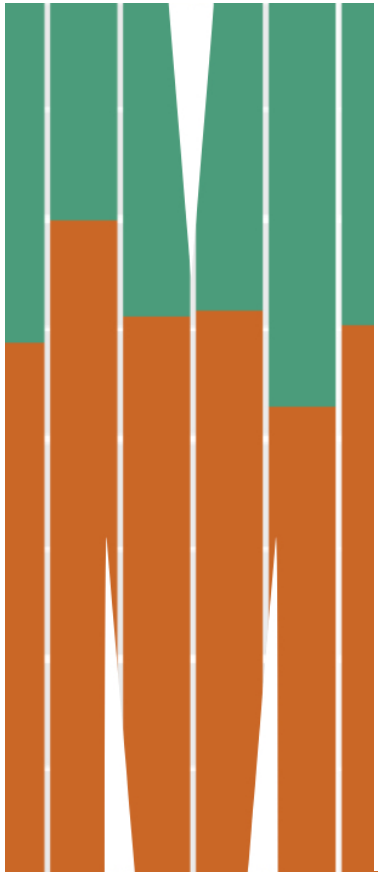


# ETC3250/5250: Introduction

## Semester 1, 2020

Professor Di Cook

Econometrics and Business Statistics
Monash University

Week 1 (a)

# Who are we?


Di Cook
*Chief examiner*

Samantha Dawson
*Teaching Associate*

Ursula Laa
*Teaching Associate*

???
*Teaching Associate*

# Textbook

James, Witten, Hastie and Tibshirani (2012) An Introduction to Statistical Learning. Springer.

http://www.statlearning.com

- Free pdf online
- Data sets in associated R package *ISLR*
- R code for examples

# Semester outline

- Week 1: Introduction to statistical learning, Chapter 2
- Week 2: Linear regression, Chapter 3
- Week 3: Resampling, Chapter 5
- Week 4: Dimension reduction, Chapter 10.2 + instructor's notes
- Week 5: Visualisation, Instructor's notes
- Week 6: Classification, Chapters 8, 7
- Week 7: Classification, Chapter 9
- Week 8: Ensembles and boosted models, Chapter 8.2
- Week 9: Regularization methods, Chapter 6
- Week 10: Model assessment, Instructor's notes
- Week 11: Clustering, Chapter 10
- Week 12: Project presentations

# Assessment

- Final exam 60%
- Four assignments, 4% each (due weeks 3, 5, 7, 9)
- Tutorial quizzes (10) 4% total (start of each tutorial)
- Project 20% (due week 12)

# Communication

- Website: https://iml.numbat.space
  - Lecture notes
  - Assignments
  - Data
- Moodle
  - Links to resources
  - Marks
  - Discussion board, questions
  - Assignment turn in

# What is business analytics?

Business analytics is the scientific process of learning from data, transforming data into insight for making better decisions

- Broader than business intelligence which focuses on describing and predicting performance.
- Broader than econometrics which typically starts from theory (hypotheses or models), and analysts assess if the data supports or refutes.
- Narrower than data science as we are primarily focusing on business problems.

# Flavours of

- Financial Analytics
- Human Resource Analytics
- Marketing Analytics
- Health Care Analytics
- Supply Chain Analytics
- Analytics for Government and Nonprofits
- Sport Analytics
- Web Analytics

# Related fields

How these other disciplines relate to business analytics

*These are my sound bites, to create some distinction but in practice there is a lot of overlap between activities*

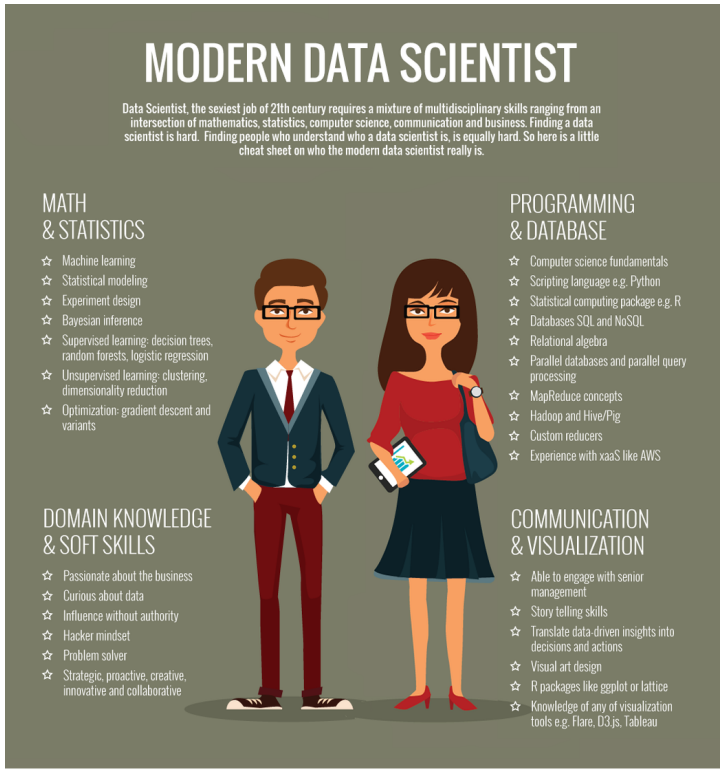
| | |
|---|---|
| Statistics measuring, controlling, and communicating uncertainty, typically with probabilistic models and antecedent hypotheses | Machine learning construction and study of predictive algorithms that improve automatically through experience |
| Data science what can the data tell us: cleaning, validation, transformation, visualisation, models, related to exploratory data analysis | Data mining algorithms for discovering patterns in data, including data storage and access, focused more on prediction |

# Top jobs

Annual job ratings can be found here

https://www.careercast.com/jobs-rated/2019-jobs-rated-report

# Skills needed



# Thinking out loud

*What sort of personality makes for an effective data scientist? Definitely curiosity…. The biggest question in data science is 'Why?' Why is this happening? If you notice that there's a pattern, ask, "Why?" Is there something wrong with the data or is this an actual pattern going on? Can we conclude anything from this pattern? A natural curiosity will definitely give you a good foundation.* -- Carla Gentry, Data Scientist at Talent Analytics

*[Data scientists are] able to think of ways to use data to solve problems that otherwise would have been unsolved, or solved using only intuition.* -- Peter Skomoroch, Former Principal Data Scientist at LinkedIn

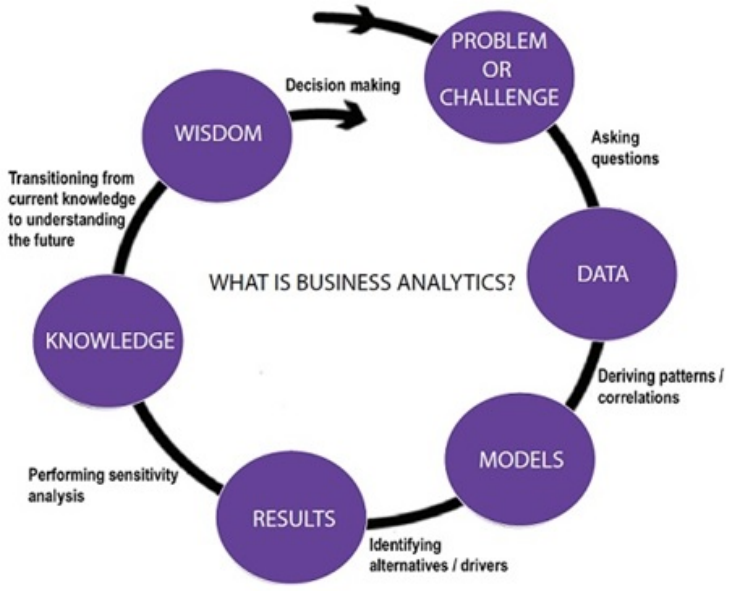
# Thinking out loud

*Always ask yourself how the data can be used to positively impact the lives around you, and use that to guide your design and development.* -- Hanji Xiong, Chief Scientist at Experian's Global DataLab

*Data analysts who don't organize their transformation pipelines often end up not being able to repeat their analyses, so the advice I would give to myself is the same advice often given to traditional scientists: make your experiments repeatable!* -- Mike Driscoll, Founder & CEO at Metamarkets

All quotes come from https://www.kdnuggets.com/2017/05/42-essential-quotes-data-science-thought-leaders.html/2 which has the links to original sources.

# The business analytics process

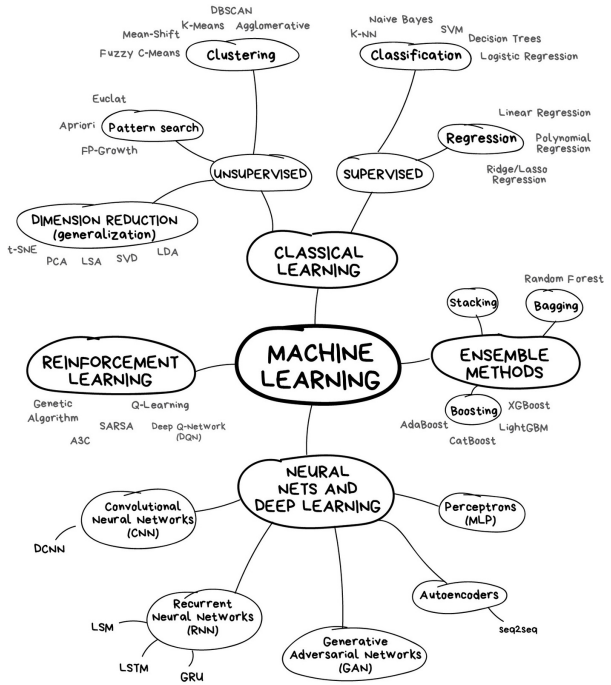



Source: http://www.stern.nyu.edu/programs-admissions/executive-education/short-courses/schedule/short-course-program-7

# no, no, no ...


Source:

https://imgs.xkcd.com/comics/machine_learning.png




Machine Learning for Everybody 😆

is a easy to read overview of the field.

http://vas3k.com/blog/machine_learning/

# Learning objectives for this class

- Select and develop appropriate models for regression, classification or clustering
- Estimate and simulate from a variety of statistical models, and measure the uncertainty of a prediction using resampling methods
- Manage large data sets in a modern software environment, and explain and interpret the analyses undertaken clearly and effectively
- Apply business analytic tools to produce innovative solutions in finance, marketing, economics and related areas

Teaching and learning approach: Two 1-hour lectures and a one 1.5 hour lab class each week for 12 weeks.

How do you do well in this class?

How do you do well in this class?

Turn up to class, summarise your notes after each, note what you understand, and what you don't 🌊

## How do you do well in this class?

Turn up to class, summarise your notes after each, note what you understand, and what you don't 🌊

Do exercises from the textbook related to material each week, check your answers with online solutions 🧗

## How do you do well in this class?

Turn up to class, summarise your notes after each, note what you understand, and what you don't 🌊

Participate actively in computer labs, work with team mates to solve problems, get best answers 🍌

Do exercises from the textbook related to material each week, check your answers with online solutions 🧗

## How do you do well in this class?

Turn up to class, summarise your notes after each, note what you understand, and what you don't 🌊

Participate actively in computer labs, work with team mates to solve problems, get best answers 🍌

Do exercises from the textbook related to material each week, check your answers with online solutions 🧗

## Common languages








## Used in this class


and the RStudio IDE

# After this course

## ETC3555 - Statistical Machine Learning

This unit covers the methods and practice of statistical machine learning for modern data analysis problems. Topics covered will include recommender systems, social networks, text mining, matrix decomposition and completion, and sparse multivariate methods. All computing will be conducted using the R programming language.

Prerequisites: ETC3250 or FIT3154

## ETC5550 - Advanced Statistical Modelling

This unit introduces extensions of linear regression models for handling a wide variety of data analysis problems. Three extensions will be considered: generalised linear models for handling counts and binary data; mixed-effect models for handling data with a grouped or hierarchical structure; and non-parametric regression for handling non-linear relationships. All computing will be conducted using R.

Prerequisites: ETC2410, ETC2420, ETC3440 or equivalent.

# 👨‍💻 Made by a human with a computer

Slides at https://iml.numbat.space.

Code and data at https://github.com/numbats/iml.

Some materials from Alison Hill and Garrett Grolemund's Introduction to Machine Learning in the Tidyverse, RStudio 2020 workshop.

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.