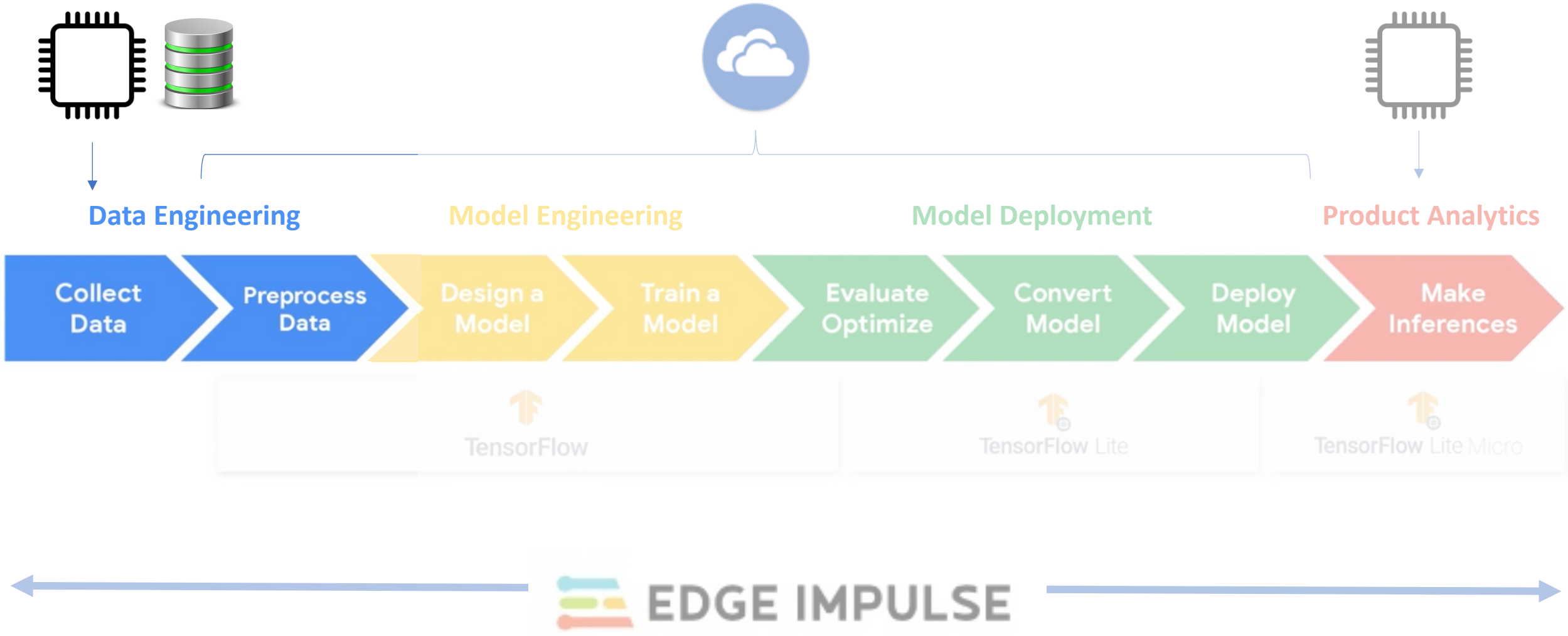
# IESTI01 - TinyML

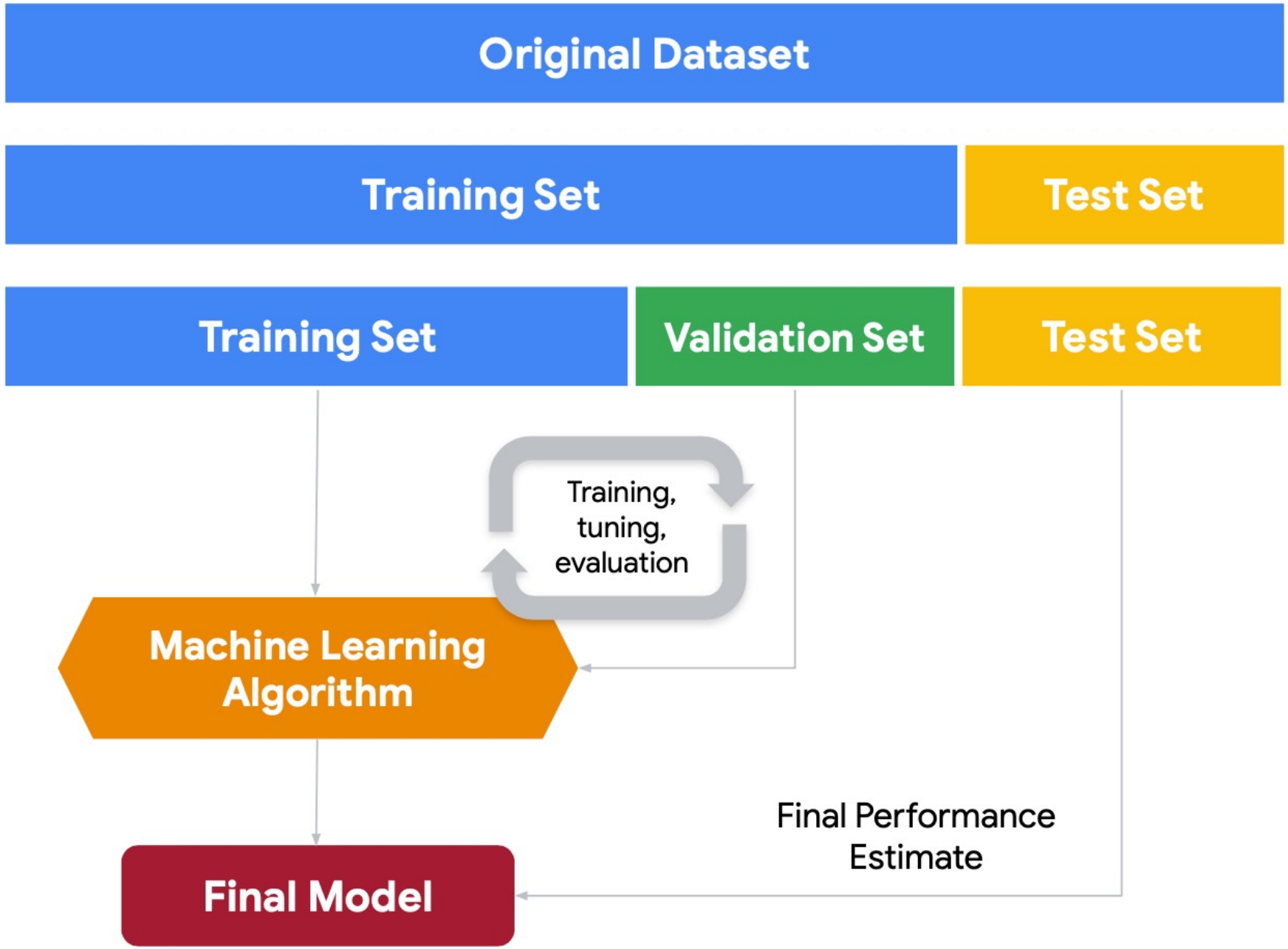
## Data Engineering for TinyML

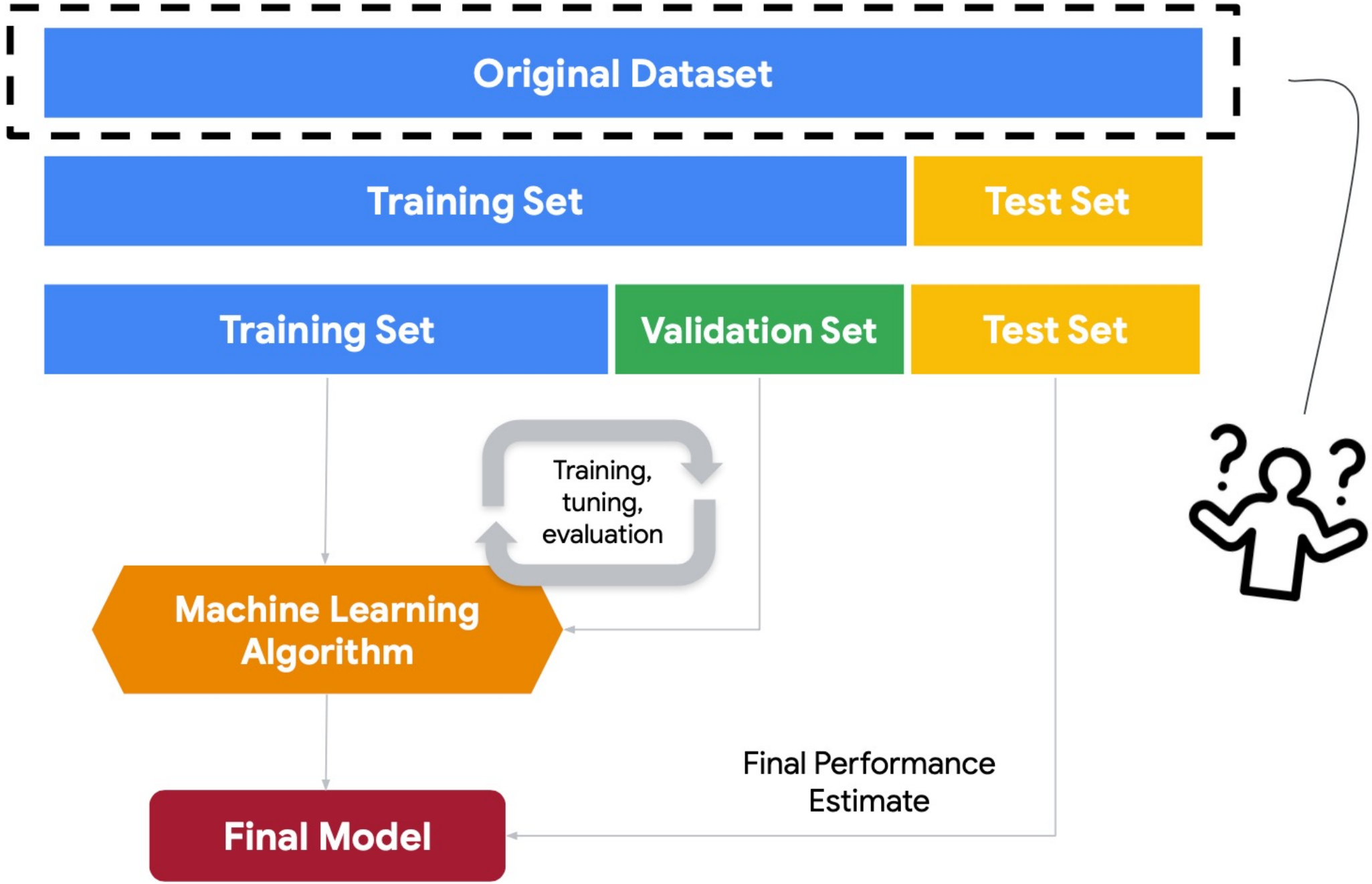
Prof. Marcelo Rovai

June 23rd, 2021

# What is data Engineering?


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
TensorFlow
TensorFlow Lite
TensorFlow Lite Micro
EDGE IMPULSE


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


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

# 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*?

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

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

# Licenses



# Data Engineering

**Requirements** → **Gathering**

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

- People
- Collection
- Labeling
- **Data sources**

# Data **sources**

- Sensors
- Crowdsourcing
- Product users
- Paid contributors

# Data Engineering

**Requirements** → **Gathering** → **Refinement**

- 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


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

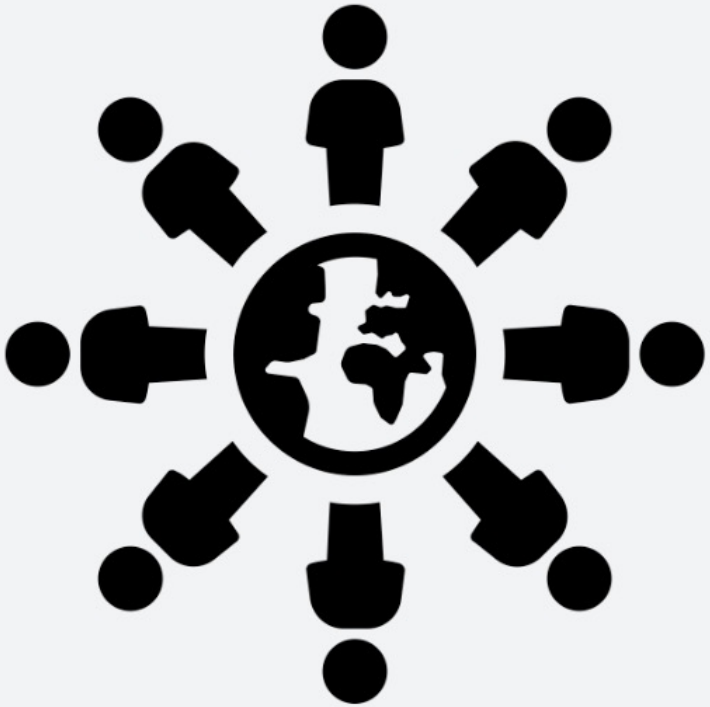
- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- **Versioning**

# Your dataset will ***evolve***

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



# Data Engineering


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

- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

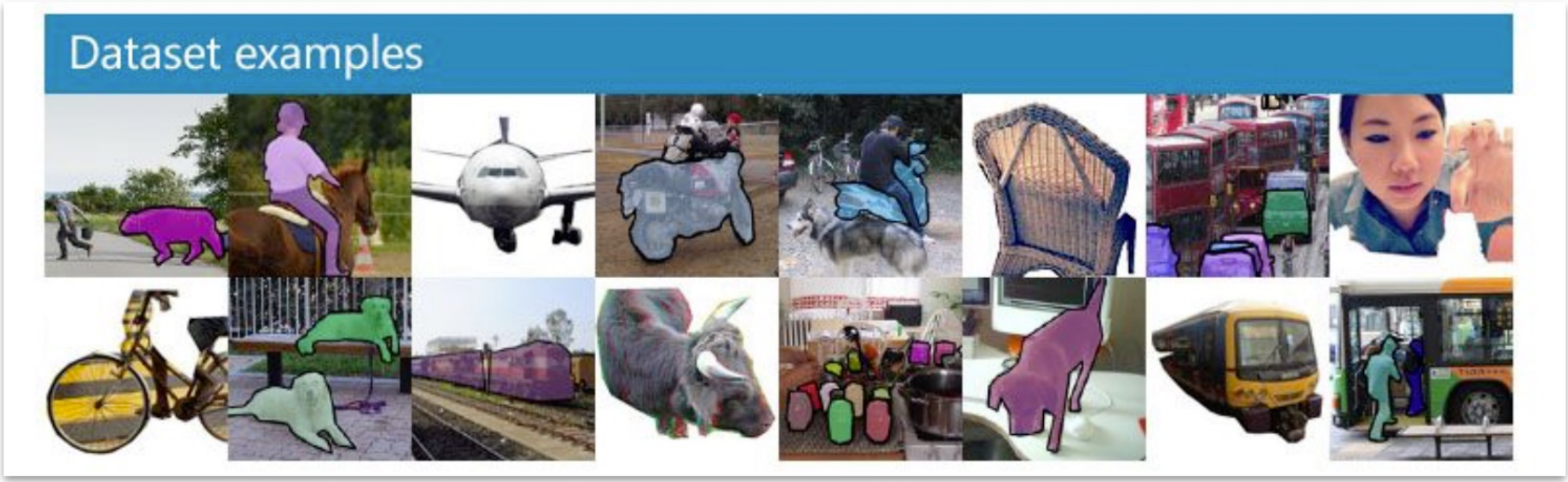
- Storage
- Security
- Errors
- Versioning

Datasets require
**significant effort**

# **Classifying** Images

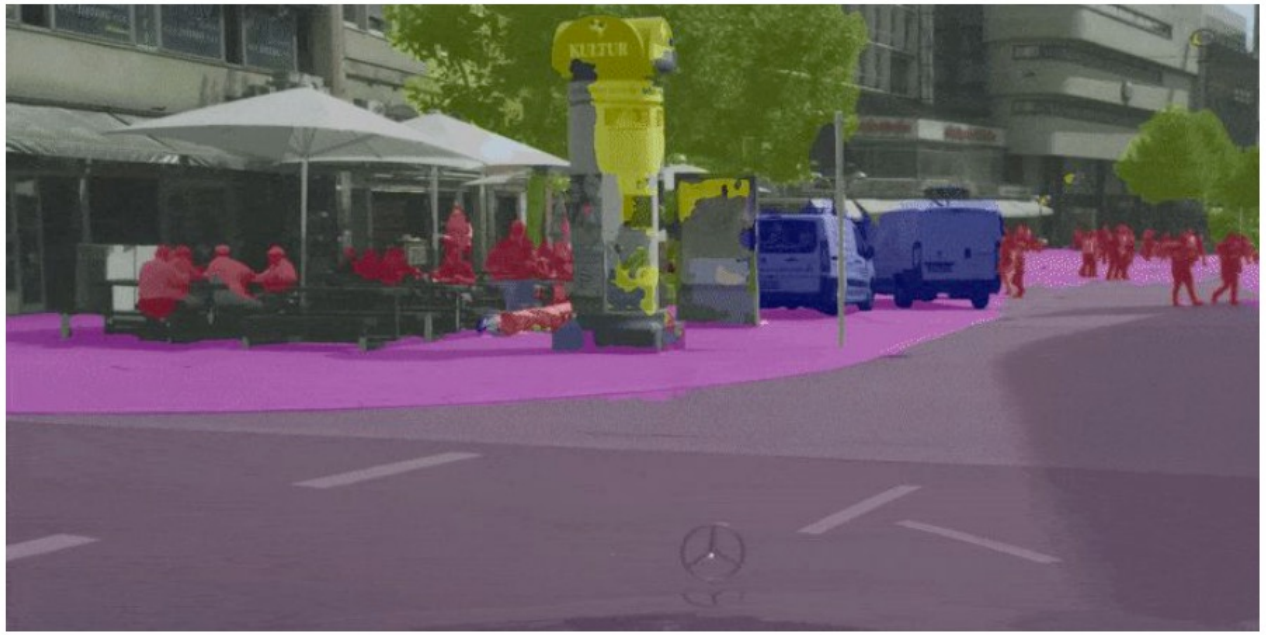
# **Detecting** Objects

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

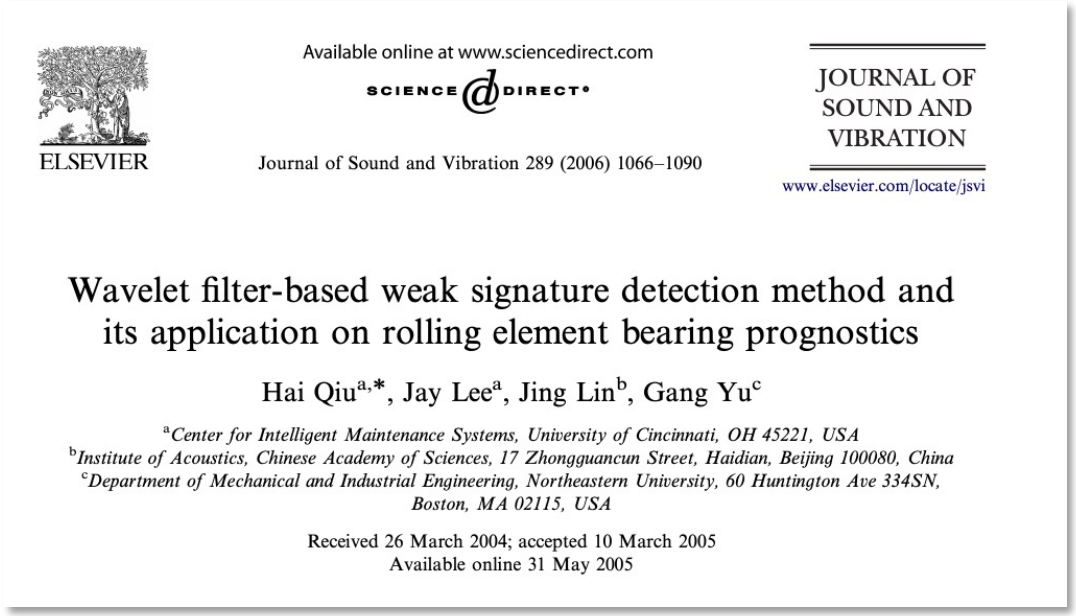


# **Classifyng** Video Motion

- **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
SCIENCE DIRECT®

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

JOURNAL OF SOUND AND VIBRATION

www.elsevier.com/locate/jsvi

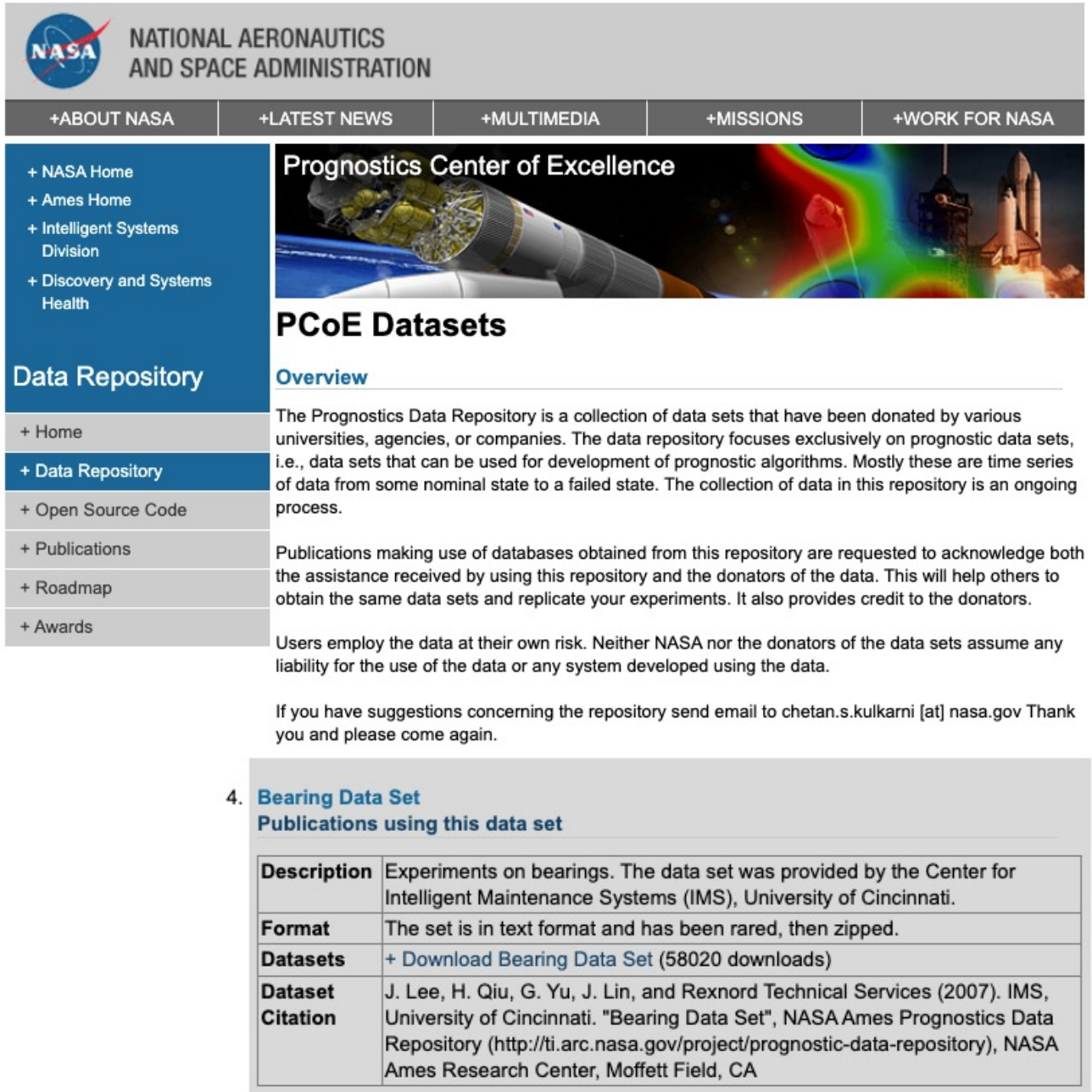
Wavelet filter-based weak signature detection method and its application on rolling element bearing prognostics

Hai Qiu[a,*], Jay Lee[a], Jing Lin[b], Gang Yu[c]

[a]*Center for Intelligent Maintenance Systems, University of Cincinnati, OH 45221, USA*
[b]*Institute of Acoustics, Chinese Academy of Sciences, 17 Zhongguancun Street, Haidian, Beijing 100080, China*
[c]*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 chetan.s.kulkarni [at] nasa.gov Thank you and please come again.

4. **Bearing Data Set**
**Publications using this data set**

| | |
|---|---|
| **Description** | Experiments on bearings. The data set was provided by the Center for Intelligent Maintenance Systems (IMS), University of Cincinnati. |
| **Format** | The set is in text format and has been rared, then zipped. |
| **Datasets** | + Download Bearing Data Set (58020 downloads) |
| **Dataset Citation** | 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 (http://ti.arc.nasa.gov/project/prognostic-data-repository), NASA Ames Research Center, Moffett Field, CA |

https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/

# 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

April 2018

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

# Data Engineering


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

- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

# Data Engineering


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

- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

# 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


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

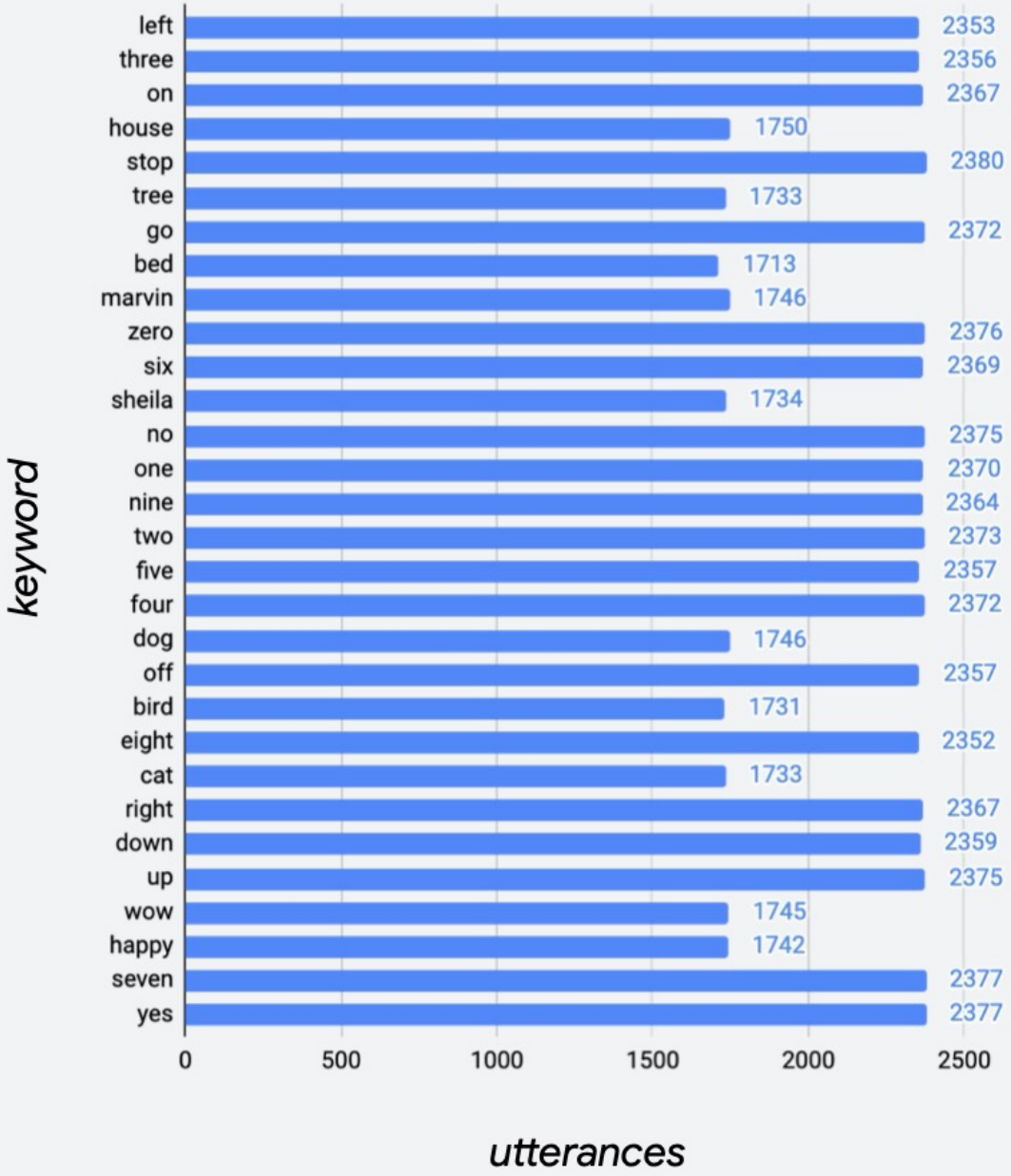
- **Data sources**
- **People**
- **Collection**
- **Labeling**

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

# 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


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

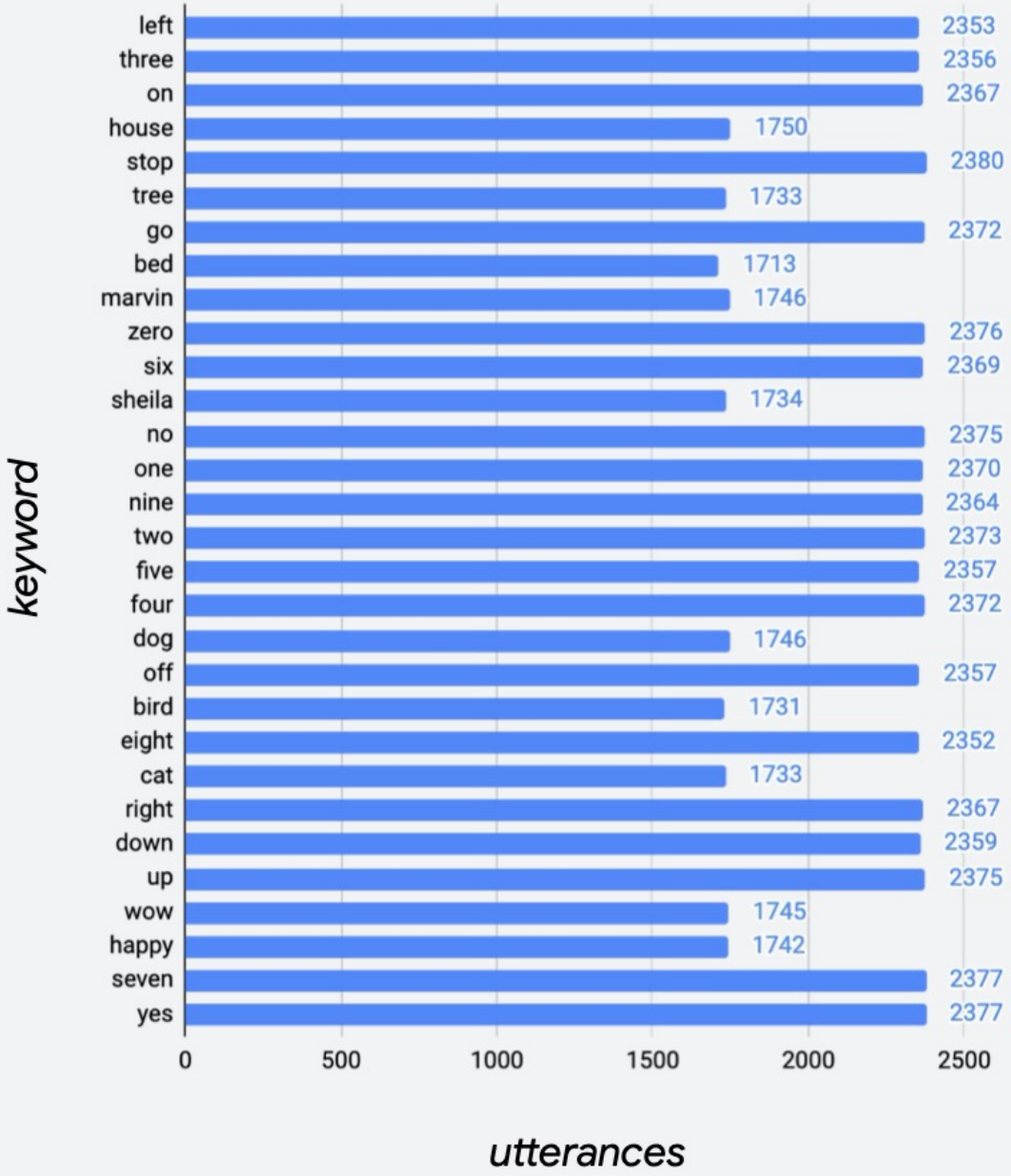
- Data sources
- People
- Collection
- Labeling

- **Processing**
- **Validation**
- **Augmentation**

- Storage
- Security
- Errors
- Versioning

# 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


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

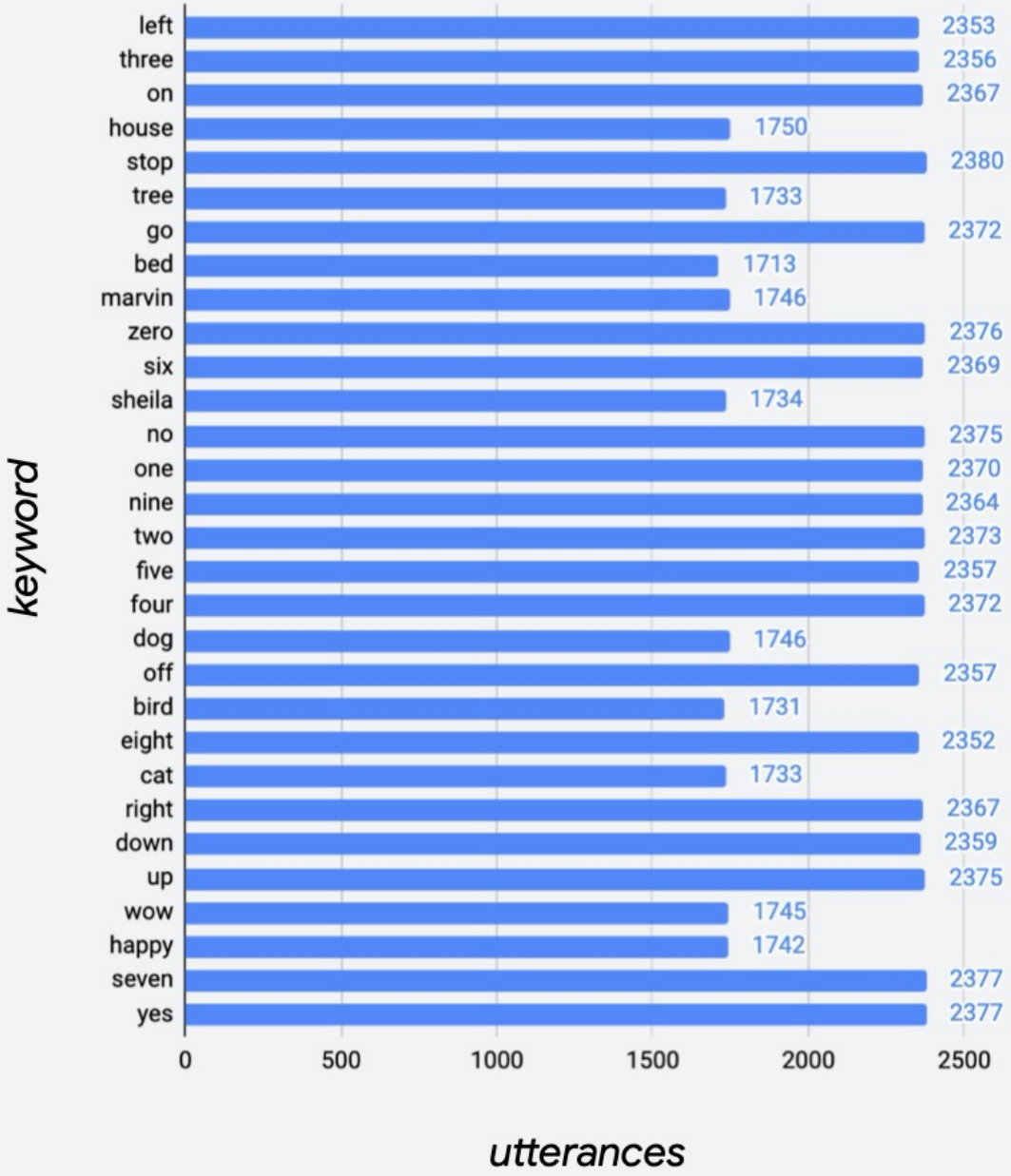
- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

- **Storage**
- **Security**
- **Errors**
- **Versioning**

# 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**?



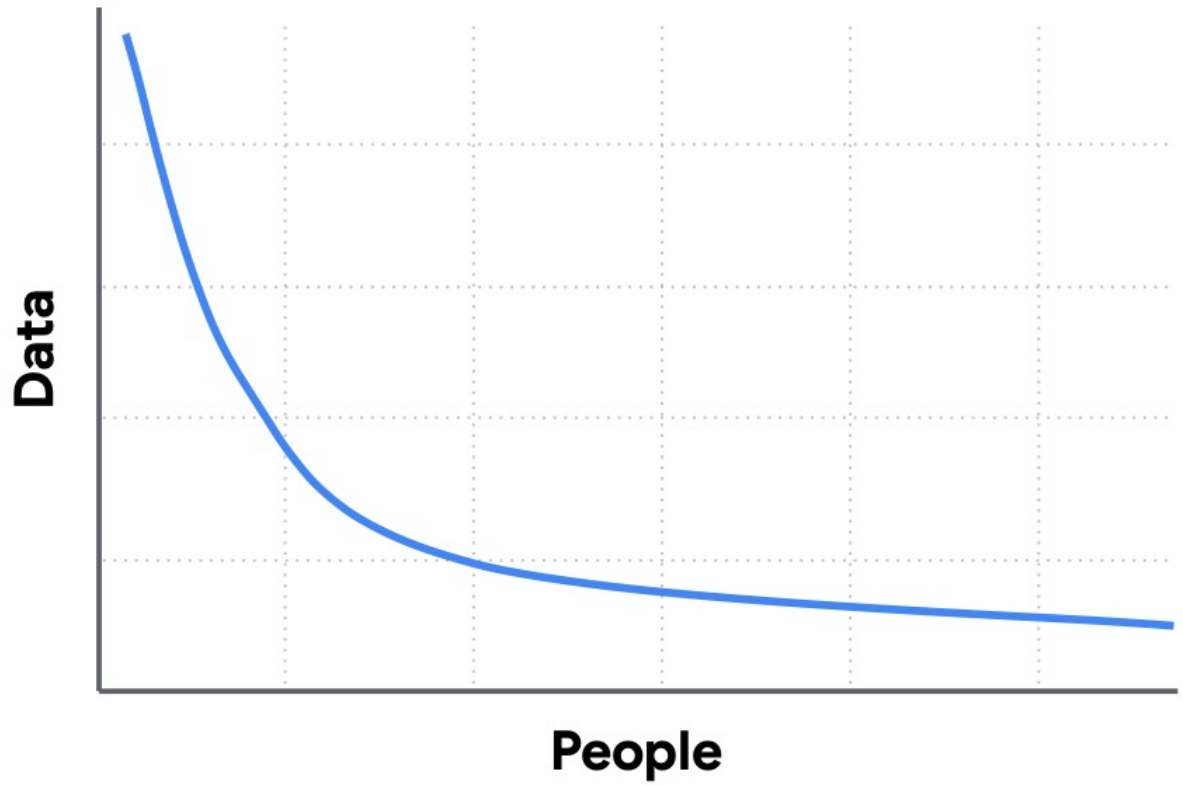
# Why is **wide availability** of data ***important***?

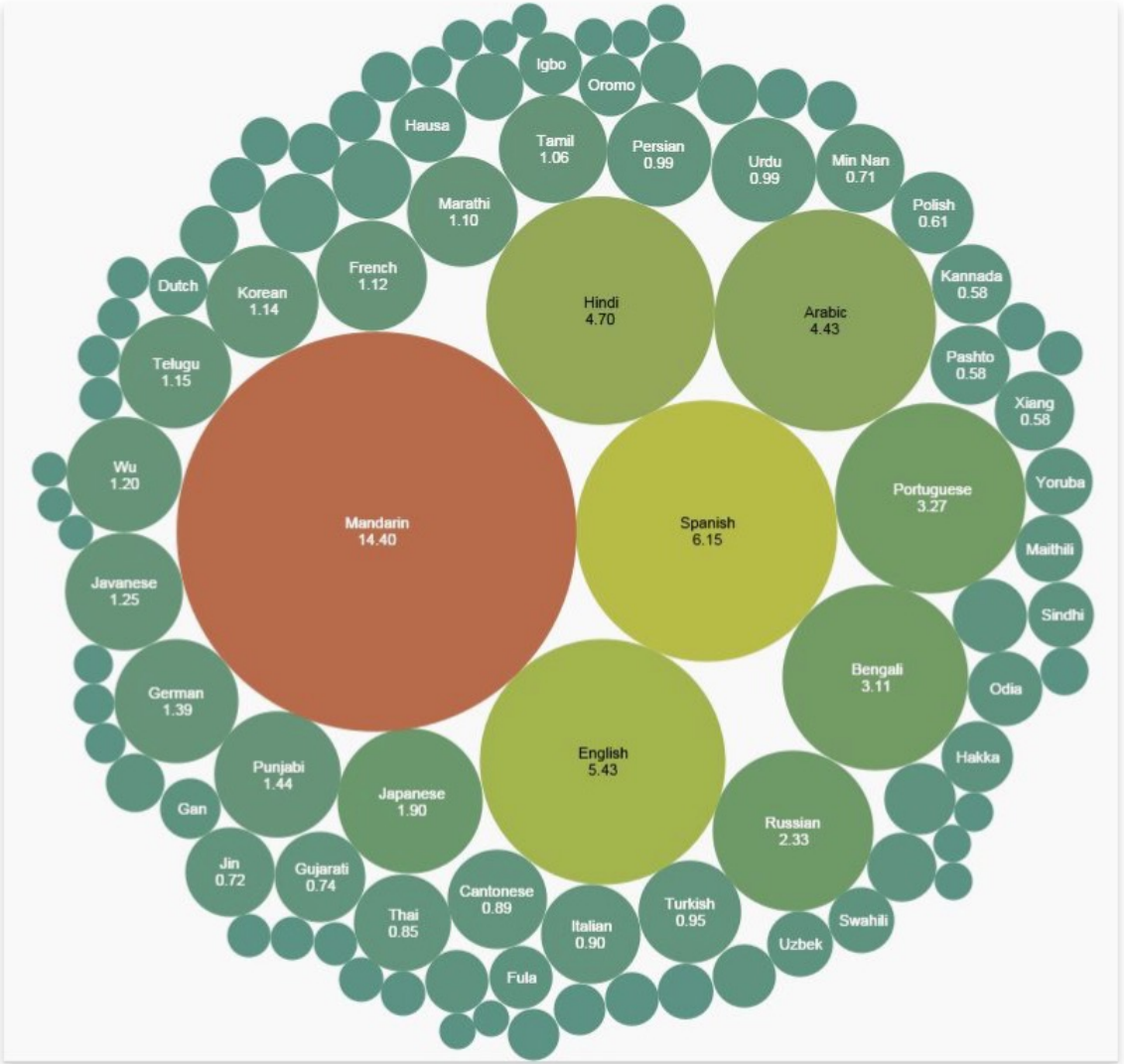
| | |
|---|---|
| **Compare** | **Without a standard dataset, no easy way to compare models** |
| **Benchmark** | **Good benchmarks: critical under *TinyML* constraints!<br>Not just accuracy: power, memory, latency** |
| **Improve** | **Speech Commands: a standard training &<br>evaluation dataset for each new KWS technique** |

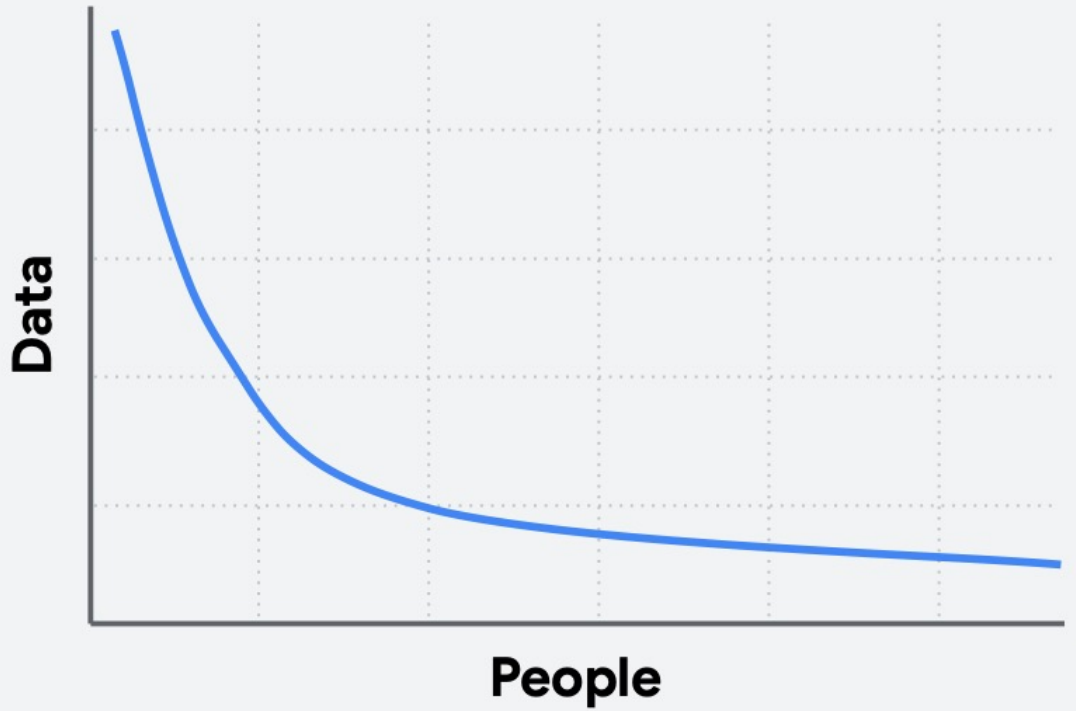
Researches using the same data: *"apple-to-apples" evaluation*

# Crowdsourcing Data

## for the Long Tail


Data
People


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
Telugu
1.15
Pashto
0.58
Xiang
0.58
Wu
1.20
Mandarin
14.40
Spanish
6.15
Portuguese
3.27
Yoruba
Maithili
Javanese
1.25
Sindhi
German
1.39
Bengali
3.11
Odia
Punjabi
1.44
English
5.43
Hakka
Japanese
1.90
Gan
Russian
2.33
Jin
0.72
Gujarati
0.74
Cantonese
0.89
Turkish
0.95
Thai
0.85
Italian
0.90
Swahili
Uzbek
Fula


Data
People

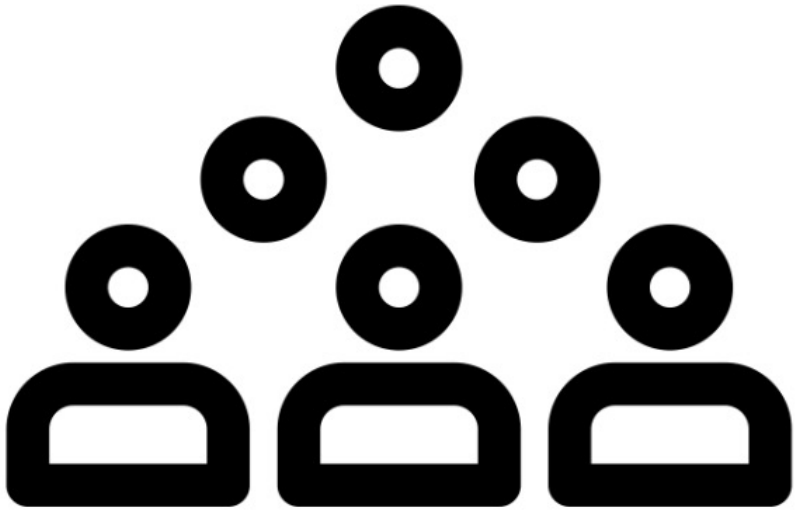
# **Cost** Model v. **Community** Model?

Limited Scale

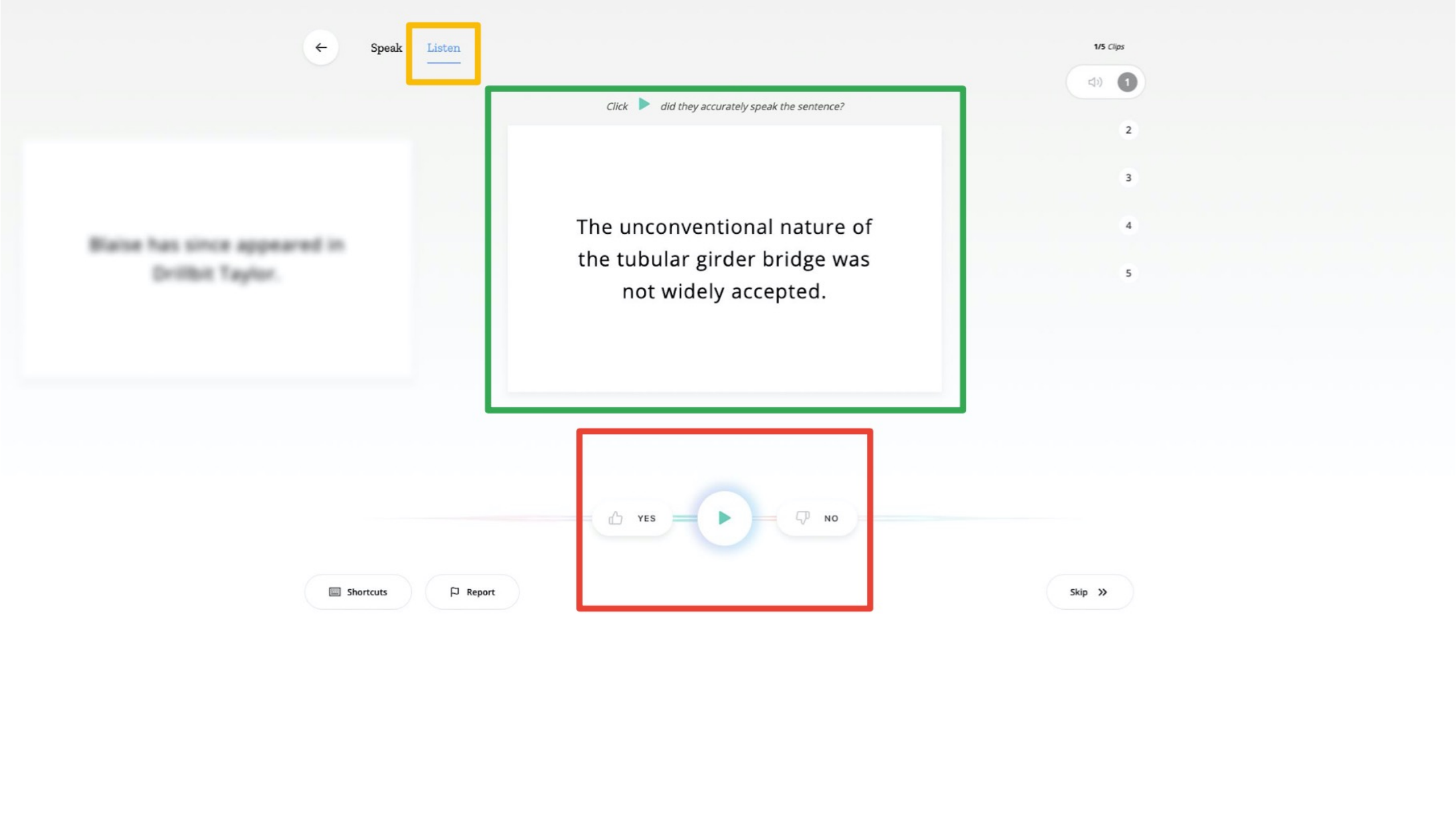
Social Good

# Common Voice

- **Crowdsourcing** platform

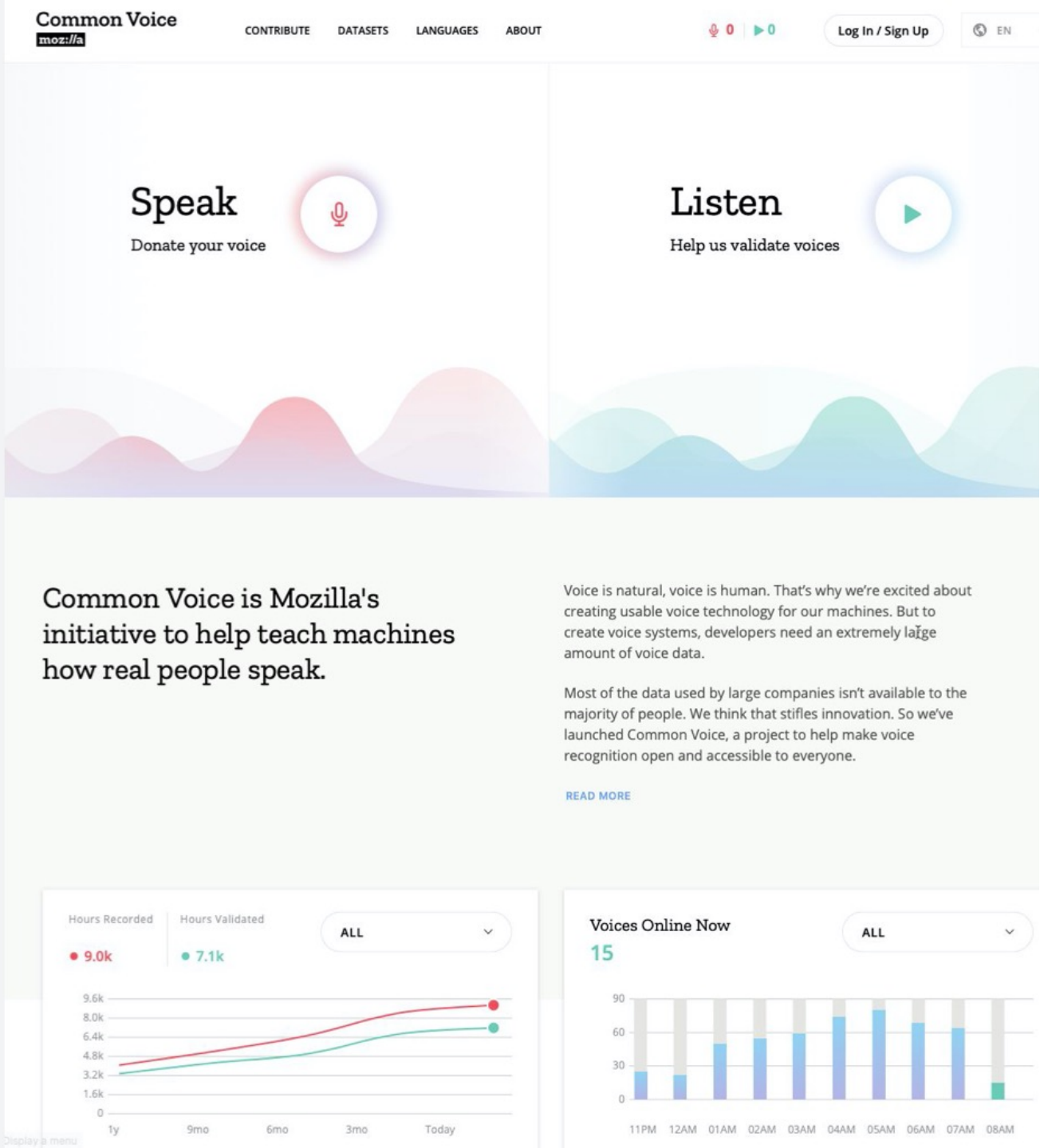


https://commonvoice.mozilla.org/en


Speak
Listen
1/5 Clips
1
2
3
4
5
Click
did they accurately speak the sentence?
The unconventional nature of the tubular girder bridge was not widely accepted.
YES
NO
Shortcuts
Report
Skip

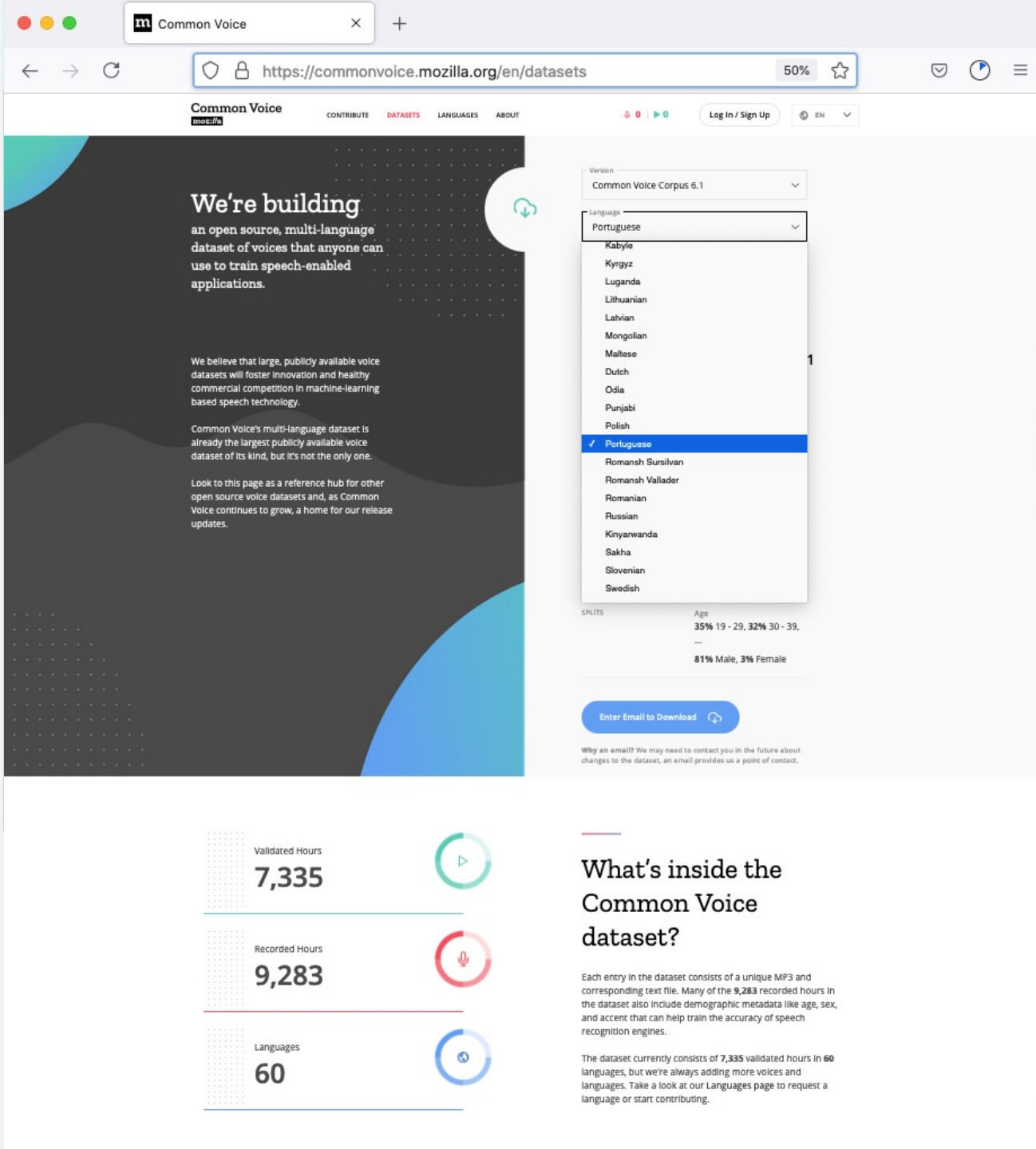
# Common Voice

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



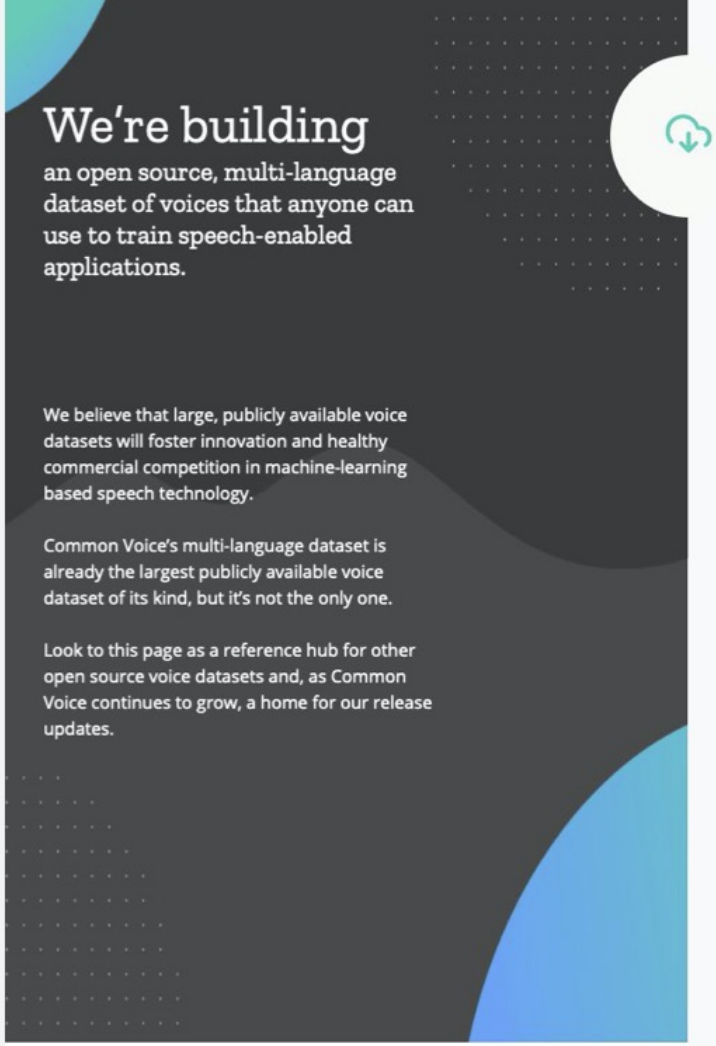
# Common Voice

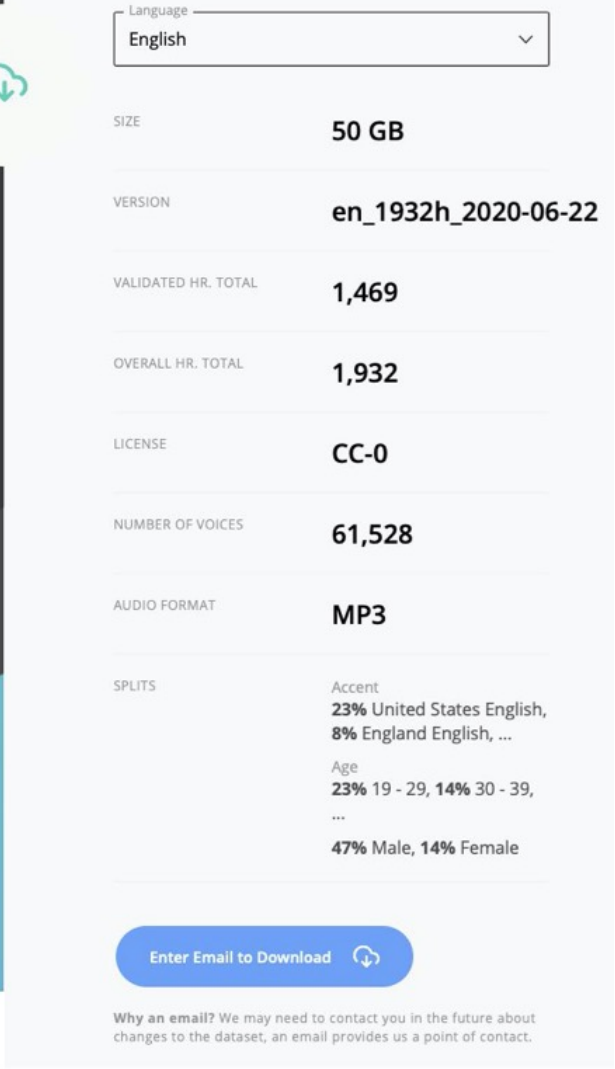
- Crowdsourcing platform
- Over 50,000 volunteers
- **54 different languages**

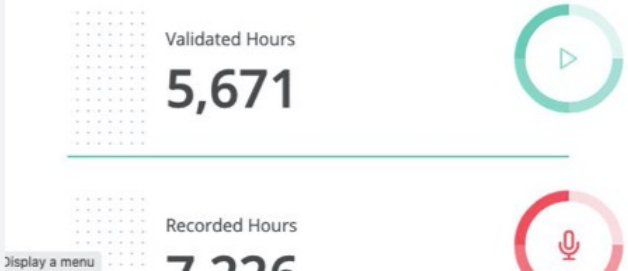


# Common Voice

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



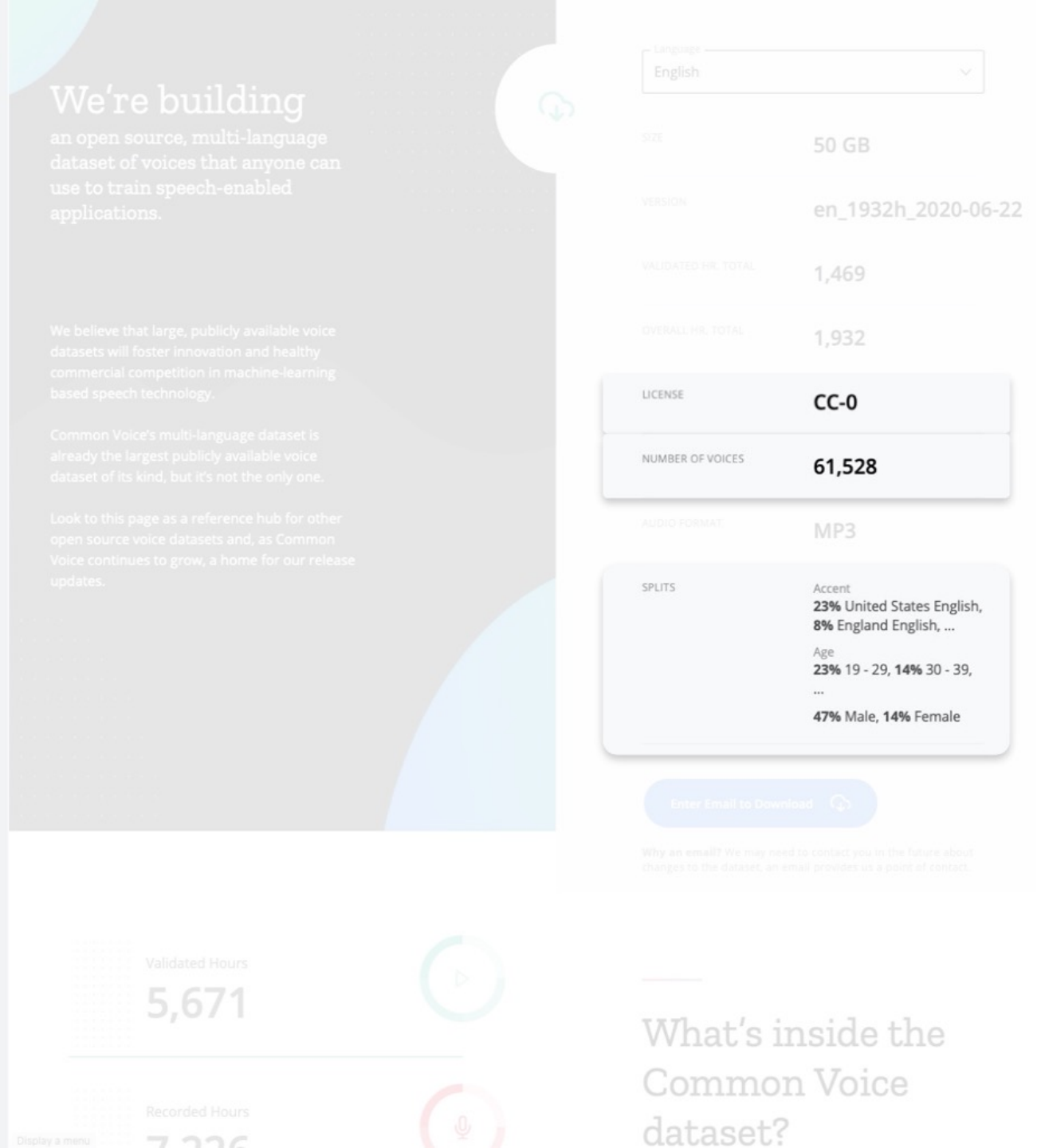

## What's inside the Common Voice dataset?

# 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

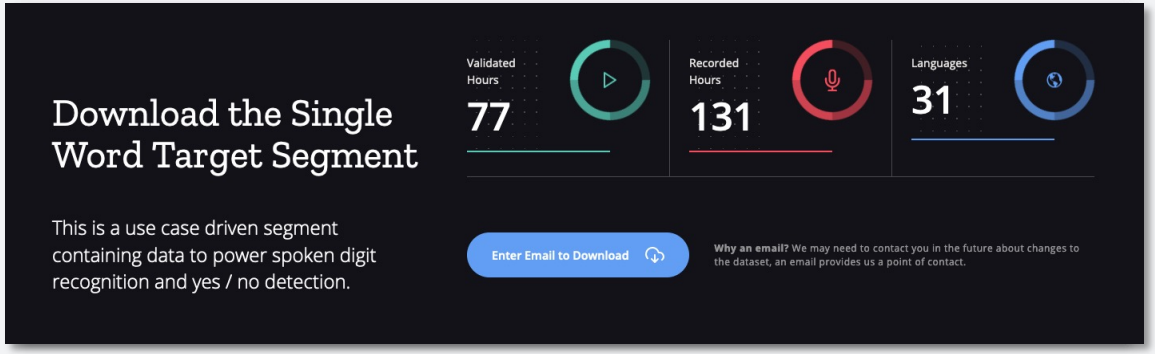
- Permissive license
- Many contributors
- Comes with metadata



# Single Word Target Segment

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

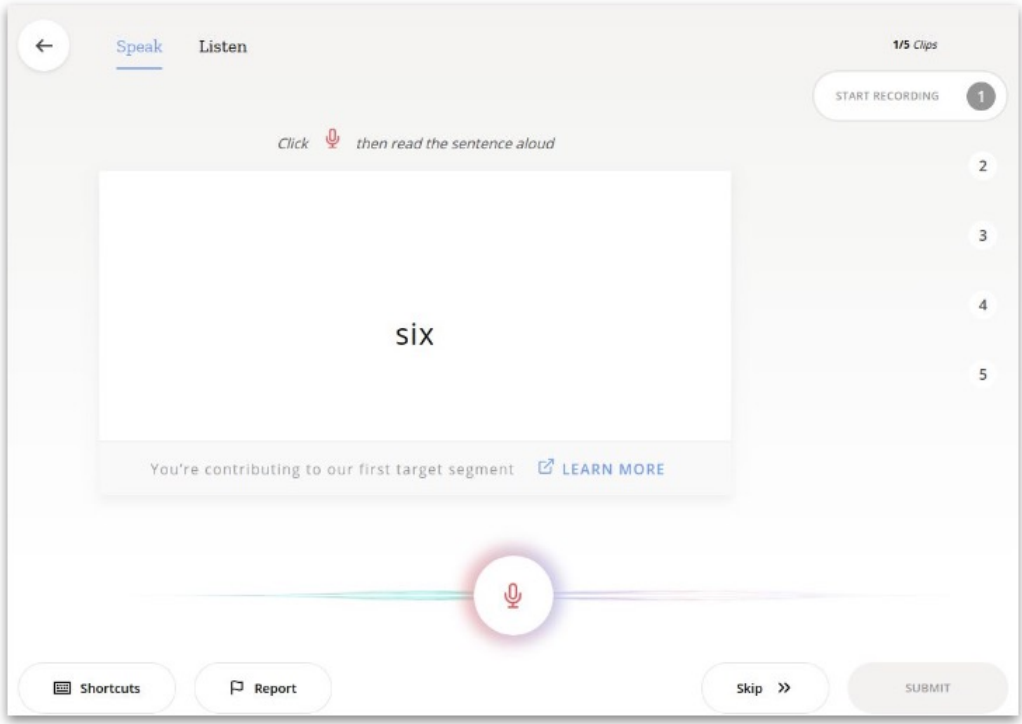
- “Yes” // “no”
- “hey” & “Firefox”
- **digits** 0–9




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**!

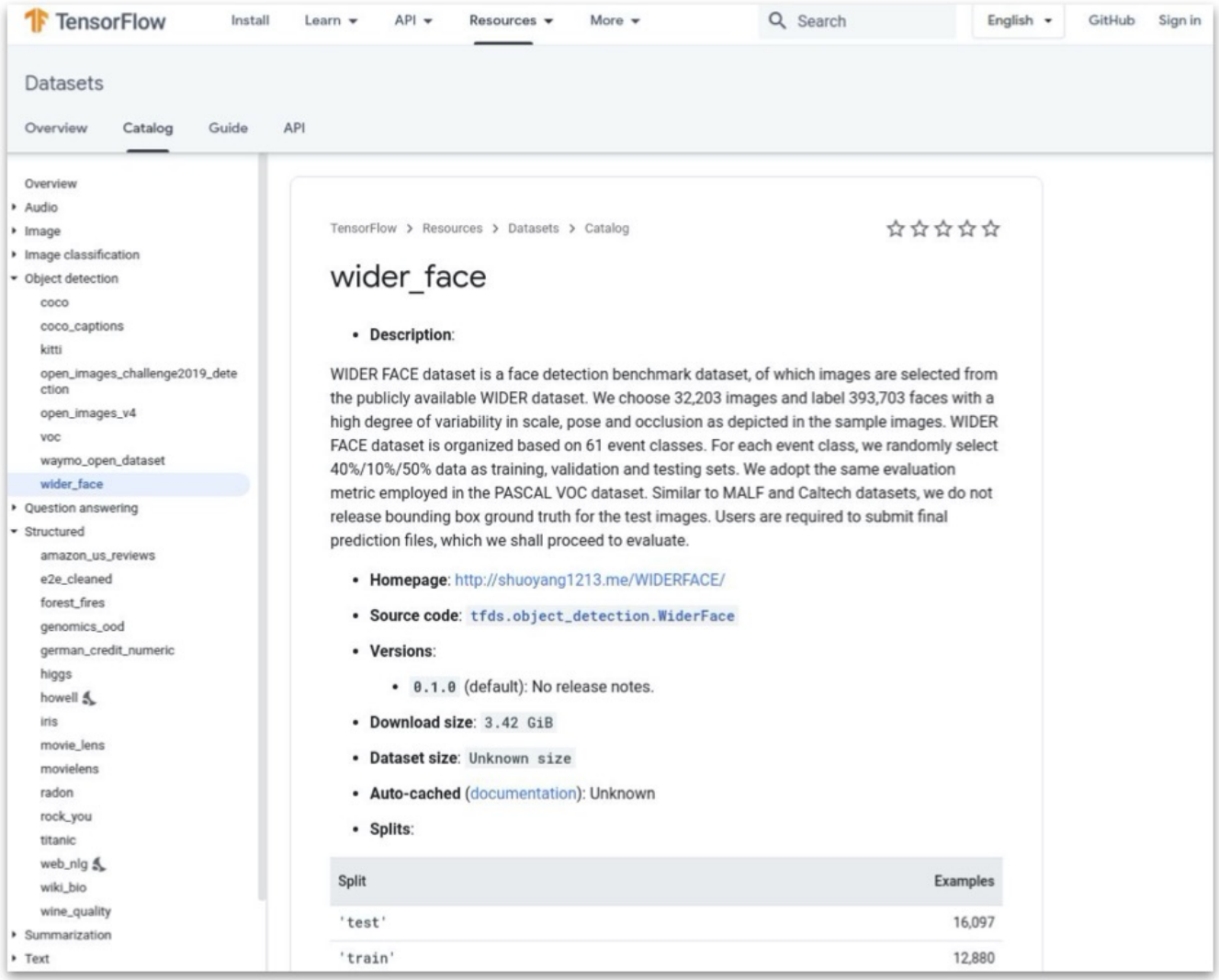
- Can we ***reuse*** existing data?

**What's available?**

**What's missing?**

# TensorFlow
## Datasets Catalog

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

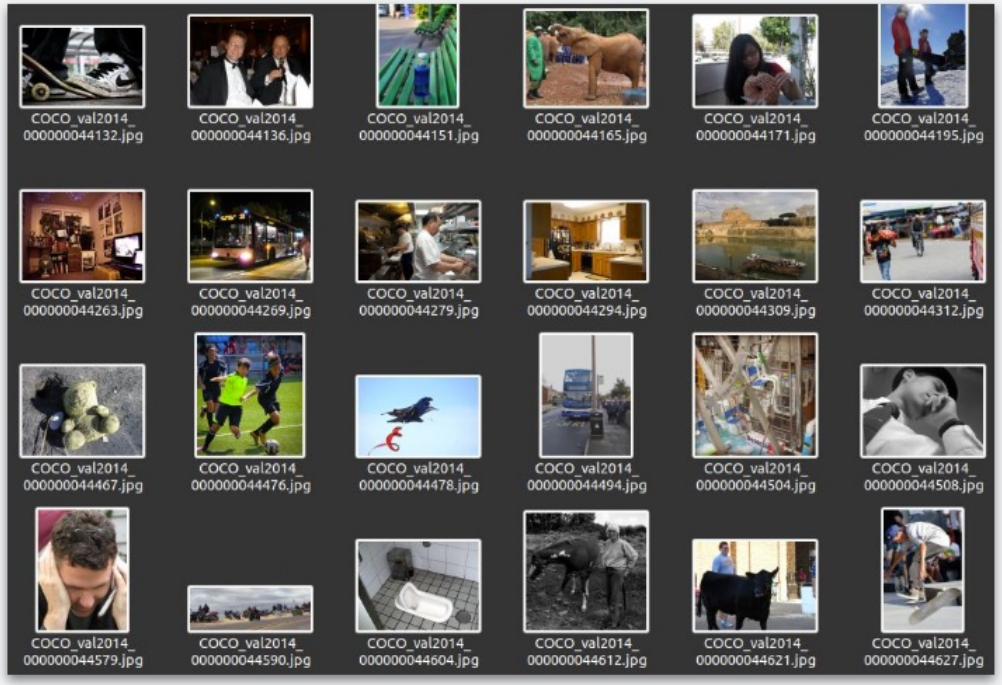



Always check license!

# 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

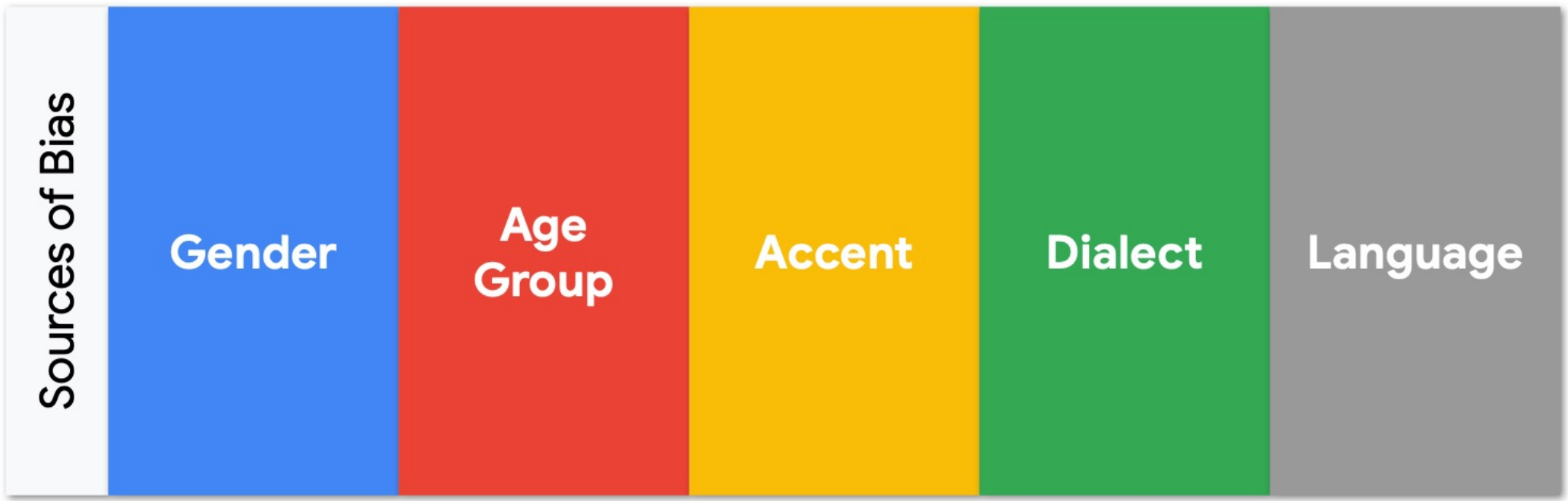


# Don’t *learn* from scratch

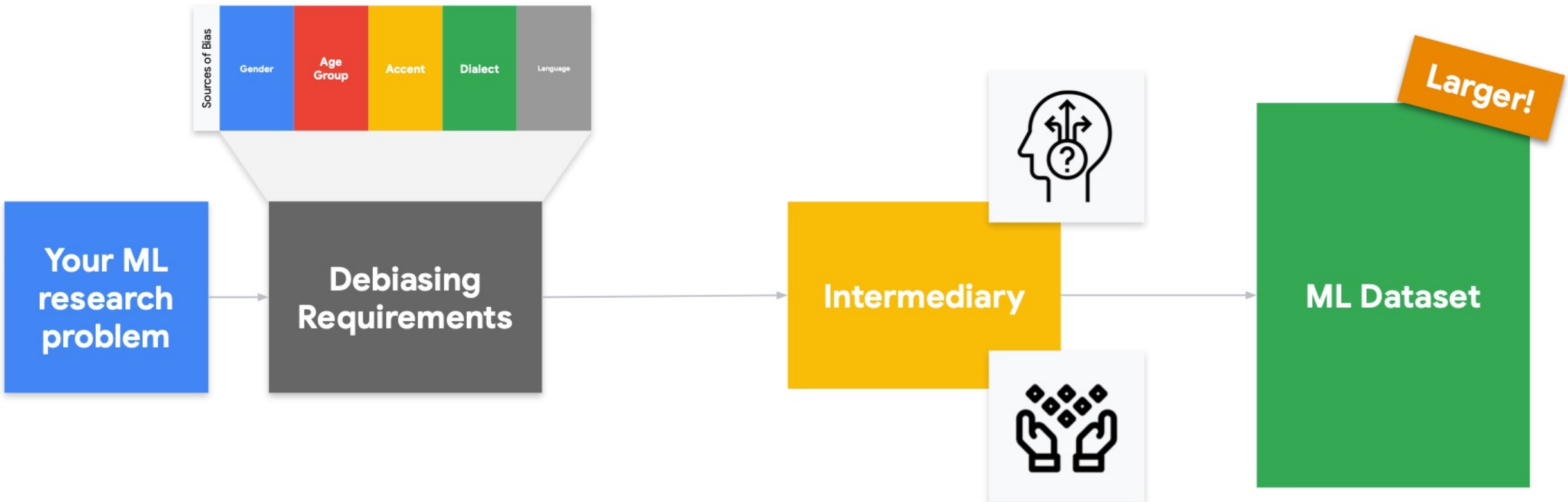
- **Transfer** learning
- **Pretrained** models: your “AI Data Labeling Assistant”
- **Generate** your own data
  - Simulations
  - ML models

# Responsible Data Collection

# Potential **Bias** in Speech Recognition



# Bias and **Market Forces**



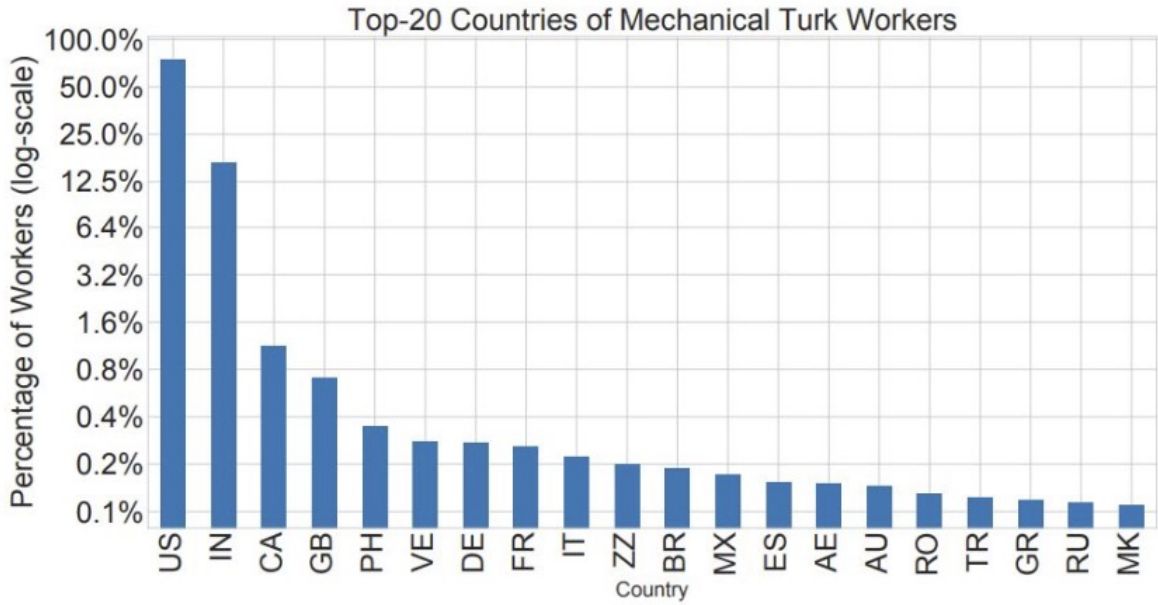
# Data Engineering and Bias

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

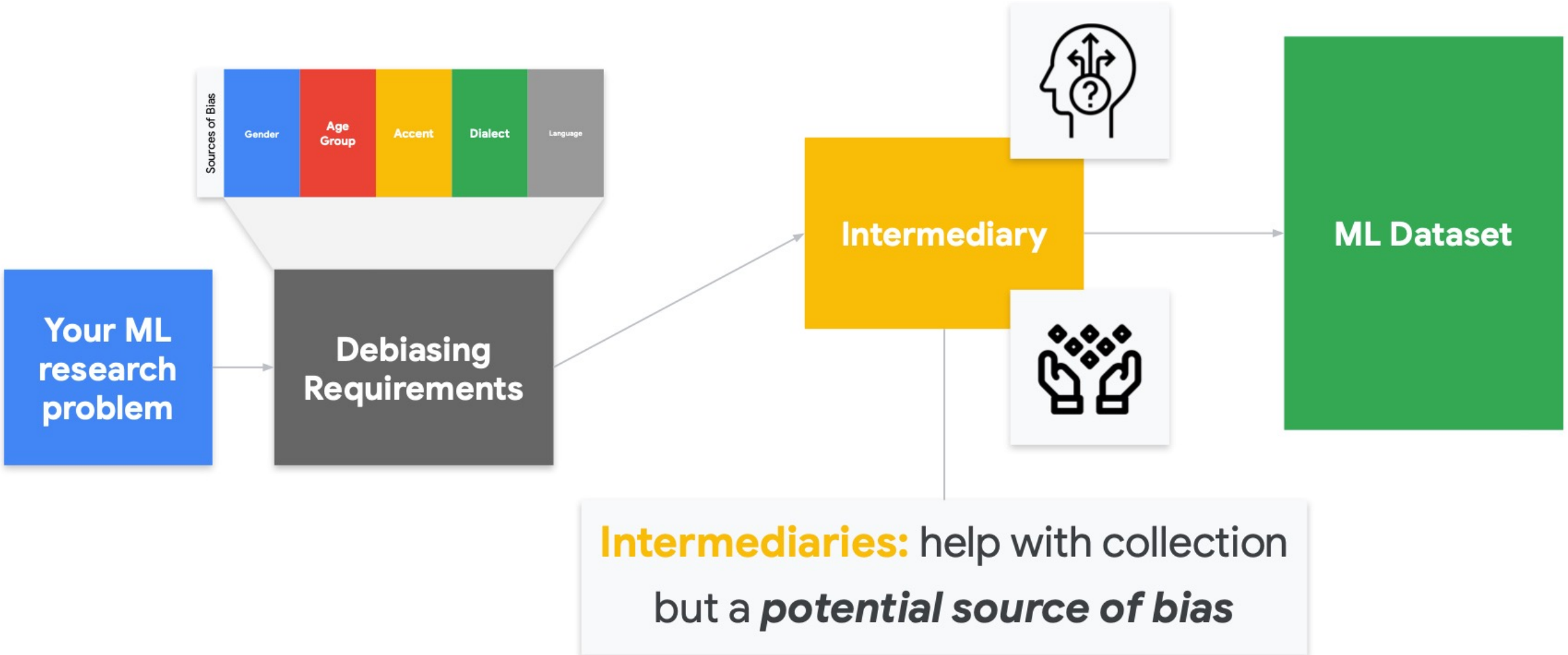
# Data Engineering and Bias

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

- **Workers are (not) unbiased**



# Bias and **Market Forces**



# 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!