# An Easy-to-Use Data Science Toolkit

Created by Sam Peel; Supervised by William J. Teahan

PRIFYSGOL BANGOR UNIVERSITY

## Background & Motivation

Some consider data science to be the fourth paradigm of scientific research[1]. Therefore, scientific disciplines would benefit from a tool that allows them to carry out empirical data-driven research in a structured and easily comprehensible graphical environment.
Although similar tools already exist, it was immediately clear that they fail to compromise between ease of use (which benefits users with limited/no programming skills) and flexibility (which benefits competent programmers).

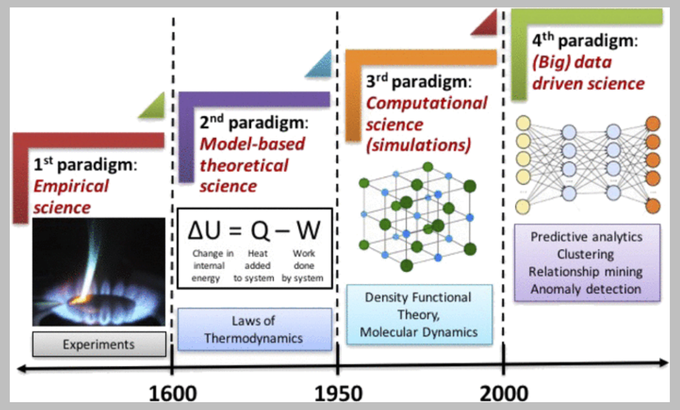


*Figure 1: Scientific Paradigms*

## Proposed Solution

Between these limiting factors there is room for a tool that bridges this gap by walking the user through every step of the data exploration process, while explaining all findings and providing options for different visualization methods and machine learning models. At the end of the project all generated code, data, visualizations and models will be organized into appropriate files, with which non-programmers can proceed to implement their models, whereas pragmatic programmers will be able to modify the generated code.

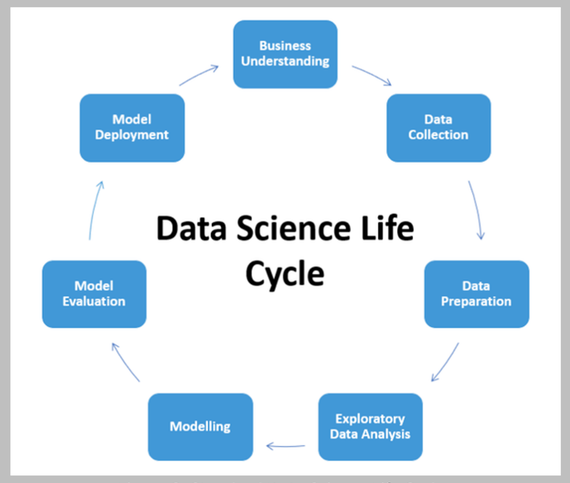


*Figure 2: Standard Data Science Life Cycle*

## Aims & Objectives

- Create a tool (using Python libraries) that allows users with varying levels of technical knowledge to carry out data-driven research.
- Help users to gain insight from their data by walking them through the process of exploring, visualizing and modelling process.
- Organize the generated visualisations and models so they can be utilised or modified.

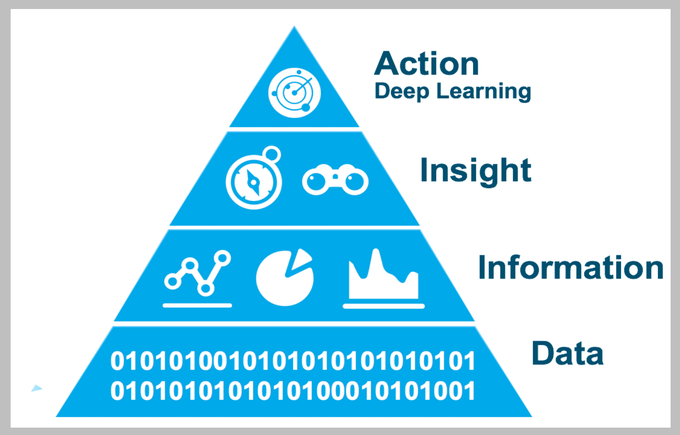


*Figure 3: Data Transformation*

## Logical Design

The program itself will be very basic, in order to aid ease of use. The main window will consist of a work area (for the user to view their data), a handful of buttons (for the user to select their desired machine learning model), and most importantly; along the right-hand side of the window there will be a glossary, which will describe patterns within the data to the user and advise the user on the benefits and hinderances of each appropriate machine learning model.

## Libraries Used

This tool has been made possible due to the vast array of comprehensive Python libraries, aimed at carrying out every step of the data science lifecycle. The user interface will be constructed using PySide6, an open-source widget toolkit. In order to import and format the data, NumPy and Pandas will be used. Moving on, the formatted data will be visualized appropriately using Matplotlib, whereby the user will be able to decide which machine learning model they would like to use, which will then be implemented using Scikit-Learn.

## Future Work

Refinements can be made to this tool by finding new intuitive ways for users to interact with their data. This would lead to a more fluid and modular environment which will promote exploration and experimentation. Additionally, by exploring new visualization techniques the ease at which data can be interpreted and insight gained may be increased. Finally, a tier system for the glossary could be added, which would vary the level of abstraction in the descriptions in order to benefit users of all knowledge levels.

*Text: 1. Hey, A.J. ed., 2009. The fourth paradigm: data-intensive scientific discovery (Vol. 1). Redmond, WA: Microsoft research.*
Figures: 1. Science Paradigms: https://aip.scitation.org/doi/full/10.1063/1.4946894 *2. Life Cycle: https://www.quora.com/What-is-the-life-cycle-of-a-data-science-project 3. Transformation:: https://aip.scitation.org/doi/full/10.1063/1.4946894*
*Logos: Python: https://en.wikipedia.org/wiki/File:Python-logo-notext.svg PySide: http://www.h-online.com/open/news/item/PySide-LGPL-Python-bindings-for-Qt-743017.htm NumPy: https://technopremium.com/blog/python-numpy/ Matplotlib: https://en.wikipedia.org/wiki/Matplotlib Scikit-Learn: https://technopremium.com/blog/python-numpy/*
*Pandas: https://en.wikipedia.org/wiki/Pandas_%28software%29*