COMP4030

DATA MODELLING AND ANALYSIS

Lecture 1: Module Overview

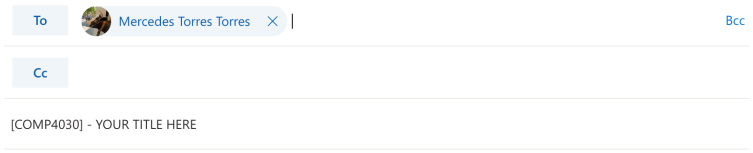
Dr Mercedes Torres Torres
mercedes.torrestorres@nottingham.ac.uk

OUTLINE

- Administrative details
- What is COMP4030?
- Teaching:
  - Outline of lectures
  - Labs and lectures
- Assessments
- Reading material
- What is this Data Modelling and Analysis?
- Examples

ADMINISTRATIVE DETAILS

- E-mail: mercedes.torrestorres@nottingham.ac.uk
- Office: B83
- "Office/Teams" Hours:
  - Wednesdays: 9:00 to 11:00
  - E-mail me for a meeting outside those hours. Give me enough time to organise the meeting (>48hr)
- Online etiquette:
  - **E-mail** and **Messages in the General COMP4030 channel are** the best way of reaching me.
  - For e-mail: Use "**[COMP4030]**" in the subject of the e-mail



ABOUT ME

- Member of:
  - Computer Vision Lab
  - Horizon Research Institute.
- My main areas of research:
  - Data Analysis, particularly data augmentation techniques
  - Computer Vision, particularly with medical or aerial imagery
  - Machine Learning, particularly Deep Learning models (for small/skewed datasets).
- Areas of applications:
  - Healthcare
  - Affective Computing
  - Cybersecurity
  - GIS

WHAT IS COMP4030 ALL ABOUT?

- Part of the AI, Modelling and Optimisation theme of the School
- Main themes:
  - The collection, analysis, and visualisation of data
  - The modelling of real-world problems with the use of data, mathematics and computers
  - The principles of optimisation to find solutions to the modelled problems
  - Range of problems that can be modelled computationally
  - Range of techniques that are suitable to analyse and solve those problems

WHAT IS COMP4030 ALL ABOUT?

- Topics covered in this module include:
  - Basic statistics
  - Data visualisation techniques
  - Basics of modelling
  - Model validation and verification
  - Data pre-processing methods
  - Basic data mining
  - Techniques for dealing with uncertainty
  - Introduction to computer simulation and model interpretation

TEACHING

- **Lectures**
  - 12 weeks; 2 hours per week
  - Mondays from 9:00 AM to 11:00 AM (Teams)
- **Labs**
  - 12 weeks; 2 hours per week
  - Wednesdays from 11:00 AM to 1:00 PM (Teams)

OUTLINE OF LECTURES

L#	Lecture Topic	Lab Topic	
1	Introduction	Introduction to R	
2	Modelling	Advanced R	**Block A:** Creating Mathematical Models
	BREAK		
3	Simulation	Modelling and Simulation in R	
4	Modelling and Simulation Exercises	Modelling and Simulation in R II	
5	Data Analysis	Data Analysis	**Block B:** Pre-processing Data
6	Visualisation and Pre-Processing	Visualisation	
7	Pre-processing II and Mining	Pre-processing + CW Intro	
	EASTER BREAK		
8	Clustering	**Lab Submission**	**Block C:** Learning/ Generating Knowledge
9	Clustering II	CW Support + Clustering demo	
	[Bank Holiday]	CW Support	
10	Classification I	CW Support	
11	Classification II	CW Support + Classif. demo	

LECTURES: OVERVIEW

- Each week, we will cover one topic and:
  - How it fits within the Data Modelling/Analysis (DMA) framework
  - Theory behind it
  - Methods to accomplish said topic
  - Examples and exercises
  - R or Weka Demos
- We will have a break after the first hour.
- **Lectures will start at 9:00AM sharp.**

LABS: OVERVIEW

- Labs complement lectures
- How to implement in R what we have seen in class
- On Monday, I will publish on Moodle:
  - **Instructions**: overview of techniques/functions to use
  - **Exercises**: designed to help you use R to solve DMA problems.
- We will build on work done in previous weeks:
  - **DO NOT leave exercises from one week to another!**
- The labs are an opportunity to ask me anything about the module/exercises/examples.

Attendance is compulsory

LECTURES IN TIME OF CORONA

- Lectures will be carried out in Teams
- Materials: All in Moodle. To support your learning, I will provide a series of materials each week:
  1. Slides: available at least 24h before class. Complement with your own notes.
  2. Exercises: when applicable, exercises to cover during the lecture. Available with slides.
  3. Recordings of the lecture: available after the lecture happens.
  4. Highlights: 3 to 5 extra videos summarising key points from the lecture. Available on Mondays.
  5. Quizzes: to help assess your understanding of that week's topic. Available on Mondays.
  6. Tell me more (TMM) videos: 5-to-10 min video exploring more in depth one of the concepts from that week, from real-world uses, to key people in their development, etc. Available on Mondays.
  7. Lab Instructions and Specs: Available on Mondays.
  8. Recordings of the lab: available after the lab happens.

LABS IN TIME OF CORONA

- Labs will take place in Teams
  - Connect to the lab as if it was a lecture, I'll be online.
- They will be organised as such:
  - **11:00 - 11:30: Introduction to the topic -** I will give a general overview on the topic of the lab, tips and advice.
  - **11:30 - 12:30: Independent work -** Time for you to work through Instructions and Questions.
    - While you don't have questions, keep working on the lab specs and exercises.
    - When you have a question, unmute yourself, raise your hand, and ask
  - **12:30 - 13:00: Live Session -** I will choose (at least) one exercise from that week's lab and solve it in real time.

GENERAL MICROSOFT TEAMS ETIQUETTE

- Mute your microphone when you join.
- Video can be turned on or off.
- When you have a question during a lab and/or lecture:
  - Use the "*Raise Hand*" function, OR
  - Write it in the chat (if you @me, I'll get a ping and notice faster), OR
  - Unmute yourself and ask.
- All sessions (lectures, labs, Q&As) will be recorded and available in Teams.

ASSESSMENT

- **Two components:**
  - 1 portfolio of lab work [25 marks]
    - Lab submission
  - 1 coursework [75 marks]
    - Report on data modelling and analysis in R and Weka

ASSESSMENT: LAB PORTFOLIO SUBMISSION

- 25 marks
- Date: 21st April 2021 **[to be confirmed] During the lab**
- Open-book submission.
- No late submissions allowed.
- Use R to:
  - Create and analyse models
  - Simulate models
  - Analyse/visualise data in depth
  - Pre-process and mine data
- Further details will be provided in Lab 6.
- **Deliverable:** a R Notebook with the answers to the questions.

ASSESSMENT: COURSEWORK

- 75 marks
- Release date: 24th of March (during Lab 7)
- Submission date: 24th of May 2021 3PM
- Submission format: electronic (via Moodle)
- Usual late-submission penalty applied: 5% per day.
- Overview:
  - A data set
  - You will be asked to analyse, mine, cluster and classify it.
  - **Deliverable:** a report and data file(s) with all relevant calculations and analysis.
- Further details will be provided in Lab 7.

COMP3040: CREDITS DISTRIBUTION

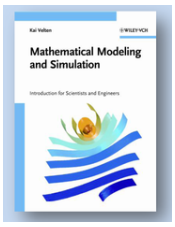
- 20 credits = 200 hours of work
  - Lectures = 11 x 2 = 22
  - Labs = 11 x 2 = 22
  - Lecture study = 10 x 2.5 = 25
  - Lab practice = 10 x 2.5 = 25
  - External reading = 12
  - Revision for lab exam = 12
  - Coursework = 82

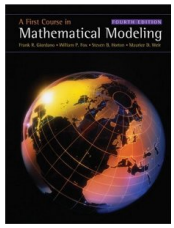
ADVICE

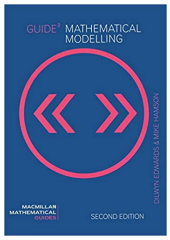
- DMA lectures are organised so every week we build on work done in the previous weeks.
  - It is a marathon, not a sprint!
- Take notes during the lectures.
  - Slides will have some of the information given in class, but not all.
- Revise weekly, instead of at the end.
- Come to all lectures and labs
- Complete all lab work
- When applicable, try out all the exercises beforehand.
- Ask questions!

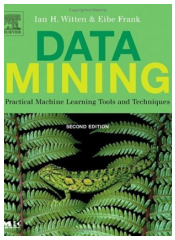
READING MATERIAL

- Main Textbooks:
  - Velten, K., *"Mathematical Modeling and Simulation: Introduction for Scientists and Engineers"*; Wiley-VCH, 2009
  - Giordano, F.R., et al, *"A First Course in Mathematical Modeling"* (4th ed.); Brooks/Cole, Cengage Learning, 2009
  - Edwards, D., Hamson M: *"Guide to Mathematical Modelling"*, Palgrave Mathematical Guides
  - Witten, I.H. and Frank, E.; *"Data Mining: Practical Machine Learning Tools and Techniques"*; Morgan Kaufmann, 2005
- Other books and articles maybe mentioned as we go along





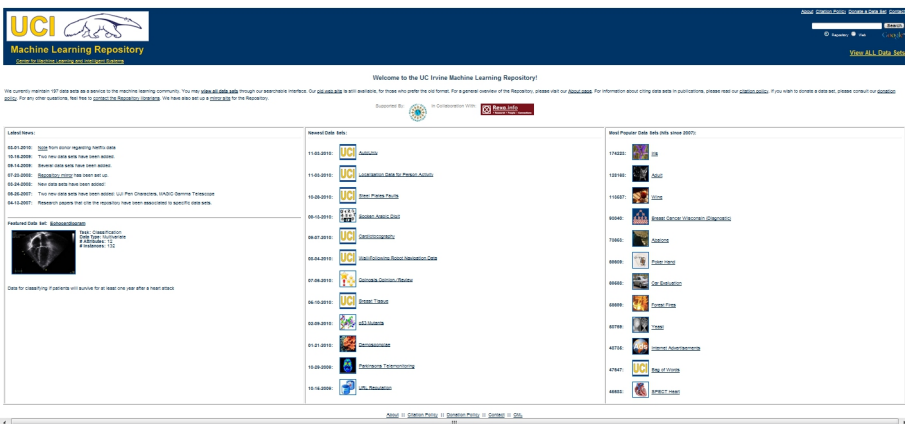




RESOURCES: DATABASE RESPOSITORY

- This is a very useful repository of datasets

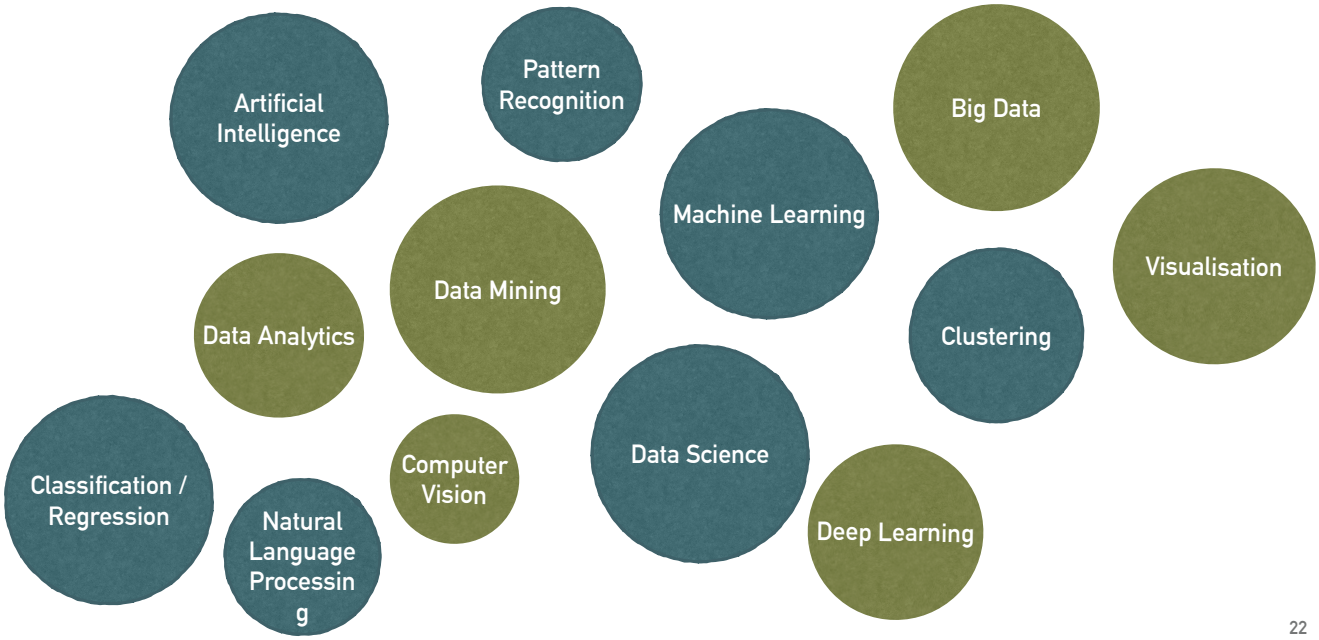
http://archive.ics.uci.edu/ml/



WHAT IS DATA MODELLING AND ANALYSIS ANYWAY?

IS IT EVEN USEFUL?

DATA MODELLING IN CONTEXT

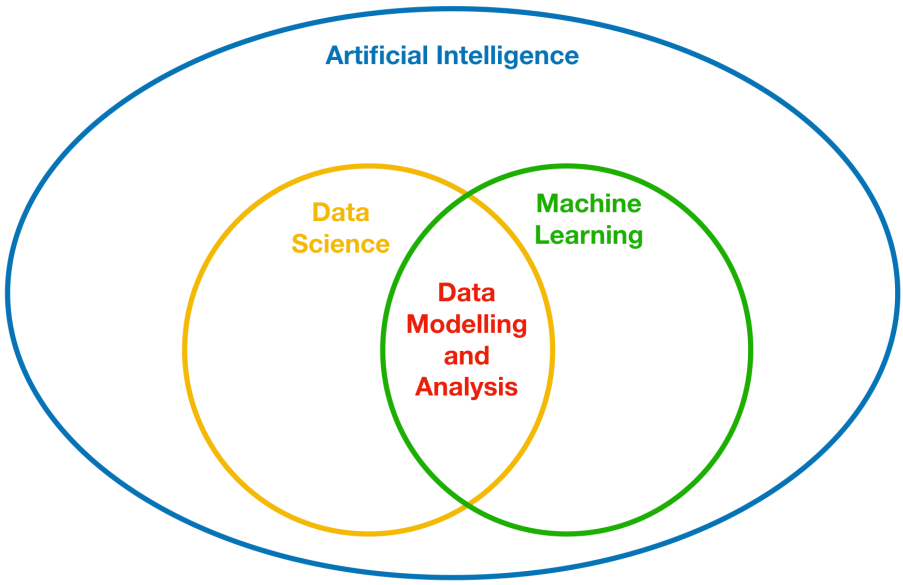


DATA MODELLING IN CONTEXT

- **Artificial Intelligence:**
  - Computer Vision
  - Natural Language Processing
- **Machine Learning:**
  - Pattern Recognition
  - Clustering
  - Classification/Regression
  - Deep Learning
- **Data Science:**
  - Big Data
  - Data Mining
  - Data Analytics / Visualisation

We are going to touch on all these topics - explicitly or implicitly.

DATA MODELLING IN CONTEXT



DATA MODELLING IN CONTEXT

- The overall goal in all of these fields and methods is the same:

To predict the world

- **This is done by:**
    1. Observing/measuring the world (and outcomes)
    2. Learning relationships between observations and outcomes
    3. Applying this knowledge to new observations

HOW ARE DATA MODELLING AND ANALYSIS METHODS USED IN THE REAL WORLD?

AND WHERE?

MODELLING EXAMPLE 1: WEST NILE VIRUS (WNV) SPREAD IN GERMANY [1]

- WNV is an arbovirus transmitted by the bites of infected mosquitoes.
- WNV can infect horses and humans, where it may cause serious illness and be fatal.
- Cycle of the infection involves birds and mosquitoes
- Scientists in Germany have developed a mathematical model that explains:
  1. How mosquitoes are infected by birds
  2. How birds are infected
- Taking into consideration local and migratory birds
- Let's see the first part of the model:

MODELLING EXAMPLE 1: WNV SPREAD IN GERMANY [1] – PARAMETERS

Symbols	Interpretation
b_B	Birth Rate of local birds
m_B	Mortality Rate of local birds
K_B	Carrying capacity of local birds
β_4	Transmission rate from infected mosquitoes to clinical local birds
β_3	Transmission rate from infected mosquitoes to subclinical local birds
γ_{BC}	Latency period of clinical local birds
γ_{BSC}	Latency period of subclinical local birds
α_4	Recovery rate of clinical local birds
α_3	Recovery rate of subclinical local birds
α_9	Recovery rate of migratory birds
γ_3	Rate at which recovered subclinical becomes infected
b_{Bm}	Birth Rate of migratory birds
m_{Bm}	Mortality Rate of migratory birds
K_{Bm}	Carrying capacity of migratory birds
β_5	Transmission rate from infected mosquito to the migratory birds
c_3	Transmission rate from infected migratory birds to the mosquito
γ_{Bm}	Latency period of migratory birds
α_9	Recovery rate of migratory birds
b_M	Birth Rate of mosquitoes
m_M	Mortality Rate of mosquitoes
K_M	Carrying capacity of mosquitoes
c_1	Transmission rate from infected subclinical local birds to the mosquitoes
c_2	Transmission rate from infected clinical local birds to the mosquitoes
γ_M	Incubation period
d_{BC}	Disease induced clinical bird death
d_{BSC}	Disease induced subclinical bird death
d_{Bm}	Disease induced subclinical bird death

MODELLING EXAMPLE 1: WNV SPREAD IN GERMANY [1] – MODELS

$$\frac{dS_M}{dt} = (b_M N_M - m_M S_M)\left(1 - \frac{S_M}{K_M}\right) - \frac{c_2 S_M I_{BC}}{K_B} - \frac{c_1 S_M I_{BSC}}{K_B}$$

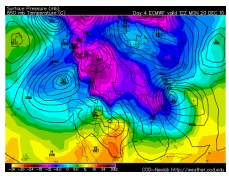
$$\frac{dE_M}{dt} = \frac{c_2 S_M I_{BC}}{K_B} + \frac{c_1 S_M I_{BSC}}{K_B} - \gamma_M E_M - m_M E_M$$

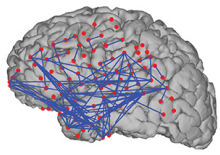
$$\frac{dI_M}{dt} = \gamma_M E_M - m_M I_M, \qquad (1)$$

where S_M, E_M and I_M represent the *susceptible*, *exposed* and *infected* part of the mosquito population, respectively. The total mosquito population is given by $N_M = S_M + E_M + I_M$. Transmission parameters are c_1 from infected subclinical local birds to the mosquitoes and c_2 from infected clinical local birds to the mosquitoes.

MATHEMATICAL MODELS IN THE REAL WORLD

- Many other examples of useful modelling:
  - Election Results/Polling [1, 2]
  - Weather forecasting [1]
  - Flooding [1, 2]
  - Prediction of biodiversity losses [1]
  - Traffic flow forecasting [1, 2, 3, 4]
  - Heart disease [1, 2]
  - Disease Spread [1, 2, 3, 4]
  - And many many more [link!]





DATA MINING EXAMPLE 1: LOAN APPLICATIONS

- Loan application questionnaire
- Statistical analysis of answers
- Outcome:
  - Accept
  - Reject
  - Borderline
- For Borderline: human judgment required
  - Borderline cases: prime customers
- Can we automate “borderline judgments”?

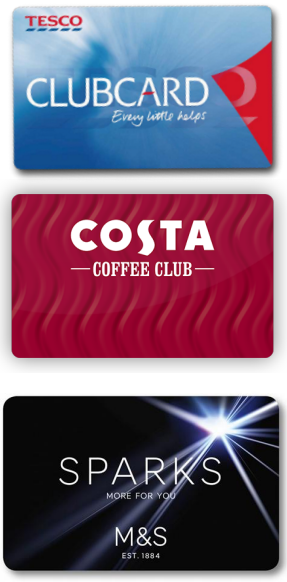
DATA MINING EXAMPLE 1: LOAN APPLICATIONS

- Real mining solution for a real company.
- Data collected from 1000 borderline customers, including “pay back”
- 20 attributes from questionnaires (e.g. age, gender, income, studies, etc.)
- Rules created for **predicting** probability of defaulting on loan
- The system was capable of predicting 2/3 of the test cases
- The company that commissioned the system was satisfied with the system because:
  - It improved the success rate of loan decisions
  - The rules adopted by the model could be used to explain to customers why a particular loan was rejected

DATA MINING EXAMPLE 2: LOYALTY CARDS

- What products tend to be bought together? Associations
- Example association:
  - customers that buy beer tend to buy chips
  - on Thursdays customers often buy nappies and beer together
- Some potential responses:
  - store layout,
  - discount on only one of the paired items,
  - coupons for the other paired item, when only one is bought, etc.

- Personal purchase histories:
  - One of the reasons for discount/rewards or “loyalty” cards
- Models built from these histories allow retailers to:
  - Make buying predictions
  - Optimise their products
  - Optimise product placements
  - Assess effects of advert campaigns
  - Offer different service
  - Etc.



TO SUMMARISE...

- COMP4030 sits at the intersection of Data Science and Machine Learning, under AI.
- The module has two main components:
  1. Theory/methods in Data Modelling and Analysis (Mathematical models, data analysis, pre-processing, machine learning)
  2. How to develop them in R.
- You will be assessed in both components.
- **Previous knowledge:**
  - Maths: matrices, functions, linear equations, numerical series.
  - Programming (general, not in R!): instructions, functions, variables, parameters,...

THE END

Questions