Custom Datasets

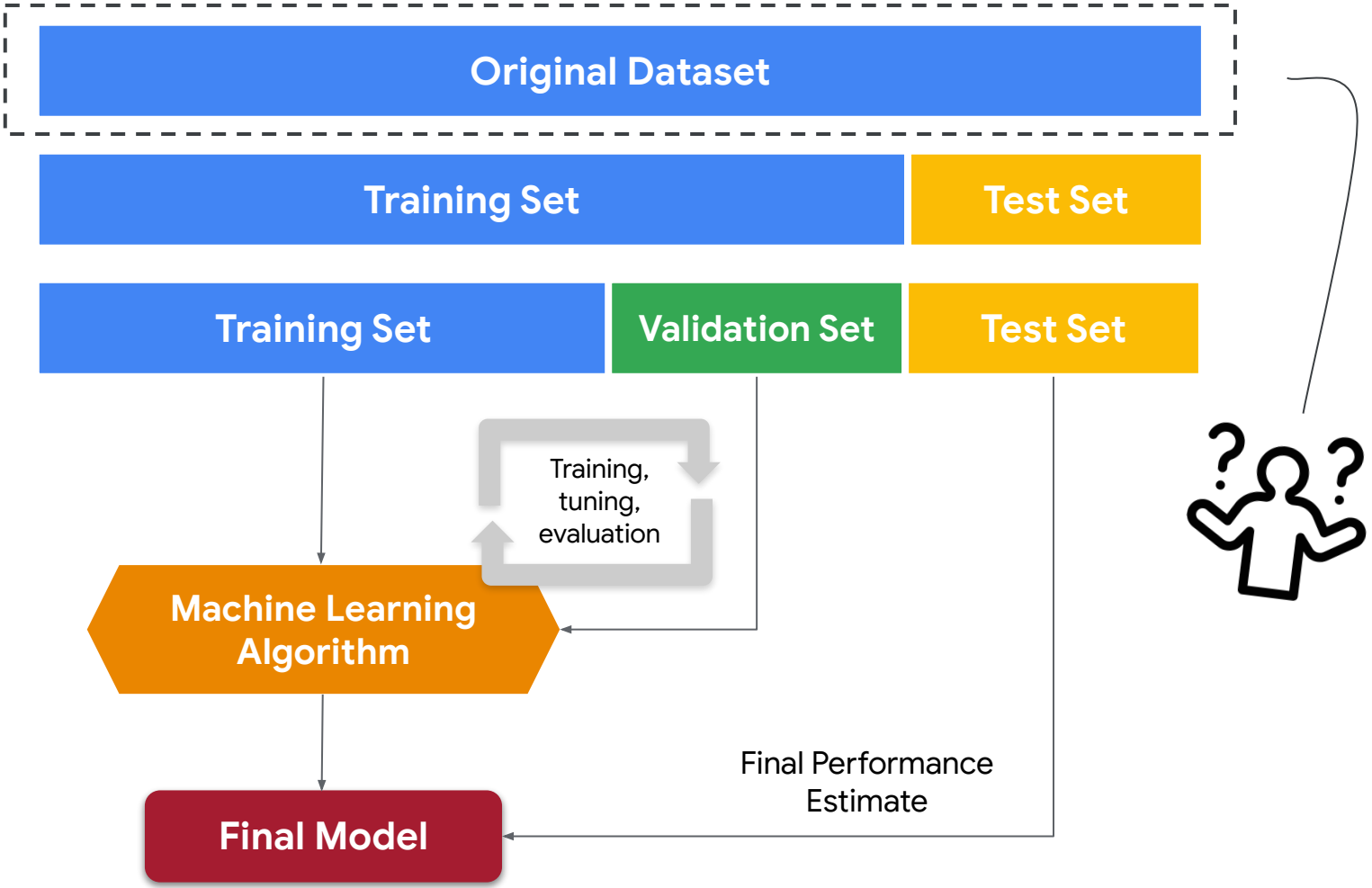
ML Workflow


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

ML Workflow


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

Collecting data for a new machine learning task


Original Dataset
Training Set
Test Set
Training Set
Validation Set
Test Set
Training, tuning, evaluation
Machine Learning Algorithm
Final Performance Estimate
Final Model

Datasets require ***significant effort***

These **massive** machine learning datasets are ***constructed by hand***

- **Common Voice**—**5000+** hours of spoken audio
- Common Objects in Context (**COCO**)—**2.5M+** labeled images
- **ImageNet**—**4M+** labeled images
- **Waymo—1,950** 20-second driving segments
- **KITTI 360**—**73KM+** of annotated driving data

Data Engineering: *how do you build your own dataset*

Data Engineering

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

- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition

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

April 2018

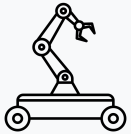
Data Collection

"yes"

"no"

Common Use

"left"
"right"
"go"
"stop"



Robotics

"one"
"two"
"four"
"six"

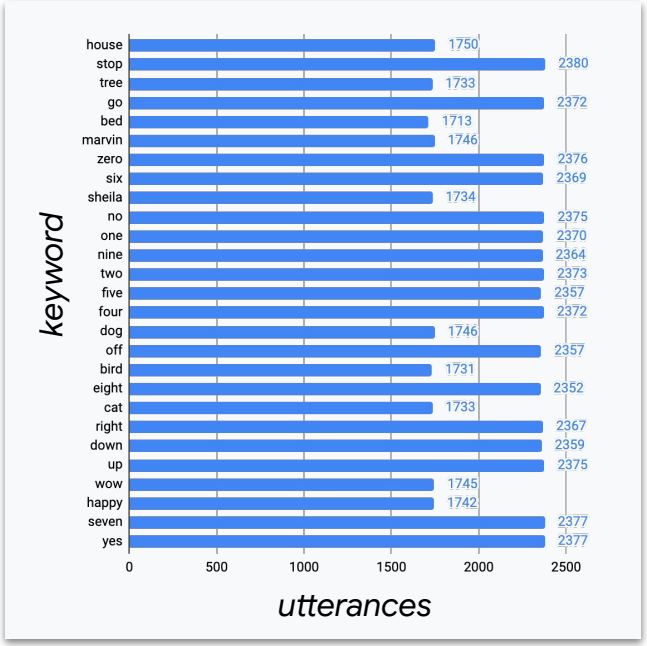
#

Numbers

Requirements Gathering Refinement Sustainment

Data Collection

- **2,618** volunteers
  - consented to have their voices redistributed
  - Variety of accents
- > 1,000 examples for **each** keyword
- **Browser-based** recording



Requirements | **Gathering** | Refinement | Sustainment

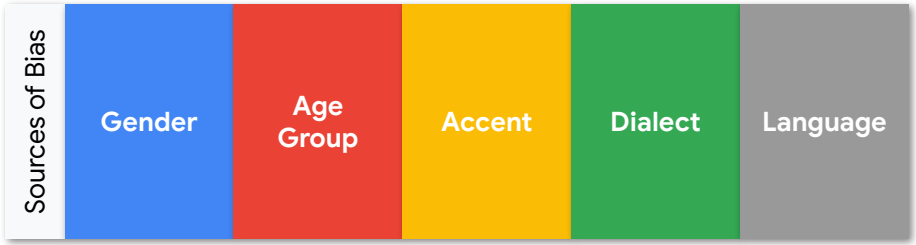
Data Validation

You'll run this tool on your custom data!

- Some data is **unusable**
  - Too quiet, wrong word, etc
- Started with **automated tools**
  - Remove low volume recordings
  - Extract loudest 1s (from 1.5sec examples)
- All 105,829 remaining utterances **manually reviewed** through crowdsourcing

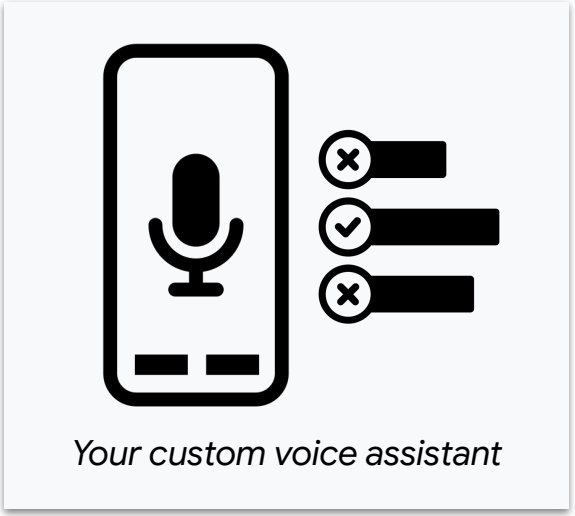
Your dataset will **evolve**

- Missing **demographics**?
- Expanding your user-base?
- Reducing **bias**?
  - Multiple sources



Collecting your **custom** dataset

- **Next assignment:** custom keyword spotting model
- Touches on many aspects of *data engineering*

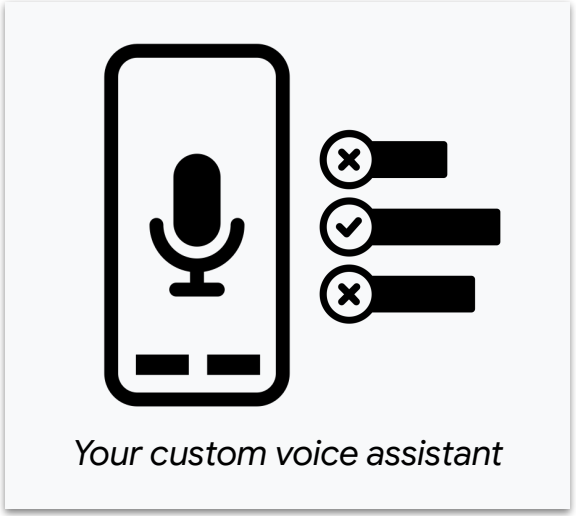


Your custom voice assistant

Requirements | Gathering | Refinement | Sustainment

Collecting your **custom** dataset

- **How much** data is needed?
- Acceptable **false positive** and **false negative** rates?
- **Impact** of errors?

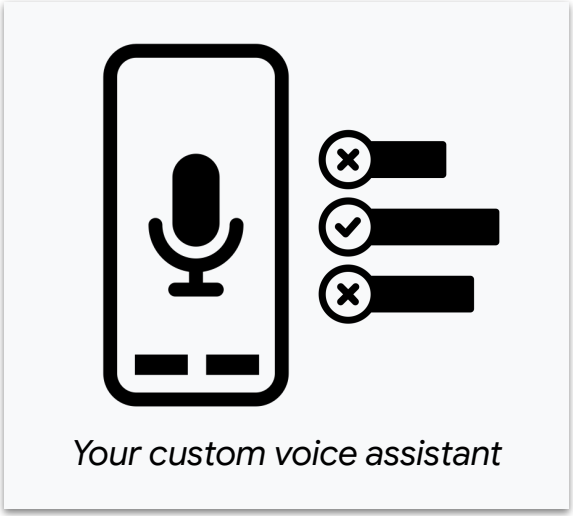


Your custom voice assistant

Requirements | **Gathering** | Refinement | Sustainment

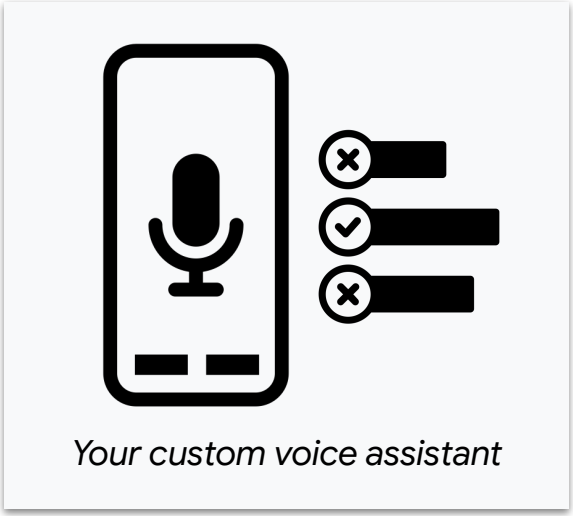
Collecting your **custom** dataset

- Recording **issues**
- Too short or clipped utterances
- Too quiet
- **Augmenting** with background noise



Collecting your **custom** dataset

- Can **others** use your trained model?
- Does the model work in **different environments**?
  - Kitchen
  - Car
  - TV/Radio in the background
  - Crowded room



Your custom voice assistant