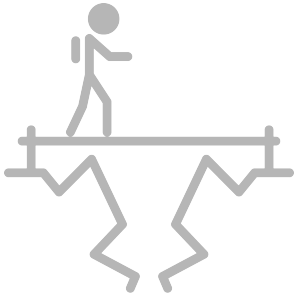
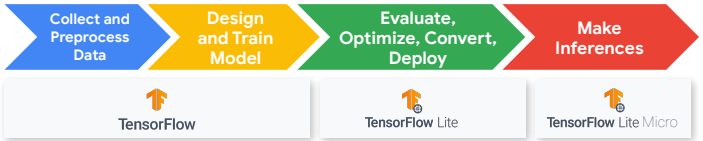
Keyword Spotting Datasets

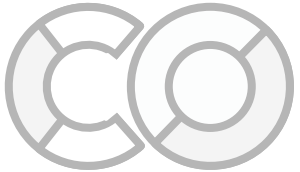
What are we going to learn?



Challenges with Keyword Spotting



The Keyword Spotting ML **Pipeline**



Hands-on training of a Keyword Spotting Model

What does it mean to have a **good** dataset?

So how do we **build** a good dataset?

- Who are the **users**?

So how do we **build** a good dataset?

- Who are the **users**?
- What do they **need**?

So how do we **build** a good dataset?

- Who are the **users**?
- What do they **need**?
- What **task** are they trying to solve?

So how do we **build** a good dataset?

- Who are the **users**?
- What do they **need**?
- What **task** are they trying to solve?
- How do they **interact** with the system?

So how do we **build** a good dataset?

- Who are the **users**?
- What do they **need**?
- What **task** are they trying to solve?
- How do they **interact** with the system?
- How does the **real world** make this hard?

Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition

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

April 2018

The **Speech Commands Dataset**

- Recorded as individual **words** not sentences

The **Speech Commands Dataset**

- Recorded as individual **words** not sentences
- 1000-4000 examples of each word

The **Speech Commands Dataset**

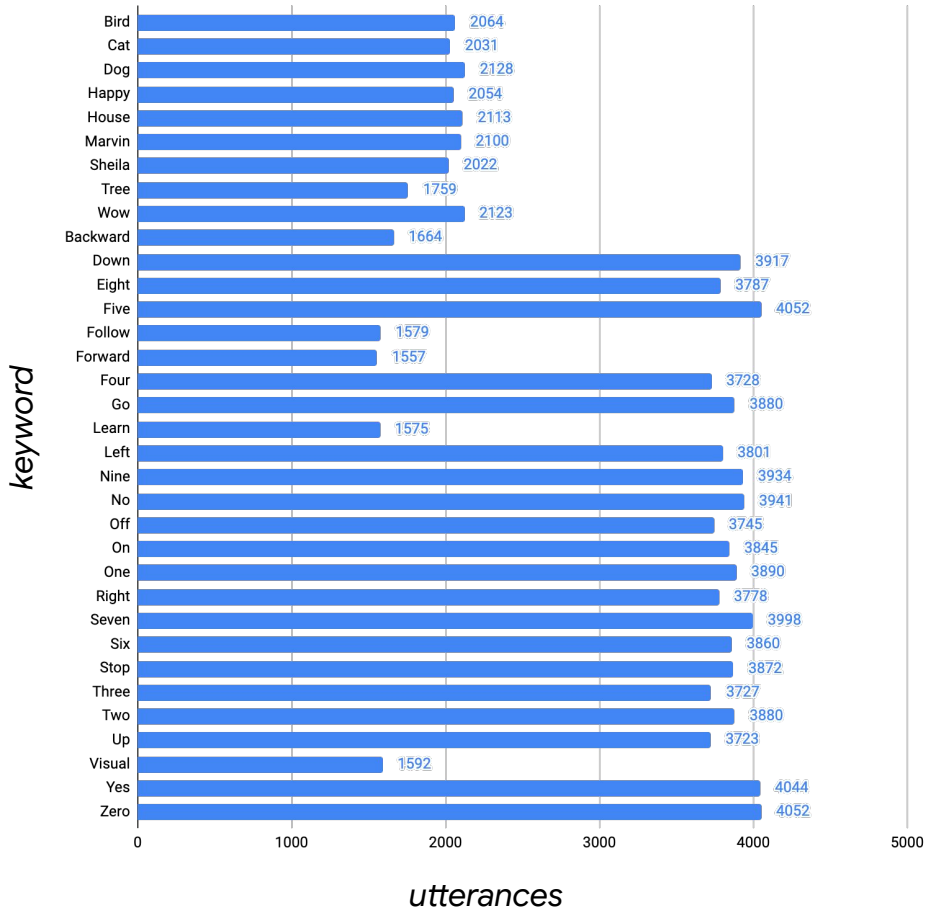
- Recorded as individual **words** not sentences
- 1000-4000 examples of each word
- >2,500 volunteers

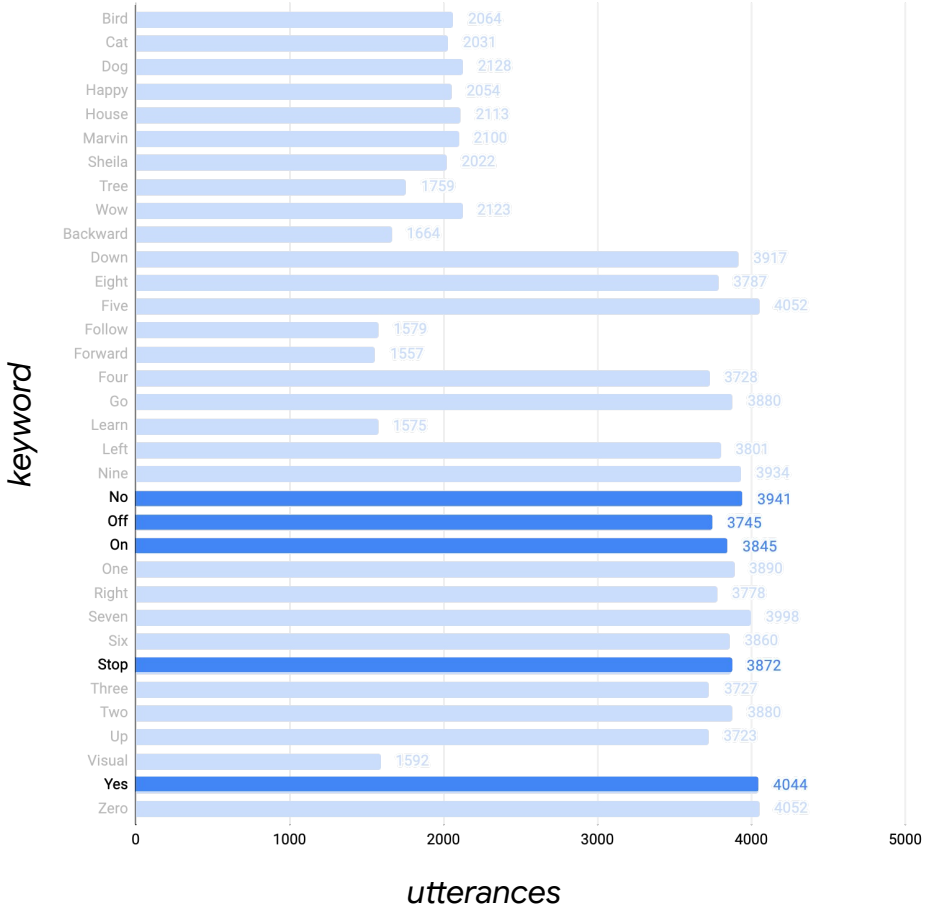
The **Speech Commands Dataset**

- Recorded as individual **words** not sentences
- 1000-4000 examples of each word
- >2,500 volunteers
- Representative of **real world audio** and includes background noise as well

The **Speech Commands Dataset**

- Recorded as individual **words** not sentences
- 1000-4000 examples of each word
- >2,500 volunteers
- Representative of **real world audio** and includes background noise as well
- **25 “IoT keywords”** + **10 “unknown words”** (with phonetic similarities: “three” vs “tree”)


keyword
Bird 2064
Cat 2031
Dog 2128
Happy 2054
House 2113
Marvin 2100
Sheila 2022
Tree 1759
Wow 2123
Backward 1664
Down 3917
Eight 3787
Five 4052
Follow 1579
Forward 1557
Four 3728
Go 3880
Learn 1575
Left 3801
Nine 3934
No 3941
Off 3745
On 3845
One 3890
Right 3778
Seven 3998
Six 3860
Stop 3872
Three 3727
Two 3880
Up 3723
Visual 1592
Yes 4044
Zero 4052
0
1000
2000
3000
4000
5000
utterances


keyword
Bird 2064
Cat 2031
Dog 2128
Happy 2054
House 2113
Marvin 2100
Sheila 2022
Tree 1759
Wow 2123
Backward 1664
Down 3917
Eight 3787
Five 4052
Follow 1579
Forward 1557
Four 3728
Go 3880
Learn 1575
Left 3801
Nine 3934
No 3941
Off 3745
On 3845
One 3890
Right 3778
Seven 3998
Six 3860
Stop 3872
Three 3727
Two 3880
Up 3723
Visual 1592
Yes 4044
Zero 4052
0
1000
2000
3000
4000
5000
utterances

Food for Thought

QC (Quality Control)

- Need to keep **only** what a human can hear
- Microphone issues
- **Noisy** backgrounds

(more on this soon)