COMP20008
Elements of data processing

Semester 1 2020

Lecture 1: Introduction

Introduction

- Lecturers: Pauline Lin (coordinator), Chris Ewin, and Uwe Aickelin
- Head tutor: Anam Khan
- Online tutor: Abby Yuan
- Tutor team:

Abby (Meng) Yuan	Joel Ong	Sadegh Motallebi
Ali Qadar	Josh Nguyen	Sara Moghaddam
Anam Khan	Lianglu Pan	Shima Rashidi
Grady Fitzpatrick	Mateen Khan	Shreyasi Datta
Hangfan Li	Mengmeng Wang	Yujing Jiang
Jack Shee	Neil Marchant	

About you


When poll is active, respond at **PollEv.com/comp20008**
Text **COMP20008** to **+61 427 541 357** once to join

What is the subject?

- Formally, Elements of Data Processing
- 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. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Chief Economist at Google

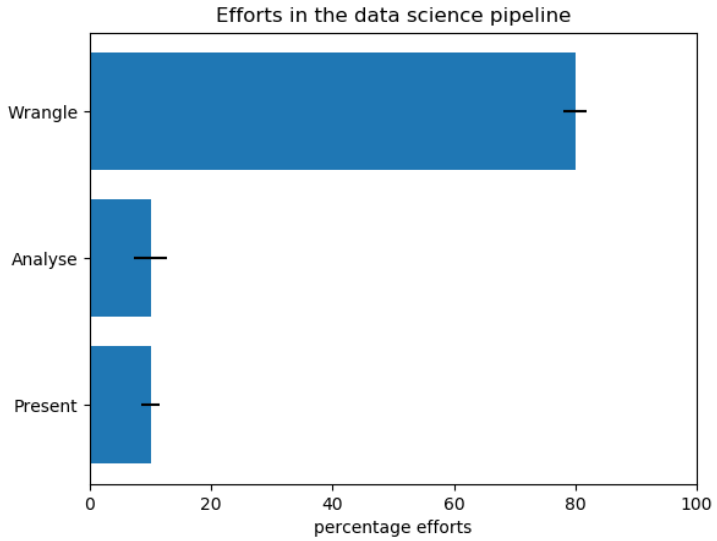
The McKinsey Quarterly, Jan 2009

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

Why is it important?

Data science

- Wrangle the data
- Analyse the data
- Present, deploy and communicate results



Why is it important

When we hear about data, we think this:

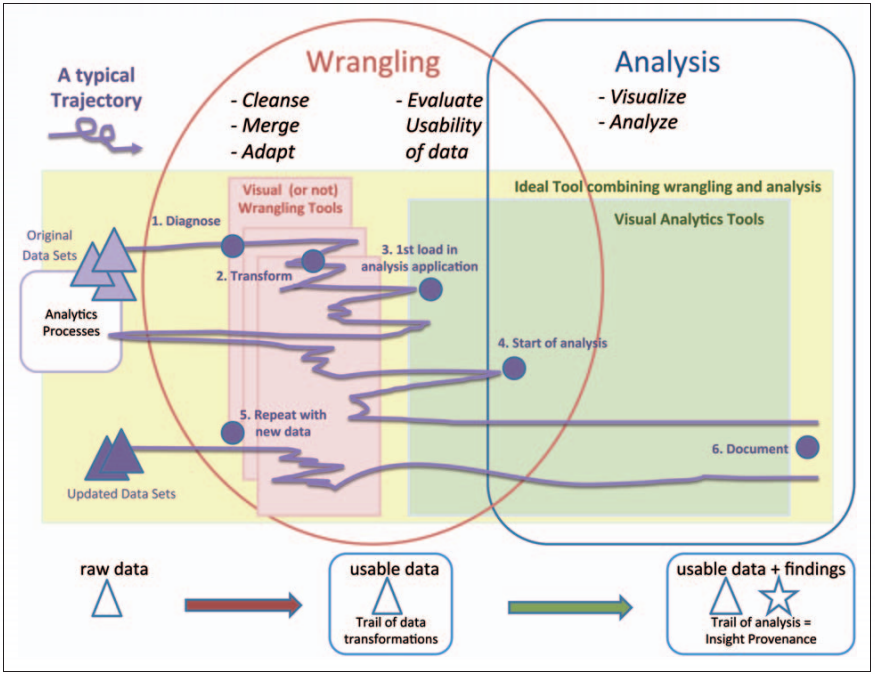
	A	B	C	D
1	customer_id	order_value	date_ordered	delivery_state
2	55321	129.99	11/02/2018	VIC
3	55165	32.88	13/02/2018	NSW
4	55235	99.00	15/02/2018	VIC
5	55121	88.98	17/02/2018	TAS
6	55193	52.00	18/02/2018	NSW
7	55113	158.65	18/02/2018	SA
8	55155	132.95	19/02/2018	VIC
9	55119	111.95	19/02/2018	QLD
10	55002	63.50	20/02/2018	QLD
11	55111	63.50	21/02/2018	VIC

Why is it important

In practice:

	A	B	C	D
1	customer_id	order_value	date_ordered	delivery_state
2	55321	129.99	11/02/2018	VIC
3	55165	32.88		NSW
4		99.00	15/02/2018	VIC
5		88.98	17/02/2018	TAS
6	55193	$52.00	18/02/2018	NSW
7	55113	15865	18/02/2001	SA
8	55155	132.95	19/02/2018	Victoria
9	55119	111.95	4.32E+04	QLD
10	-1	63.50	02/20/2018	N/A
11	55111	0.00	21/02/2018	VIC

Why is it important

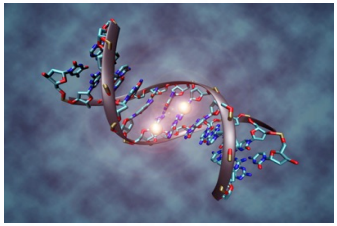


Research directions in Data Wrangling:
visualisations and transformations for crediible data. S. Kandel et al, Information Visualisation 10(4), 2011.

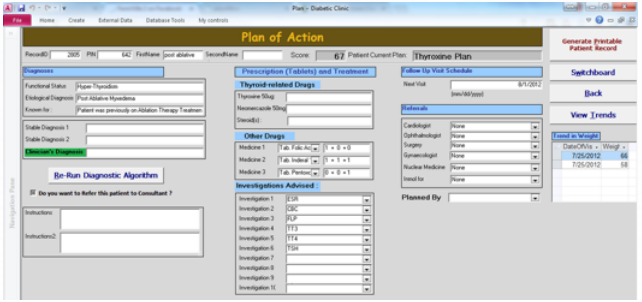
Data and health

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

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



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



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

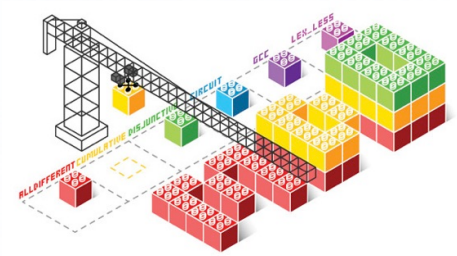
https://upload.wikimedia.org/wikipedia/commons/c/ca/Fitibit_Flex.jpg

Data and business (au.yahoo.com)

Data and education

massively online education

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



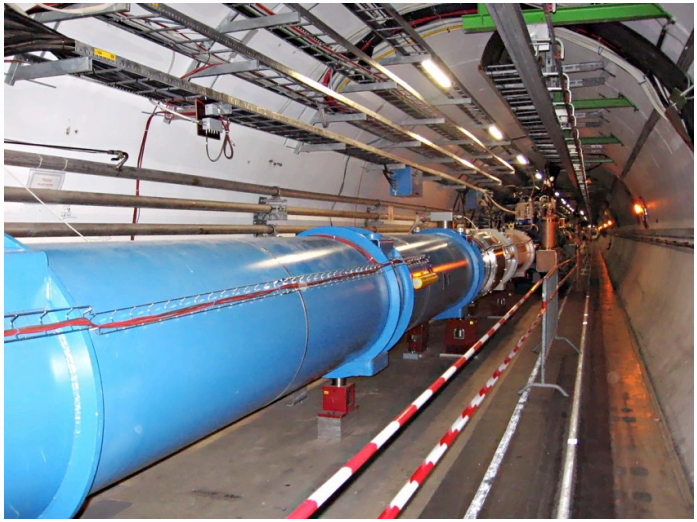
Modelling discrete optimisation

Data and science

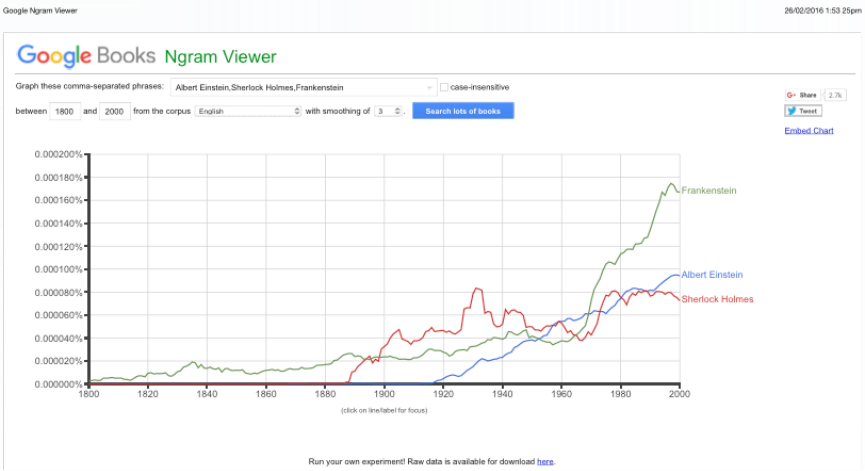
CERN

Large hadron collider

1000 terabytes/second



Data and 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 food

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

Data and government

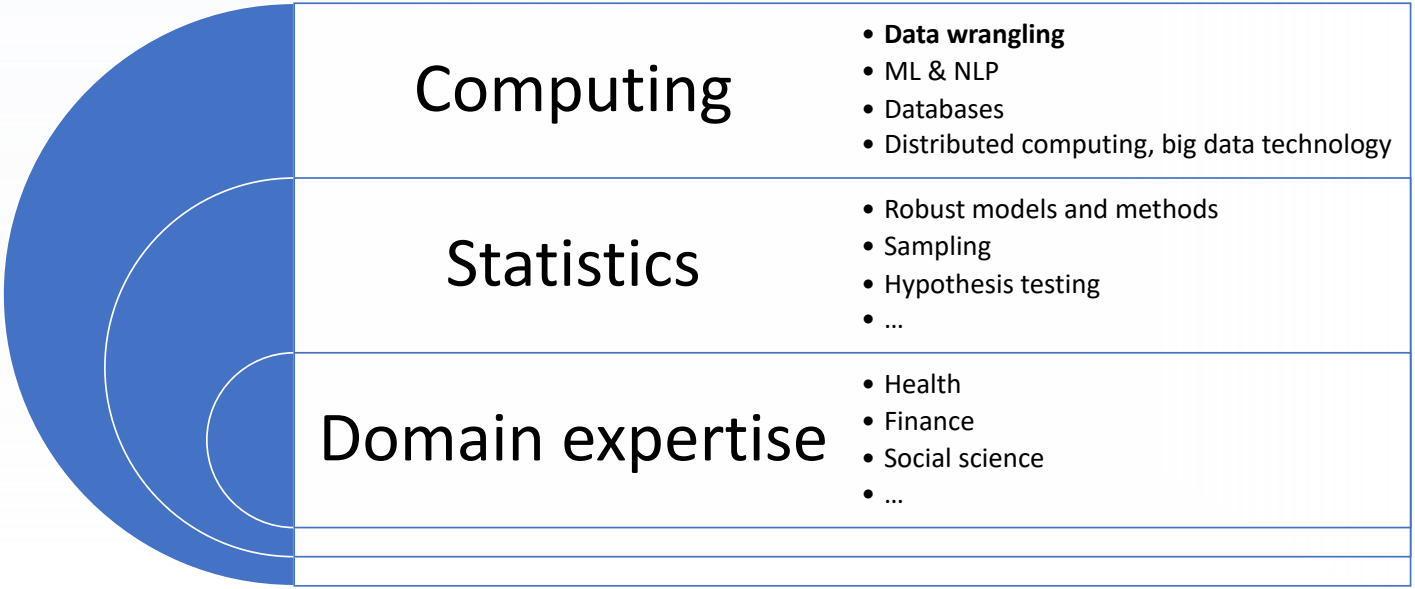
- www.data.vic.gov.au
- data.melbourne.vic.gov.au
- AURIN (Australian Urban Research Infrastructure Network)
- data.gov.au

Data and …

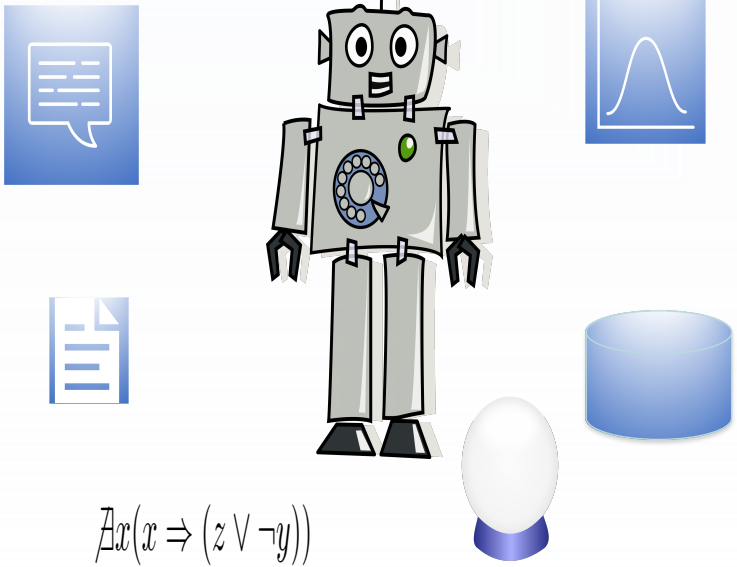
Data science

- “In God we trust. All others must bring data.” ~ W. Edwards Deming, statistician
- “Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.” ~ Josh Wills, Director of Data Engineering at Slack
- “Torture the data, and it will confess to anything.” ~ Ronald Coase, Economics, Nobel Prize Laureate
- “Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.” - Mike Loukides, VP, O’Reilly Media.
- https://www.youtube.com/watch?v=jbkSRLYSojo

Data science landscape



AI—“Intelligent machines and software”

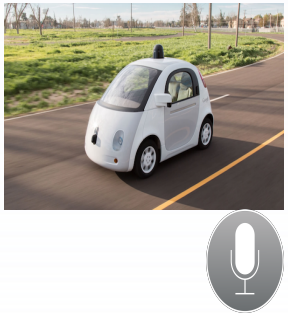


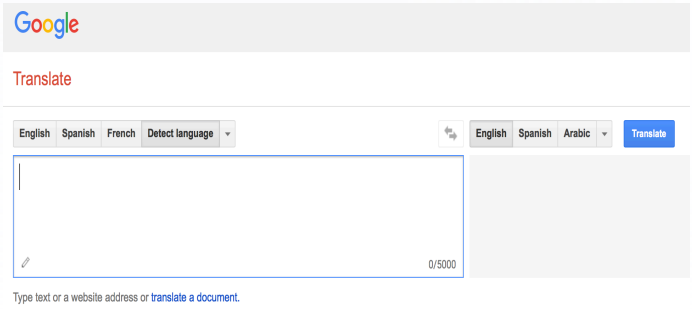
$$\exists x(x \Rightarrow (z \vee \neg y))$$

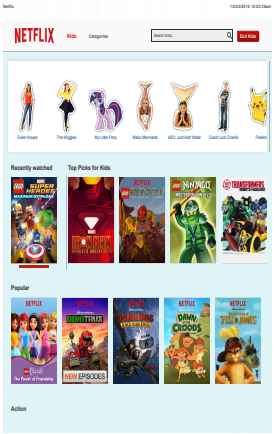
AI – data-driven, intelligent systems

"AI is the new electricity" – Andrew Ng

How would you like AI to change your life in 2030?







What is the subject all about?

Preprocessing & visualisation (6 lectures): Weeks 1 – 5

- Data types and data formats
- Visualization methods with outlier detection
- Preprocessing, data exploration, data cleaning including missing data
- Linkage and integration – managing relationships between multiple datasets
- **Week 6, no lectures**

Analysis (introductory) (8 lectures): Weeks 7 – 10

- Correlations
- Basic prediction techniques
- Feature engineering & dimensionality reduction

Social, ethical and privacy issues (3 lectures): Weeks 11 –12

- K-anonymity, l-diversity,
- Privacy & ethics

COMP20008 is a building block for data science and artificial intelligence

What is not covered?

Advanced analysis of data

Advanced predictive analytics, machine learning, information retrieval & search engines and data mining algorithms

- COMP30027 Machine Learning

Relational databases

- INFO20003 Database Systems

Level and expectations

- You have completed COMP10001; knowledge of Python
- Will rely on using a range of basic mathematics for some topics
  - logarithms, powers, distances, mean, median, etc.
- Subject is at 2nd year level
  - More independence
  - Will need to be comfortable looking up documentation for Python library functions

Programming for data wrangling

- Python programming in JupyterLab (we will be using)
  - Python
    - Good for combining data wrangling activities into a larger pipeline
    - Good library support for scientific and machine learning extensions.
  - JupyterLab
    - “next-generation web-based user interface for Project Jupyter.”
    - “JupyterLab enables you to work with documents and activities such as Jupyter notebooks, text editors, terminals, and custom components in a flexible, integrated, and extensible manner”
    - https://jupyterlab.readthedocs.io/en/stable/getting_started/overview.html

Lectures and workshops

- Lectures
  - For presentation of principles
  - 2 lectures Friday 9 – 10am and 10 – 11am
  - Available in LMS
- Workshops (one per week)
  - A mixture of practical programming exercises and tutorial
  - Start in week 1 – purpose is to set up environment and revise Python
  - Please come prepared
- Online workshops (supporting students affected by the travel ban)
  - 4 Zoom workshops available

Subject support

- Canvas LMS
  - Official announcements, e.g. assessment specifications
  - Supply of materials, e.g. lectures, workshop materials
- Discussion forum:
  - We will use '**Ed**'; will invite you all to join in week 1
  - Ask (and answer) questions here – preferred channel
- Consultation with head tutor
  - Monday 10:00-11:00 am 7.02 DMD
- Consultation by appointment
  - Send an email

Student representatives

- We need 2 volunteers to act as “student representatives” for the subject, with the following responsibilities
  - Be aware of issues and concerns affecting student learning
  - Raise concerns with staff on behalf of all students
  - 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 Anam Khan (anam.khan@unimelb.edu.au) if you are interested

Please contact them with any feedback you have about the subject

Assessments

Subject mark will be made up of

- Final exam: 50%
- Project work during semester: 50%
  - Programming project 1: 20%
  - Programming project 2: 20%
  - Oral presentation: 10%

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

Reminder

- 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.

Textbooks?

- None !
  - No single textbook includes all topics we cover
  - Material needed will be covered in the lectures and the references provided
- There are several 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.

Background reading

- 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