CS249r: Data Engineering

Sep. 18, 2023

Goals for today

1. Course logistics
2. Lecture
3. Paper discussions
4. Guest speaker

Course Logistics

Class Schedule

		Start	~End
Lecture		12:45:00 PM	1:40:00 PM
Break		1:40:00 PM	1:45:00 PM
Paper Discussions	Paper 1	1:45:00 PM	2:05:00 PM
	Paper 2	2:05:00 PM	2:25:00 PM
Break		2:25:00 PM	2:30:00 PM
Guest Lecture		2:30:00 PM	3:30:00 PM

Office Hours

Matthew Stewart

Postdoctoral Researcher, Harvard University

Office Hours: 5-6 PM Thursday (SEC 1.412)

Ikechukwu Uchendu

Computer Science, PhD Student, Harvard University

Office Hours: 9-10 AM Thursday (SEC 1.412)

Jason Jabbour

Computer Science, PhD Student, Harvard University

Office Hours: 5-6 PM Tuesday

Jessica Quaye

Computer Science, PhD Student, Harvard University

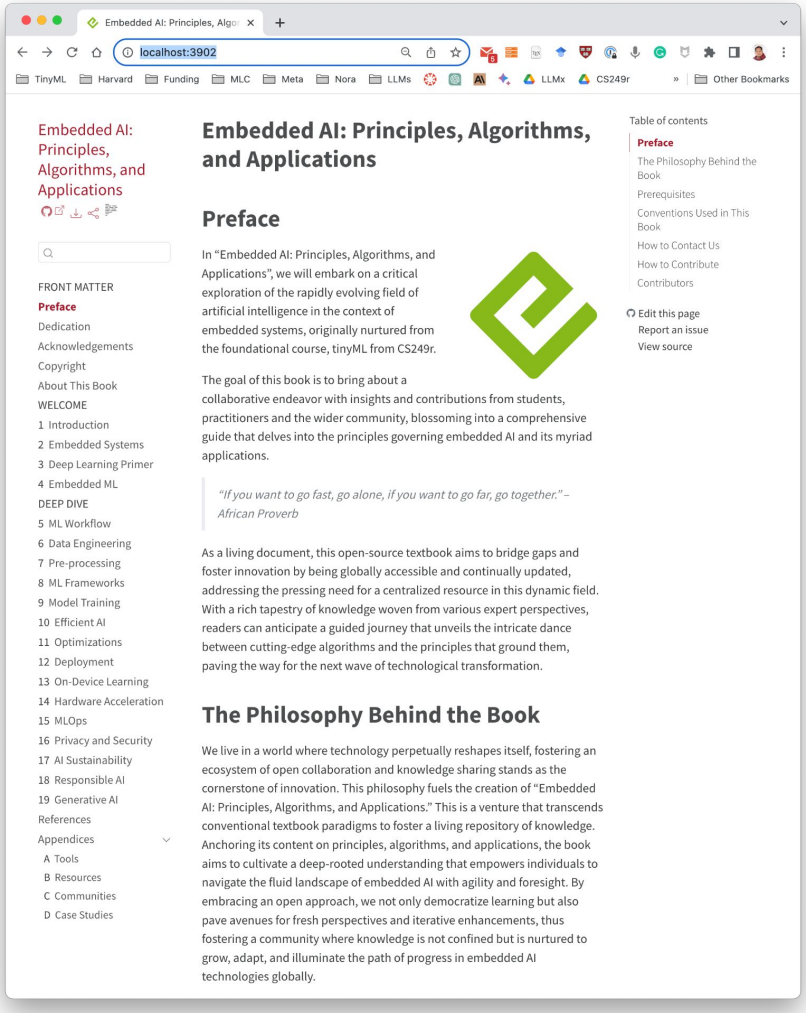
Office Hours: 1-2 PM Tuesday (SEC 5.403)

Paper Discussion Sign-up Sheet

1. Please take a few minutes to sign up for next week's paper discussion:
2. Submit to Canvas:
    a. Paper Reading Group - 1 for your first time
    b. Paper Reading Group - 2 for your second time

Book

- Goal:
  - Open source “living” textbook
- Putting in “boiler plate” material
- Welcome to read through the first few chapters of background
  - Improve it!
- Link



Book Chapter Contributions

- Link to sign-up sheet
- 3-4 people sign up
- Produce at least 10 pages of content
  - Brainstorm sections within a chapter
  - Meet with Matthew Stewart and Prof. VJ to discuss sections
  - Work collaboratively to improve the content/sections
- Grading
  - Content quality
  - Topic coverage
  - Figures/References
  - Peer Review

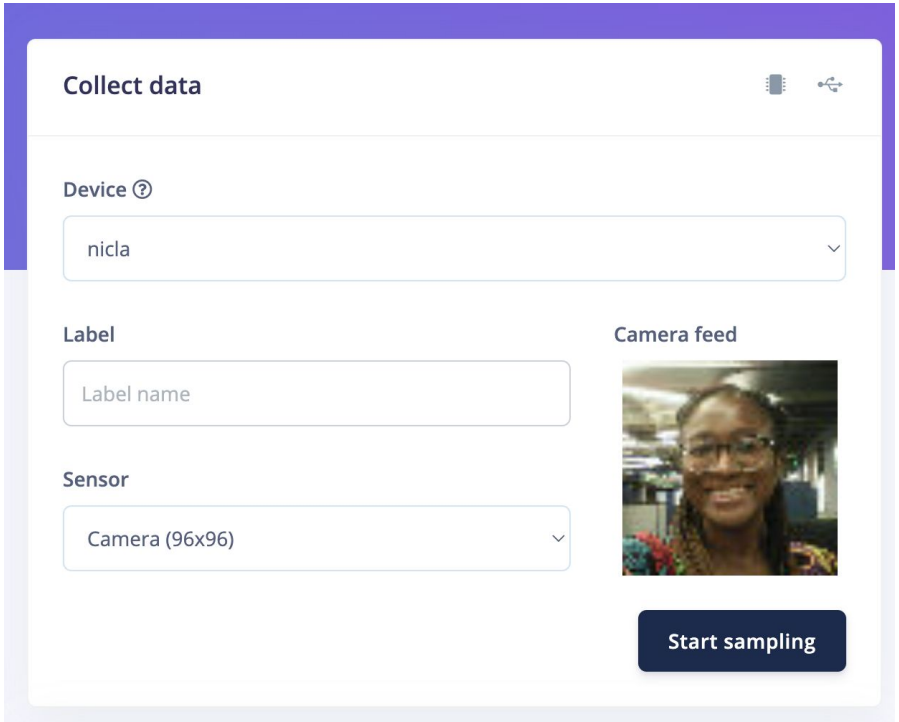
Assignment 1: Setup Arduino Nicla Vision + Data Engineering for Person Detection

- TinyML workflow
- Data Engineering
- Edge Impulse
- Transfer Learning

Assignment 1: Setup Arduino Nicla Vision + Data Engineering for Person Detection

- **Part 1**:
  - **What**: You will **set up the Arduino Nicla Vision** and submit an image of yourself taken with the Nicla. Your image will be included in the testing dataset for Part 2.
  - **How:** You will use the Edge Impulse pipeline to bootstrap your system.

Assignment 1: Setup Arduino Nicla Vision

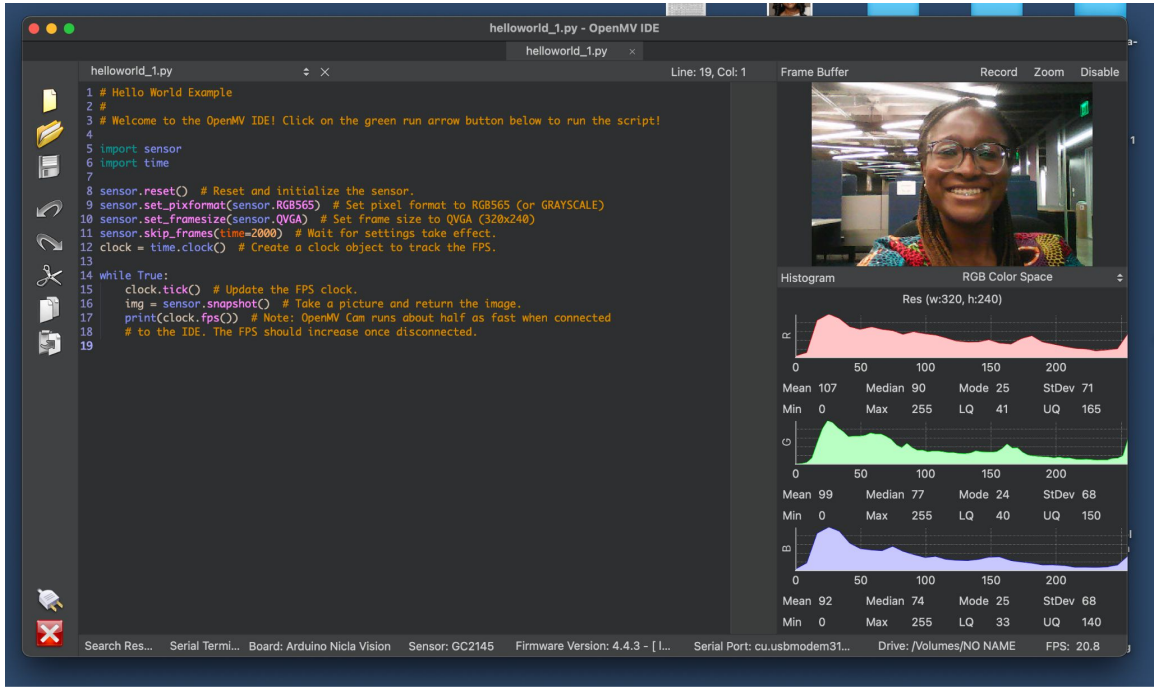


Assignment 1: Setup Arduino Nicla Vision + Data Engineering for Person Detection

- **Part 1**:
  - **What**: You will **set up the Arduino Nicla Vision** and submit an image of yourself taken with the Nicla. Your image will be included in the testing dataset for Part 2.
  - **How:** You will use the Edge Impulse pipeline to bootstrap your system.

- **Part 2**:
  - **What:**You will **gather training data for a person detection model**.
  - **How:** You will use Edge Impulse + OpenMV tools to get your training data.

Assignment 1: Collect Training Data for Person Detection Model



Lecture


amazon

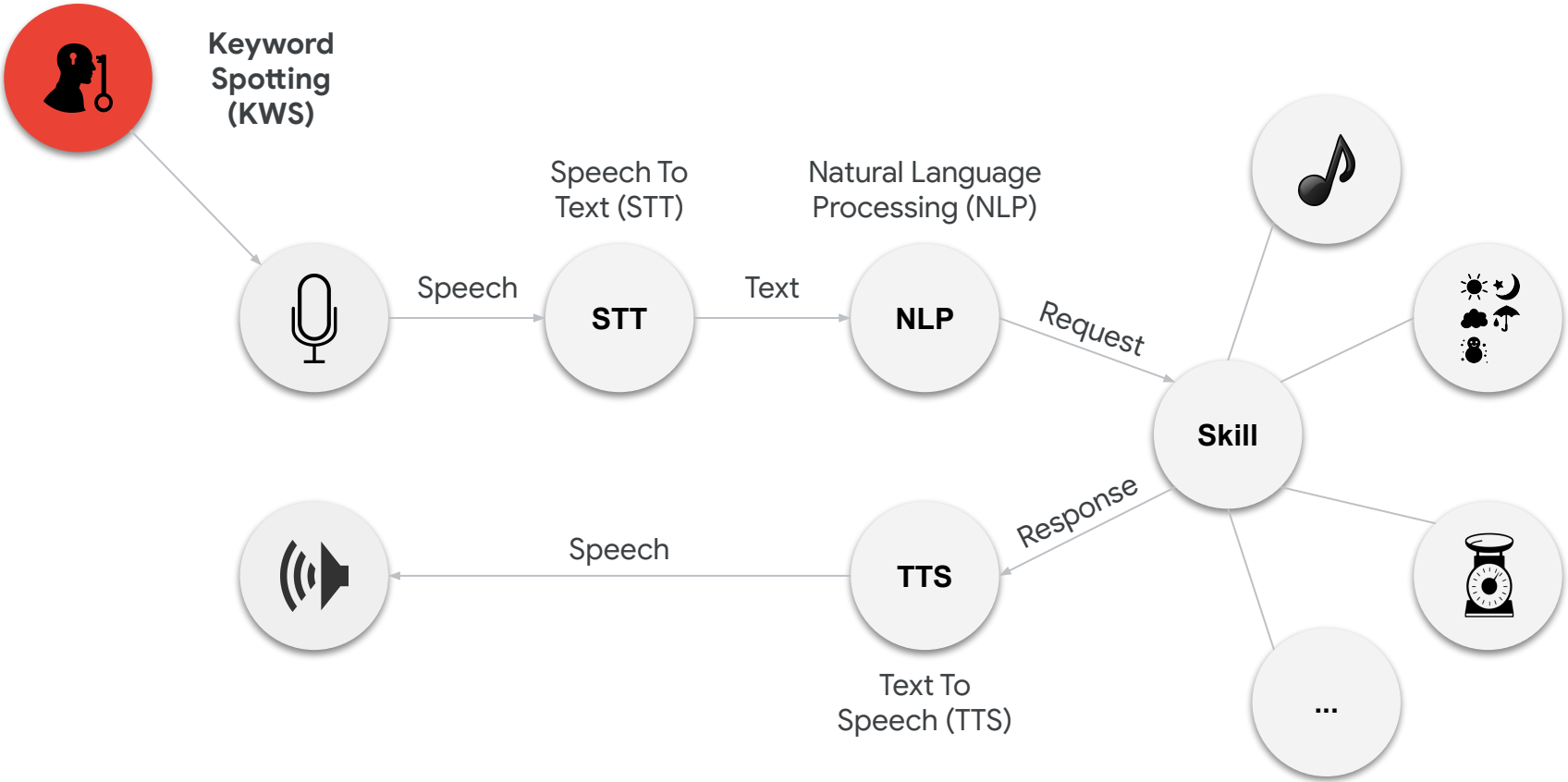
Keyword Spotting v. General Speech Recognition

- **Keyword spotting** is one of the most successful examples of **TinyML**
  - Low-power, continuous, on-device, machine learning
  - Common Voice SWTS expands keyword spotting to more languages

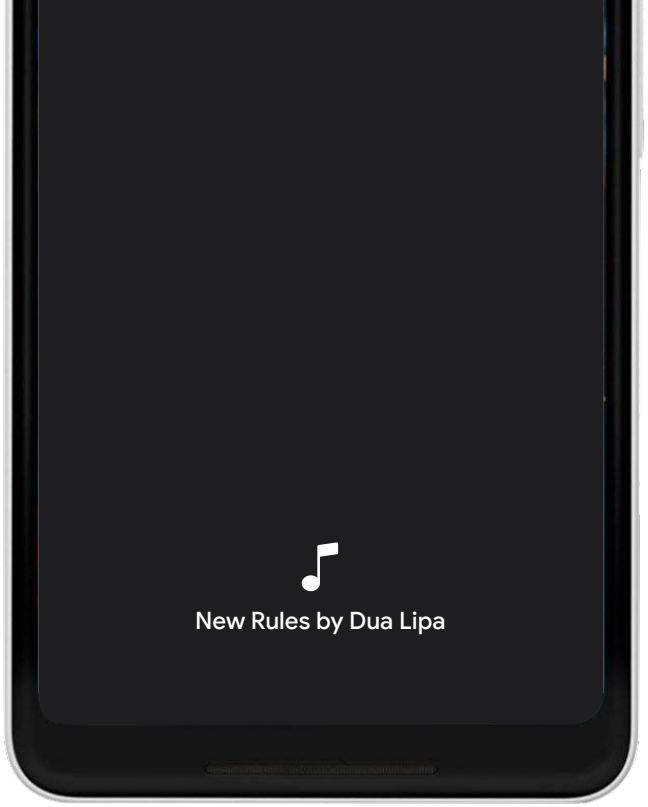
Keyword Spotting v. General Speech Recognition

- **Keyword spotting** is one of the most successful examples of **TinyML**
  - Low-power, continuous, on-device
  - Common Voice SWTS expands keyword spotting to more languages

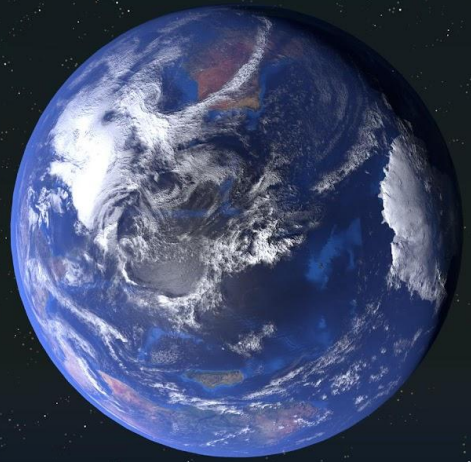
- **General ASR** still requires **larger, power-hungry models**
  - But it can run on mobile devices (offline dictation on smartphones)


Keyword Spotting (KWS)
Speech To Text (STT)
Natural Language Processing (NLP)
Speech
STT
Text
NLP
Request
Skill
Response
Speech
TTS
Text To Speech (TTS)
...

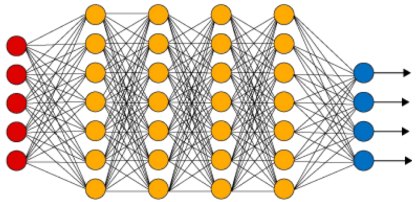
Now Playing
by Google

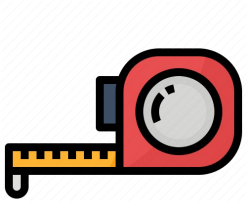


Imagine a world where every device has a voice assistant in *your* favorite language.

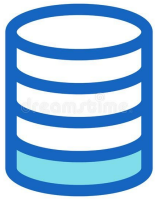


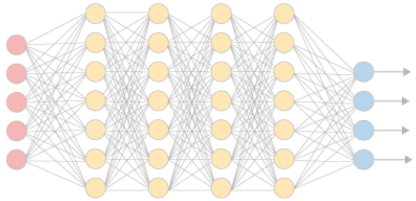

你好
HALLO
안녕
HOLA
नमस्ते
CIAO
HELLO
γεια
こんにちは
привет
BONJOUR
مرحبا
OLÁ

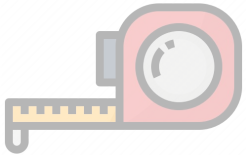




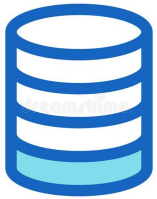
Collect Data → Preprocess Data → Design a Model → Train a Model → Evaluate Optimize → Convert Model → Deploy Model → Make Inferences

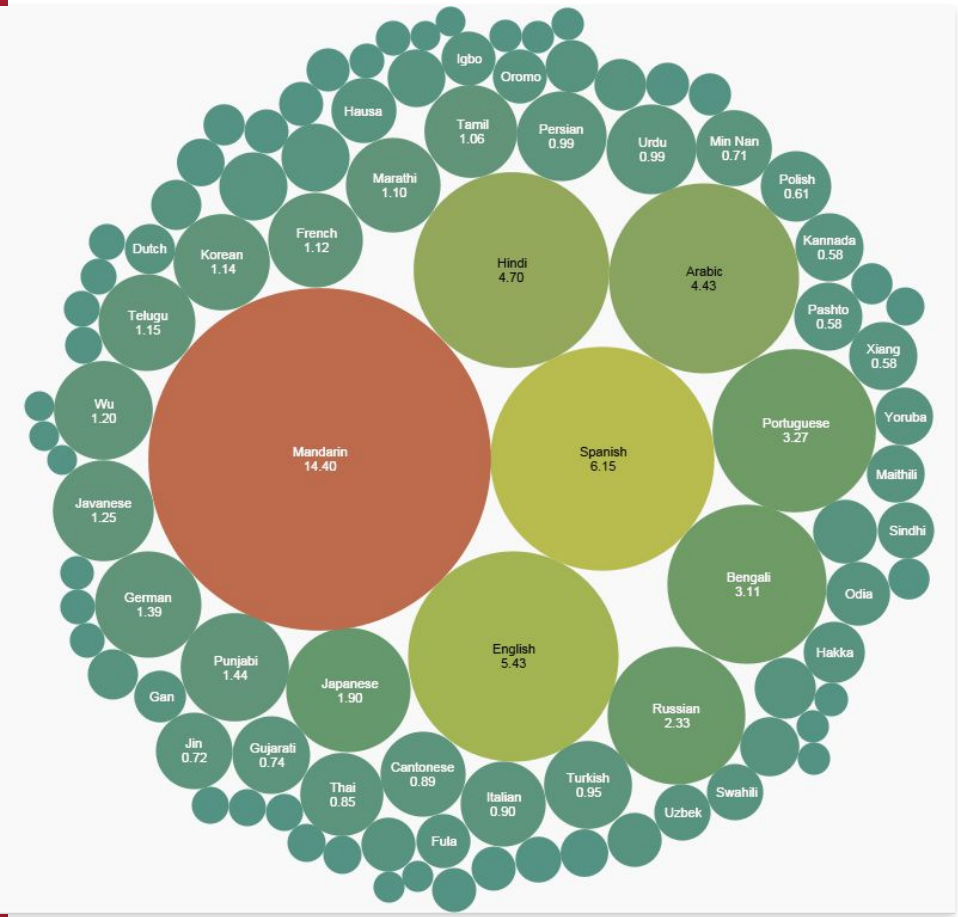




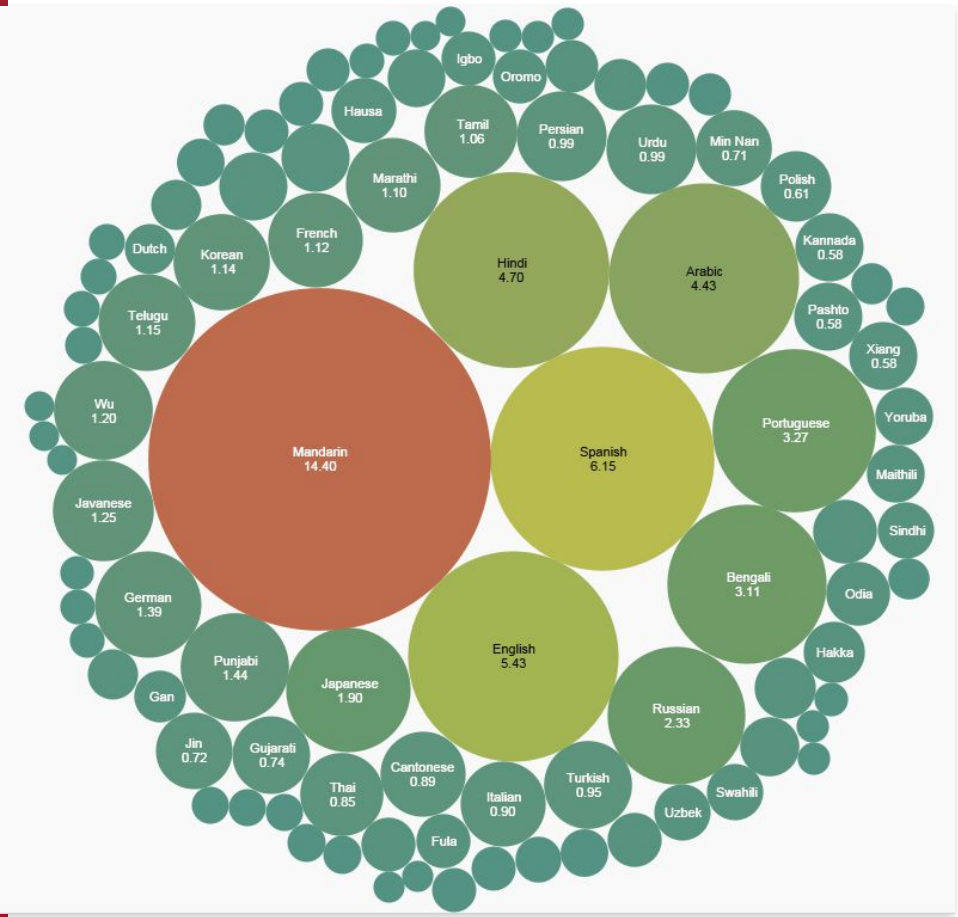


Collect Data → Preprocess Data → Design a Model → Train a Model → Evaluate Optimize → Convert Model → Deploy Model → Make Inferences

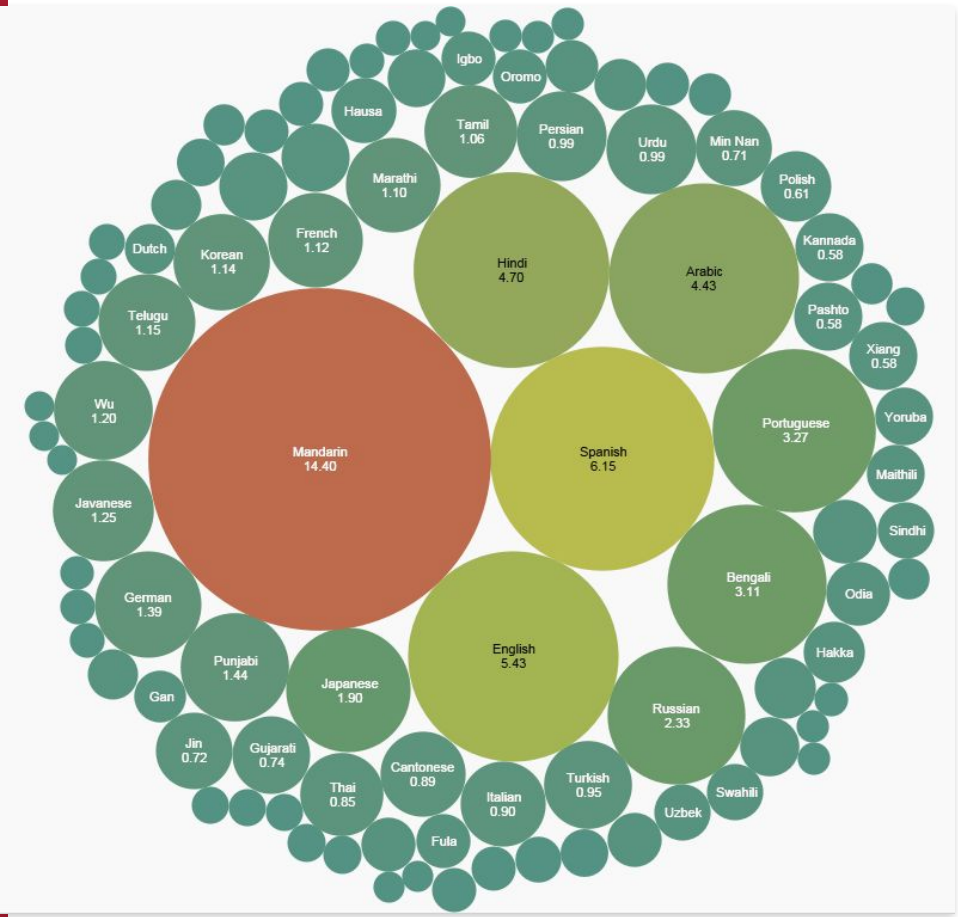


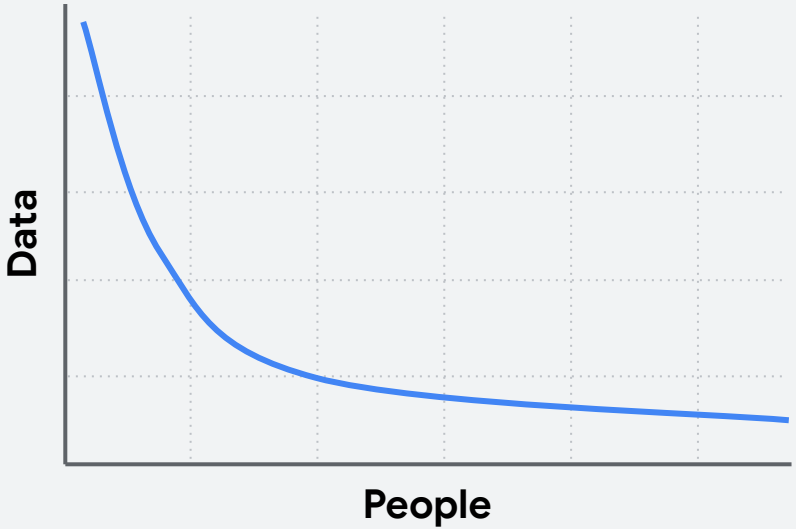


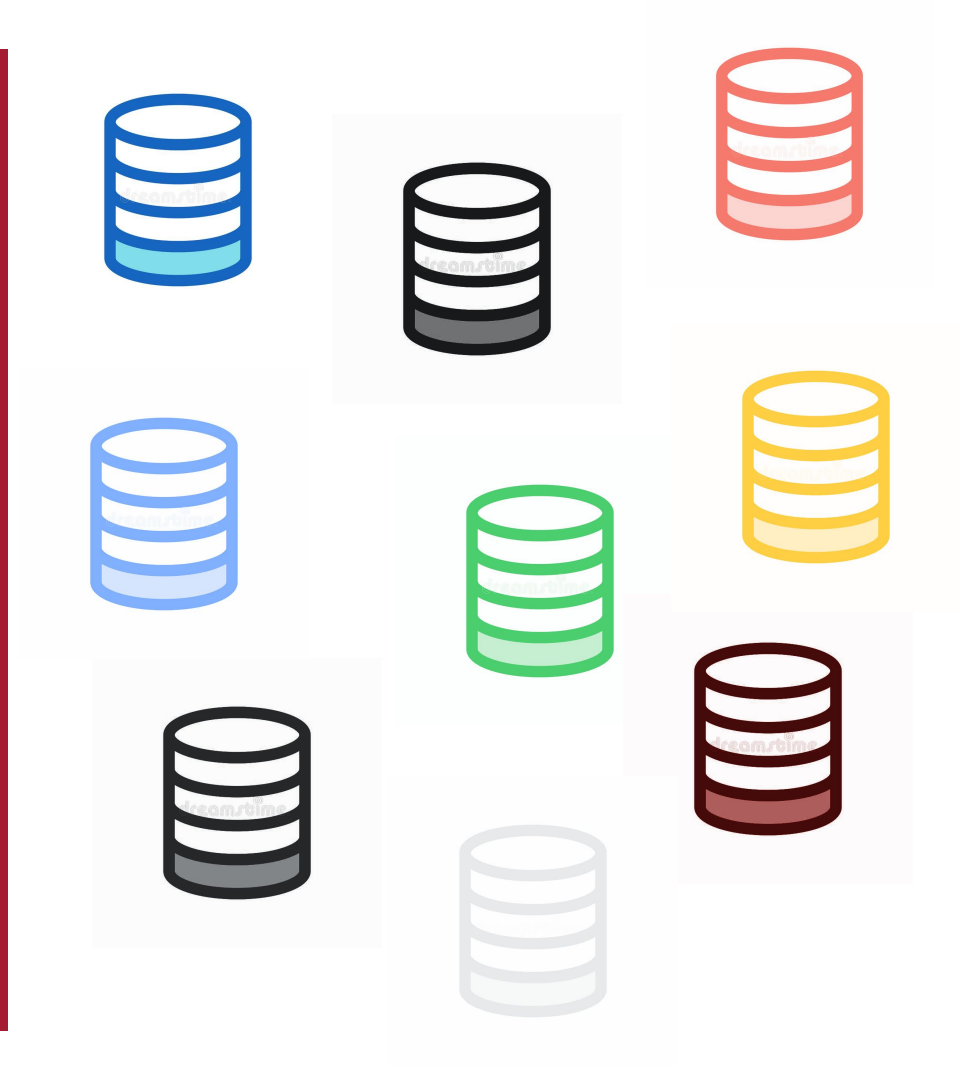
- Speech commands for the **whole planet**?

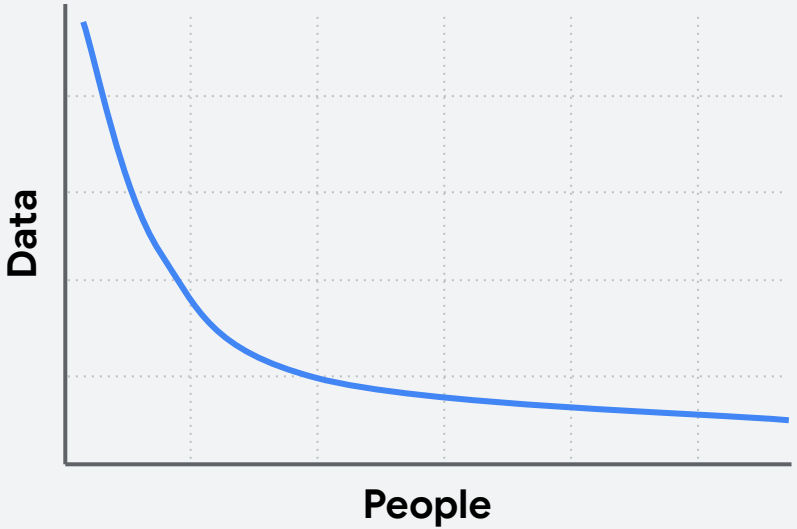


- Speech commands for the **whole planet**?
- For **more than** just voice assistants


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
Pashto
0.58
Telugu
1.15
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
Hakka
Punjabi
1.44
Japanese
1.90
Russian
2.33
Gan
Jin
0.72
Gujarati
0.74
Cantonese
0.89
Turkish
0.95
Swahili
Thai
0.85
Italian
0.90
Uzbek
Fula


Data
People




Data
People

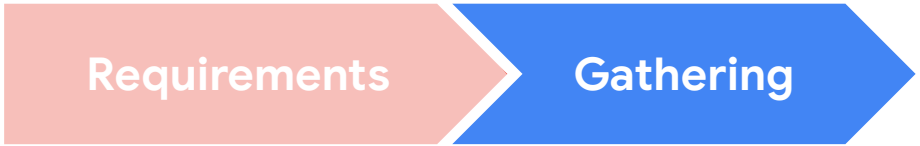
How can we democratize ML?

Data Engineering

Requirements

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

Data Engineering



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

- People
- Collection
- Labeling
- Data sources

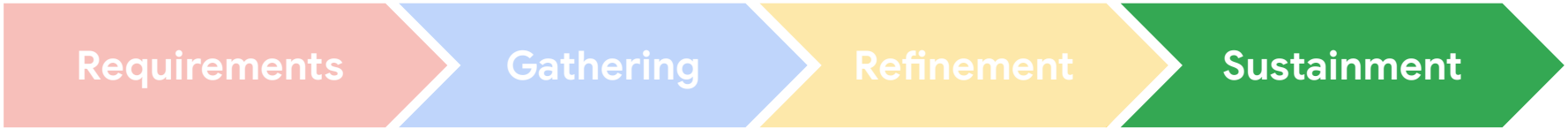
Data Engineering

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

- People
- Collection
- Labeling
- Data sources

- Processing
- Validation
- Augmentation

Data Engineering



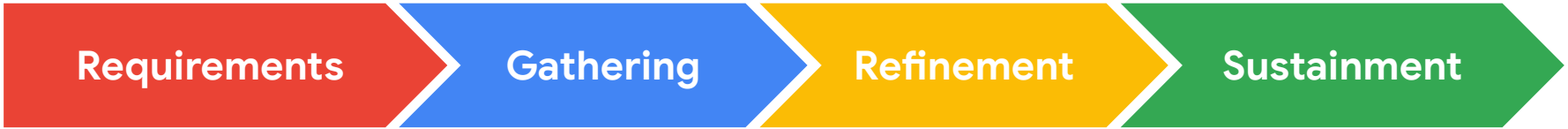
- Problem definition
- Permissions & rights
- Machine & human usable format

- People
- Collection
- Labeling
- Data sources

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

Data Engineering



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

- People
- Collection
- Labeling
- Data sources

- Processing
- Validation
- Augmentation

- Storage
- Security
- Errors
- Versioning

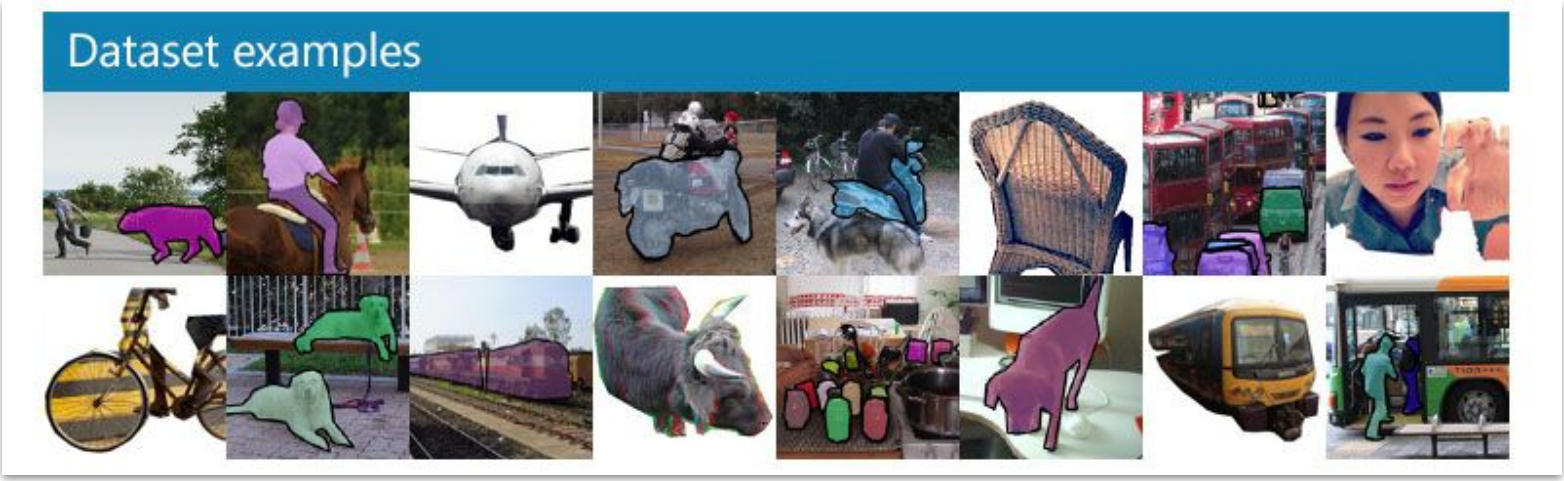
Data engineering requires significant capital & effort.

Classifying Images

(Deng et al., 2009)

Detecting Objects

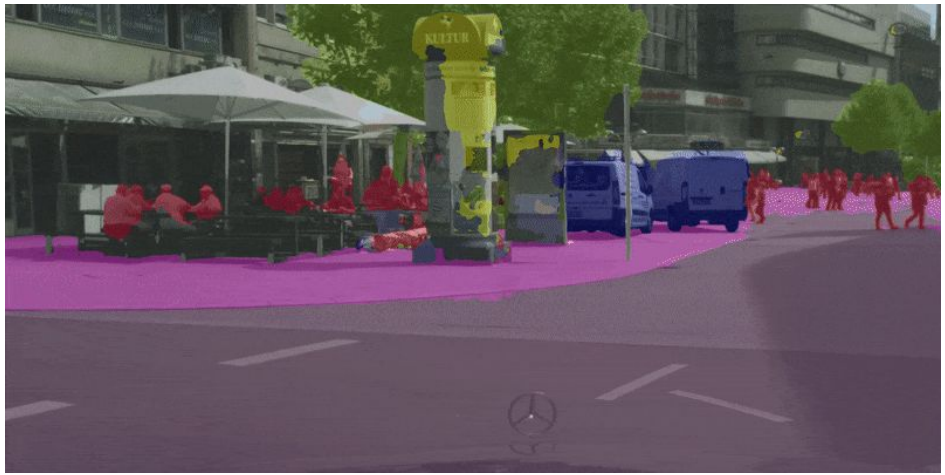
- Common Objects in Context (**COCO**)—**2.5M+** segmented images



(Lin et al., 2014)

Datasets require ***significant effort***

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

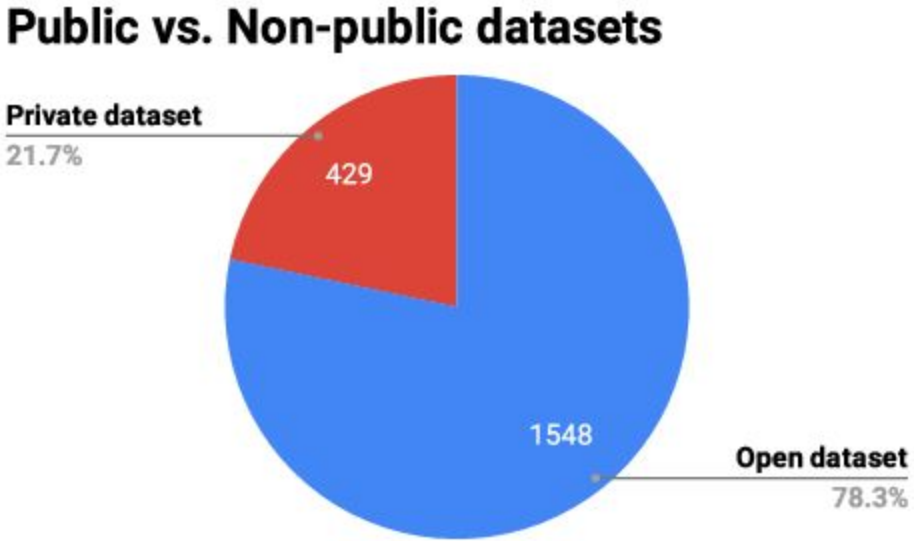


Datasets require ***significant effort***

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

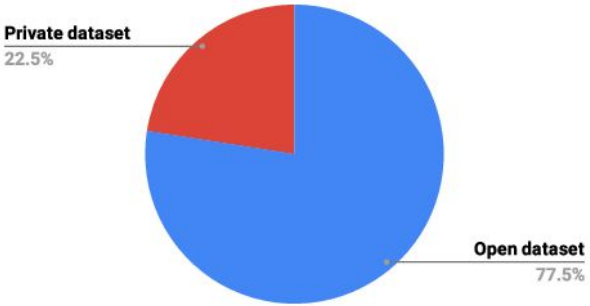
- **Common Voice**—**5000+** hours of spoken audio
- Common Objects in Context (**COCO**)—**2.5M+** labeled images
- **ImageNet**—**4M+** labeled images
- **Waymo—1,950** 20-second driving segments
- **KITTI 360**—**73KM+** of annotated driving data

Data Engineering is costly and tedious.

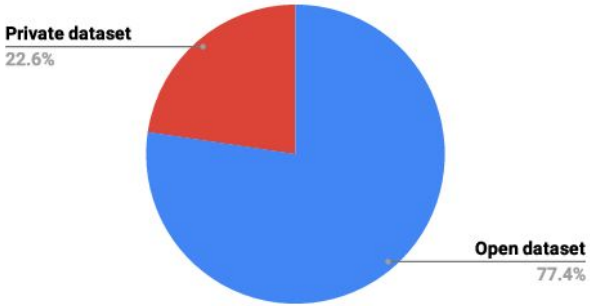


(Reddi et al., 2021)

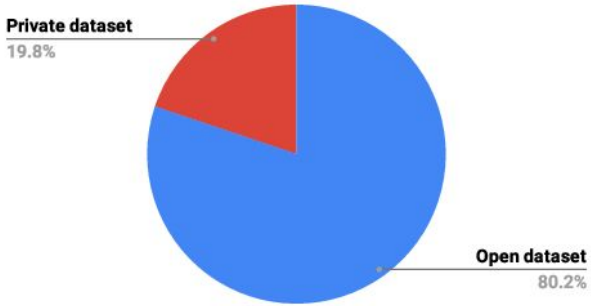
Facebook



Google



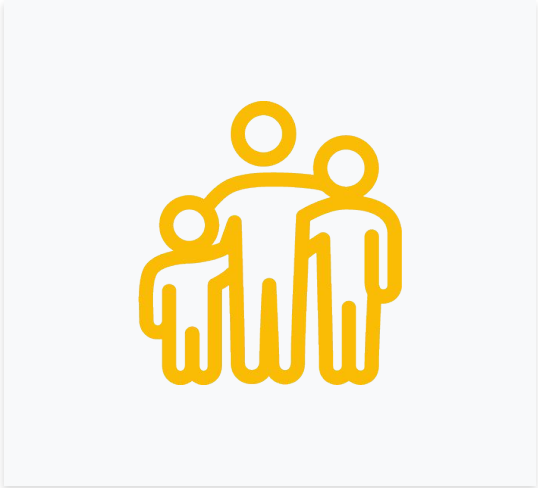
Microsoft



(Reddi et al., 2021)

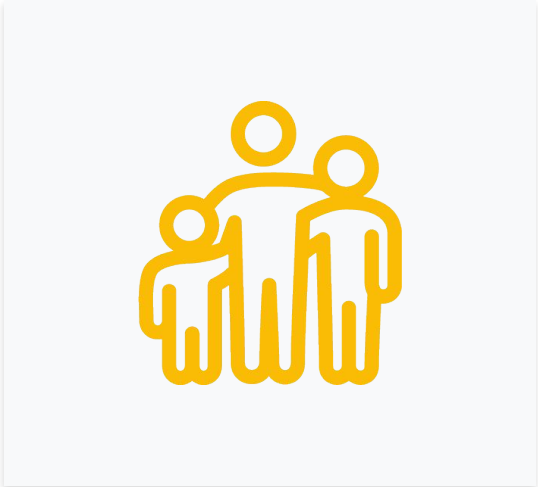
How do we build 1000 words for 1000 languages?

Cost Model v. **Community** Model?



Cost Model v. **Community** Model?

Limited Scale



Social Good

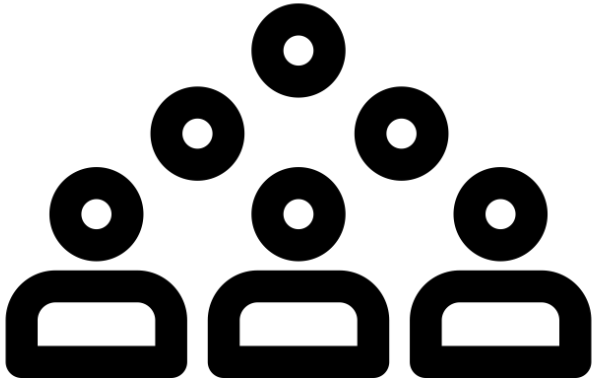
Common Voice

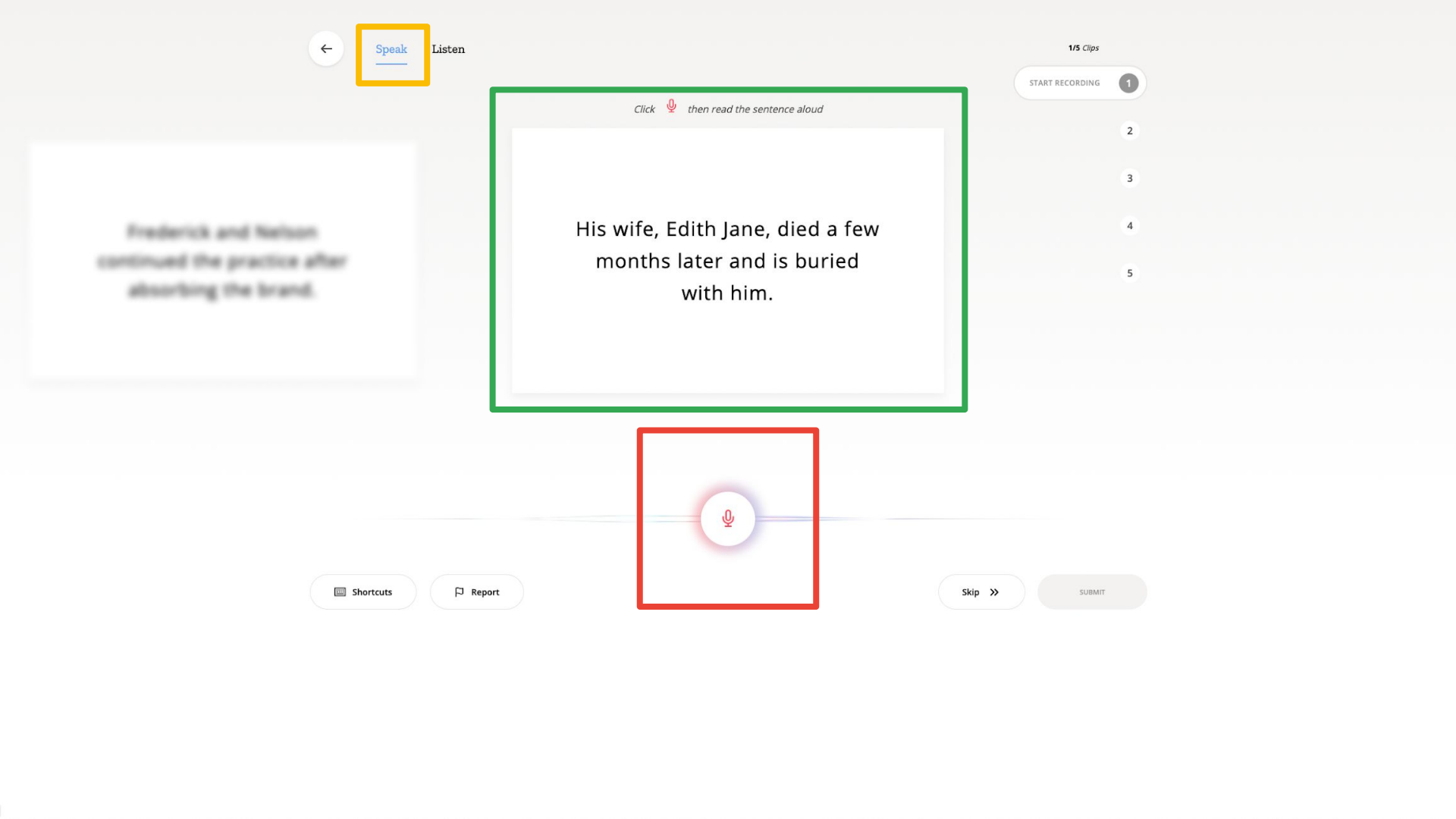
https://commonvoice.mozilla.org

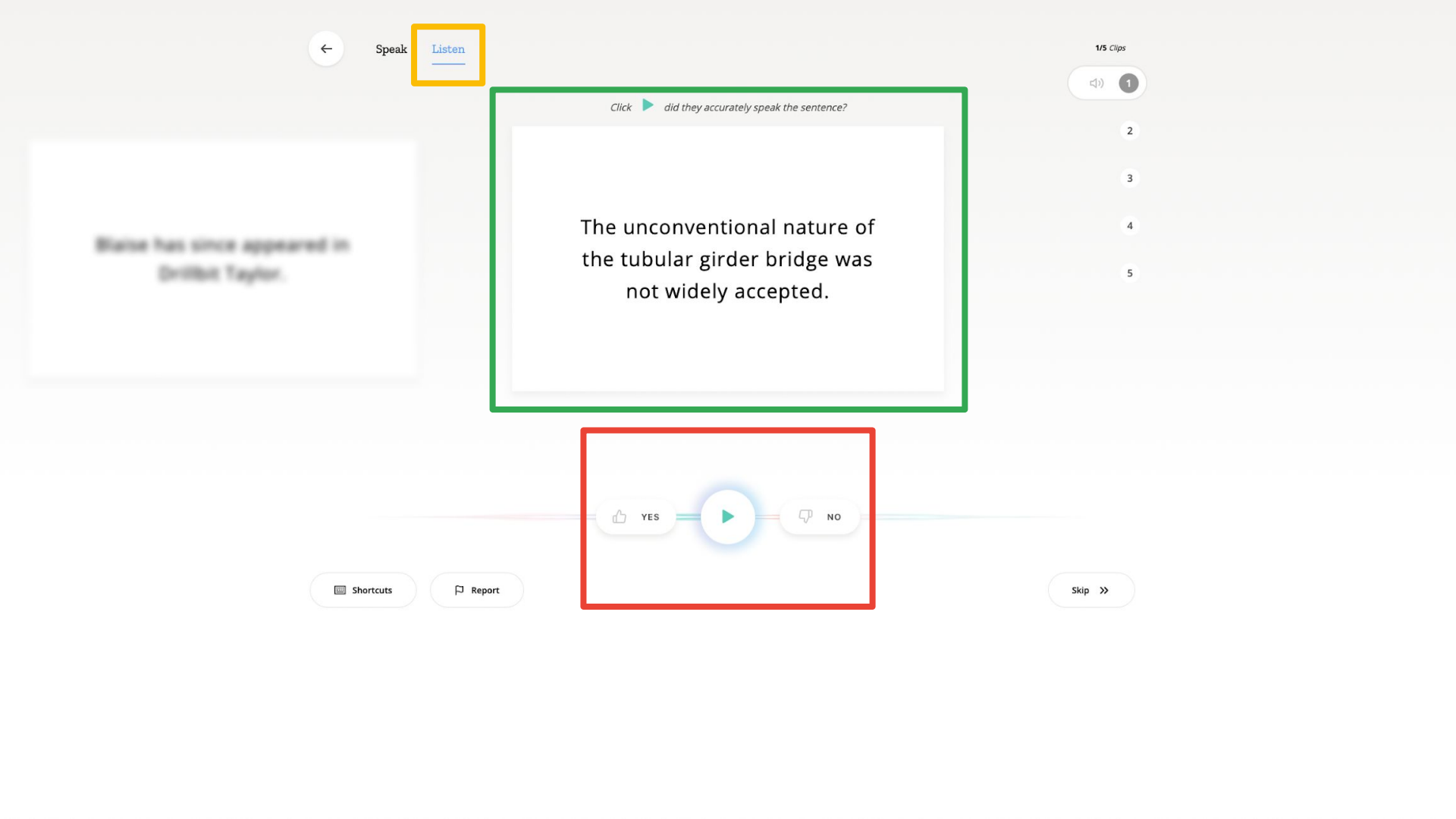
(Ardila et al., 2019)

Common Voice

- **Crowdsourcing** platform

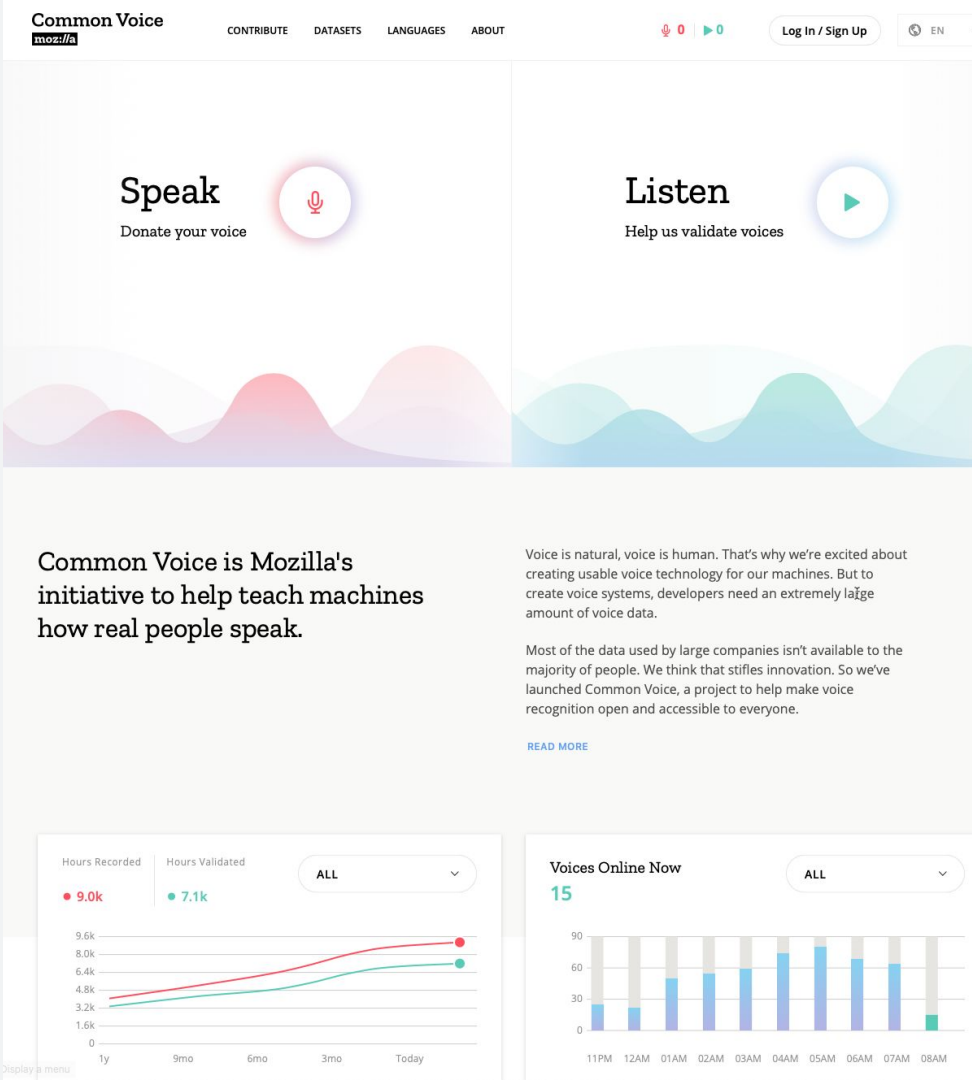



Speak
Listen
1/5 Clips
START RECORDING
1
2
3
4
5
Click
then read the sentence aloud
His wife, Edith Jane, died a few months later and is buried with him.
Shortcuts
Report
Skip
SUBMIT


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

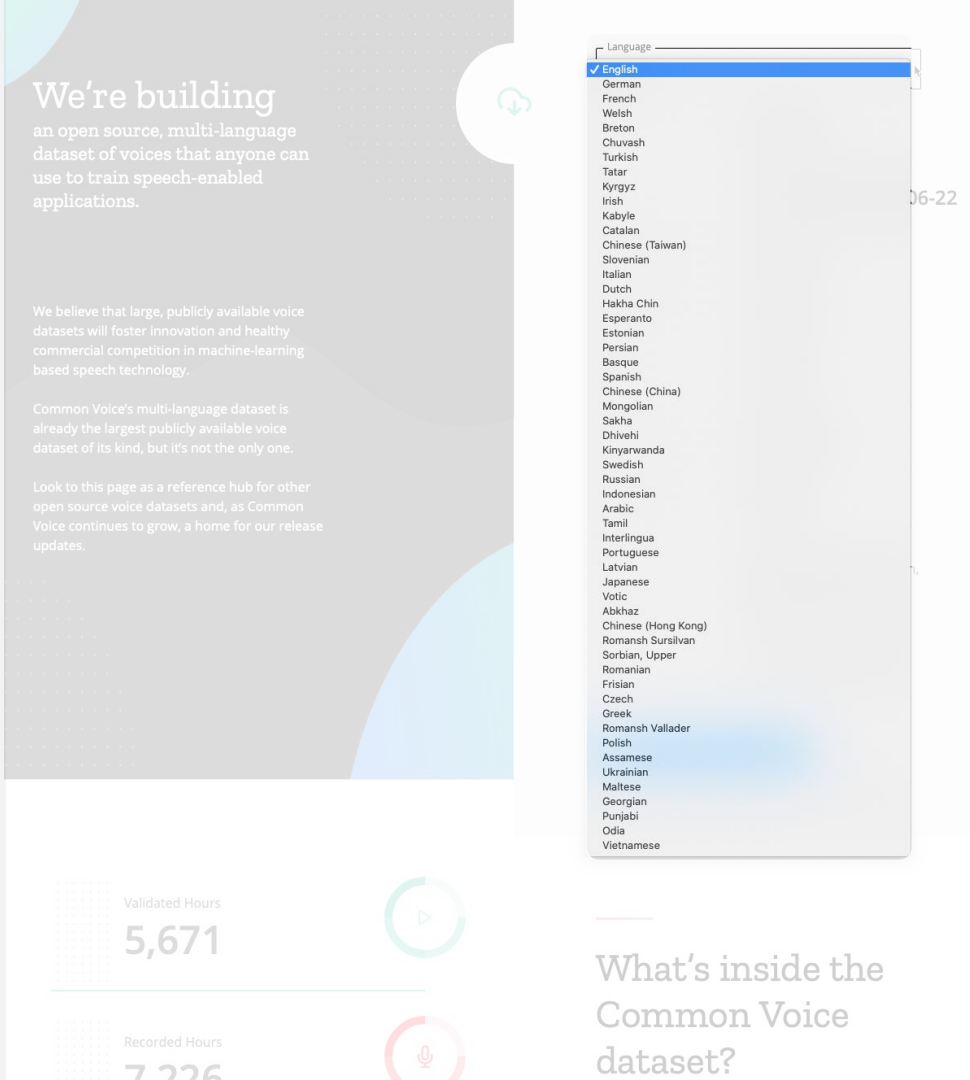
Common Voice

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



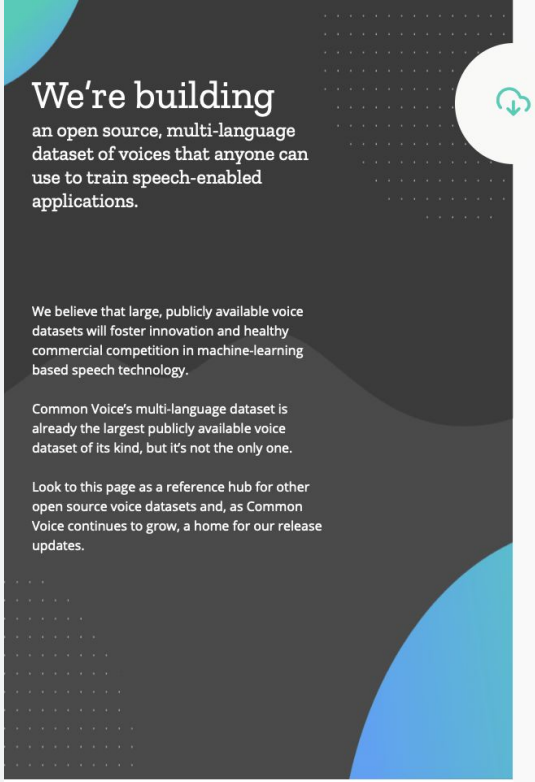
Common Voice

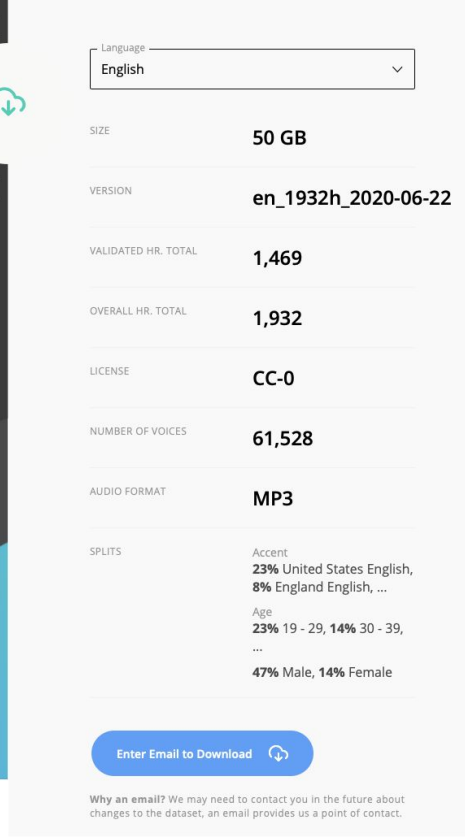
- **Crowdsourcing** platform
- Over **50,000 volunteers**
- **54 different languages**

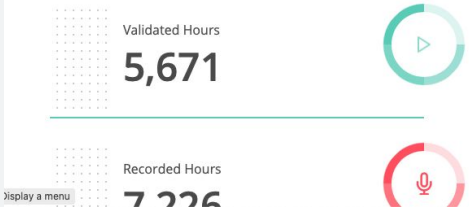


Common Voice

- **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?

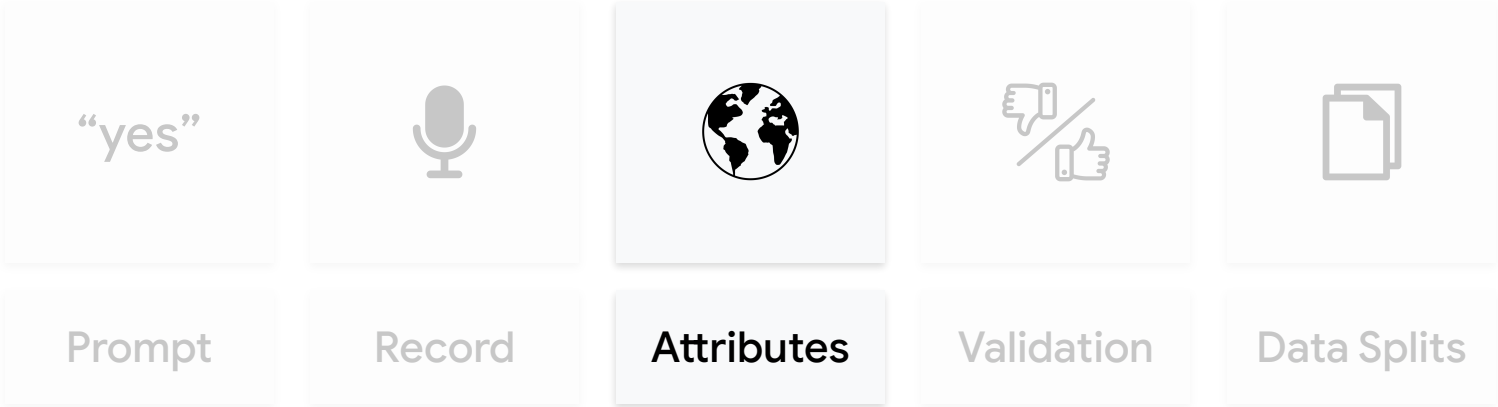
Common Voice **Tasks**

“yes”				
Prompt	Record	Attributes	Validation	Data Splits

Common Voice **Tasks**

“yes”				
Prompt	**Record**	Attributes	Validation	Data Splits

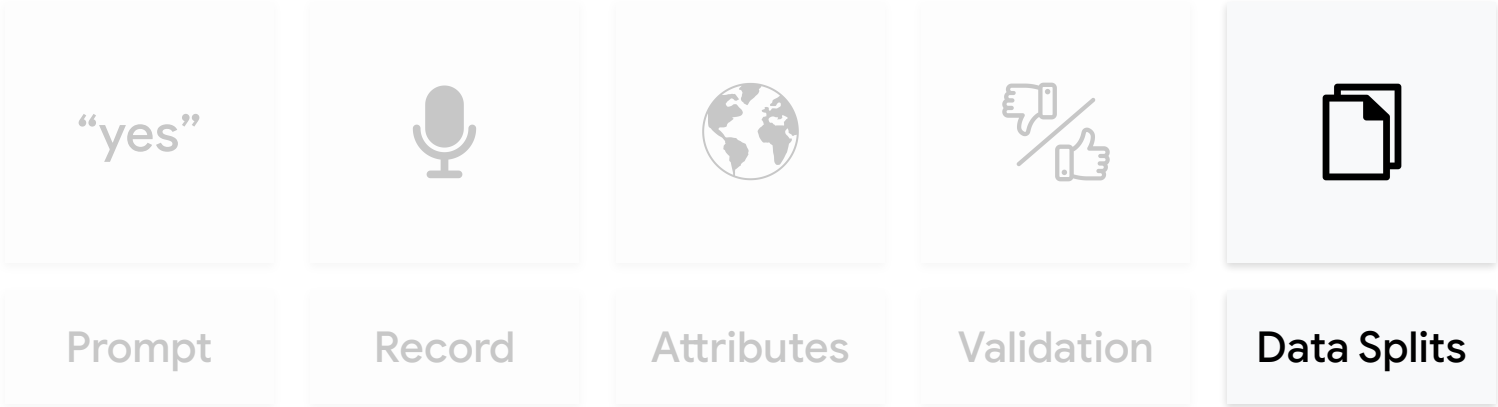
Common Voice **Tasks**



Common Voice **Tasks**

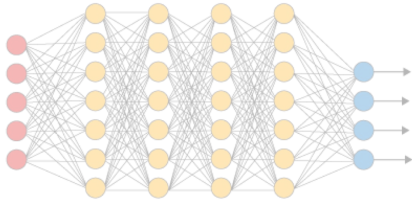
“yes”				
Prompt	Record	Attributes	**Validation**	Data Splits

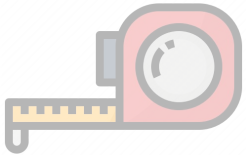
Common Voice **Tasks**



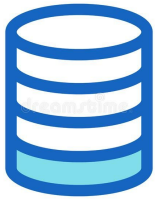
Common Voice **Tasks**

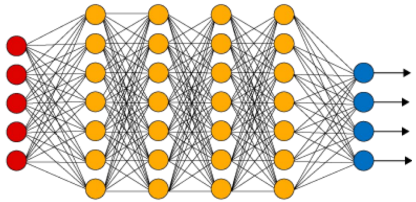
“yes”				
Prompt	Record	Attributes	Validation	Data Splits

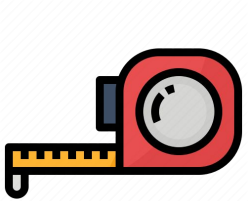




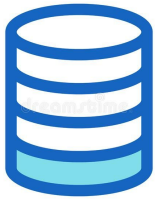
Collect Data → Preprocess Data → Design a Model → Train a Model → Evaluate Optimize → Convert Model → Deploy Model → Make Inferences







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



How do we democratize the data engineering pipeline?

Automatic Keyword Dataset Generation

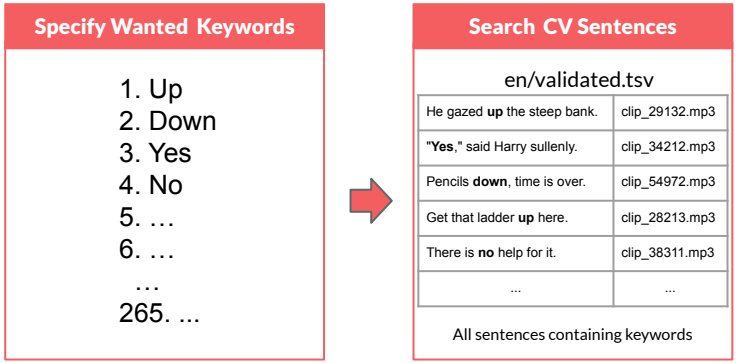
Specify Wanted Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

265. ...

Automatic Keyword Dataset Generation



Automatic Keyword Dataset Generation

Specify Wanted Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

265. …

Search CV Sentences

en/validated.tsv

He gazed **up** the steep bank.	clip_29132.mp3
"**Yes**," said Harry sullenly.	clip_34212.mp3
Pencils **down**, time is over.	clip_54972.mp3
Get that ladder **up** here.	clip_28213.mp3
There is **no** help for it.	clip_38311.mp3
…	…

All sentences containing keywords

Extract Keyword Utterances

Input 5-10 second mp3

*"… gazed **up** the steep…"*

up

Output

1 second wav, silence-padded

Automatic Keyword Dataset Generation

Specify Wanted Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

265. …

Search CV Sentences

en/validated.tsv

He gazed **up** the steep bank.	clip_29132.mp3
"**Yes**," said Harry sullenly.	clip_34212.mp3
Pencils **down**, time is over.	clip_54972.mp3
Get that ladder **up** here.	clip_28213.mp3
There is **no** help for it.	clip_38311.mp3
…	…

All sentences containing keywords

Extract Keyword Utterances

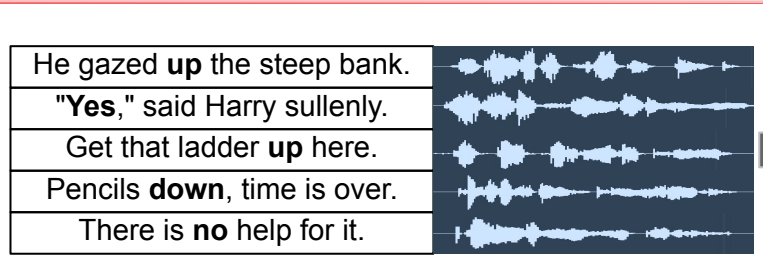
Input 5-10 second mp3

*"… gazed **up** the steep…"*

up

Output

1 second wav, silence-padded



Estimate Per-Word Timing

Extract Keywords

Large Keyword Dataset

4.3M Utterances
3,126 Keywords
22 Languages

Add

([Mazumder et al., 2021](#))

Automatic Keyword Dataset Generation

Specify Wanted Keywords

1. Up
2. Down
3. Yes
4. No
5. …
6. …

…

265. …

Search CV Sentences

en/validated.tsv

He gazed **up** the steep bank.	clip_29132.mp3
"Yes," said Harry sullenly.	clip_34212.mp3
Pencils **down**, time is over.	clip_54972.mp3
Get that ladder **up** here.	clip_28213.mp3
There is **no** help for it.	clip_38311.mp3
…	…

All sentences containing keywords

Extract Keyword Utterances

Input 5-10 second mp3

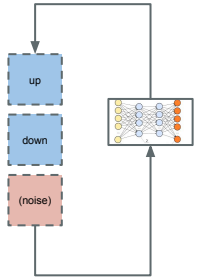
*"… gazed **up** the steep…"*

up

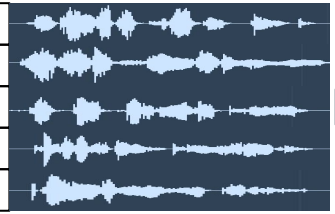
Output

1 second wav, silence-padded

Train KWS Model



He gazed **up** the steep bank.
"Yes," said Harry sullenly.
Get that ladder **up** here.
Pencils **down**, time is over.
There is **no** help for it.



Estimate Per-Word Timing

Extract Keywords

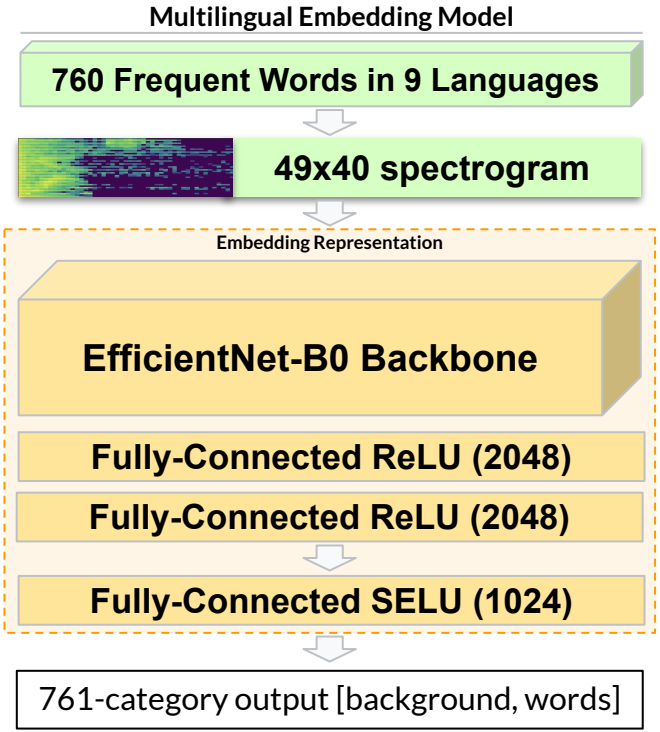
Large Keyword Dataset

4.3M Utterances
3,126 Keywords
22 Languages

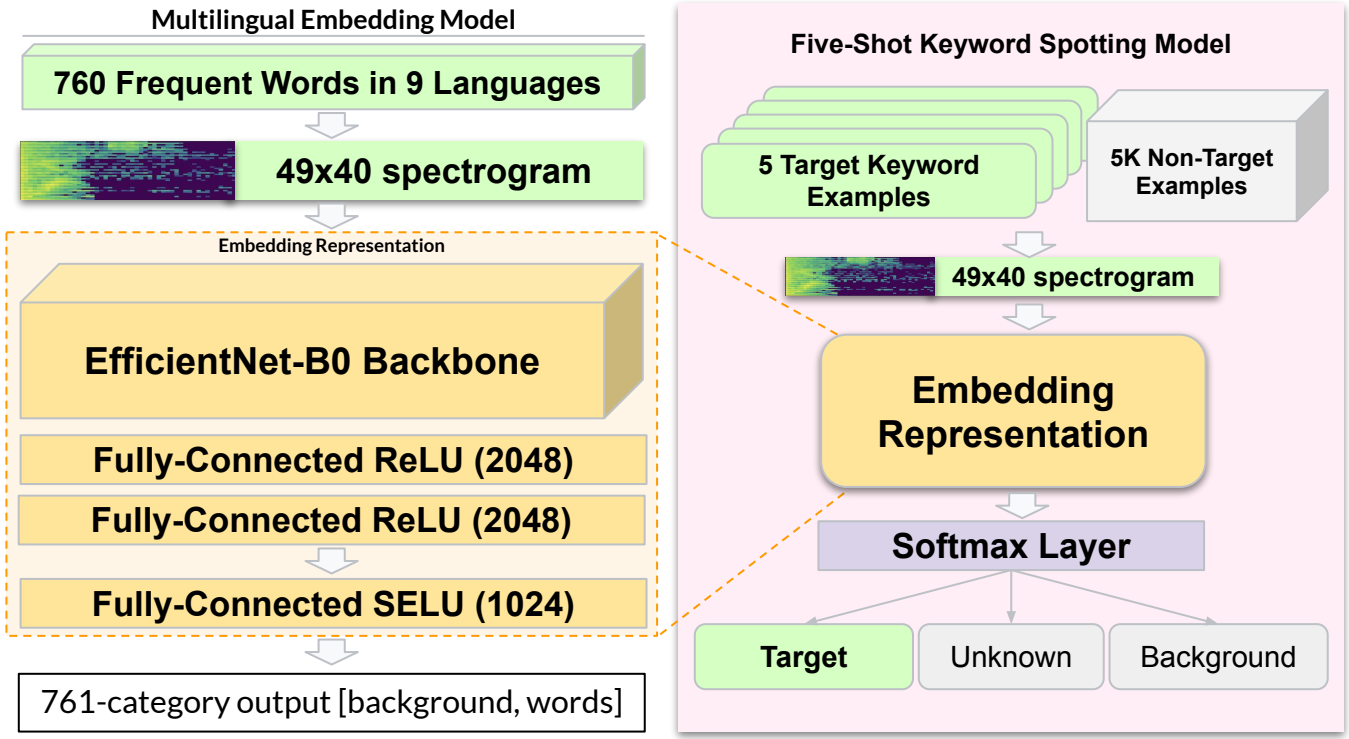
Add

(Mazumder et al., 2021)

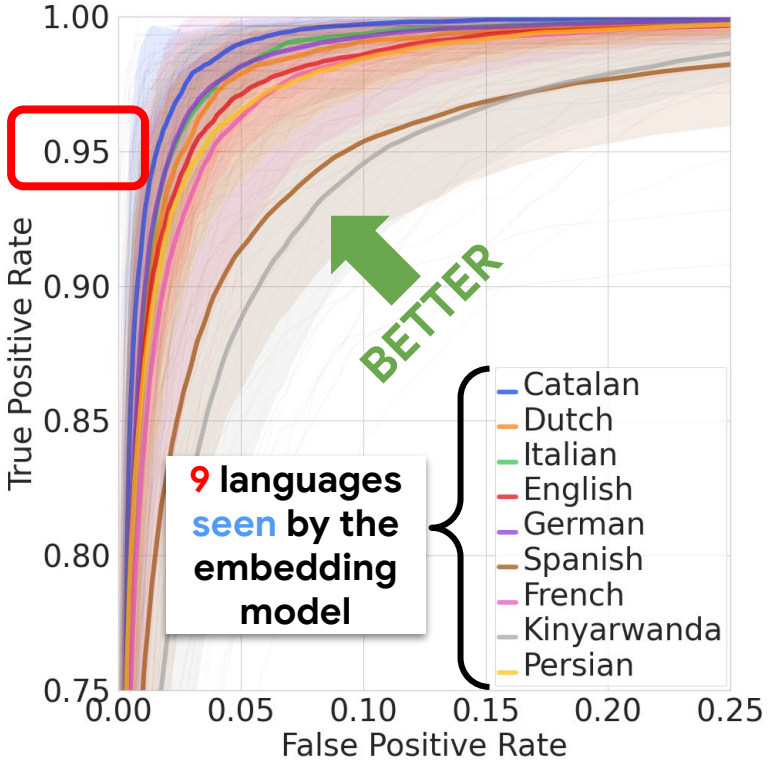
Nine-Language Embedding Model



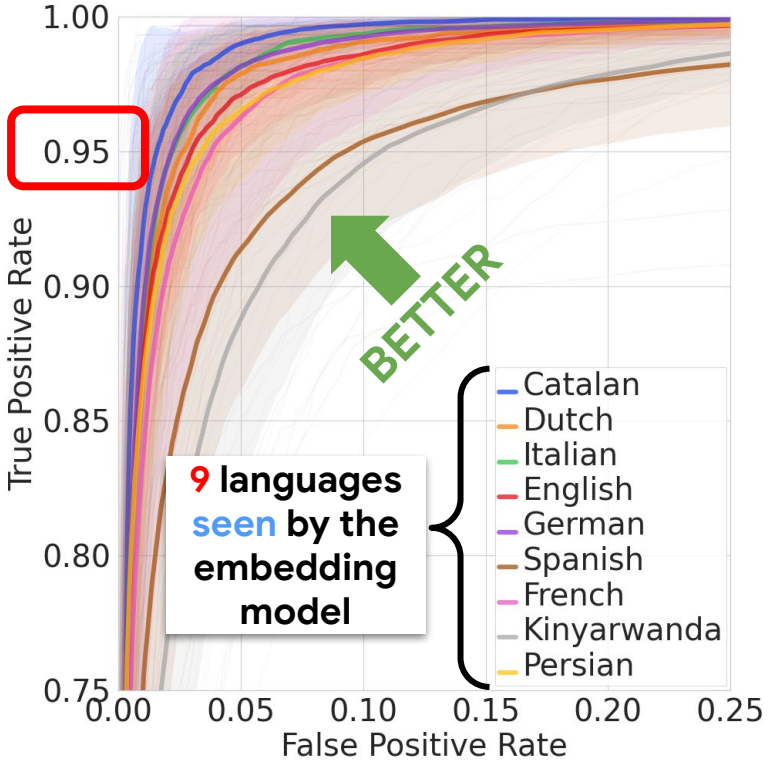
Nine-Language Embedding Model

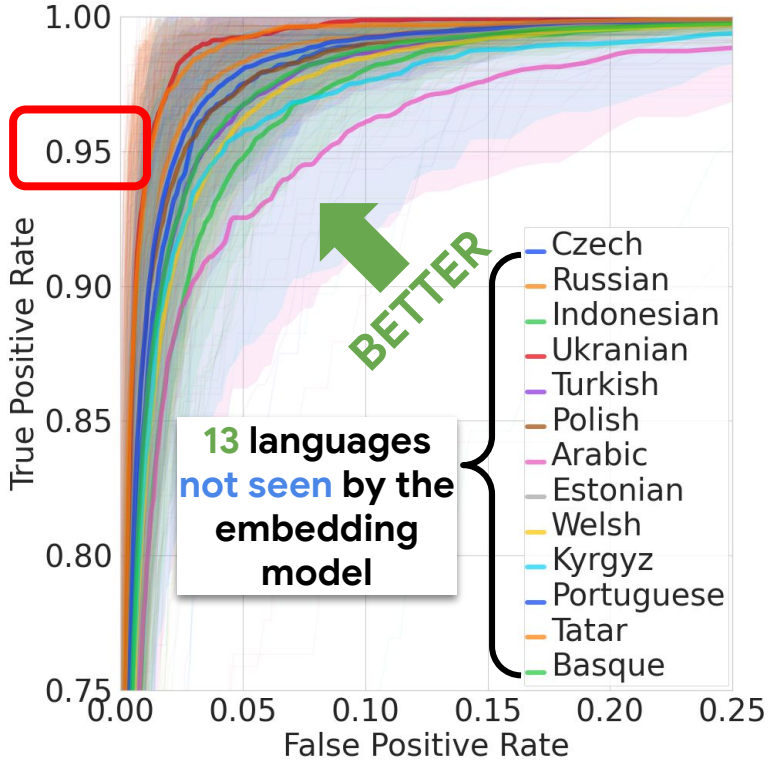


Generalizing to **Any** Language

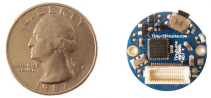


Generalizing to **Any** Language

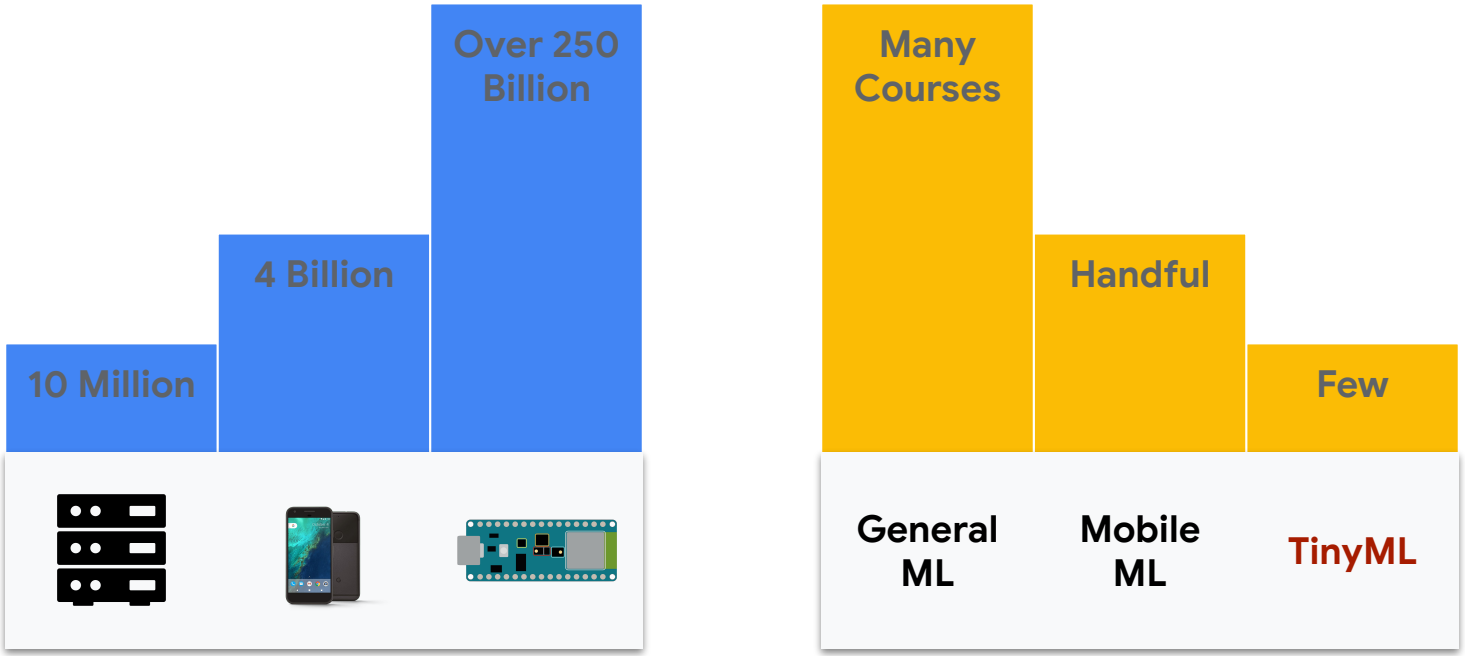





10:00
5G
Monday, Aug 3
68°F
5



How do we widen access to applied machine learning?


Over 250 Billion
4 Billion
10 Million
Many Courses
Handful
Few
General ML
Mobile ML
TinyML

Welcome to the Tiny Machine Learning Open Education Initiative (TinyMLx)

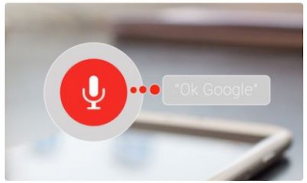
TinyML is a cutting-edge field that brings the transformative power of machine learning (ML) to the performance- and power-constrained domain of embedded systems. Successful deployment in this field requires intimate knowledge of applications, algorithms, hardware, and software.

Take an edX Course to Lean More **Build and Teach your own TinyML Course**

Take one of our exciting courses!

Foundations of TinyML

Focusing on the basics of machine learning and embedded systems, such as smartphones, this course will introduce you to the "language" of TinyML.



Applications Of TinyML

Get the opportunity to see TinyML in practice. You will see examples of TinyML applications, and learn first-hand how to train these models for tiny applications such as keyword spotting, visual wake words, and gesture

Deploying TinyML

Learn to program in TensorFlow Lite for microcontrollers so that you can write the code, and deploy your model to your very own tiny microcontroller. Before you know it, you'll be implementing an entire TinyML application.

An Applied Machine Learning Journey

Course 1

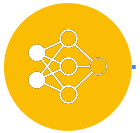
Fundamentals of TinyML

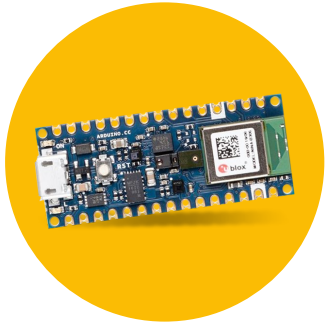
Course 2

Applications of TinyML

Course 3

Deploying TinyML






ARDUINO
TINY MACHINE
LEARNING KIT
edX
TensorFlow

Scribing

Topics to Scribe

- Data Engineering Challenges
- Data Engineering Collection Methods
- Democratizing the Data Engineering Pipeline

Paper Discussions

Paper 1: Model Cards for Model Reporting

Link

Paper 2: Multilingual Spoken Words Corpus

Link

Guest Speaker

Guest Speaker – Kasia Chmielinski

Kasia is the Co-Founder of the Data Nutrition Project and will dive into the importance of understanding datasets beyond numbers. We will learn why nutritional labels are crucial for responsible AI and how we can interpret them effectively.

With hands-on experience mitigating bias in AI and building tools for responsible data systems, Kasia brings a unique and crucial perspective to our exploration of tiny machine learning.