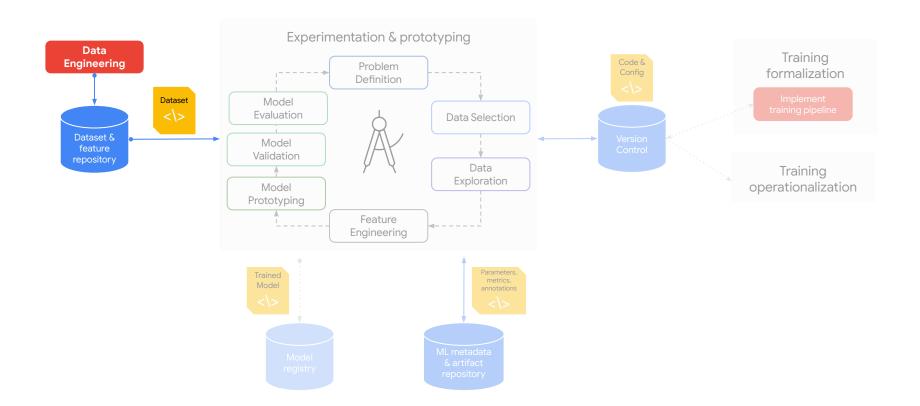
ML Development: Data Engineering



MLOps: ML Development



The MLOps **Personas**



ML Engineer



ML Researcher



Data Scientist



Data Engineer



Software Engineer



DevOps



Business Analyst

Requirements

Gathering

Refinement

Requirements

Gathering

Refinement

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

Requirements

Gathering

Refinement

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

- Data sources
- People
- Collection
- Labeling

Requirements

Gathering

- Data sources
- People
- Collection
- Labeling

Refinement

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

- Processing
- Validation
- Augmentation

Requirements

Gathering

- Processing Validation
- Augmentation

Refinement

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

- Data sources
- People
- Collection
- Labeling

- Storage
- Security
- **Errors**
- Versioning

Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition

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

April 2018

Requirements

Gathering

Refinement

Requirements

Gathering

Refinement

- English KWS dataset
- Permissively licensed for research and use for commercial apps

Requirements

Gathering

Refinement

- English KWS dataset
- Permissively licensed for research and use for commercial apps
- 2,618volunteers
- Variety of accents
- > 1K examples per word

Requirements

Gathering

Refinement

- English KWS dataset
- Permissively licensed for research and use for commercial apps
- 2,618volunteers
 - Variety of accents
- > 1K examples per word

- Remove quiet or low volume recordings
- Manually review through crowdsourcing

Requirements

English KWS dataset

Permissively licensed

for research and use

for commercial apps

volunteers

Gathering

Variety of accents

2.618

• > 1K examples per word

Refinement

- Remove quiet or low volume recordings
- Manually review through crowdsourcing

- V1 had 10 words
- V2 added 25 more words
- Includes train. validation and test dataset splits

