COMP20008 Elements of Data Processing

Semester 1 2017

James Bailey

Who?

- Lecturers
  - James Bailey (baileyj@unimelb.edu.au)
  - Office: DMD 7.09 (level 7 of Doug McDonell)

  - Elham Naghizade (e.naghi@unimelb.edu.au)

- James and Elham are 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 Elham, please start the subject line with COMP20008

My Background

- My background
  - Data mining
  - Machine learning
  - Databases
  - Data science …
  - Example projects
    - Health (medical emergency prediction, liver transplants,)
    - Education (student attrition in MOOCs, intelligent tutoring systems)
    - Bioinformatics (cell tracking, genomics)
    - …

Who?

- Tutors
  - Donia Malekian (head tutor)
    dmalekian@student.unimelb.edu.au
  - Qingyu Chen (tutor)
  - Florin Schimbinschi (tutor)

About You

- 137 students

- Most popular subjects concurrently studied
  - COMP20007 Design of Algorithms
  - SWEN20003 Object Oriented Software Development
  - INFO20003 Database Systems
  - MAST2004 Probability
  - MAST20026 Real Analysis

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, …)

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

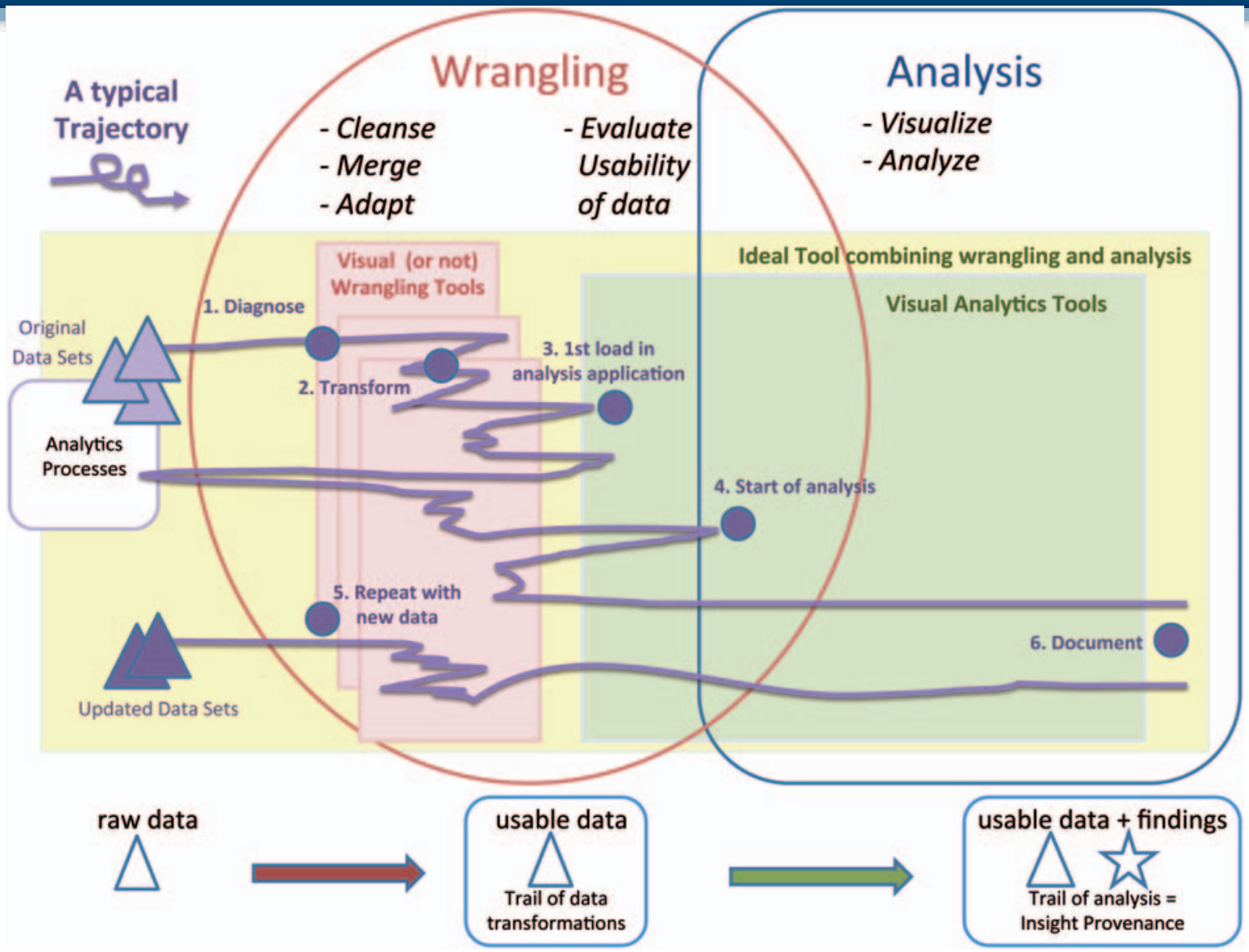
The McKinsey Quarterly, Jan 2009

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 ……



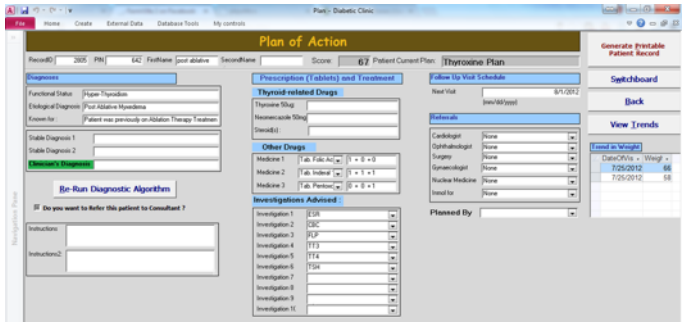
Research directions in Data Wrangling: visualisations and transformations for crediible data. S. Kandel et al, Information Visualisation 10(4), 2011.

Data and health

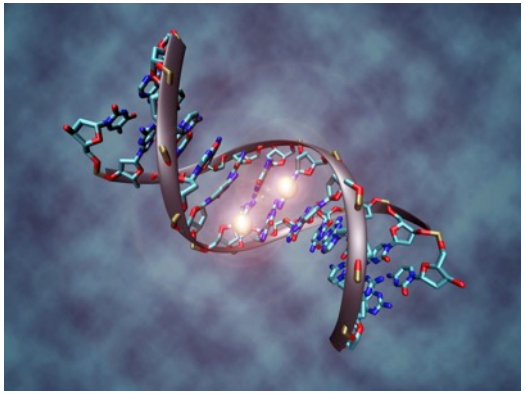
- Gene sequences
- Mobile data
- Electronic medical records
- Insurance claims
- Imaging results
- GP data
- Prescription data
- Social media
- …….

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

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



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

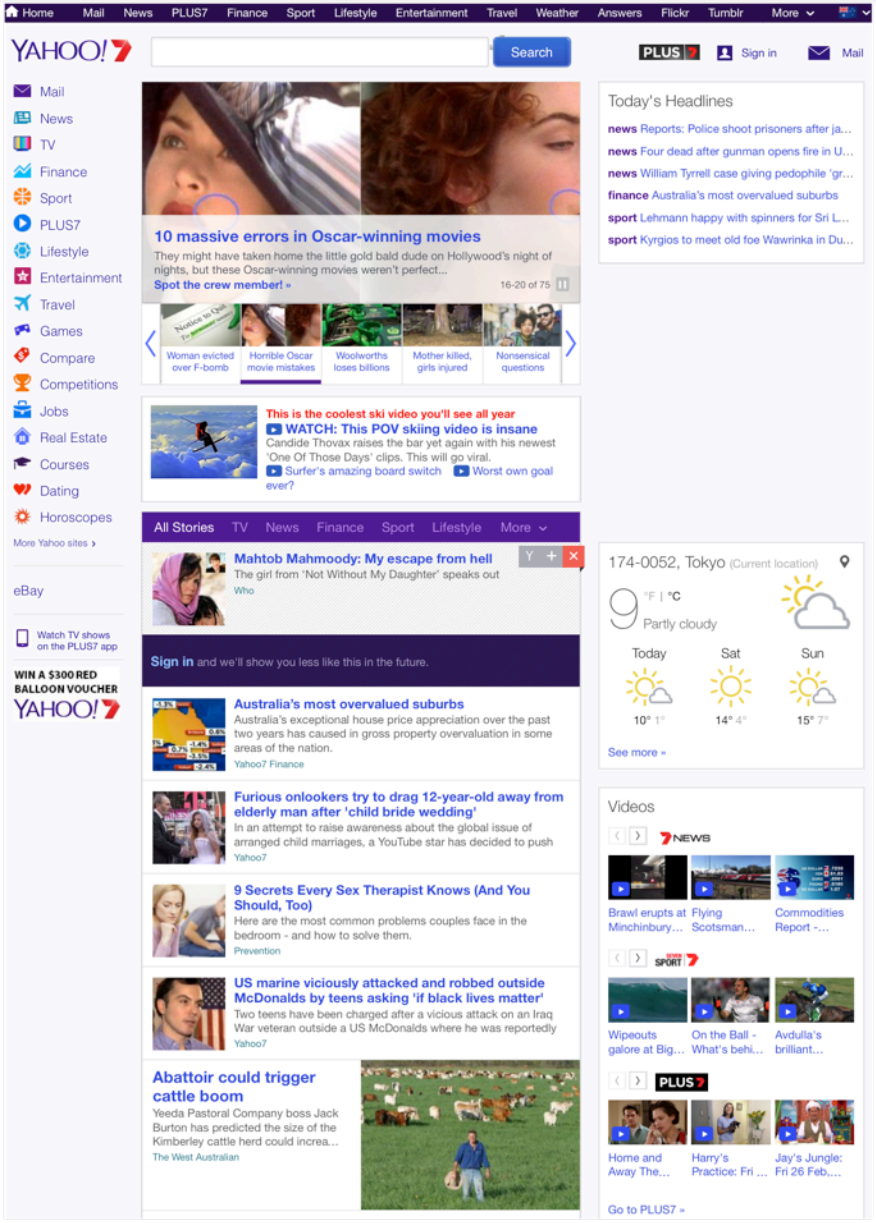


https://upload.wikimedia.org/wikipedia/commons/8/80/DNA_methylation.jpg

Data and government

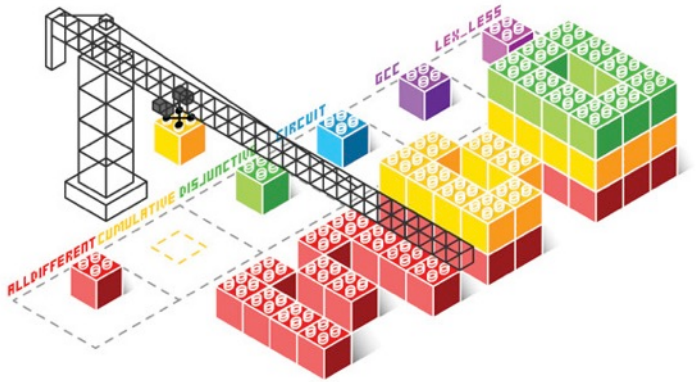
- www.data.vic.gov.au
  - 6122 datasets
- data.melbourne.vic.gov.au
  - 152 datasets
- AURIN (Australian Urban Research Infrastructure Network)
  - 1200 datasets
- data.gov.au
  - 15000 datasets

Data and business (au.yahoo.com)



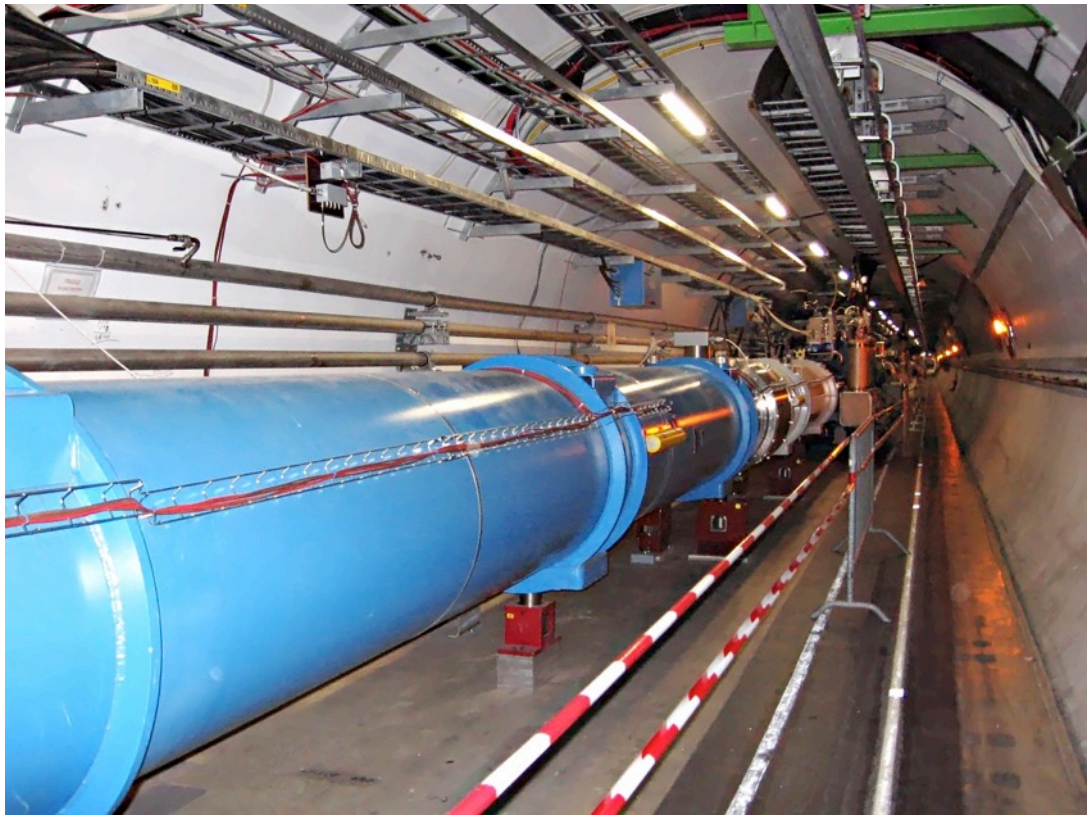
Data and massively online education

- MOOCs (>20 at Unimelb)
  - Video viewing behaviour
  - Quizzes
  - Discussion forum
  - Assignments
  - Interventions to improve learning



Modelling discrete optimisation

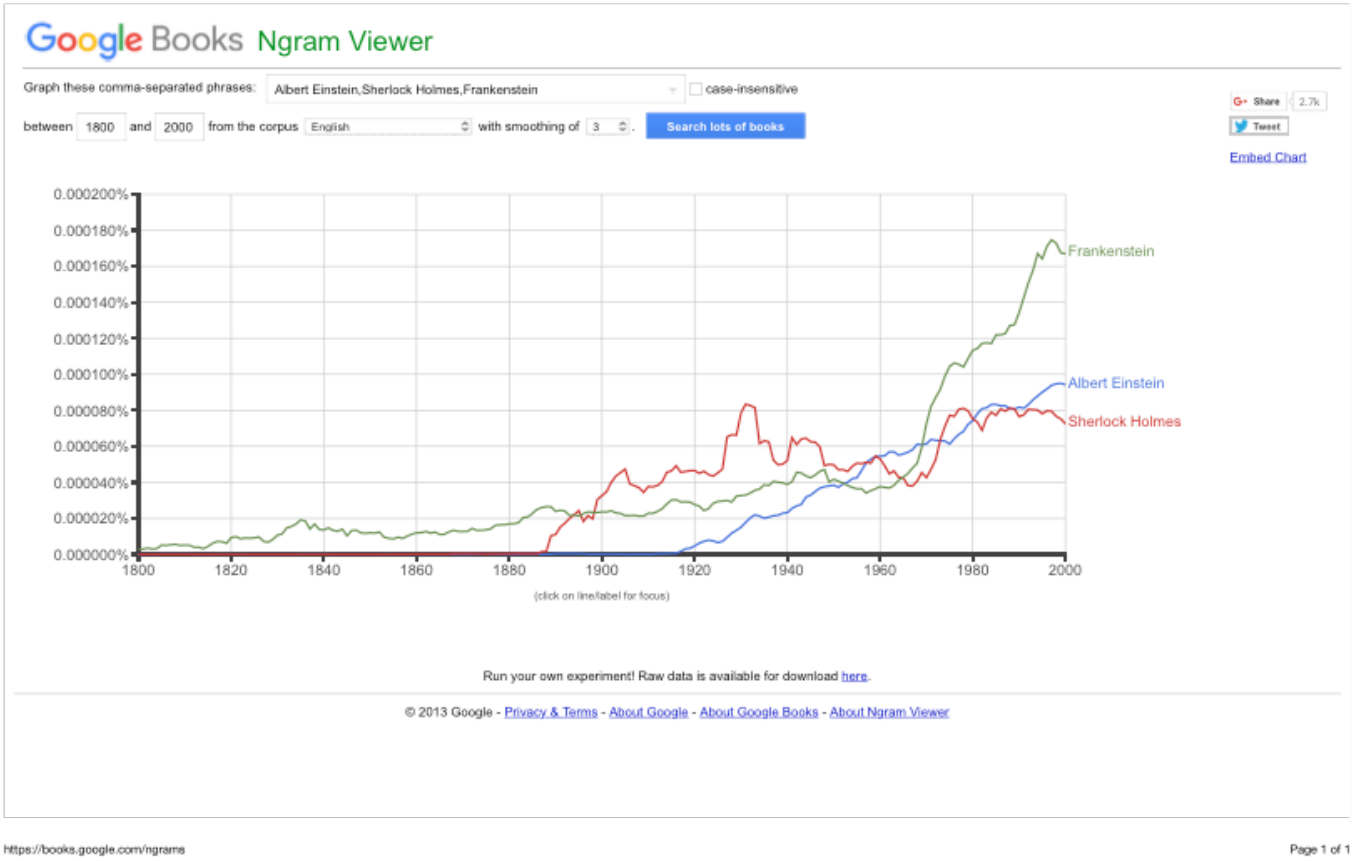
Data and physics



CERN

- Large hadron collider
- 1000 terabytes/second

Data and the humanities



- 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)

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** (5 lectures): Weeks 8-10
  - noSQL and cloud, 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, Good Friday holiday (no lecture), two guest lectures (Scott Thomson from Google, Richard Sinnott from CIS)

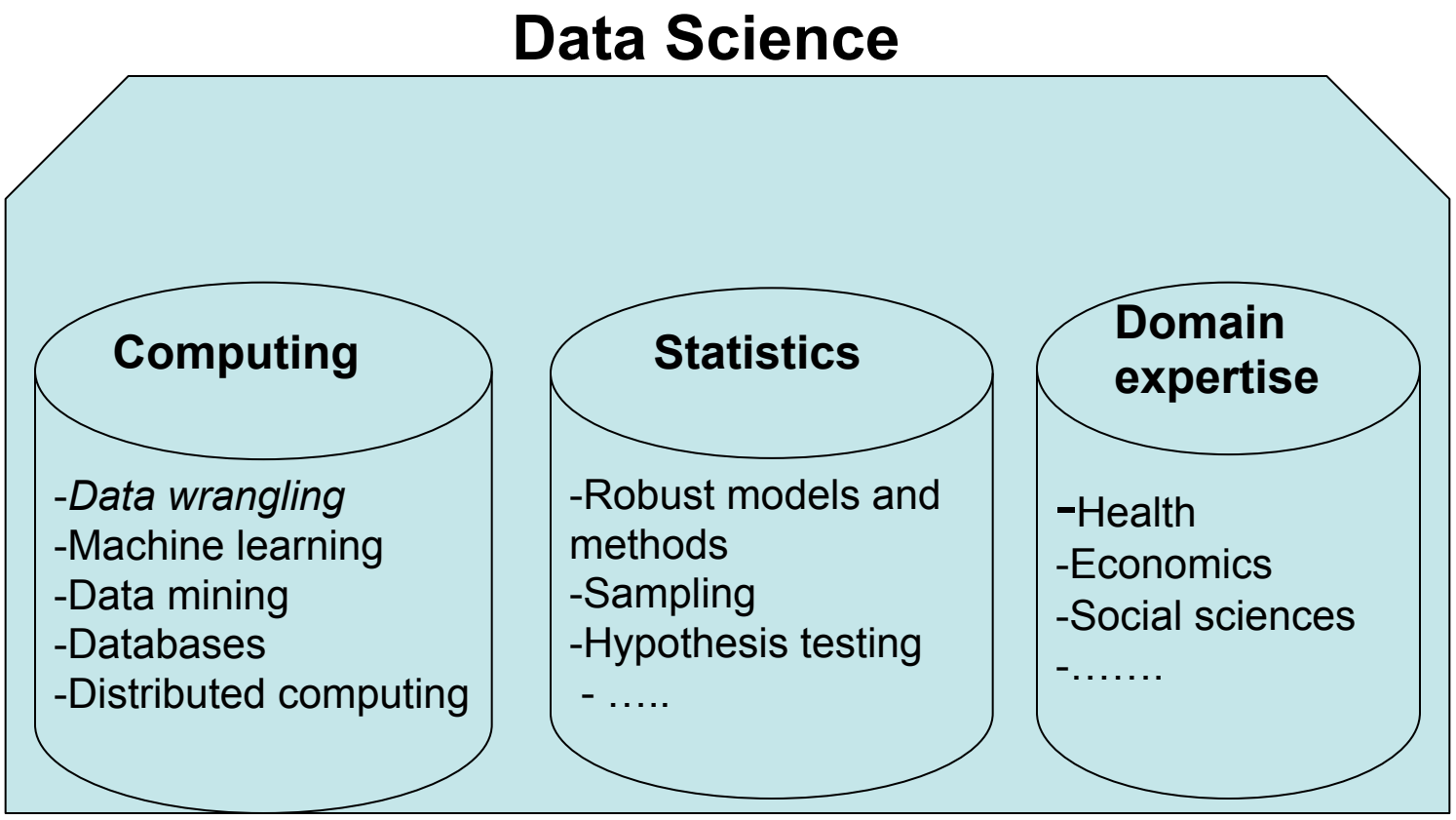
What's not covered?

- 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. There may be some slight overlap between COMP20008 and COMP30027
- Relational databases
  - See instead INFO20003 Database Systems

Level of the subject

- Assumes you have completed COMP10001 and COMP10002
  - Knowledge of programming (in Python) and algorithms
- Material will be pitched at *2nd year level*
- *You cannot gain credit for both COMP20008 and INFO20002 (Foundations of informatics)*
  - Only take **one** of these, **not both**

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”

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 for the first time last year (2016)
  - Will broadly follow last year's syllabus, but making a range of revisions to specifics of the lecture notes and workshops
  - Exam structure and difficulty will be similar

- 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 Monday (9:05-9:55)
    - 9:00-10:00 Friday (9:05-9:55)
    - Recorded using Lectopia
  - 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
  - For workshops in Alan-Gilbert-111, we will be using Python under the Mac OS X environment
  - For the workshop in Alice Hoy-236, we will use Python under Windows environment
  - 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

Python

- We will be using Python 3

- Get a copy of Python for your machine at home
  - http://www.python.org/download/
  - The Anaconda distribution can also be particularly convenient
    - https://www.continuum.io/downloads

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 (15%)
    - Phases 2-4: Data wrangling investigation on an open dataset (pick one you like and find patterns/insights)
      - Phase 2: Concept formulation and initial investigation (12%)
      - Phase 3: Report (13%)
      - Phase 4: Oral presentation in workshop (10%)

Assessment

- 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 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 (Python 3)
- 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

- Post a question to the LMS forum
- Talk to the lecturer after the lecture
- 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 Resources 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 subhect and run a feedback session immediately beforehand to poll the student body.
  - Email James 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 Python 3
- Read through next week's workshop (to be released Wednesday 1 March)
- 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