00. Introduction

1. Data wrangling

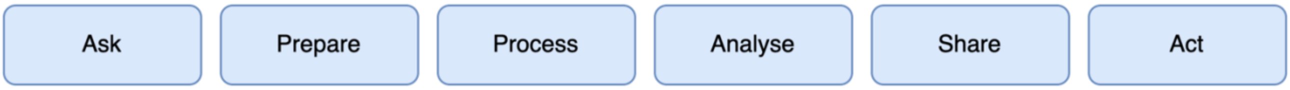
Data wrangling captures the activities performed on digital data to make them more usable or valuable for downstream consumption. Synonyms for wrangling include munging, cleaning, scrubbing, aggregation, integration, profiling, ETL/ELT, transformation and processing.

Data from source rarely, if ever, is in the desired format or contains no errors. Data wrangling is necessary to ensure data is fit for downstream analysis or visualisation.

Data wrangling is underappreciated and often unknown to consumers of data. It takes up most of the time when working with data, yet it gets the least amount of glory. Data wrangling is like janitorial work. It is a valuable skill essential to all data project because it brings value to the data.

2. The data life cycle

A generic data life cycle follows these phases:



Ask

Prepare

Process

Analyse

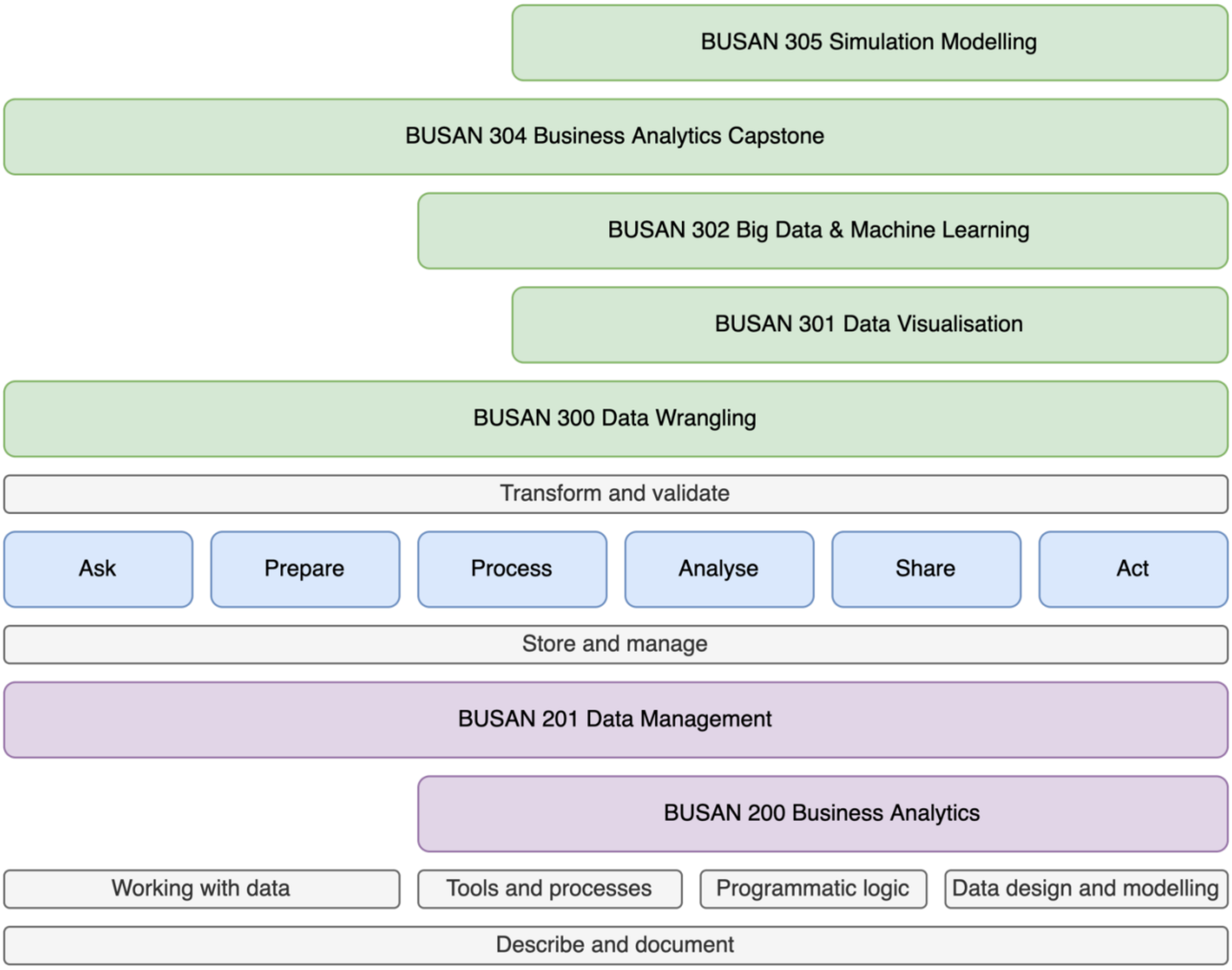
Share

Act

Early in the life cycle, data is raw and considered less valuable and usable. Moving to the end of the life cycle, data is refined and it could be transformed into a more valuable and usable form (e.g. ready to be fed into production systems that use data in an automated fashion).

Transform and validate	Store and manage	Describe and document	Ask: understand the problem
			Prepare: identify and acquire the data from its source
			Process: clean and transform the data for its intended use (aka data wrangling)
			Analyse: harvest knowledge and insight from the processed data
			Share: communicate and distribute the harvested knowledge and insight
			Act: provide recommendation to solve the problem and transform data for future use

The structure of the business analytics major roughly follows this life cycle:


BUSAN 305 Simulation Modelling
BUSAN 304 Business Analytics Capstone
BUSAN 302 Big Data & Machine Learning
BUSAN 301 Data Visualisation
BUSAN 300 Data Wrangling
Transform and validate
Ask
Prepare
Process
Analyse
Share
Act
Store and manage
BUSAN 201 Data Management
BUSAN 200 Business Analytics
Working with data
Tools and processes
Programmatic logic
Data design and modelling
Describe and document

Stage II courses are in purple; stage III courses are in green; grey boxes represent skills, tools and techniques used throughout the entire life cycle.

2.1 Exercise / Discussion

Could you source, wrangle, analyse, interpret, visualise and communicate:

- Your last 5 years' spending habits?
- Your academic performance from first year of university until now?
- The historic NZ CPI trends?
- Your weekly schedule?
- Any other data you have access to?

3. Focus of this course

3.1 Content

This course covers 4 major topics in data wrangling:

01. Tabular data
    - Structured; spreadsheet format; columns/rows; DSV; text-based; Excel as primary environment/tool
02. Web data

  - Semi-structured; XML and JSON format; human-readable machine-consumed; text-based; basic retrieval; basic queries

03. Big data
    - 5Vs; semi-structured; NoSQL databases; MongoDB as primary environment/tool; basic queries

04. Wrangling text with pattern matching
    - Regular expressions (aka RegEx)

3.2 Data format

Digital data exists in many formats. Common formats for structured data are:

- relational databases—covered in prerequisite courses
- tabular (spreadsheets; column-row format)—covered in BUSAN 201 and this course

Common formats for semi-structured data are:

- plain text files—introduced in BUSAN 201
- XML—introduced in BUSAN 201 and revisited in this course
- JSON—introduced in BUSAN 201 and revisited in this course
- non-relational (NoSQL) databases—covered in this course

Unstructured data include raw text files, images, videos, and audio files—introduced in BUSAN 201.

Background reading for students who have not completed BUSAN 201 is outlined in Canvas.

4. Course software

4.1 Microsoft Excel

To work with tabular data, we will use **Microsoft Excel**. It is a ubiquitous business tool. The techniques covered in this course, though specifically taught using Excel, are transferrable to other equivalently powerful spreadsheet software, e.g. Google Sheets.

UoA students are entitled to a free Microsoft Office 365 licence (which includes Excel). Details here: https://www.library.auckland.ac.nz/services/it-essentials/computer-facilities/software-purchase-or-download/office-365

While Excel runs on both platforms, the Windows version is far superior feature- and performance-wise. The macOS version still lags in feature-parity compared with the Windows version. Having said that, students with Mac will not be disadvantaged in this course from an assessment point of view.

4.2 MongoDB

To work with big data, we will use **MongoDB** and **Robo3T**. MongoDB is chosen for its popularity of use in industry. Alternative NoSQL databases to MongoDB have similar features and aims, but their usage is unique. While techniques covered in this course will be specific to MongoDB, the concepts are transferrable when dealing with any semi-structured/NoSQL data.

MongoDB uses client-server architecture. The server is called “MongoDB”. The client that comes with MongoDB is called “MongoDB Shell” and is not student-friendly since it is command-line based. The recommended MongoDB client for this course is Robo3T.

4.3 Text editor

To work with plain text, any text editor that supports regular expressions will suffice.

- VSCode or VSCodium
- Atom or Notepad++

4.4 XPath query environment

To run XPath queries students could use **VSCode/VSCodium** with an addon called **XML Tools**.

- VSCode/VSCodium with XMLTools addon
- When using the OGGB PCs, you will need to install the XML Tools addon every time you log in.

4.5 Other tools listed in the course outline

There are other tools listed in the course outline: Firefox/Chrome Developer Tools and Command Prompt / Terminal. Details will be provided when they are used.

4.6 Optional: clipboard manager

A clipboard manager is a tool that remembers multiple text clippings (copied with Ctrl/⌘+c). They are helpful for data wrangling-type tasks:

- ClipX or Flycut

5. Before the next lecture

Have the software installed, ready to use in lecture:

- Microsoft Excel
- VSCode/VSCodium and/or Atom/Notepad++

6. Document change history

v1.0 2022-01-25

- Initial release