JAMES COOK UNIVERSITY
AUSTRALIA

Week 6

MA5821 – Advanced Statistical Methods for Data Scientists

Dr Mostafa Shaikh
mostafa.shaikh@jcu.edu.au

online.jcu.edu.au

Cairns
Singapore
Townsville

Agenda and announcement

- Week 5 Quiz
- Time series
- Model Selection

Michael Cohen
Masters in AI and Software Engineering

Michael.Cohen@jcu.edu.au
Michael.Cohen@thinkplace.com.au

Announcement – SAS VA outages are being resolved. Check announcements for updated Due Date for the capstone project.

Reminder – Get started on writing your report for the capstone project.

Week 5 Quiz

Question 1 — 10 points

Which method of growing decision trees is the most informative and hands-on approach?

Choose at least one correct answer

- (A) Growing decision trees automatically
- (B) Growing decision trees autonomously
- (C) Growing decision trees interactively — *Correct answer*

Why is it the most informative?

Why is that important?

Why is it the most hands on?

Week 5 Quiz

Question 2 — 10 points

The best split for an input is a split that yields what?

Choose at least one correct answer

- (A) Maximal tree
- (B) Contingency table
- (C) Depth adjustment
- (D) The highest logworth — *Correct answer*

The best split for an input is the split that yields the highest logworth

The logworth statistic is used for pruning or growing a tree. It is defined as the –log(*p*-value)

Typically, if the logworth is greater than 2, then the variable that is used in the branch is significant and should be included in the tree.

Week 5 Quiz

Question 3

10 points

Which of the following statements is **true** regarding model assessment?

Choose at least one correct answer

- (A) Data splitting can be used only on data with continuous targets
- (B) The validation data set is used to calculate the parameter estimates and validate the model
- (C) Assessing the performance of the model on the data that you used to fit the model usually leads to an optimistically biased assessment *Correct answer*
- (D) Small differences in performance on the training data set versus the validation data set usually indicate overfitting

A. False. Data can be split on multiple data types.

B. False. If you are calculating parameters with it, it is no longer validation set

C. True.

D. False. Overfitting usually indicates a large difference between the training data set performance and the validation data set performance

Week 5 Quiz

Question 4

10 points

Decision Tree models use pruning to adjust model complexity and avoid the potential problem known as what?

Choose at least one correct answer

- (A) Overfitting — *Correct answer*
- (B) Accuracy
- (C) Concordance
- (D) Misclassification

The maximal tree represents the most complicated model you are willing to construct from a set of training data.

To avoid potential overfitting, many predictive modelling procedures offer some mechanism for adjusting model complexity.

For decision trees, this process is known as pruning

Remember the bias – variance trade-off

Week 5 Quiz

Question 5 — 10 points

Which of the following do SAS Visual Analytics stopping rules help to avoid?

Choose at least one correct answer

- A Logworth
- B Orphan nodes — *Correct answer*
- C Missing values
- D Probability

An orphan node is one with no parent.

Common parameters used in stopping rules include:

(a) the minimum number of records in a leaf;

(b) the minimum number of records in a node prior to splitting; and

(c) the depth (i. e. , number of steps) of any leaf from the root node. SAS Visual Analytics stopping rules help to control growth

Week 6 Warm up quiz

Question 2 2/5

The training data should always be larger than the validation data.

- ○ True
- ○ False

Submit

Why would it be true?

Why would it be false?

Time Series

Forecasting Horizons

- Long Term
  - 5+ years into the future
  - R&D, plant location, product planning
  - Principally judgement-based
- Medium Term
  - 1 season to 2 years
  - Aggregate planning, capacity planning, sales forecasts
  - Mixture of quantitative methods and judgement
- Short Term
  - 1 day to 1 year, less than 1 season
  - Demand forecasting, staffing levels, purchasing, inventory levels
  - Quantitative methods

Introduction to time series

- Regression analysis useful in short-term forecasting, but flawed
- A better approach: base the forecast of a variable on its own history
  - Avoids need to specify a causal relationship and to predict the values of explanatory variables
- *time series* methods for inferencing and forecasting

Why time series data are different from other data

- Data are not independent
  - Much of the statistical theory relies on the data being independent and identically distributed
- Large samples sizes are good, but long time series are not always the best
  - Series often change with time, so bigger isn't always better
- Time series vs Stochastic process:
  - A time series data is a collection of time-value data-point pairs.
  - A stochastic process is a mathematical model or a mathematical description of a distribution of time series.
  - A stochastic process is a model to generate a time series
- Time series represent a stationary model (ie, mean, variance and autocorrelation structure do not change over time)
  - First order differencing: computing the difference between consecutive observations
  - Seasonal differencing, e.g., for monthly data, computing the difference between an observation and the observation 12 time periods ago
  - Transformations: e.g., taking the log transformation of the series

Time series: Definitions, Applications and Techniques

JAMES COOK UNIVERSITY AUSTRALIA

An ordered sequence of values of a variable at equally spaced time intervals.

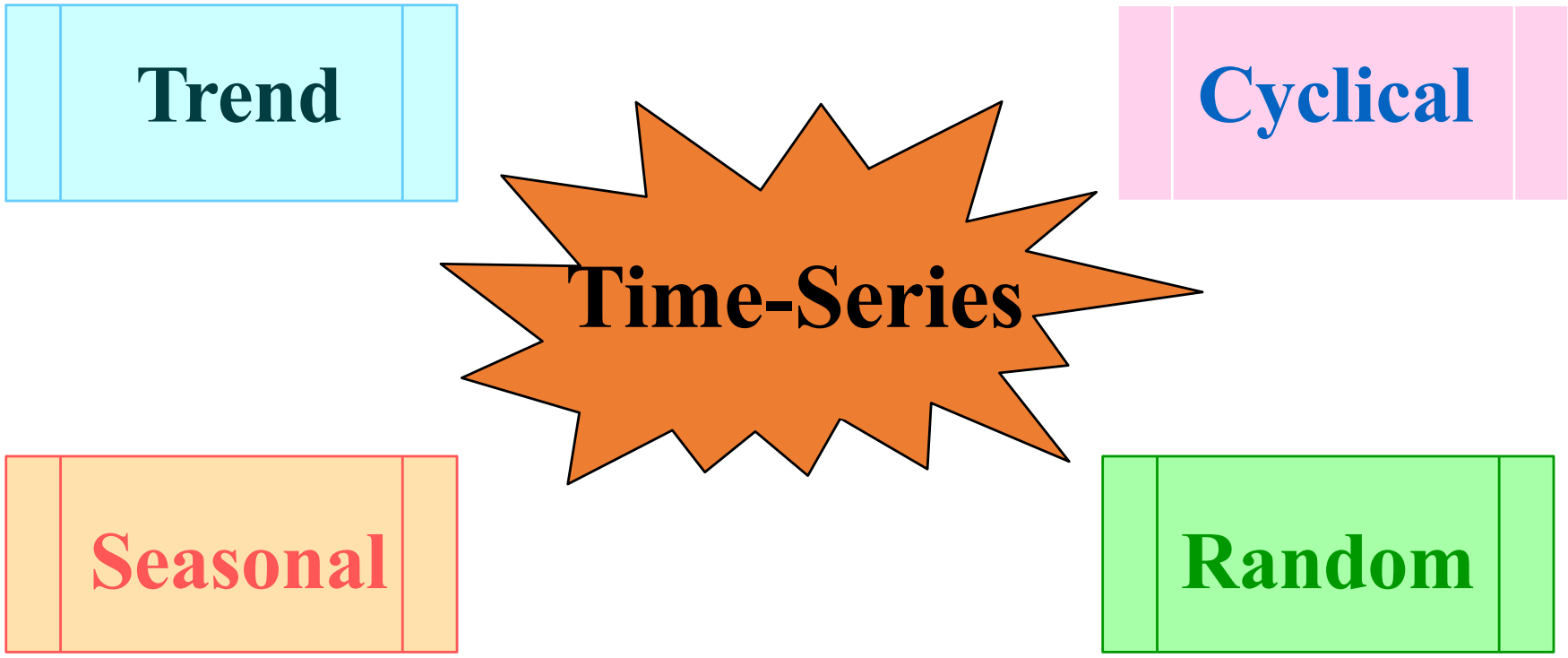
The usage is twofold or two approaches:

- Obtain an understanding of the underlying forces and structure that produced the observed data (**Inference based: What happened in the past?)**
- Fit a model and proceed to forecasting, monitoring or even feedback and feedforward control (**Forecasting based: What is likely to happen in the future?)**

Techniques:

- ARIMA (Autoregressive Integrated Moving Average)
- Damped trend exponential smoothing
- Linear exponential smoothing
- Seasonal exponential smoothing
- Simple exponential smoothing
- Winters method (additive)
- Winters method (multiplicative)

Time-Series Components

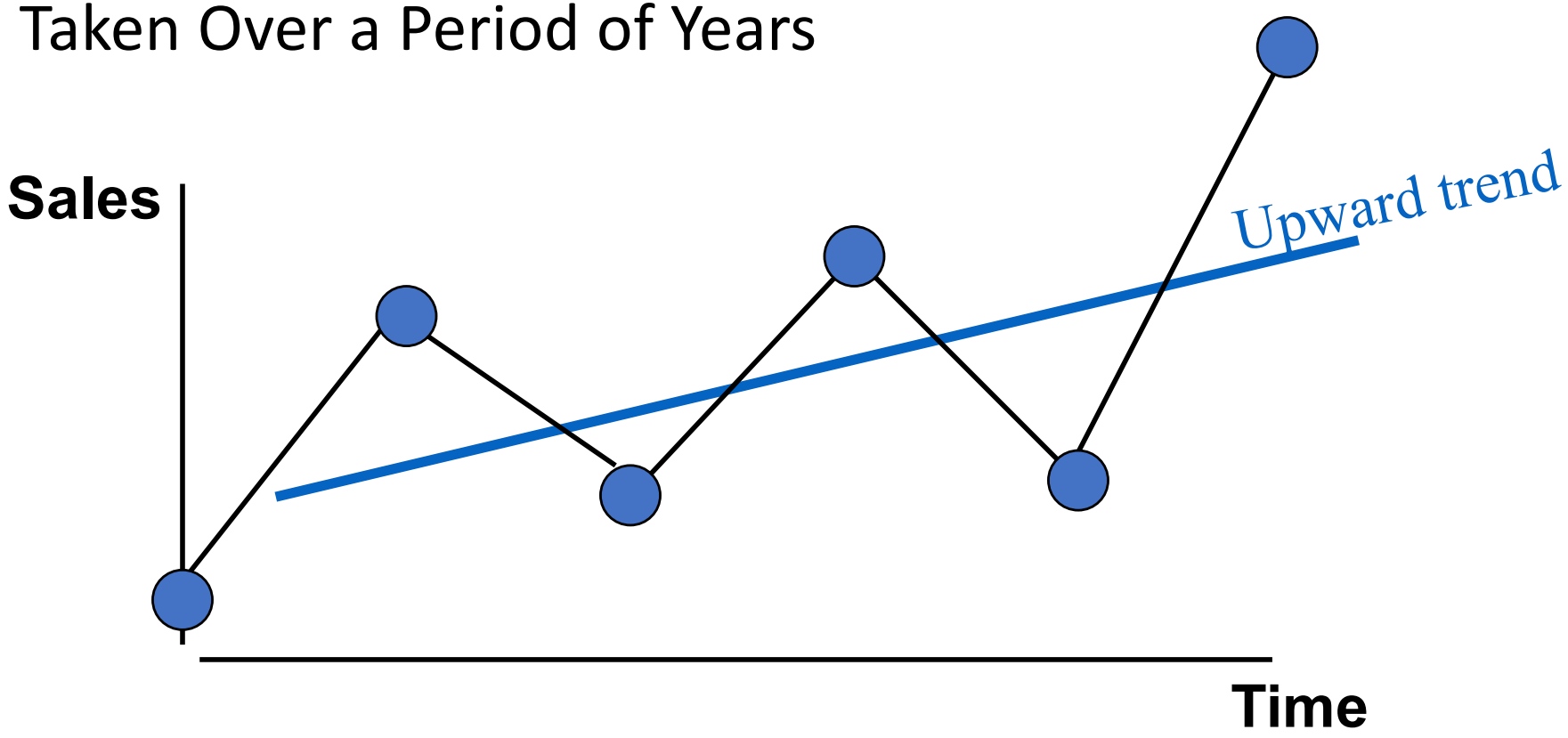


Trend (T_t)

- Trend: the long-term patterns or movements in the data.
- Overall or persistent, long-term upward or downward pattern of movement.
- The trend of a time series is not always linear.

Trend Component

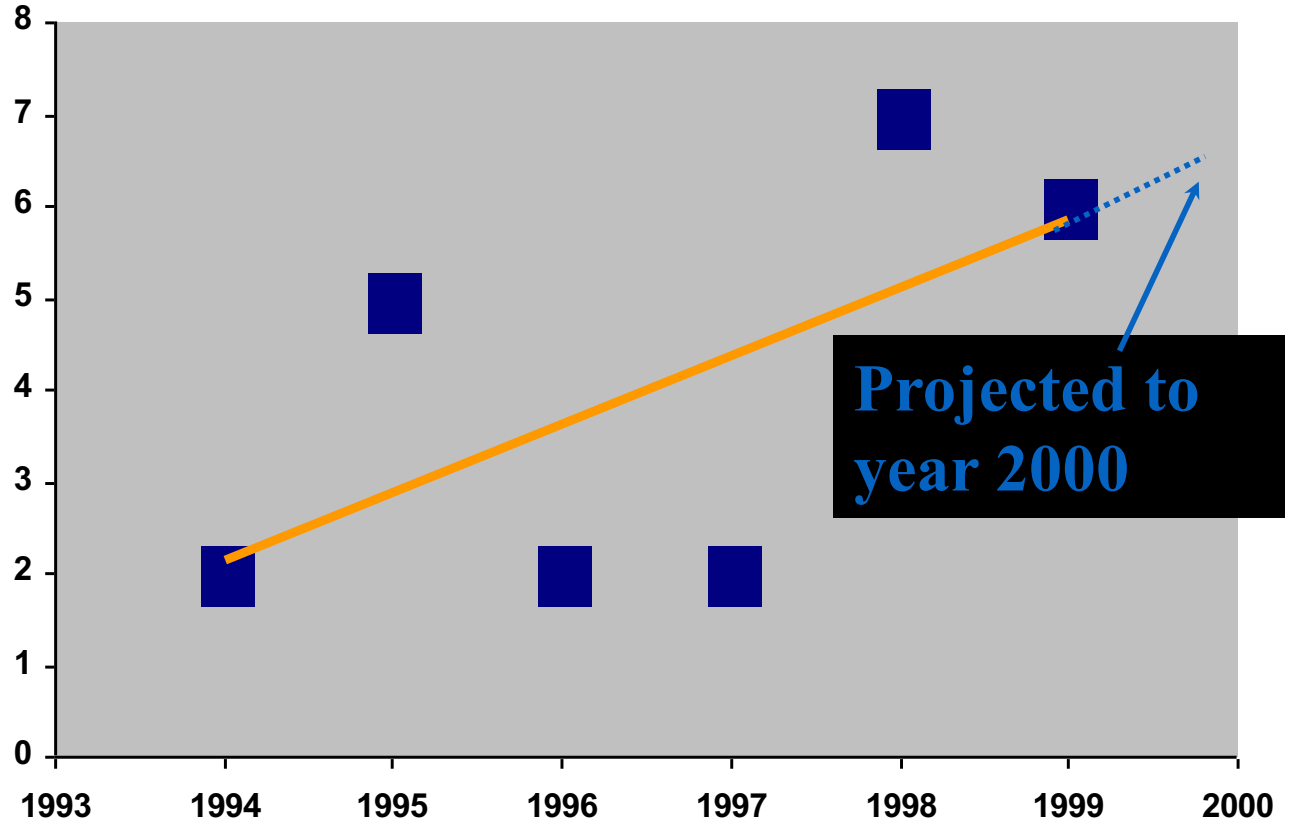
- Overall Upward or Downward Movement
- Data Taken Over a Period of Years



The Linear Trend Model

Year	Coded	Sales
94	0	2
95	1	5
96	2	2
97	3	2
98	4	7
99	5	6

$$\hat{Y}_i = b_0 + b_1 X_i = 2.143 + .743 X_i$$



Excel Output

	Coefficients
Intercept	2.14285714
X Variable	0.74285714

The Quadratic Trend Model

Year	Coded	Sales
94	0	2
95	1	5
96	2	2
97	3	2
98	4	7
99	5	6

$$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2$$

	Coefficients
Intercept	2.85714286
X Variable 1	-0.3285714
X Variable 2	0.21428571

Excel Output

$$\hat{Y}_i = 2.857 - 0.33 X_i + .214 X_i^2$$

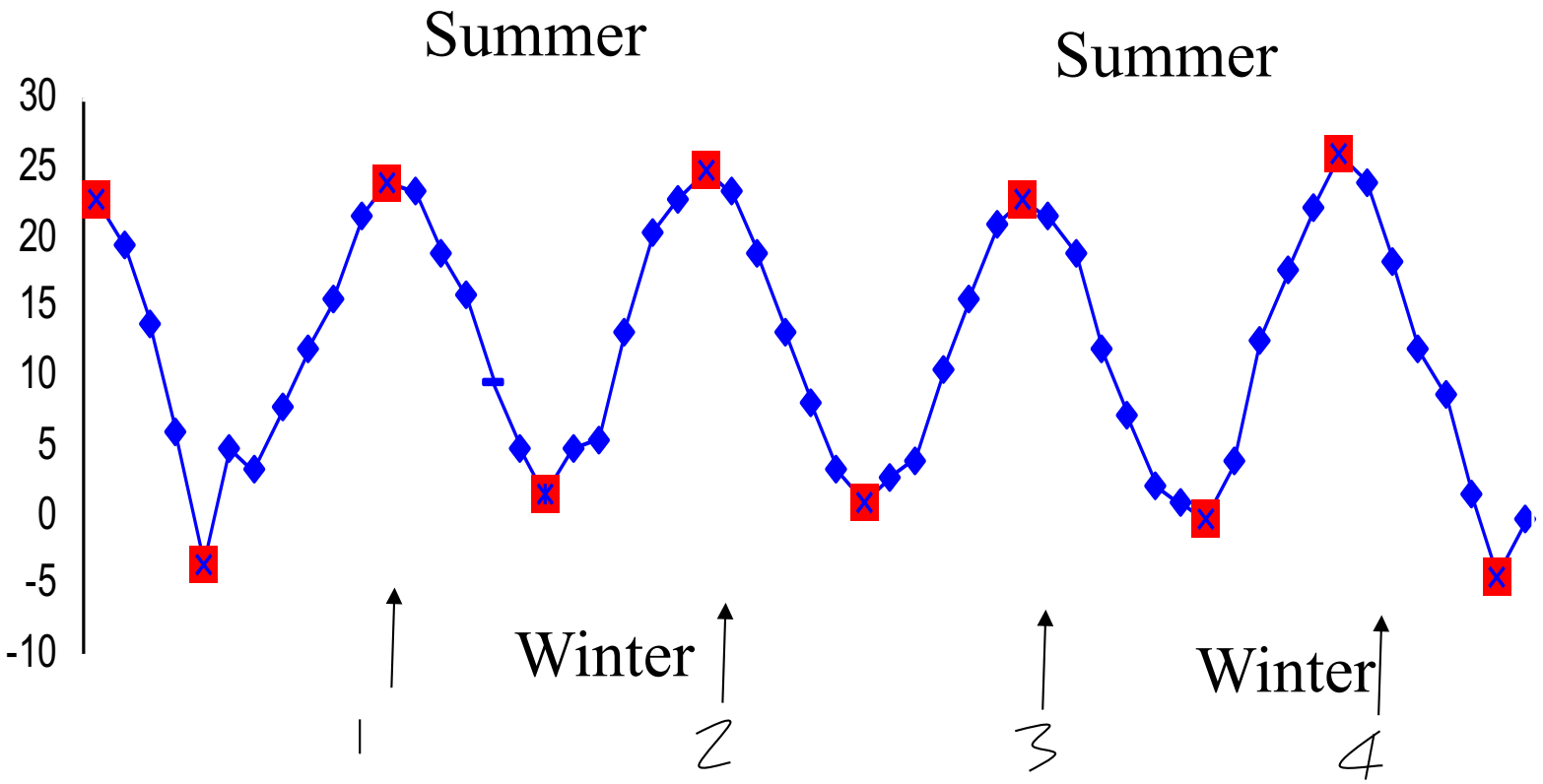
Seasonal variation (St)

- Regular periodic fluctuations that occur within year.

Examples:

- Consumption of heating oil, which is high in winter, and low in other seasons of year.
- Gasoline consumption, which is high in summer when most people go on vacation.

Seasonal variation (S_t)



Causes of Seasonal Effects

- Possible causes are
  - Natural factors
  - Administrative or legal measures
  - Social/cultural/religious traditions (e.g., fixed holidays, timing of vacations)

Cyclical variation (C_t)

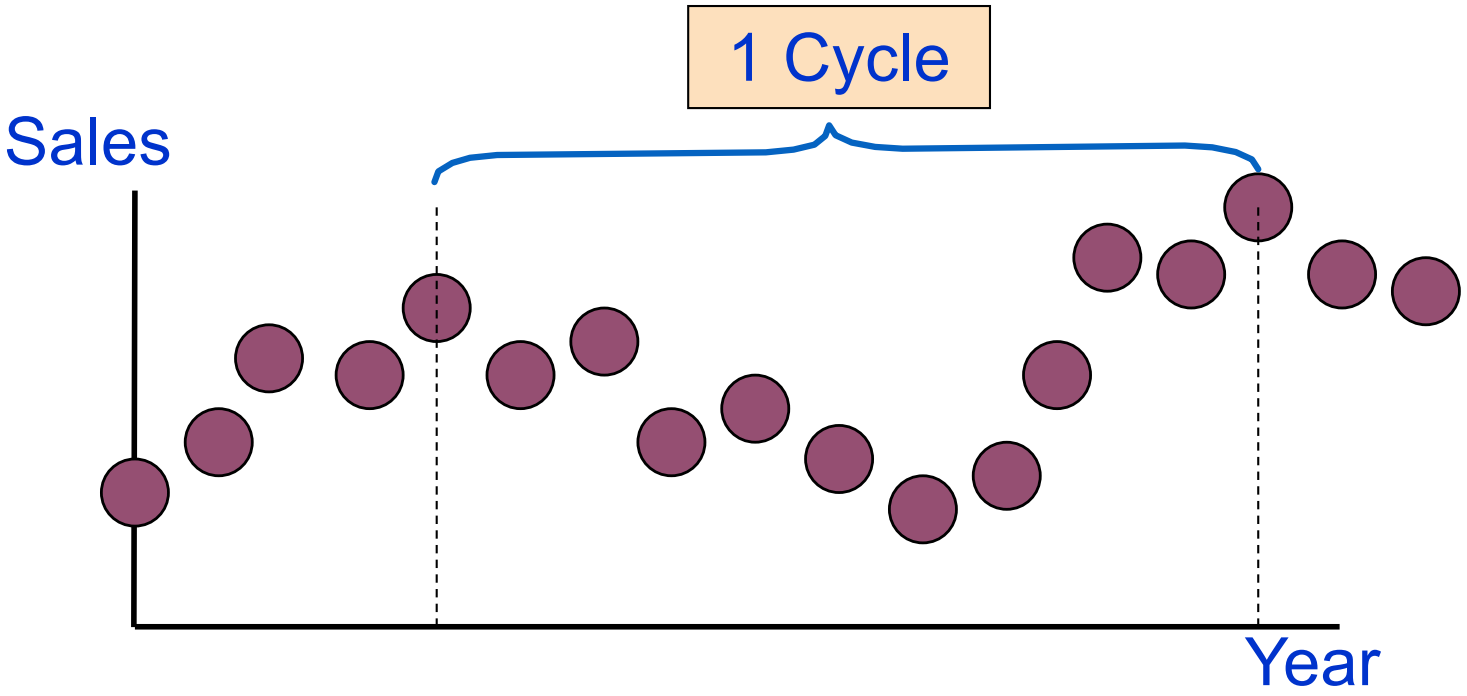
- Cyclical variations are similar to seasonal variations. Cycles are often irregular both in height of peak and duration.

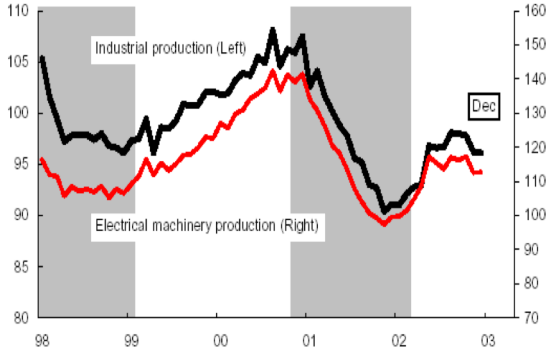
Examples:

- Long-term product demand cycles.
- Cycles in the monetary and financial sectors. (Important for economists!)

Cyclical Component

- Long-term wave-like patterns (upward or downward swings)
- Regularly occurs but may vary in length
- Often measured peak to peak or trough to trough
- Usually Lasts 2 - 10 Years





Irregular/Random Component

- Unpredictable, random, “residual” fluctuations
- Due to random variations of
  - Nature
  - Accidents or unusual events
  - Unseasonable weather
  - Sampling error
  - Non-sampling error
- “Noise” in the time series
- Short Duration and Non-repeating

Smoothing techniques

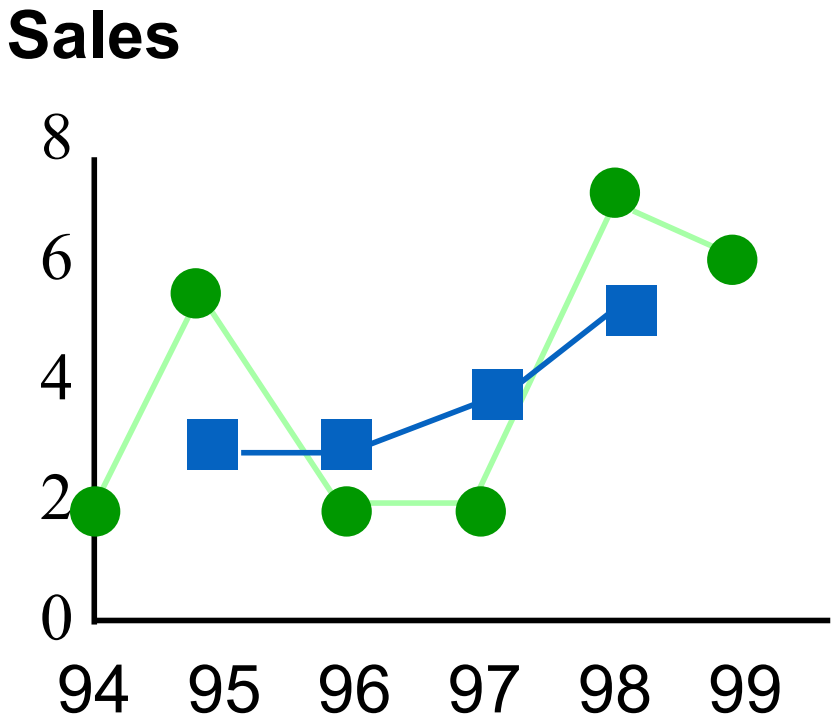
- Smoothing helps to see overall patterns in time series data.
- Smoothing techniques smooth or “iron” out variation to get the overall picture.
- There are several smoothing techniques of time series.
  - Moving average.
  - Exponential smoothing

Moving Averages

- Used for smoothing
- A series of arithmetic means over time
- Result dependent upon choice of L (length of period for computing means)
- Examples:
  - For a 3 year moving average, L = 3
  - For a 5 year moving average, L = 5
  - Etc.

Moving Average Example Solution

Year	Response ●	Moving Ave ■
1994	2	NA
1995	5	3
1996	2	3
1997	2	3.67
1998	7	5
1999	6	NA



The Exponential Smoothing Model

- **Exponential smoothing** weighs recent observations more than older ones.

$$S_t = \alpha x_t + (1 - \alpha) S_{t-1}$$

- Where α (the **smoothing constant)** is some number between zero and one.
- S_t is the **smoothed value** of the observations (our "best guess" as to the value of the mean)
- The forecast $F_{t+1} = S_t$
- Why is it called "Exponential"?
  - $S_t = \alpha y_{t-1} + (1-\alpha)[\alpha y_{t-2} + (1-\alpha)S_{t-2}] = \alpha y_{t-1} + \alpha(1-\alpha)y_{t-2} + (1-\alpha)^2 S_{t-2}$.

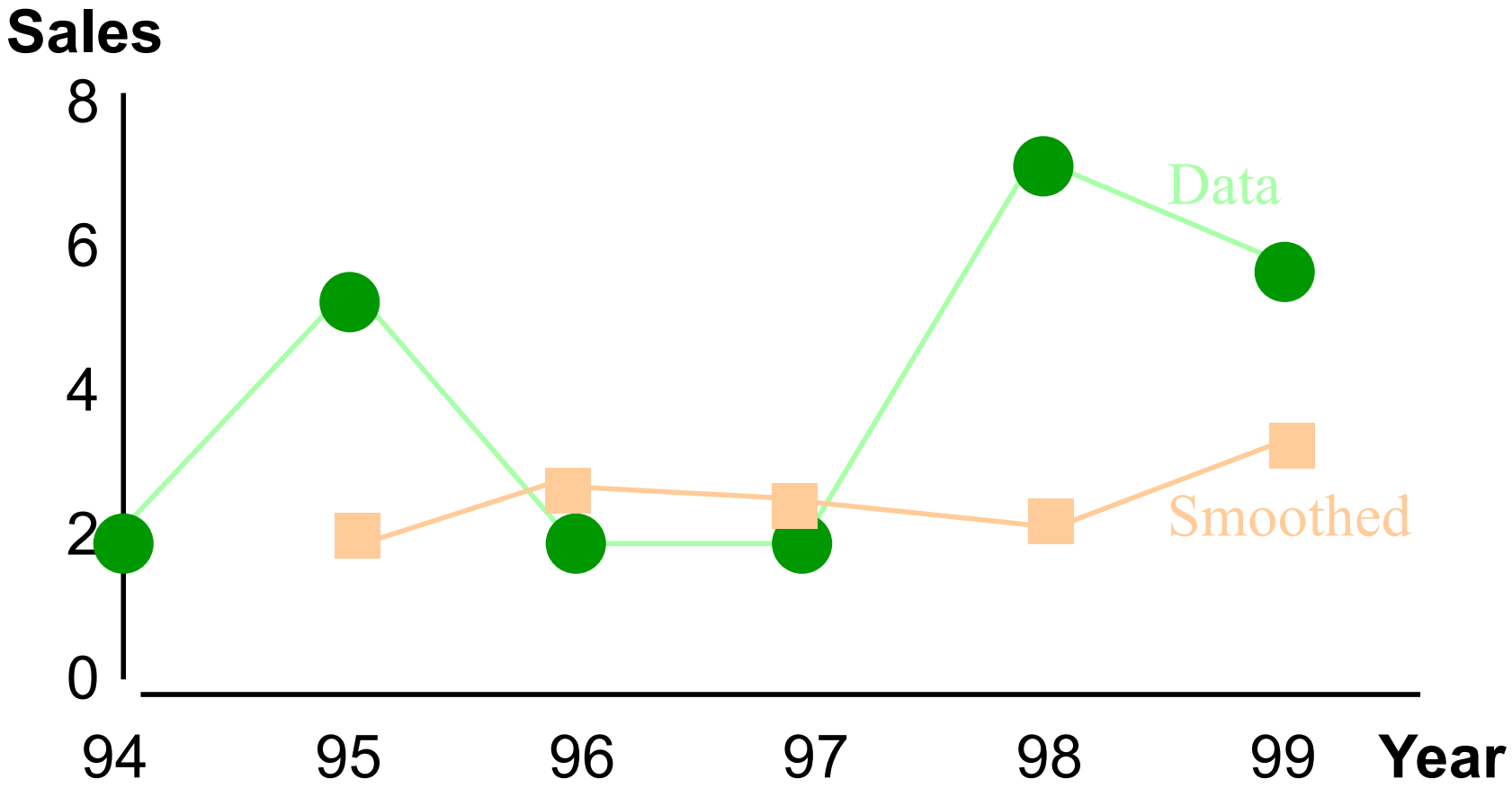
$$S_t = \alpha \sum_{i=1}^{t-2} (1-\alpha)^{i-1} y_{t-i} + (1-\alpha)^{t-2} S_2, \quad t \geq 2.$$

Exponential Weight: Example

$$E_i = WY_i + (1-W)E_{i-1}$$

Year	Response	Smoothing Value (W = .2)	Forecast
1994	2	2	NA
1995	5	(.2)(5) + (.8)(2) = 2.6	2
1996	2	(.2)(2) + (.8)(2.6) = 2.48	2.6
1997	2	(.2)(2) + (.8)(2.48) = 2.384	2.48
1998	7	(.2)(7) + (.8)(2.384) = 3.307	2.384
1999	6	(.2)(6) + (.8)(3.307) = 3.846	3.307

Exponential Weight: Example Graph



Holt, Holt-Winters method

$$S_t = \alpha y_{t-1} + (1-\alpha)S_{t-1} \quad 0 < \alpha \leq 1 \quad t \geq 3$$

(Holt, 1957, non-seasonal time series no trend)

$$S_t = \alpha y_t + (1-\alpha)(S_{t-1} + b_{t-1}) \quad 0 \leq \alpha \leq 1$$

$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1} \quad 0 \leq \gamma \leq 1$$

$$F_{t+1} = S_t + b_t$$

$$F_{t+m} = S_t + mb_t$$

$$b_1 = y_2 - y_1$$

$$b_1 = \frac{1}{3}[(y_2 - y_1) + (y_3 - y_2) + (y_4 - y_3)]$$

$$b_1 = \frac{y_n - y_1}{n-1}$$

(Holt, 1958, non-seasonal time series with trend)

$$S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) \quad \text{OVERALL SMOOTHING}$$

$$b_t = \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1} \quad \text{TREND SMOOTHING}$$

$$I_t = \beta \frac{y_t}{S_t} + (1-\beta)I_{t-L} \quad \text{SEASONAL SMOOTHING}$$

$$F_{t+m} = (S_t + mb_t)I_{t-L+m} \quad \text{FORECAST},$$

- y is the observation
- S is the smoothed observation
- b is the trend factor
- I is the seasonal index
- F is the forecast at m periods ahead
- t is an index denoting a time period

(Winters, 1965, with seasonality time series)

Exponential Smoothing with Trend and Cyclical Factors

- the exponential smoothing model with cyclical/ seasonality are two types:
  - *additive model (*an implicit assumption that the different components affected the time series additively)

$$\text{Data} = \text{Seasonal effect} + \text{Trend} + \text{Cyclical} + \text{Residual}$$

  - *multiplicative model (*seasonal and other effects act proportionally on the series is equivalent to a multiplicative model)

$$\text{Data} = (\text{Seasonal effect}) \times \text{Trend} \times \text{Cyclical} \times \text{Residual}$$

$$\begin{aligned}\log(\text{Data}) &= \log(\text{Seasonal effect} \times \text{Trend} \times \text{Cyclical} \times \text{Residual}) \\ &= \log(\text{Seasonal effect}) + \log(\text{Trend}) \\ &\quad + \log(\text{Cyclical}) + \log(\text{Residual})\end{aligned}$$

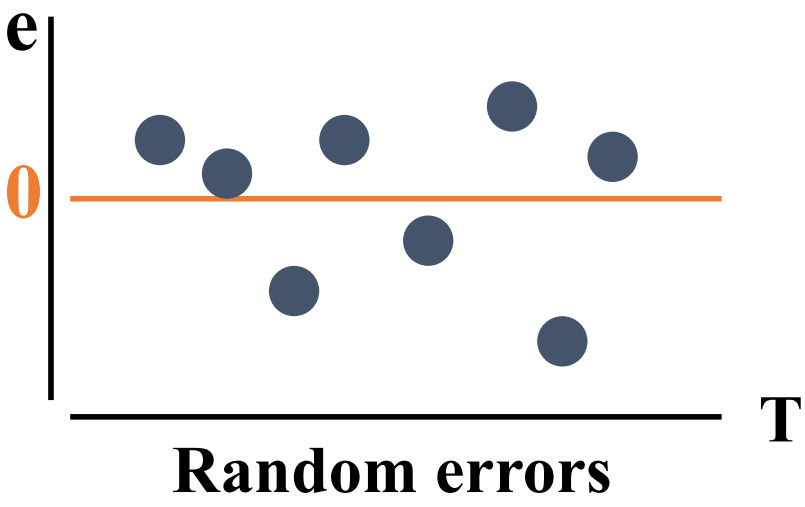
Autoregressive Modeling

- Used for Forecasting
- Takes Advantage of Autocorrelation
  - 1st order - correlation between consecutive values
  - 2nd order - correlation between values 2 periods apart
- Autoregressive Model for *pth* order:

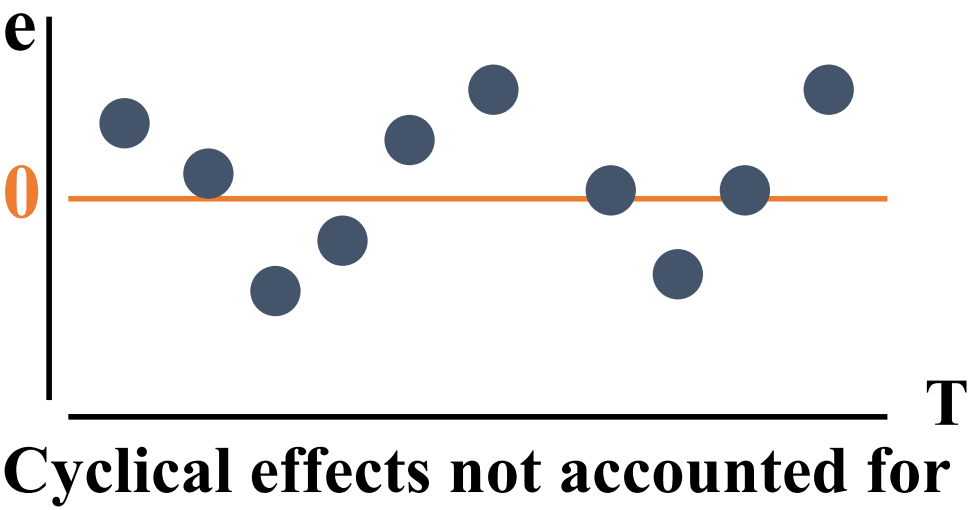
Random Error

$$Y_i = A_0 + A_1 Y_{i-1} + A_2 Y_{i-2} + \bullet\bullet\bullet + A_p Y_{i-p} + \delta_i$$

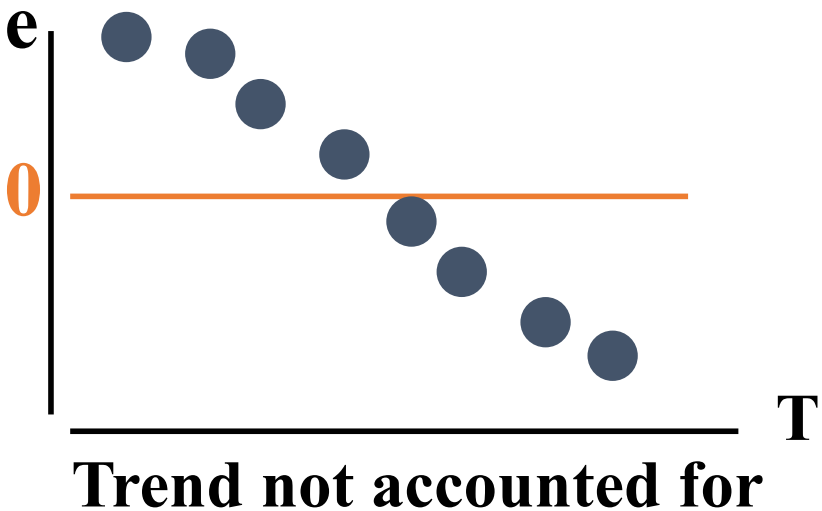
Residual Analysis



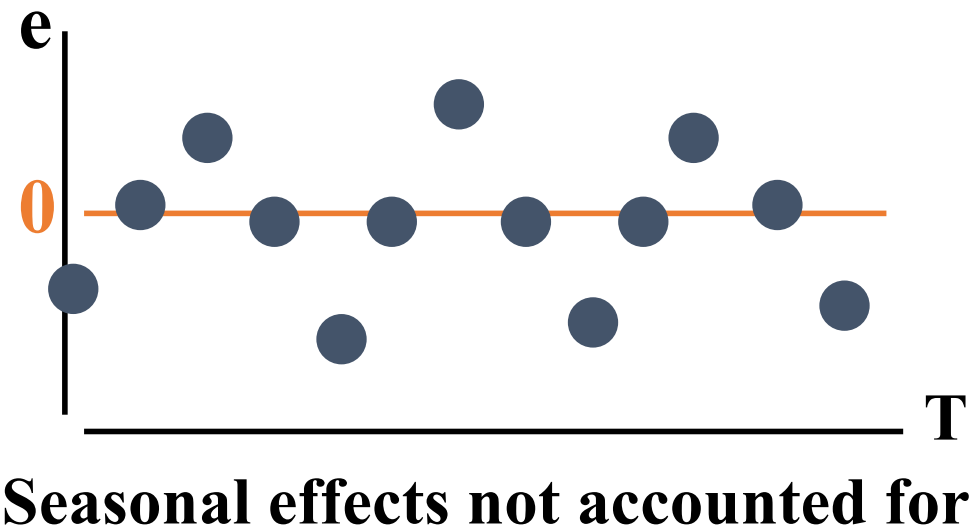
Random errors



Cyclical effects not accounted for



Trend not accounted for



Seasonal effects not accounted for

Measures of Forecast Accuracy

- MSE: the Mean Squared Error between forecast and actual
- MAD: the Mean Absolute Deviation between forecast and actual
- MAPE: the Mean Absolute Percent Error between forecast and actual

$$\text{MSE} = \frac{1}{(u-v+1)} \sum_{t=u}^{v} (F_t - x_t)^2$$

$$\text{MAD} = \frac{1}{(u-v+1)} \sum_{t=u}^{v} |F_t - x_t|$$

$$\text{MAPE} = \frac{1}{(u-v+1)} \sum_{t=u}^{v} \left| \frac{F_t - x_t}{x_t} \right|$$

Selecting a Forecasting Model

JAMES COOK UNIVERSITY AUSTRALIA

- Perform Residual Analysis
  - Look for pattern or direction
- Measure Sum Square Errors - SSE (residual errors)
- Measure Residual Errors Using MAD
- Use Simplest Model
  - Principle of Parsimony

Summary

- Moving averages and exponential smoothing are widely used for short-term forecasting
- By making projections from past data, these methods assume that the future will resemble the past
- the exponential smoothing procedure is sophisticated enough to permit representations of a linear trend and a cyclical factor in its calculations
- Exponential smoothing procedures are adaptive
- The smoothing factor should be chosen to trade off stability and responsiveness in an appropriate manner
- SAS community article: https://communities.sas.com/t5/SAS-Communities-Library/Forecasting-is-a-Snap-in-SAS-Visual-Analytics-8-2-on-SAS-Viya-3/ta-p/457167

Model Selection

Model Selection

Which of the following would be the best model for selection?

A. The simplest model with the best performance on the training data
B. The simplest model with the best performance on the validation data
C. The most complex model with the best performance on the training data
D. The most complex model with the performance on the validation data

The training data should always be larger than the validation data.

True
False

Which of the following statements are true?

A. Model management is a key part of good business analytics.
B. Models should be evaluated before, during and after deployment.
C. New models will always replace old ones as new observations come in.

Decisions Require high accuracy or low misclassification
Rankings Require high concordance or low discordance
Estimates Require low (average) squared error

Model Selection

focus on the validation fit statistics that are appropriate to the type of prediction that you are interested in

Prediction Type	Validation Fit Statistic	Direction
decision	misclassification	smallest
	average profit / loss	largest / smallest
	Kolmogorov-Smirnov statistic	largest
ranking	ROC index (concordance)	largest
	Gini coefficient	largest
estimate	average squared error	smallest
	Schwarz Bayesian Criterion	smallest
	log-likelihood	largest

Capstone report

JAMES COOK UNIVERSITY AUSTRALIA

- Title: concise and accurate reflection of the contents
- Abstract: sum(Background, Purpose, Method, Finding, Conclusion)
- Introduction: Scenario/Use cases for investigative efforts, Motivation, Background with reference, Key objectives/hypotheses, important issues you address
- Data: You know what is your input ….
- Method: How did you solve the problem using SAS VA
  - Data representation; Exploratory visualisation using SAS VA; Unstructured to Structured data; Data cleaning; Type conversion; Missing value imputation (informative missingness); Assessment criteria; Data subset selection; grouping and/or subsampling; Group-based data summarisation; Variable selection and/or transformation; Modelling and comparison in SAS VA
- Result & Discussion: What is your finding and why this is interesting
  - What are the main outcomes and why you expect that; Are they aligned with your interest you initially proposed in introduction, why or why not? How the result aligns with the goals? What is your main finding(s) and what does that imply to the audience?
- Conclusion: Conclusive remarks with assumption, future work
  - Final remarks on your findings; What makes this interesting/useful for now and for future work; why do you think your finding(s) support/reject original objective or hypothesis; any limitation/assumption of current work
- Writing: coherent; flow; spelling; grammar; references; tables and figures; formatting