Course 2: Production ML Systems

Module 3: Designing Adaptable ML Systems

Lesson Title: **Introduction**

Presenter: Max Lotstein

Format: Talking Head

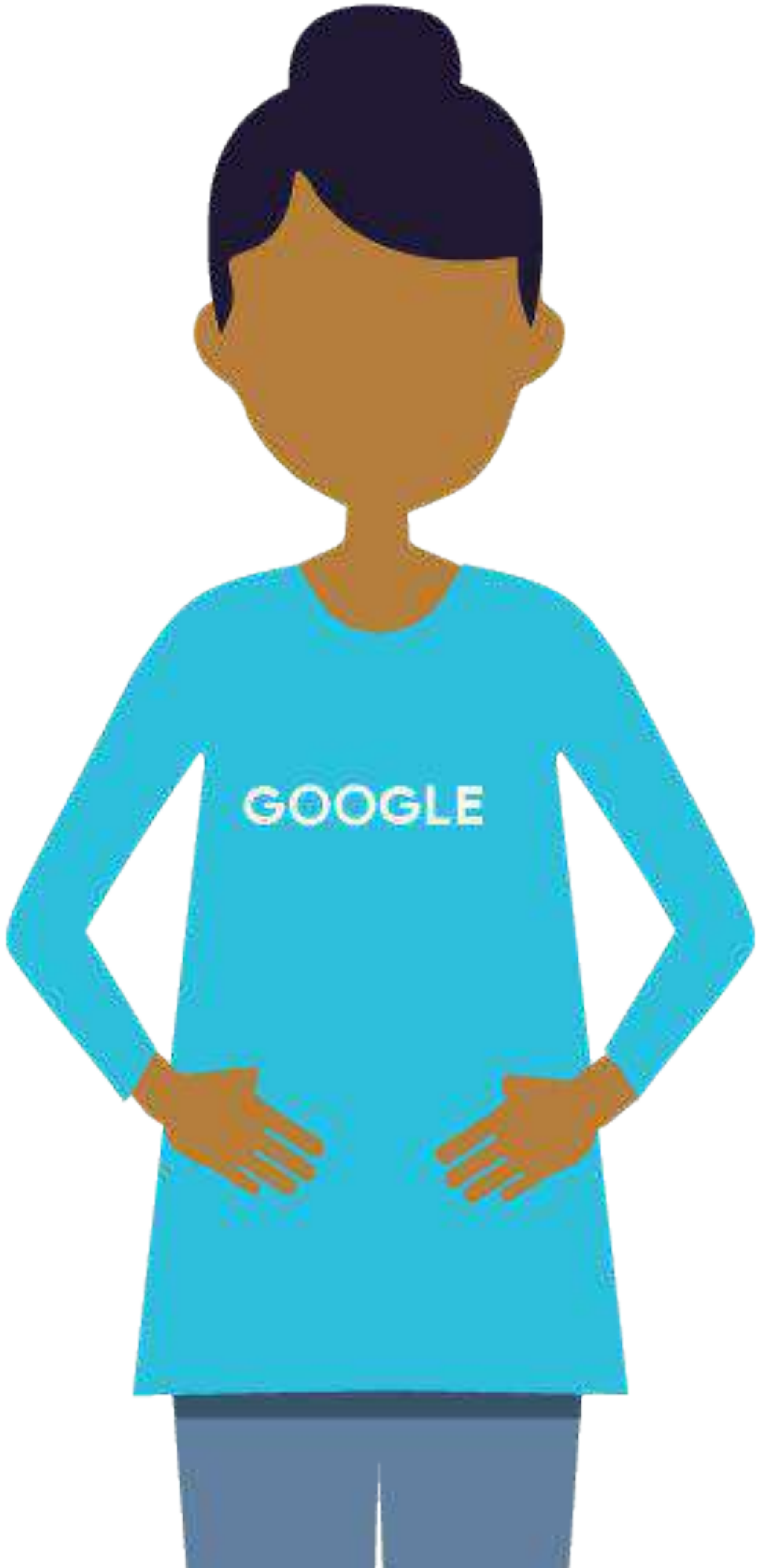
Video Name: T-PSML-O_3_l1_introduction

Google Cloud

Designing Adaptable ML Systems

Advanced Machine Learning on GCP

Max Lotstein



Learn how to...

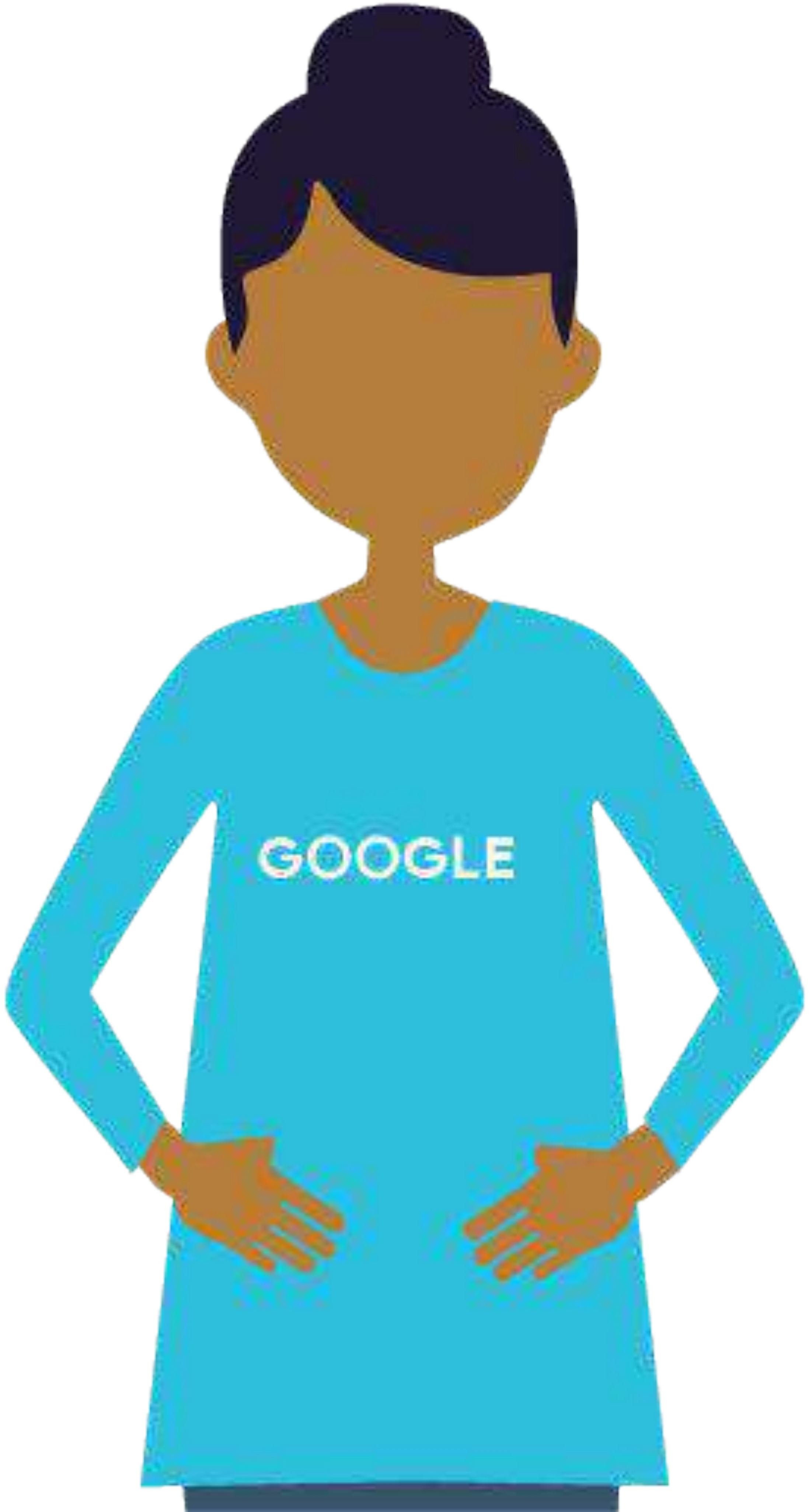
Recognize various data dependencies

Make cost-conscious engineering decisions

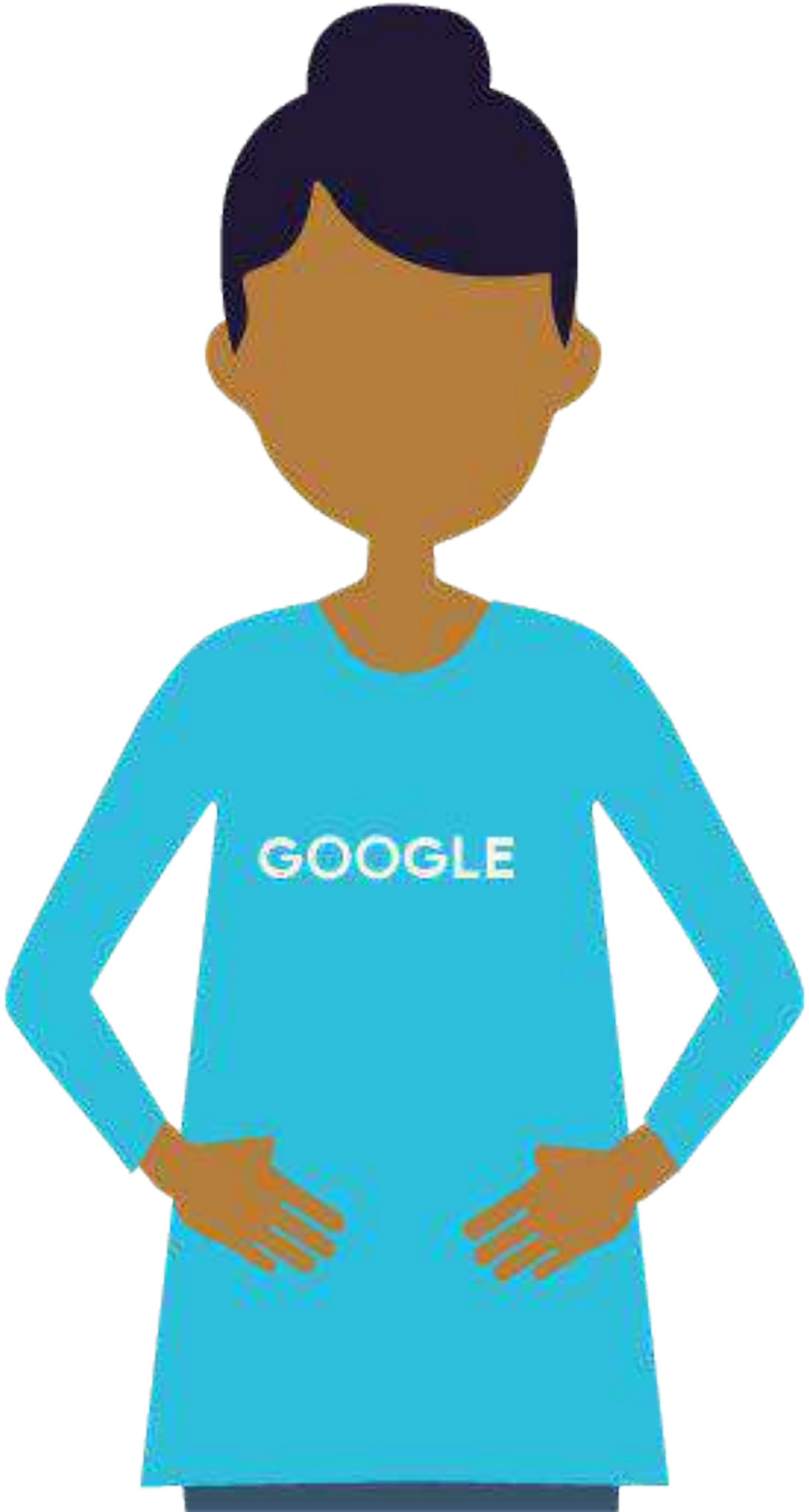
Mitigate model pollution

Implement a pipeline that is immune to one type of dependency

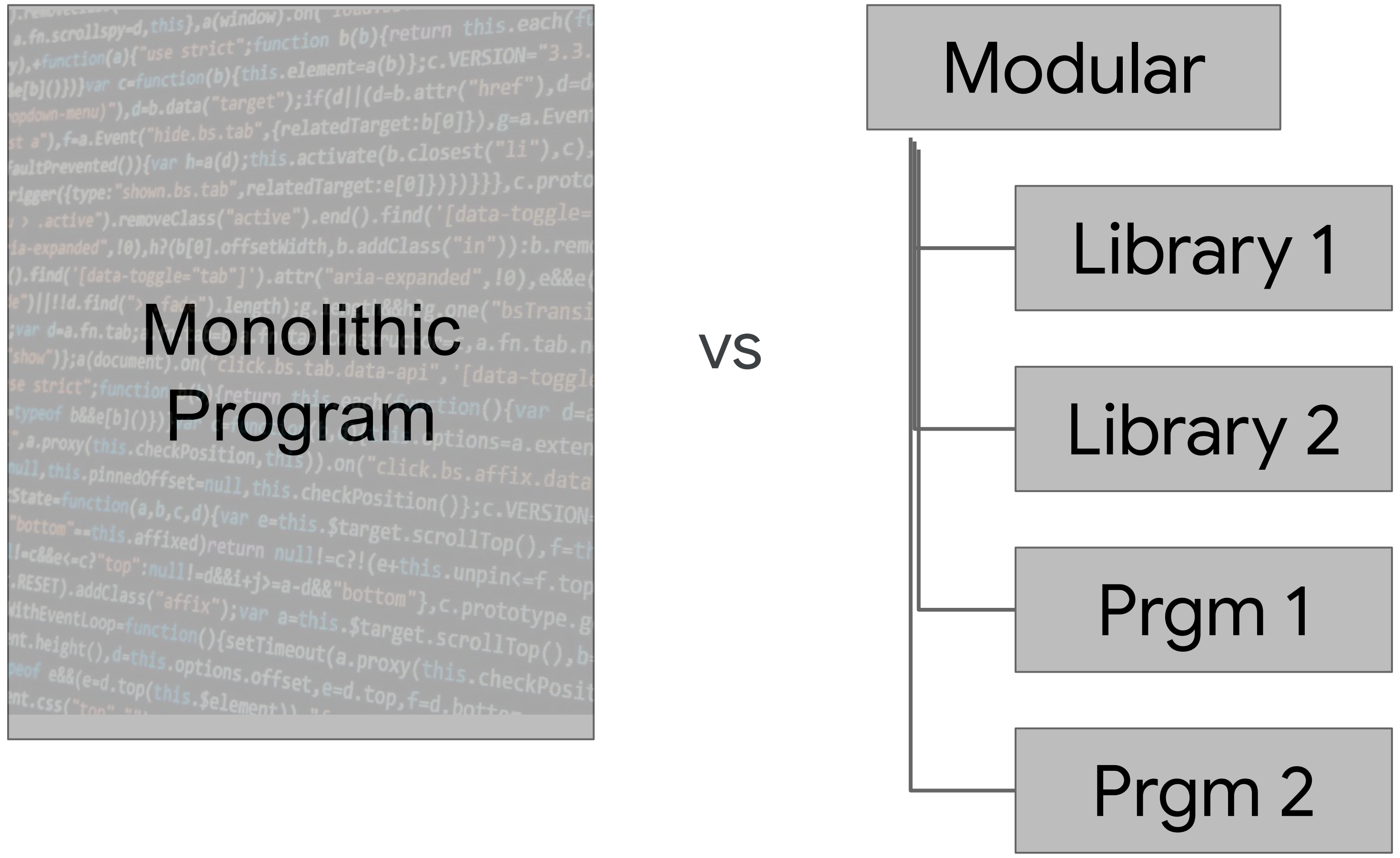
Debug the causes of observed model behavior



Few Programs are Islands



Few Programs are Islands



Monolithic Program

vs

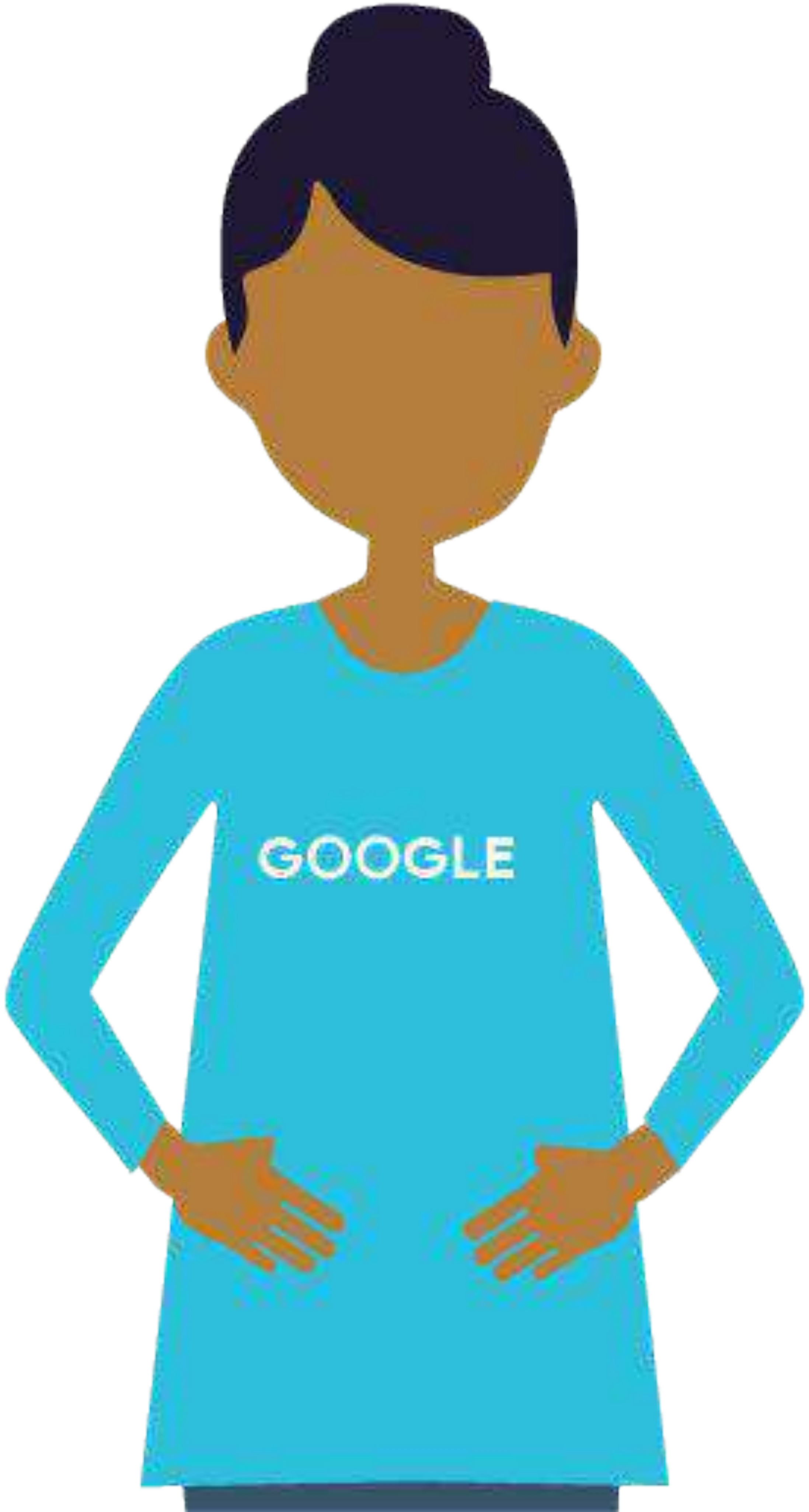
Modular

Library 1

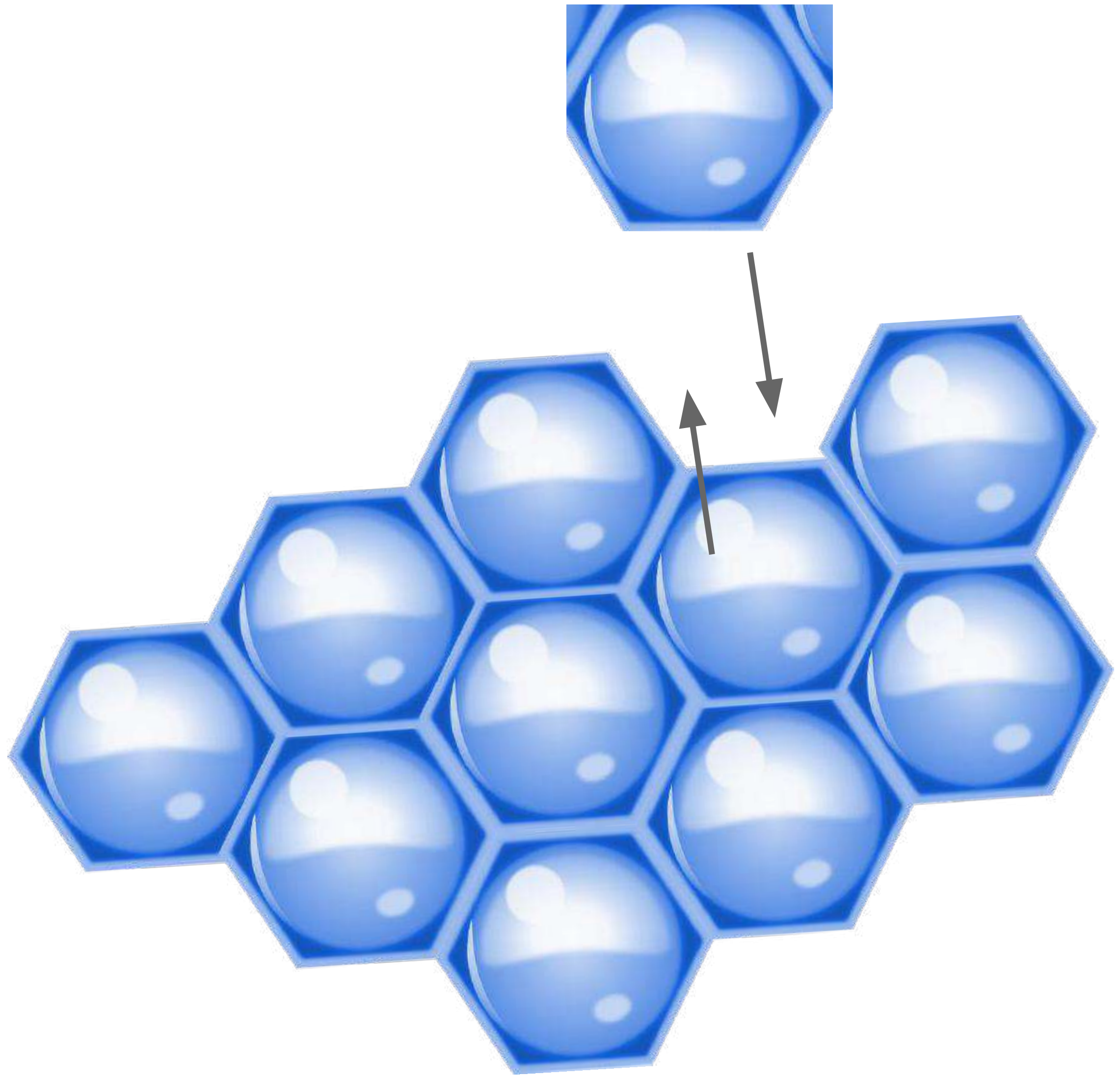
Library 2

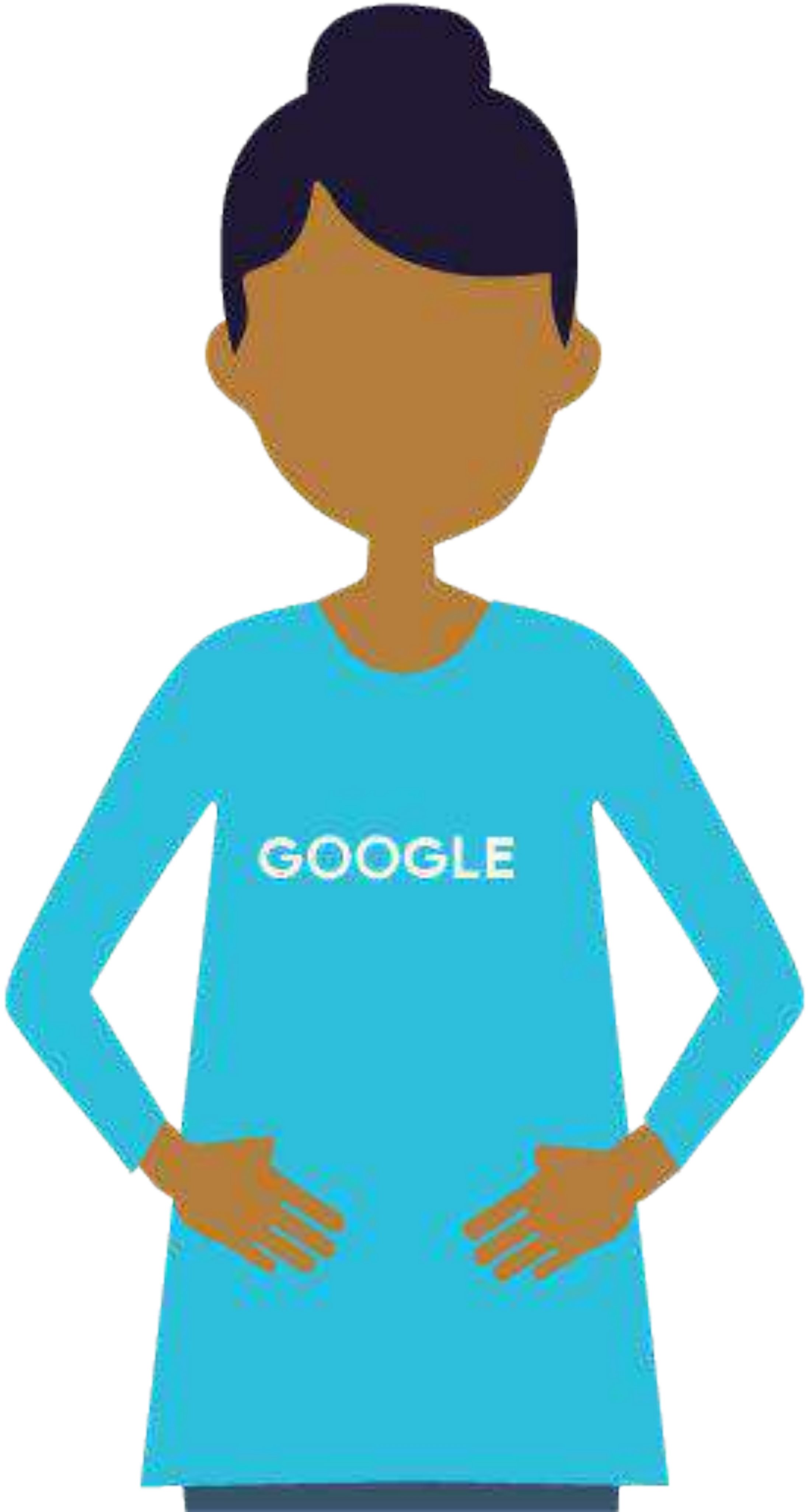
Prgm 1

Prgm 2

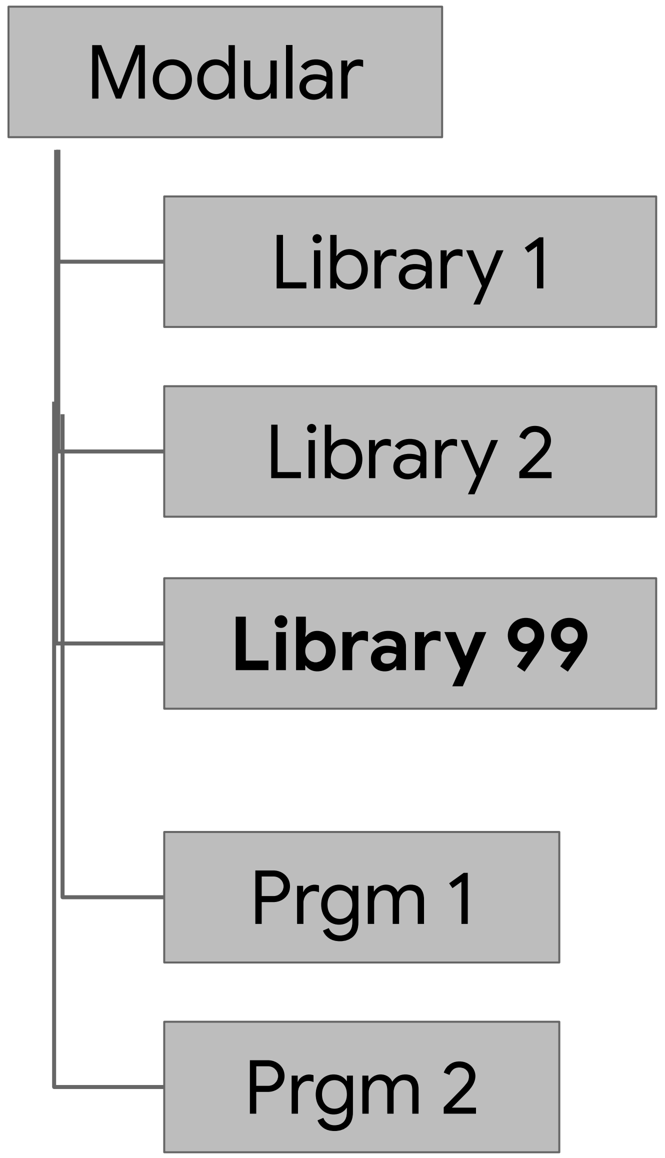


Modular Is More Maintainable



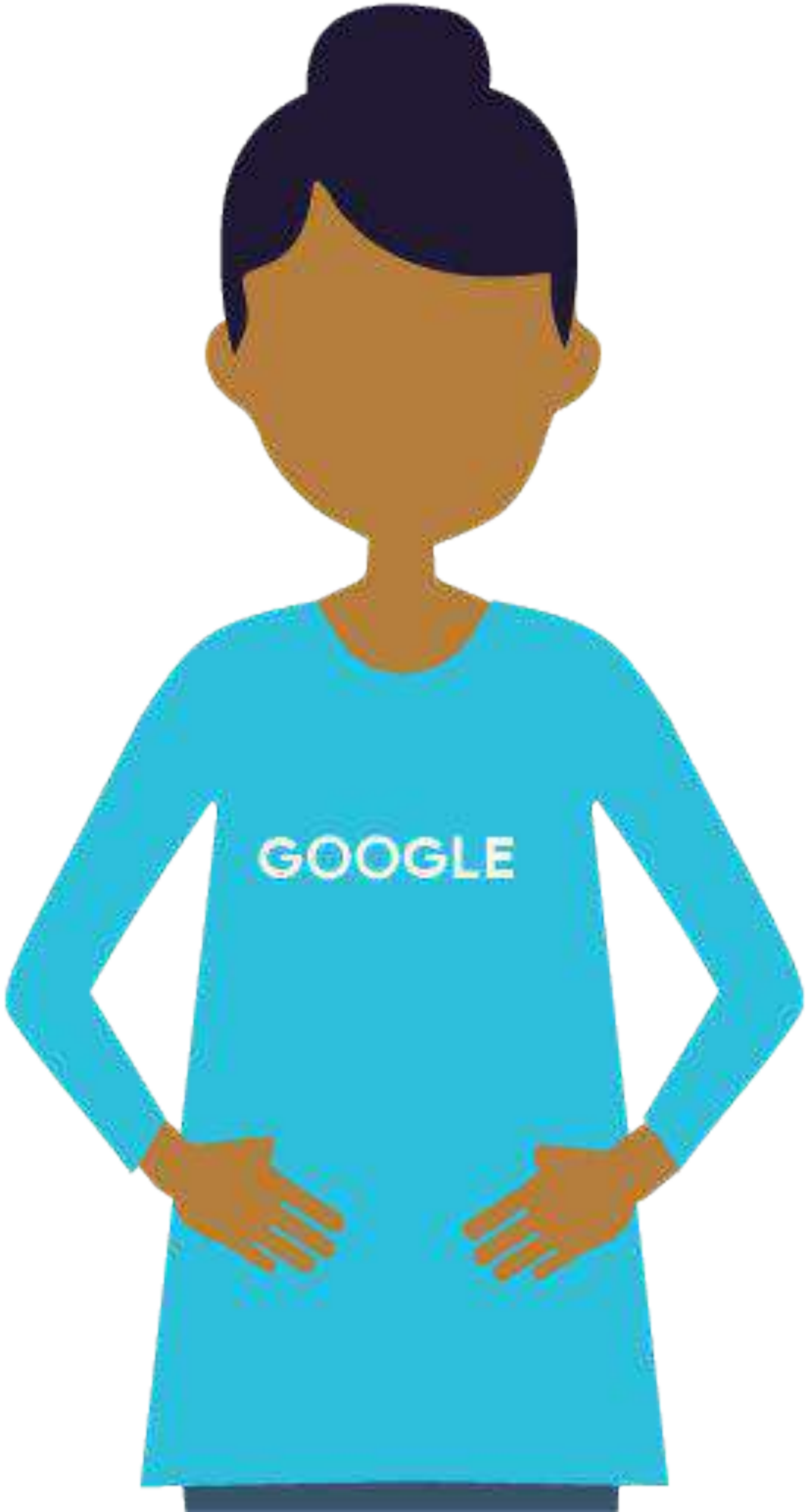


Dependency Management Is Manageable

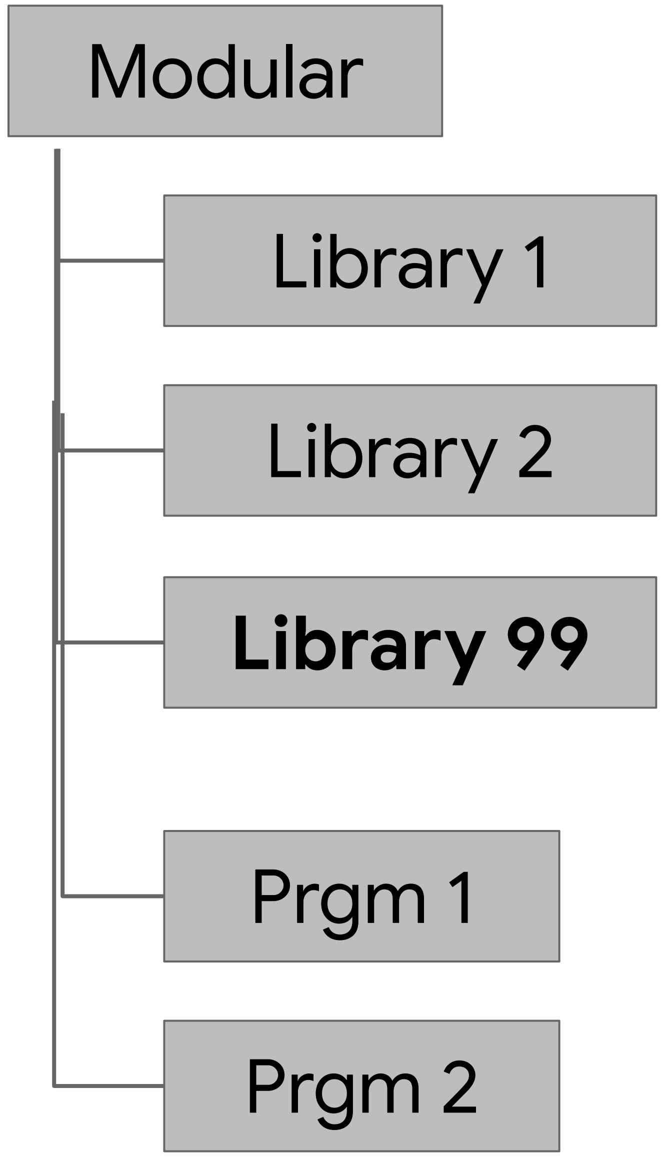


Modular

- Library 1
- Library 2
- **Library 99**
- Prgm 1
- Prgm 2



Dependency Management Is Manageable



Modular

Library 1

Library 2

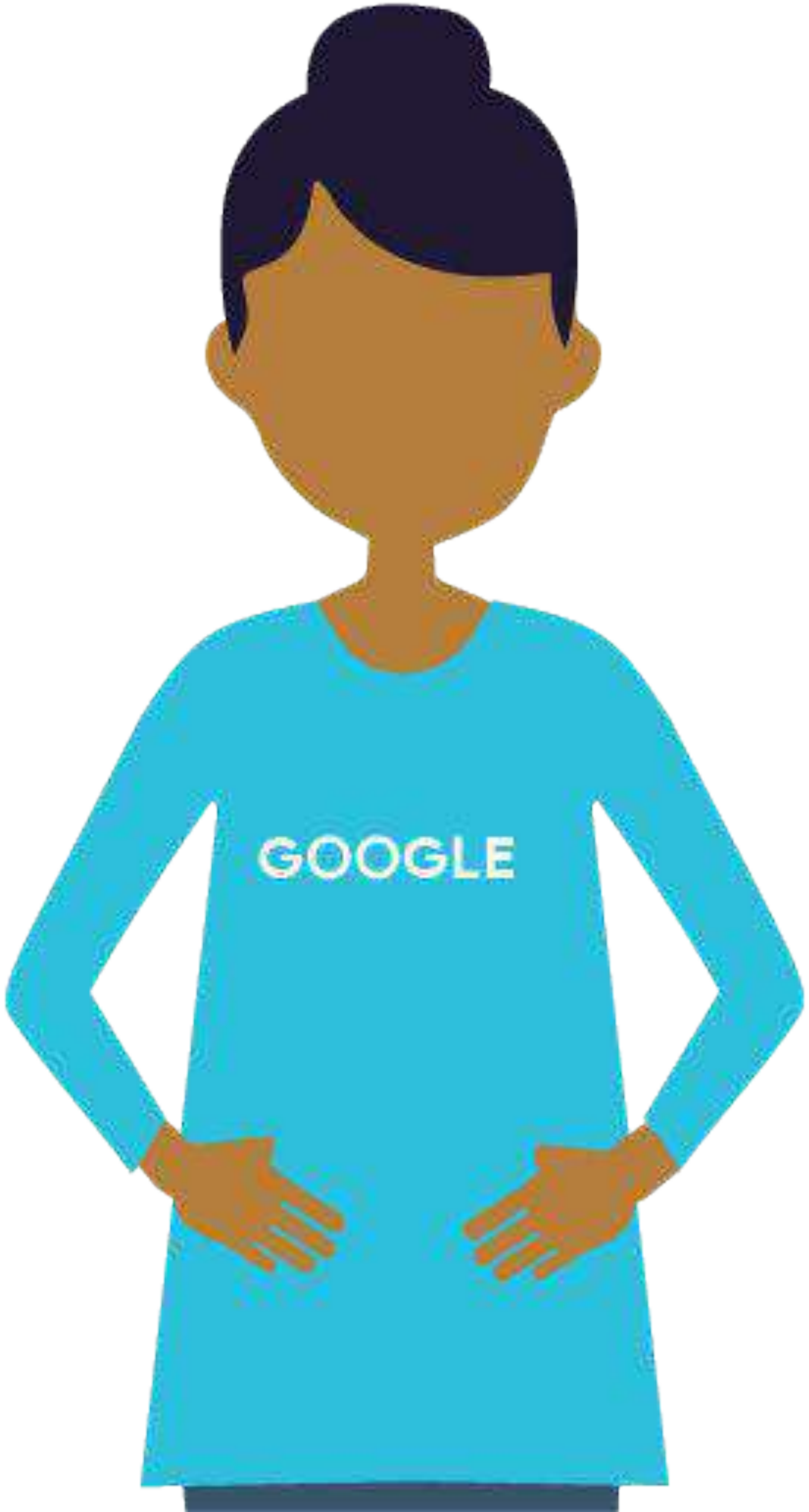
Library 99

Prgm 1

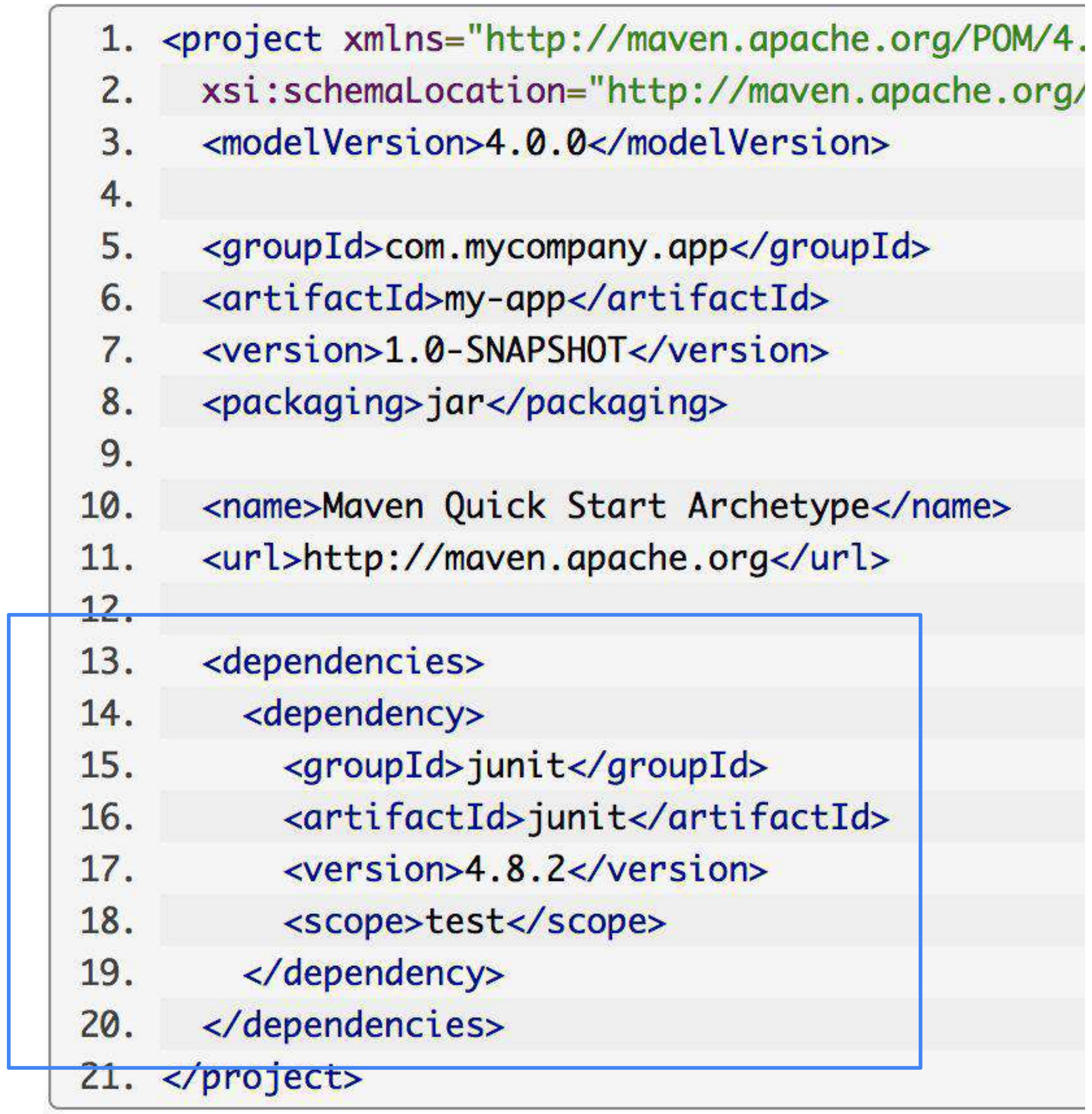
Prgm 2

Maven™

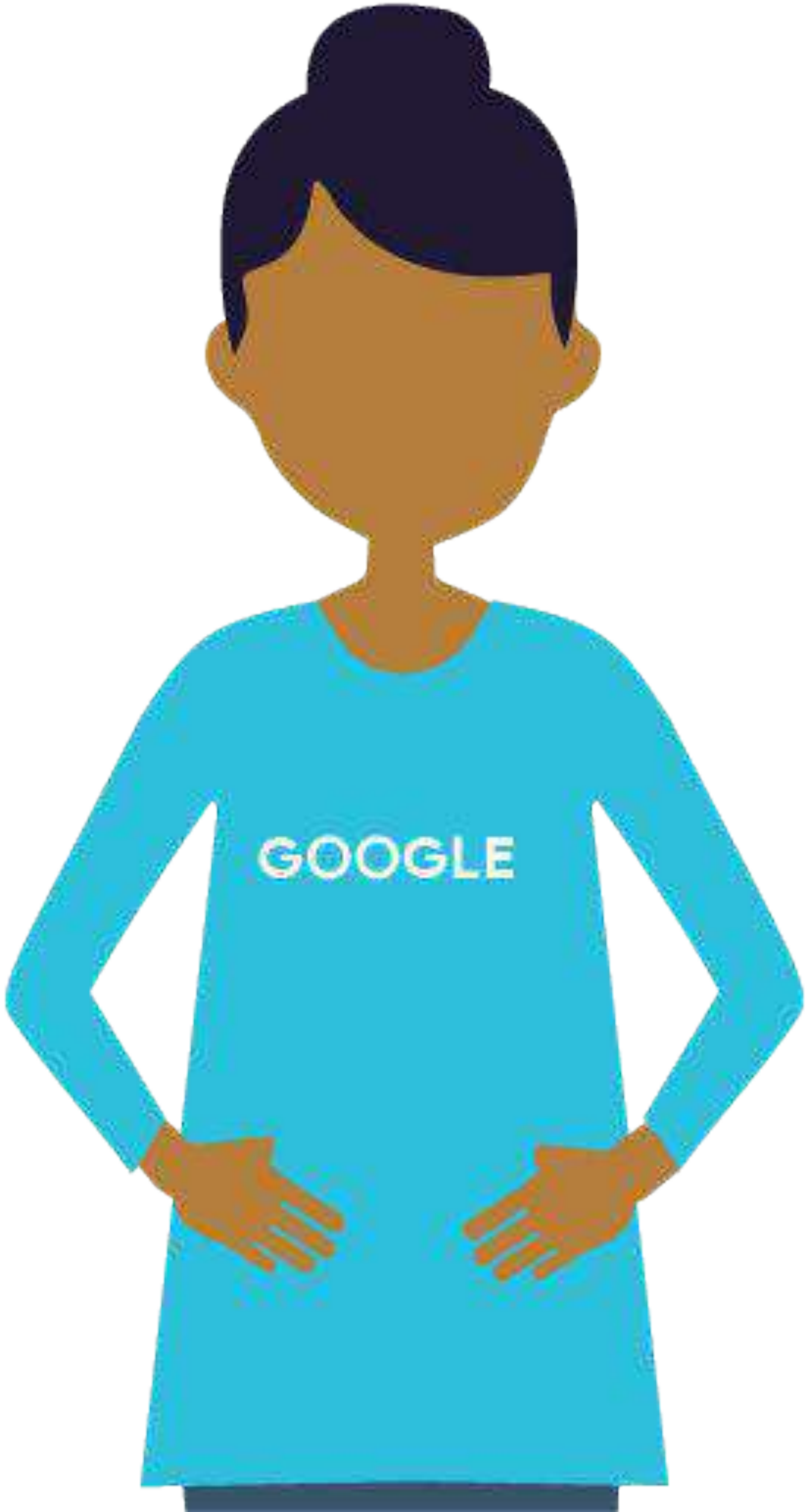
PIP



Explicit Dependencies Make Life Easier



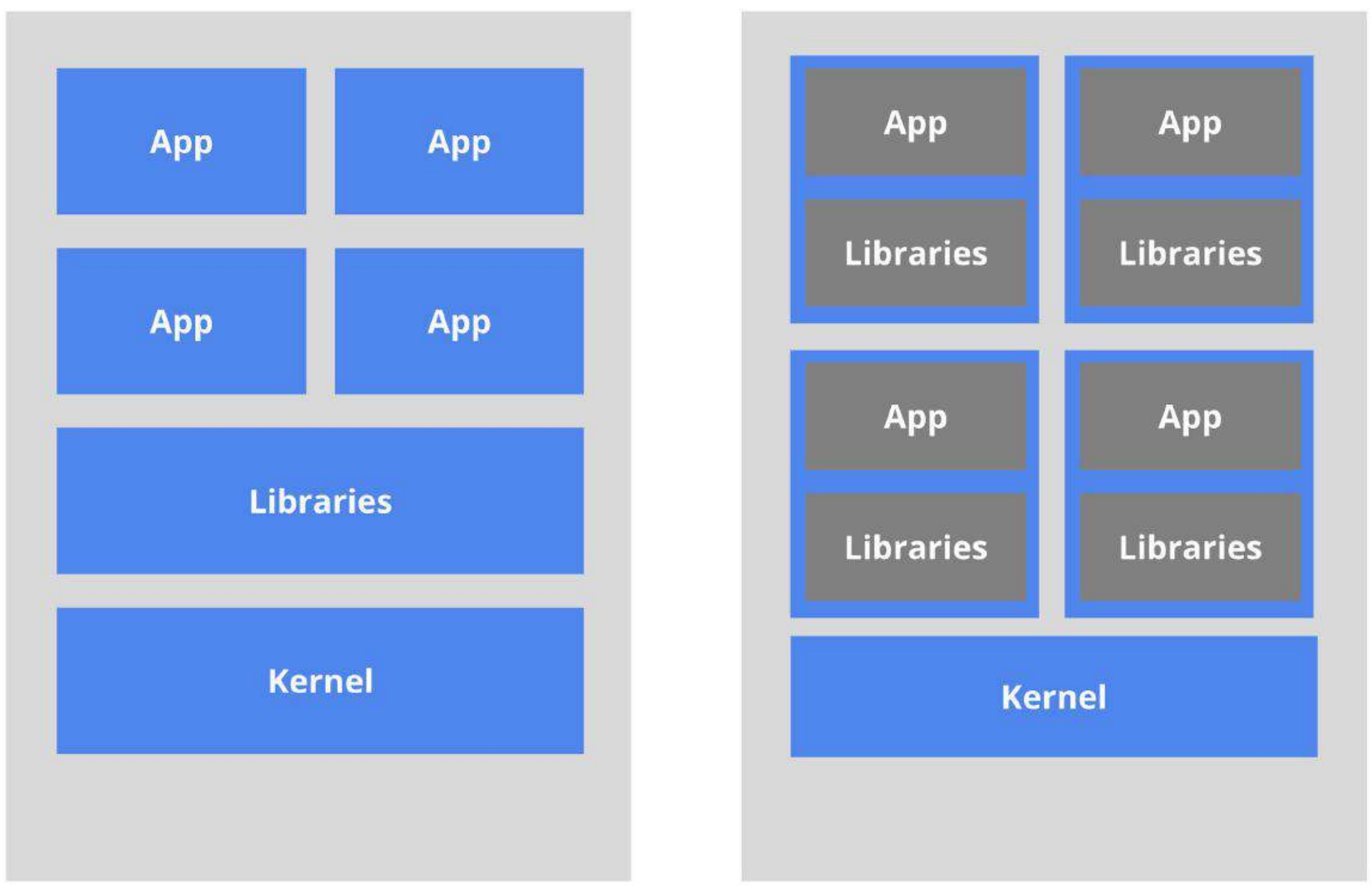
```
<project xmlns="http://maven.apache.org/POM/4.
  xsi:schemaLocation="http://maven.apache.org/
  <modelVersion>4.0.0</modelVersion>

  <groupId>com.mycompany.app</groupId>
  <artifactId>my-app</artifactId>
  <version>1.0-SNAPSHOT</version>
  <packaging>jar</packaging>

  <name>Maven Quick Start Archetype</name>
  <url>http://maven.apache.org</url>

  <dependencies>
    <dependency>
      <groupId>junit</groupId>
      <artifactId>junit</artifactId>
      <version>4.8.2</version>
      <scope>test</scope>
    </dependency>
  </dependencies>
</project>
```



Containers eliminate infrastructure dependencies

The old way: Applications on host

The new way: Deploy containers



App

App

App

App

Libraries

Kernel

App

Libraries

App

Libraries

App

Libraries

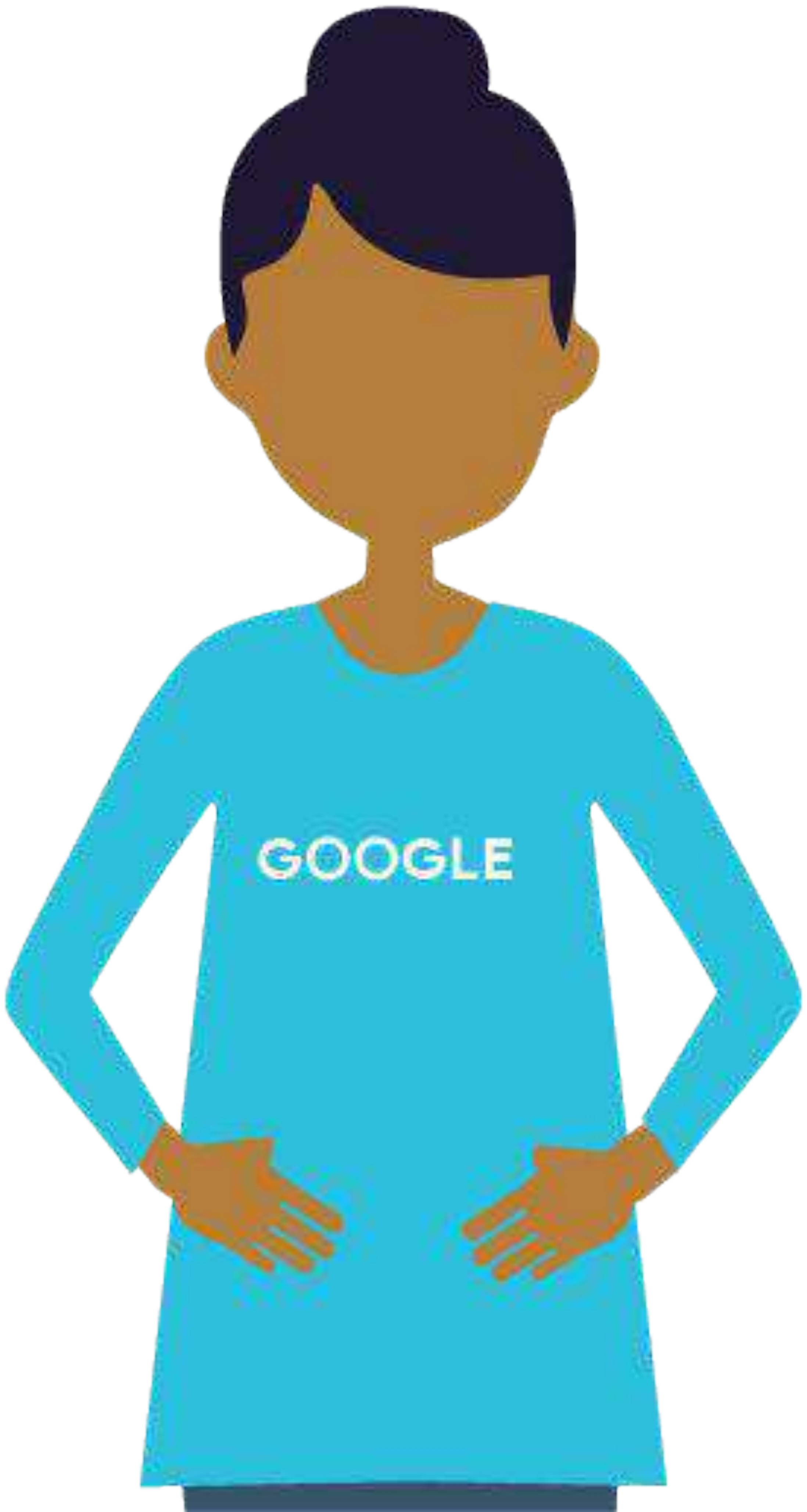
App

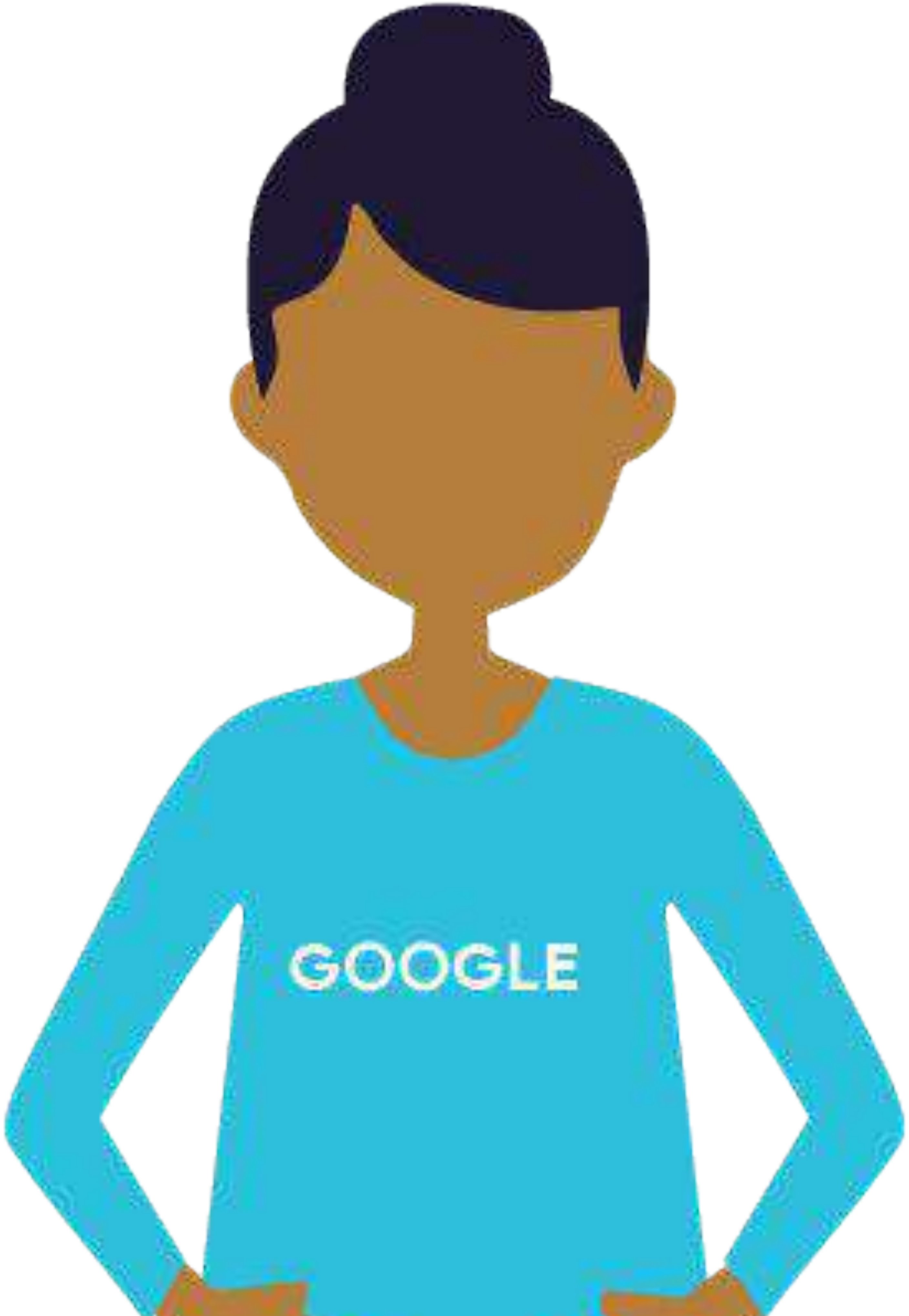
Libraries

Kernel

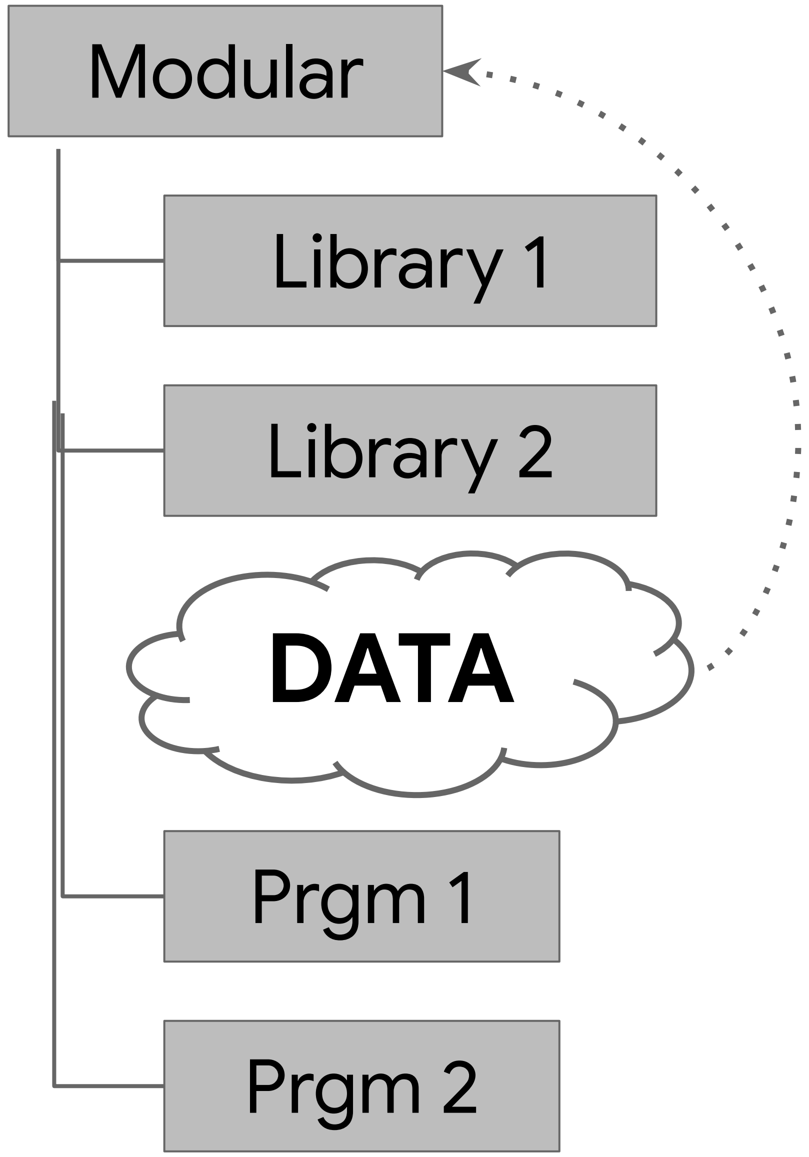
Heavyweight, non-portable
Relies on OS package manager

Small and fast, portable
Uses OS-level virtualization




GOOGLE

Data: the Dependency Outside the Codebase



Modular

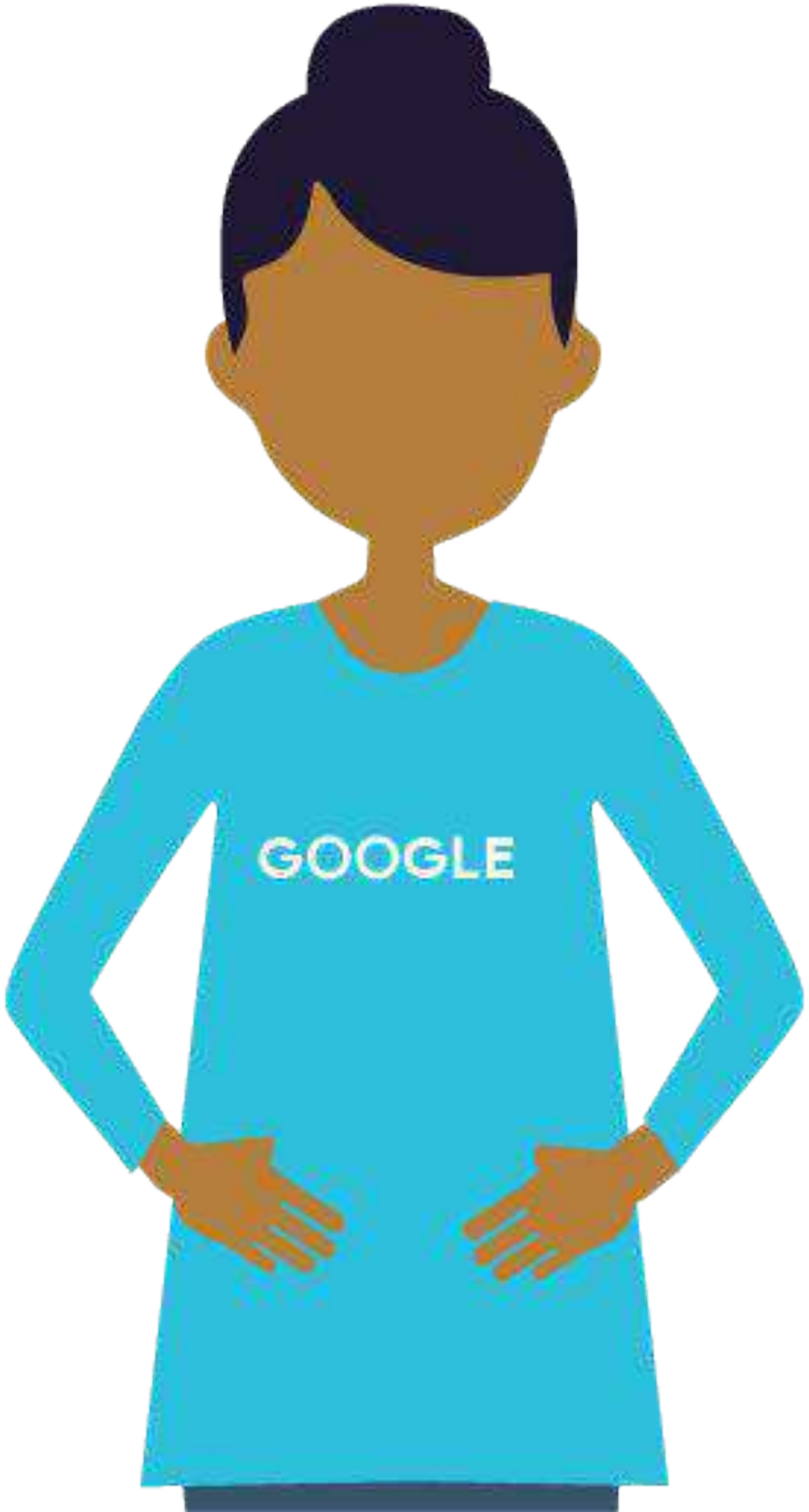
Library 1

Library 2

DATA

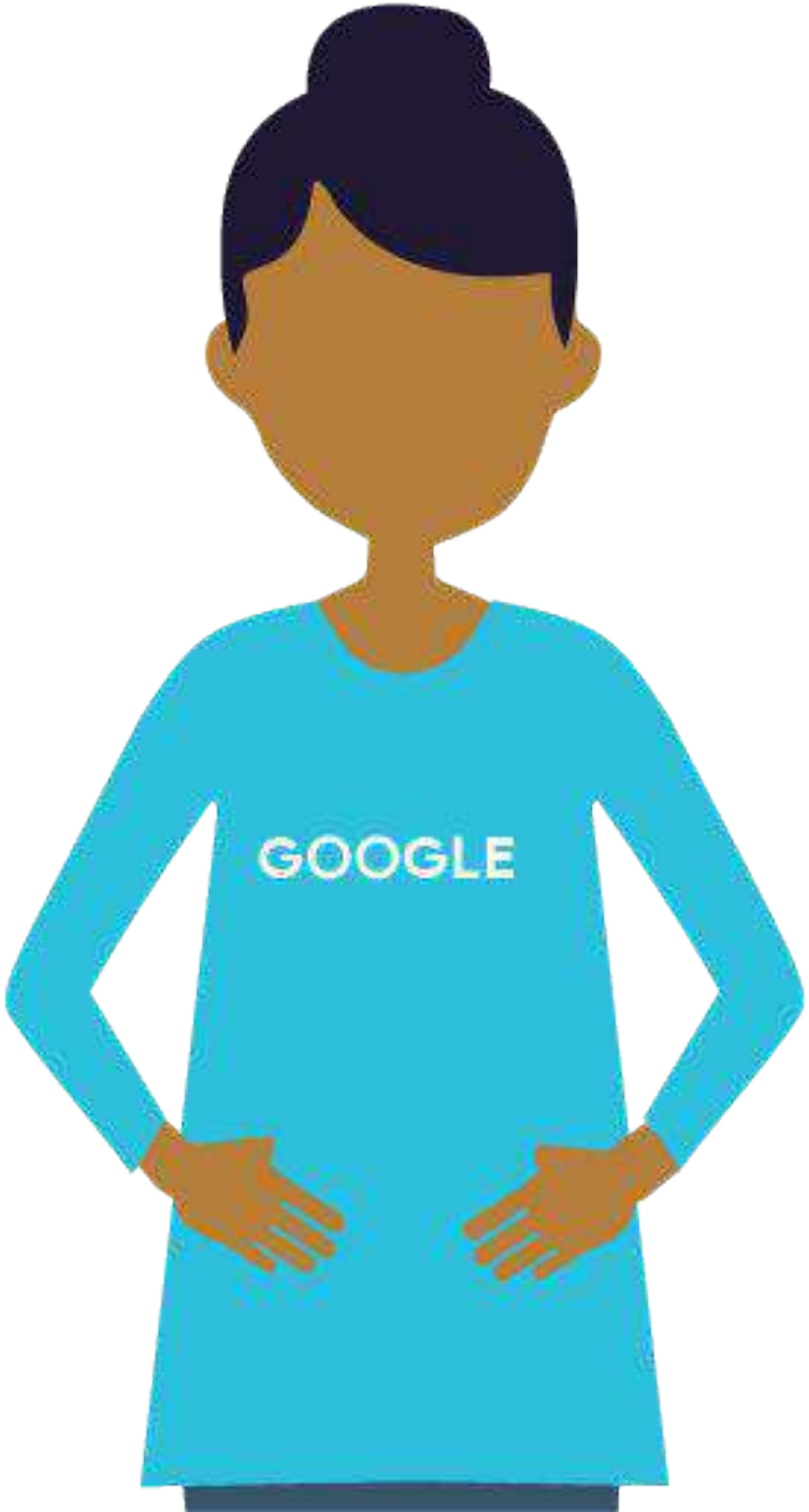
Prgm 1

Prgm 2



Mismanaged Dependencies are Costly

- X Losses in prediction quality
- X Decreases to system stability
- X Decreases in team productivity



Course 2: Production ML Systems

Module 3: Designing Adaptable ML Systems

Lesson Title: **Adapting to Data**

Presenter: Max Lotstein

Format: Talking Head

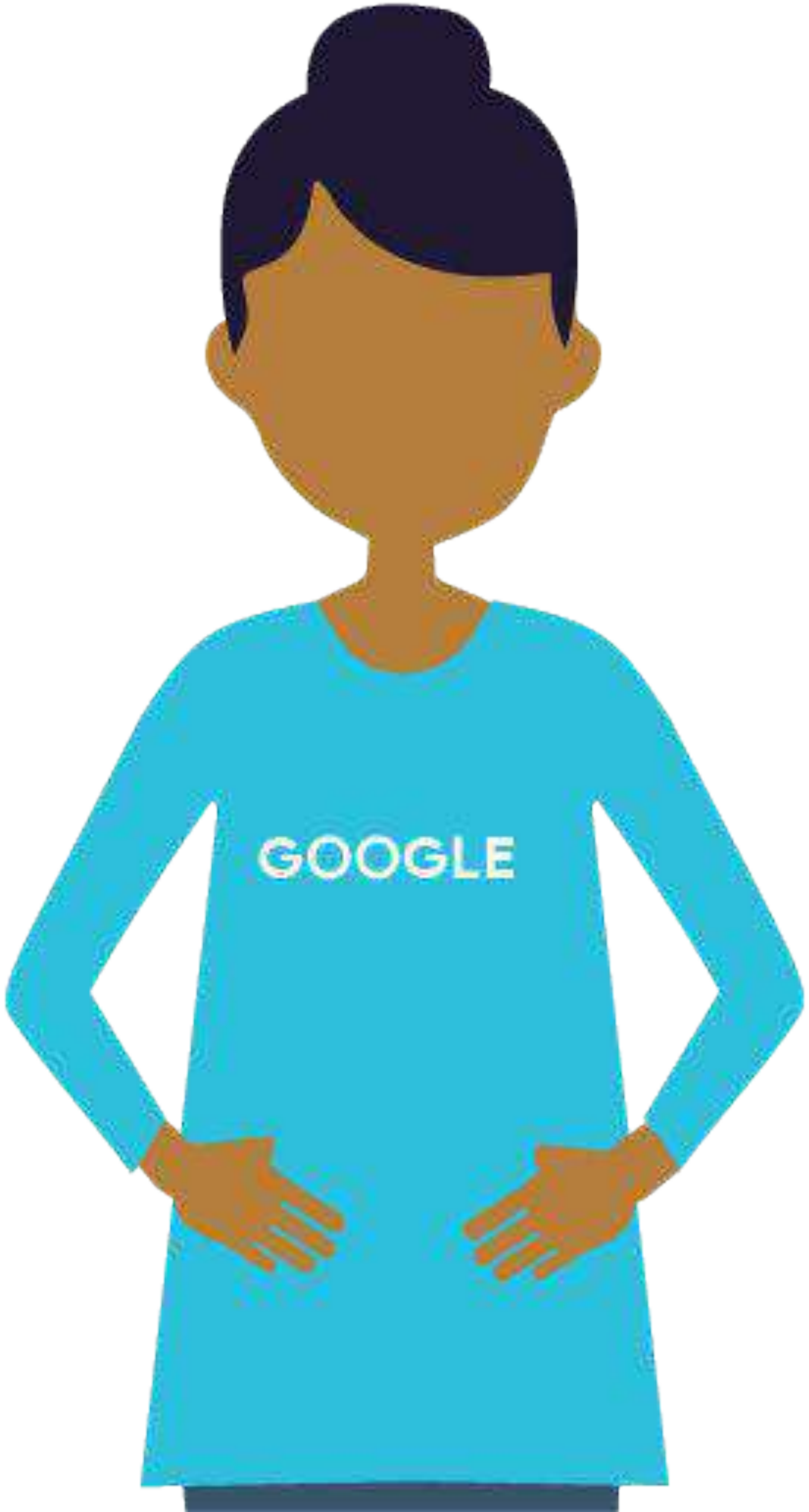
Video Name: T-PSML-O_3_l3_adapting_to_data

Agenda

Adapting to Data

Mitigating Training-Serving Skew Through Design

Debugging a Production Model

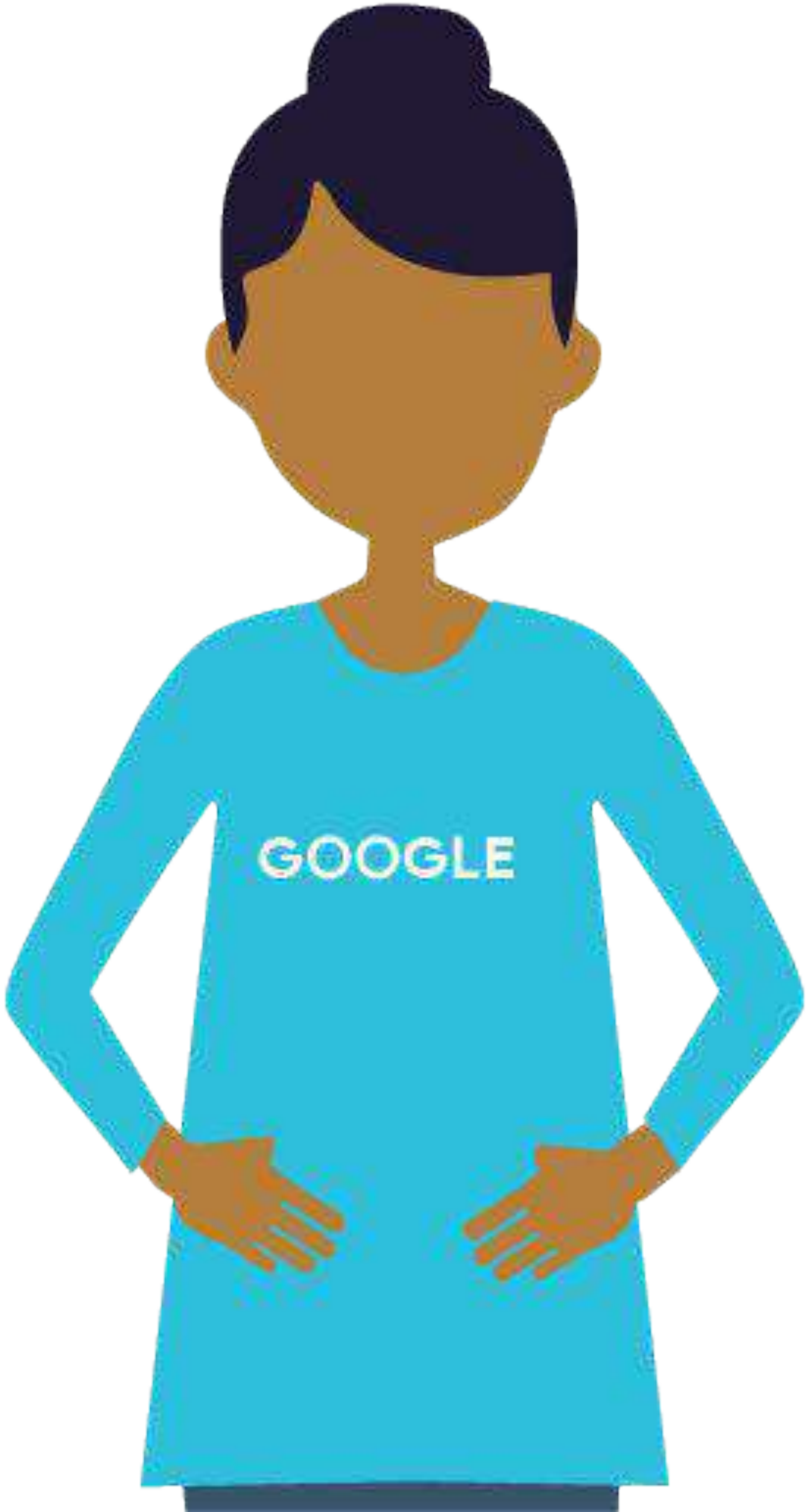


Agenda

Adapting to Data

Mitigating Training-Serving Skew Through Design

Debugging a Production Model

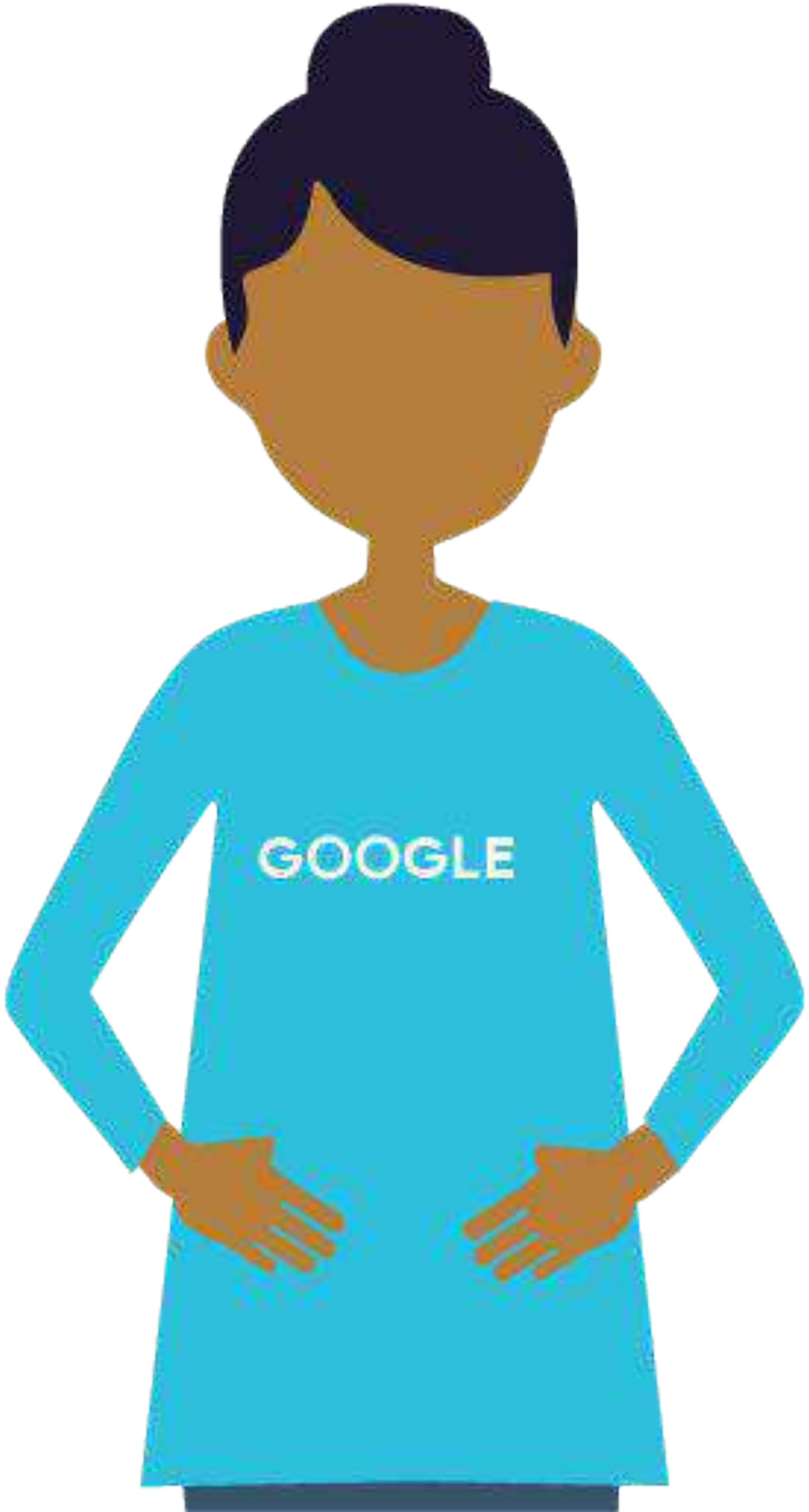


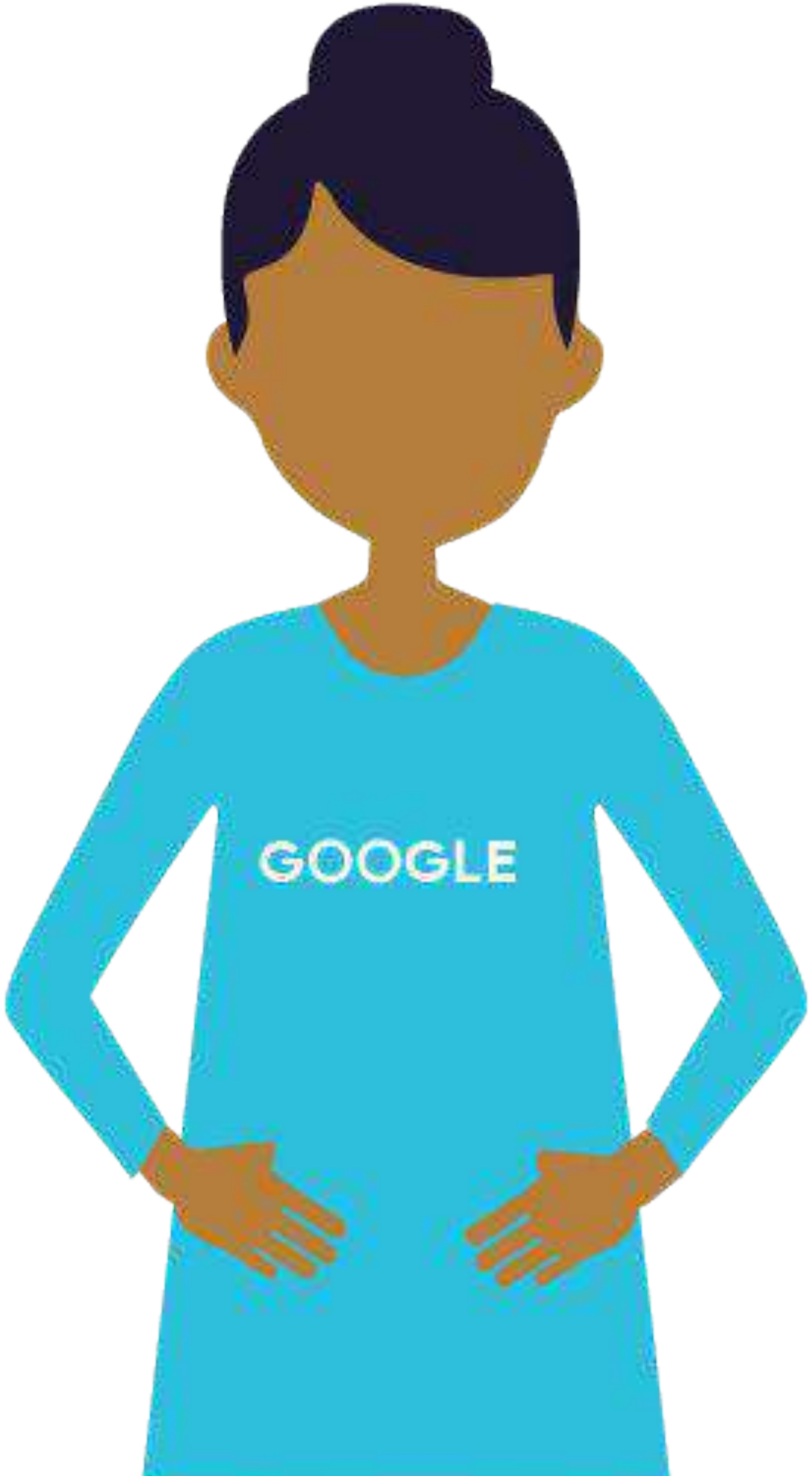
Agenda

Adapting to Data

Mitigating Training-Serving Skew Through Design

Debugging a Production Model



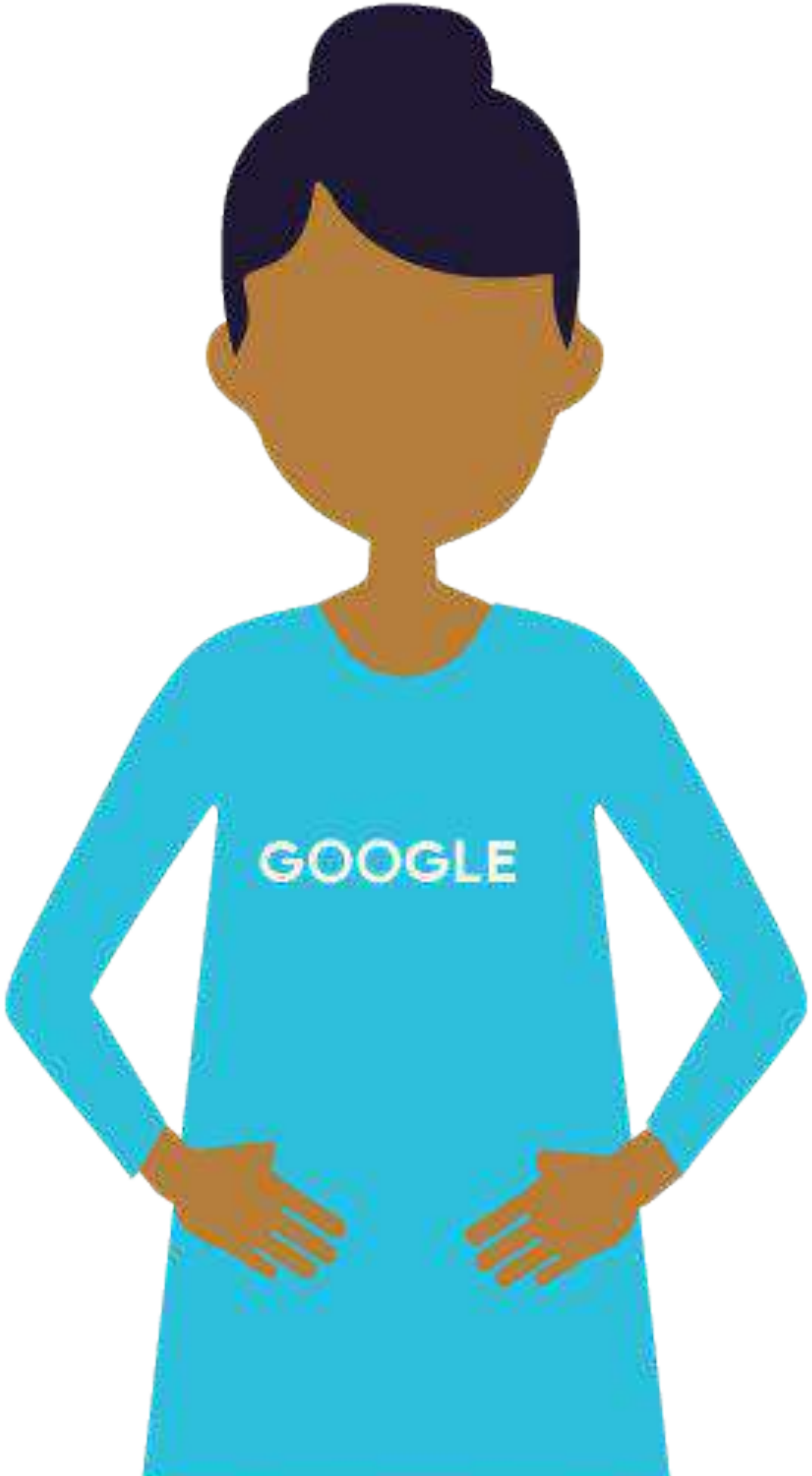


Agenda

Adapting to Data

Mitigating Training-Serving Skew Through Design

Debugging a Production Model

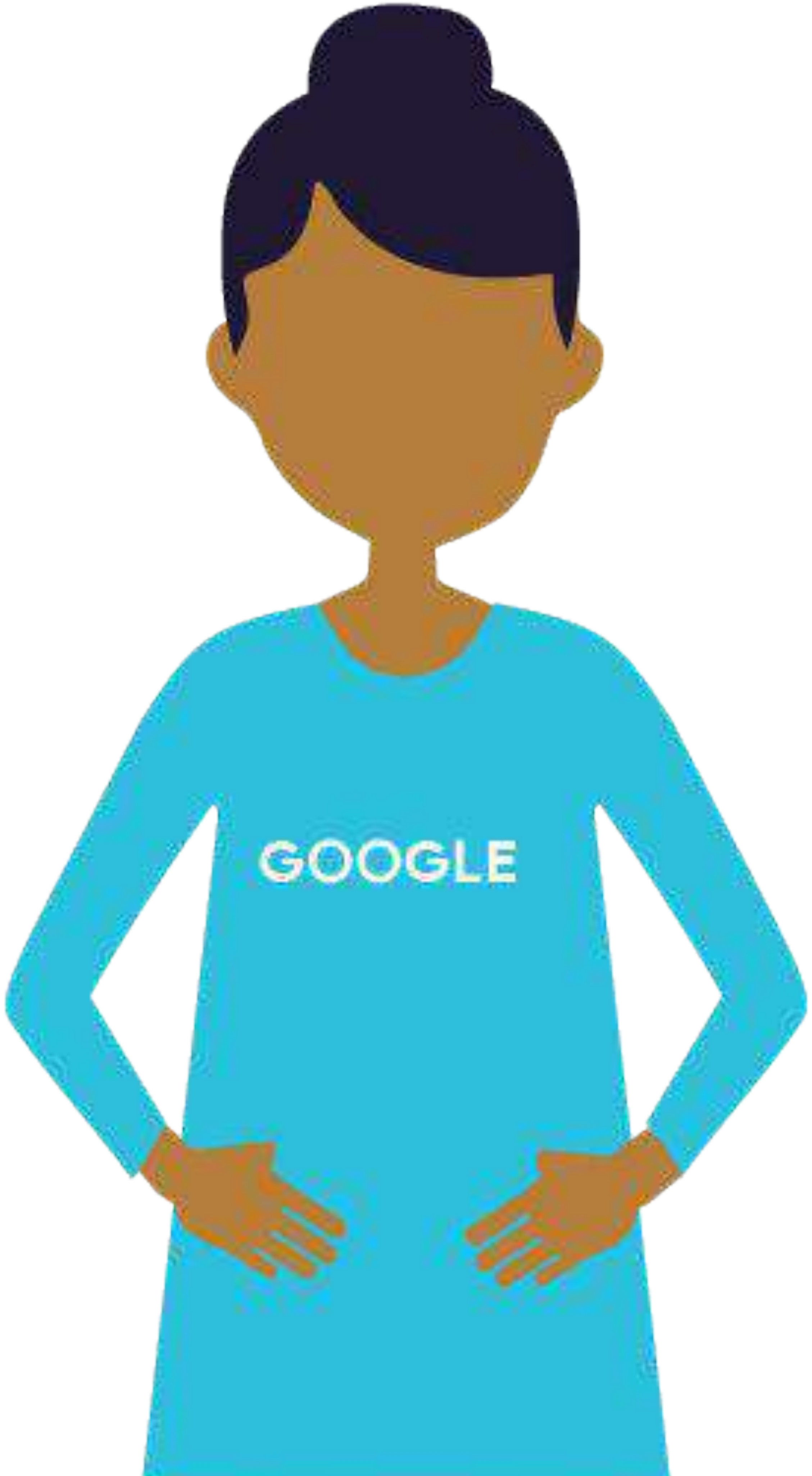


Agenda

Adapting to Data

Mitigating Training-Serving Skew Through Design

Debugging a Production Model

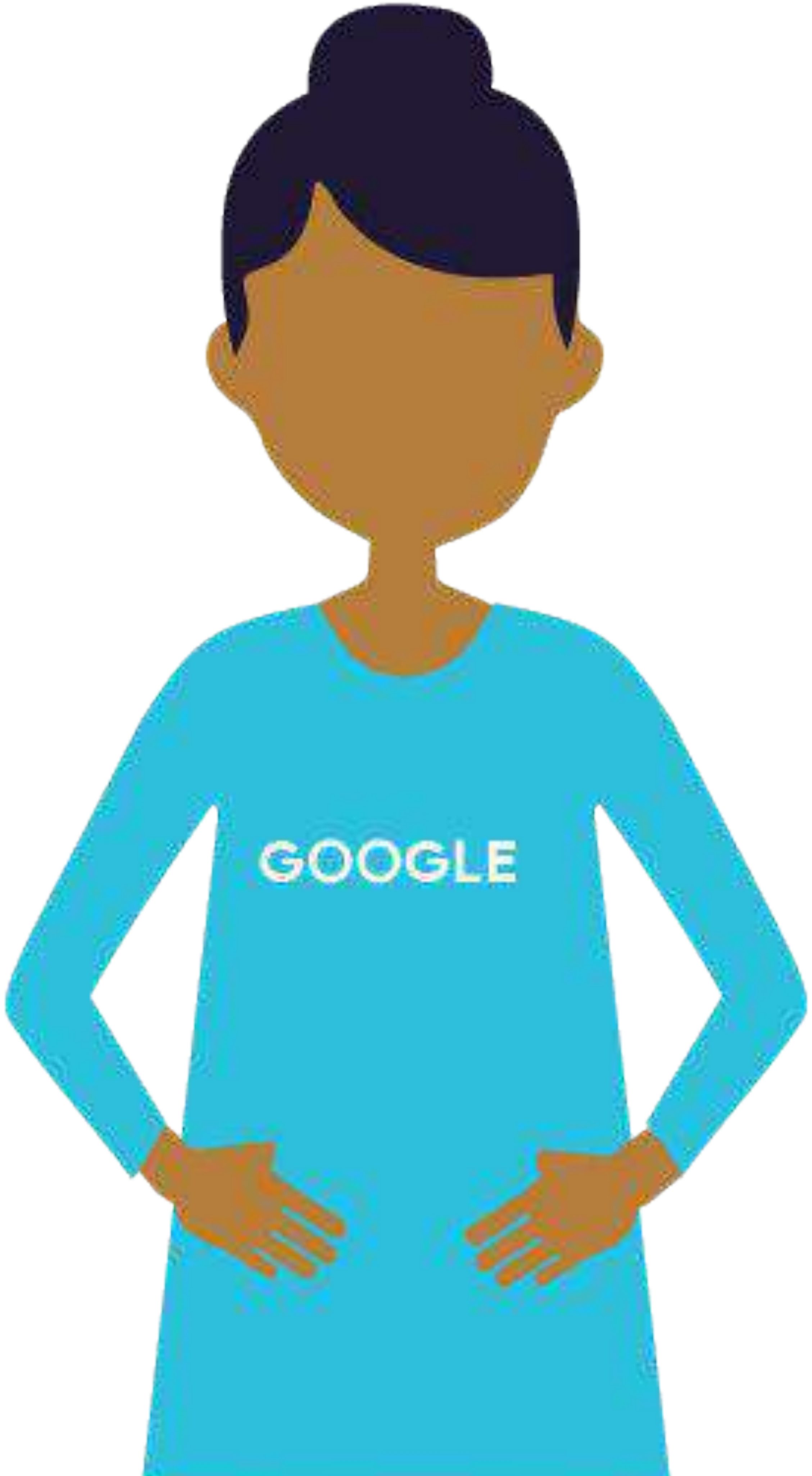


Agenda

Adapting to Data

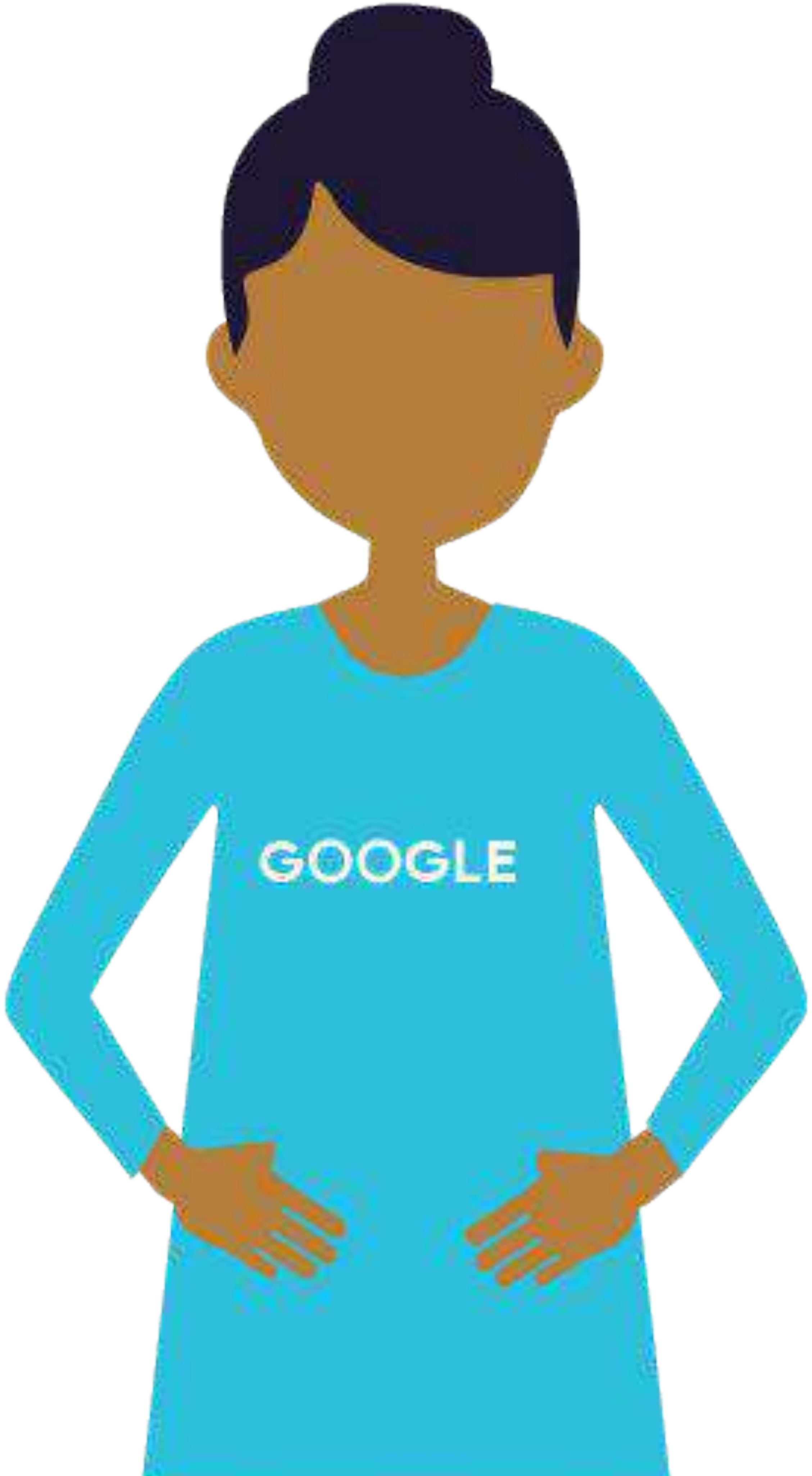
Mitigating Training-Serving Skew Through Design

Debugging a Production Model



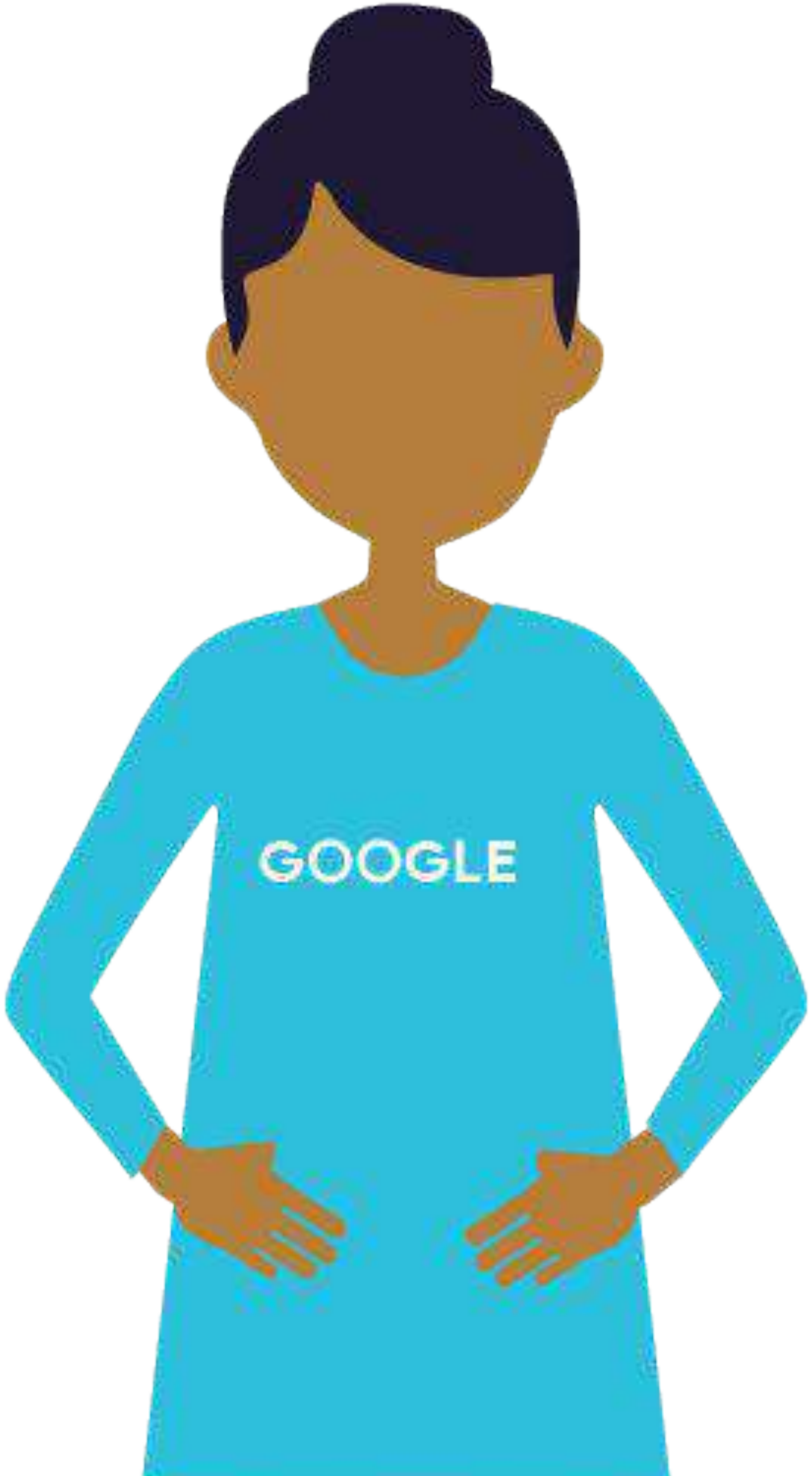
Which of these is least likely to change?

1. An upstream model
2. A data source maintained by another team
3. The relationship between features and labels
4. The distribution of inputs

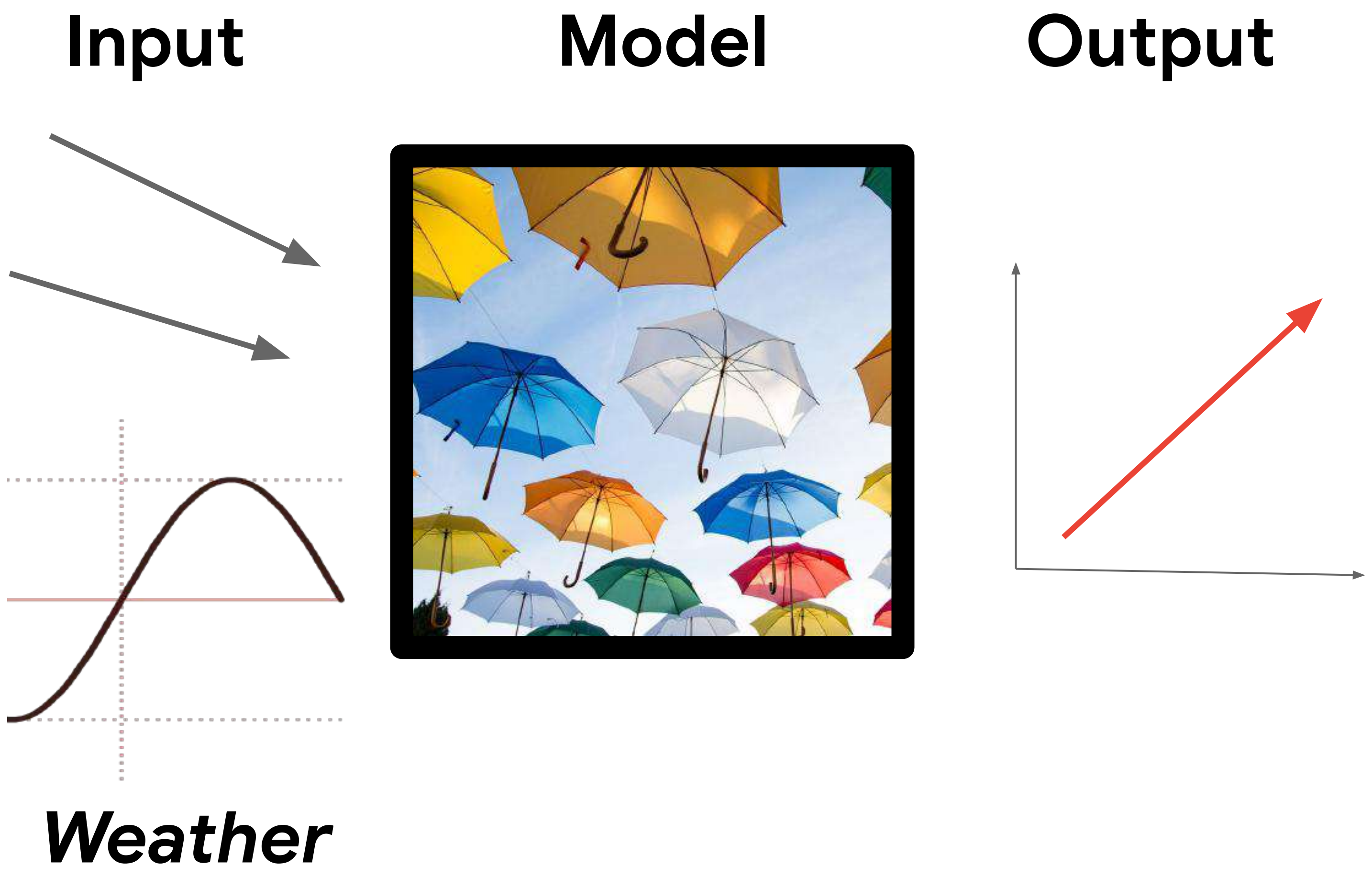


Which of these is least likely to change?

1. **An upstream model**
2. **A data source maintained by another team**
3. **The relationship between features and labels**
4. **The distribution of inputs**



Decoupled upstream data producers

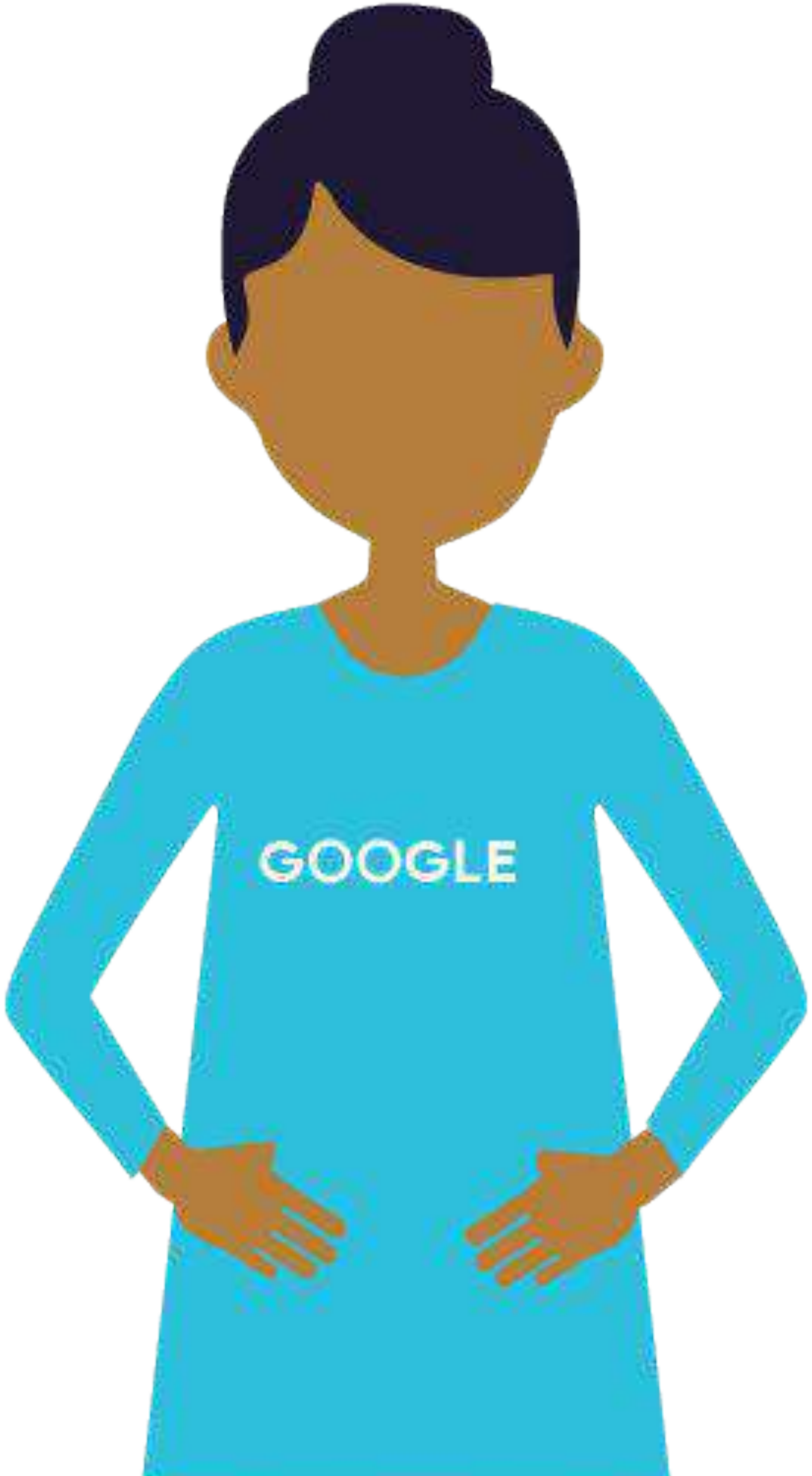


Input

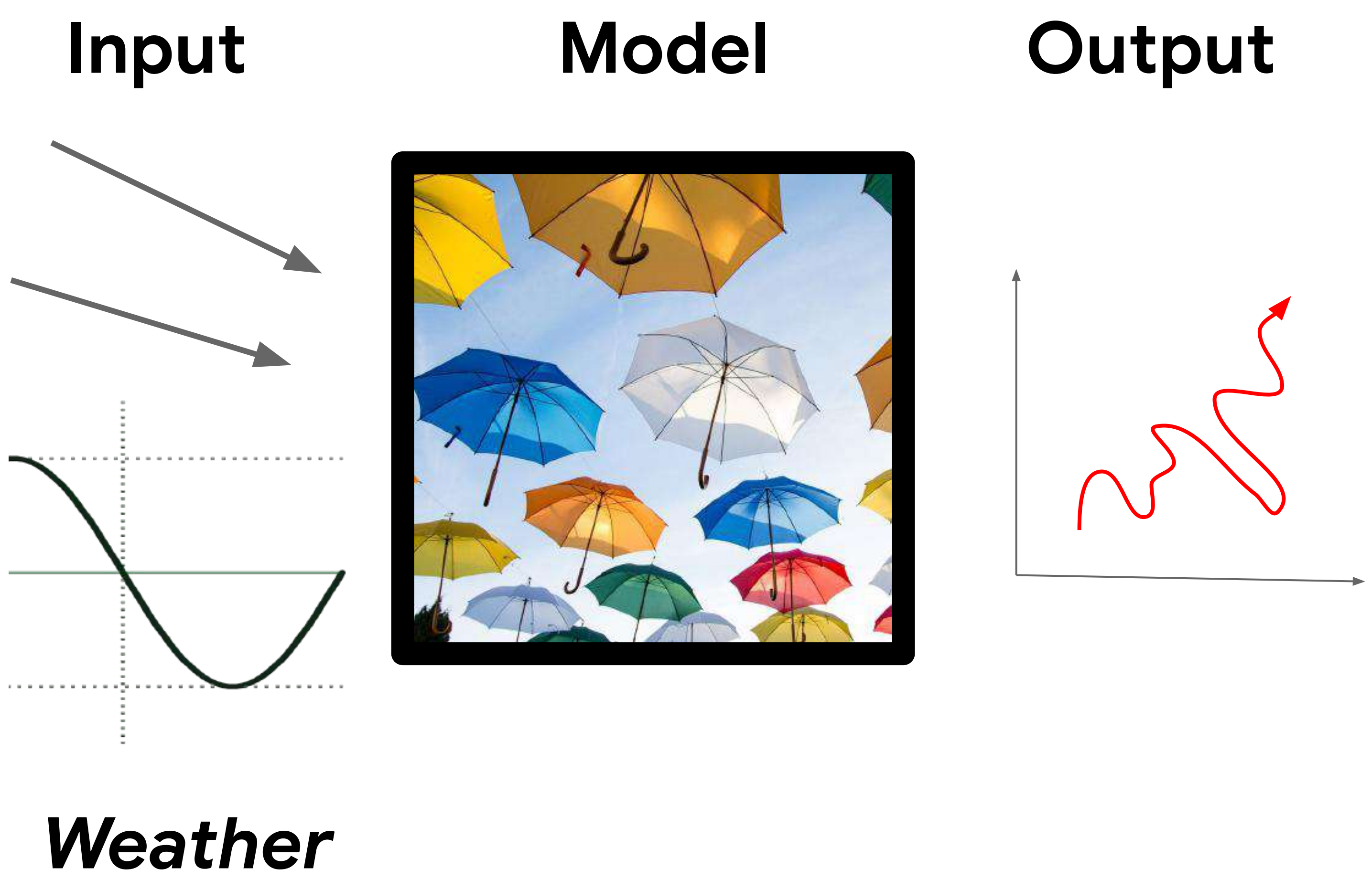
Model

Output

Weather



Decoupled upstream data producers

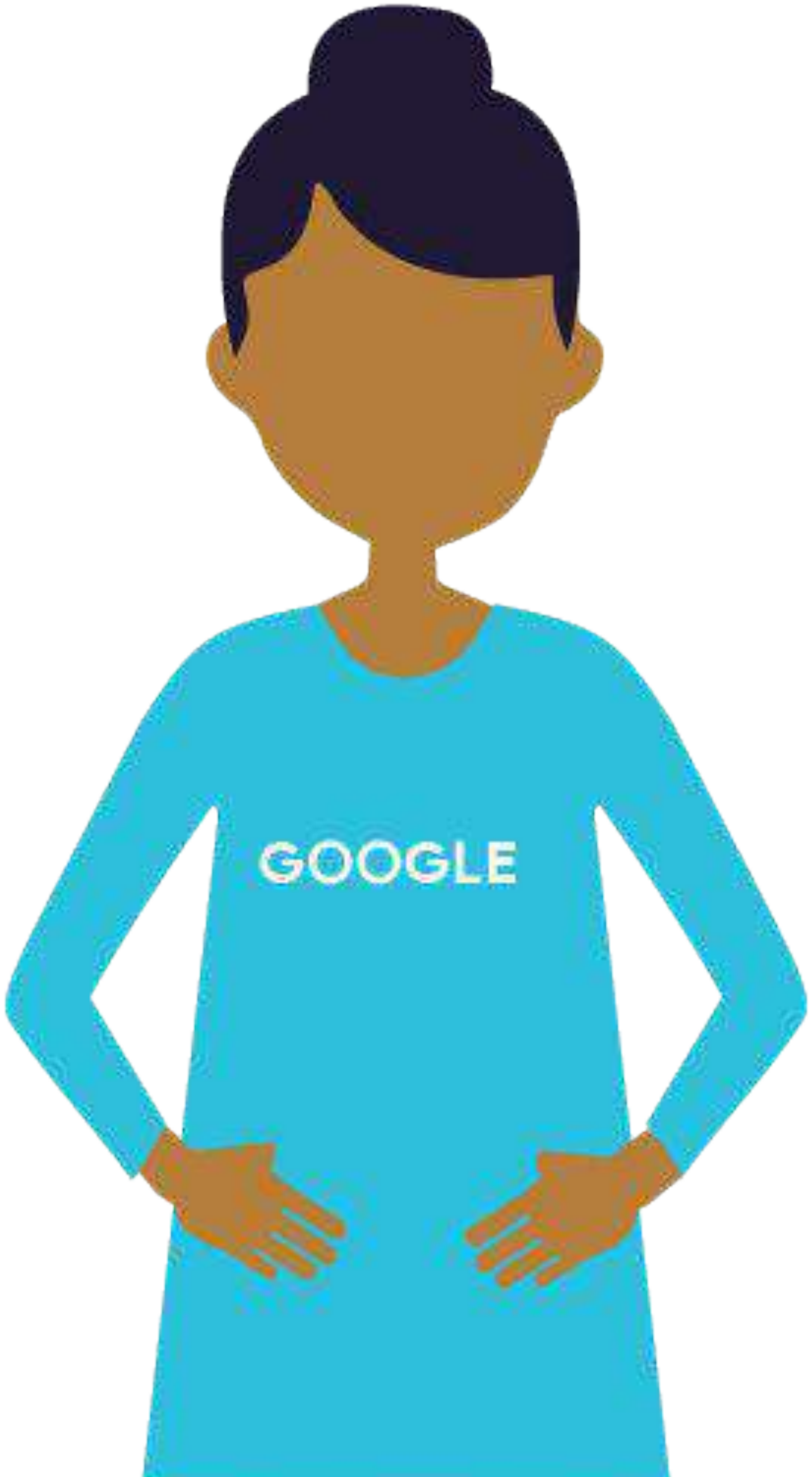


Input

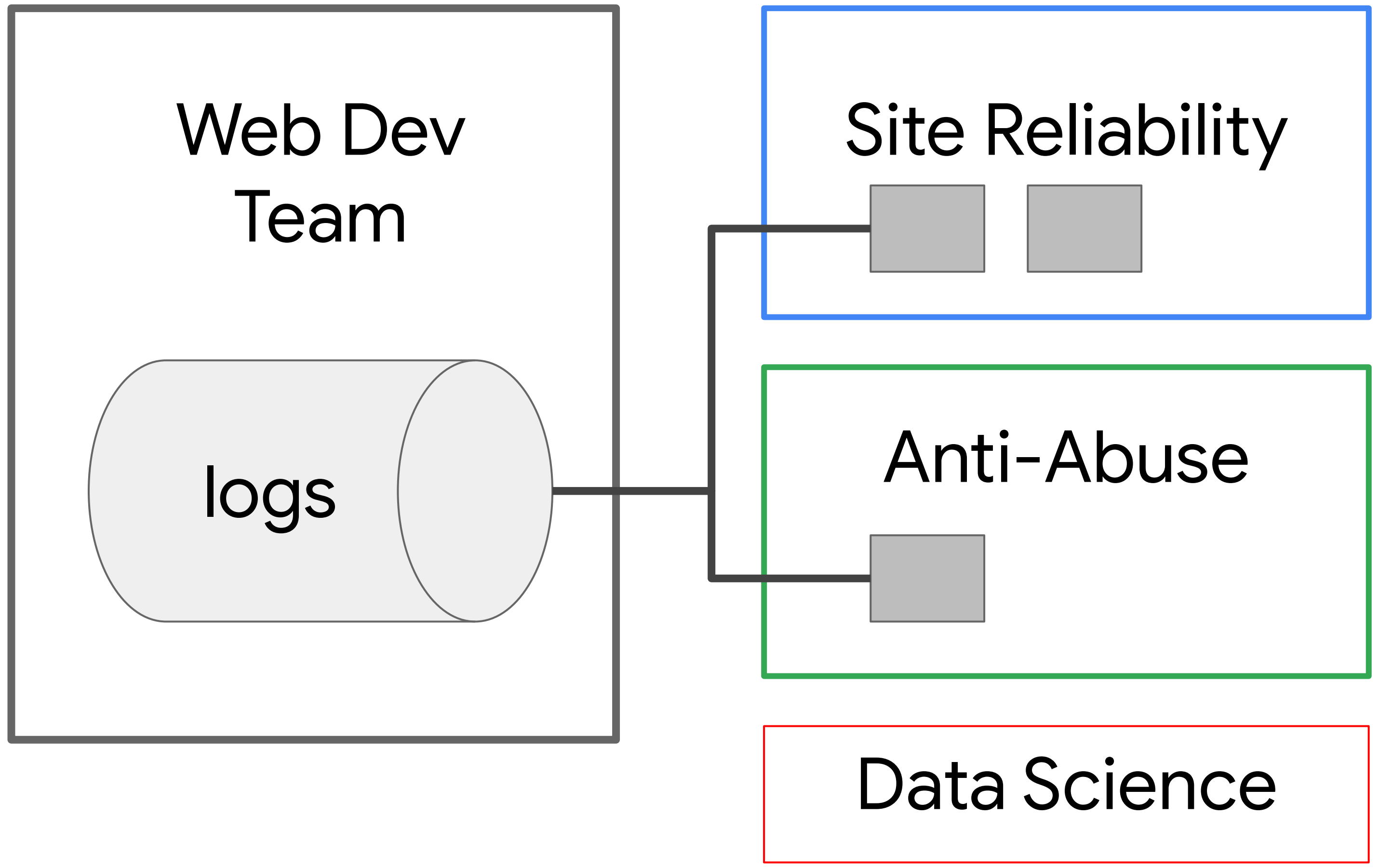
Model

Output

Weather



Decoupled upstream data producers



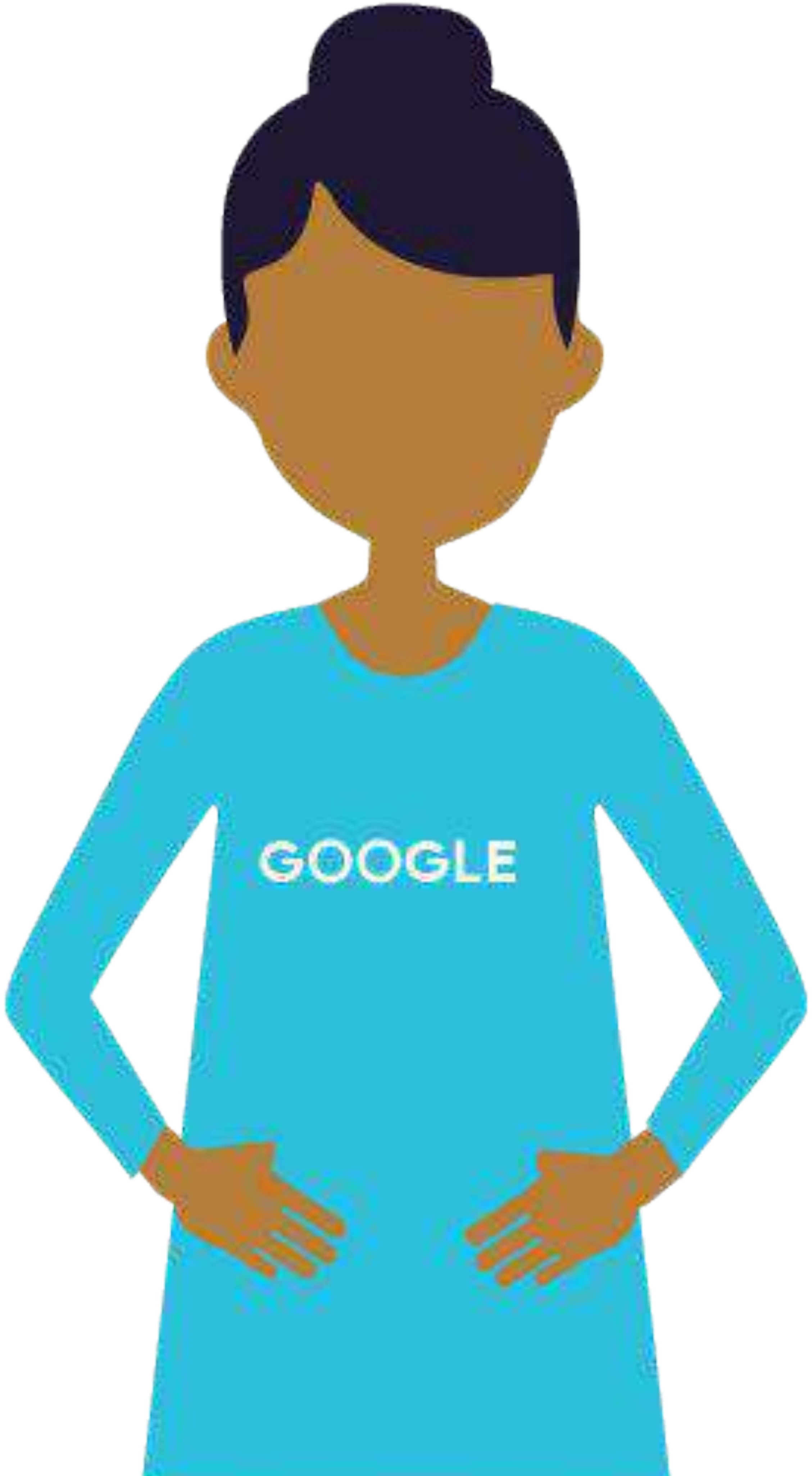
Web Dev Team

logs

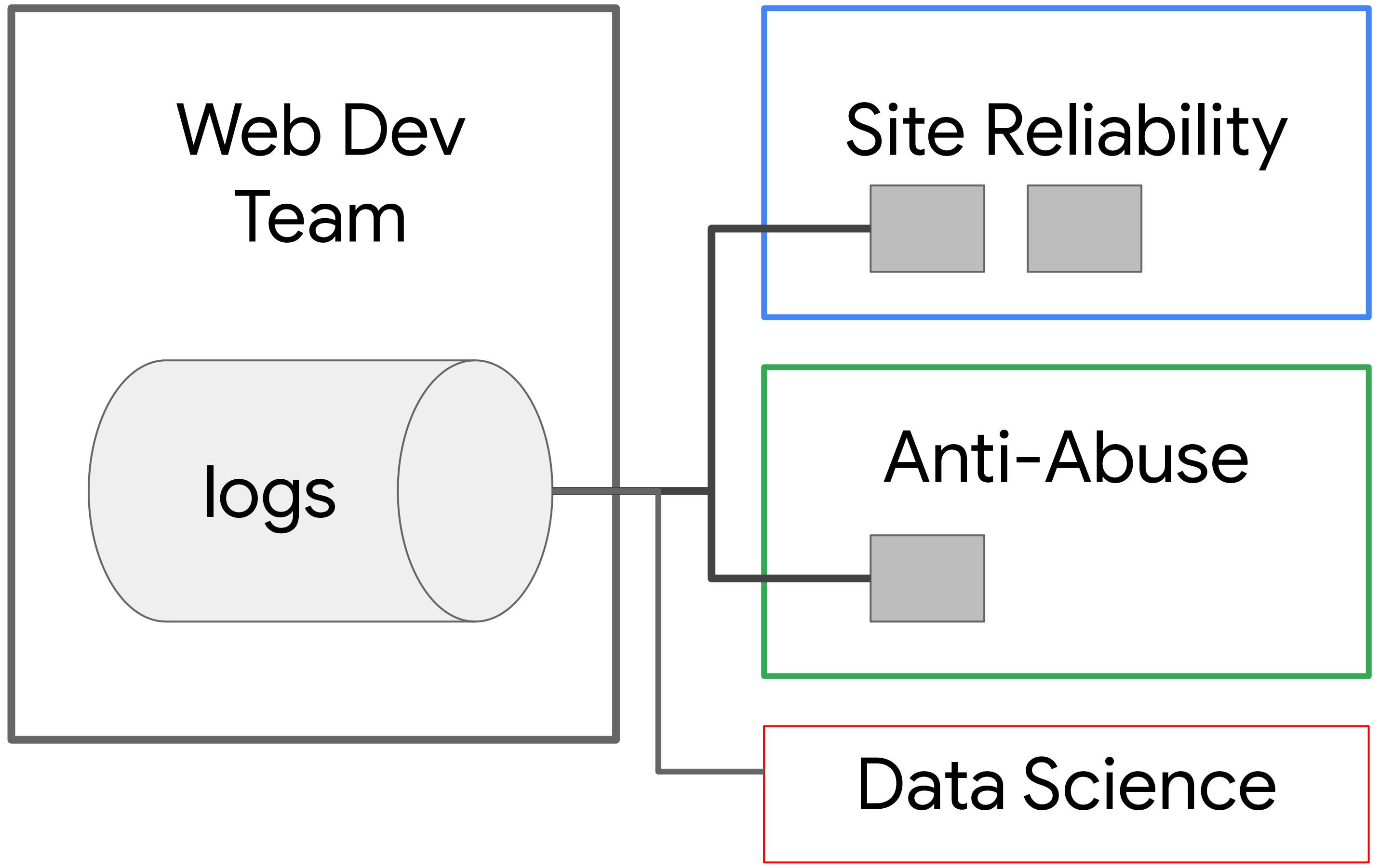
Site Reliability

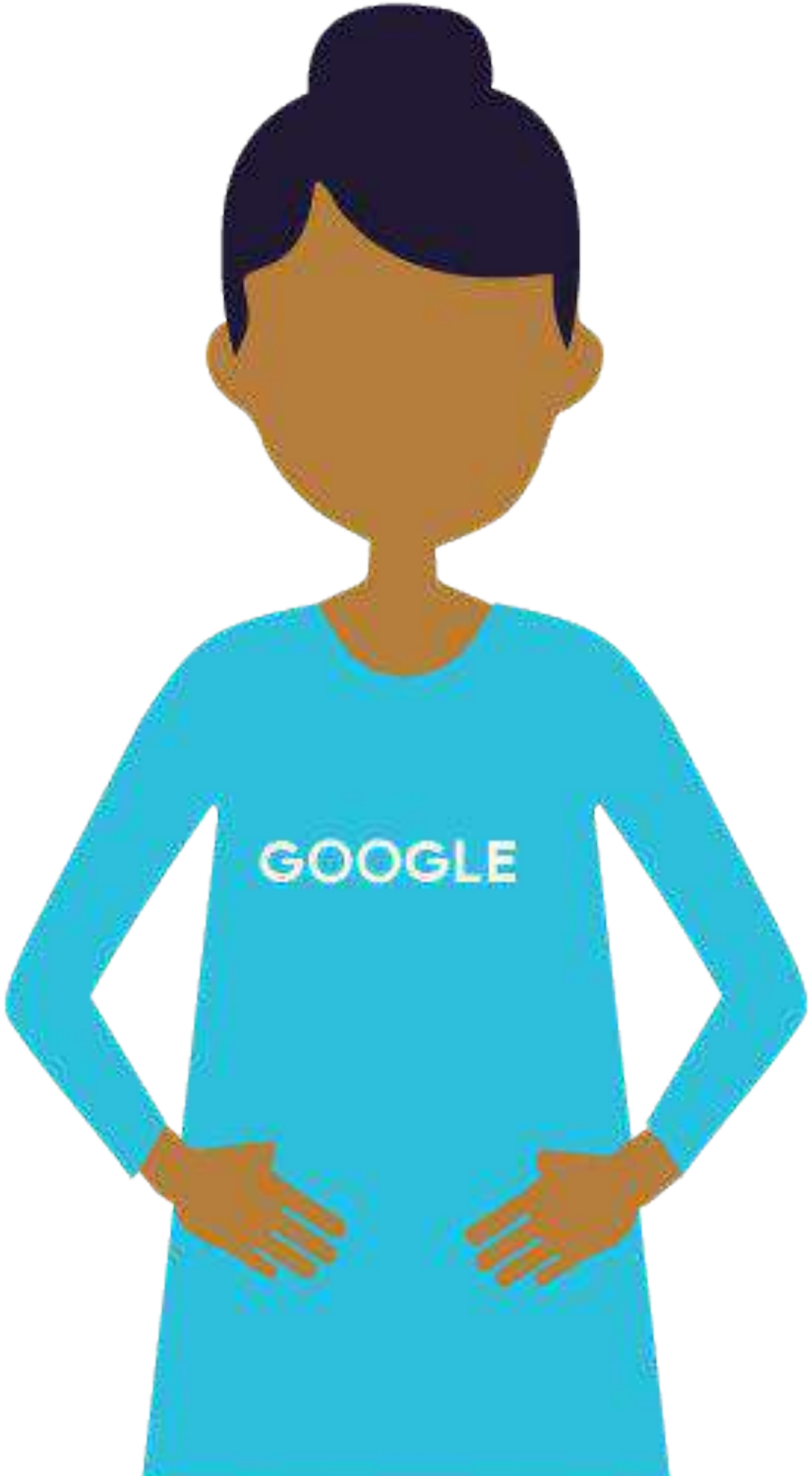
Anti-Abuse

Data Science

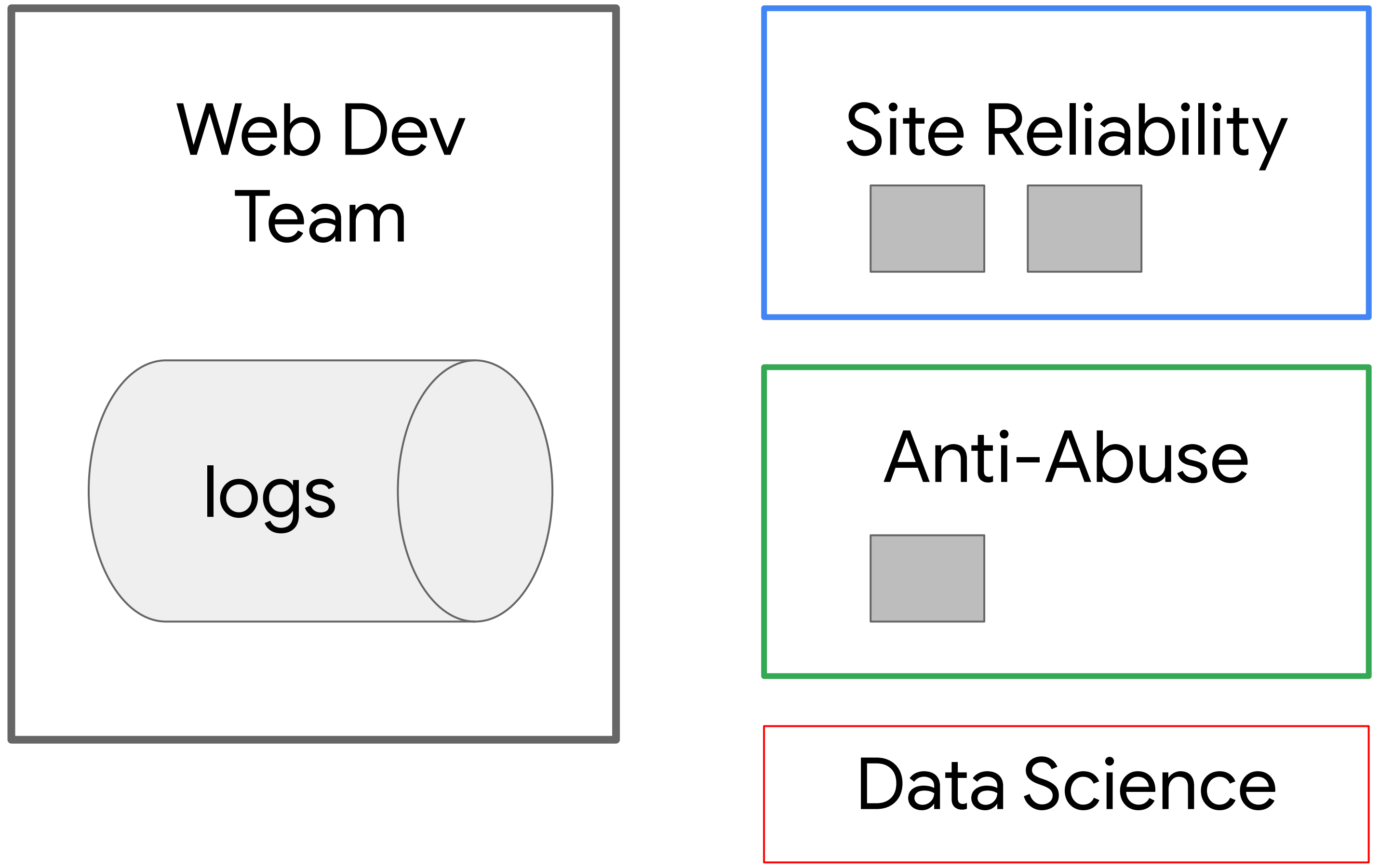


Decoupled upstream data producers





Decoupled upstream data producers



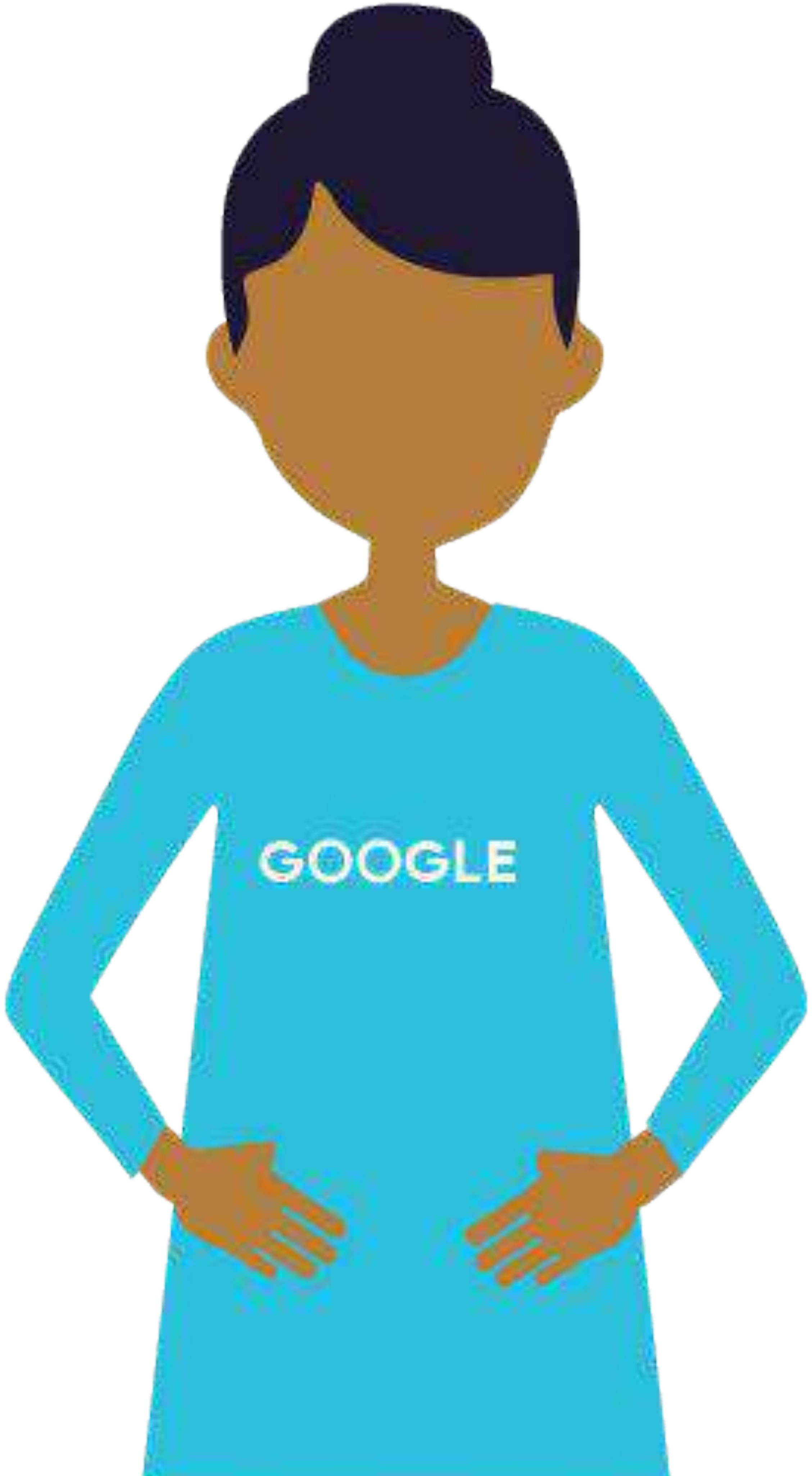
Web Dev Team

logs

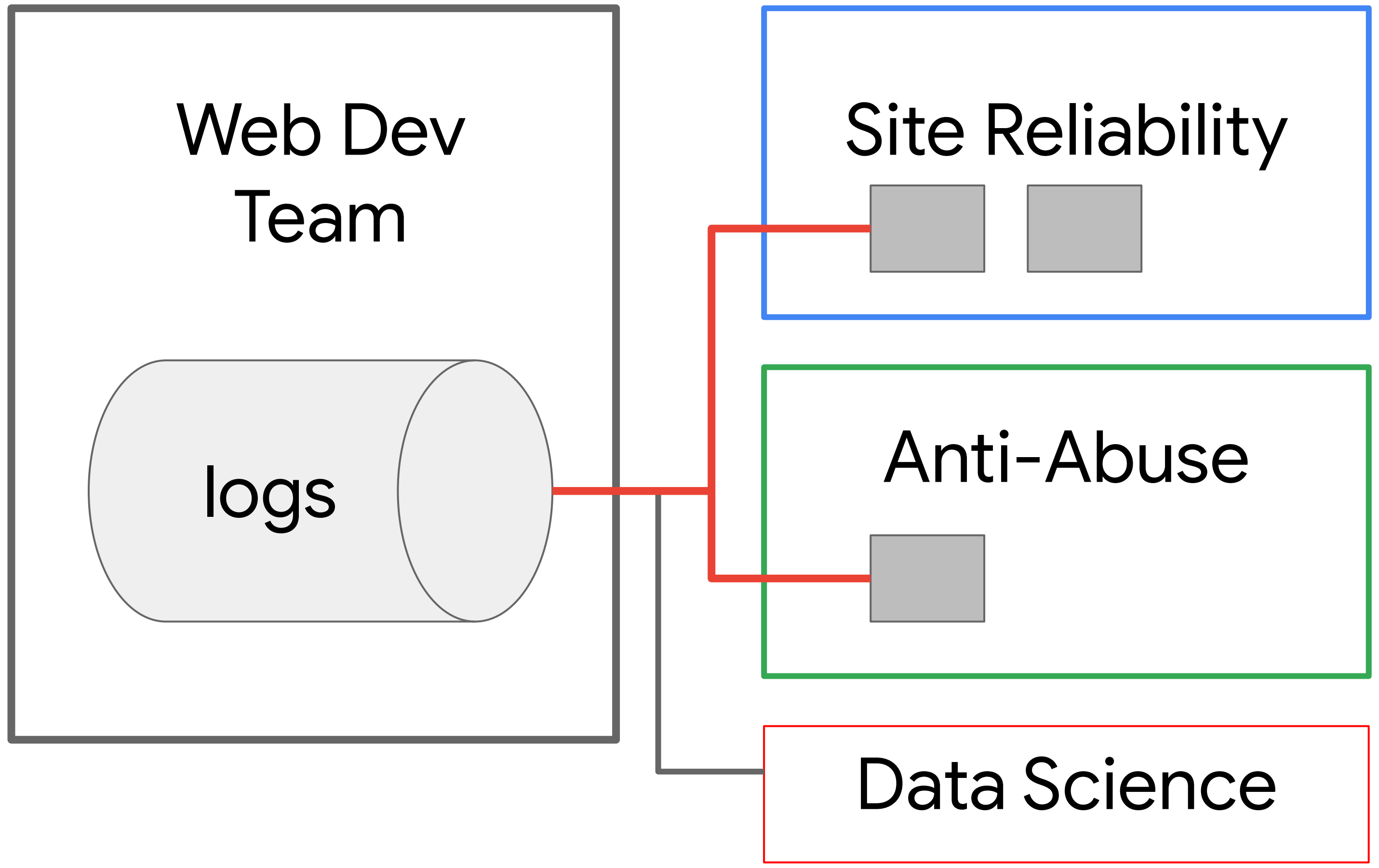
Site Reliability

Anti-Abuse

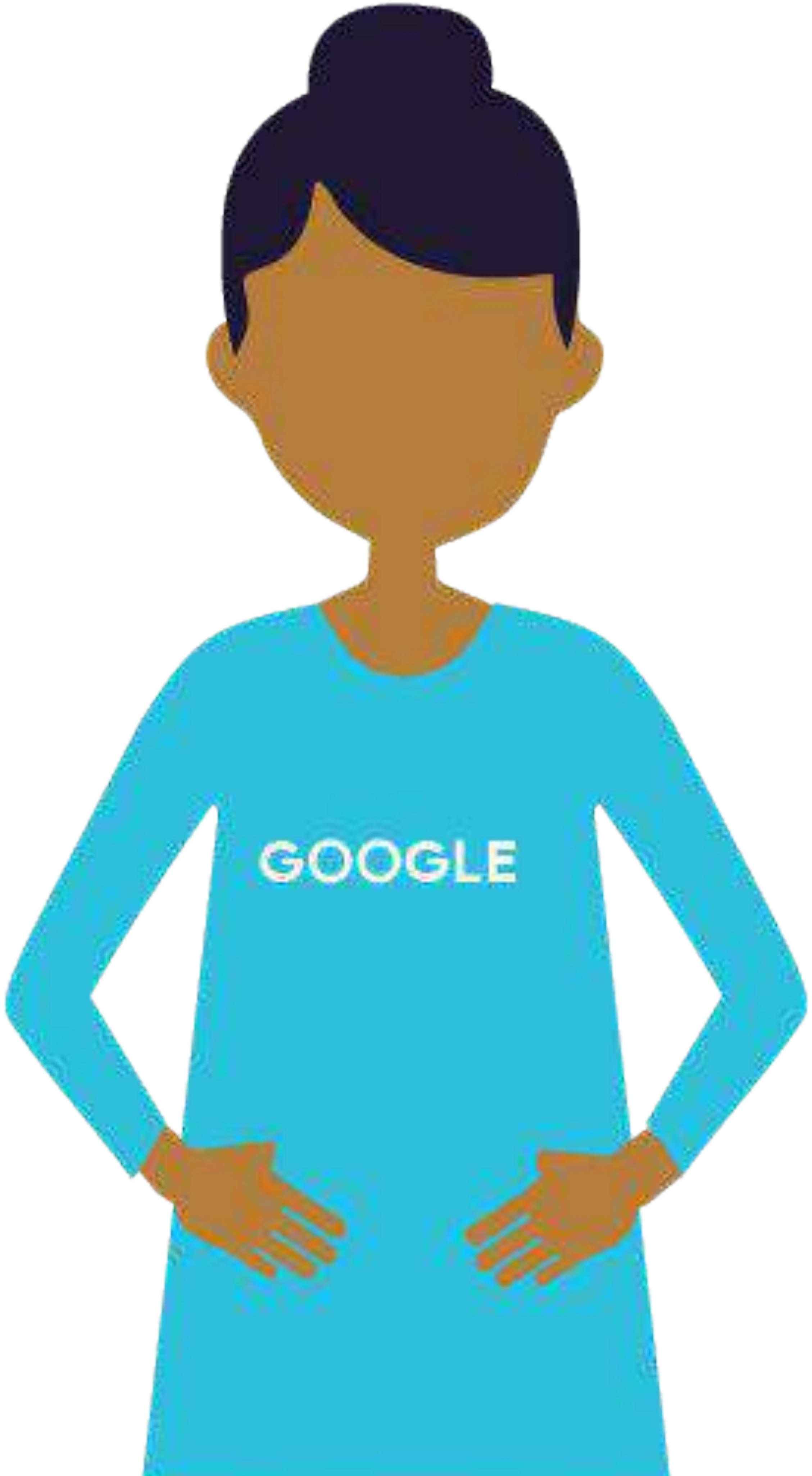
Data Science

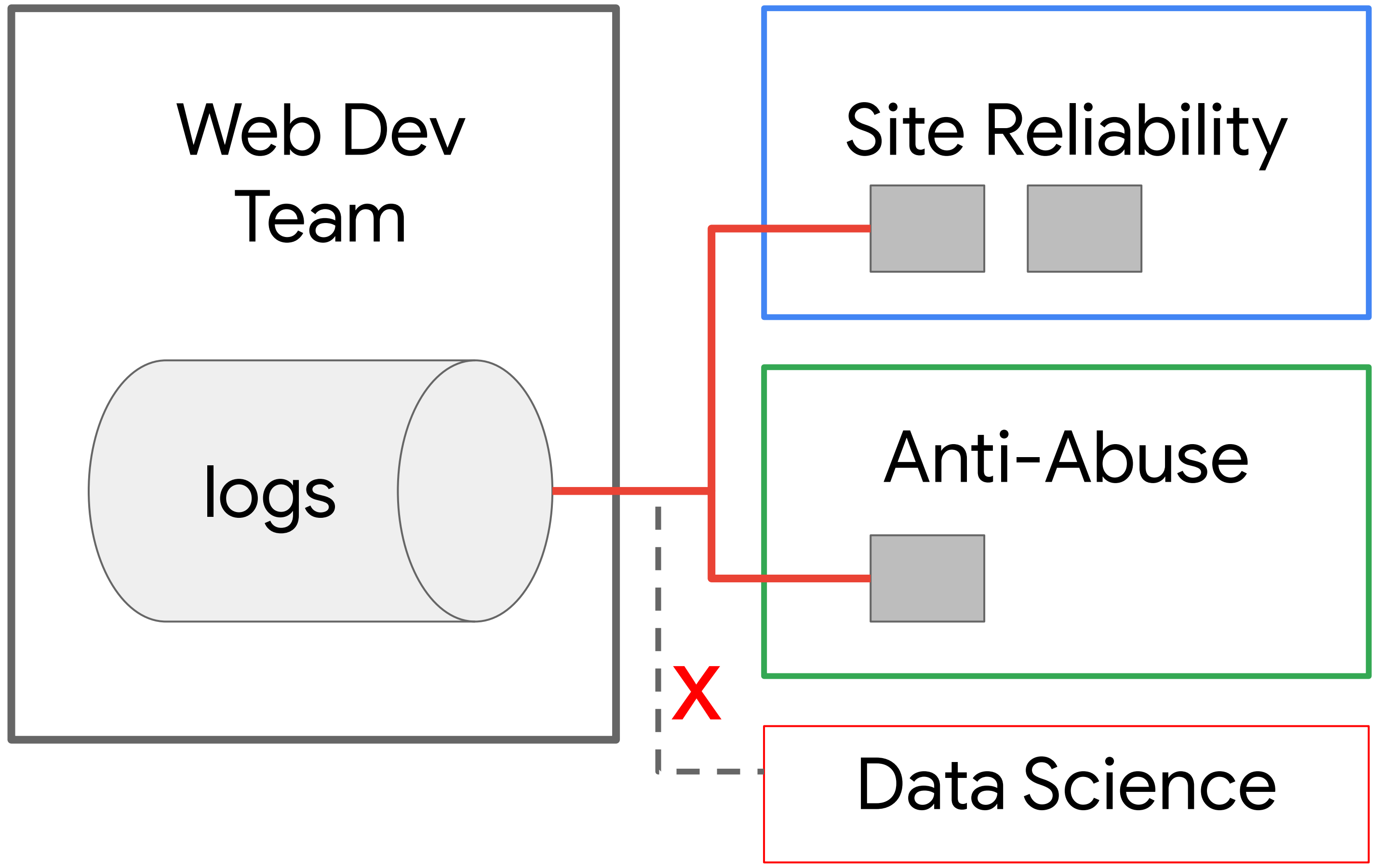


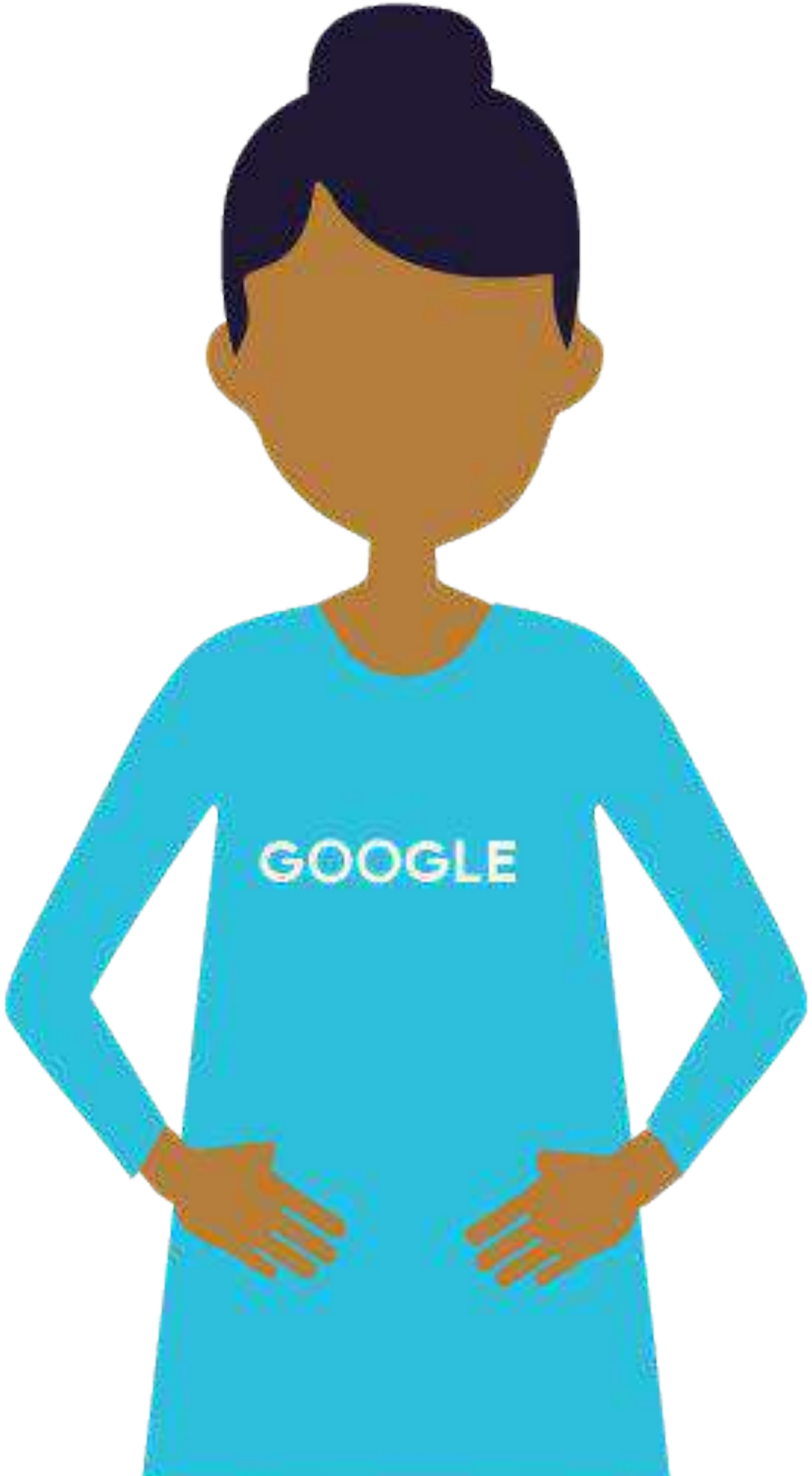
Decoupled upstream data producers



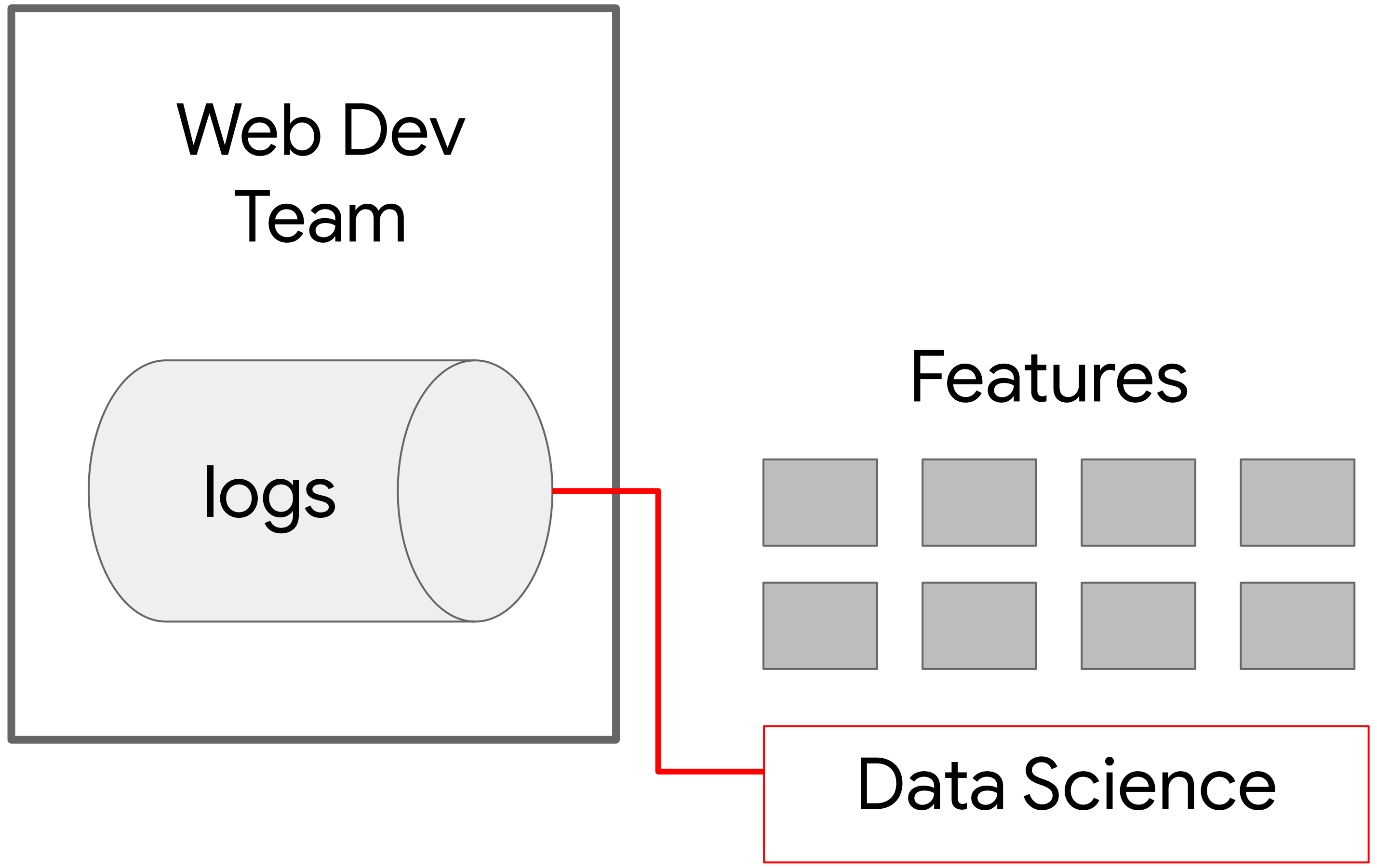
Decoupled upstream data producers

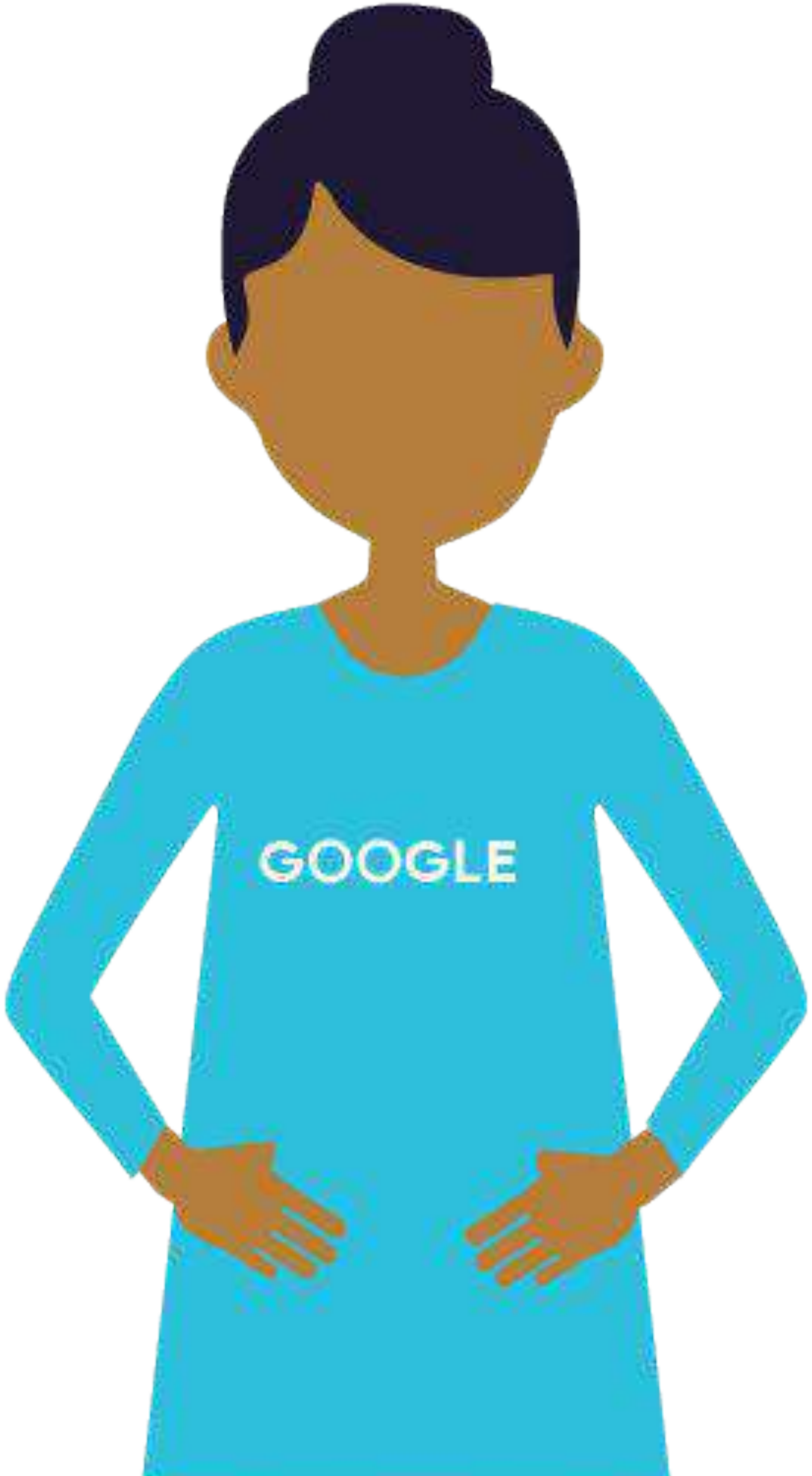




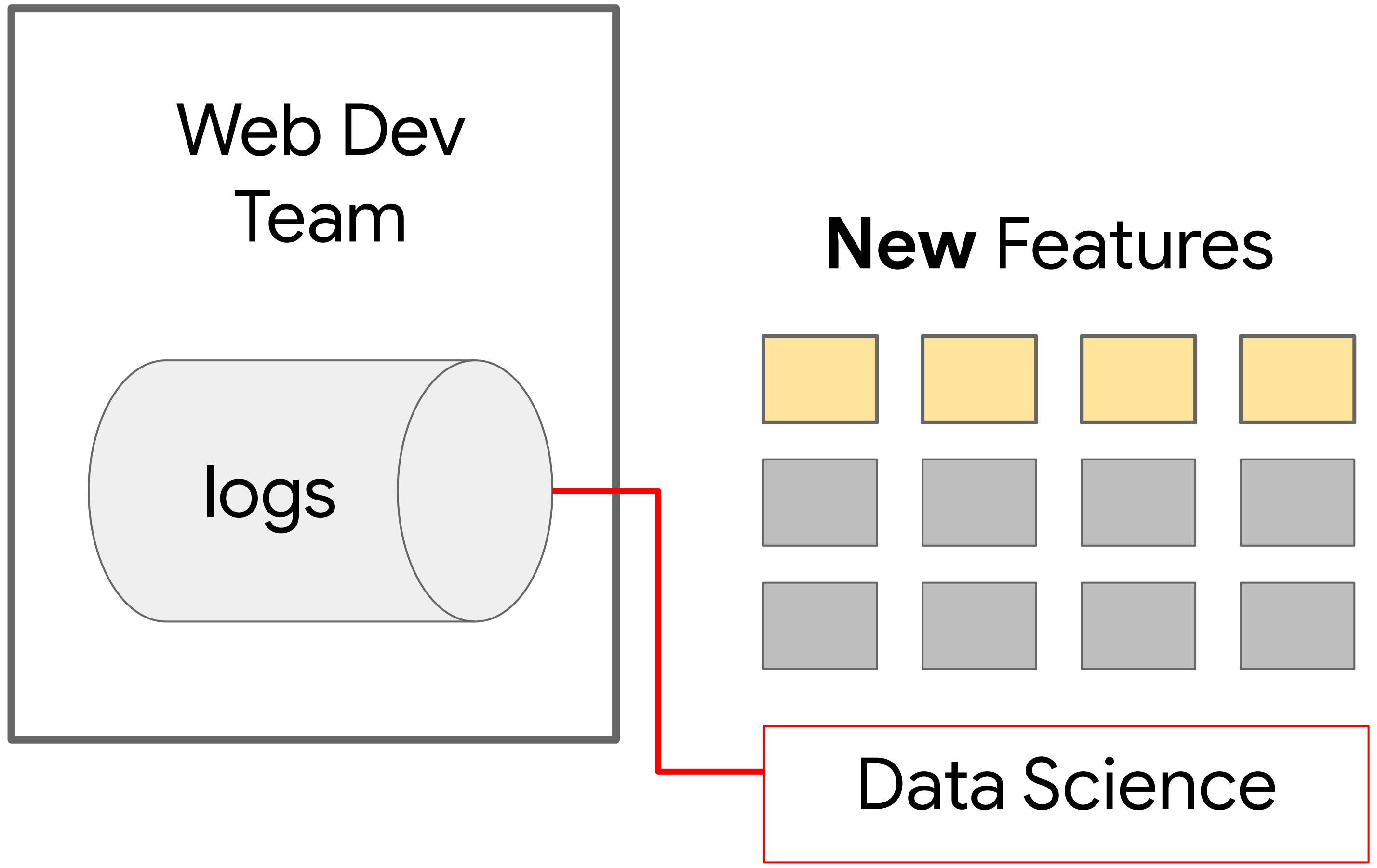


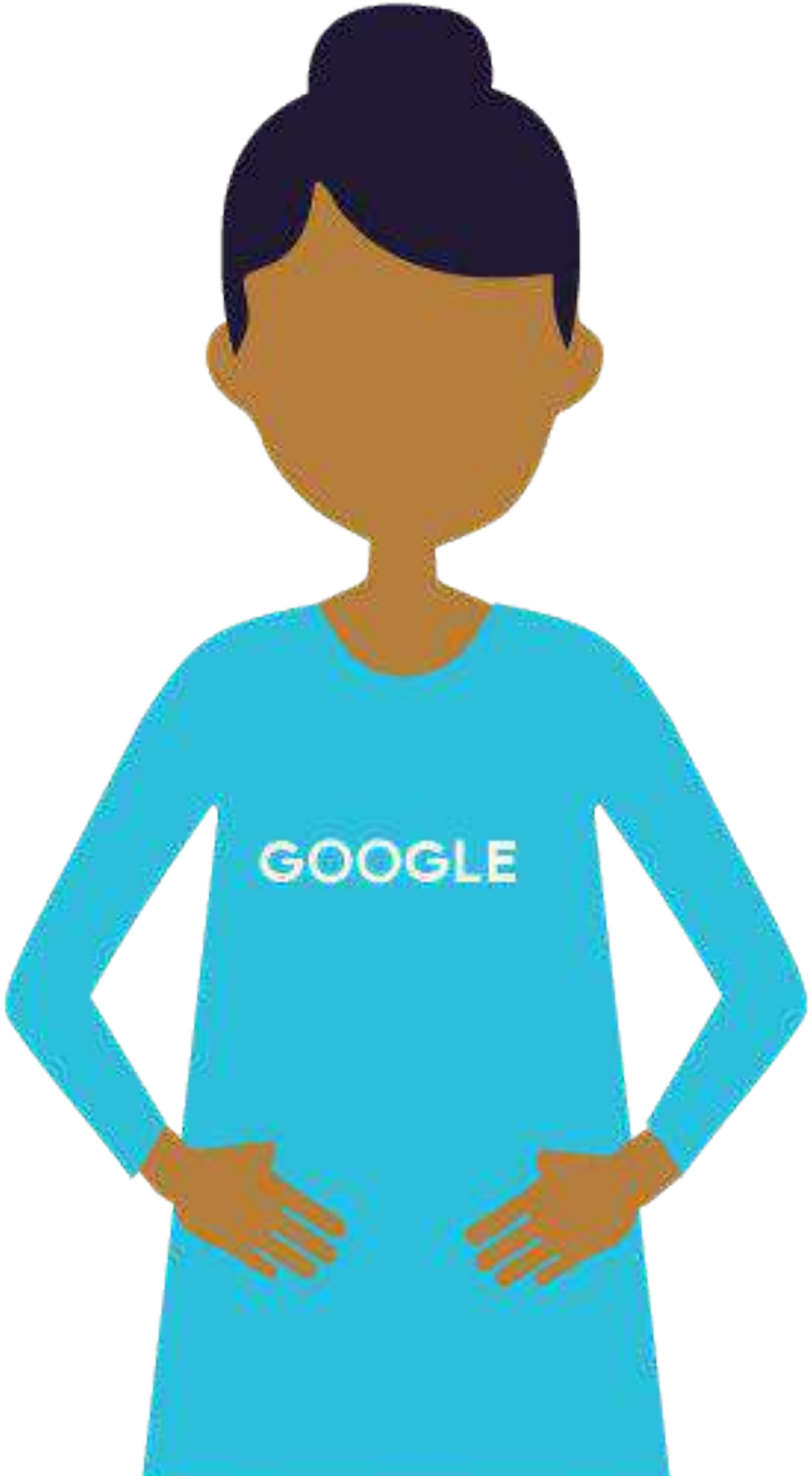
Underutilized data dependencies



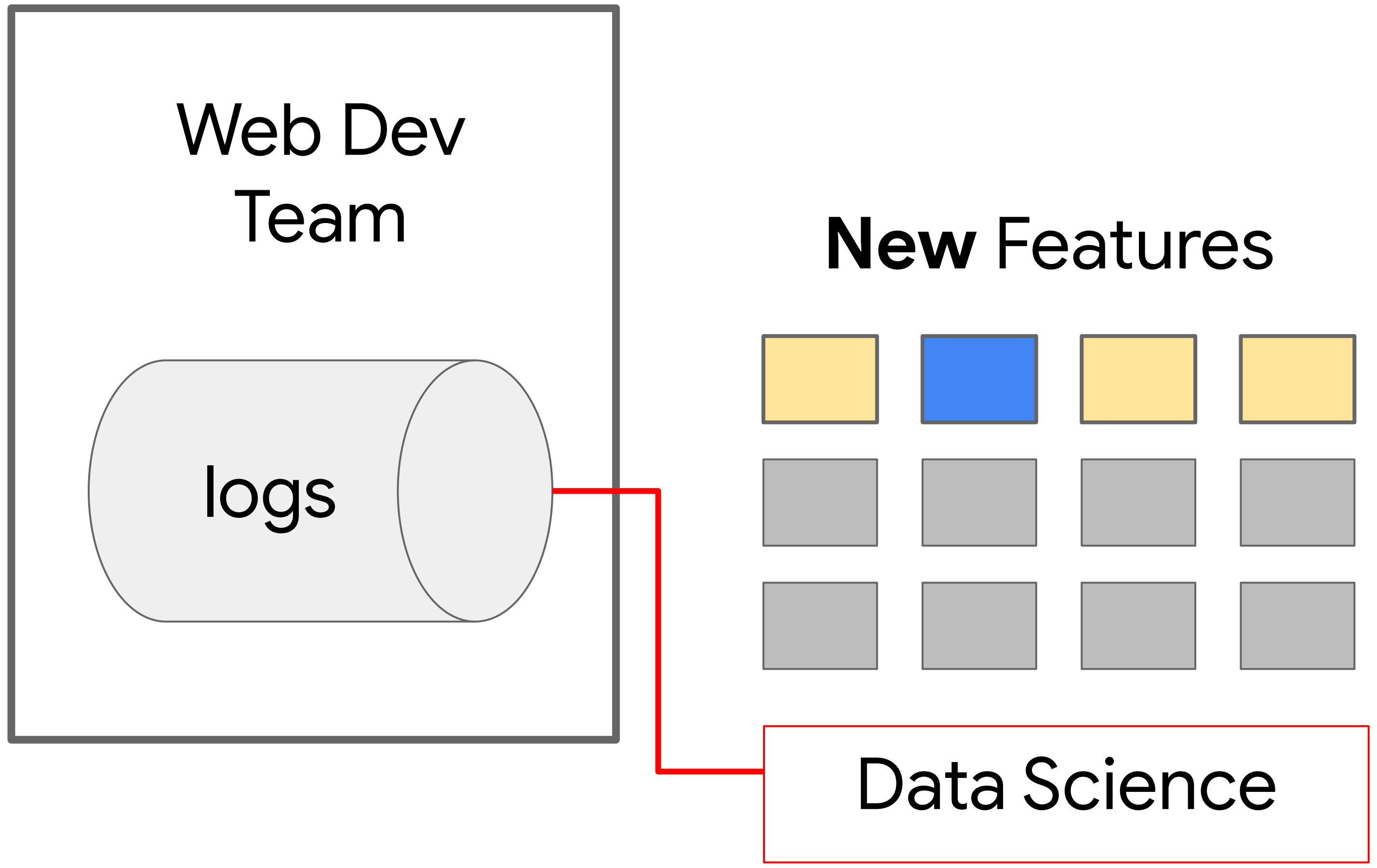


Underutilized data dependencies





Underutilized data dependencies

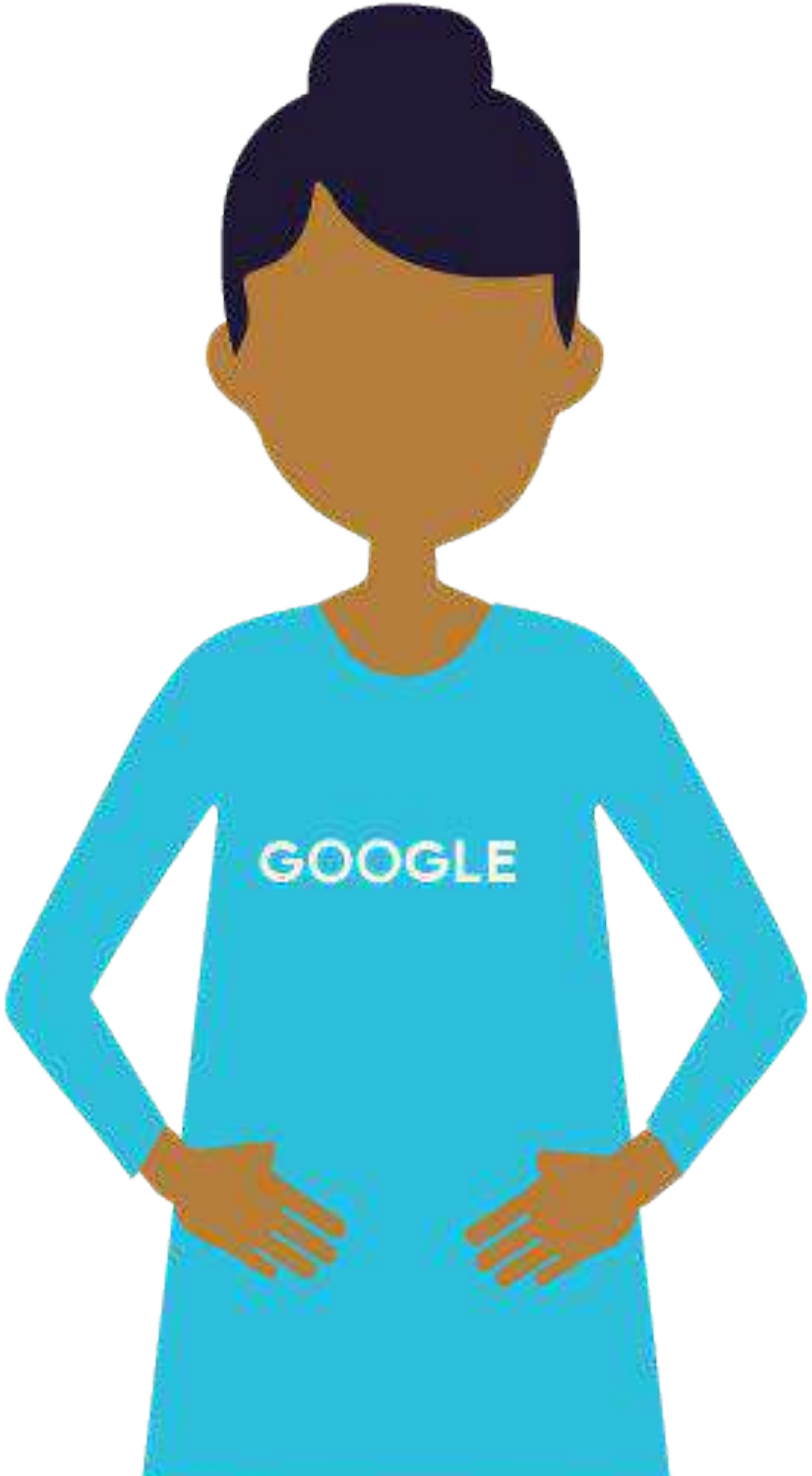


Web Dev Team

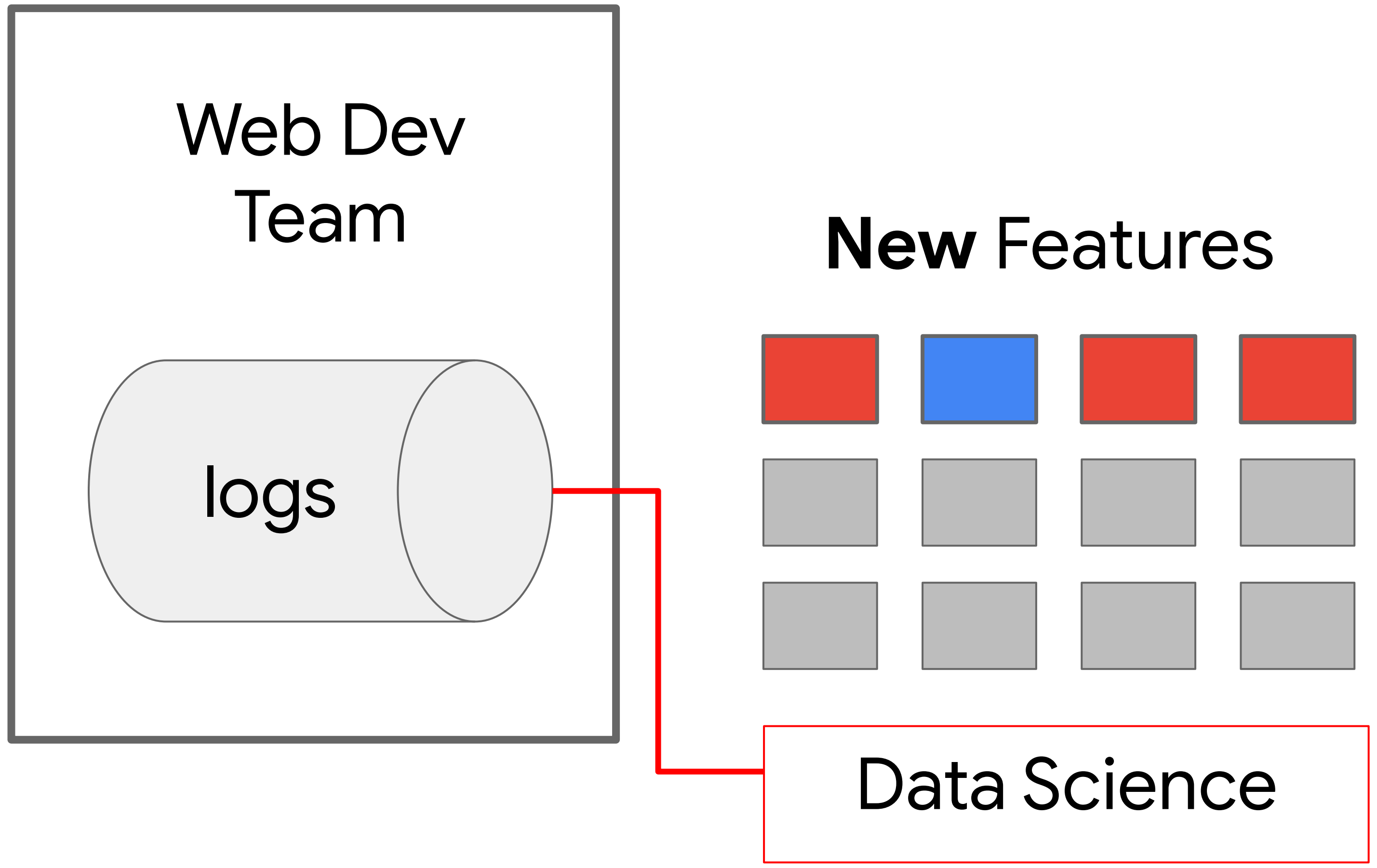
logs

New Features

Data Science



Underutilized data dependencies

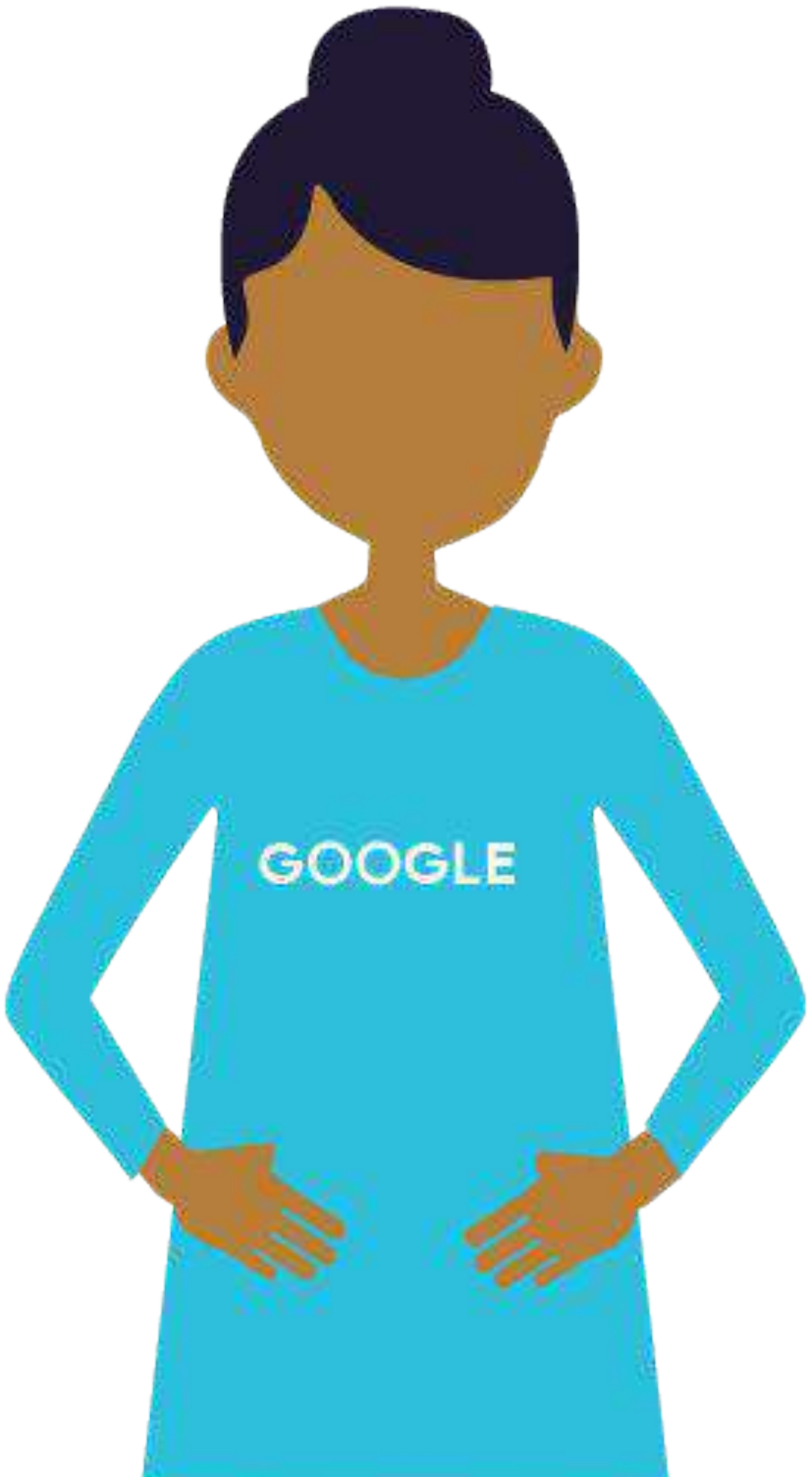


Web Dev Team

logs

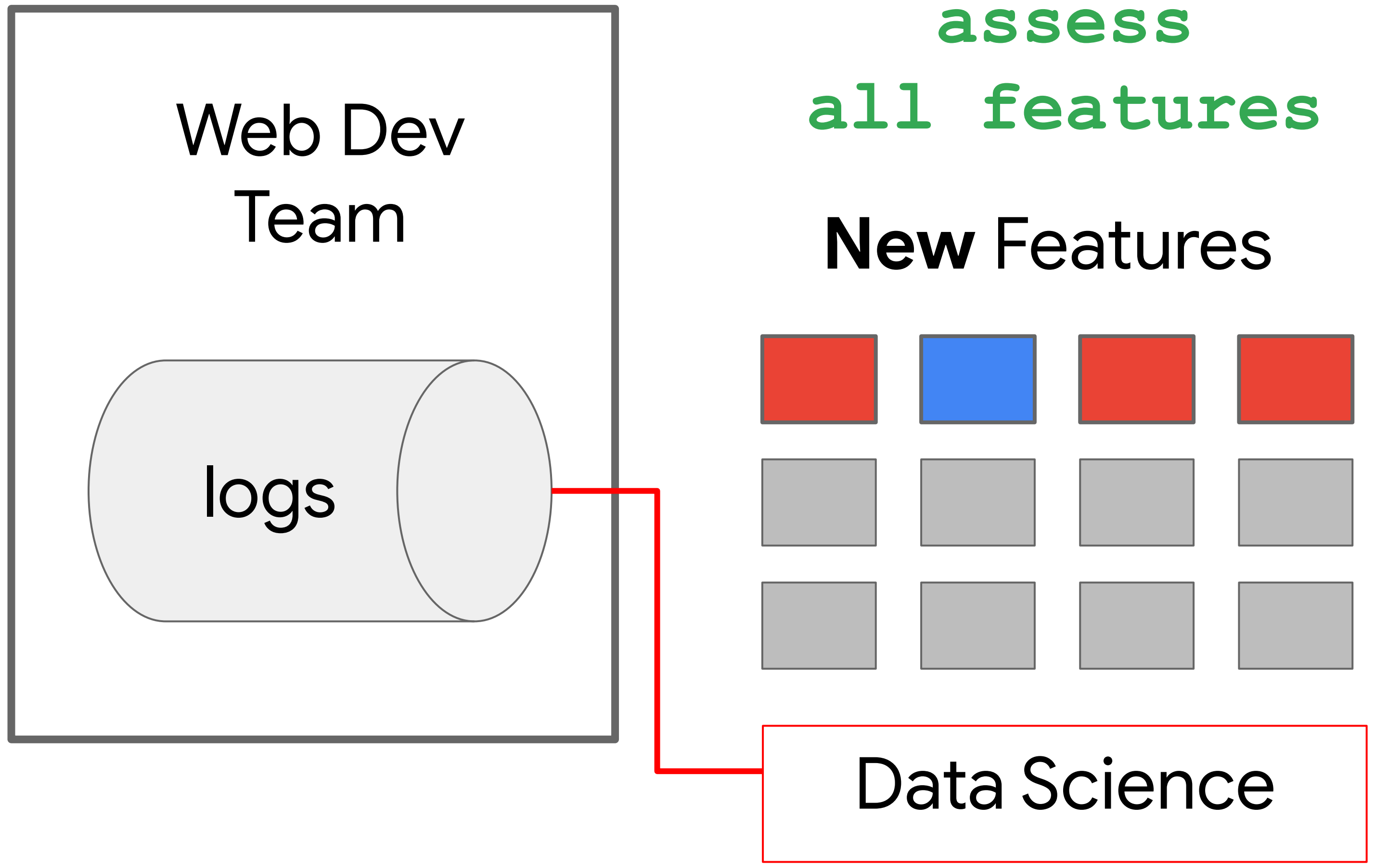
New Features

Data Science



Underutilized data dependencies

Rigorously assess all features



Web Dev Team

logs

New Features

Data Science

Course 2: Production ML Systems

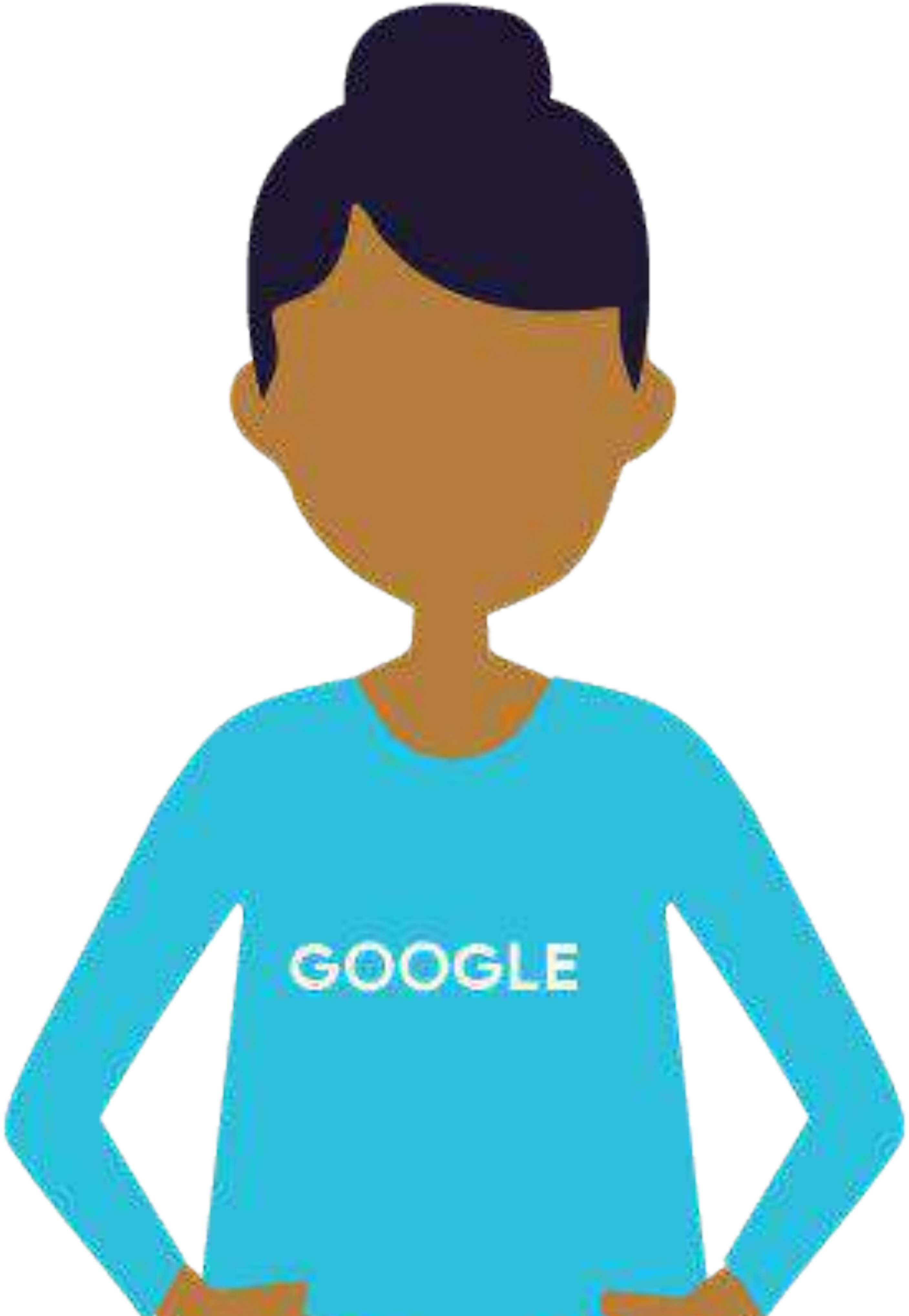
Module 3: Designing Adaptable ML Systems

Lesson Title: **Adapting to Data: Changing Distributions**

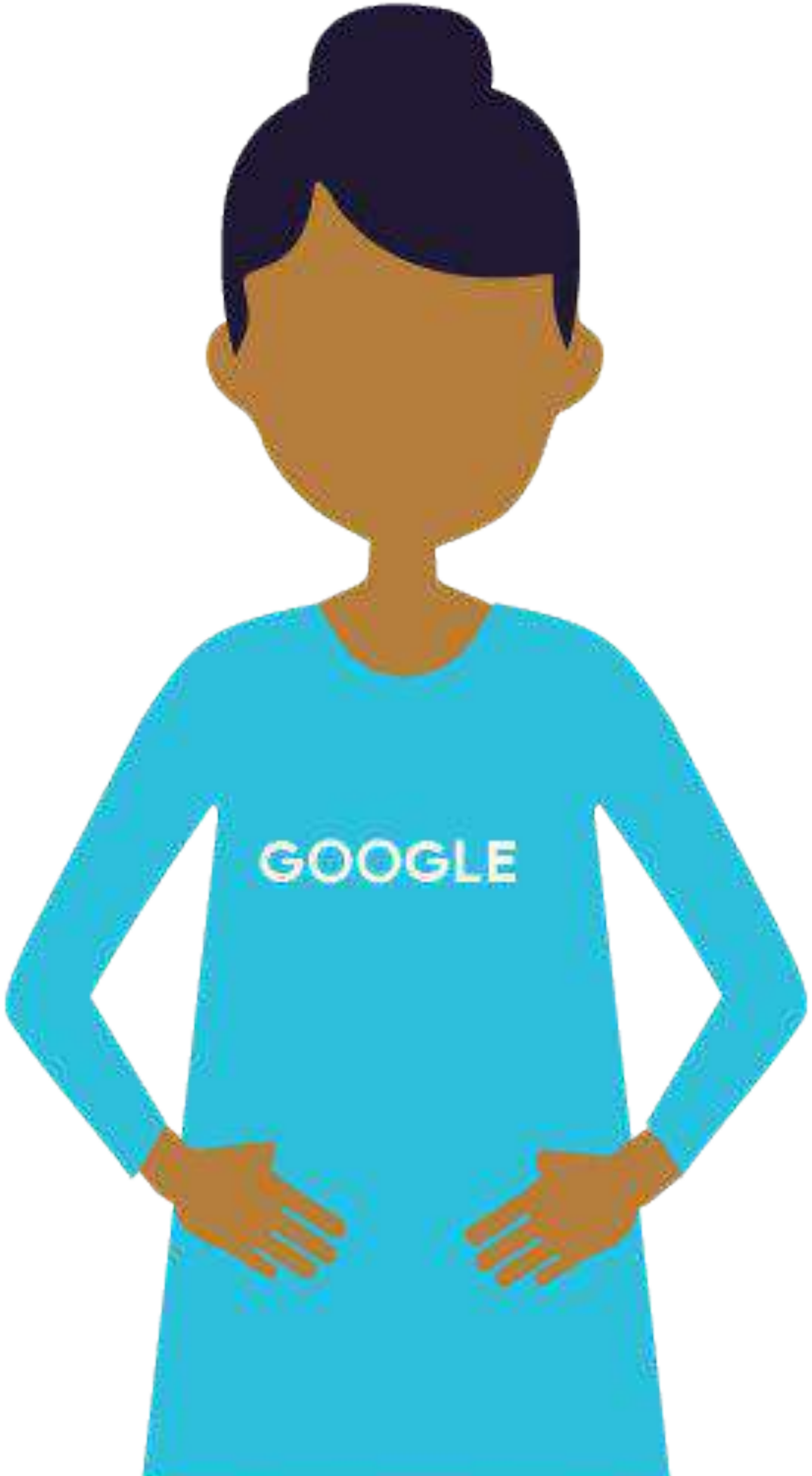
Presenter: Max Lotstein

Format: Talking Head

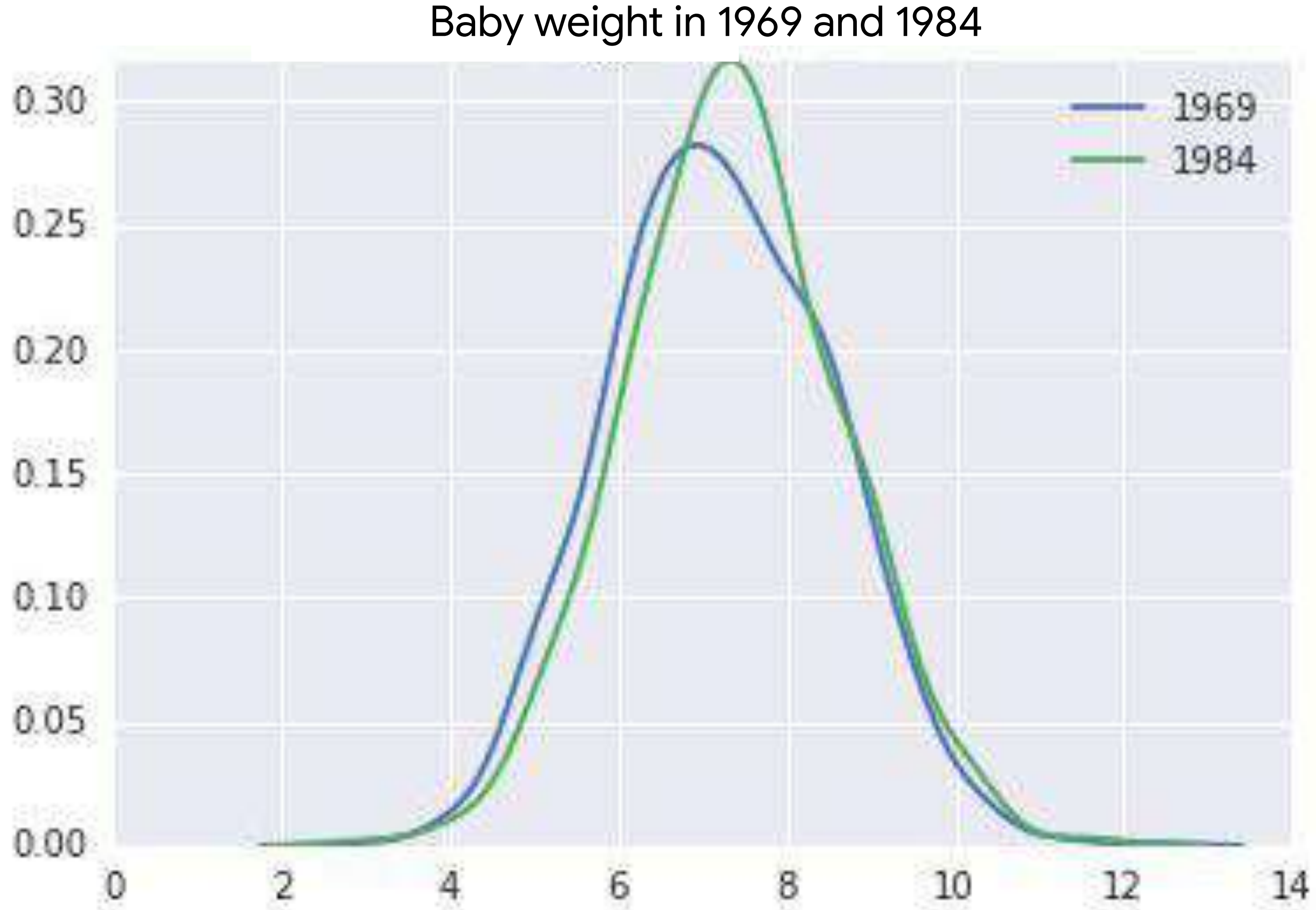
Video Name:
T-PSML-O_3_l4_adapting_to_data:_changing_distributions

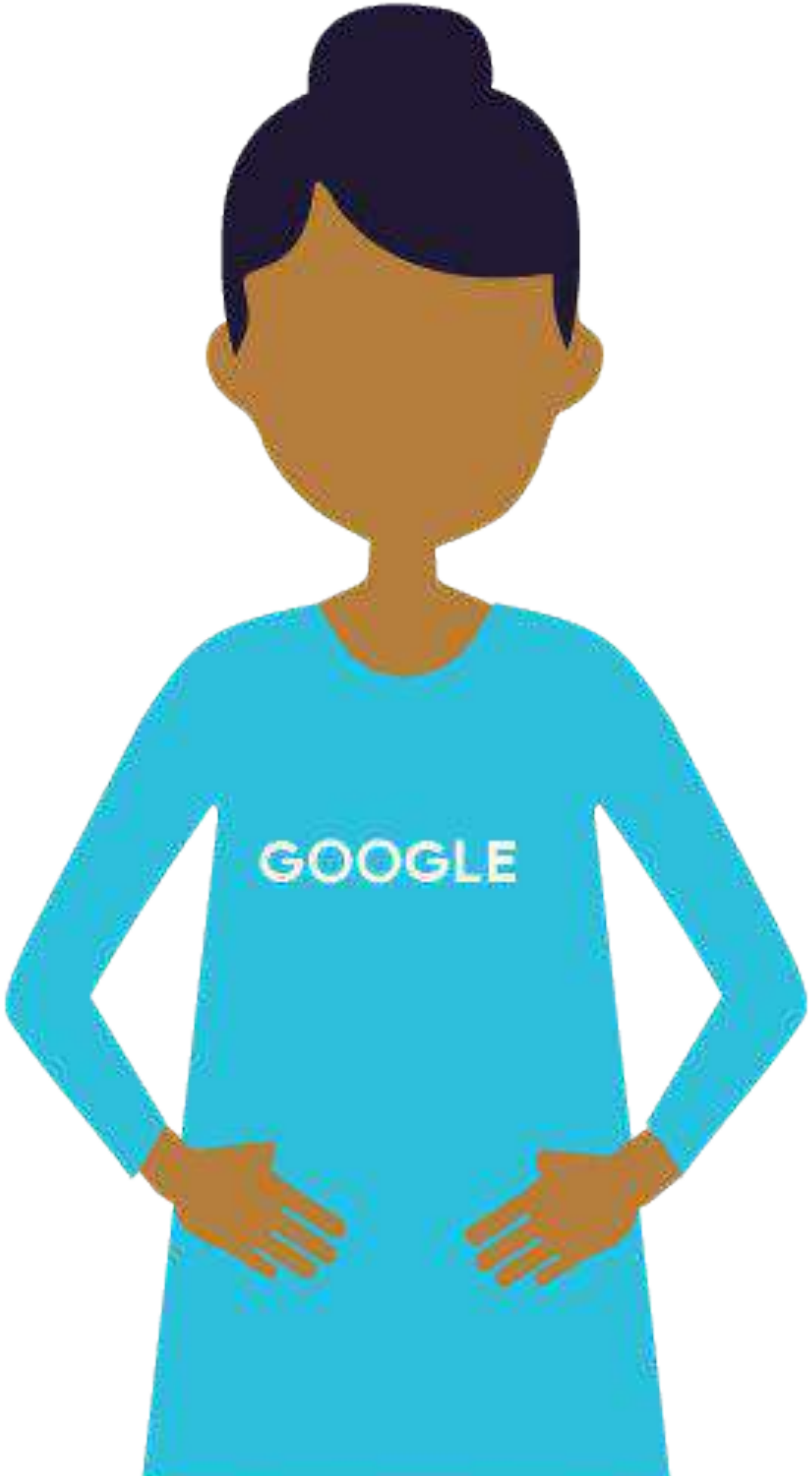


GOOGLE

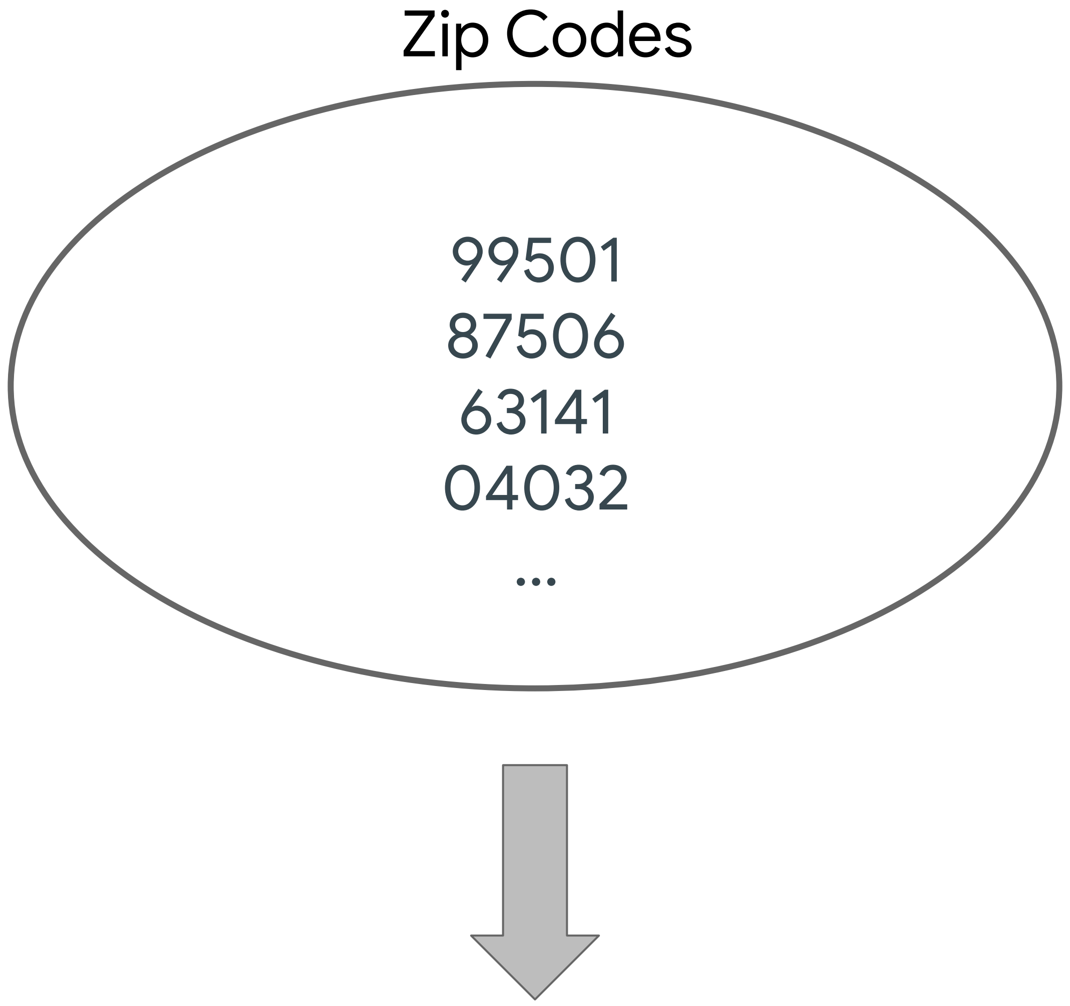


Distributions change





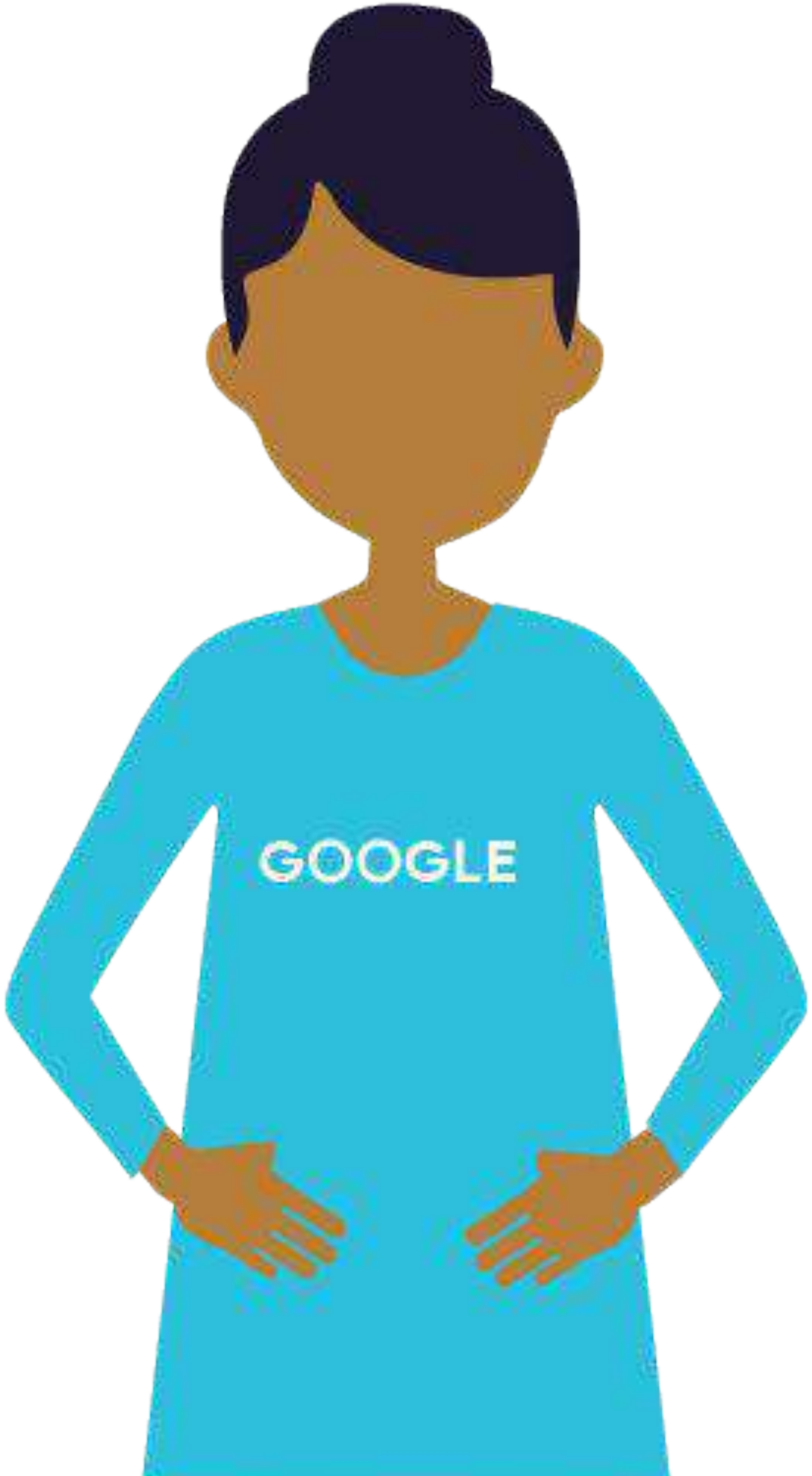
Distributions change



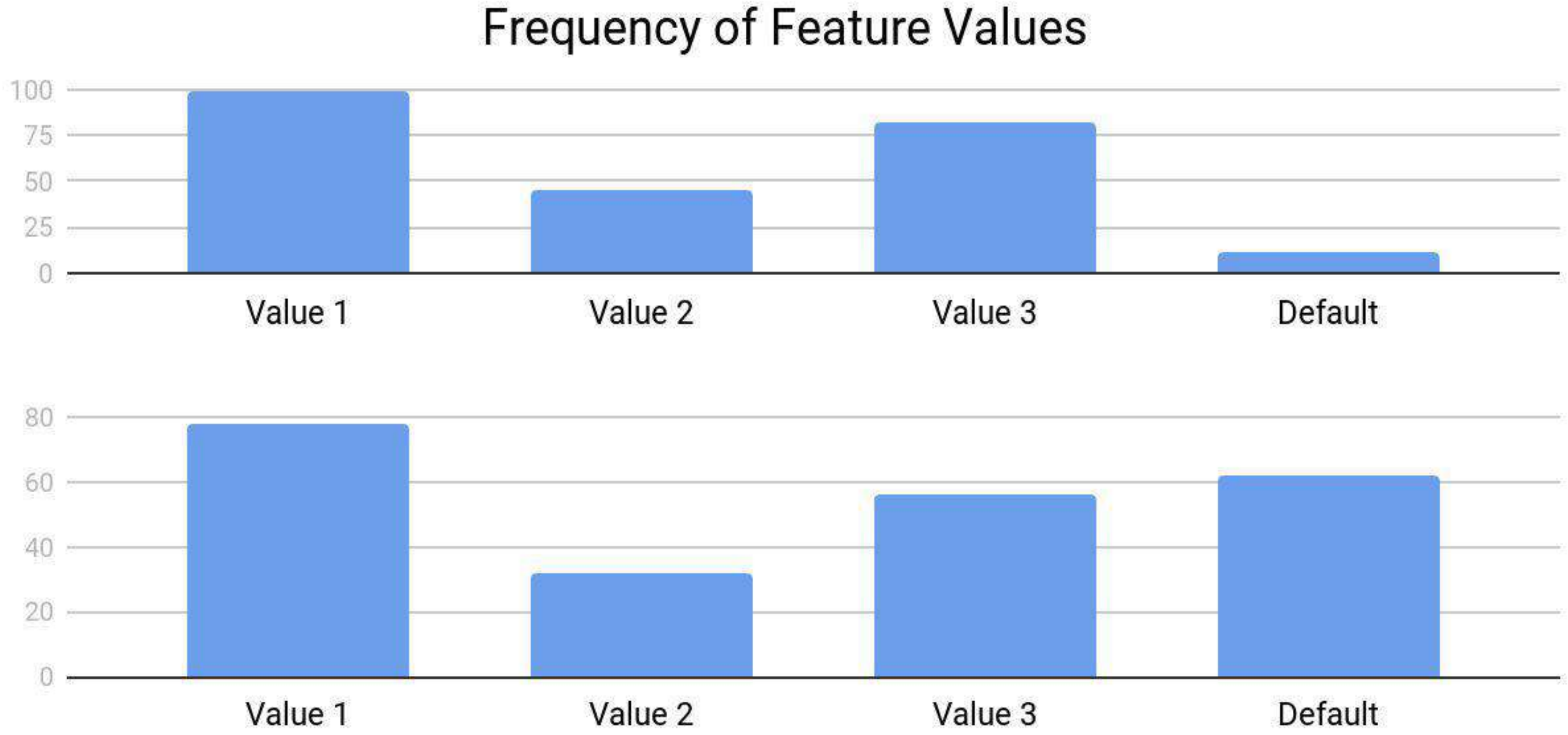
Zip Codes

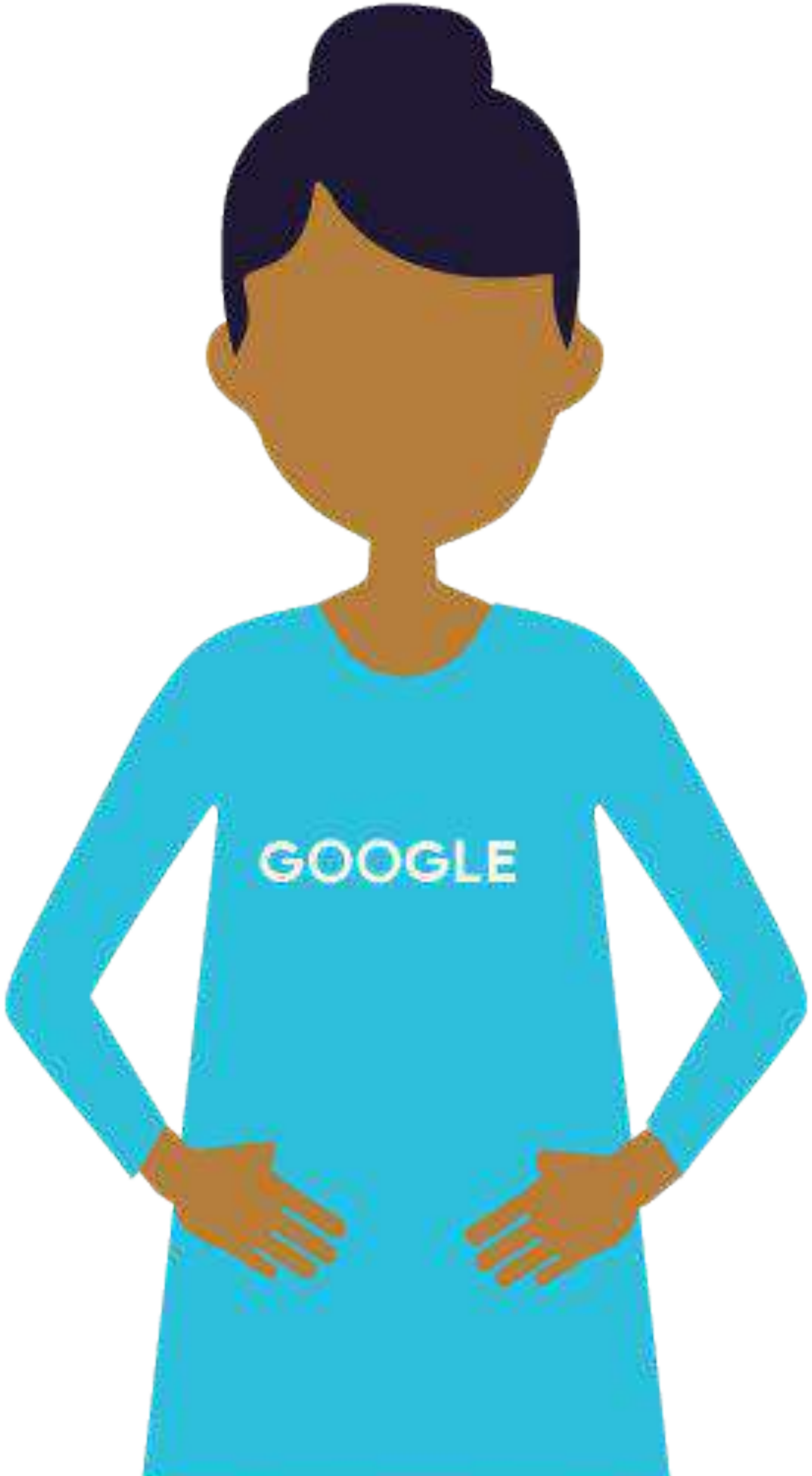
99501
87506
63141
04032
...

$$\{element1, element2, element3, ...\}$$

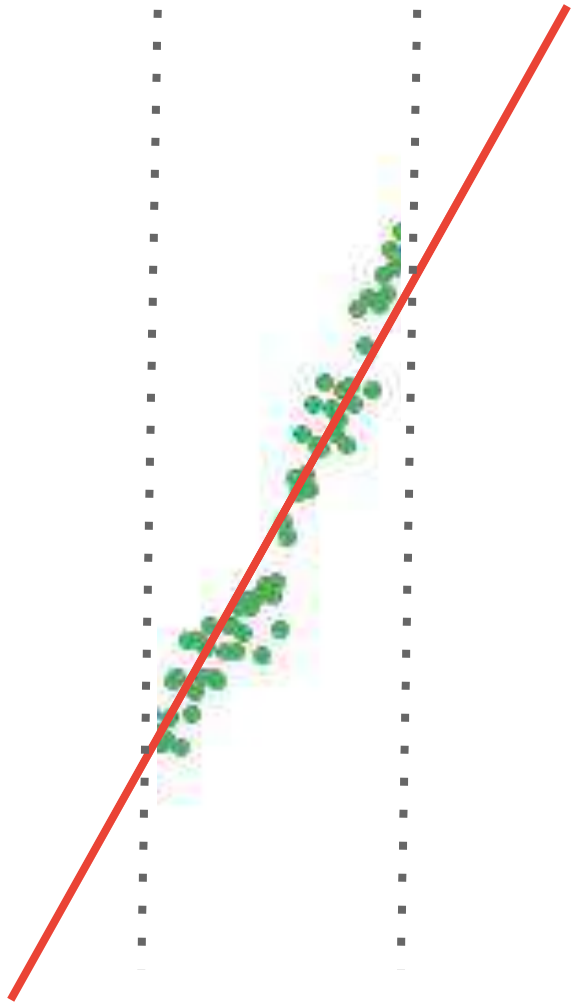


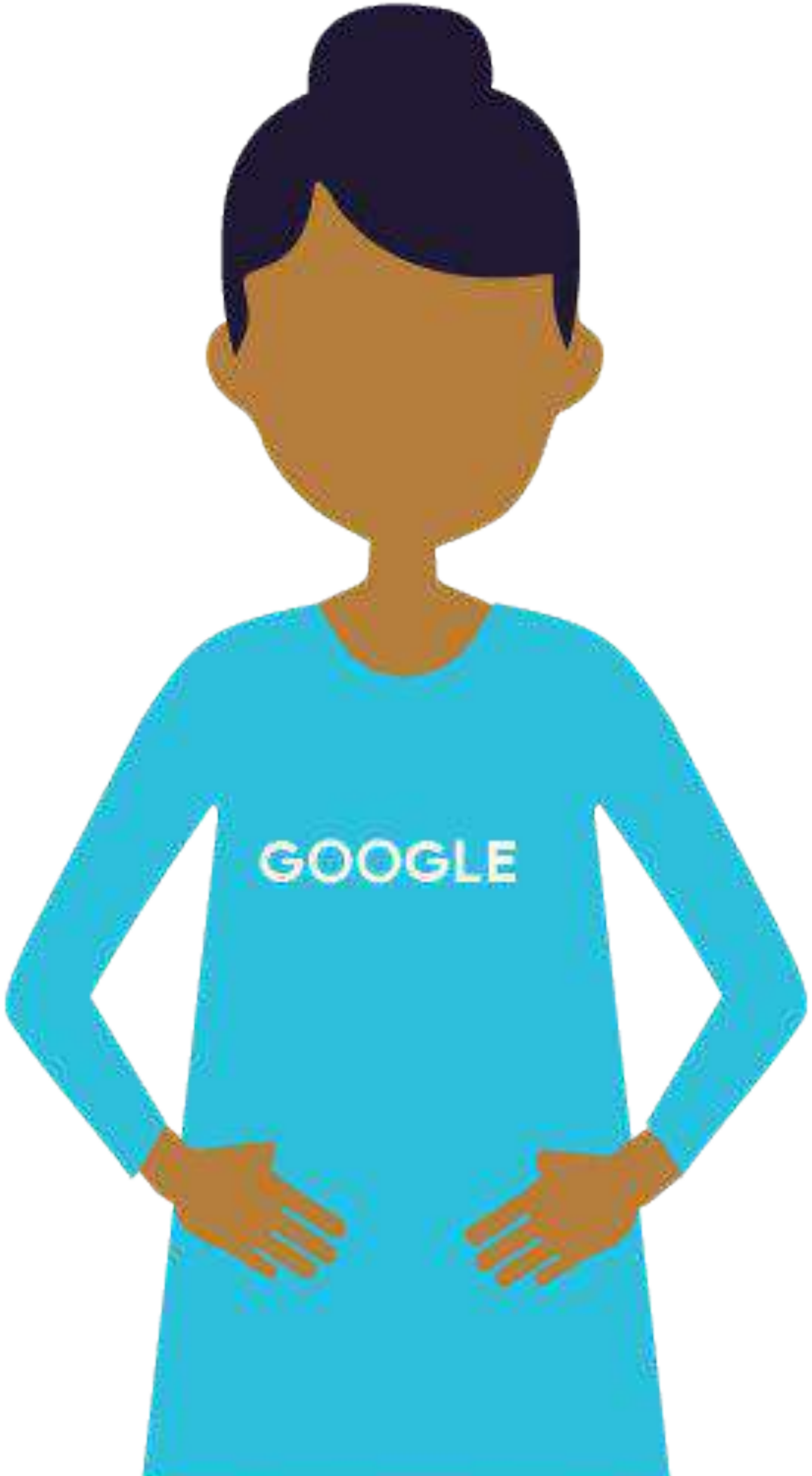
Distributions change



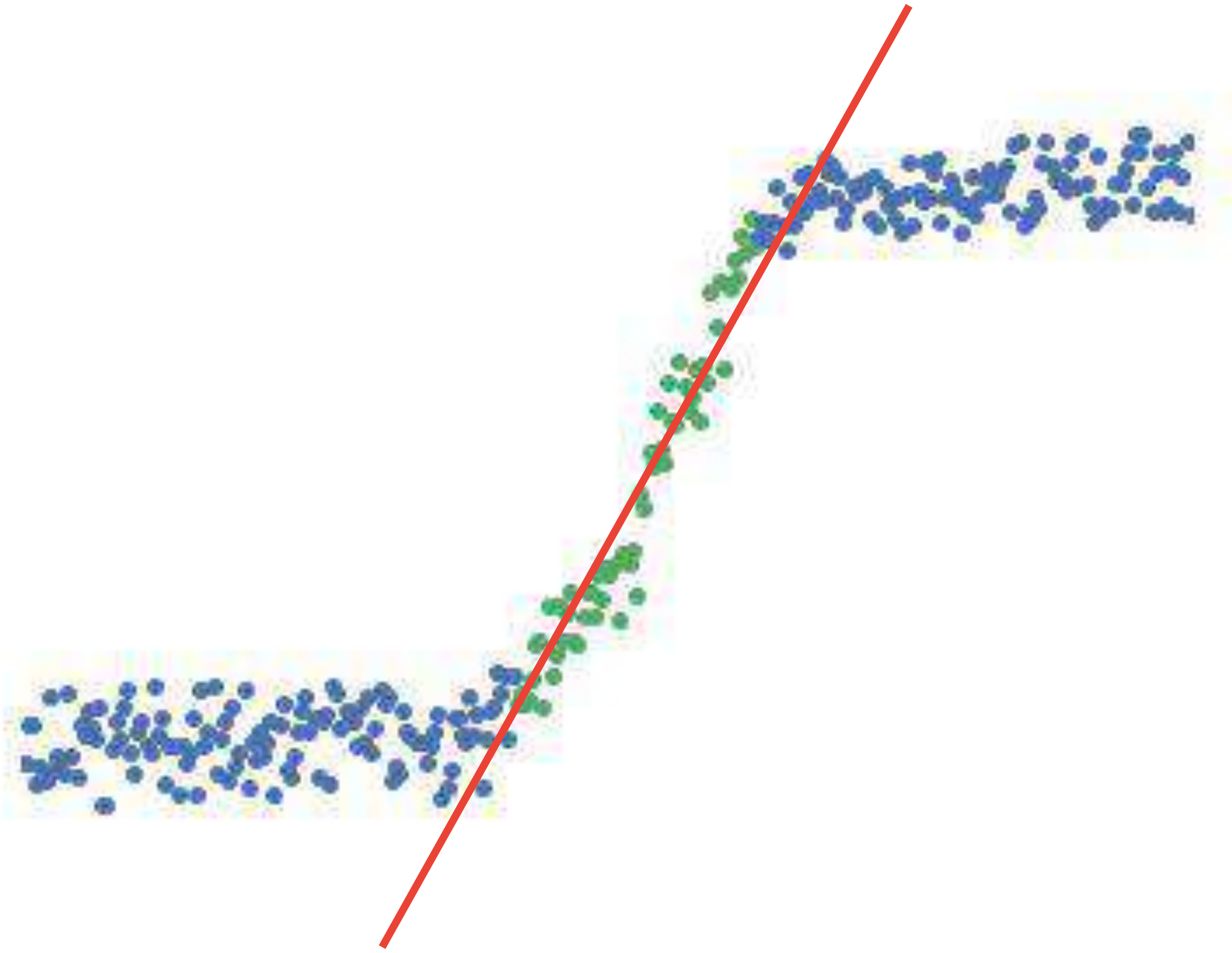


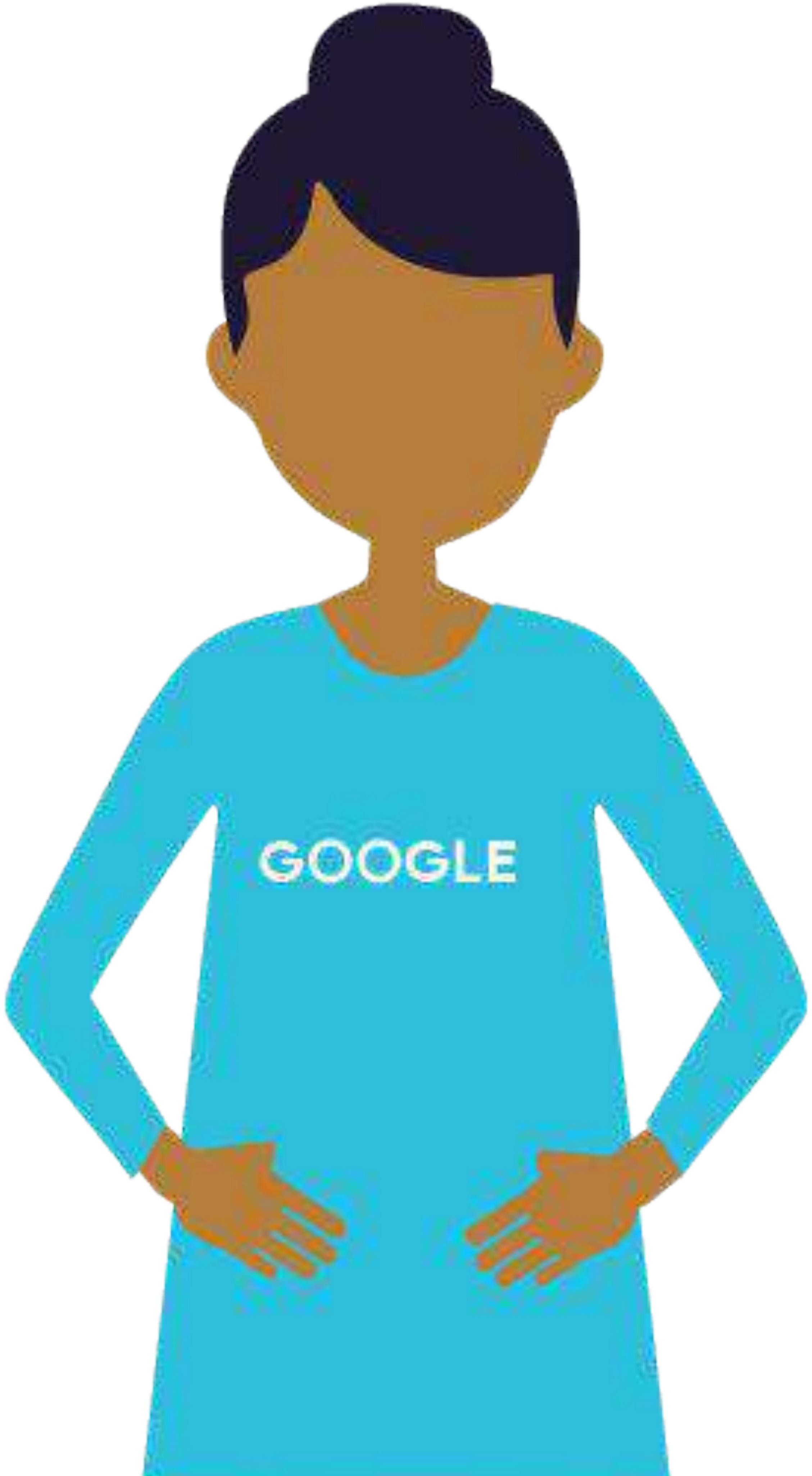
Distributions change





Distributions change





Distributions change

- Monitor descriptive statistics for your inputs and outputs
- Monitor your residuals as a function of your inputs
- Use custom weights in your loss function to emphasize data recency
- Use dynamic training architecture and regularly retrain your model

Course 2: Production ML Systems

Module 3: Designing Adaptable ML Systems

Lesson Title: **Adapting to Data: Lab**

Presenter: Max Lotstein

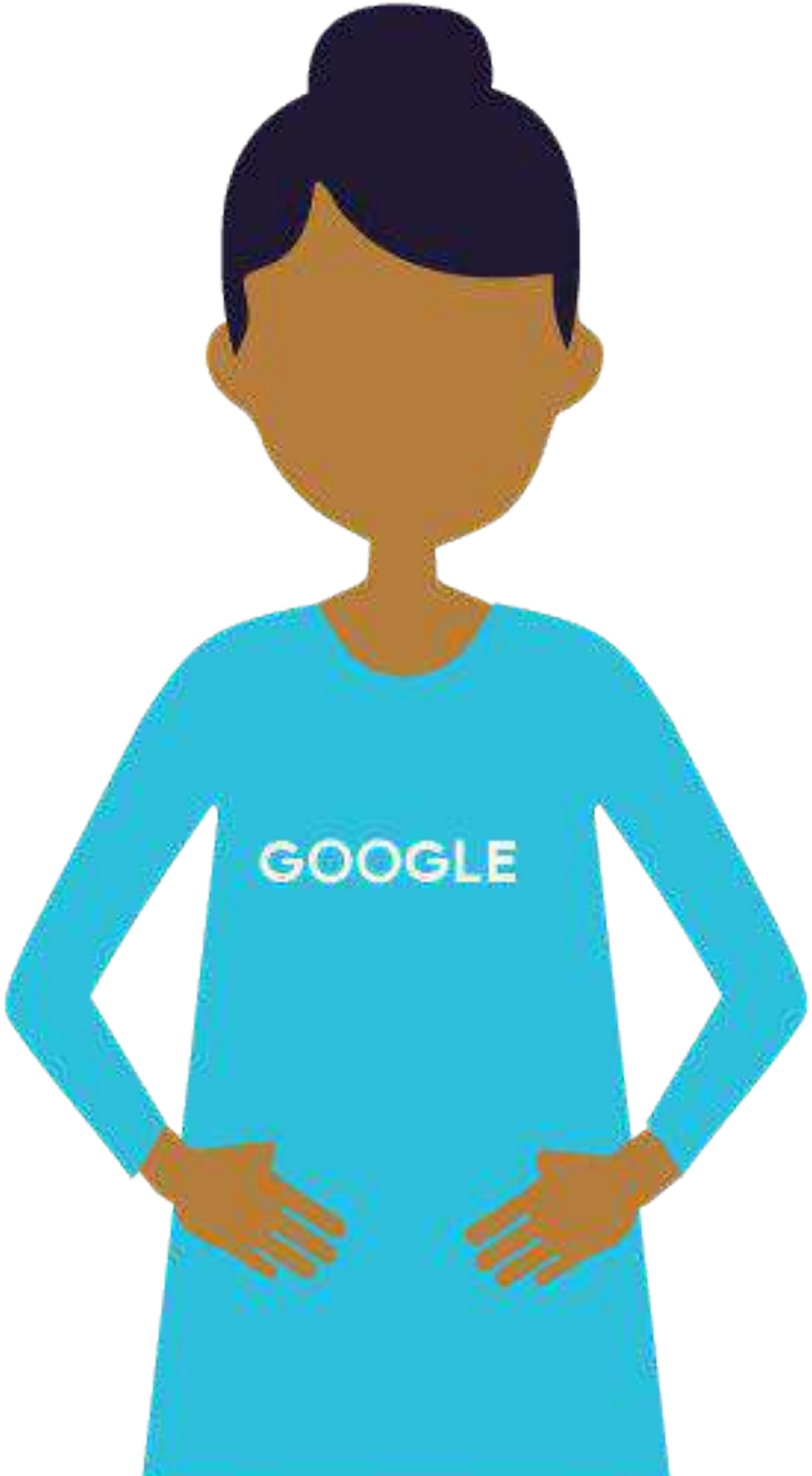
Format: Talking Head

Video Name: T-PSML-O_3_l5_adapting_to_data:_lab_intro

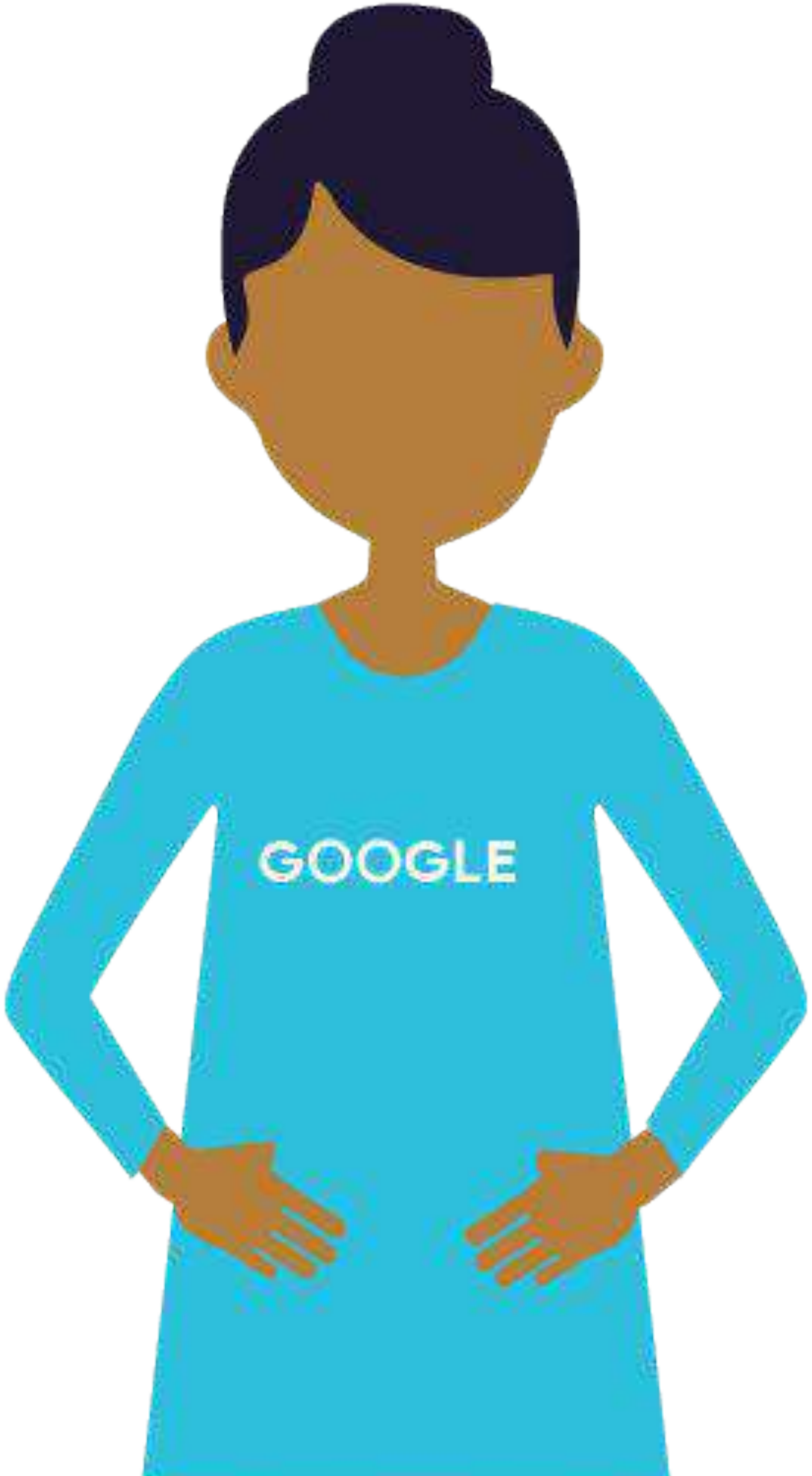
Lab

Making Good ML Engineering Investments

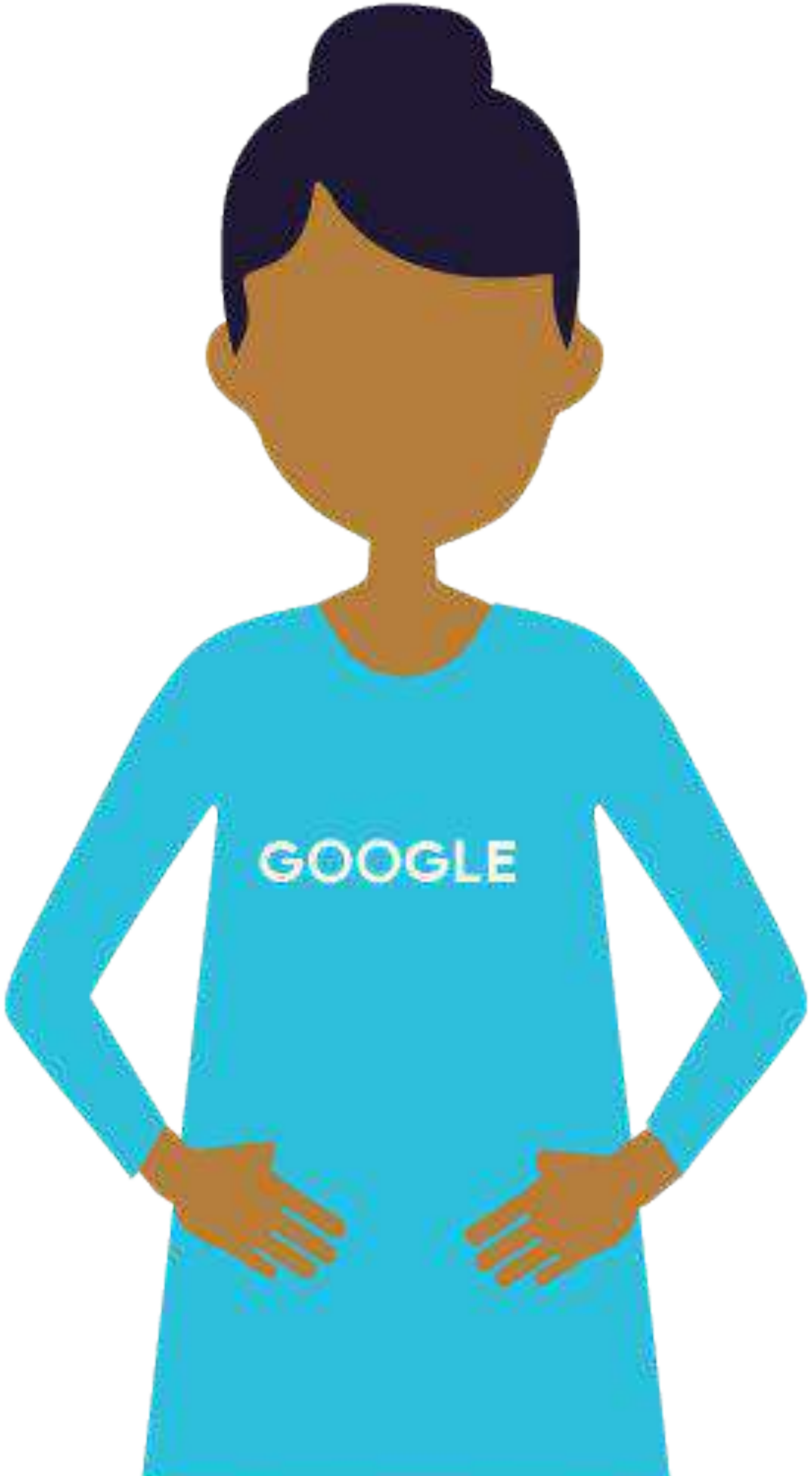
Max Lotstein



Scenario 1: Code Sprint



Scenario 2: A Gift Horse



Course 2: Production ML Systems

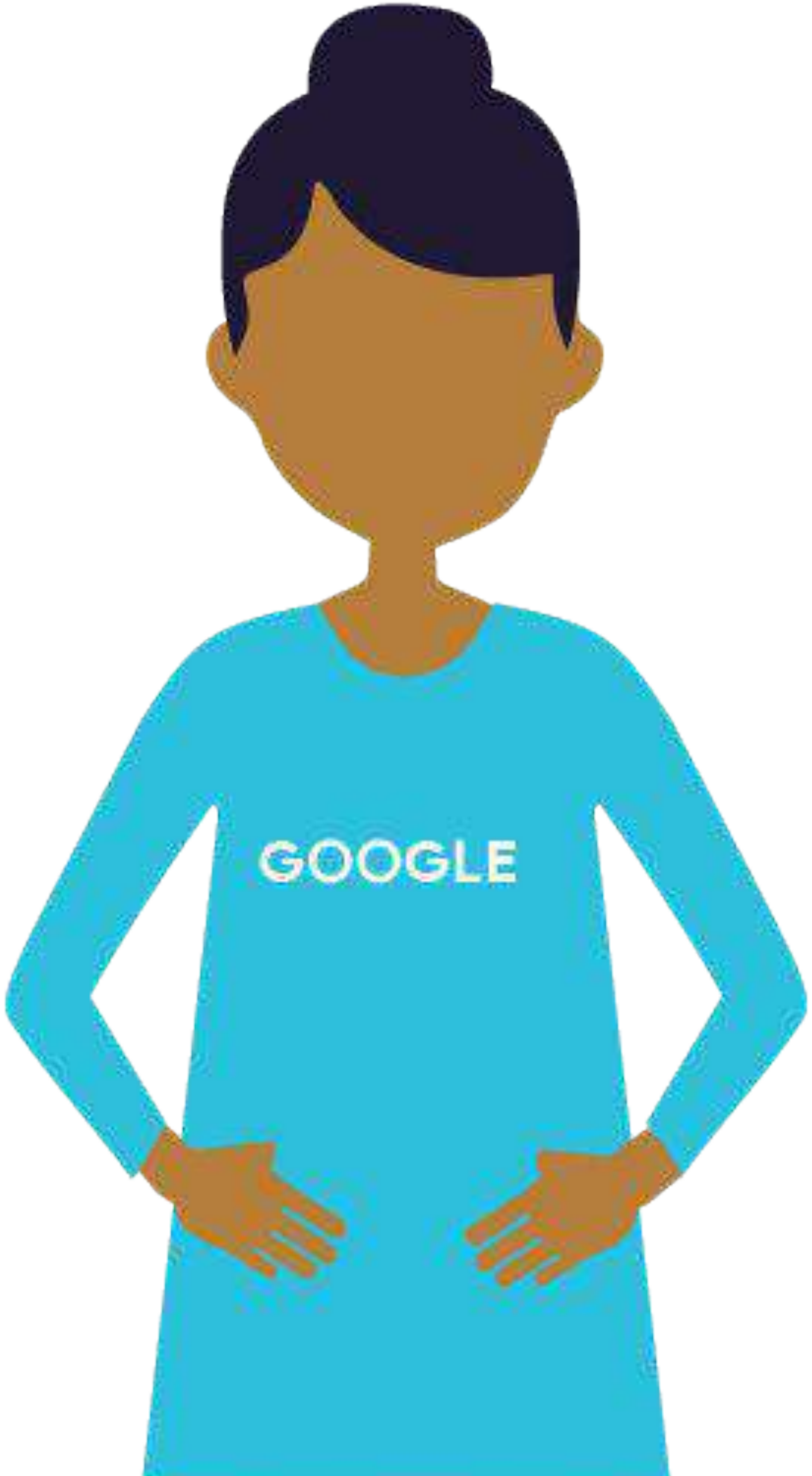
Module 3: Designing Adaptable ML Systems

Lesson Title: **Adapting to Data: Right and Wrong Decisions**

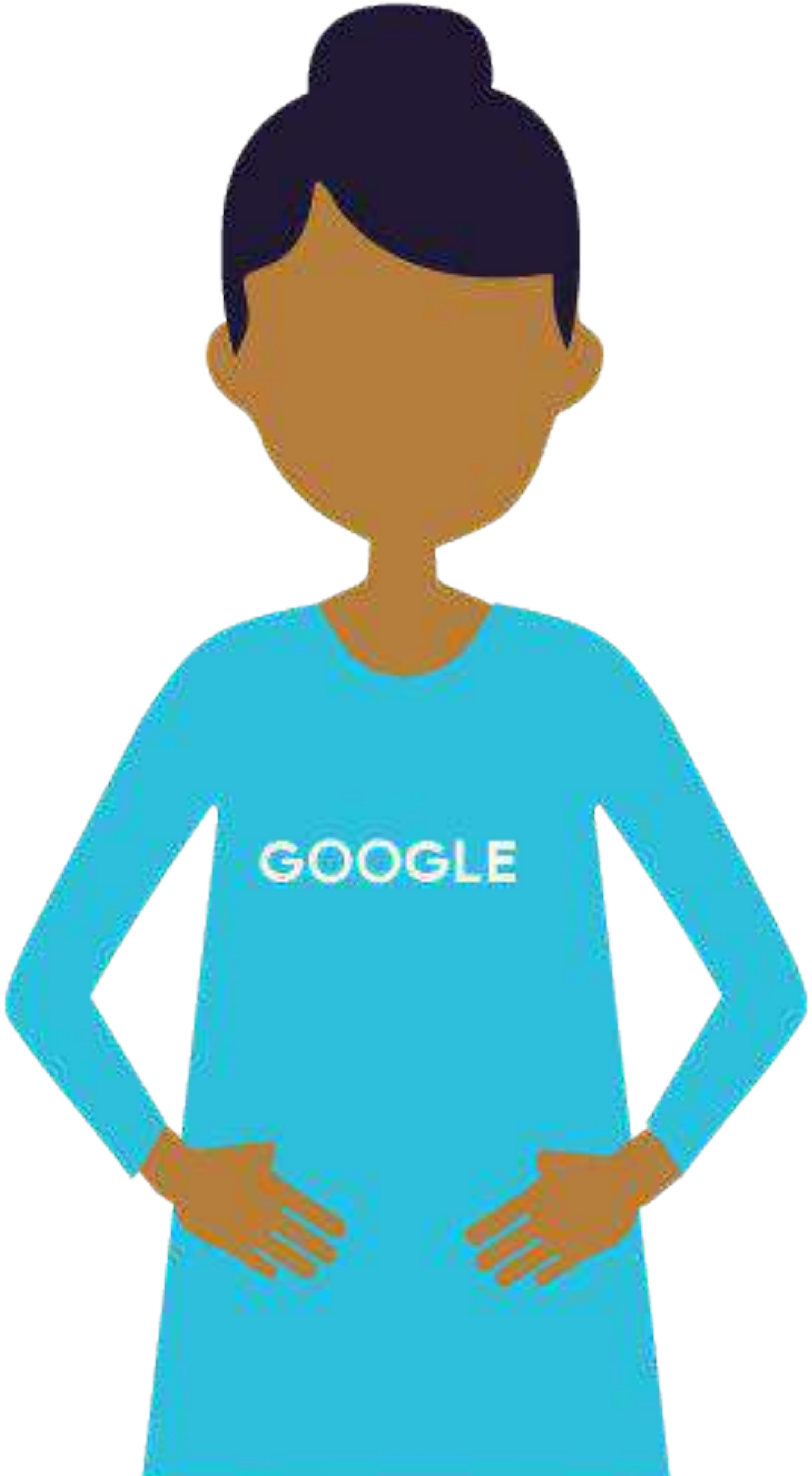
Presenter: Max Lotstein

Format: Talking Head

Video Name:
T-PSML-O_3_l7_adapting_to_data:_right_and_wrong_decisions



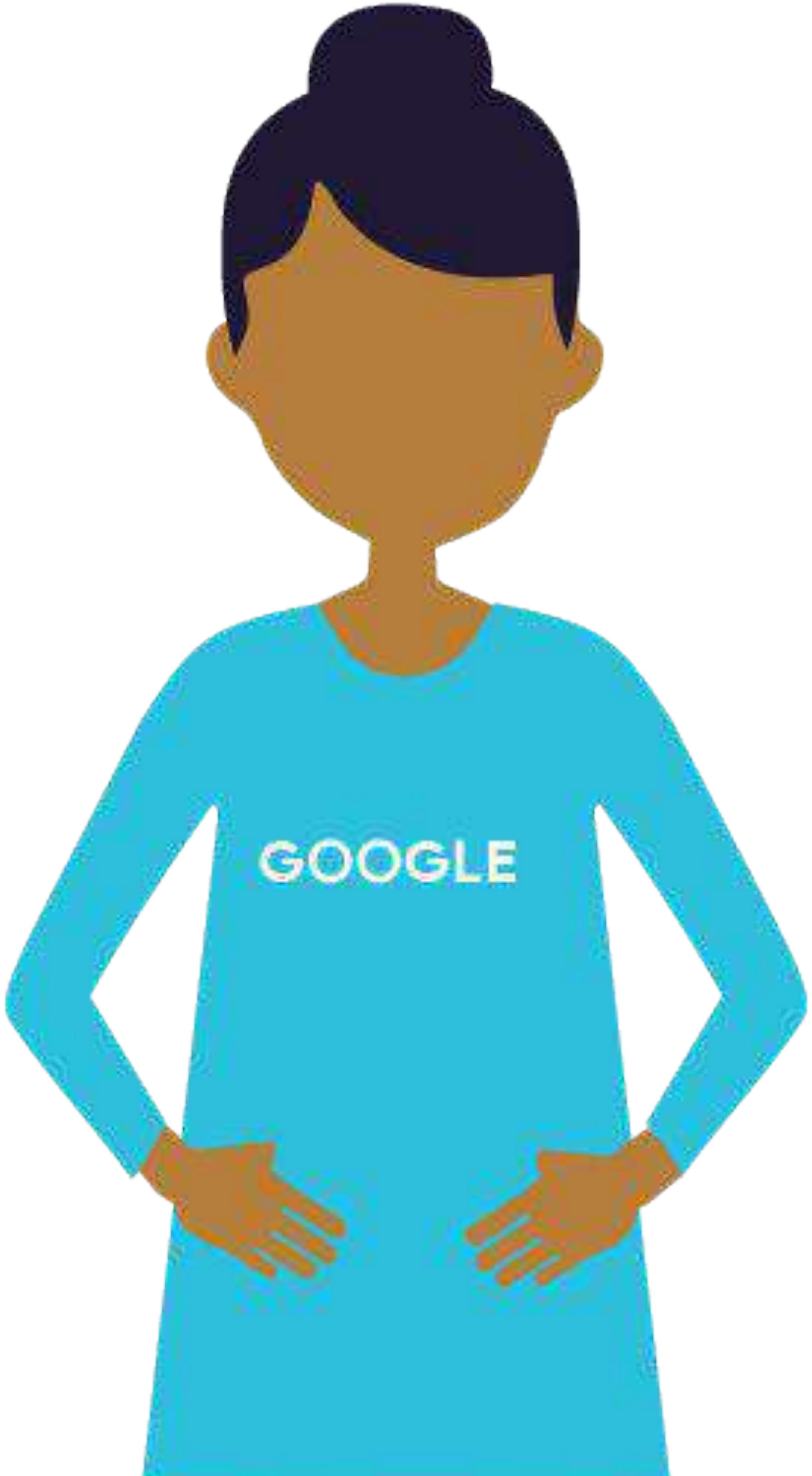
Right and Wrong Data Decisions



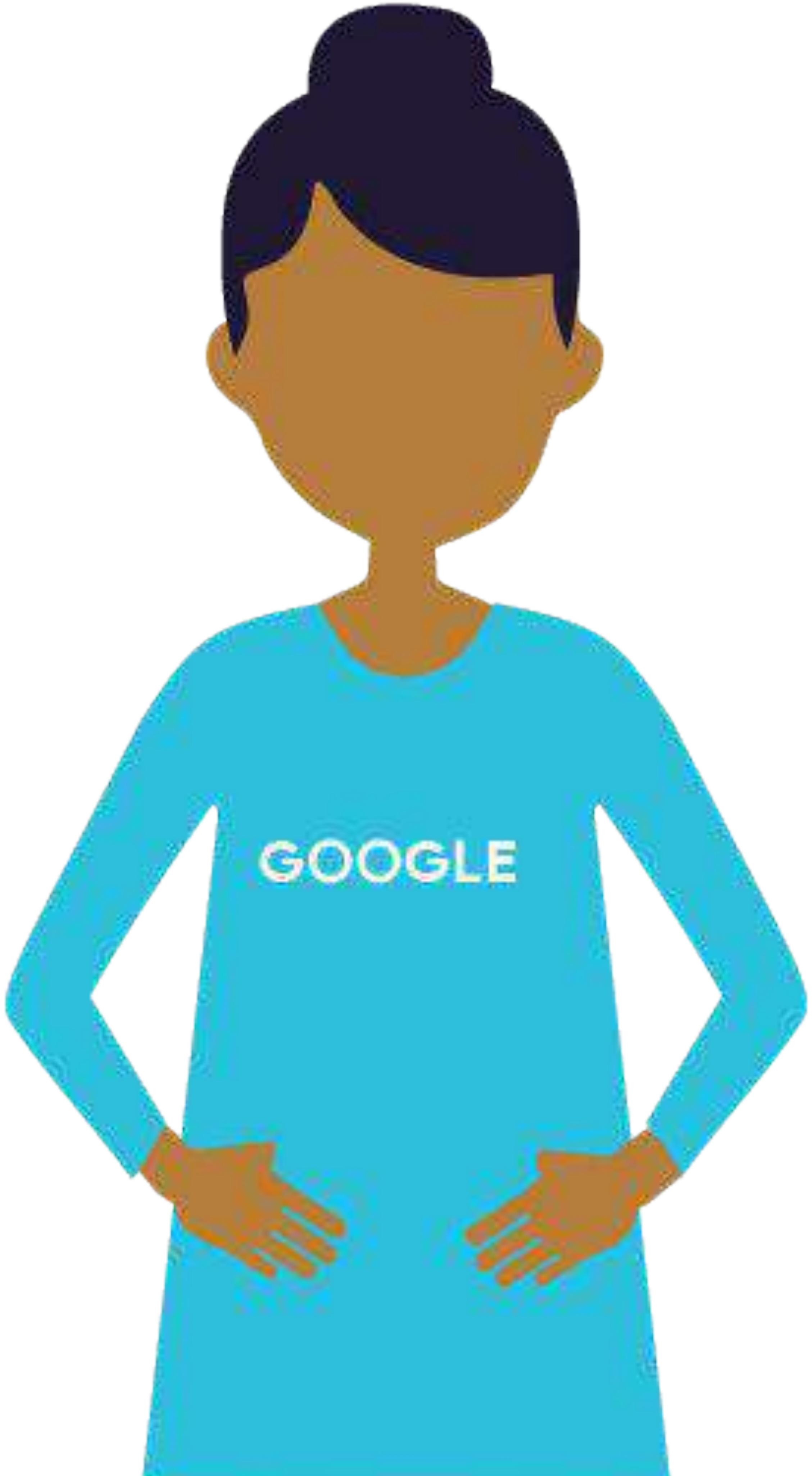
Right and Wrong Data Decisions

- patient age
- gender
- prior medical conditions
- hospital name
- vital signs
- test results

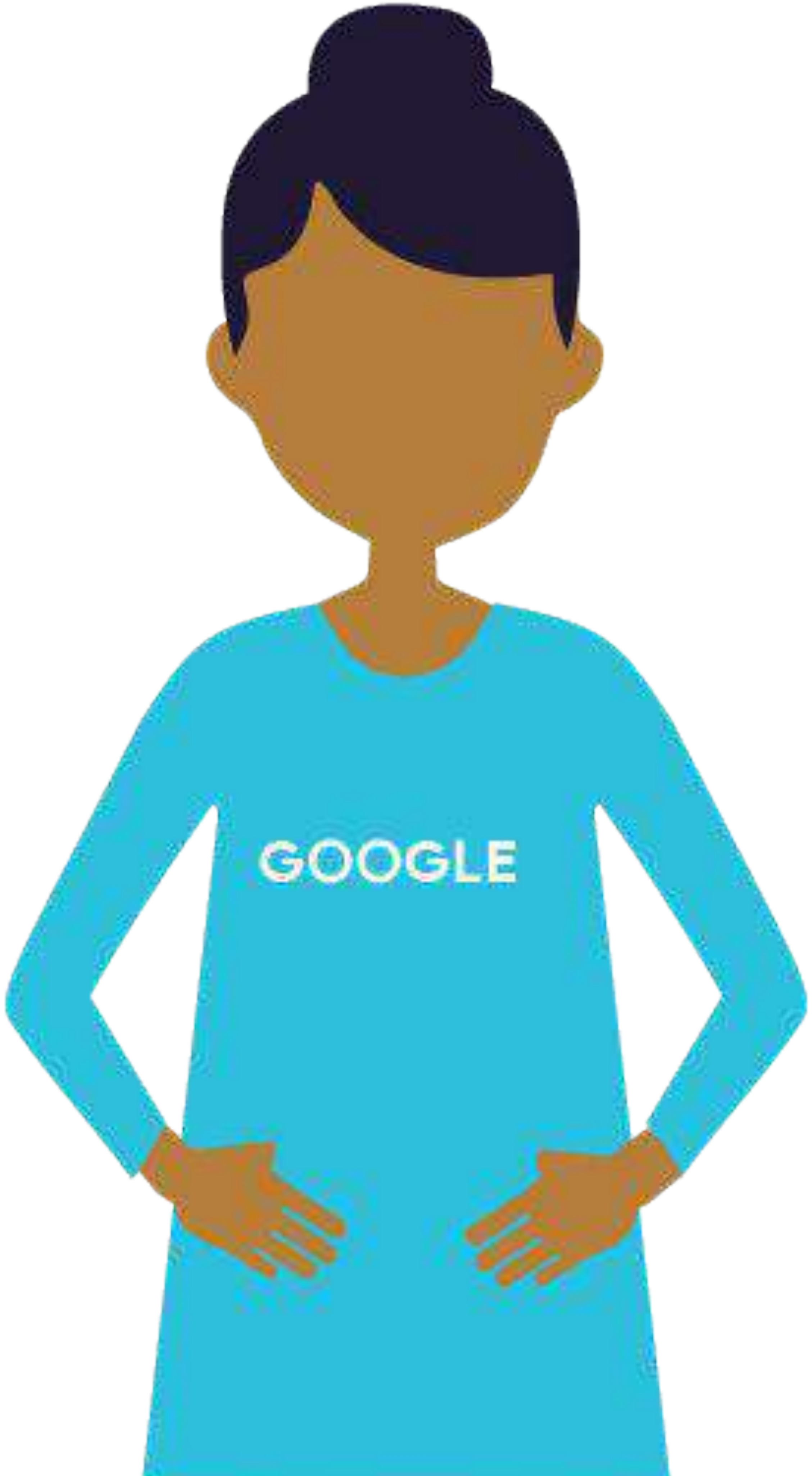
Data Leakage



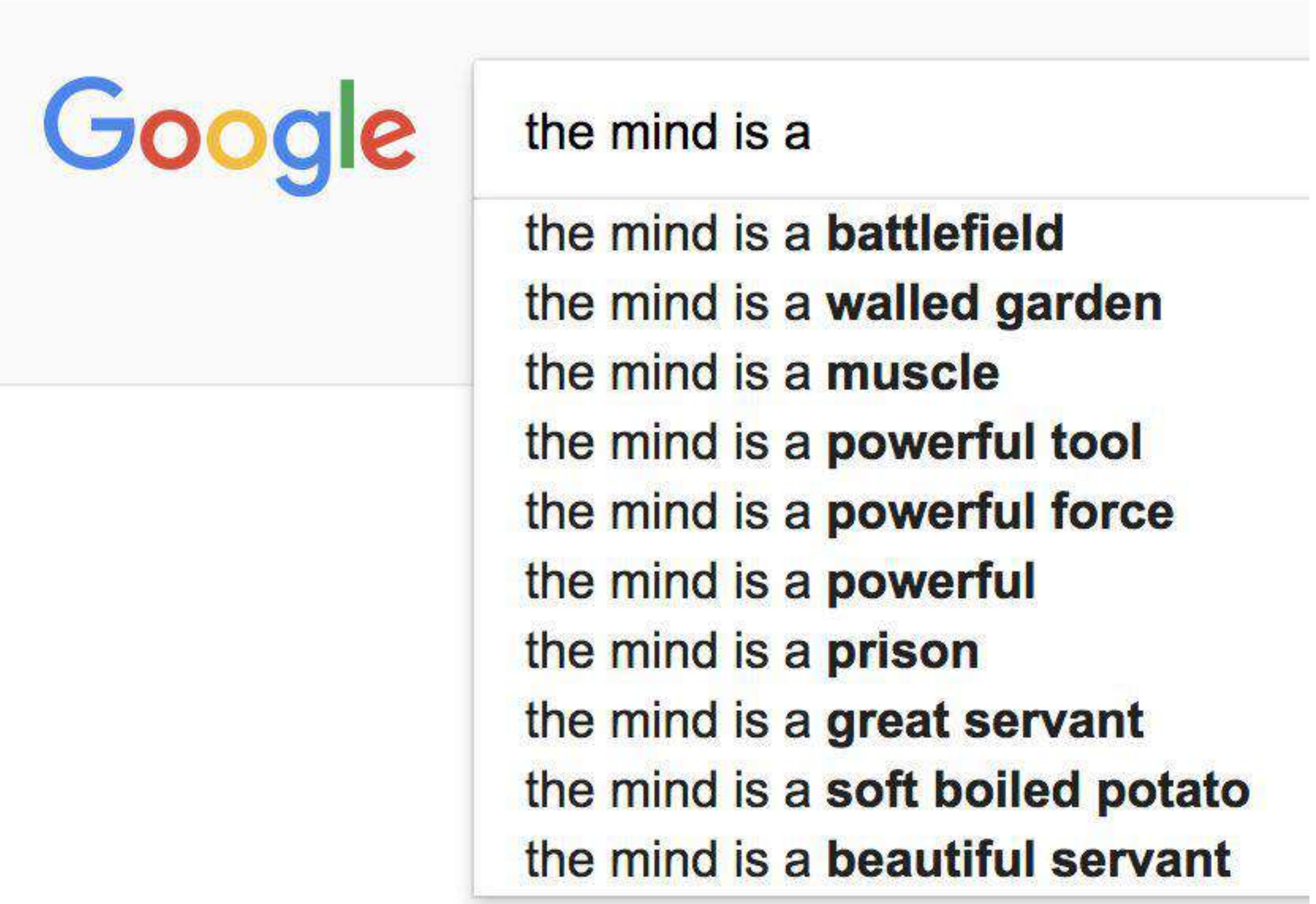
https://upload.wikimedia.org/wikipedia/commons/5/5f/Beth_Israel_Deaconess_Medical_Center_East_Campus.jpg



Predict political affiliation from metaphors

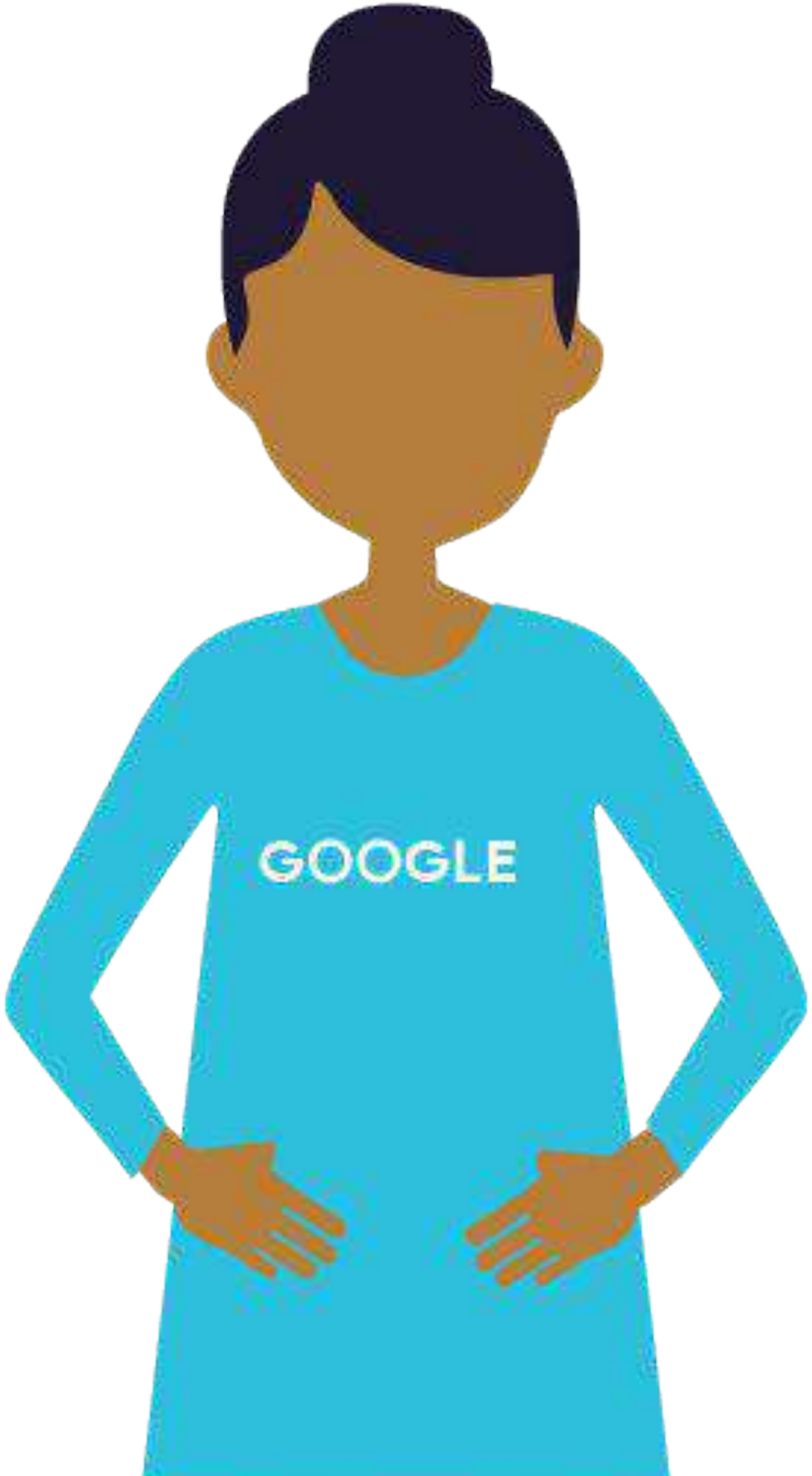


Predict political affiliation from metaphors

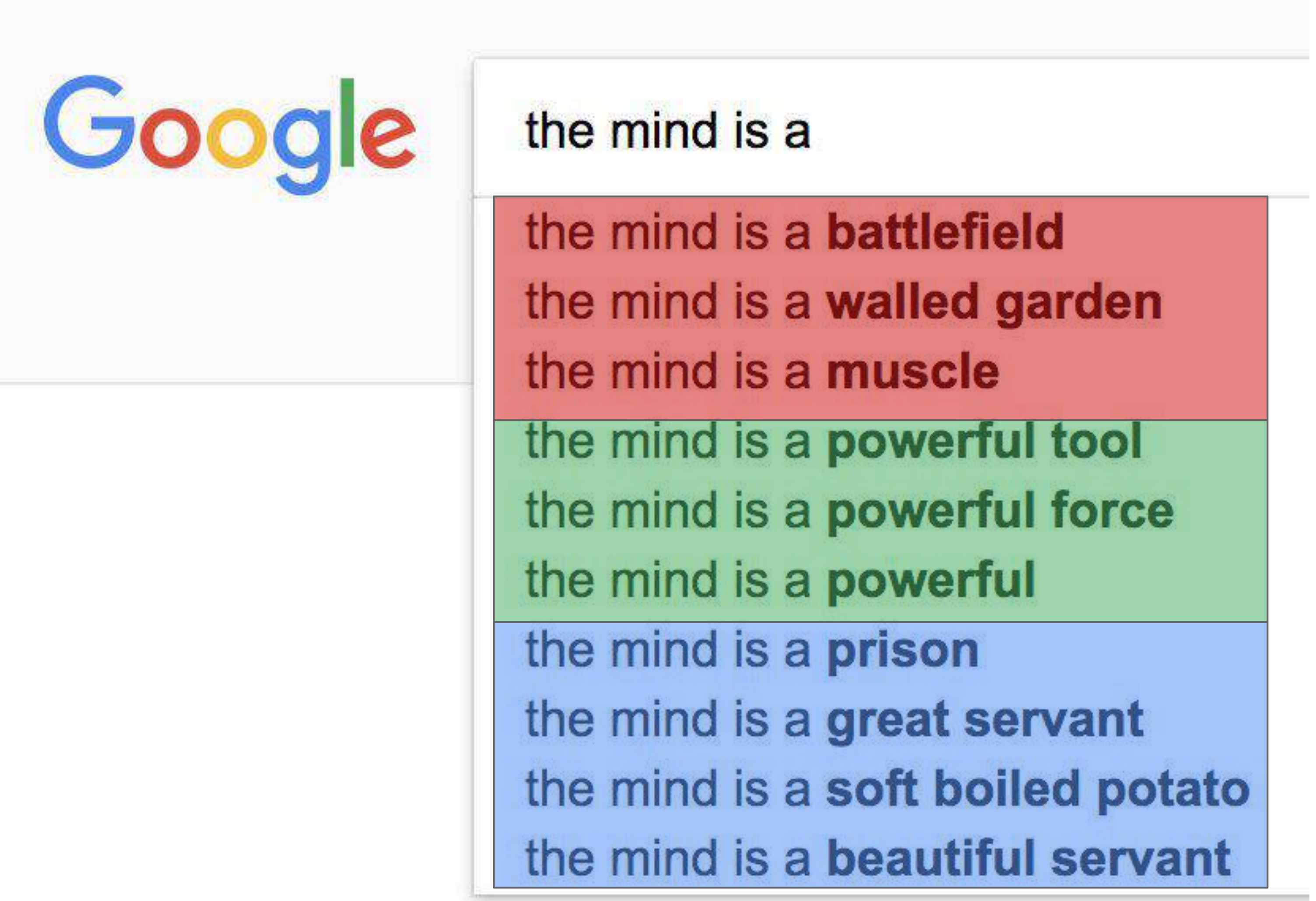


the mind is a

- the mind is a **battlefield**
- the mind is a **walled garden**
- the mind is a **muscle**
- the mind is a **powerful tool**
- the mind is a **powerful force**
- the mind is a **powerful**
- the mind is a **prison**
- the mind is a **great servant**
- the mind is a **soft boiled potato**
- the mind is a **beautiful servant**

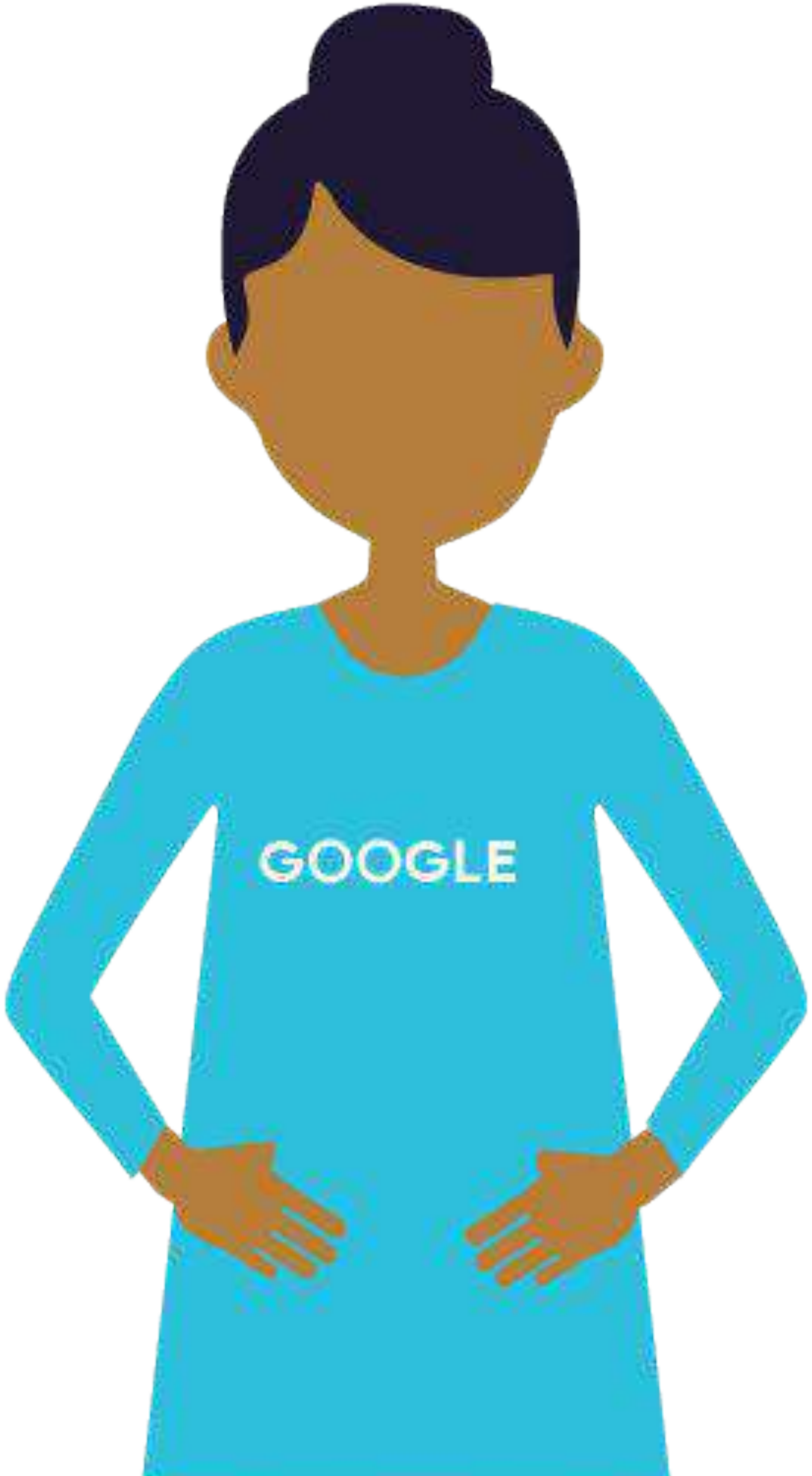


Predict political affiliation from metaphors



the mind is a

the mind is a **battlefield**
the mind is a **walled garden**
the mind is a **muscle**
the mind is a **powerful tool**
the mind is a **powerful force**
the mind is a **powerful**
the mind is a **prison**
the mind is a **great servant**
the mind is a **soft boiled potato**
the mind is a **beautiful servant**



Training Set

Swift Blake Defoe

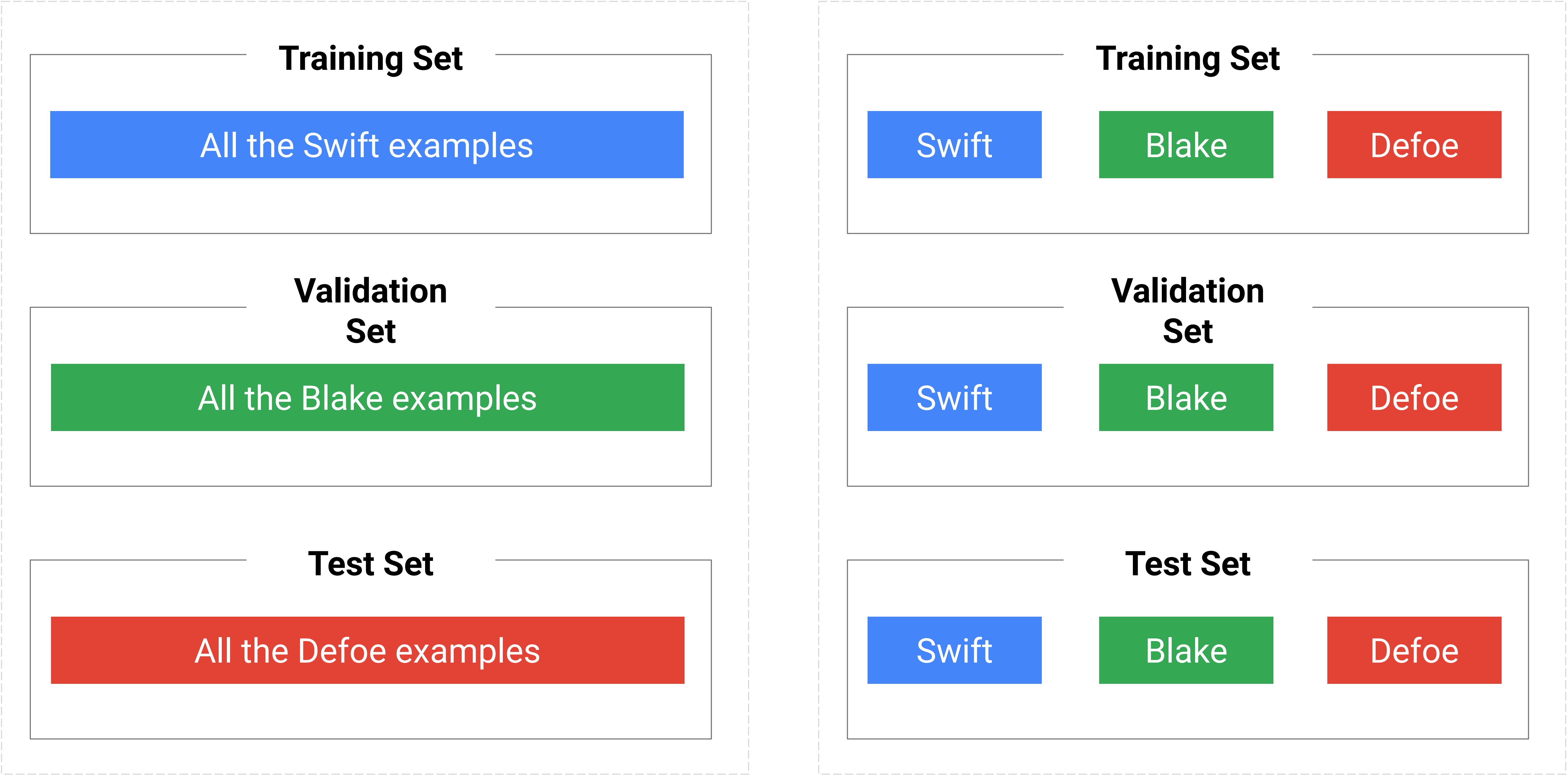
Validation Set

Swift Blake Defoe

Test Set

Swift Blake Defoe

Solution: Cross-contamination; you have to split by author



Training Set

All the Swift examples

Validation Set

All the Blake examples

Test Set

All the Defoe examples

Training Set

Swift Blake Defoe

Validation Set

Swift Blake Defoe

Test Set

Swift Blake Defoe

Course 2: Production ML Systems

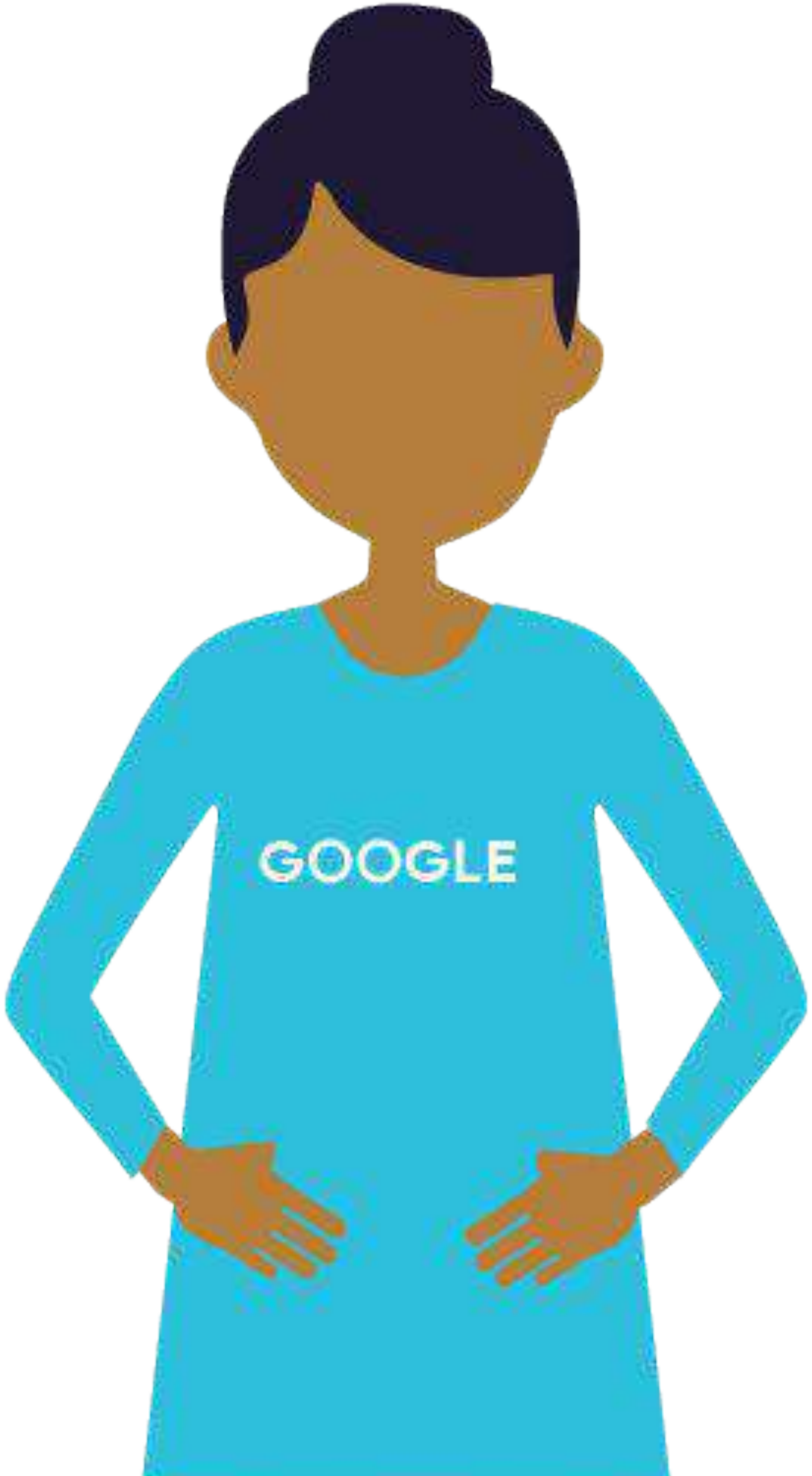
Module 3: Designing Adaptable ML Systems

Lesson Title: **Adapting to Data: System Failure**

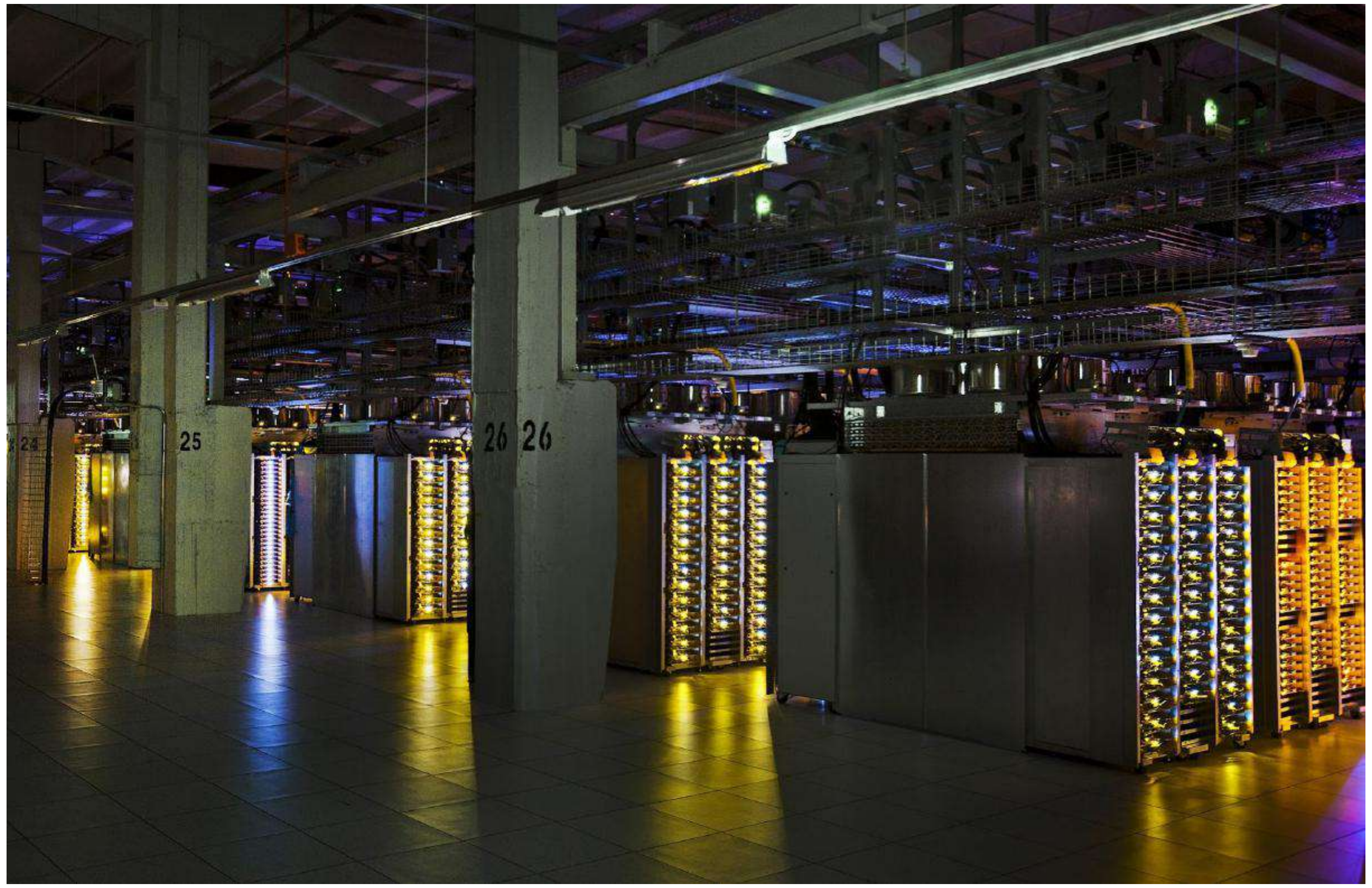
Presenter: Max Lotstein

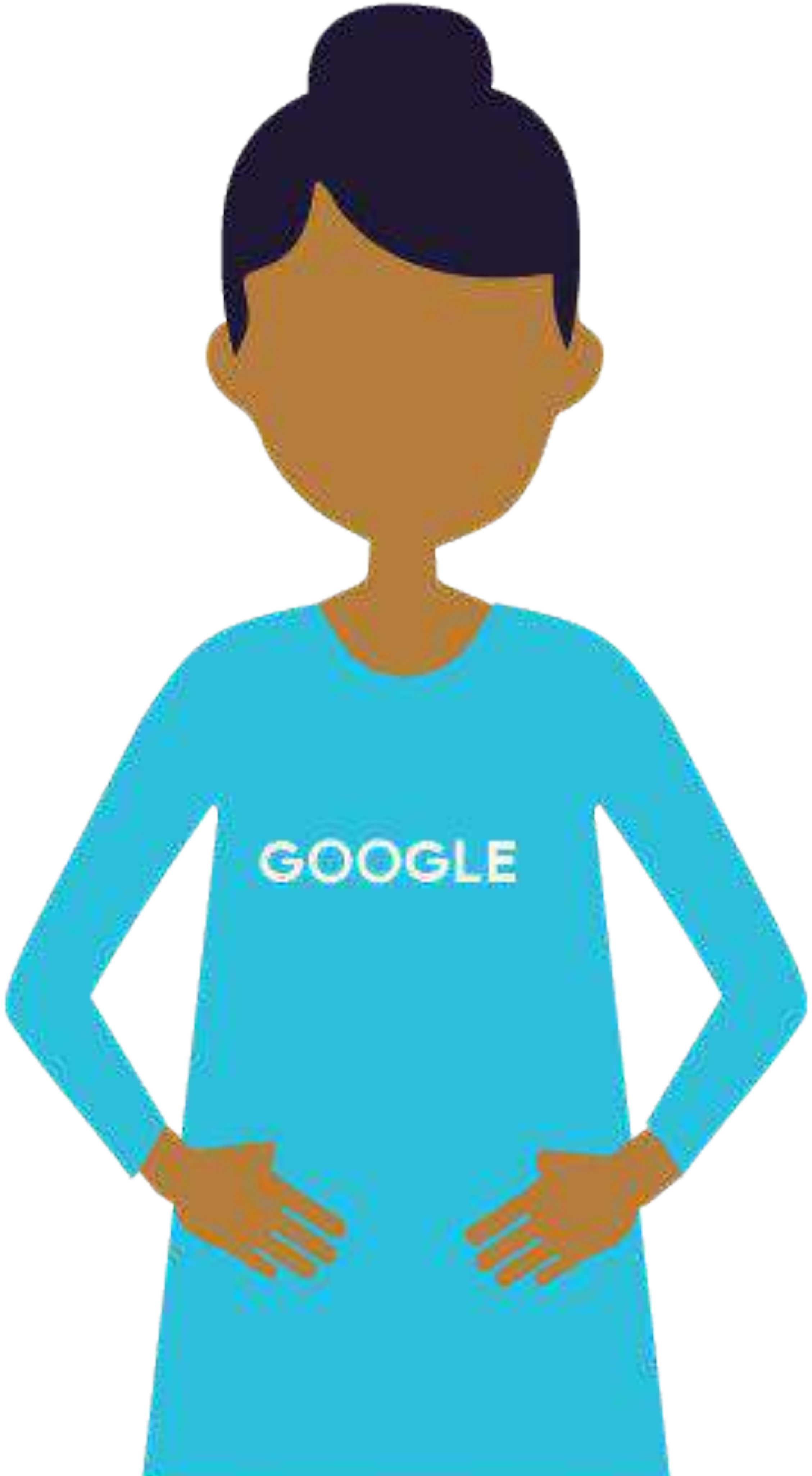
Format: Talking Head

Video Name: T-PSML-O_3_l8_adapting_to_data:_system_failure

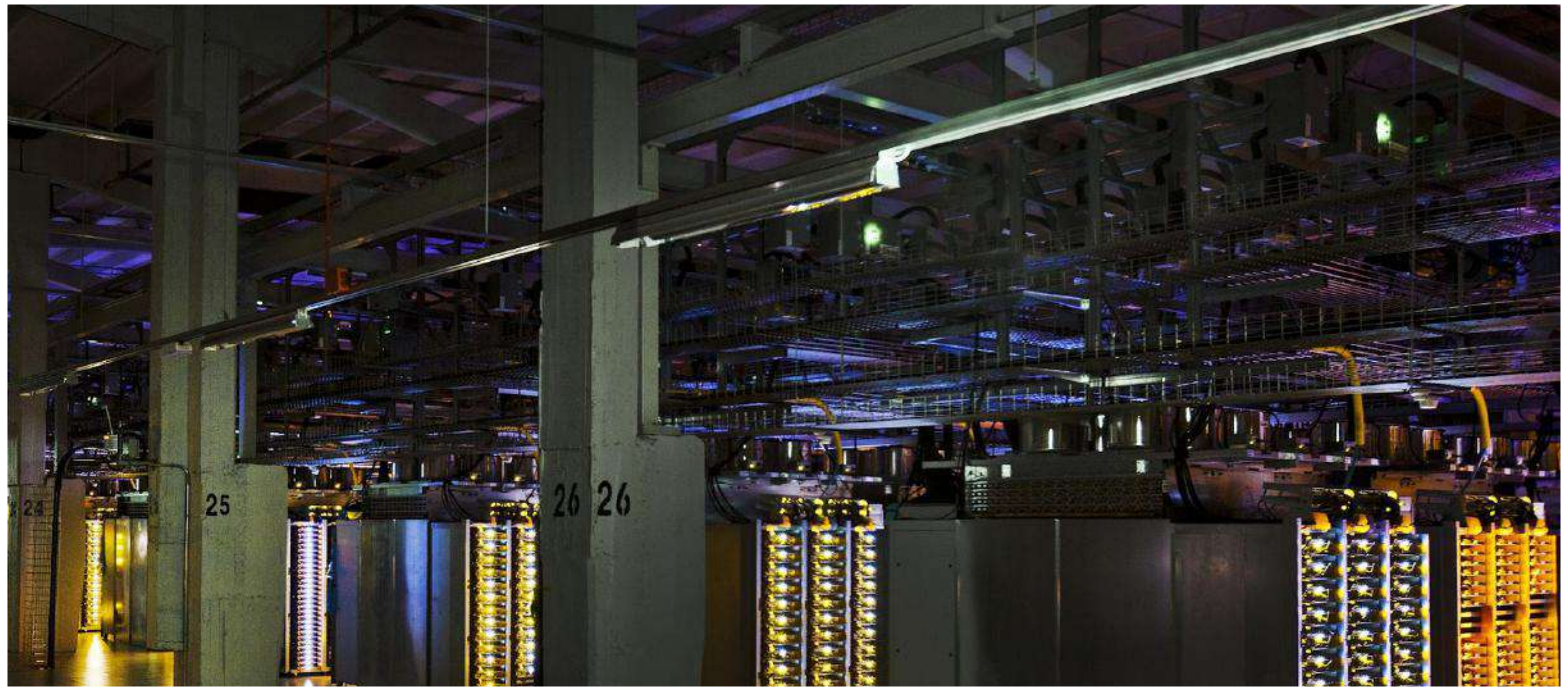


Systems Fail

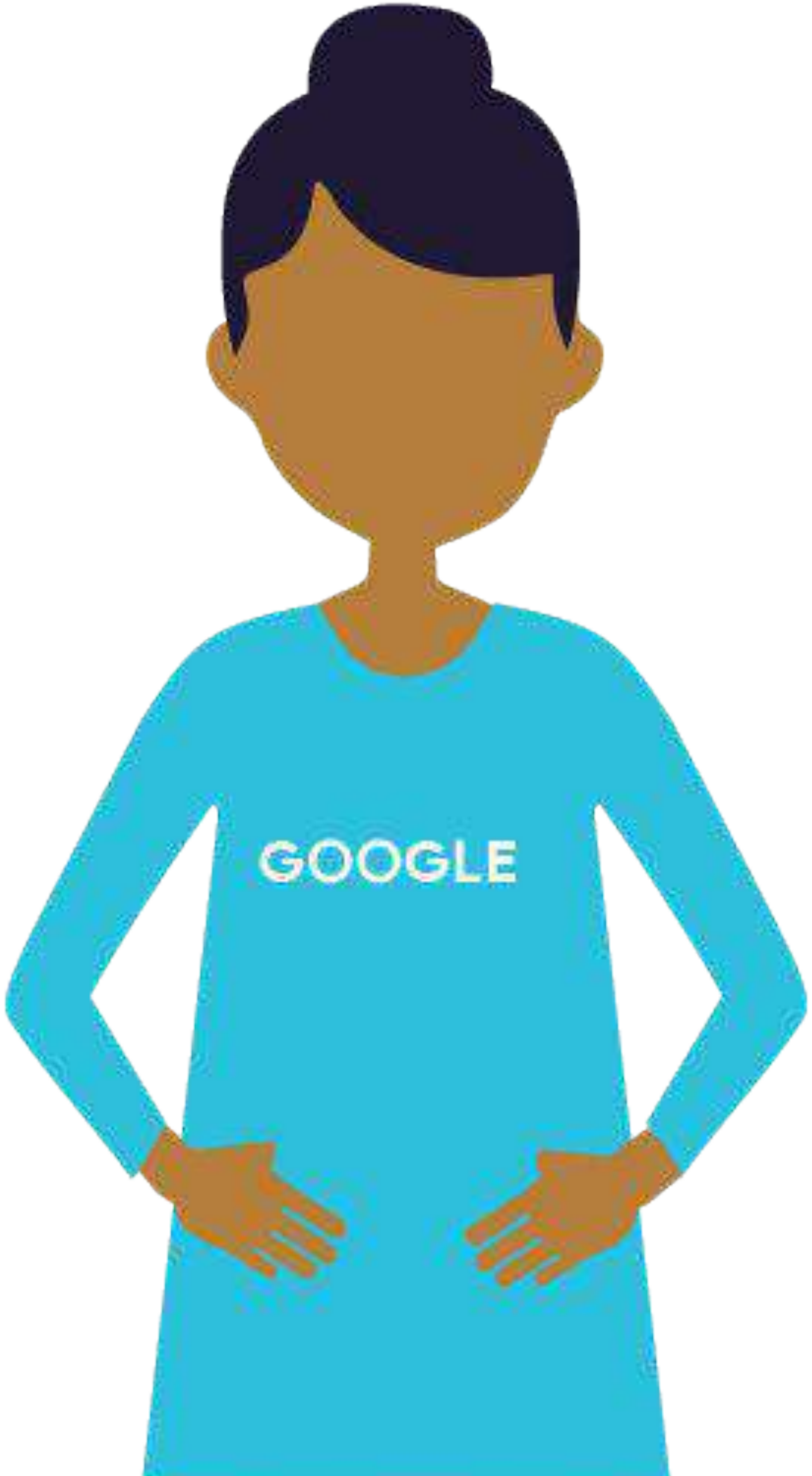




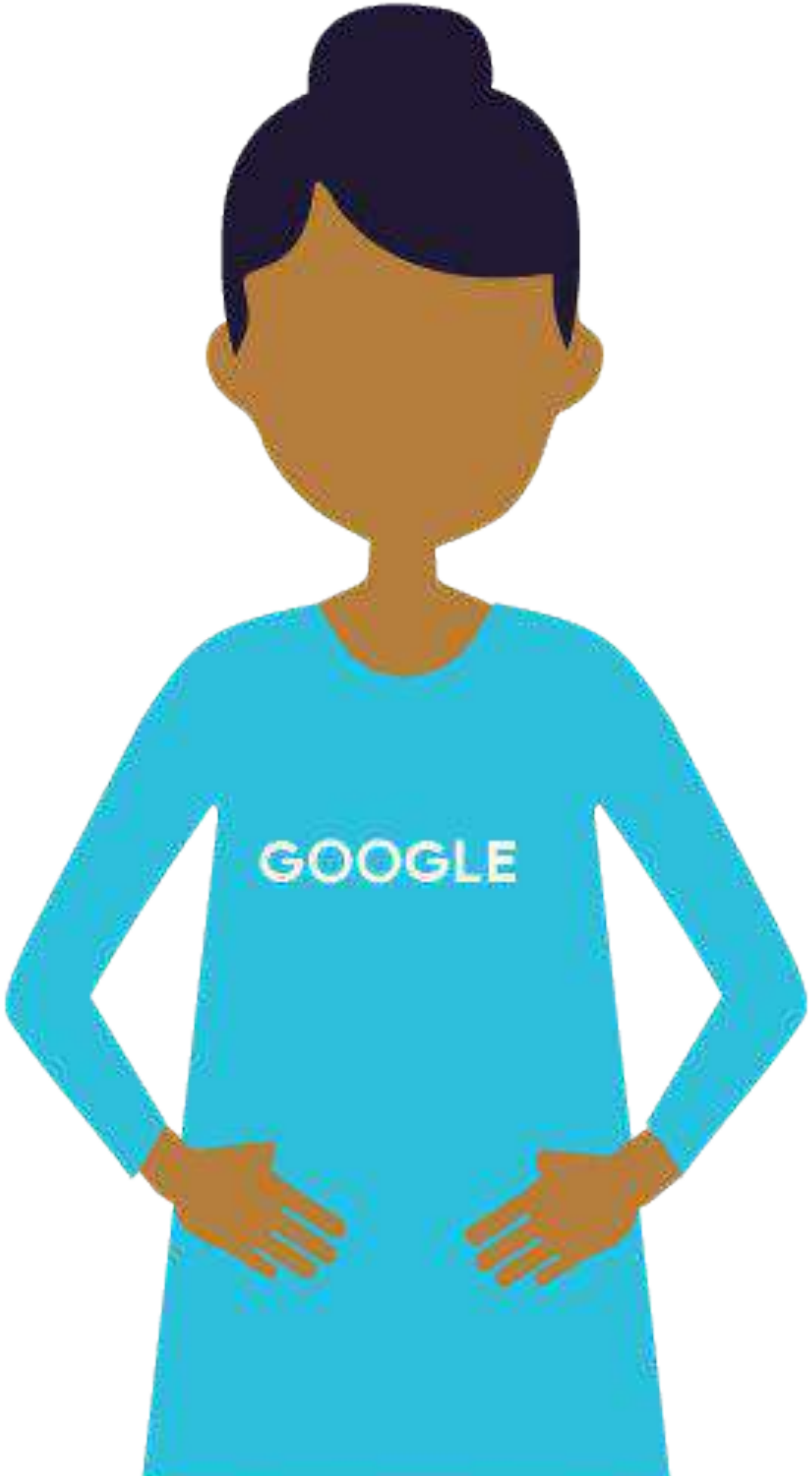
Systems Fail



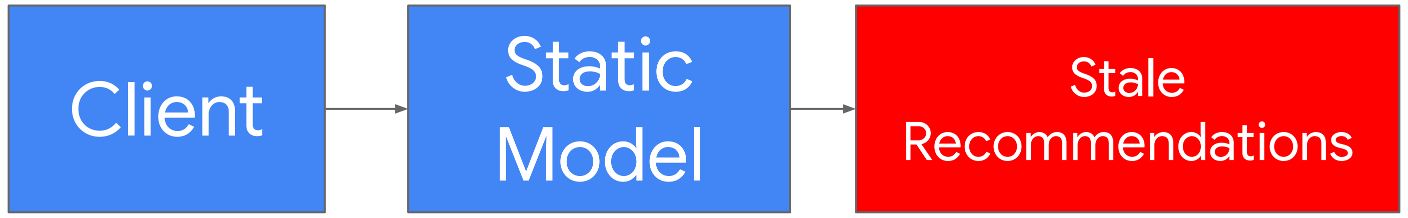
Rollback Initiated
Version 1.0.1
Three Months old



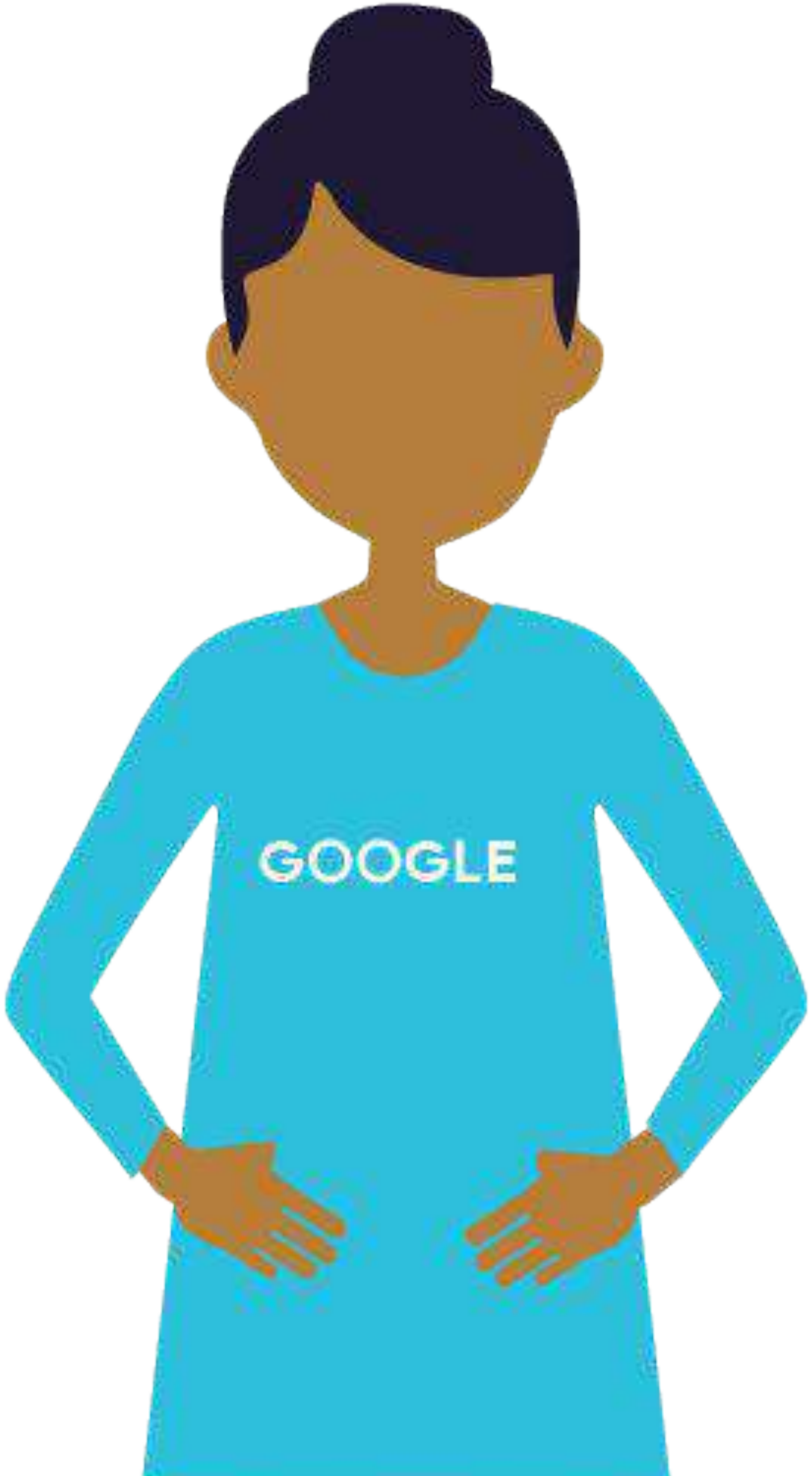
Feedback Loops



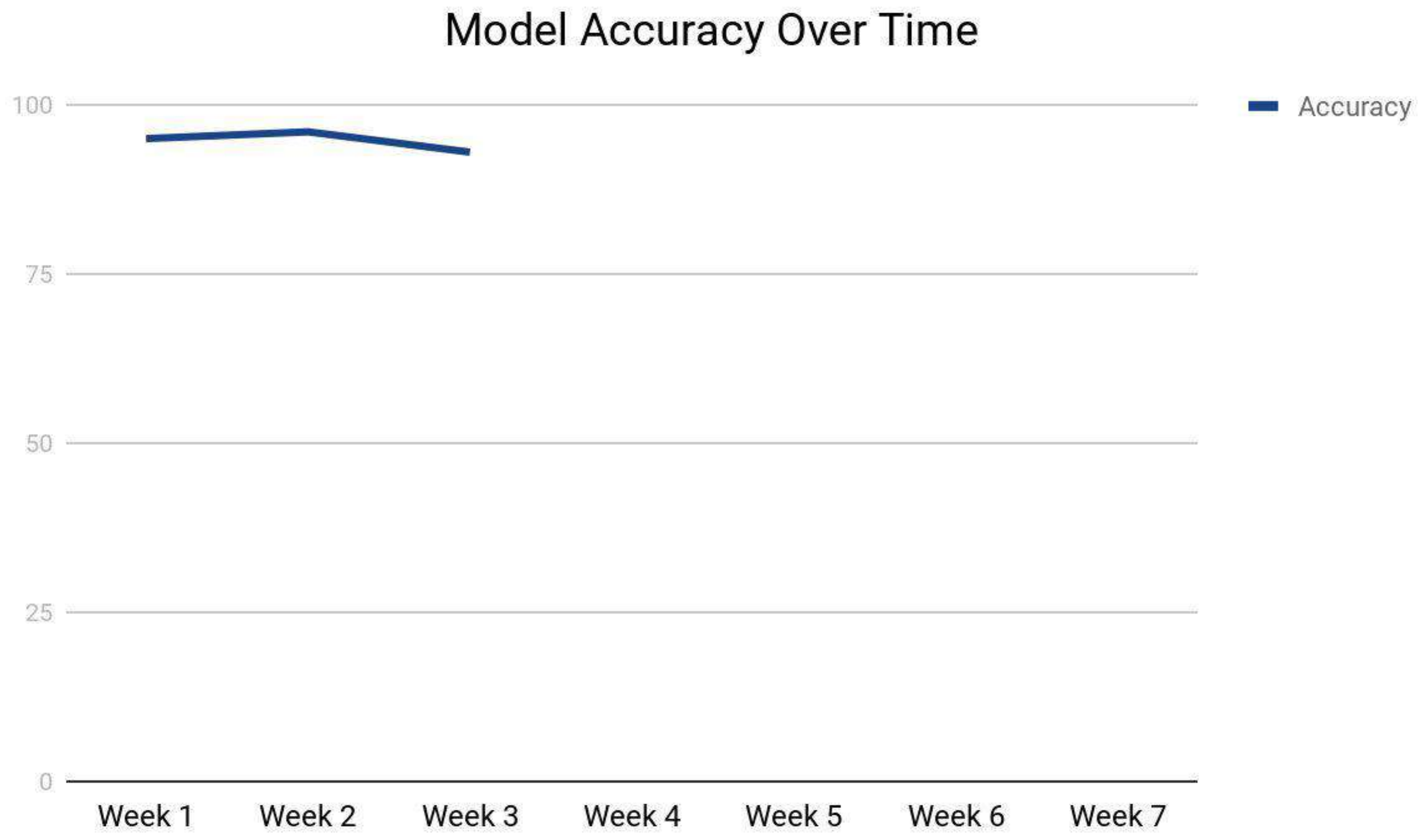
Feedback Loops

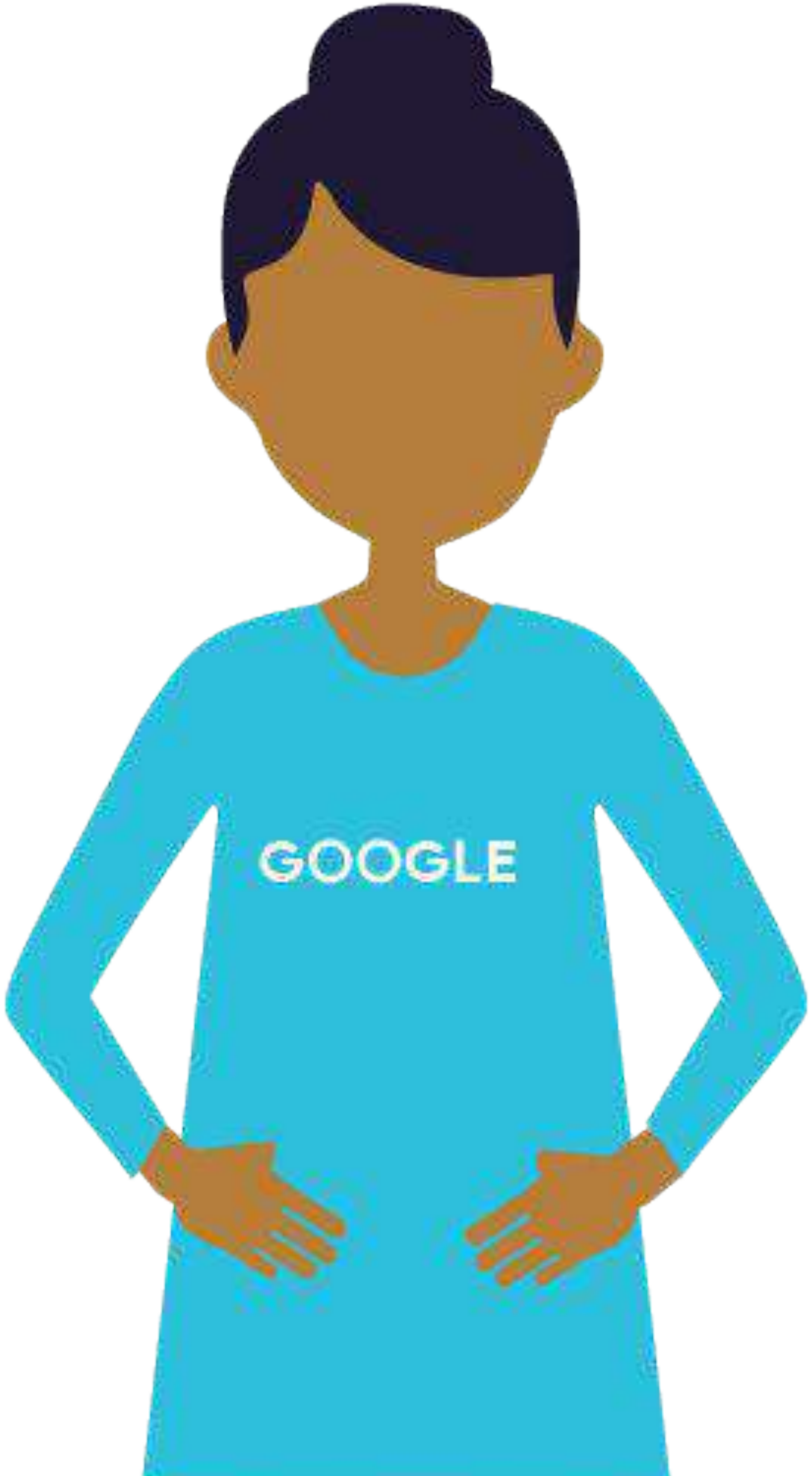


Client → Static Model → Stale Recommendations

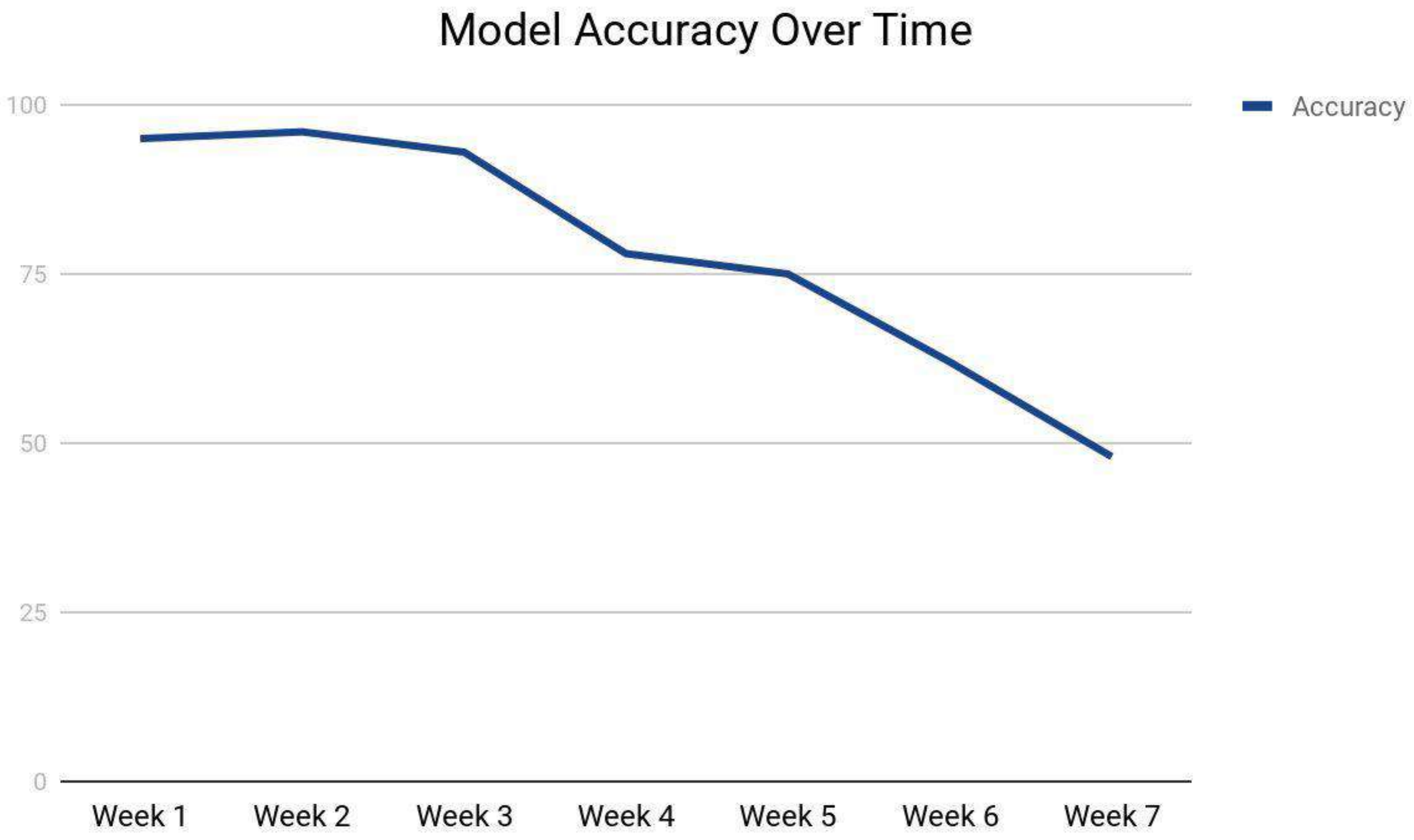


Feedback Loops





Feedback Loops



Course 2: Production ML Systems

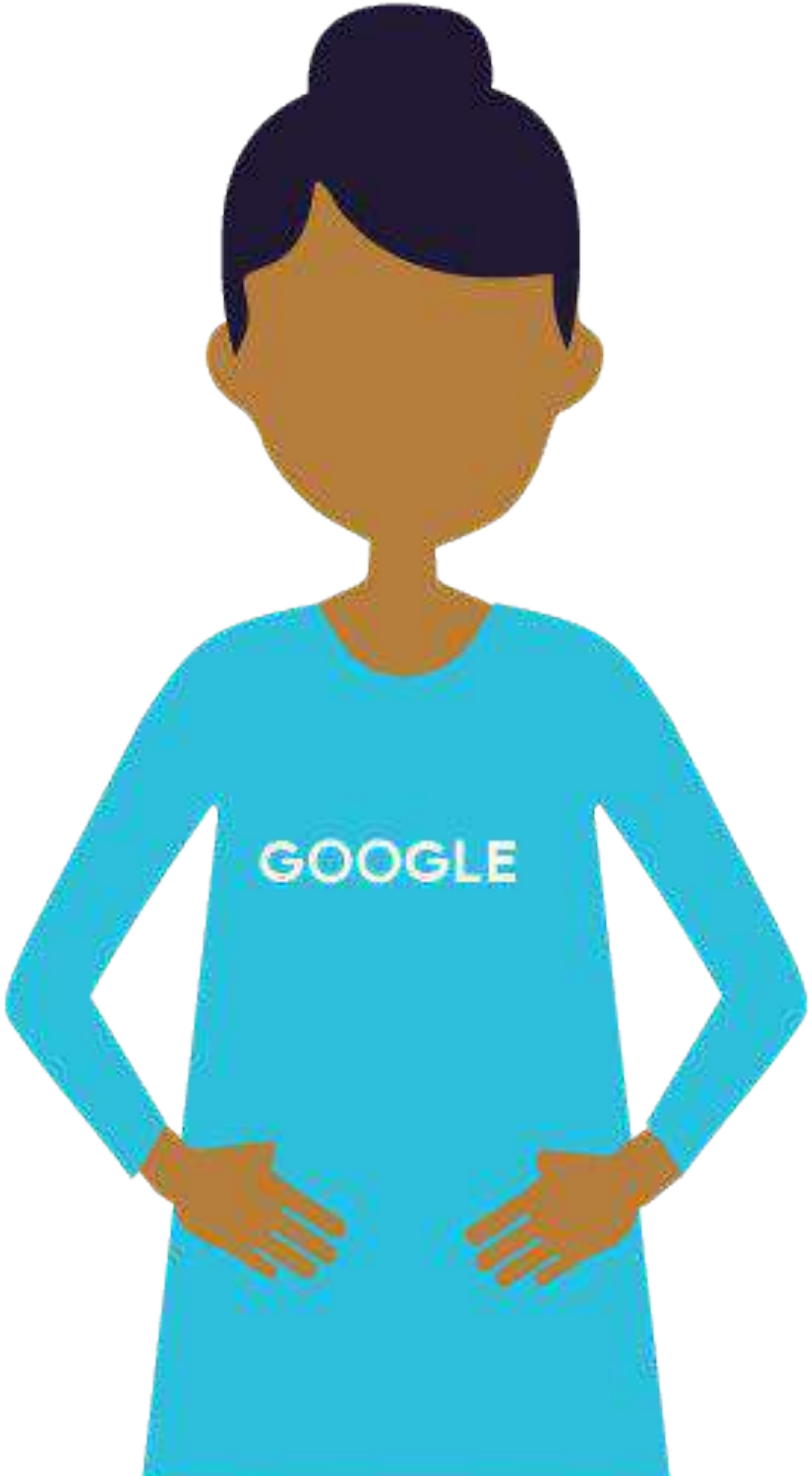
Module 3: Designing Adaptable ML Systems

Lesson Title: **Adapting to Data: Summary**

Presenter: Max Lotstein

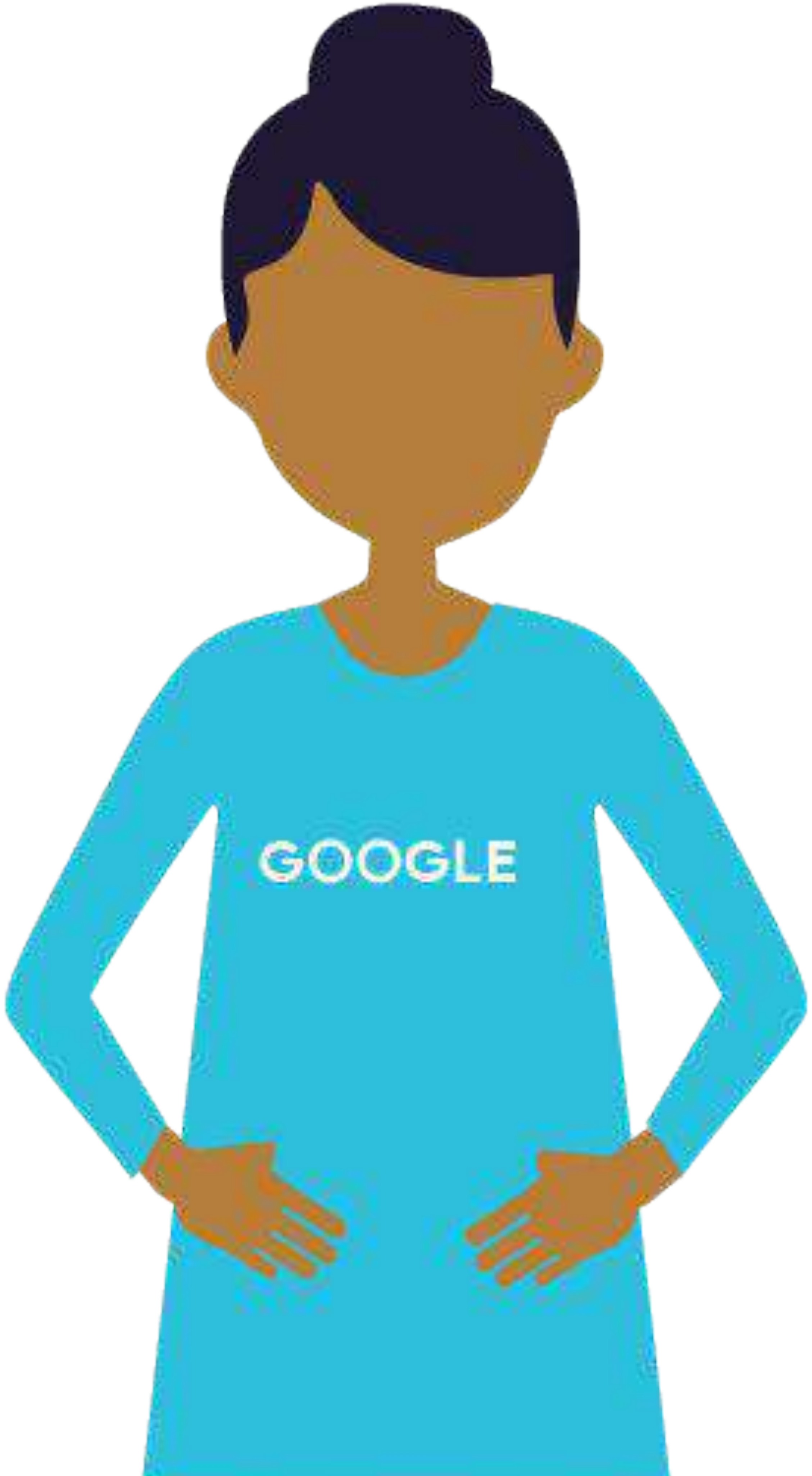
Format: Talking Head

Video Name: T-PSML-O_3_l9_adapting_to_data:_summary



Adapting to Data

- Assess all data sources and features based on both cost and benefit before including into the model
- Communicate with upstream data producers to make your needs known
- Replicate critical data sources
- Monitor descriptive statistics for your inputs and outputs



Adapting to Data

- Monitor your residuals as a function of your inputs
- Use custom weights in your loss function to emphasize data recency
- Use dynamic training architecture and regularly retrain your model
- You get what you optimize for

Course 2: Production ML Systems

Module 3: Designing Adaptable ML Systems

Lesson Title: **Mitigating Training-Serving Skew Through Design**

Presenter: Max Lotstein

Format: Talking Head

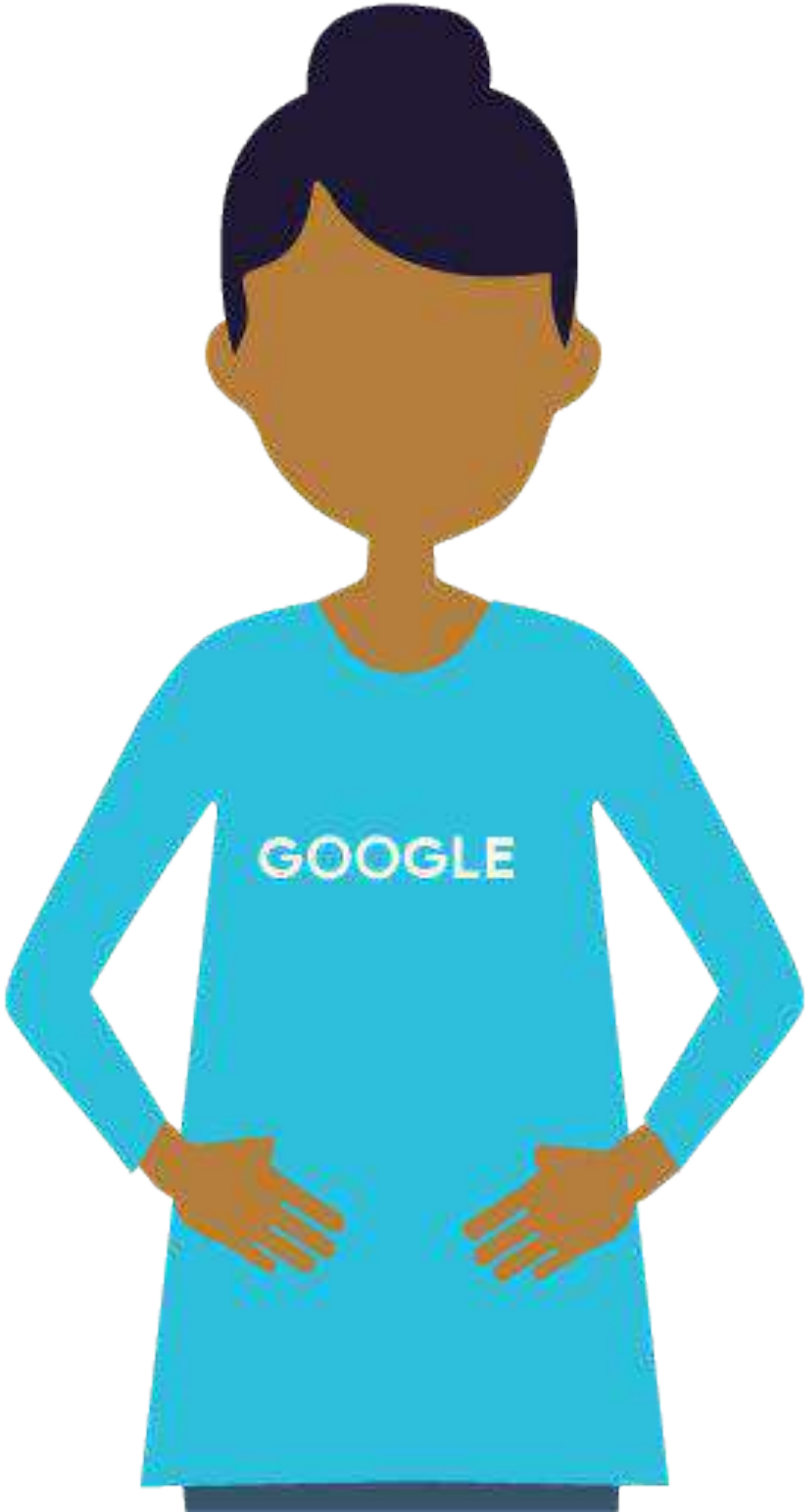
Video Name:
T-PSML-O_3_l10_mitigating_training-serving_skew_through_design

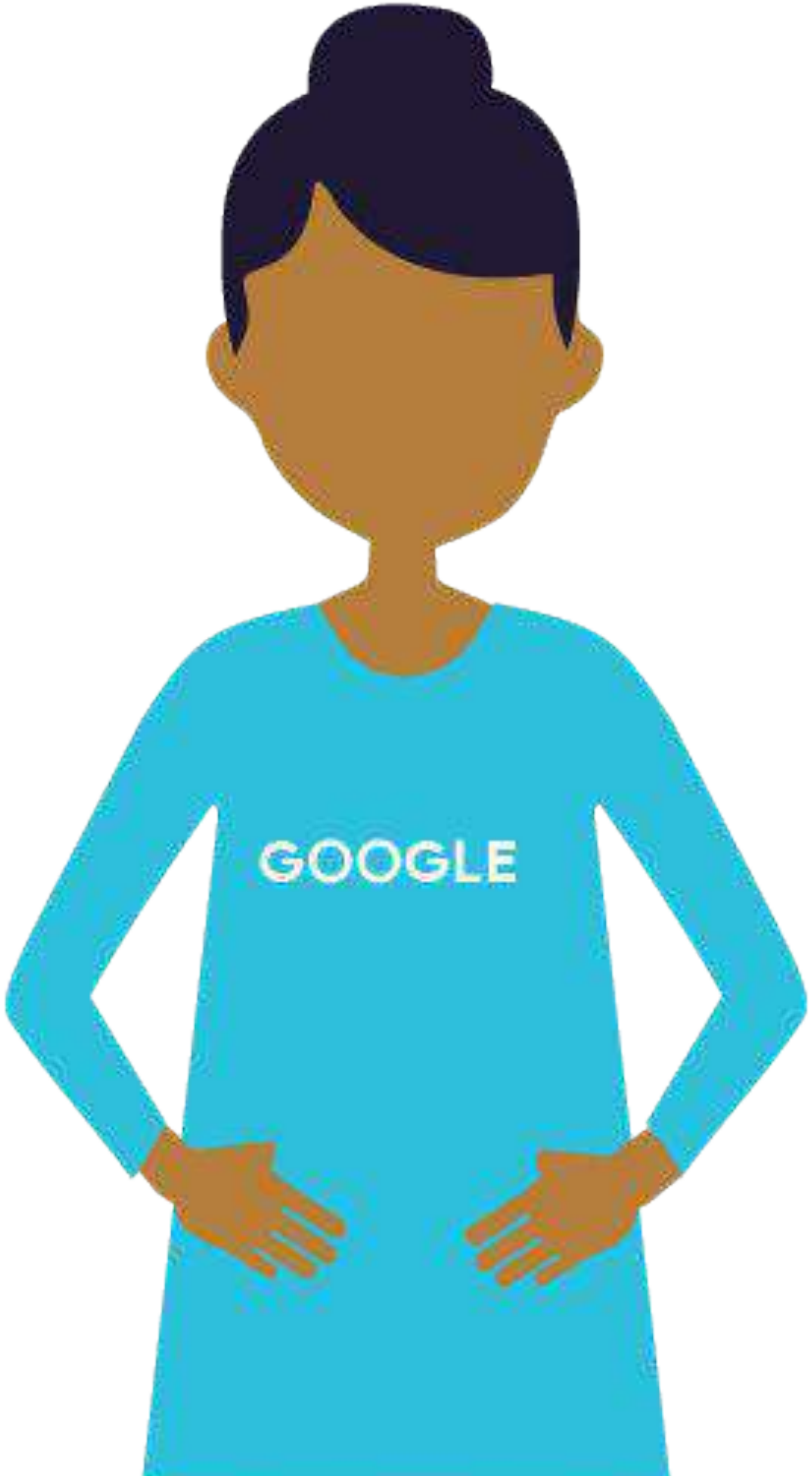
Agenda

Adapting to Data

Mitigating Training-Serving Skew Through Design

Debugging a Production Model



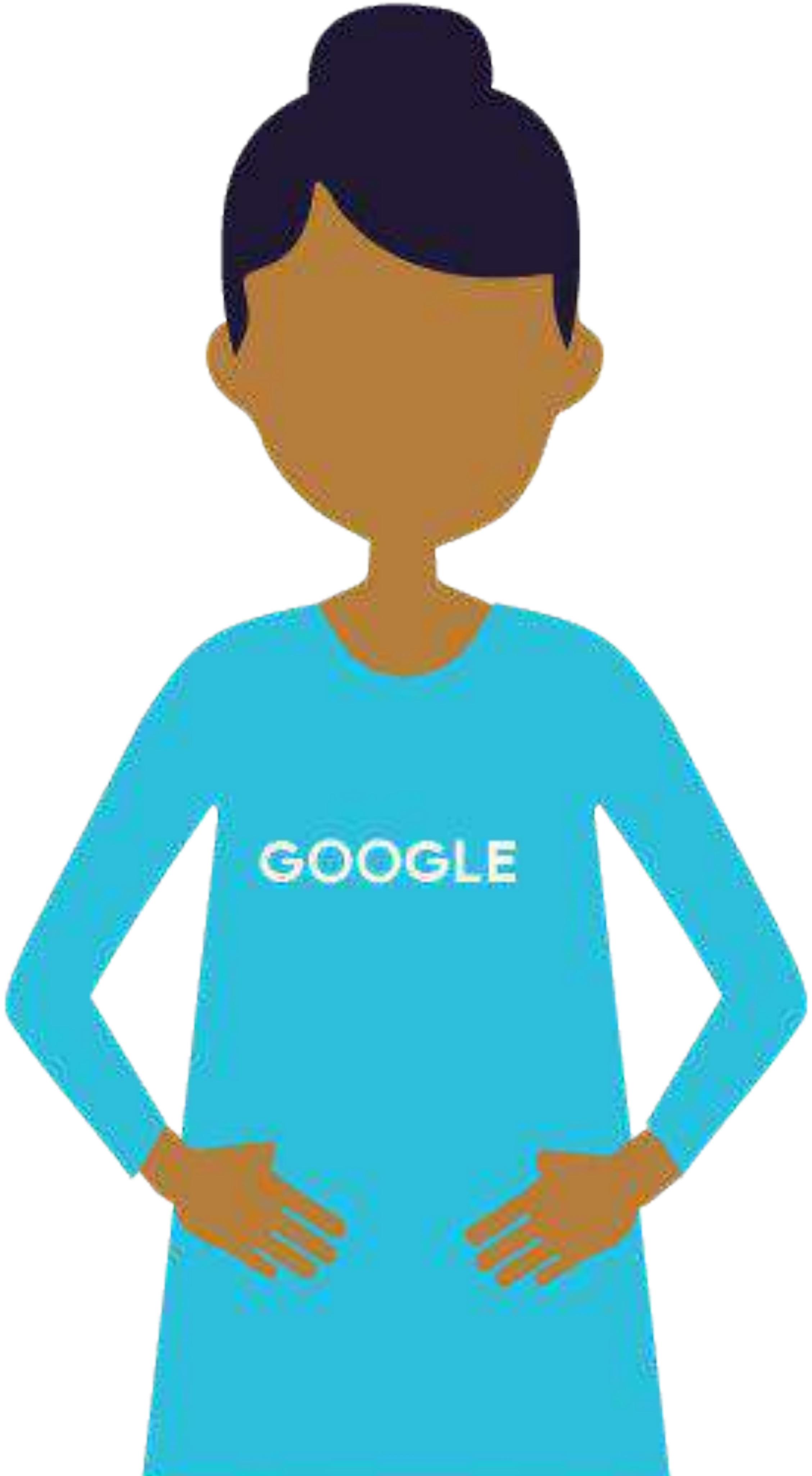


Agenda

Adapting to Data

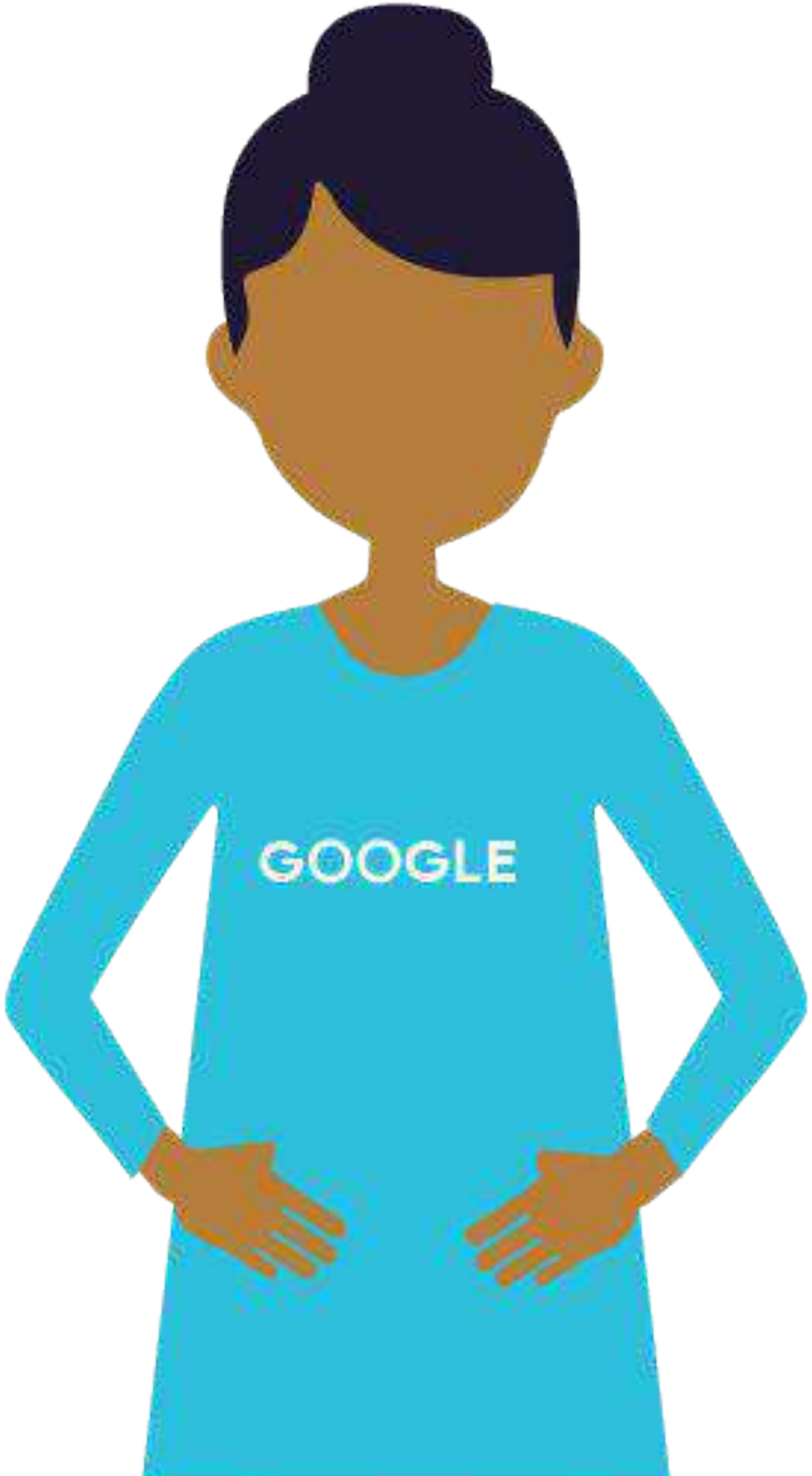
Mitigating Training-Serving Skew Through Design

Debugging a Production Model



Training/Serving Skew

1. A discrepancy between how you handle data in the training and serving pipelines
2. A change in the data between when you train and when you serve.
3. A feedback loop between your model and your algorithm.



How Code Can Create Training/Serving Skew

- Different library versions that are functionally equivalent but optimized differently
- Different library versions that are not functionally equivalent
- Re-implemented functions

Course 2: Production ML Systems

Module 3: Designing Adaptable ML Systems

Lesson Title: **Lab Intro: Serving ML Predictions in batch and real-time**

Presenter: Max Lotstein

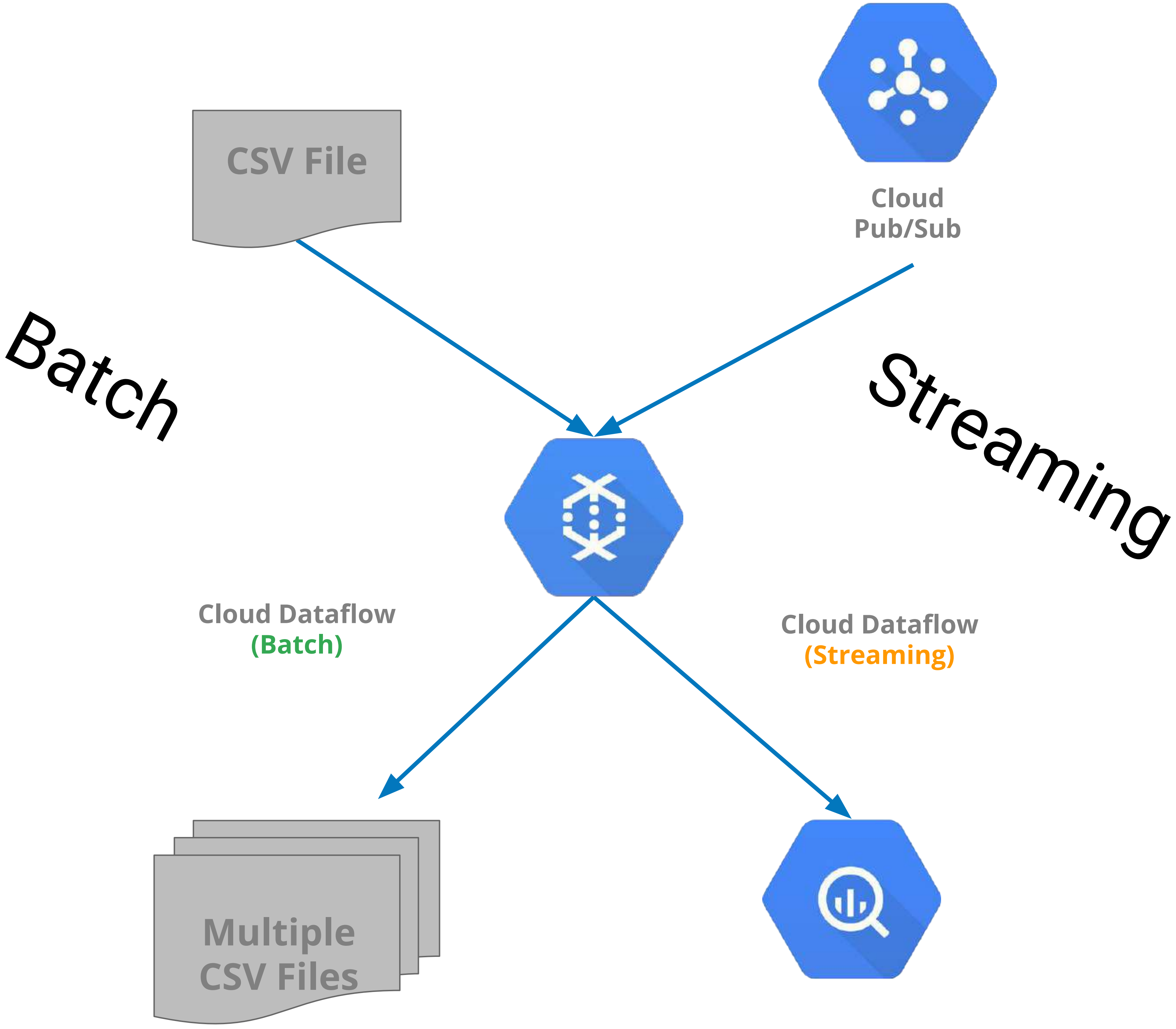
Format: Talking Head

Video Name:
T-PSML-O_3_l11_lab_intro:_serving_ml_predictions_in_batch_and_real-time

Lab

Serving ML Predictions in batch and real-time

Max Lotstein



CSV File

Cloud
Pub/Sub

Batch

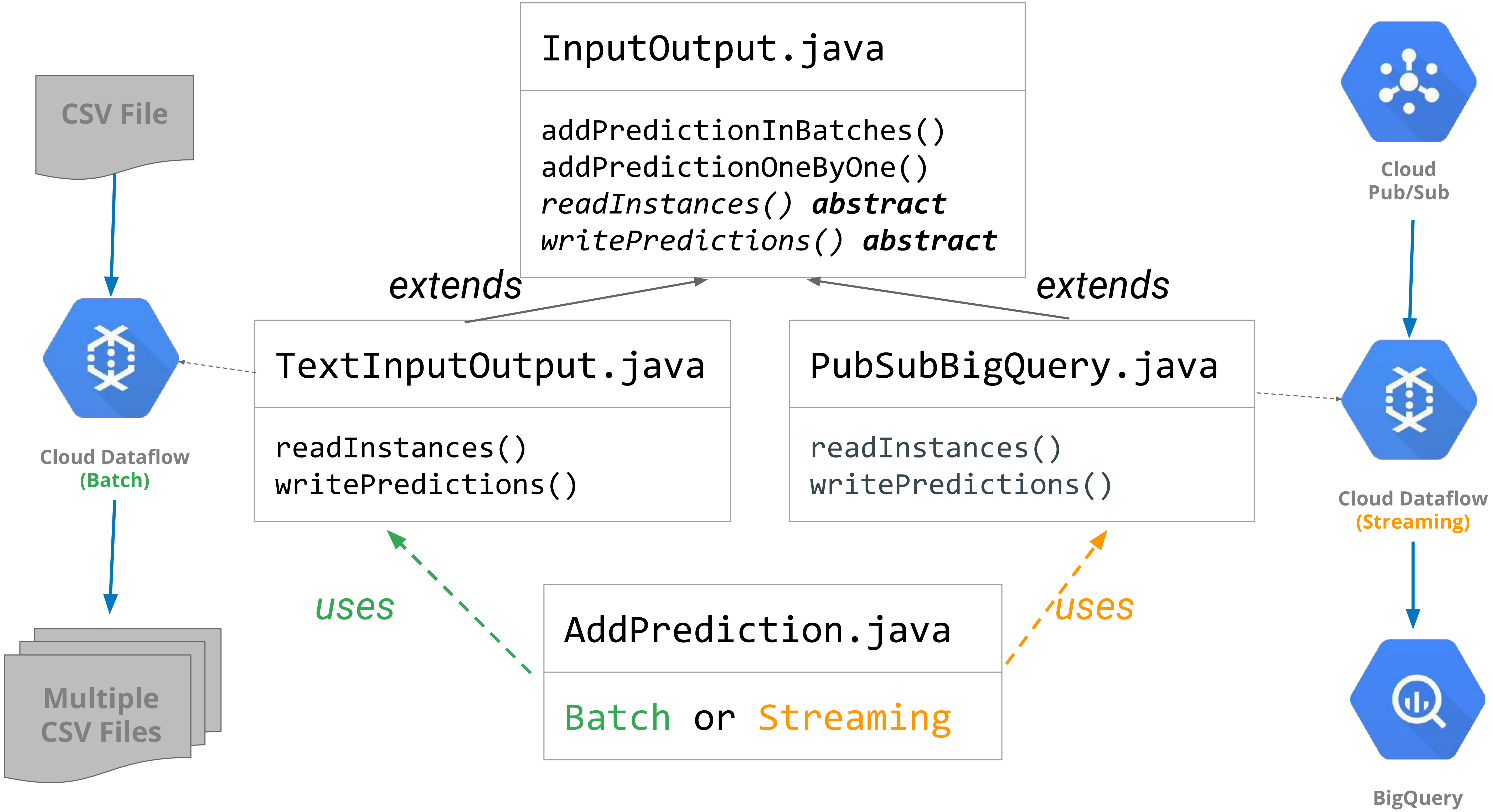
Streaming

Cloud Dataflow
(Batch)

Cloud Dataflow
(Streaming)

Multiple
CSV Files

Lab: Serving ML Predictions in batch and real-time



Course 2: Production ML Systems

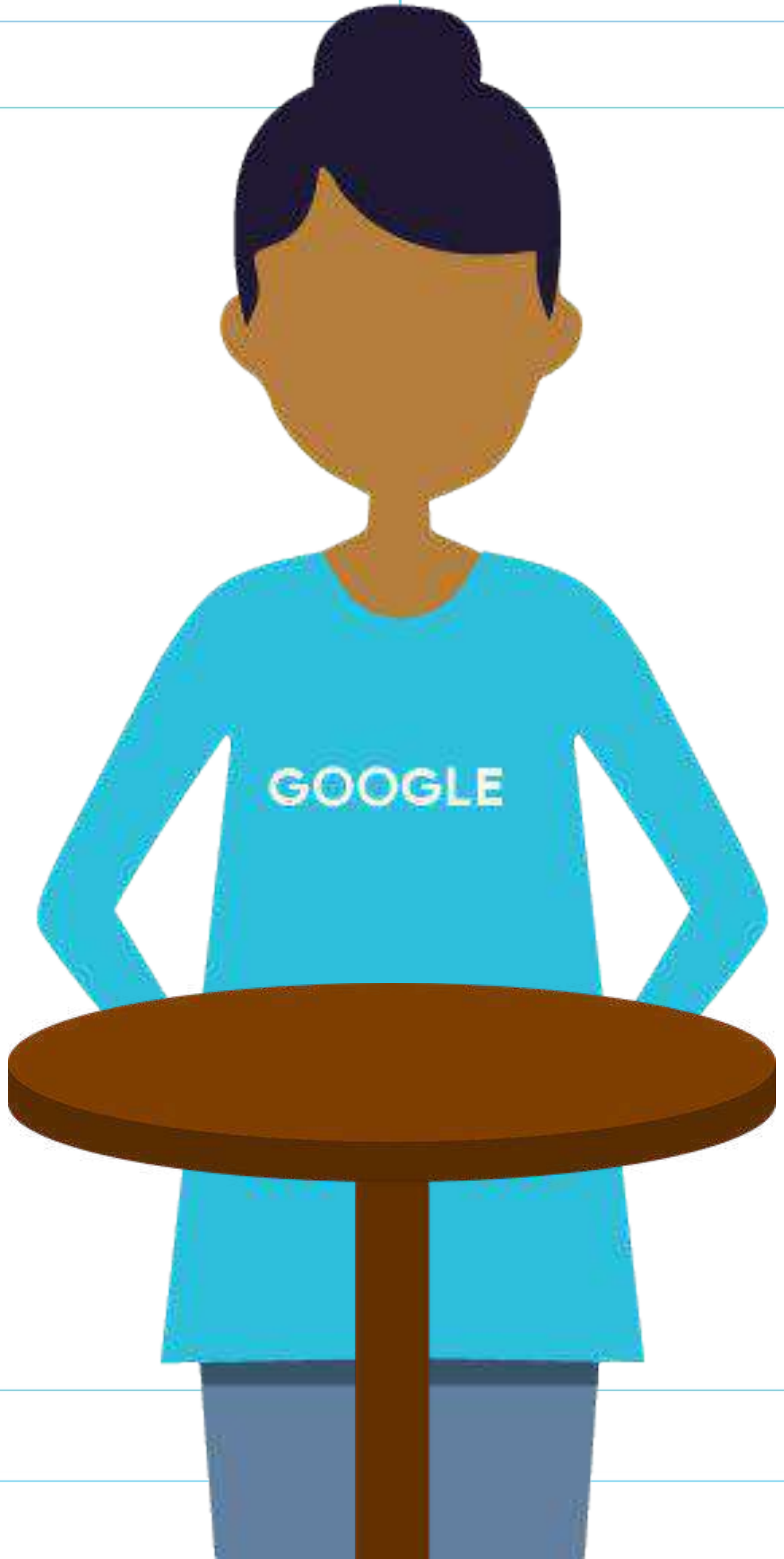
Module 3: Designing Adaptable ML Systems

Lesson Title: **Lab Solution: Serving ML Predictions in batch and real-time**

Presenter: Max Lotstein

Format: Talking Head

Video Name:
T-PSML-O_3_l12_lab_solution:_serving_ml_predictions_in_batch_and_real-time



Title Safe >

< Action Safe

Course 2: Production ML Systems

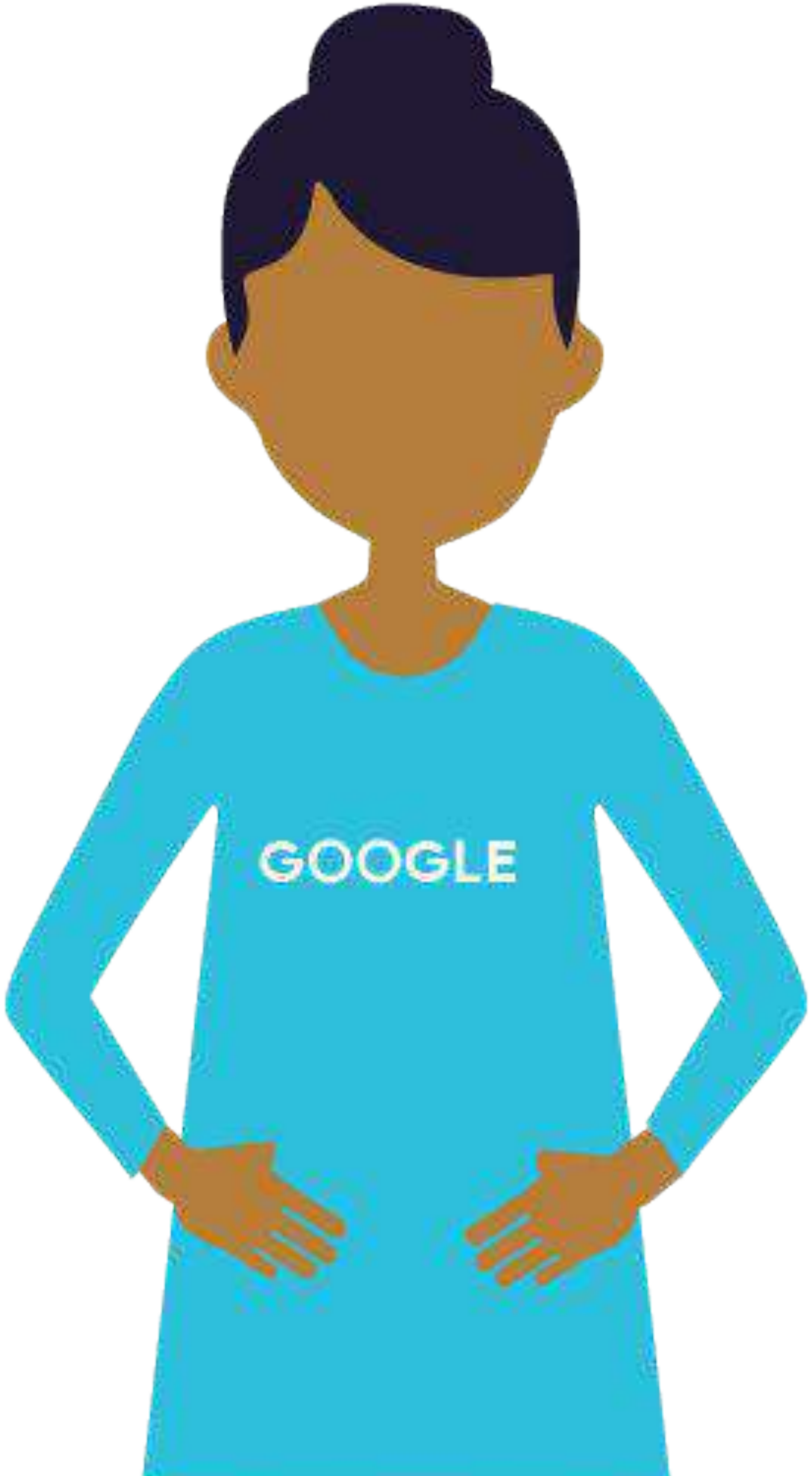
Module 3: Designing Adaptable ML Systems

Lesson Title: **Debugging a Production Model**

Presenter: Max Lotstein

Format: Talking Head

Video Name: T-PSML-O_3_l13_debugging_a_production_model



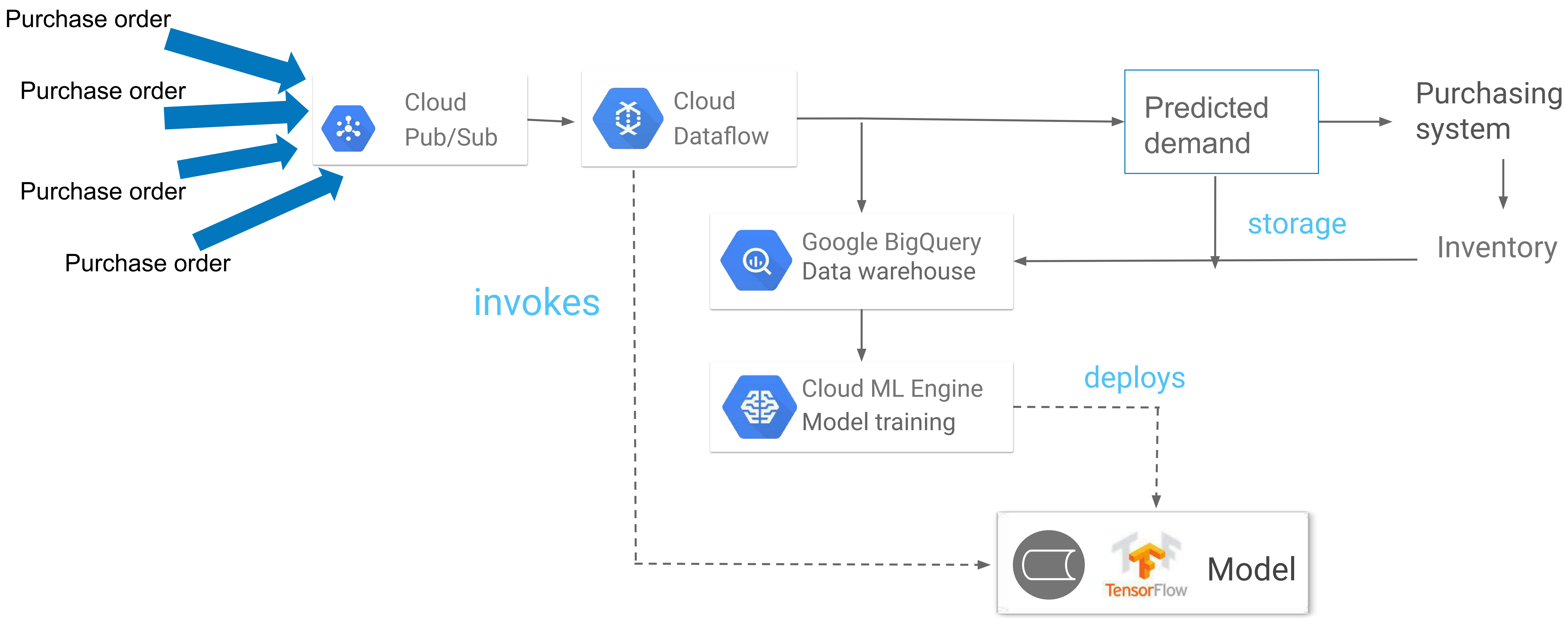
Agenda

Adapting to Data

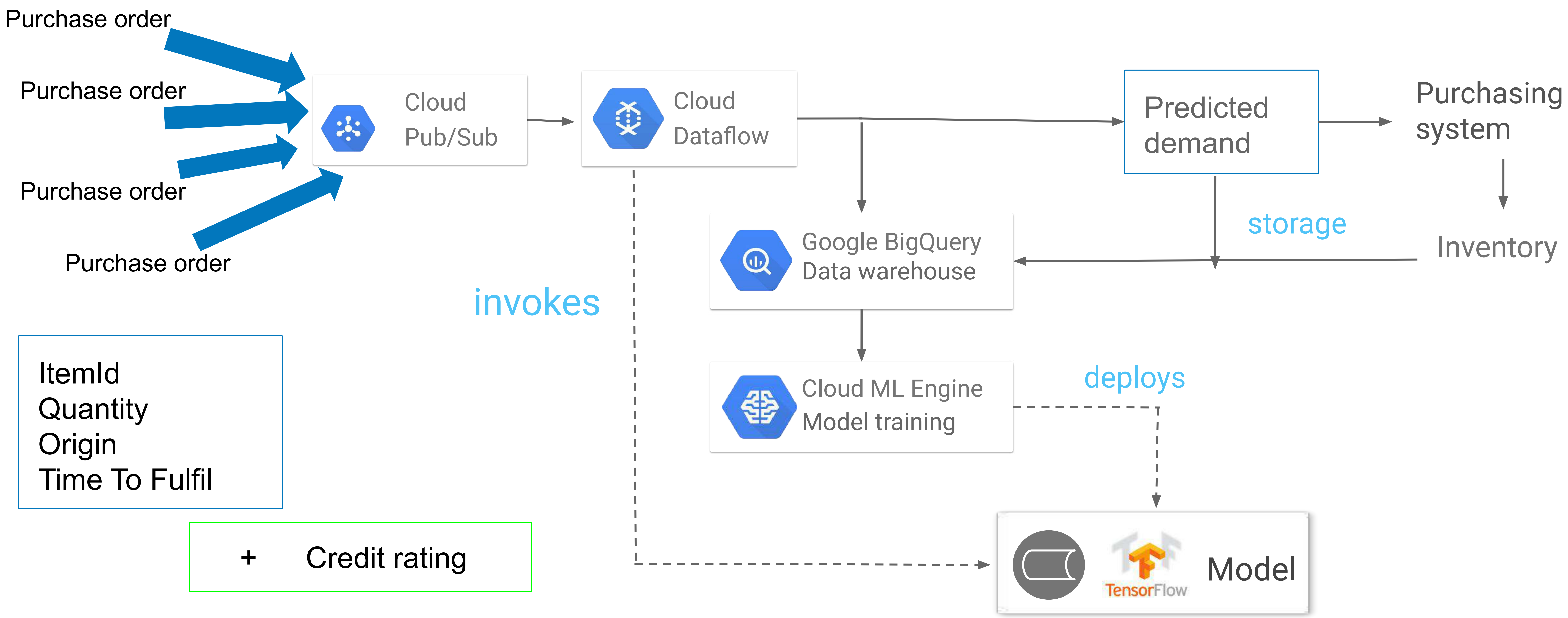
Mitigating Training-Serving Skew Through Design

Debugging a Production Model

Predicting Widget Demand

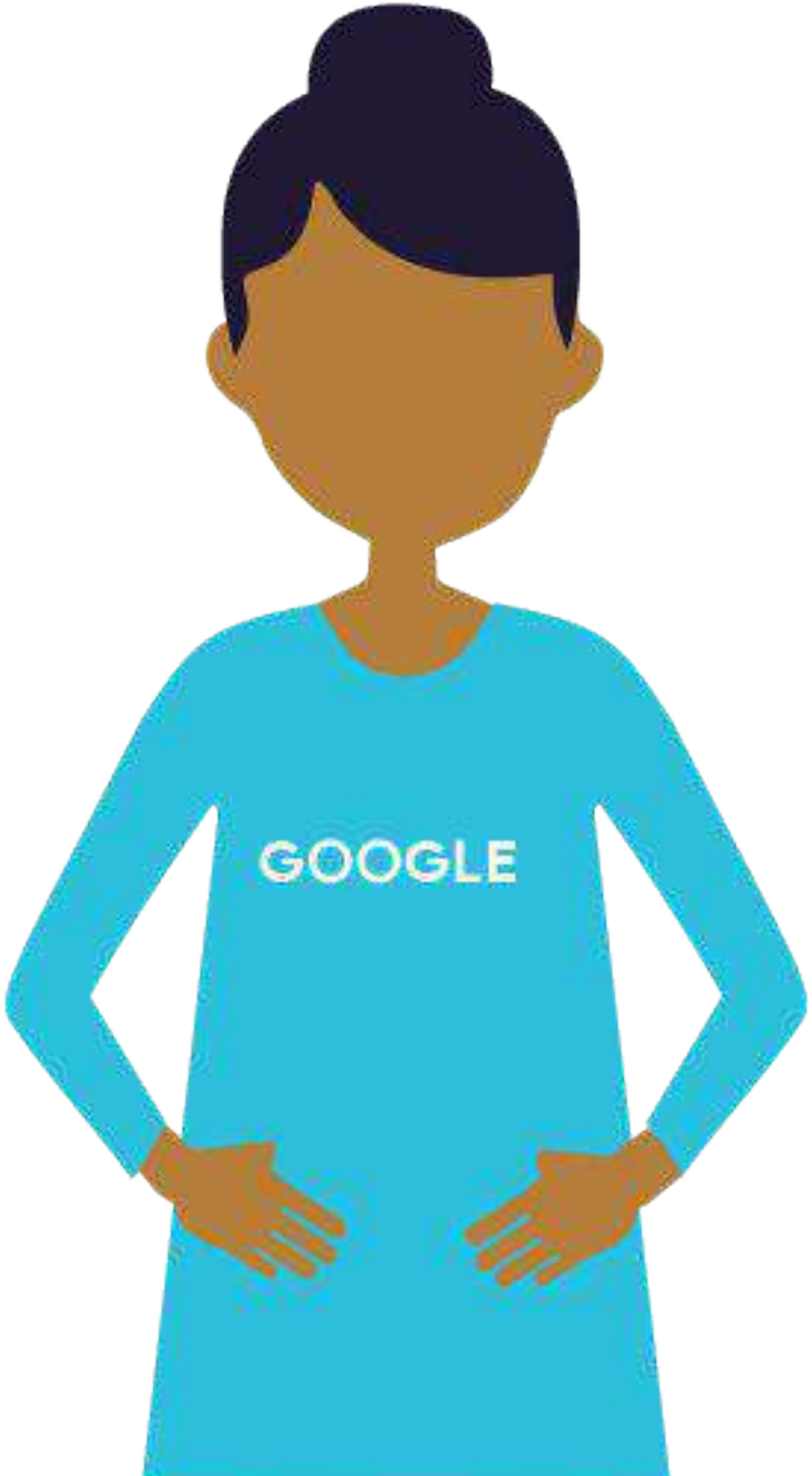


Along comes a new feature

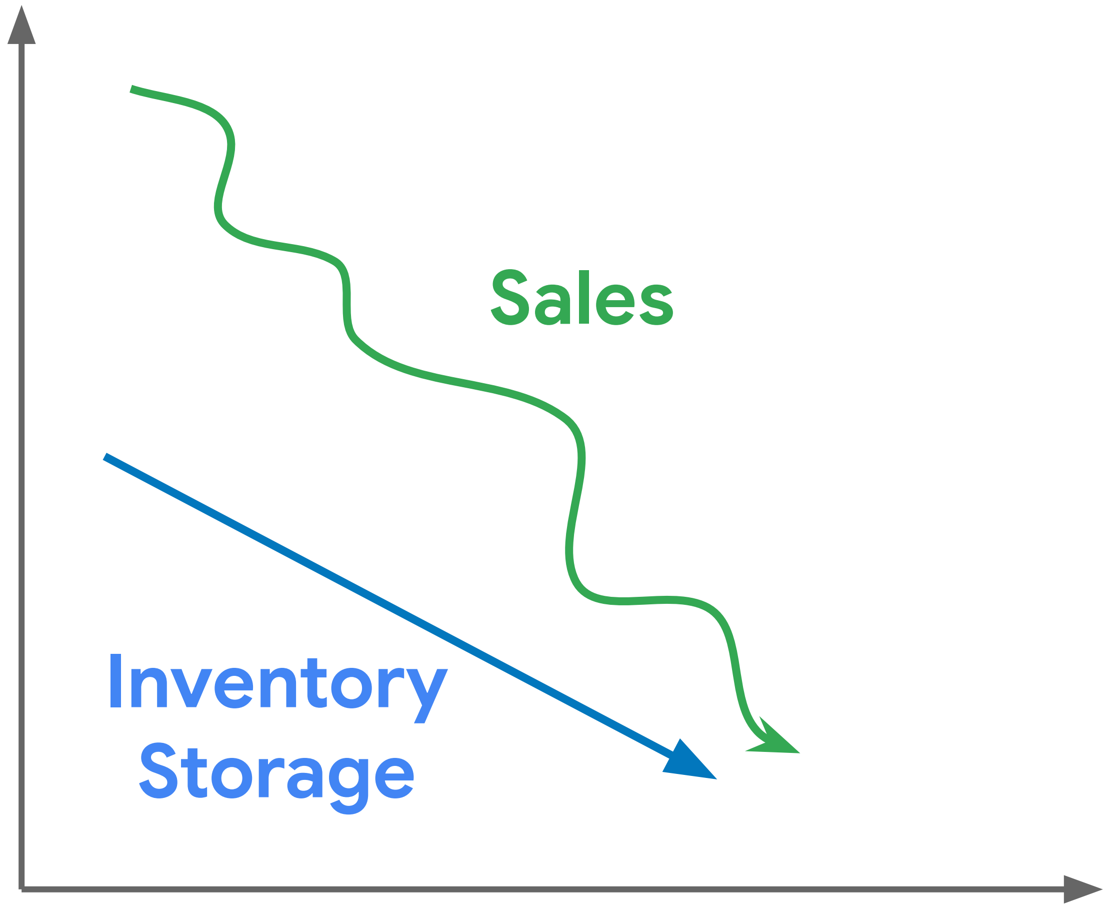


ItemId
Quantity
Origin
Time To Fulfil

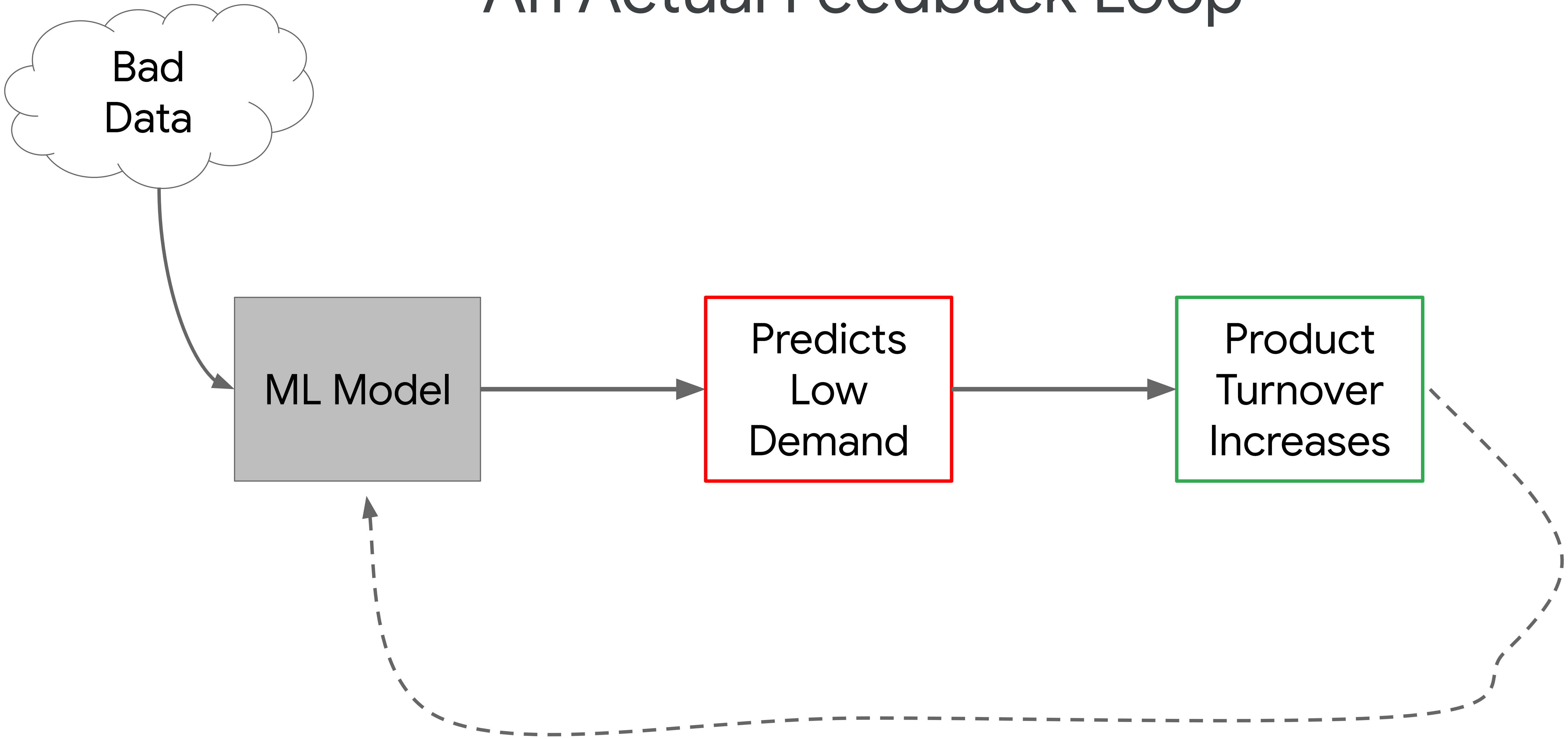
+ Credit rating

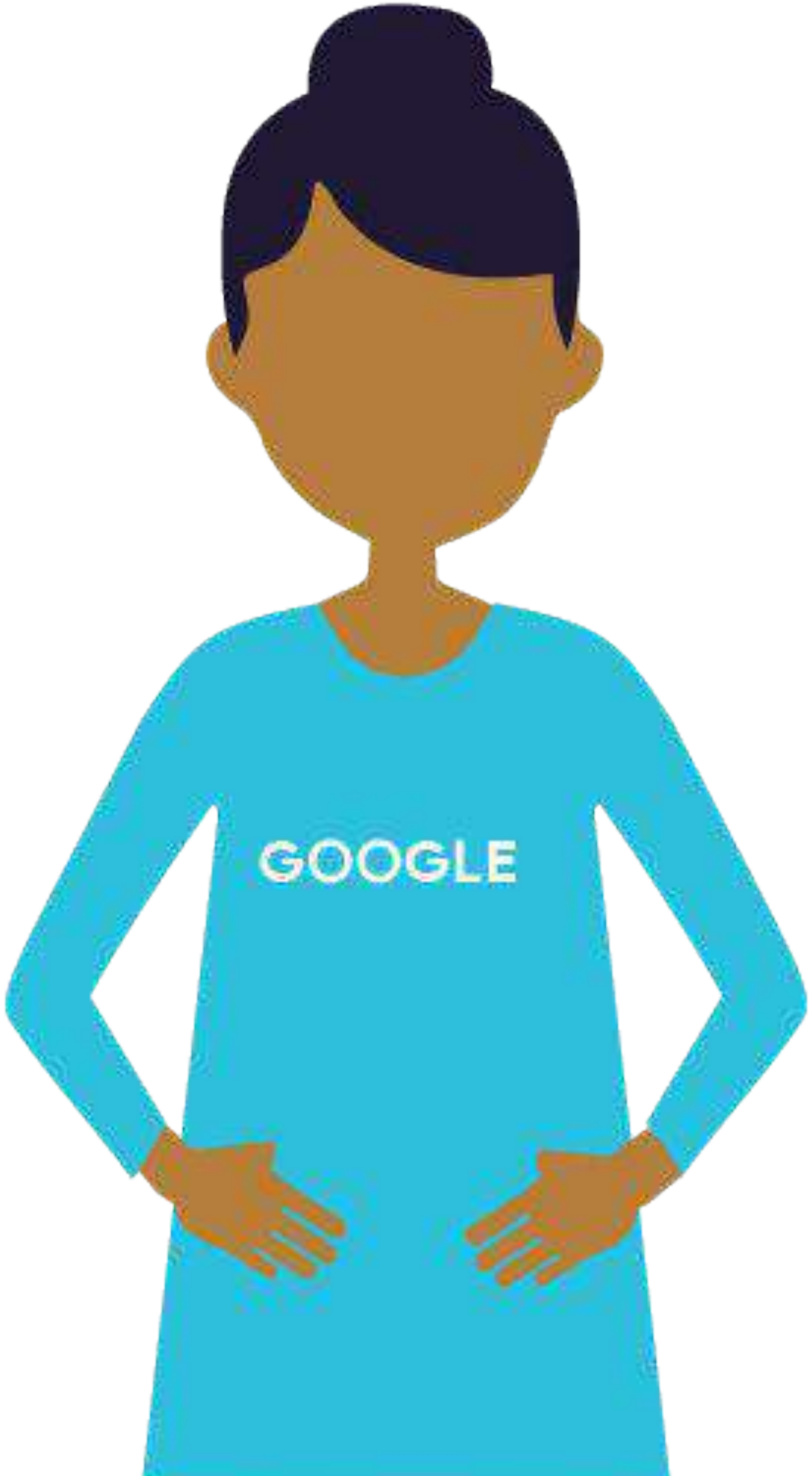


Business Catastrophe 1

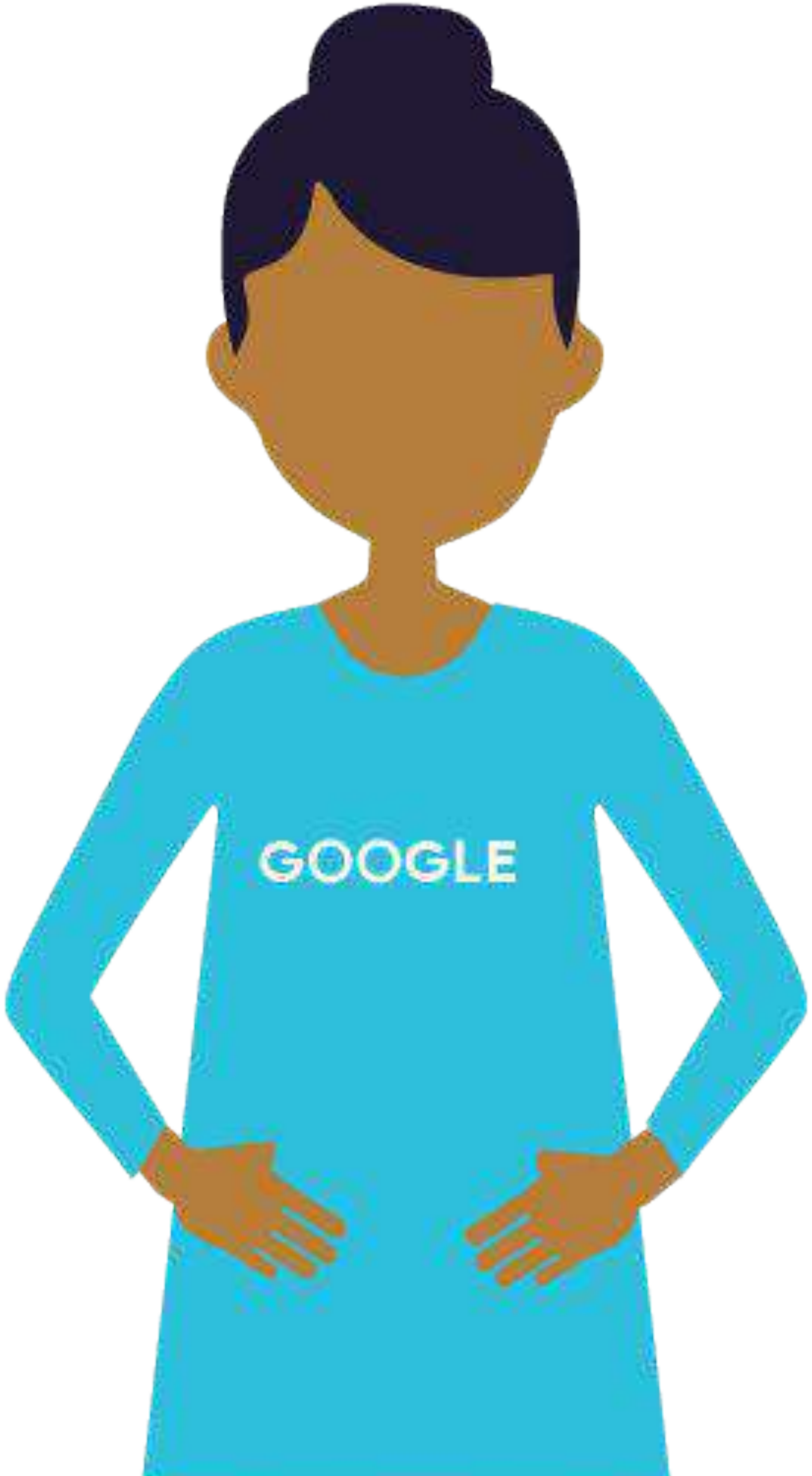


An Actual Feedback Loop

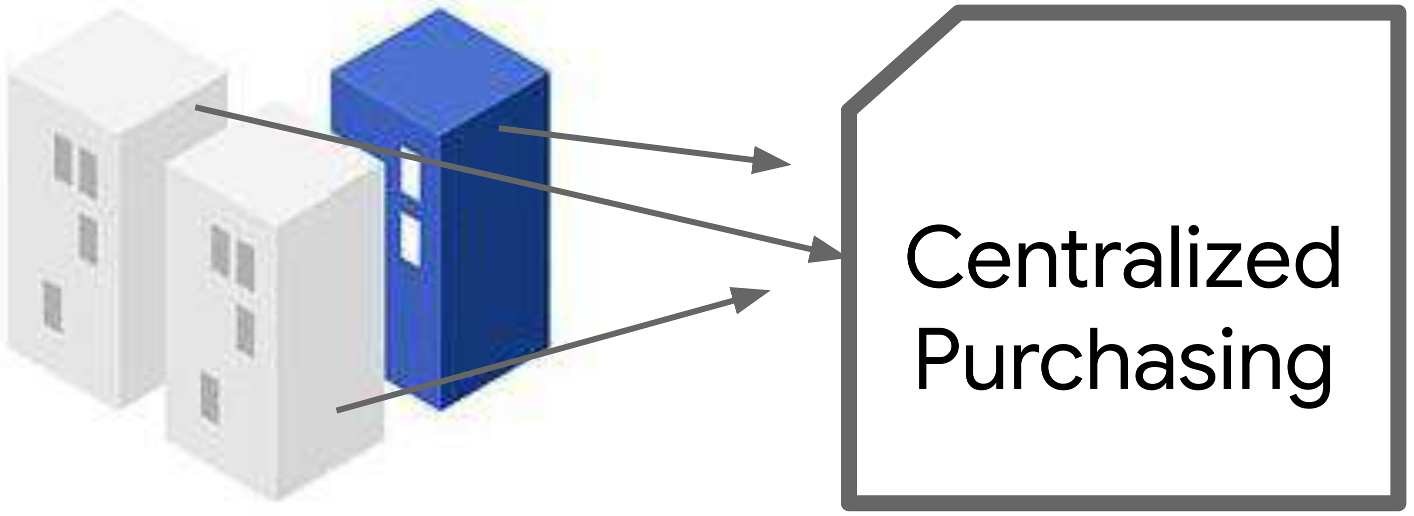


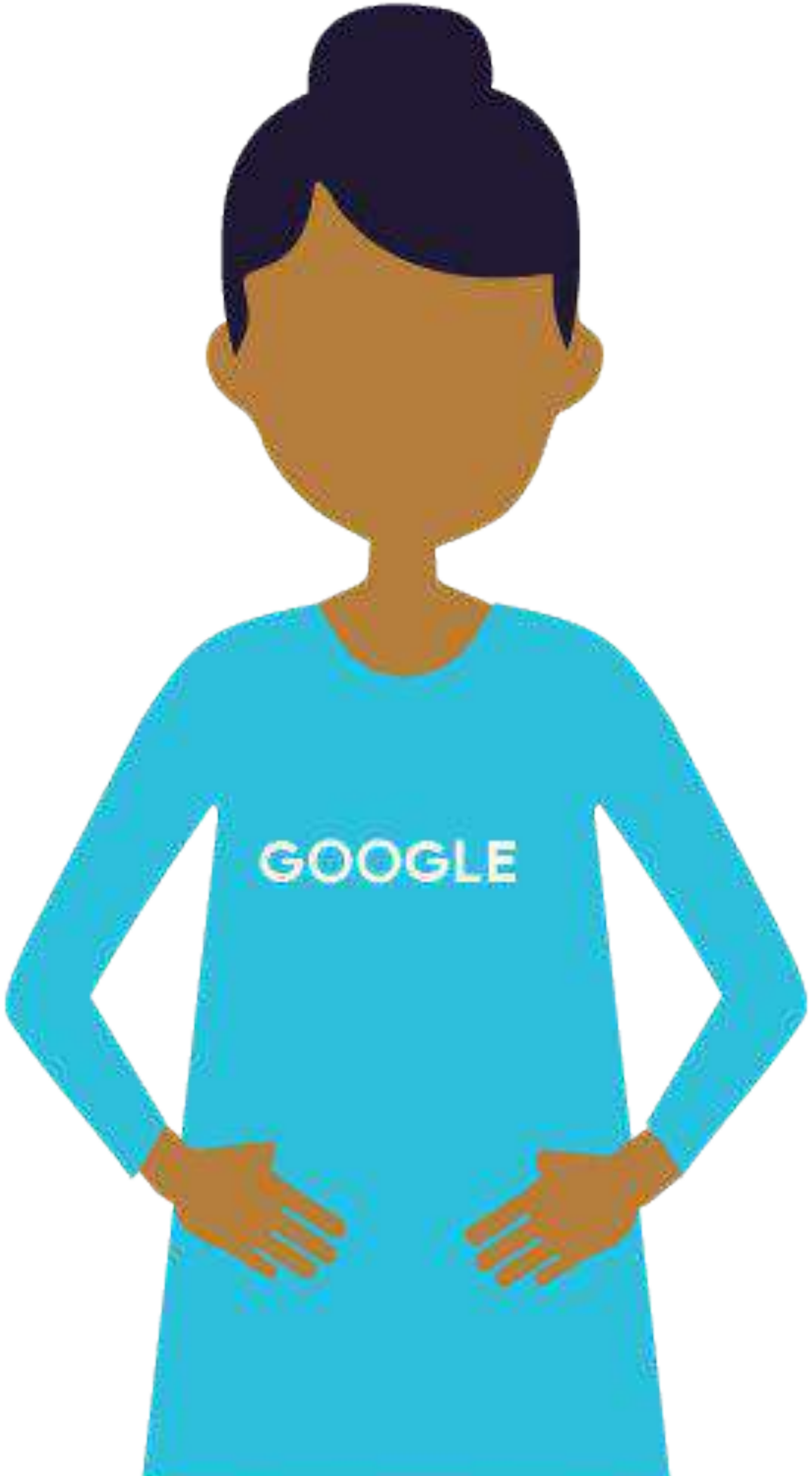


Business Catastrophe 2

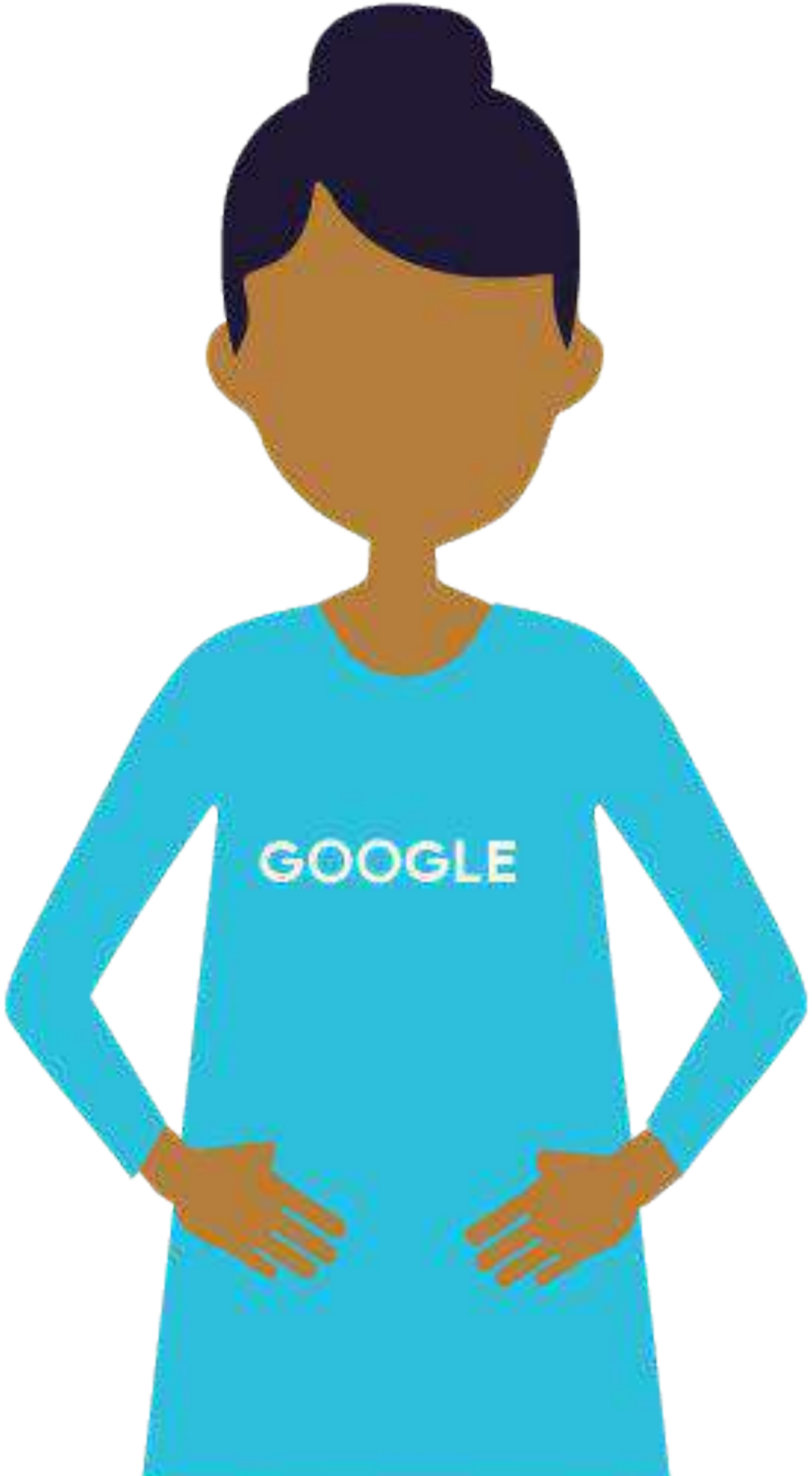


Business Catastrophe 2

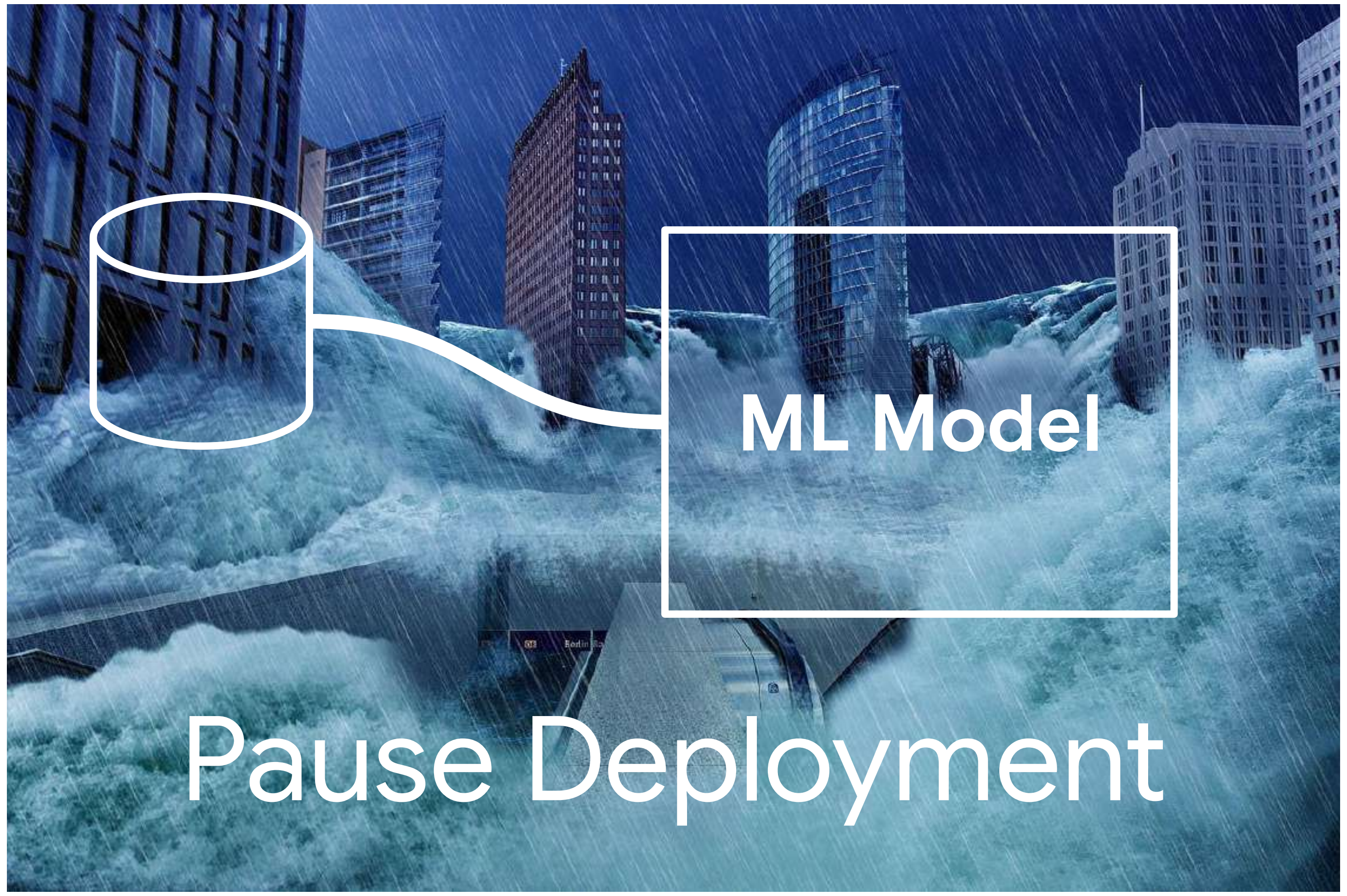




Business Catastrophe 3



Business Catastrophe 3



Course 2: Production ML Systems

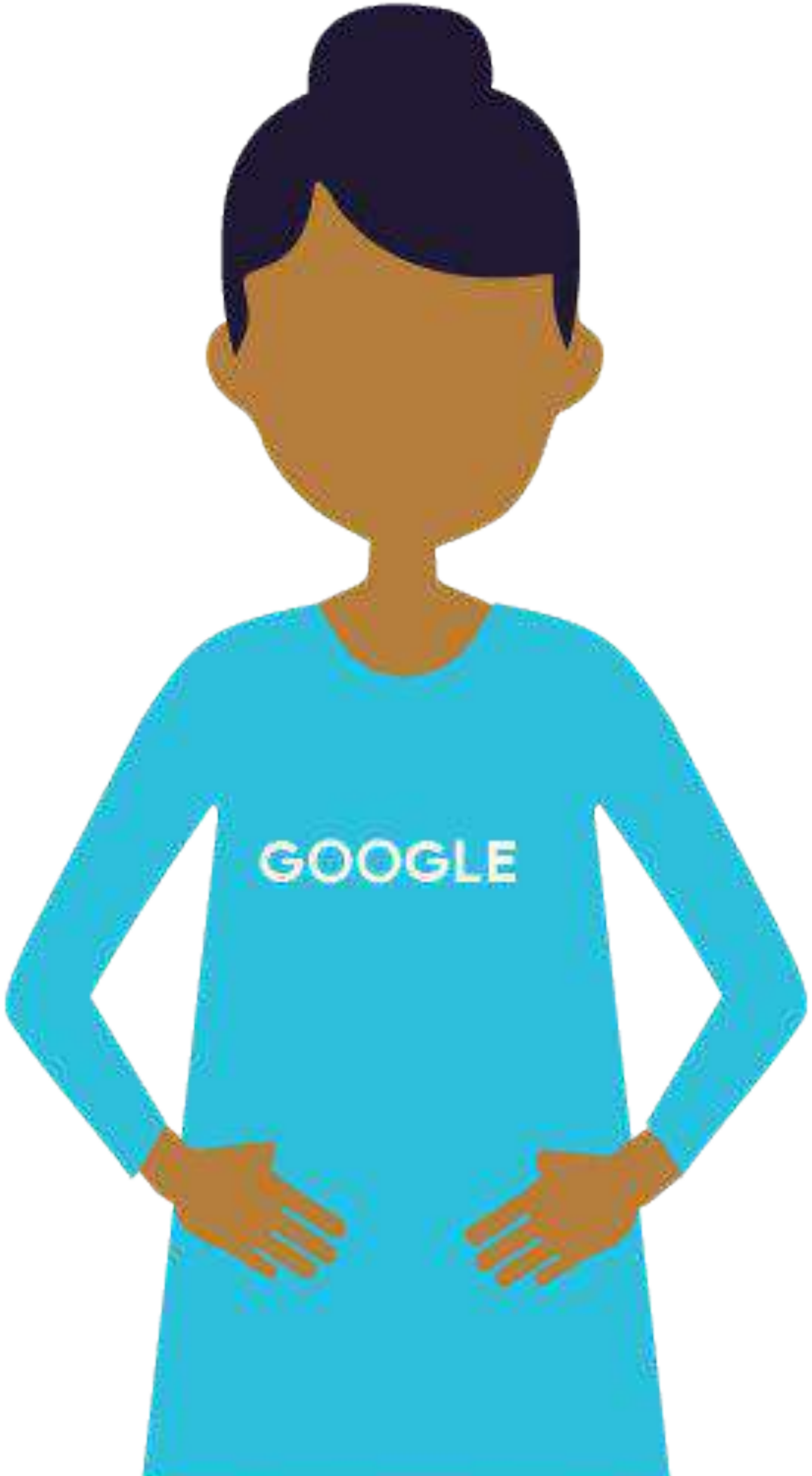
Module 3: Designing Adaptable ML Systems

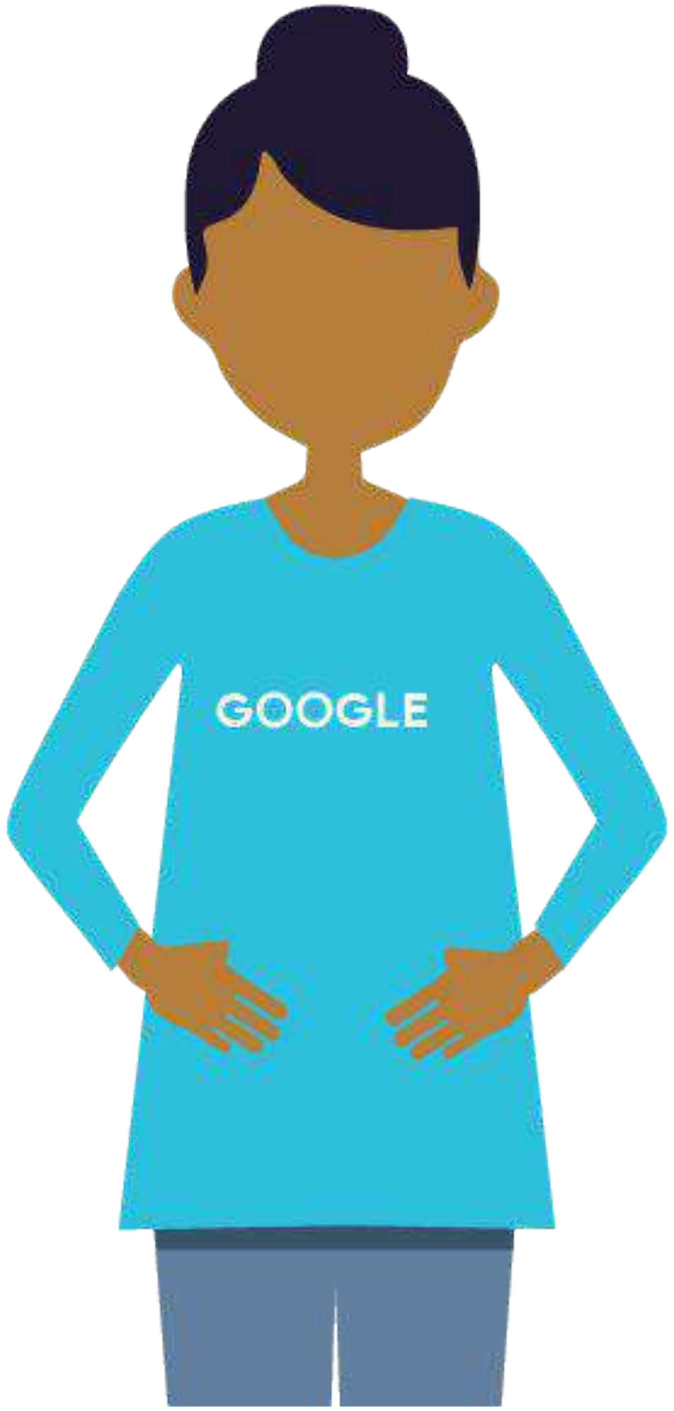
Lesson Title: **Module Summary**

Presenter: Max Lotstein

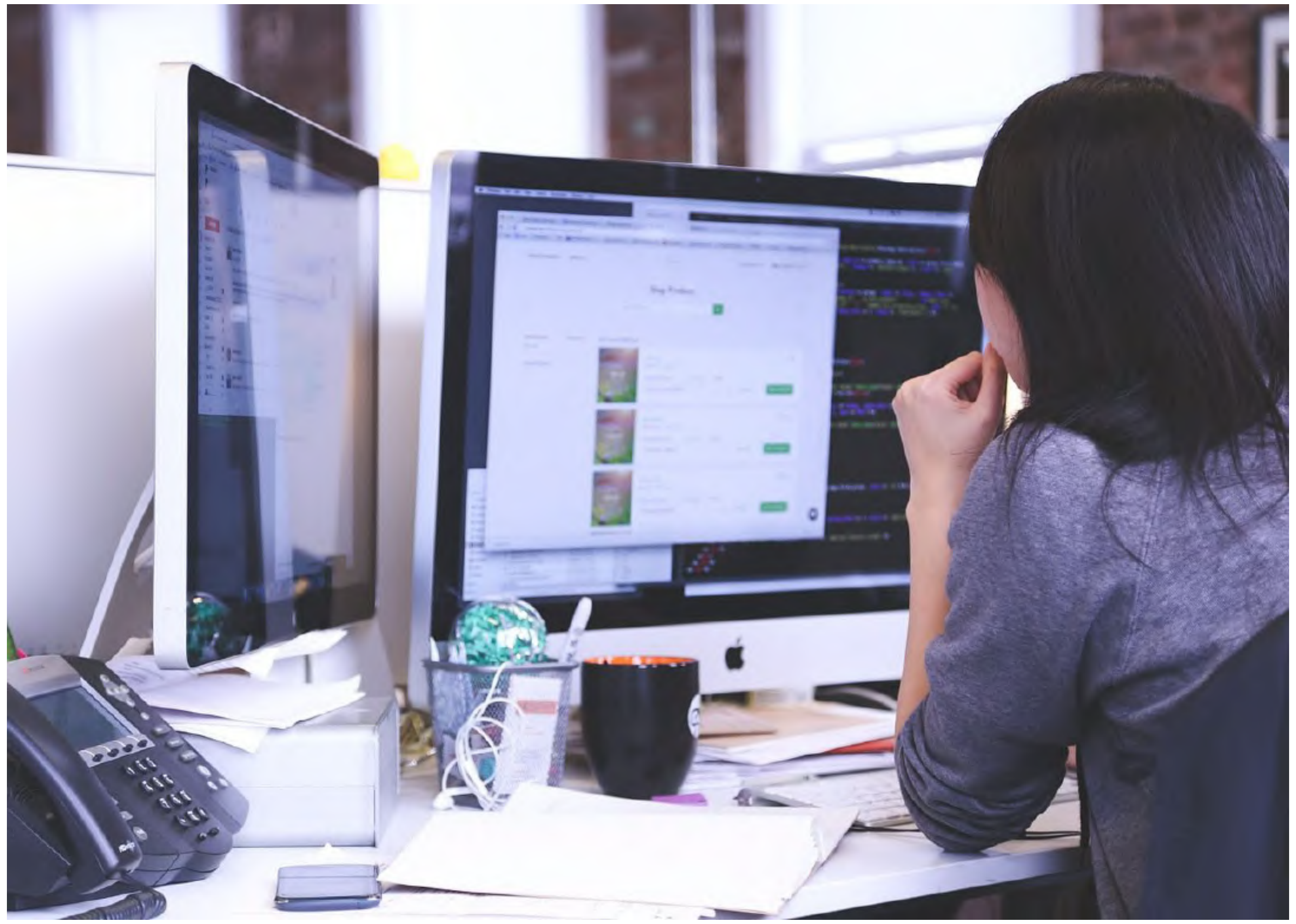
Format: Talking Head

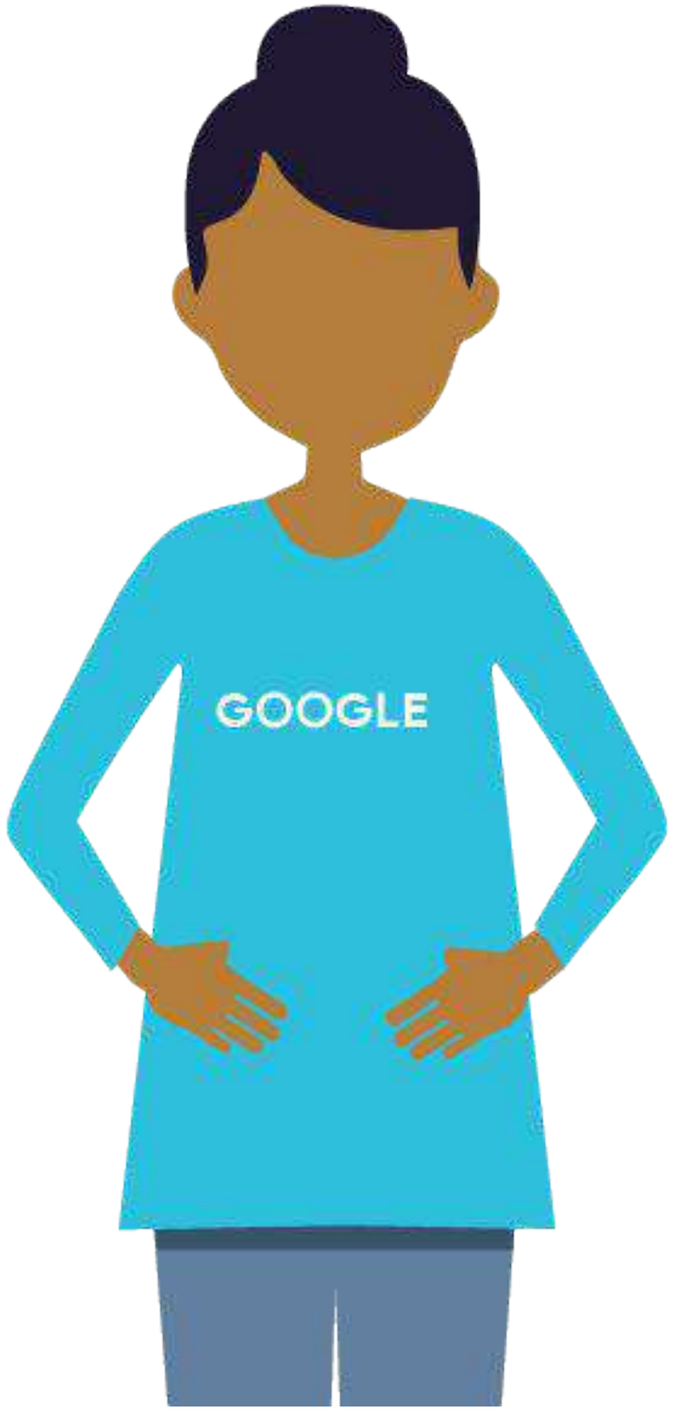
Video Name: T-PSML-O_3_l14_module_summary


GOOGLE



Keep humans in the loop





Prioritize maintainability

Get ready to roll back