Continuous Training: Data Validation

MLOps: Continuous Training

Dataset Sources

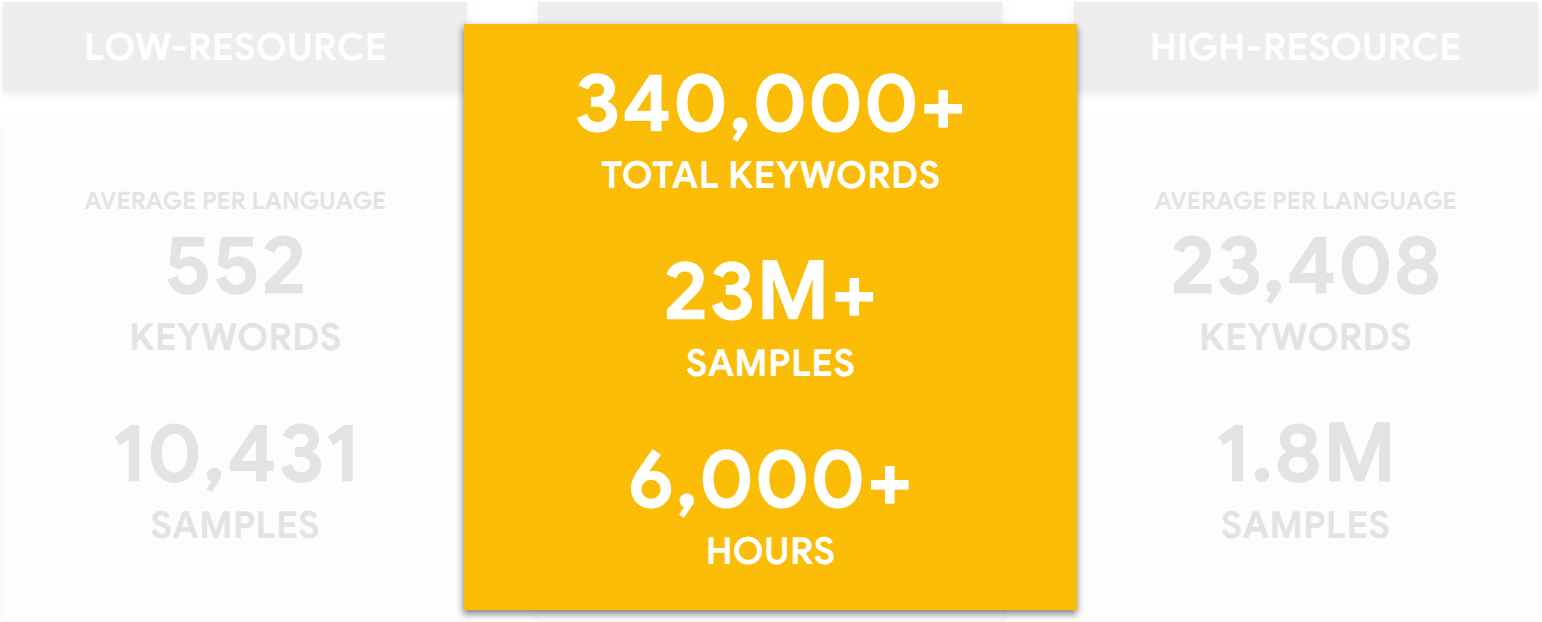
- **2,618** volunteers
  - consented to have their voices redistributed
  - Variety of accents
- > 1,000 examples for **each** keyword
- Started with **automated tools**
  - Remove low volume recordings
  - Extract loudest 1s (from 1.5sec examples)
- All 105,829 remaining utterances **manually reviewed** through crowdsourcing

Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition

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

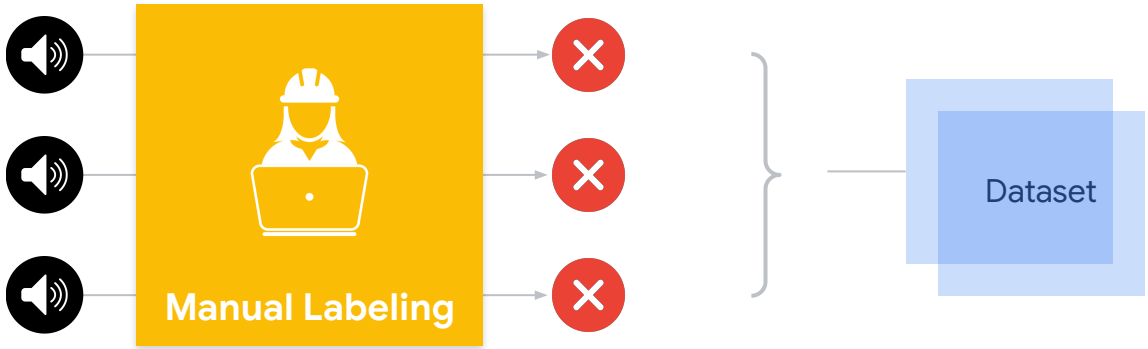
April 2018

Keywords per language

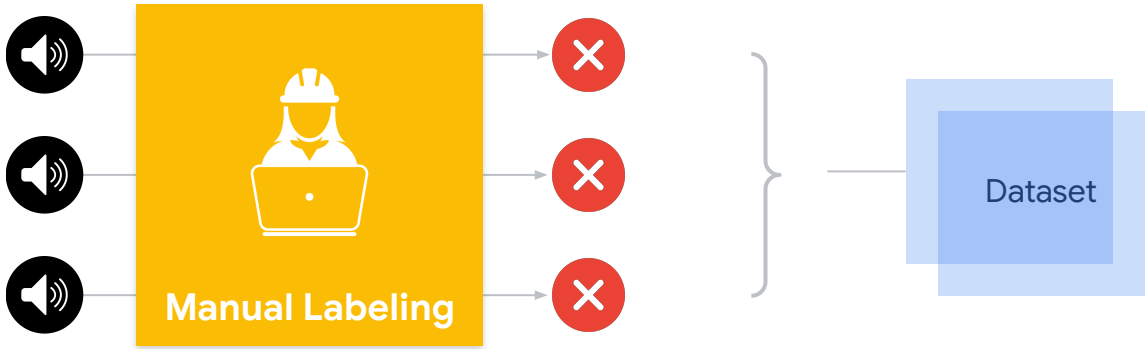


Validating the MSWC Samples

Validating the MSWC Samples

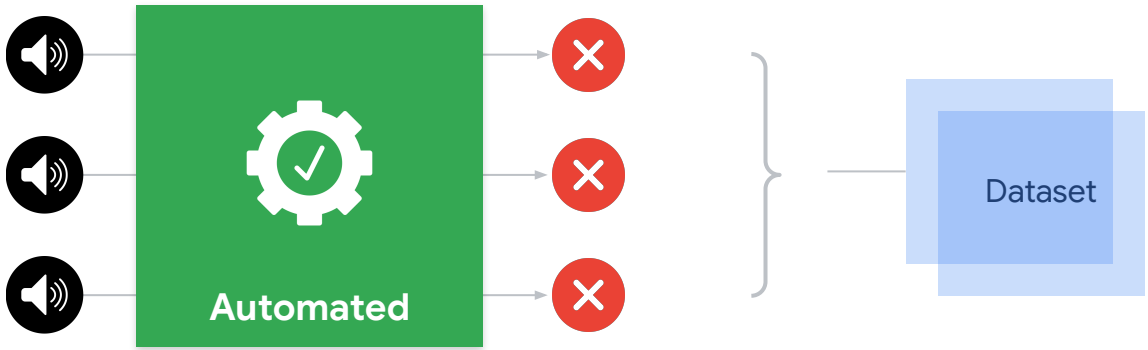


Validating the MSWC Samples



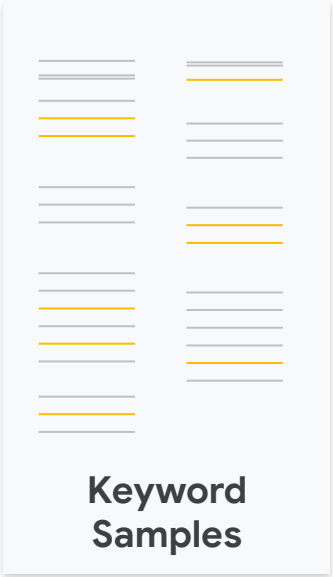
Manual validation is infeasible for all 350,000 keywords and 6,000+ hours of data

Validating the MSWC Samples



Self-supervised **nearest-neighbor** anomaly detection

- Filters **outliers** by distance
- User **tunable** threshold



Self-supervised **nearest-neighbor** anomaly detection

- Filters **outliers** by distance
- User **tunable** threshold

Self-supervised **nearest-neighbor** anomaly detection

- Filters **outliers** by distance
- User **tunable** threshold

Self-supervised **nearest-neighbor** anomaly detection

- Filters **outliers** by distance
- User **tunable** threshold

Self-supervised **nearest-neighbor** anomaly detection

- Filters **outliers** by distance
- User **tunable** threshold

NEAREST 50 CLIPS

CLIP ERROR RATE
2.9%

FARTHEST 50 CLIPS

CLIP ERROR RATE
23.5%

BY DISTANCE TO CLUSTER CENTER

Continuous Validation

In a continuous training pipeline:

- **Same criteria** or new and old data?
  - “Validation drift” can happen as a result of changes in personnel
- Concept drift
  - previously validated samples can become invalid
  - old data should be revisited on occasion