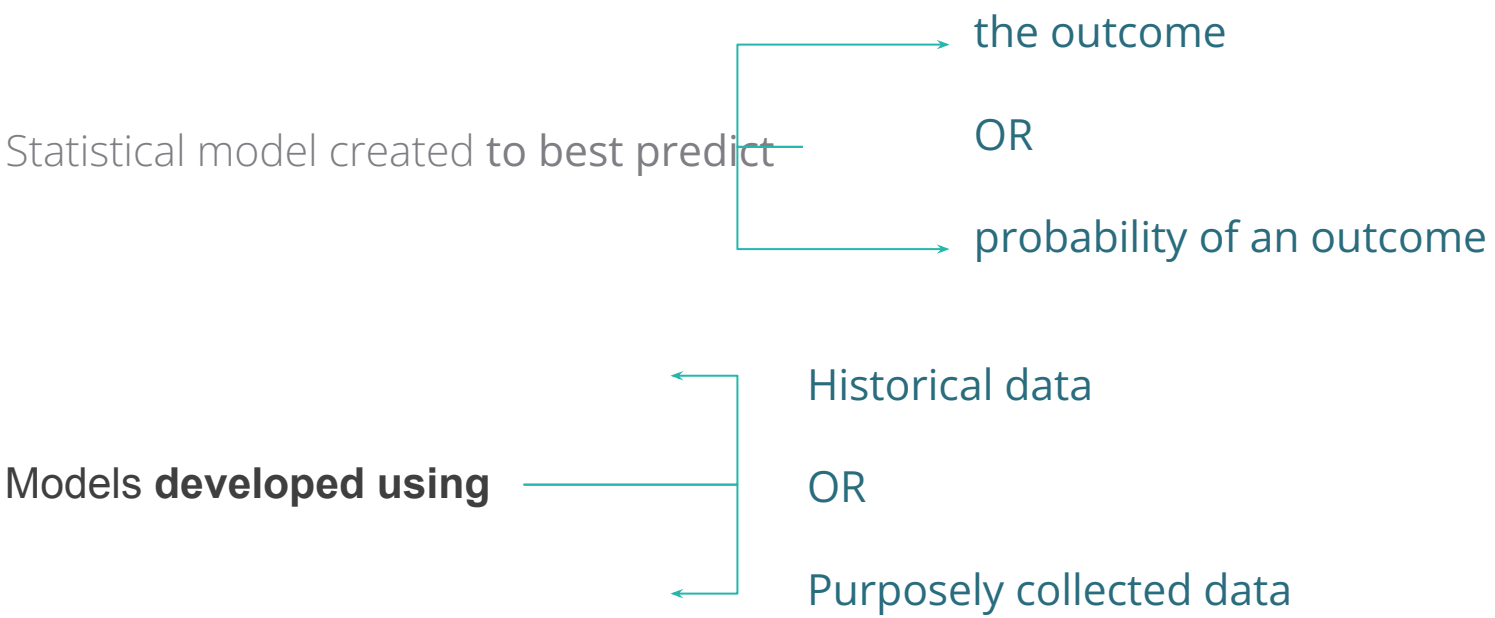
Introduction to Predictive Modelling

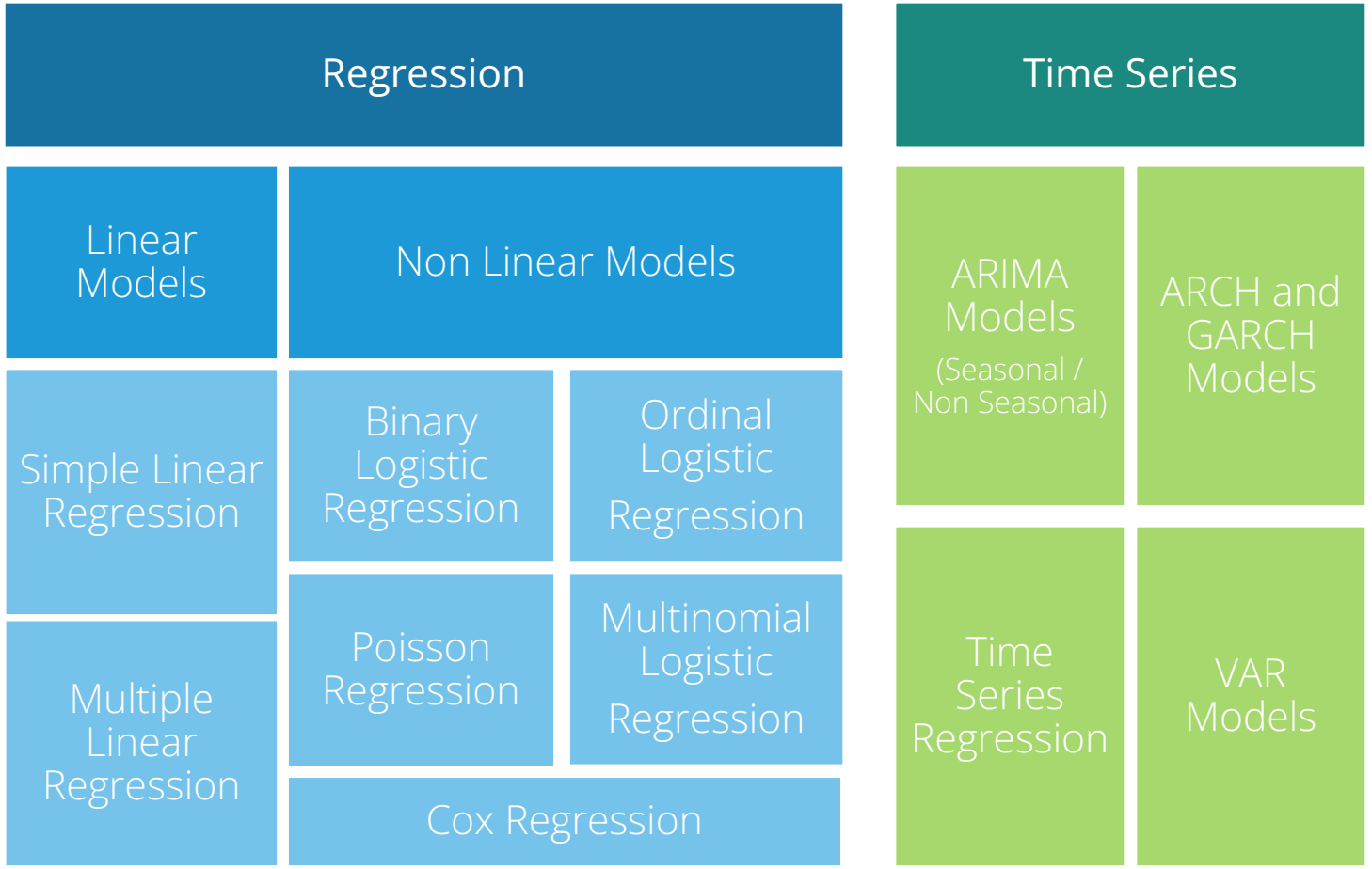
Contents

What is Predictive modelling?

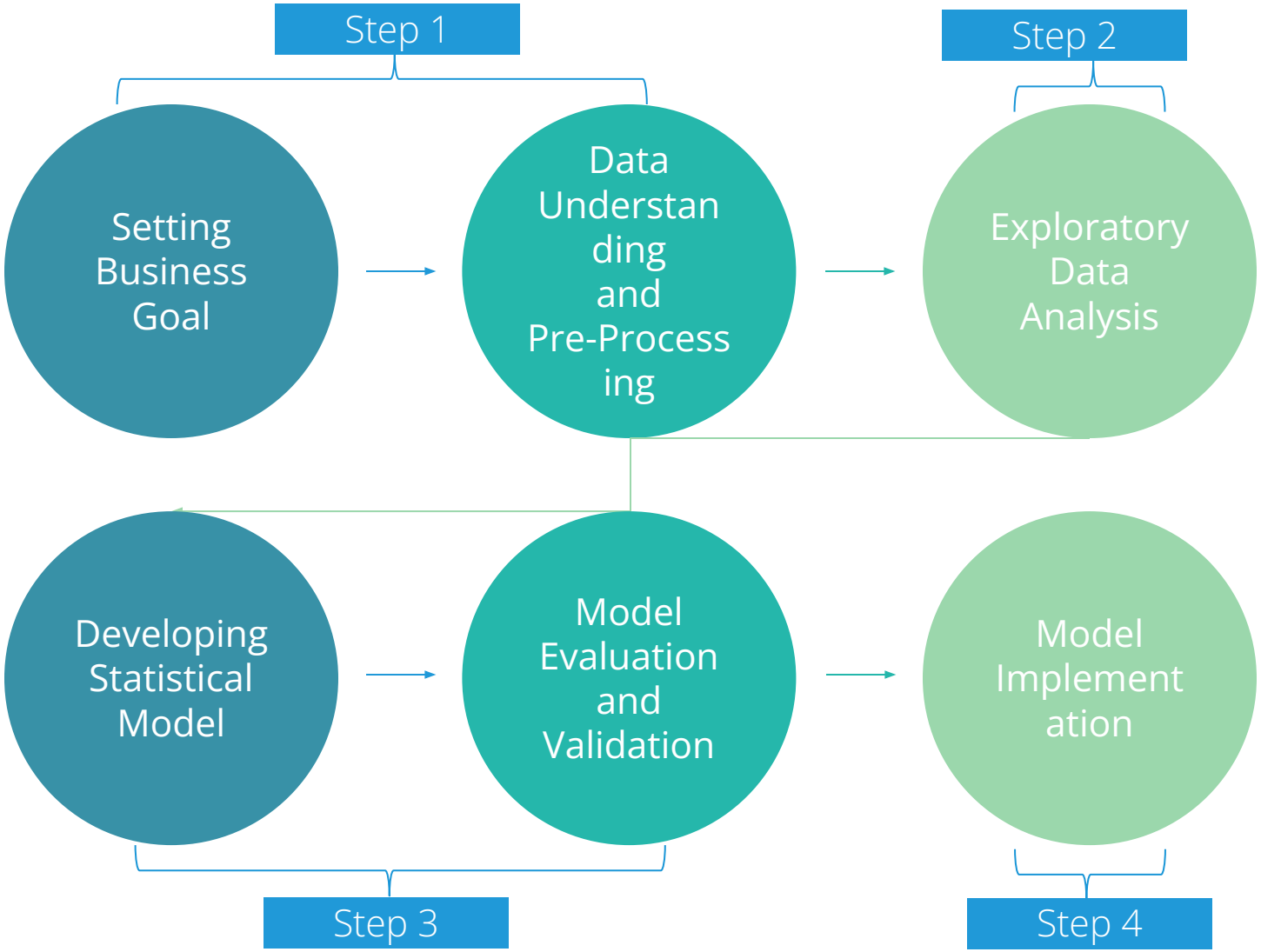


Predictive analytics is used in **financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals, sports and several other fields**

Predictive modelling – Important Techniques



Predictive modelling – General Approach



Step 1 – Data Understanding and Pre-processing

Data Understanding and Pre Processing

- Data Understanding
  - Understanding data dimension, variable types, variable relationships
- Converting raw data to usable data

  Data cleaning by checking for and handling:
  - Missing values
  - Inconsistencies
- Transforming variables
- Feature engineering
  - Using domain knowledge to create new features or variables which can be used in the model

Pre-Processing:

- Grouping / Factoring / Segmentation / Reduction

Step 2 – Exploratory Data Analysis

Exploratory Data Analysis

- **Performing exploratory data analysis** using:
  - Frequency tables
  - Cross tables
  - Descriptive statistics
  - Visualizations
  - Correlation matrix

Step 3 – Model Identification , Selection and Validation

Model Selection and Validation

- **Model identification and selection** is based on:
  - Study objective
  - Type of dependent variable
  - Checking different statistics / decision criteria based on models (Eg. p-value, R^2, AIC, etc.) i.e diagnostic checks
  - Automatic search procedures
- **Cross Validation**
  - Splitting data into training data and test data
  - Checking predictive ability of model on new data
  - Comparison of results with theoretical expectations, empirical results and simulation results

Step 4 – Model Implementation

Model Implementation

- **Drawing inferences from the model results** by :
  - Building equations using only the coefficients of significant variables
  - Mapping the model chosen with the existing system
- Fitting the model on new data and generating predictions

- **Observing values of Predictors :**
  - In a spreadsheet or
  - Web application or any other user interface or
  - Integrating the current systems

Get an Edge!

Any predictive model is developed on historical data. Sample size and data dimension are key determinants for a good model

- If sample size is too small, model may not give good insight about the relationship among the variables.
- Also, if the data has large number of variables (columns) but few observations (rows), we are essentially trying to learn too much from a small sample. Results from models developed using such data will be erratic. **The rule of thumb for appropriate sample dimension is that observations should be 10 times the number of variables.** For instance, if we wish to study the relationship of 8 variables, then we must have more than 80 observations.

Quick Recap

In this session, we gained knowledge on the concept of **predictive modelling :**

Predictive modelling	• Used to predict the outcome or the probability of an outcome • Models developed using historical data or purposely collected data
Important Statistical models	• Regression Models- Linear and non-linear • Time Series Models
General Approach in Predictive modelling	• Data understanding and pre-processing • Exploratory data analysis • Model selection and validation • Model Implementation