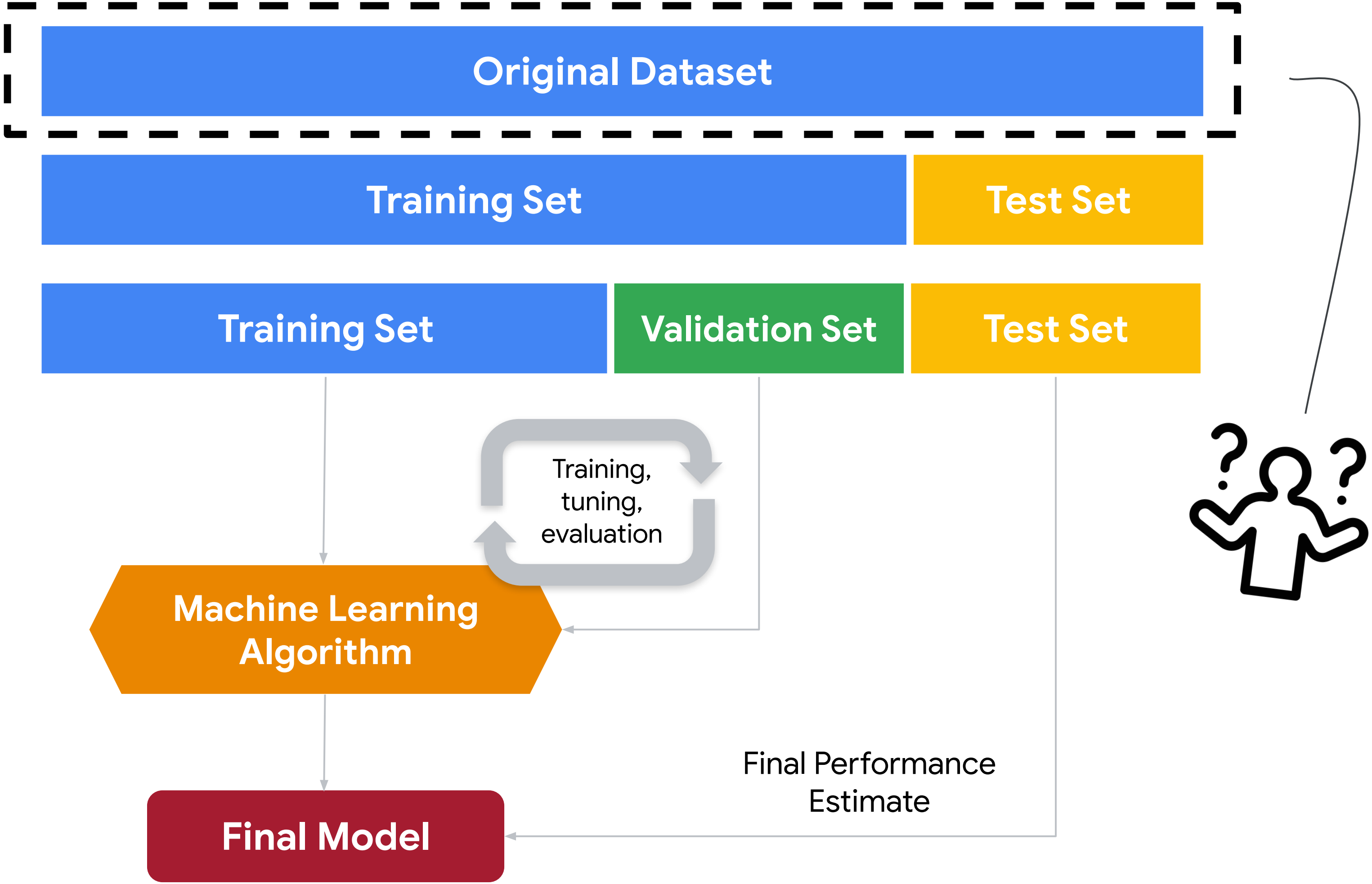
What is Data Engineering?

A supervised AI is trained on a corpus of training data.


Original Dataset
Training Set
Test Set
Training Set
Validation Set
Test Set
Training,
tuning,
evaluation
Machine Learning Algorithm
Final Model
Final Performance Estimate

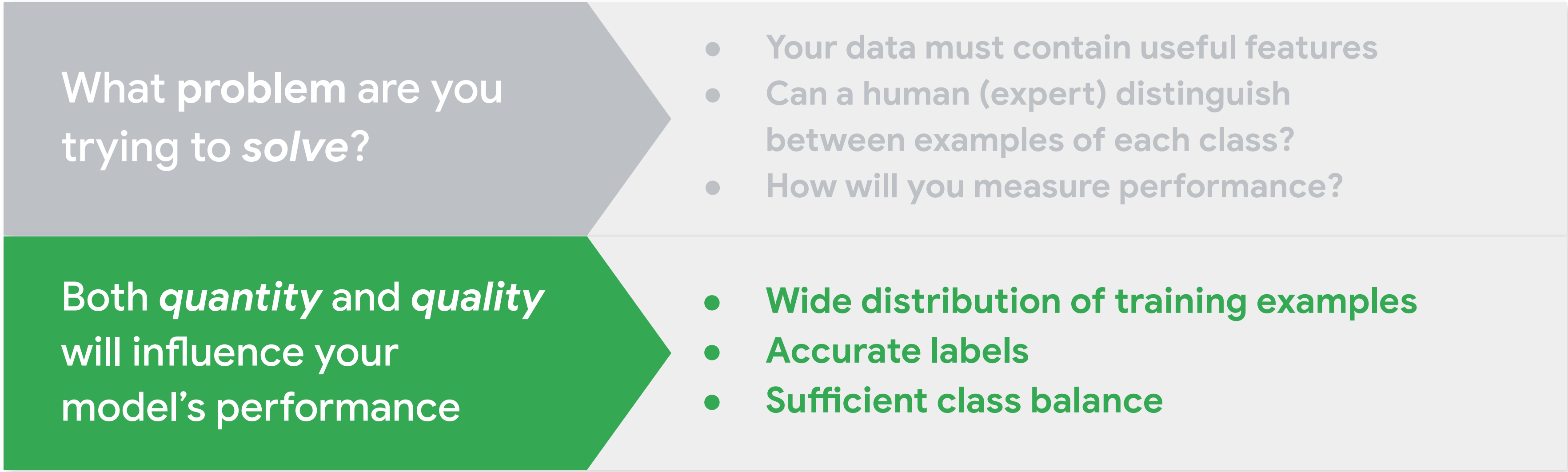
Data Engineering is all about **datasets**

Good Data is Necessary for Accuracy

What problem are you trying to *solve*?

- Your data must contain useful features
- Can a human (expert) distinguish between examples of each class?
- How will you measure performance?

Good Data is Necessary for Accuracy



What **problem** are you trying to *solve*?

- Your data must contain useful features
- Can a human (expert) distinguish between examples of each class?
- How will you measure performance?

Both *quantity* and *quality* will influence your model's performance

- Wide distribution of training examples
- Accurate labels
- Sufficient class balance

Data Engineering

Requirements

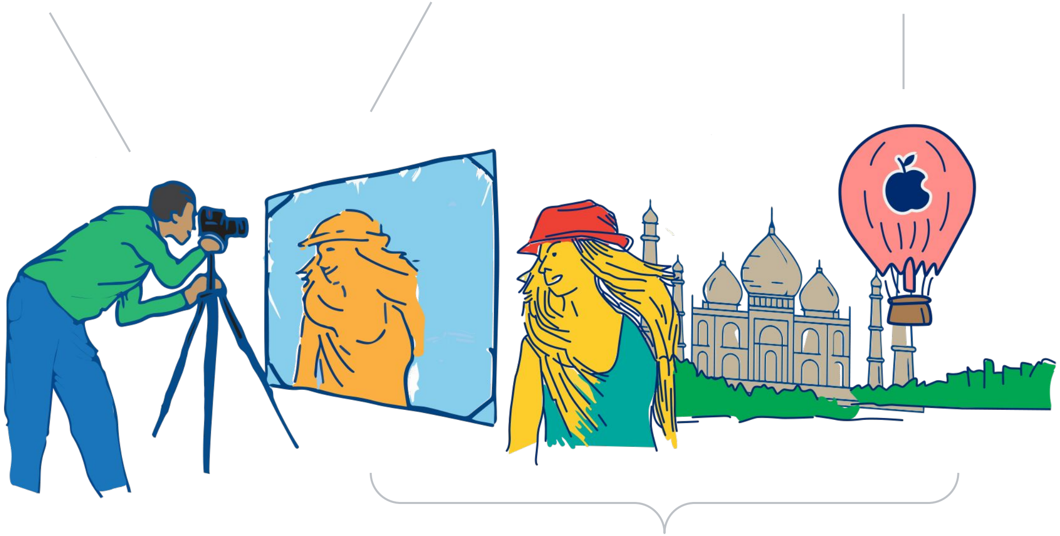
- Problem definition
- Machine & human usable format
- **Permissions & rights**

Data **isn't free** to use

Where does your data **originate**?

- Open?
- Copyrighted?
- Licensed?
- Product users?

What’s Yours and What’s **Not** Yours



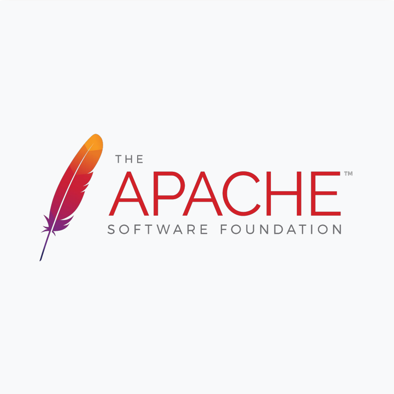
Author / Owner

Creative Work

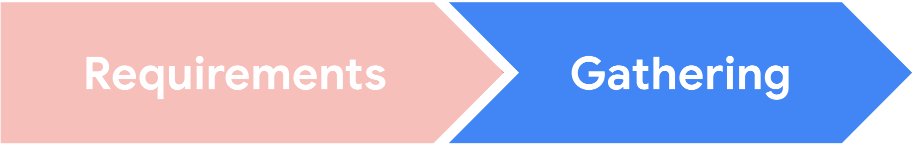
Trademarked Logo

Copyrighted

Licenses



Data Engineering



Requirements

Gathering

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

- People
- Collection
- Labeling
- **Data sources**

Data **sources**

- Sensors
- Crowdsourcing
- Product users
- Paid contributors

Data Engineering

Requirements

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

Gathering

- Data sources
- People
- Collection
- Labeling

Refinement

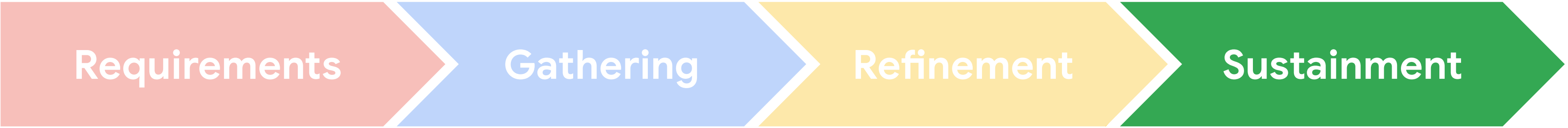
- Processing
- Augmentation
- **Validation**

Some data is ***unusable***

How will you **verify** the data you collected?

- **Manually** (time, cost)
- **Automation**
- Domain expertise
    - disputes / disagreements

Data Engineering



Requirements

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

Gathering

- Data sources
- People
- Collection
- Labeling

Refinement

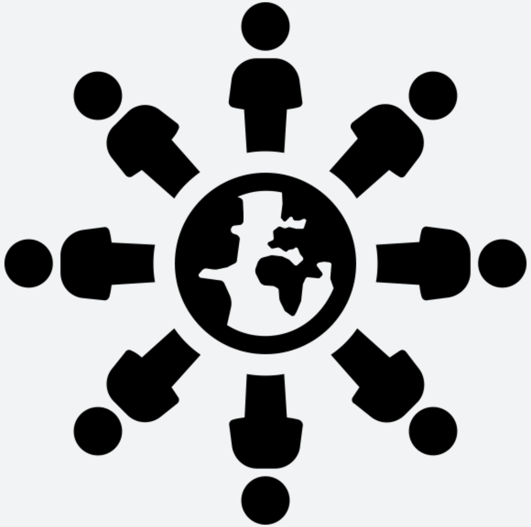
- Processing
- Validation
- Augmentation

Sustainment

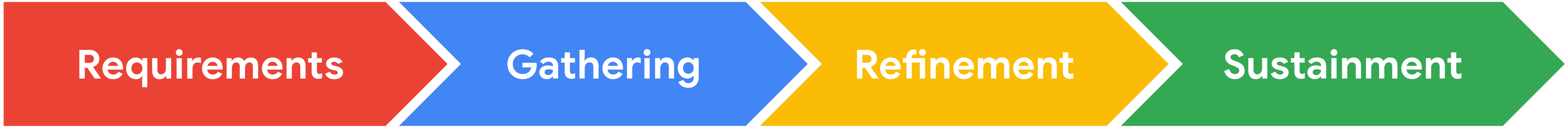
- Storage
- Security
- Errors
- **Versioning**

Your dataset will ***evolve***

- Missing **demographics**?
- **Expanding** your user-base?



Data Engineering



Requirements

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

Gathering

- Data sources
- People
- Collection
- Labeling

Refinement

- Processing
- Validation
- Augmentation

Sustainment

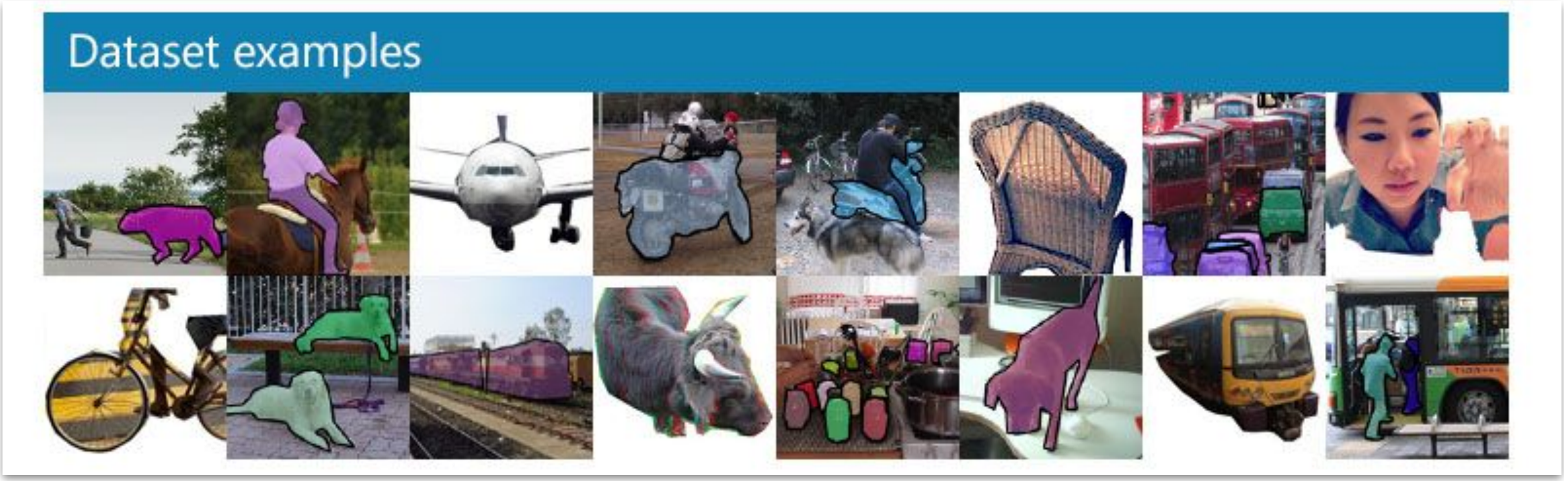
- Storage
- Security
- Errors
- Versioning

Datasets require **significant effort**

Classifying Images

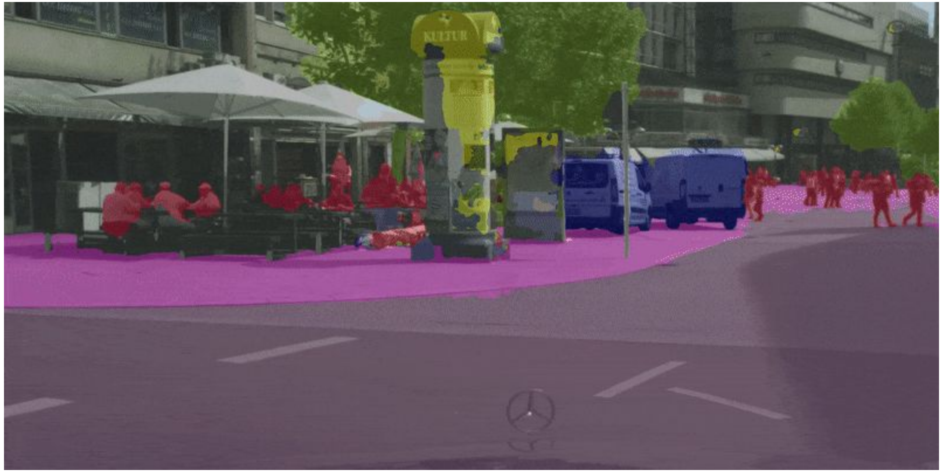
Detecting Objects

- Common Objects in Context (**COCO**)—**2.5M+** segmented images



Datasets require ***significant effort***

- **Waymo**—**1,950** 20-second driving segments (cameras, LIDAR, labels)
- **KITTI 360**—**73KM+** of annotated driving data



Datasets require ***significant effort***

These **massive** machine learning datasets are ***constructed by hand***

- **Common Voice**—**5000+** hours of spoken audio
- Common Objects in Context (**COCO**)—**2.5M+** labeled images
- **ImageNet**—**4M+** labeled images
- **Waymo—1,950** 20-second driving segments
- **KITTI 360**—**73KM+** of annotated driving data

Data Engineering: *How to build your own dataset?*

How do you build your own datasets for ***TinyML***?