

# IESTI01 - TinyML

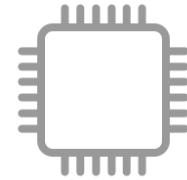
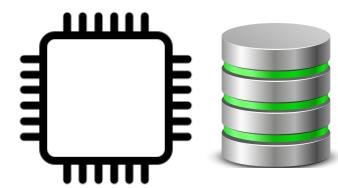
Data Engineering for TinyML

Prof. Marcelo Rovai

June 23<sup>rd</sup>, 2021



# What is data Engineering?



### Data Engineering

Collect Data

Preprocess Data

Design a Model

Train a Model

Evaluate Optimize

Convert Model

Deploy Model

Make Inferences

### Model Engineering



TensorFlow

### Model Deployment

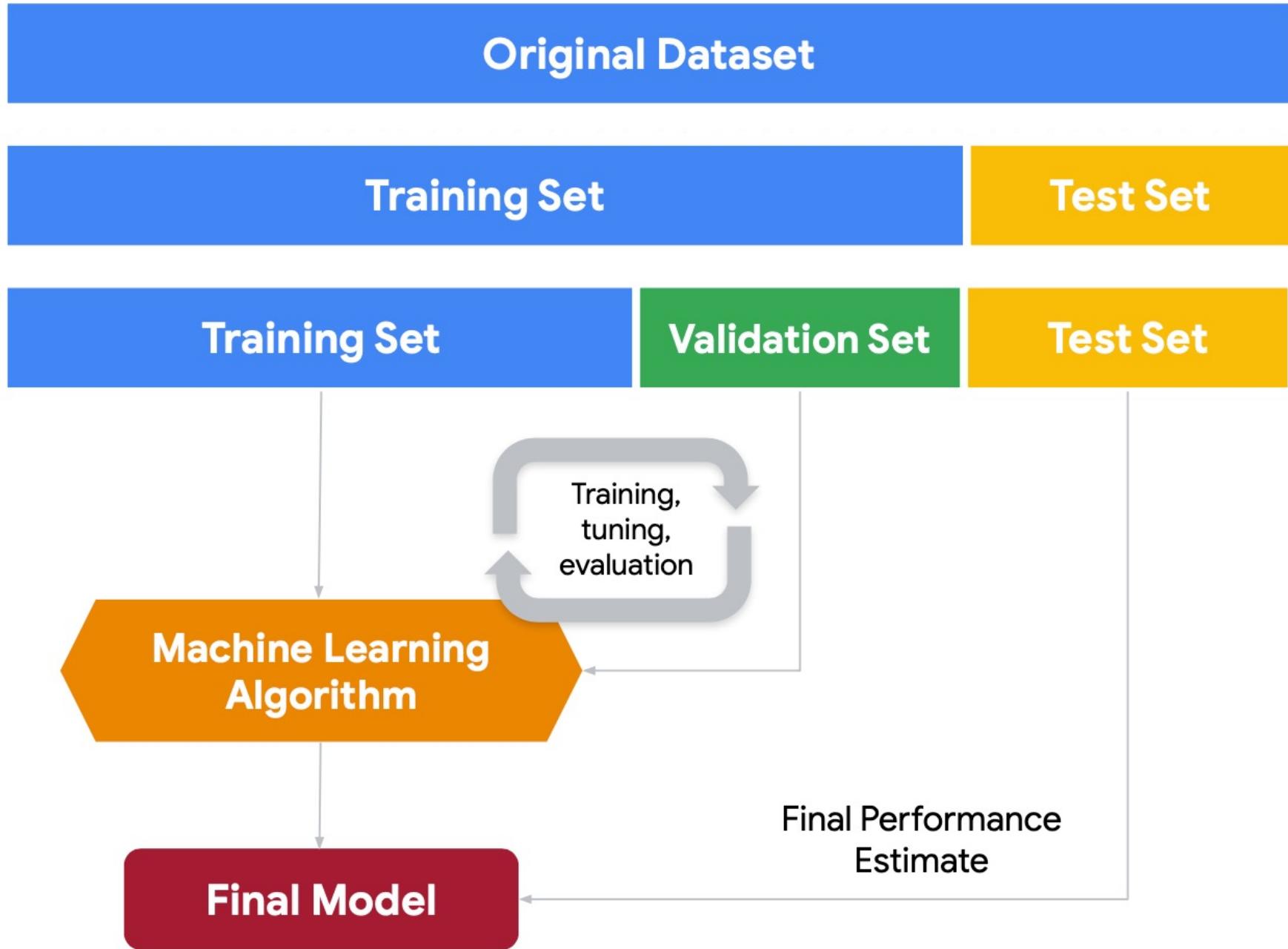


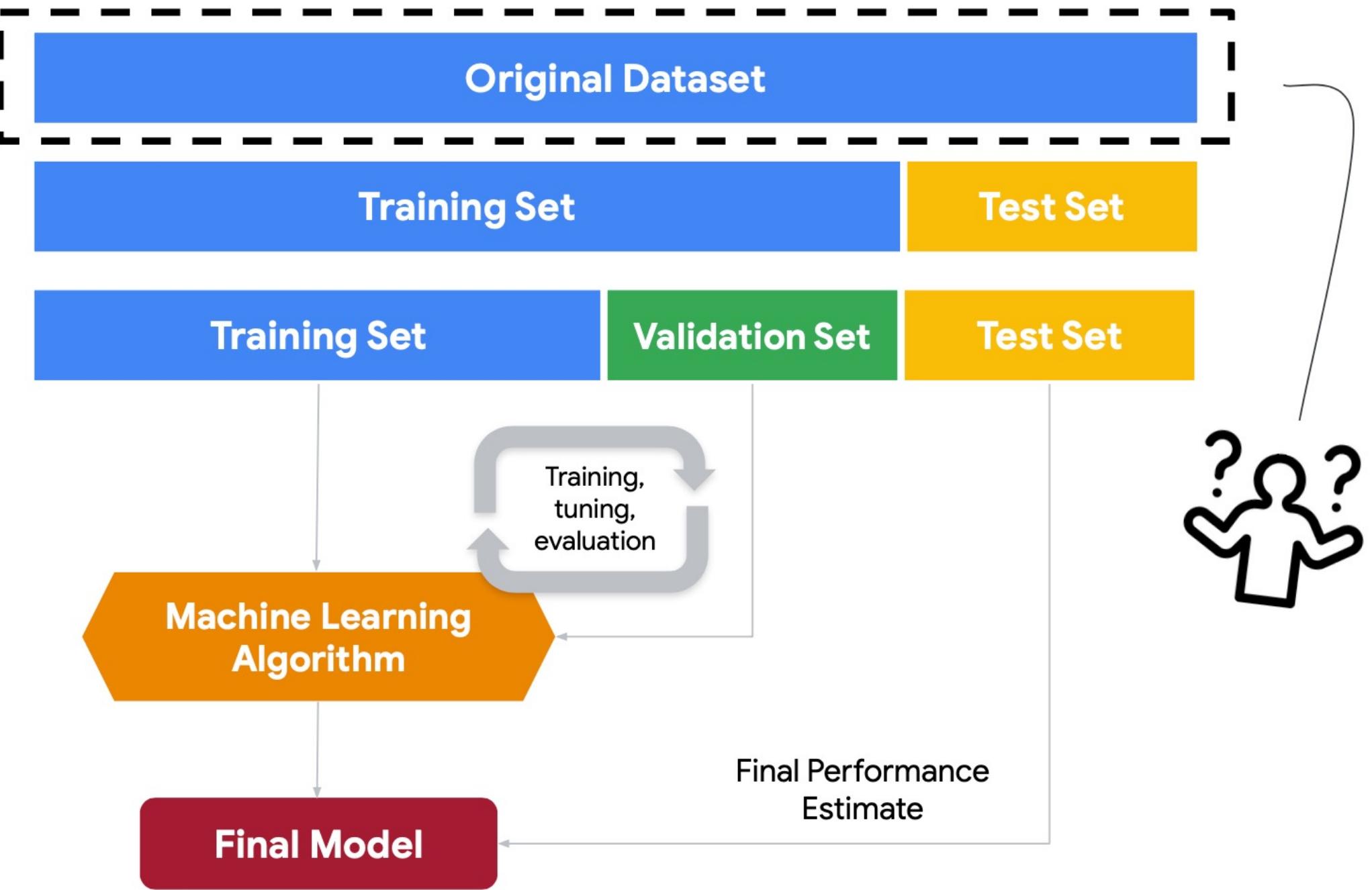
TensorFlow Lite



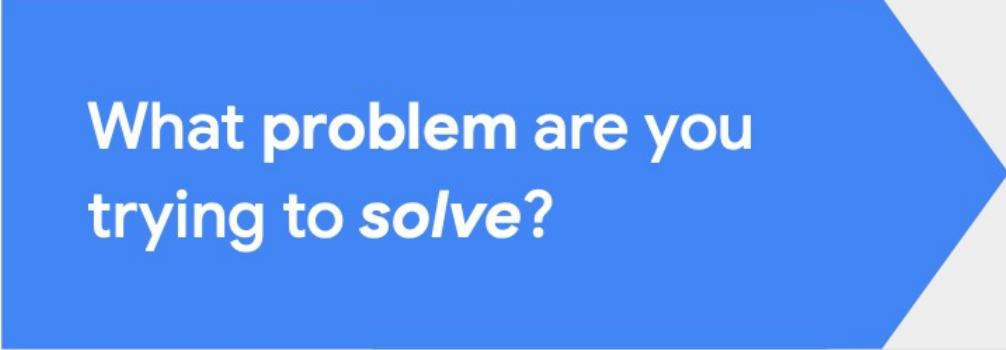
TensorFlow Lite Micro

 EDGE IMPULSE





# Good Data is Necessary for Accuracy



What problem are you  
trying to **solve**?

- Your data must contain useful features
- Can a human (expert) distinguish between examples of each class?
- How will you measure performance?

# Good Data is Necessary for Accuracy

What problem are you trying to *solve*?

Both **quantity** and **quality** will influence your model's performance

- Your data must contain useful features
  - Can a human (expert) distinguish between examples of each class?
  - How will you measure performance?
- 
- **Wide distribution of training examples**
  - **Accurate labels (Ground Truth)**
  - **Sufficient class balance**

# Data Engineering

## Requirements

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

# Data isn't free to use

Where does your data **originate?**

- Open?
- Copyrighted?
- Licensed?
- Product users?

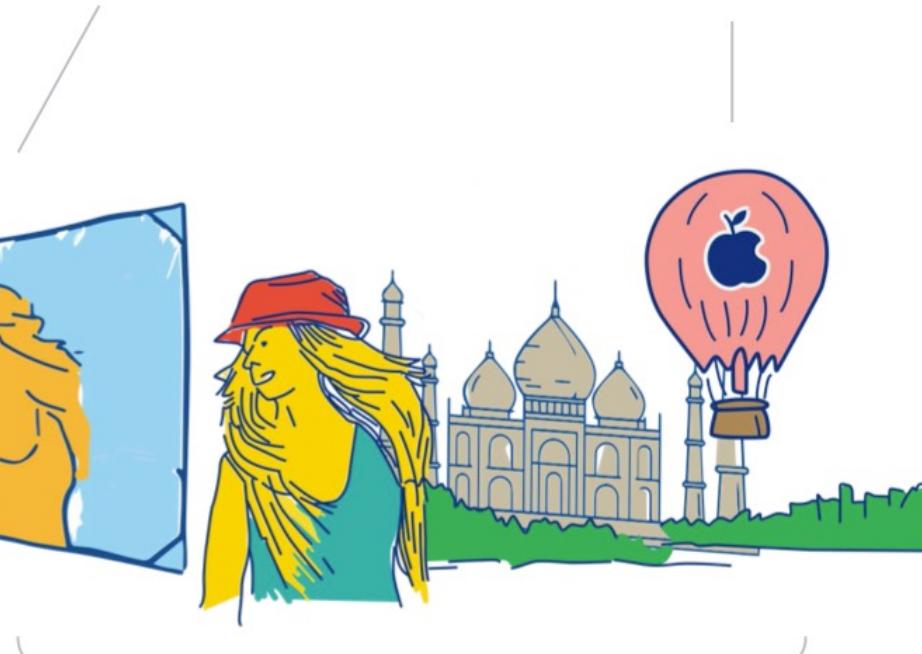


# What's Yours and What's Not Yours

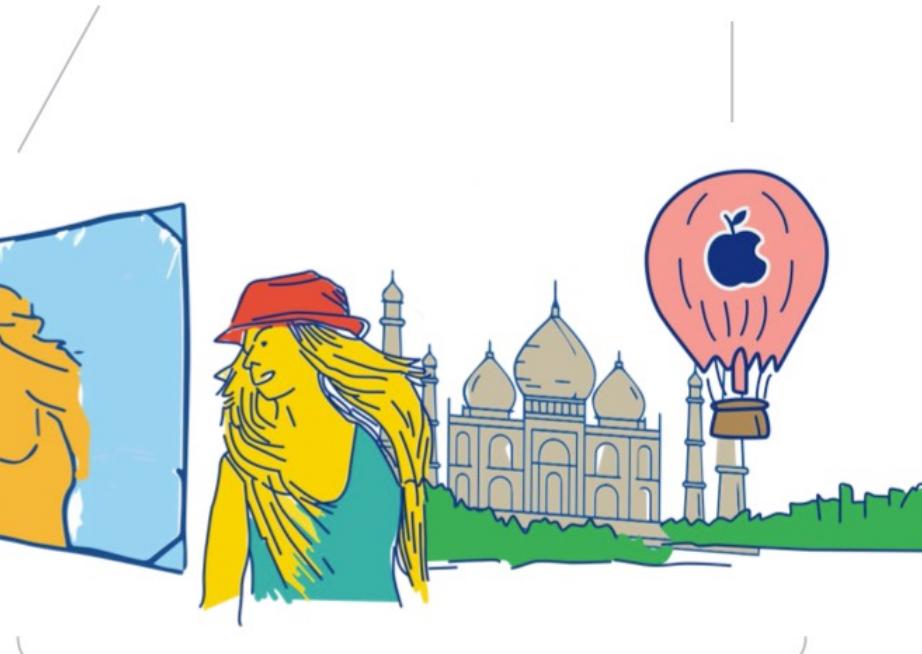
Author / Owner



Creative Work

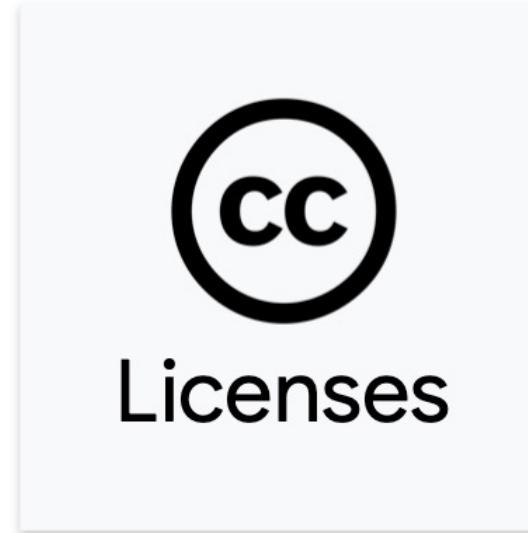


Trademarked Logo



Copyrighted

# Licenses



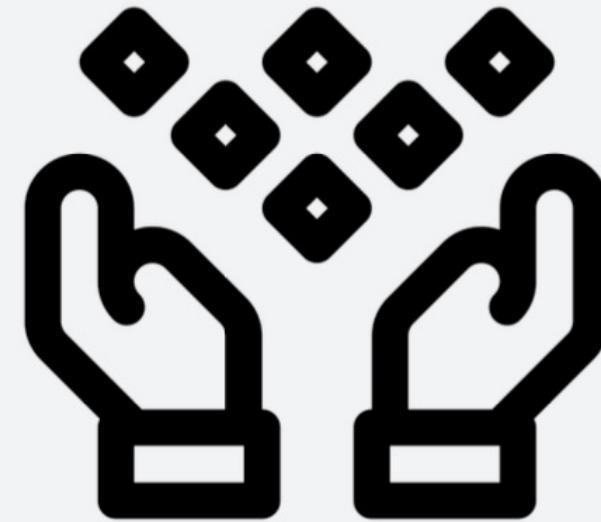
# Data Engineering



- Problem definition
- Permissions & rights
- Machine & human  
usable format
- People
- Collection
- Labeling
- **Data sources**

# Data sources

- Sensors
- Crowdsourcing
- Product users
- Paid contributors



# Data Engineering



- Problem definition
- Permissions & rights
- Machine & human  
usable format
- Data sources
- People
- Collection
- Labeling
- Processing
- Augmentation
- **Validation**

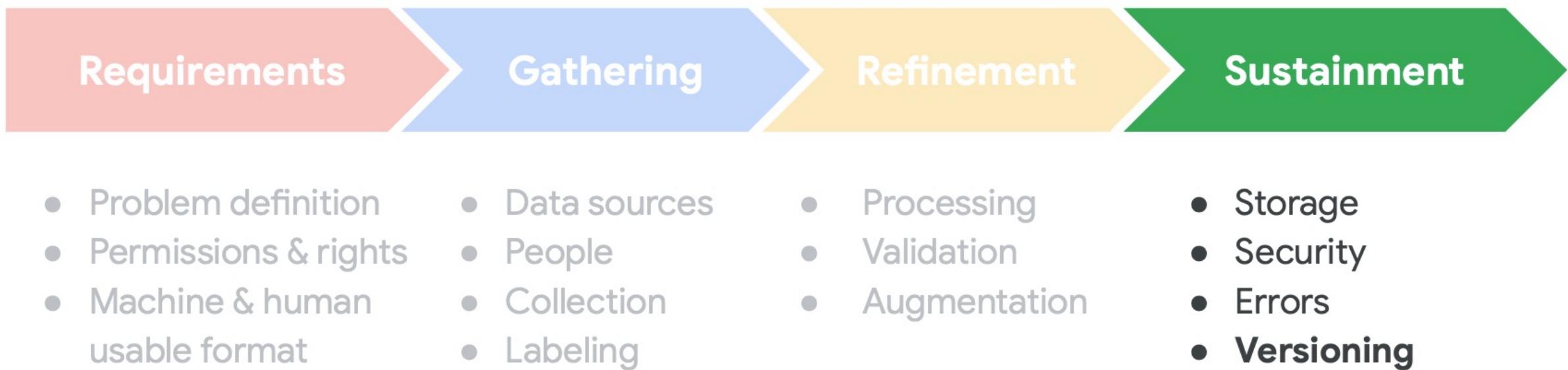
# Some data is *unusable*

How will you **verify** the data you collected?

- **Manually** (time, cost)
- **Automation**
- Domain expertise
  - disputes / disagreements

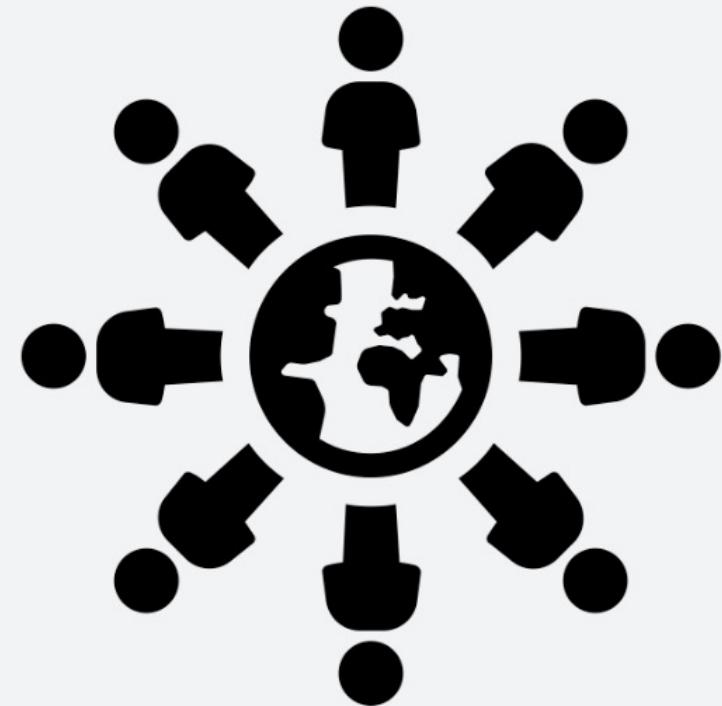


# Data Engineering

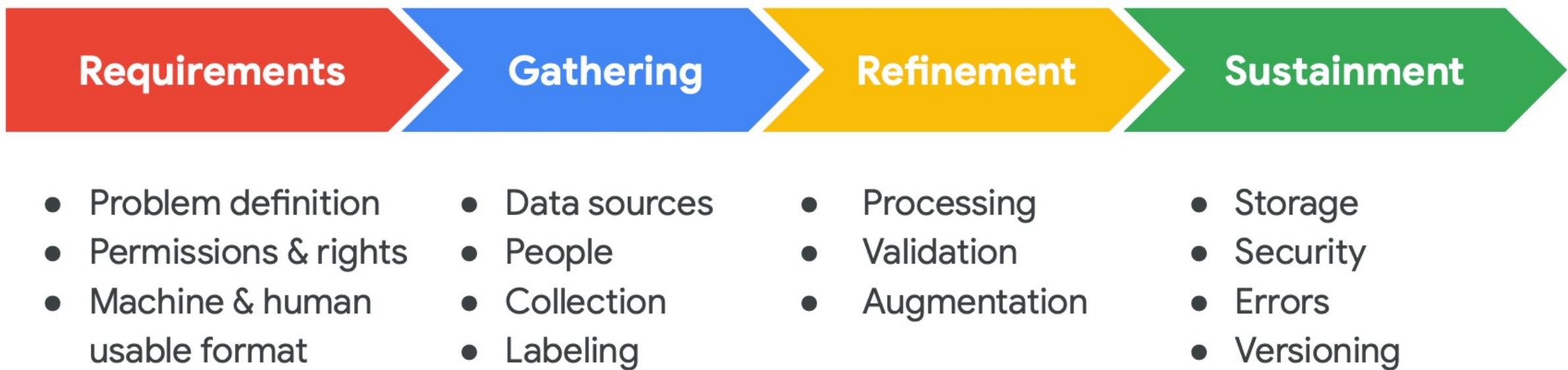


# Your dataset will **evolve**

- Missing **demographics**?
- **Expanding** your user-base?



# Data Engineering



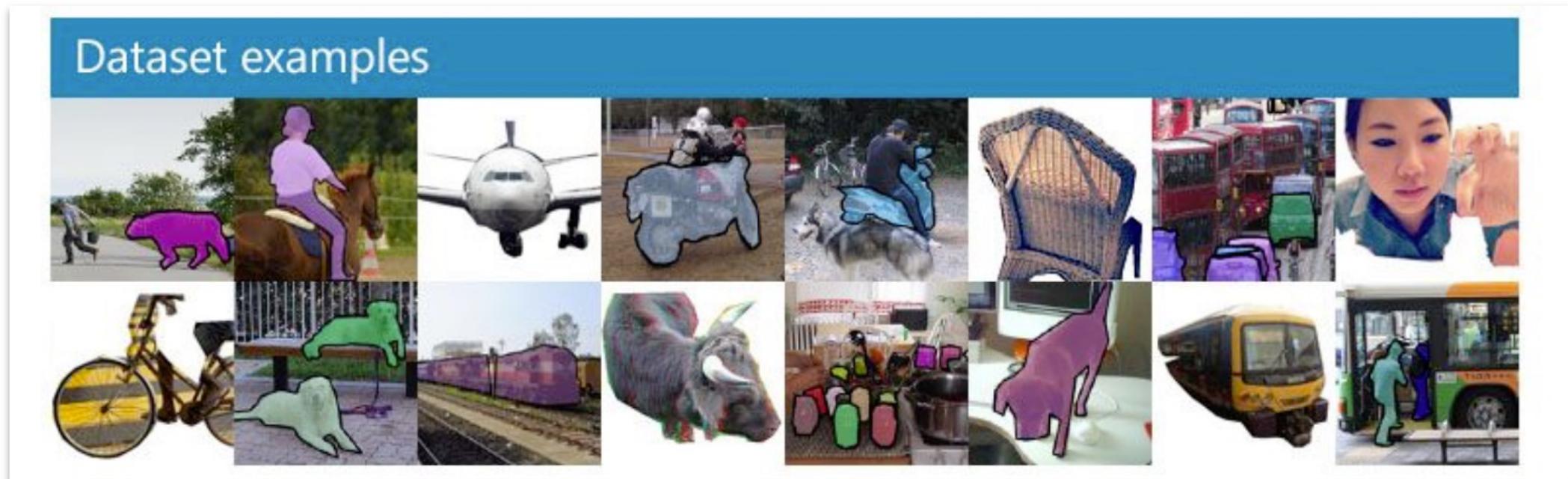
Datasets require  
**significant effort**

# Classifying Images



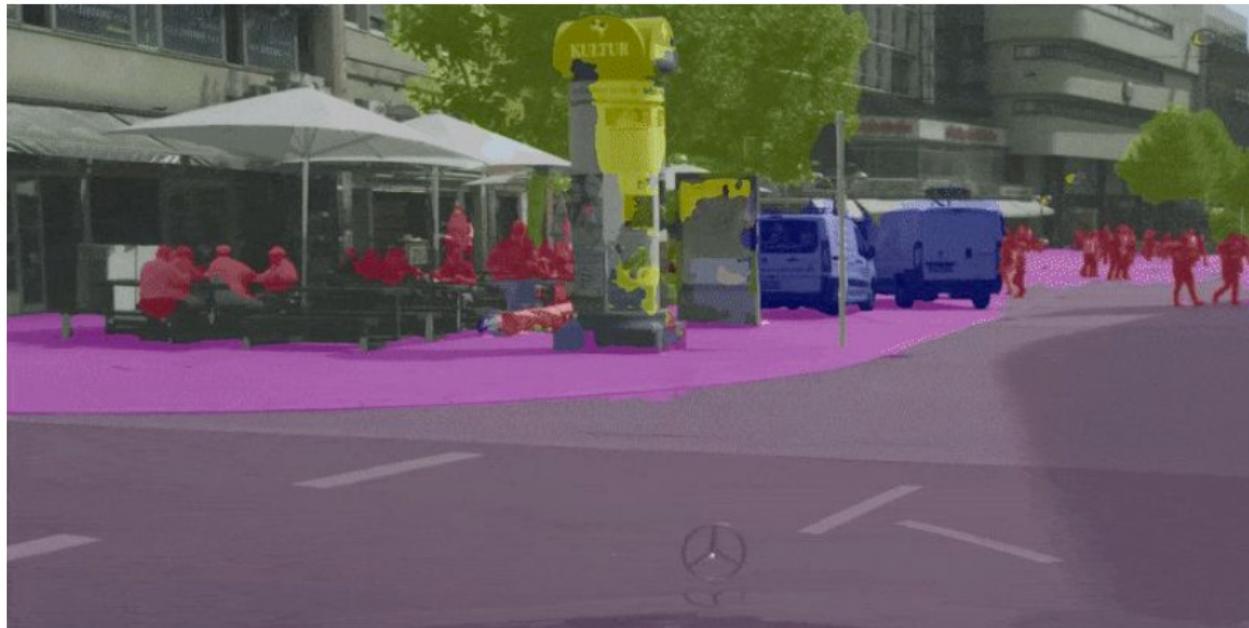
# Detecting Objects

- Common Objects in Context (**COCO**)—**2.5M+** segmented images



# Classifying Video Motion

- Waymo—**1,950** 20-second driving segments (cameras, LIDAR, labels)
- KITTI 360—**73KM+** of annotated driving data



# Detecting Anomalies



ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)



Journal of Sound and Vibration 289 (2006) 1066–1090

JOURNAL OF  
SOUND AND  
VIBRATION

[www.elsevier.com/locate/jsvi](http://www.elsevier.com/locate/jsvi)

## Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics

Hai Qiu<sup>a,\*</sup>, Jay Lee<sup>a</sup>, Jing Lin<sup>b</sup>, Gang Yu<sup>c</sup>

<sup>a</sup>Center for Intelligent Maintenance Systems, University of Cincinnati, OH 45221, USA

<sup>b</sup>Institute of Acoustics, Chinese Academy of Sciences, 17 Zhongguancun Street, Haidian, Beijing 100080, China

<sup>c</sup>Department of Mechanical and Industrial Engineering, Northeastern University, 60 Huntington Ave 334SN, Boston, MA 02115, USA

Received 26 March 2004; accepted 10 March 2005

Available online 31 May 2005

[Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics](#)

NATIONAL AERONAUTICS  
AND SPACE ADMINISTRATION

+ABOUT NASA    +LATEST NEWS    +MULTIMEDIA    +MISSIONS    +WORK FOR NASA

+ NASA Home    + Ames Home    + Intelligent Systems Division    + Discovery and Systems Health

**Data Repository**

+ Home    **+ Data Repository**    + Open Source Code    + Publications    + Roadmap    + Awards

**Prognostics Center of Excellence**

**PCoE Datasets**

**Overview**

The Prognostics Data Repository is a collection of data sets that have been donated by various universities, agencies, or companies. The data repository focuses exclusively on prognostic data sets, i.e., data sets that can be used for development of prognostic algorithms. Mostly these are time series of data from some nominal state to a failed state. The collection of data in this repository is an ongoing process.

Publications making use of databases obtained from this repository are requested to acknowledge both the assistance received by using this repository and the donators of the data. This will help others to obtain the same data sets and replicate your experiments. It also provides credit to the donators.

Users employ the data at their own risk. Neither NASA nor the donators of the data sets assume any liability for the use of the data or any system developed using the data.

If you have suggestions concerning the repository send email to [chetaan.s.kulkarni \[at\] nasa.gov](mailto:chetaan.s.kulkarni@nasa.gov) Thank you and please come again.

**4. Bearing Data Set**

**Publications using this data set**

<b>Description</b>	Experiments on bearings. The data set was provided by the Center for Intelligent Maintenance Systems (IMS), University of Cincinnati.
<b>Format</b>	The set is in text format and has been zipped.
<b>Datasets</b>	<a href="#">+ Download Bearing Data Set (58020 downloads)</a>
<b>Dataset Citation</b>	J. Lee, H. Qiu, G. Yu, J. Lin, and Rexnord Technical Services (2007). IMS, University of Cincinnati. "Bearing Data Set", NASA Ames Prognostics Data Repository ( <a href="http://ti.arc.nasa.gov/project/prognostic-data-repository">http://ti.arc.nasa.gov/project/prognostic-data-repository</a> ), NASA Ames Research Center, Moffett Field, CA

# 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
- **NASA Bearing Dataset** - **+60K** 1-second vibration of 4 bearings

**Data Engineering:** *How to build your own dataset?*

# Speech Commands

## A TinyML dataset example

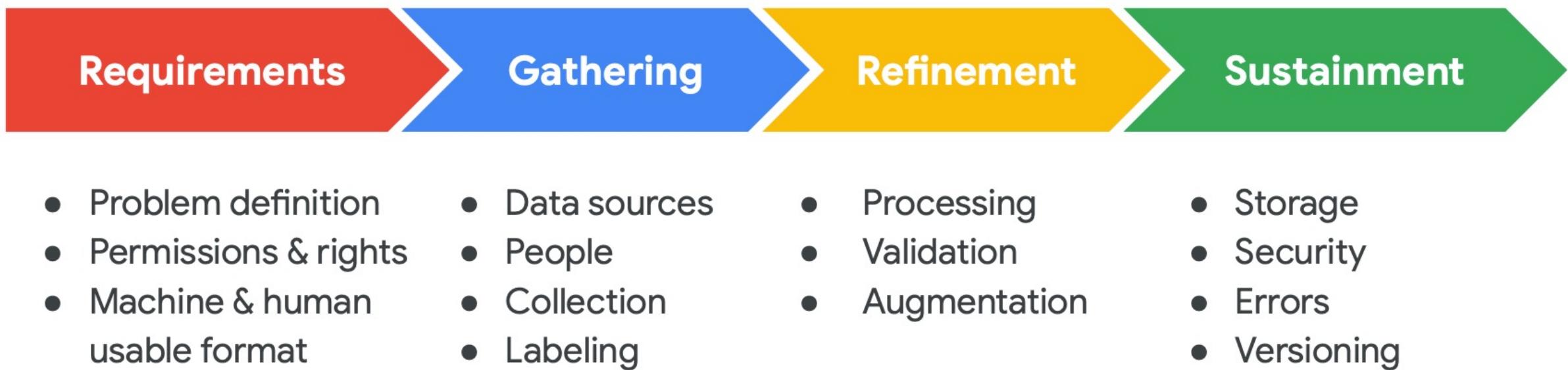
# Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition

Pete Warden  
Google Brain  
Mountain View, California  
[petewarden@google.com](mailto:petewarden@google.com)

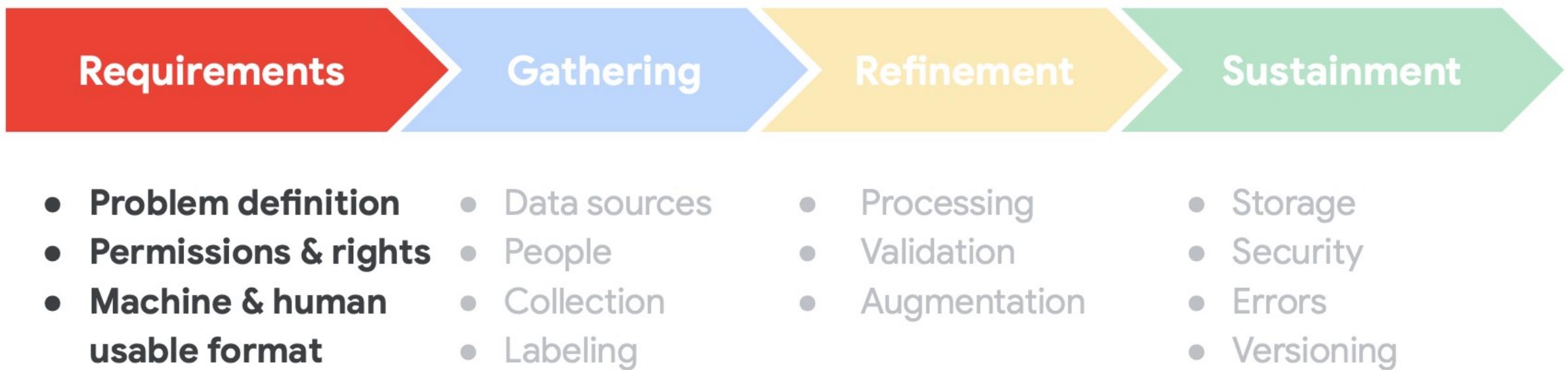
April 2018

<https://arxiv.org/pdf/1804.03209.pdf>

# Data Engineering



# Data Engineering



# Requirements

- Need for a **public** dataset for keyword spotting (KWS)
- Previous datasets **not KWS-focused**
- Google, Apple, Amazon use **proprietary datasets**
- **Speech Commands:**
  - Permissively licensed
  - Research & commercial use ok
- **Established new standard**

# Requirements

“yes”



“no”



*Common Use*

“left”

“right”

“go”

“stop”



*Robotics*

“one”

“two”

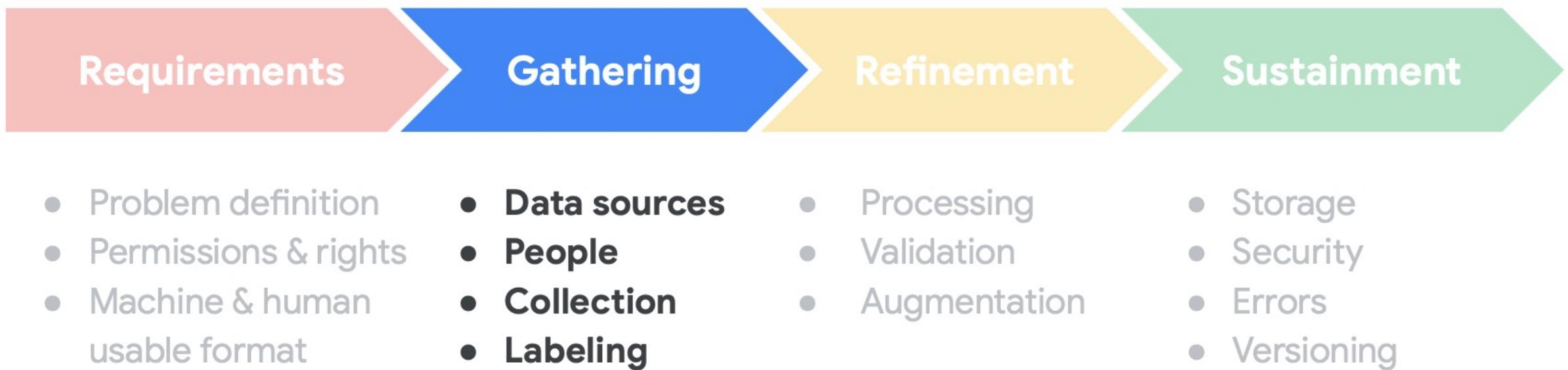
“four”

“six”



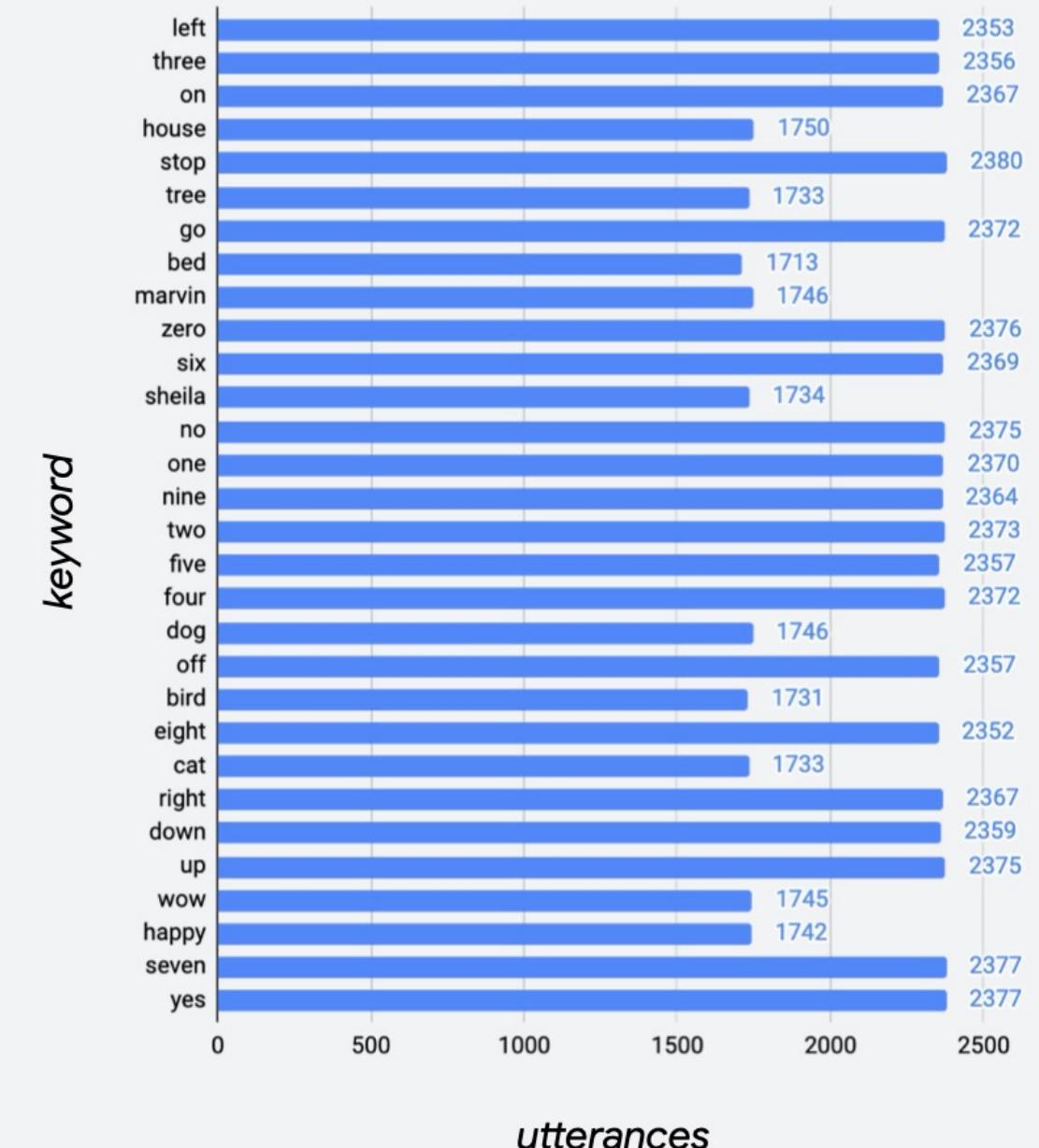
*Numbers*

# Data Engineering

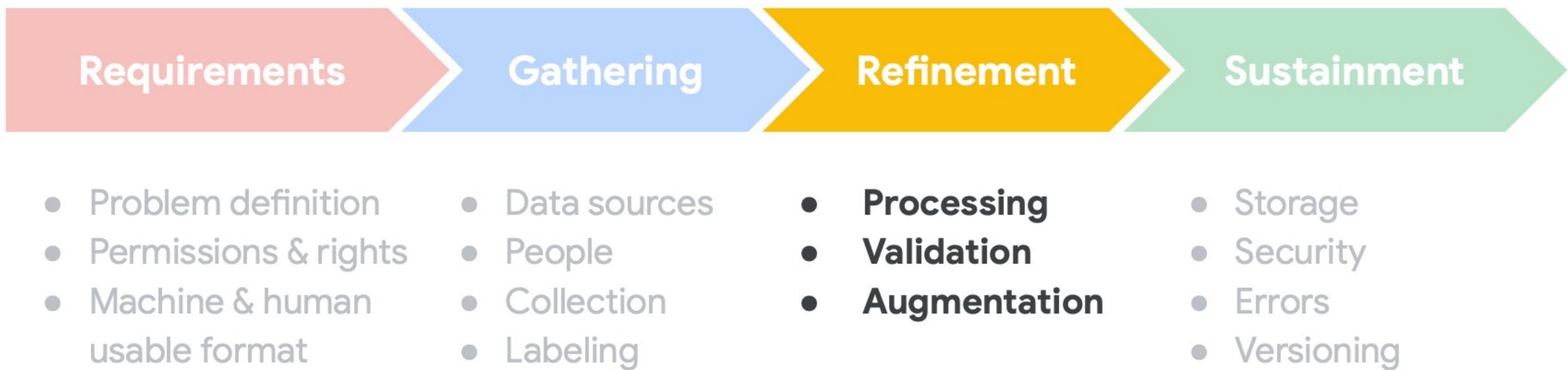


# Data Collection

- **2,618** volunteers
  - consented to have their voices redistributed
  - Variety of accents
- > 1,000 examples for **each** keyword
- **Browser-based**  
(no app to install)



# Data Engineering

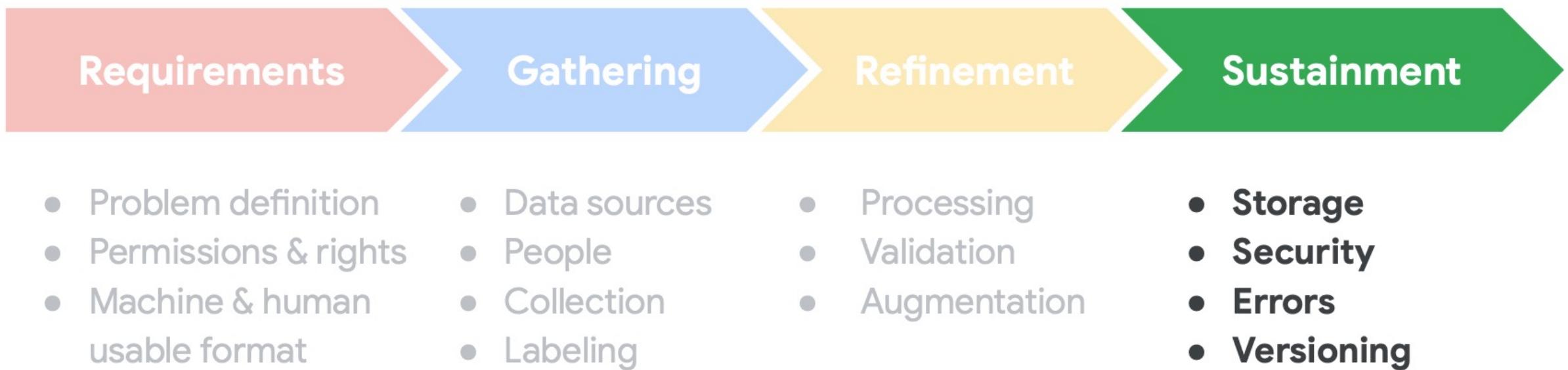


# Data Validation

- 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

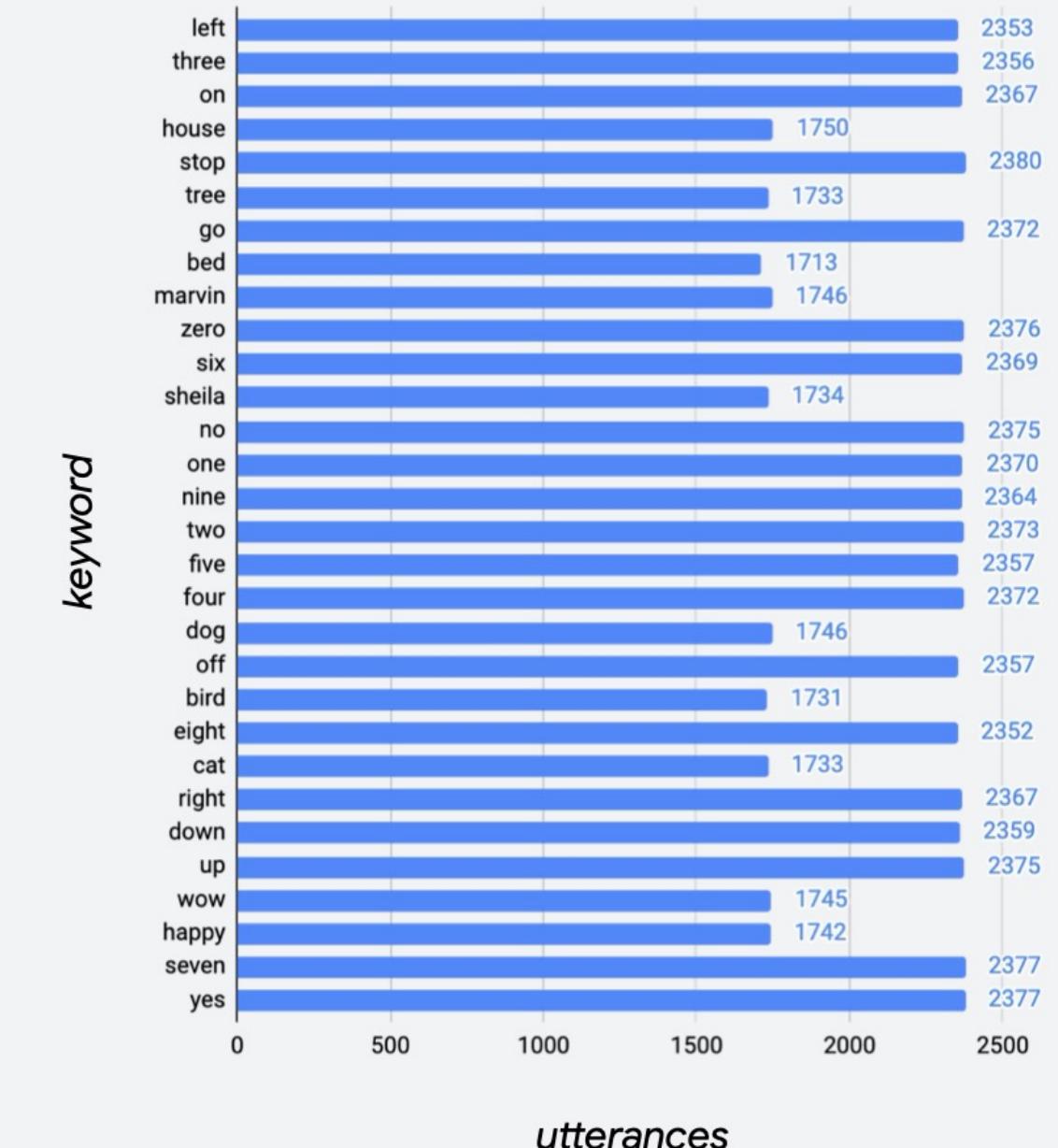


# Data Engineering



# Sustaining KWS Research

- Speech Commands is now in **v2**
  - **Expanded to 35 keywords** from original 10
- Includes train/validation/test splits
- Expand to **new languages?**



# Why is wide availability of data *important*?

**Compare**

Without a standard dataset, no easy way to compare models

**Benchmark**

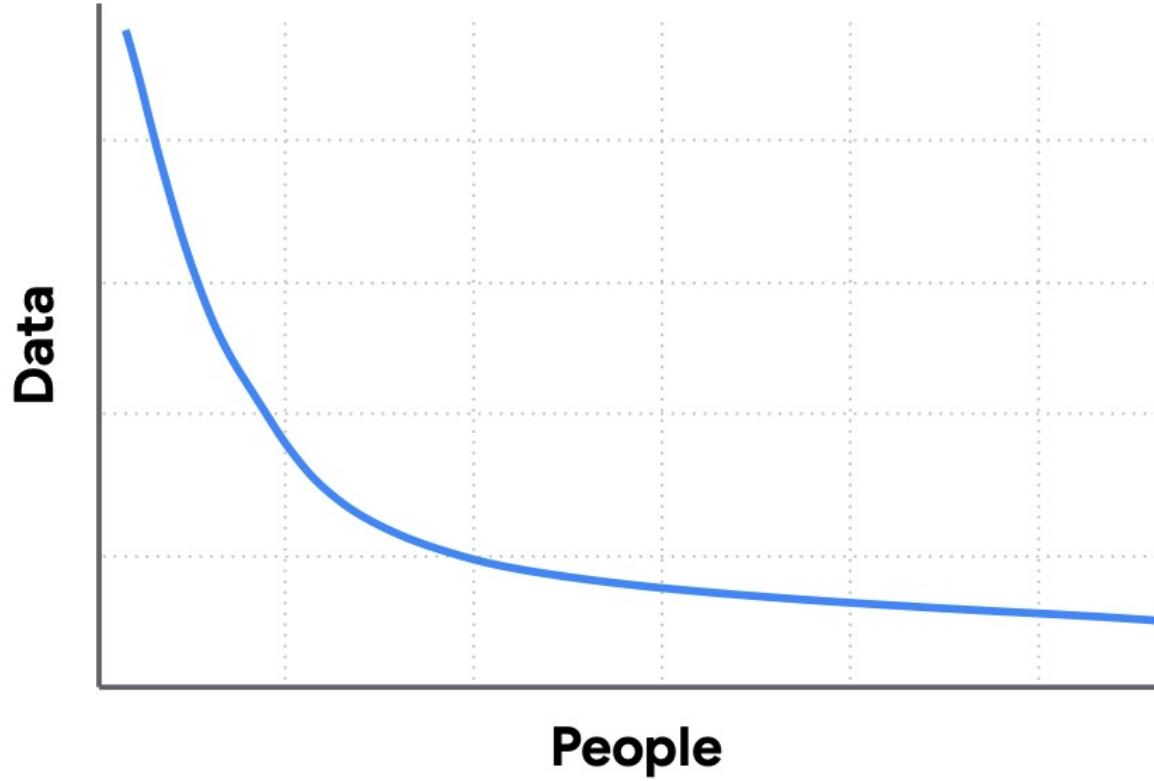
Good benchmarks: critical under *TinyML* constraints!  
Not just accuracy: power, memory, latency

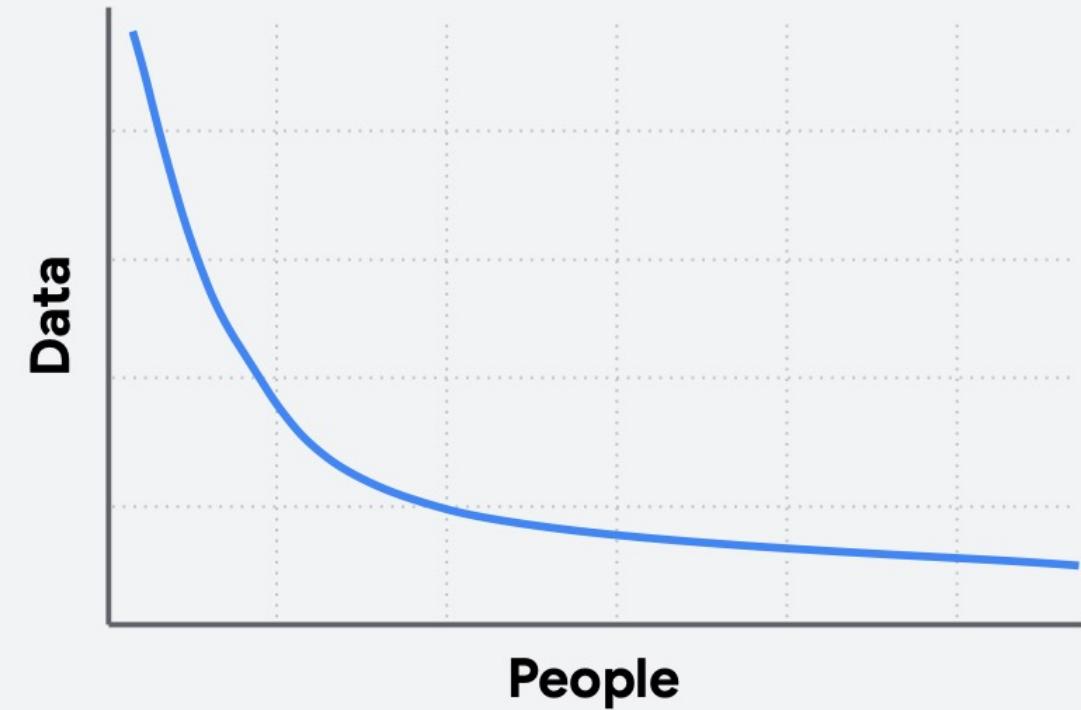
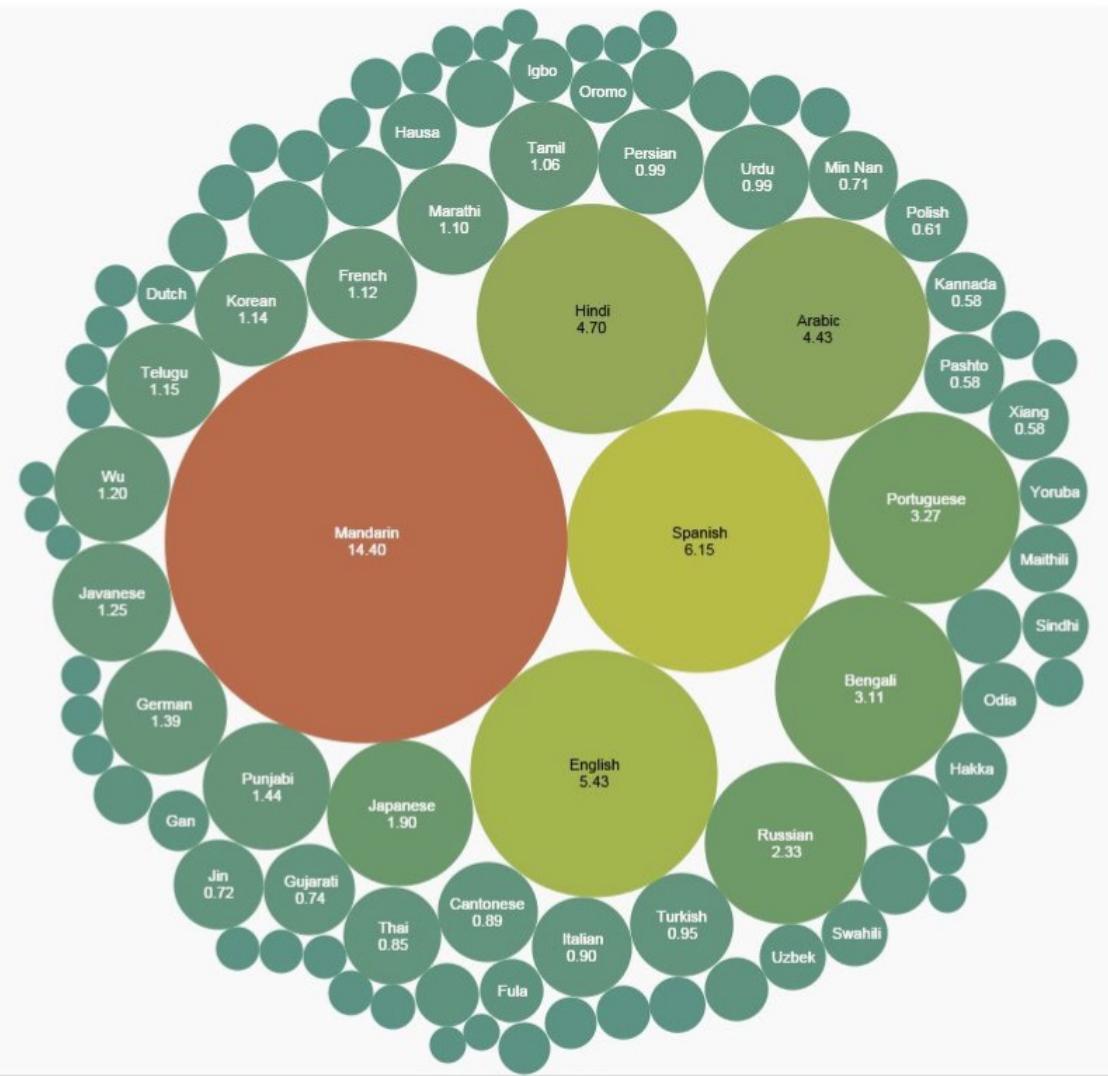
**Improve**

Speech Commands: a standard training & evaluation dataset for each new KWS technique

Researches using the same data: “apple-to-apples” evaluation

# Crowdsourcing Data for the Long Tail





# Cost Model v. Community Model?



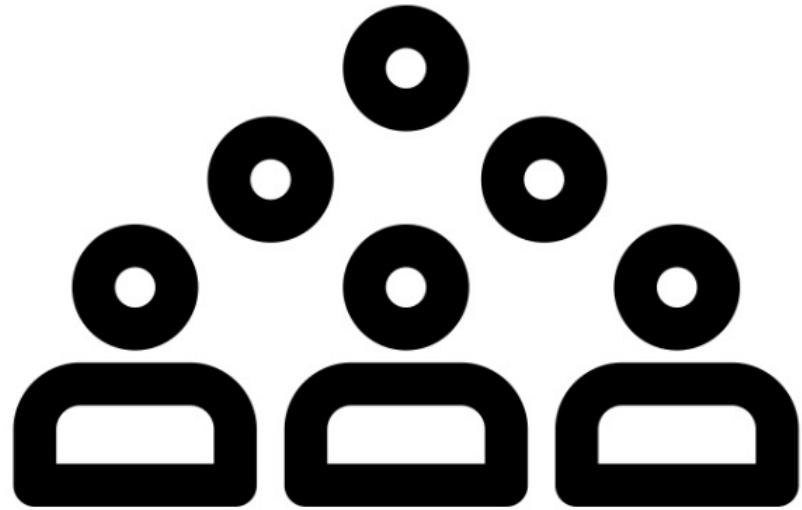
Limited Scale



Social Good

# Common Voice

- **Crowdsourcing** platform



<https://commonvoice.mozilla.org/en>



Speak

Listen

1/5 Clips



1

2

3

4

5

Bianca has since appeared in  
Drillbit Taylor.

Click did they accurately speak the sentence?

The unconventional nature of  
the tubular girder bridge was  
not widely accepted.



YES



NO

Shortcuts

Report

Skip &gt;&gt;

# Common Voice

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

Common Voice  
moz://a

CONTRIBUTE DATASETS LANGUAGES ABOUT

0 0 Log In / Sign Up EN

## Speak

Donate your voice



## Listen

Help us validate voices





Common Voice is Mozilla's initiative to help teach machines how real people speak.

Voice is natural, voice is human. That's why we're excited about creating usable voice technology for our machines. But to create voice systems, developers need an extremely large amount of voice data.

Most of the data used by large companies isn't available to the majority of people. We think that stifles innovation. So we've launched Common Voice, a project to help make voice recognition open and accessible to everyone.

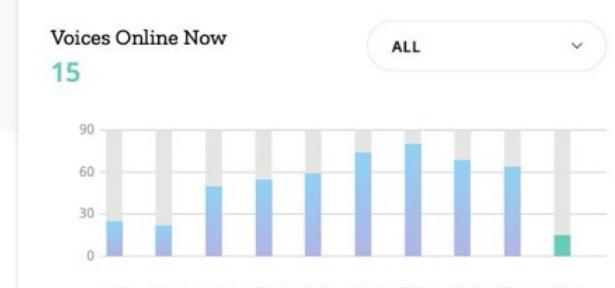
[READ MORE](#)

Hours Recorded ● 9.0k Hours Validated ● 7.1k ALL



Display a menu

Voices Online Now 15 ALL



11PM 12AM 01AM 02AM 03AM 04AM 05AM 06AM 07AM 08AM

# Common Voice

- Crowdsourcing platform
- Over 50,000 volunteers
- 54 different languages

The screenshot shows the Mozilla Common Voice website at <https://commonvoice.mozilla.org/en/datasets>. The page displays information about the dataset, including its purpose, size, and demographic details. A sidebar on the right lists various languages available in the dataset, with Portuguese selected. The bottom section provides statistics on validated and recorded hours and the number of languages included.

We're building  
an open source, multi-language  
dataset of voices that anyone can  
use to train speech-enabled  
applications.

We believe that large, publicly available voice  
datasets will foster innovation and healthy  
commercial competition in machine-learning  
based speech technology.

Common Voice's multi-language dataset is  
already the largest publicly available voice  
dataset of its kind, but it's not the only one.

Look to this page as a reference hub for other  
open source voice datasets and, as Common  
Voice continues to grow, a home for our release  
updates.

Version: Common Voice Corpus 6.1  
Language: Portuguese  
Kabyle  
Kyrgyz  
Luganda  
Lithuanian  
Latvian  
Mongolian  
Maltese  
Dutch  
Odia  
Punjabi  
Polish  
✓ Portuguese  
Romansh Sursilvan  
Romansh Vallader  
Romanian  
Russian  
Kinyarwanda  
Sakha  
Slovenian  
Swedish

SPLITS: 35% 19 - 29, 32% 30 - 39, ...  
AGE: 81% Male, 3% Female

Validated Hours: 7,335  
Recorded Hours: 9,283  
Languages: 60

Enter Email to Download

Why an email? We may need to contact you in the future about changes to the dataset, an email provides us a point of contact.

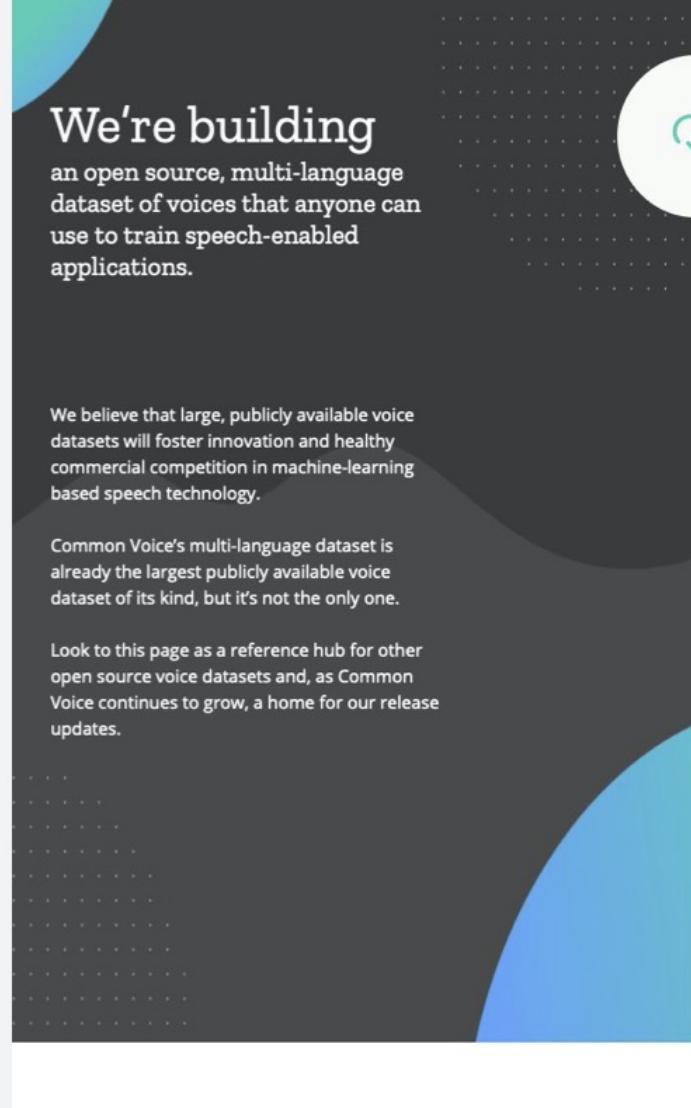
## What's inside the Common Voice dataset?

Each entry in the dataset consists of a unique MP3 and corresponding text file. Many of the 9,283 recorded hours in the dataset also include demographic metadata like age, sex, and accent that can help train the accuracy of speech recognition engines.

The dataset currently consists of 7,335 validated hours in 60 languages, but we're always adding more voices and languages. Take a look at our Languages page to request a language or start contributing.

# Common Voice

- **Crowdsourcing platform**
- **Over 50,000 volunteers**
- **54 different languages**
- **Goal: speech recognition for *all languages* on the planet**



Display a menu

What's inside the Common Voice dataset?

# File Structure

- **Valid**
  - At least 2 people listen to them, and the majority of those listeners say the audio matches the text
- **Invalid**
  - At least 2 listeners, and the majority say the audio does not match the clip
- **Other**
  - All other clips, i.e., fewer than 2 votes, or those that have equal valid and invalid votes, are labelled “other”



# Interesting Attributes

- **Permissive license**
- **Many contributors**
- **Comes with metadata**

We're building an open source, multi-language dataset of voices that anyone can use to train speech-enabled applications.

We believe that large, publicly available voice datasets will foster innovation and healthy commercial competition in machine-learning based speech technology.

Common Voice's multi-language dataset is already the largest publicly available voice dataset of its kind, but it's not the only one.

Look to this page as a reference hub for other open source voice datasets and, as Common Voice continues to grow, a home for our release updates.

Language English

SIZE 50 GB

VERSION en\_1932h\_2020-06-22

VALIDATED HR. TOTAL 1,469

OVERALL HR. TOTAL 1,932

LICENSE CC-0

NUMBER OF VOICES 61,528

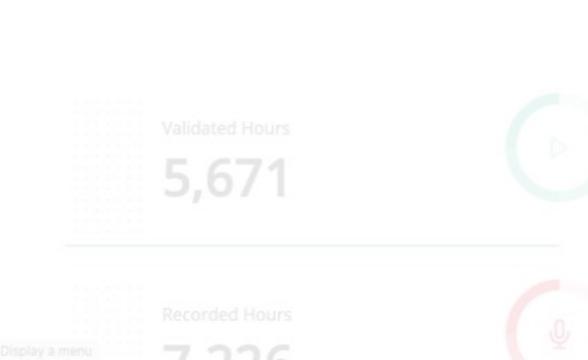
AUDIO FORMAT MP3

SPLITS

- Accent 23% United States English, 8% England English, ...
- Age 23% 19 - 29, 14% 30 - 39, ...
- Gender 47% Male, 14% Female

Enter Email to Download

Why an email? We may need to contact you in the future about changes to the dataset; an email provides us a point of contact.

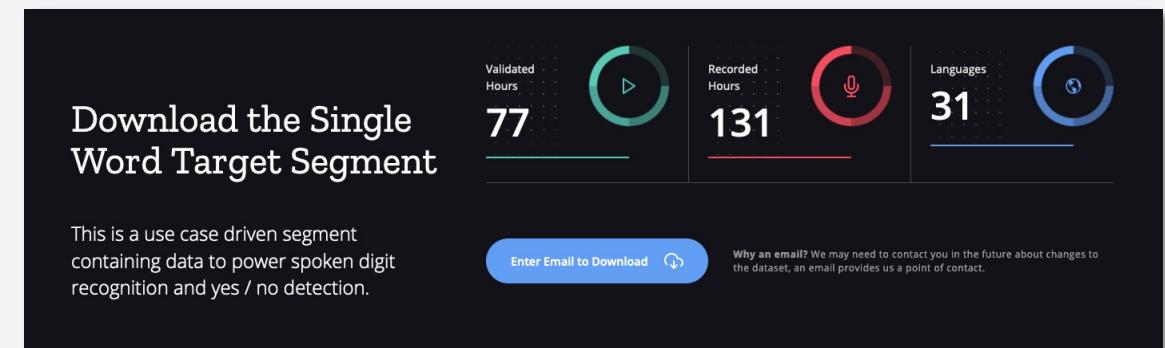


What's inside the Common Voice dataset?

# Single Word Target Segment

A *speech commands-style* dataset for **18 languages**

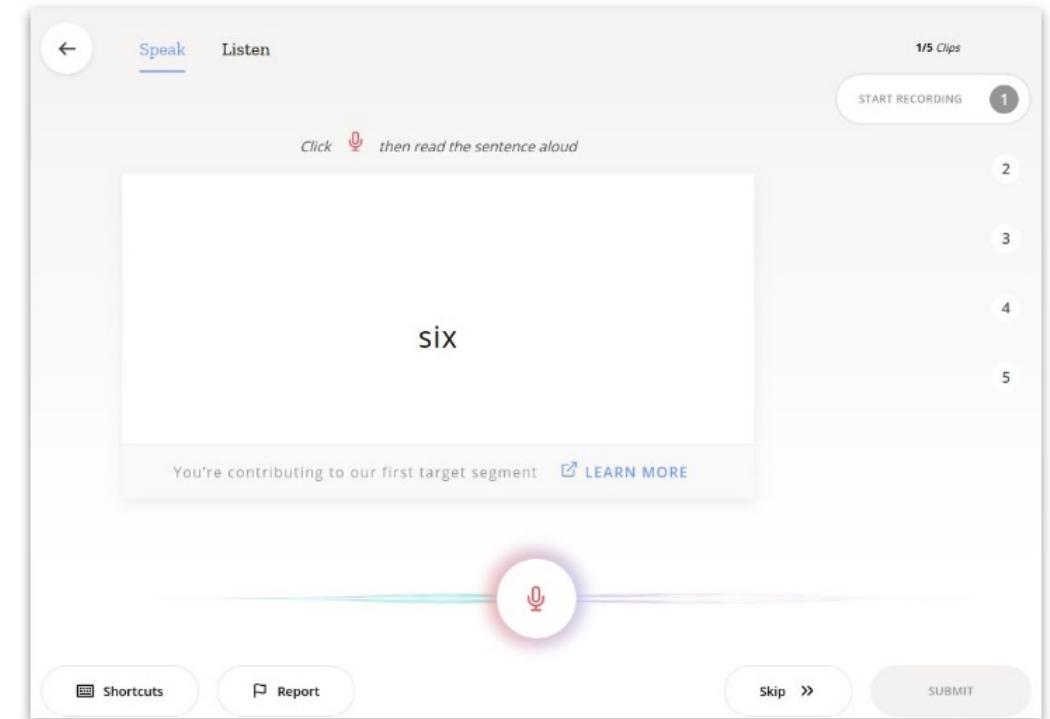
- “Yes” // “no”
- “hey” & “Firefox”
- **digits** 0-9



<https://commonvoice.mozilla.org/en/datasets>

# ASR Diversity and Reach

- **Common Voice**
  - Permissive license
  - Minority languages
- **Ease-of-use, wide reach**
  - Browser-based
  - Community can add new languages
- **You can contribute!**



# Common Voice Data Structure

“yes”



Prompt

Record

Attributes

Validation

Data Splits

# Using Existing Datasets for **TinyML**

# Don't **collect** from scratch

Data collection is **difficult!**

- Can we **reuse** existing data?

What's available?

What's missing?

# TensorFlow

## Datasets Catalog

Audio  
Image  
*Image Classification*  
*Object Detection*  
*Question Answering*  
*Structured Summarization*  
Text  
*Translate*  
Video



Screenshot of the TensorFlow Datasets Catalog page:

The page title is "Datasets". The navigation bar includes "Install", "Learn", "API", "Resources", "More", "Search", "English", "GitHub", and "Sign in".

The "Catalog" tab is selected. The sidebar lists various datasets:

- Overview
- ▶ Audio
- ▶ Image
- ▶ Image classification
- ▶ Object detection
  - coco
  - coco\_captions
  - kitti
  - open\_images\_challenge2019\_detection
  - open\_images\_v4
  - voc
  - waymo\_open\_dataset
  - wider\_face
- ▶ Question answering
- ▶ Structured
  - amazon\_us\_reviews
  - e2e\_cleaned
  - forest\_fires
  - genomics\_cod
  - german\_credit\_numeric
  - higgs
  - howell
  - iris
  - movie\_lens
  - movielens
  - radon
  - rock\_you
  - titanic
  - web\_nlg
  - wiki\_bio
  - wine\_quality
- ▶ Summarization
- ▶ Text

The "wider\_face" dataset is highlighted in blue. The main content area shows the following details for "wider\_face":

TensorFlow > Resources > Datasets > Catalog

★★★★★

### wider\_face

- Description:**

WIDER FACE dataset is a face detection benchmark dataset, of which images are selected from the publicly available WIDER dataset. We choose 32,203 images and label 393,703 faces with a high degree of variability in scale, pose and occlusion as depicted in the sample images. WIDER FACE dataset is organized based on 61 event classes. For each event class, we randomly select 40%/10%/50% data as training, validation and testing sets. We adopt the same evaluation metric employed in the PASCAL VOC dataset. Similar to MALF and Caltech datasets, we do not release bounding box ground truth for the test images. Users are required to submit final prediction files, which we shall proceed to evaluate.
- Homepage:** <http://shuoyang1213.me/WIDERFACE/>
- Source code:** [tfds.object\\_detection.WiderFace](#)
- Versions:**
  - 0.1.0 (default): No release notes.
- Download size:** 3.42 Gib
- Dataset size:** Unknown size
- Auto-cached (documentation):** Unknown
- Splits:**

Split	Examples
'test'	16,097
'train'	12,880

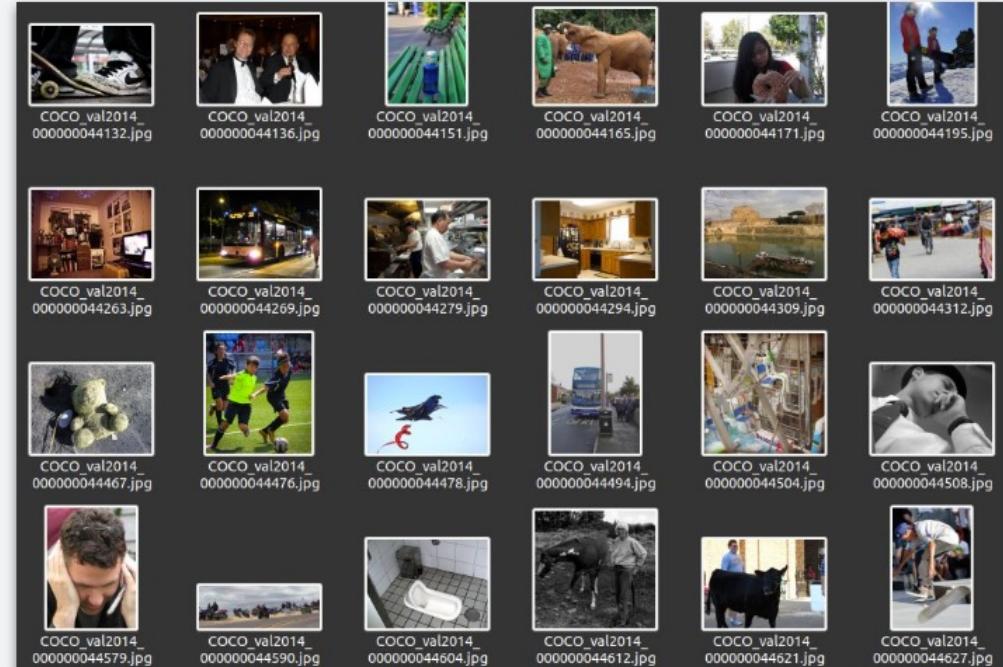
Always check **license!**

# TinyML

## Person Detection

- **Visual Wake Words**: a new dataset built from Common Objects in Context (**COCO**)
    - **people v. no people**

# Repurposing existing datasets for **TinyML** tasks is a powerful concept

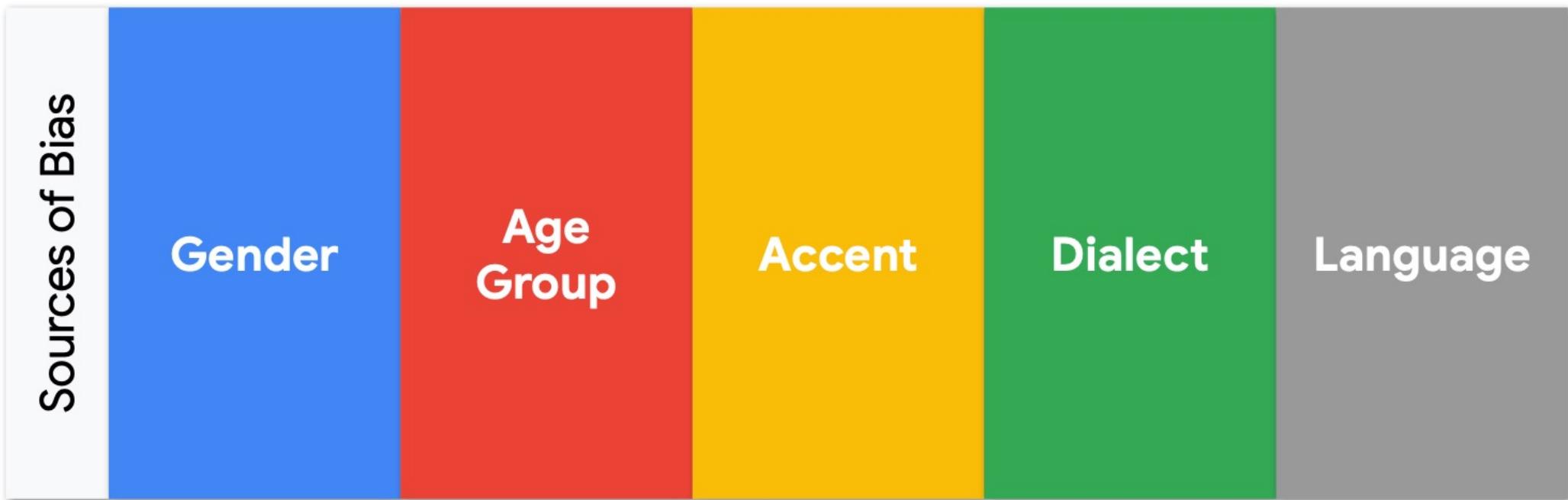


# **Don't *learn* from scratch**

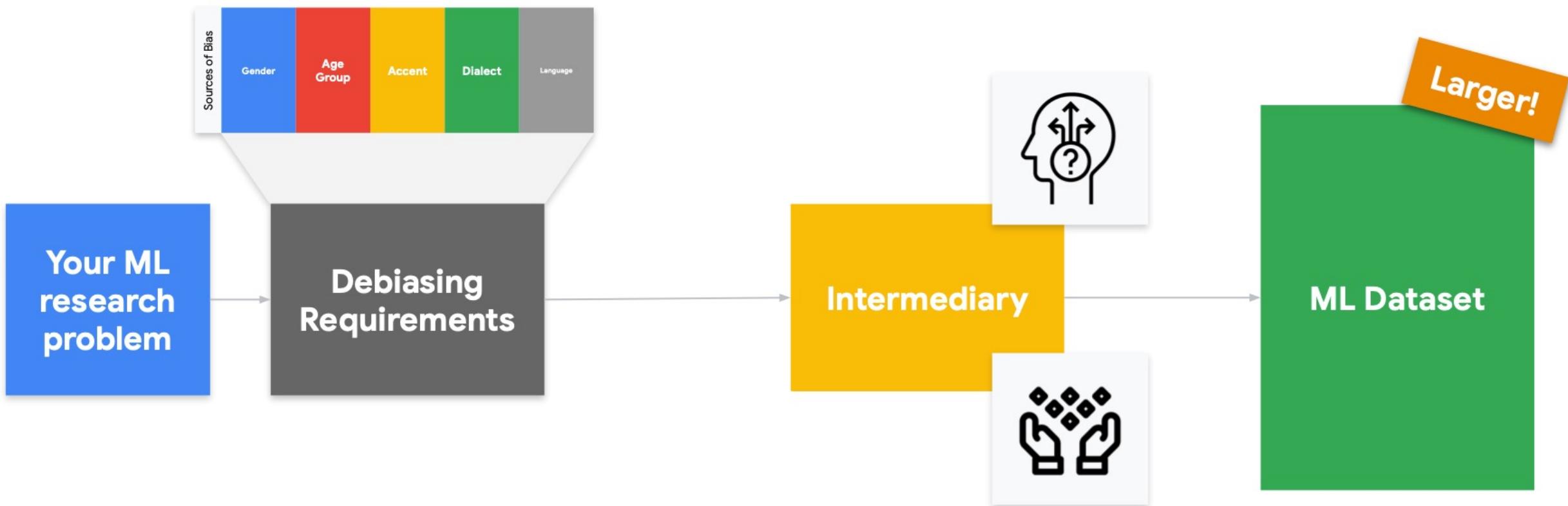
- **Transfer** learning
- **Pretrained** models: your “AI Data Labeling Assistant”
- **Generate** your own data
  - Simulations
  - ML models

# Responsible Data Collection

# Potential Bias in Speech Recognition



# Bias and Market Forces



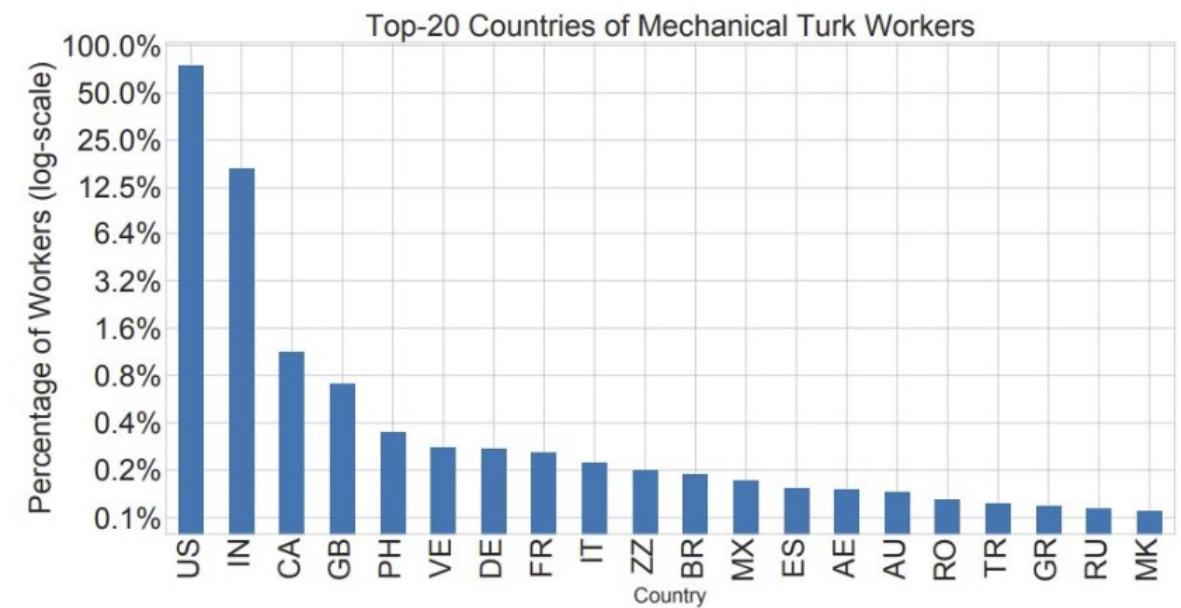
# Data Engineering and Bias

- Amazon Mechanical Turk is a crowdsourcing platform used for labeling data for ML tasks

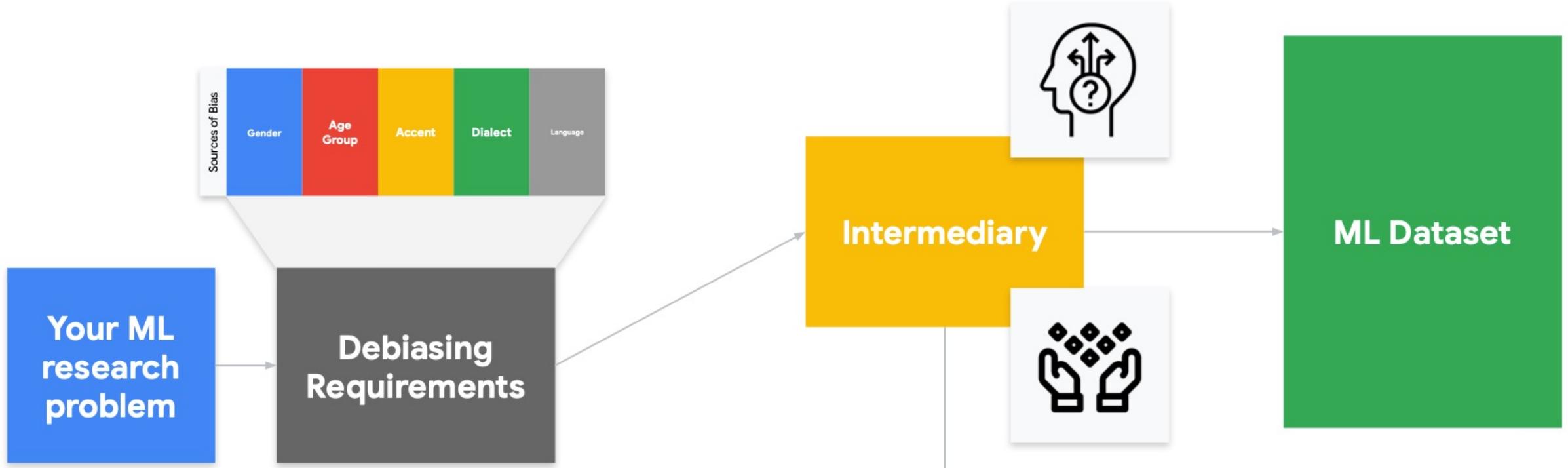


# Data Engineering and Bias

- Amazon Mechanical Turk is a crowdsourcing platform used for labeling data for ML tasks
- Workers are **(not) unbiased**



# Bias and Market Forces



**Intermediaries:** help with collection  
but a *potential source of bias*

# How can we work to avoid bias in our dataset?

Biases are just one of the aspects that we've touched on the data engineering. There are many other aspects.

We should get an accurate model, not just from a moral standpoint, but from **overall quality of experience**.

Building a dataset is quite complicated, if we get the **dataset wrong**, we can get a model that works just well, but not working in **the right way**.

Just because the "**Colab says**" we got a certain accuracy does not mean that it's actually doing its job well from a TinyML application standpoint.

# Reading Material

# Main references

- [Harvard School of Engineering and Applied Sciences - CS249r: Tiny Machine Learning](#)
- [Professional Certificate in Tiny Machine Learning \(TinyML\) – edX/Harvard](#)
- [Introduction to Embedded Machine Learning \(Coursera\)](#)
- [Text Book: "TinyML" by Pete Warden, Daniel Situnayake](#)

I want to thank [Shawn Hymel](#) and [Edge Impulse](#), [Laurence Moroney from Google](#), [Harvard professor Vijay Janapa Reddi](#), Ph.D. student [Brian Plancher](#) and their staff for preparing the excellent material on TinyML that is the basis of this course at UNIFEI.

The IESTI01 course is part of the [TinyML4D](#), an initiative to make TinyML education available to everyone globally.

**Thanks**  
And stay safe!

