

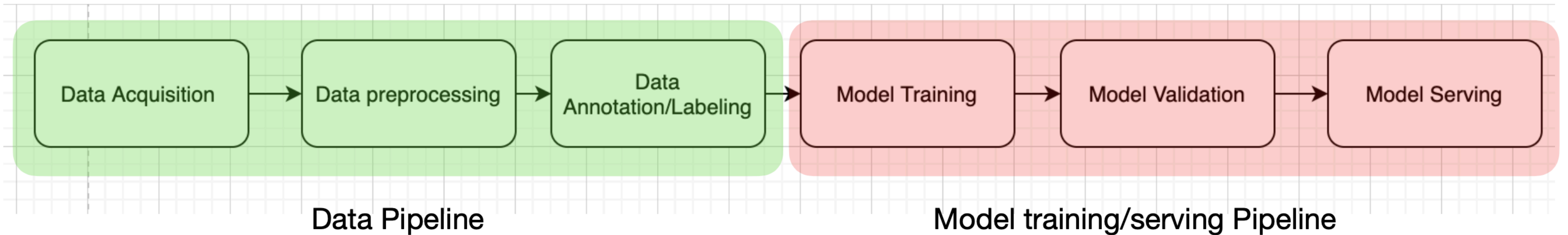
Data in deep Learning

Perumadura De Silva

Content

- Introduction to Data Pipeline
- Structured vs Unstructured Data
- Data types classification
- Descriptive statistics basics
- Importance of data in deep learning
- Improving poor datasets

Introduction to Data Pipeline



- Data Pipeline:
 - Responsible for preparing the data suitable for model consumption
- Training/ Serving Pipeline
 - Train a set of models by utilising the data feed from the data pipeline
 - Select a model through model validation:
 - Check how the model performs on different unseen data
 - Check if the model satisfies the resource requirements for serving
- Serve the model: Let the consumers use the model for e.g. classification in production level (e.g. RestAPI calls)

Introduction to Data Pipeline

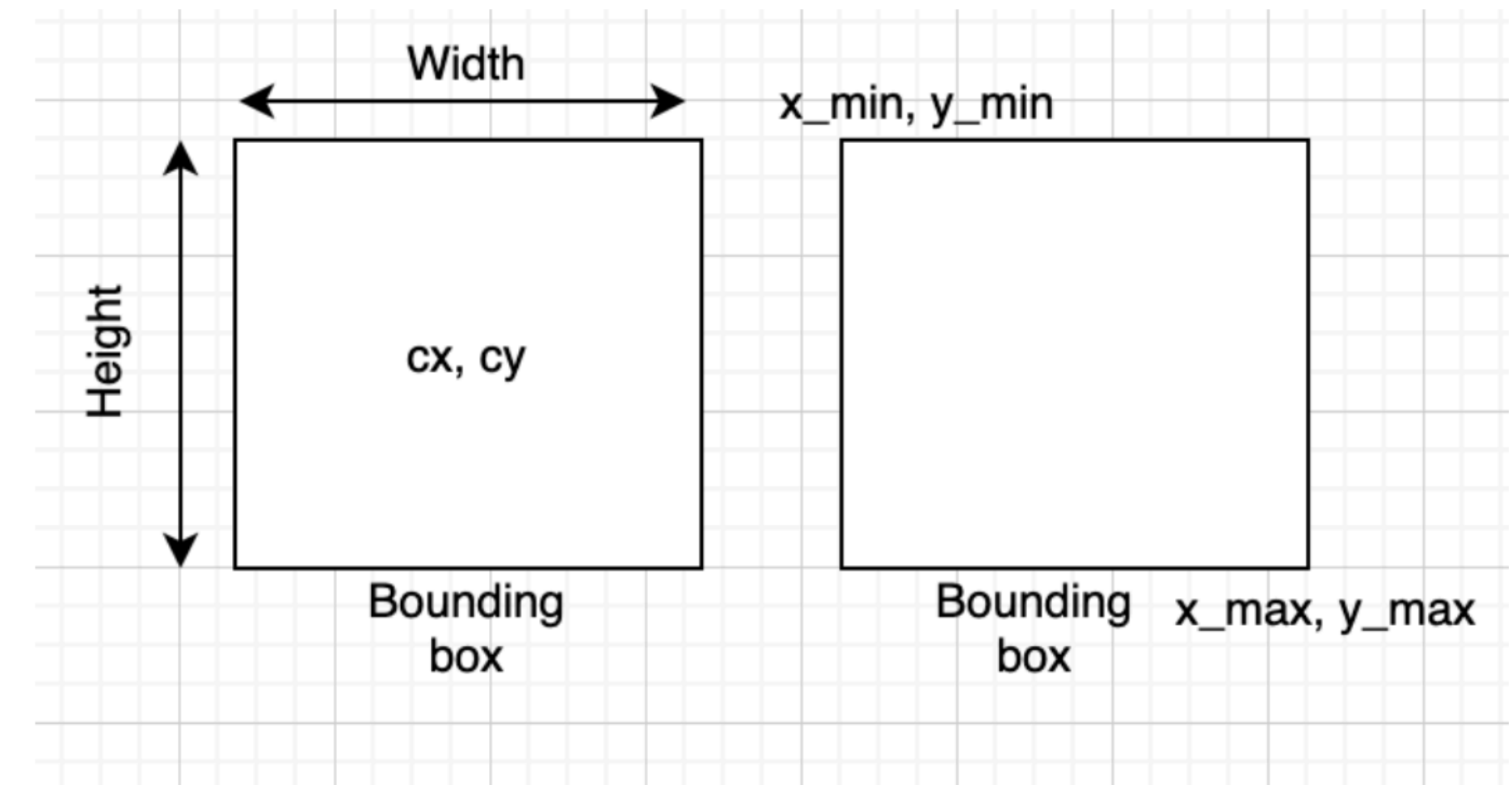
- Data acquisition:
 - Images and Videos: From Cameras
 - Text data: Web scrapping, Output of an Optical Character Recognition (OCR) system
 - Audio data: From Microphone
 - Tabular data: From sales or shipping records, inventory data

Introduction to Data Pipeline

- Data Preprocessing:
 - This step of data preparation is completely task dependent
 - E.g: Implement image denoising or deblurring techniques to improve the image quality for image classification
 - But if the task is to denoise an image, we do not follow the image quality enhancement step
- Typical preprocessing steps are image quality enhancement, filling the missing data through approximation

Introduction to Data Pipeline

- Data annotation/labelling:
 - In the supervised learning case:
 - deep learning (DL) models need inputs and their corresponding labels
 - For classification task data must be labeled with the corresponding class
 - E.g: cat_image.jpg – cat, dog_image.jpg – dog ...
 - For object detection task, images/videos must be annotated:
 - class_name: cat
 - bounding_box_coordinates :
 - [center_x, center_y, box_height, box_width]
 - [x_min, y_min, x_max, y_max]



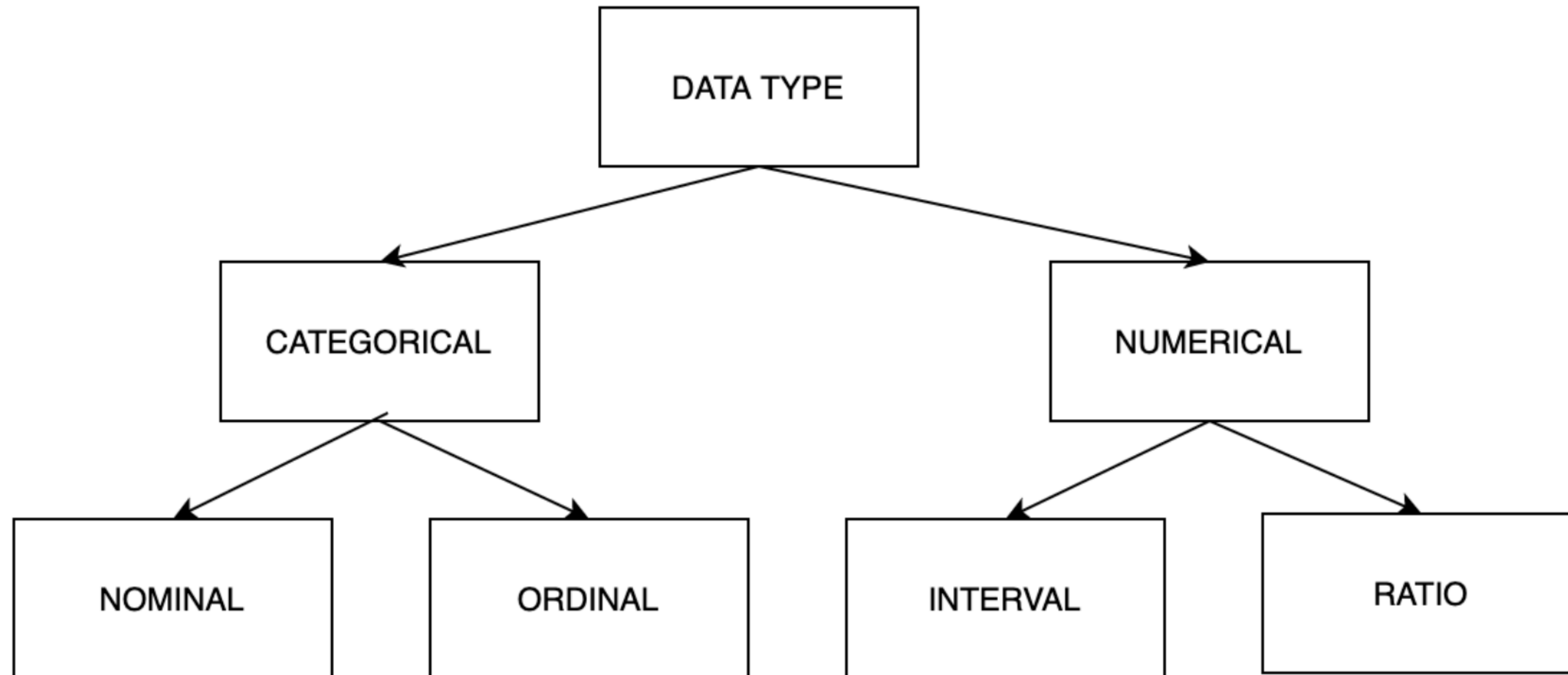
Introduction to Data Pipeline

- List of image annotation software:
 - Free:
 - <https://github.com/tzutalin/labelImg>
 - <https://github.com/cgvict/roLabelImg>
 - <https://labelstud.io>
 - <https://www.makesense.ai/>
 - Commercial:
 - <https://labelbox.com/>
 - <https://prodi.gy/>

Structured vs Unstructured Data

- Unstructured data:
 - Data that does not provide any information about the content directly
 - E.g: Images, videos, email and pdf data etc.
- Structured data:
 - Data that is highly organised and searchable in a database
 - E.g: Data entry with key, value pairs

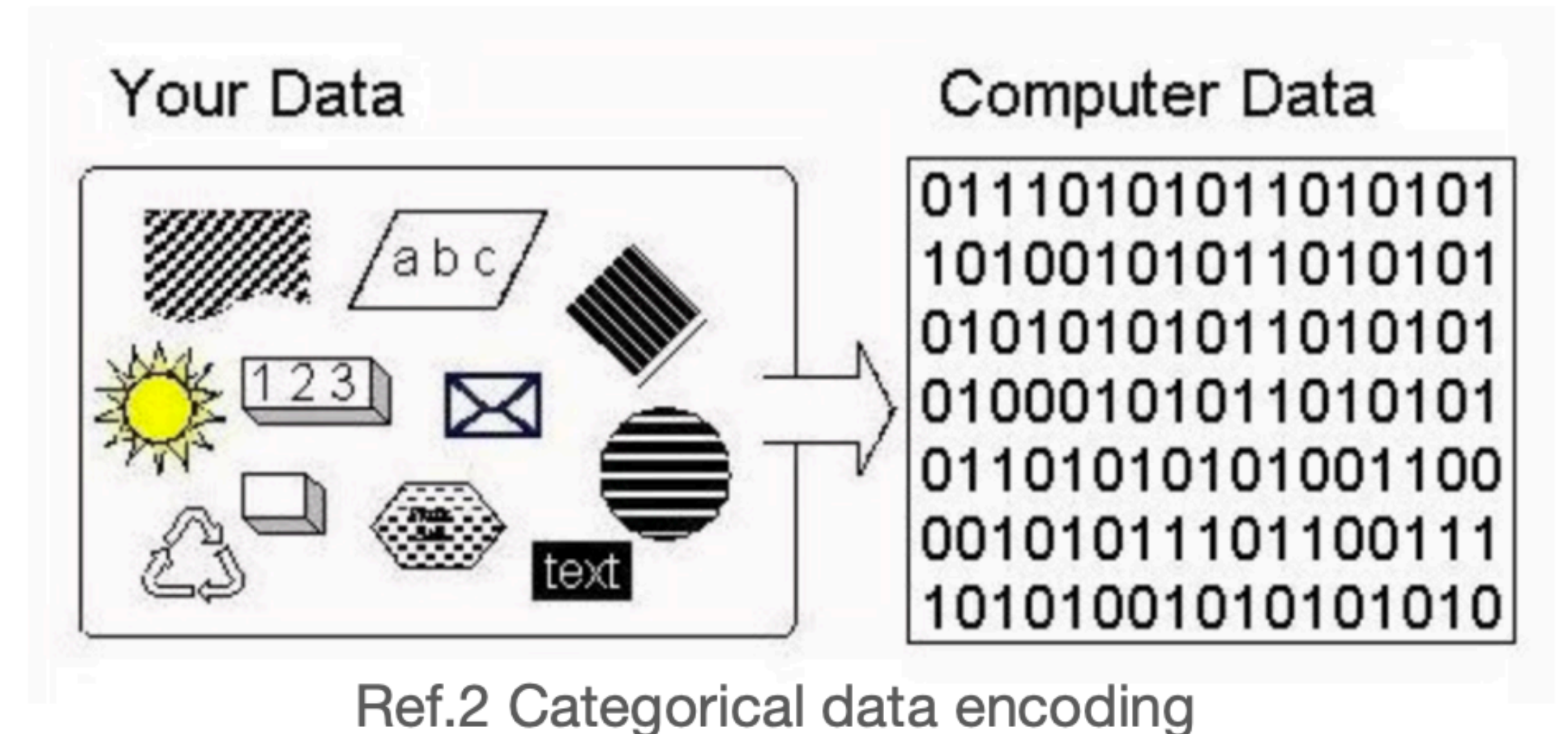
Data type classification



Classification of data types

Data type classification

- Categorical data:
 - Data that is organised into categories according to their characteristics,
 - This data can be represented in numerical values but have no mathematical meaning
 - Do numerical encoding on categorical data for machine understanding
 - E.g. 1-Red, 2-Green, 3-Blue



Data type classification

- Categorical data:
 - Nominal Data:
 - Represented as discrete values
 - Data labelling is done using nominal data
 - **Does not** have an order in the data
 - Ordinal Data:
 - Similar to nominal data but has an order
 - The distance between two units are **not** measurable

Data type classification

- Numerical Data:
 - Data that are represented using numerical values
 - The distance between two data points are measurable

.

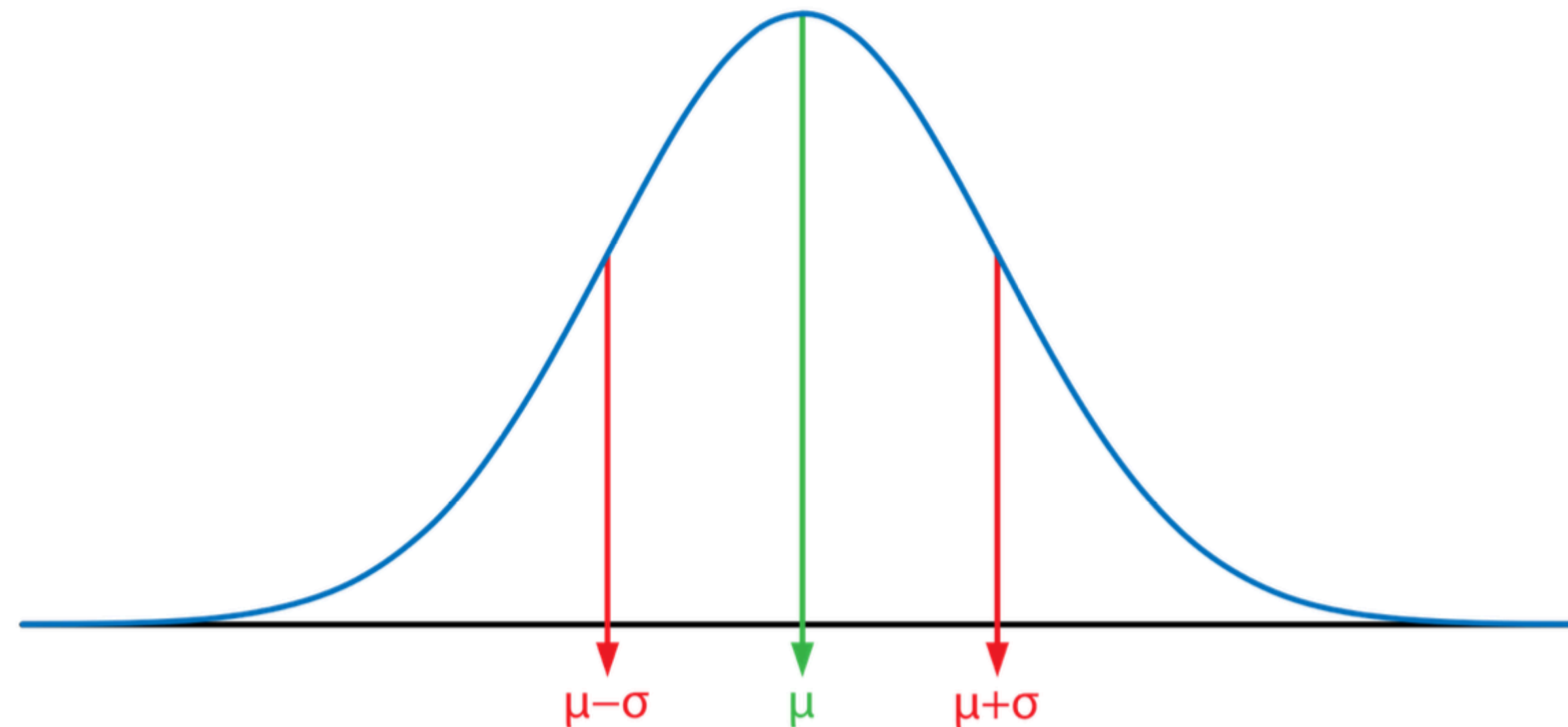
Data type classification

- Numerical Data:
 - Interval Data:
 - Data that takes a numerical value and has an order
 - But interval data does not have an absolute “0”
 - E.g. The temperature in °C or °F
 - Ratio Data:
 - Similar to interval data, but has an absolute “0”
 - E.g. The temperature in Kelvin (K), Height, Length

.

Descriptive statistics basics

- Descriptive statistics, describe, summarise and organise the data
- Data is represented using a normal distribution
 - Properties: Bell-shaped symmetrical, Centred and unimodal (one peak)



Ref.3 Normal distribution with mean μ and standard deviation σ

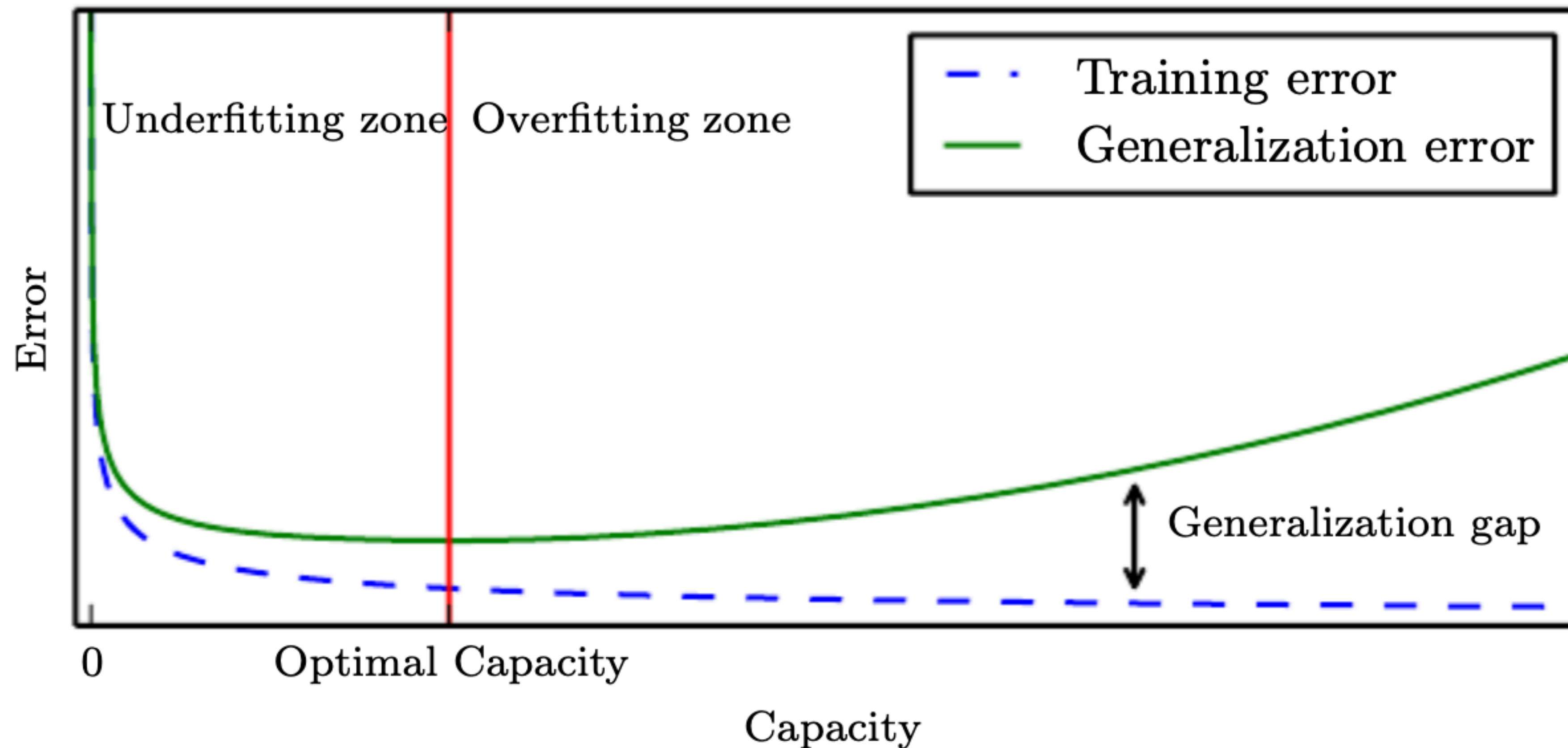
Descriptive statistics basics

- Central Tendency:
 - The descriptors for central trends: mean, median and mode
 - Mean: Simple average of the data
 - Mode: The most occurring category of the data
 - Median: Mid value of the data or the 50th percentile
 - Median is not much affected by the outliers
 - In a normal distribution: $\text{Mean} = \text{Mode} = \text{Median}$
- Variability of data
 - The variability of data measures how much the data spreads
 - Methods: Variance, Standard deviation, Interquartile range (IQR)

Colab Exercise

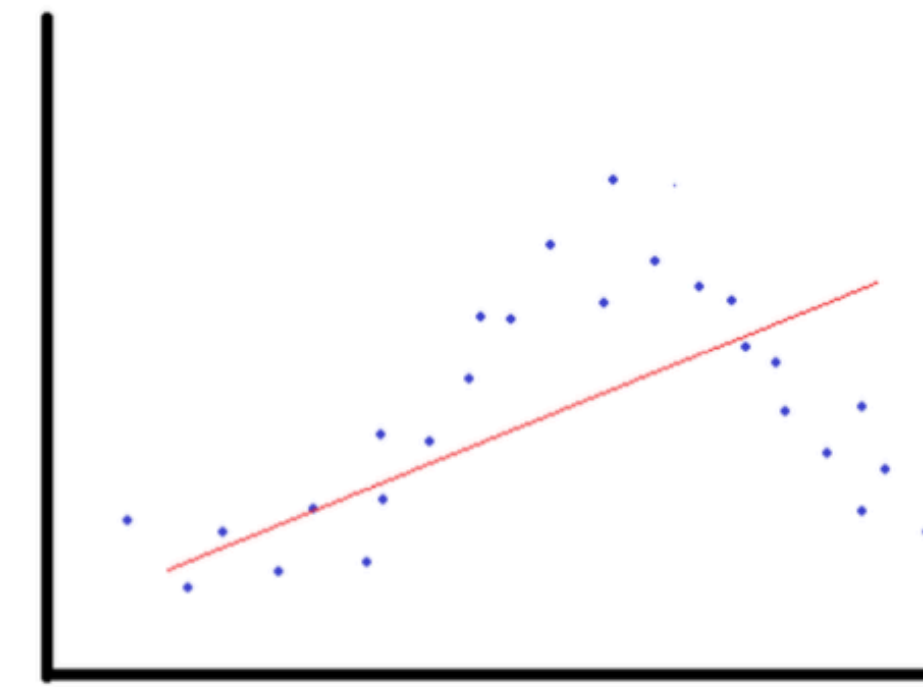
Model Optimality

- Model Capacity vs Prediction error

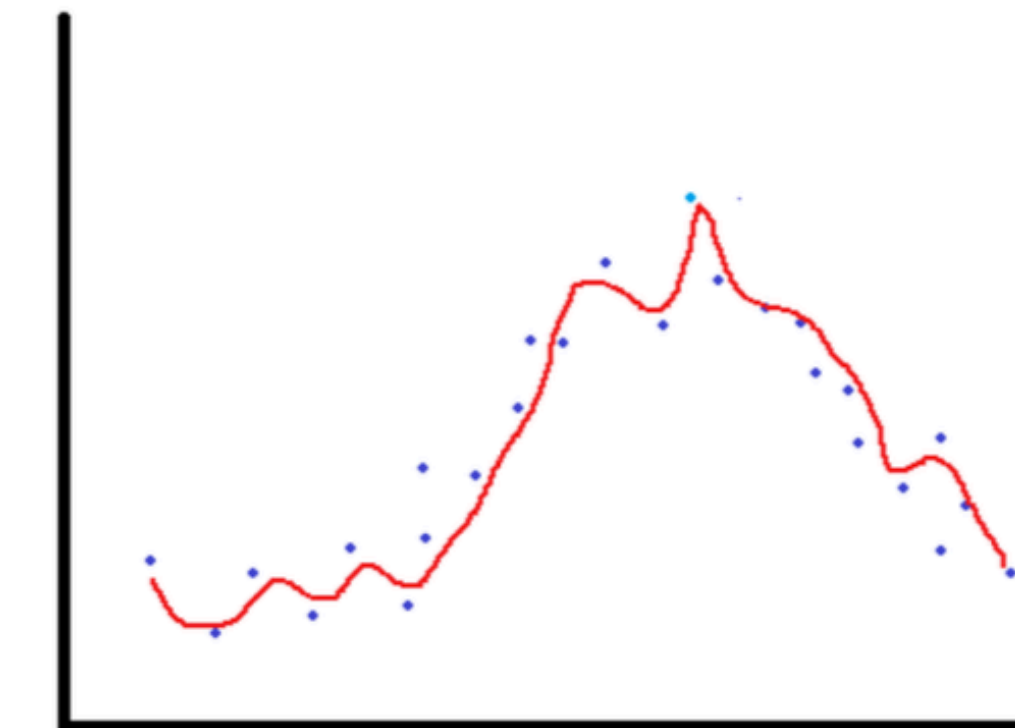


Importance of Data in Deep Learning

- Two main properties of a machine learning model prediction are:
 - Bias:
 - How much the prediction deviates from the true value.
 - Happens when solving a complex task using a simple model
 - Higher bias leads to underfitting
 - Indicated by high training and validation error gap
 - Variance:
 - The variability in the model predictions
 - Happens when the model learn the patterns by remembering
 - Higher variance leads to overfitting
 - Indicated by very high gap between training and validation error



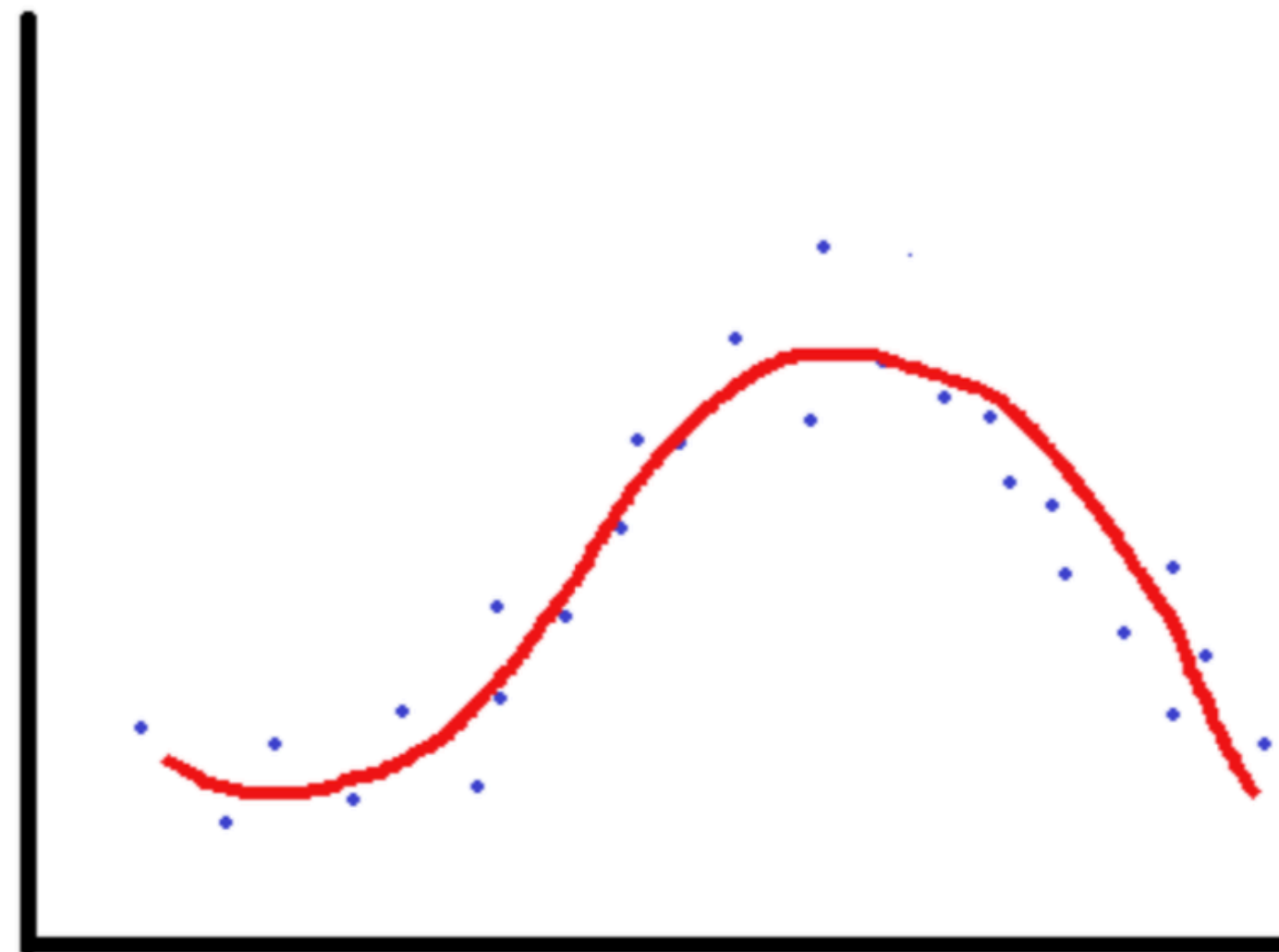
Ref.3 Model with high bias (Underfitting)



Ref.3 Model with high variance (Overfitting)

Importance of Data in Deep Learning

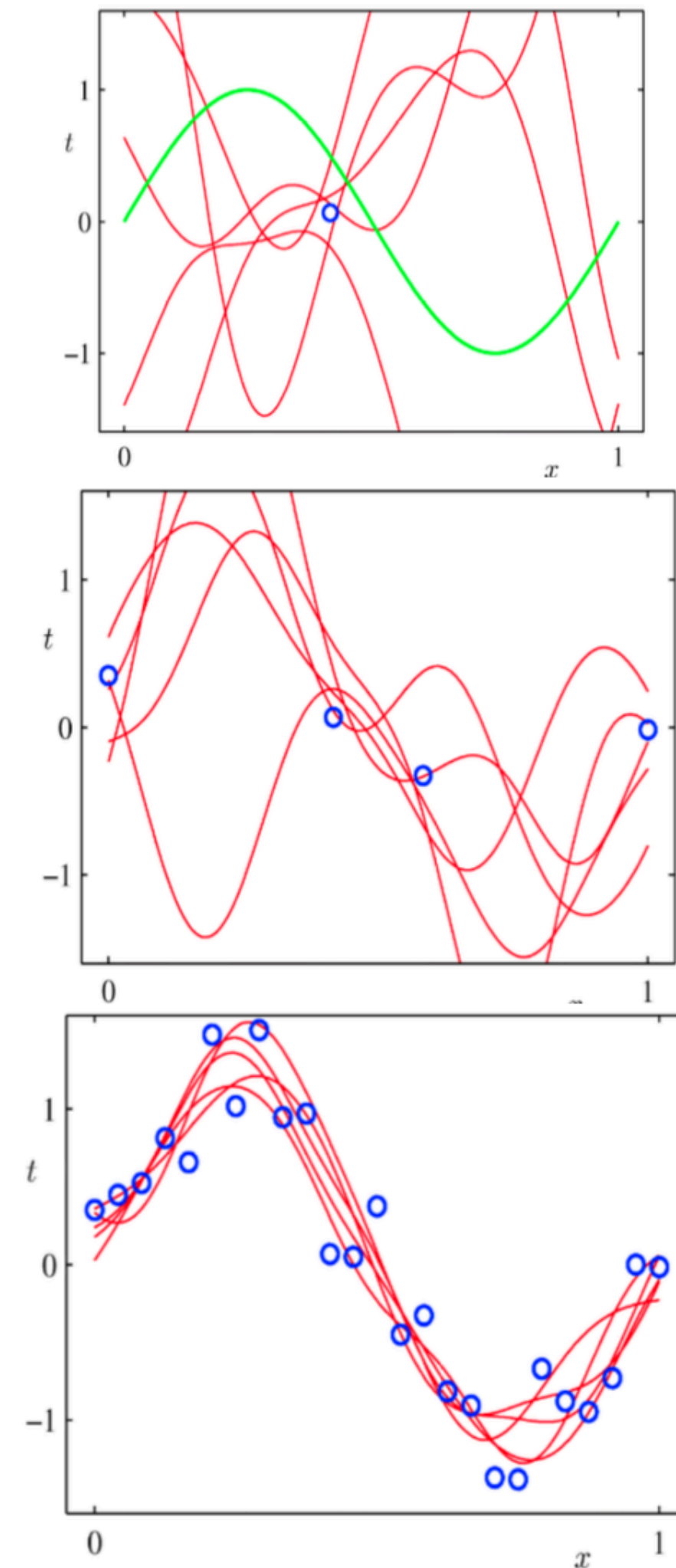
- So when designing a model for a given task, the idea is to find a balance between bias and variance (there will be always tradeoffs)



Ref.3 Model with a good fit

Importance of Data in Deep Learning

- To see the importance of data, let's consider one data point:
- Many models can approximate this datapoint easily
- Now when we increase the number of data points:
- We only have the models that can approximate the function that is used to generate the data



Ref.4 Reduction in number of possible models as the number of data points increases

Improving poor datasets


- A poor data set could have following issues:
 - Many missing values (in tabular or time series datasets)
 - Class imbalance in classification case (image or time series datasets)
 - Small datasets (could lead to model overfitting)
 - Too many distortions in data (e.g: noise, blur and skew in image data)

Improving poor datasets

- Solutions to improve poor datasets (some methods):
 - Missing values: Fill the missing values by means of Imputation
 - Follow: <https://www.kdnuggets.com/2020/06/missing-values-dataset.html>
 - Class imbalance in classification:
 - Undersample the over represented classes
 - Oversample the under represented classes through duplication or augmentation
 - Small dataset:
 - Increase the number of data through data duplication or augmentation
 - Too many distorted data:
 - Apply preprocessing such as denoising, deblurring and skew correction on distorted data

Improving poor datasets

- The class imbalance case:

Class-1	1000		Class-1	1000
Class-2	300		Class-2	1000
Class-3	10		Class-3	1000
Class-4	2		Class-4	1000
Class-5	860		Class-5	1000

- Ideally there should be equal number of data samples for each class
- This imbalance could lead to a bad generalisation error in the corresponding classes
- Therefore oversampling of the under represented classes is required Typically oversample to the most over represented class if possible

Improving poor datasets

- Image data oversampling techniques:
- Oversample through data duplication:
 - This technique adds no new information to the dataset and could lead to model overfitting
- Oversample through data augmentations:
 - Generate a slightly changed image so that the image features are different
 - The features are changed by introducing distortions such as e.g. rotation, flipping, addition of noise and blur to the image

Improving poor datasets

- Class balancing by means of augmentation:
 - Visualise the current state of class balance
 - Calculate the required number of images per class
 - Define an augmentation strategy
 - Generate the required number of images using the defined augmentation strategy
 - Update the dataset
 - Verify the class balance through visualisation

Reference

- Ref.1: <https://helpx.adobe.com/photoshop/key-concepts/skew.html>
- Ref.2: <https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/>
- Ref.3: <https://towardsdatascience.com/bias-and-variance-in-machine-learning-b8019a5a15bc>
- Ref.4: <https://towardsdatascience.com/breaking-the-curse-of-small-datasets-in-machine-learning-part-1-36f28b0c044d>