

# QBUS6840 Predictive Analytics

## Lecture 1: Introduction

Discipline of Business Analytics  
University of Sydney Business School

# Table of contents

What is QBUS6840 about?

Admin stuff

Introduction to Forecasting

# Outline

What is QBUS6840 about?

Admin stuff

Introduction to Forecasting

# Business Analytics

- ▶ Businesses are increasing investment in business infrastructure, and collecting massive data everyday
- ▶ More and more companies, government organizations use and need business analytics
- ▶ Business analytics refers to skills, methodologies and technologies that help extracting useful knowledge from business data, which is important for data-driven decision making.

The good news: many opportunities for you

- ▶ Companies are creating positions for data analytics or data sciences in recent years
- ▶ See a recent survey from the Society of Human Resource Management (SHRM)

## PRESS RELEASES

## SHRM Survey: Data Analysis Skills in Demand for Jobs of the Future

Workforce planning will be critical as 59% of employers expect to increase number of jobs requiring data analysis skills in next 5 years

Nov 14, 2016

**ALEXANDRIA, Va.** — Start with a growing demand for workers who understand data analysis and an already tight job market. Add in the need for specialized data analysis skills.

The result: The Society for Human Resource Management (SHRM) reported today that the demand for workers with data analysis skills is expected to continue growing during the next five to 10 years.

SHRM's *Jobs of the Future: Data Analysis Skills* report, sponsored by the [American Statistical Association](#), found that 59 percent of U.S. organizations expect to increase the number of positions requiring data analysis skills in the next five years.

# SHRM Survey Key Findings

- ▶ 59% of organizations expect to increase the number of positions requiring data analysis skills over the next five years.
- ▶ 78% of organizations who had hired for data analysis positions in the last 12 months reported difficulty recruiting for these positions.
- ▶ You can do an analysis with data  
<https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Pages/data-analysis-skills.aspx>

# Key Finding Snapshots

## Definition

**Data analysis skills:** the ability to gather, analyze and draw practical conclusions from data, as well as communicate data findings to others.

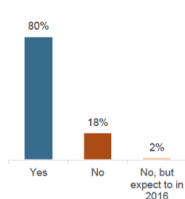
## Examples

- Data analyst
- Data scientist
- Statistician
- Market research analyst
- Financial analyst
- Research manager

## Prevalence of Positions Requiring Data Analysis Skills

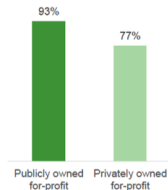
Four out of five organizations (80%) currently had positions that required data analysis skills, and another 2% expected to create such positions in 2016. Publicly owned for-profit organizations were more likely than privately owned for-profit organizations to have data analysis positions.

### Require Data Analysis Skills



n = 398

### Comparisons by Sector



*Note: Only statistically significant differences are shown. Dark bars are statistically larger than light bars.*

Jobs of the Future: Data Analysis Skills ©SHRM 2016

7

# Key Finding Snapshots

## HR People + Strategy Research: *Big Data*

- 53% of HR departments use big data to help make strategic decisions.

### How HR uses big data:

- 71% sourcing, recruitment or selection
- 63% identifying causes of turnover and/or employee retention strategies or trends
- 61% managing talent and performance

### Why organizations do not use big data:

- 51% lack of knowledge/expertise
- 30% not enough data collected/available

## Functional Areas for Positions Requiring Data Analysis Skills

The majority of organizations (71%) had data analysis positions within the accounting and finance department. At least one-half of organizations required data analysis skills for positions within human resources (54%) or business and administration (50%) units.



Note: n = 300. Percentages do not total 100% due to multiple response options.

Jobs of the Future: Data Analysis Skills ©SHRM 2016

12

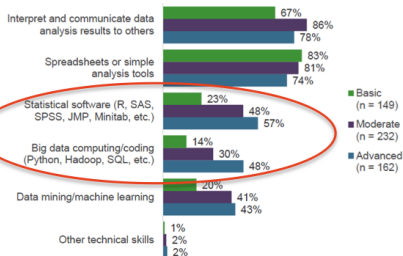


# Key Finding Snapshots

**60%**

of organizations had positions that require the ability to interpret and communicate data analysis results, but did not require gathering or analyzing data.

## Technical Skills Required or Preferred at Each Skill Level



Note: Percentages do not total 100% due to multiple response options.

Jobs of the Future: Data Analysis Skills ©SHRM 2016

22

# Key Finding Snapshots

## Comparisons by Organization Sector

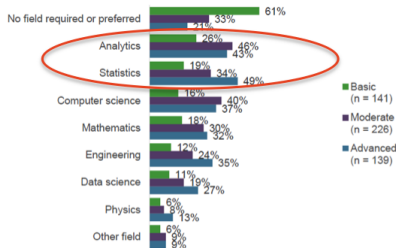
### Engineering at the Advanced Level

Privately owned for-profit organizations (53%) were more likely than nonprofit organizations (14%) to require or prefer a degree in engineering at the advanced level.

*Note: Only statistically significant differences are shown.*

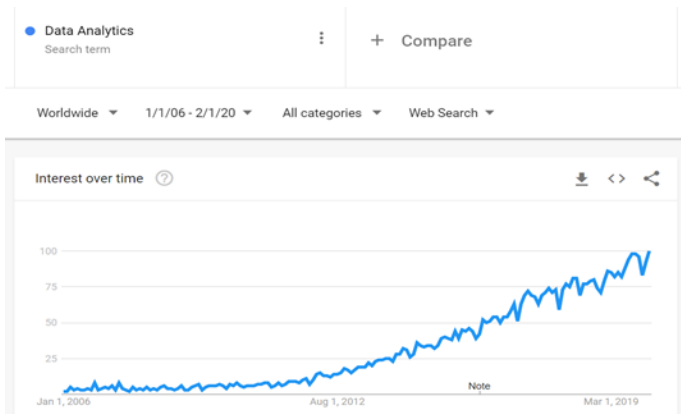
## Fields of Study Required or Preferred, by Skill Level

At the basic skill level, three out of five organizations (61%) did not require or prefer a particular field of study. For moderate level skilled jobs, many organizations were looking for a degree in analytics, computer science or statistics. Advanced level skilled positions required similar fields, but more organizations needed candidates who studied engineering, data science and physics.

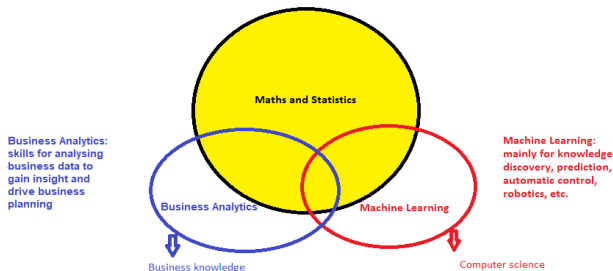


*Note: Percentages do not total 100% due to multiple response options.*

# Google Trend



# Business Analytics



- ▶ Business analytics includes a set of many data analysis techniques (regression, classification, clustering, artificial neural networks, etc), each designed for different type of data (cross-sectional data, time series data, text data ,etc).
- ▶ In this course, we focus on techniques for analysing time series data, and focus on forecasting

# Time series data

- ▶ A time series is a time stamped sequence of observations on a variable.
  - ▶ Weekly unit sales of a product.
  - ▶ Unemployment rate in Australia each quarter.
  - ▶ Daily production levels of a product.
  - ▶ Average annual temperature in Sydney.
  - ▶ Monthly water level in Warragamba Dam.
  - ▶ 5 minute prices for CBA stock on the ASX.
- ▶ Mathematically we denote a time series as, finite or **infinite** in time,

$$Y_0, Y_1, Y_2, \dots, Y_t, \dots, Y_n, Y_{n+1}, \dots$$

# Examples of questions we try to answer in this unit

- ▶ What is the underlying pattern in the yearly GDP time series?
- ▶ How can we forecast the electricity demand in NSW next year?
- ▶ Can we forecast the variation (called volatility) of the stock return of an asset tomorrow?
- ▶ Is the forecast produced by a colleague of yours sensible/accurate?
- ▶ etc.

## So, what is QBUS6840 about?

- ▶ offers a survey of main statistical methodologies for analysis of business time series data
- ▶ provides the tools necessary to extract information required for specific tasks such as forecasting, quantifying prediction uncertainty.
- ▶ Emphasis will be given to business applications of predictive analytics methods using modern software tools.
- ▶ Emphasis on state-of-the-art forecasting methods, recent research in this area

# Learning objectives

At the completion of the unit, students

- ▶ Understand the characteristics of time-series data in order to analyse real business data of this form
- ▶ Select and use an appropriate technique to predict the future behaviour of business variables of interest
- ▶ Be fluent in using computational tools to assist carrying out your analysis and generating visualisation
- ▶ Can identify advantages and limitations of each method
- ▶ Learn something about Python
- ▶ Can present and write about their findings effectively, communicate effectively with highly technical data scientists, can supervise business analytics projects.



# Outline

What is QBUS6840 about?

Admin stuff

Introduction to Forecasting

# Textbook

- ▶ Free Online Textbook (main textbook used in the first half of the unit)
  - ▶ Rob J Hyndman and George Athanasopoulos, "Forecasting: principles and practice", an open access book available at [www.otexts.org/fpp2/](http://www.otexts.org/fpp2/)
- ▶ Deep Learning book <https://www.deeplearningbook.org/>
- ▶ Another textbook (recommended)
  - ▶ BOK: Bowerman, O'Connell and Koehler, "Forecasting, Time Series, and Regression," Duxbury Applied Series, 4th edition, 2005.

## Other References (not required)

- ▶ *Forecasting: Methods and Applications*, 3rd Edition, 1998 by Makridakis, Wheelwright, and Hyndman, Wiley, ISBN 0-471-53233-9. (MWH)
- ▶ *Business Forecasting*, 9th edition, 2008 by Hanke and Wichern, ISBN 0-13-500933-2.
- ▶ *Forecasting practice and Process for Demand Management*, 2006 by Levenbach H. and Cleary, J. P., ISBN 0-534-26268-6, Thomson.
- ▶ *Applied Econometric Time Series*, by Walter Enders, 2nd Edition, Wiley 2004.
- ▶ *Introduction to Time Series and Forecasting*, 2nd edition, 2002 by Brockwell, P. J., and Davis, R. A. Springer-Verlag: New York, ISBN 0-387-95351-5.

# Teaching Material

- ▶ Comprehensive lecture notes are given weekly
- ▶ Information on the Canvas site  
<https://canvas.sydney.edu.au>. Check the Canvas at least weekly!
- ▶ Tutorial material
- ▶ Some materials I will give during the course which may be of interest to you




# Notes







- ▶ Statistical terminologies are shown in blue when first introduced, e.g., autoregression, recurrent neural network
- ▶ Words in red or *italic* are used to emphasize what being said.
- ▶ This is an applied course - we focus more on applications, explaining how the methods work, how to use them, their advantages and limits, etc. Theoretical part and maths are also sometimes discussed briefly.
- ▶ Slides with “\*” are rather technical. You are strongly encouraged to go through them, but they are not tested in the exams.

# The teaching team

- ▶ Lectures: by A/Prof. Minh-Ngoc Tran (coordinator)
- ▶ Tutorials: coordinated by two head tutors
  - ▶ Rangika Peiris
  - ▶ Nick Nguyen
- ▶ See Canvas for the other tutors' details, tutorial timetable, consultation times, etc

# Assessments

Type	Description	Weight	Due	Length
Final exam (Record+) 	Final exam Exam	50%	Formal exam period	2 hours
	Outcomes assessed: <a href="#">LO1</a> <a href="#">LO2</a> <a href="#">LO3</a> <a href="#">LO4</a>			
In-semester test (Record+) 	Mid-Semester Exam In-semester Exam	25%	Week 08	1.5 hours
	Outcomes assessed: <a href="#">LO1</a> <a href="#">LO2</a> <a href="#">LO3</a> <a href="#">LO4</a>			
Assignment 	Group assignment Assignment	25%	Week 13	TBD
	Outcomes assessed: <a href="#">LO1</a> <a href="#">LO2</a> <a href="#">LO3</a> <a href="#">LO4</a> <a href="#">LO5</a>			

 = group assignment   = Type B final exam   = Type B in-semester exam 

Note: the due date and time are subject to change.

# Academic Honesty Module

- ▶ Academic Honesty module may commence from Week 2
- ▶ All undergraduate and postgraduate students commencing their degrees required to complete this new online module
- ▶ Students who must undertake the module should go <http://canvas.sydney.edu.au>
- ▶ Students must achieve 100% in all tests before the new module considers the student to have passed the module
- ▶ It takes around one hour to complete the module.
- ▶ Students must pass the Academic Honesty, or an Absent Fail (AF) grade will be given for the entire unit



# Plagiarism

- ▶ **Be careful to avoid Plagiarism**
- ▶ What, Why and How: <https://sydney.edu.au/students/academic-dishonesty-and-plagiarism.html>
- ▶ Plagiarism means presenting work that is not your own without acknowledging the original source of the work. It doesn't matter whether you do this on purpose or accidentally.
- ▶ Never copy other stuff
- ▶ Turnitin system compares your assignment reports with other assignments, reports, journals etc (almost openly available sources)
- ▶ **Do not use ghost writing!** Expecting a severe punishment if you hire a ghostwriter to do your assignments

# Group Project

- ▶ Each group consisting of 5 students
- ▶ Groups will be formed in a self-choice way on Canvas. Please register a group with your favourite fellow students on Canvas, not necessarily in the same classes
- ▶ Any remaining students will be randomly grouped, or added to some formed groups
- ▶ A group leader of each group will be automatically assigned by Canvas
- ▶ Attend group activities and complete the Group Project as a team
- ▶ Please refer to group project materials

# Communication

- ▶ The best option is to talk to the teaching team during or after classes, at consultation times
- ▶ Using emails is also an option
  - ▶ Only send emails from your USyd official email address, don't send emails from your work email addresses or other personal email addresses
  - ▶ Always mention your full name and student ID

# The Software for the Unit

- ▶ Python
  - ▶ Free and Works on PCs, Mac, Unix/Linux
  - ▶ Does statistical modelling, visualisation and programming environment
  - ▶ Can be used for almost all models to be discussed in this class
  - ▶ pandas, statsmodel, keras packages needed for this course
- ▶ Matlab (not supported by us)
  - ▶ Licensed: USyd provides a license to enrolled students and can be downloaded and installed on personal computers, download it from <https://sydney.edu.au/students/forms/student-it/matlab-use-agreement.html>
  - ▶ Does statistical modelling, visualisation and programming environment
  - ▶ Powerful Statistics Toolbox and Time Series Toolbox
- ▶ Excel is not enough to complete most data analysis tasks in this course

# Some Tips

- ▶ Learning by doing
- ▶ You become a better problem solver by solving problems. Focus your efforts on the assignments and tutorials.
- ▶ Ask questions in the lectures, consultation times, and by e-mail.
- ▶ Discuss the materials with your classmates/colleagues. Ask for help.
- ▶ Look for answers and extra readings on the Internet, especially about programming issues.
- ▶ We focus on technical materials, but remember the profile of the data scientist: common sense, communication and data visualisation skills are very important.

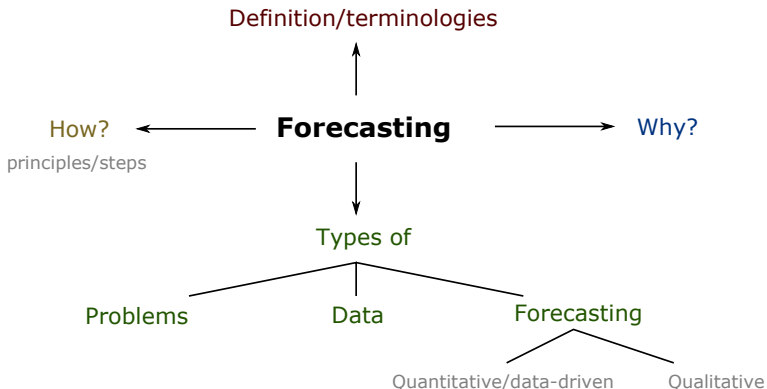
# Outline

What is QBUS6840 about?

Admin stuff

Introduction to Forecasting

# Introduction to Forecasting



- ▶ Reading: <https://otexts.com/fpp2/intro.html>

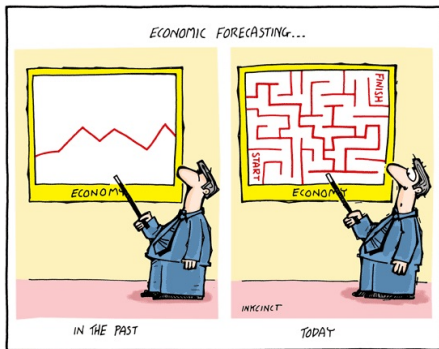
# Defining predictive modeling

- ▶ **Predictive modeling**: the process of developing a mathematical/statistical tool or model that generates an accurate prediction.
- ▶ Sometimes we only care about the predictions themselves. But it is also often essential to establish and communicate the uncertainty in the predictions, as this is often important to account for the risk in decision making.
- ▶ Prediction vs Interpretation.
  - ▶ Interpretation (or **statistical inference**) focuses on interpreting and understanding what has happened
  - ▶ Prediction is to forecast what might happen in future.



# Forecasting

- ▶ A **forecast** is a prediction of what might happen in the future. **Forecasting** is the process of making a forecast.
- ▶ This this course, the terms prediction and forecast are the same, and used exchangeably.
- ▶ Forecasting influences business and economic decision making, planning, policy setting, etc.



2011-503 © INKCINCT Cartoons [www.inkcinct.com.au](http://www.inkcinct.com.au)

# Importance of forecasting

- ▶ Governments need to forecast unemployment, economic growth, expected revenues from income taxes, etc. to formulate policies.
- ▶ Companies need to forecast demand, sales, consumer preferences in strategic planning.
- ▶ Banks/investors/financial analysts need to forecast financial returns, risk or volatility.
- ▶ University administrators need to forecast enrollments to plan for facilities and for faculty recruitment
- ▶ Retail stores need to forecast demand to control inventory levels, hire employees and provide training
- ▶ Sports organisations need to project sports performance, crowd figures, club gear sales, revenues, etc. in the coming season.

# Types of problems and data

- ▶ Regression problems and classification problems.
- ▶ Time series data (the focus of this unit) and cross-sectional data (also considered in this unit, but not much).

# Regression problems

- ▶ In regression problems we want to predict a *numerical* outcome.
  - ▶ “How many copies will this book sell?”
  - ▶ “What will inflation be next month?”
  - ▶ “How much will my house sell for in the current market?”
  - ▶ “How much is this customer going to spend in my website today, given that he is going to purchase something?”
  - ▶ “How many tourists are going to visit NSW within the next years?”
- ▶ Examples of regression models/algorithms: linear regression, penalized linear regression, partial least squares, neural networks, regression trees, etc.

# Classification Problems

- ▶ Classification involves mapping your data points into a finite set of labels or the probabilities for each label. Some examples:
  - ▶ “Will someone click on this ad?” 0 or 1 (no or yes)
  - ▶ “What is this news article about?” politics, sports, culture ...
  - ▶ “What number is this? (image recognition)” 0, 1, 2, ...
  - ▶ “Is this message spam” 0 or 1
  - ▶ “Is this transaction fraudulent?” 0 or 1
  - ▶ “Is the customer going to leave the service?” 0 or 1
- ▶ Examples of techniques for predicting labels: logistic regression, k-nearest neighbours, naive Bayes, discriminant analysis, classification trees, support vector machines, etc.
- ▶ We will not cover classification problems in this course, but you can study this in the Data Mining course (QBUS6810) and/or Machine Learning course (QBUS6850)

# Cross-sectional data

- ▶ Cross-sectional data are values observed “at one point” in time.
  - ▶ Starting salary and WAM for graduates in 2015
  - ▶ For the used Commodores, data were collected on each car’s age, condition, odometer reading and on whether the seller is an individual or a dealer.
  - ▶ Annual return on Fortune 500 company stocks in 2018.
  - ▶ Votes for or against Labor party in 2019 Federal election.

## Example: car emissions

Model	Engine (litres)	City (mpg)	Highway (mpg)	Carbon (tons CO2 per year)
Chevrolet Aveo	1.6	25	34	6.6
Chevrolet Aveo 5	1.6	25	34	6.6
Honda Civic	1.8	25	36	6.3
Honda Civic Hybrid	1.3	40	45	4.4
Honda Fit	1.5	27	33	6.1
Honda Fit	1.5	28	35	5.9
Hyundai Accent	1.6	26	35	6.3
Kia Rio	1.6	26	35	6.1
Nissan Versa	1.8	27	33	6.3
Nissan Versa	1.8	24	32	6.8
Pontiac G3 Wave	1.6	25	34	6.6
Pontiac G3 Wave 5	1.6	25	34	6.6
Pontiac Vibe	1.8	26	31	6.6
Saturn Astra 2DR Hatchback	1.8	24	30	6.8
Saturn Astra 4DR Hatchback	1.8	24	30	6.8
Scion xD	1.8	26	32	6.6
Toyota Corolla	1.8	27	35	6.1
Toyota Matrix	1.8	25	31	6.6
Toyota Prius	1.5	48	45	4.0
Toyota Yaris	1.5	29	35	5.9

# Cross sectional predictions

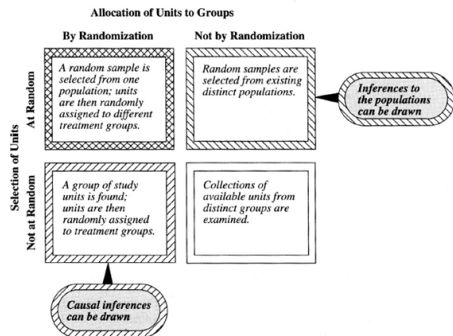
- ▶ Predict the carbon footprint (tons of CO<sub>2</sub> per year) for other similar vehicles that are not included in the above table.
- ▶ First estimate the effects of the predictors.
- ▶ Then, forecast its carbon footprint.



# What type of generalisation does your data support?

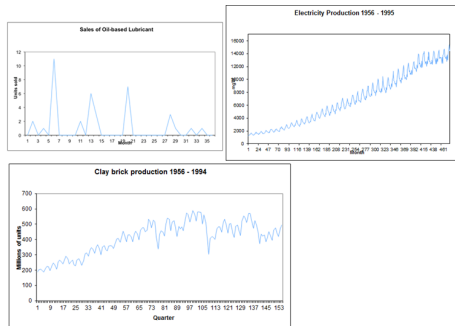
- For purposes of this course, we will be dealing with observational data and will only be interested in predicting a numerical variable of interest, not establishing causality, which is the job of econometrics or economics

Statistical inferences permitted by study designs



# Time series plots and forecasting

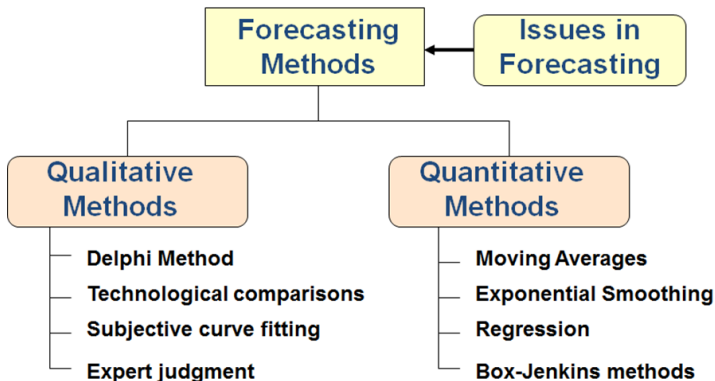
- ▶ Top left plot is monthly sales of a lubricant. It is challenging to forecast because...
- ▶ Top right plot is monthly electricity production. It is easier to forecast because..
- ▶ Bottom plot is quarterly sales of clay bricks in Australia. It is challenging to forecast because...



# Types of forecasting

- ▶ Qualitative (judgmental) forecasting.
- ▶ Quantitative (data based) forecasting.
- ▶ Combine of the two: judgmentally adjusted statistical forecasting

# Types of forecasting



# Judgmental forecasting

- ▶ Expert opinion (subjective)
  - ▶ often used when there is a lack of historical data
  - ▶ used in conjunction with data-based forecasting
- ▶ Delphi method: a popular judgmental forecasting method
  - ▶ Invented in 1950s by Helmer and Dalkey
  - ▶ Assumption: forecasts from a group is more accurate than those from individuals
  - ▶ Stages: forming panel, setting tasks, initial expert views, feedback to experts, aggregating expert views for forecasting
- ▶ Subjectively extending previous patterns into future

# The Delphi method

1. Panel of experts BUT never meet as a group.
2. Use questionnaire to obtain forecasts from all participants
3. Summarize and re-distribute results to all participants PLUS appropriate new questions
4. Summarize again, refining forecasts and conditions, and develop new questions.
5. Repeat Step 4 if necessary. Distribute the final results to all participants.

# Quantitative Forecasting

- ▶ Based on historical data
- ▶ Use formal econometric or statistical forecasting methods.
  - ▶ Project previous patterns into future using a statistical model.
  - ▶ Time series modelling.
  - ▶ Regression modelling.
- ▶ Is the focus of this course
  - ▶ time series forecast
  - ▶ regression forecast

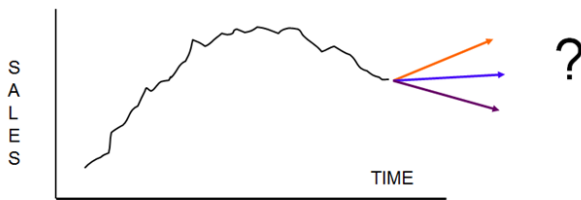


# Time series forecasts

- ▶ A class of forecasting techniques based on time series data analysis: AR, ARMA, Recurrent neural networks, etc.
- ▶ Project previous patterns into future using a formal statistical model e.g.

$$\text{Sales}_t = f(\text{Sales}_{t-1}, \text{Sales}_{t-2}, \dots, \text{Sales}_{t-24}) + \epsilon_t$$

- ▶ Only concerned with forecasting, not reasons why the variable changes



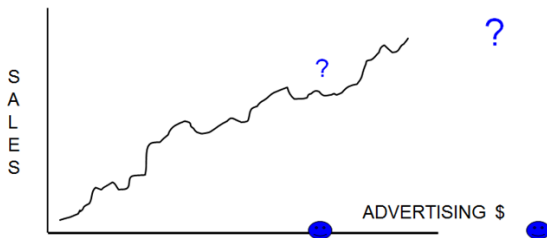


# Regression forecasts

- ▶ A class of forecasting techniques based on regression modelling
- ▶ Use a formal statistical regression model

$$\text{Price}_t = f(\text{Season}_t; \text{Demand}_t; \text{GeneratorVendor}_t) + \epsilon_t$$

- ▶ Can assess other quantities related to price changes
- ▶ Can do scenario forecasting



# The process of forecasting

- ▶ Formulation of the business problem
- ▶ Gathering information
- ▶ Preliminary data analysis
- ▶ Choose and test models.
- ▶ Using and evaluating a forecasting model (forecast, assess forecasts, and implement decision making).

This can be and should be an iterative process of discovery: (i) to refine the forecasting process, (ii) to take into account new information; the business environment is always changing

# Formulation of the business problem

## ► Formulation

- The first step is to understand the business problem to be solved. Often, raising the business question is as important as finding the solution
- What exactly needs forecasting? Can the variable of interest even be forecast?
- How will forecasts be used?

## ► Principles

- Use experts knowledge/previous studies to examine if forecasting is considered possible.
- Use theory to guide the search for possible explanatory factors.
- Communicate with all involved in data collection, decision making, etc. to properly structure problem definition.

# Gathering information

- ▶ What information is available for the forecast problem?
- ▶ What kind of data is needed? Where to find the data?  
customer database, transaction database, etc.
- ▶ Often, it's necessary to invest a lot of money in data collection. Sometimes, it's necessary to estimate and compare the costs and benefits of each data source

# Preliminary data analysis

- ▶ Each statistical data analysis technique requires a specific form of data it uses: numerical, categorical, tabular format,...
- ▶ In the preliminary data analysis phase, the raw data might be converted into the required forms
- ▶ Cleaning data, checking typos, checking unusual observations,...
- ▶ Visualize the data to get some insights

# Choose and test models

- ▶ From the problem formulation and preliminary data analysis step, try to work out the predictive models that are suitable for the data you have
- ▶ There are an extensive list of predictive postulated models: ARMA, Recurrent Neural Networks, etc. It's important to select an appropriate model (or set of models).
- ▶ Principles
  - ▶ We must be careful not to overfit by using excessively complex models that pick up noise in the training sample instead of underlying predictive patterns: [Occam's Razor principle](#)
  - ▶ *All models are wrong, but some are useful* - George Box. No single model or method will always be best. We should consider a wide variety of techniques.
  - ▶ Combining predictions from different models might work best than any single model in isolation.

# Using and evaluating a forecasting model

- ▶ How will the forecasts be used?
- ▶ Put the forecasts into real use in order to realize some return on investment

# Where to find time series data on the web? Australian Data

- ▶ Data Libraries
- ▶ OZDasl - Australian Data and Story Library
- ▶ ANU Social Science Data Archives
- ▶ University of Sydney Library Databases Collection
- ▶ Original Data Sources
- ▶ Reserve Bank of Australia
- ▶ Australian Bureau of Statistics
- ▶ Penn World Tables ? Australia
- ▶ Time Series Data Library (Hyndman)
- ▶ Datastream International



# Where to find time series data on the web? US Data

- ▶ Yahoo Finance Database
- ▶ Data and Story Library — CMU
- ▶ Bureau of Economic Analysis
- ▶ Economagic - times series data
- ▶ FRED - Federal Reserve Economic Data
- ▶ White House - Economic Statistics Brieng Room
- ▶ NBER - National Bureau of Economic Research
- ▶ NBER - Marriage and Divorce Data
- ▶ ICensus - Statistical Abstracts
- ▶ Census Data
- ▶ ICPSR - Interuniversity Consortium for Social and Political Research
- ▶ Panel Survey of Income Dynamics (PSID)
- ▶ Bureau of Labor Statistics
- ▶ Survey of Income and Program Participation (SIPP)
- ▶ National Center for Health Statistics
- ▶ Statistics in Sport (American Statistical Association)

# Elements of a Good Forecast



# We're human after all!

- ▶ Forecasting techniques you're going to learn in this course are powerful. But they are not magic! They require good data, proper understanding and validation.
- ▶ Forecaster's creativity, business knowledge, common sense and human judgment are essential for their success.

**Happy forecasting!**