Week 1: Introduction to Data Science & Databases

Most of you probably already know python. Following the "lab1_PyIntro.ipynb" python notebook that you can download from SurreyLearn in the lab 1 material section will either introduce you to python or refresh your knowledge. To run this notebook, you will need to download also the "lab1_mpg.csv" in the same folder. Load the python notebook and follow the instructions to understand every section in the code.

Other interesting tools in data science include the following:

- Wolfram Alpha computational knowledge engine, which processes natural language-like queries through a mix of algorithms and pre-digested data sources. Check it out at http://www.wolframalpha.com/.

- On Big Data Analytics, Hadoop and Spark are the famous parallel processing systems using Java and C++.

- Matlab is for processing matrices, with many linear algebra, statistical and machine learning toolboxes. It is propriety but GNU Octave is an open-source alternative.

- R is open source packages for statisticians. Many of its functions are available in Python, and you can call R library functions from Python if not available.

- Excel and other spreadsheet programs include a variety of functions for exploratory data analysis and data manipulations that is worth searching deep for before writing your own code.

- Google Books NGram aims at compiling all books and show the statistics of using phrases as it evolves over time, with links to books mentioning these phrases. .this enables any data scientist to check the context of using phrases and what they mean and how it evolves. Checkout https://books.google.com/ngrams

Finding and working with data will require connecting to a database. SQLite is native to python. You can introduce yourself to it by following the "lab1_db.ipynb" python notebook that you can download from the same section in SurreyLearn.

Data comes in other Pvarious formats to be both computer parse-able and human readable.
The most interesting file formats are:

- CSV (comma separated value) files: covered in the "lab1_PyIntro.ipynb"
- XML (eXtensible Markup Language): covered in the "lab1_db.ipynb"
- JSON (JavaScript Object Notation): covered in the "lab1_db.ipynb"

More helpful pandas syntax can be found in their Intro to Data Structures documentation.