Data Analysis Technique

Unit - 3

Feature Generation

- Feature Generation (also known as feature construction or feature engineering) is the process of transforming features into new features that better relate to the target.
- This can involve mapping a feature into a new feature using a function like log, or creating a new feature from one or multiple features using multiplication or addition.
- Feature Generation can improve model performance when there is a feature interaction.
- Two or more features interact if the combined effect is (greater or less) than the sum of their individual effects.

https://www.turintech.ai/feature-generation-what-it-is-and-how-to-do-it/

Feature Generation (contd.)

- It is possible to make interactions with three or more features, but this tends to result in diminishing returns.
- Feature Generation is often overlooked as it is assumed that the model will learn any relevant relationships between features to predict the target variable.
- However, the generation of new flexible features is important as it allows us to use less complex models that are faster to run and easier to understand and maintain.

Feature Selection

- Feature selection, one of the main components of feature engineering, is the process of selecting the most important features to input in machine learning algorithms.
- Feature selection chooses optimal set of features from all available.

Why do we need feature selection?

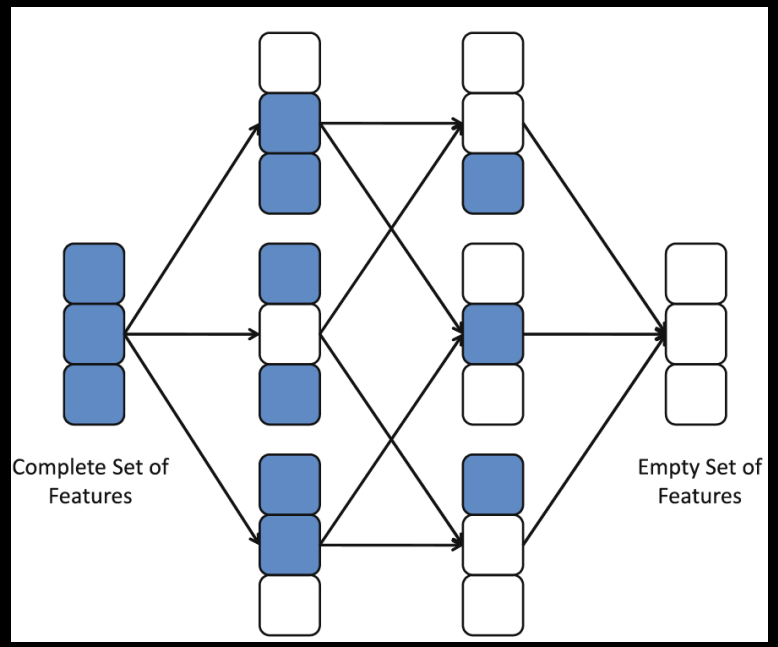
- Remove irrelevant data/noise
- Increase predictive accuracy of learning model
- Reduce Dimensionality
- Increase learning efficiency / reduce training time
- Reduce the complexity of resulting model by reducing the dimensionality of data as well as model..

Perspectives of Feature Selection

- Search for the best subset of features
- Criteria for evaluating different subsets

Perspectives of Feature Selection

Search the best subset of features



- Feature Selection can be considered as a search problem, where each state of the search space corresponds to a concrete subset of features selected.
- The selection can be represented as a binary array, with each element corresponding to the value 1, if the feature is currently selected by the algorithm and 0, if it does not occur.
- There should be a total of 2M subsets where M is the number of features of a data set.

Perspectives of Feature Selection

Search the best subset of features

Search Directions

- **Sequential Forward Generation (SFG)**: It starts with an empty set of features S. As the search starts, features are added into S according to some criterion that distinguish the best feature from the others. S grows until it reaches a full set of original features. The stopping criteria can be a threshold for the number of relevant features m or simply the generation of all possible subsets in brute force mode.

- **Sequential Backward Generation (SBG)**: It starts with a full set of features and, iteratively, they are removed one at a time. Here, the criterion must point out the worst or least important feature. By the end, the subset is only composed of a unique feature, which is considered to be the most informative of the whole set. As in the previous case, different stopping criteria can be used.

Perspectives of Feature Selection

Search the best subset of features

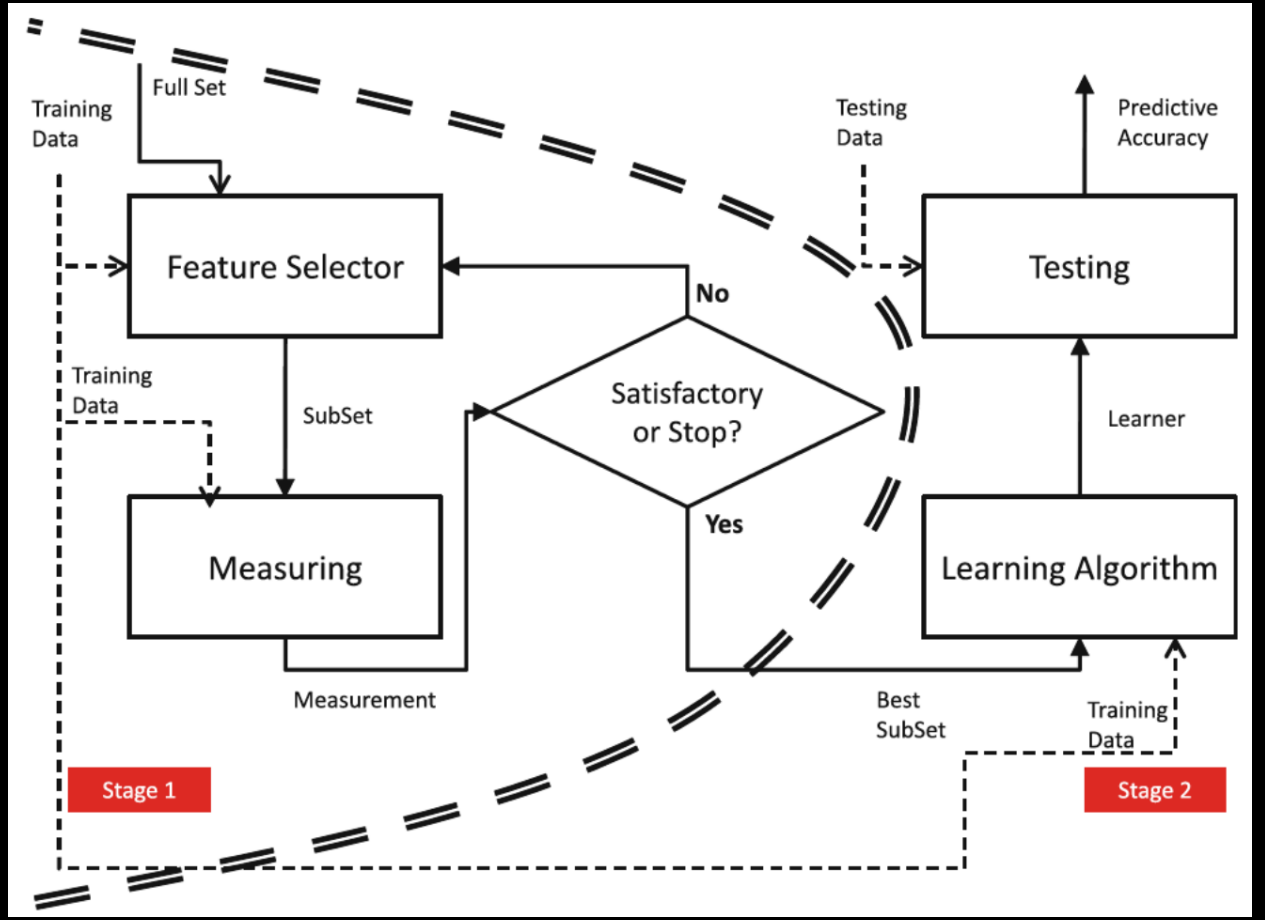
Search Directions

- **Bidirectional Generation (BG)**: Begins the search in both directions, performing SFG and SBG concurrently. They stop in two cases:
  1. when one search finds the best subset comprised of m features before it reaches the exact middle, or
  2. both searches achieve the middle of the search space. It takes advantage of both SFG and SBG.
- **Random Generation (RG)**: It starts the search in a random direction. The choice of adding or removing a features is a random decision. RG tries to avoid the stagnation into a local optima by not following a fixed way for subset generation. Unlike SFG or SBG, the size of the subset of features cannot be stipulated..

Perspectives of Feature Selection - Filters

- measuring uncertainty, distances, dependence or consistency is usually cheaper than measuring the accuracy of a learning process. Thus, filter methods are usually faster.
- it does not rely on a particular learning bias, in such a way that the selected features can be used to learn different models from different DM techniques.
- it can handle larger sized data, due to the simplicity and low time complexity of the evaluation measures.

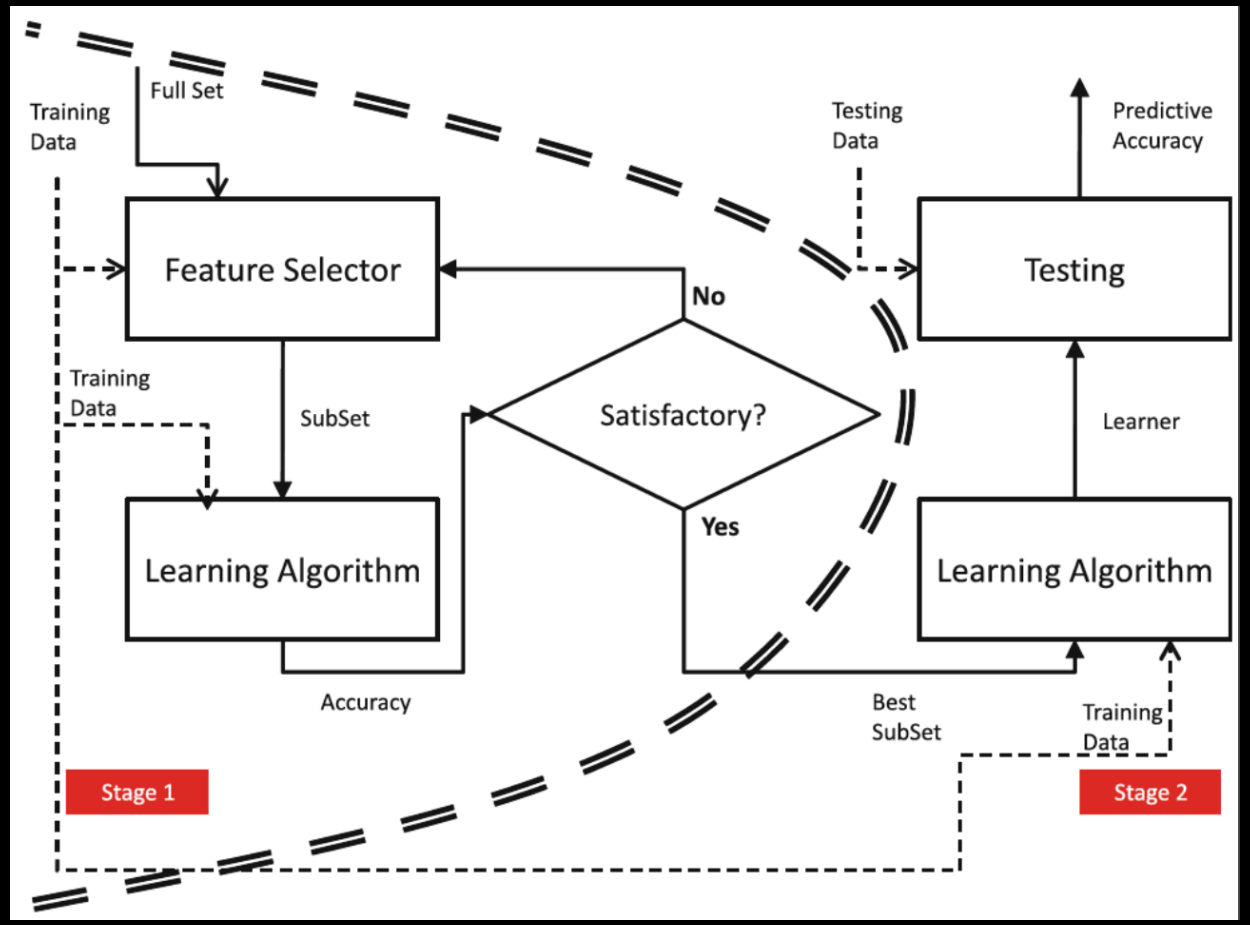
Perspectives of Feature Selection – Filters



Perspectives of Feature Selection - Wrappers

- can achieve the purpose of improving the particular learner's predictive performance.
- usage of internal statistical validation to control the overfitting, ensembles of learners and hybridizations with heuristic learning like Bayesian classifiers or Decision Tree induction.
- filter models cannot allow a learning algorithm to fully exploit its bias, whereas wrapper methods do.

Perspectives of Feature Selection – Wrappers



Predictive Analytics

- Predictive analytics is the practice of using data, statistical algorithms, and machine learning techniques to predict future outcomes or behavior.
- Predictive analytics is a type of data analytics that uses historical data to make predictions about future outcomes.
- It enables organizations to gain valuable insights, make informed decisions, and anticipate future trends.
- Predictive analytics can be used to identify risks and opportunities, make better decisions, and improve efficiency.
- Predictive analytics is often used in business, but it can also be used in other areas, such as healthcare, education, and government.

How does predictive analytics work?

- Predictive analytics uses a variety of techniques to analyze data, including:
  - Statistical modeling
  - Machine learning
  - Data mining
- These techniques are used to identify patterns in the data and to build models that can be used to make predictions.

Key Steps in Predictive Analytics

a) **Problem Definition:**
   - Clearly define the problem or objective you want to solve using predictive analytics.

   Identify the specific outcomes or behaviors you want to predict.

b) **Data Collection and Preparation:**
   - Gather relevant data from various sources, such as databases, APIs, or external datasets.
   - Clean, preprocess, and transform the data to ensure its quality and suitability for analysis.

c) **Feature Selection and Engineering:**
   - Select the most relevant features (variables) that have a significant impact on the outcome.
   - Create new features or transform existing ones to improve predictive power.

Key Steps in Predictive Analytics

d) Model Selection and Training:

- Choose an appropriate predictive model based on the nature of the problem and data.
- Split the data into training and testing sets.
- Train the model using the training data, adjusting model parameters as needed.

e) Model Evaluation and Validation:

- Evaluate the model's performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score).
- Validate the model by testing its performance on the unseen testing data.

f. Deployment and Monitoring:

- Deploy the predictive model into production or real-world applications.
- Continuously monitor the model's performance and update it if necessary.

Benefits of Predictive Analytics

- Predictive analytics can provide a number of benefits, including:
  - Improved decision-making
  - Increased efficiency
  - Reduced risk
  - Increased profits
  - Enhanced Customer Experience
- Predictive analytics can help businesses to:
  - Target marketing campaigns more effectively
  - Identify potential customers
  - Reduce fraud
  - Improve customer service

Applications of Predictive Analytics

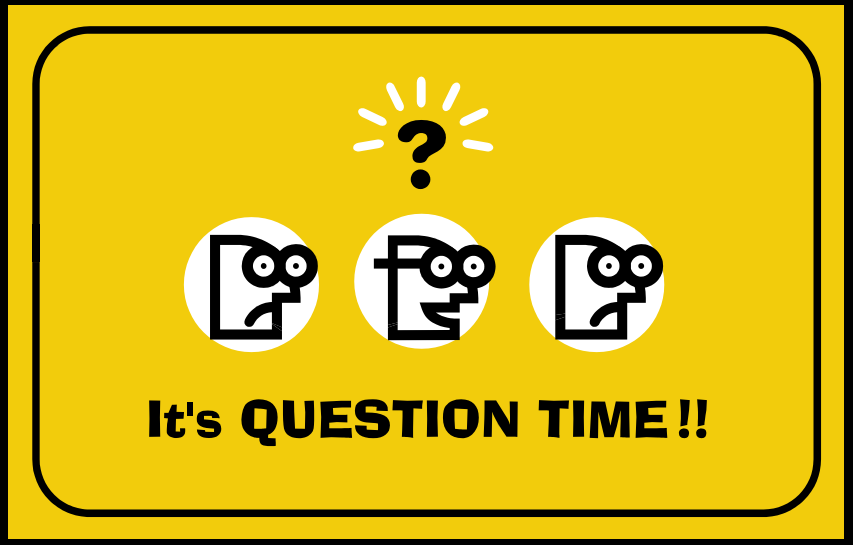
- Customer Churn Prediction: Forecasting customer attrition to take proactive retention measures.
- Sales Forecasting: Predicting future sales based on historical data and market trends.
- Fraud Detection: Identifying fraudulent activities or transactions in real-time.
- Demand Forecasting: Estimating future demand for products or services.
- Predictive Maintenance: Anticipating equipment failures and optimizing maintenance schedules.
- Risk Assessment: Assessing risks and making predictions in finance, insurance, and healthcare.

Time Series Analysis

Time series and Trend analysis

A time series consists of a set of observations measured at specified, usually equal, time interval.

Time series analysis attempts to identify those factors that exert influence on the values in the series.



Time series analysis is a basic tool for forecasting. Industry and government must forecast future activity to make decisions and plans to meet projected changes.

An analysis of the trend of the observations is needed to acquire an understanding of the progress of events leading to prevailing conditions.

The trend is defined as the long term underlying growth movement in a time series.

Accurate trend spotting can only be determined if the data are available for a sufficient length of time.

Forecasting does not produce definitive results. Forecasters can and do get things wrong from election results and football scores to the weather.

Time series examples

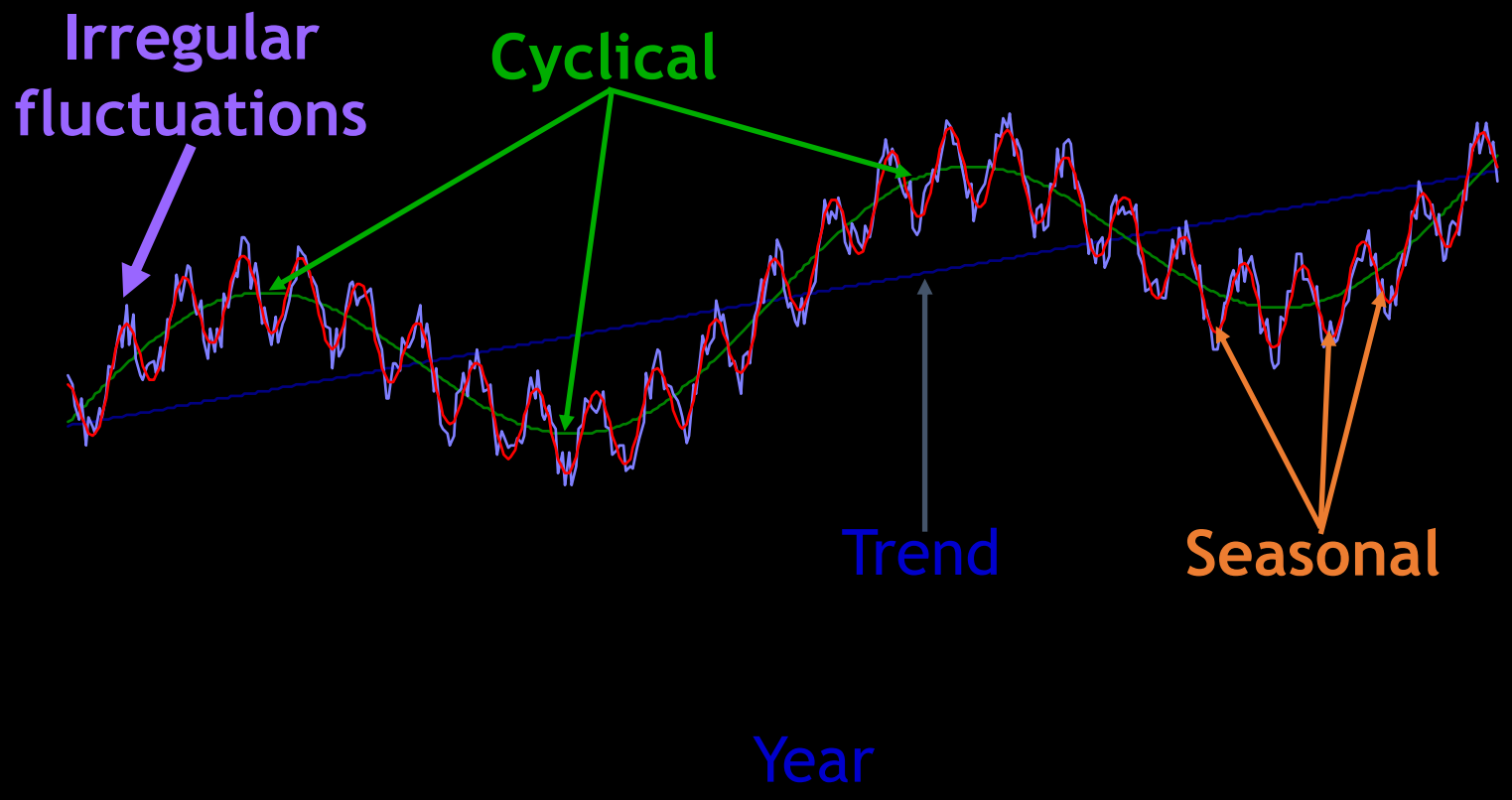
- Sales data
- Gross national product
- Share prices
- Exchange rate
- Unemployment rates
- Population
- Foreign debt
- Interest rates

Time series components

Time series data can be broken into these four components:

1. Secular trend
2. Seasonal variation
3. Cyclical variation
4. Irregular variation

Components of Time-Series Data



Predicting long term trends without smoothing?
What could go wrong?
Where do you commence your prediction from the bottom of a variation going up or the peak of a variation going down……

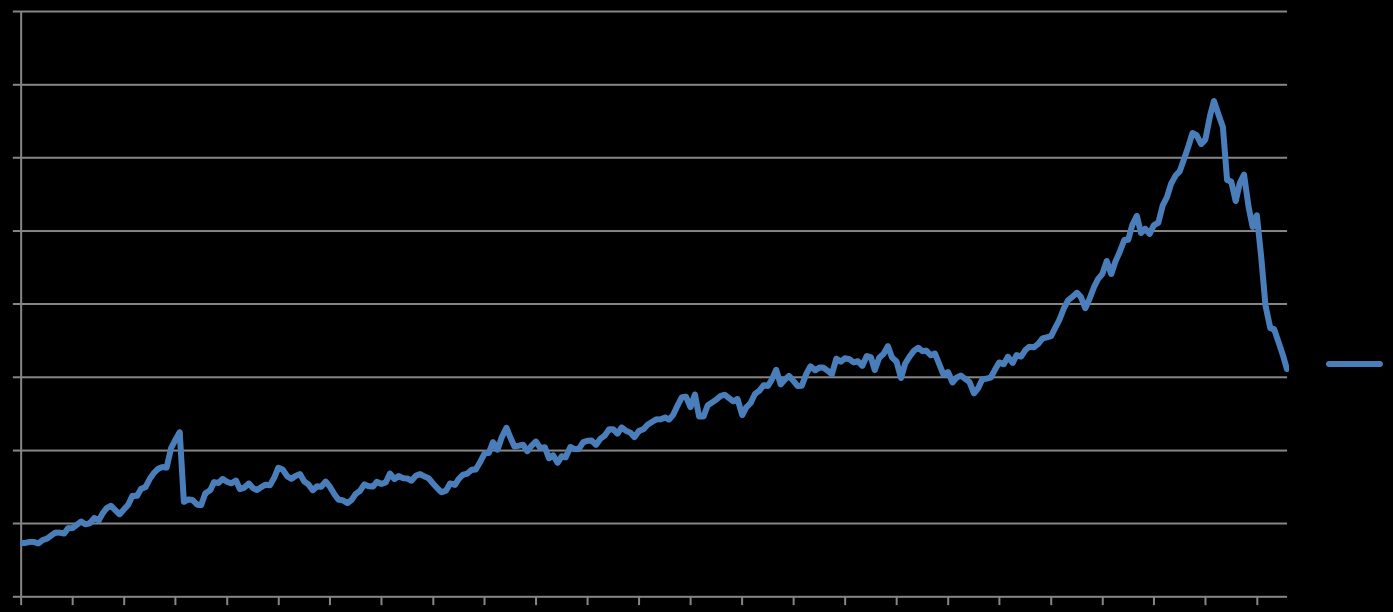
1. Secular Trend

This is the long term growth or decline of the series.

- In economic terms, long term may mean >10 years
- Describes the history of the time series
- Uses past trends to make prediction about the future
- Where the analyst can isolate the effect of a secular trend, changes due to other causes become clearer

Secular Trend

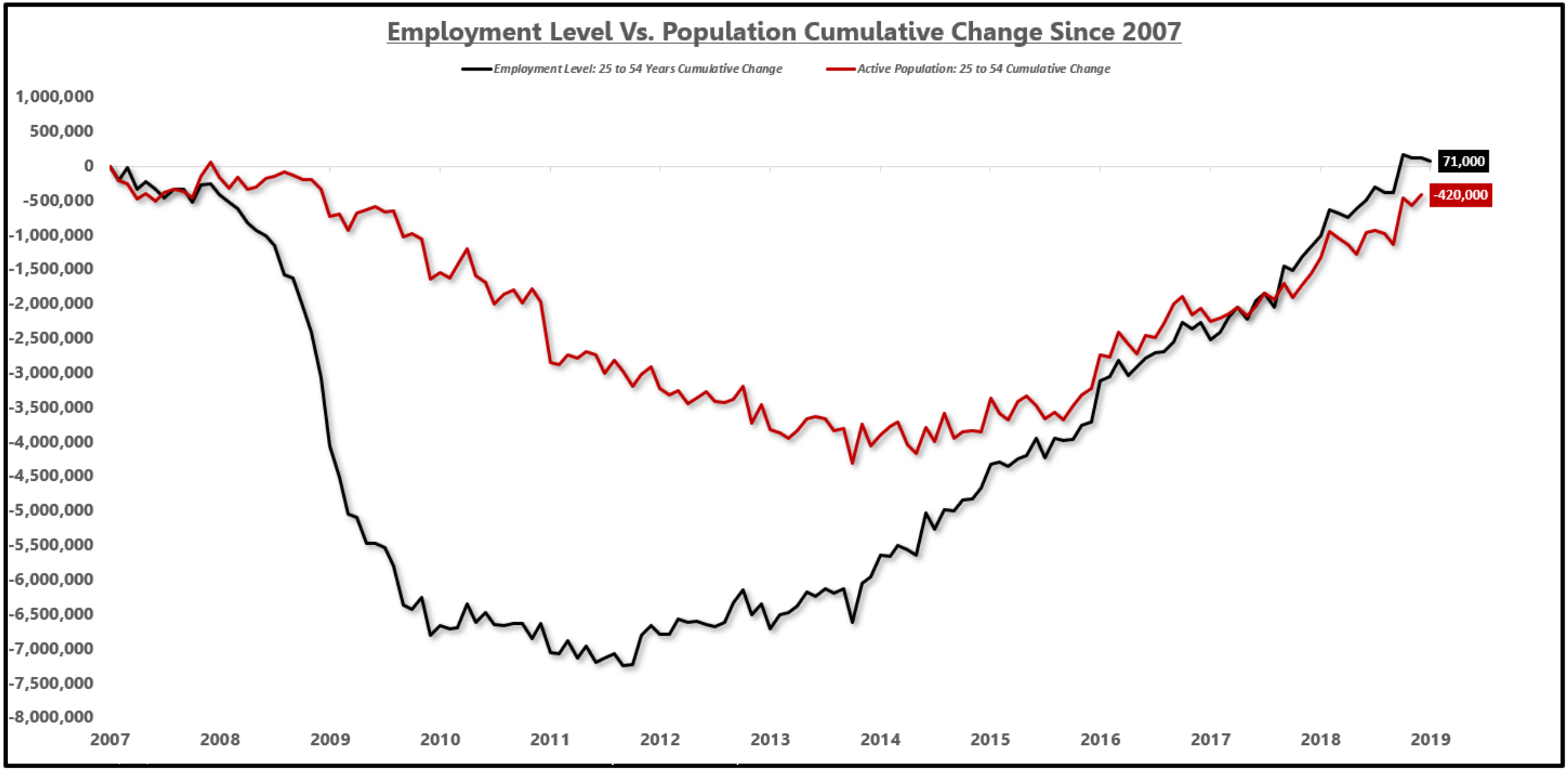
A secular trend identifies the underlying trend (direction) of the data: – increasing, decreasing or remaining constant. It is the long term direction of the data, usually described by the "line of best fit". And is deduced over a large number of periods. The following chart is a long term graph of the ASX200.



Look out

While trend estimates are often reliable, in some instances the usefulness of estimates is reduced by:

- by a high degree of irregularity in original or seasonally adjusted series or
- by abrupt change in the time series characteristics of the original data


Employment Level Vs. Population Cumulative Change Since 2007
Employment Level: 25 to 54 Years Cumulative Change
Active Population: 25 to 54 Cumulative Change
1,000,000
500,000
0
-500,000
-1,000,000
-1,500,000
-2,000,000
-2,500,000
-3,000,000
-3,500,000
-4,000,000
-4,500,000
-5,000,000
-5,500,000
-6,000,000
-6,500,000
-7,000,000
-7,500,000
-8,000,000
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
71,000
-420,000

2. Seasonal Variation

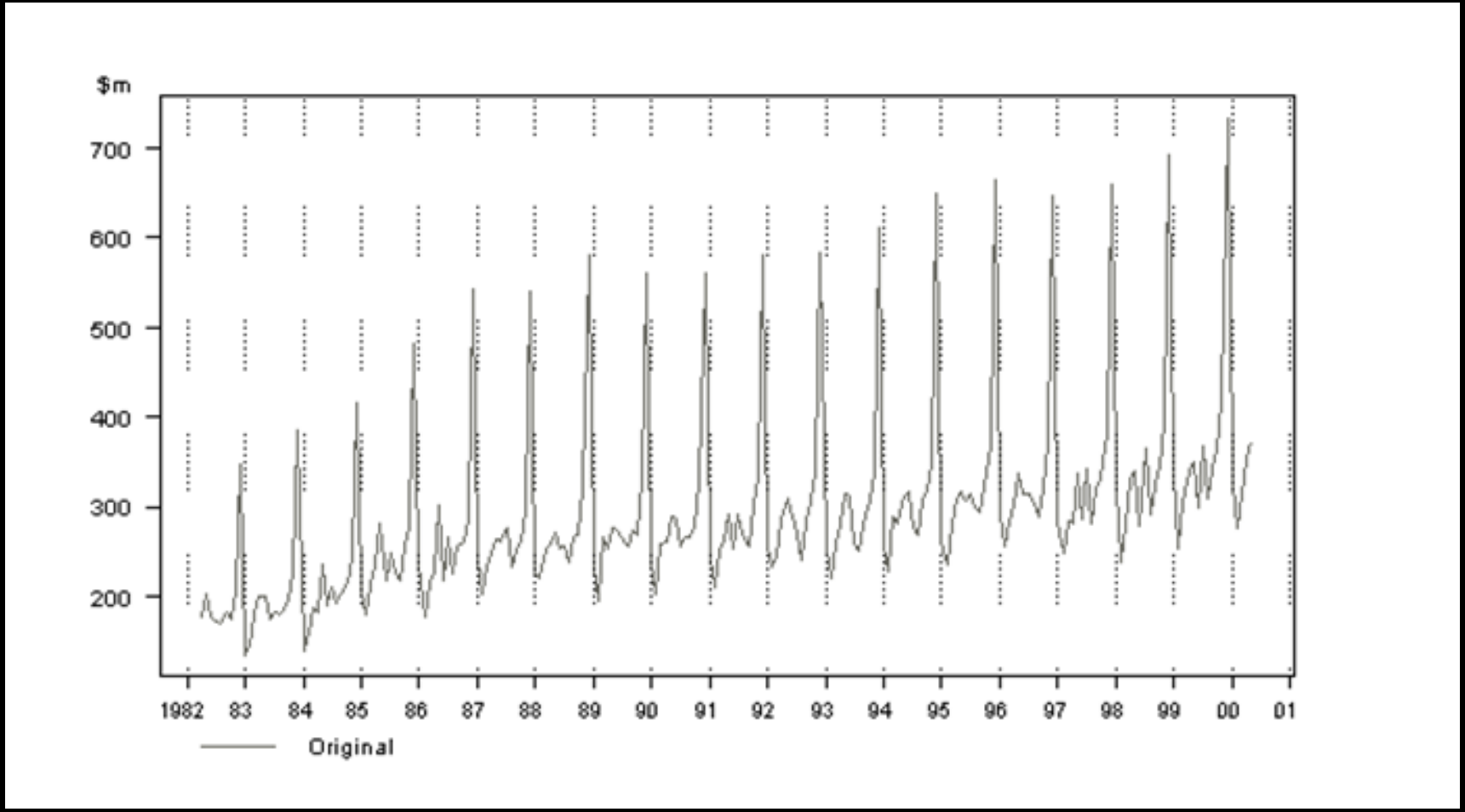
The seasonal variation of a time series is a pattern of change that **recurs** regularly over time.

Seasonal variations are usually due to the differences between seasons and to festive occasions such as Easter and Christmas.

Examples include:

- Air conditioner sales in Summer
- Heater sales in Winter
- Flu cases in Winter
- Airline tickets for flights during school vacations

Monthly Retail Sales in NSW Retail Department Stores



Seasonal Movement

Seasonal movement refers to regular periodic fluctuations that occur in each time period – yearly, monthly, daily. Some examples are speciality cards for Valentine's Day, monthly travel passes and off-peak heating.

Seasonal variations greatly impact on the outcomes of recorded data and often belie the underlying trend. Businesses need to identify the seasonal impact:

- So that a measurement (index) can be used to adjust the expected outcome.

- In order to recognise the direction of the underlying trend.

3. Cyclical variation

Cyclical variations also have recurring patterns but with a longer and more **erratic time scale** compared to Seasonal variations.

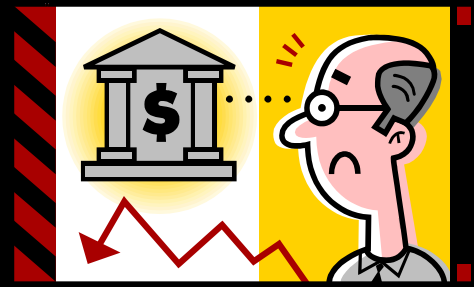
The name is quite misleading because these cycles can be far from regular and it is usually impossible to predict just how long periods of expansion or contraction will be.

There is no guarantee of a regularly returning pattern.

Cyclical variation

Example include:

- Floods
- Wars
- Changes in interest rates
- Economic depressions or recessions
- Changes in consumer spending

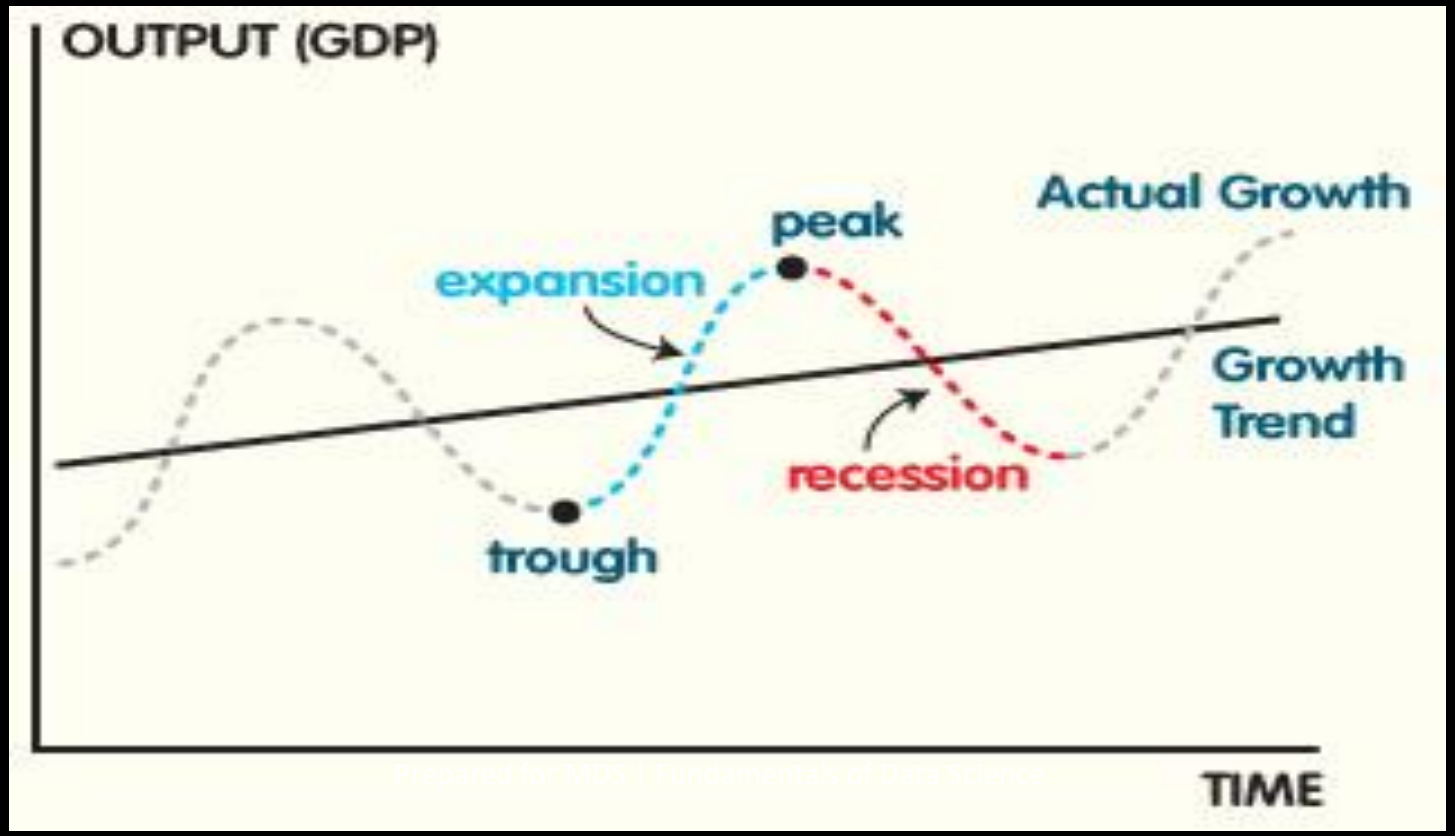


Cyclical Movement

This reflects the level of business activity and economic movement over time by fluctuating patterns, known as the economic cycle. These variations measure periods of expansion and contraction in industry and the economy. Their regularity and intensity are not predictable, however certain economic indicators contribute to their existence – level of investment, confidence in the economy, GDP, trade indexes and government policy.

Cyclical variation

This chart represents an economic cycle, but we know it doesn't always go like this. The timing and length of each phase is not predictable.



4. Irregular variation

An irregular (or random) variation in a time series occurs over varying (usually short) periods.

It follows no pattern and is by nature unpredictable.

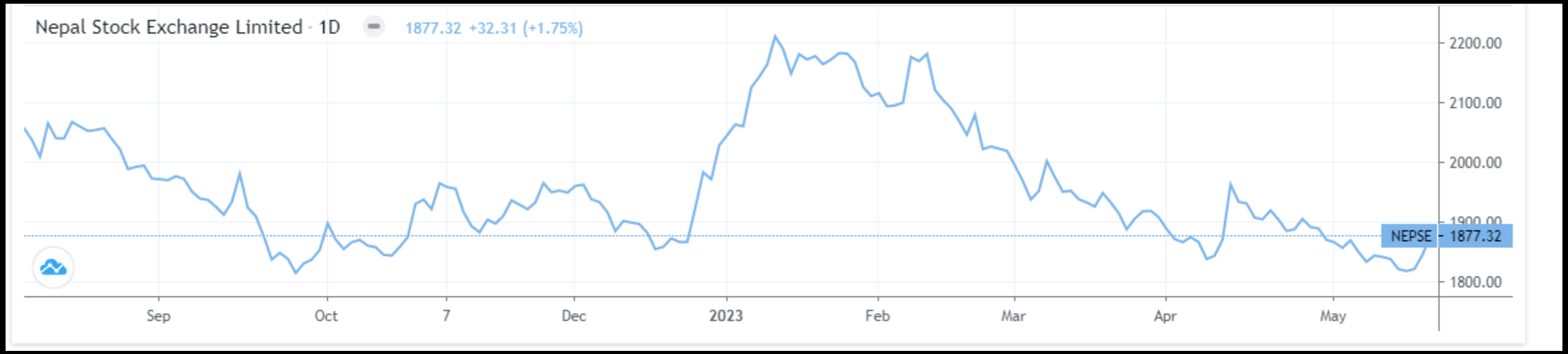
It usually occurs randomly and may be linked to events that also occur randomly.

Irregular variation cannot be explained mathematically.

Irregular variation

If the variation cannot be accounted for by secular trend, season or cyclical variation, then it is usually attributed to irregular variation. Example include:

- Sudden changes in interest rates
- Collapse of companies
- Natural disasters
- Sudden shift s in government policy
- Dramatic changes to the stock market
- Effect of Middle East unrest on petrol prices


Nepal Stock Exchange Limited · 1D
1877.32 +32.31 (+1.75%)
2200.00
2100.00
2000.00
1900.00
NEPSE
1877.32
1800.00
Sep
Oct
7
Dec
2023
Feb
Mar
Apr
May

Irregular Movements

These patterns refer to random variations that impact greatly on the level of business activity, often called natural variation. Some examples are extreme weather patterns (flood, fire, cyclone), extreme business variation (stock market crash, drop in $A), political climate (sudden elections, wars, death of a leader), and industry changes (pilot strikes, waterside strikes.). The resulting patterns will exert a great pressure on the predicted underlying trends and for this reason must be accounted for when planning for the future. However, the irregular movements are unpredictable.

Time Series Model - Examples

- Autoregressive (AR) model
- Moving average (MA) model
- Autoregressive moving average (ARMA) model
- Autoregressive integrated moving average (ARIMA) model
- Seasonal autoregressive integrated moving average (SARIMA) model
- Vector autoregressive (VAR) model
- Vector error correction (VECM) model

Autoregressive (AR) model

- Autoregressive (AR) models are defined as regression models in which the dependent or response variable is a linear function of past values of the dependent/response variable.
- The order of an autoregressive model is denoted as 'p', which represents the number of lags used to predict the current value.
- For example, if p=0, then it means that we are predicting the current time-step (t) based on the previous time-step (t-0).
- If p=n, then we are predicting time-step (t) based on n past time-steps.
- The general form of an autoregressive model can be represented as:

$$Y_t = c + \phi_p Y_{t-p} + \varepsilon_t$$

Moving average (MA) model

- A moving average (MA) is a type of model used for time-series forecasting.
- The moving average models are primarily used for stationary data, the data where we don’t see significant trends or seasonality.
- There are two different kinds of moving average model. They are Simple Moving Average (SMA) and Weighted Moving Average model.

Moving average (MA) model

- Simple Moving Average (SMA)
  - a type of moving Average model that uses a fixed number of data points for the averaging calculation.
  - Easy to calculate and can be implemented in wide in almost any language

$$\hat{y}_{t+1} = \frac{y_t + y_{t-1} + \ \ldots + y_{t-k+1}}{k}$$

- Weighted Moving Average (SMA)
  - a type of moving Average model that uses a weighting scheme to give more importance to more recent data points.
  - This type of MA can be used for time-series forecasting, and can help to reduce the impact of older data points on the average.

Time Series Analysis - Applications

- Time series analysis is used for non-stationary data—things that are constantly fluctuating over time or are affected by time.
- Industries like finance, retail, and economics frequently use time series analysis because currency and sales are always changing.
- Stock market analysis is an excellent example of time series analysis in action, especially with automated trading algorithms.
- Likewise, time series analysis is ideal for forecasting weather changes, helping meteorologists predict everything from tomorrow's weather report to future years of climate change.

Time Series Analysis - Applications

- Weather data
- Rainfall measurements
- Temperature readings
- Heart rate monitoring (EKG)
- Brain monitoring (EEG)
- Quarterly sales
- Stock prices
- Automated stock trading
- Industry forecasts
- Interest rates

End of the Chapter