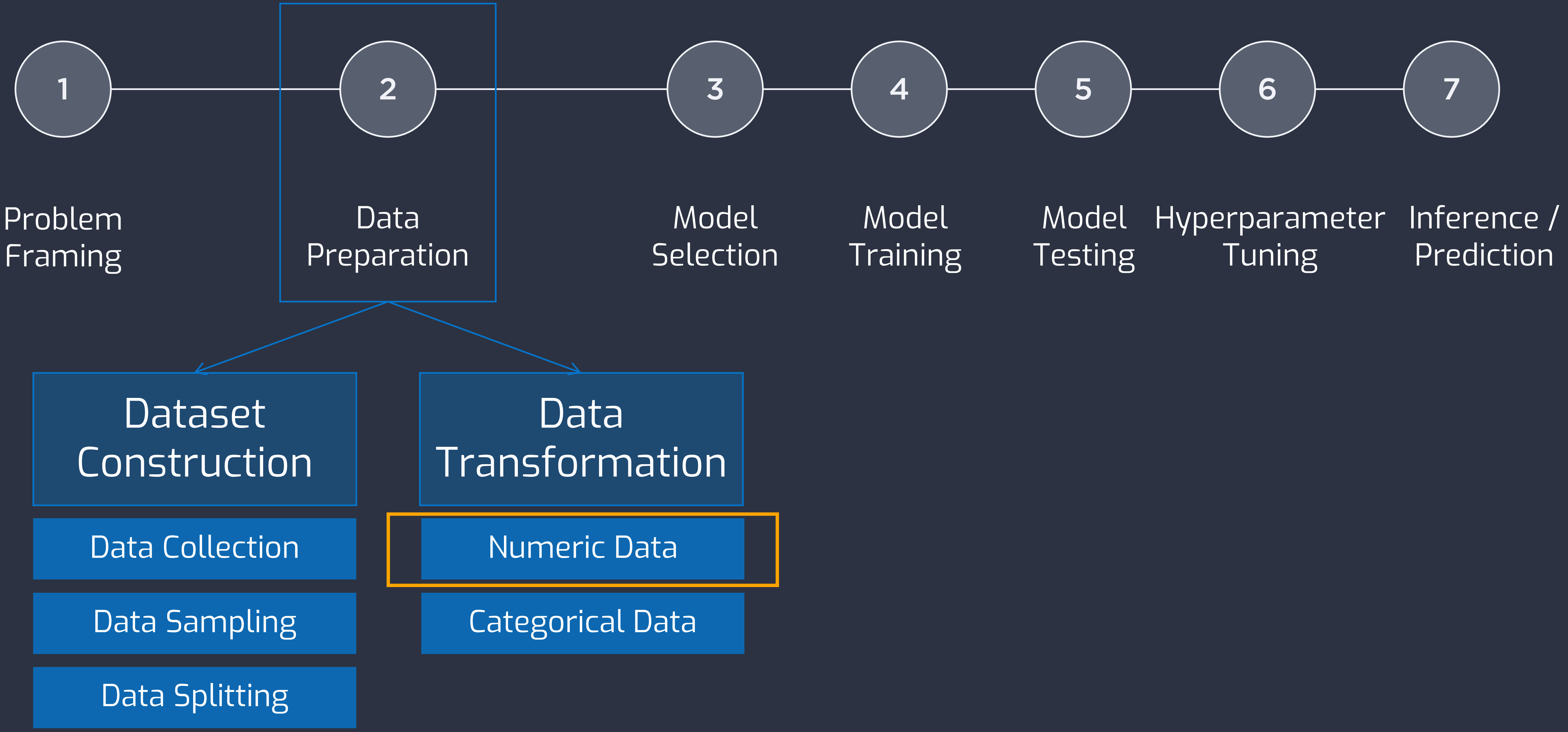
COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

# Transforming Numeric Data

Dr SHI Lei

Durham University

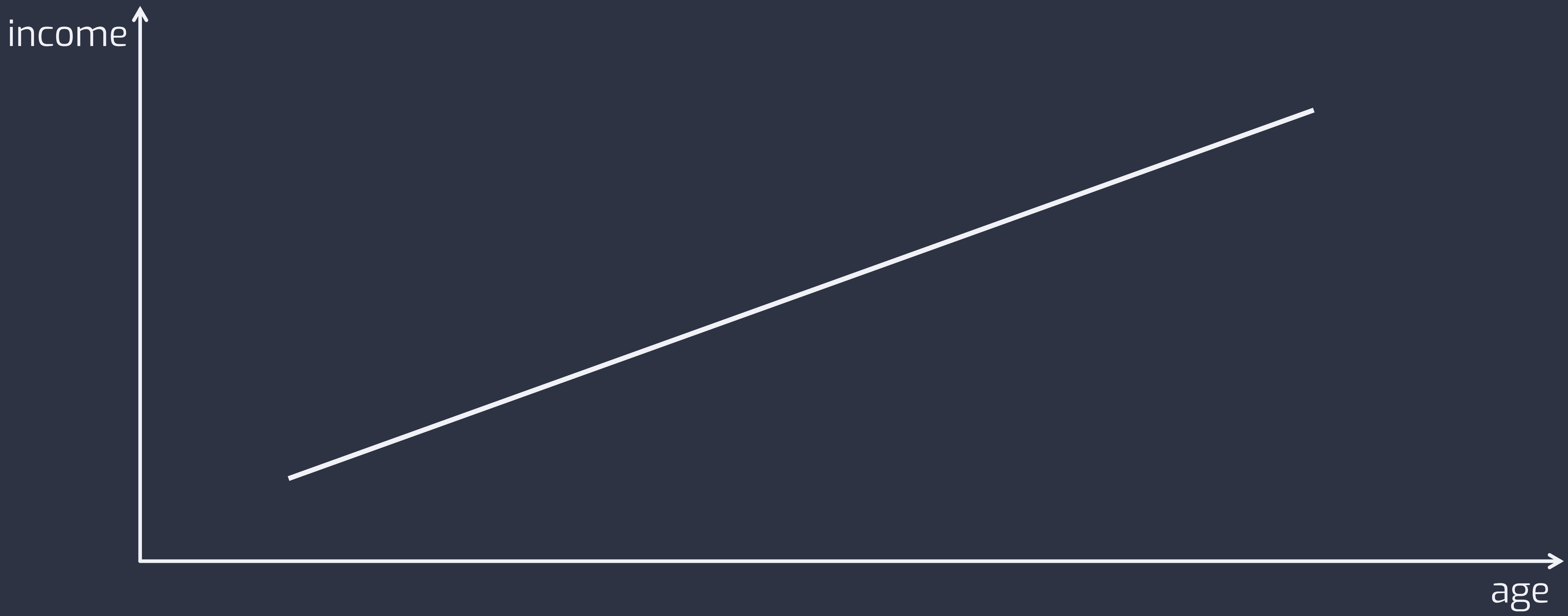
learning lab


1
2
3
4
5
6
7
Problem Framing
Data Preparation
Model Selection
Model Training
Model Testing
Hyperparameter Tuning
Inference / Prediction
Dataset Construction
Data Transformation
Data Collection
Data Sampling
Data Splitting
Numeric Data
Categorical Data

# Learning Objectives

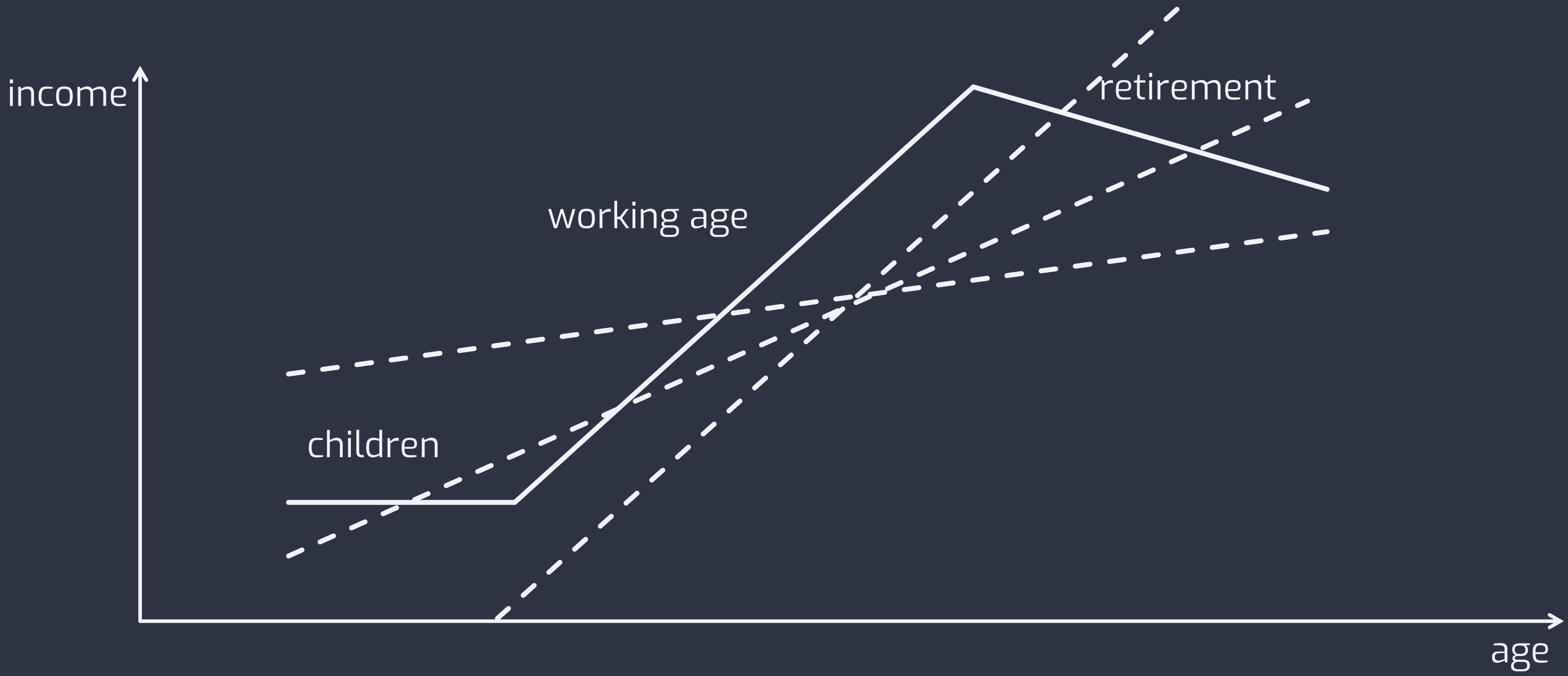
- Understand what is data binning and how to use it.
- Understand what is feature scaling and how to use it.
- Understand difference between feature scaling techniques.

EXAMPLE. use age to predict income



What could go wrong with this approach?

**EXAMPLE.** use age to predict income

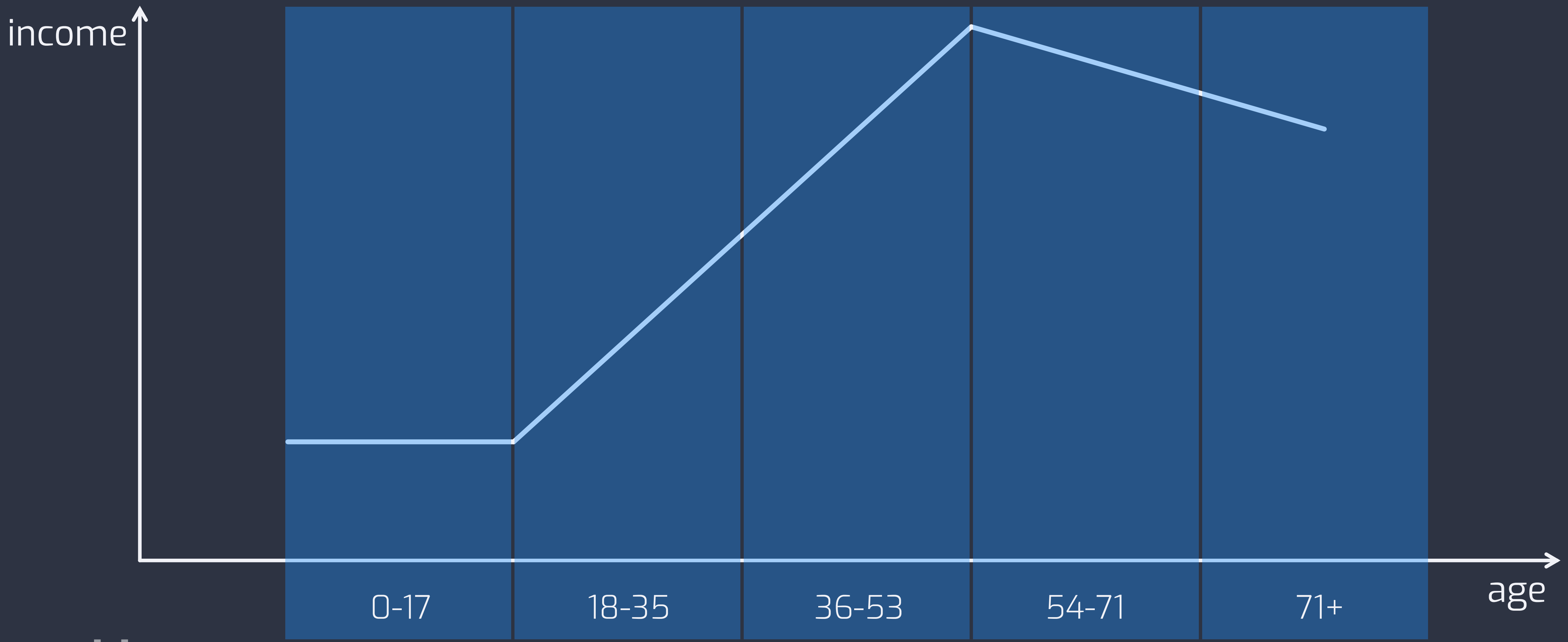


# Data Binning

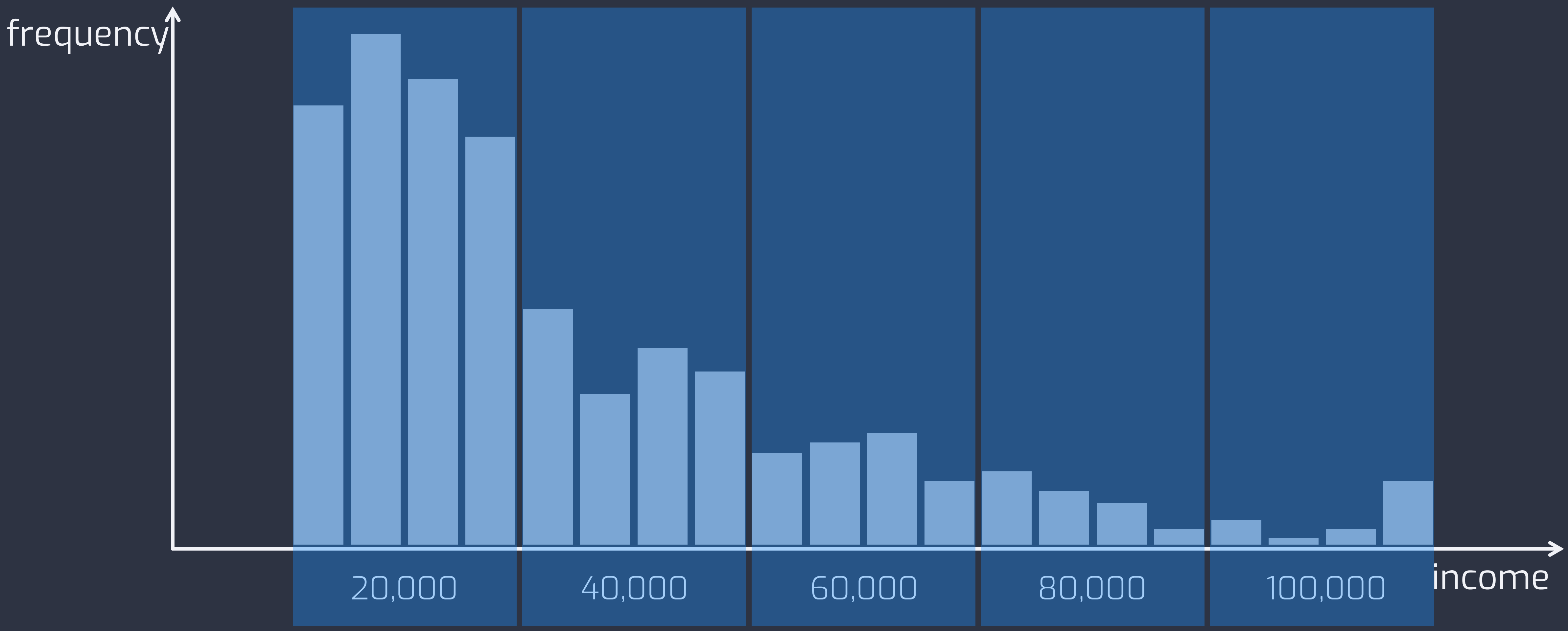
# Data Binning

**EXAMPLE.** use age to predict income

Transform numeric features into categorical ones based on range it falls into.

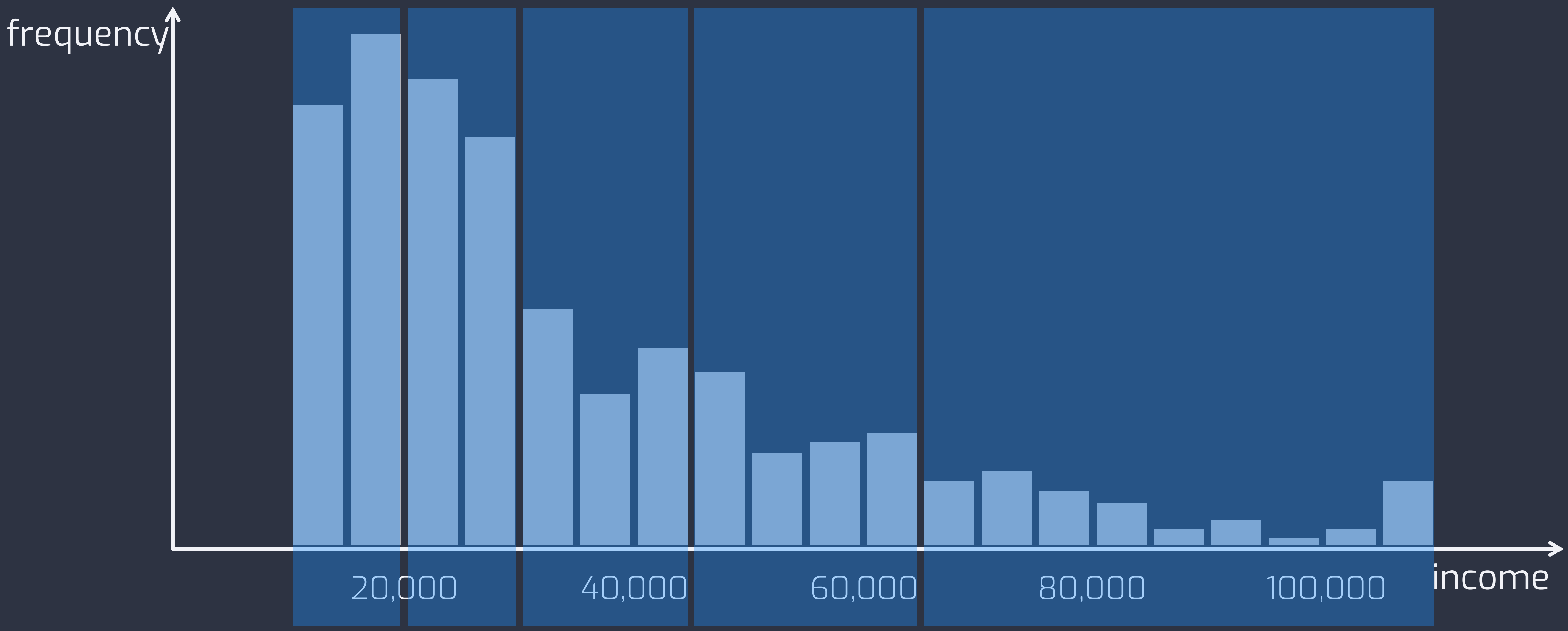


# Data Binning




Equally Spaced Binning (Fixed-Width Binning)

# Data Binning




~~Equally Spaced Binning (Fixed-Width Binning)~~

Quantile Binning

# Data Binning

When choosing data binning techniques, we must be clear about how to set the boundaries and which type of binning we want to use.

## Equally Spaced Binning

The boundaries are fixed and encompass the same range (e.g. age range: 18-35, 35-53,54-71). In this case, some bins may contain a lot more instances, and others may contain very few.

## Quantile Binning

The boundaries are not fixed and encompass a wide or narrow range (e.g. income: ~20k, 20k~30k, 30k~45k, 45k~65k, 65k~110k). In this case, each bin contains equal (or similar) number of instances.

## EXAMPLE. Features in very different ranges

A dataset containing two features, age and income.

Age ranges from 18-71; income ranges from 22,000-92,000


Income will intrinsically influence the result much more.

But it's not necessary that income is more important as a predictor than age.

# Feature Scaling

# Feature Scaling

- Can help transform the values of numeric features to be on a similar scale without distorting differences in ranges of values.

| | | |
|---|---|---|
| Min-Max Normalisation | Mean Normalisation | Standardisation |
| Unit-Length Scaling | Log Scaling | Clipping |

- It is necessary for many machine learning algorithms, e.g., many classifiers calculate the distance between 2 instances by e.g. Euclidean distance, so if one of the features is in a much larger range, it will dominate the distance.

- Gradient Descent converges faster with feature scaling (to cover later).

# Min-Max Normalisation

# Feature Scaling - Min-Max Normalisation

- The simplest technique to scale features in similar ranges.
- Can be used to rescale feature into the range of `[0, 1]`, via

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

e.g. $age \in \boxed{[18,71]} \longrightarrow age' \in \left[\frac{18-18}{71-18}, \frac{71-18}{71-18}\right] = \boxed{[0,1]}$

$income \in \boxed{[22000,92000]} \longrightarrow \text{income}' \in \left[\frac{22000-22000}{92000-22000}, \frac{92000-22000}{92000-22000}\right] = \boxed{[0,1]}$

- Can be used to rescale feature into other ranges e.g. `[-1, 1]`, depending on the nature of data and the learning algorithms to be used.

# Mean Normalisation

# Feature Scaling - Mean Normalisation (Normalisation)

- Rescale the feature values around the mean value.

$$x' = \frac{x - \mathrm{mean}(x)}{\max(x) - \min(x)}$$

# Standardisation

# Feature Scaling - Standardisation (Z-score Normalisation)

- Rescale the feature values to have zero-mean and unit-variance, i.e. $\mu = 0$ & $\sigma=1$

$$x' = \frac{x - \text{mean}(x)}{sd(x)}$$

or

$$z = \frac{x-\mu}{\sigma}$$

- Commonly used in many machine learning algorithms e.g. K-Nearest Neighbours and Support Vector Machines, Principal Component Analysis, Clustering, LASSO and Ridge regressions.

- Not necessary to machine learning algorithms which are not sensitive to the magnitude of features, e.g. Logistic Regression, Naive Bayes, and Tree-based algorithms such as Decision Tree, Random Forest and Gradient Boosting.

## Min-Max Normalisation

$$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$$

VS

## Standardisation

$$x' = \frac{x - \text{mean}(x)}{sd(x)}$$

- Min-Max Normalisation can generate smaller standard deviations than Standardisation, so can scale out data to be more concentrated around the mean value.
- Min-Max Normalisation doesn’t require Gaussian distribution, so good for K-Nearest Neighbours and Neural Networks, but it doesn’t handle well outliers; whereas Standardisation can help with cases where data follows Gaussian distribution, and it can better deal with outliers and facilitate convergence for e.g. Gradient Descent.
- We can always try fitting model to raw, normalised and standardised data and then compare their performances for the best results.

# Unit-Length Scaling

# Feature Scaling - Unit-Length Scaling

- Rescale the components of a feature vector, so the complete vector's length is one.

$$x' = \frac{x}{\|x\|}$$

if $\vec{x} = (x_1, x_1, \dots, x_n)$ then $\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$

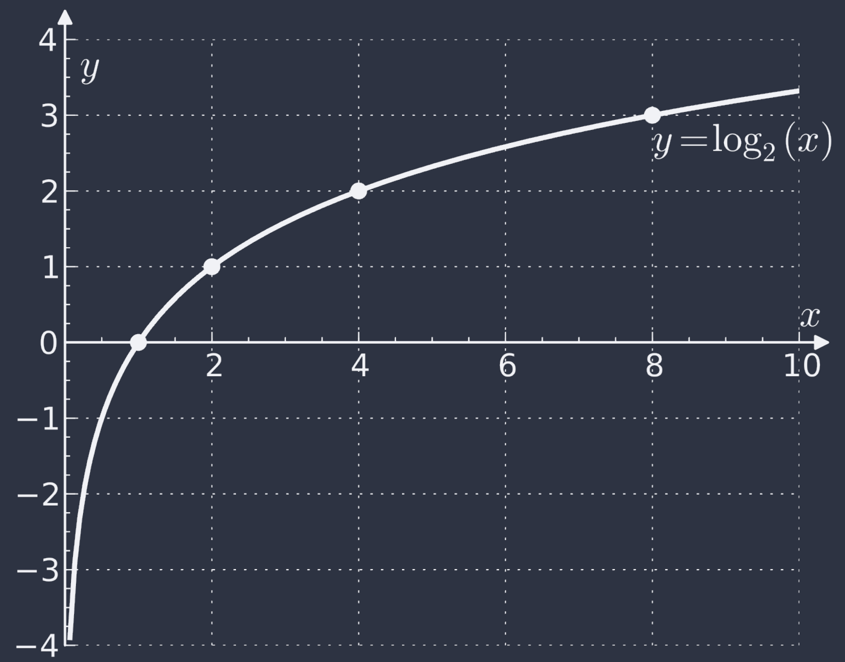
- This is to normalise N-dimensional vector features to have unit length (length 1), similar to normalising 1-dementional features to have a range of (0,1).

# Log Scaling

# Feature Scaling - Log Scaling

- Rescale feature values to a narrow range.
- May make skewed numeric features to become normally distributed.

$$x' = \log(x)$$




- The output of log function for positive values increases very slowly, and so higher values are marginalised more as compared to lower values.
- Very useful when dataset has many instances sharing a small range of values, but very few instances sharing a large range of values.

# Clipping

# Feature Scaling - Clipping

- Caps all the feature values which are either above a specific max value or below a specific min value.
- Formula: `set max/min values to avoid outliers`.
- To be used when dataset containing extreme outliers.

    e.g. clip all height values above 2 meters to be exact 2 meters.

# Takeaway Points

- Binning to transform numeric features into categorical ones based on range it falls into.
- Quantile Binning to avoid some bins containing much more data than other bins.
- Feature Scaling to transform numeric features to be on a similar scale without distorting differences in ranges of values.
- Different Feature Scaling techniques for different data and learning algorithms.