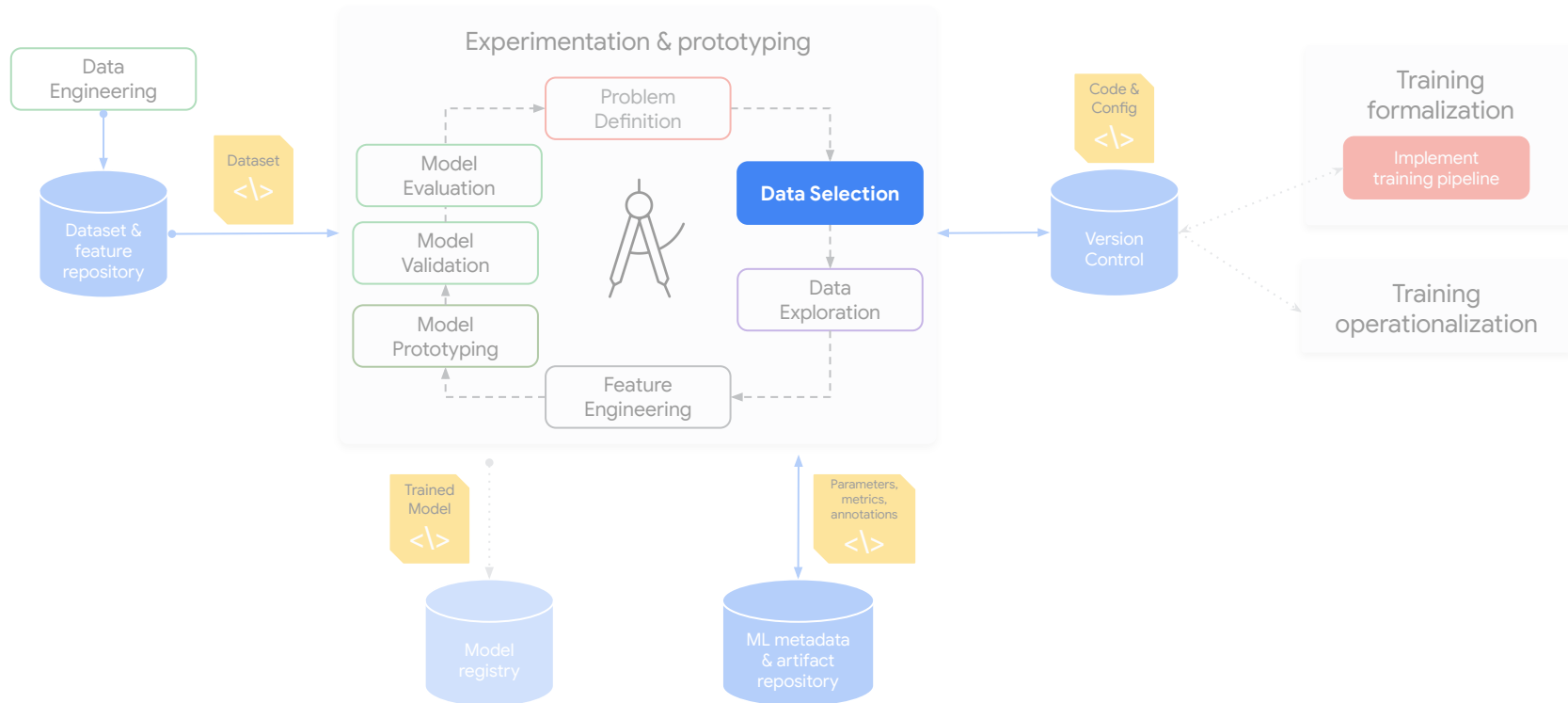


ML Development: Data Selection



MLOps: ML Development



The MLOps Personas



ML
Engineer



ML
Researcher



Data
Scientist



Data
Engineer



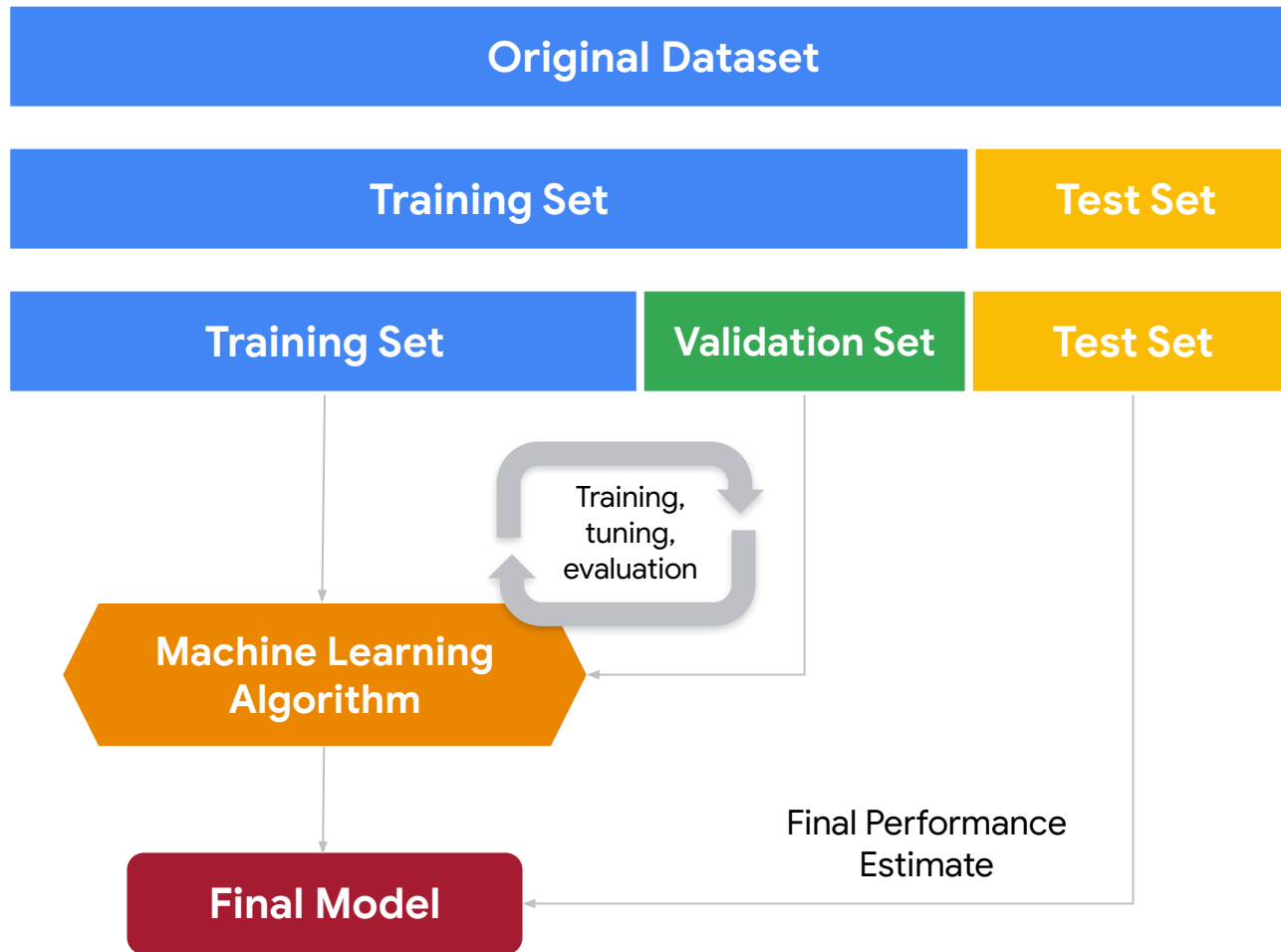
Software
Engineer

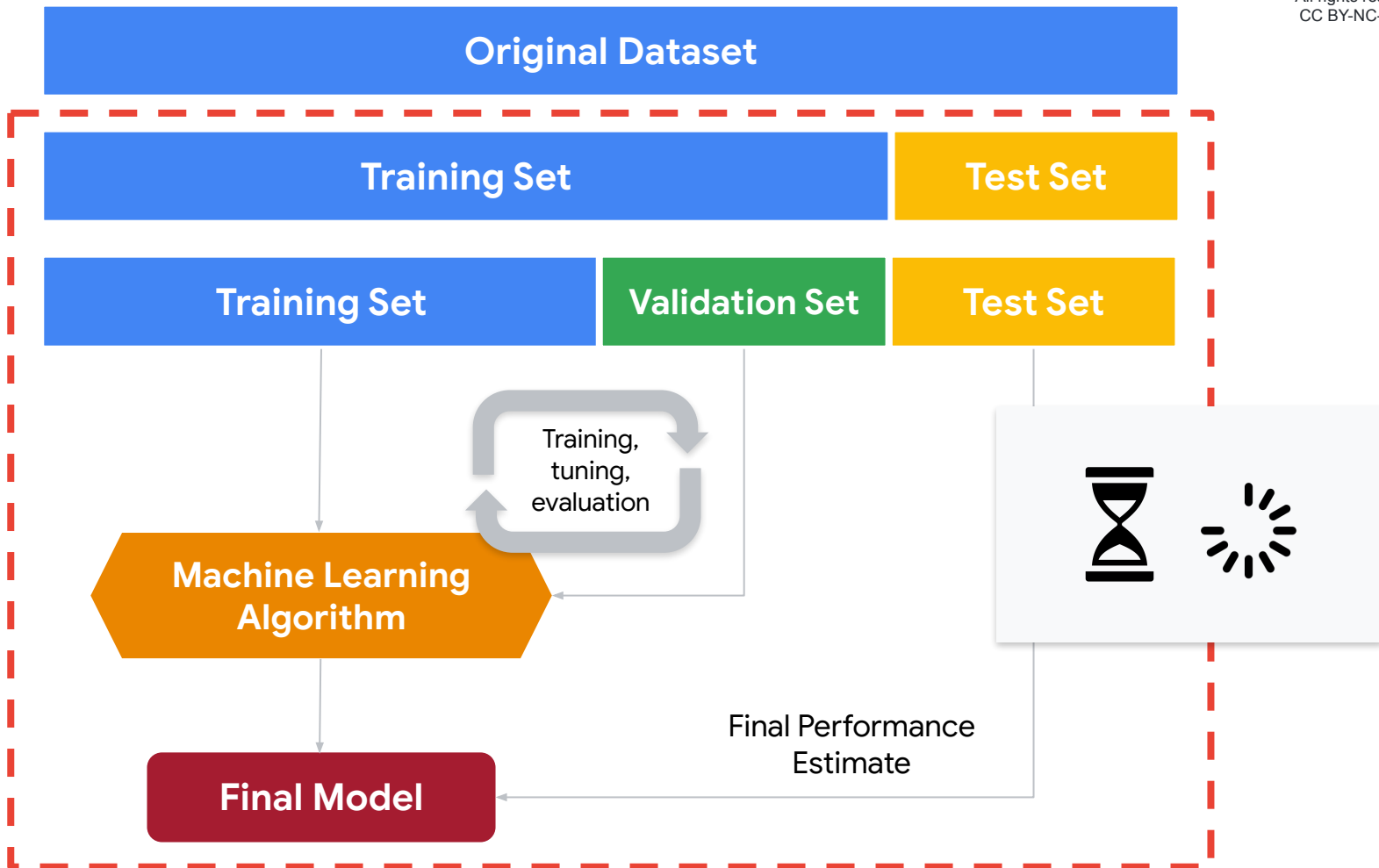


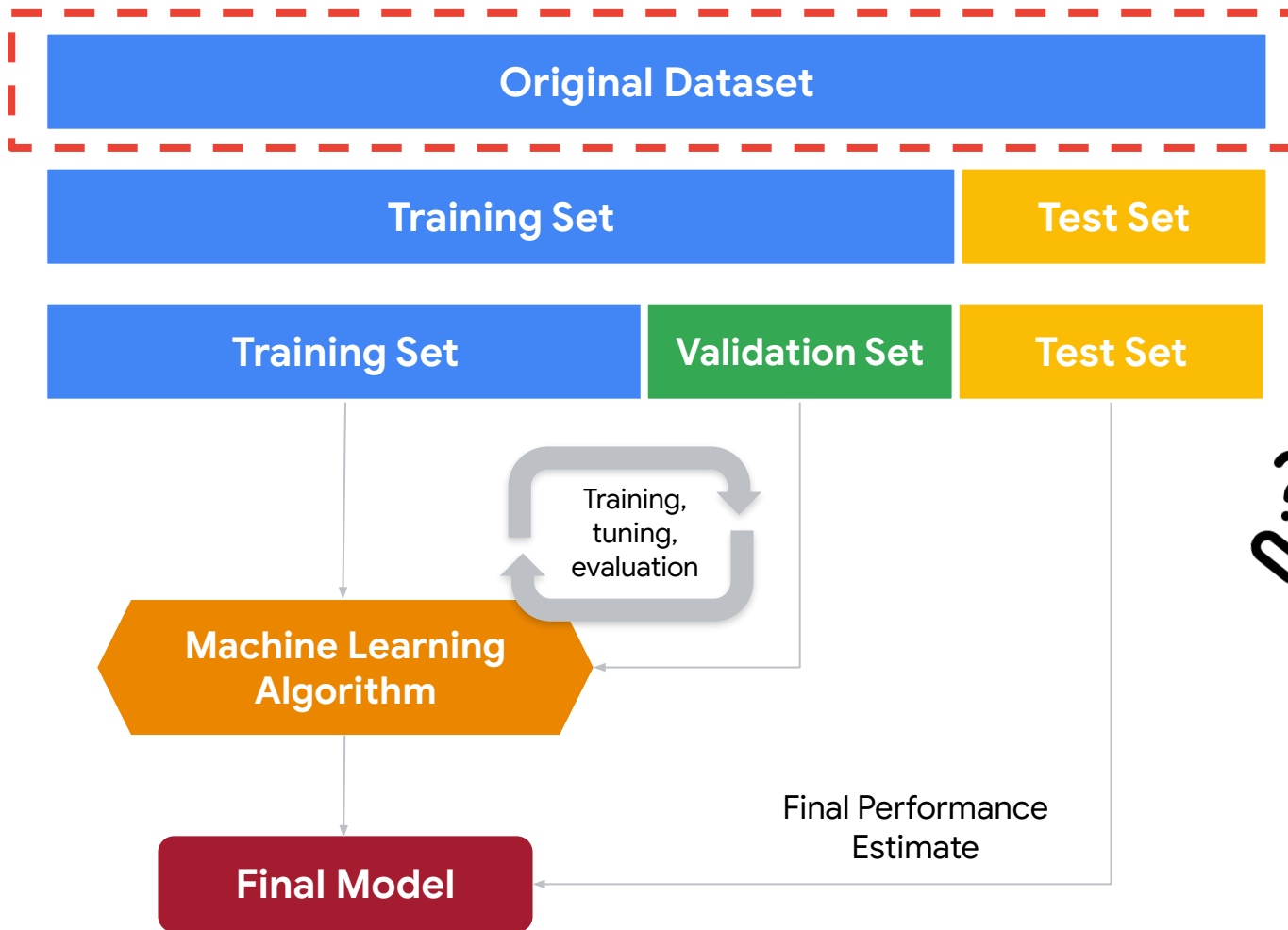
DevOps



Business
Analyst







Dataset Sources

- What **relevant datasets** are available?

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

TensorFlow Datasets Catalog

The screenshot shows the TensorFlow Datasets Catalog interface. The left sidebar lists various dataset categories: 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, a2a_cleaned, forest_fires, genomics_poa, german_credit_numeric, higgs, howell, iris, movie_lens, movielens, radon, rock_you, titanic, web_nlg, wiki_bio, wine_quality, Summarization, and Text. The main content area displays the details for the **wider_face** dataset. It includes a description, homepage, source code, versions, download size, dataset size, auto-cached status, and splits.

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 MAF 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

Dataset Sources

- What **relevant datasets** are available?

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

Mozilla Common Voice

The screenshot shows the Mozilla Common Voice website. The header includes navigation links: CONTRIBUTE, DATASETS, LANGUAGES, and ABOUT. There are also links for Log In / Sign Up and a language selector set to EN. The main content area features a large heading "We're building an open source, multi-language dataset of voices that anyone can use to train speech-enabled applications." followed by two paragraphs of text. The first paragraph states: "We believe that large, publicly available voice datasets will foster innovation and healthy commercial competition in machine-learning based speech technology." The second paragraph states: "Common Voice's multi-language dataset is already the largest publicly available voice dataset of its kind, but it's not the only one." Below this is a reference hub link. On the right side, there is a sidebar with a search icon and a list of dataset details. The details include: Version (Common Voice Corpus 7.0), Language (English), DATE (2021-07-21), SIZE (65 GB), VERSION (en_2637h_2021-07-21), VALIDATED HR. TOTAL (2,015), OVERALL HR. TOTAL (2,637), LICENSE (CC-0), NUMBER OF VOICES (75,879), AUDIO FORMAT (MP3), and SPLITS (Accent: 23% United States English, 8% England English, ...; Age: 24% 19 - 29, 13% 30 - 39, ...; 45% Male, 15% Female).

Version	Language	DATE	SIZE	VERSION	VALIDATED HR. TOTAL	OVERALL HR. TOTAL	LICENSE	NUMBER OF VOICES	AUDIO FORMAT	SPLITS
Common Voice Corpus 7.0	English	2021-07-21	65 GB	en_2637h_2021-07-21	2,015	2,637	CC-0	75,879	MP3	Accent: 23% United States English, 8% England English, ... Age: 24% 19 - 29, 13% 30 - 39, ... 45% Male, 15% Female

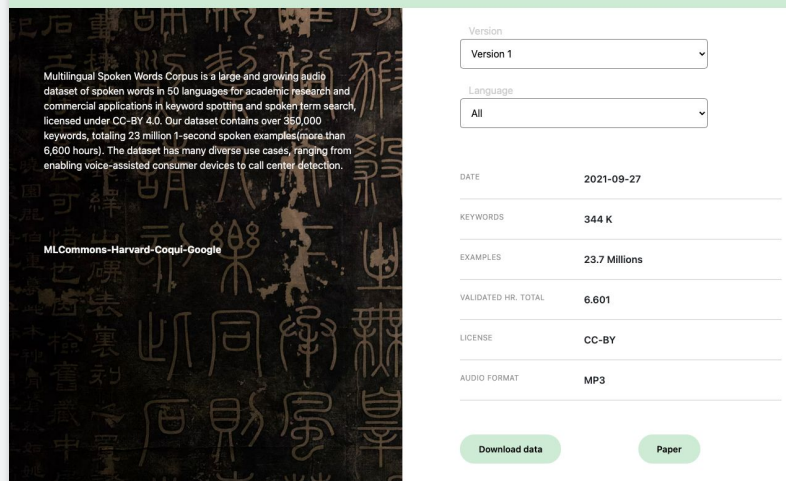
Dataset Sources

- What **relevant datasets** are available?

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

ML • Commons

Multilingual Spoken Words Corpus



Multilingual Spoken Words Corpus is a large and growing public dataset of spoken words in 50 languages for academic research and commercial applications in keyword spotting and spoken term search, licensed under CC-BY 4.0. Our dataset contains over 350,000 keywords, totaling 23 million 1-second spoken examples (more than 6,600 hours). The dataset has many diverse use cases, ranging from enabling voice-assisted consumer devices to call center detection.

MLCommons-Harvard-Cogut-Google

Version	Version 1
Language	All
DATE	2021-09-27
KEYWORDS	344 K
EXAMPLES	23.7 Millions
VALIDATED HR. TOTAL	6,601
LICENSE	CC-BY
AUDIO FORMAT	MP3

[Download data](#) [Paper](#)

Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?



What problem are you trying to *solve*?

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

Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
 - **2,618** volunteers
 - ◆ Variety of accents
 - 1,000+ examples for **each** word
 - All 105,829 remaining utterances **manually reviewed** through crowdsourcing

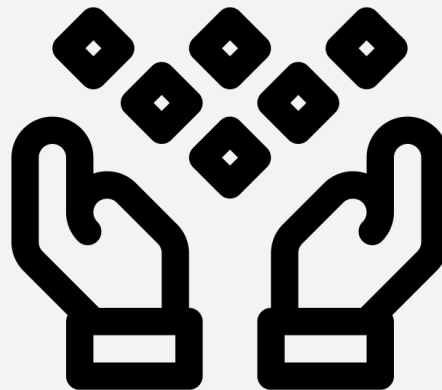
Speech Commands: A Dataset for Limited-Vocabulary Speech
Recognition

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

April 2018

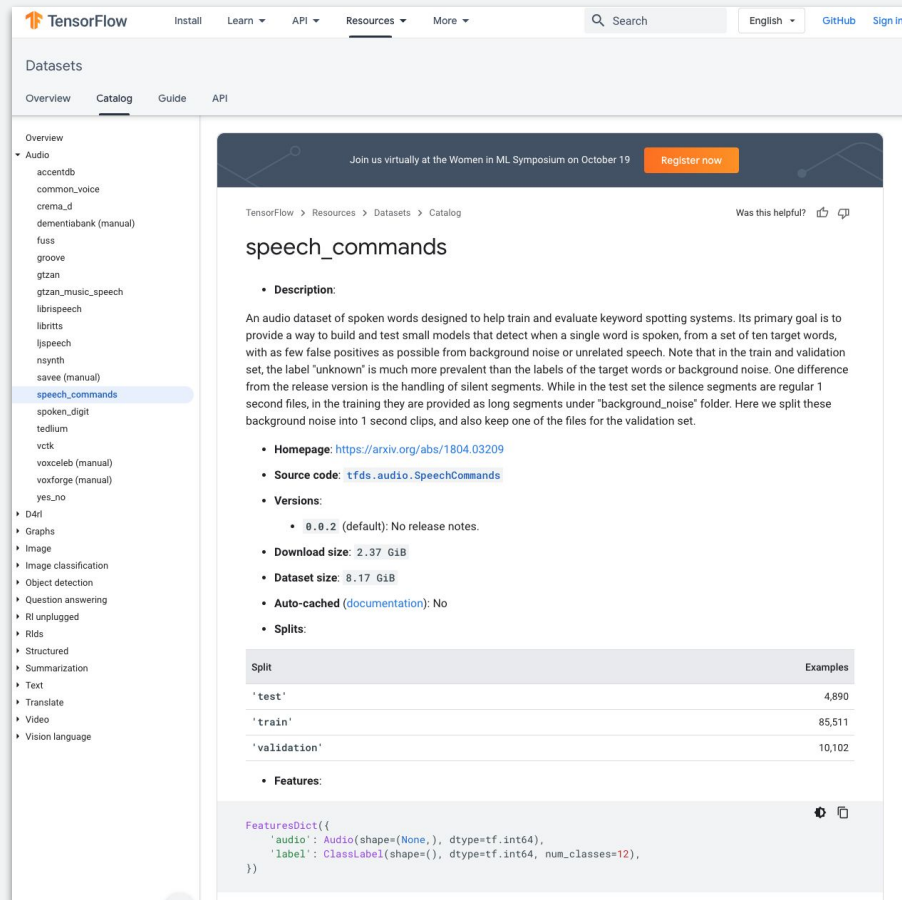
Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?



Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?



The screenshot shows the TensorFlow Datasets Catalog page for the 'speech_commands' dataset. The page is organized into a sidebar with categories like Audio, Image, Text, and Video. The 'speech_commands' dataset is highlighted under the Audio category. The main content area provides a detailed description of the dataset, including its purpose for keyword spotting, its structure (train, validation, and test splits), and its features (audio and label). A table shows the dataset splits: 'test' (4,890 examples), 'train' (85,511 examples), and 'validation' (10,102 examples). The page also includes a 'FeaturesDict' snippet and a 'Was this helpful?' feedback prompt.

TensorFlow

Install Learn API Resources More

Search English GitHub Sign in

Datasets

Overview Catalog Guide API

Overview

- Audio
 - accentdb
 - common_voice
 - crema_d
 - dementiabank (manual)
 - fuss
 - groove
 - gtzan
 - gtzan_music_speech
 - librispeech
 - libritts
 - ljspeech
 - nsynth
 - savee (manual)
 - speech_commands**
 - spoken_digit
 - tedium
 - vctk
 - voxceleb (manual)
 - voxforge (manual)
 - yes_no
- D4rl
- Graphs
- Image
 - Image classification
 - Object detection
- Question answering
- RI unplugged
- Rlds
- Structured
- Summarization
- Text
 - Translate
- Video
 - Vision language

Join us virtually at the Women in ML Symposium on October 19 [Register now](#)

TensorFlow > Resources > Datasets > Catalog Was this helpful? [👍](#) [👎](#)

speech_commands

- Description:**

An audio dataset of spoken words designed to help train and evaluate keyword spotting systems. Its primary goal is to provide a way to build and test small models that detect when a single word is spoken, from a set of ten target words, with as few false positives as possible from background noise or unrelated speech. Note that in the train and validation set, the label "unknown" is much more prevalent than the labels of the target words or background noise. One difference from the release version is the handling of silent segments. While in the test set the silence segments are regular 1 second files, in the training they are provided as long segments under "background_noise" folder. Here we split these background noise into 1 second clips, and also keep one of the files for the validation set.
- Homepage:** <https://arxiv.org/abs/1804.03209>
- Source code:** [tfds.audio.SpeechCommands](#)
- Versions:**
 - 0.0.2 (default): No release notes.
- Download size:** 2.37 GiB
- Dataset size:** 8.17 GiB
- Auto-cached** ([documentation](#)): No
- Splits:**

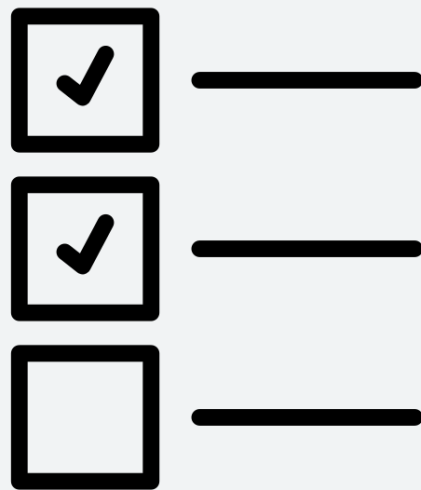
Split	Examples
'test'	4,890
'train'	85,511
'validation'	10,102

- Features:**

```
FeaturesDict({  
  'audio': Audio(shape=(None,), dtype=tf.int64),  
  'label': ClassLabel(shape=(), dtype=tf.int64, num_classes=12),  
})
```

Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?
- What **properties** are in this dataset?

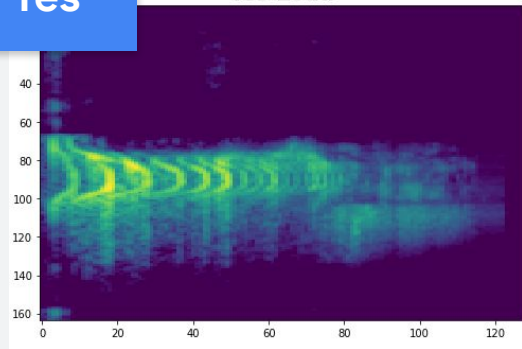


Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?
- What **properties** are in this dataset?
- Does the data need to be **labeled**?

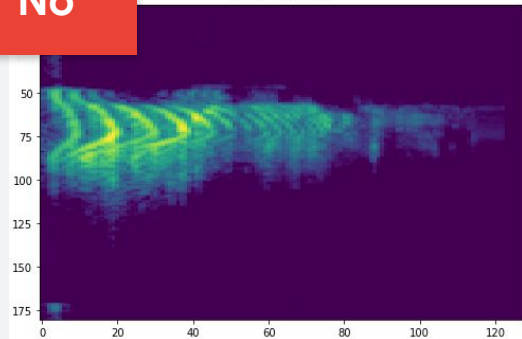
Yes

Yes Loud



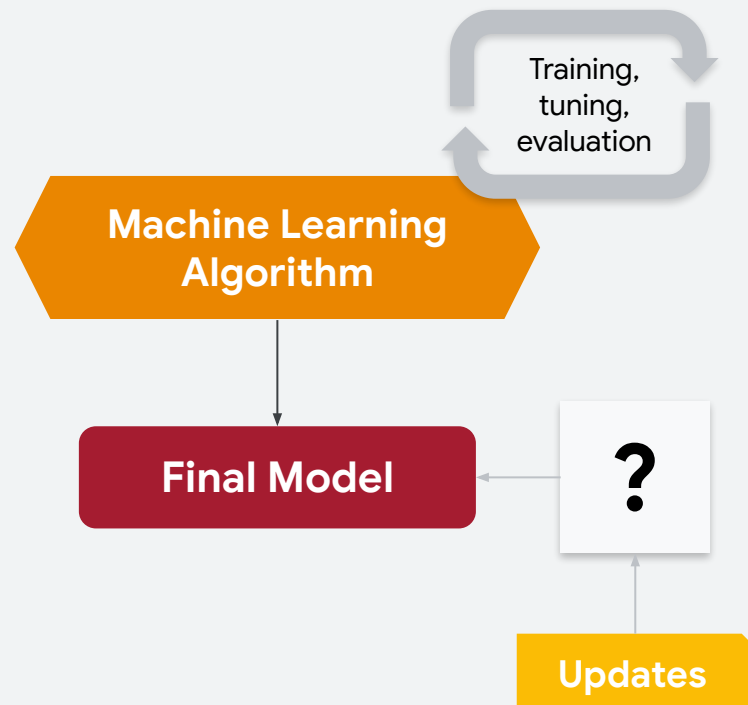
No

No Loud



Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?
- What **properties** are in this dataset?
- Does the data need to be **labeled**?
- How will it be **updated** post training?



Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?
- What **properties** are in this dataset?
- Does the data need to be **labeled**?
- How will it be **updated** post training?
- What are the **licenses** and terms of use?
 - Open?
 - Copyrighted?
 - Licensed?
 - Product users?



Dataset Sources

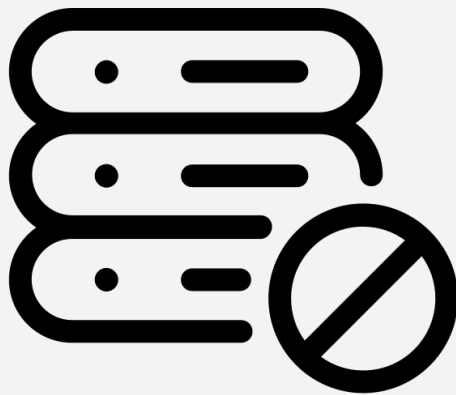
- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?
- What **properties** are in this dataset?
- Does the data need to be **labeled**?
- How will it be **updated** post training?
- What are the **licenses** and terms of use?
- Is there **personally identifiable information** (PII) in the dataset?



Security & Privacy

Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?
- What **properties** are in this dataset?
- Does the data need to be **labeled**?
- How will it be **updated** post training?
- What are the **licenses** and terms of use?
- Is there **personally identifiable information** (PII) in the dataset?
- Are there **restricted features** that cannot be used for commercial applications?



Dataset Sources

- What **relevant datasets** are available?
- Is this data **accurate and reliable**?
- Can stakeholders get **access to data**?
- What **properties** are in this dataset?
- Does the data need to be **labeled**?
- How will it be **updated** post training?
- What are the **licenses** and terms of use?
- Is there **personally identifiable information** (PII) in the dataset?
- Are there **restricted features** that cannot be used for commercial applications?

