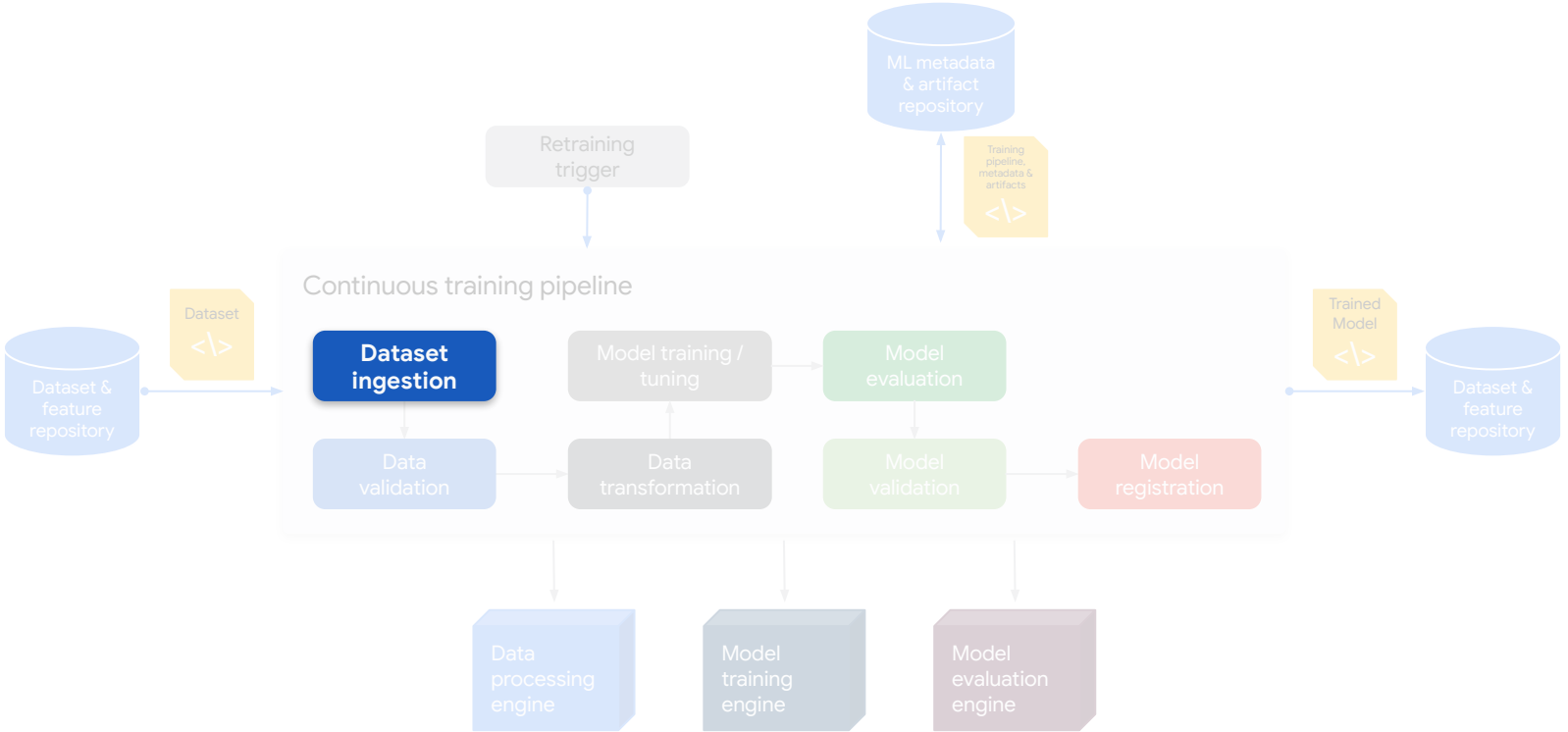
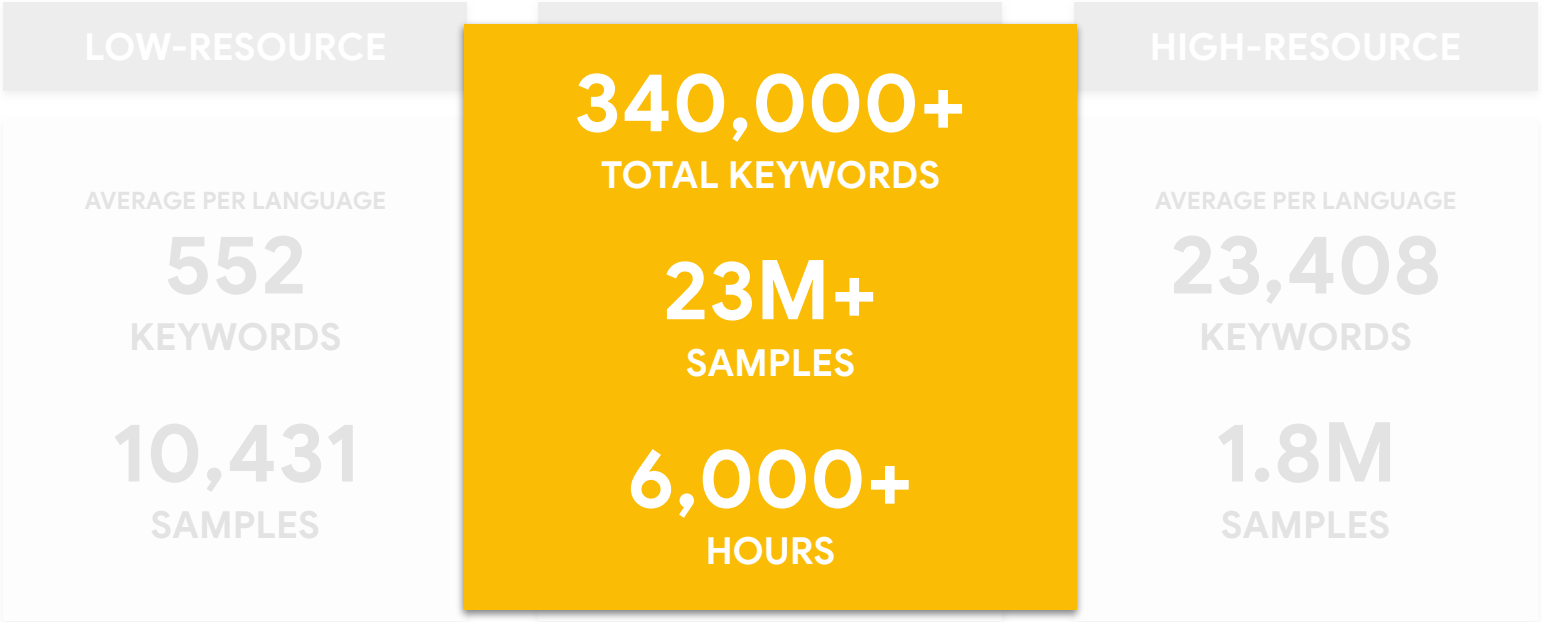
Continuous Training: Data Ingestion

MLOps: Continuous Training

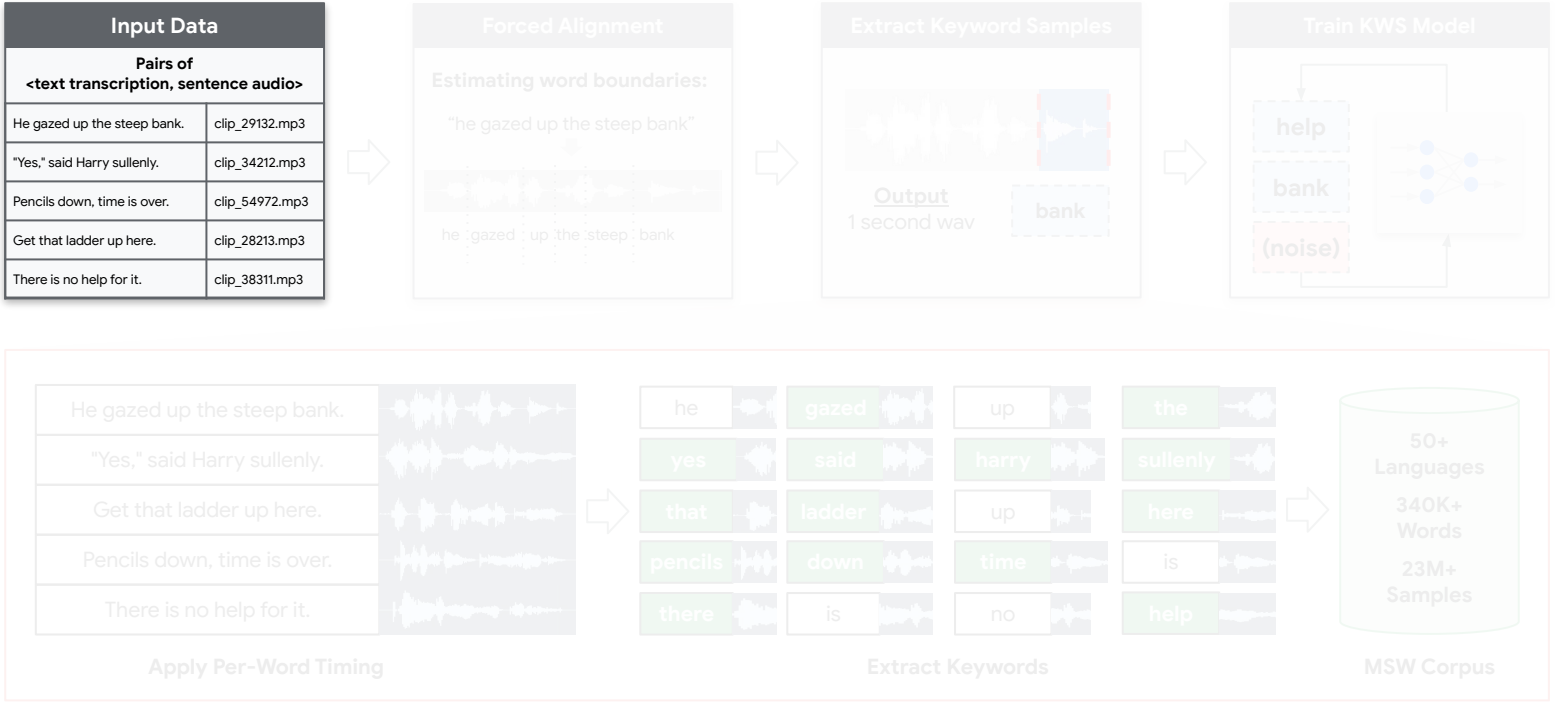


Keywords per language



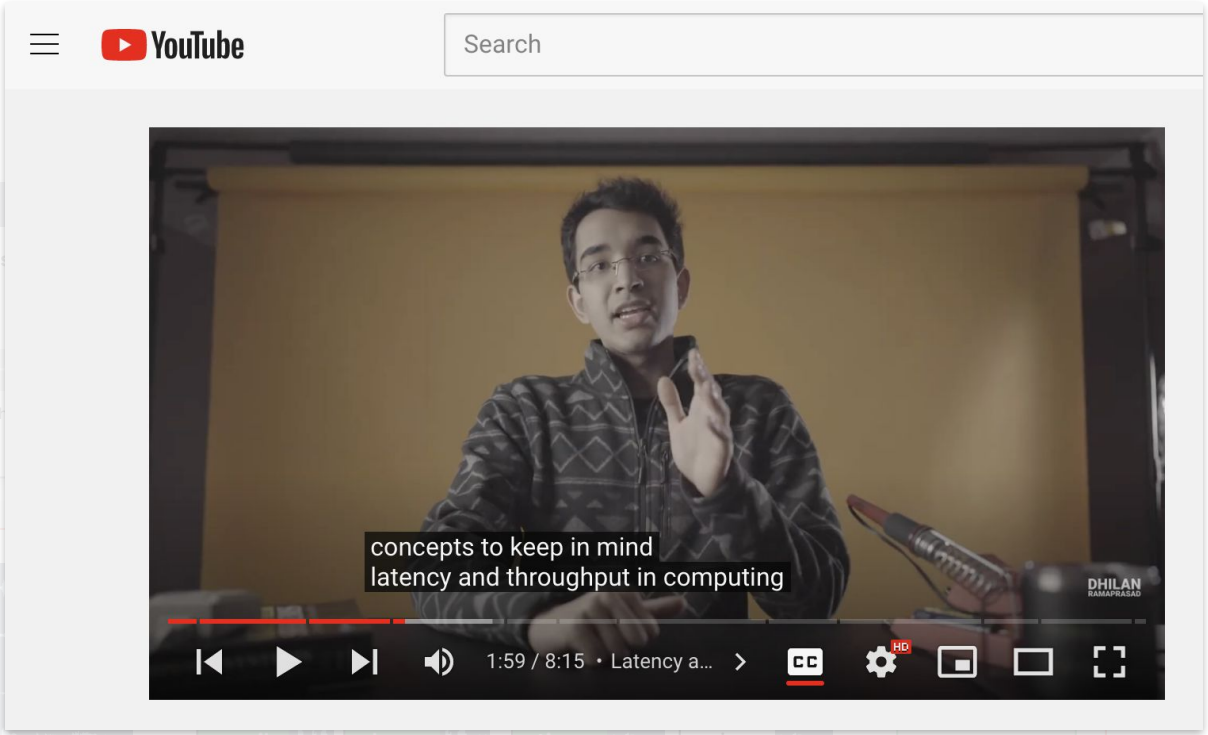
Input Data

Input Data	
Pairs of <text transcription, sentence audio>	
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

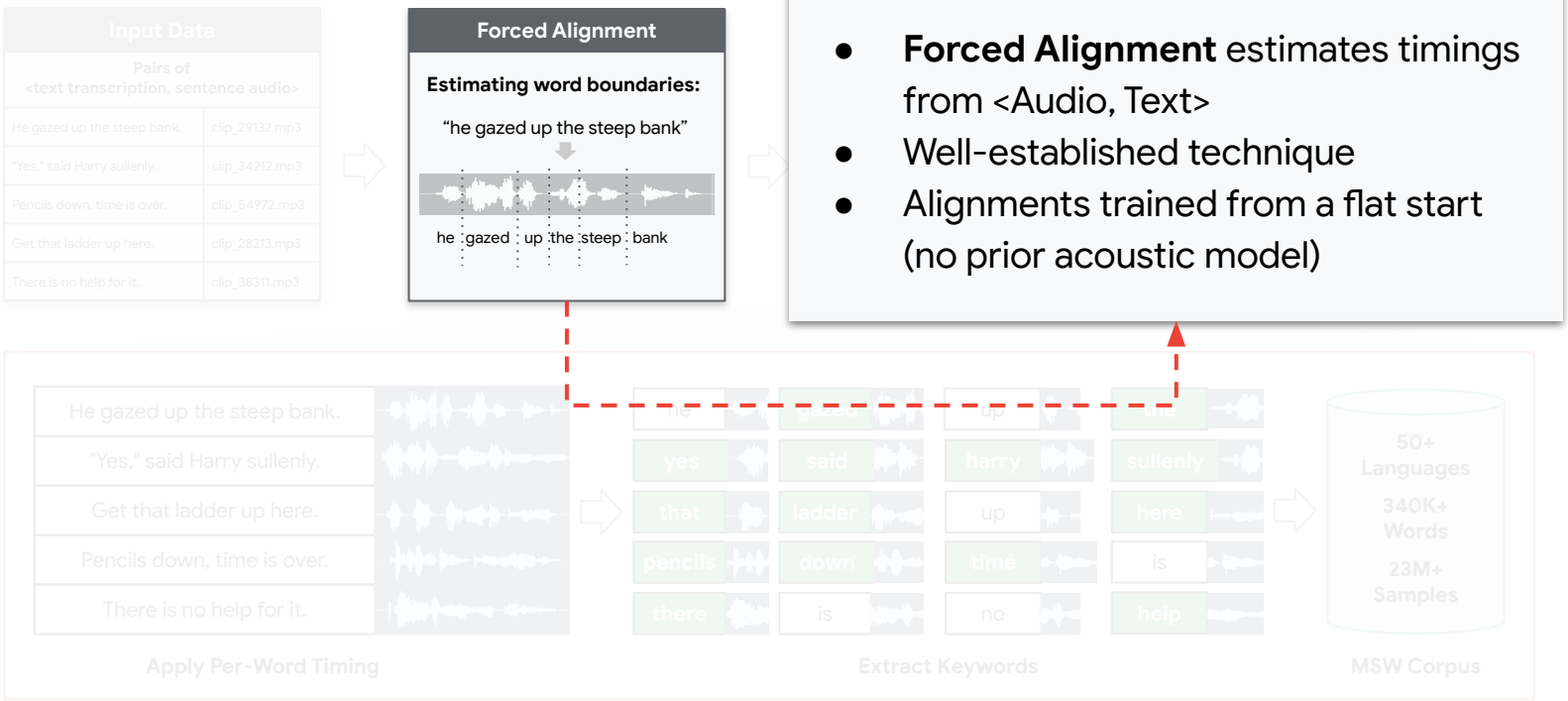


Input Data

Input Data	
Pairs of <text transcription, sentence audio>	
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



Forced Alignment

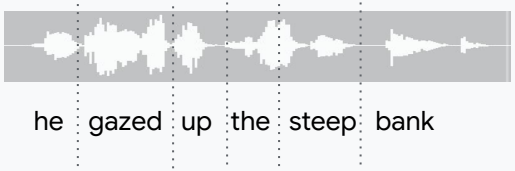


Forced Alignment

Forced Alignment

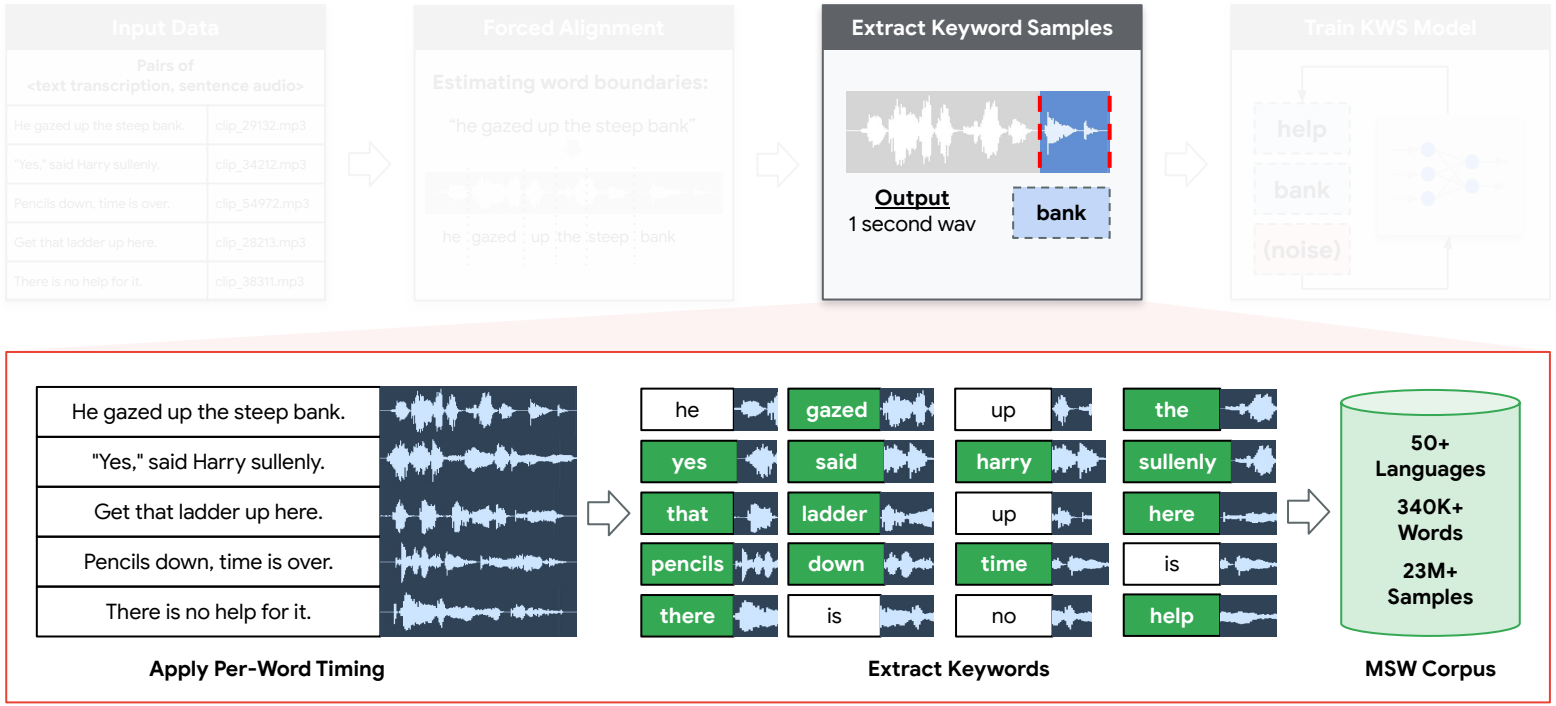
Estimating word boundaries:

"he gazed up the steep bank"



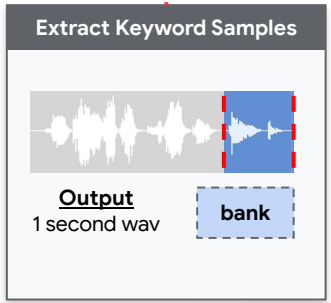
Start Timestamp	End Timestamp	Word
0.000	0.501	he
0.501	1.120	gazed
1.120	1.496	up
...	...	...

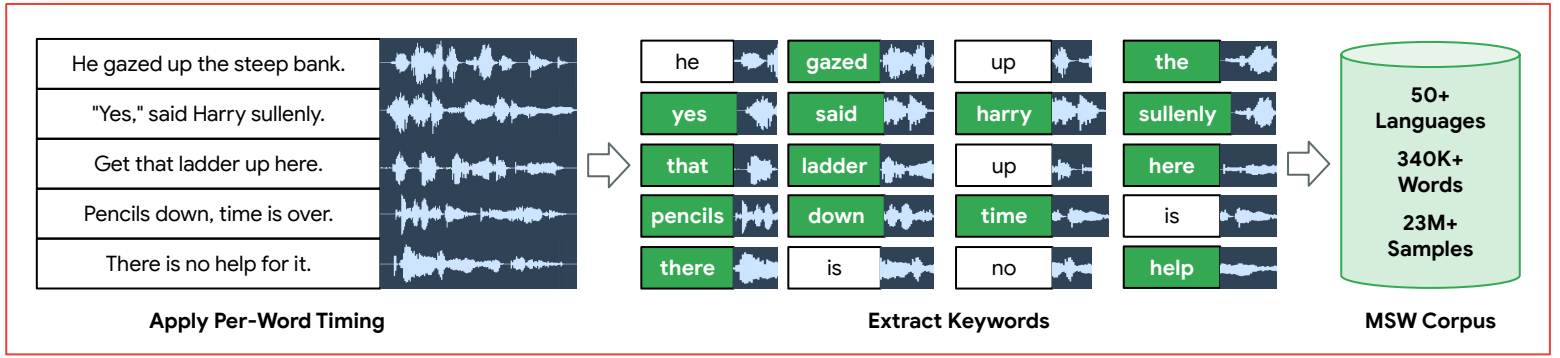
Inclusion Criteria



Inclusion Criteria

- Minimum character length of **three**
  - Coarse stop-word filter
  - More precise forced alignment word boundaries
- Minimum of **five** samples per word





Train the KWS Model

Data Ingestion: **Key Challenges**

1. Tracking sample **provenance** (where each sample was sourced).
    i. Detecting and removing systemic issues
    ii. Identifying high-importance sample (data that leads to largest accuracy gains)

Data Ingestion: **Key Challenges**

1. Tracking sample **provenance** (where each sample was sourced).
    i. Detecting and removing systemic issues
    ii. Identifying high-importance sample (data that leads to largest accuracy gains)
2. Training data **updates**
    a. Add **more data** to undersampled classes or classes that perform worst
    b. Combat **data-drift** (discard non-representative data)
    c. Maintaining **data splits**
        i. Training samples in V1 should not become testing samples in V2
        ii. Errors can poison evaluation metrics