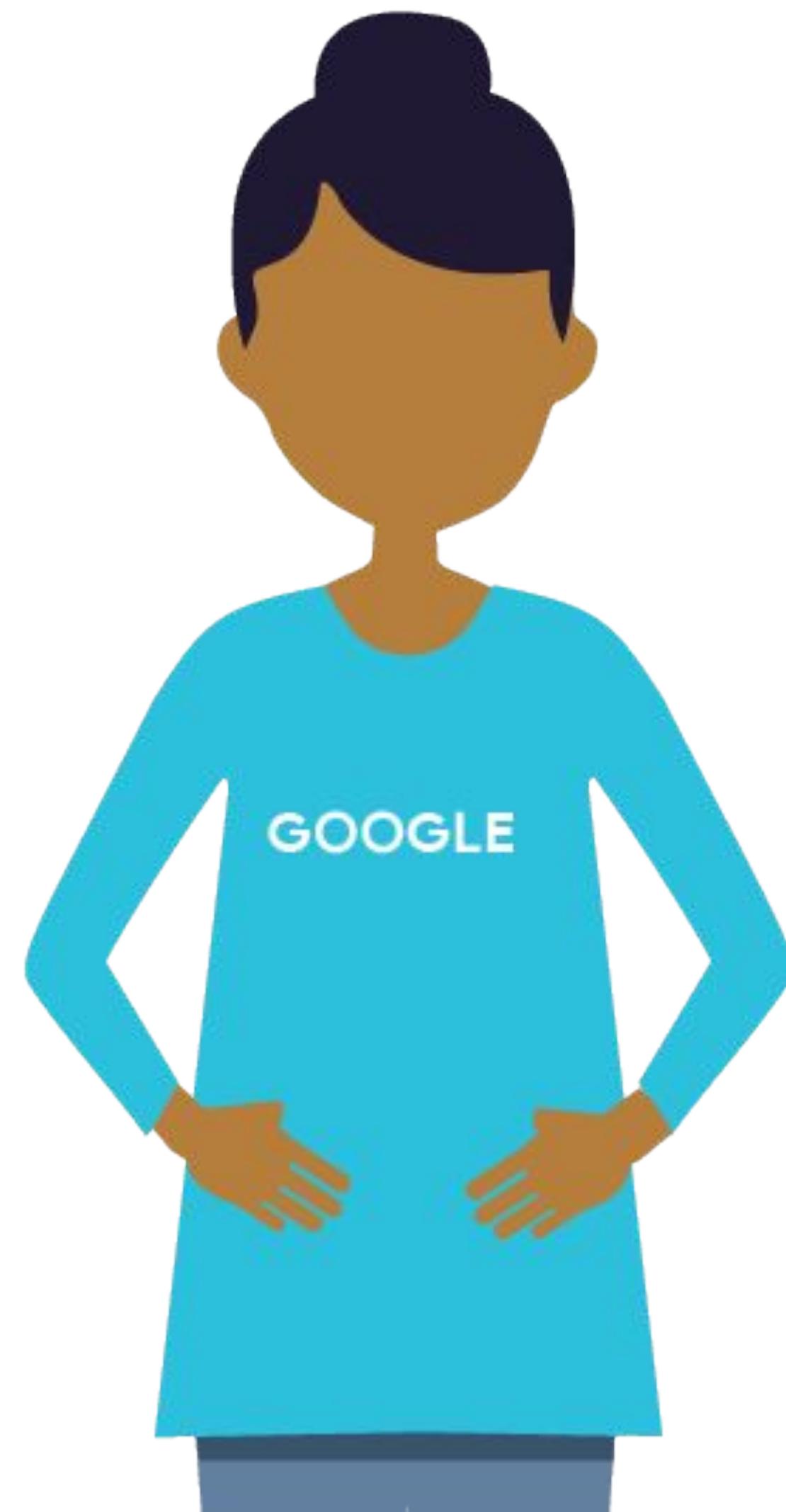


Transform your graph to
remove nodes you don't use
in prediction



fold_batch_norms:

Remove Muls for
batch norm



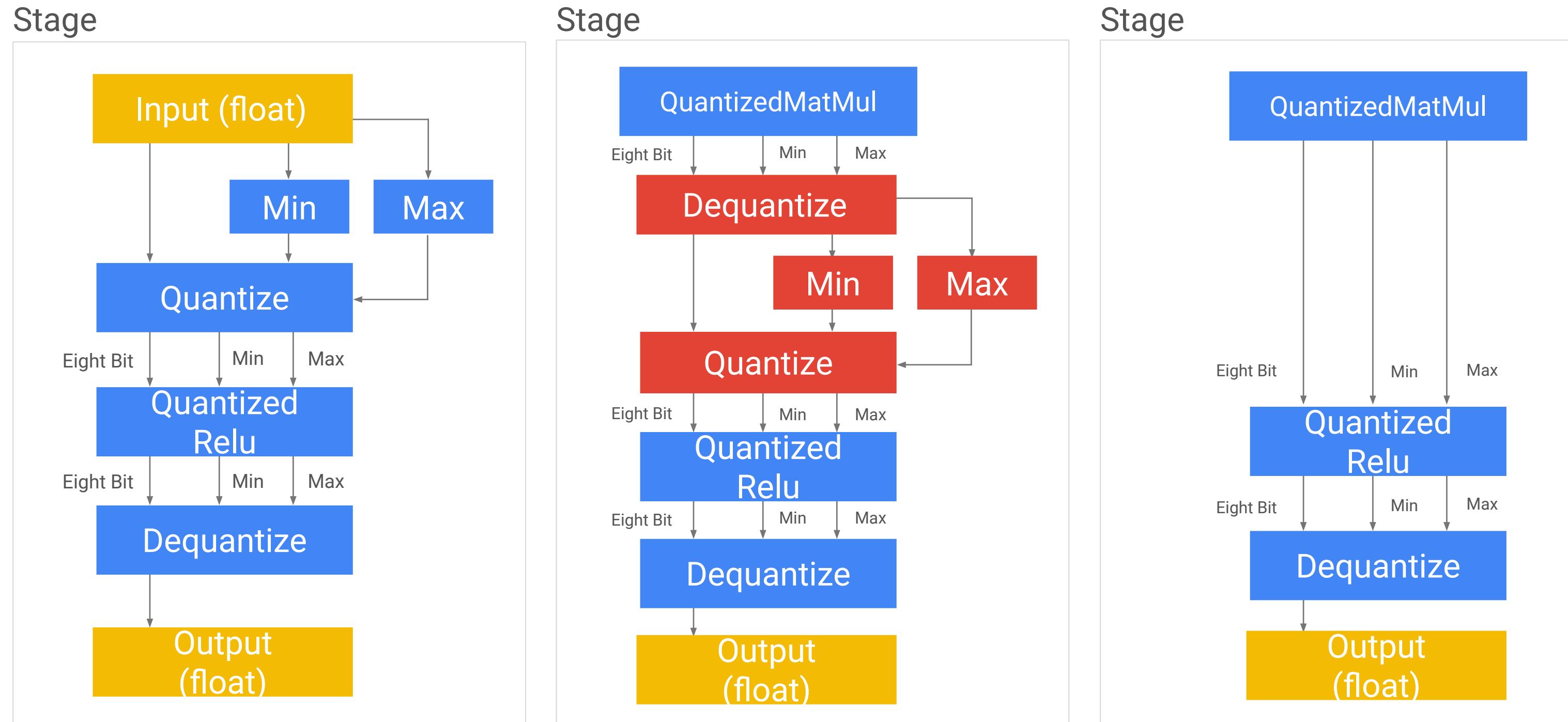
Transform your graph to
remove nodes you don't use
in prediction



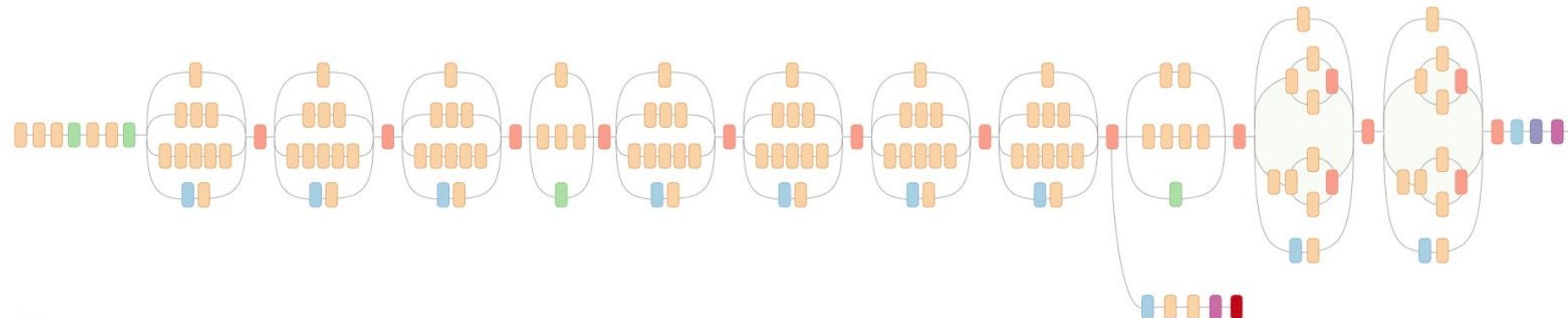
quantize_weights
quantize_nodes

Add quantization

Quantizing weights and calculations boosts performance

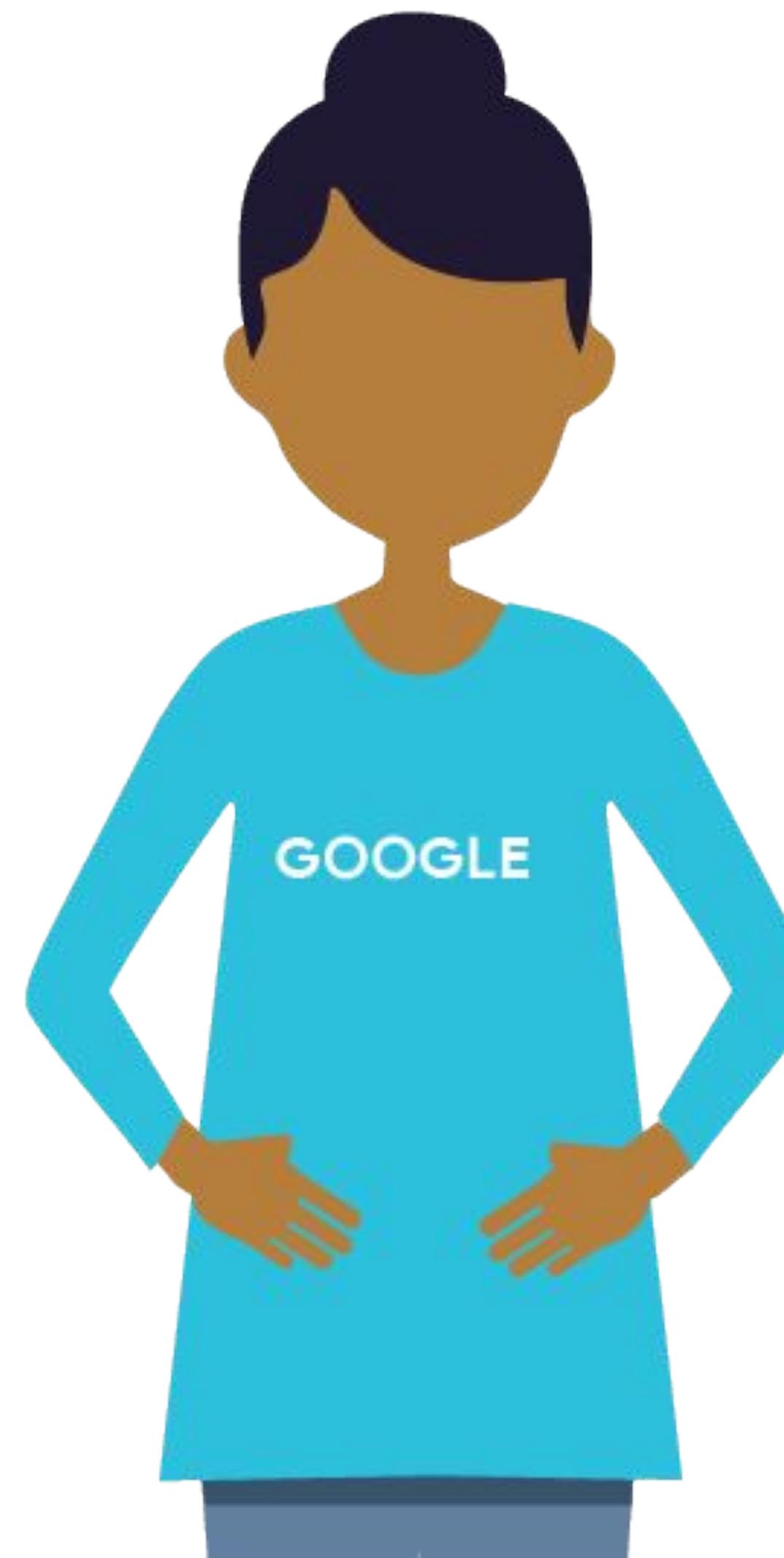


After these optimizations, the neural network is 75% smaller

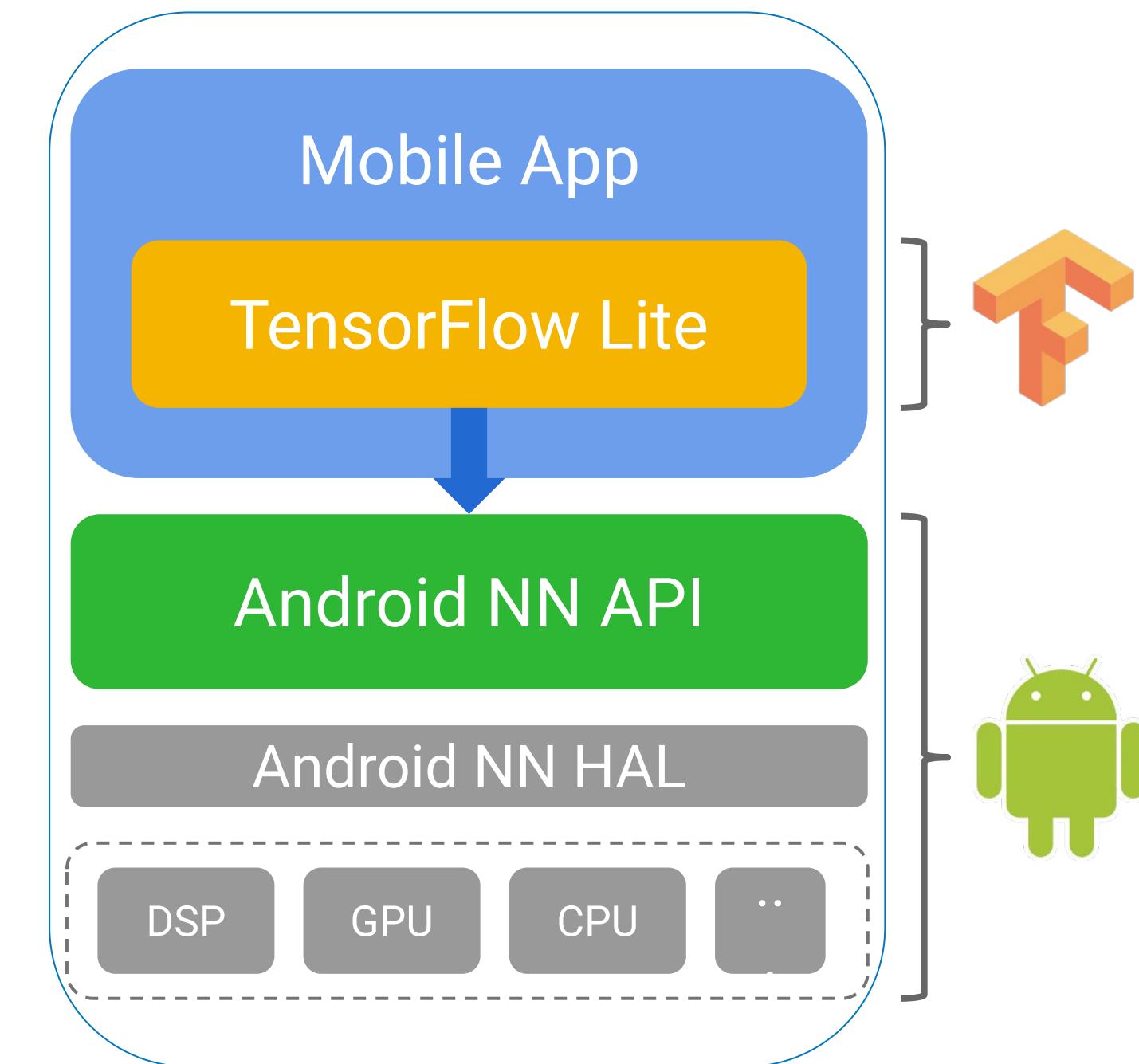


- Convolution
- AvgPool
- MaxPool
- Concat
- Dropout
- Fully connected
- Softmax

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

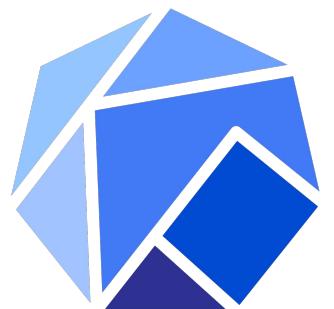
Presenter: Val

Video Name: T-PSML-O_5_l8_summary

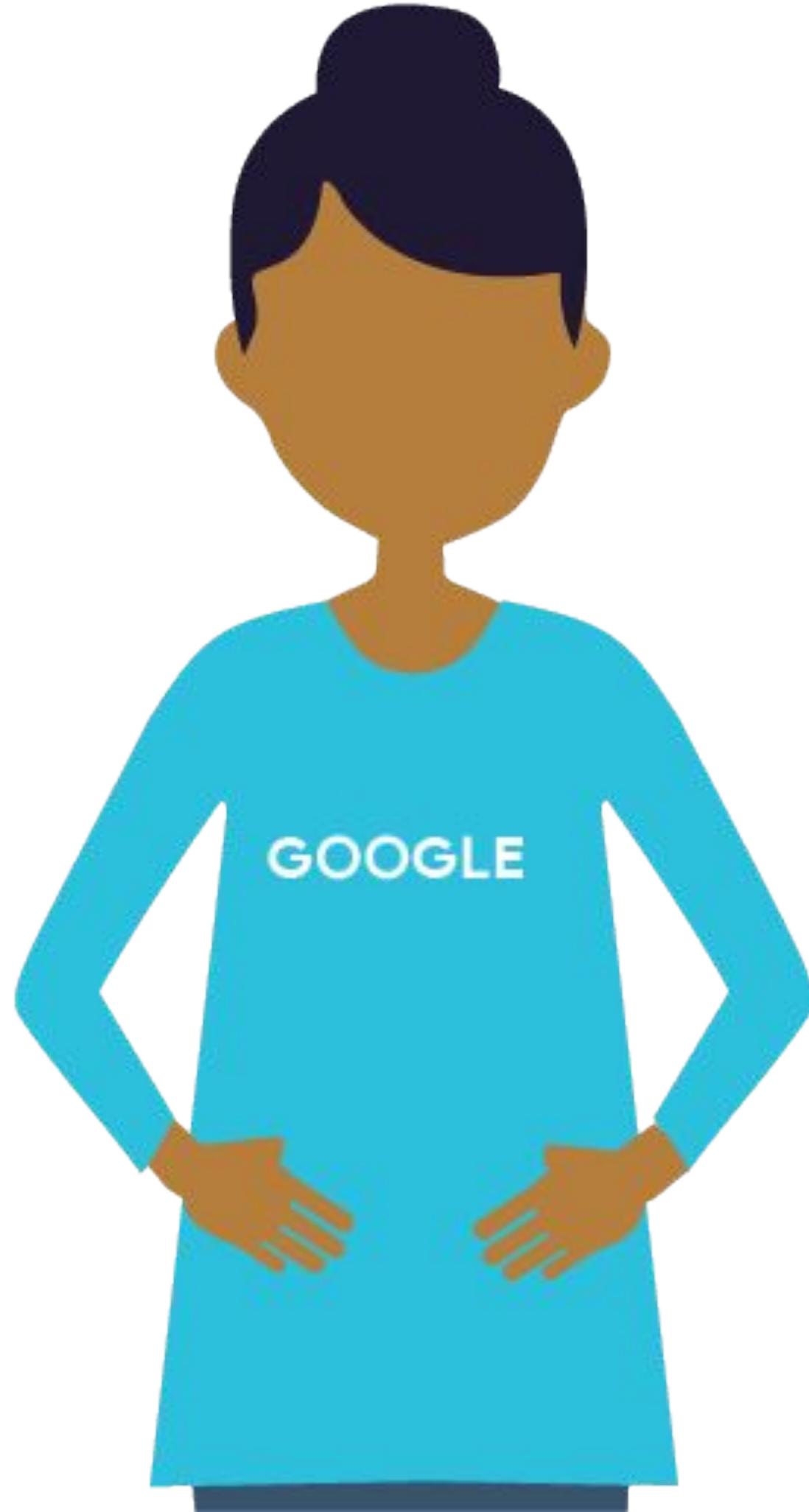
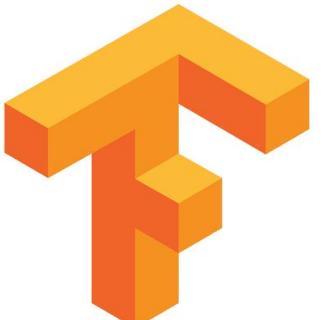
Summary

Build hybrid cloud machine learning models

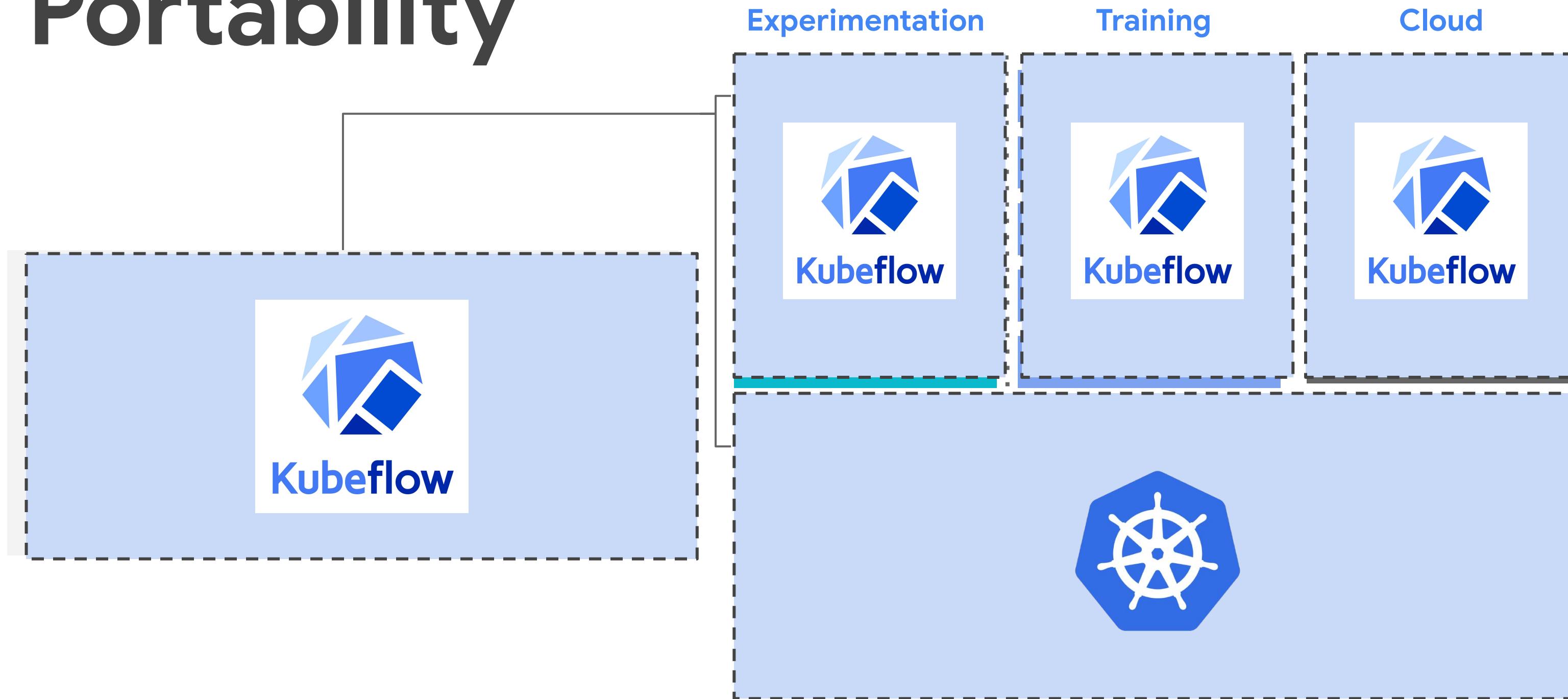
Optimize TensorFlow graphs for mobile



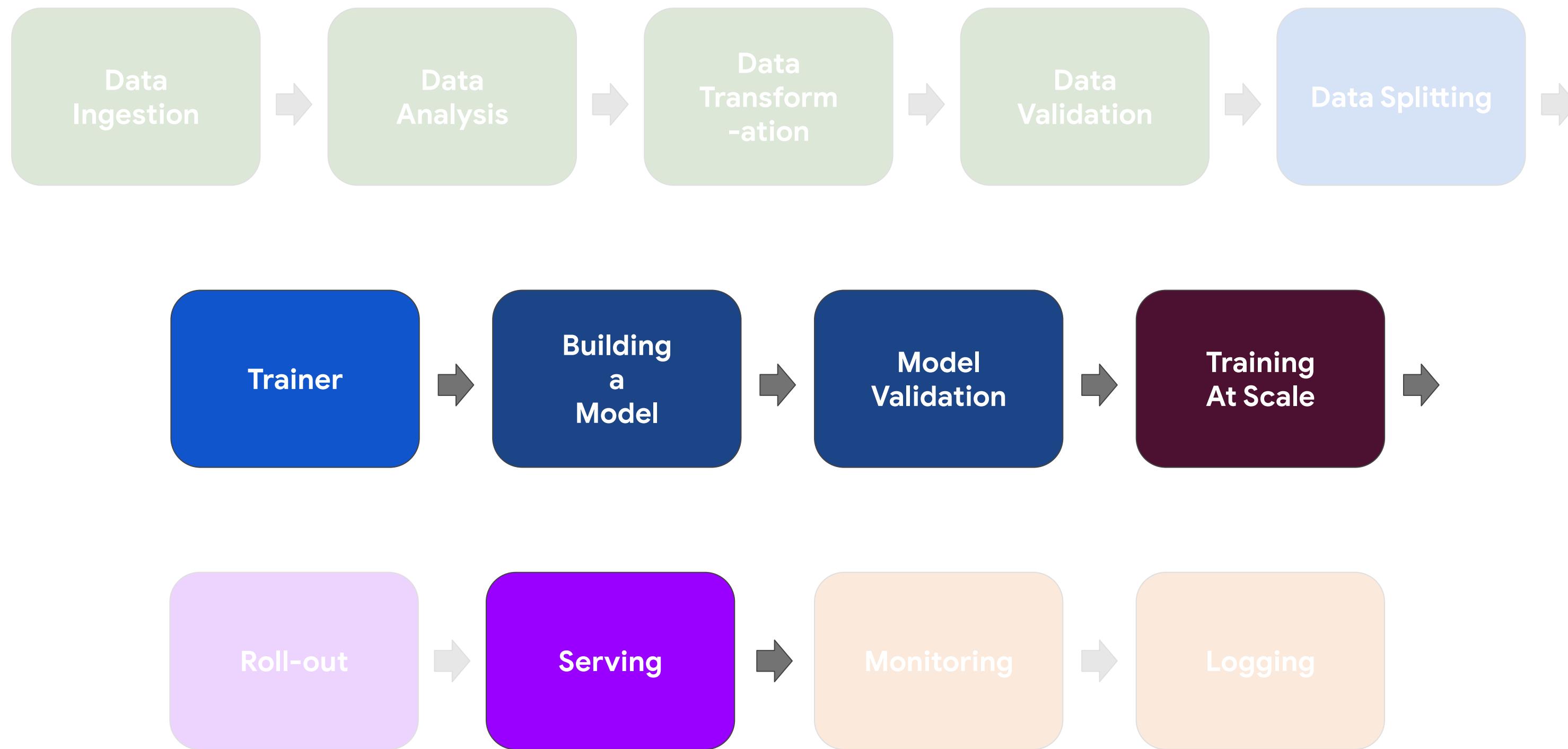
Kubeflow

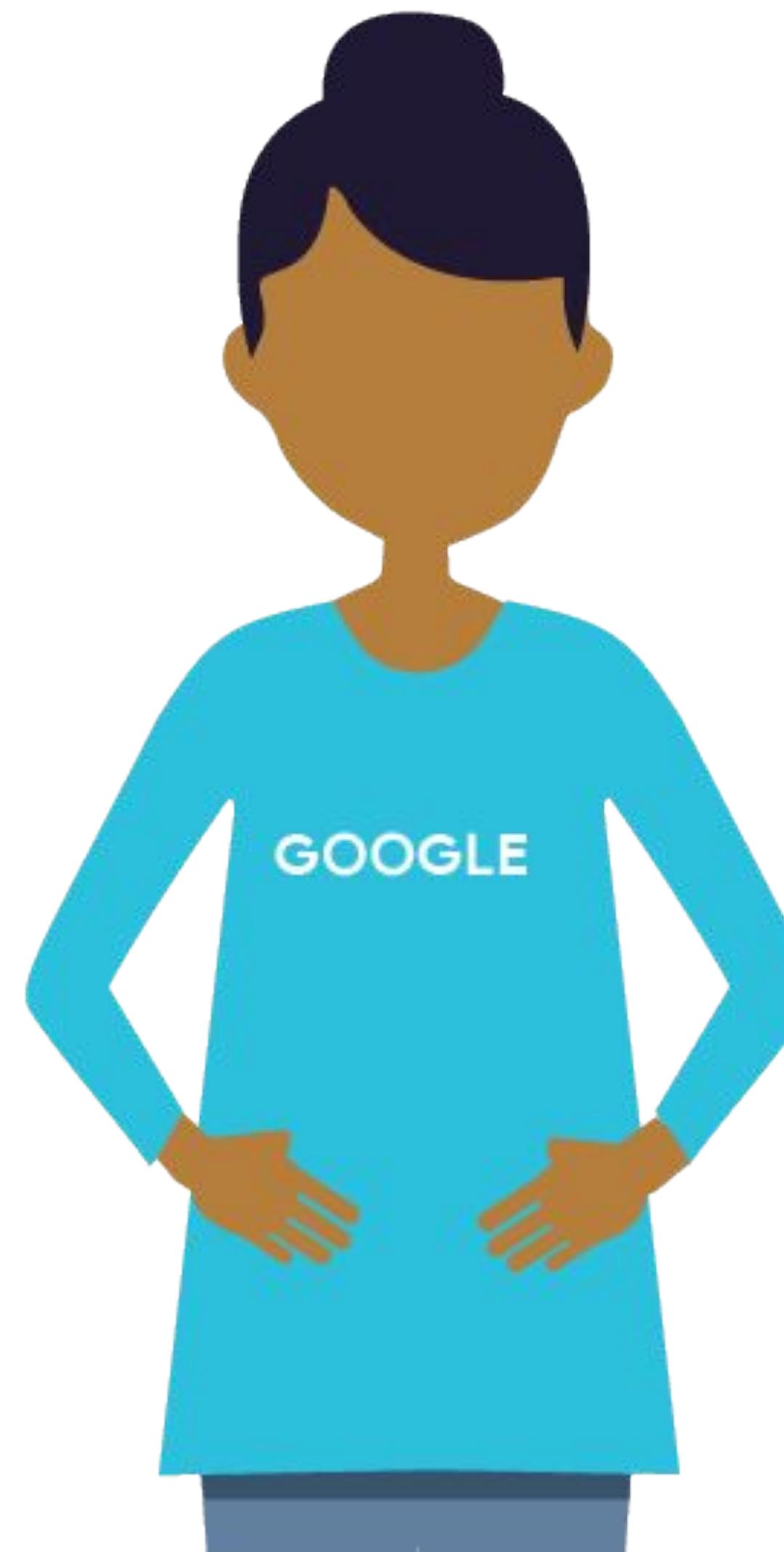


Portability



What's in the box?

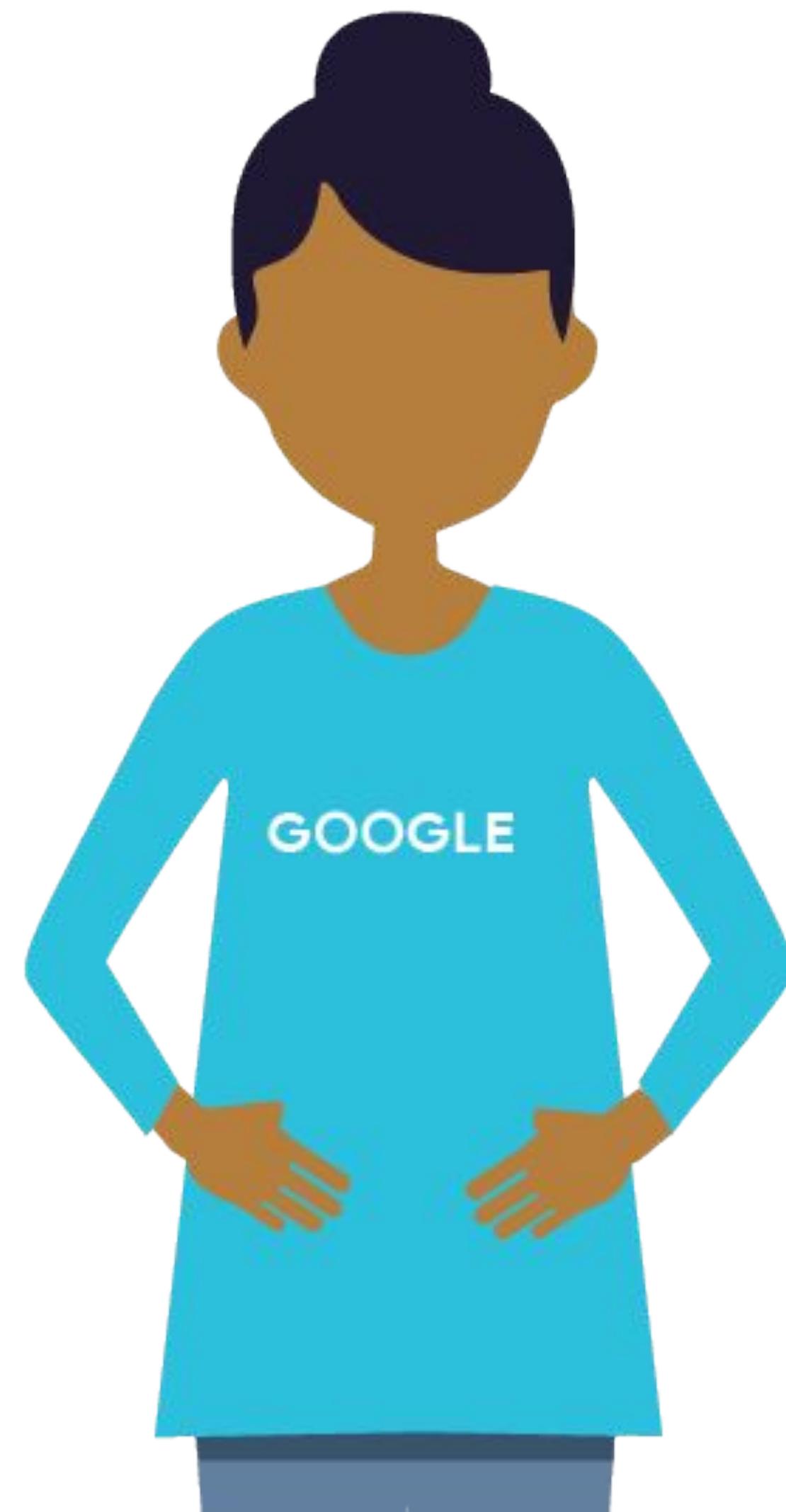




Freezing a graph can do load
time optimization



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