DMV
Putting it all together

suzanne.little@dcu.ie

Today

What have we learnt about Data Management and Visualisation?

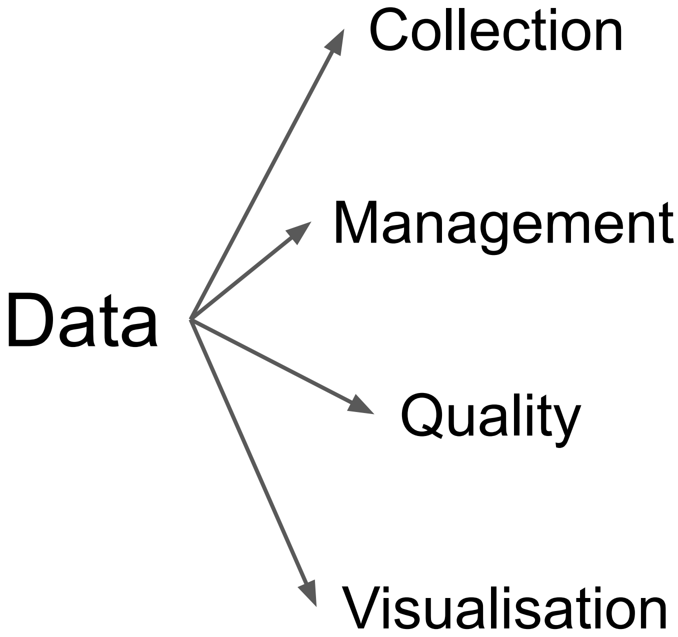
Where does all this fit in a Data Management Lifecycle or Pipeline?

What tools/skills should you now be aware of?

The exam

Module Course Outline

Module Content Outline



Data

Collection

Management

Quality

Visualisation

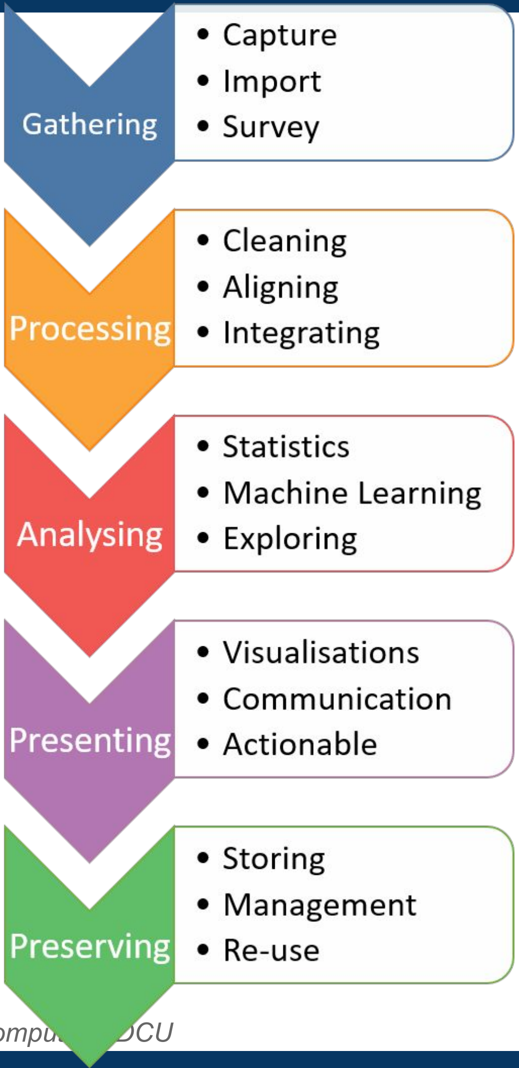
Outcomes

1. Analyse the requirements of applications handling large datasets.
2. Demonstrate an ability to efficiently structure a large dataset.
3. Implement data quality measures.
4. Identify and implement appropriate data visualization techniques.

Module Content Outline

1. Introduction : A Generic Pipeline
2. What is data?
3. Describing data
4. Cleaning & finding data
5. Data visualisation: Communication
6. Data visualisation: Making charts
7. Data visualisation: Humans
8. Data visualisation: Design
9. Storing data
10. Data privacy

Data Analytics Pipeline



Gathering
- Capture
- Import
- Survey

Processing
- Cleaning
- Aligning
- Integrating

Analysing
- Statistics
- Machine Learning
- Exploring

Presenting
- Visualisations
- Communication
- Actionable

Preserving
- Storing
- Management
- Re-use

Open Data, Web services (REST)
Formats like CSV, JSON

Spreadsheets (Excel, GoogleSheets)
GoogleRefine (OpenRefine)
Python (Juypter notebooks)

Exploratory visualisation - Jupyter,
Tableau, Spreadsheets

Python: Matplotlib, Bokeh or Seaborn

Relational (e.g, SQL based)
Document or NoSQL (e.g., MongoDB)
Linked Data & Metadata

Data Management and Visualisation <u>Skills</u>

Data Skills

Understand the pipeline or life cycle of data analytics

Deconstruct a dataset:

- what is the format of the data?
- what are the types of the attributes/columns?
- where might errors have crept in?
- what metadata is available?

Data Management Skills

Plan how a dataset could be stored - what questions need to be asked, what technologies or computer apps may be suitable (uses of different database technologies)

Understand the concept of data quality - how can it be measured, what causes bad data, how can processes be managed to improve data quality

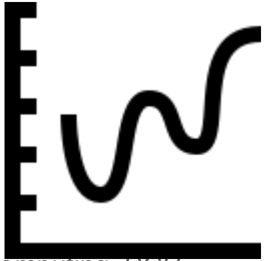
Methods for data cleaning, use tools to find data errors or artefacts

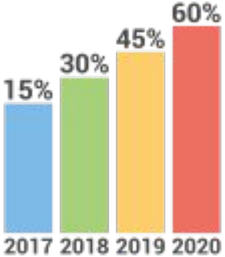
Understand basic concepts for data protection & privacy

Data Visualisation skills

Plan a visualisation - choose a graph type, understand communication and human perception, use design “rules” for elements like colour, layout, perception, attention

Read, evaluate & deconstruct a visualisation - critically assess how communication, human perception and design rules have (or have not!) been used





The Exam

On campus, online exam

- Three hours
- Online, Open Book exam hosted on loop in the labs
- The exam is worth 75% of the module total and will be marked out of 100
- 10-15 questions on loop quiz (40%)
- 4 longer questions (choose 3 to answer) in a word document (.docx), download & enter your answers in the spaces indicated
- Upload a single document, word or pdf format to loop

IMPORTANT logistics

- Know your login details for the school of computing labs! Check beforehand.
- Windows or Linux? Use windows for consistency.
- Come early to the exam so you can get settled and logged in as soon as you are allowed to. First login can take a bit longer as your profile is configured.
- You may bring pen & paper but you can’t use your own computer or mobile device.
- The exam is invigilated. No communication of any kind.

Open book?

You can access any of the class notes on loop

You can refer to any online materials

You can (try!) to google for answers (but that’s a waste of time)

You cannot “phone a friend” - no email, no chat, slack, discord, message boards, irc, shared documents, morse code, semaphore or any other form of communication.

No use of ChatGPT or similar

Content?

- 10-15 short questions on loop (multichoice, choose most appropriate, match the topics) → 40%
- 4 longer questions in a word document that you download. Choose any 3 to answer. If you answer all 4 then the lowest mark will be discarded → 60%
- A sample quiz is available now on loop plus a copy of the 2020/2021 exam paper (in a more traditional format).

Answering the longer questions (not essays!)

The computers will have Excel, Word, Powerpoint, OpenRefine, Python (Jupyter), R and you can access web-based tools like Google Sheets, Colab etc. This should be sufficient for the questions.

You may need to: download a dataset and find some errors/artefacts; suggest a visualisation method; sketch some ideas; critique a visualisation; or identify data characteristics.

You won’t be asked to do anything bigger than what was expected in the labs.

Recall the types of discussion activities that we had in class.

Academic Integrity

All students are expected to abide by DCU’s Academic Integrity and Plagiarism Policy

By submitting an exam, in addition to not using unauthorised or unreferenced material, you declare

1. that all of the work is your own;
2. that you did not seek whole or partial solutions -- or any other input -- for any part of your submission from others; and
3. that you did not and will not discuss, exchange, share, or publish complete or partial solutions for this exam or any part of it during the exam session.

It is important to note that both providing and receiving material can be breaches of academic integrity.

Don’t copy content from web pages for the exam unless you also provide a reference

Exams at DCU

Bring your student card, check your room number

For CSC1143/DMV only you need your phone for 2FA if you want to access the loop materials but it must be silent and face-down on the desk once you’ve logged in

Be early so you can be seated in time. Exam will finish at 12:30 regardless of when you start

Problems? https://www.dcustudentlife.ie/help-support/exam-support

Some tips ….

Read questions carefully & check the marks available

If a question asks to suggest **a** method, don’t list them all!

If a question asks for the **most suitable**, don’t overthink it!

Bullet points are acceptable (and preferred)

State any assumptions you make when giving your answer

Resources

- Documents, slides & videos on loop
  - Especially the final slide (usually) titled “Resources”
  - Eg. “Introduction to Metadata”, Tony Gill, Anne J. Gilliland, Maureen Whalen, and Mary S. Woodley http://www.getty.edu/research/publications/electronic_publications/intrometadata/ - Only the chapters on “Metadata and the Web” and “Practical Principals for Metadata Creation and Maintenance”
- DCU library ebooks

Books or eBooks

Data Types: Principles of Data Science, Sinan Ozdemir (2016), Chapters 2,

Data Cleaning: Data Mining and Predictive Analytics (Chapter 2, ignore sections on normalisation), https://ebookcentral-proquest-com.dcu.idm.oclc.org/lib/dcu/reader.action?ppg=52&docID=1895687&tm=1539596548831

“The Data Science Handbook”, Field Cady (2017), ebook, Chapters 1, 2, 4, 5 & 12.

Books or eBooks

Data Visualisation

- John Dimarco, “Digital Design for Print and Web An Introduction to Theory, Principles, and Techniques”, (Part 1 only),
- Andy Kirk, “Data Visualisation - a successful design process” (2012)
- Cole Nussbaumer Knaflic, “Storytelling with data”,

How to study for DMV?

- Review the materials -- know where to find information for each topic or skill
  - Create your own “structure” of the content (index)
- Review (or complete) lab exercises to practise skills
- Practise basics of tools like python, spreadsheets or similar

Finally ...

“... data visualization is not an exact science. There is rarely, if ever, a single right answer or single best solution. It is much more about using **heuristic methods to determine the most satisfactory solutions**.”

p20, Andy Kirk, Data Visualisation - a successful design process (2012)

Finally ...

Keep learning

- Follow people on Twitter/X (@DatavisDaily, @visualisingdata, ...)
- Read and criticise graphs and charts in online publications
- Learn a new tool or library (Tableau, PowerBI, online tools, ...)
- Explore Kaggle or Datacamp
- Practise!