Continuous Training: Data Processing Engine

MLOps: Continuous Training

The MLOps **Personas**

ML Engineer

ML Researcher

Data Scientist

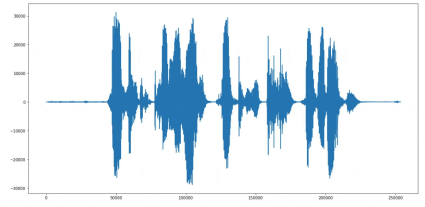
Data Engineer

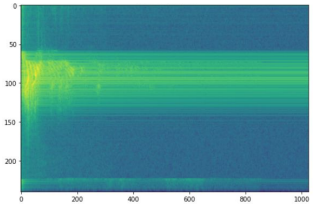
Software Engineer

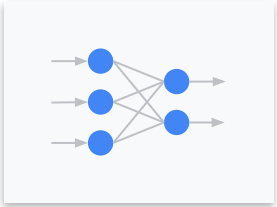
DevOps

Business Analyst

Keyword Spotting



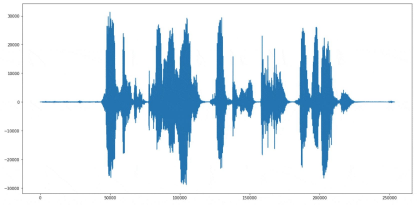


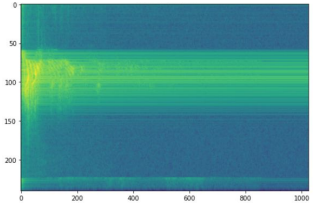


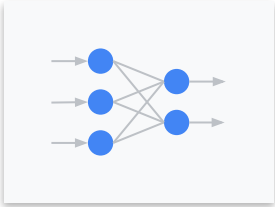
“Yes” - 0.91
“No” - 0.09

- Voice assistants are **ubiquitous**

Keyword Spotting



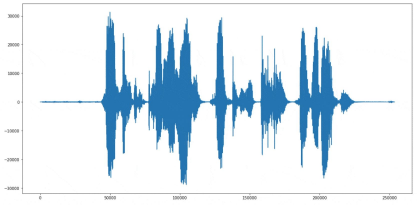


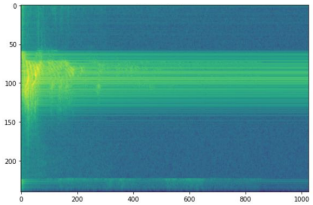


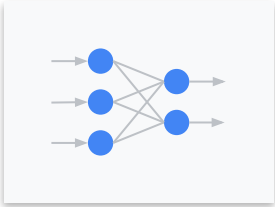
“Yes” - 0.91
“No” - 0.09

- Voice assistants are **ubiquitous**
- **Limited** vocabulary and languages

Keyword Spotting



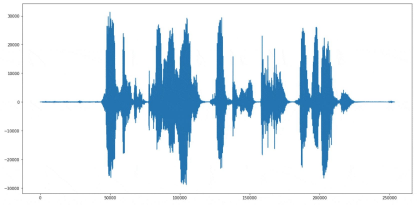


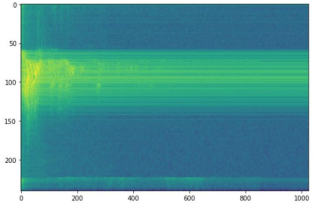


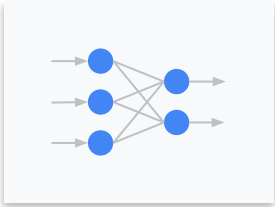
“Yes” - 0.91
“No” - 0.09

- Voice assistants are **ubiquitous**
- **Limited** vocabulary and languages
- Datasets require **thousands** of training examples ***per keyword***

Keyword Spotting







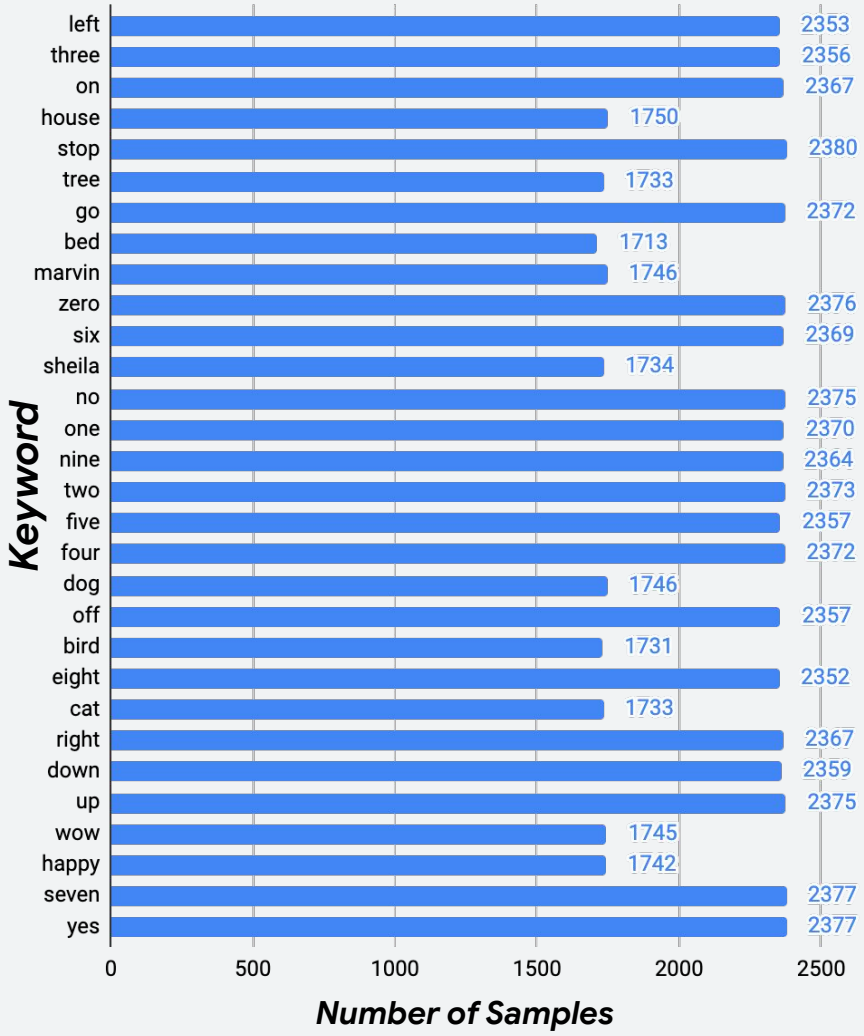
“Yes” - 0.91
“No” - 0.09

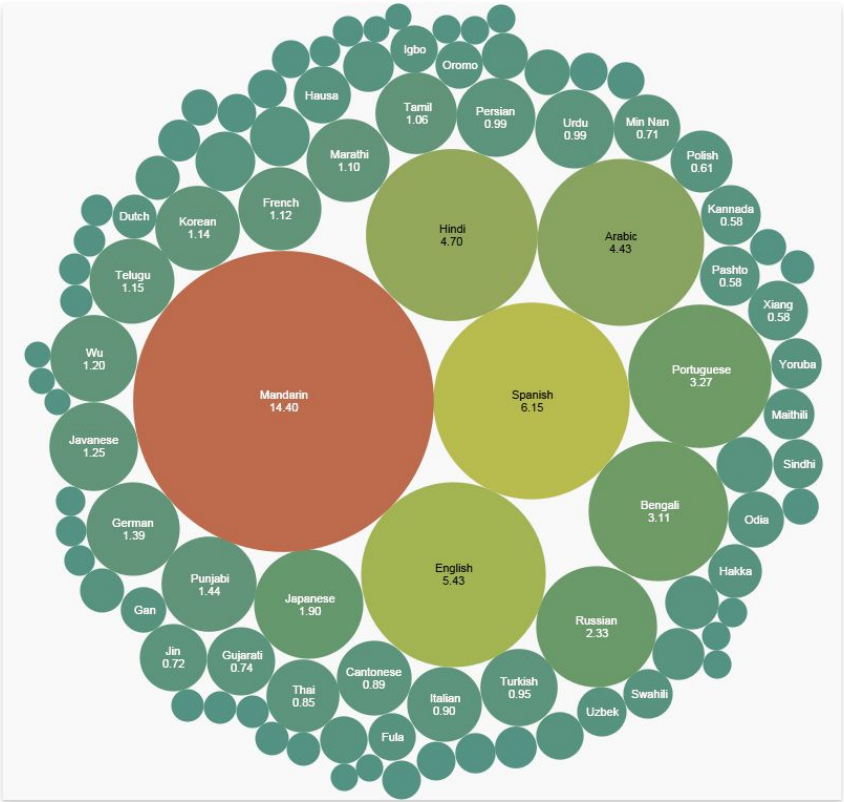
- Voice assistants are **ubiquitous**
- **Limited** vocabulary and languages
- Datasets require **thousands** of training examples ***per keyword***

Goal: make voice interfaces available to a **worldwide** audience

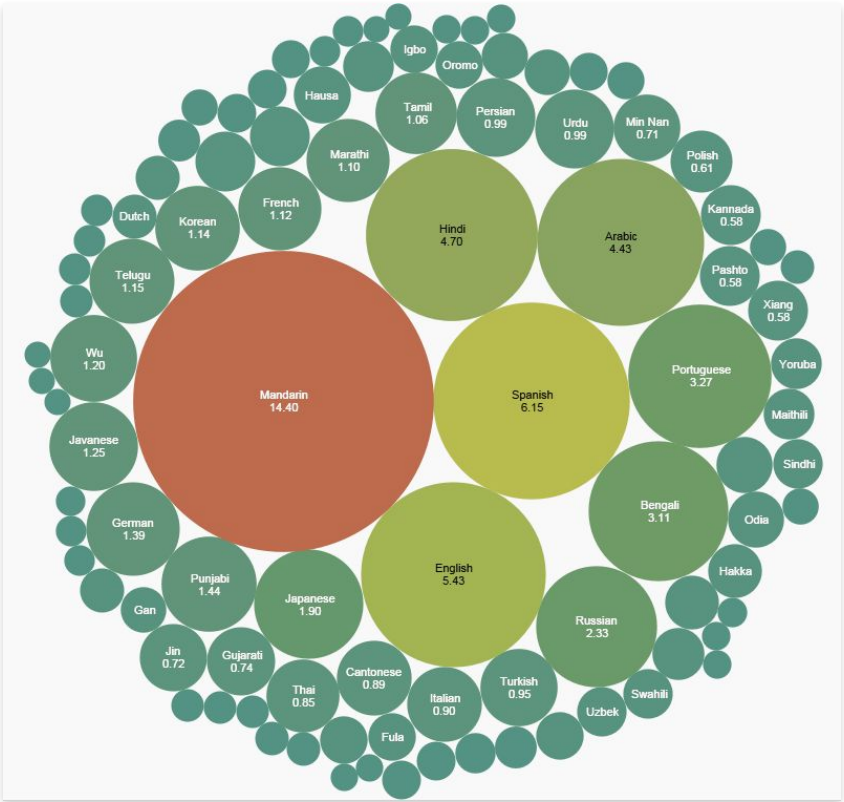
Google Speech Commands (**GSC**)

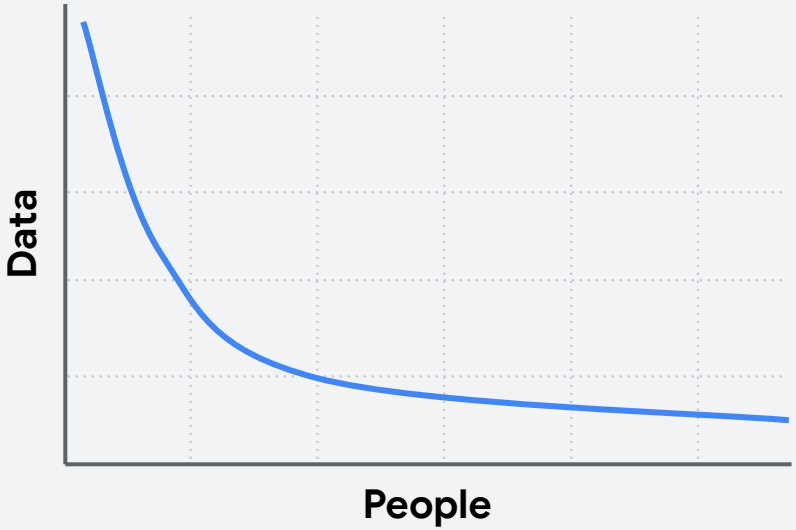
- **English** dataset
- **35** spoken words
- **105,000** audio samples
- **Individually recorded** words





- Speech commands for the **whole planet**?
- For **more than** just voice assistants


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
Bengali
3.11
German
1.39
Odia
English
5.43
Punjabi
1.44
Hakka
Japanese
1.90
Gan
Russian
2.33
Jin
0.72
Gujarati
0.74
Cantonese
0.89
Thai
0.85
Turkish
0.95
Italian
0.90
Swahili
Uzbek
Fula


Data
People

Data Engineering

Data Engineering

Requirements

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

Data Engineering

Requirements → **Gathering**

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

- Data sources
- People
- Collection
- Labeling

Data Engineering

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

- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

Data Engineering

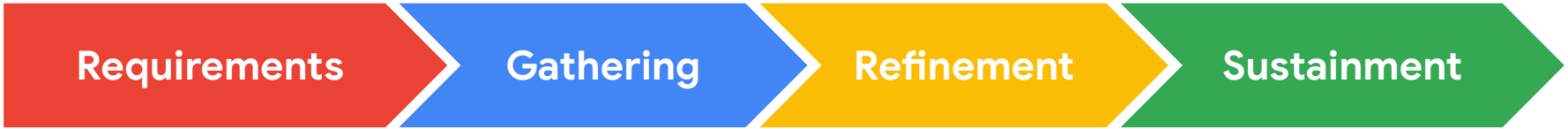
- Problem definition
- Machine & human usable format
- Permissions & rights

- Data sources
- People
- Collection
- Labeling

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

Data Engineering **for Speech Commands**



- English KWS dataset
- Permissively licensed for research and use for commercial apps

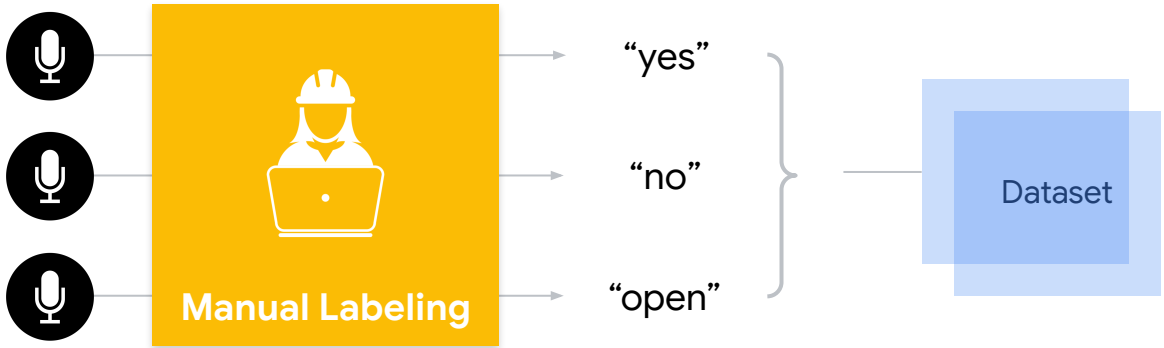
- 2,618 volunteers
- Variety of accents
- > 1K examples per word

- Remove quiet or low volume recordings
- Manually review through crowdsourcing

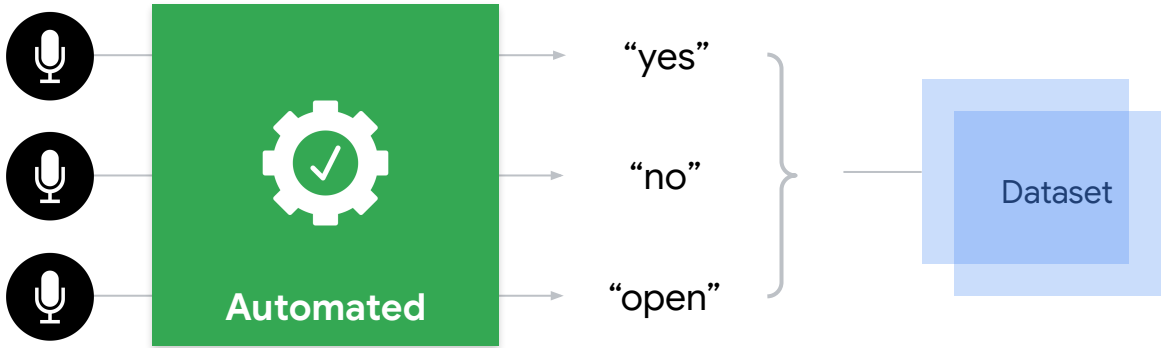
- V1 had 10 words
- V2 added 25 more words
- Includes train, validation and test dataset splits

How do we **scale** past this?

Labeling is Expensive



How do we **automate**?



Multilingual Spoken Words Corpus

Spoken Words in **50 Languages**

Spoken Words in **50 Languages**

HOURS	CATEGORIZATION	LANGUAGES
<10	Low-Resource	Arabic, Assamese, Breton, Chuvash, Chinese (zh-CN), Dhivehi, Frisian, Georgian, Guarani, Greek, Hakha Chin, Hausa, Interlingua, Irish, Latvian, Lithuanian, Maltese, Oriya, Romanian, Sakha, Slovenian, Slovak, Sursilvan, Tamil, Vallander, Vietnamese

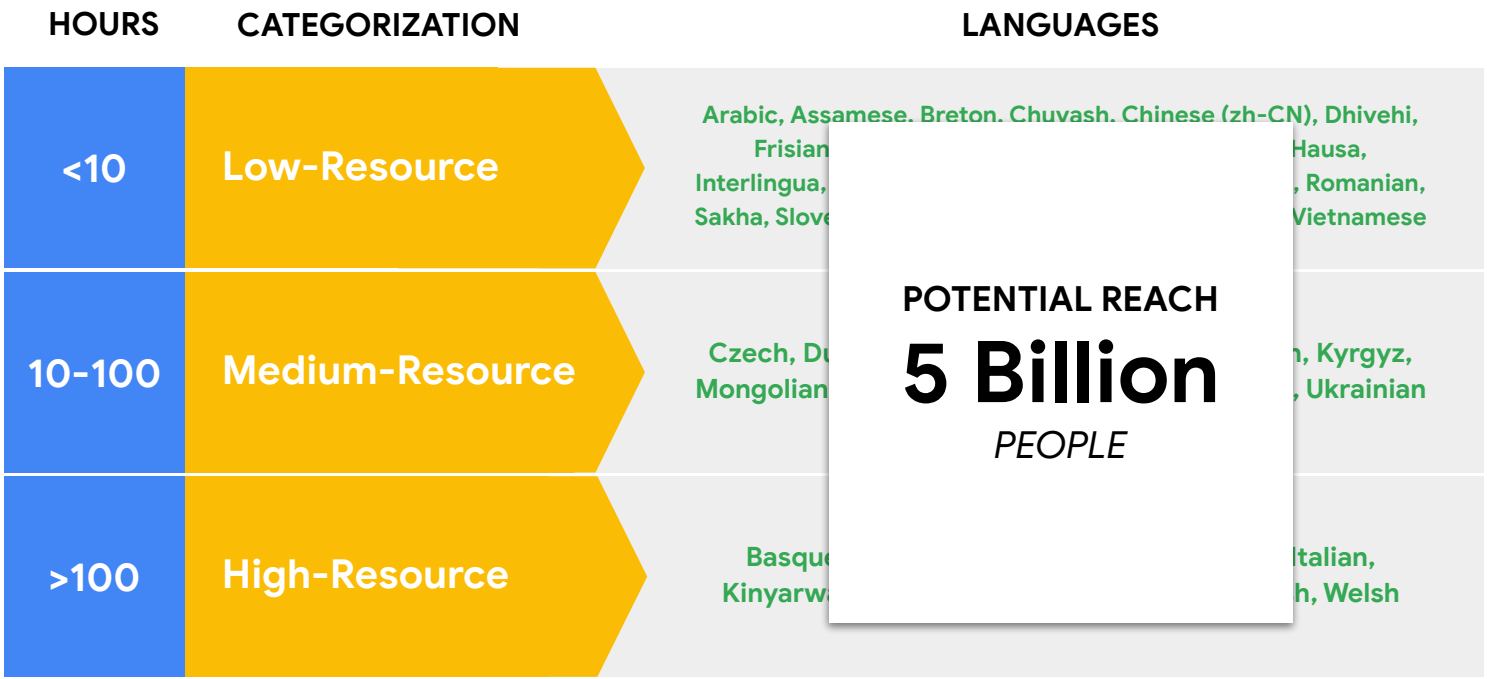
Spoken Words in **50 Languages**

HOURS	CATEGORIZATION	LANGUAGES
<10	Low-Resource	Arabic, Assamese, Breton, Chuvash, Chinese (zh-CN), Dhivehi, Frisian, Georgian, Guarani, Greek, Hakha Chin, Hausa, Interlingua, Irish, Latvian, Lithuanian, Maltese, Oriya, Romanian, Sakha, Slovenian, Slovak, Sursilvan, Tamil, Vallander, Vietnamese
10-100	Medium-Resource	Czech, Dutch, Estonian, Esperanto, Indonesian, Kyrgyz, Mongolian, Portuguese, Swedish, Tatar, Turkish, Ukrainian

Spoken Words in **50 Languages**

HOURS	CATEGORIZATION	LANGUAGES
<10	Low-Resource	Arabic, Assamese, Breton, Chuvash, Chinese (zh-CN), Dhivehi, Frisian, Georgian, Guarani, Greek, Hakha Chin, Hausa, Interlingua, Irish, Latvian, Lithuanian, Maltese, Oriya, Romanian, Sakha, Slovenian, Slovak, Sursilvan, Tamil, Vallander, Vietnamese
10-100	Medium-Resource	Czech, Dutch, Estonian, Esperanto, Indonesian, Kyrgyz, Mongolian, Portuguese, Swedish, Tatar, Turkish, Ukrainian
>100	High-Resource	Basque, Catalan, English, French, German, Italian, Kinyarwanda, Persian, Polish, Russian, Spanish, Welsh

Spoken Words in **50 Languages**



Keywords per language

LOW-RESOURCE	MEDIUM-RESOURCE	HIGH-RESOURCE
AVERAGE PER LANGUAGE	AVERAGE PER LANGUAGE	AVERAGE PER LANGUAGE
552 KEYWORDS	4,054 KEYWORDS	23,408 KEYWORDS
10,431 SAMPLES	109,597 SAMPLES	1.8M SAMPLES

https://mlcommons.org/words