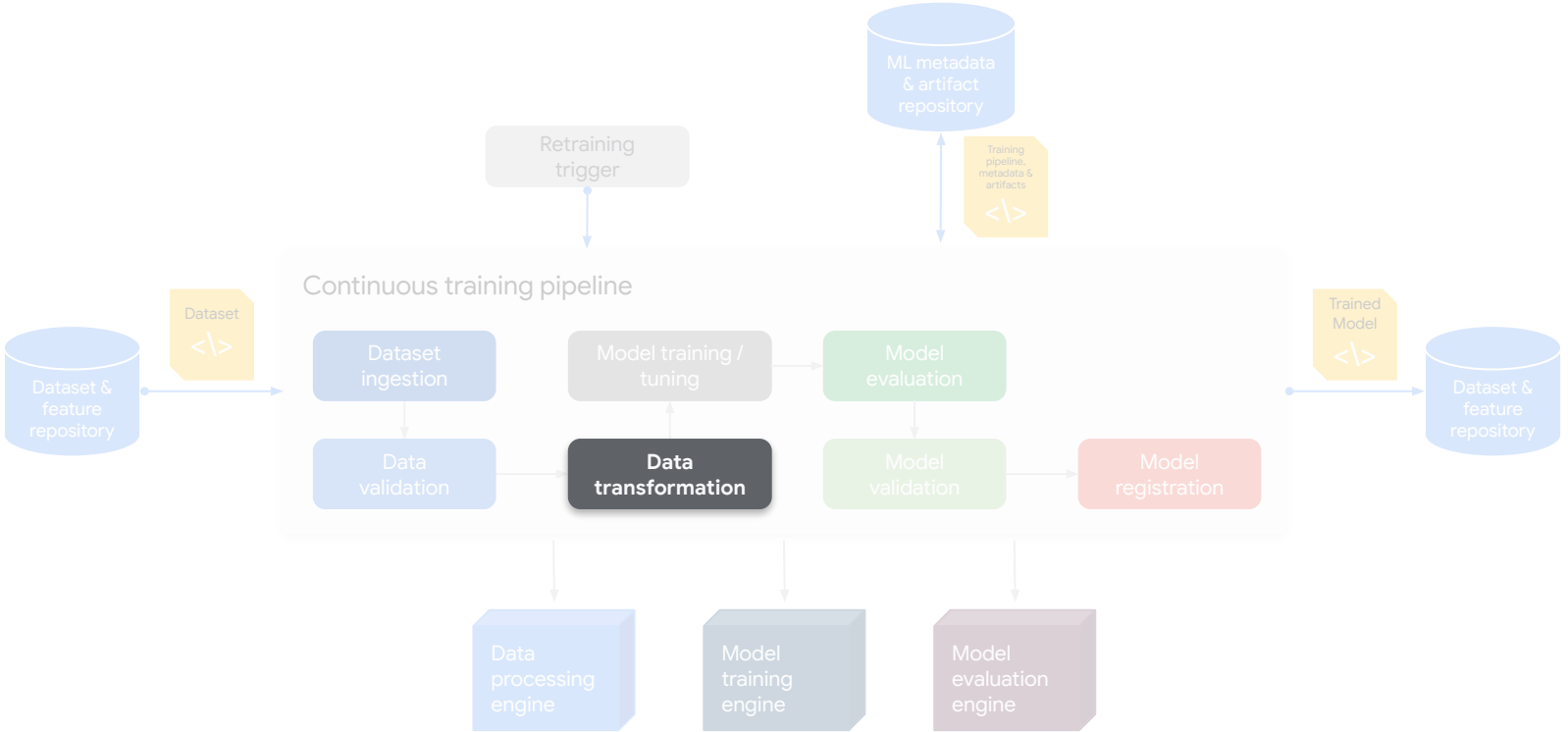
Continuous Training: Data Transformation

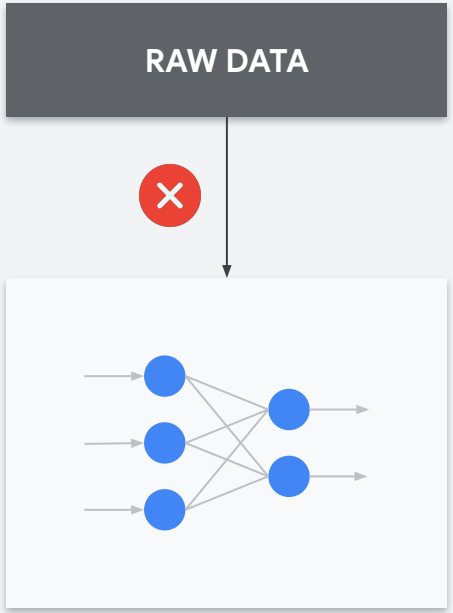
MLOps: Continuous Training



Feature preprocessing

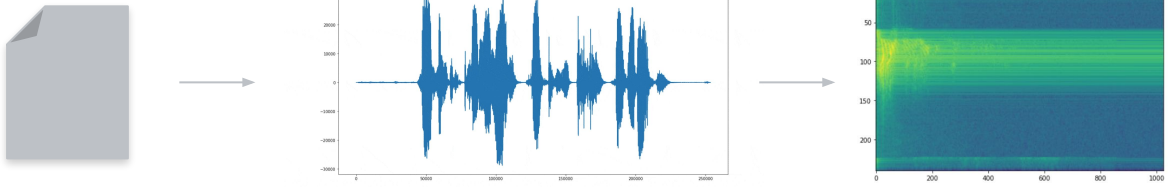
Examples:

- **Stack slices** for 3D convolution in medical imaging
- Extract **Mel filter banks** for speech features
- Extract **ImageNet features** via a pretrained network (VGG16, ResNet) for fine tuning or few-shot learning



Models often **can't ingest** raw training data formats

MSWC Data Transformation



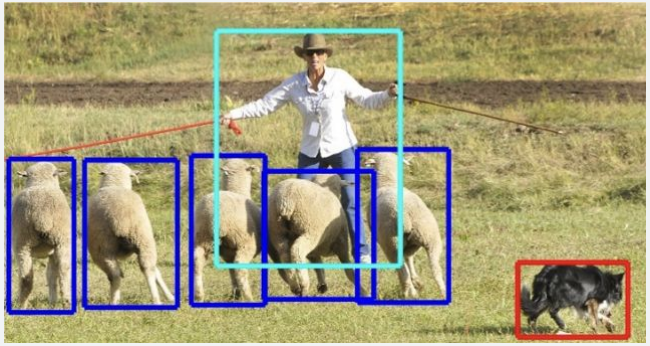
Compressed .opus audio file format

1-second 16KHz single-channel audio waveform

49x40 TensorFlow Lite Micro Frontend Spectrogram

Label Transformation

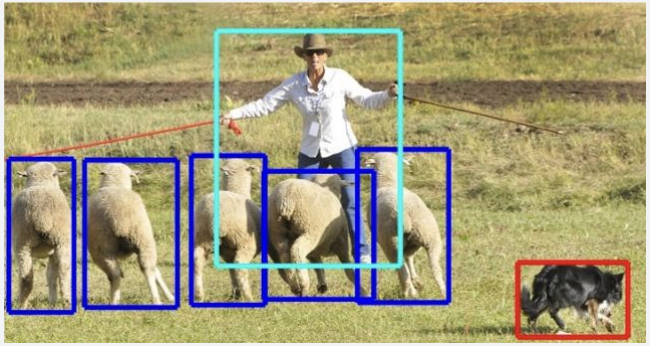
- Training data can come from multiple sources, datasets, and labeling tools
  - Ensure all samples use the **same format**

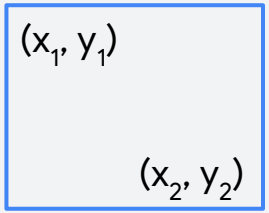


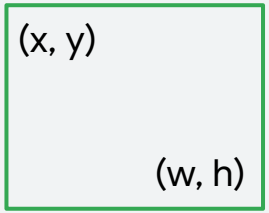
Microsoft COCO: Common Objects in Context

Label Transformation

- Training data can come from multiple sources, datasets, and labeling tools
  - Ensure all samples use the **same format**
- Bounding box format examples:
  - (x_1, y_1) and (x_2, y_2) pairs
  - (x, y) corner and (w,h)
  - (x, y) center and (w,h)
- Pick one and make sure **all sources conform**



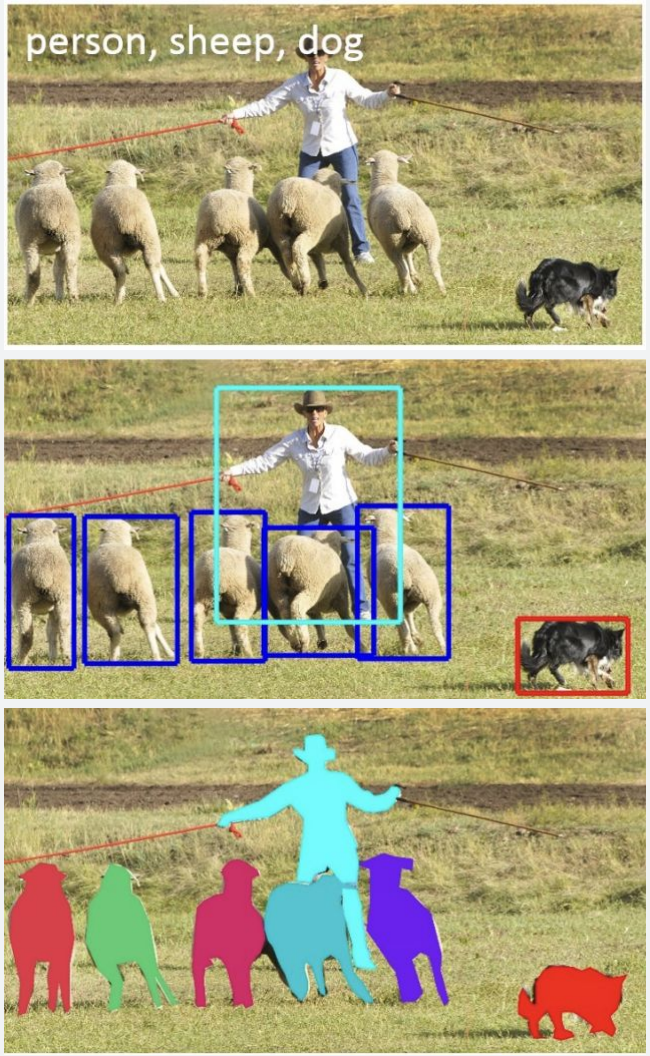




Microsoft COCO: Common Objects in Context

Adapting data for model expressivity

- Data transformation might involve targeting new formats
  - e.g., converting from bounding boxes to **segmentation maps**



Microsoft COCO: Common Objects in Context

Continuous Training Considerations

- Repurpose existing data for **new** training regimes
  - e.g., **converting** categorical ImageNet data to **contrastive pairs** for self-supervised learning