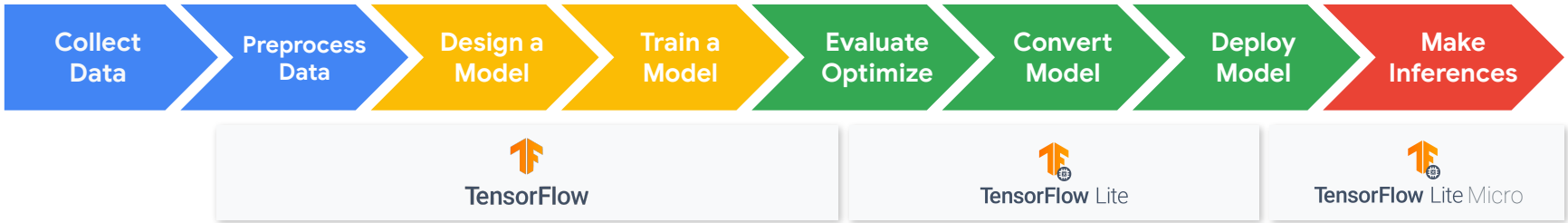
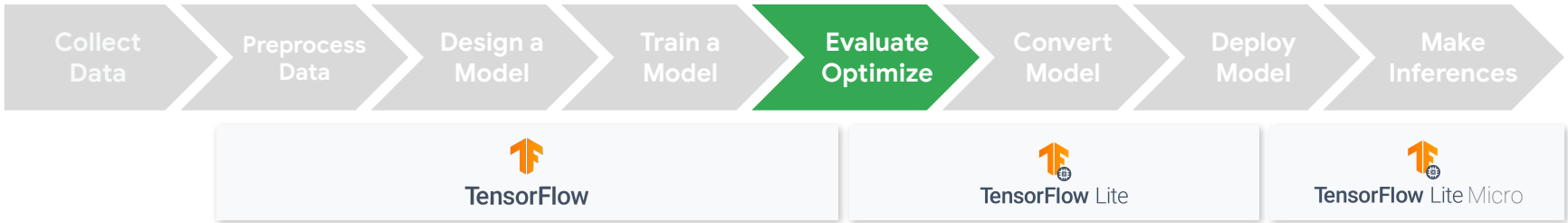
Metrics for KWS


Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro


Collect Data
Preprocess Data
Design a Model
Train a Model
Evaluate Optimize
Convert Model
Deploy Model
Make Inferences
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro

What **metrics** matter?

Accuracy

Efficiency

Beyond Model Metrics

What **metrics** matter?

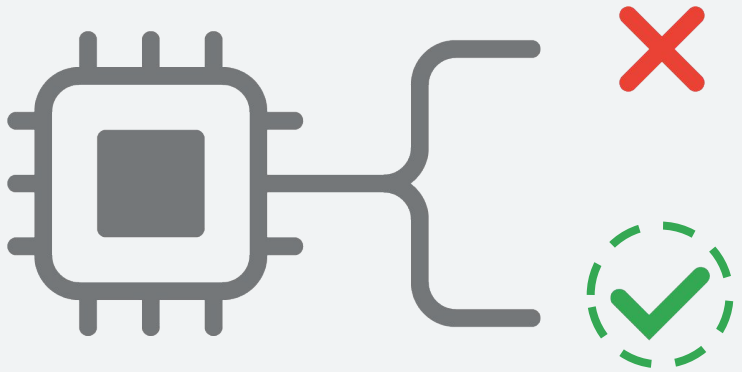
Accuracy

Efficiency

Beyond Model Metrics

False Positive

Did **NOT** say keyword but device **DOES** trigger



REAL

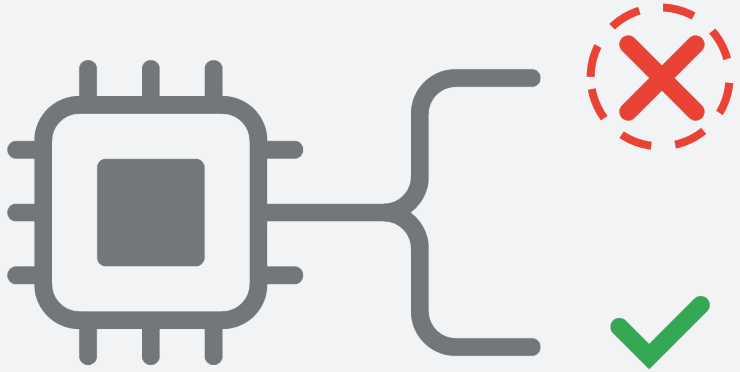
ERROR

False Negative

DID say keyword

but device **DOES**

NOT trigger

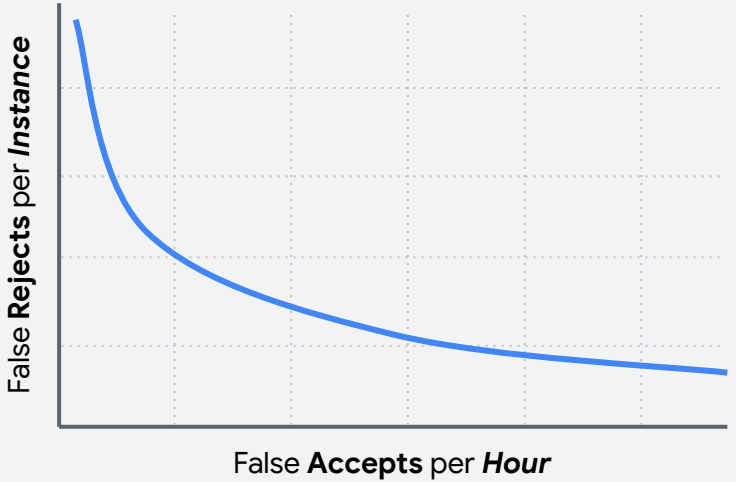


REAL	✓

ERROR

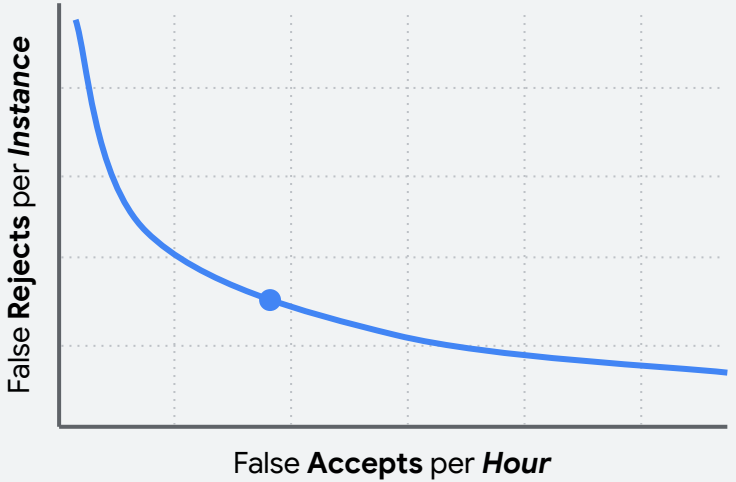
False Positive and **False Negative**

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



False Positive and **False Negative**

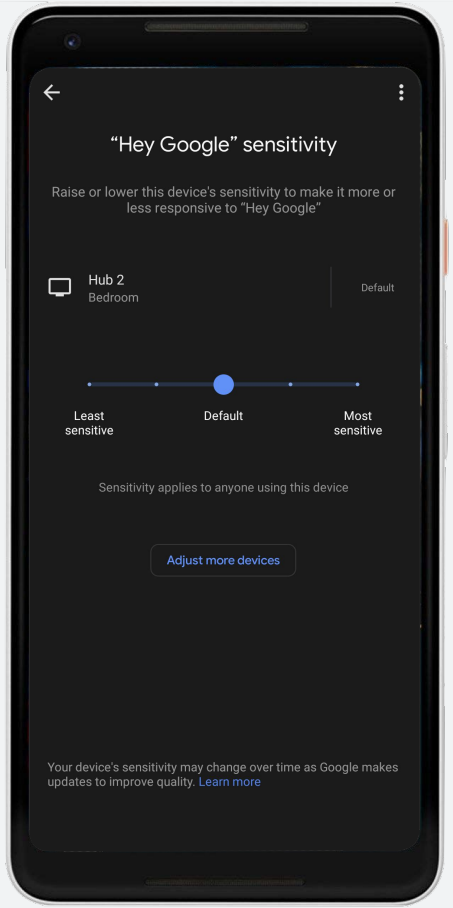
- Accuracy is measured as a tradeoff between **false accept rate** (FAR) and **false reject rate** (FRR)
- Choose an **operating point**

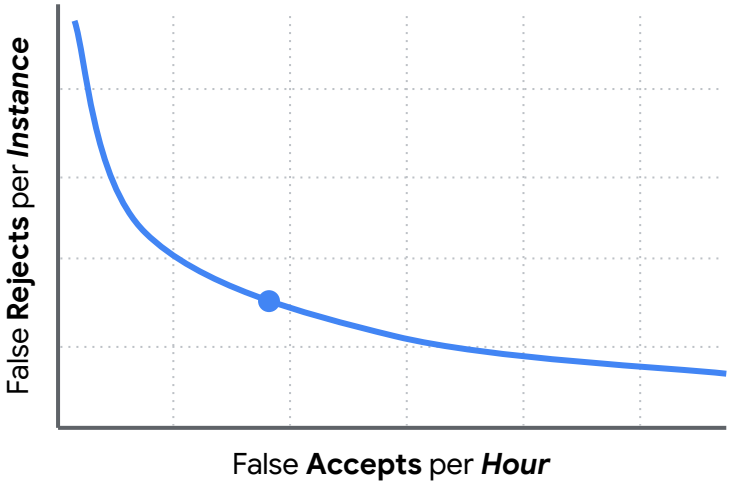


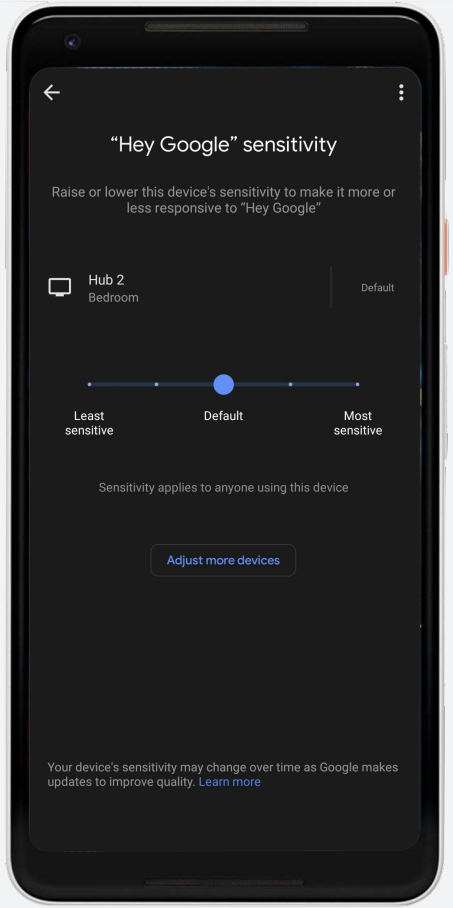
Operating Point

False Accept Rate and ***False Reject Rate*** are measured on **particular audio**—however

- Your phone might be in your pocket, purse, or backpack most of the day
- Your smart speaker might be next to a TV, or next to where your family eats, or in a relatively quiet bedroom
- You might not be a native speaker of English




False Rejects per Instance
False Accepts per Hour


"Hey Google" sensitivity
Raise or lower this device's sensitivity to make it more or less responsive to "Hey Google"
Hub 2
Bedroom
Default
Least sensitive
Default
Most sensitive
Sensitivity applies to anyone using this device
Adjust more devices
Your device's sensitivity may change over time as Google makes updates to improve quality. Learn more

Latency

- Model must be f**ast enough** to keep up with the speech input
- The model must run fast enough to be **responsive** to the end user

What **metrics** matter?

Accuracy

Efficiency

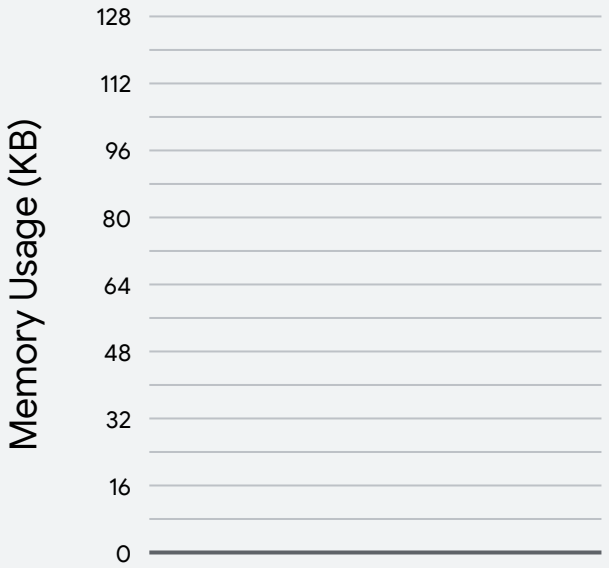
Beyond Model Metrics

Latency

- Model must be f**ast enough** to keep up with the speech input
- The model must run fast enough to be **responsive** to the end user
- But it must run efficiently on a *small* processor **TinyML**

Memory Usage

- Need to be **resource aware**
- **Less** compute
- **Less** memory
- Use **quantization**



Memory Usage

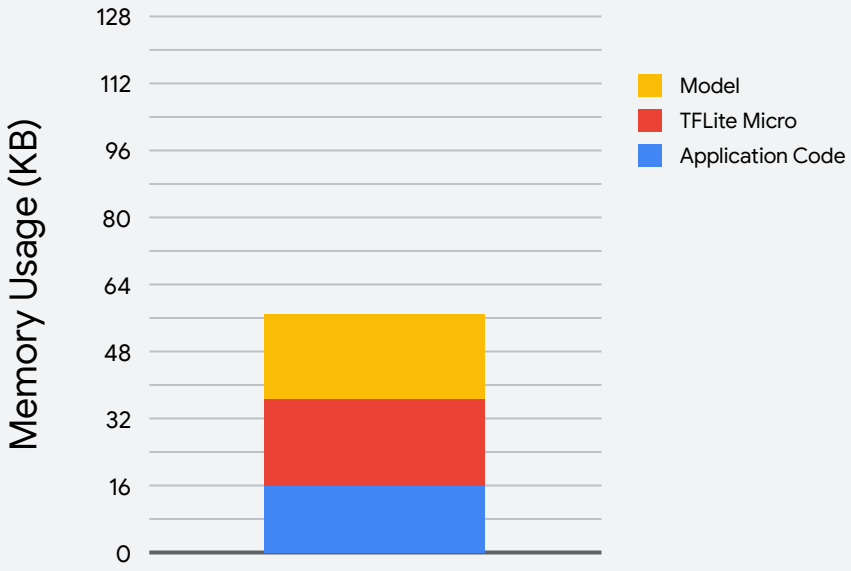
- Need to be **resource aware**
- **Less** compute
- **Less** memory
- Use **quantization**

Memory Usage

- Need to be **resource aware**
- **Less** compute
- **Less** memory
- Use **quantization**

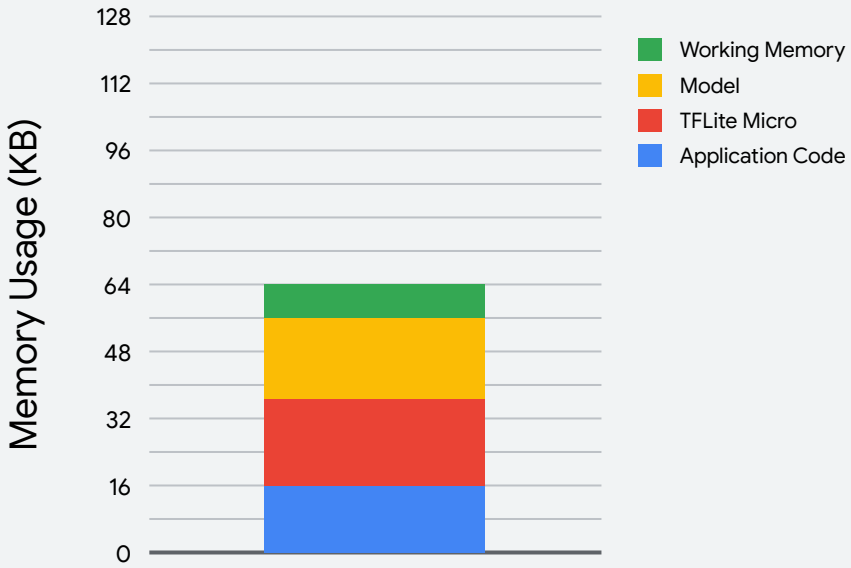
Memory Usage

- Need to be **resource aware**
- **Less** compute
- **Less** memory
- Use **quantization**



Memory Usage

- Need to be **resource aware**
- **Less** compute
- **Less** memory
- Use **quantization**

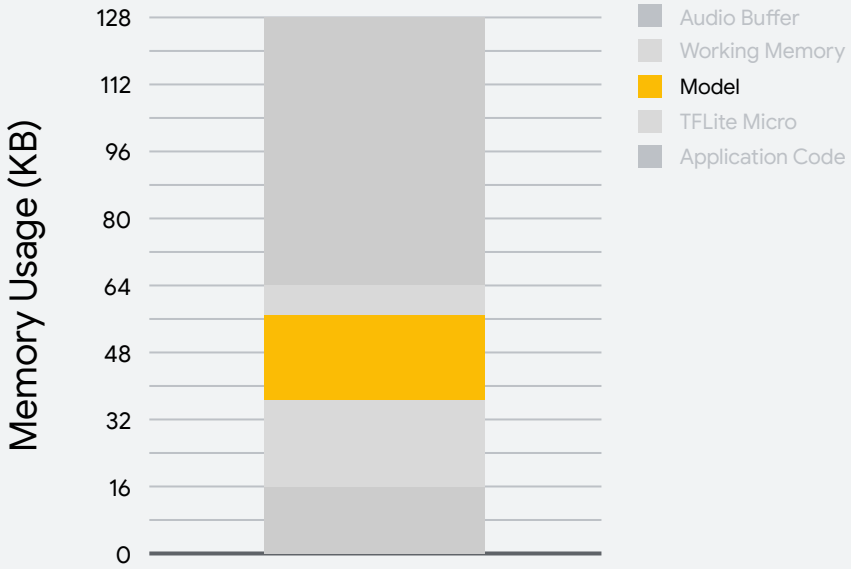


Memory Usage

- Need to be **resource aware**
- **Less** compute
- **Less** memory
- Use **quantization**

Memory Usage

- Need to be **resource aware**
- **Less** compute
- **Less** memory
- Use **quantization**



What **metrics** matter?

Accuracy

Efficiency

Beyond Model Metrics

Beyond Model Metrics

Think about **quality of experience (QoE)**

What really defines that? It isn't just whether the model is performing well on a given dataset. It is about the **user experience**: how can we **assess** that?

Beyond Model Metrics

Think about **quality of experience (QoE)**

What really defines that? It isn't just whether the model is performing well on a given dataset. It is about the **user experience**: how can we **assess** that?

- Have diverse **users** to test against?
- Test in different **backgrounds**?
- Add **noise** while training the models?

So how can we **improve** our KWS Application?