CSC1158
Data Processing & Visualisation

Suzanne Little
suzanne.little@dcu.ie

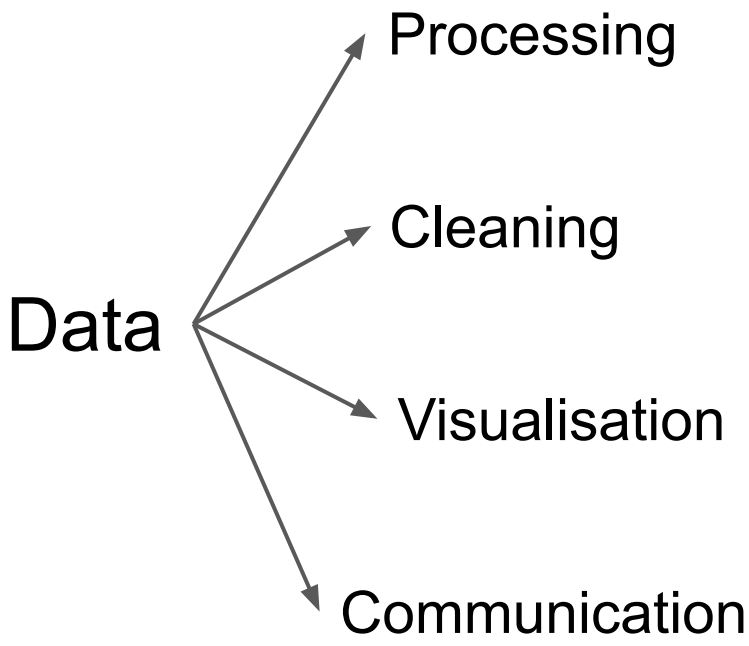
Today

- Module Outline
- Assessment
- Resources and Support
- Content Overview
- A Data Analytics Pipeline
- Git

CA273? CA273A? DS?

CSC1158 or DPV

CA273 Data Processing & Visualisation



CA273 Topics (probable order over 12 weeks)

1. What is data?
2. Where is data?
3. Manipulating & importing data
4. Cleaning & documenting data
5. Big data & the Internet
6. Visualisation: Communication
7. Visualisation: Encoding data
8. Visualisation: Design
9. Data wrangling
10. Bringing it all together

Skills/Tools

- Git, Markdown
- Python (pandas, numpy)
- Jupyter notebooks
- Scripting (bash)
- Spreadsheets
- HTML & scraping data
- Creating graphs and charts

How will DPV be delivered?

2 x 2hr lab-based session per week (with some exceptions)

Very practical but some important background concepts

Materials provided on loop and via gitlab.computing.dcu.ie

100% CA

Exercises and practise via datacamp

Remember this is a 7.5 credit module so you'll need to work on your skills outside of the 4 hours per week

Assessment - 100% continuous assessment

A1	Git project - clone, edit, commit a skill report on DataCamp work	gitlab	Tue Sep 24th 6pm	9%
A2	Quiz based on content from weeks 1-3	loop	Thu Oct 3rd 10am (during class)	12%
A3	Data cleaning assignment, commit to gitlab	gitlab	Fri Oct 18th 5pm	12%
A4	Quiz based on content from weeks 5-8	loop	Thu Nov 7th 10am (during class)	12%
A5	Visualisation Assignment & Presentation	upload	Fri Nov 29th 23:59	20%
A6	Visualisation Critique Assignment	upload	Dec 6th 09:00	20%
A7	Skill review and report, commit to gitlab	gitlab	Dec 20th 09:00	15%

Some tips for CSC1158

It's intentionally very practical but also designed for you to develop ways to **"teach yourself"**.

Try to **practise** a little bit outside the labs every week -- exercises, assessment, datacamp.

I want you to **become faster** at manipulating data and files. Can you do things without using the mouse? Learn the shortcuts.

There's also some background on operating systems, files and the Internet.

DW & Data Mining (Mark Roantree) is "front-loaded", DPV is therefore lighter for the first few weeks with more assessment at the end.

Resources and Support

Loop and gitlab (gitlab.computing.dcu.ie)

- main source of communication, notes, resources
- check your student email!

DCU Library

- Online ebooks (see next slide and Loop for list)

Myself

- Email is best for contacting me (suzanne.little@dcu.ie)
- Please include [CSC1158] or [DPV] in the subject

Resources

Books available in the DCU library (online ebooks):

- “Principles of Data Science”, Sinan Ozdemir (2024)
- “The Data Science Handbook”, Field Cady (2017)
- Kirk: Data Visualisation

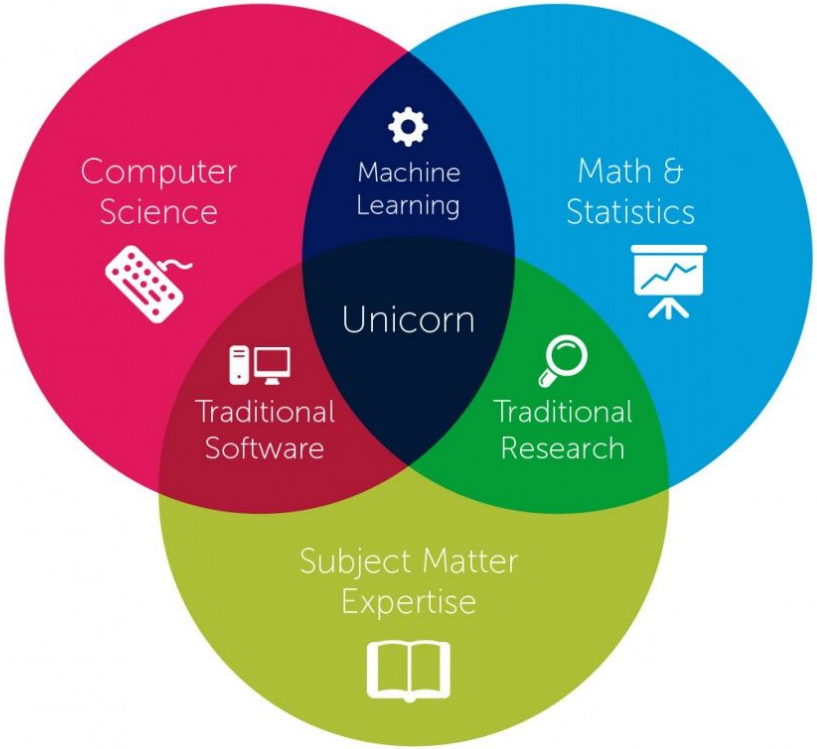
Monitor: https://www.packtpub.com/free-learning

DataCamp: http://datacamp.com

Academic license gives you free access for 6 months

Questions?

Data Science



https://indico.io/blog/data-science-buzzwords-demystified/

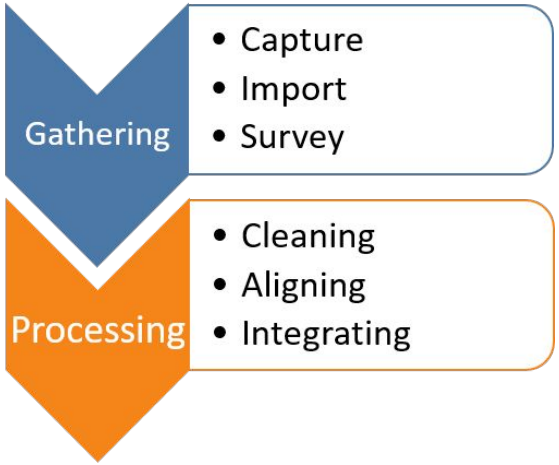
Surveys suggest that **80%** of effort in a data analytics project is in

finding, cleaning and reorganising

https://www.infoworld.com/article/3228245/data-science/the-80-20-data-science-dilemma.html

Today

- Module Outline
- Assessment
- Resources and Support
- Content Overview
- A Data Analytics Pipeline
- Data types
- First task



Sources: websites, user surveys, sensors,

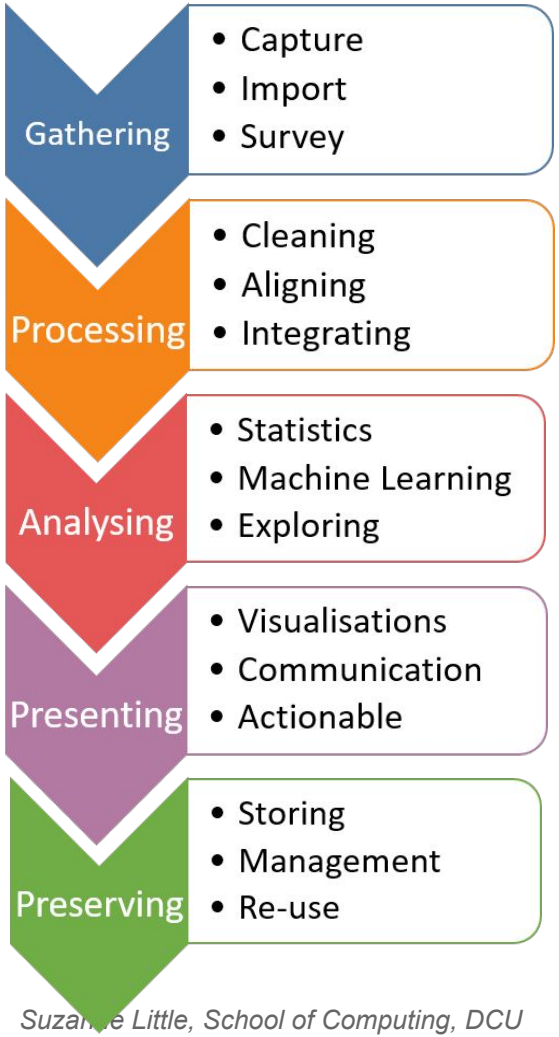
computers are simple,
people are complicated

Gathering
- Capture
- Import
- Survey

Processing
- Cleaning
- Aligning
- Integrating

Sources: websites, user surveys, sensors,

https://www.youtube.com/watch?v=TkR9QWuoyJ8



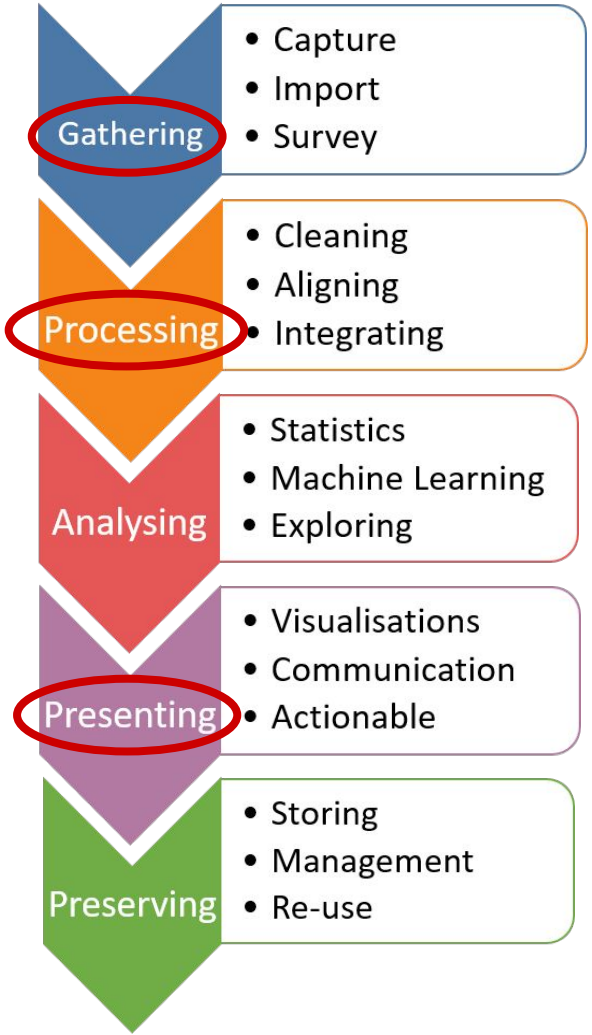
Sources: websites, user surveys, sensors,

computers are simple,
people are complicated

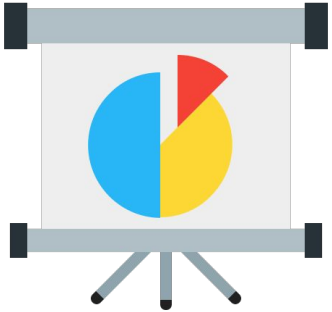
number crunching, discovering patterns

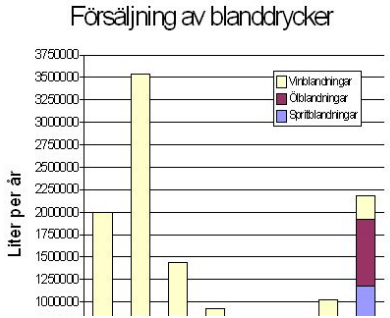
What is your message?
Who is your audience?

Data storage, libraries (indexing),
recording your processes

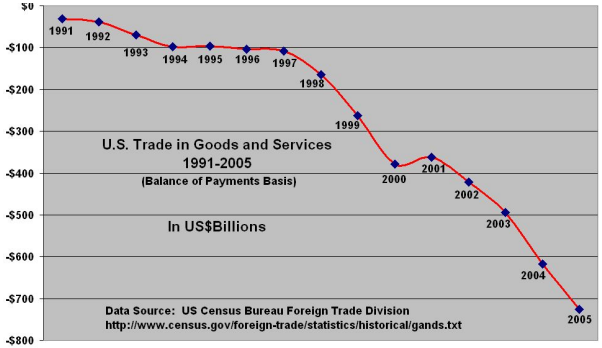

Gathering
• Capture
• Import
• Survey
Processing
• Cleaning
• Aligning
• Integrating
Analysing
• Statistics
• Machine Learning
• Exploring
Presenting
• Visualisations
• Communication
• Actionable
Preserving
• Storing
• Management
• Re-use

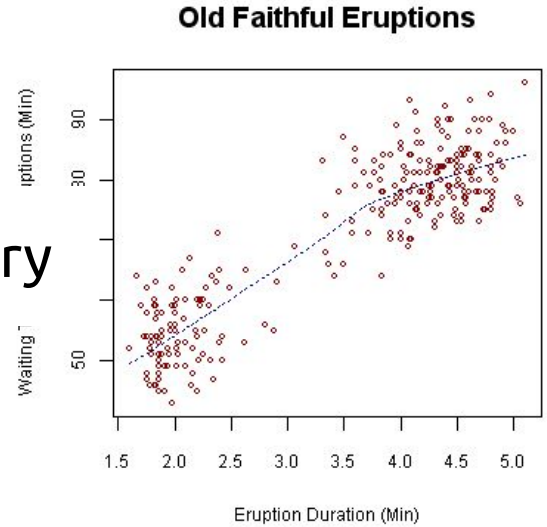
Why do we visualise data?

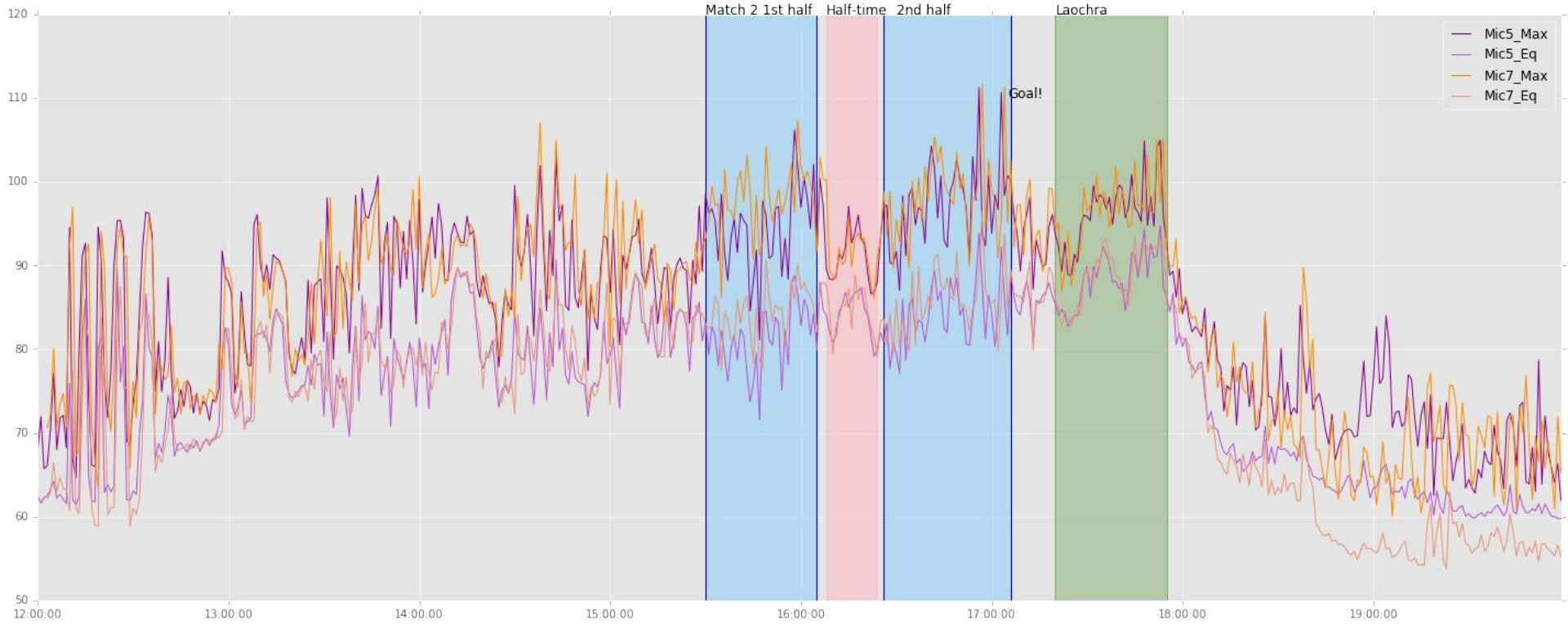




Exploratory vs Explanatory

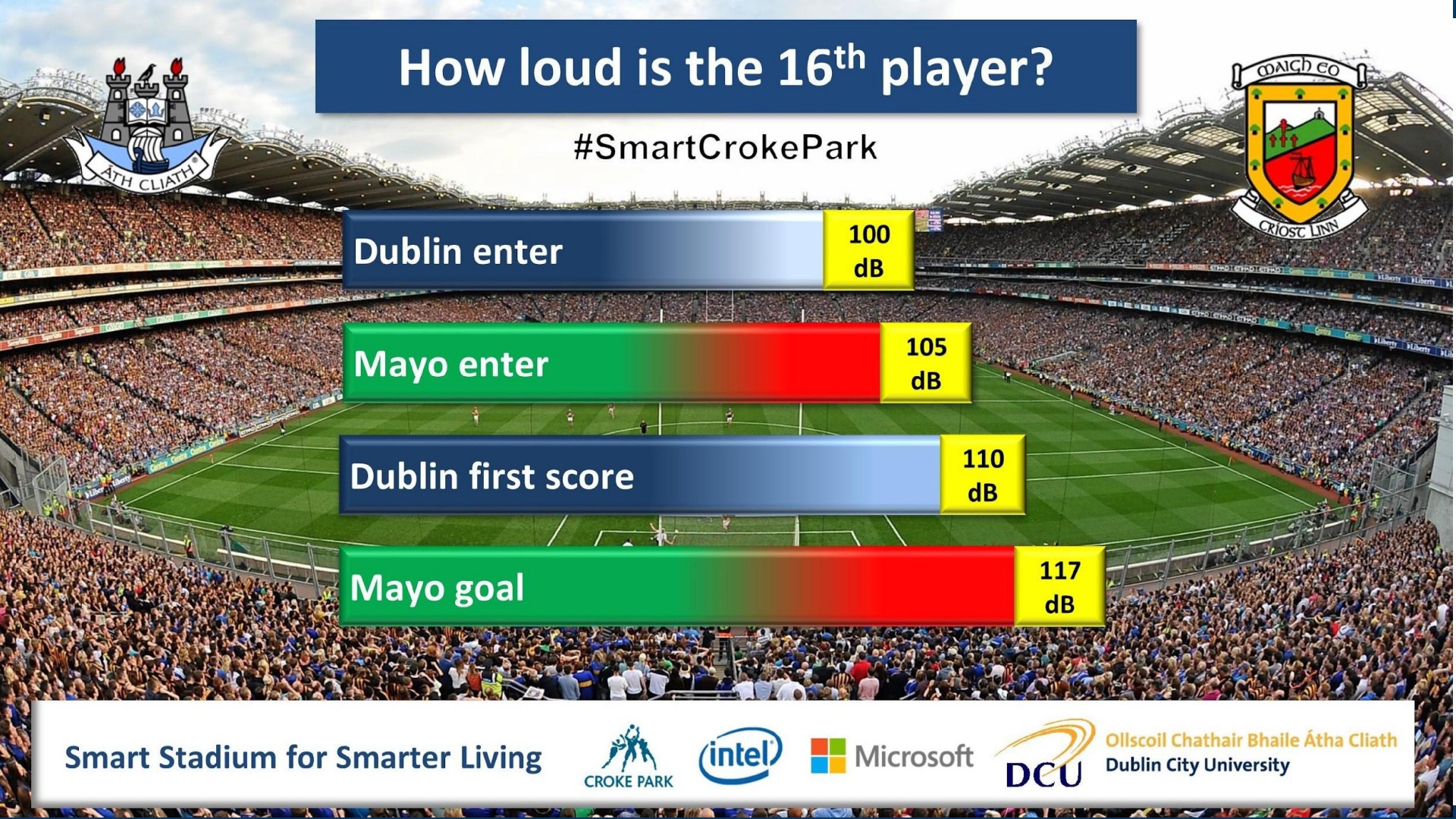





Match 2 1st half
Half-time
2nd half
Laochra
Goal!
Mic5_Max
Mic5_Eq
Mic7_Max
Mic7_Eq
120
110
100
90
80
70
60
50
12:00:00
13:00:00
14:00:00
15:00:00
16:00:00
17:00:00
18:00:00
19:00:00

How loud is the 16th player?

#SmartCrokePark



Smart Stadium for Smarter Living

CROKE PARK

intel

Microsoft

DCU Ollscoil Chathair Bhaile Átha Cliath Dublin City University

Gitlab & Your skills portfolio

Assignment 1 & 7 - Using Git & Skill Portfolio

Assignment 1 (Sep 24th @ 6pm) is to demonstrate your ability to use git (gitlab.computing.dcu.ie)

Assignment 7 (Dec 20th) is a collection of skill portfolios that you will save in git over the course of the semester -- 6 of the following topics will be marked.

- Git
- Spreadsheets
- Pandas
- Python Visualisation (matplotlib, seaborn or bokeh)
- Scripting
- Jupyter Notebooks (not python specific)
- OpenRefine
- Cleaning with Python
- Spreadsheets (Advanced)
- HTML

How skilled are you?

On a scale of 1 to 10, how would you rate your python programming ability?

What is 1? What is 10?

Knowledge? Potential? Efficiency?

skill

/skɪl/

noun

the ability to do something well; expertise.
"difficult work, taking great skill"

Similar: expertise | skilfulness | expertness | adeptness | adroitness

verb

train (a worker) to do a particular task.
"there is a lack of basic skilling"

Your skill record

https://gitlab.computing.dcu.ie/slittle/csc1158-skills

Summary: [One sentence summary of what the tool is useful for or what it does]

Data formats in: [where relevant!]
Data formats out: [where relevant!]

Three tips: [one of which must be advanced -- ie, likely to be found in a later chapter or section on datacamp]
1.
2.
3.

Examples of use: [minimal program sample, link to notebook or document or screenshot]

Contribution to data analytics pipeline: [refer to the talk in week 1 for the phases]

Comment on your skill level: [self-rating on this skill and what you'd like to learn to do better or what you'll do next]

References: [links to any sources used or just to helpful reference materials]

Assignment 1: using git

Git project - clone, edit, commit a skill report on Git

Worth 9%

You need to:

1. clone csc1158-skills repository
2. edit to add your details
3. create git.md from template.md (ie, copy contents to a new file)
4. fill in your first skill report on git
5. commit (with a comment) before **6pm Sep 24th**

Git

Next week

Keep practising Git, datacamp course on git is very helpful → Assignment 1

Next week: Data types, Introduction to tabular data and spreadsheets

Prior to next week read Chapter 2 “Data Types” of “Principles of Data Science”, Sinan Ozdemir (2024), https://dcu.primo.exlibrisgroup.com/permalink/353DCU_INST/jrp0g3/alma991005580550807206