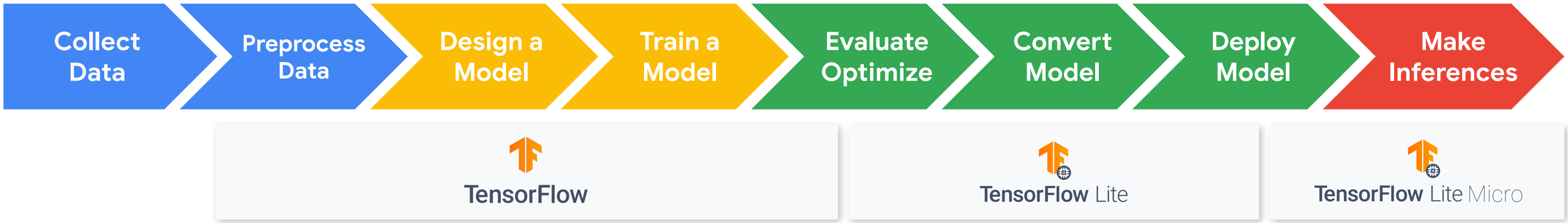
Takeaways from Course 2

Tiny Machine Learning Apps

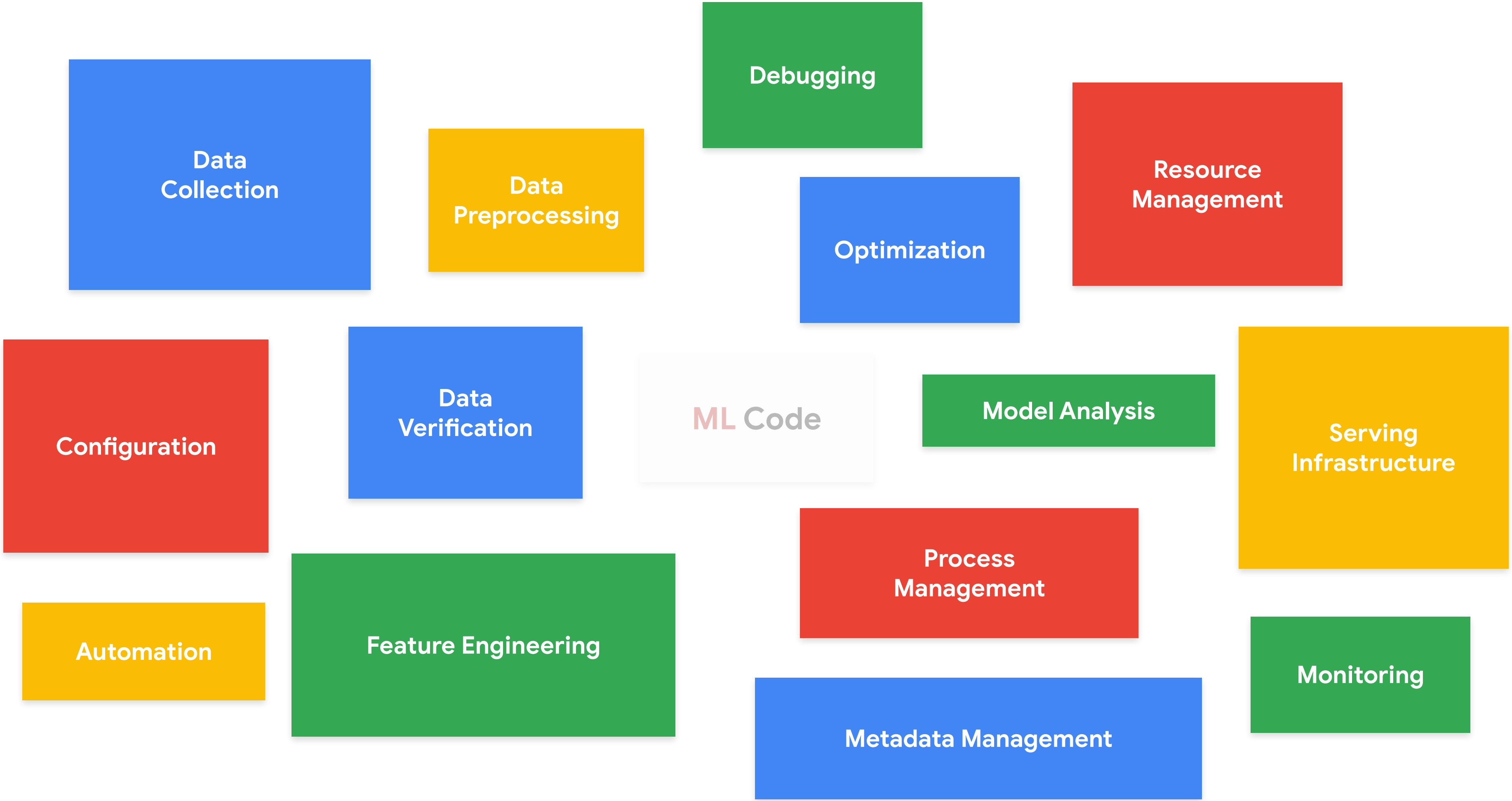
Tiny Machine Learning Apps

Collect Data → Preprocess Data → Design a Model → Train a Model → Evaluate Optimize → Convert Model → Deploy Model → Make Inferences

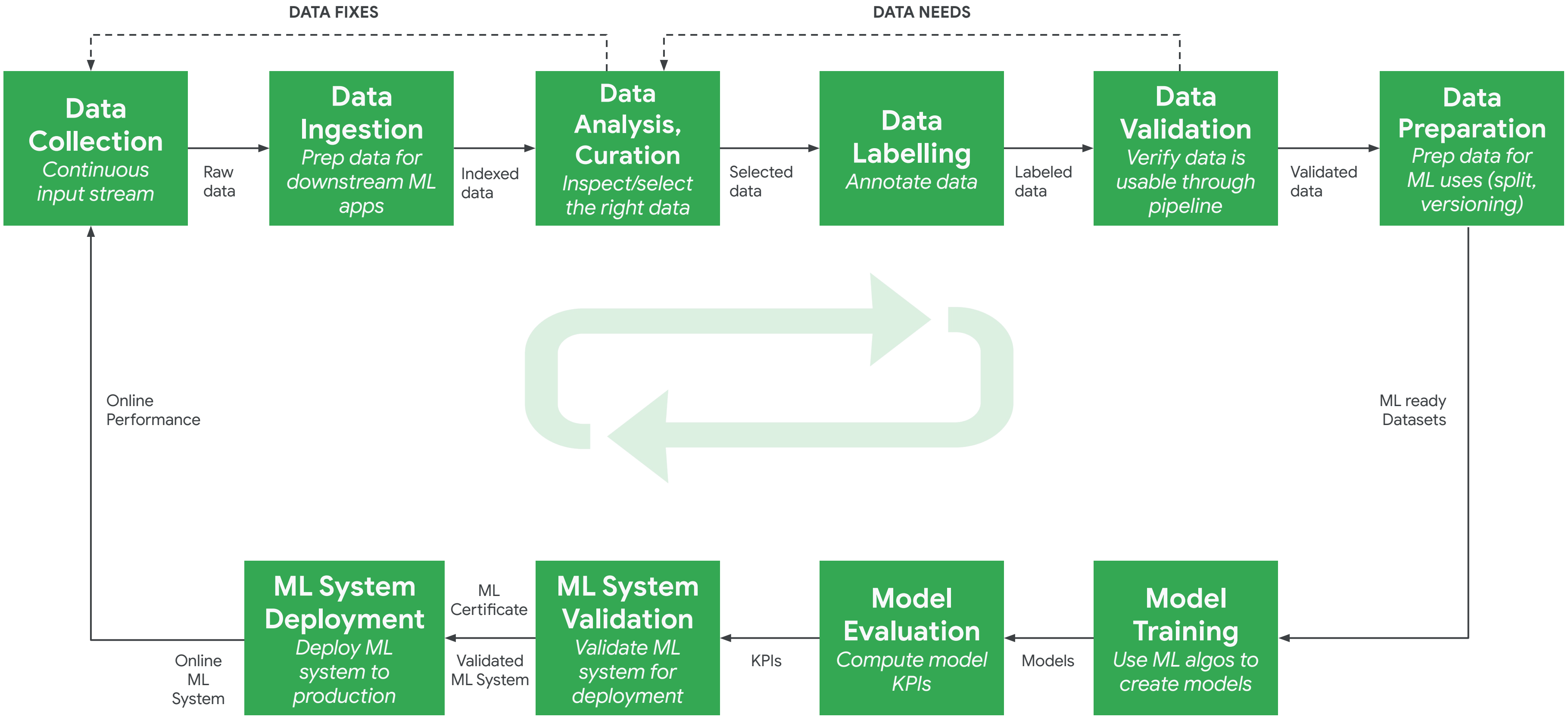
Tiny Machine Learning Apps



ML Code


Data Collection
Data Preprocessing
Debugging
Resource Management
Optimization
Configuration
Data Verification
ML Code
Model Analysis
Serving Infrastructure
Process Management
Automation
Feature Engineering
Monitoring
Metadata Management

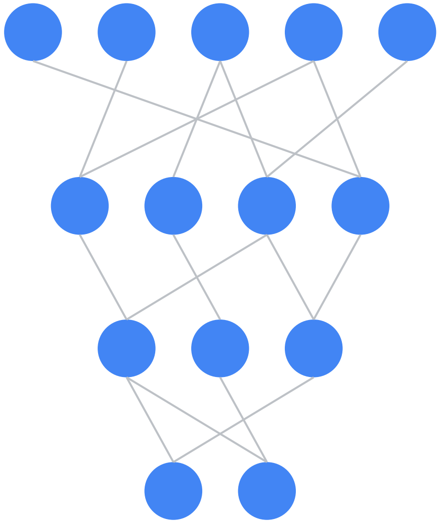
Life cycle of ML

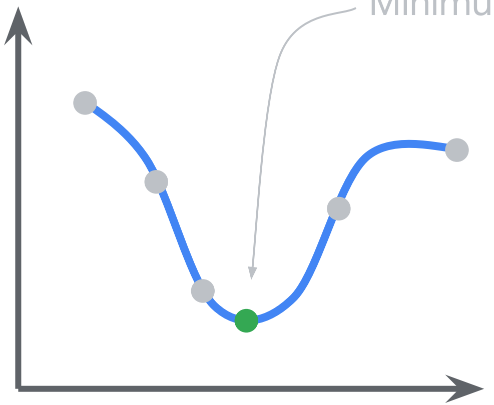


Acoustic Sensors
Ultrasonic, **Microphones**, Geophones, Vibrometers

Image Sensors
Thermal, **Image**

Motion Sensors
Gyroscope, Radar, **Accelerometer**





Course 2: End-to-end **TinyML** application design

AI Infrastructure

Data Engineering

Model Engineering

Model Deployment

Product Analytics

Collect Data → Preprocess Data → Design a Model → Train a Model → Evaluate Optimize → Convert Model → Deploy Model → Make Inferences

Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition

Pete Warden
Google Brain
Mountain View, California
petewarden@google.com

April 2018

Data Engineering

Requirements

- Problem definition
- Permissions & rights
- Machine & human usable format

Gathering

- Data sources
- People
- Collection
- Labeling

Refinement

- Processing
- Validation
- Augmentation

Sustainment

- Storage
- Security
- Errors
- Versioning

Data **Collection** and **Processing**

Someone scraped 40,000 Tinder selfies to make a facial dataset for AI experiments

Natasha Lomas @riptari / 7:21 PM EDT • April 28, 2017

Update: A Tinder spokesperson has now provided the following statement:

> We take the security and privacy of our users seriously and have tools and systems in place to uphold the integrity of our platform. It's important to note that Tinder is free and used in more than 190 countries, and the images that we serve are profile images, which are available to anyone swiping on the app. We are always working to improve the Tinder experience and continue to implement measures against the automated use of our API, which includes steps to deter and prevent scraping.

> This person has violated our terms of service (Sec. 11) and we are taking appropriate action and investigating further.

Recall: **Don't *collect*** from scratch

Data collection is **difficult**!

- Can we ***reuse*** existing data?

What's available?

What's missing?

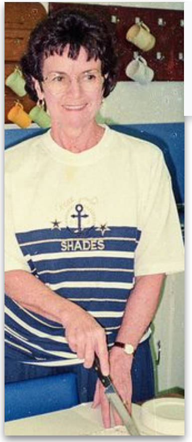
Visual Wake Words **Dataset**

Visual Wake Words Dataset

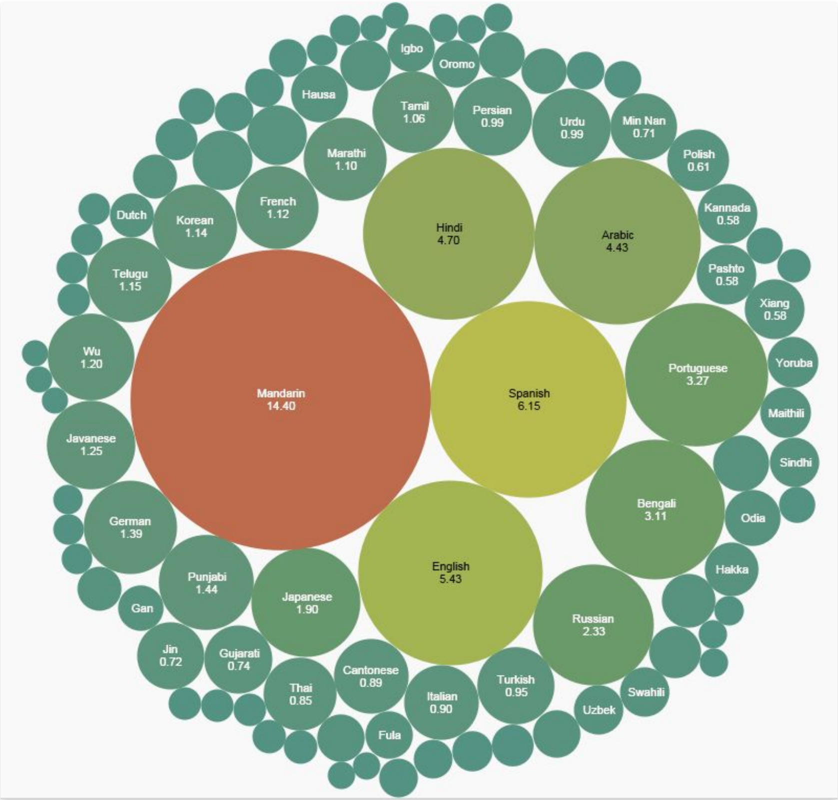
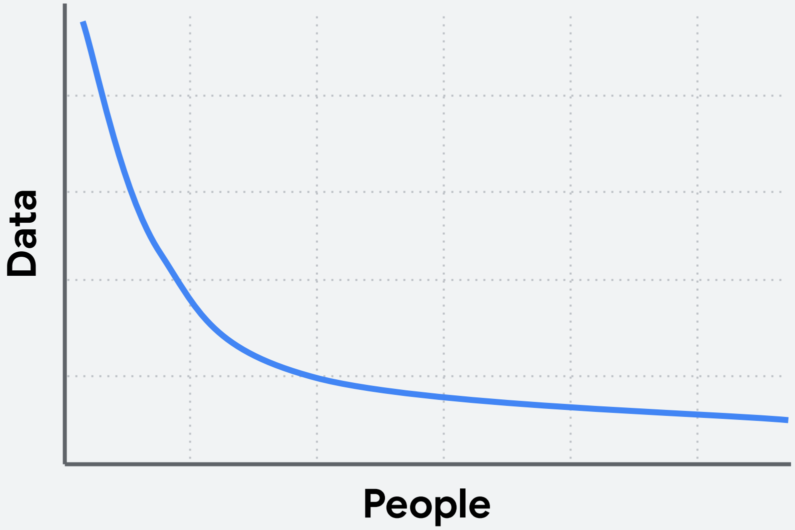
Aakanksha Chowdhery, Pete Warden, Jonathon Shlens,
Andrew Howard, Rocky Rhodes
Google Research
{chowdhery, petewarden, shlens, howarda, rocky}@google.com

Visual Wake Words **Dataset**

***Label:* "person"**

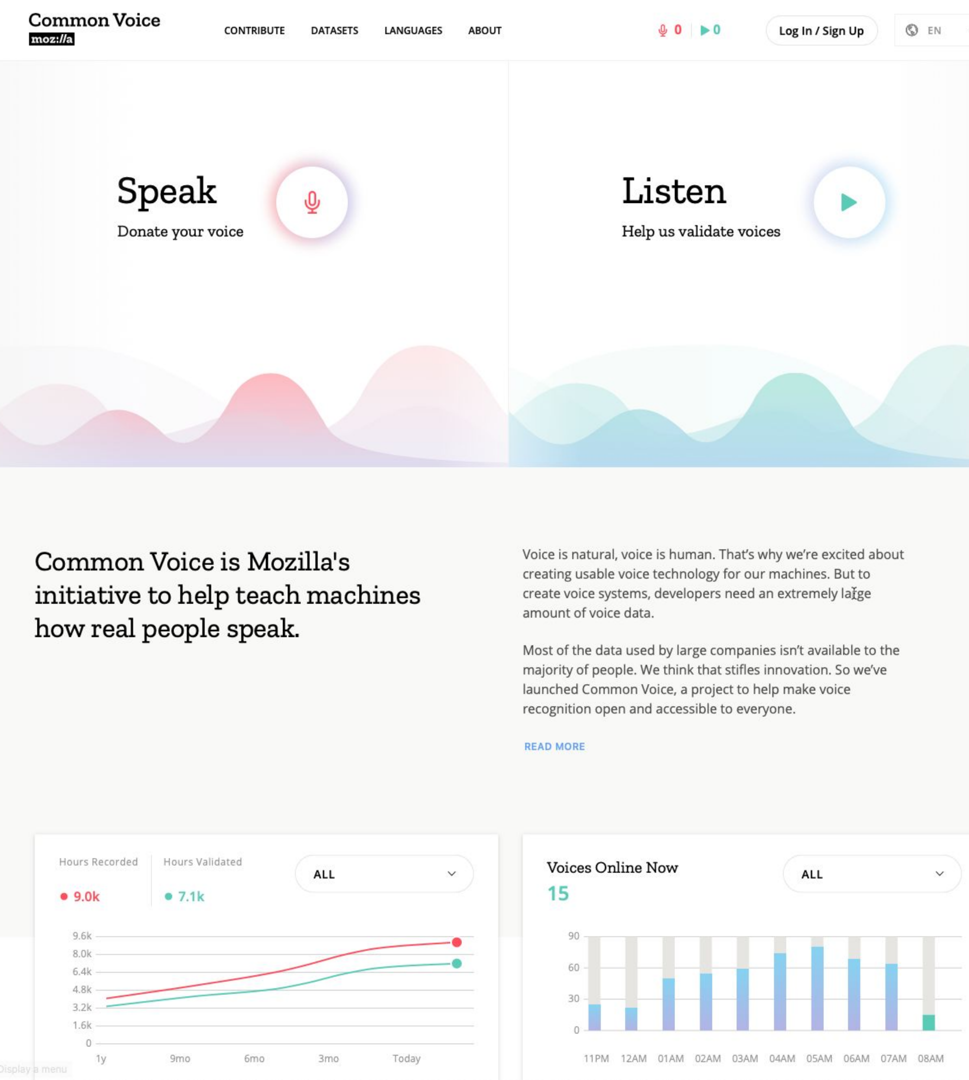


***Label:* "person"**

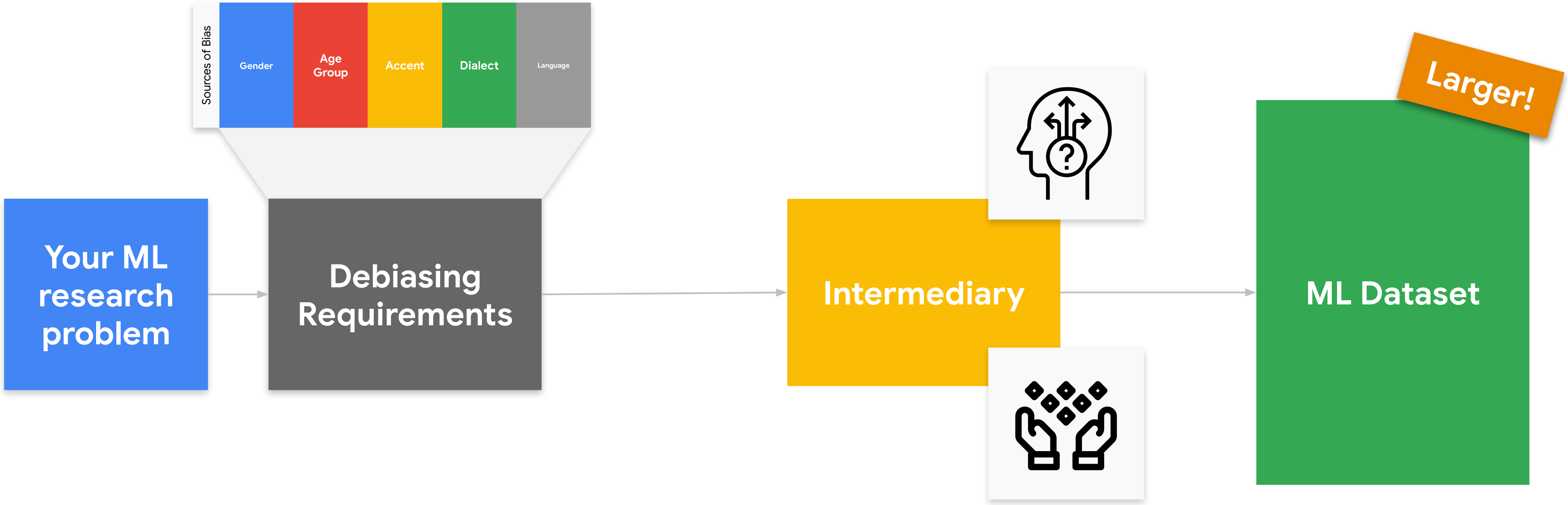

Igbo
Oromo
Hausa
Tamil 1.06
Persian 0.99
Urdu 0.99
Min Nan 0.71
Polish 0.61
Marathi 1.10
French 1.12
Dutch
Korean 1.14
Hindi 4.70
Arabic 4.43
Kannada 0.58
Pashto 0.58
Xiang 0.58
Telugu 1.15
Wu 1.20
Mandarin 14.40
Spanish 6.15
Portuguese 3.27
Yoruba
Maithili
Javanese 1.25
Sindhi
Bengali 3.11
Odia
German 1.39
English 5.43
Hakka
Punjabi 1.44
Japanese 1.90
Gan
Russian 2.33
Jin 0.72
Gujarati 0.74
Thai 0.85
Cantonese 0.89
Italian 0.90
Turkish 0.95
Swahili
Uzbek
Fula

Data
People

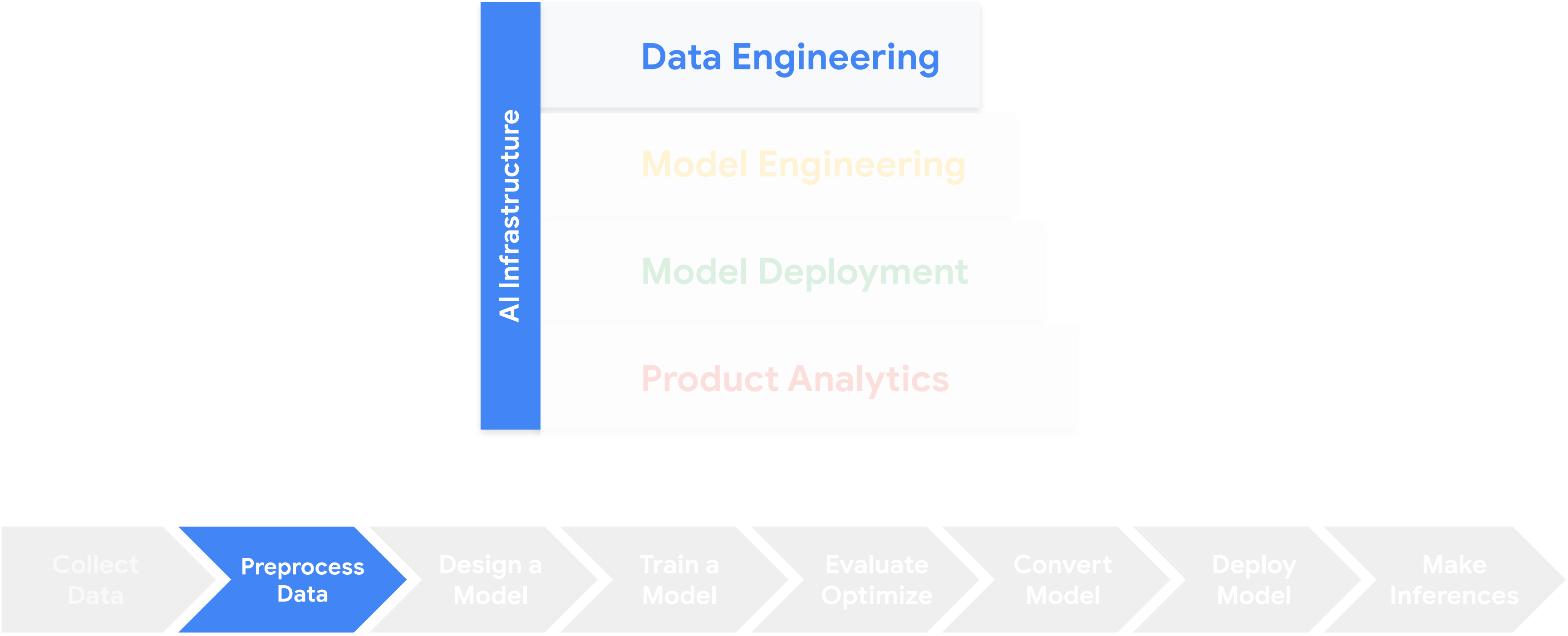
Common Voice

- Crowdsourcing platform
- Over **50,000 volunteers**



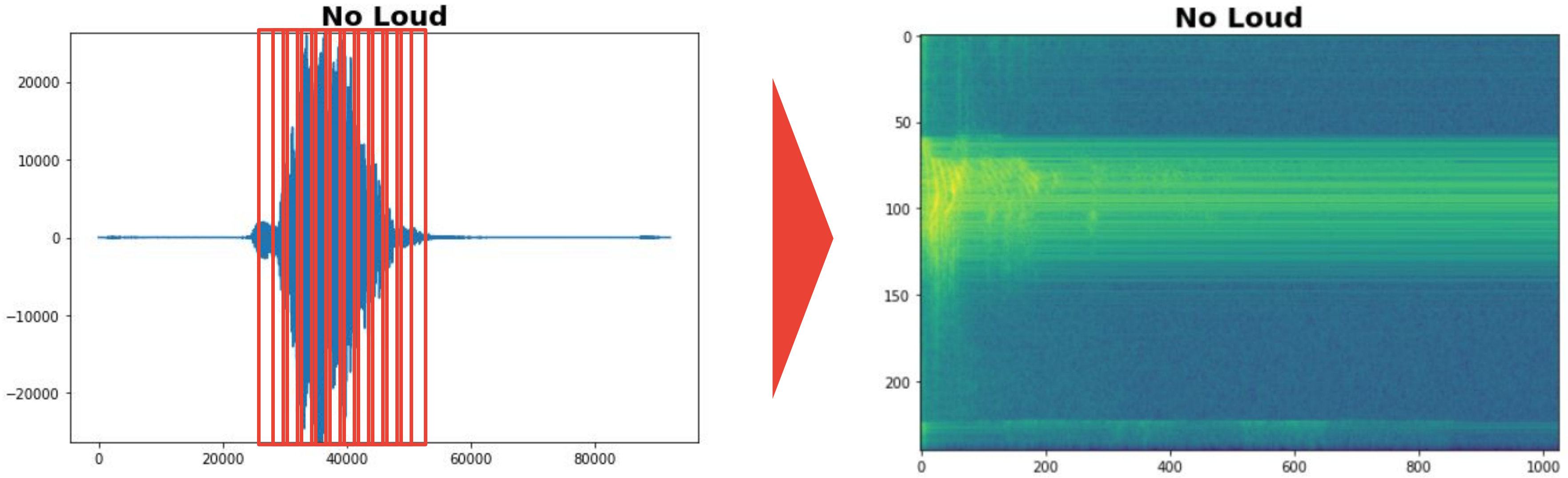
Bias and **Market Forces**

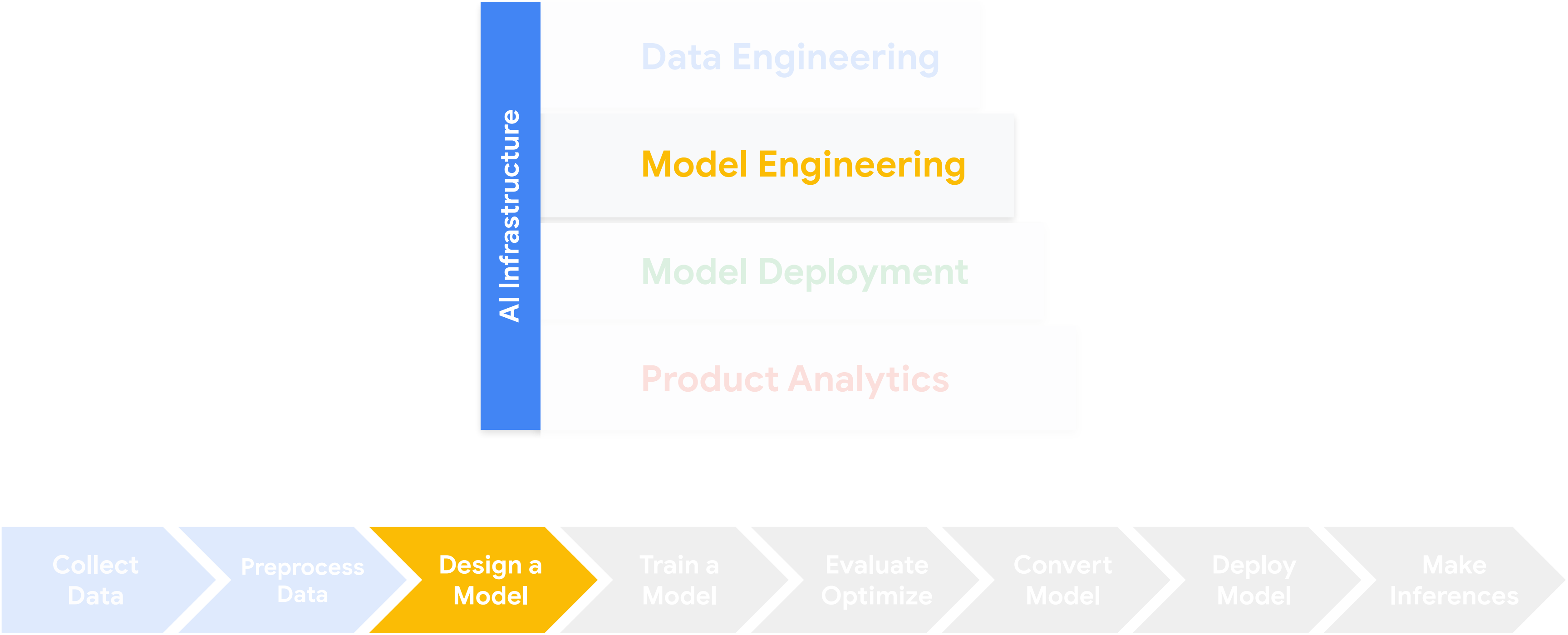




AI Infrastructure
Data Engineering
Model Engineering
Model Deployment
Product Analytics
Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences

Data Preprocessing: **Spectrograms**

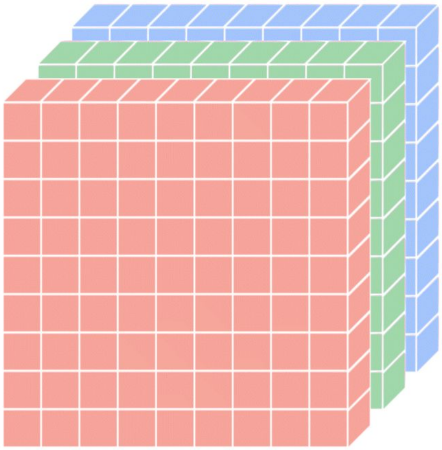


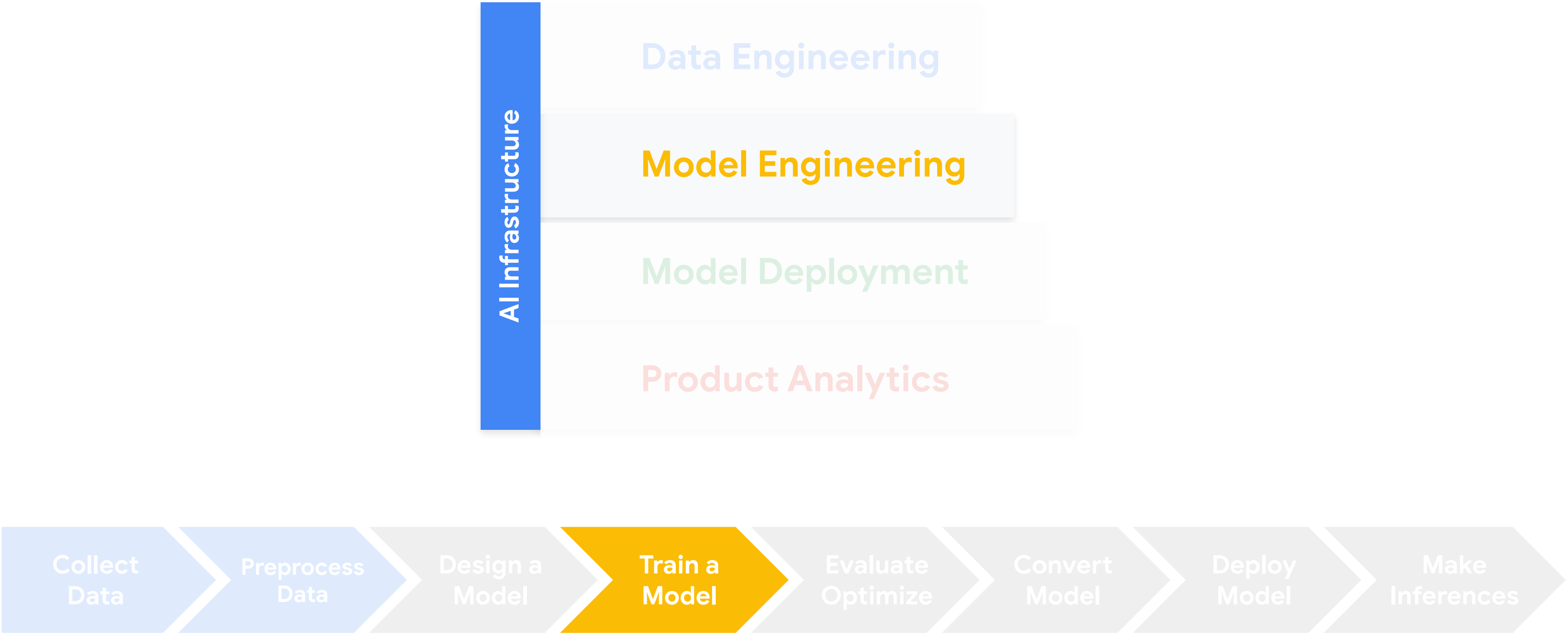

AI Infrastructure
Data Engineering
Model Engineering
Model Deployment
Product Analytics
Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences

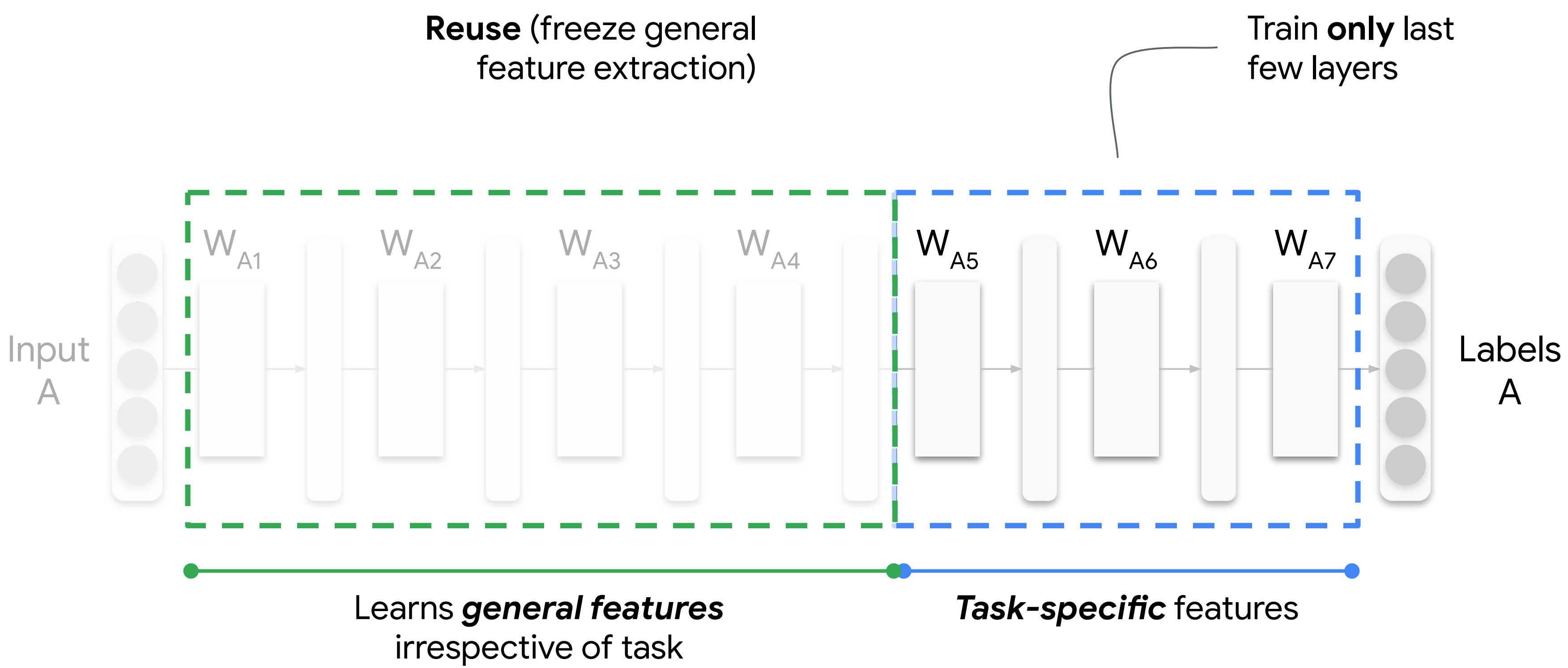
separable

Depthwise Convolution **(*3 Channel*—*e.g., RGB*)**

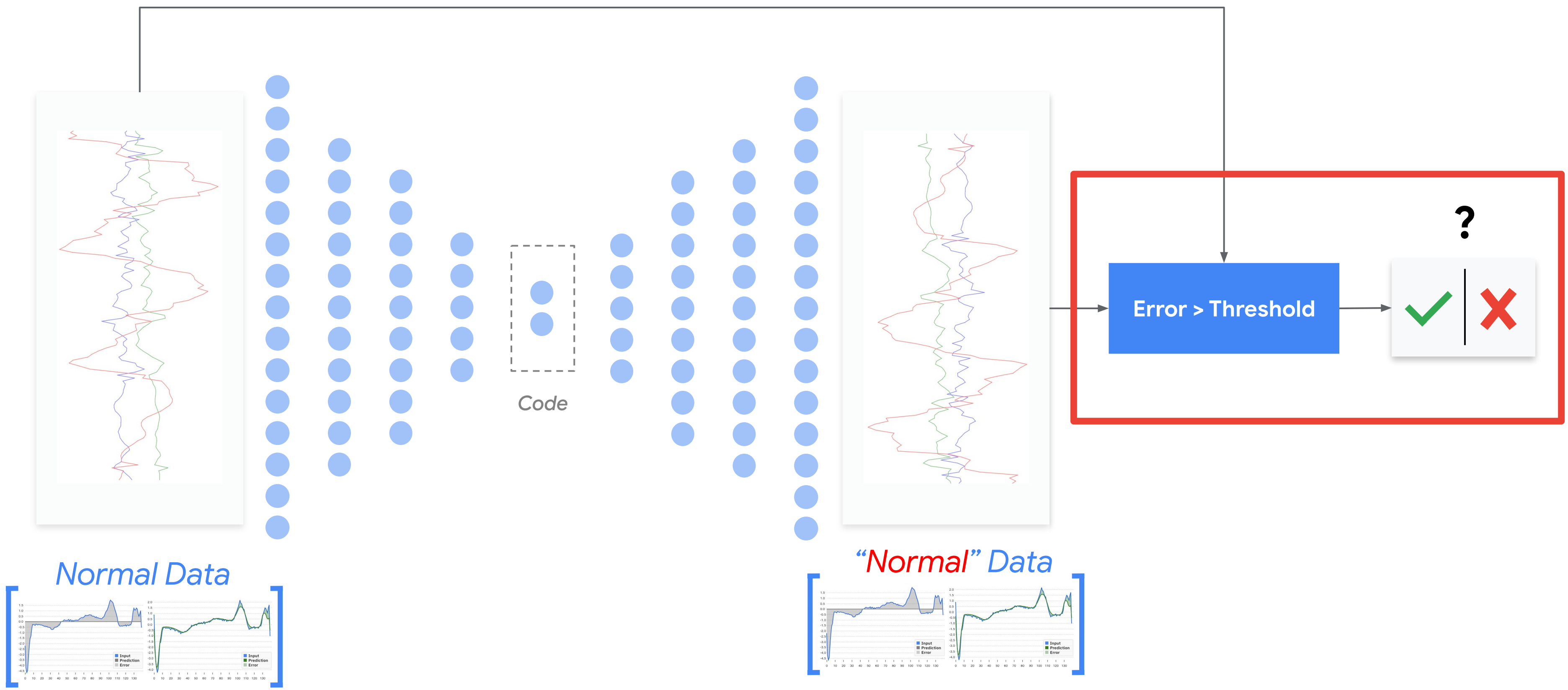
includes **pointwise** conv

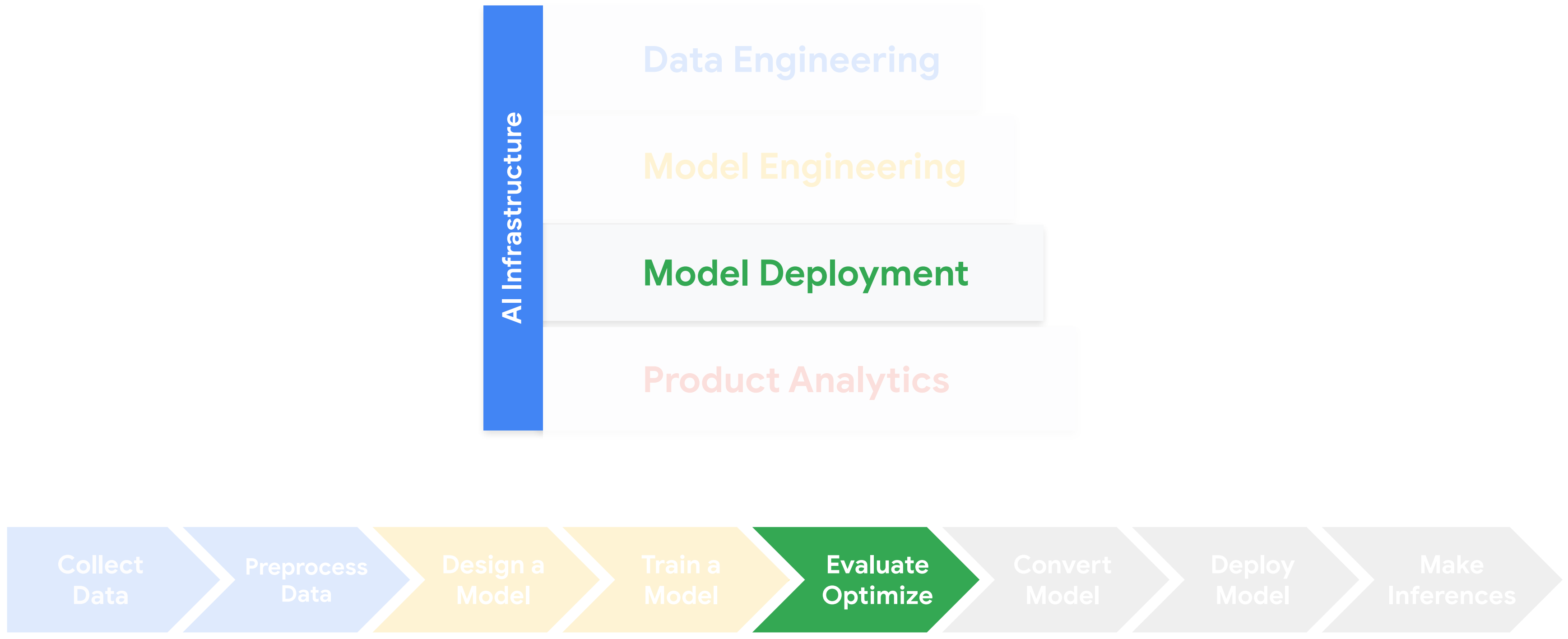


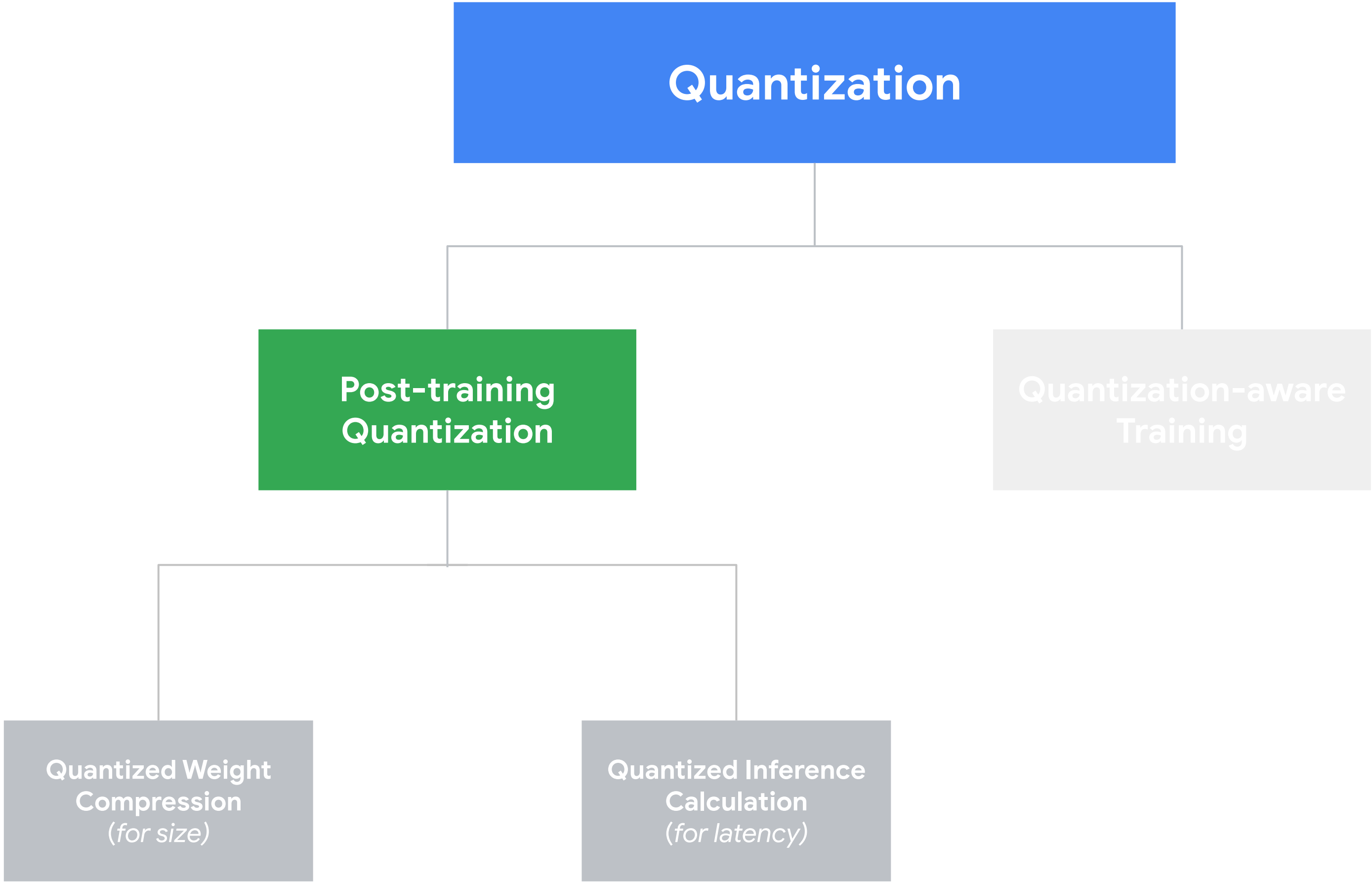

AI Infrastructure
Data Engineering
Model Engineering
Model Deployment
Product Analytics
Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences

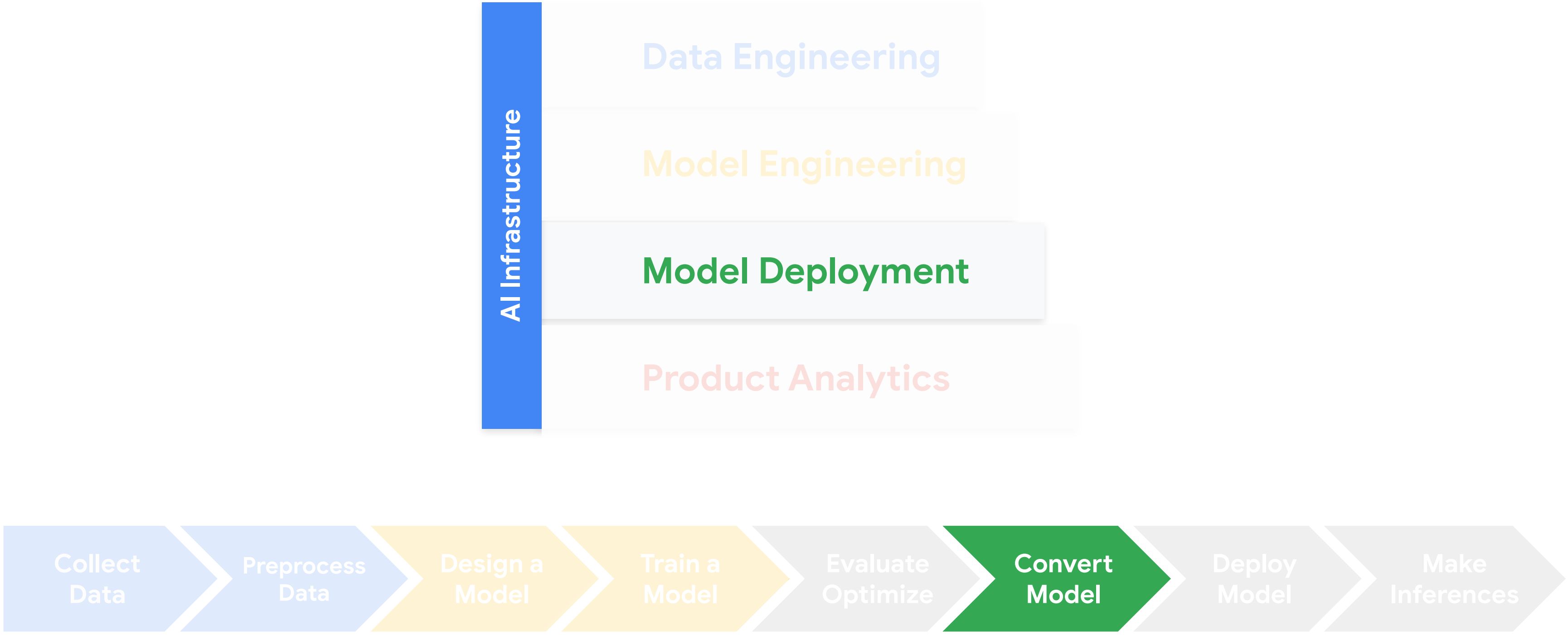


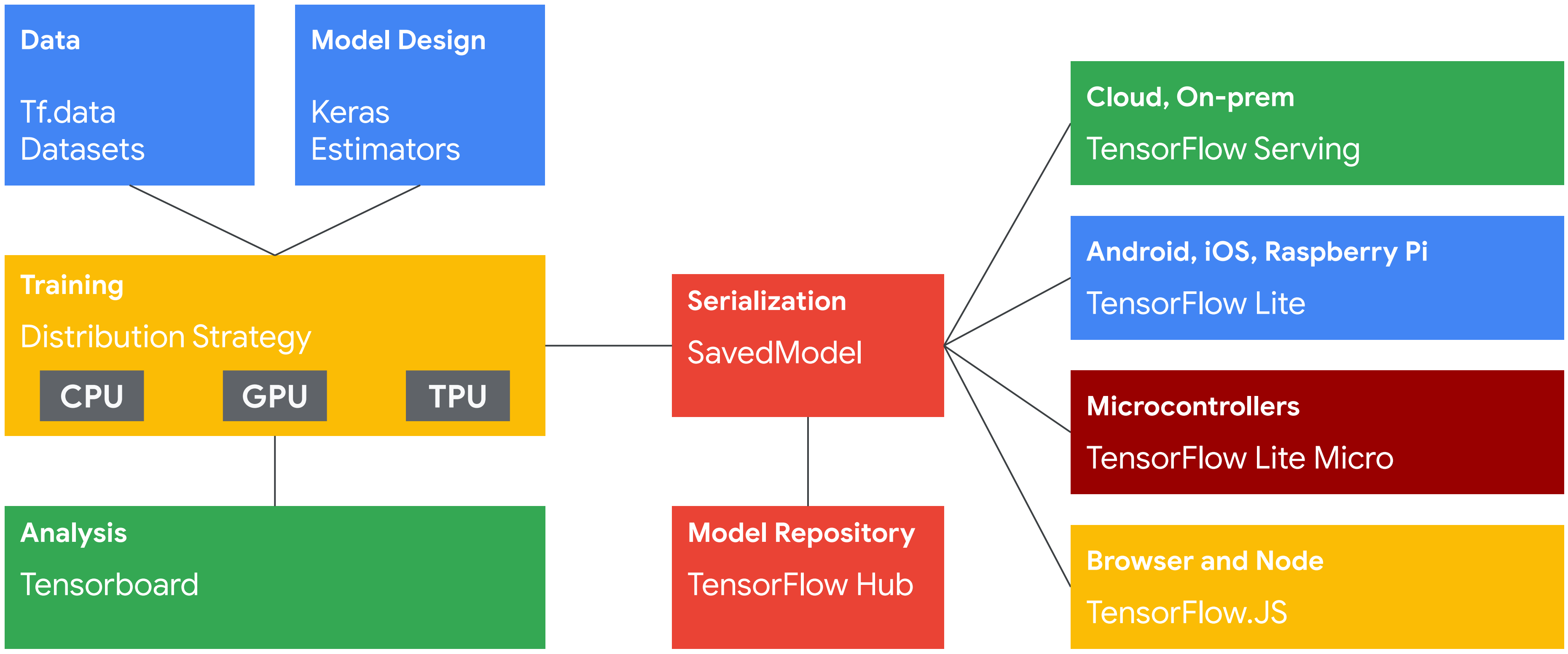
Reuse (freeze general feature extraction)
Train **only** last few layers
Input A
W_{A1}
W_{A2}
W_{A3}
W_{A4}
W_{A5}
W_{A6}
W_{A7}
Labels A
Learns ***general features*** irrespective of task
Task-specific features

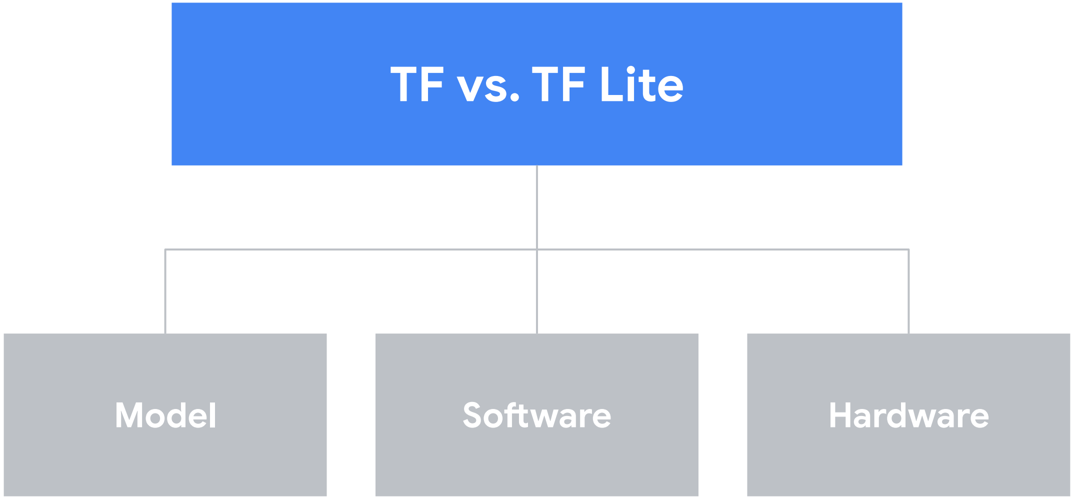

Code
Error > Threshold
?
Normal Data
"Normal" Data

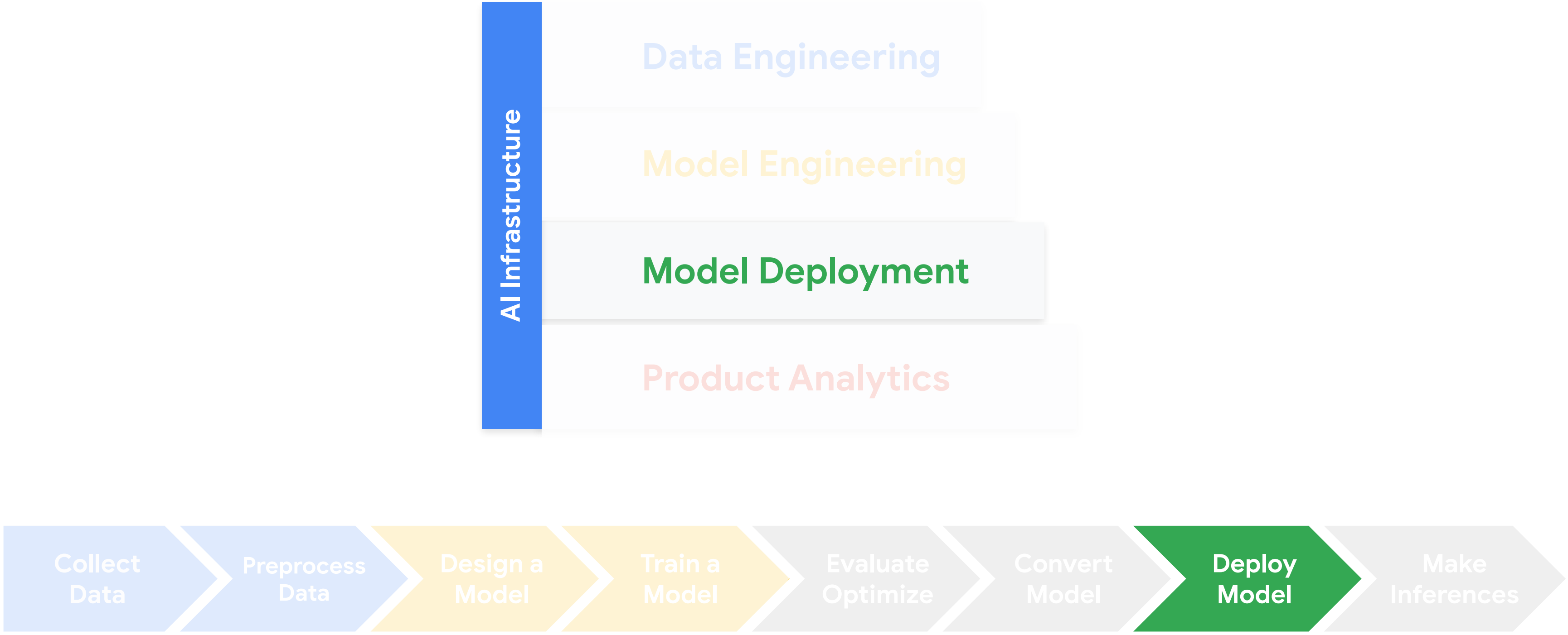

AI Infrastructure
Data Engineering
Model Engineering
Model Deployment
Product Analytics
Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences


Quantization
Post-training Quantization
Quantization-aware Training
Quantized Weight Compression
(for size)
Quantized Inference Calculation
(for latency)


AI Infrastructure
Data Engineering
Model Engineering
Model Deployment
Product Analytics
Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences


Data
Tf.data
Datasets
Model Design
Keras
Estimators
Training
Distribution Strategy
CPU
GPU
TPU
Analysis
Tensorboard
Serialization
SavedModel
Model Repository
TensorFlow Hub
Cloud, On-prem
TensorFlow Serving
Android, iOS, Raspberry Pi
TensorFlow Lite
Microcontrollers
TensorFlow Lite Micro
Browser and Node
TensorFlow.JS


TF vs. TF Lite
Model
Software
Hardware


AI Infrastructure
Data Engineering
Model Engineering
Model Deployment
Product Analytics
Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences

Common Metrics

Accuracy

Quantitative

Efficiency

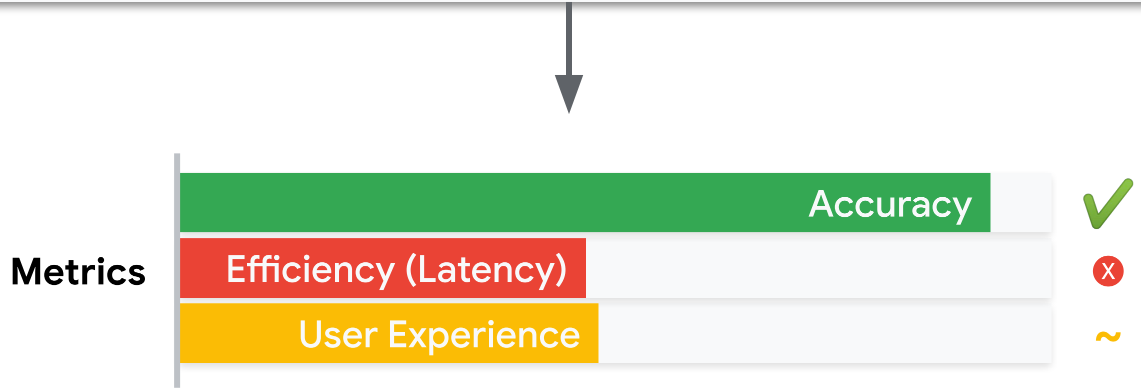
Quantitative

User Experience

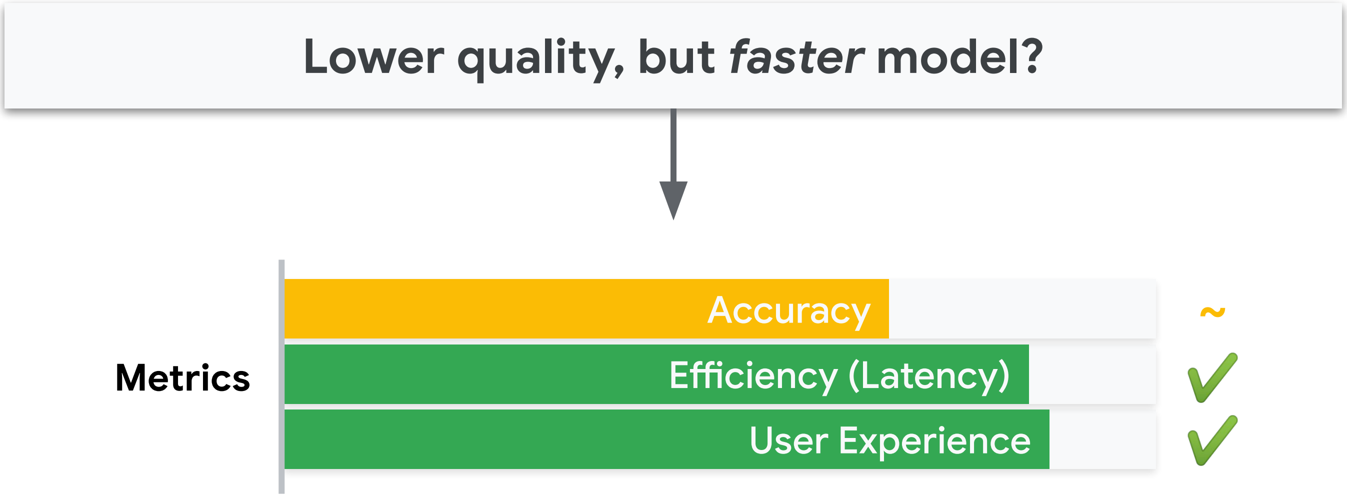
Qualitative

Latency

Accurate but *SLOW* model?



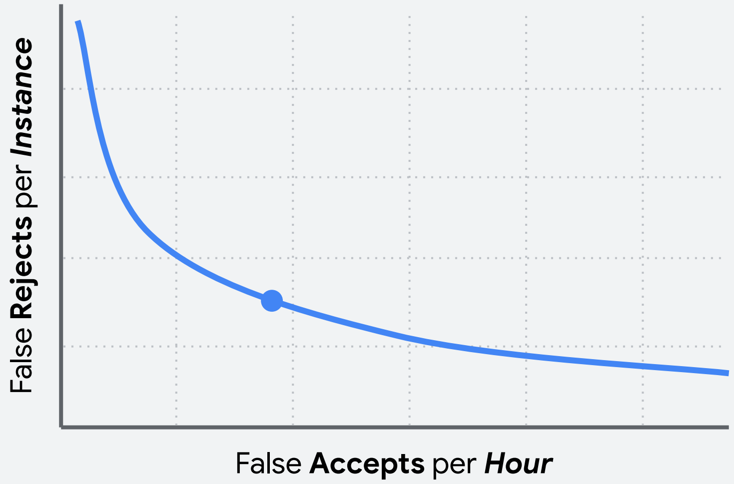
Latency



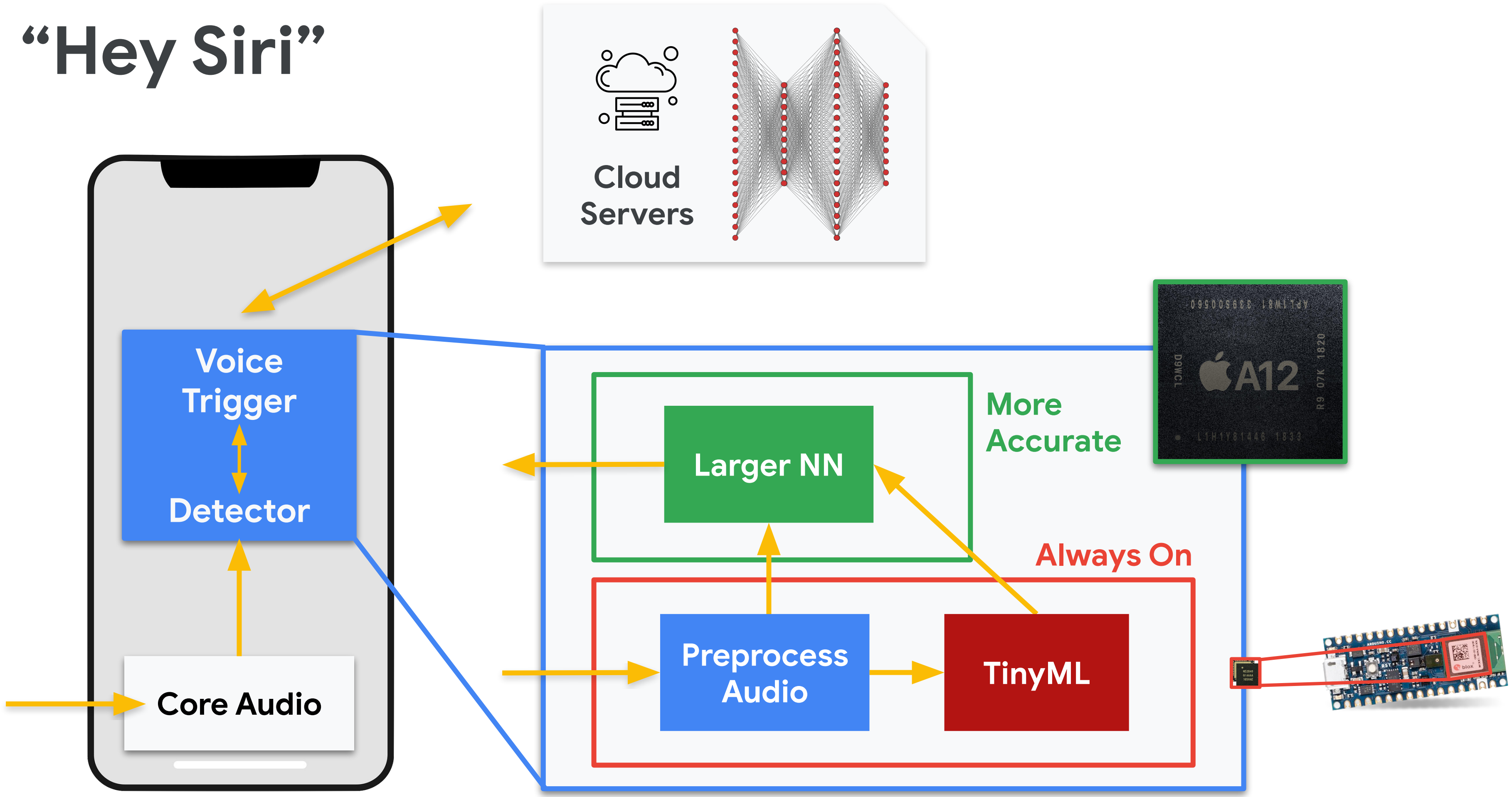
Lower quality, but *faster* model?
Metrics
Accuracy
~
Efficiency (Latency)
User Experience

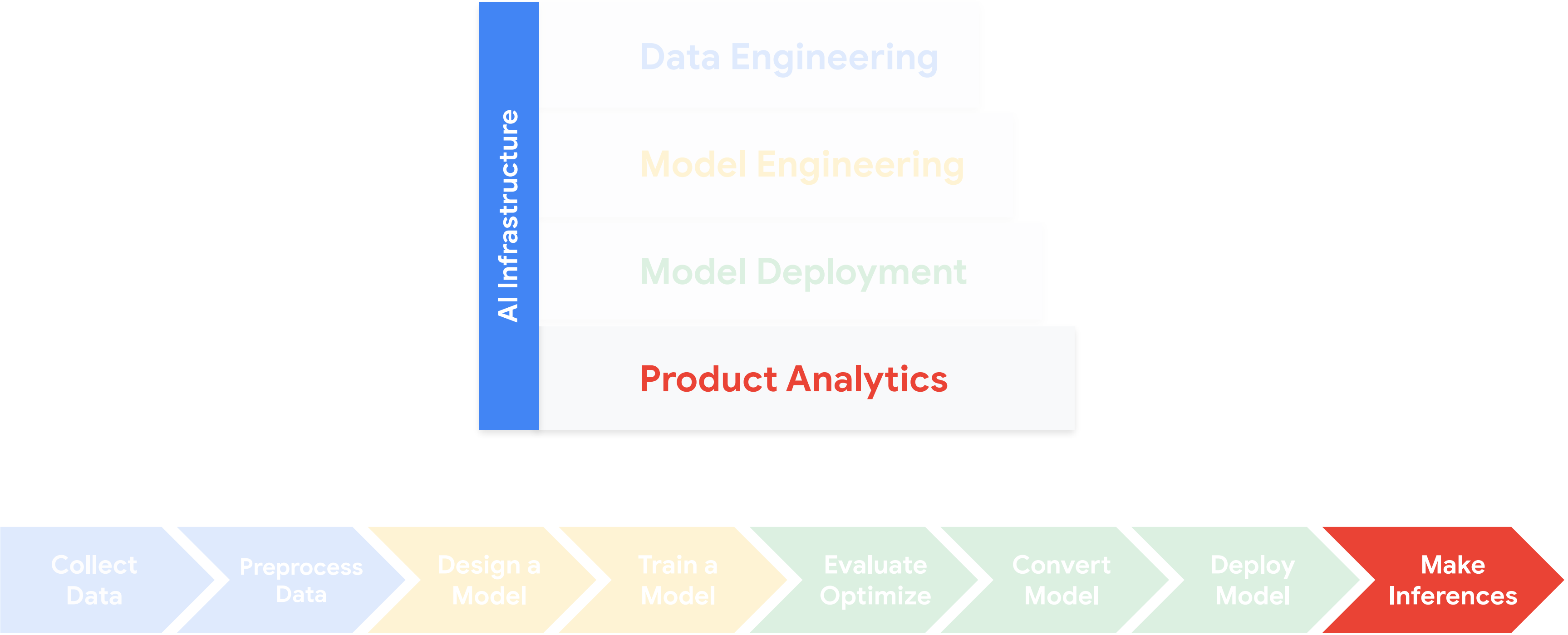
False Positive and **False Negative**

- Accuracy is measured as a tradeoff between **false accept rate** (FAR) and **false reject rate** (FRR)



“Hey Siri”


Cloud
Servers
Voice
Trigger
Detector
Core Audio
Larger NN
More
Accurate
Always On
Preprocess
Audio
TinyML
A12

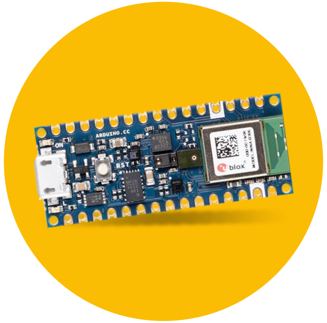

AI Infrastructure
Data Engineering
Model Engineering
Model Deployment
Product Analytics
Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences

Course Sequence

Course 1
Fundamentals of TinyML

Course 2
Applications of TinyML

Course 3
Deploying TinyML



Learning

You will learn how to deploy models on a real microcontroller. Along the way you will explore the challenges unique to and amplified by **TinyML** (e.g., preprocessing, post-processing, dealing with resource constraints).