

Yá'át'ééh 

CRESTLEX 3.0

CREating Effective STEM
Learning EXperiences

with Navajo Tech





I'm Monsuru!

I am an Assistant Professor
at **Navajo Technical University**



I'm Brian!

I am a Robotics PhD Candidate and
teach AI/ML at **Harvard University**.

Using Zoom

- **Ask questions**
 - Use the Zoom chat
 - Raise your hand
 - Interrupt us!
- Cameras & Microphone
 - Camera on if you'd like
 - **Mic muted** unless speaking

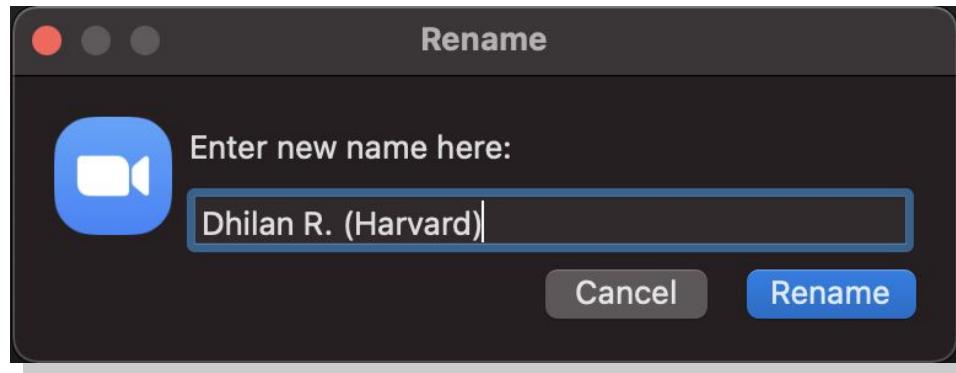


Participants (1)

Dhilan Rama... (Host, me) Unmute More

Rename

Edit Profile Picture



Our website!

tinymlx.org/CRESTLEX3/

home base for **all information!**

Workshop Agenda

Morning Session (9–11am)

The Future of AI (with Laurence)

How ML Works? (with Dhilan)

ML in the Navajo Nation (with NTU)

Responsible AI (with Susan)

Afternoon Session (12pm–2pm)

Experimenting with AI (with Dhilan)

Exploring ML (with Jenny)

Build it! Your own app (with Jenny)

What's Next?

What's Next?

(with guest student panel)

(with instructors)

Workshop Agenda

Morning Session (9–11am)

The Future of AI (with Laurence)

How ML Works? (with Dhilan)

ML in the Navajo Nation (with NTU)

Responsible AI (with Susan)

Afternoon Session (12pm–2pm)

Experimenting with AI (with Dhilan)

Exploring ML (with Jenny)

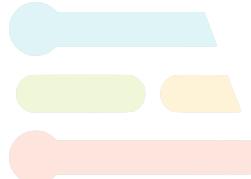
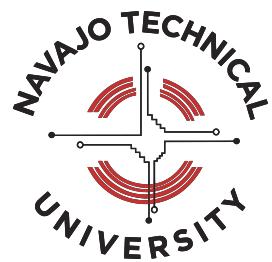
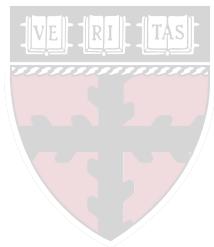
Build it! Your own app (with Jenny)

What's Next?

(with guest student panel)

What's Next?

(with instructors)



A quick recap!



AI and ML for today and tomorrow...



Laurence Moroney
@l moroney

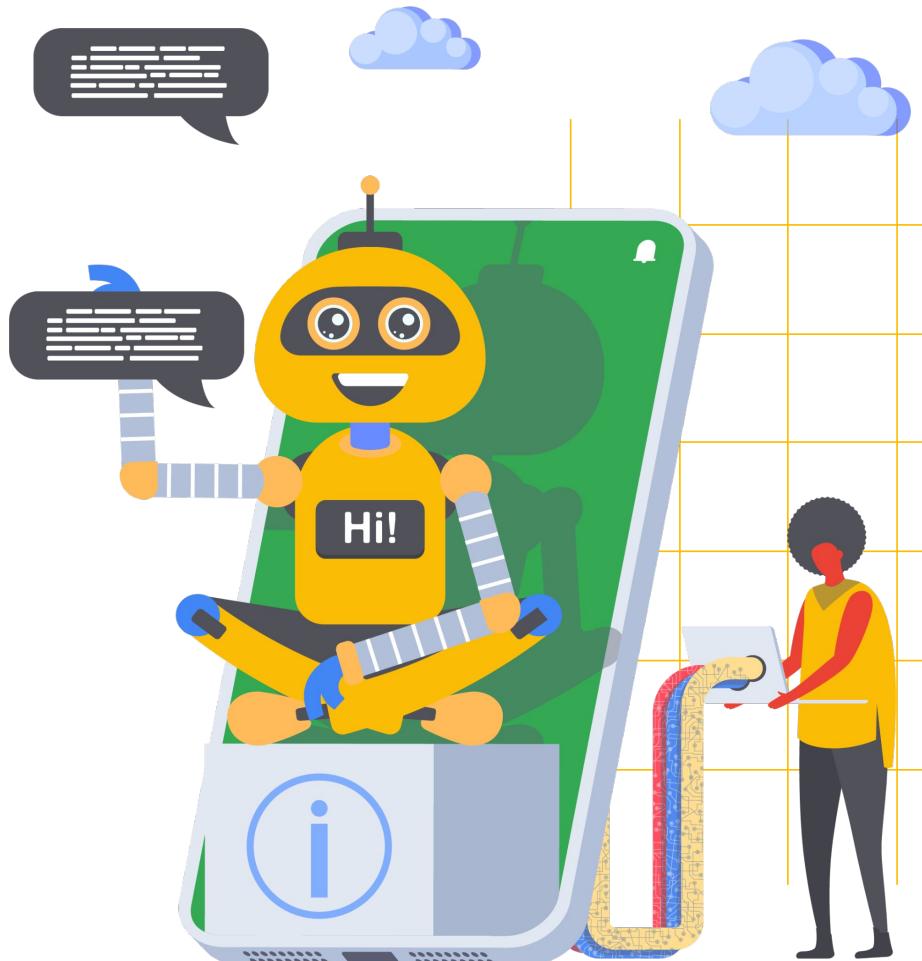
Experimenting with AI

with Dhilan



How ML works?

with Dhilan





deploy

to your tiny **devices!**

Review what we've learned

Machine learning provides a computer with data, **rather than explicit instructions**. Using these data, the computer learns to **recognize patterns** and becomes able to execute tasks on its own.

Training the machine



WE PROVIDE

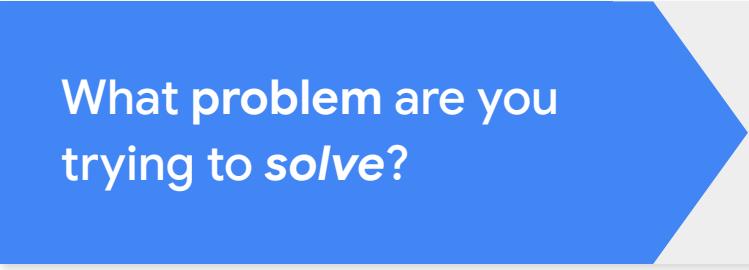
LABELS

INPUTS



RULES

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 variety of training examples
- Correct labels (answers)
- Good Balance (e.g., dog, cat, random)

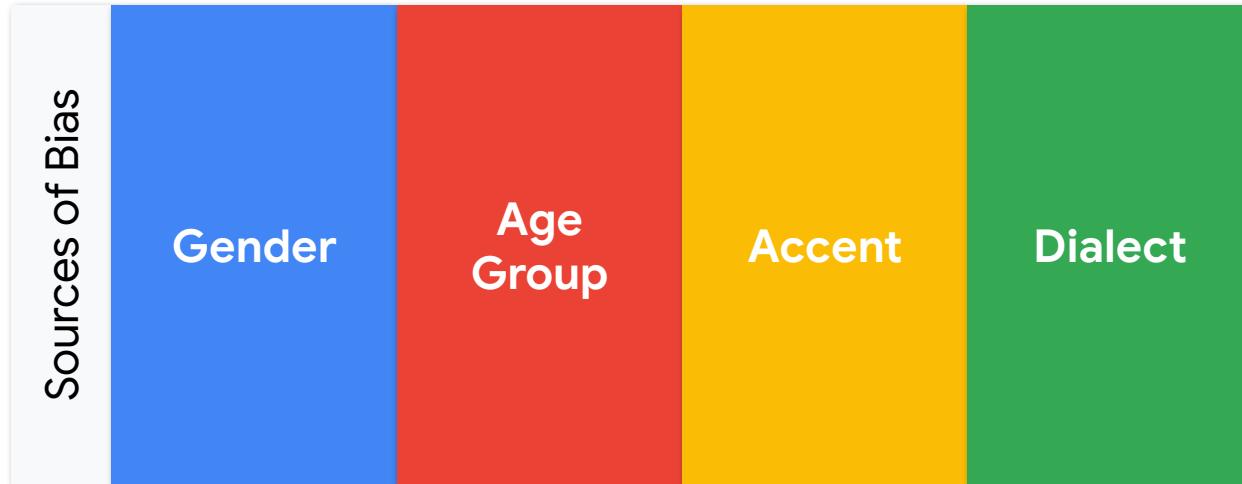
What do 50 million drawings look like?



What do 50 million drawings look like?



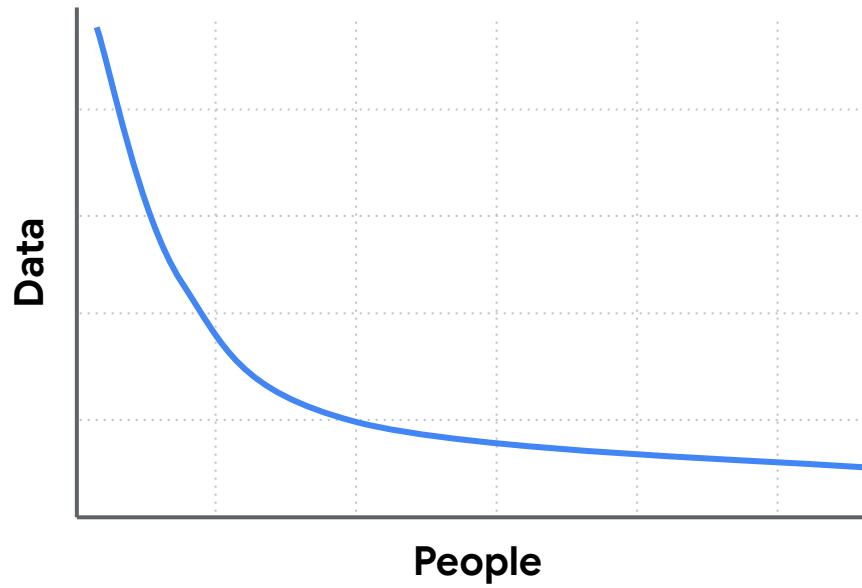
Potential **Bias** in Speech Recognition

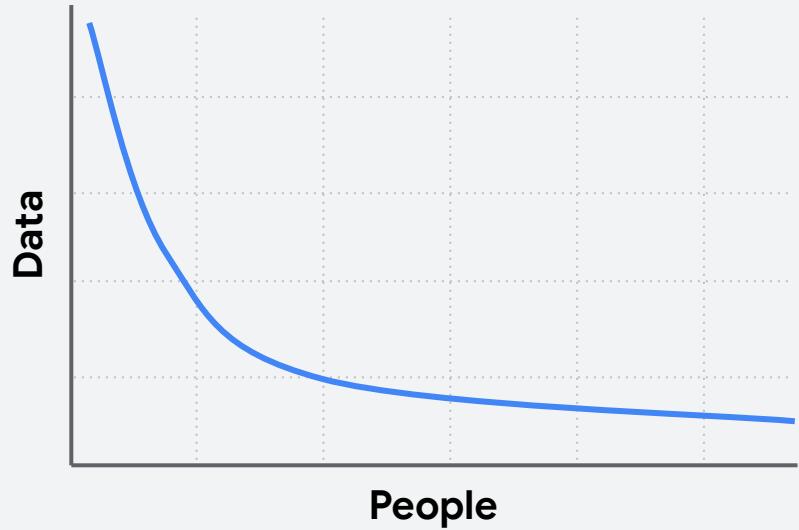
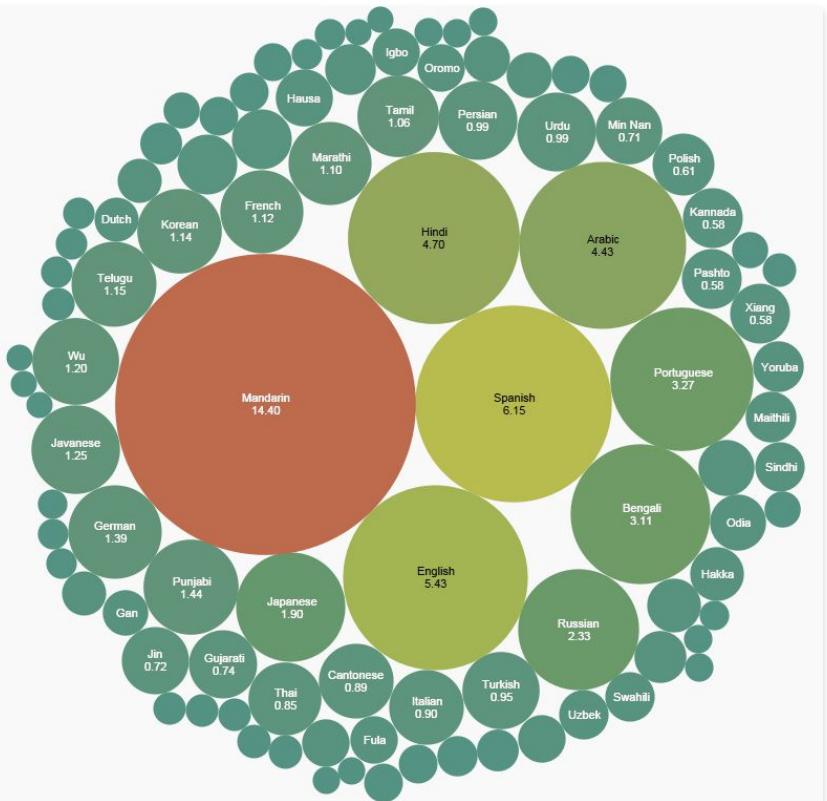


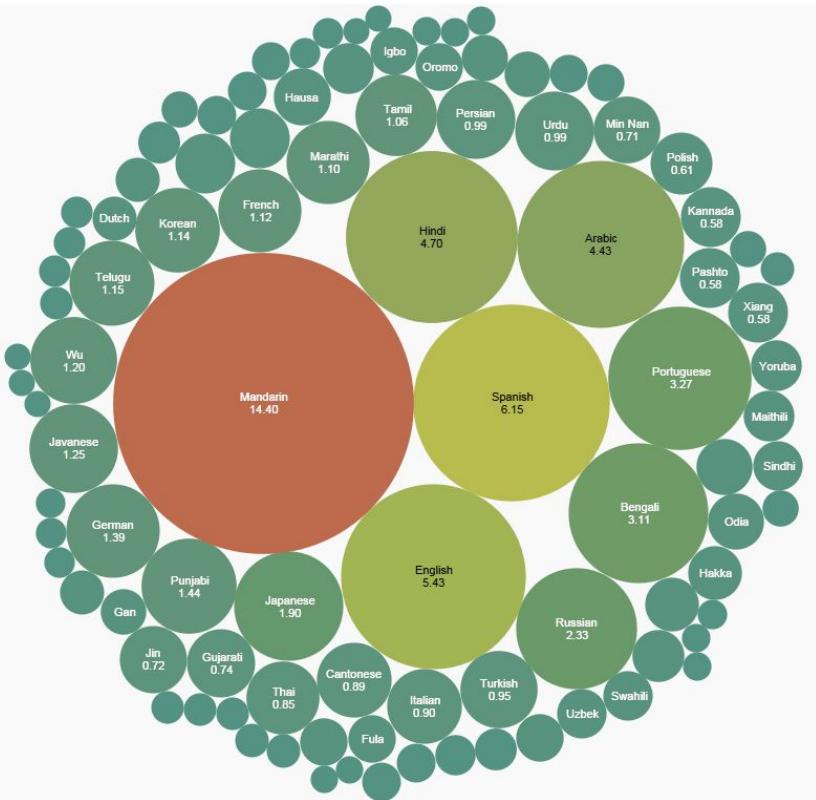
what's missing?

datasets are imperfect

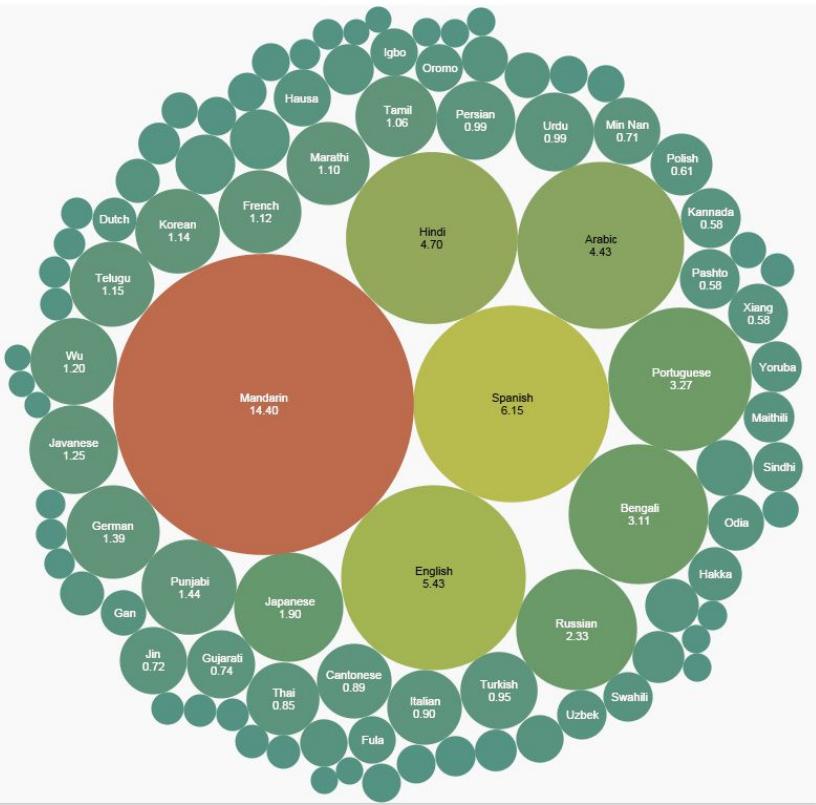
why? how can we **fix them?**



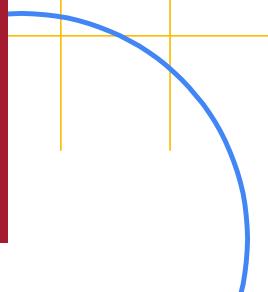
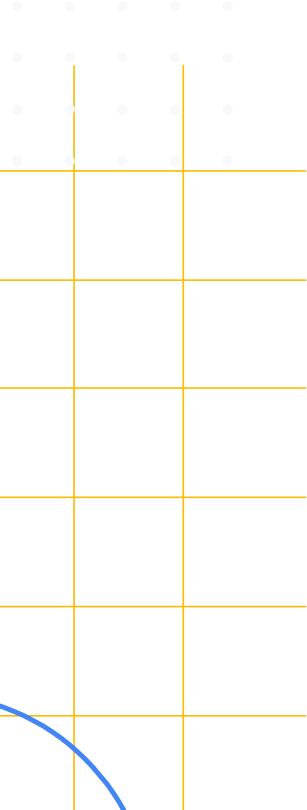




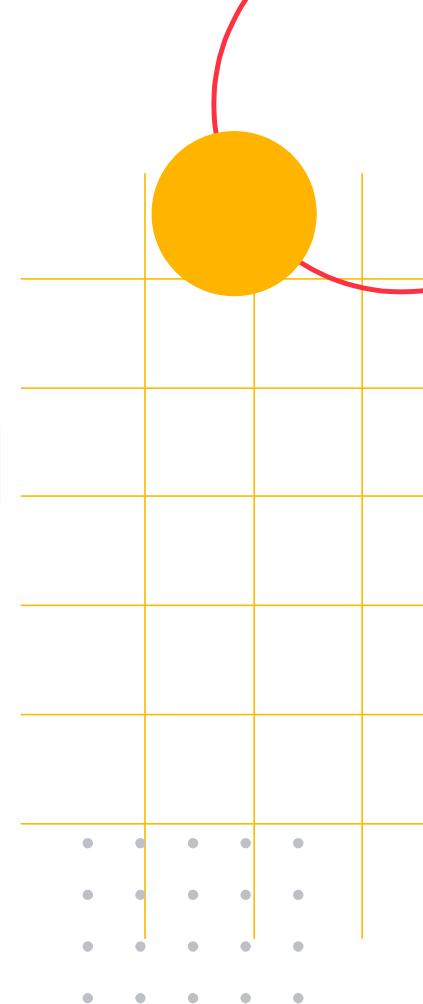
- Speech commands for the **whole planet?**



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



Crowdsourcing to the rescue!

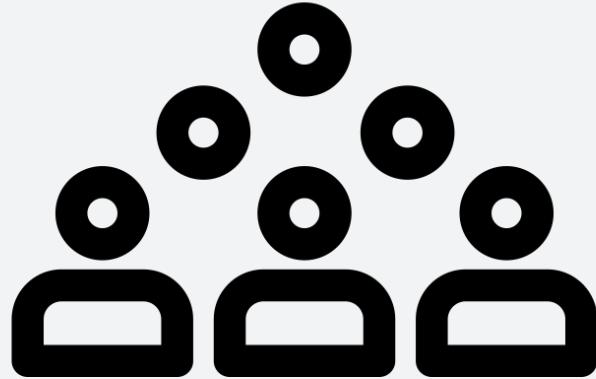


Common Voice

<https://commonvoice.mozilla.org>

Common Voice

- **Crowdsourcing** platform



Frederick and Nelson
continued the practice after
aborting the brand.

←

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

Bianca has since appeared in
Brilliant 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 >>

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

Help us validate voices

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**

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.

Validated Hours
5,671

Recorded Hours
7,226

Language

English
German
French
Welsh
Breton
Chuvash
Turkish
Tatar
Kyrgyz
Irish
Kabyle
Catalan
Chinese (Taiwan)
Slovenian
Italian
Dutch
Hakha Chin
Esperanto
Estonian
Persian
Basque
Spanish
Chinese (China)
Mongolian
Sakha
Dhivehi
Kinyarwanda
Swedish
Russian
Indonesian
Arabic
Tamil
Interlingua
portuguese
Latvian
Japanese
Votic
Abkhaz
Chinese (Hong Kong)
Romansh Sursilvan
Sorbian, Upper
Romanian
Frisian
Czech
Greek
Romansh Vallader
Polish
Assamese
Ukrainian
Maltese
Georgian
Punjabi
Odia
Vietnamese

What's inside the Common Voice dataset?

Common Voice

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

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.

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,
...
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?

File Structure

- **Valid**
 - At least 2 people listen to them, and the majority of those listeners say the audio matches the text



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



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

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

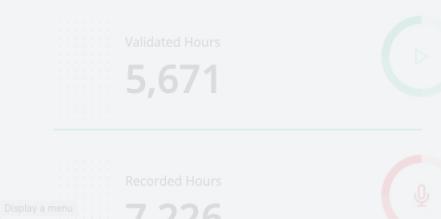
NUMBER OF VOICES 61,528

AUDIO FORMAT MP3

SPLITS Accent
23% United States English,
8% England English, ...
Age
23% 19 - 29, 14% 30 - 39,
...
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?

Interesting Attributes

- Permissive license
- Many contributors

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, ... 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.



Display a menu

What's inside the Common Voice dataset?

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.

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, ...
	47% Male, 14% Female

Enter Email to Download



What's inside the Common Voice dataset?

Common Voice

Download the Single Word Target Segment

This is a use case driven segment containing data to power spoken digit recognition, yes / no detection, and wakeword testing data for [Firefox Voice](#).



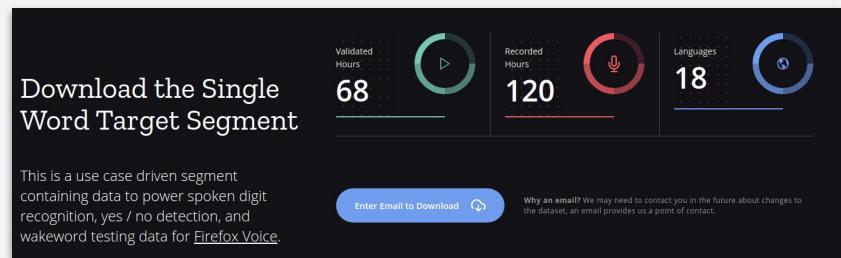
Enter Email to Download

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.

Single Word Target Segment

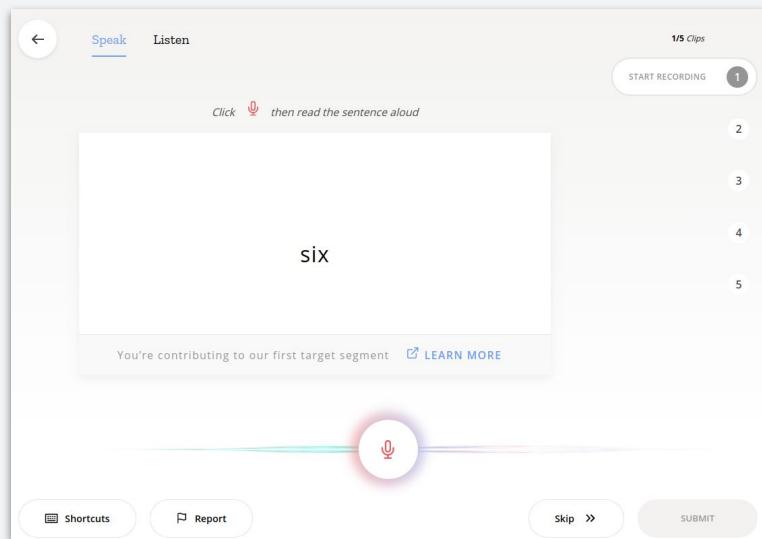
A *speech commands-style* dataset for **18 languages**

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



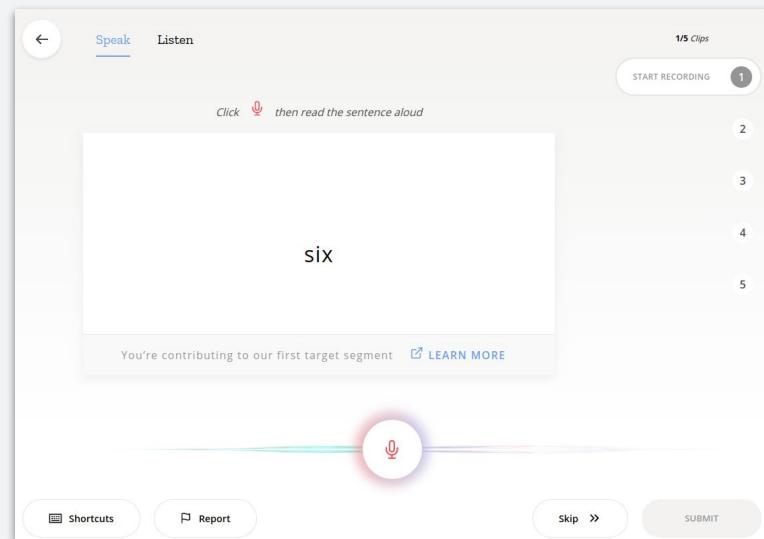
Speech Recognition Diversity and Reach

- **Common Voice**
 - Permissive license
 - Minority languages



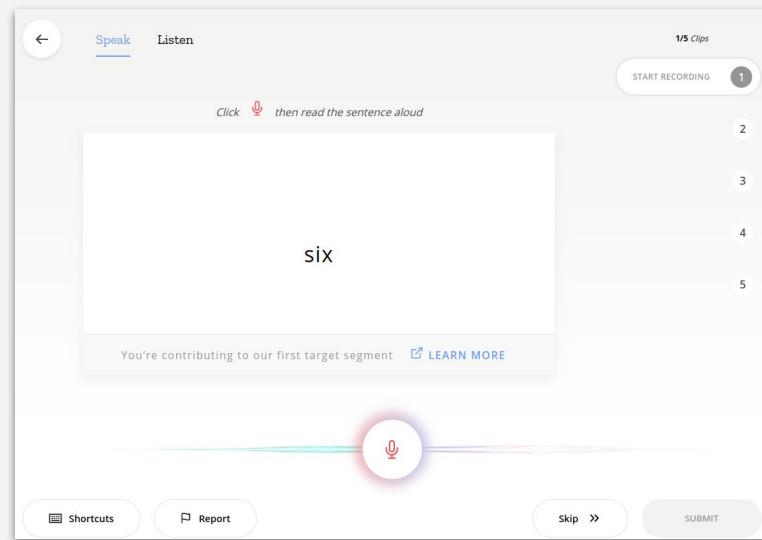
Speech Recognition Diversity and Reach

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



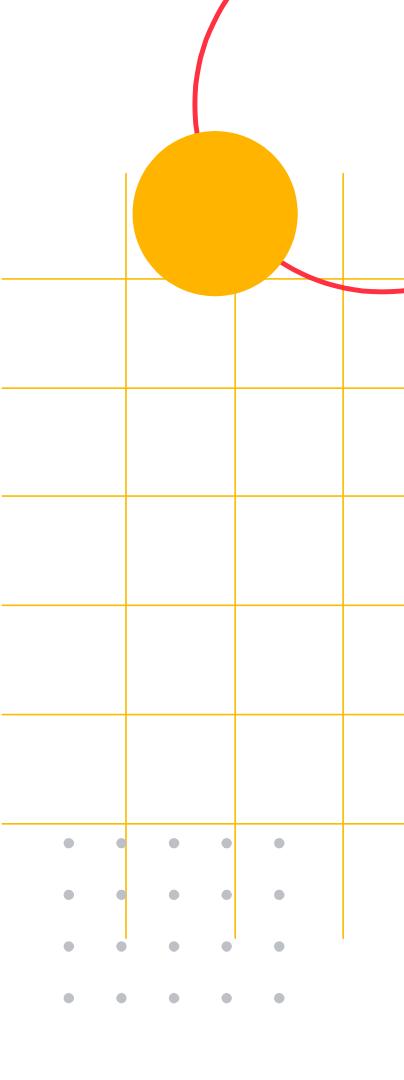
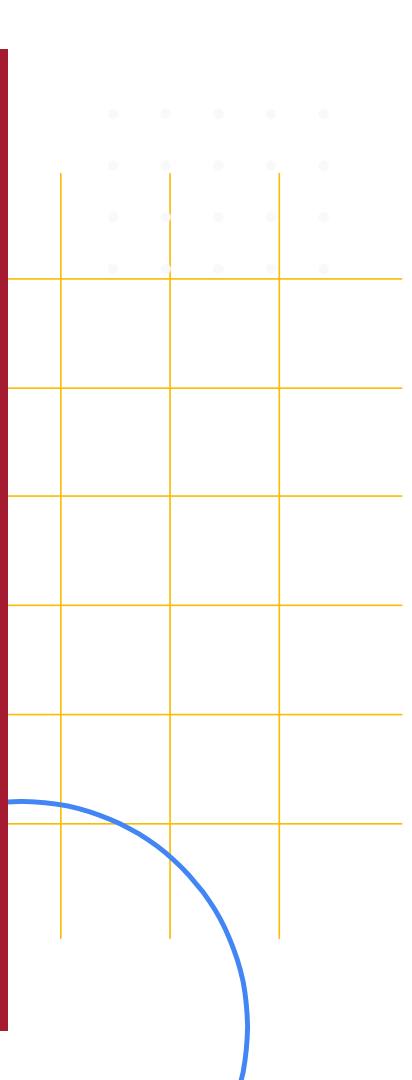
Speech Recognition Diversity and Reach

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



Common Voice

<https://commonvoice.mozilla.org>



today!

ML in the Navajo Nation

How can **you** use ML
in your community!