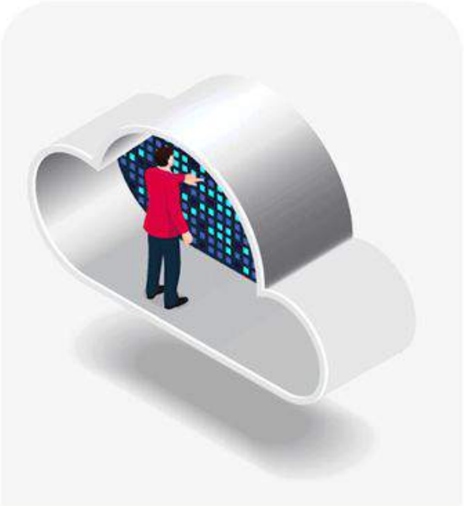
05 Develop a Data-Centric Perspective

Topic Outline

- Data Collection Techniques
  - Acquisition, Annotation and Improvement of Existing Data
  - Importance of Skilful Data Labeller
- Data Annotation Techniques for Images
  - Bounding Box, Segmentation
  - Sample Deep Learning Application
  - General Guidelines on Best Practices for Labelling
- Data Fallacies
- Quality Assurance
- Adoption of AI into business

Data Collection Techniques

Data Acquisition	Data Annotation	Data Enhancement
		
Collecting dataset for training machine learning models.	**Labelling the collected dataset** to determine the tags and position of objects inside an image	Techniques on **tuning the existing dataset** when acquiring and labelling new data is not the best option

Data Acquisition

- To collect datasets that can be used to train AI algorithms
- 3 approaches:

Task	Approach	Explanation
Data discovery	Sharing	Focus on collaborative analysis or publishing on the web, or both.
	Searching	Mainly designed for searching dataset.
Data augmentation		Introduce variance into the existing data
Data generation	Crowdsourcing	Employing crowd workers to accomplish tasks that cannot be automated.
	Synthetic Data	Generate synthetic data

Data Annotation

- Task for labelling the dataset using selected tags based on the use case.
- In most times, labelling is usually done along with data acquisition.

Category	Approach	Explanation
Use existing labels	Self-labelled	Generate more labels by trusting one's own predictions.
Crowd-based	Crowdsourcing	Labelling task done by workers who are not necessarily labelling experts

Data Enhancement

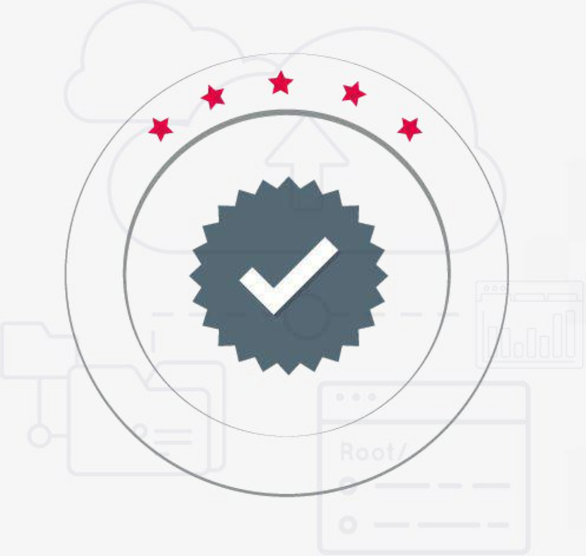
- These techniques are used when the developed model does not provide good prediction and when acquiring new labelled data is not a feasible option

Task	Techniques	Explanation
Improve Data	Data Cleaning	Removing the noise within the dataset either manually or automatically.
	Re-labelling	Manual inspection and provide new labels where necessary

Importance of Skilful Data Labeller

To aid in the regulation of data quality

To accelerate the development of machine learning models



To build the competency of human employees for big data-specific work domain

Data Annotation Techniques for Images

Lets start with images

Grey images

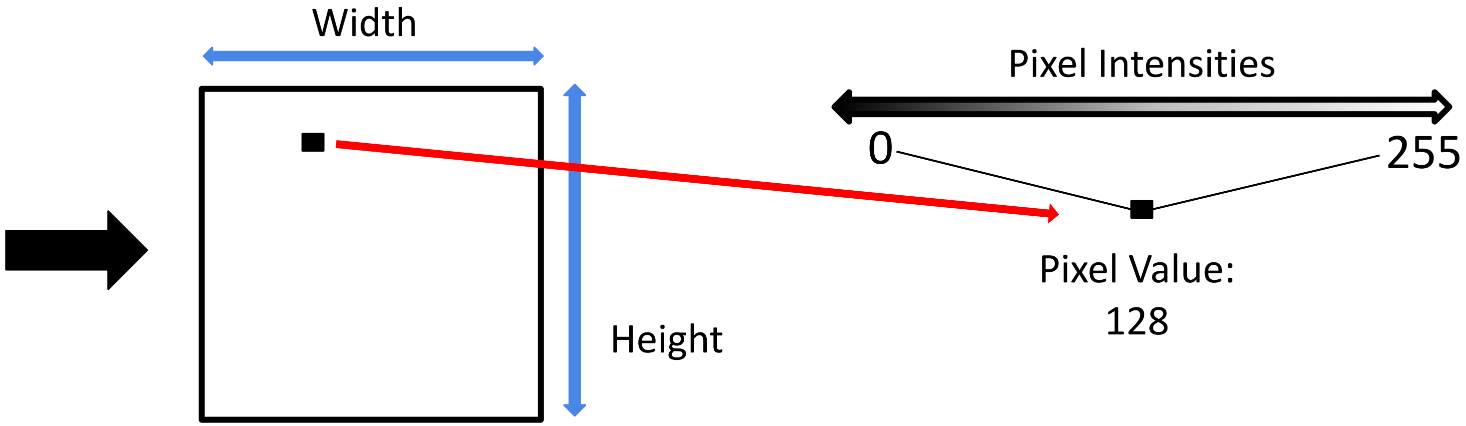
Colour images

Prerequisite

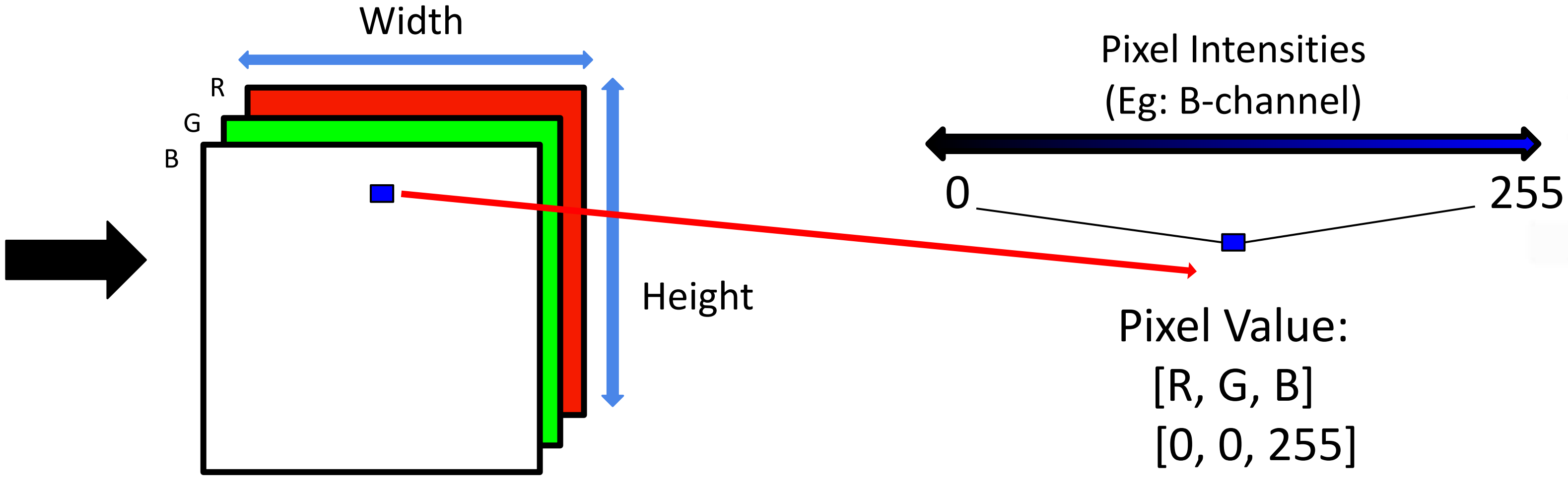
Pixel is a physical point in an image

For an 8 bit image, the pixel depth allow 256 intensities (2^8) for a channel

Grey image

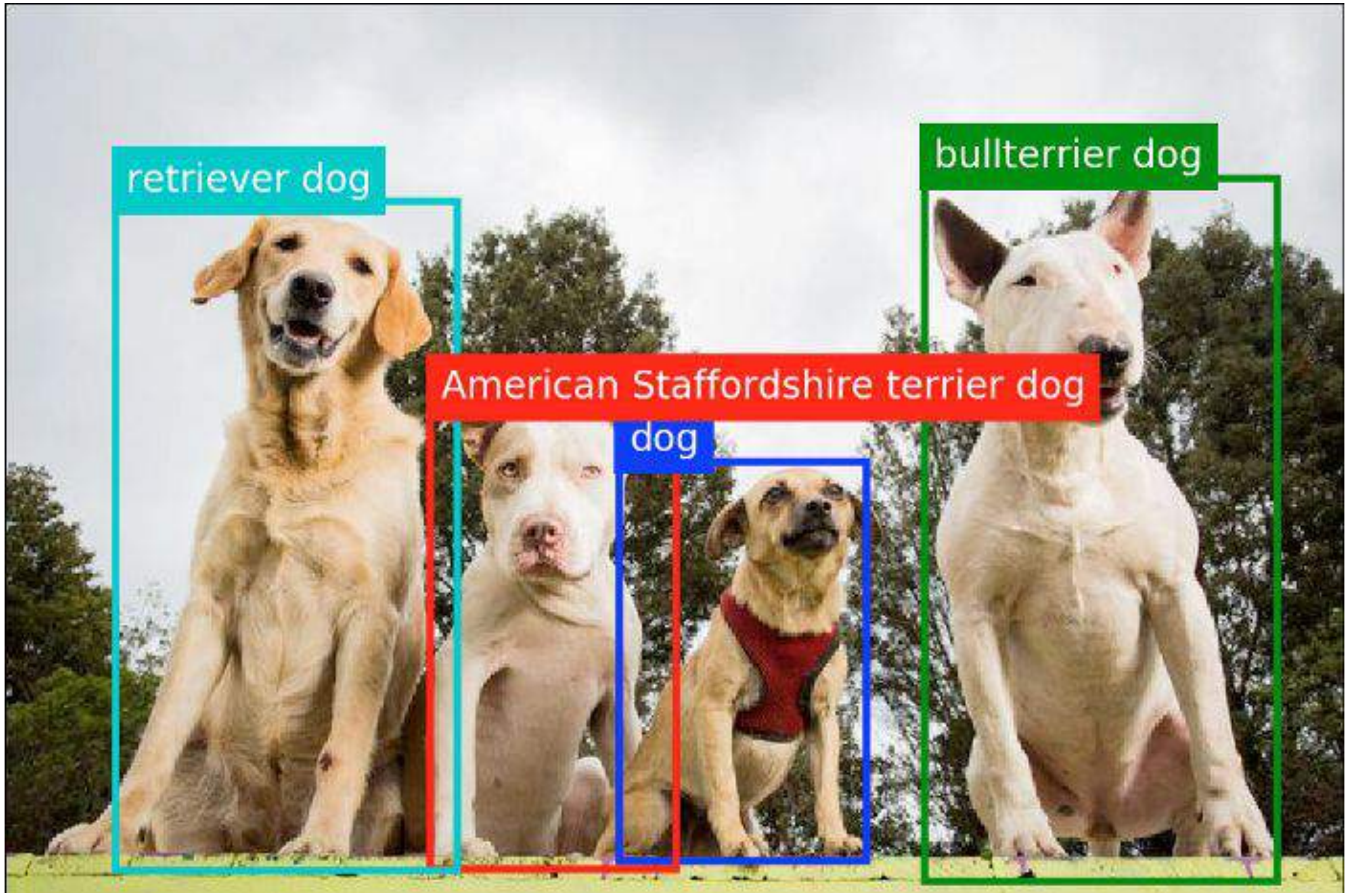


Colour image



Prerequisite

The goal of labelling is to detect the item and location of the object of interest

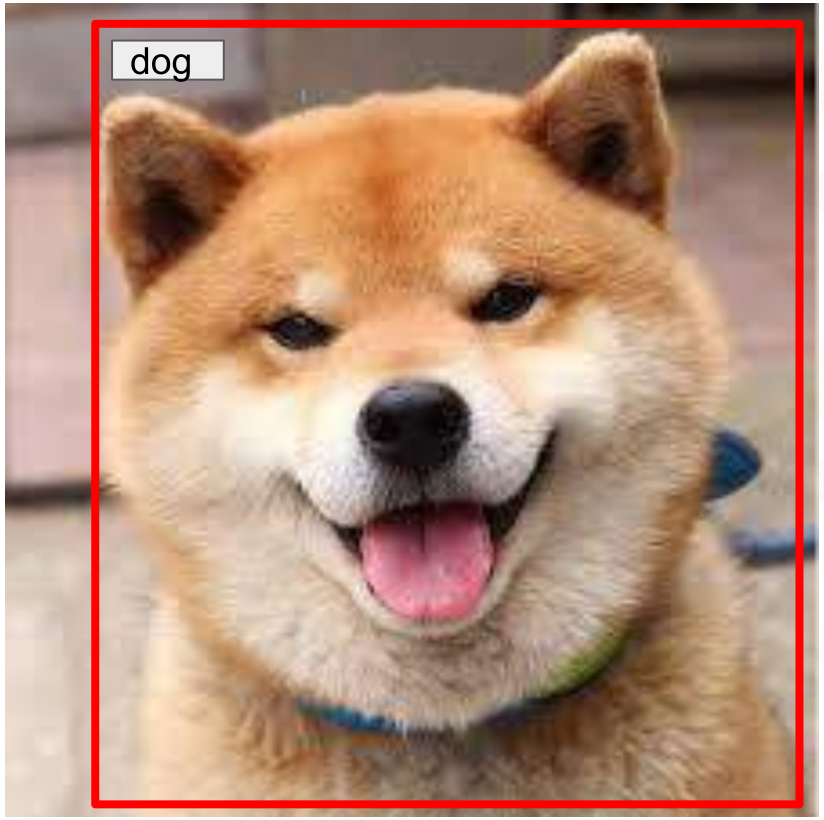


Reference: https://medium.com/ibm-watson/dont-miss-your-target-object-detection-with-tensorflow-and-watson-488e24226ef3

Overview of Labelling Techniques for Images

Bounding Box

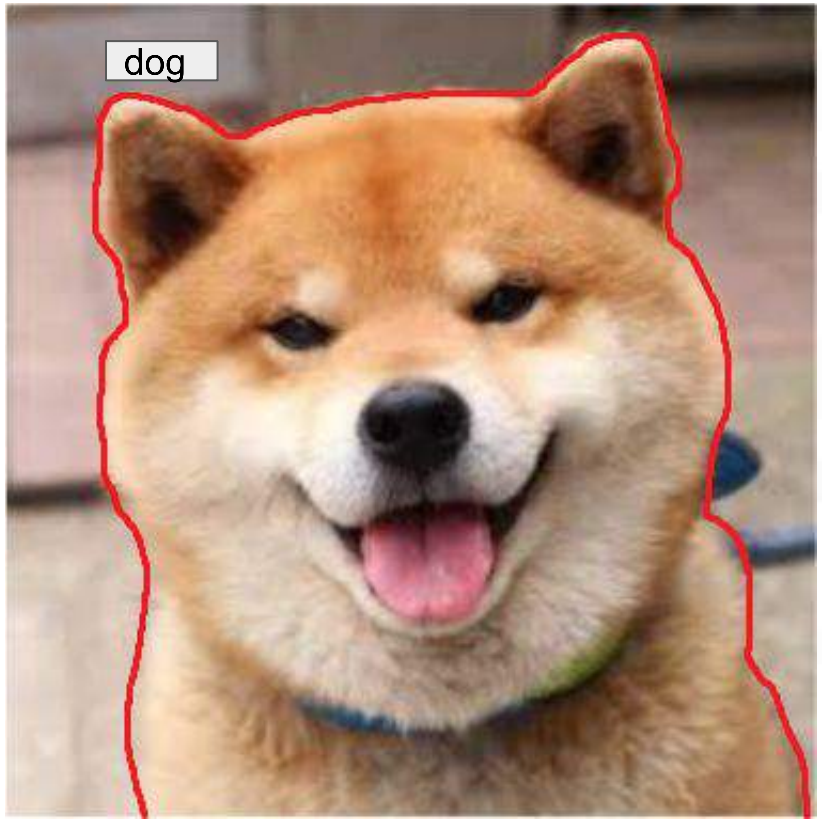
Rectangular border around the object of interest



Overview of Labelling Techniques for Images

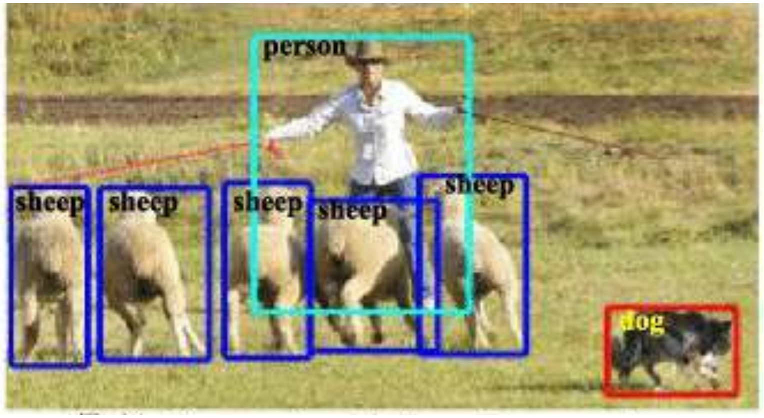
Irregular Shapes

Pixel based identification providing exact outline of the object of interest

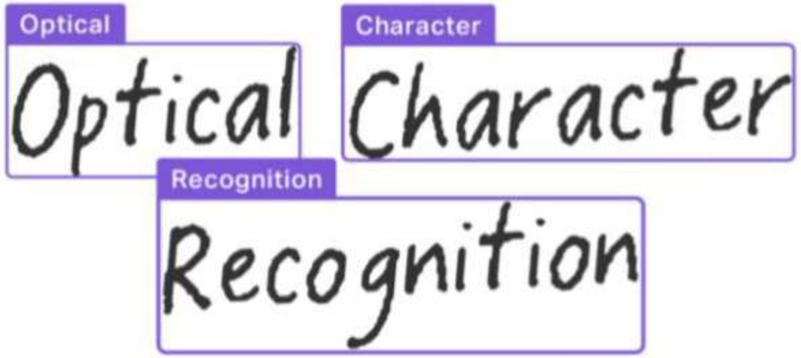


Sample Deep Learning Application

Bounding Box



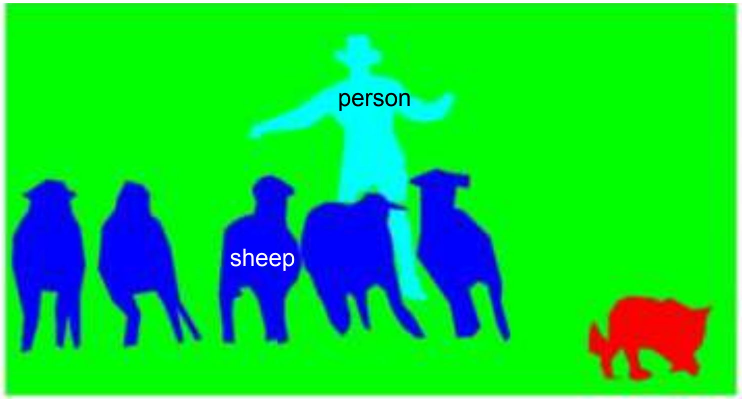
Object Detection



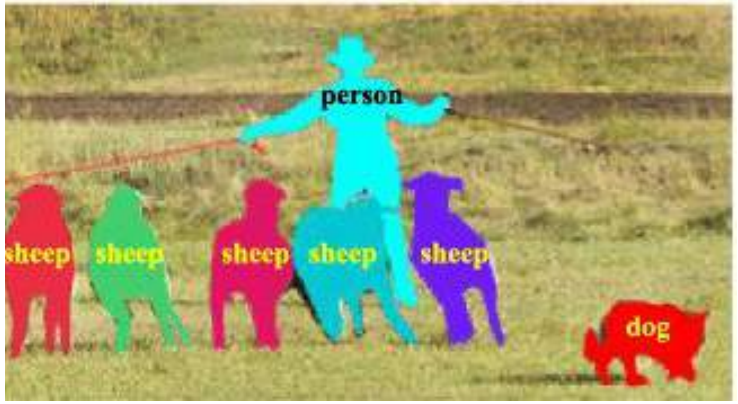
Text Recognition

Sample Deep Learning Application

Irregular Shapes



Semantic Segmentation



Instance Segmentation

Source: https://www.researchgate.net/figure/Recognition-problems-related-to-generic-object-detection-a-image-level-object_fig5_327550187

General Guidelines on Best Labelling Practices

To create a high quality dataset, labelling images carefully and accurately is important. General guidelines are as follows:-

1. **Label the full object**
   a. The bounding shape should enclose the object completely
2. **Every object of interest in the image should be labelled**
   a. Avoid false negatives in our model
3. **Detailed label names**
   a. Specific label names can always be combined but general labels cannot be discretized without relabelling
   b. Eg: faber_pen, parker_pen ✓ pen ✗
4. **Tight boundaries**
   a. Help the model to identify the regions of interest precisely.
   b. Careful not to cut off a portion of the object during labelling
5. **Label occluded objects**
   a. Use imagination to draw boundaries on the full object, including the parts that are not visible

Data Fallacies to Avoid in Supervised Learning

Data **Fallacy**

X **FALLACY: The more the data, the better the results in any given scenario**

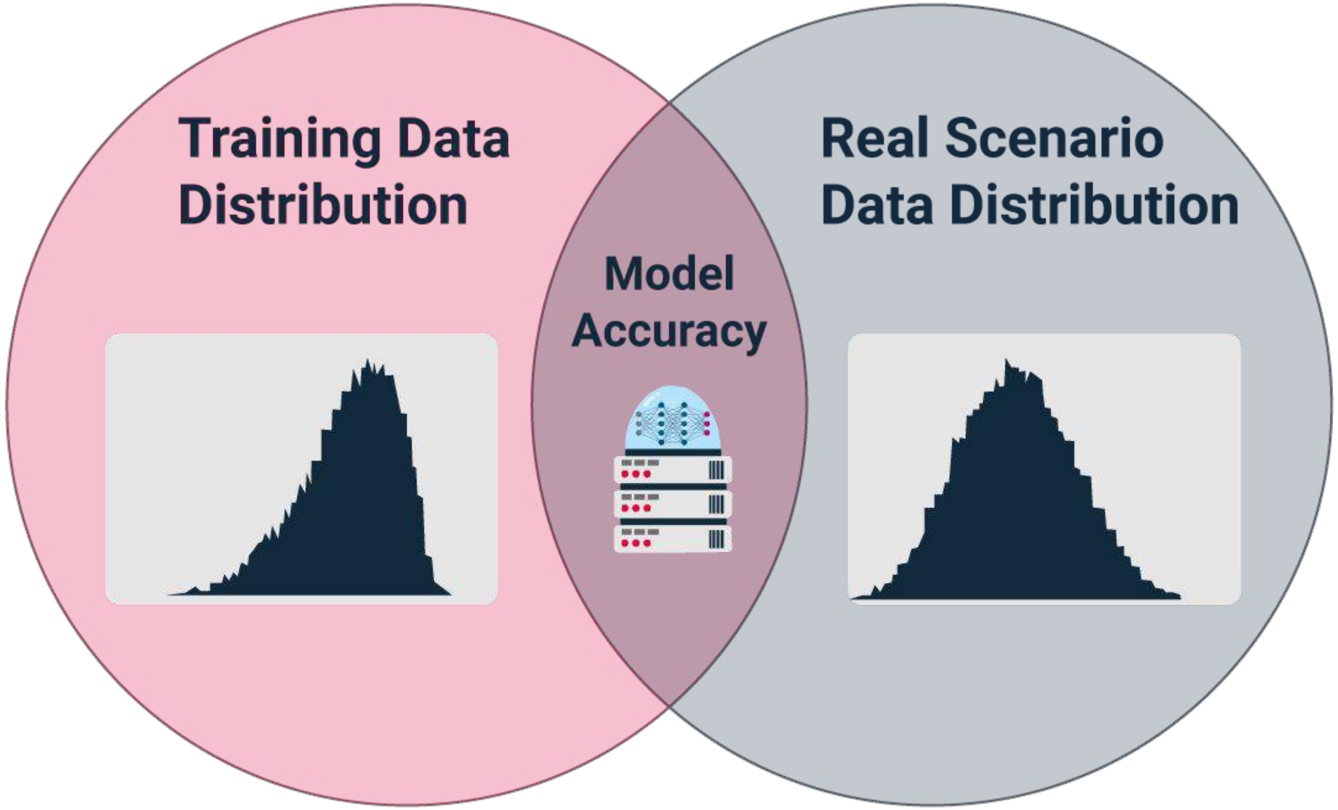
More data will generally improve model accuracy. Yet this might not always be the case.

For supervised learning, model accuracy depends on

- **Relevance of dataset**
- **Encapsulation of intra-class data variation**
- **Quality of labels**
- **Other factors**

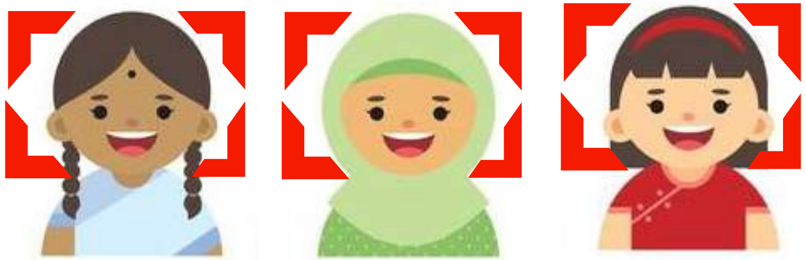
Relevance of Dataset

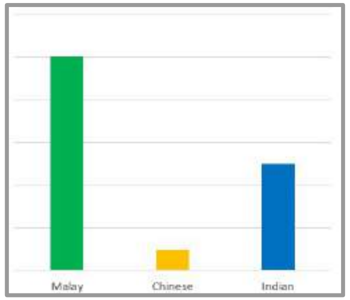
- **The distribution of training data has to be close to the real prediction environment.**
    - In other words, the training data have to encapsulate the gist of the problem statement.



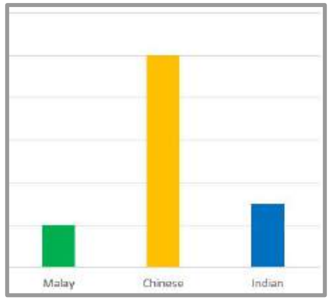
Relevance of **Dataset**

Use Case: Identification of races (Malay, Chinese, Indian) through face detection

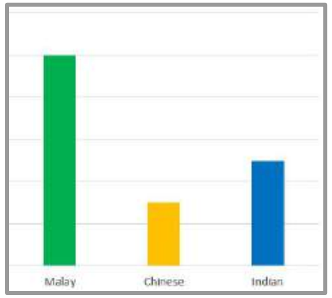




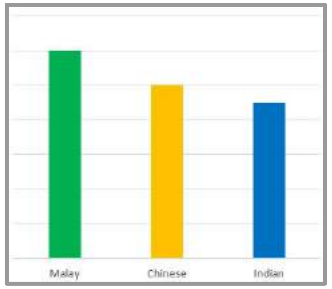
Scenario 1



Scenario 2



Scenario 3



Scenario 4

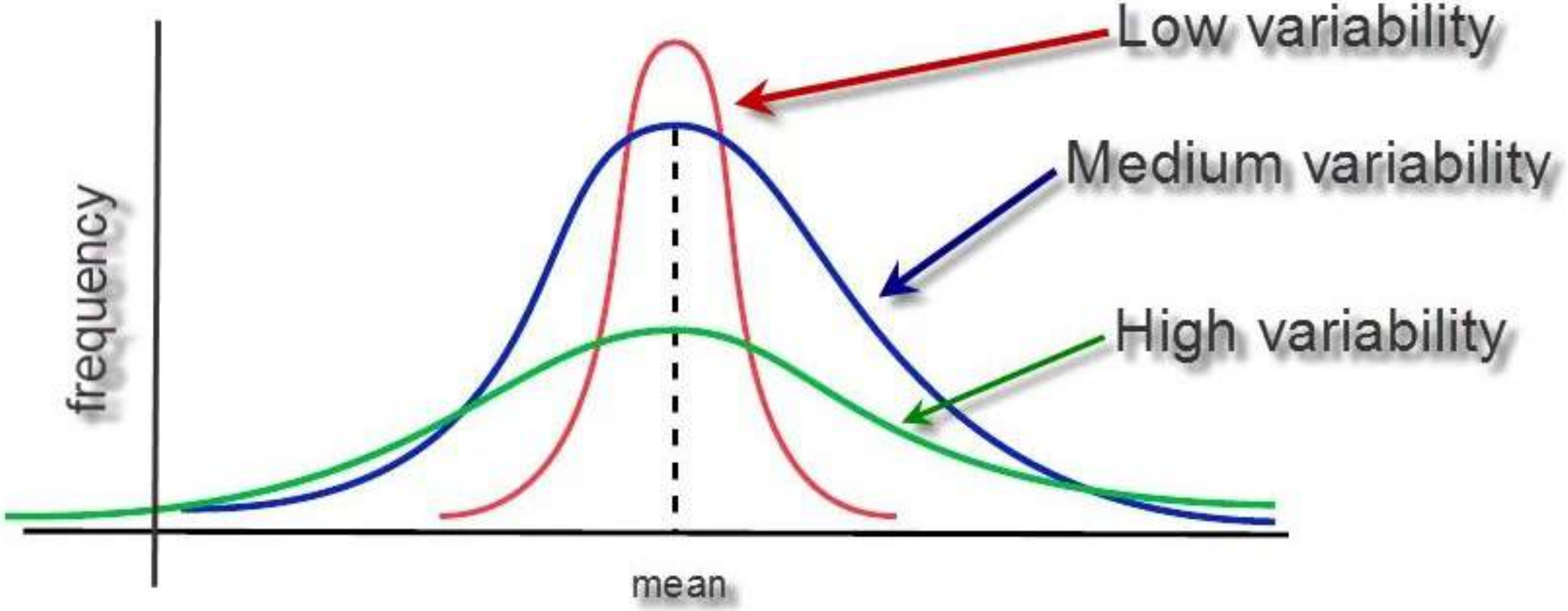
Bad baseline for modelling

Ideal baseline for modelling

Encapsulation of Intra-Class Data Variation

- **The dataset should capture variation of forms within a class.**
  - Variation describes how widely data are spread out from the center of a distribution
  - Intuitively, data of the same class inherits certain variability.

The taller curve has less dispersion
The flatter curve has more dispersion



Encapsulation of Intra-Class Data Variation

Use Case: Dog and Cat Classification

Dog class

Cat class

Test Input

Which class?

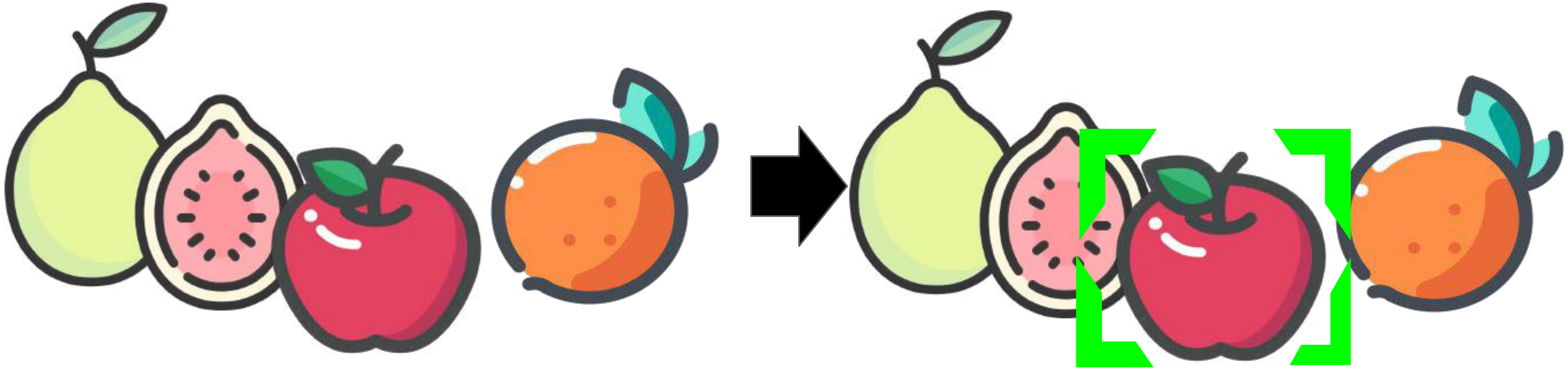
Cat class

Quality of Labels

- **Computers do not intuitively understand problem statements**
  - The quality of the ROI (Eg - bounding boxes) supplied during training has a high impact on a model's ability to **detect** and **recognise** objects.

- **Definition of quality in this context:**
  How well does the ROI capture the object of interest.

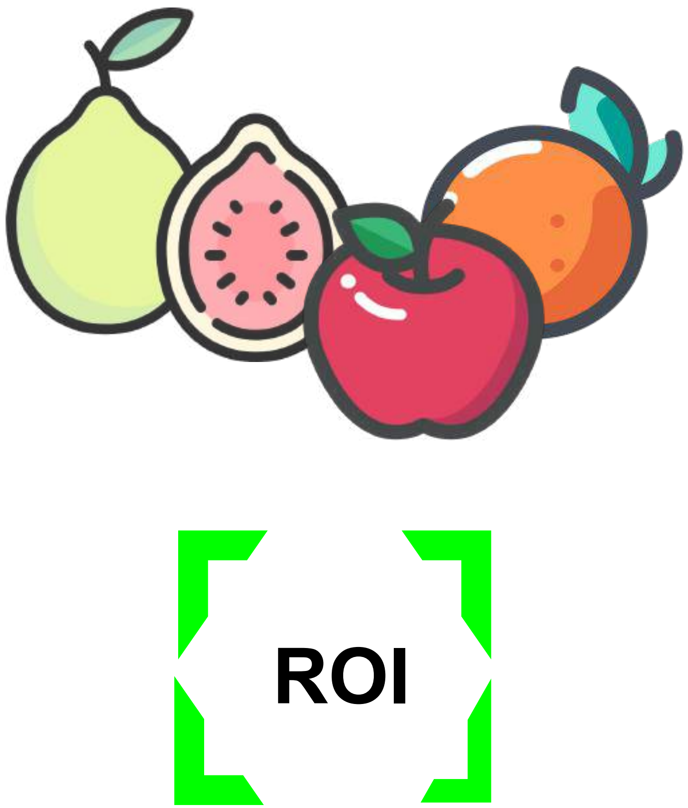
AI in Agriculture

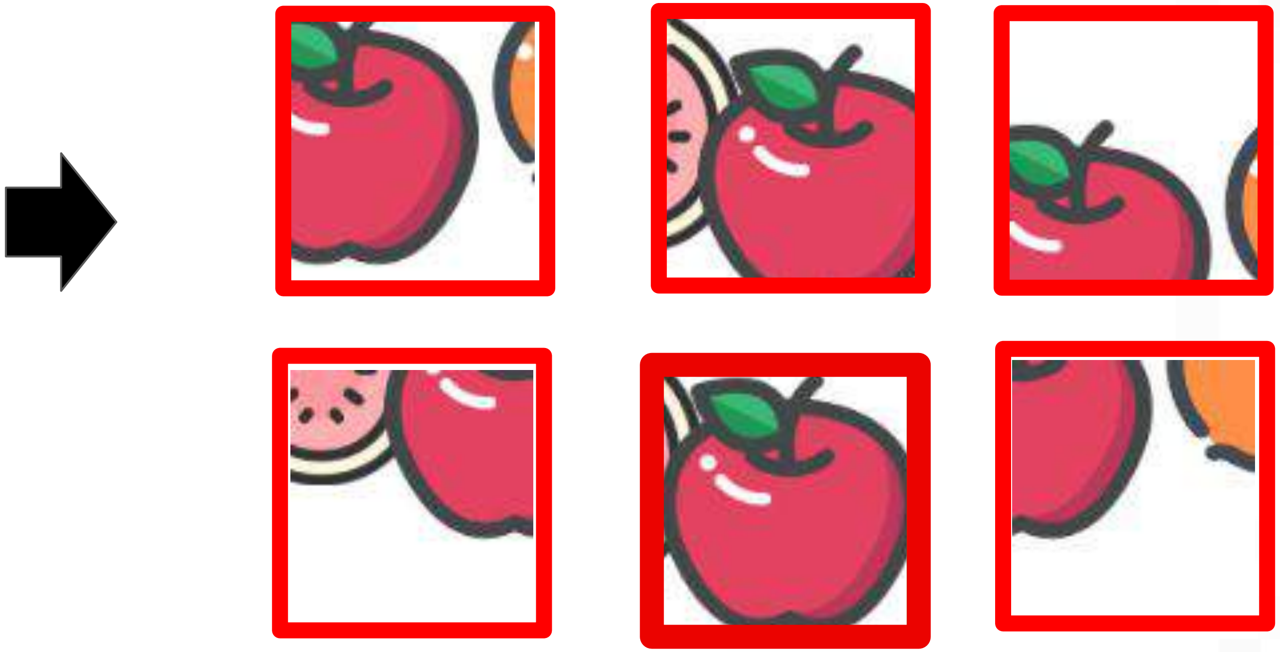


Quality of Labels

Region of interest should be centered on the object

(Will discuss more about this in hands-on session)





Quality of Labels

- **Labels of bad quality will take a significant impact on the AI workflow.**
- In particular, it will take a toll on
  - **AI modelling**
    - the algorithm in training unable to reach an optimal result due to ambiguity in the labels
  - **Deployment**
    - the AI model (trained) on the skewed labelled data providing incorrect output, impacting the business output

CERTIFAI
powered by Skymind

Quality of **Labels**

Various factors decide the quality of labels.

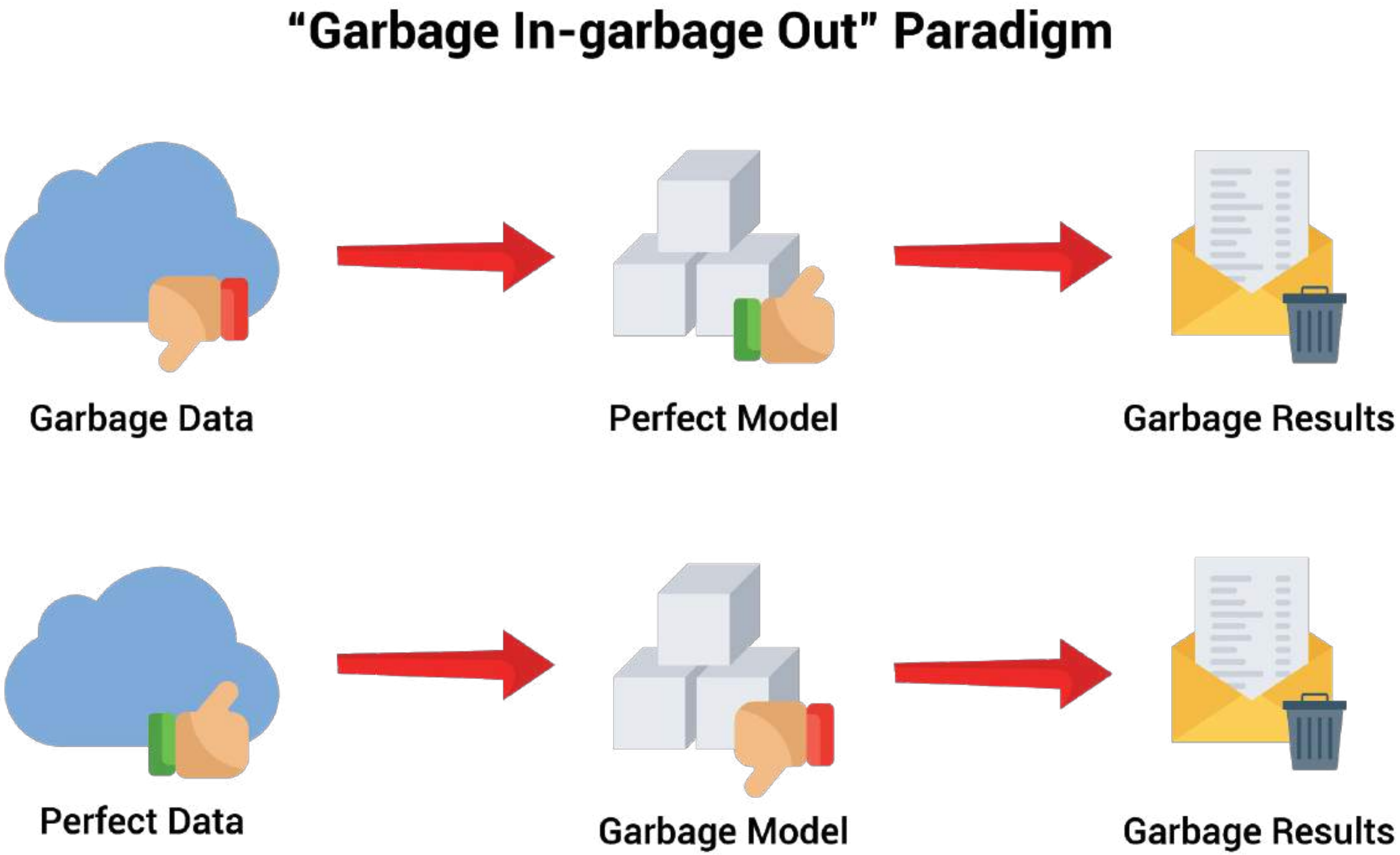
- **People** who responsible for the labelling tasks
- **Processes** in the routine work.
- **Products / infrastructures** to support labelling work.

Classifai is a labelling tool that addresses these three key factors.

CLASSIFAI

Take-away

O **FACT:** Data “healthiness” has to be considered.



Quality Assurance Techniques

Quality Assurance Techniques

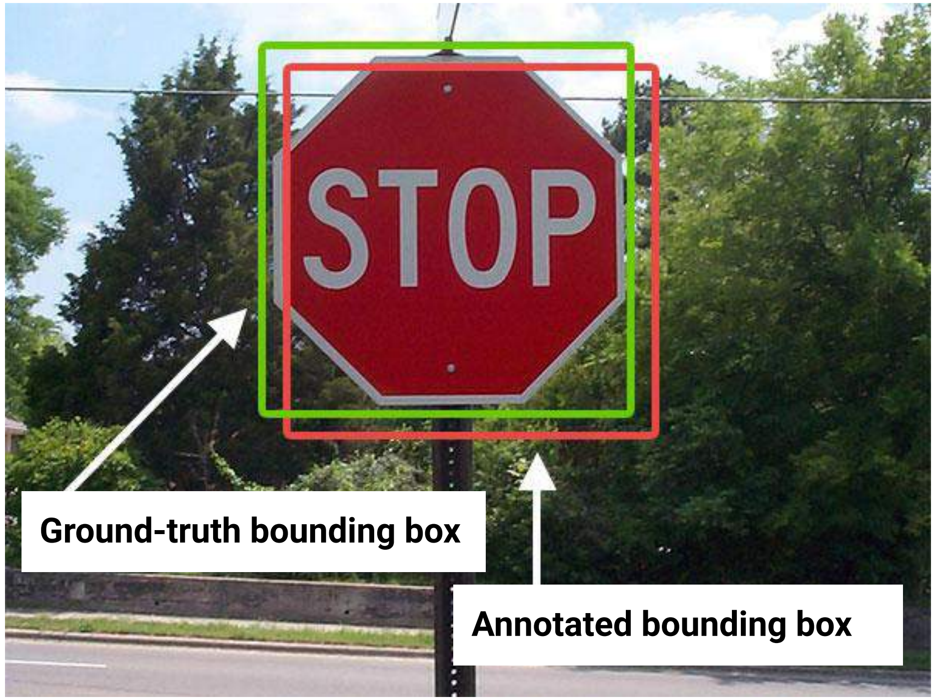
- As mentioned in the previous part, quality of labelling output drives the success of AI projects.
- Quality of a labelling process may be evaluated using the methods below:

1. Ground Truth Checking
2. Batch Sampling Review
3. Majority Agreement / Consensus

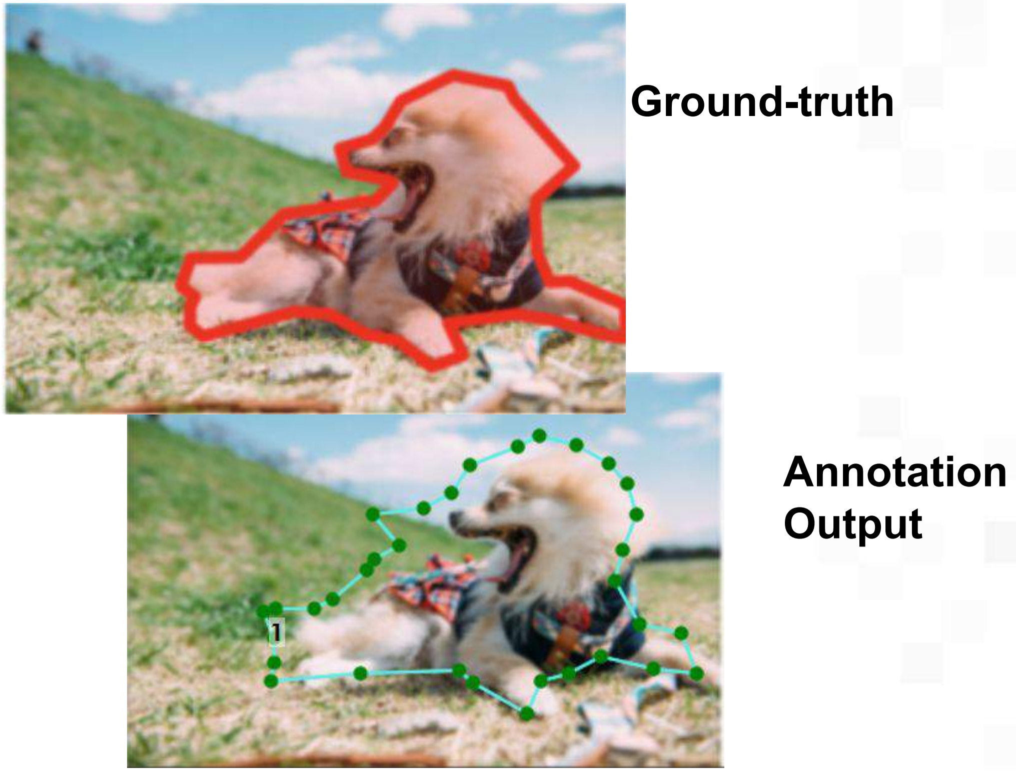
Quality Assurance Techniques

Method 1: Ground Truth Checking

- Comparison of the annotation to the ground truth. The quality of the label is determined with a metric called Intersection over Union (IoU).
- **Drawback:**
    - Time Consuming



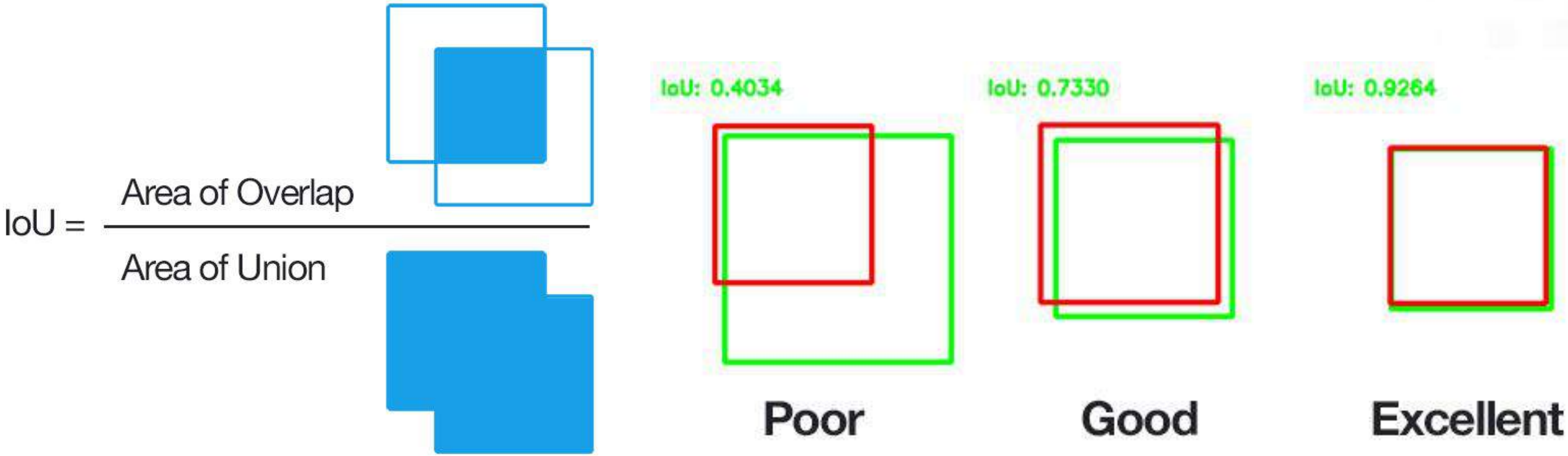
Bounding Box Detection



Segmentation

Quality Assurance Techniques

Intersection over Union (IoU)



Quality Assurance Techniques

Method 2: Batch Sampling Review

- Member of the QA team randomly draws and manually evaluates the quality of labels.
- **Drawback**:
    - Possibility of overlooking severely problematic samples

Quality Assurance Techniques

Method 3: Majority Agreement/ Consensus

- Labels that are considered as high quality are defined based on the convention that majority of the labelled datasets abide by.

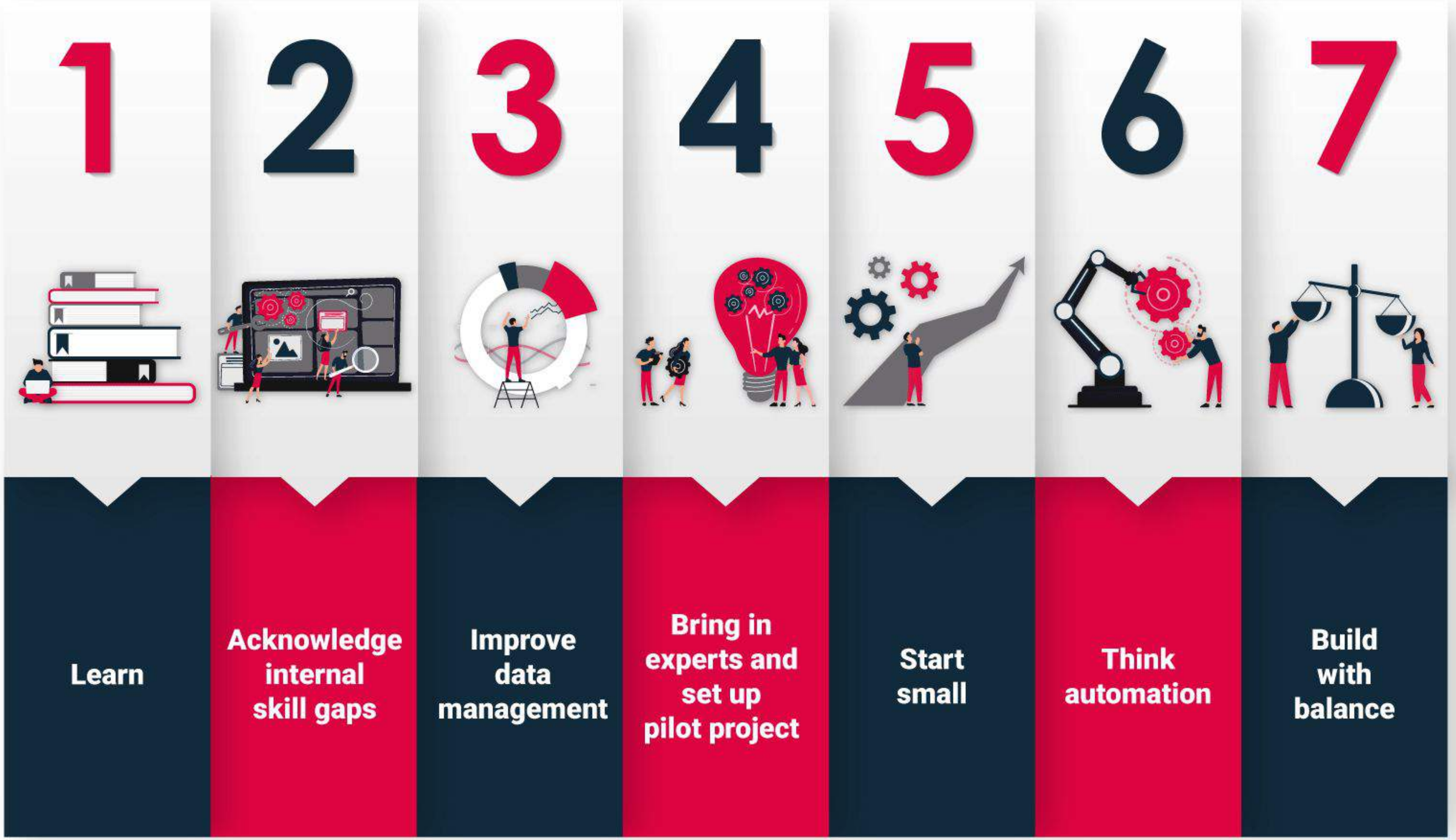
- **Drawback**:
    - Subjective approach

Adoption of AI into Business

How do you prepare for adopting AI solutions inside your business?

- Here are 7 steps to aid you in that process.

CERTIFAI
powered by Skymind



Topic Summary

- Data collecting strategies include data acquisition, data annotation and data enhancement.
- **Skilful data labellers** are essential for controlling data quality, accelerating the development of AI models, and facilitating workers with specialised data tasks.
- Understanding **data fallacies** aids in avoiding low model performance while formulating an inference.
- The **quality of labels** is determined by people, processes and products.