

THE DATA SCIENCE LIFECYCLE

Lecturer: Mark Gillan

THE PPDAC



The PPDAC cycle is a framework to follow when asking and answering real-world problems using data.

1. **Problem:** Identify the question being solved
2. **Plan:** Decide how to answer the question
3. **Data:** Collect, store and secure data
4. **Analysis:** Look for patterns and insights
5. **Conclusion:** Interpret, communicate and implement findings



PROBLEM



Identify the right question to solve

- How much, or how many?
- Which category or group does this belong?
- Is this weird?
- Which is the best option to choose?



PLAN

- What data is needed to solve this question?
- Where will this data come from?
- Access to data?
- Will there be sufficient volume of data to provide a robust answer?
- Where and how will the data be stored?
- Are there any ethical implications?



DATA



- Collect and Store the Data Securely
- Data Quality Checks
- Data Understanding
- Data Dictionary Creation



ANALYSIS

- Preparation
- Manipulation
- Visualisation
- Data Modelling
- Validation of Findings



CONCLUSION



- Summarise and Communicate
- How does the data answer the original problem statement?
- Are there any aspects of the original question that has not been addressed?
- What are the conclusions?
- What could be done differently next time?
- What should happen next?



IMPORTANT

- The planning phase of a data science project is critically important.
- Often, similar analyses may have been carried out in the past.
- Not only can time be saved by not repeating existing work, but previous analyses can also be built upon and expanded.



REMINDER: THE PPDAC



The PPDAC cycle is a framework to follow when asking and answering real-world problems using data.

1. **Problem:** Identify the question being solved
2. **Plan:** Decide how to answer the question
3. **Data:** Collect, store and secure data
4. **Analysis:** Look for patterns and insights
5. **Conclusion:** Interpret, communicate and implement findings

