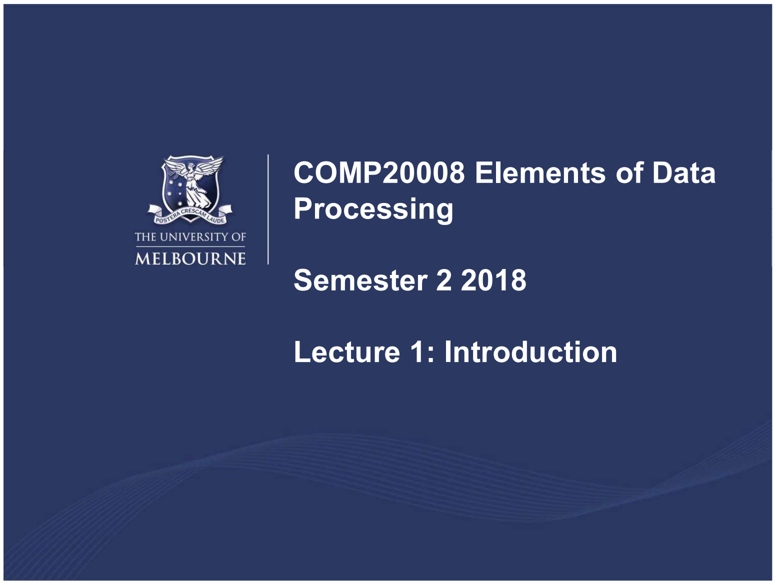
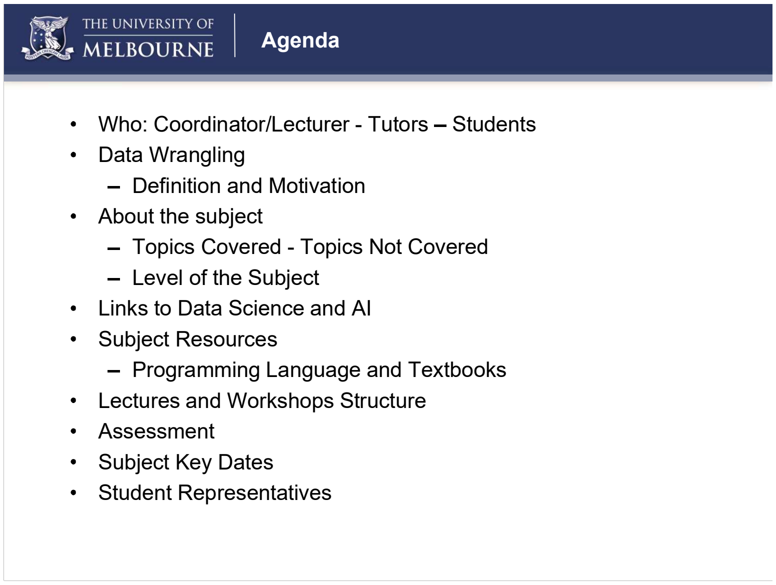


COMP20008 Elements of Data Processing

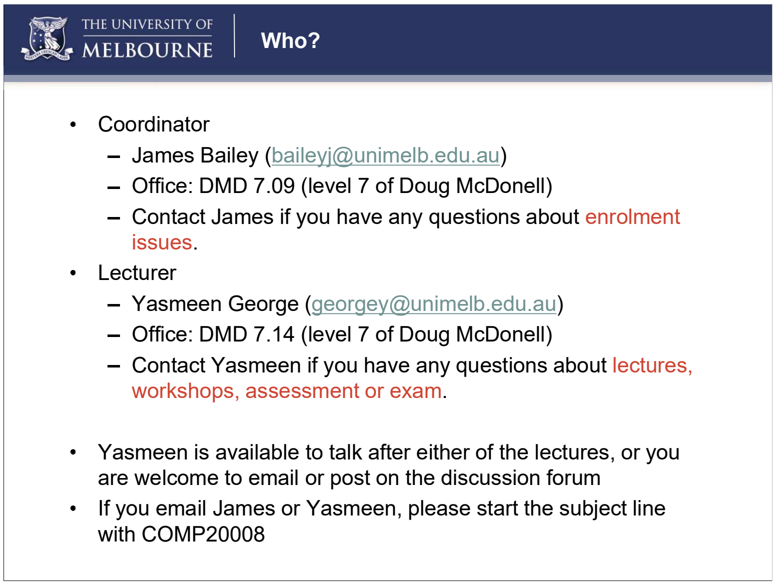
Semester 2 2018

Lecture 1: Introduction



Agenda

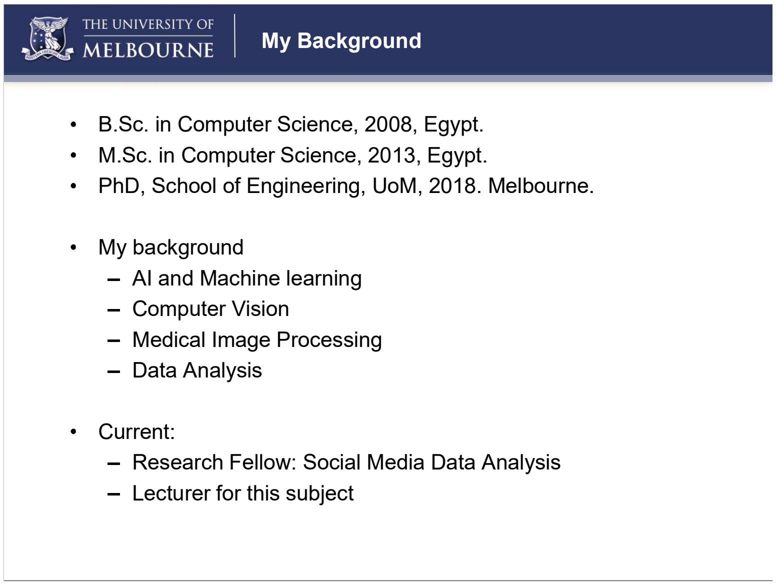
- Who: Coordinator/Lecturer - Tutors – Students
- Data Wrangling
  - Definition and Motivation
- About the subject
  - Topics Covered - Topics Not Covered
  - Level of the Subject
- Links to Data Science and AI
- Subject Resources
  - Programming Language and Textbooks
- Lectures and Workshops Structure
- Assessment
- Subject Key Dates
- Student Representatives



Who?

- Coordinator
  - James Bailey (baileyj@unimelb.edu.au)
  - Office: DMD 7.09 (level 7 of Doug McDonell)
  - Contact James if you have any questions about enrolment issues.
- Lecturer
  - Yasmeen George (georgey@unimelb.edu.au)
  - Office: DMD 7.14 (level 7 of Doug McDonell)
  - Contact Yasmeen if you have any questions about lectures, workshops, assessment or exam.

- Yasmeen is available to talk after either of the lectures, or you are welcome to email or post on the discussion forum
- If you email James or Yasmeen, please start the subject line with COMP20008



My Background

- B.Sc. in Computer Science, 2008, Egypt.
- M.Sc. in Computer Science, 2013, Egypt.
- PhD, School of Engineering, UoM, 2018. Melbourne.

- My background
  - AI and Machine learning
  - Computer Vision
  - Medical Image Processing
  - Data Analysis

- Current:
  - Research Fellow: Social Media Data Analysis
  - Lecturer for this subject

Who?

- Tutors
  - Winn Chow
  - Namrata Srivastava
  - Sobia Amjad
  - Anam Khan
  - Ilya Verenich

About You

- ~275 students
- Students from
  - Bachelor of Science
  - Diploma in Informatics
  - Bachelor of Design
  - Bachelor of Commerce
  - …

What is the subject called?

- Formally, *Elements of Data Processing*
- But we will refer to it as **Data Wrangling.**
- Wrangle: "to control and care for (horses, cattle, etc) on a ranch"

https://en.wikipedia.org/wiki/National_Finals_Rodeo#/media/File:Luke_2004-05-19.jpg

Data Wrangling

- *Data wrangling*: the process of organising, converting, mapping data from one format into another. This may include activities such as data integration, enrichment, aggregation, structuring, storage, visualisation and publishing.
- *Data wrangler*: the person who does the wrangling (transforming data, integrating from multiple sources, overseeing quality issues, visualising, …)

Data Science

The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. …. You also want to be able to visualize the data, communicate the data, and utilize it effectively. But I do think those skills - of being able to access, understand, and communicate the insights you get from data analysis - are going to be extremely important.

Hal Varian, Chief Economist at Google

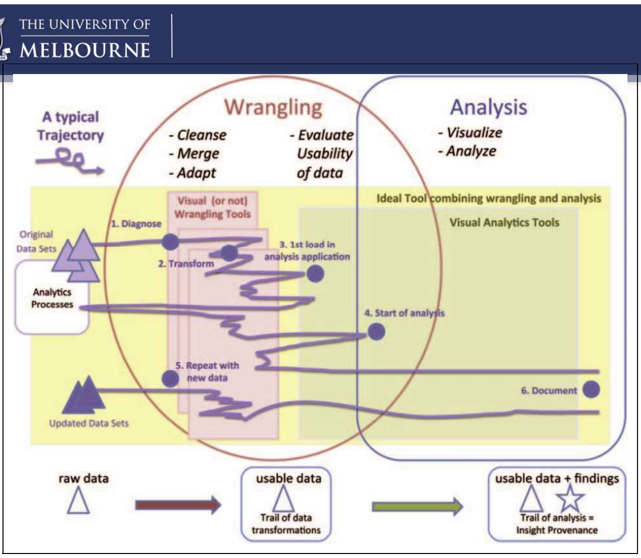
The McKinsey Quarterly, Jan 2009

https://upload.wikimedia.org/wikipedia/commons/2/26/Hal_Varian.jpg

Why is it important?

- Data Science
  - Wrangle the data (80%)
  - Analyse the data (<20%)
  - Present, deploy and communicate results (<20%)

- Most of the effort is spent on data wrangling ……



Research directions in Data Wrangling: visualisations and transformations for crediible data. S. Kandel et al, Information Visualisation 10(4), 2011.

Data and health

- Gene sequences
- Mobile data
- Electronic medical records
- Insurance claims
- Imaging results
- GP data
- Prescription data
- Social media
- …….

https://upload.wikimedia.org/wikipedia/commons/c/ca/Fitibit_Flex.jpg

https://upload.wikimedia.org/wikipedia/commons/e/ee/MRI-Philips.JPG

Plan of Action
Thyroxine Plan
Re-Run Diagnostic Algorithm
Generate Printable Patient Record
Switchboard
Back
View Trends

https://upload.wikimedia.org/wikipedia/commons/b/b7/Thyroid_Clinic_plan.png

https://upload.wikimedia.org/wikipedia/commons/8/80/DNA_methylation.jpg

THE UNIVERSITY OF MELBOURNE

Data and government

- www.data.vic.gov.au
- data.melbourne.vic.gov.au
- AURIN (Australian Urban Research Infrastructure Network)
- data.gov.au

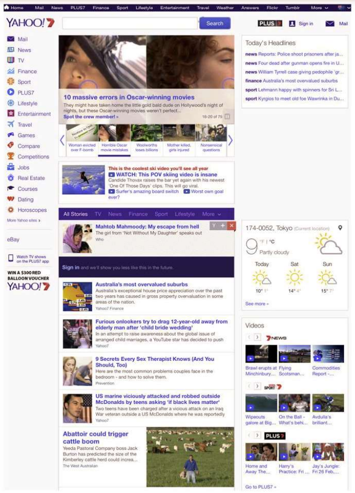
THE UNIVERSITY OF MELBOURNE

Data science for social good

- Tackle social challenges using data science techniques
  - Partnerships with government, NGOs, companies, …..
- Examples
  - https://dssg.uchicago.edu/projects/

THE UNIVERSITY OF MELBOURNE

Data and business (au.yahoo.com)



10 massive errors in Oscar-winning movies
They might have taken home the little gold bald dude on Hollywood's night of nights, but these Oscar-winning movies weren't perfect...
Spot the crew member! »
16-20 of 75

Woman evicted over F-bomb | Horrible Oscar movie mistakes | Woolworths loses billions | Mother killed, girls injured | Nonsensical questions

This is the coolest ski video you'll see all year
WATCH: This POV skiing video is insane
Candide Thovax raises the bar yet again with his newest 'One Of Those Days' clips. This will go viral.
Surfer's amazing board switch | Worst own goal ever?

Mahtob Mahmoody: My escape from hell
The girl from 'Not Without My Daughter' speaks out
Who

Sign in and we'll show you less like this in the future.

Australia's most overvalued suburbs
Australia's exceptional house price appreciation over the past two years has caused in gross property overvaluation in some areas of the nation.
Yahoo7 Finance

Furious onlookers try to drag 12-year-old away from elderly man after 'child bride wedding'
In an attempt to raise awareness about the global issue of arranged child marriages, a YouTube star has decided to push
Yahoo7

9 Secrets Every Sex Therapist Knows (And You Should, Too)
Here are the most common problems couples face in the bedroom - and how to solve them.
Prevention

US marine viciously attacked and robbed outside McDonalds by teens asking 'if black lives matter'
Two teens have been charged after a vicious attack on an Iraq War veteran outside a US McDonalds where he was reportedly
Yahoo7

Abattoir could trigger cattle boom
Yeeda Pastoral Company boss Jack Burton has predicted the size of the Kimberley cattle herd could increa...
The West Australian

Today's Headlines
news Reports: Police shoot prisoners after ja...
news Four dead after gunman opens fire in U...
news William Tyrell case giving pedophile 'gr...
finance Australia's most overvalued suburbs
sport Lehmann happy with spinners for Sri L...
sport Kyrgios to meet old foe Wawrinka in Du...

174-0052, Tokyo (Current location)
9 °F | °C
Partly cloudy

Today	Sat	Sun
10° 1°	14° 4°	15° 7°

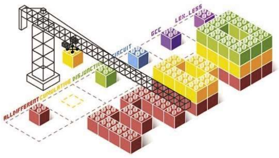
See more »

Videos
Brawl erupts at Minchinbury... | Flying Scotsman... | Commodities Report -...
Wipeouts galore at Big... | On the Ball - What's behi... | Avdulla's brilliant...
Home and Away The... | Harry's Practice: Fri ... | Jay's Jungle: Fri 26 Feb....
Go to PLUS7 »

THE UNIVERSITY OF MELBOURNE

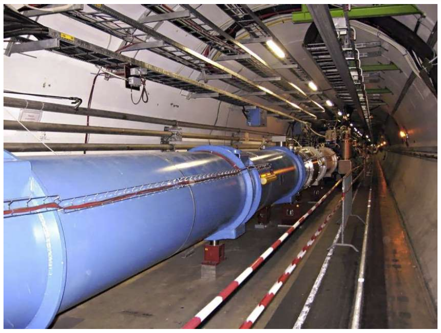
Data and massively online education

- MOOCs (>30 at Unimelb)
  - Video viewing behaviour
  - Quizzes
  - Discussion forum
  - Assignments
  - Interventions to improve learning



Modelling discrete optimisation

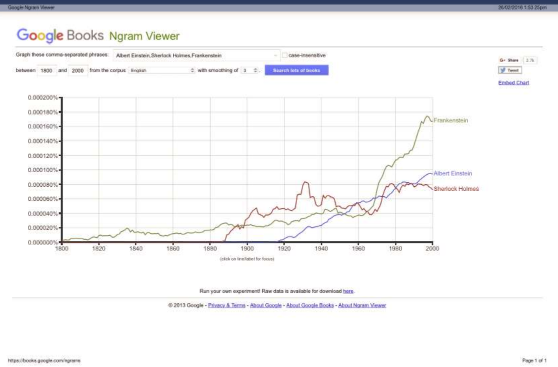
Data and physics



CERN

- Large hadron collider
- 1000 terabytes/second

Data and the humanities



Google Ngram Viewer 26/02/2016 1:53 25pm

Google Books Ngram Viewer

Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein □ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3.

Run your own experiment! Raw data is available for download here.

© 2013 Google - Privacy & Terms - About Google - About Google Books - About Ngram Viewer

https://books.google.com/ngrams Page 1 of 1

- https://books.google.com/ngrams
  - the more frequently an irregular verb is used, the less likely it is to be regularized over time (Aiden and Michel)

Food analytics

- Analyse food samples
- Extract DNA from a food sample
- Identify ingredients, look for contamination or other suprises

- E.g. Hot dog analytics
  - https://s3-us-west-2.amazonaws.com/clearlabs.web/production/public/assets/img/hotdog-report-consumer.pdf

Data and sport

https://upload.wikimedia.org/wikipedia/commons/3/34/BDS_West_2010-11-26.jpg

- Video analysis
- Wearables, GPS tracking, heart rate
- Skin patch behind ear, mouthguard sensors

What is the subject all about?

- **Preprocessing** (4 lectures): Weeks 1-3
  - Data types and processing, data cleaning including outliers, missing data
- **Visualisation** (3 lectures): Weeks 3-4
  - Plotting and visualisation methods, clustering, dimensionality reduction
- **Analysis** (4 lectures): Weeks 5-7
  - Correlations, basic prediction techniques
- **Infrastructure and Distributed** (4 lectures): Weeks 8-10
  - Data linkage and integration, blockchain
- **Social, ethical and privacy issues** (3 lectures): Weeks 11-12
  - K-anonymity, l-diversity, location privacy, ethics

- Additionally, there is an introductory lecture (today), final lecture, three guest lectures (Scott Thomson from Google, Richard Sinnott from CIS, James Bailey from CIS)

What's not covered?

- Complex analysis of data

- Predictive analytics, machine learning, information retrieval technology and data mining algorithms (we will just get a taste)
  - These are covered in more depth in Machine Learning COMP30027. Relational databases
  - See instead INFO20003 Database Systems

Level of the subject

- Assumes you have completed COMP10001 (or equivalent)
  - Assumes knowledge of programming (in Python)
  - Assumes knowledge of Maths basics
    - Logarithms, power, Euclidean distance, matrices and vectors, mean, median, sigma notation $\sum$

- Material will be pitched at 2^{nd} *year level*

- *You cannot gain credit for both COMP20008 and INFO20002 (Foundations of Informatics)*

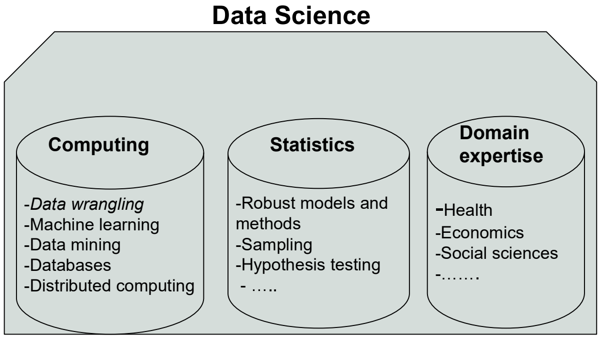
Links to data science and AI

- This subject is a building block for data science and artificial intelligence

- From this morning's paper (SMH)
  - *The question used to be, 'what will you do when you grow up?' The question now is, 'what will you do when the robots grow up?'*
  - *"The typical university student of the future will be a 40-something robot-programmer embarking on their third micro-degree."*

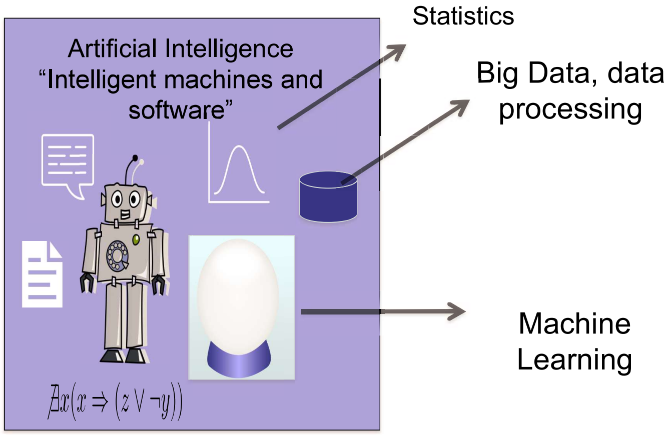
- For a high level idea of future directions
  - https://www.youtube.com/watch?v=4kiBETz1V7o

Data Science landscape

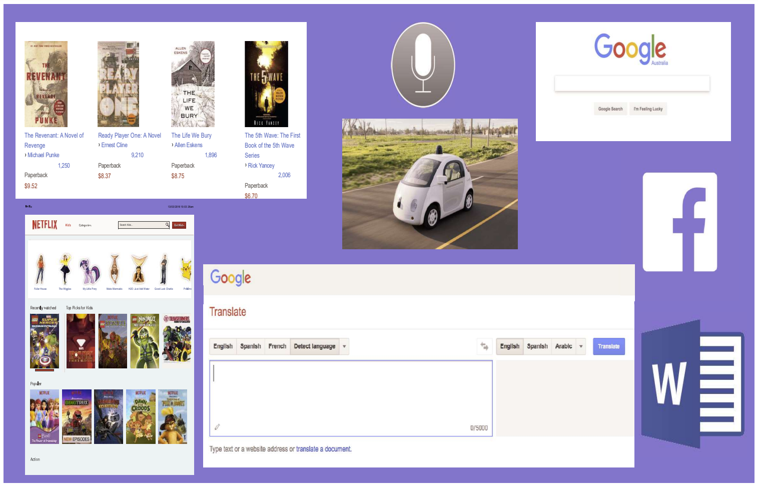


This subject (COMP20008) is one of the subjects leading to a data science major in BSc. A pre-requisite for 3rd year "Machine Learning"

Artificial Intelligence, Big Data, and Machine Learning


Artificial Intelligence
"Intelligent machines and software"
Statistics
Big Data, data processing
Machine Learning

"AI is the new electricity" ---- Andrew Ng
Data-driven, intelligent systems



Programming for data wrangling

- Python (we will be using)
  - Fully fledged, multi purpose programming language
  - Good for combining data wrangling activities into a larger pipeline of production or web development
  - Good library support for scientific and machine learning extensions
- R (we will not be using)
  - More of a statistics focus
  - Large community support

Textbooks

- None !
  - No single textbook includes all topics we cover
  - Do not need to purchase any textbook. Material needed will be covered in the lectures and the references provided
- There exist a number of practically oriented books on data wrangling using python. We will adapt some exercises from these for the workshops. You do not need to purchase these books.
  - Data wrangling with python: Tips and tools to make your life easier. Jacqueline Kazel and Katharine Jarmul. Published by O'Reilly 2015
  - Data science from scratch: First principles with python. Joel Grus, Published by O'Reilly 2015
  - Python for data analysis. Wes Mckinney, Published by O'Reilly 2013.

Previous offerings and materials

- Subject was offered in 2016, 2017 and 2018-SM1
  - 2018-SM2 will be very similar to 2018-SM1 offering
  - Exam structure and difficulty will be similar
  - Assessments will be a bit different
- Lectures and workshop content will be posted to the LMS. Typically an early draft of the next lecture will be available a few days before (labelled *draft*). This will then be replaced by the final lecture content just before the lecture is delivered.
- Lecture recordings will be available through the LMS

Schedule

- A combination of lectures and workshops
  - Lectures:
    - Presentation of principles
    - 9:00-10:00 Wednesday (9:05-9:55)
    - 16:15-17:15 Friday (16:20-17:05)
    - Recorded using Echo
      - New features in Echo – can anonymously mark points of confusion in lecture video
    - No Python programming in lectures!
  - Workshops (one per week)
    - 2 hours
    - A mixture of tutorial and programming lab
    - Start in Week 2 (*NO WORKSHOPS THIS WEEK*)

Workshops

- Workshops will include programming exercises on the lab computers, using Jupyter notebooks
  - The tutor will make every effort to provide advice tailored to the computing environment being used in that workshop
    - We might not be able to provide advice if you choose to use a different environment, or use your own laptop
- Your expectation:
  - Workshops materials should be uploaded one week before actual tutorials.
- Our expectation:
  - You come to the workshop prepared. An attempt to solve the exercises would make the workshops run more efficiently!

Python

- We will be using Jupyter Notebook Python 3
  - The Anaconda distribution can also be particularly convenient
    - https://www.continuum.io/downloads
- Get prepared for workshop week 2
  - Pandas library is one of the most preferred tools for data scientists to do data manipulation and analysis
  - Matplotlib for data visualization
  - NumPy, the fundamental library for scientific computing in Python on which Pandas was built.
  - https://pandas.pydata.org/pandas-docs/stable/tutorials.html

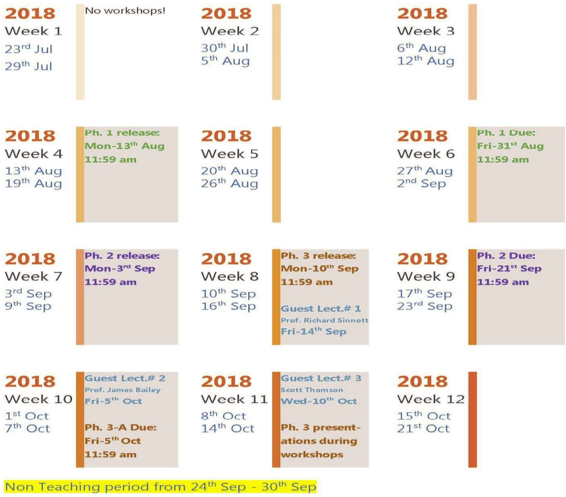
Assessment

- Your subject mark will be made up of
  - Final exam: 50%
  - Project work during semester (staged project): 50%
    - Phase 1: Python data wrangling warmup exercises (20% - week 4-6)
    - Phase 2: Python data wrangling exercises (15% - week 7-9)
    - Phases 3: Data wrangling investigation on an open dataset (will be flexible in what to use) (15% - week 8-11)
      - Phase 3-A: Code (5%)
      - Phase 3-B: Oral presentation in workshop (10%)

Assessment Cont.

- There are two hurdles for passing the subject
  - You must achieve at least 20/50 for the final exam
  - You must achieve at least 20/50 for the workshop presentation + project work
  - If you fail either component, you will fail the overall subject
- And of course you must get at least 50/100 overall
- Assessable content includes material from the lectures, workshops and assignments
  - During semester, will progressively release a study guide describing the key concepts to focus on

Subject Key Dates



2018 Week 1 23rd Jul 29th Jul	No workshops!	2018 Week 2 30th Jul 5th Aug		2018 Week 3 6th Aug 12th Aug	
2018 Week 4 13th Aug 19th Aug	Ph. 1 release: Mon-13th Aug 11:59 am	2018 Week 5 20th Aug 26th Aug		2018 Week 6 27th Aug 2nd Sep	Ph. 1 Due: Fri-31st Aug 11:59 am
2018 Week 7 3rd Sep 9th Sep	Ph. 2 release: Mon-3rd Sep 11:59 am	2018 Week 8 10th Sep 16th Sep	Ph. 3 release: Mon-10th Sep 11:59 am Guest Lect.# 1 Prof. Richard Sinnott Fri-14th Sep	2018 Week 9 17th Sep 23rd Sep	Ph. 2 Due: Fri-21st Sep 11:59 am
2018 Week 10 1st Oct 7th Oct	Guest Lect.# 2 Prof. James Bailey Fri-5th Oct Ph. 3-A Due: Fri-5th Oct 11:59 am	2018 Week 11 8th Oct 14th Oct	Guest Lect.# 3 Scott Thomson Wed-10th Oct Ph. 3 presentations during workshops	2018 Week 12 15th Oct 21st Oct	

Non Teaching period from 24th Sep - 30th Sep

Subject Workload

- Around 14 hours per week
  - Workshop (2 hours attendance + 2 hours follow up)
  - Lectures (2 hours attendance + 3 hours follow up)
  - Assignments (5 hours on average)

Getting Started

- Check out the LMS
  - www.lms.unimelb.edu.au
- Brush up on Python 3 (Jupyter Notebook)
- Revise the Maths notations and basics uploaded on LMS-week1
- Lecture slides, lectures recordings and code examples will be made available from the lectures/workshops page on the LMS
- Take a look at the discussion forum – please use for general questions and for project related questions

Getting Help

- Talk to the lecturer after the lecture
  - Office Hours: Wednesdays and Fridays after the lecture
- Post a question to the LMS forum
- Talk to your tutor/demonstrator during workshop time
- Consultation by appointment – send an email

Remember

- Never share any examinable code with your fellow students (not on the forums, not via email, not via shared machines,….)
- Review carefully the Academic Honesty section on the COMP20008 Academic Integrity page of the LMS.

Student volunteers

- We need 2 volunteers to act as “student representatives” for the subject, with the following responsibilities
  - Keep finger on pulse of the student body
  - (possibly) act as go-between between students and teaching staff
  - Attend a Staff-Student Liaison Committee meeting in the middle of semester to report on issues with the subject and run a feedback session immediately beforehand to poll the student body.
  - Email Yasmeen if you are interested

Before the next Lecture

- Don’t go to a workshop this week
- Check that you can access the LMS site
- Install Anaconda Python 3: Instructions on LMS-week1
- Read through the getting started materials on LMS-week1
- Read through next week’s workshop
- Read the following background articles on data wrangling
  - Six core data wrangling activities
    - http://www.datanami.com/2015/09/14/six-core-data-wrangling-activities/
  - Research directions in Data Wrangling: visualisations and transformations for crediible data. S. Kandel et al, Information Visualisation 10(4), 2011.
    - http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf
  - Data wrangling for big data: challenges and opportunities
    - https://openproceedings.org/2016/conf/edbt/paper-94.pdf